UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes

Institut Denis Poisson

THÈSE présentée par :

Julien WEIBEL

soutenue le : **8 novembre 2024**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Mathématiques**

Graphons de probabilités, limites de graphes pondérés aléatoires et chaînes de Markov branchantes cachées

THÈSE dirigée par :

M. Romain Abraham	Professeur, Université d'Orléans
M. Jean-François Delmas	Ingénieur général des ponts, des eaux et des fôrets, École Nationale des Ponts et Chaussées

RAPPORTEURS :

M. Bartłomiej Błaszczyszyn	Directeur de recherche, INRIA, École Normale Supérieure
Mme. Catherine Matias	Directrice de recherche, CNRS, Sorbonne Université

JURY :

M. Romain Abraham	Professeur, Université d'Orléans
M. Bartłomiej Błaszczyszyn	Directeur de recherche, INRIA, École Normale Supérieure
M. Jean-François Delmas	Ingénieur général des ponts, des eaux et des fôrets, École Nationale des Ponts et Chaussées
Mme. Catherine Matias	Directrice de recherche, CNRS, Sorbonne Université, Présidente du jury
M. Pierre-André Zitt	Maître de conférences, Université Gustave Eiffel

Graphons de probabilités, limites de graphes pondérés aléatoires et chaînes de Markov branchantes cachées

Julien Weibel

Institut Denis Poisson, Université d'Orléans, France

Thèse sous la direction de Romain Abraham et Jean-François Delmas

julien.weibel@normalesup.org

8 novembre 2024

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes celles et ceux qui m'ont accompagné et soutenu tout au long de ce parcours de thèse.

Je tiens tout d'abord à exprimer ma profonde gratitude envers mes directeurs de thèse, Romain Abraham et Jean-François Delmas. Merci de m'avoir fait découvrir des sujets de recherche aussi riches et passionnants. J'apprécie sincèrement le temps que vous avez consacré à nos discussions, tant sur le plan mathématique que personnel. Grâce à vous, j'ai pu développer et affiner mon intuition, cultiver mon goût pour l'élégance mathématique, et renforcer ma rigueur scientifique. Je retiendrai votre approche complémentaire redoutablement efficace et les conseils de rédaction de Jean-François pour des textes « plus secs ». Votre accompagnement a été déterminant dans mon parcours, et je vous en suis très reconnaissant.

Je tiens également à remercier Bartłomiej (Bartek) Błaszczyszyn et Catherine Matias, qui m'ont fait l'honneur de rapporter cette thèse, ainsi que Pierre-André Zitt, pour sa participation au jury.

Ces trois années de thèse ont été particulièrement agréables grâce à l'accueil chaleureux de l'Institut Denis Poisson (IDP) et du CERMICS, qui m'ont offert un cadre de travail stimulant et bienveillant. Je remercie notamment Marie-France, Anne, et Marie de l'IDP, ainsi qu'Isabelle et Stéphanie du CERMICS, pour leur efficacité et leur bienveillance.

Je tiens également à remercier tous les doctorants et post-docs de l'IDP et du CERMICS. Ces années ont été d'autant plus enrichissantes et agréables grâce aux moments passés ensemble. Du côté de l'IDP, merci donc aux anciens : Alexis, Émile, Grégoire, Léo, Maxime, Mohamed, Ouassim, Rana, Rita, Sonia et Titouan. Merci aussi aux nouveaux : Bruno, Cisse, Emma et Paguiel. Du côté du CERMICS, un merci tout particulier aux membres du deuxième avec qui j'ai partagé de nombreuses discussions enrichissantes : Emanuele, Kacem, Léo, Nerea, Oscar, Roberta, Seta et Zoé. Merci aussi à tous les autres membres du bureau du deuxième pour la bonne ambiance au quotidien : Camila, Clément, Coco, Edoardo, Fabian, Faten, Hélène, Héloïse, Hervé, Jonathan, Laetitia, Louis, Luca, Mathis, Michel, Paul, Raian, Thibault, Vitor et Yue. Je n'oublie pas non plus tous ceux du troisième avec qui nous avons partagé de nombreuses et agréables pauses : Alberic, Alfred, Alicia, Amandine, Antonin, Charlotte, Clément, Éloïse, Epiphane, Étienne, Fabian, Giulia, Guido, Hadrien, Jean, Guillaume, Laurent, Louis, Louis-Pierre, Mathias, Noé, Pierre, Raphaël, Régis, Renato, Rutger, Shiva, Simon et Solal. Et merci aussi aux anciens : Benoît, Cyrille, Dylan, Inass, Gaspard, Guillaume, Léo, Maël, Rémi, Sébastien et Thomas. Merci à tous !

Enfin, je remercie chaleureusement mes amis et les membres de ma famille pour leur soutien. Merci en particulier à mes amis de l'ÉNS et aux amis ayant rejoint ce groupe : Simon, Octavie, Éléonore, Rémy LG., Côme, Vincent, Louis, Rémy C., Stéphane, Léo et Amaury. Un immense merci aussi à mes parents et à ma sœur Élodie, pour leur soutien constant et inestimable.

Résumé

Les graphes sont des objets mathématiques qui servent à modéliser tout type de réseaux, comme les réseaux électriques, les réseaux de communications et les réseaux sociaux. Formellement un graphe est composé d'un ensemble de sommets et d'un ensemble d'arêtes reliant des paires de sommets. Les sommets représentent par exemple des individus, tandis que les arêtes représentent les interactions entre ces individus. Dans le cas d'un graphe pondéré, chaque arête possède un poids ou une décoration pouvant modéliser une distance, une intensité d'interaction, une résistance. La modélisation de réseaux réels fait souvent intervenir de grands graphes qui ont un grand nombre de sommets et d'arêtes.

La première partie de cette thèse est consacrée à l'introduction et à l'étude des propriétés des objets limites des grands graphes pondérés : les graphons de probabilités. Ces objets sont une généralisation des graphons introduits et étudiés par Lovász et ses co-auteurs dans le cas des graphes sans poids sur les arêtes. À partir d'une distance induisant la topologie faible sur les mesures, nous définissons une distance de coupe sur les graphons de probabilités. Nous exhibons un critère de tension pour les graphons de probabilités lié à la compacité relative dans la distance de coupe. Enfin, nous prouvons que cette topologie coïncide avec la topologie induite par la convergence en distribution des sous-graphes échantillonnés.

Dans la deuxième partie de cette thèse, nous nous intéressons aux modèles markoviens cachés indexés par des arbres. Nous montrons la consistance forte et la normalité asymptotique de l'estimateur de maximum de vraisemblance pour ces modèles sous des hypothèses standards. Nous montrons un théorème ergodique pour des chaînes de Markov branchantes indexés par des arbres avec des formes générales. Enfin, nous montrons que pour une chaîne stationnaire et réversible, le graphe ligne est la forme d'arbre induisant une variance minimale pour l'estimateur de moyenne empirique parmi les arbres avec un nombre donné de sommets.

Mots clés : graphes pondérés aléatoires, réseaux stochastiques, graphons de probabilités, modèles markoviens cachés indexés par des arbres, chaînes de Markov branchantes, processus de branchement

Abstract

Graphs are mathematical objects used to model all kinds of networks, such as electrical networks, communication networks, and social networks. Formally, a graph consists of a set of vertices and a set of edges connecting pairs of vertices. The vertices represent, for example, individuals, while the edges represent the interactions between these individuals. In the case of a weighted graph, each edge has a weight or a decoration that can model a distance, an interaction intensity, or a resistance. Modeling real-world networks often involves large graphs with a large number of vertices and edges.

The first part of this thesis is dedicated to introducing and studying the properties of the limit objects of large weighted graphs : probability-graphons. These objects are a generalization of graphons introduced and studied by Lovász and his co-authors in the case of unweighted graphs. Starting from a distance that induces the weak topology on measures, we define a cut distance on probability-graphons. We exhibit a tightness criterion for probability-graphons related to relative compactness in the cut distance. Finally, we prove that this topology coincides with the topology induced by the convergence in distribution of the sampled subgraphs.

In the second part of this thesis, we focus on hidden Markov models indexed by trees. We show the strong consistency and asymptotic normality of the maximum likelihood estimator for these models under standard assumptions. We prove an ergodic theorem for branching Markov chains indexed by trees with general shapes. Finally, we show that for a stationary and reversible chain, the line graph is the tree shape that induces the minimal variance for the empirical mean estimator among trees with a given number of vertices.

Keywords : random weighted graphs, stochastic networks, probability-graphons, hidden Markov models indexed by trees, branching Markov chains, branching processes

Sommaire

Liste des figures

Première partie

Introduction

Chapitre 1

Grands graphes denses et leurs objets limites, les graphons

1.1 Une brève introduction aux graphes finis

On fait usuellement remonté le début de la théorie des graphes à l'étude par Leonhard Euler [Eul36] en 1735-1741 du problème des septs ponts de Königsberg, qui consiste à trouver un chemin en boucle passant exactement une fois par chacun des septs ponts.

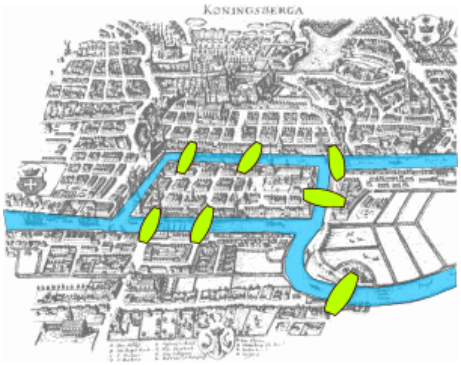


FIGURE 1.1 – Giuşcă Bogdan. (2005) Représentation graphique du problème des sept ponts de Königsberg. Wikimedia commons. `https://upload.wikimedia.org/wikipedia/commons/5/5d/Konigsberg_bridges.png`

Les graphes sont des objets mathématiques qui servent à modéliser tout type de réseaux. Les réseaux apparaissent naturellement dans un grand nombre de contexte divers, nous en donnons quelques exemples : les réseaux internets et de communications [FFF99, BAJ00, LW06], les réseaux de transport [RB05, BCM09], les réseaux électriques [AAN04, SAL18], ou bien encore les réseaux sociaux [AB02, New03]. Mentionnons également les processus épidémiologiques [DM10, KMS17] où le graphe modélise le réseau d'interactions entre les individus susceptibles de s'infecter les uns et les autres. Et enfin mentionnons les réseaux biologiques [BS02, Lew09] qui peuvent prendre différentes forme comme par exemple un graphe résumant les interactions protéines-protéines.

Formellement, un graphe $G = (V, E)$ est composé d'un ensemble de *sommets* $V(G) = V$ et d'un ensemble d'*arêtes* $E(G) = E$ qui est un sous-ensemble de $V \times V$. Selon le contexte, les sommets représentent les individus, les croisements ou les nœuds qui composent le réseau, tandis que les arêtes représentent les interactions entres des individus, les routes entre des croisements, ou les capacités, résistances ou débits entre des nœuds. Lorsque l'ensemble E est symétrique, on dit que le graphe G est *symétrique* ou *non-orienté* et que les arêtes sont *non-orientées*. Sinon, on dit que le graphe G et les arêtes sont *orientés*, et les arêtes ne peuvent alors être parcourues que dans un seul sens. Voir la Figure 1.2 pour un exemple d'un graphe non-orienté et d'un graphe orienté. Dans notre cas, nous n'autoriserons pas les arêtes allant d'un sommet vers lui-même, et que l'on appelle des *boucles* (ou *"self-loops"* en anglais). Usuellement, un

graphe sans boucles est dit *simple*. Ici, tous les graphes seront simples et donc nous ne le préciserons pas. Nous dirons qu'un graphe est *étiqueté* (ou *numéroté* ou *ordonné*) si ces sommets sont ordonnés, et qu'un graphe est *non-étiqueté* si ce n'est pas le cas.

FIGURE 1.2 – Un exemple d'un graphe étiqueté non-orienté (à gauche) et d'un graphe étiqueté orienté (à droite).

Notons également que l'on peut associer à chaque arête un poids ou une étiquette pour venir rajouter de l'information quantitative ou qualitative (e.g. distance, résistance, capacité, fréquence d'interaction, contenu de messages, etc.), et l'on dit alors que le graphe est *pondéré* ou *étiqueté*. Nous y reviendrons dans une section ultérieure. Mentionnons aussi les multi-graphes qui sont des graphes où l'on autorise deux sommets à être reliés par plusieurs arêtes simultanément, et qui peuvent être vu comme des graphes pondérés par des poids entiers.

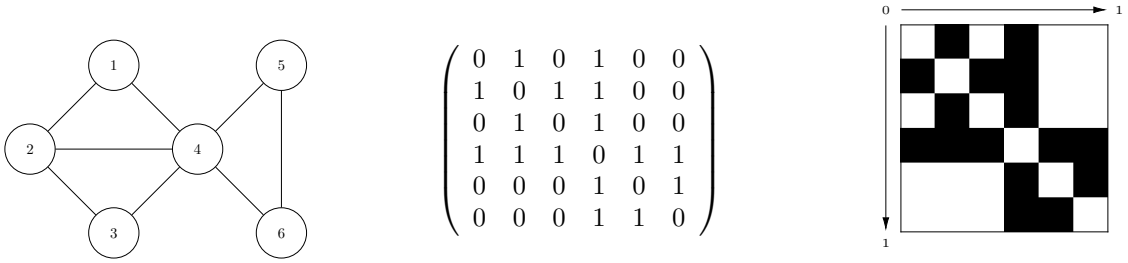
Pour une arête (u, v) d'un graphe, on appelle *extrémités* de l'arête les sommets u et v. Le *degré* d'un sommet du graphe est le nombre d'arête du graphe dont ce sommet est une extrémité. Lorsque le graphe est orienté, pour chaque arête (u, v) (qui est donc orientée, et que l'on appelle parfois aussi *arc*) on distingue sa *source* u de sa *destination* v. De plus, pour un sommet d'un graphe orienté, on définit pour ce sommet son *degré sortant* (resp. *degré entrant*) comme étant le nombre d'arêtes dont il est la source (resp. la destination).

Lorsque le graphe G comporte un nombre fini de sommets $v(G)$, on dit qu'il est *fini*. Pour un graphe fini G avec $v(G) = n$ sommets, quitte à numéroter arbitrairement les sommets, on peut alors supposer que V est égal à $[n] := \{1, \cdots, n\}$, et on définit sa *matrice d'adjacence* A qui est de taille $n \times n$ par :

$$A_{i,j} = \begin{cases} 1 & \text{si } (i,j) \in E \\ 0 & \text{sinon.} \end{cases}$$

Un graphe fini est caractérisé par son nombre de sommets et par sa matrice d'adjacence. Pour comparer les matrices d'adjacence de graphes qui ont des nombres de sommets différents, on peut les représenter dans l'espace plus grand des fonctions de $[0,1]^2$ à valeurs dans $\{0,1\}$. Pour cela, on associe à chaque matrice d'adjacence A d'un graphe fini à n sommets son image pixelisée en découpant le carré unité $[0,1]^2$ en n^2 petits carrés de côté $1/n$, et en inscrivant sur ces petits carrés la valeur des coefficients de la matrice A (sur les figures qui suivent, carré blanc pour la valeur 0, et carré noir pour la valeur 1).

Dans la Figure 1.3, on donne un exemple de graphe symétrique avec sa matrice d'adjacence et l'image pixelisée correspondante. Notons que pour l'image pixelisée, l'origine est placée dans le coin en haut à gauche afin d'être en accord avec la convention de numérotation des coefficients d'une matrice.



$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

FIGURE 1.3 – Un exemple de graphe, sa matrice d'adjacence et son image pixelisée.

Nous verrons dans la section suivante que ces représentations en images pixelisées permettent de donner une bonne intuition de la notion de convergence pour les suites de grands graphes aléatoires, ainsi qu'une bonne intuition de leurs objets limites (i.e. les graphons, qui seront définis formellement en Section 1.3).

Notons qu'un graphe fini avec $v(G) = n$ sommets a au maximum n^2 arêtes s'il est orienté, et a au maximum $\binom{n}{2}$ arêtes s'il est non-orienté. L'indice de densité ρ d'un graphe G est défini comme le ratio entre le nombre d'arêtes $e(G)$ du graphe G et le nombre d'arêtes maximales possibles ; pour un graphe non-orienté c'est donc :

$$\rho(G) = \frac{e(G)}{\binom{n}{2}} = \frac{2e(G)}{v(G)(v(G)-1)},$$

et de manière similaire pour un graphe orienté. On dit qu'une suite de graphes finis $(G_n)_{n\in\mathbb{N}}$ est une suite de graphes *denses* (resp. *creux*, ou *"sparse"* en anglais) si la suite de leurs indices de densités $(\rho(G_n))_{n\in\mathbb{N}}$ converge vers une constante strictement positive (resp. vers 0). Ce critère de suite de graphes denses permet d'éviter que les images pixelisées associées aux graphes ne deviennent quasiment intégralement blanches lorsque n tend vers l'infini (et donc que la limite ne soit triviale). Notons également que pour une suite de graphes denses $(G_n)_{n\in\mathbb{N}}$ avec $v(G_n) = n$, le degré moyen d'un sommet croit linéairement avec n.

1.2 Modèles de graphes aléatoires et heuristique de leurs limites

Nous allons maintenant introduire quelques modèles de graphes aléatoires. Nous allons également en profiter pour donner une heuristique de convergence pour estimer la limite de ces différents modèles de graphes aléatoires en utilisant les images pixelisées introduites dans la section précédente.

Le modèle d'Erdős-Rényi : Le modèle de graphe aléatoire le plus simple et l'un des plus étudiés est celui introduit par d'Erdős et Rényi [ER59, ER60, ER61a, ER61b] en 1959, mais en réalité déjà étudié avant eux par Gilbert [Gil59] en 1959. Ce modèle possède deux paramètres : le nombre de sommets n, et une probabilité $p \in [0, 1]$. On construit ensuite un graphe aléatoire (symétrique ou non) qui possède n sommets et dont les arêtes sont présentes indépendamment les unes des autres avec probabilité p.

Notons qu'Erdős et Rényi ont également considéré une variante de ce modèle où l'on fixe le nombre d'arêtes k, et où le graphe symétrique est uniformément distribué sur l'ensemble des graphes à n sommets et k arêtes.

Dans la Figure 1.4, on présente les images pixelisées des simulations du modèle d'Erdős-Rényi pour $n = 10$, 50 et 100 sommets et avec une probabilité de présence d'arêtes de $1/2$. On remarque que les images pixelisées de ces graphes aléatoires semblent converger « visuellement » vers le carré unité uniformément gris dont le niveau de gris est $1/2$ sur l'échelle allant de blanc pour 0 à noir pour 1. En particulier, notons que cette convergence « visuelle » est nécessairement plus faible que la convergence L^1.

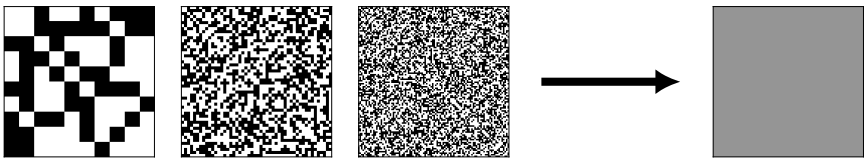


FIGURE 1.4 – Heuristique de convergence pour les images pixelisées du graphe aléatoire d'Erdős-Rényi (à gauche de la flèche) avec paramètre $p = 1/2$ et $n = 10$, 50 et 100 sommets (de gauche à droite). La limite (à droite de la flèche) est un carré uniformément gris dont le niveau de gris est $1/2$ sur l'échelle allant de blanc pour 0 à noir pour 1.

Le modèle à blocs stochastiques : Il existe également une généralisation inhomogène du modèle d'Erdő-Rényi que l'on appelle le *modèle à blocs stochastiques* (en anglais, *"stochastic block model"* ou SBM). Ce modèle est apparu indépendamment dans différentes communautés scientifiques sous différents noms : SBM en statistique et en machine learning [HLL83], *"planted partition model"* en informatique théorique [Bop87, BCLS87, DF89], et *"inhomogeneous random graph"* dans la littérature mathématique [BJR07]. Ces dernières années, la terminologie qui semble dominer est SBM.

Le modèle à blocs stochastiques possède plusieurs paramètres : un nombre de sommets n, un nombre de types ou de communautés k pour les sommets, une mesure de probabilité ν sur l'ensemble des types $[k] = \{1, \cdots, k\}$, et une matrice $(p_{i,j})_{1\leq i,j\leq n}$ de taille $n \times n$ dont les coefficients sont des probabilités dans $[0, 1]$. Le modèle à blocs stochastiques permet de générer un graphe aléatoire (symétrique ou non)

de la manière suivante : chacun des n sommets i reçoit indépendamment des autres un type aléatoire X_i dans $[k]$ distribué selon ν ; puis, conditionnellement au vecteur de types $(X_i)_{1\leq i\leq n}$, chaque arête (i,j) est ajoutée ou non indépendamment des autres avec probabilité p_{X_i,X_j}. Notons qu'en général les variables de types $(X_i)_{1\leq i\leq n}$ servant à la construction du graphe ne sont pas observées (on parle alors de variables latentes ou cachées).

Une application importante du modèle à blocs stochastiques est la détection de communauté [Abb18] où l'on cherche à inférer les différentes communautés des sommets à partir de l'observation seule des arêtes du graphes. Nous en reparlerons plus tard.

Dans la Figure 1.5, on présente les images pixelisées de simulations du modèle à blocs stochastiques pour $n = 10$, 50 et 100 sommets, avec deux communautés (k=2) de même taille (i.e. les types sont équiprobables) et dont la matrice des probabilités de présence d'arêtes $(p_{i,j})_{1\leq i,j\leq 2}$ a pour coefficients $p_{1,1} = p_{2,2} = 3/4$ (intra-communautaire) et $p_{1,2} = p_{2,1} = 1/4$ (extra-communautaire). On remarque que les images pixelisées de ces graphes aléatoires semblent converger « visuellement » vers l'image pixelisée associée à la matrice $(p_{i,j})_{1\leq i,j\leq 2}$. Notons que la définition d'image pixelisée associée à une matrice d'adjacence d'un graphe (dont les coefficients sont 0 ou 1) peut facilement être étendue à une matrice dont les coefficients sont à valeurs dans $[0,1]$.

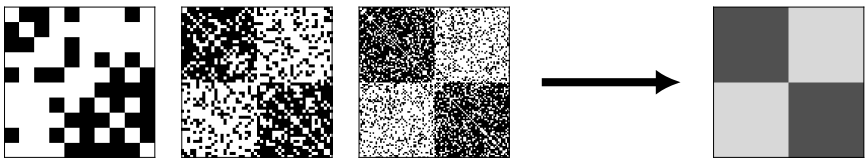


FIGURE 1.5 – Heuristique de convergence pour les images pixelisées pour $n = 10$, 50 et 100 sommets (de gauche à droite) du modèle à blocs stochastiques avec deux communautés de même taille, et dont les arêtes sont présentes avec probabilité 3/4 à l'intérieur des communautés et 1/4 entre les communautés. La limite est l'image pixelisé de la matrice paramètre du modèle $(p_{i,j})_{1\leq i,j\leq 2}$.

Modèle à attachement uniforme croissant : Le modèle à attachement uniforme croissant permet de construire de manière itérative une suite de graphes aléatoires symétriques $(G_n)_{n\geq 1}$. On commence par le graphe G_1 contenant un unique sommet. Puis, à chaque étape, on définit un nouveau graphe G_n à partir du précédent en ajoutant un nouveau sommet et en reliant chaque paire de sommets distincts qui n'était pas déjà connectés à l'étape précédente par une arête indépendamment des autres paires avec probabilité $1/n$.

Dans la Figure 1.6, on présente les images pixelisées de simulations de graphes aléatoires à attachement uniforme croissant pour $n = 10$, 50 et 100 sommets. On remarque une fois de plus que les images pixelisées de ces graphes semblent converger « visuellement » vers vers le graphe en nuance de gris de la fonction W de $[0,1]^2$ dans $[0,1]$ définie par $W(x,y) = 1 - \max(x,y)$ pour tout $x, y \in [0,1]$.

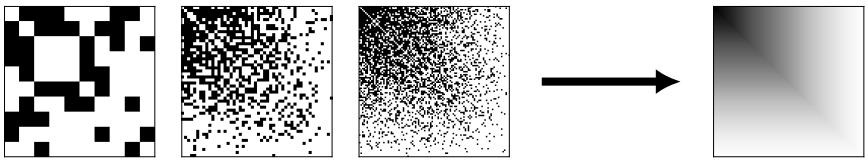


FIGURE 1.6 – Heuristique de convergence pour les images pixelisées pour $n = 10$, 50 et 100 sommets (de gauche à droite) du modèle à attachement uniforme croissant. La limite est le graphe de la fonction $W(x,y) = 1 - \max(x,y)$ en niveau de gris sur l'échelle allant de blanc pour la valeur 0 à noir pour la valeur 1.

La re-numérotation des sommets ne change pas la limite : Les exemples précédents pourraient laisser penser que nous disposons toujours d'un ordre sur l'ensemble des sommets d'un graphe. Hors, pour les réseaux issus du monde réel, il n'existe pas en pratique d'ordre canonique sur l'ensemble de leurs sommets. La convergence de graphes que nous avons présentée plus haut sous forme d'heuristique doit donc être insensible à la re-numérotation ou permutation des sommets. Nous allons également voir

que si la convergence dépendait de la numérotation des sommets, alors l'utilisation d'une numérotation des sommets qui rassemble les sommets semblables est nécessaire pour pouvoir obtenir une limite non triviale.

Pour illustrer notre choix d'une convergence invariante par re-numérotation des sommets, considérons deux exemples. Dans le premier exemple (Figure 1.7), on considère des graphes bipartites complets, c'est à dire dont les sommets sont répartis en deux groupes de même taille, et toutes les arêtes reliant un sommet d'un groupe à un sommet de l'autre groupe sont présentes (mais aucune arête reliant deux sommets d'un même groupe n'est présente). Dans la Figure 1.7, on présente les images pixelisées des graphes biparties complets avec $n = 10$, 50 et 100 sommets en considérant deux possibilités de tri des sommets : par groupes dans la première ligne, et en alternant les groupes dans la deuxième ligne. Dans le premier cas, la limite de la convergence « visuelle » est évidente (avec ou sans re-numérotation) : c'est l'image pixelisée associée à la matrice 2×2 avec des 0 sur la diagonale et des 1 sur l'anti-diagonale. Dans le deuxième cas, la limite pour une convergence « visuelle » sans invariance par re-numérotation est une image uniformément grise de niveau de gris 1/2, et ne conserve donc que la densité moyenne d'arêtes. Tandis que pour une convergence avec invariance par re-numérotation, la limite dans le deuxième cas est la même que dans le premier cas.

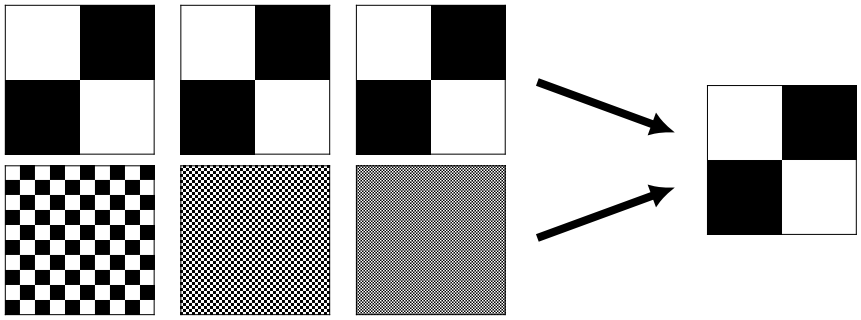


FIGURE 1.7 – Heuristique de convergence pour les images pixelisées pour $n = 10$, 50 et 100 sommets (de gauche à droite) pour le graphe bipartite complet. Les sommets sont triés par groupes dans la première ligne, et sont en alternance de groupes dans la deuxième ligne. La re-numérotation des sommets ne change pas la limite.

Considérons maintenant un deuxième exemple. Dans la Figure 1.8, on présente les images pixelisées de simulations du modèle à blocs stochastiques sous les mêmes paramètres que dans la Figure 1.5 plus haut (et toujours avec $n = 10$, 50 et 100 sommets). Les sommets sont triés par communauté dans la première ligne et sont triés selon une permutation aléatoire dans la seconde ligne. Comme nous l'avons déjà remarqué avec la Figure 1.5, la première ligne de la Figure 1.8 converge « visuellement » vers l'image pixelisée de la matrice $(p_{i,j})_{1 \leq i,j \leq 2}$ des probabilités d'arêtes (paramètre du modèle). Pour une convergence sans invariance par re-numérotation, la deuxième ligne présente une situation similaire au modèle d'Erdős-Rényi et on obtient une convergence « graphique » vers une image uniformément grise dont le niveau de gris est la densité moyenne d'arêtes (ici 1/2). Tandis que pour une convergence avec invariance par re-numérotation, on obtient la même limite pour les deux lignes.

Ainsi, remarquons que cette invariance de la limite par re-numérotation des sommets est là pour permettre de détecter les structures sous-jacentes et les caractéristiques propres des sommets (par exemple, « âge », type ou numéro de communauté) influençant la formation d'arêtes. Mais cette propriété d'invariance par re-numérotation sert aussi à éviter certains cas triviaux où la seule information donnée par la limite serait la densité d'arête moyenne dans le graphe.

Quelques autres modèles de graphes aléatoires : L'étude des réseaux issus du monde réel (comme le réseau internet ou les réseaux sociaux) a permis d'exhiber de nouvelles propriétés différentes de celles du graphe d'Erdős-Réniy. Nous présentons quelques-unes de ces propriétés parmi les plus étudiées.

Une première propriété partagées par de nombreux réseaux réels est d'être des réseaux dits « petits mondes » (*"small worlds"* en anglais), c'est à dire que les distances typiques entre deux sommets sont petites. Les réseaux petits mondes ont été introduits par Watts et Strogatz [WS98] en 1998 (voir aussi [Dur06, Chapitre 5] ou [van24, Partie III]). Une croyance populaire qui illustre bien le concept de réseaux

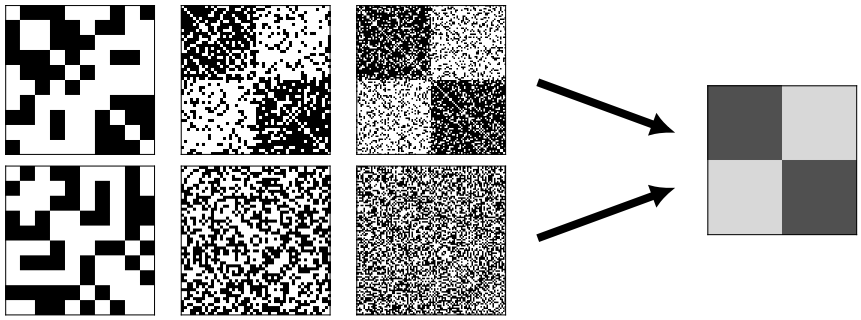


FIGURE 1.8 – Heuristique de convergence pour les images pixelisées pour $n = 10$, 50 et 100 sommets (de gauche à droite) du modèle à blocs stochastiques avec deux communautés de même taille, et dont les arêtes sont présentes avec probabilité 3/4 à l'intérieur des communautés et 1/4 entre les communautés. Les sommets sont triés par communautés d'appartenance dans la première ligne, alors qu'ils sont dans un ordre aléatoire dans la seconde ligne. La limite est l'image pixelisé de la matrice paramètre du modèle $(p_{i,j})_{1\leq i,j\leq 2}$ avec ou sans re-numérotation des sommets.

petits mondes est que chaque individu serait connecté à n'importe quel autre individu par une chaîne de connaissances mutuelles de longueur au plus 6 (voir [Dur06, Section 1.3] pour une discussion sur cette croyance et d'autres similaires).

Une autre propriété des réseaux réels est qu'ils sont souvent des *réseaux invariants d'échelle* (*"scale-free networks"* en anglais), c'est à dire que la distribution $(P(k))_{k\in\mathbb{N}}$ du degré des sommets suit une loi puissance (ou *"power law"* en anglais) $P(k) \sim k^{-\gamma}$ avec γ un paramètre est typiquement dans $(2, 3)$. Le terme de réseaux invariants d'échelle est du aux propriétés d'auto-similarité de ces réseaux. Cette propriété d'invariance d'échelle de certains réseaux réels a été observée pour la première fois par Barabási et Albert [BA99] en 1999 (voir aussi [Dur06, Section 1.4] ou [van16, Section 1.3]).

Les graphes représentant des réseaux réels ont également un caractère *inhomogène*, c'est à dire que la densité d'arêtes (ou probabilité d'avoir une arête) n'est pas uniforme dans le graphe et peut dépendre de caractéristiques propres aux sommets extrémaux de l'arête. Nous donnons quelques exemples de telles caractéristiques inspirées des réseaux réels pouvant influer sur la probabilité de formation d'arêtes. Nous avons déjà rencontré un exemple de graphes inhomogènes avec le modèle à blocs stochastiques où les caractéristiques des sommets étaient leur appartenance à l'une communauté. Les sommets peuvent également avoir des temps de « naissance » différents induisant des sommets anciens et nouveaux aux propriétés différentes. Les sommets peuvent aussi avoir une position géographique qui encourage la formation d'arête avec des sommets proches géographiquement plutôt qu'avec des sommets éloignés. Enfin, avec comme inspiration des réseaux biologiques ou épidémiologiques, des traits comportementaux ou génétiques peuvent également être des caractéristiques importantes des sommets.

Citons quelques modèles de graphes aléatoires exhibant certaines de ces propriétés.

- Le modèle de configuration donne un graphe aléatoire où les degrés des sommets sont fixés (paramètre du modèle). Ce modèle a été introduit indépendamment par Bender et Canfield [BC78] en 1978 et par Bollobás [Bol80] en 1980 entre autre pour étudier les graphes réguliers (dont tous les sommets ont le même degré qui est fixé). Pour plus de détails voir [JŁR00, Chapitre 9], [FK15, Chapitre 11] ou [van16, Chapitre 7].
- Le modèle de Watts et Strogatz [WS98] donne un graphe aléatoire avec des degrés (des sommets) bornés, un grand nombre de triangles et un petit diamètre (i.e. la distance maximale entre deux sommets), et donc est petit monde. On part d'un graphe circulaire avec n sommets où chaque sommet est relié à ses k plus proches voisins à gauche et à droite, et on reconnecte chaque arête avec probabilité p à un nouveau sommet en conservant l'une des extrémités. Pour plus de détails voir [Dur06, Section 5.1].
- Le modèle à attachement préférentiel a été introduit par Barabási et Albert [BA99] et produit des graphes aléatoires invariant d'échelle. Ce modèle est construit de manière récursive, en partant d'un unique sommet, à chaque étape de temps un nouveau sommet est ajouté et est relié par une arête à d sommets choisis aléatoirement proportionnellement à leurs degrés. Notons que l' « âge » d'un

sommet (l'étape de temps où il est rajouté) influence son degré. Pour plus de détails voir [Dur06, Chapitre 4], [FK15, Chapitre 19] ou [van16, Chapitre 8].

- Les graphes aléatoires géométriques (parfois aussi appelé modèle à espace latent, ou *"latent space model"* en anglais) sont des modèles de graphes aléatoires où chaque sommet possède une position sur un espace géométrique latent et est relié par une arête uniquement aux sommets à moins d'une certaine distance r. Pour plus de détails voir [Pen03].
- Le modèle de graphe aléatoire exponentiel (parfois aussi appelé modèle p^*) est un modèle inspiré de la physique statistique : on part d'une loi de référence (e.g. la loi uniforme sur l'ensemble des graphes avec n sommets), puis l'on change la loi en pondèrent la probabilité de tirer chaque graphe par un facteur exponentiel. Ce facteur exponentiel peut par exemple faire intervenir des termes de dépendance entre les arêtes et même une dépendance à des caractéristiques des sommets. Cela permet d'obtenir des graphes aux caractéristiques proches de celles observées pour les réseaux réels. Ce modèle a été introduit par Holland et Leinhardt [HL81] en 1981, et a connu plusieurs améliorations depuis [Str86, FS86, AWC99, PN04]. Pour plus de détails voir [RPKL07] ou [LKR13].

1.3 Les graphons, objets limites des grands graphes denses

Dans cette section, nous allons désormais nous intéresser aux graphons, qui ont été introduits par Lovász et Szegedy [LS06] en 2006. Les graphons apparaissent comme les objets limites des suites de grands graphes denses. Ces graphes sont trop grands pour être représentés entièrement dans les applications ciblées. L'idée est donc de passer d'une représentation combinatoire donnée par le graphe à une représentation en continuum infini donnée par le graphon. Nous présenterons les propriétés importantes des graphons. Notons que les principaux résultats sur les graphons ont été réunis dans une monographie écrite par Lovász [Lov12] en 2012.

Notons également que Diaconis et Janson [DJ08] ont montré que la théorie des graphons et des limites de graphes est liée à la théorie des tableaux échangeables de variables aléatoires introduite par Aldous [Ald81] et Hoover [Hoo79].

1.3.1 Les graphons

Un *graphon* est une fonction mesurable symétrique W de $[0,1]^2$ dans $[0,1]$. De manière usuelle, on identifie deux graphons qui sont égaux presque partout. On dénote par $\mathcal{W}$ l'espace des graphons. Un graphon peut donc être vu comme une image pixelisée en nuance de gris, voir les exemples de la section précédente. On peut également voir le graphon W comme un graphe pondéré infini dont l'ensemble des sommets est l'intervalle $[0,1]$ et dont les arêtes sont pondérés par des poids dans $[0,1]$ qui représente la densité locale moyenne d'interaction (i.e. d'arêtes).

Une classe importante de graphons est celle des *fonctions étagées* (ou *step-functions* en anglais). Un graphon W est une fonction étagée s'il existe une partition (mesurable) finie $\{S_1, \cdots, S_k\}$ de l'intervalle $[0,1]$ telle que W est constant sur chaque ensemble $S_i \times S_j$ pour $1 \leq i,j \leq k$. Remarquons que les graphons fonctions étagées représentent les modèles à blocs stochastiques vu à la section précédente : pour la partition $\{S_1, \cdots, S_k\}$ on choisit des intervalles dont les longueurs sont données par la mesure $\nu = (\nu(1), \cdots, \nu(k))$ sur l'ensemble des types $[k]$, puis on définit $W_{S_i \times S_j} \equiv p_{i,j}$ pour $1 \leq i,j \leq k$.

À tout graphe fini G avec $n = v(G)$ sommets, on peut associer un graphon W_G fonction étagée via une construction proposée par [BCL$^+$08, Section 3.1]. Pour cela, on découpe l'intervalle $[0,1]$ en n intervalles $(J_1, \cdots, J_n)$ de même longueur $1/n$ (définies par $J_i = [\frac{i-1}{n}, \frac{i}{n})$ pour $1 \leq i \leq n$), et on définit W_G pour tout $x \in J_i$, $y \in J_j$ et $1 \leq i,j \leq n$ par :

$$W_G(x,y) = \begin{cases} 1 & \text{si } (i,j) \in E(G) \\ 0 & \text{sinon,} \end{cases}$$

où l'on suppose sans perte de généralité que $V(G) = [n]$. Remarquons que le graphe du graphon W_G (en tant que fonction) est l'image pixelisée associée au graphe G. Notons que l'ordre des sommets de G peut changer la définition de W_G en changeant l'ordre des intervalles. Cependant, comme les graphes que nous considérons sont non-ordonnés (i.e. il n'y a pas d'ordre sur les sommets), nous verrons dans la prochaine

section qu'en travaillant à permutation près des sommets nous pouvons identifier ces différentes valeurs de W_G.

Les graphons permettent également de définir un modèle de graphes aléatoires inhomogènes à variables latentes $\mathbb{G}(n, W)$ que l'on appelle *W-graphe aléatoire* et dont les paramètres sont le nombre de sommets n et un graphon W. Pour cela, on commence par générer les variables latentes ou types des sommets $X_1, \cdots, X_n$ selon des lois uniformes sur l'intervalle $[0, 1]$. Puis, conditionnellement aux types des sommets $X_1, \cdots, X_n$, chaque arête est ajoutée au graphe indépendamment des autres avec probabilité $W(X_i, X_j)$. Notons que le modèle de W-graphe aléatoire généralise le modèle à blocs stochastiques dans le cas où l'espace de types des sommets est quelconque et potentiellement infini. Notons également qu'il existe des résultats de concentration pour ces W-graphes aléatoires, voir [BCL$^+$06a] ou [Lov12, Chapitre 10].

Enfin, mentionnons qu'il est possible de remplacer l'espace $[0, 1]$ par un espace probabilisé $(\Omega, \mathcal{A}, \mu)$ quelconque et de définir des graphons W comme des fonctions mesurables symétriques de Ω^2 dans $[0, 1]$. Cela n'offre pas vraiment plus de généralité, mais permet dans certains contextes d'application de représenter de manière plus naturelles les types ou traits des sommets (e.g. âge, position géographique, traits génétiques, appartenance à des communautés). Notons que travailler avec des graphons construits sur un espace probabilisé quelconque introduit quelques difficultés techniques. Par simplicité, dans la suite, nous nous concentrons sur le cas où $\Omega = [0, 1]$ muni de la mesure de Lebesgue.

1.3.2 Distance de coupe et convergence pour les suites de graphes denses ou de graphons

Nous allons désormais introduire la distance de coupe qui permet de définir précisément la notion de convergence pour les graphons et les graphes denses, et que nous avions évoquée sous forme d'heuristique plus haut. Cette notion de convergence a été initialement introduite par Borgs, Chayes, Lovász, Sós et Vesztergombi [BCL$^+$06b, BCL$^+$08].

La distance de coupe est basé sur la *norme de coupe* $\|\cdot\|_\square$ qui a été introduite par Frieze et Kanan [FK99] en 1999. Dans le cas d'une matrice carrée A de taille $n \times n$ avec $n \in \mathbb{N}^*$ (e.g. A est la différence des matrices d'adjacence de deux graphes à n sommets), sa norme de coupe est définie par :

$$\|A\|_\square = \frac{1}{n^2} \max_{S,T \subset [n]} \left| \sum_{i \in S, j \in T} A_{i,j} \right|.$$

Un couple de sous-ensembles $S, T \subset [n]$ s'appellent une *coupe* du graphe. La norme de coupe s'interprète donc comme le maximum sur toutes les coupes possibles $S, T \subset [n]$ du graphe du nombre ou de la somme des poids $A_{i,j}$ des arêtes (i, j) traversant cette coupe de S vers T. La définition de la norme de coupe s'étend naturellement aux graphons. Pour toute fonction mesurable bornée W de $[0, 1]^2$ dans $\mathbb{R}$ (et en particulier si W est un graphon), sa norme de coupe est définie par :

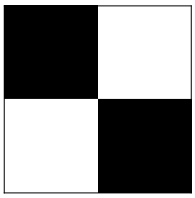
$$\|W\|_\square = \sup_{S,T \subset [0,1]} \left| \int_{S \times T} W(x, y) \, \mathrm{d}x\mathrm{d}y \right|,$$

où le supremum est pris sur l'ensemble des sous-ensembles mesurables $S, T \subset [0, 1]$. Ainsi, pour un graphe fini G, sa matrice d'adjacence A et son graphon associé W_G donnent la même valeur $\|A\|_\square = \|W_G\|_\square$ pour la norme de coupe.

Notons que la norme de coupe est plus petite que la norme L^1 : on a $\|W\|_\square \leq \|W\|_1$. Cependant, la norme L^1 est trop restrictive pour la convergence qui nous intéresse : si l'on reprend l'exemple heuristique précédent du graphe d'Erdős-Rényi de paramètre $1/2$, la norme L^1 empêche toute convergence car les graphons associés à ces graphes aléatoires $(G_n)_{n \in \mathbb{N}}$ sont à valeurs dans $\{0, 1\}$ et donc $\|W_{G_n} - W\|_1 = 1/2$ pour W le graphon constant égal à $1/2$. Tandis que nous verrons que la convergence a lieu pour la norme de coupe.

Mentionnons le lemme de régularité faible (qui intervient dans les preuves de nombreuses propriétés pour les graphons, et est une variante du lemme de partition de Szemerédi) qui pour la norme de coupe entre tout graphon W et son approximation par un graphon fonction étagée fournit une borne uniforme en le nombre d'éléments de sa partition de $[0, 1]$, voir [FK99] ou [Lov12, Chapitre 9].

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$



$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

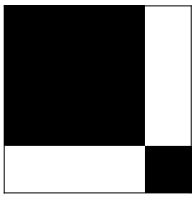


FIGURE 1.9 – Matrices d'adjacence et graphons associés pour deux graphes distincts. Notons que l'on peut passer d'un graphon à l'autre en "re-numérotant" via une bijection mesurable ne préservant pas la mesure. Cependant, aucune permutation des sommets ne permet d'obtenir un de ces deux graphes à partir de l'autre. De plus, notons que la densité moyenne d'arêtes $\int_{[0,1]^2} W(x,y)\,\mathrm{d}x\mathrm{d}y$ n'est pas la même pour ces deux graphons, tandis que les deux graphes ont le même nombre d'arêtes.

Comme nous l'avions évoqué plus haut sous forme d'heuristique, la notion de convergence pour les graphes denses et graphons que nous considérons ici est à ré-étiquetage (ou re-numérotation) près des sommets. Notons que pour le ré-étiquetage des graphes, une propriété importante des permutations des sommets est de préserver la masse / la mesure, voir Figure 1.9. Pour les graphons, il est donc naturel de remplacer les permutations des sommets d'un graphe fini par des bijections mesurables de $[0,1]$ préservant la mesure. De plus, les graphes bipartites de tailles différentes mais ayant le même graphon associé présentés en Figure 1.7 (ligne du haut) indiquent qu'on peut en fait relâcher l'hypothèse de bijectivité.

Le *ré-étiquetage* d'un graphon W par une application (non nécessairement bijective) mesurable φ préservant la mesure de $[0,1]$ donne un nouveau graphon W^φ défini par $W^\varphi(x,y) = W(\varphi(x),\varphi(y))$ pour tout $x,y \in [0,1]$. On peut alors définir la *(pseudo-)distance de coupe* $\delta_\square$, qui est une version à ré-étiquetage près de la norme de coupe $\|\cdot\|_\square$, pour tous U,W deux applications mesurables bornées de $[0,1]^2$ dans $\mathbb{R}$ (et en particulier si U et W sont des graphons) par :

$$\begin{aligned}\delta_\square(U,W) &= \min_{\varphi,\psi} \|U^\varphi - W^\psi\|_\square \\ &= \min_{\pi\in\mathcal{M}} \sup_{S,T\subset[0,1]^2} \left| \int_{S\times T} (U(x,y) - W(r,s))\ \pi(\mathrm{d}x,\mathrm{d}r)\pi(\mathrm{d}y,\mathrm{d}s) \right|,\end{aligned}$$

où pour la première égalité le minimum est pris sur l'ensemble des applications mesurables φ et ψ préservant la mesure de $[0,1]$, et où pour la second égalité $\mathcal{M}$ est l'ensemble des mesures de couplages entre deux copies de $[0,1]$ munie de la mesure de Lebesgue (i.e. des mesures de couplages entre les distributions de probabilités des types de sommets de deux graphons), le supremum est pris sur l'ensemble des sous-ensemble mesurable $S,T \subset [0,1]^2$, et l'intégrale est sur $(x,r) \in S$ et $(y,s) \in T$. Voir [BR09, Lemme 2.6] ou [Lov12, Théorème 8.13] pour l'équivalence entre ces deux expressions et le fait que les minima sont atteints.

Notons également que la définition de la distance de coupe $\delta_\square$ peut s'étendre au cas où chacun des deux graphons est construit sur un espace probabilisé distincts pour les types des sommets, voir [Jan13].

Notons que deux graphons distincts peuvent être à distance nulle pour $\delta_\square$ (e.g. un graphon fonction étagée et sa copie où on réordonne les blocs de la partition). En identifiant les graphons qui sont à distance nulle pour la distance de coupe $\delta_\square$, on obtient un espace quotient $\widetilde{\mathcal{W}}$ qui est l'*espace des graphons non-étiquetés*. La pseudo-distance de coupe $\delta_\square$ induit une distance sur l'espace $\widetilde{\mathcal{W}}$ que l'on appelle encore distance de coupe et que l'on note encore $\delta_\square$. De plus, la distance de coupe $\delta_\square$ rend l'espace $\widetilde{\mathcal{W}}$ compact.

Théorème 1.3.1 ([LS07, Theorem 5.1])**.** *L'espace des graphons non-étiquetés $\widetilde{\mathcal{W}}$ muni de la distance de coupe $\delta_\square$ est un espace métrique compact.*

Notons que la suite de graphes aléatoires $(\mathbb{G}(n,W))_{n\in\mathbb{N}^*}$ échantillonnés à partir d'un graphon W converge presque sûrement vers ce graphon W pour la distance de coupe $\delta_\square$, voir [BCL$^+$06a] ou [Lov12, Chapitre 10]. En particulier, on obtient que l'espace des graphons non-étiquetés $\widetilde{\mathcal{W}}$ est le complété de l'espace des graphes finis (non-étiquetés) pour la distance de coupe $\delta_\square$. Enfin, on obtient également que la convergence pour la distance de coupe coincide avec la converge heuristique « visuelle » présenté en Section 1.2 dans le cas du graphe d'Erdős-Rényi et du modèle à blocs stochastique. De plus, notons que pour la distance de coupe $\delta_\square$ on retrouve l'égalité entre les graphons des lignes du haut et du bas dans les Figures 1.7 and 1.8.

1.3.3 Densités d'homomorphisme et lien avec la distance de coupe

Nous allons maintenant relier la convergence pour la distance de coupe $\delta_\square$ à une autre notion de convergence pour les graphes denses introduite avant elle par Lovász et Szegedy [LS06]. Cette autre notion de convergence repose sur les densités d'homomorphisme de graphes et est parfois appelée convergence à gauche.

Notons qu'une troisième notion de convergence (appelée convergence à droite, reposant également sur les densités d'homomorphismes et ayant des liens avec des concepts de physique statistique) équivalente aux deux autres a été étudiée par Borgs, Chayes, Lovász, Sós et Vesztergombi [BCL$^+$12] (voir aussi [Lov12, Chapitre 12]).

Notons également qu'on peut exprimer de nombreux paramètres de graphes via les nombres et densités d'homomorphismes, voir [FLS06] ou [Lov12, Section 5.3] pour des exemples. Enfin, notons qu'il existe pour les paramètres de graphes un équivalent de la théorie des approximations de Taylor où les polynômes sont remplacées par des combinaisons linéaires de densités d'homomorphisme, voir [DGKR15].

Les nombres et densités d'homomorphisme de graphes

Soient F et G deux graphes. Un *homomorphisme (de graphe)* φ de F dans G est une application de $V(F)$ dans $V(G)$ préservant les arêtes, c'est à dire que pour chaque arête $(u,v) \in E(F)$, on a aussi $(\varphi(u),\varphi(v)) \in E(G)$. L'homomorphisme de graphe est donc une application préservant la structure d'adjacence de graphe définie par les arêtes. On note $\mathrm{Hom}(F,G)$ l'ensemble des homomorphismes de F dans G, et on appelle $|\mathrm{Hom}(F,G)|$ le *nombre d'homomorphisme* de F dans G.

Dans certains cas, on peut vouloir se restreindre au sous-ensemble $\mathrm{Inj}(F,G) \subset \mathrm{Hom}(F,G)$ des homomorphismes injectifs (i.e. qui envoient les sommets de F sur des sommets distincts de G). On peut également se restreindre au sous-ensemble $\mathrm{Ind}(F,G) \subset \mathrm{Inj}(F,G)$ des homomorphismes induits qui préserve les relations d'adjacence et de non-adjacence, c'est à dire que $(u,v) \in E(F)$ si et seulement si $(\varphi(u),\varphi(v)) \in E(G)$. On appelle $|\mathrm{Inj}(F,G)|$ le *nombre d'homomorphisme injectif* de F dans G, et $|\mathrm{Ind}(F,G)|$ le *nombre d'homomorphisme induit* de F dans G.

Usuellement, on fixe le graphe F qui est vu comme un « petit » graphe, et on fait varier le graphe G qui est vu comme un « grand » graphe. Les nombres d'homomorphisme $|\mathrm{Hom}(F,G)|$, $|\mathrm{Inj}(F,G)|$ et $|\mathrm{Ind}(F,G)|$ sont donc des paramètres de graphes qui compte le nombre de copies étiquetées de F dans G avec plus ou moins de restrictions : avec ($|\mathrm{Hom}(F,G)|$) ou sans ($|\mathrm{Inj}(F,G)|$ et $|\mathrm{Ind}(F,G)|$) répétitions possibles des sommets de G, et en préservant ($|\mathrm{Ind}(F,G)|$) ou non ($|\mathrm{Hom}(F,G)|$ et $|\mathrm{Inj}(F,G)|$) la relation de non-adjacence. Notons que dans ce contexte, le graphe F est parfois aussi appelé *motif* ou *graphlet*. Donnons un exemple : si F est un triangle (trois sommets tous reliés deux-à-deux par des arêtes), alors $|\mathrm{Hom}(F,G)| = |\mathrm{Inj}(F,G)| = |\mathrm{Ind}(F,G)|$ est égal au nombre de triangles dans G multiplié par 6 (pour tenir compte des permutations des sommets). Dans la Figure 1.10, on présente un exemple de calcul pour deux graphes donnés de ces différents nombres d'homomorphismes.

FIGURE 1.10 – Un exemple des nombres d'homomorphismes de graphes où les graphes F et G sont respectivement les graphes de gauche et de droite sur la figure. On trouve $\mathrm{Hom}(F,G) = 30$, $\mathrm{Inj}(F,G) = 18$, et $\mathrm{Ind}(F,G) = 4$. Faisons quelques remarques sur les décomptes liés au choix pour l'image de (b,a,c). Les choix $(1,5,4)$ et $(4,5,1)$ sont comptés indépendamment dans chacun de ces nombres d'homomorphismes. Le choix $(2,1,3)$ est compté pour $\mathrm{Hom}(F,G)$ et $\mathrm{Inj}(F,G)$ mais pas pour $\mathrm{Ind}(F,G)$ à cause de l'arête $(2,3)$. Enfin, le choix non-injectif $(2,1,2)$ est compté pour $\mathrm{Hom}(F,G)$ mais pas pour $\mathrm{Inj}(F,G)$ et $\mathrm{Ind}(F,G)$.

Remarque 1.3.2. Certains auteurs étudient les comptes de sous-graphes $\chi_F(G)$ plutôt que les nombres d'homomorphismes, la seule différence étant le fait de ne pas compter plusieurs fois le même sous-graphe

à cause des permutations possibles de ces sommets, c'est à dire :

$$\chi_F(G) = \frac{|\mathrm{Ind}(F,G)|}{|\mathrm{Aut}(F)|},$$

où $|\mathrm{Aut}(F)|$ est le nombre d'isomorphisme de graphe (i.e. de morphisme de graphe induit bijectif) d'un graphe F dans lui-même (voit [BR09, Section 2.1] pour plus de détails.)

Comme le nombre d'homomorphisme $|\mathrm{Hom}(F,G)|$ croit avec la taille du graphe G, pour espérer trouver une limite, on travaille plutôt avec sa version normalisée $t(F,G)$, que l'on appelle *densité d'homomorphisme* et qui est définie par :

$$t(F,G) = \frac{|\mathrm{Hom}(F,G)|}{n^p},$$

où l'on note $p = v(F)$ et $n = v(G)$ le nombre de sommets respectifs des graphes F et G, et n^p est le nombre d'applications de $V(F)$ dans $V(G)$. De la même manière, on définit la *densité d'homomorphisme injectif* t_{inj} et la *densité d'homomorphisme induit* t_{ind} par :

$$t_{\mathrm{inj}}(F,G) = \frac{|\mathrm{Inj}(F,G)|}{\binom{n}{p}}, \qquad \text{et} \qquad t_{\mathrm{ind}}(F,G) = \frac{|\mathrm{Ind}(F,G)|}{\binom{n}{p}},$$

où $\binom{n}{p}$ est le nombre d'applications injectives de $V(F)$ dans $V(G)$. Remarquons que la densité d'homomorphisme induit $t_{\mathrm{ind}}(F,G)$ peut s'interpréter comme la probabilité qu'en tirant p sommets de G uniformément sans remise, on obtienne un sous-graphe étiqueté de G qui soit une copie de F. Pour la densité d'homomorphisme injectif $t_{\mathrm{inj}}(F,G)$, on autorise ce sous-graphe à avoir plus d'arêtes que celles de F. Et pour la densité d'homomorphisme $t(F,G)$, le tirage des sommets est de plus fait avec remise.

On observe facilement que ces différentes notions d'homomorphismes sont reliées entre elles, voir [LS06, paragraphe 2.4] ou [Lov12, Section 5.2.3]. Pour F, G deux graphes finis avec p et n sommets respectivement, en bornant la probabilité de collision lors du tirage des sommets, on obtient :

$$|t(F,G) - t_{\mathrm{inj}}(F,G)| \leq \frac{1}{n}\binom{p}{2}. \tag{1.1}$$

On rappelle que $e(F) = |E(F)|$ est le nombre d'arêtes d'un graphe F. De plus, en utilisant un principe d'inclusion-exclusion (que l'on peut interpréter comme une formule d'inversion de Möbius), on obtient :

$$t_{\mathrm{inj}}(F,G) = \sum_{F' \supset F} t_{\mathrm{ind}}(F,G) \quad \text{et} \quad t_{\mathrm{ind}}(F,G) = \sum_{F' \supset F} (-1)^{e(F')-e(F)} t_{\mathrm{inj}}(F,G), \tag{1.2}$$

où $F' \supset F$ signifie que $V(F') = V(F)$ et $E(F) \subset E(F')$. Ainsi, si les densités d'homomorphisme induit t_{ind} sont plus précises, elles contiennent autant d'information que les densités d'homomorphisme t qui sont plus simples à étudier et sur lesquelles nous allons donc nous concentrer.

Il a été montré dans [Lov67] (voir aussi [Lov12, Théorème 5.29, Section 5.4]) que les nombres d'homomorphisme caractérisent un graphe : si $|\mathrm{Hom}(F,G_1)| = |\mathrm{Hom}(F,G_2)|$ pour tout graphe fini F, alors $G_1 = G_2$. Tandis que les densités d'homomorphismes ne caractérisent pas totalement un graphe : si $G(p)$ est obtenu à partir de G en remplaçant chaque sommet par p sommets jumeaux (c'est à dire p copies reliées aux mêmes sommets par arêtes), alors $t(F,G(p)) = t(F,G)$ pour tout graphe fini F. Cependant, c'est le seul défaut de caractérisation par les densités d'homomorphisme (voir [Lov12, Théorème 5.32, Section 5.4]) : si $t(F,G_1) = t(F,G_2)$ pour tout graphe fini F, alors il existe un graphe fini G et des entiers p_1, p_2 tels que $G_1 \cong G(p_1)$ et $G_2 \cong G(p_2)$ où $\cong$ dénote la relation d'isomorphisme de graphes (i.e. l'existence d'un isomorphisme de graphes entre les deux graphes). En particulier, cela signifie que G_1 et G_2 ont la même image pixelisée (à permutation près des sommets).

Les densités d'homomorphisme pour les graphons

La définition des différentes densités d'homomorphisme s'étend de manière naturelle aux graphons en utilisant la correspondance avec les graphes finis, voir [LS06]. Pour tout graphe fini F et tout graphon W, on définit donc les *densités d'homomorphisme* par :

$$t(F,W) = t_{\mathrm{inj}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{(i,j)\in E(F)} W(x_i,x_j) \prod_{i\in V(F)} \mathrm{d}x_i$$

$$\text{et} \qquad t_{\mathrm{ind}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{(i,j)\in E(F)} W(x_i,x_j) \prod_{(i,j)\notin E(F)} (1-W(x_i,x_j)) \prod_{i\in V(F)} \mathrm{d}x_i.$$

Remarquons que les densités d'homomorphisme sont invariantes par ré-étiquetage du graphon W, c'est à dire $t(F,W^\varphi) = t(F,W)$ et $t_{\mathrm{ind}}(F,W^\varphi) = t_{\mathrm{ind}}(F,W)$ pour tout graphon W et toute application mesurable φ préservant la mesure de $[0,1]$. Notons également que pour tout graphe fini G, on a $t_{\mathrm{inj}}(F,G) = t_{\mathrm{inj}}(F,W_G)$.

La raison pour laquelle les densités d'homomorphisme injectif $t_{\mathrm{inj}}(F,W)$ coïncident pour les graphons avec les densités d'homomorphisme $t(F,W)$ est dû au fait que la probabilité que deux sommets tirés uniformément sur $[0,1]$ soient égaux est nulle. Intuitivement, les tirages de sommets avec ou sans remise coïncident quand le nombre de sommets est infini (considérer (1.1) quand n tend vers l'infini). De plus, les formules reliant les densités d'homomorphisme injectif et induit t_{inj} et t_{ind} présentées en (1.2) restent valables pour les graphons.

On dit qu'une suite de graphes finis $(G_n)_{n\in\mathbb{N}}$ est *convergente* si pour tout graphe fini F, la suite $(t(F,G_n))_{n\in\mathbb{N}}$ converge dans $\mathbb{R}$. Le résultat principal de Lovász et Szegedy [LS06] pour la convergence en terme de densités d'homomorphisme est le suivant, garantissant l'existence d'un graphon limite pour toute suite convergente de graphes denses.

Théorème 1.3.3 ([LS06, Theorem 2.2])**.** *Pour toute suite de graphes convergente $(G_n)_{n\in\mathbb{N}}$, il existe un graphon W tel que :*

$$\lim_{n\to\infty} t(F,G_n) = t(F,W), \qquad \textit{pour tout graphe fini } F.$$

Enfin, terminons par le résultat de Borgs, Chayes, Lovász, Sós et Vesztergombi [BCL$^+$08] donnant l'équivalence entre la convergence pour la distance de coupe $\delta_\square$ et la notion de convergence définie par les densités d'homomorphisme.

Théorème 1.3.4 ([BCL$^+$08, Theorem 3.7])**.** *Pour toute suite de graphons $(W_n)_{n\in\mathbb{N}}$ et tout graphon W, les propositions suivantes sont équivalentes.*

(i) $\lim_{n\to\infty} t(F,W_n) = t(F,W)$*, pour tout graphe fini F.*

(ii) $\lim_{n\to\infty} \delta_\square(W_n,W) = 0$.

(iii) Pour tout $k\in\mathbb{N}$, la suite de graphes aléatoires $(\mathbb{G}(k,W_n))_{n\in\mathbb{N}}$ converge en distribution vers $\mathbb{G}(k,W)$, c'est à dire, on a $\lim_{n\to\infty} \mathbb{P}(\mathbb{G}(k,W_n)=F) = \mathbb{P}(\mathbb{G}(k,W)=F)$ pour tout graphe fini F à $v(F)=k$ sommets.

1.3.4 Applications des graphons

Ces dernières années, les graphons ont été utilisés dans de nombreux contextes d'application : les méthodes d'estimation non-paramétriques et les algorithmes pour les réseaux massifs [BC17], les modèles épidémiologiques SIS [DDZ22, DDZ23], l'étude des propriétés de transférabilité pour les réseaux de neurones graphiques (en anglais, *"graph neural network"* ou GNNs) [KBV21, KV23]. La limite d'échelle des graphes pondérés convergeant vers un graphon L^3 (i.e. une fonction mesurable L^3 de $[0,1]^2$ dans $\mathbb{R}_+$) est étudiée dans [BBB$^+$23], ce résultat étant la généralisation de la limite d'échelle pour le graphe aléatoire d'Erdös-Rényi dans la fenêtre critique (i.e. pour une probabilité de présence d'arête $p_n = n^{-1} + \lambda n^{-4/3}$ dépendant de n, où $p_n^c = n^{-1}$ est la probabilité critique pour l'émergence d'une composante « géante » du graphe de taille linéaire en n). De plus, il y a eu des développements récents dans l'étude des systèmes à champ moyen utilisant les graphons pour représenter l'hétérogénéité des interactions dans la population du système : les jeux stochastiques et leurs équilibres de Nash [CH21, LS22], la dynamique d'opinion sur un graphon [AN22], l'apprentissage par renforcement multi-agent coopératif [HWYZ23], les systèmes de particules en interaction [BWZ23, BK24], pour n'en citer que quelques-uns. Voir aussi le récent survey [AD24].

1.3.5 Autres objets limites pour les graphes

Les suites de graphes creux, dont la densité moyenne d'arêtes tend vers 0, convergent toutes vers le graphon nul. Pour la théorie des graphons, tous les graphes creux sont donc considérés comme équivalent au graphe vide, ce qui ne donne aucune information sur la structure de ces graphes. Dans cette section, nous présentons quelques autres théories permettant de traiter le cas des graphes creux.

Dans le cas des graphes de degrés bornés (parfois aussi appelés régime *"very sparse"* en anglais), leur structure est bien décrite par la théorie de Benjamini-Schramm [BS01] (parfois appelé convergence locale), et par un de ses raffinements, la convergence locale-globale [BR11, HLS14]. L'objet limite de la convergence locale-globale peut être représenté par un graphe borélien dont les degrés sont bornés (appelé un *graphing*) et satisfaisant certaines propriétés de mesurabilité (voir aussi [Lov12, Section 18.3]).

Plusieurs théorie existe pour essayer d'étendre la convergence des graphons dans le cas de graphes creux de densités intermédiaires, c'est à dire dont le nombre d'arêtes est sur-linéaire ($\Omega(n)$) mais sous-quadratiques ($o(n^2)$). Une première piste développée par Bollobás et Riordan [BR09] en 2009 consiste à re-normaliser la matrice d'adjacence pour obtenir une densité d'arêtes effective constante, et traduit l'intuition que deux graphes ayant des densités différentes peuvent néanmoins être structurellement similaires. Intuitivement, cette idée de re-normalisation pour tout de même approximer des graphes creux par un graphon est l'analogue du modèle d'Erdös-Rényi avec une probabilité de présence d'arêtes p_n dépendant de n comme $p_n = p \log n/n$. Cette théorie a ensuite été améliorée par Borgs, Chayes, Cohn et Zhao [BCCZ19, BCCZ18] en autorisant les graphons a être des fonctions L^p au lieu de juste borné, ce qui permet de plus grandes fluctuations dans les degrés des sommets (par exemple des graphes creux avec certaines parties denses). Ces L^p-graphons permettent entre autre d'analyser les graphes invariants d'échelle.

En parallèle, plusieurs auteurs [VR15, Jan16, CF17, BCCH18] ont proposé de traiter le cas des graphes creux de densité intermédiaire en considérant des graphons construits sur un espace σ-fini pour les types des sommets (au lieu d'un espace de probabilité), et que l'on appelle *graphexes* ou *graphon processes*. Les graphexes permettent de générer des graphes aléatoires avec un nombre dénombrable de sommets et un nombre fini d'arêtes via un processus ponctuel de Poisson sur l'espace de types des sommets. Voir également [BCCL19, BCCV19, VR19, BCDS21, Jan22] pour des graphexes un peu plus généraux. Notons que les L^p-graphons et graphexes traitent de différents graphes creux selon que la cause du faible nombre d'arêtes est uniformément répartie ou non entre les sommets, voir l'introduction de [BCCH18] pour plus de détails.

Plus récemment, dans le but d'unifier les différentes théories de limites de graphes, Backhausz et Szegedy [BS22] ont introduit un nouvel objet, le *graphop* (contraction de graphe et opérateur) qui correspond à l'opérateur de transition sur un graphe. Voir aussi [KLS19, NODM20] pour d'autres tentatives d'unifier les différentes théories de limites de graphes.

1.4 Les graphes pondérés

La plupart des exemples de réseaux issus du monde réel que nous avons donnés en Section 1.1 sont en fait des réseaux pondérés : chaque lien entre deux nœuds du réseaux possède une information supplémentaire ou un poids qui représente selon le contexte une distance, une intensité d'interaction, une résistance ou une capacité. Il nous faut donc compléter nos graphes en rajoutant des poids ou des décorations sur les arêtes. Voir la Figure 1.11 pour un exemple de réseau d'interaction protéine-protéine : les sommets sont les protéines et différents types d'interactions sont représentées par des arêtes de différentes couleurs ; notons que la base de données STRING [SKK$^+$23] permet de générer des graphes similaires avec une grande variété de protéines. Voir aussi la Figure 1.12 qui représente un graphe social d'interactions entre utilisateurs sur un post du site web "Reddit" : les sommets sont les utilisateurs, et la taille des arêtes orientés représente le nombre de messages envoyés d'un utilisateur à un autre, tandis que la couleur représente la teneur du message (sur une échelle allant de négatif à positif).

Formellement, soit $\mathbf{Z}$ un espace polonais (i.e. un espace métrique complet séparable) qui contient les décorations possibles des arêtes (e.g. $\mathbf{Z} = \mathbb{R}$ ou $\mathbb{R}_+$ pour des poids). Un *graphe pondéré* (ou *graphe décoré*, ou $\mathbf{Z}$-graphe) $G = (V, E, \phi)$ est la donnée d'un graphe (V, E) (pour la structure d'adjacence) et d'une fonction de décoration ϕ de E dans $\mathbf{Z}$ qui à chaque arête $e \in E$ associe une décoration $\phi(e)$ dans $\mathbf{Z}$. (Ces graphes sont parfois aussi appelés *graphes étiquetés* au sens où les arêtes sont étiquetées, et non

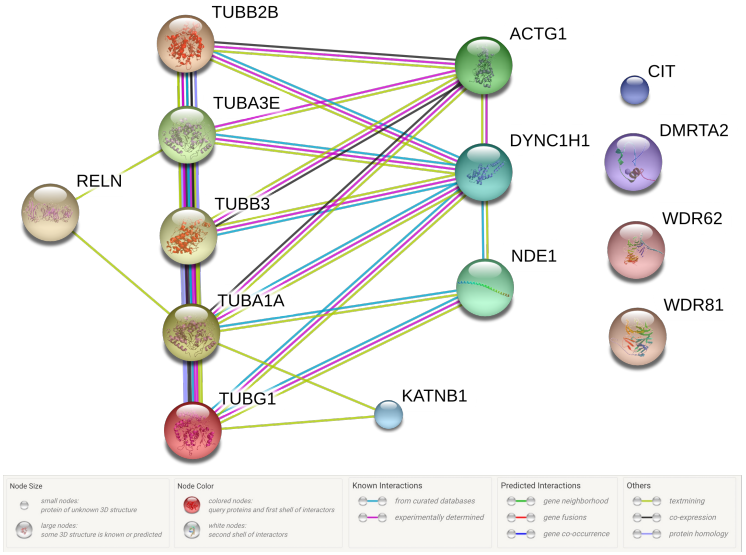


FIGURE 1.11 – Brainist. (2018) STRING protein-protein interaction networks of proteins associated with microlissencephaly STRING database : `http://version10.string-db.org/` Wikimedia commons. `https://commons.wikimedia.org/wiki/File:MLIS-STRING10.png`

pas comme précédemment où les sommets étaient numérotés.) Notons que l'on retrouve les multi-graphes pour $\mathbf{Z} = \mathbb{N}$.

Il est parfois plus naturel ou plus simple d'attribuer aux arêtes absentes du graphe une décoration commune qui est un élément neutre ou un point cimetière ∂ (possiblement rajouté dans ce but) de $\mathbf{Z}$ (e.g. le poids neutre 0 lorsque $\mathbf{Z} = \mathbb{R}$) ; ceci permet de considérer le graphe comme complet.

Il existe de nombreux modèles de graphes aléatoires pondérés. Les modèles les plus simples consistent à prendre un modèle de graphes aléatoires (non pondérés, comme ceux introduits en Section 1.2) et à rajouter des poids indépendants et uniformément distribués sur les arêtes. Par exemple, des modèles de configurations où les arêtes reçoivent des poids aléatoires indépendants distribués selon des lois exponentielles ont été considérés pour étudier la percolation de premier passage [BVDHH10], le temps de "flooding" (premier temps où tout les sommets du graphe reçoivent une information partie d'un unique sommet) [ADL13], ou encore le diamètre du graphe (maximum de la distance pondérée entre deux sommets du graphe, où la distance pondéré est le minimum du poids total des arêtes d'un chemin reliant les deux sommets) [AL15] Des modèles de graphes géométriques aléatoires (chaque sommet à une variable latente qui est une position dans un espace géométrique) où les sommets et les arêtes reçoivent des poids indépendants et identiquement distribués ont été considérés pour étudier la distance pondéré entre deux sommets choisis aléatoirement de manière uniforme [KL20].

Dans les modèles plus complexes de graphes aléatoires pondérés, la distribution du poids de chaque arête peut dépendre des caractéristiques ou variables latentes des sommets extrémaux. Dans [Gar09], l'auteur étudie des graphes aléatoires où chaque sommet u reçois un poids aléatoire w_u dans $(0, 1)$, puis chaque paire de sommets $\{u, v\}$ est reliée par une arête dont le poids est distribué selon une loi géométrique de paramètre $w_u w_v$. Un autre exemple est le *modèle à blocs stochastiques pondérés* (en anglais, *"weighted stochastic blocs models"*) où la distribution du poids des arêtes dépend des numéros de communauté des deux sommets extrémaux, et qui a été étudié pour résoudre les problèmes de détection de communauté [LMX15] (voir également [XML14] pour des modèles plus généraux où les décorations des arêtes viennent d'un espace compact), d'estimation exacte des communautés [JL15], et pour obtenir des bornes sur le nombre de sommets mal-classifiés [XJL20].

De plus, récemment, dans [HV23], les auteurs ont étudié la limite du poids total de l'arbre couvrant minimal (en anglais, minimum spanning tree ou MST) pour une suite de graphes aléatoires pondérés.

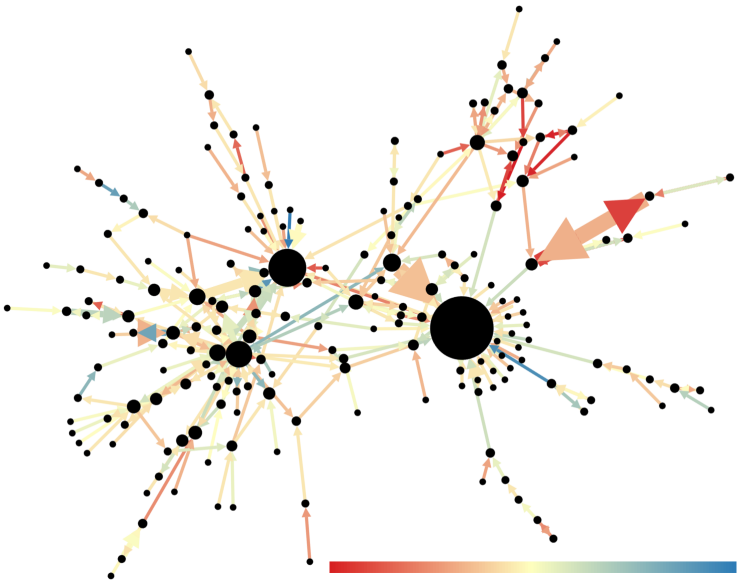


FIGURE 1.12 – Siobhán Grayson. (2017) Reddit is a social news aggregation, web content rating, and discussion website. This is a network visualisation of one submission from the subreddit called 'skeptic'. Each participant is represented by a black circle (node) and the coloured lines connecting nodes together (edges) represent the responses of participants. Node size represents 'in-degree', arrows indicate the edge direction, edge width the number of interactions, whilst the colour scales according to how negative (red) to positive (blue) a response is. Altogether, the image provides an example of how sentiment analysis, a natural language processing technique, can be used to enhance social network analysis of discussion forums. Wikimedia commons. `https://commons.wikimedia.org/wiki/File:Network_visualisation_incorporating_sentiment_analysis_of_the_subreddit_%27skeptic%27_from_Reddit.png`

1.5 Contribution : les graphons de probabilités

Dans cette section, nous présentons de manière synthétique les résultats du Chapitre 3 (qui correspond à la prépublication [ADW23]) : l'introduction et l'étude des *graphons de probabilités* (en anglais, *"probability-graphons"*) qui sont les objets limites pour les suites de grands graphes denses pondérés.

Motivés par les exemples des sections précédentes, nous considérerons les *graphons de probabilités* comme des limites possibles des grands graphes pondérés ; ils sont définis comme des applications de $[0,1]^2$ vers l'espace des mesures de probabilité $\mathcal{M}_1(\mathbf{Z})$ sur un espace polonais $\mathbf{Z}$ muni de sa tribu borélienne $\mathcal{B}(\mathbf{Z})$. Pour un graphe pondéré (possiblement aléatoire) $G = (V, E, \phi)$ dont les poids sont à valeurs dans $\mathbf{Z}$, on l'identifie avec le $\mathcal{M}_1(\mathbf{Z})$-graphe $G' = (V, E, \phi')$ où pour toute arête $e \in E$, le poids $\phi(e)$ est remplacé par la masse de Dirac $\phi'(e) = \delta_{\phi(e)}$ en ce poids (où δ_z dénote la masse de Dirac en $z \in \mathbf{Z}$). Précisons que si G est un graphe pondéré aléatoire, alors G' est un $\mathcal{M}_1(\mathbf{Z})$-graphe aléatoire dont chaque décoration d'arête est une masse de Dirac en un point qui lui est possiblement aléatoire. La construction du graphon associé W_G présentée en Section 1.3.1 s'étend alors naturellement au cas des graphes pondérés et des $\mathcal{M}_1(\mathbf{Z})$-graphes (comme nous le verrons en Section 1.5.2).

De plus, les graphons de probabilités permettent de généraliser les modèles de graphes aléatoires pondérés présentés dans la Section 1.4. À partir d'un graphon de probabilité W, on définit le W-graphe aléatoire $\mathbb{G}(n, W)$ de manière similaire à la construction présentée en Section 1.3.1. Pour cela, on commence par générer les variables latentes ou types des sommets $X_1, \cdots, X_n$ selon des lois uniformes sur l'intervalle $[0,1]$. Puis, conditionnellement aux types des sommets $X_1, \cdots, X_n$, chaque arête (i,j) reçoit un poids aléatoire $Z_{i,j}$ indépendamment des autres arêtes et distribué selon la mesure de probabilité $W(X_i, X_j; \cdot)$.

Dans cette section, nous allons définir une généralisation de la distance de coupe pour les graphons de probabilités. Nous verrons que la topologie induite par cette distance caractérise la convergence en distribution des sous-graphes échantillonnés $(\mathbb{G}(k, W_n))_{n\in\mathbb{N}}$ pour tout $k \in \mathbb{N}^*$ pour les suites de graphons de probabilité $(W_n)_{n\in\mathbb{N}}$.

1.5.1 Résumé des résultats et de la littérature proche

Dans le Chapitre 3, nous étudions les propriétés topologiques de l'espace des graphons de probabilités $\mathcal{W}_1$ lorsque $\mathbf{Z}$ est un espace polonais général : l'espace $\mathcal{W}_1$ est un espace topologique polonais et nous donnons des distances de coupe « naturelles » sur $\mathcal{W}_1$ qui sont complètes. L'une des principales difficultés est que l'espace des mesures de probabilité $\mathcal{M}_1(\mathbf{Z})$ peut être muni de nombreuses distances qui induisent la topologie de la convergence faible (i.e. la convergence étroite ou convergence en distribution des mesures de probabilité), chacune d'elles donnant lieu à une distance de coupe différente sur $\mathcal{W}_1$. Nous prouvons que la topologie induite sur $\mathcal{W}_1$ ne dépend pas du choix initial de la distance sur $\mathcal{M}_1(\mathbf{Z})$, à condition que cette distance satisfasse certaines conditions générales simples, et en particulier si cette distance est quasi-convexe (une propriété généralisant la convexité d'une norme). Cependant, nous soulignons que ces distances de coupe ne sont pas toutes complètes. Nous vérifions également que cette topologie caractérise la convergence en distribution des sous-graphes échantillonnés avec des poids aléatoires sur les arêtes ou, de manière équivalente, la convergence des densités d'homomorphismes des sous-graphes décorés par $C_b(\mathbf{Z})$ (l'espace des fonctions continues bornées de $\mathbf{Z}$ dans $\mathbb{R}$). De manière similaire au cadre des graphons, nous prouvons la convergence en distribution des grands sous-graphes pondérés échantillonnés $(\mathbb{G}(n, W))_{n\in\mathbb{N}^*}$ à partir d'un graphon de probabilité W vers ce dernier. Nous fournissons également un critère de compacité pour étudier la convergence des graphes pondérés vers les graphons de probabilités ; ce critère est un analogue du critère de tension pour les mesures de probabilités et généralise la condition de compacité dans [KR11] pour les multi-graphes (i.e. quand $\mathbf{Z} = \mathbb{N}$).

Lorsque $\mathbf{Z}$ est compact, cette question a été étudiée dans [LS10] et dans [Lov12, Section 17.1] en utilisant la convergence des densités d'homomorphismes de sous-graphes décorés avec des fonctions réelles définies sur $\mathbf{Z}$, voir aussi [KR11] sur les multi-graphes où $\mathbf{Z} = \mathbb{N}$, mais les propriétés métriques de l'ensemble des graphons de probabilités $\mathcal{W}_1$ n'ont été établies que lorsque $\mathbf{Z}$ est fini, voir [FOSU16]. Le travail [KLS22] est une extension de [LS10] où $\mathcal{M}_1(\mathbf{Z})$ est remplacé par l'espace dual $\mathcal{Z}$ d'un espace de Banach séparable $\mathcal{B}$. La norme introduite sur l'espace des graphons à valeurs dans $\mathcal{Z}$ implique la convergence des densités d'homomorphismes des sous-graphes décorés par $\mathcal{B}$, cependant il n'y a pas d'équivalence *a priori*. Comme $\mathcal{M}_1(\mathbf{Z})$ est un sous-ensemble du dual de $C_b(\mathbf{Z})$, cette approche couvre une partie de notre cadre lorsque $C_b(\mathbf{Z})$ est séparable, c'est-à-dire, lorsque $\mathbf{Z}$ est compact (voir Section 3.2 dans le Chapitre 3).

Concomitamment à notre travail, dans [AD23], les auteurs ont étudié les équations à champ moyen sur de grands graphes à poids réels modélisant des interactions avec un noyau de probabilité de $[0,1]^2$ vers $\mathcal{M}_1(\mathbb{R})$, l'ensemble des mesures de probabilité sur $\mathbb{R}$, mais ils n'ont pas étudié les propriétés topologiques de l'ensemble de ces noyaux de probabilité.

En conclusion, nous pensons que le cadre unifié développé ici est facile à utiliser et permettra d'utiliser les graphons de probabilités pour étudier les grands graphes (aléatoires) pondérés.

1.5.2 Définition des graphons de probabilités

Commençons donc par définir les *graphons de probabilités*, qui sont un analogue des graphons pour les graphes pondérés. Pour éviter toute confusion, dans le reste de ce chapitre nous dirons *graphons à valeurs réelles* au lieu de graphons. Nous considérons le cas général où les graphes pondérés prennent leurs poids d'arêtes dans un espace polonais $\mathbf{Z}$ (e.g. $\mathbb{Z}$, $\mathbb{R}$ ou $\mathbb{R}^d$), ce qui couvre également le cas des graphes décorés, des multi-graphes (graphes avec éventuellement plusieurs arêtes entre deux sommets) pour $\mathbf{Z} = \mathbb{N}$, et des graphes dynamiques (où les poids des arêtes évoluent au fil du temps) pour par exemple $\mathbf{Z} = \mathbb{R}^{\mathbb{N}}$ ou $\mathbf{Z} = \mathcal{C}([0,1])$ (l'espace des fonctions continues de $[0,1]$ dans $\mathbb{R}$).

Nous définissons un *graphon de probabilités* comme un noyau de probabilité $W : [0,1]^2 \to \mathcal{M}_1(\mathbf{Z})$, où $\mathcal{M}_1(\mathbf{Z})$ est l'espace des mesures de probabilité sur $\mathbf{Z}$.

Définition 1.5.1 (Graphon de probabilités)**.** *Un* graphon de probabilités *est une application W de $[0,1]^2$ dans $\mathcal{M}_1(\mathbf{Z})$ telle que :*

(i) W est une mesure de probabilité *en* dz *: pour tout $(x,y) \in [0,1]^2$, $W(x,y;\cdot)$ est dans $\mathcal{M}_1(\mathbf{Z})$.*

(ii) W est mesurable *en (x,y) : pour tout sous-ensemble mesurable $A \subset \mathbf{Z}$, la fonction $(x,y) \mapsto W(x,y;A)$ définie sur $[0,1]^2$ est mesurable.*

Un graphon de probabilités peut être interprété comme suit : le poids aléatoire z d'une arête entre deux sommets de type x et y dans $[0,1]$ est distribué selon la mesure de probabilité $W(x,y;\mathrm{d}z)$. En particulier, le cas spécial $\mathbf{Z} = \{0,1\}$ permet de retrouver les graphons à valeurs réelles : en effet, tout graphon à valeurs réelles $w : [0,1]^2 \to [0,1]$ peut être représenté comme un graphon de probabilités dont les valeurs sont des lois de Bernoulli :

$$W(x,y;\cdot) = w(x,y)\delta_1 + (1-w(x,y))\delta_0,$$

où δ_z dénote la masse de Dirac située en z. Mentionnons qu'il est possible de définir les graphons de probabilités sur un espace de probabilité plus général $(\Omega, \mathcal{A}, \mu)$ que $[0,1]$ pour les types de sommet, voir la Remarque 3.3.4 dans le Chapitre 3 pour les détails. Notons que le Point (ii) dans la Définition 1.5.1 est équivalente à la mesurabilité de W en tant qu'application de $[0,1]^2$ dans $\mathcal{M}_1(\mathbf{Z})$ muni de la topologie faible (i.e. la topologie de la convergence faible), voir la Section 3.2 du Chapitre 3.

Dans le Chapitre 3, nous définissons et étudions également les propriétés des *noyaux à valeurs mesures signées* qui sont des fonctions mesurables bornées (en masse totale/norme de variation totale) $W : [0,1]^2 \mapsto \mathcal{M}_\pm(\mathbf{Z})$ dont les valeurs sont des mesures signées (de masse totale finie), mais par soucis de brièveté, nous nous concentrons principalement sur les graphons de probabilités dans cette introduction. Notons qu'une des difficultés est que la topologie faible n'est en général pas métrisable sur l'espace des mesures signées $\mathcal{M}_\pm(\mathbf{Z})$, voir la Section 3.2 du Chapitre 3.

Comme les graphons de probabilités sont des fonctions mesurables, nous identifions les graphons de probabilités qui sont égaux pour presque tout $(x,y) \in [0,1]^2$, et nous dénotons par $\mathcal{W}_1$ l'espace des graphons de probabilités. De plus, comme nous considérons des graphes pondérés non-étiquetés (c'est-à-dire dont les sommets sont non-ordonnés), nous devons considérer les graphons de probabilités à « ré-étiquetage » près : pour une application mesurable préservant la mesure $\varphi : [0,1] \to [0,1]$ (application de ré-étiquetage pour les graphons de probabilités), nous définissons $W^\varphi(x,y;\cdot) = W(\varphi(x),\varphi(y);\cdot)$; nous disons que deux graphons de probabilités sont faiblement isomorphes s'il existe des applications préservant la mesure $\varphi, \psi : [0,1] \to [0,1]$ telles que $U^\varphi(x,y) = W^\psi(x,y)$ pour a.e. $(x,y) \in [0,1]^2$. Nous dénotons par $\widetilde{\mathcal{W}_1}$ l'espace des graphons de probabilités où l'on identifie les graphons de probabilités qui sont faiblement isomorphes.

Comme nous l'avons remarqué dans la section précédente, Nous pouvons toujours supposer que les graphes pondérés sont des graphes complets en ajoutant toutes les arêtes manquantes et en leur attribuant un poids / décoration ∂ qui est un point cimetière ajouté à $\mathbf{Z}$. Tout graphe pondéré G peut être représenté comme un graphon de probabilités W_G de la manière suivante : dénotons par n le nombre de sommets de G et divisons l'intervalle unité $[0,1]$ en n intervalles $I_1, \cdots, I_n$ de longueurs égales, alors W_G est défini pour $(x,y) \in I_i \times I_j$ comme $W_G(x,y;\cdot) = \delta_{\phi(i,j)}$, où $\phi(i,j)$ est le poids sur l'arête (i,j) dans G. Notons que les graphes pondérés peuvent être orientés ou non-orientés ; dans le cas des graphes pondérés non-orientés, leurs objets limites sont des graphons de probabilités symétriques, c'est-à-dire des graphons de probabilités W tels que $W(x,y;\cdot) = W(y,x;\cdot)$.

1.5.3 La distance de coupe pour les graphons de probabilités et ses différentes propriétés

Les différents choix pour la distance de coupe

Bien qu'il existe une distance usuelle sur l'ensemble des réels $\mathbb{R}$, ce n'est pas le cas pour les mesures de probabilité, les mesures ou les mesures signées dotées de la topologie faible. Dans ce chapitre et dans le Chapitre 3, notons que *mesure* signifiera toujours mesure positive. (Rappelons qu'en général, la topologie faible n'est pas métrisable sur les mesures signées, voir la Section 3.2 du Chapitre 3.) Présentons quelques distances et normes couramment utilisés. Commençons par la distance de Prohorov $d_\mathcal{P}$ qui peut être définie pour deux mesures $\mu, \nu \in \mathcal{M}_+(\mathbf{Z})$ par :

$$d_\mathcal{P}(\mu,\nu) = \inf\{\varepsilon > 0 : \forall A \in \mathcal{B}(\mathbf{Z}),\ \mu(A) \le \nu(A^\varepsilon) + \varepsilon \quad \text{et} \quad \nu(A) \le \mu(A^\varepsilon) + \varepsilon\},$$

où $A^\varepsilon = \{x \in \mathbf{Z} : \exists y \in A, d_0(x,y) < \varepsilon\}$ (où d_0 est une distance induisant la topologie sur $\mathbf{Z}$) est le ε-voisinage ouvert de A. Mentionnons également la norme de Kantorovitch-Rubinstein $\|\cdot\|_{\mathrm{KR}}$ (parfois également appelée norme de Lipschitz bornée, ou "bounded Lipschitz norm" en anglais) et la norme de

Fortet-Mourier $\|\cdot\|_{\mathrm{FM}}$ définies sur les mesures signées (mais métrisant la topologie faible sur les mesures) pour $\mu \in \mathcal{M}_{\pm}(\mathbf{Z})$ par :

$$\|\mu\|_{\mathrm{KR}} = \sup\left\{\int_{\mathbf{Z}} f \,\mathrm{d}\mu : f \text{ est 1-lipschitzienne et } \|f\|_\infty \leq 1\right\},$$
$$\|\mu\|_{\mathrm{FM}} = \sup\left\{\int_{\mathbf{Z}} f \,\mathrm{d}\mu : f \text{ est lipschitzienne et } \|f\|_\infty + \mathrm{Lip}(f) \leq 1\right\},$$

où $\|f\|_\infty = \sup_{x\in\mathbf{Z}} |f(x)|$ est la norme infinie et $\mathrm{Lip}(f)$ est la plus petite constante $L > 0$ telle que f est L-lipschitzienne. Nous appelons *suite déterminant la convergence* toute suite $\mathcal{F} = (f_k)_{k\in\mathbb{N}} \subset C_b(\mathbf{Z})$ telle que pour toute suite de mesures $(\mu_n)_{n\in\mathbb{N}}$ et mesure μ, si $\lim_{n\to\infty} \int_{\mathbf{Z}} f_k \,\mathrm{d}\mu_n = \int_{\mathbf{Z}} f_k \,\mathrm{d}\mu$ pour tout $k \in \mathbb{N}$, alors $(\mu_n)_{n\in\mathbb{N}}$ converge faiblement vers μ. Nous utilisons également une norme $\|\cdot\|_{\mathcal{F}}$ basée sur une suite déterminant la convergence $\mathcal{F} = (f_k)_{k\in\mathbb{N}} \subset C_b(\mathbf{Z})$ et définie par :

$$\|\mu\|_{\mathcal{F}} = \sum_{k\in\mathbb{N}} 2^{-k} |\mu(f_k)|.$$

Voir la Section 3.3.8 du Chapitre 3 pour différentes propriétés et relations de ces distances. Pour $\mathrm{m} \in \{\mathrm{KR}, \mathrm{FM}, \mathcal{F}\}$, nous notons d_{m} la distance induite par la norme N_{m}.

Pour définir un analogue de la norme de coupe pour les graphons de probabilités, nous devons d'abord choisir une distance d_{m} qui métrise la topologie faible sur l'espace des mesures de sous-probabilité $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (i.e. des mesures de masse totale au plus 1) ; nous définissons alors la *distance de coupe* $d_{\square,\mathrm{m}}$ pour les graphons de probabilités comme :

$$d_{\square,\mathrm{m}}(U, W) = \sup_{S,T\subset[0,1]} d_{\mathrm{m}}\Big(U(S\times T;\cdot), W(S\times T;\cdot)\Big),$$

où le supremum est pris sur tous les sous-ensembles mesurables S et T de $[0,1]$, et où $W(S\times T;\cdot) = \int_{S\times T} W(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y$ est une mesure de sous-probabilité et de même pour U. De plus, si la distance d_{m} est dérivée d'une norme N_{m} définie sur l'espace des mesures signées $\mathcal{M}_{\pm}(\mathbf{Z})$, alors la distance de coupe $d_{\square,\mathrm{m}}$ est dérivée de la *norme de coupe* $N_{\square,\mathrm{m}}$ définie sur les noyaux à valeurs mesures signées :

$$N_{\square,\mathrm{m}}(W) = \sup_{S,T\subset[0,1]} N_{\mathrm{m}}\Big(W(S\times T;\cdot)\Big).$$

Dénotons par $\tau_1 : (x,y) \mapsto x$ et $\tau_2(x,y) \mapsto y$ les projections sur chacune des deux composantes de $[0,1]^2$ vers $[0,1]$. Pour un graphon de probabilité $W \in \mathcal{W}_1$, un indice $i \in \{1,2\}$ et deux sous-ensembles mesurables $S, T \subset [0,1]^2$, on définit :

$$W^{\tau_i}(S\times T;\cdot) = \int_{S\times T} W(\tau_i(x,r), \tau_i(y,s);\cdot)\ \pi(\mathrm{d}x,\mathrm{d}r)\pi(\mathrm{d}y,\mathrm{d}s),$$

où l'intégrale est sur $(x,r) \in S$ et $(y,s) \in T$. Nous définissons ensuite la *distance de coupe non-étiquetée* $\delta_{\square,\mathrm{m}}$ sur l'espace des graphons de probabilités non-étiquetés $\widetilde{\mathcal{W}}_1$ comme :

$$\begin{aligned}\delta_{\square,\mathrm{m}}(U,W) &= \min_{\varphi,\psi} d_{\square,\mathrm{m}}(U^\varphi, W^\psi)\\ &= \min_{\pi\in\mathcal{M}} \sup_{S,T\subset[0,1]^2} d_{\mathrm{m}}\Big(U^{\tau_1}(S\times T;\cdot), W^{\tau_2}(S\times T;\cdot)\Big),\end{aligned}$$

où pour la première égalité le minimum est pris sur l'ensemble des applications mesurables φ et ψ préservant la mesure de $[0,1]$, et où pour la second égalité $\mathcal{M}$ est l'ensemble des mesures de couplages entre deux copies de $[0,1]$ munie de la mesure de Lebesgue, le supremum est pris sur l'ensemble des sous-ensemble mesurable $S, T \subset [0,1]^2$. Dans le Chapitre 3, voir la Proposition 3.3.18 pour des expressions alternatives de $\delta_{\square,\mathrm{m}}$ (y compris la preuve que le minimum existe pour la deuxième expression) et voir le Théorème 3.3.17 qui énonce que $\delta_{\square,\mathrm{m}}$ est bien une distance sur $\widetilde{\mathcal{W}}_1$. Remarquons que l'indice m permet de rappeler la dépendance de $\delta_{\square,\mathrm{m}}$ et $d_{\square,\mathrm{m}}$ par rapport à d_{m}. Dans la Proposition 3.4.13 du Chapitre 3, nous prouvons un équivalent du lemme de régularité faible pour les graphons de probabilités.

L'équivalence des topologies

Nous définissons la notion de distance quasi-convexe, qui généralise la convexité d'une norme.

Définition 1.5.2 (Distance quasi-convexe)**.** *Soit (X, d) un espace métrique qui est un sous-ensemble convexe d'un espace vectoriel. La distance d est* quasi-convexe *si pour tous $x_1, x_2, y_1, y_2 \in X$ et pour tout $\alpha \in [0,1]$, nous avons :*

$$d(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2) \leq \max(d(x_1, y_1), d(x_2, y_2)).$$

En particulier, toute distance (sur un sous-ensemble convexe d'un espace vectoriel) qui dérive d'une norme est quasi-convexe. De plus, les distances que nous avons présentées précédemment sont toutes quasi-convexes (voir le Lemme 3.3.21 dans le Chapitre 3).

Lemme 1.5.3. *Les distances $d_{\mathcal{P}}$, d_{KR}, d_{FM} et $d_{\mathcal{F}}$ sont toutes quasi-convexes sur $\mathcal{M}_+(\mathbf{Z})$.*

Un fait intéressant est que sous certaines conditions sur d_{m} (y compris le cas où d_{m} est quasi-convexe), la topologie induite par la distance de coupe associée $\delta_{\square,\mathrm{m}}$ ne dépend pas du choix particulier de d_{m}. Le théorème suivant est un cas particulier de deux résultats du Chapitre 3, le Théorème 3.5.5 ainsi que le Corollaire 3.4.14.

Théorème 1.5.4. *Les distances de coupe $\delta_{\square,\mathrm{m}}$, où d_{m} est une distance quasi-convexe sur $\mathcal{M}_{\leq 1}(\mathbf{Z})$ qui induit la topologie faible, induisent la même topologie sur l'espace des graphons de probabilités non-étiquetés $\widetilde{\mathcal{W}}_1$.*

En particulier, les distances de coupe $\delta_{\square,\mathcal{P}}$, $\delta_{\square,\mathrm{KR}}$, $\delta_{\square,\mathrm{FM}}$ et $\delta_{\square,\mathcal{F}}$ induisent la même topologie sur l'espace des graphons de probabilités non-étiquetés $\widetilde{\mathcal{W}}_1$.

Le critère de tension pour la convergence

Rappelons que $\mathbf{Z}$ est un espace polonais. Nous affirmons maintenant que $\widetilde{\mathcal{W}}_1$ est également polonais pour la distance $\delta_{\square,\mathcal{P}}$ (mais pas toujours pour $\delta_{\square,\mathcal{F}}$!), et nous renvoyons au Théorème 3.5.10 du Chapitre 3 pour d'autres distances.

Théorème 1.5.5. *L'espace des graphons de probabilités non-étiquetés $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathcal{P}})$ est un espace métrique polonais.*

Nous donnons un analogue du théorème de Prohorov avec un critère de tension pour les graphons de probabilités. Nous disons qu'un sous-ensemble de graphons de probabilités $\mathcal{K} \subset \widetilde{\mathcal{W}}_1$ est *tendu* si l'ensemble des mesures de probabilité $\{M_W : W \in \mathcal{K}\}$ est tendu (au sens des mesures de probabilité), où l'application $W \in \widetilde{\mathcal{W}}_1 \mapsto M_W \in \mathcal{M}_1(\mathbf{Z})$ est définie par :

$$M_W(\cdot) = W([0,1]^2; \cdot).$$

Notons que ce critère de tension pour les graphons de probabilités rappelle le critère de tension pour les mesures aléatoires, voir [Kal17, Theorem 4.10]. Le résultat suivant est une conséquence du Théorème 3.5.1 et de la Proposition 3.5.2 dans le Chapitre 3 ainsi que du Corollaire 3.4.14 lui aussi dans le Chapitre 3.

Théorème 1.5.6 (Propriété de compacité)**.** *Considérons la topologie sur $\widetilde{\mathcal{W}}_1$ du Théorème 1.5.4.*

(i) Si une suite d'éléments de $\widetilde{\mathcal{W}}_1$ est tendue, alors elle admet une sous-suite convergente.

(ii) Un sous-ensemble $\mathcal{K} \subset \widetilde{\mathcal{W}}_1$ est relativement compact si et seulement si il est tendu.

(iii) Si $\mathbf{Z}$ est compact, alors l'espace $\widetilde{\mathcal{W}}_1$ est compact.

En particulier, pour $\mathbf{Z} = \{0,1\}$ on retrouve la compacité de l'espace des graphons à valeurs réelles énoncée dans le Théorème 1.3.1.

1.5.4 Échantillonnage à partir d'un graphon de probabilités et lien avec la distance de coupe

Convergence des W-graphes aléatoires

Enfin, nous relions la topologie de la distance de coupe $\delta_{\square,\mathrm{m}}$ avec l'échantillonnage de sous-graphes. Les graphons de probabilité permettent de définir des modèles de graphes aléatoires pondérés (le modèle de W-graphe aléatoire) qui généralisent le modèle à blocs stochastiques pondérés, et qui jouent le rôle de sous-graphes échantillonnés pour les graphons de probabilité. Le W-graphe aléatoire (ou sous-graphe échantillonné à partir de W) de taille k, noté $\mathbb{G}(k,W)$, a deux paramètres : un nombre de sommets k et un graphon de probabilité W pour les poids des arêtes. Le graphe aléatoire $\mathbb{G}(k,W)$ est défini comme suit : tout d'abord, soit $X_1,\cdots,X_k$ k variables aléatoires (les « types des sommets ») indépendants et uniformément distribués sur $[0,1]$; ensuite, étant donné $X_1,\cdots,X_k$, chaque arête reçoit indépendamment un poids, où le poids de l'arête (i,j) est distribué selon $W(X_i,X_j;\cdot)$.

Nous démontrons également la convergence presque sûr des sous-graphes échantillonnés pour la topologie du Théorème 1.5.4, voir dans le Chapitre 3 le Théorème 3.6.13 ainsi que le Corollaire 3.5.6.

Théorème 1.5.7. *Soit W un graphon de probabilité. Alors, p.s. la suite des sous-graphes échantillonnés $(\mathbb{G}(k,W))_{k\in\mathbb{N}^*}$ converge vers W pour la topologie du Théorème 1.5.4.*

Pour prouver ce théorème, nous adaptons le schéma de preuve de [Lov12, Sections 10.5 et 10.6] en nous appuyant sur les premier et second lemmes d'échantillonnage pour les graphons à valeurs réelles. La preuve est réalisée en utilisant la distance de coupe $\delta_{\square,\mathcal{F}}$ en raison des bonnes propriétés d'approximation de $\|\cdot\|_{\mathcal{F}}$.

Les densités d'homomorphisme pour les graphons de probabilités

Rappelons les densités d'homomorphismes pour les graphons à valeurs réelles définies dans la Section 1.3.3. Dans le cas des graphes pondérés et des graphons de probabilité, nous devons remplacer l'absence / présence des arêtes (qui est à valeur 0-1) par des fonctions tests de $C_b(\mathbf{Z})$ décorant les arêtes de F. Ainsi, nous définissons la *densité d'homomorphisme* d'un $\mathcal{G}$-graphe F^g qui est un graphe fini $F=(V,E)$ dont les arêtes sont décorées par une famille de fonctions $g=(g_e)_{e\in E}$ d'un sous-ensemble $\mathcal{G}\subset C_b(\mathbf{Z})$ (en pratique, nous considérons seulement les cas $\mathcal{G}=C_b(\mathbf{Z})$ ou $\mathcal{G}=\mathcal{F}\subset C_b(\mathbf{Z})$ une suite déterminant la convergence), dans un graphon de probabilité W comme :

$$t(F^g,W)=M_W^F(g):=\int_{[0,1]^V}\prod_{(i,j)\in E}W(x_i,x_j;g_{i,j})\prod_{i\in V}\mathrm{d}x_i,$$

où $W(x,y;f)=\int_{\mathbf{Z}}f(z)\,W(x,y;\mathrm{d}z)$. De plus, M_W^F définit une mesure sur $\mathbf{Z}^E$, que nous notons encore M_W^F, définie par $M_W^F(\otimes_{e\in E}g_e)=M_W^F(g)$ pour $g=(g_e)_{e\in E}$. Notons que lorsque F est le graphe complet avec k sommets, M_W^F est la mesure jointe de tous les poids des arêtes du graphe aléatoire $\mathbb{G}(k,W)$, et caractérise ainsi le graphe aléatoire $\mathbb{G}(k,W)$.

Lien entre l'échantillonnage et la distance de coupe

Dans le Lemme 3.7.5 de comptage et le Lemme 3.7.7 de comptage faible du Chapitre 3, nous prouvons que la norme de coupe $\|\cdot\|_{\square,\mathcal{F}}$ permet de contrôler les densités d'homomorphismes. Réciproquement, dans le Lemme 3.7.8 de comptage inverse du Chapitre 3, nous prouvons que la norme de coupe $\|\cdot\|_{\square,\mathcal{F}}$ peut être contrôlée par les densités d'homomorphismes. En particulier, la topologie de la distance de coupe s'avère être exactement la topologie de la convergence en distribution pour les sous-graphes échantillonnés de toute taille donnée ; le résultat suivant est une conséquence directe du Théorème 3.7.11 du Chapitre 3.

Théorème 1.5.8 (Caractérisation de la topologie)**.** *Soit $(W_n)_{n\in\mathbb{N}}$ et W des graphons de probabilité non-étiquetés de $\widetilde{\mathcal{W}_1}$. Les propriétés suivantes sont équivalentes :*

(i) *$(W_n)_{n\in\mathbb{N}}$ converge vers W pour la topologie du Théorème 1.5.4.*

(ii) *$\lim_{n\to\infty}t(F^g,W_n)=t(F^g,W)$ pour tout $C_b(\mathbf{Z})$-graphe F^g.*

(iii) *$\lim_{n\to\infty}t(F^g,W_n)=t(F^g,W)$ pour tout $\mathcal{F}$-graphe F^g, pour une suite déterminant la convergence $\mathcal{F}$ donnée.*

(iv) Pour tout $k \geq 2$, la suite de sous-graphes échantillonnés $(\mathbb{G}(k, W_n))_{n\in\mathbb{N}}$ converge en distribution vers $\mathbb{G}(k, W)$.

Nous pouvons maintenant revenir au problème initial de trouver un objet limite pour une suite convergente de graphes pondérés $(G_n)_{n\in\mathbb{N}}$; ici, convergente signifie que pour tout $k \geq 2$, la suite $(\mathbb{G}(k, G_n) = \mathbb{G}(k, W_{G_n}))_{n\in\mathbb{N}}$ des sous-graphes échantillonnés de taille k (définie ci-dessus) converge en distribution (vers un certain graphe aléatoire limite). Notons que le critère de tension pour une suite de graphons de probabilité $(W_n)_{n\in\mathbb{N}}$ peut être reformulé de manière équivalente en termes de tension de la suite $(\mathbb{G}(2, W_n))_{n\in\mathbb{N}}$ des sous-graphes échantillonnés de taille 2. Ainsi, la convergence en distribution de la suite $(\mathbb{G}(2, G_n))_{n\in\mathbb{N}}$ implique sa tension, et donc la tension de la suite des graphons de probabilité $(W_{G_n})_{n\in\mathbb{N}}$. Ensuite, le Théorème 1.5.6 garantit l'existence d'un graphon de probabilité W et d'une sous-suite $(W_{G_{m_n}})_{n\in\mathbb{N}}$ de la suite $(W_{G_n})_{n\in\mathbb{N}}$ qui converge pour la distance de coupe $\delta_{\square,\mathcal{F}}$ vers W. Enfin, le Théorème 1.5.8 garantit que pour tout $k \geq 2$, la suite $(\mathbb{G}(k, G_n))_{n\in\mathbb{N}}$ converge en distribution vers $\mathbb{G}(k, W)$.

En conséquence, les graphons de probabilité sont exactement les objets limites pour les suites de graphes pondérés (possiblement aléatoire) $(G_n)_{n\in\mathbb{N}}$ dont le nombre de sommets tend vers l'infini et telle que pour chaque taille $k \geq 2$, la suite des sous-graphes échantillonnés $(\mathbb{G}(k, G_n))_{n\in\mathbb{N}}$ converge en distribution. Notons que lorsque le nombre de sommets ne tend pas vers l'infini, la limite est simplement un graphe pondéré.

Remarque 1.5.9 (Extension aux poids sur les sommets)**.** Le cadre que nous avons développé pour les graphons de probabilité pourrait facilement être étendu pour ajouter des poids sur les sommets, ou de manière équivalente pour permettre des boucles auto-référantes (i.e. des arêtes reliant un sommet à lui-même). Dans ce cas, les graphes pondérés et les graphons de probabilité ont un noyau à deux variables (graphon de probabilité) W^{e} pour les poids des arêtes comme précédemment, et un noyau à une variable $W^{\mathrm{v}} : [0, 1] \to \mathcal{M}_1(\mathbf{Z})$ pour les poids des sommets. Notons que dans ce cas, pour préserver la consistance de l'information entre les poids des sommets et des arêtes, nous choisissons d'utiliser la même application préservant la mesure $\varphi : [0, 1] \to [0, 1]$ pour les deux noyaux W^{v} et W^{e} lors du ré-étiquetage.

1.6 Perspectives

Limites locale et d'échelle de l'arbre couvrant minimal (MST) d'une suite de graphes pondérés aléatoires

Comme mentionné en Section 1.4, récemment, dans [HV23], les auteurs ont étudié la limite du poids total de l'arbre couvrant minimal (MST) pour une suite de graphes aléatoires pondérés. En suivant ce qui a été fait pour l'arbre couvrant uniforme dans [HNT18, ANS24], on s'attend à ce que les limites locale et d'échelle du MST soient directement construites à partir de la limite des graphes aléatoires pondérés, c'est à dire à partir d'un graphon de probabilités.

Détection de communauté pour le modèle à blocs stochastiques pondérés ou décorés (en anglais, weighted SBMs et labeled SBMs)

Un problème très étudié sur le modèle à blocs stochastiques est l'estimation des communautés, c'est à dire le fait d'estimer les variables cachés de communauté des sommets, et se divise selon le degré de parcimonie du graphe en deux cas : détection de communauté ($\Theta(n)$ arêtes) et estimation exacte des communautés ($\Theta(n \log n)$ arêtes) ; voir le survey d'Abbe [Abb18] pour plus de détails.

Dans le cas où les arêtes sont pondérées ou décorées, comme mentionné en Section 1.4, le modèle à blocs stochastiques pondérés ou décorés (en anglais, "weighted SBMs" et "labeled SBMs") a été étudié dans le cas symétrique (seulement deux distributions de poids / décorations partagées par toutes les arêtes selon qu'elles soient intra- ou extra-communautaires) pour résoudre le problème d'estimation des communautés. Dans le cas d'un espace fini de poids / décorations, le seuil de faisabilité (théorique et algorithmique) a été exhibé et démontré pour la détection de communauté [LMX15] et l'estimation exacte des communautés [JL15]. De plus, le nombre de sommets mal-classifiés a été étudié dans le cas de poids réels [XJL20] et de poids / décorations dans un espace mesurable général [ADL22]. Notons que ces différents résultats font intervenir un rapport signal-bruit (en anglais, "signal-to-noise" ratio ou SNR) qui repose sur la divergence de Rényi d'ordre 1/2. Notons que pour deux mesures de probabilité discrètes $P = (p_i)_{1\leq i\leq n}$

et $Q = (q_i)_{1 \leq i \leq n}$, la divergence de Rényi d'ordre 1/2 est définie par $D_{1/2}(P\|Q) = -2 \log \sum_{i=1}^{n} \sqrt{p_i q_i}$ (une généralisation de la divergence de Kullback-Leibler et qui est liée à la distance de Hellinger).

Il serait donc intéressant d'étendre les résultats de détection de communauté et d'estimation exacte des communauté au cas de distributions de poids / décorations sur un espace mesurable quelconque. Et il serait également intéressant d'étendre tous ces résultats au cas des modèles à blocs stochastiques pondérés ou décorés non-symétriques.

Éstimation de graphons de probabilités à partir de graphes échantillonnés

Dans le cas de l'estimation pour la norme L^2 des graphons à valeurs réelles à partir de graphes échantillonnés, plusieurs méthodes ont été étudiées sous l'hypothèse de régularité que le graphon est hölderien (possiblement par blocs). (Notons que pour la norme de coupe, [KV19] suggère que le meilleur estimé est directement le graphe échantillonné.) La méthode d'estimation la plus étudiée [ACC13, WO13, GLZ15, KTV17, LG24] consiste à approximer le graphon par une fonction étagée, ou de manière équivalente par un modèle à bloc stochastique, ceci nécessitant une étape d'estimation des blocs (ou groupes) de sommets comme dans la détection de communauté. Notons qu'il existe une variante de cette méthode [BCS15, BCSZ18] pour une estimation privée limitant l'information individuelle révélée sur chaque sommet. Une autre méthode d'estimation est la méthode spectrale [Cha15, Jia18] consistant à ne garder que les valeurs propres au dessus d'un certain seuil dans la décomposition en valeurs singulières de la matrice d'adjacence. Voir aussi le récent survey de Gao et Ma [GM21] pour plus de détails sur ces différentes méthodes et le lien avec la détection de communauté. Il serait également intéressant d'étudier l'estimation de graphons de probabilités à partir de graphes échantillonnés.

Chapitre 2

Modèles markoviens cachés indexés par des arbres

Repartons du problème de détection de communauté pour le modèle à blocs stochastiques pondérés ou décorés présenté en ouverture à la fin du Chapitre 1. Le lien entre ce problème et celui de reconstruction sur les arbres avec des poids ou décorations sur les arêtes à été étudié dans [HLM12, LMX15]. Commençons donc par définir les arbres, et nous reviendrons après au problème de reconstruction sur les arbres et commenter son lien avec la détection de communauté.

2.1 Une brève introduction sur les arbres

2.1.1 Définition formelle des arbres

Rappelons que les graphes ont été définis dans la Section 1 du Chapitre 1. On dit qu'un graphe est *connexe* si pour toute paire de sommets du graphe est reliée par un chemin. On dit qu'un graphe est *acyclique* s'il n'existe pas de chemin en cycle dans le graphe. Un *arbre* est un graphe connexe acyclique. Pour un graphe connexe (et donc aussi pour un arbre), la distance de graphe entre deux sommets est le plus petit nombre d'arêtes d'un chemin séparant ces deux sommets. Notons que le caractère acyclique d'un arbre implique que pour toute paire de sommets, il existe un unique chemin de longueur minimale. Notons également qu'un graphe fini avec n sommets possède $n-1$ arêtes.

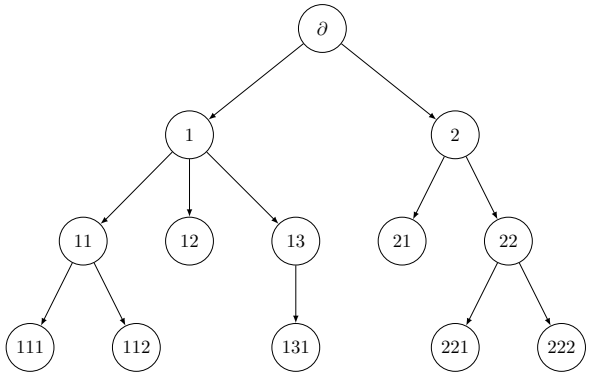


FIGURE 2.1 – Exemple d'arbre enraciné planaire. Notons que les feuilles ne sont pas toutes dans la dernière génération : les sommets 12 et 21 sont des feuilles et sont dans le génération 2.

Un arbre est dit *enraciné* s'il existe un sommet distingué noté ∂ que l'on appelle alors la *racine*. Un arbre enraciné T peut être décomposé en la suite de ses générations : la racine ∂ est le seul sommet de la génération 0, les sommets voisins de la racine forment la génération 1, et plus généralement les sommets à distance k de la racine forme la génération k. Pour un sommet $u \in T$ de la k-ième génération, ses voisins dans la génération $k+1$ sont appelés ses *enfants*, et si $k \geq 1$ (i.e. si u n'est pas la racine) son unique voisin

dans la génération $k-1$ est appelé son *parent* et est noté $\mathrm{p}(u)$. Notons que seule la racine ∂ n'a pas de parent. Des sommets partageant le même parent sont appelés des sommets *frères*. Un sommet qui n'a pas d'enfants est appelé une *feuille*. Nous notons $h(u)$ la *hauteur* du sommet $u \in T$, c'est-à-dire sa distance à la racine δ. Un arbre enraciné T est dit *ordonné* ou *planaire* si les enfants de chaque sommet sont munis d'un ordre. Par exemple, pour un sommet $u \in T$ ayant d enfants, cela signifie que les enfants de u sont ordonnés en une suite $u_1, \cdots, u_d$. Dans la Figure 2.1, nous présentons un exemple d'arbre enraciné planaire. Un autre exemple d'arbre enraciné planaire que nous reverrons par la suite est l'arbre binaire complet infini où chaque sommet a exactement deux enfants, voir la Figure 2.2.

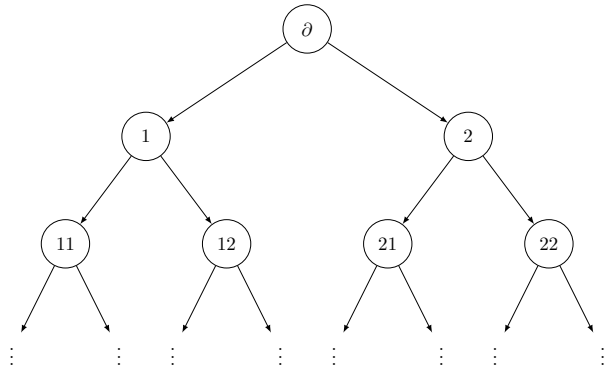


FIGURE 2.2 – L'arbre binaire complet est un exemple d'arbre enraciné planaire infini où tous les sommets ont exactement deux enfants.

Pour tout $k \in \mathbb{N}$ et $u \in T$ de hauteur $h(u) \geq k$, on note $\mathrm{p}^k(u) \in T$ le k-ième ancêtre de u qui est obtenu en appliquant k fois successivement la fonction parent p en partant de u ; de plus, u est appelé un *descendant* de $\mathrm{p}^k(u)$. Pour un sommet $u \in T$, on note $T(u)$ le sous-arbre de T enraciné en u et composé des descendants de u. Pour deux sommets $u, v \in T$, on note $u \wedge v$ le sommet de T qui est l'*ancêtre commun le plus récent* (en anglais, *"most recent common ancestor"*), ou plus simplement par abus l'*ancêtre commun*, de u et v. Notons que la distance de graphe d sur T entre deux sommets u et v, qui est le nombre d'arêtes de (l'unique) plus court chemin de u à v, peut s'écrire $d(u,v) = h(u) + h(v) - 2h(u \wedge v)$. Pour un arbre enraciné T, on note G_n la n-ième génération de l'arbre T et $T_n = \cup_{k=0}^{n} G_k$ l'arbre T jusqu'à la génération n incluse. Si cet arbre T est fini, on appelle sa *hauteur* le numéro de la dernière génération non vide, c'est-à-dire le maximum de $h(u)$ pour $u \in T$. Pour une suite $x = (x_u, u \in T)$ indexés par T, par simplicité, pour tout sous-ensemble $A \subset T$ (possiblement infini), nous notons $x_A = (x_u, u \in A)$.

Dans cette thèse, nous ne considérons que des arbres où chaque sommet a un nombre au plus dénombrable d'enfants (et même le plus souvent un nombre fini). Dans ce cas, tout sommet u d'un arbre enraciné planaire T peut être décrit de manière unique par la notation de Neveu [Nev86] qui encode le chemin de la racine ∂ jusqu'à u. Nous définissons la notation de Neveu de manière récursive : un sommet $u \in T$ de hauteur $n \geq 1$ est encodé par la suite d'entiers $(u_{(i)})_{1 \leq i \leq n} \in (\mathbb{N}^*)^n$ si u est le $u_{(n)}$-ième enfants de sont parent $\mathrm{p}(u)$ et que $\mathrm{p}(u)$ est encodé par $(u_{(i)})_{1 \leq i \leq n-1}$; la représentation de la racine ∂ est la suite vide unique élément de $(\mathbb{N}^*)^0$. Ainsi, si un sommet $u \in T$ de hauteur n est encodé par $(u_{(i)})_{1 \leq i \leq n}$, alors pour suivre le chemin de ∂ vers u : on part de $u_0 = \partial$, puis à l'étape $i \in \{1, \cdots, n\}$ on avance du sommet courant u_{i-1} vers son $u_{(i)}$-ième enfant u_i, et à la fin de l'étape n on arrive à $u_n = u$. Notons que dans les Figures 2.1 et 2.2, les sommets sont étiquetés par leur notation de Neveu.

Définissons l'arbre d'Ulam-Harris-Neveu $\mathbb{T}^\infty = \cup_{n \in \mathbb{N}} (\mathbb{N}^*)^n$ qui est l'ensemble des suites de longueur fini à valeurs dans $\mathbb{N}^*$. La racine ∂ de $\mathbb{T}^\infty$ est l'unique sommet de $(\mathbb{N}^*)^0$. Pour tout sommet $u = (u_{(i)})_{1 \leq i \leq n}$ de $\mathbb{T}^\infty$, sont sommet parent est $\mathrm{p}(u) = (u_{(i)})_{1 \leq i \leq n-1}$. L'arbre d'Ulam-Harris-Neveu $\mathbb{T}^\infty$ est donc un arbre enraciné planaire infini. Notons que $\mathbb{T}^\infty$ possède un nombre dénombrable de sommets. En utilisant la notation de Neveu, tout arbre enraciné planaire peut donc être vu comme un sous-arbre de l'arbre d'Ulam-Harris-Neveu $\mathbb{T}^\infty$ partageant la même racine ∂.

2.1.2 L'arbre de Bienaymé-Galton-Watson

Nous allons maintenant introduire un modèle d'arbre aléatoire très connu et étudié : l'arbre de Bienaymé-Galton-Watson. Cet arbre aléatoire permet de représenter la généalogie du processus introduit par Bienaymé [Bie45] en 1845 puis indépendamment par Galton et Watson [GW74] en 1874 pour étudier la disparition des noms de familles nobles. Voir [Ken75, Bac11] pour des informations historiques sur le processus de Bienaymé-Galton-Watson. Le processus de Bienaymé-Galton-Watson fait partie de la famille des *processus de branchement* qui servent entre autre à modéliser l'évolution de populations et ont de nombreuses applications en biologie, voir [HJV07, KA15]. Pour les principaux résultats de la théorie des processus de branchement, voir [Har63, AN72].

Soit $\mathrm{p} = (p_n)_{n\in\mathbb{N}}$ une distribution de probabilité sur $\mathbb{N}$ qui est la distribution du nombre d'enfants d'un individu de la population. Soit $(\xi_u)_{u\in\mathbb{T}^\infty}$ une famille de variables aléatoires indépendante et identiquement distribué de loi p. L'arbre aléatoire de Bienaymé-Galton-Watson T de loi de reproduction p est un sous-arbre aléatoire de $\mathbb{T}^\infty$ définie de manière récursive comme suit. La racine ∂ est dans T, puis pour tout $u \in T$ et tout $j \in \{1, \dots, \xi_u\}$ on a $uj \in T$, avec la convention $\{1, \cdots, \xi_u\} = \emptyset$ si $\xi_u = 0$.

On note $|A|$ le cardinal d'un ensemble fini A. Rappelons que l'on dénote G_n la n-ième génération de l'arbre T. Le processus de Bienaymé-Galton-Watson de loi de reproduction p est alors définie par $Z_n = |G_n|$ pour tout $n \in \mathbb{N}$. Soit $(\tilde{\xi}_{n,i})_{n\in\mathbb{N},i\in\mathbb{N}^*}$ une famille de variables aléatoires indépendante et identiquement distribué de loi p. Le processus $(Z_n)_{n\in\mathbb{N}}$ a même loi que le processus $(\tilde{Z}_n)_{n\in\mathbb{N}}$ défini de manière récursive par :

$$\tilde{Z}_0 = 1, \qquad \text{et} \qquad \tilde{Z}_{n+1} = \sum_{i=1}^{\tilde{Z}_n} \tilde{\xi}_{n,i}.$$

Supposons que $p_1 < 1$ pour éviter le cas trivial. On dit que le processus de Bienaymé-Galton-Watson est *sous-critique* (resp. *critique*, resp. *sur-critique*) si le nombre moyen d'enfants par individu $m := \sum_{k\in\mathbb{N}} k\, p_k$ satisfait $m < 1$ (resp. $m = 1$, resp. $m > 1$). Un résultat célèbre est que la population de l'arbre de Bienaymé-Galton-Watson s'éteint presque sûrement en temps fini (i.e. T est fini) dans les cas sous-critique et critique, mais survit (i.e. T est infini) avec probabilité positive dans le cas sur-critique, voir [Har63, Chapitre 1] ou [AN72, Section I.A].

2.1.3 De la détection de communauté avec poids au problème de reconstruction sur les arbres

Revenons désormais au problème de reconstruction sur les arbres avec poids sur les arêtes et à son lien avec la détection de communauté pour les modèles à blocs stochastiques pondérés.

Dans le problème de reconstruction sur un arbre enraciné (sans poids sur les arêtes), on diffuse un bit d'information de la racine vers les feuilles de l'arbre, et à partir de l'information reçue par les feuilles on cherche à estimer le bit envoyé par la racine. Plus formellement pour $\varepsilon \in (0, 1)$ et T un arbre enraciné, on définit un processus $(X_u, u \in T)$ indexé par l'arbre T à valeurs dans $\{0, 1\}$ où X_∂ est distribué selon une loi de Bernoulli de paramètre $1/2$ et $X_u = X_{\mathrm{p}(u)}$ avec probabilité $1 - \varepsilon$ pour tout $u \in T \setminus \{\partial\}$. Rappelons que G_n désigne la n-ième génération de l'arbre T et $T_n = \cup_{k=0}^n G_k$ l'arbre T jusqu'à la n-ième génération (incluse). L'objectif est d'estimer X_∂ à partir de $X_{G_n} := (X_u : u \in G_n)$.

Soient μ et ν deux mesures de probabilités. Dans le problème de reconstruction sur un arbre enraciné avec poids sur les arêtes, en plus de X_{G_n}, on observe pour chaque sommet $u \in T_n \setminus \{\partial\}$ un poids aléatoire $Y_{\mathrm{p}(u),u}$ sur l'arête $(\mathrm{p}(u), u)$ qui est indépendant des autres arêtes conditionnellement à $(X_{\mathrm{p}(u)}, X_u)$, et est distribué selon μ si $X_{\mathrm{p}(u)} = X_u$ et selon ν sinon. L'objectif est alors d'estimer X_∂ à partir de X_{G_n} et de $Y_{T_n} := (Y_{(\mathrm{p}(u),u)}, u \in T_n \setminus \{\partial\})$. Dans ce contexte, la *reconstruction* est dite *faisable* si $\lim_{n\to\infty} \mathbb{E}\big[\big|\mathbb{E}[X_\partial \,|\, X_{G_n}, Y_{T_n}] - 1/2\big|\big] > 0$, et *infaisable* sinon. Dans [HLM12], les auteurs exhibent un *rapport signal-bruit* (SNR ou "*signal-to-noise ratio*" en anglais) τ pour lequel la détection est faisable pour $\tau > 1$ et infaisable pour $\tau \leq 1$, voir [HLM12, Théorème 3]. Dans le cas sans poids sur les arêtes (ou pour $\mu = \nu$, ce qui fait que les poids n'apportent aucun information) et où chaque sommet de l'arbre a D enfants, ce rapport signal-bruit s'écrit $\tau = D(1 - 2\varepsilon)^2$ et est le fameux seuil de Kesten-Stigum [KS66], voir aussi [BRZ95, EKPS00].

Considérons le problème de détection de communauté sur le modèle à blocs stochastiques pondérés avec n sommets, deux communautés de même taille, une probabilité de présence d'arête a/n (resp. b/n) et une distribution μ (resp. ν) pour le poids de l'arête (si celle-ci est présente) entre deux sommets au sein

d'une même communauté (resp. de deux communautés différentes). Pour ce modèle à blocs stochastiques pondérés, notons H_n le graphe aléatoire, $(\tilde{X}_u, u \in H_n)$ les variables aléatoires donnant les numéros de communauté des sommets, et $(\tilde{Y}_{u,v}, E(H_n))$ les poids aléatoires des arêtes. Le lien entre les problèmes de reconstruction sur l'arbre et de détection de communauté et le fait qu'il partage le même seuil est du au fait suivant, voir [LMX15, preuve du Théorème 5] dans le cas pondéré et [Abb18, Section 4] dans le cas non-pondéré. Pour un sommet u de H_n (choisit aléatoirement ou non), le voisinage de u jusqu'à une distance $O(\log n)$ peut être couplé avec probabilité tendant vers 1 avec le processus de diffusion d'un bit d'information sur un arbre de Bienaymé-Galton-Watson de loi de reproduction la distribution de Poisson de paramètre $(a+b)/2$ et avec $\varepsilon = b/(a+b)$, voir [LMX15, Lemme 3].

Motivé par le lien entre ces deux problèmes, et par le fait que le processus $((X_u, Y_{\mathrm{p}(u),u}), u \in T \setminus \{\partial\})$ est un processus markovien caché, dans cette seconde partie de la thèse nous allons nous intéresser aux processus markoviens et markoviens cachés indexés par des arbres (nous définirons ces termes dans les sections suivantes).

2.2 Chaînes de Markov classiques et indexées par des arbres

2.2.1 Chaînes de Markov classiques

Commençons par rappeler la définition d'une chaîne de Markov classique, c'est-à-dire indexée par $\mathbb{N}$, introduite par Markov [Mar06, Mar04] en 1906. Pour deux variables aléatoires X et Y construites sur un même espace de probabilité, on note $\mathcal{L}(X \,|\, Y)$ la loi de X conditionnellement à Y. Un processus aléatoire $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov si pour tout $n \in \mathbb{N}$, on a :

$$\mathcal{L}(X_{n+1} \,|\, X_0, \cdots, X_n) = \mathcal{L}(X_{n+1} \,|\, X_n).$$

Cette relation est appelée *propriété de Markov*. L'idée est que conditionnellement à l'état présent, le passé et le futur du processus sont indépendants. Voir la Figure 2.3 qui illustre les relations de dépendance entre les variables d'une chaîne de Markov.

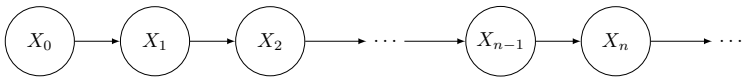


FIGURE 2.3 – Graphe de dépendance (markovienne) des variables d'une chaîne de Markov.

Dans cette thèse, nous considérons des chaînes de Markov à valeurs dans un espace métrique $\mathcal{X}$ muni de sa tribu borélienne $\mathcal{B}(\mathcal{X})$. Nous notons $\mathcal{M}_1(\mathcal{X})$ l'espace des mesures de probabilités sur $\mathcal{X}$ et $C_b(\mathcal{X})$ l'espace des fonctions continues bornées de $\mathcal{X}$ dans $\mathbb{R}$. Rappelons que $\mathcal{M}_1(\mathcal{X})$ muni de la topologie de la convergence faible est un espace métrique (voir Section 5 du Chapitre 1). La plupart des chaînes de Markov considérées dans cette thèse sont dites *homogènes*, c'est-à-dire que pour tout $n \in \mathbb{N}$ et tout $x \in \mathcal{X}$, nous avons :

$$\mathcal{L}(X_{n+1} \,|\, X_n = x) = \mathcal{L}(X_1 \,|\, X_0 = x).$$

Un *noyau de transition (de probabilité)* Q sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ est une application mesurable de $\mathcal{X}$ dans $\mathcal{M}_1(\mathcal{X})$ muni de la topologie de la convergence faible (et donc de la tribu borélienne associé), c'est-à-dire :

(i) Q est à valeurs *mesures de probabilité* : pour tout $x \in \mathcal{X}$, $Q(x;\cdot)$ est dans $\mathcal{M}_1(\mathcal{X})$.

(ii) Q est *mesurable* en x : pour tout ensemble mesurable $A \in \mathcal{B}(\mathcal{X})$, la fonction $x \mapsto Q(x;A)$ définie sur $\mathcal{X}$ est mesurable.

Notons la ressemblance avec les graphons de probabilités, voir la Définition 1.5.1 du Chapitre 1. La loi d'une chaîne de Markov homogène peut donc être définie à partir d'une mesure de probabilité initiale ν pour X_0 et d'un noyau de transition Q sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ pour tout $n \in \mathbb{N}$ par :

$$\forall f \in C_b(\mathcal{X}^{n+1}), \quad \mathbb{E}\big[f(X_0, \cdots, X_n)\big] = \int_{\mathcal{X}^{n+1}} f(x_0, \cdots, x_n)\ \nu(\mathrm{d}x_0) \prod_{k=1}^{n} Q(x_{k-1}; \mathrm{d}x_k),$$

ce que l'on note plus simplement :

$$\mathbb{P}(X_0 \in \mathrm{d}x_0, \cdots, X_n \in \mathrm{d}x_n) = \nu(\mathrm{d}x_0) \prod_{k=1}^{n} Q(x_{k-1}; \mathrm{d}x_k).$$

Pour les principaux résultats et propriétés sur les chaînes de Markov, voir [MT10, DMPS18].

Mentionnons également qu'une utilisation très répandues des chaînes de Markov est pour l'estimation via les méthodes de *Monte Carlo par chaînes de Markov* (MCMC ou *"Markov Chain Monte Carlo"* en anglais). Les méthodes de *Monte Carlo* permettent d'estimer l'intégrale d'une fonction intégrable f de $\mathbb{R}^d$ dans $\mathbb{R}$ pour $d \in \mathbb{N}^*$ contre une mesure de probabilité μ sur $\mathbb{R}^d$ en exploitant la *loi des grands nombres* : on échantillonne des variables aléatoires $X_1, \cdots, X_n$ indépendantes et distribuées selon μ, et la loi des grands nombres (voir par exemple [Kal02, Théorème 4.23]) nous assure que lorsque n est grand on a :

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \approx \mathbb{E}[f(X_1)] = \int_{\mathbb{R}^d} f \mathrm{d}\mu. \tag{2.1}$$

Cependant, lorsqu'il n'est pas possible d'échantillonner des variables de loi μ (par exemple si cette loi μ est absolument continue sous la mesure de Lebesgue avec une densité sous trop coûteuse à calculer), alors on peut remplacer la méthode de Monte Carlo par une méthode de Monte Carlo par chaînes de Markov : on échantillonne des variables aléatoires $(X_n)_{n\in\mathbb{N}}$ qui forment une chaîne de Markov de loi invariante μ dont la loi initiale et le noyau de transition sont plus simples à calculer, et un équivalent de la loi des grands nombres pour les chaînes de Markov nous donne que (2.1) est valable dans ce cas. Pour plus de détails sur les méthodes de Monte Carlo par chaînes de Markov, voir [BGJM11, LCL13].

2.2.2 Chaînes de Markov indexées par des arbres

Les *processus de Markov branchants* (ou *processus markoviens branchants* ou *chaîne de Markov branchante*) sont une généralisation des chaînes de Markov où le processus est indexé par un arbre enraciné au lieu d'être indexé par $\mathbb{N}$. Ces processus font partie de la famille des *processus de branchement* qui sont entre autres très utilisés pour décrire l'évolution et la croissance d'une population, voir [HJV07, KA15]. Un processus de Markov branchant $X = (X_u, u \in T)$ à valeurs dans un espace métrique $\mathcal{X}$ est donc un processus aléatoire indexé par un arbre enraciné T et vérifiant la *propriété de Markov* : des sommets frères (i.e. qui possèdent le même parent) prennent des valeurs indépendantes conditionnellement à la valeur de leur sommet parent. La propriété de Markov pour le processus de Markov branchant peut également se réécrire en faisant intervenir l'ancêtre commun $u \wedge v$ de deux sommets u et v de la manière suivante :

$$\mathcal{L}(X_u \,|\, X_v, X_{u\wedge v}) = \mathcal{L}(X_u \,|\, X_{u\wedge v}).$$

Rappelons que $T(u)$ dénote le sous-arbre de T des descendants du sommet u (u inclus). Pour un sommet $u \in T \setminus \{\partial\}$, la propriété de Markov donne également :

$$\mathcal{L}(X_u \,|\, X_{T\setminus T(u)}) = \mathcal{L}(X_u \,|\, X_{\mathrm{p}(u)}),$$

et même mieux :

$$\mathcal{L}(X_{T(u)} \,|\, X_{T\setminus T(u)}) = \mathcal{L}(X_{T(u)} \,|\, X_{\mathrm{p}(u)}).$$

La Figure 2.4 présente le graphe des relations de dépendance des variables d'un processus de Markov branchant indexé par un arbre binaire complet infini. Notons la ressemblance entre les graphes des Figures 2.4 et 2.2. En particulier, le graphe de dépendance des variables d'un processus de Markov branchant indexé par un arbre enraciné est précisément l'arbre indexant le processus. Pour un deuxième exemple, si on considérait un processus de Markov branchant indexé par l'arbre de la Figure 2.1, alors le graphe de dépendance des variables du processus serait là aussi l'arbre indexant le processus. Une autre façon de voir la propriété de Markov pour un processus de Markov branchant $X = (X_u, u \in T)$ est que si on conditionne par la valeur de X_u pour $u \in T$, alors le processus X devient indépendant entre les composantes connexes de $T \setminus \{u\}$, voir la Figure 2.5.

Dans la suite de cette thèse, nous considérons uniquement des processus de Markov branchants $X = (X_u, u \in T)$ *homogènes*, c'est-à-dire tels que pour tout $u, v \in T \setminus \{\partial\}$ et tout $x \in \mathcal{X}$, nous avons :

$$\mathcal{L}(X_u \,|\, X_{\mathrm{p}(u)} = x) = \mathcal{L}(X_v \,|\, X_{\mathrm{p}(v)} = x).$$

Plus formellement, on définit processus de Markov branchant (homogène) de la manière suivante.

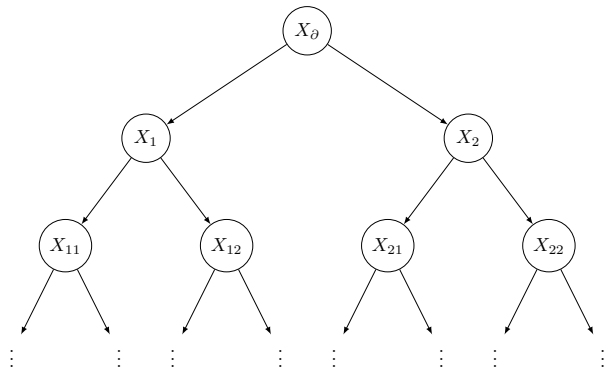


FIGURE 2.4 – Graphe de dépendance des variables d'un processus de Markov branchant ou chaîne de Markov indexée par un arbre binaire complet.

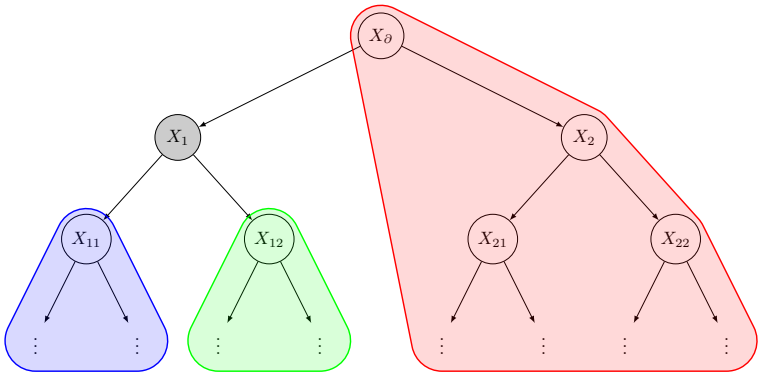


FIGURE 2.5 – Illustration de la propriété de Markov pour un processus de Markov branchant $X = (X_u, u \in T)$ indexée par un arbre binaire complet T. En conditionnant par X_1 (en gris), cela revient à rendre le processus X indépendant entre les trois composantes connexes de $T \setminus \{1\}$, c'est-à-dire $X_{T(11)}$, $X_{T(12)}$ et $X_{T\setminus T(1)}$ (respectivement en bleu, vert et rouge) sont indépendants conditionnellement à X_1.

Définition 2.2.1 (processus de Markov branchant)**.** *Un processus stochastique $X = (X_u, u \in T)$ est appelé un* processus de Markov branchant (homogène) *avec noyau de transition (de probabilité) Q sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et mesure (de probabilité) initiale ν sur $\mathcal{X}$ si pour tout sous-arbre fini A de T contenant ∂, on a :*

$$\mathbb{P}(X_A \in \mathrm{d}x_A) = \nu(\mathrm{d}x_\partial)\, \Pi_{u \in A\setminus\{\partial\}} Q(x_{\mathrm{p}(u)}; \mathrm{d}x_u).$$

Sans perte de généralité, comme T est un sous-arbre de l'arbre d'Ulam-Harris-Neveu $\mathbb{T}^\infty$, on peut toujours construire le processus de Markov branchant $X = (X_u, u \in \mathbb{T}^\infty)$ indexé par $\mathbb{T}^\infty$ puis le restreindre à T. Ceci permet de considérer le cas où l'arbre enraciné T est un arbre aléatoire ne dépendant pas du processus $X = (X_u, u \in \mathbb{T}^\infty)$ (par exemple si T est un arbre de Bienaymé-Galton-Watson défini en Section 2.1.2).

Les théorèmes limites, tels que la loi des grands nombres (parfois également appelée théorème ergodique dans les contextes markoviens), sont des outils importants pour étudier les propriétés d'une population, telles que la distribution des traits. La loi des grands nombres pour les processus de Markov branchants a été étudiée pour des valeurs à la fois discrètes et générales [AK98a, AK98b]. Pour étudier le vieillissement cellulaire, une version plus générale de la loi forte des grands nombres pour une classe plus large de fonctions test et pour des cellules filles non indépendantes a été prouvée dans [Guy07], voir aussi [DM10] pour une extension aux arbres de Bienaymé-Galton-Watson de degré borné et [Ban19] pour une extension à un environnement variable dans le temps et à une distribution de descendance dépendant des traits. Dans le Chapitre 4 (qui correspond à la prépublication [Wei24b]), nous présentons un théorème ergodique pour une large classe de fonctions test comme dans [Guy07], et pour des processus de Markov branchants où la reproduction est indépendante des traits individuels mais où l'arbre généalogique de la

population peut avoir une forme arbitraire (nous résumons ces résultats en Section 2.4.1).

2.3 Modèles markoviens cachés classiques et indexés par des arbres

2.3.1 Modèles markoviens cachés classiques

Définition des Modèles markoviens cachés classiques

Un *modèle markovien caché* (HMM ou *"hidden Markov model"* en anglais) ou *chaîne de Markov cachée* (HMC ou *"hidden Markov chain"* en anglais) est composé d'un processus caché et d'un processus observé. Le processus caché est une chaîne de Markov (classique) $X = (X_n)_{n\in\mathbb{N}}$. Notons que le processus caché est parfois appelé processus latent dans la littérature. Conditionnellement au processus caché X, le processus observé $Y = (Y_n)_{n\in\mathbb{N}}$, est composé de variables aléatoires indépendantes Y_n qui ne dépendent que de X_n pour tout $n \in \mathbb{N}$, c'est-à-dire :

$$\forall n \in \mathbb{N}, \qquad \mathcal{L}(Y_n \,|\, X) = \mathcal{L}(Y_n \,|\, X_n).$$

Voir la Figure 2.6 qui illustre les relations de dépendance entre les variables d'une chaîne de Markov cachée. Les HMMs ont été introduits pour la première fois par Baum et Petrie dans [BP66] et ont été popularisés par le tutoriel de Rabiner [Rab89]. Depuis lors, les HMMs ont été utilisés dans une grande variété d'applications telles que la reconnaissance vocale [YD15], la bio-informatique [Kos01], la finance [ME14], et l'analyse de séries temporelles [ZM09] ; voir aussi [BFA22] pour une référence plus globale sur les applications des HMMs.

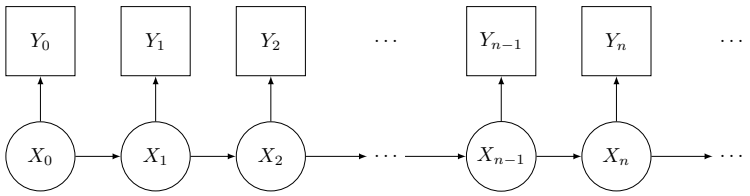


FIGURE 2.6 – Graphe de dépendance des variables d'une chaîne de Markov cachée. Les variables observées sont représentées dans des carrés, et les variables cachées dans des cercles.

Dans cette thèse, nous considérons le cas où les processus cachés et observés sont à valeurs dans des espaces métriques généraux $\mathcal{X}$ et $\mathcal{Y}$, respectivement. Tous les HMMs (X, Y) que nous considérons dans cette thèse sont dits *homogènes* (et nous ne le préciserons donc plus), c'est-à-dire que le processus caché X est une chaîne de Markov homogène et que pour tout $n \in \mathbb{N}$ et tout $x \in \mathcal{X}$, nous avons :

$$\mathcal{L}(Y_n \,|\, X_n = x) = \mathcal{L}(Y_0 \,|\, X_0 = x).$$

Pour toute suite $(x_n)_{n\in\mathbb{N}}$, par simplicité, on note $x_{i:j} = (x_n)_{i\le n\le j}$ pour $i \le j$. La loi d'un HMM peut donc être définie à partir d'une mesure (de probabilité) initiale ν pour X_0, d'un noyau de transition Q sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et d'un noyau de transition G sur $(\mathcal{X}, \mathcal{B}(\mathcal{Y}))$ pour tout $n \in \mathbb{N}$ par :

$$\mathbb{P}(X_{0:n} \in \mathrm{d}x_{0:n}, Y_{0:n} \in \mathrm{d}y_{0:n}) = \nu(\mathrm{d}x_0) \prod_{k=1}^{n} Q(x_{k-1}; \mathrm{d}x_k) \prod_{k=0}^{n} G(x_k; \mathrm{d}y_k).$$

Pour les principaux résultats et propriétés sur les HMMs, voir [CMR05].

Estimation par maximum de vraisemblance pour les HMMs

Dans cette thèse, nous nous intéressons au cas où la loi du HMM fait partie d'une famille de loi dépendant d'un paramètre θ, c'est-à-dire que la mesure initiale ν_θ et les noyaux de transition Q_θ et G_θ dépendent du paramètre θ. Par exemple, si l'espace des états cachés $\mathcal{X}$ est fini et que Y_n conditionné à X_n est une variable aléatoire gaussienne pour chaque $n \in \mathbb{N}$, alors θ pourrait paramétrer la matrice de transition du processus caché ainsi que la moyenne et la variance de la distribution gaussienne associée à

chaque état caché. Nous notons Θ l'ensemble des paramètres θ possibles que nous supposons être un sous-espace compact de $\mathbb{R}^d$ avec $d \in \mathbb{N}^*$. L'objectif est alors d'estimer le vrai paramètre $\theta^\star$ du HMM à partir de l'observation seule des variables $Y_{0:n}$ pour $n \in \mathbb{N}$. Pour ce faire, une méthode très répandue est d'utiliser le principe d'*estimation par maximum de vraisemblance* introduit par Fisher [Fis22, Fis92] en 1922 et qui consiste à choisir un ou le paramètre θ maximisant la *vraisemblance* des données observées, c'est-à-dire la probabilité ou la densité de probabilité contre une mesure de référence (commune à tous les paramètres dans Θ) sous ce paramètre θ des données observées. Pour des détails historiques sur l'estimation par maximum de vraisemblance, voir [Ald97]. Supposons que les noyaux de transition G_θ admettent une densité g_θ par rapport à une mesure de référence commune μ, c'est-à-dire que g_θ est une fonction mesurable de $\mathcal{X} \times \mathcal{Y}$ dans $\mathbb{R}_+$ telle que pour tout $x \in \mathcal{X}$ et tout $A \in \mathcal{B}(\mathcal{Y})$, on a $G_\theta(x; A) = \int_{\mathcal{Y}} g_\theta(x, y)\ \mu(\mathrm{d}y)$. Pour tout $n \in \mathbb{N}$, la *vraisemblance* de données observées $y_{0:n} \in \mathcal{Y}^{n+1}$ sous la loi de paramètre $\theta \in \Theta$ est alors définie par :

$$\ell_n(\theta; y_{0:n}) := \log \int_{\mathcal{X}^{n+1}} g_\theta(x_k, y_k)\ \nu_\theta(\mathrm{d}x_0) \prod_{k=1}^{n} Q_\theta(x_{k-1}; \mathrm{d}x_k)$$

En remplaçant $y_{0:n}$ par les données observées $Y_{0:n}$ et en maximisant la vraisemblance, on peut alors définir l'*estimateur du maximum de vraisemblance* (MLE ou *"maximum likelihood estimator"* en anglais) $\hat{\theta}_n$, qui est une variable aléatoire dépendant de $Y_{0:n}$, par :

$$\hat{\theta}_n = \hat{\theta}_n(Y_{0:n}) \in \mathrm{argmax}_{\theta \in \Theta}\, \ell_n(\theta; Y_{0:n}). \tag{2.2}$$

Notons que sous certaines conditions de régularité du modèle (par exemple si la vraisemblance est une fonction continue de θ), le maximum est atteint est l'ensemble argmax est bien non vide. Notons également que l'ensemble argmax dans (2.2) n'est pas nécessairement unique, auquel cas nous sélectionnons un paramètre θ de l'ensemble des argmax de manière mesurable (ce qui est possible, voir [BS96, Proposition 7.33]).

En pratique, l'estimation du maximum de vraisemblance pour les HMMs s'appuie souvent sur des méthodes numériques itératives pour approcher le MLE. Ces méthodes sont souvent basées sur l'algorithme d'*espérance-maximisation* (EM ou "*expectation-maximization*" en anglais) qui est un algorithme pour les modèles avec des données manquantes et a été popularisé par Dempster et al. [DLR77] dans un article pionnier. Pour les HMMs avec un espace d'états cachés fini, la première présentation d'une stratégie complète d'espérance-maximisation est due à Baum et al. [BPSW70], et est le bien connu algorithme "forward-backward" ou algorithme de Baum-Welch. Le cas plus complexe des HMMs avec un espace d'états cachés continu a été étudié plus tard dans les années 1990 en utilisant des méthodes basées sur des simulations de Monte Carlo par chaînes de Markov, voir [DJS95, CL95, DK97, KSC98]. Pour plus de détails sur les algorithmes d'espérance-maximisation et "forward-backward" ainsi que leurs approximations stochastiques, voir [CMR05, Chapitres 10 et 11].

Propriétés asymptotique du MLE pour les HMMs

Deux propriétés statistiques importantes du MLE à étudier sont la *consistance forte* et la *normalité asymptotique*. On dit que le MLE $\hat{\theta}_n$ est *fortement consistant* si presque sûrement il converge vers le vrai paramètre $\theta^\star$. On dit que le MLE $\hat{\theta}_n$ est *asymptotiquement normal* si $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ a des fluctuations normales, c'est-à-dire si on a la convergence en distribution suivante :

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, M),$$

où $\mathcal{N}(0, M)$ est une distribution gaussienne centré de matrice de covariance M et où M est généralement l'inverse de la matrice d'information de Fisher du modèle.

Les propriétés statistiques du MLE pour les HMMs ont été étudiées pour la première fois dans [BP66], qui a prouvé la consistance et la normalité asymptotique dans le cas où les processus cachés et observés ne peuvent prendre qu'un nombre fini de valeurs. Ces résultats ont ensuite été successivement étendus dans une série d'articles [Ler92, BRR98, JP99, LGM00, DM01]. Une extension de ces résultats pour les HMMs avec auto-régression (c'est-à-dire, lorsque conditionnellement à la chaîne de Markov cachée, le processus observé est une chaîne de Markov d'ordre s inhomogène pour un certain $s \in \mathbb{N}$) a été développée plus tard dans [DMR04], qui a prouvé, sous des hypothèses plus faibles, la consistance forte et la normalité

asymptotique du MLE pour les HMMs auto-régressifs avec un espace d'états cachés compact et avec un régime éventuellement non stationnaire. Les méthodes utilisées dans [DMR04] reposent sur l'expression du logarithme de la vraisemblance comme une fonction additive d'une chaîne de Markov étendue avec un passé infini, grâce à la stationnarité, et en utilisant l'ergodicité géométrique de cette chaîne étendue (l'extension au régime non stationnaire est ensuite faite séparément). La méthode de [DMR04] a été adaptée dans [KS19] sous des hypothèses similaires pour permettre aux densités de transition du processus caché de prendre des valeurs nulles. Depuis l'article [DMR04], la consistance forte du MLE a été prouvée sous des hypothèses plus faibles dans [GL06, DMOVH11, DRS16], mais aucune généralisation n'a été faite pour la normalité asymptotique du MLE.

2.3.2 Modèles markoviens cachés indexés par des arbres

Un *modèle markovien caché indexé par un arbre* ou *chaîne de Markov branchante cachée* (en anglais, *hidden Markov tree* ou HMT) est une généralisation des chaînes de Markov cachées présentées en Section 2.3.1 dans la cas où le processus est indexé par un arbre enraciné. Les HMTs ont été introduits pour la première fois dans [CNB98] pour prendre en compte la dépendance multi-échelle des coefficients d'ondelettes dans le traitement statistique du signal, avec des applications dans le traitement d'images [RCBK00, CB01, DWB08, SH17]. Par la suite, les HMTs ont été utilisés dans plusieurs contextes d'application tels que le traitement du langage naturel [GOB13, KDM13], la cartographie des inondations [XJS18], l'imagerie médicale [MBY+12, HYG17, HBSLLB+17], la modélisation de la croissance des plantes [DGCC05], et la bio-informatique [OCB+09, BWX13, NSK20].

Le HMT est donc composé d'un processus caché et d'un processus observé. Le processus caché $X = (X_u, u \in T)$ est un processus de Markov branchant indexé par un arbre enraciné T à valeurs dans un espace métrique $\mathcal{X}$, comme défini en Section 2.2.2. Conditionnellement au processus caché X, le processus observé $Y = (Y_u, u \in T)$, à valeurs dans un autre espace métrique $\mathcal{Y}$, est composé de variables aléatoires indépendantes Y_u qui ne dépendent que de X_u pour tout $u \in T$, c'est-à-dire :

$$\forall u \in T, \qquad \mathcal{L}(Y_u \,|\, X) = \mathcal{L}(Y_u \,|\, X_u).$$

La Figure 2.7 présente le graphe des relations de dépendance des variables d'un processus de Markov branchant indexé par un arbre binaire complet infini.

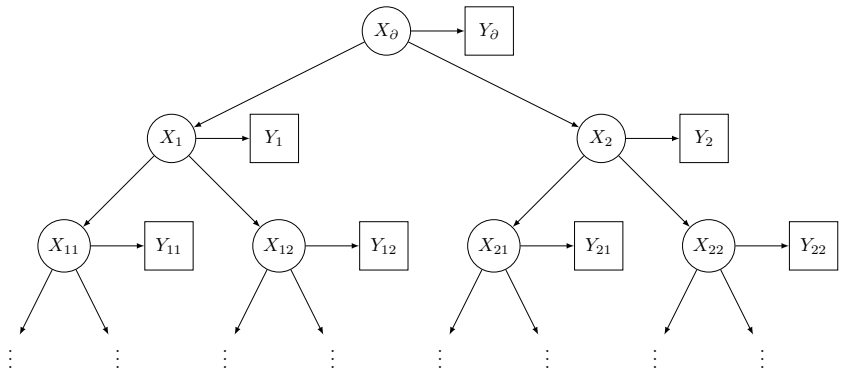


FIGURE 2.7 – Graphe de dépendance des variables d'une chaîne de Markov cachée indexée par un arbre binaire complet. Les variables observées sont représentées dans des carrés, et les variables cachées dans des cercles.

Notons que les graphes de dépendance des Figures 2.6 et 2.7 montrent que les HMMs et les HMTs peuvent être vu comme des processus de Markov branchants (comparer ces deux figures avec la Figure 2.4) en considérant chaque sommet associé à une variable observée Y_u (resp. Y_n) comme un enfant du sommet associé à sa variable cachée correspondante X_u (resp. X_n). Ce fait permet de mieux comprendre comment exploiter la propriété de Markov pour le HMT. Par exemple, la propriété de Markov pour le HMT implique que pour tout $k \in \mathbb{N}^*$, tout $u \in T$ de hauteur $h(u) \leq k$, et tout sous-ensemble $A \subset T$, on a :

$$\mathcal{L}(X_u \,|\, Y_A, X_{\mathrm{p}^k(u)}) = \mathcal{L}(X_u \,|\, Y_{A \cap T(\mathrm{p}^{k-1}(u))}, X_{\mathrm{p}^k(u)}), \tag{2.3}$$

résultat que nous réutiliserons dans le Chapitre 5. Voir la Figure 2.8 qui illustre l'utilisation de la propriété de Markov du HMT et permet d'obtenir (2.3), et notons la ressemblance de cette figure avec la Figure 2.5 qui illustre la propriété de Markov pour le processus de Markov branchant.

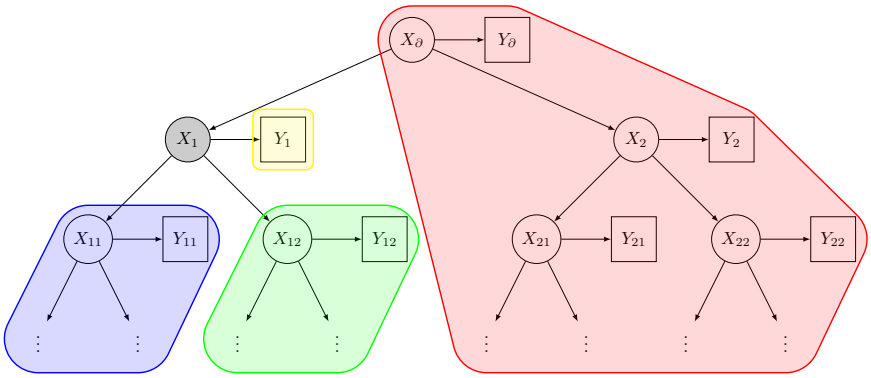


FIGURE 2.8 – Illustration de la propriété de Markov pour un HMT $(X, Y) = ((X_u, Y_u), u \in T)$ indexée par un arbre binaire complet T. En conditionnant sur X_1 (en gris), cela revient à rendre le processus (X, Y) indépendant entre les quatre composantes connexes de l'arbre de dépendance de la Figure 2.7 privé du sommet X_1 c'est-à-dire Y_1, $(X_{T(11)}, Y_{T(11)})$, $(X_{T(12)}, Y_{T(12)})$ et $(X_{T\setminus T(1)}, Y_{T\setminus T(1)})$ (respectivement en jaune, bleu, vert et rouge) sont indépendants.

Comme dans le cas des HMMs, tous les HMTs (X, Y) que nous considérons dans cette thèse sont dits *homogènes* (et nous ne le préciserons donc plus), c'est-à-dire que le processus caché X est une processus de Markov branchant homogène et que pour tout $n \in \mathbb{N}$ et tout $x \in \mathcal{X}$, nous avons :

$$\mathcal{L}(Y_n \,|\, X_n = x) = \mathcal{L}(Y_\partial \,|\, X_\partial = x).$$

Dans cette thèse, nous considérons le cas où les processus cachés et observés sont à valeurs dans des espaces métriques généraux $\mathcal{X}$ et $\mathcal{Y}$, respectivement.

Définition 2.3.1 (processus de Markov caché branchant ou HMT)**.** *Un processus stochastique* $(X, Y) = ((X_u, Y_u), u \in T)$ *est appelé un* processus de Markov caché branchant (homogène) *avec noyau de transitions (de probabilité)* Q *sur* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ *et* G *sur* $(\mathcal{X}, \mathcal{B}(\mathcal{Y}))$ *et mesure (de probabilité) initiale* ν *sur* $\mathcal{X}$ *si pour tout sous-arbre fini* A *de* T *contenant* ∂*, on a :*

$$\mathbb{P}(X_A \in \mathrm{d}x_A, Y_A \in \mathrm{d}y_A) = \nu(\mathrm{d}x_\partial) \prod_{u \in A\setminus\{\partial\}} Q(x_{\mathrm{p}(u)}; \mathrm{d}x_u) \prod_{u \in A} G(x_u; \mathrm{d}y_u).$$

Dans le Chapitre 5 (qui correspond à la prépublication [Wei24a]), nous adaptons la méthode de preuve de [DMR04] pour prouver la consistance forte et la normalité asymptotique du MLE dans le cas des HMTs (voir la Section 2.4.2 pour un résumé de ces résultats). Notons que pour l'approximation numérique du MLE pour les HMTs, l'algorithme "forward-backward" doit être remplacé par l'algorithme "upward-downward" développé dans [CNB98]. Voir aussi [DGG04] pour des formules récursives "upward-downward" alternatives qui peuvent gérer les problèmes de soupassement arithmétique (*"underflow"* en anglais) de manière implicite.

2.4 Contributions

Dans la Section 2.4.1, nous présentons de manière synthétique les résultats du Chapitres 3 (qui correspond à la prépublication [Wei24b]) : un théorème ergodique pour les processus de Markov branchants indexés par des arbres de formes arbitraires ainsi que l'étude de la forme d'arbre minimisant la variance de l'estimateur de moyenne empirique. Dans la Section 2.4.2, nous présentons de manière synthétique les résultats du Chapitres 4 (qui correspond à la prépublication [Wei24a]) : la consistance forte et la normalité asymptotique de l'estimateur du maximum de vraisemblance pour des modèles markoviens cachés indexés par des arbres binaires.

2.4.1 Contribution : Théorème ergodique pour les chaînes de Markov branchantes indexées par des arbres de formes arbitraires

Pour simplifier, dans cette introduction, nous nous restreignons au cas où le noyau de transition Q du processus de Markov branchant X est *ergodique* (resp. *uniformément ergodique*), c'est-à-dire que pour toute fonction continue bornée f sur $\mathcal{X}$, nous avons pour tout $x \in \mathcal{X}$ que $\lim_{n\to\infty} |Q^n f(x) - \langle \mu, f\rangle| = 0$ (resp. $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |Q^n f(x) - \langle \mu, f\rangle| = 0$) où μ est l'unique mesure invariante de Q.

Pour un sous-ensemble fini (non vide) $A \subset \mathbb{T}^\infty$ et une fonction f sur $\mathcal{X}$, nous définissons la moyenne empirique normalisée :

$$\bar{M}_A(f) = \frac{1}{|A|} \sum_{u\in A} f(X_u).$$

Notre objectif est d'étudier le comportement asymptotique de la moyenne empirique normalisée lorsque les moyennes sont effectuées sur une suite $(A_n)_{n\in\mathbb{N}}$ de sous-ensembles finis de $\mathbb{T}^\infty$ dont la taille tend vers l'infini. Par exemple, l'ensemble A_n peut être la n-ième génération G_n d'un arbre T, ou bien T_n, l'arbre T jusqu'à la génération n.

Dans ce but, nous avons besoin de certaines hypothèses géométriques sur la suite de sous-ensembles finis $(A_n)_{n\in\mathbb{N}}$, qui stipulent que les sommets sont éloignés les uns des autres avec une grande probabilité.

Hypothèse 1 (Géométrique)**.** *Soit $A_n \subset \mathbb{T}^\infty$ pour $n \in \mathbb{N}$ un ensemble fini (non vide), et soit U_n et V_n indépendants et uniformément distribués sur A_n. En notant d la distance de graphe sur $\mathbb{T}^\infty$, pour tout $k \in \mathbb{N}$, nous avons :*

$$\mathbb{P}\big(d(U_n, V_n) \le k\big) = \frac{1}{|A_n|^2} \sum_{u,v\in A_n} \mathbb{1}_{\{d(u,v)\le k\}} \underset{n\to\infty}{\longrightarrow} 0.$$

Soulignons que l'Hypothèse 1 implique que $\lim_{n\to\infty} |A_n| = \infty$.

Ensuite, nous devons soit supposer que Q est uniformément ergodique, soit que Q est ergodique et que la suite $(A_n)_{n\in\mathbb{N}}$ satisfait la condition suivante, stipulant que le dernier ancêtre commun de deux sommets est proche de la racine avec une forte probabilité. Rappelons que l'on note $h(u)$ la hauteur d'un sommet u et $u \wedge v$ l'ancêtre commun de deux sommets u et v (voir Section 1).

Hypothèse 2 (Ancestrale)**.** *Pour tout $n \in \mathbb{N}$, soient U_n et V_n indépendants et uniformément distribués sur A_n. La suite de variables aléatoires $(h(U_n \wedge V_n))_{n\in\mathbb{N}}$ est tendue, c'est-à-dire que pour tout $\varepsilon > 0$, il existe $k \in \mathbb{N}$ tel que $\mathbb{P}(h(U_n \wedge V_n) > k) < \varepsilon$ pour n suffisamment grand.*

Notons que les Hypothèses 1 et 2 sont similaires aux Hypothèses 2.(b) et 2.(a), respectivement, considérées dans [Ban19] dans le cas où A_n est la n-ième génération de l'arbre.

Remarque 2.4.1 (Quelques conditions suffisantes pour les Hypothèses 1 et 2, voir la Section 4.3 du Chapitre 4)**.** L'Hypothèse 1 est toujours satisfaite pour les arbres de Cayley (i.e. les arbres où chaque sommet a exactement D enfants pour $D \ge 2$) ainsi que pour les arbres à degré borné (voir le Lemme 4.3.1 du Chapitre 4). Les Hypothèses 1 et 2 sont satisfaites pour les arbres sphériquement symétriques (i.e. pour tout $n \in \mathbb{N}$, tous les sommets de G_n ont le même nombre d'enfants $D_n \ge 2$) lorsque $A_n = G_n$ (voir le Lemme 4.3.4 du Chapitre 4). Dans le Lemme 4.3.5 du Chapitre 4, nous prouvons que les Hypothèses 1 et 2 sont satisfaites pour les arbres de Bienaymé-Galton-Watson sur-critiques conditionnés à la non-extinction lorsque $A_n = G_n$ ou T_n.

Nous pouvons maintenant formuler le théorème ergodique pour les processus de Markov branchants indexés par des arbres de formes arbitraires. Dans la Section 4.2 du Chapitre 4, nous prouvons le Théorème 4.2.2 du Chapitre 4, une version plus générale de ce théorème.

Théorème 2.4.2 (Théorème ergodique pour les processus de Markov branchants indexés par des arbres de formes arbitraires)**.** *Soit $(A_n)_{n\in\mathbb{N}}$ une suite de sous-ensembles finis de $\mathbb{T}^\infty$ qui satisfait l'Hypothèse 1. Soit X un processus de Markov branchant indexé par $\mathbb{T}^\infty$ à valeurs dans $\mathcal{X}$ dont le noyau de transition Q est ergodique. Supposons de plus que soit Q est uniformément ergodique, soit $(A_n)_{n\in\mathbb{N}}$ satisfait l'Hypothèse 2. Alors, pour toute fonction continue bornée $f \in C_b(\mathcal{X})$ sur $\mathcal{X}$, nous avons :*

$$\bar{M}_{A_n}(f) = \frac{1}{|A_n|} \sum_{u\in A_n} f(X_u) \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} \langle \mu, f\rangle.$$

Remarque 2.4.3. Nous discutons la principale différence entre le Théorème 2.4.2 et la loi des grands nombres pour les processus de Markov branchants démontrée dans [Guy07]. Les résultats dans [Guy07] s'appliquent aux processus de Markov où les sommets enfants peuvent avoir des distributions non indépendantes lorsqu'on conditionne sur leur sommet parent, tandis que dans notre cas elles doivent être indépendantes. En échange, nos résultats permettent une plus grande flexibilité sur la forme de l'arbre généalogique de la population : par exemple, plus de flexibilité sur le nombre d'enfants de chaque nœud (y compris les arbres de Bienaymé-Galton-Watson avec un degré non borné), ou même permettre au nombre d'enfants d'un nœud d'augmenter avec le temps (e.g. le degré de la racine peut croître comme $\log n$ avec $|A_n| = n$, i.e. une condensation lente). De plus, dans nos résultats, la moyenne empirique peut être effectuée sur une grande variété de sous-ensembles de l'arbre, et pas seulement sur la n-ième génération.

Enfin, motivés par des considérations sur les méthodes de Monte-Carlo par chaînes de Markov, nous étudions dans la Section 4.4 du Chapitre 4 la variance de l'estimateur de moyenne empirique $\bar{M}_A(f)$ et sa dépendance par rapport à la forme de A. Nous effectuons un calcul exact de la variance dans le cas où le noyau de transition Q induit un opérateur compact auto-adjoint sur $L^2(\mu)$, ce qui prouve la proposition suivante concernant une comparaison de variances non-asymptotique.

Proposition 2.4.4 (Le graphe ligne a une variance minimale)**.** *Soit μ une mesure invariante pour Q, et supposons que le noyau de transition Q induit un opérateur compact auto-adjoint sur $L^2(\mu)$. Soit X un processus de Markov branchant sur $\mathbb{T}^\infty$ avec un noyau de transition Q et une distribution initiale $\nu = \mu$. Soit f une fonction non-constante dans $L^2(\mu)$.*

Nous avons que $\mathbb{E}\left[\bar{M}_A(f)\right] = \langle \mu, f \rangle$ pour tout sous-ensemble fini $A \subset \mathbb{T}^\infty$, et ainsi l'estimateur de moyenne empirique n'a pas de biais. De plus, le minimum de l'application $A \mapsto \mathrm{Var}(\bar{M}_A(f))$ parmi les sous-arbres de $\mathbb{T}^\infty$ de cardinal donné est atteint par le graphe ligne (i.e. la chaîne de Markov).

Remarquons que, comme Q est un noyau de transition (de probabilité), son spectre en tant qu'opérateur sur $L^2(\mu)$ est un sous-ensemble de $[-1, 1]$. De plus, notons que lorsque $f \in \mathrm{Ker}(Q) \oplus \mathrm{Ker}(Q - I)$, alors la valeur de $\mathrm{Var}(\bar{M}_T(f))$ ne dépend pas de la forme de l'arbre T. Nous prouvons également que lorsque $f \in \mathrm{Ker}(Q + I)$, alors la valeur de $\mathrm{Var}(\bar{M}_T(f))$ est minimale parmi les sous-arbres de taille n lorsque T a une 2-coloration bipartite équilibrée, et pour $n \geq 5$, le graphe ligne n'est pas le seul arbre avec une 2-coloration bipartite équilibrée.

Remarque 2.4.5. Dans la Proposition 4.1.4 du Chapitre 4 nous montrons également sous les hypothèses de la Proposition 2.4.4 le fait suivant : lorsque $n \geq 5$, le graphe ligne est le seul sous-arbre de taille n atteignant le minimum de l'application $A \mapsto \mathrm{Var}(\bar{M}_A(f))$ si et seulement si $f \notin \mathrm{Ker}(Q) \oplus \mathrm{Ker}(Q - I) \oplus \mathrm{Ker}(Q + I)$.

Ainsi, si l'on veut approximer $\langle \mu, f \rangle$, utiliser une chaîne de Markov branchante n'améliore pas le taux de convergence par rapport à une chaîne de Markov classique.

2.4.2 Contribution : Propriétés asymptotiques de l'estimateur de maximum de vraisemblance pour les modèles markoviens cachés indexés par les arbres binaires

Dans cette section et dans le Chapitre 5, les modèles markoviens cachés (défini en Section 2.3.2) sont tous indexés par l'arbre enraciné binaire complet infini que nous notons T.

Considérons un processus HMT dont la distribution est décrite par un paramètre qui est un vecteur θ, c'est-à-dire que le noyau de transition Q_θ entre les variables cachées et le noyau de transition G_θ des variables cachées aux variables observées dépendent tous deux de θ. Notre objectif est d'estimer le vrai paramètre $\theta^\star$ du processus HMT parmi un ensemble compact de paramètres possibles $\Theta \subset \mathbb{R}^d$, pour un certain entier d, en utilisant uniquement la connaissance du processus observé Y sur n générations de l'arbre. Notons que, comme nos hypothèses à venir impliquent des propriétés de perte de mémoire exponentielle uniforme pour la distribution initiale, nous ne pouvons pas essayer d'estimer la distribution initiale. Ainsi, nous supposons que la distribution de la variable cachée de la racine X_∂ est une mesure inconnue ζ qui ne dépend pas de θ. Notons par $\mathbb{P}_{\theta^\star,\zeta}$ la distribution de probabilité du HMT sous le vrai paramètre $\theta^\star$ lorsque la distribution initiale inconnue de X_∂ est ζ. Quand ζ est l'unique mesure invariante de Q_θ (i.e. dans le cas stationaire), nous notons $\mathbb{P}_{\theta^\star}$ à la place de $\mathbb{P}_{\theta^\star,\zeta}$.

Pour estimer le vrai paramètre $\theta^\star$ du HMT, nous utilisons l'estimateur du maximum de vraisemblance (MLE). Nous travaillons avec la vraisemblance conditionnée à l'état caché de la racine X_∂. La raison est que le calcul de la distribution stationnaire du processus joint (X, Y), et donc également de la vraie vraisemblance, n'est pas calculable dans les applications typiques. Notons que l'idée de conditionner sur l'état caché initial a déjà été utilisée dans [DMR04] pour les HMMs avec la même motivation, et que le conditionnement sur les observations initiales dans les séries temporelles remonte au moins à [MW43]. Rappelons que T_n dénote l'arbre T jusqu'à la n-ième génération incluse. Ainsi, pour toute valeur $x \in \mathcal{X}$, nous notons par $\ell_{n,x}(\theta)$ la log-vraisemblance sous le paramètre θ du processus observé $(Y_u, u \in T_n)$ jusqu'à la n-ième génération de l'arbre T conditionnellement à $X_\partial = x$ (voir (5.7) à la page 125 du Chapitre 5 pour la définition exacte). Ensuite, pour toute valeur $x \in \mathcal{X}$, nous définissons le MLE $\hat{\theta}_{n,x}$ comme le maximiseur de $\ell_{n,x}$ sur Θ (voir (5.33) à la page 135 du Chapitre 5 pour la définition exacte).

Notre objectif est d'étudier les propriétés asymptotiques du MLE. Nous prouvons la consistance forte et la normalité asymptotique du MLE dans le cas stationnaire respectivement dans les Sections 5.3 et 5.4 du Chapitre 5. Nous étendons ensuite ces résultats au cas non stationnaire dans la Section 5.5 du Chapitre 5. Dans nos résultats, l'espace des états cachés $\mathcal{X}$ et l'espace des états observés $\mathcal{Y}$ sont tous deux des espaces métriques généraux. Nous prouvons nos résultats sous les mêmes hypothèses utilisées que dans [DMR04] et dans [CMR05, Chapitre 12] pour les HMMs, avec les hypothèses d'intégrabilité L^1 et L^2 remplacées respectivement par des hypothèses d'intégrabilité L^2 et L^4, pour accommoder les hypothèses plus fortes nécessaires dans les théorèmes ergodiques pour les chaînes de Markov branchantes. Voir la Remarque 2.4.11 ci-dessous pour une discussion sur les principales différences entre le cas des HMMs tel que dans [DMR04, CMR05] et le cas des HMTs que nous développons dans le Chapitre 5.

Nous commençons par énoncer la consistance forte du MLE sous des hypothèses standard pour les HMMs. Suivant [DMR04], nous supposons un modèle entièrement dominé, c'est-à-dire que les noyaux de transition Q_θ et G_θ admettent des densités q_θ et g_θ par rapport à des mesures communes λ et μ, respectivement (voir l'Hypothèse 6 dans le Chapitre 5). Nous supposons également (voir l'Hypothèse 7 dans le Chapitre 5) :

$$0 < \sigma^- \leq \inf_{x,x' \in \mathcal{X}} q_\theta(x, x') \leq \sup_{x,x' \in \mathcal{X}} q_\theta(x, x') \leq \sigma^+ < \infty. \tag{2.4}$$

Cette hypothèse est plutôt forte car elle impose une connexion complète pour l'espace caché, voir [KS19] pour une extension de la méthode dans [DMR04] pour les HMMs où q_θ peut prendre des valeurs nulles. Néanmoins, cette hypothèse implique les propriétés de pertes de mémoire exponentielles uniformes avec un taux de mélange $\rho := 1 - \sigma^-/\sigma^+$ de la distribution initiale conditionnée aux observations $(Y_u, u \in T_n)$. Les autres hypothèses sont des hypothèses de régularité plus standards pour les densités q_θ et g_θ (voir les Hypothèses 7-10 dans le Chapitre 5), et l'identifiabilité du modèle. Nous pouvons maintenant énoncer la consistance forte du MLE sous ces hypothèses, voir Théorèmes 5.3.11 et 5.5.1 du Chapitre 5 pour les énoncés précis dans les cas stationnaire et non stationnaire, respectivement.

Théorème 2.4.6 (Consistance forte du MLE)**.** *Sous ces hypothèses de modèle entièrement dominé dont les densités vérifient* (2.4) *et d'autres hypothèses de régularité plus standards, et sous l'hypothèse d'identifiabilité du modèle, pour tout $x \in \mathcal{X}$, le MLE $\hat{\theta}_{n,x}$ est fortement consistant, c'est-à-dire que la suite $(\hat{\theta}_{n,x})_{n\in\mathbb{N}}$ converge $\mathbb{P}_{\theta^\star,\zeta}$-presque sûrement vers le vrai paramètre $\theta^\star \in \Theta$.*

Pour prouver la normalité asymptotique du MLE, en plus des hypothèses utilisées dans le Théorème 2.4.6, nous avons besoin d'hypothèses d'existence et de régularité pour le gradient et la hessienne des densités de transition q_θ et g_θ (voir les Hypothèses 11-13 du Chapitre 5). Notons par $\mathcal{I}(\theta^\star)$ la matrice d'information de Fisher limite du modèle (voir (5.54) à la page 144 du Chapitre 5 pour une définition précise). La preuve de la normalité asymptotique dans le cas non stationnaire est une extension du cas stationnaire. La preuve de la normalité asymptotique dans le cas stationnaire suit un argument standard pour la normalité asymptotique du MLE qui repose sur le Théorème 2.4.6 et les Théorèmes 2.4.7 et 2.4.8 ci-dessous.

Le théorème suivant, que nous prouvons uniquement dans le cas stationnaire, énonce que le score normalisé $|T_n|^{-1/2} \nabla_\theta \ell_{n,x}(\theta^\star)$ présente des fluctuations normales asymptotiques avec pour matrice de covariance $\mathcal{I}(\theta^\star)$, voir le Théorème 5.4.3 du Chapitre 5 pour l'énoncé précis. Notons que l'hypothèse supplémentaire dans le Théorème 2.4.7 (qui n'est pas présente dans le cas des HMMs), que $\rho < 1/\sqrt{2}$ pour le taux de mélange ρ du processus HMT, provient des bornes d'approximation utilisées dans la preuve de ce théorème. Voir Remarque 2.4.10 ci-dessous pour une discussion sur cette condition sur ρ.

Théorème 2.4.7 (Normalité asymptotique du score normalisé)**.** *Sous les hypothèses du Théorème 2.4.6, et des hypothèses d'existence et de régularité du gradient et de la hessienne des densités de transitions du modèle (voir les Hypothèses 11-13 du Chapitre 5), et sous l'hypothèse que $\rho < 1/\sqrt{2}$ pour le taux de mélange ρ du processus HMT, dans le cas stationnaire nous avons :*

$$|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)) \quad \textit{sous } \mathbb{P}_{\theta^\star},$$

où $\mathcal{N}(0, M)$ désigne la distribution gaussienne centrée avec matrice de covariance M.

Le théorème suivant énonce la convergence locale uniforme $\mathbb{P}_{\theta^\star,\zeta}$-p.s. de l'information observée normalisée $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta)$ vers la matrice d'information de Fisher $\mathcal{I}(\theta^\star)$, voir les Théorèmes 5.4.6 et 5.5.2 du Chapitre 5 pour les énoncés précis dans les cas stationnaire et non stationnaire, respectivement. Notons que dans ce théorème, nous avons besoin de l'hypothèse plus forte $\rho < 1/2$ pour le taux de mélange ρ du processus HMT car nous utilisons des bornes d'approximation plus restrictives dans la preuve de ce théorème que celles utilisées dans la preuve du Théorème 2.4.7.

Théorème 2.4.8 (Convergence de l'information observée normalisée)**.** *Sous les hypothèses du Théorème 2.4.7 sur le modèle HMT et sous l'hypothèse que $\rho < 1/2$ pour le taux de mélange ρ du processus HMT, pour tout $x \in \mathcal{X}$, nous avons :*

$$\lim_{\delta\to 0}\lim_{n\to\infty} \sup_{\theta\in\Theta\,:\,\|\theta-\theta^\star\|\le\delta} \left\| -|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta) - \mathcal{I}(\theta^\star)\right\| = 0 \quad \mathbb{P}_{\theta^\star,\zeta}\text{-}p.s.$$

En particulier, en combinant les Théorèmes 2.4.6 et 2.4.8, nous obtenons que l'information observée normalisée $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\hat\theta_{n,x})$ évaluée au MLE $\hat\theta_{n,x}$ est un estimateur fortement consistant de la matrice d'information de Fisher $\mathcal{I}(\theta^\star)$.

Comme annoncé ci-dessus, en suivant un argument standard pour la normalité asymptotique du MLE, les Théorèmes 2.4.6, 2.4.7 et 2.4.8 impliquent le théorème suivant qui énonce que le MLE présente des fluctuations normales asymptotiques avec pour matrice de covariance $\mathcal{I}(\theta^\star)^{-1}$. Voir les Théorèmes 5.4.7 et 5.5.5 du Chapitre 5 pour les énoncés précis dans les cas stationnaire et non stationnaire, respectivement.

Théorème 2.4.9 (Normalité asymptotique du MLE)**.** *Sous les hypothèses du Théorème 2.4.7 sur le modèle HMT que $\theta^\star$ est un point intérieur de Θ, que la matrice d'information de Fisher $\mathcal{I}(\theta^\star)$ est non-singulière, et sous l'hypothèse que $\rho < 1/2$ pour le taux de mélange ρ du processus HMT, nous avons la convergence en distribution suivante :*

$$|T_n|^{1/2}\big(\hat\theta_n - \theta^\star\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1}) \quad \textit{sous } \mathbb{P}_{\theta^\star,\zeta},$$

où $\mathcal{N}(0, M)$ désigne la distribution gaussienne centrée avec matrice de covariance M.

Notons que l'argument standard utilisé dans la preuve du Théorème 2.4.9 implique que nous avons la convergence jointe en distribution suivante :

$$\Big(|T_n|^{1/2}\big(\hat\theta_n - \theta^\star\big),\, |T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star)\Big) \xrightarrow[n\to\infty]{(d)} \big(\mathcal{I}(\theta^\star)^{-1/2}\,G,\, \mathcal{I}(\theta^\star)^{1/2}\,G\big) \quad \text{sous } \mathbb{P}_{\theta^\star},$$

où G est une variable aléatoire gaussienne distribuée selon $\mathcal{N}(0, I_d)$ avec I_d la matrice identité de dimension $d \times d$, et $\mathcal{I}(\theta^\star)^{1/2}$ est une matrice racine carrée de la matrice $\mathcal{I}(\theta^\star)$.

La remarque suivante est une discussion sur les conditions concernant le taux de mélange ρ du processus HMT qui apparaît dans les Théorèmes 2.4.9, 2.4.7 et 2.4.8.

Remarque 2.4.10 (Sur les conditions sur le taux de mélange ρ)**.** Notons que dans le théorème central limite pour les chaînes de Markov branchantes, trois régimes avec des comportements asymptotiques différents (et des termes de normalisation différents) pour $\rho < 1/\sqrt{2}$, $\rho = 1/\sqrt{2}$ et $\rho > 1/\sqrt{2}$ ont été observés dans [BPD22a], correspondant à une compétition entre le taux de mélange ergodique ρ et le taux de branchement 2 dans l'arbre binaire T, voir aussi [Ath69, BPDG14, BPD22b]. Cependant, la condition sur ρ disparaît lorsque l'on considére des incréments de martingales dans le théorème limite centrale pour les chaînes de Markov branchantes, voir [Guy07, BDSG09, DM10].

Dans notre cas, la condition $\rho < 1/\sqrt{2}$ sur le taux de mélange ρ qui apparaît dans le Théorème 2.4.7 est due aux bornes de couplage et au regroupement des termes utilisés dans la preuve du Lemme 5.4.2 du Chapitre 5 (les bornes majorantes à la fin de la preuve n'ajoutent qu'un facteur multiplicatif constant). Le fait de savoir si la convergence dans le Théorème 2.4.7 est également valable lorsque $\rho \geq 1/\sqrt{2}$ reste une question ouverte. Néanmoins, notons que la preuve du Théorème 2.4.7 repose sur la décomposition du score $\nabla_\theta \ell_{n,x}(\theta)$ en une somme d'incréments de martingales, ce qui pourrait indiquer qu'une convergence est possible pour $\rho \geq 1/\sqrt{2}$.

De plus, la condition plus stricte $\rho < 1/2$ sur le taux de mélange ρ qui apparaît dans le Théorème 2.4.8, et donc dans le Théorème 2.4.9, est due aux bornes de couplage du Lemme 5.4.16 et au regroupement des termes utilisés dans la preuve du Lemme 5.4.17 du Chapitre 5 (les bornes supérieures dans le reste de la preuve n'ajoutent qu'un facteur multiplicatif constant). Le fait de savoir si la convergence dans les Théorèmes 2.4.8 et 2.4.9 est également valable avec $\rho \geq 1/2$ reste une question ouverte. Notons également que la condition $\rho < 1/2$ est utilisée lors de la démonstration de l'extension du Théorème 2.4.9 au cas non stationnaire pour construire un couplage entre un processus HMT stationnaire et un processus HMT non stationnaire, voir le Lemme 5.5.3 du Chapitre 5.

Dans la remarque suivante, nous discutons des principales différences entre le cas des HMMs tel que dans [DMR04, CMR05] et le cas des HMTs que nous développons dans le Chapitre 5.

Remarque 2.4.11 (Sur les principales différences avec le cas des HMMs)**.** Dans les cas des HMMs et des HMTs, l'étude de la log-vraisemblance repose sur sa décomposition en une somme d'incréments, puis sur l'extension dans le cas stationnaire du « passé » vu par chaque variable. Cependant, alors que le « passé » étendu ne s'étend que vers l'arrière dans le cas des HMMs, le « passé » étendu dans le cas des HMT est un sous-arbre qui s'étend également latéralement en raison des topologies différentes entre la ligne $\mathbb{Z}$ et l'arbre binaire, voir la Figure 2.9 pour une illustration. Voir également les Sections 5.2.4 et 5.3.1 du Chapitre 5 pour la définition de ces « passé » et « passé » étendu. De plus, en raison de l'énumération des sommets de l'arbre dans l'ordre de parcours en largeur (en anglais *"breadth first search order"*), ces « passé » étendus n'ont pas les mêmes « formes » pour tous les sommets, voir la Section 5.2.4 du Chapitre 5. Notons également que le « passé » étendu infini d'un sommet repose sur une « épine dorsale » aléatoire infinie de descendants gauche / droite (voir la Figure 2.10), ce qui ajoute une source d'aléatoire supplémentaire à la « forme » du « passé ».

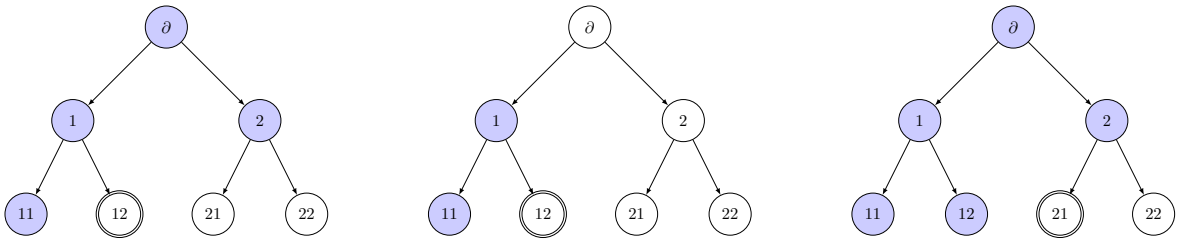


FIGURE 2.9 – Illustration des sous-arbres de « passé » et de « passé » tronqué dans le cas du HMT. Les sommets en bleus sont ceux faisant partie du « passé » (resp. du « passé » tronqué) du sommet dans un double cercle. De gauche à droite, en utilisant la notation de Neveu, nous avons le sous-arbre de « passé » du sommet 12, le sous-arbre de « passé » tronqué de hauteur 1 du sommet 12, et le sous-arbre de « passé » du sommet 21. Notons que les sous-arbres de « passés » des sommets 12 et 21 n'ont pas la même « forme » : ces sous-arbres n'ont pas le même nombre de sommets bien que les deux sommets 12 et 21 sont dans la même génération.

En outre, contrairement au cas des HMMs, l'étalement latéral du « passé » de chaque sommet dans le cas des HMTs implique que les incréments de log-vraisemblance avec des « passé » étendus infinis ne forment pas un processus de Markov branchant. Pour cette raison, nous devons travailler avec des incréments de log-vraisemblance dont le « passé » est limité à une hauteur de sous-arbre commune fixée, et ne s'étendre au « passé » infini qu'à la limite. Pour prouver la convergence pour des sommes d'incréments de log-vraisemblance avec des « passé » tronqués qui ont des formes différentes, nous devons développer de nouveaux théorèmes ergodiques pour les chaînes de Markov branchantes et des fonctions dépendant du voisinage (i.e. pour chaque sommet, la fonction dépend des variables du voisinage de ce sommet), voir la Section 5.2.4 et l'Annexe 5.A du Chapitre 5.

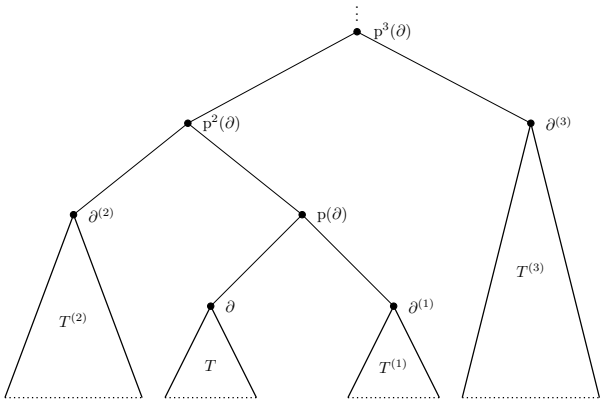


FIGURE 2.10 – Illustration de la construction de l'« épine dorsale » aléatoire infinie de descendants gauche / droite de la racine ∂. On rajoute une suite infinie d'ancêtres $(\mathrm{p}^k(\partial))_{k\geq 1}$ de la racine ∂. Puis, pour chaque $k \in \mathbb{N}$, le sommet $\mathrm{p}^k(\partial)$ est l'enfant gauche (resp. droit) de $\mathrm{p}^{k+1}(\partial)$ avec probabilité 1/2. Enfin, on greffe une copie $T^{(k)}$ de racine $\partial^{(k)}$ sur $\mathrm{p}^k(\partial)$ pour former en arbre binaire complet.

Dans la preuve de la normalité asymptotique du score normalisé, le score est décomposé en une somme d'incréments de martingales qui ne sont plus stationnaires dans le cas des HMTs à cause du fait que les « passés » des sommets ont des formes différentes. Ainsi, pour appliquer le théorème central limite pour les martingales, nous devons d'abord vérifier la convergence pour les variations quadratiques de la suite d'incréments de martingales et la condition de Lindeberg. De plus, le calcul des bornes d'approximation pour les incréments utilisés pour décomposer le score et l'information observée est plus complexe et impose des conditions sur la valeur du taux de mélange ρ, comme déjà discuté dans la Remarque 2.4.10. Cela implique également que le schéma de preuve de la convergence de la matrice d'information observée doit être modifié car nous ne pouvons pas avoir une convergence presque sûre pour tous les incréments simultanément, et nous devons nous appuyer sur la convergence L^2 à la place.

Enfin, comme discuté dans la Section Section 2.3.1, les résultats pour les HMMs dans [DMR04] couvrent les processus avec auto-régression (rappelons, c'est-à-dire lorsque, conditionnellement à la chaîne de Markov cachée, le processus observé est une chaîne de Markov inhomogène d'ordre s pour un certain $s \in \mathbb{N}$). Nos résultats pour les HMTs sont énoncés pour des processus sans auto-régression. Cependant, comme notre approche adapte le schéma de preuve de [DMR04], notons qu'avec des modifications simples de nos preuves, nous pourrions permettre l'auto-régression dans les processus HMTs.

2.5 Perspectives

Agrégats et division cellulaire

Dans la vie d'une cellule, il arrive que certaines protéines produites par la cellule ne se plient pas correctement. Ces protéines deviennent alors inopérantes et finissent par s'assembler formant des agrégats qui « parasitent » la cellule réduisant son activité et sa capacité de reproduction [MK23, WHF$^+$23]. Il a été observé que lors de la division cellulaire, l'une des deux cellules filles hérite de la plupart des agrégats de la cellule mère [LMD$^+$08, TMB10]. Ce phénomène de ségrégation asymétrique permettant de créer des cellules saines qui assurent la survie de l'espèce.

Comme mentionné en Section 2.2.2, des processus de Markov branchants autorisant une dépendance de la distribution des variables de deux sommets frères conditionnellement à la valeur de leur sommet parent, et appelés *processus de Markov bifurquants*, ont été considérés dans [Guy07, DM10] pour modéliser la division cellulaire asymétrique. Un *processus de Markov bifurquant* indexé par l'arbre binaire complet T est donc défini par une mesure (de probabilité) initiale ν sur $\mathcal{X}$ et un noyau de transition P sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}^2))$ pour tout $n \in \mathbb{N}$ par :

$$\mathbb{P}(X_{T_n} \in \mathrm{d}x_{T_n}) = \nu(\mathrm{d}x_\partial) \, \Pi_{u \in T_{n-1}} P(x_u; \mathrm{d}x_{u1}, \mathrm{d}x_{u2}).$$

Une généralisation possible à étudier des résultats du Chapitre 5 sur le MLE du HMT pourrait donc être le cas où le processus caché X est remplacé par un processus de Markov bifurquant. Ensuite, comme les

agrégats ne sont observables que de manières indirectes par l'intermédiaire de l'activité cellulaire, il serait intéressant de voir si ce phénomène peut être modéliser par un HMT où le processus cachés représente la naissance des agrégats et leur héritage par les cellules filles. Cela donnerait également une motivation biologique pour un modèle avec lequel faire des simulations numériques de la convergence du MLE.

Généralisations des résultats du Chapitre 5 sur le MLE du HMT pour d'autres géométries d'arbres généalogiques

Une autre voie à étudier serait de garder les processus HMTs comme définis en Section 2.3.2, mais de généraliser les résultats du Chapitre 5 au cas où le processus est indexé par un autre arbre (possiblement aléatoire) que l'arbre binaire complet, par exemple l'arbre (aléatoire) de Bienaymé-Galton-Watson. Une extension des résultats de [Guy07, DM10] pour des processus de Markov branchants où la distribution du nombre de descendants de chaque individu dépend des ses traits, ce qui implique que la forme de l'arbre dépend du processus markovien, a été étudiée dans [Ban19]. Ainsi, une troisième généralisation possible les résultats du Chapitre 5 serait de considérer un cadre similaire à [Ban19] où la forme de l'arbre dépend du processus HMT.

Détection de communauté pour le modèle à blocs stochastiques pondérés ou décorés (en anglais, weighted SBMs et labeled SBMs)

Comme évoqué en introduction de ce chapitre et en Section 2.1.3, le problème de reconstruction sur les arbres avec des poids ou décorations sur les arêtes est lié au problème de détection de communauté pour le modèle à blocs stochastiques pondérés ou décorés présenté en ouverture à la fin du Chapitre 1, voir [HLM12, LMX15]. Ainsi, les modèles markoviens cachés indexés par des arbres pourraient être un outil pour étudier le seuil de faisabilité pour la détection de communauté sur des modèles à blocs stochastiques pondérés avec des poids dans des espaces probabilisés généraux.

Limites locale et d'échelle de l'arbre couvrant minimal (MST) d'une suite de graphes pondérés aléatoires

Revenons également sur les limites locale et d'échelle de l'arbre couvrant minimal (MST) d'une suite de graphes pondérés aléatoires présentées en ouverture à la fin du Chapitre 1. En suivant ce qui a été fait pour l'arbre couvrant uniforme dans [HNT18, ANS24], on s'attend à ce que la construction de l'arbre limite pour les limites locale et d'échelle du MST puissent s'interpréter comme un modèle markovien caché indexé par un arbre faisant intervenir un graphon de probabilités pour les distributions des poids des arêtes.

Deuxième partie

Résultats

Chapitre 3

Probability-graphons : Limits of large dense weighted graphs

The material for this chapter has been released in [ADW23] and is currently under review. Note that Appendix 3.A gathers proofs that are straightforward modifications of proofs existing in the literature and were not included in [ADW23] for brevity.

Abstract: *We introduce probability-graphons which are probability kernels that generalize graphons to the case of weighted graphs. Probability-graphons appear as the limit objects to study sequences of large weighted graphs whose distribution of subgraph sampling converge. The edge-weights are taken from a general Polish space, which also covers the case of decorated graphs. Here, graphs can be either directed or undirected. Starting from a distance d_{m} inducing the weak topology on measures, we define a cut distance on probability-graphons, making it a Polish space, and study the properties of this cut distance. In particular, we exhibit a tightness criterion for probability-graphons related to relative compactness in the cut distance. We also prove that under some conditions on the distance d_{m}, which are satisfied for some well-know distances like the Prohorov distance, and the Fortet-Mourier and Kantorovitch-Rubinstein norms, the topology induced by the cut distance on the space of probability-graphons is independent from the choice of d_{m}. Eventually, we prove that this topology coincides with the topology induced by the convergence in distribution of the sampled subgraphs.*

2020 Mathematics Subject Classification— 05C80, 60B10

Key words and phrases— random graphs, stochastic networks, graphons, probability-graphons, dense graph limits, weighted graphs, decorated graphs

3.1 Introduction

3.1.1 Motivation and literature review

Networks appear naturally in a wide variety of context, including for example: biological networks [BS02, Lew09], epidemics processes [DM10, KMS17], electrical power grids [AAN04] and social networks [AB02, New03]. Most of those problems involve large dense graphs, that is graphs that have a large number of vertices and a number of edges that scales as the square of the number of vertices. Those graphs are too large to be represented entirely in the targeted applications. The idea is then to go from a combinatorial representation given by the graph to an infinite continuum representation.

In the case of unweighted graphs (i.e. graphs without edge-weights), a theory was developed to study the asymptotic behaviour of large dense graphs, with the limit objects being the so-called *graphons*. The properties of graphons were studied in a series of articles started by [LS06, FLS06, BJR10, BCL$^+$08, BCL$^+$12]. We shall refer to the monograph [Lov12] which exposes in details the theory of graphons developed in this series of articles. Graphons can be used to define models of random graphs with latent vertex-type variables (called W-random graphs) generalizing the Erdös-Rényi graph and the stochastic block model (SBM). The space of graphons can be equipped with the so-called *cut distance*, making it a compact space, and whose topology is that of the convergence in distribution for all sampled subgraphs, or equivalently of the convergence for subgraph homomorphism densities.

In recent years, graphons have been used in several application context: non-parametric estimation methods and algorithms for massive networks [BC17], SIS epidemic models [DDZ22], the study of transferability properties for Graph Neural Networks [KV23]. Furthermore, there has been recent developments in the study of mean-field systems using graphons: stochastic games and their Nash equilibria [LS22], opinion dynamic on a graphon [AN22], cooperative multi-agent reinforcement learning [HWYZ23], to cite a few.

However, most real-world phenomenon on the above networks involve weighted networks, where each edge in the graph carries additional information such as intensity or frequency of interaction, or transfer capacity.

There exists many models of random weighted graphs. For example configuration models with edges having independent exponential weights have been considered in [BVDHH10, ADL13, AL15], see also [Gar09, HG13] where the distribution of the weight of an edge depends on the types of its end-points. Random geometric graphs with vertices and edges having independent Gaussian weights have been considered in [AMGM18].

Weighted SBMs (sometimes also called labeled SBMs), in which each edge independently receives a random weight whose distribution depends on the community labels of its end-points, have been studied to solve community detection in [LMX13] (see also [XML14] for more general models where vertex-labels come from a compact space), and exact community recovery in [JL15], and to get bounds on the number of misclassified vertices in [YP16, XJL20]. Note that weighted SBMs correspond to a special case of the probability-graphons we study in this article where the space of vertex-labels is finite (they correspond to the stepfunction probability-graphons we define in Section 3.3).

Concomitantly to our work, in [AD23], the authors studied mean-field equations on large real-weighted graphs modeling interactions with a probability kernel from $[0,1]^2$ to $\mathcal{M}_1(\mathbb{R})$ the set of probability measures on $\mathbb{R}$, but they did not study the topological properties of the set of those probability kernels. Recently, in [HV23], the authors studied the limit of the total weight of the minimum spanning tree (MST) for a sequence of random weighted graphs. Following what has been done for the uniform spanning tree in [HNT18, ANS24], one expects the local and scaling limits of the MST to be directly constructed from the limit of the random weighted graphs.

Motivated by those examples, we shall consider *probability-graphons* as possible limits of large weighted graphs; they are defined as maps from $[0,1]^2$ to the space of probability measures $\mathcal{M}_1(\mathbf{Z})$ on a Polish space $\mathbf{Z}$. When $\mathbf{Z}$ is compact, this question has been considered in [LS10] and in [Lov12, Section 17.1] using convergence of homomorphism densities of subgraphs decorated with real functions defined on $\mathbf{Z}$, see also [KR11] on multigraphs where $\mathbf{Z} = \mathbb{N}$, but the metric properties of the set of probability-graphons $\mathcal{W}_1$ have only been established when $\mathbf{Z}$ is finite, see [FOSU16]. The work [KLS22] is an extension of [LS10] where $\mathcal{M}_1(\mathbf{Z})$ is replaced by the dual space $\mathcal{Z}$ of a separable Banach space $\mathcal{B}$. As $\mathcal{M}_1(\mathbf{Z})$ is a subset of the dual of $C_b(\mathbf{Z})$, this approach covers our setting when $C_b(\mathbf{Z})$ is separable, that is, $\mathbf{Z}$ compact (see Section 3.2 below). The norm introduced on the space of $\mathcal{Z}$-valued graphons therein implies the convergence of homomorphisms densities of $\mathcal{B}$-decorated sub-graphs, however there is no equivalence *a priori*.

In this paper we study the topological properties of the space of probability-graphons $\mathcal{W}_1$ when $\mathbf{Z}$ is a general Polish space: the space $\mathcal{W}_1$ is a Polish topological space and we give "natural" cut distances on $\mathcal{W}_1$ which are complete. One of the main difficulty is that the space of probability measures $\mathcal{M}_1(\mathbf{Z})$ can be endowed with many distances which induce the topology of weak convergence, each of them giving rise to a different cut distance on $\mathcal{W}_1$. We prove that the topology induced on $\mathcal{W}_1$ does not depend on the initial choice of the distance on $\mathcal{M}_1(\mathbf{Z})$, provided this distance satisfies some simple general conditions, and in particular when this distance is quasi-convex (a property that generalizes the convexity of a norm). However, we stress that not all of these cut distances are complete. We also check that this topology characterizes the convergence in distribution of the sampled subgraphs with random weights on the edges or equivalently the convergence of the homomorphism densities of $C_b(\mathbf{Z})$-decorated subgraphs. Similarly to the graphon setting, we prove the convergence in distribution of large sampled weighted subgraphs from a probability-graphon W to itself. We also provide a tightness criterion for studying the convergence of weighted graphs towards probability-graphons; this criterion generalizes the tightness condition in [KR11] for multigraphs where $\mathbf{Z} = \mathbb{N}$.

In conclusion, we believe that the unified framework developed here is easy-to-work-with and will allow to use probability-graphons to study large (random) weighted graphs.

3.1.2 New contribution

Through the article, *measure* will always be used to denote a positive measure.

Definition of probability-graphons

In this article, we define an analogue of graphons for weighted graphs, which we call *probability-graphons*, and study their properties. To avoid any confusion, in the rest of the article we say *real-valued graphons* instead of graphons. We consider the general case where weighted graphs take their edge-weights in a Polish space $\mathbf{Z}$ (e.g. $\mathbb{Z}$, $\mathbb{R}$ or $\mathbb{R}^d$), which thus also covers the case of decorated graphs, multi-graphs (graphs with possibly multiple edges between two vertices) and dynamical graphs (where edge-weights evolve over time).

We define a *probability-graphon* as a probability kernel $W : [0,1]^2 \to \mathcal{M}_1(\mathbf{Z})$, where $\mathcal{M}_1(\mathbf{Z})$ is the space of probability measures on $\mathbf{Z}$. A probability-graphon can be interpreted as follows: for two "vertex type" x and y in $[0,1]$, the weight z of an edge between two vertices of type x and y is distributed as the probability measure $W(x,y;\mathrm{d}z)$. In particular, the special case $\mathbf{Z} = \{0,1\}$ allows to recover real-valued graphons: as any real-valued graphon $w : [0,1]^2 \to [0,1]$ can be represented as a probability-graphon $W(x,y;\cdot) = w(x,y)\delta_1 + (1-w(x,y))\delta_0$, where δ_z denote the Dirac mass located at z. Let us mention that it is possible to define the probability-graphons on a more general probability space $(\Omega, \mathcal{A}, \mu)$ than $[0,1]$ for the vertex-types, see Remark 3.3.4 for details. In this article, we also define and study the properties of *signed measure-valued kernels* which are bounded (in total mass/total variation norm) measurable functions $W : [0,1]^2 \mapsto \mathcal{M}_\pm(\mathbf{Z})$ whose values are signed measures, but for brevity we mainly focus on probability-graphons in this introduction.

As probability-graphons are measurable functions, we identify probability-graphons that are equal for almost every $(x,y) \in [0,1]^2$, and we denote by $\mathcal{W}_1$ the space of probability-graphons. Moreover, as we consider weighted graphs that are unlabeled (that is vertices are unordered), we need to consider probability-graphons up to "relabeling": for a measure-preserving map $\varphi : [0,1] \to [0,1]$ (relabeling map for probability-graphons), we define $W^\varphi(x,y;\cdot) = W(\varphi(x),\varphi(y);\cdot)$; we say that two probability-graphons are weakly isomorphic if there exists measure-preserving maps $\varphi, \psi : [0,1] \to [0,1]$ such that $U^\varphi = W^\psi$ for a.e. $(x,y) \in [0,1]^2$. We denote by $\widetilde{\mathcal{W}}_1$ the space of probability-graphons where we identity probability-graphons that are weakly isomorphic.

We can always assume that weighted graphs are complete graphs by adding all missing edges and giving them a weight/decoration ∂ which is a cemetery point added to $\mathbf{Z}$. Any weighted graph G can be represented as a probability-graphon W_G in the following way: denote by n the number of vertices of G and divide the unit interval $[0,1]$ into n intervals $I_1, \cdots, I_n$ of equal lengths, then W_G is defined for $(x,y) \in I_i \times I_j$ as $W_G(x,y;\cdot) = \delta_{M(i,j)}$, where $M(i,j)$ is the weight on the edge (i,j) in G. Note that weighted graphs can be either directed or undirected, in the case of undirected weighted graphs their limit objects are symmetric probability-graphons, that is probability-graphons W such that $W(x,y;\cdot) = W(y,x;\cdot)$.

The cut distance for probability-graphons and its properties

While there is a usual distance on the field of reals $\mathbb{R}$, this is not the case for probability measures, measures or signed measures endowed with the weak topology. Some commonly used distances include the Prohorov distance $d_{\mathcal{P}}$ which can be defined on measures, and the Kantorovitch-Rubinstein norm $\|\cdot\|_{\mathrm{KR}}$ (sometimes also called the bounded Lipschitz norm) and the Fortet-Mourier norm $\|\cdot\|_{\mathrm{FM}}$ defined on signed measures but metrizing the weak topology on measures. (Note that in general the weak topology is not metrizable on signed measures, see Section 3.2 below.) We also use a norm $\|\cdot\|_{\mathcal{F}}$ based on a convergence determining sequence $\mathcal{F} \subset C_b(\mathbf{Z})$. See Section 3.3.8 for definition of those distances. To define an analogue of the cut norm for probability-graphons, we first need to choose a distance d_{m} that metrizes the weak topology on the space of sub-probability measures $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (i.e. measures with total mass at most 1); we then define the *cut distance* $d_{\square,\mathrm{m}}$ for probability-graphons as:

$$d_{\square,\mathrm{m}}(U,W) = \sup_{S,T\subset[0,1]} d_{\mathrm{m}}\Big(U(S\times T;\cdot), W(S\times T;\cdot)\Big),$$

where the supremum is taken over all measurable subsets S and T of $[0,1]$, and where $W(S\times T;\cdot) = \int_{S\times T} W(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y$ is a sub-probability measure and similarly for U. Moreover, if the distance d_{m} is

derived from a norm N_{m} defined on the space of signed measures $\mathcal{M}_{\pm}(\mathbf{Z})$, then the cut distance $d_{\square,\mathrm{m}}$ derives from the *cut norm* $N_{\square,\mathrm{m}}$ defined on signed measure-valued kernels:

$$N_{\square,\mathrm{m}}(W) = \sup_{S,T\subset[0,1]} N_{\mathrm{m}}\Big(W(S\times T;\cdot)\Big).$$

We then define the unlabeled *cut distance* $\delta_{\square,\mathrm{m}}$ on the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$ as:

$$\delta_{\square,\mathrm{m}}(U,W) = \inf_{\varphi} d_{\square,\mathrm{m}}(U,W^{\varphi}) = \min_{\varphi,\psi} d_{\square,\mathrm{m}}(U^{\varphi},W^{\psi}),$$

where the infimum is taken over all measure-preserving maps φ and ψ, see Proposition 3.3.18 for alternative expressions of $\delta_{\square,\mathrm{m}}$ (including proof that the minimum exist for the second expression) and see Theorem 3.3.17 that states that $\delta_{\square,\mathrm{m}}$ is indeed a distance on $\widetilde{\mathcal{W}}_1$. In Proposition 3.4.13, we prove an equivalent of the weak regularity lemma for probability-graphons.

We define the notion of a quasi-convex distance, which generalizes the convexity of a norm.

Definition 3.1.1 (Quasi-convex distance)**.** *Let (X,d) be a metric space which is a convex subset of a vector space. The distance d is* quasi-convex *if for all $x_1,x_2,y_1,y_2\in X$ and all $\alpha\in[0,1]$, we have:*

$$d(\alpha x_1+(1-\alpha)x_2,\alpha y_1+(1-\alpha)y_2)\leq \max(d(x_1,y_1),d(x_2,y_2)).$$

In particular, any distance (on a convex subset of a vector space) which derive from a norm is quasi-convex. Moreover, the Prohorov distance $d_{\mathcal{P}}$ is quasi-convex (see Lemma 3.3.21).

Lemma 3.1.2. *The distances $d_{\mathcal{P}}$, d_{KR}, d_{FM} et $d_{\mathcal{F}}$ are all quasi-convex on $\mathcal{M}_{+}(\mathbf{Z})$.*

An interesting fact is that under some conditions on d_{m}, including the case when d_{m} is quasi-convex), the topology induced by the associated cut distance $\delta_{\square,\mathrm{m}}$ does not depend on the particular choice of d_{m}. The following proposition is a particular case of Theorem 3.5.5 together with Corollary 3.4.14.

Proposition 3.1.3. *The distance $\delta_{\square,\mathrm{m}}$, where d_{m} is a quasi-convex distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ that induces the weak topology, all induce the same topology on the space of probability-graphons $\widetilde{\mathcal{W}}_1$.*

Recall that $\mathbf{Z}$ is a Polish space. We now state that $\widetilde{\mathcal{W}}_1$ is also Polish for the distance $\delta_{\square,\mathcal{P}}$ (but not for $\delta_{\square,\mathcal{F}}$!), and we refer to Theorem 3.5.10 for other distances.

Theorem 3.1.4. *The space of probability-graphons $(\widetilde{\mathcal{W}}_1,\delta_{\square,\mathcal{P}})$ is a Polish metric space.*

We prove an analogue of Prohorov's theorem with a tightness criterion for probability-graphons. We say that a subset of probability-graphons $\mathcal{K}\subset\widetilde{\mathcal{W}}_1$ is *tight* if the set of probability measures $\{M_W : W\in\mathcal{K}\}$ is tight (in the sense of probability measures), where $M_W(\cdot)=W([0,1]^2;\cdot)$. The next result is consequence of Theorem 3.5.1 and Proposition 3.5.2 as well as Corollary 3.4.14.

Theorem 3.1.5 (Compactness property)**.** *Consider the topology on $\widetilde{\mathcal{W}}_1$ from Proposition 3.1.3.*

(i) If a sequence of elements of $\widetilde{\mathcal{W}}_1$ is tight, then it has a converging subsequence.

(ii) A subset $\mathcal{K}\subset\widetilde{\mathcal{W}}_1$ is relatively compact is and only if it is tight.

(iii) If $\mathbf{Z}$ is compact, then the space $\widetilde{\mathcal{W}}_1$ is compact.

Sampling from probability-graphons and its link with the cut distance

Finally, we link the topology of the cut distance $\delta_{\square,\mathrm{m}}$ with subgraph sampling. The probability-graphons allow to define models of random weighted graphs (the W-random graph model) which generalize weighted SBM random graphs, and which plays the role of sampled subgraphs for probability-graphons. The W-random graph (or sampled subgraph of size k) $\mathbb{G}(k,W)$ has two parameters, a number of vertices k and a probability-graphon W for edge-weights, and is defined as follows: first let $X_1,\cdots,X_k$ be k independent random "vertex-types" uniformly distributed over $[0,1]$; then given $X_1,\cdots,X_k$, each edge receive a weight independently, where the weight of the edge (i,j) is distributed as $W(X_i,X_j;\cdot)$.

We also provide the a.s. convergence of sampled subgraphs for the topology from Proposition 3.1.3, see Theorem 3.6.13 together with Corollary 3.5.6.

Theorem 3.1.6 (Convergence of sampled subgraphs)**.** *Let W be a probability-graphon. Then, a.s. the sequence of sampled subgraphs $(\mathbb{G}(k,W))_{k\in\mathbb{N}^*}$ converges to W for the topology from Proposition 3.1.3.*

To prove this theorem, we adapt the proof scheme of [Lov12, Sections 10.5 and 10.6] relying on the first and second sampling lemmas for real-valued graphons. The proof is done using the cut distance $\delta_{\square,\mathcal{F}}$ because of the good approximations properties of $\|\cdot\|_{\mathcal{F}}$.

In the case of unweighted graphs, the homomorphism numbers $\hom(F,G)$ count the number of occurence of a graph F (often called a *motif* or a *graphlet*) as an induced subgraph of G, and their normalized counterparts, the homomorphism densities $t(F,G)$ allow to characterize a graph (up to relabeling and twin-vertices expansion), and also characterize the topology on real-valued graphons. In the case of weighted graphs and probability-graphons, we need to replace absence/presence of edges (which is 0-1 valued) by test functions from $C_b(\mathbf{Z})$ decorating the edges. Hence, we define the *homomorphism density* of a $\mathcal{G}$-graph F^g which is a finite graph $F=(V,E)$ whose edges are decorated with a family of functions $g=(g_e)_{e\in E}$ from a subset $\mathcal{G}\subset C_b(\mathbf{Z})$ (in practice, we only consider the cases $\mathcal{G}=C_b(\mathbf{Z})$ or $\mathcal{G}=\mathcal{F}\subset C_b(\mathbf{Z})$ a convergence determining sequence), in a probability-graphon W as:

$$t(F^g,W)=M_W^F(g):=\int_{[0,1]^V}\prod_{(i,j)\in E}W(x_i,x_j;g_{i,j})\prod_{i\in V}\mathrm{d}x_i,$$

where $W(x,y;f)=\int_{\mathbf{Z}}f(z)\,W(x,y;\mathrm{d}z)$. Moreover, M_W^F defines a measure on $\mathbf{Z}^E$ (which we still denote by M_W^F) which is defined by $M_W^F(\otimes_{e\in E}g_e)=M_W^F(g)$ for $g=(g_e)_{e\in E}$. Note that when F is the complete graph with k vertices, M_W^F is the joint measure of all the edge-weights of the random graph $\mathbb{G}(k,W)$, and thus characterizes the random graph $\mathbb{G}(k,W)$.

In the counting Lemma 3.7.5 and the weak counting Lemma 3.7.7, we prove that the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ allows to control the homomorphism densities. Conversely, in the inverse counting Lemma 3.7.8, we prove that the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ can be controlled by the homomorphism densities. In particular, the topology of the cut distance turns out to be exactly the topology of convergence in distribution for sampled subgraphs of any given size; the next result is a direct consequence of Theorem 3.7.11.

Theorem 3.1.7 (Characterization of the topology)**.** *Let $(W_n)_{n\in\mathbb{N}}$ and W be unlabeled probability-gra-phons from $\widetilde{\mathcal{W}_1}$. The following properties are equivalent:*

(i) *$(W_n)_{n\in\mathbb{N}}$ converges to W for the topology from Proposition 3.1.3.*

(ii) *$\lim_{n\to\infty}t(F^g,W_n)=t(F^g,W)$ for all $C_b(\mathbf{Z})$-graph F^g.*

(iii) *$\lim_{n\to\infty}t(F^g,W_n)=t(F^g,W)$ for all $\mathcal{F}$-graph F^g, for some convergence determining sequence $\mathcal{F}$.*

(iv) *For all $k\geq 2$, the sequence of sampled subgraphs $(\mathbb{G}(k,W_n))_{n\in\mathbb{N}}$ converges in distribution to $\mathbb{G}(k,W)$.*

Now, we can turn back to the initial problem of finding a limit object for a convergent sequence of weighted graphs $(G_n)_{n\in\mathbb{N}}$; here convergent means that for all $k\geq 2$, the sequence $(\mathbb{G}(k,G_n)=\mathbb{G}(k,W_{G_n}))_{n\in\mathbb{N}}$ of sampled subgraphs of size k (defined above) converges in distribution (to some limit random graph). Note that the tightness criterion for a sequence of probability-graphons $(W_n)_{n\in\mathbb{N}}$ can be equivalently rephrased as tightness of the sequence $(\mathbb{G}(2,W_n))_{n\in\mathbb{N}}$ of sampled subgraphs of size 2. Hence, the convergence in distribution of the sequence $(\mathbb{G}(2,G_n))_{n\in\mathbb{N}}$ implies its tightness, and thus the tightness of the sequence of probability-graphons $(W_{G_n})_{n\in\mathbb{N}}$. Then, Theorem 3.1.5 guarantees the existence of a probability-graphon W which is a sub-sequential limit of the sequence $(W_{G_n})_{n\in\mathbb{N}}$ in the cut distance $\delta_{\square,\mathcal{F}}$, and then Theorem 3.1.7 guarantees that for all $k\geq 2$, the sequence $(\mathbb{G}(k,G_n))_{n\in\mathbb{N}}$ converges in distribution to $\mathbb{G}(k,W)$.

As a consequence, probability-graphons are precisely the limit objects for sequences of weighted graphs $(G_n)_{n\in\mathbb{N}}$ (and also for random weighted graphs) whose number of vertices goes to infinity (otherwise the limit would simply be a weighted graph) and such that for each size $k\geq 2$, the sequence of sampled subgraphs $(\mathbb{G}(k,G_n))_{n\in\mathbb{N}}$ converges in distribution.

Remark 3.1.8 (Extension to vertex-weights)**.** The framework we have developed for probability-gra-phons could easily be extended to add weights on the vertices, or equivalently to allow for self-loops (i.e. edges linking a vertex to itself). In this case, weighted graphs and probability-graphons have a two-variable kernel (probability-graphon) W^{e} for edge-weights as before, and a one-variable kernel $W^{\mathrm{v}}:[0,1]\to\mathcal{M}_1(\mathbf{Z})$ for vertex-weights. Note that this implies, as expected, that the same measure-preserving map $\varphi:[0,1]\to[0,1]$ must be used for both kernels W^{v} and W^{e} when relabeling.

3.1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 3.2, we define some notations used throughout the paper, and remind some properties of the weak topology on the space of signed measures. In Section 3.3, we define probability-graphons and signed-measure valued kernels, we then define the cut distance and the cut norm and study their properties, and we also give some exemple of distances with the Prohorov distance $d_{\mathcal{P}}$, the Kanrorovitch-Rubinstein and Fortet-Mourier norms $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$, and the norm $\|\cdot\|_{\mathcal{F}}$ based on a convergence determining sequence. In Section 3.4, we define the steppings of a probability-graphon (which are stepfunction approximations corresponding to conditional expectations on $[0,1]^2$), we define the tightness criterion for probability-graphons, and we prove the weak regularity property of the cut distance. In Section 3.5, we prove the theorem linking the tightness criterion with relative compactness for the cut distance, we prove that under some conditions the topology of the cut distance does not depend on the choice of the initial distance d_{m}, and we prove that the space of probability-graphons with the cut distance is a Polish space. In Section 3.6, we define the subgraph $\mathbb{G}(k,W)$ sampled from a probability-graphon W, we then prove approximation bound in the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ between probability-graphons and their sampled subgraphs. In Section 3.7, we prove the counting lemmas linking the cut distance with the homomorphism densities, and prove that the topology induced by the cut distance coincides with the topology of convergence in distribution for all the sampled subgraphs.

Contents

3.2 Notations and topology on the space of signed measures

Through the article, *measure* will always be used to denote a positive measure.

Let $\mathbb{N} = \mathbb{Z}_+$ be the set of non-negative integers, $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$ the set of positive integers, and, for $n \in \mathbb{N}^*$, we define the integer set $[n] = \{1, \dots, n\}$. For $k \in \mathbb{N}^*$, the set $[0,1]^k$ is endowed with the Borel σ-field and the Lebesgue measure λ_k; and we write λ for λ_k when the context is clear. The supremum of a real-valued function f defined on $[0,1]^k$ is denoted by $\|f\|_\infty = \sup_{x\in[0,1]^k} f(x)$.

Let d be a distance on a topological space $(X, \mathcal{O})$.

(i) The distance d is continuous w.r.t. the topology $\mathcal{O}$ if the identity map from $(X, \mathcal{O})$ to (X, d) is continuous.

(ii) The distance d is sequentially continuous w.r.t. the topology $\mathcal{O}$ if for any sequence $(x_n)_{n\in\mathbb{N}}$ in X which converges to some limit x for the topology $\mathcal{O}$, we also have that $\lim_{n\to\infty} d(x_n, x) = 0$.

Let d and d' be two distances on a space X. We say that d' is continuous (resp. uniformly continuous) w.r.t. d if the identity map from (X, d) to (X, d') is continuous (resp. uniformly continuous).

Remark 3.2.1. If the topology $\mathcal{O}$ is metrizable (i.e. can be generated by a distance on the space X), then the topology on X induced by the distance d is equivalent to $\mathcal{O}$ if and only if for every sequence with values in X, convergence for d is equivalent to convergence for $\mathcal{O}$ (see [Eng89, Theorem 4.1.2]). Moreover, when the topology is metrizable, then topological notions and their sequential counterparts coincides (e.g. compact and sequentially compact sets, closed and sequentially closed sets, see [Eng89, Proposition 4.1.1 and Theorem 4.1.17]).

Remark 3.2.2. For a function, continuity always implies sequential continuity; and the converse is also true when the topology is metrizable.

A map $\varphi : \Omega_1 \to \Omega_2$ between two probability spaces $(\Omega_i, \mathcal{A}_i, \pi_i)$, $i = 1, 2$, is measure-preserving if it is measurable and if for every $A \in \mathcal{A}_2$, $\pi_2(A) = \pi_1(\varphi^{-1}(A))$. In this case, for every measurable non-negative function $f : \Omega_2 \to \mathbb{R}$, we have:

$$\int_{\Omega_1} f(\varphi(x))\ \pi_1(\mathrm{d}x) = \int_{\Omega_2} f(x)\ \pi_2(\mathrm{d}x). \tag{3.1}$$

We denote by $S_{[0,1]}$ the set of bijective measure-preserving maps from $[0,1]$ with the Lebesgue measure to itself, and by $\bar{S}_{[0,1]}$ the set of measure-preserving maps from $[0,1]$ with the Lebesgue measure to itself.

Let $(\mathbf{Z}, \mathcal{O}_\mathbf{Z})$ be some (non-empty) Polish space, and let $\mathcal{B}(\mathbf{Z})$ be the Borel σ-field on $\mathbf{Z}$ generated by the topology $\mathcal{O}_\mathbf{Z}$. We denote by $C_b(\mathbf{Z})$ the space of real-valued continuous bounded functions on $(\mathbf{Z}, \mathcal{O}_\mathbf{Z})$. We denote by $\mathcal{M}_\pm(\mathbf{Z})$ the space of finite signed measures on $(\mathbf{Z}, \mathcal{B}(\mathbf{Z}))$; $\mathcal{M}_+(\mathbf{Z})$ the subspace of measures; $\mathcal{M}_{\leq 1}(\mathbf{Z})$ the subspace of measures with total mass at most 1; and $\mathcal{M}_1(\mathbf{Z})$ the subspace of probability measures. We have:

$$\mathcal{M}_1(\mathbf{Z}) \subset \mathcal{M}_{\leq 1}(\mathbf{Z}) \subset \mathcal{M}_+(\mathbf{Z}) \subset \mathcal{M}_\pm(\mathbf{Z}).$$

For a signed measure $\mu \in \mathcal{M}_\pm(\mathbf{Z})$, we remind the definition of the Hahn-Jordan decomposition $\mu = \mu^+ - \mu^-$ where $\mu^+, \mu^- \in \mathcal{M}_+(\mathbf{Z})$ are mutually singular measures (that is $\mu^+(A) = 0$ and $\mu^-(A^c) = 0$ for some measurable set A), as well as the total variation measure of μ which is defined as $|\mu| = \mu^+ + \mu^- \in \mathcal{M}_+(\mathbf{Z})$. Note that for a measure $\mu \in \mathcal{M}_+(\mathbf{Z})$, we simply have $|\mu| = \mu$. For a signed-measure $\mu \in \mathcal{M}_\pm(\mathbf{Z})$ and a real-valued measurable function f defined on $\mathbf{Z}$, we write $\mu(f) = \langle \mu, f\rangle = \int f\,\mathrm{d}\mu = \int_\mathbf{Z} f(x)\,\mu(\mathrm{d}x)$ the integral of f w.r.t. μ whenever it is well defined. For a signed measure $\mu \in \mathcal{M}_\pm(\mathbf{Z})$, we denote by

$\|\mu\|_\infty = \mu^+(\mathbf{Z}) + \mu^-(\mathbf{Z})$ its total mass, which is also equal to the supremum of $\mu(f)$ over all measurable functions f with values in $[-1, 1]$.

We endow $\mathcal{M}_\pm(\mathbf{Z})$ with the topology of weak convergence, that is the smallest topology for which the maps $\mu \mapsto \mu(f)$ are continuous for all $f \in C_b(\mathbf{Z})$. In particular, a sequence of signed measures $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to some $\mu \in \mathcal{M}_\pm(\mathbf{Z})$ if and only if, for every function $f \in C_b(\mathbf{Z})$, we have $\lim_{n\to+\infty} \mu_n(f) = \mu(f)$. Let us recall that $\mathcal{M}_+(\mathbf{Z})$ and $\mathcal{M}_1(\mathbf{Z})$ endowed with the topology of weak convergence are Polish spaces.

Remark 3.2.3 (The weak topology on $\mathcal{M}_\pm(\mathbf{Z})$)**.** The topology of weak convergence on the set of signed measures $\mathcal{M}_\pm(\mathbf{Z})$ is equivalent to the weak-$*$ topology on $\mathcal{M}_\pm(\mathbf{Z})$ seen as a subspace of the topological dual of $C_b(\mathbf{Z})$ (see the paragraph after Definition 3.1.1 in [Bog18]). As usual in probability theory, this topology will be simply called the weak topology (this is also consistent with [Bog18]).

We recall that a sequence of $[0, 1]$-valued functions $\mathcal{F} = (f_k)_{k\in\mathbb{N}}$ in $C_b(\mathbf{Z})$, with $f_0 = \mathbb{1}$ the constant function equal to one, is:

(i) **Separating** if for every measures μ, ν from $\mathcal{M}_\pm(\mathbf{Z})$ (or equivalently just from $\mathcal{M}_+(\mathbf{Z})$) such that for every $k \in \mathbb{N}$, $\mu(f_k) = \nu(f_k)$, then $\mu = \nu$.

(ii) **Convergence determining** if for every $(\mu_n)_{n\in\mathbb{N}}$ and μ measures from $\mathcal{M}_+(\mathbf{Z})$ such that we have $\lim_{n\to+\infty} \mu_n(f_k) = \mu(f_k)$ for all $k \in \mathbb{N}$, then $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to μ.

Notice that a convergence determining sequence is also separating. A sequence of functions is separating if and only if it separates the points of $\mathbf{Z}$ (see [EK09, Theorem 3.4.5]). There always exists a convergence determining sequence on Polish spaces, see [Bog18, Corollary 2.2.6] or the proof of Proposition 3.4.4 in [EK09] (which are stated for probability measures but can be extended to finite positive measures as we required that $\mathbb{1}$ belongs to $\mathcal{F}$). Note that there does not exist a convergence determining sequence for $\mathcal{M}_\pm(\mathbf{Z})$ as the weak topology is not metrizable on $\mathcal{M}_\pm(\mathbf{Z})$ (see Remark 3.2.6 below).

Remark 3.2.4 (The Borel σ-field on $\mathcal{M}_\pm(\mathbf{Z})$)**.** By [Bog18, Corollary 5.1.9], the Borel σ-field on $\mathcal{M}_\pm(\mathbf{Z})$, associated with the weak topology, is countably generated and can be generated by either:

— the family of maps $\mu \mapsto \mu(f_n)$ where the sequence $(f_n)_{n\in\mathbb{N}}$ of functions from $C_b(\mathbf{Z})$ is separating;

— the family of maps $\mu \mapsto \mu(B)$ where $B \in \mathcal{A}$ and the subset $\mathcal{A} \subset \mathcal{B}(\mathbf{Z})$ is countable and generates the whole σ-field $\mathcal{B}(\mathbf{Z})$ (such subset $\mathcal{A}$ always exists, see [Bog07b, Corollary 6.7.5]).

Note that the Borel σ-field of a Polish space is generated by any family of Borel functions that separates points (see [Bog07b, Theorem 6.8.9]).

Furthermore, the maps $\mu \mapsto \mu^+$ and $\mu \mapsto \mu^-$ (and thus also $\mu \mapsto |\mu|$) are measurable (see [DF64, Theorem 2.8] and Remark 3.2.4). As a consequence, the map $\mu \mapsto \|\mu\|_\infty$ is also measurable (in fact it is even lower semicontinuous by [Bog18, Theorem 2.7.4]). Note that $\mathcal{M}_1(\mathbf{Z})$ and $\mathcal{M}_+(\mathbf{Z})$ are closed, and thus measurable, subsets of $\mathcal{M}_\pm(\mathbf{Z})$.

We define the following two important properties for subsets of signed measures, which are related to relative compactness (see Lemma 3.2.8 below).

Definition 3.2.5. *Let $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$ be a subset of signed measures.*

(i) The set $\mathcal{M}$ is bounded *(in total variation) if:*

$$\sup_{\mu\in\mathcal{M}} \|\mu\|_\infty < +\infty.$$

(ii) The set $\mathcal{M}$ is tight *if for all $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{Z}$ such that:*

$$\sup_{\mu\in\mathcal{M}} |\mu|(K^c) \leq \varepsilon.$$

Remark 3.2.6 (On the compact sets and metrizability of the weak topology)**.** Recall that $\mathbf{Z}$ is a Polish space. We stress that the weak topology on signed measures is not metrizable unless it coincides with the strong topology (see [Var58, Theorem 4.1]), which happens only when the initial space $\mathbf{Z}$ is finite (see [Bog18, Proposition 3.1.8]).

Moreover, the closed norm ball $\{\mu \in \mathcal{M}_\pm(\mathbf{Z}) : \|\mu\|_\infty \leq 1\}$ of $\mathcal{M}_\pm(\mathbf{Z})$ is metrizable if and only if $\mathbf{Z}$ is compact (see [Bog18, Proposition 3.1.8 and Theorem 3.1.9]).

Let $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$. The following properties are equivalent (see [Bog18, Theorems 2.3.4 and 3.1.9]):

(i) $\mathcal{M}$ is weakly compact (i.e. $\mathcal{M}$ is compact for the weak topology);

(ii) $\mathcal{M}$ is sequentially weakly compact (that is every sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{M}$ has a subsequence that converges to some limit $\mu \in \mathcal{M}$);

(iii) $\mathcal{M}$ is compact for the sequential weak topology (for which sets are closed if and only if they are closed under weak convergence).

Moreover, when any of those is true, $\mathcal{M}$ is tight, bounded, and metrizable in the weak topology. Furthermore, the Kantorovitch-Rubinstein and Fortet-Mouriet norms $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$ (defined in Section 3.3.8) can be used to generate the weak topology on a weakly compact set (see [Bog18, Remark 3.2.5]).

Nevertheless, the weak topology on the unit sphere $\{\mu \in \mathcal{M}_\pm(\mathbf{Z}) : \|\mu\|_\infty = 1\}$ of $\mathcal{M}_\pm(\mathbf{Z})$ is always metrizable with a complete metric, making the unit sphere a Polish space, however, the Kantorovitch-Rubinstein and Fortet-Mouriet norms $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$ do not provide a complete metrization in this case (see [Bog18, Theorem 3.2.8]).

Remark 3.2.7 (On the compactness of $\mathcal{M}_1(\mathbf{Z})$)**.** Let $\mathcal{M}$ be either $\mathcal{M}_1(\mathbf{Z})$, $\mathcal{M}_{\leq 1}(\mathbf{Z})$ or the closed norm ball $\{\mu \in \mathcal{M}_\pm(\mathbf{Z}) : \|\mu\|_\infty \leq 1\}$ of $\mathcal{M}_\pm(\mathbf{Z})$. Then, $\mathcal{M}$ is weakly compact if and only if $\mathbf{Z}$ is compact.

We give a short proof of this statement. As $\mathcal{M}_1(\mathbf{Z})$ is closed in $\mathcal{M}_\pm(\mathbf{Z})$ for the weak topology, if $\mathcal{M}$ is weakly compact, then $\mathcal{M}_1(\mathbf{Z})$ is also weakly compact, and thus $\mathbf{Z}$ is compact by [Var58, Theorem 3.4]. Conversely, if $\mathbf{Z}$ is compact, then by [Bog18, Theorem 1.3.3], we know that $\mathcal{M}_\pm(\mathbf{Z})$ (endowed with the weak topology) is the topological dual space of $C_b(\mathbf{Z})$ (endowed with the uniform convergence topology), thus using Banach-Alaoglu theorem (see [Bog18, Theorem 1.3.6]), we get that the closed unit norm-ball of $\mathcal{M}_\pm(\mathbf{Z})$, and thus $\mathcal{M}$, are compact for the weak topology.

We recall the following result, which is an equivalent of Prohorov's theorem for signed measures.

Lemma 3.2.8 (Prohorov's theorem for signed measures, [Bog18, Theorems 2.3.4 and 3.1.9])**.** *Let $\mathbf{Z}$ be a Polish space, and let $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$ be a subset of signed measures on $\mathbf{Z}$. Then the following conditions are equivalent:*

(i) $\mathcal{M}$ is relatively sequentially compact, that is every sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{M}$ contains a subsequence which weakly converges in $\mathcal{M}_\pm(\mathbf{Z})$.

(ii) $\mathcal{M}$ is relatively compact for the weak topology, that is the closure of $\mathcal{M}$ is compact for the weak topology.

(iii) The family $\mathcal{M}$ is tight and bounded.

Remark 3.2.9 (On the weak sequential topology)**.** When the space $\mathbf{Z}$ is infinite, the weak topology does not coincide with the weak sequential topology on $\mathcal{M}_\pm(\mathbf{Z})$ (but recall from Remark 3.2.6 that their compact sets are the same). Recall that if the space $\mathbf{Z}$ is compact, then the unit norm ball of $\mathcal{M}_\pm(\mathbf{Z})$ is metrizable, and thus the weak topology and the weak sequential topology coincide on it. However, if the space $\mathbf{Z}$ is non-compact, then the weak topology and the weak sequential topology do not coincide on the unit norm ball of $\mathcal{M}_\pm(\mathbf{Z})$.

We give a short proof of those statements according to $\mathbf{Z}$ being compact or not.

(i) Remind that when $\mathbf{Z}$ is an infinite compact space (for instance $\mathbf{Z} = [0, 1]$), the Banach space $C_b(\mathbf{Z})$ is infinite-dimensional and separable (using Stone-Weierstrass theorem), and its topological dual is $(C_b(\mathbf{Z}))^* = \mathcal{M}_\pm(\mathbf{Z})$ (see [Bog18, Theorem 1.3.3]). Thus, using [HS96, Theorem 2.5], we get the existence of a countable subset which is weak sequentially closed yet weak dense in $\mathcal{M}_\pm(\mathbf{Z})$. In particular, the weak sequential topology and the weak topology do not coincide on $\mathcal{M}_\pm(\mathbf{Z})$.

(ii) Assume that the space $\mathbf{Z}$ is non-compact. Thus, $\mathbf{Z}$ contains a countable closed subset F whose points are at mutual distances uniformly bounded away from zero. By [Bog18, Remark 3.1.7], the weak topology on $\mathcal{M}_\pm(F)$ for a closed subset F coincides with the trace of the weak topology on the whole space. By [Bog18, Section 3.1, p. 102], $\mathcal{M}_\pm(F)$ is homeomorphic to ℓ^1 both endowed with their weak topology, weak convergence on ℓ^1 is equivalent to norm convergence, and the weak topology on ℓ^1 is not sequential, even on the unit norm ball. Hence, the weak topology on $\mathcal{M}_\pm(\mathbf{Z})$ is not sequential, even on the unit norm ball.

We define the notion of a quasi-convex distance, which generalizes the convexity of a norm.

Definition 3.2.10 (Quasi-convex distance)**.** *Let (X, d) be a metric space which is a convex subset of a vector space. The distance d is* quasi-convex *if for all $x_1, x_2, y_1, y_2 \in X$ and all $\alpha \in [0,1]$, we have:*

$$d(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2) \leq \max(d(x_1, y_1), d(x_2, y_2)).$$

In particular, any distance (on a convex subset of a vector space) which derive from a norm is quasi-convex.

Lemma 3.2.11. *Let d_{m} be distance on $\mathcal{M}_\epsilon(\mathbf{Z})$ with $\epsilon \in \{+, \pm\}$ which is quasi-convex and sequentially continuous with respect to the weak topology. Then, d_{m} is uniformly continuous with respect to $\|\cdot\|_\infty$ on $\mathcal{M}_\epsilon(\mathbf{Z})$.*

Proof. We shall simply consider the case $\mathcal{M} = \mathcal{M}_+(\mathbf{Z})$, the other case being simpler. We first check that for all $\mu \in \mathcal{M}$ and $\varepsilon > 0$, there exists $\eta > 0$ such that for all $\nu \in \mathcal{M}$, we have that $\|\mu - \nu\|_\infty < \eta$ implies $d_{\mathrm{m}}(\mu, \nu) < \varepsilon$. As d_{m} is sequentially continuous w.r.t. the weak topology, it is also (sequentially) continuous w.r.t. the strong topology. Let $\mu \in \mathcal{M}$ and $\varepsilon > 0$. Then, the set $\{\nu \in \mathcal{M} \,:\, d_{\mathrm{m}}(\mu, \nu) < \varepsilon\}$ is an open set of $\mathcal{M}$ containing μ both for d_{m} and for the strong topology. Thus, it contains a neighborhood of μ for the strong topology $\{\nu \in \mathcal{M} \,:\, \|\mu - \nu\|_\infty < \eta\}$ for $\eta > 0$ small enough. This proves the claim.

As d_{m} is quasi-convex and $\mathcal{M}$ is a cone, for $\mu, \nu \in \mathcal{M}$ we have:

$$d_{\mathrm{m}}(\mu, \mu+\nu) = d_{\mathrm{m}}\left(\frac{1}{2}\cdot(2\mu+0), \frac{1}{2}\cdot(2\mu+2\nu)\right) \leq \max(d_{\mathrm{m}}(2\mu, 2\mu), d_{\mathrm{m}}(0, 2\nu)) = d_{\mathrm{m}}(0, 2\nu).$$

Let $\varepsilon > 0$ be fixed. We choose $\eta \in (0,1)$ such that $\|\nu\|_\infty < \eta$, with $\nu \in \mathcal{M}$, implies $d_{\mathrm{m}}(0, \nu) < \varepsilon$. Let $\mu, \nu \in \mathcal{M}$ be such that $\|\mu - \nu\|_\infty < \eta/2$. Let $\lambda' = \mu + \nu$ and f (resp. g) the density of μ (resp. ν) with respect to λ'. We set $\pi = \min(f, g)\,\lambda'$, $\mu' = (f-g)_+\,\lambda'$ and $\nu' = (f-g)_-\,\lambda'$ so that $\pi, \mu', \nu' \in \mathcal{M}$, $\mu = \pi + \mu'$ and $\nu = \pi + \nu'$. Since $\mu' - \nu' = \mu - \nu$ and μ' and ν' are mutually singular, we deduce that $\|\mu'\|_\infty + \|\nu'\|_\infty < \eta/2$. We get:

$$\begin{aligned} d_{\mathrm{m}}(\mu, \nu) = d_{\mathrm{m}}(\pi+\mu', \pi+\nu') &\leq d_{\mathrm{m}}(\pi, \pi+\mu') + d_{\mathrm{m}}(\pi, \pi+\nu') \\ &\leq d_{\mathrm{m}}(0, 2\mu') + d_{\mathrm{m}}(0, 2\nu') \\ &\leq 2\varepsilon. \end{aligned}$$

Hence, the distance d_{m} is uniformly continuous with respect to $\|\cdot\|_\infty$ on $\mathcal{M}$. □

3.3 Measured-valued graphons and the cut distance

In Section 3.3.1, we introduce the measure-valued graphons, which are a generalization of real-valued graphons (i.e. $[0,1]$-valued measurable functions defined on $[0,1]^2$). We refer to the monography [Lov12] on real-valued graphons for more details. In Sections 3.3.2, 3.3.3 and 3.3.4, we introduce the cut distance, and its unlabeled variant, on the space of measure-valued graphons which are analogous to the ones for real-valued graphons (see [Lov12, Chapter 8]). In Section 3.3.5, we define a weak isomorphism relation for measure-valued graphons based on this distance. Then, in Section 3.3.6, we give an alternative combinatorial formulation of the cut distance for stepfunctions.

3.3.1 Definition of measure-valued graphons

We start by defining measure-valued kernels and graphons which are a generalization of real-valued kernels and graphons. Recall that $\mathbf{Z}$ is a Polish space and $\mathcal{M}_\pm(\mathbf{Z})$ is the space of *finite* signed measures.

Definition 3.3.1 (Signed measure-valued kernels)**.** *A* signed measure-valued kernel *or* $\mathcal{M}_\pm(\mathbf{Z})$-valued kernel *is a map W from $[0,1]^2$ to $\mathcal{M}_\pm(\mathbf{Z})$, such that:*

(i) W is a signed-measure *in* dz*: for every $(x, y) \in [0,1]^2$, $W(x, y; \cdot)$ belongs to $\mathcal{M}_\pm(\mathbf{Z})$.*

(ii) W is measurable *in (x, y): for every measurable set $A \subset \mathbf{Z}$, the function $(x, y) \mapsto W(x, y; A)$ defined on $[0,1]^2$ is measurable.*

(iii) W is bounded*:*

$$\|W\|_\infty := \sup_{x,y\in[0,1]} \|W(x, y; \cdot)\|_\infty < +\infty. \tag{3.2}$$

We denote by $\mathcal{W}_1$ (resp. $\mathcal{W}_{\leq 1}$, resp. $\mathcal{W}_+$, resp. $\mathcal{W}_\pm$) the space of probability measure-valued kernels or simply probability-graphons (resp. sub-probability measure-valued kernels, resp. measure-valued kernels, resp. signed measure-valued kernels), where we identify kernels that are equal a.e. on $[0,1]^2$, with respect to the Lebesgue measure. Then, (3.2) should be read with an essential supremum instead of a supremum. In what follows, we always assume for simplicity that we choose representatives of measure-valued kernels such that $\|W\|_\infty$ is also the essential supremum of $(x,y) \mapsto \|W(x,y;\cdot)\|_\infty$.

For $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$, we denote by $\mathcal{W}_\mathcal{M}$ the subset of signed measure-valued kernel $W \in \mathcal{W}_\pm$ which are $\mathcal{M}$-valued: $W(x,y;\cdot) \in \mathcal{M}$ for every $(x,y) \in [0,1]^2$.

Remark 3.3.2 (On real-valued kernels)**.** Let $\mathbf{Z} = \{0,1\}$ be equipped with the discrete topology. Every real-valued graphon w can be represented using a probability-graphon W defined for every $x,y \in [0,1]$ by $W(x,y;\mathrm{d}z) = w(x,y)\delta_1(\mathrm{d}z) + (1-w(x,y))\delta_0(\mathrm{d}z)$, where δ_z is the Dirac mass located at z. In particular we have that $w(x,y) = W(x,y;\{1\})$ for $x,y \in [0,1]$.

Let $W \in \mathcal{W}_\pm$ be a signed measure-valued kernel. Define the map $W^+ : [0,1]^2 \to \mathcal{M}_+(\mathbf{Z})$ to be the positive part of W, i.e. for every $(x,y) \in [0,1]^2$, $W^+(x,y;\cdot)$ is the positive part of the measure $W(x,y;\cdot)$. Similarly define $W^- : [0,1]^2 \to \mathcal{M}_+(\mathbf{Z})$ the negative part of W; and then define $|W| = W^+ + W^-$ the total variation of W and $\|W\| = |W|(\mathbf{Z})$ the total mass of W.

Lemma 3.3.3 (The positive part W^+ of a kernel)**.** *The maps W^+, W^- and $|W|$ are all measure-valued kernels, and the map $\|W\| : (x,y) \mapsto \|W(x,y;\cdot)\|_\infty$ is measurable.*

Proof. The statements for $|W|$ and $\|W\|$ are immediate consequences of the statements for W^+ and W^-; and as the proof for W^+ and W^- are similar, we only need to prove that W^+ is a measure-valued kernel. It is immediate that W^+ is bounded and that for every $(x,y) \in [0,1]^2$, $W^+(x,y;\cdot)$ is a measure in $\mathcal{M}_+(\mathbf{Z})$. Thus, we are left to prove the measurability of W^+ in (x,y). By [DF64, Proposition 2.1] and Remark 3.2.4, a signed measure-valued kernel U is measurable in (x,y) (i.e. for every $A \in \mathcal{B}(\mathbf{Z})$, the map $(x,y) \mapsto U(x,y;A)$ is measurable) if and only if the map $(x,y) \mapsto U(x,y;\cdot)$ is measurable from $[0,1]^2$ (with its Borel σ-field) to $\mathcal{M}_\pm(\mathbf{Z})$ equipped with the Borel σ-field generated by the weak topology. By [DF64, Theorem 2.8], the map $\mu \mapsto \mu^+$, that associate to a signed measure the positive part of its Hahn-Jordan decomposition, is measurable from $\mathcal{M}_\pm(\mathbf{Z})$ to $\mathcal{M}_+(\mathbf{Z})$ both endowed with the Borel σ-field generated by the weak topology. Considering the composition of W and $\mu \mapsto \mu^+$, we get that W^+ is measurable in (x,y) and is thus a measure-valued kernel. □

Remark 3.3.4 (Probability-graphons $W : \Omega \times \Omega \to \mathcal{M}_1(\mathbf{Z})$)**.** Similarly to the case of real-valued graphons, it is possible to replace the vertex-type space $[0,1]$ by any standard probability space $(\Omega, \mathcal{A}, \pi)$ that might be more appropriate to represent vertex-types for some applications, and to consider probability-graphons of the form $W : \Omega \times \Omega \to \mathcal{M}_1(\mathbf{Z})$. We recall that a standard probability space $(\Omega, \mathcal{A}, \pi)$ is a probability space such that there exists a measure-preserving map $\varphi : [0,1] \to \Omega$, where $[0,1]$ is endowed with the Borel σ-field and the Lebesgue measure. In particular, every Polish space endowed with its Borel σ-field is a standard probability space. As an example, the space $[0,1]^2$ equipped with the Borel σ-field and the Lebesgue measure λ_2 is a standard probability space; we will reuse this fact later.

Using the measure preserving map φ, it is then possible to consider an unlabeled version W^φ of W constructed on $\Omega' = [0,1]$, and to modify the definition of the cut distance $\delta_{\square,\mathrm{m}}$ similarly as in [Jan13, Theorem 6.9] to allow each probability-graphons to be constructed on different standard probability spaces. For simplicity, in this article we only consider the equivalent case where all probability-graphons are constructed on $\Omega = [0,1]$.

Remark 3.3.5 (Symmetric kernels)**.** We shall consider non-symmetric measure-valued kernels and probability-graphons in order to handle directed graphs whose adjacency matrices are thus *a priori* non-symmetric. We say that a measure-valued kernel or graphon W is symmetric if for a.e. $x,y \in [0,1]$, $W(x,y;\cdot) = W(y,x;\cdot)$.

We define stepfunctions measure-valued kernel which are often used for approximation.

Definition 3.3.6 (Signed measure-valued stepfunctions)**.** *A signed measure-valued kernel $W \in \mathcal{W}_\pm$ is a* stepfunction *if there exists a finite partition of $[0,1]$ into measurable (possibly empty) sets, say $\mathcal{P} = \{S_1, \cdots, S_k\}$, such that W is constant on the sets $S_i \times S_j$, for $1 \leq i,j \leq k$. We say that W and the partition $\mathcal{P}$ are* adapted *to each other. We write $|\mathcal{P}| = k$ the number of elements of the partition $\mathcal{P}$.*

3.3.2 The cut distance

We define a distance and a norm on signed measure-valued graphons and kernels, called the *cut distance* and the *cut norm* respectively which are analogous to the cut norm for real-valued graphons and kernels, see [Lov12, Chapter 8]. For a signed measure-valued kernel $W \in \mathcal{W}_\pm$ and a measurable subsets $A \subset [0,1]^2$, we denote by $W(A;\cdot)$ the signed measure on $\mathbf{Z}$ defined by:

$$W(A;\cdot) = \int_A W(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y.$$

Definition 3.3.7 (The cut distance $d_{\square,\mathrm{m}}$)**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}$ a convex subset of $\mathcal{M}_\pm(\mathbf{Z})$ containing the zero measure. The associated* cut distance $d_{\square,\mathrm{m}}$ *is the function defined on $\mathcal{W}_{\mathcal{M}}^2$ by:*

$$d_{\square,\mathrm{m}}(U,W) = \sup_{S,T\subset[0,1]} d_{\mathrm{m}}\Big(U(S\times T;\cdot), W(S\times T;\cdot)\Big), \tag{3.3}$$

where the supremum is taken over all measurable subsets S and T of $[0,1]$.

Notice that the right-hand side of (3.3) is well defined as $\mathcal{M}$ contains the zero measure (and thus if U belongs to $\mathcal{W}_{\mathcal{M}}$ then $U(A;\cdot)$ belongs to $\mathcal{M}$).

Definition 3.3.8 (The cut norm $N_{\square,\mathrm{m}}$)**.** *The* cut norm $N_{\square,\mathrm{m}}$ *associated with a norm N_{m} on $\mathcal{M}_\pm(\mathbf{Z})$ is the function defined on $\mathcal{W}_\pm$ by:*

$$N_{\square,\mathrm{m}}(W) = \sup_{S,T\subset[0,1]} N_{\mathrm{m}}\Big(W(S\times T;\cdot)\Big),$$

where the supremum is taken over all measurable subsets S and T of $[0,1]$.

The next proposition states that the cut distance (resp. norm) is indeed a distance (resp. norm); its extension to distances on $\mathcal{M}_+(\mathbf{Z})$ and $\mathcal{M}_\pm(\mathbf{Z})$ is immediate.

Proposition 3.3.9 ($d_{\square,\mathrm{m}}$ is a distance, $N_{\square,\mathrm{m}}$ is a norm)**.** *The cut distance $d_{\square,\mathrm{m}}$ associated with a distance d_{m} on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$) is a distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$). The cut norm $N_{\square,\mathrm{m}}$ associated with a norm N_{m} on $\mathcal{M}_\pm(\mathbf{Z})$ is a norm on $\mathcal{W}_\pm$.*

Moreover, when the distance d_{m} on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$) derives from a norm N_{m} on $\mathcal{M}_\pm(\mathbf{Z})$, then the distance $d_{\square,\mathrm{m}}$ derives also from the norm $N_{\square,\mathrm{m}}$.

Proof. Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (the proof for the case $\mathcal{M}_+(\mathbf{Z})$ is similar). It is clear that $d_{\square,\mathrm{m}}$ is symmetric and satisfies the triangular inequality. Thus, we only need to prove that $d_{\square,\mathrm{m}}$ is separating. Let U and W be two probability-graphons such that $d_{\square,\mathrm{m}}(U,W) = 0$. Then, for every measurable subsets $S,T \subset [0,1]$, we have $U(S\times T;\cdot) = W(S\times T;\cdot)$. Let $\mathcal{F} = (f_k)_{k\in\mathbb{N}}$ be a separating sequence. For every $k\in\mathbb{N}$, and for every measurable subsets $S,T\subset[0,1]$, we have that $U(S\times T; f_k) = W(S\times T; f_k)$. This implies that $U(x,y;f_k)\,\mathrm{d}x\mathrm{d}y = W(x,y;f_k)\,\mathrm{d}x\mathrm{d}y$ for all $k\in\mathbb{N}$. Hence, we deduce that for all $k\in\mathbb{N}$, $U(x,y;f_k) = W(x,y;f_k)$ for almost every $(x,y)\in[0,1]^2$. Thus, $U(x,y;\cdot) = W(x,y;\cdot)$ for almost every $(x,y)\in[0,1]^2$. This implies that $d_{\square,\mathrm{m}}$ is separating on $\mathcal{W}_1$, and thus a distance on $\mathcal{W}_1$.

The proof for the cut norm is similar. The proof of the last part of the proposition is clear. □

3.3.3 Graphon relabeling, invariance and smoothness properties

The analogue of graph relabelings for graphons are measure-preserving maps. Recall the definition of a measure-preserving map from Section 3.2, and in particular (3.1). Recall $\tilde{S}_{[0,1]}$ denotes the set of measure-preserving (measurable) maps from $[0,1]$ to $[0,1]$ endowed with the Lebesgue measure, and $S_{[0,1]}$ denotes its subset of bijective maps.

The relabeling of a signed measure-valued kernel W by a measure-preserving map φ, is the signed measure-valued kernel W^φ defined for every $x,y\in[0,1]$ and every measurable set $A\subset\mathbf{Z}$ by:

$$W^\varphi(x,y;A) = W(\varphi(x),\varphi(y);A) \quad \text{for } x,y\in[0,1] \text{ and } A\subset\mathbf{Z} \text{ measurable.}$$

We say that a subset $\mathcal{K}\subset\mathcal{W}_\pm$ is *uniformly bounded* if:

$$\sup_{W\in\mathcal{K}} \|W\|_\infty < +\infty. \tag{3.4}$$

Definition 3.3.10 (Invariance and smoothness of a distance on kernels)**.** *Let d be a distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$). We say that the distance d is:*

(i) ***Invariant****: if $d(U,W) = d(U^\varphi, W^\varphi)$ for every bijective measure-preserving map $\varphi \in S_{[0,1]}$ and $U, V \in \mathcal{W}_1$ (resp. U, V belongs to $\mathcal{W}_+$ or $\mathcal{W}_\pm$).*

(ii) ***Smooth****: if a.e. weak convergence implies convergence for d, that is, if $(W_n)_{n\in\mathbb{N}}$ and W are kernels from $\mathcal{W}_1$ (resp. kernels from $\mathcal{W}_+$ or $\mathcal{W}_\pm$ that are uniformly bounded and) such that for a.e. $(x,y) \in [0,1]^2$, $W_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$ as $n \to \infty$, then $\lim_{n\to\infty} d(W_n, W) = 0$.*

We say that a norm N on $\mathcal{W}_\pm$ is invariant (resp. smooth) if its associated distance d on $\mathcal{W}_\pm$ is invariant (resp. smooth).

We shall see in Section 3.3.8 some examples of distances d_{m} for which the associated cut distance $d_{\square,\mathrm{m}}$ is invariant and smooth. The invariance property from Definition 3.3.10 is always satisfied by the cut distance, and thus also by the cut norm.

Lemma 3.3.11 ($d_{\square,\mathrm{m}}$ is invariant)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$, resp. $\mathcal{M}_\pm(\mathbf{Z})$). Then the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$, resp. $\mathcal{W}_\pm$) is invariant.*

Proof. For a signed measure-valued kernel W, a bijective measure-preserving map $\varphi \in S_{[0,1]}$, and measurable sets $S, T \subset [0,1]$, we have thanks to (3.1):

$$\int_{S\times T} W^\varphi(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y = \int_{S\times T} W(\varphi(x),\varphi(y);\cdot)\,\mathrm{d}x\mathrm{d}y = \int_{\varphi(S)\times\varphi(T)} W(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y.$$

Hence, taking the supremum over every measurable sets $S, T \subset [0,1]$, we get that the cut distance $d_{\square,\mathrm{m}}$ is invariant. □

When a smooth distance on $\mathcal{W}_1$ or $\mathcal{W}_+$ derives from a distance on $\mathcal{M}_1(\mathbf{Z})$ or $\mathcal{M}_+(\mathbf{Z})$, we have the following result.

Lemma 3.3.12 (Smoothness and the weak topology)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$) such that the distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) is smooth. Then, the distance d_{m} is continuous w.r.t. the weak topology on $\mathcal{M}_1(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$).*

Proof. Let $(\mu_n)_{n\in\mathbb{N}}$, and μ be measures from $\mathcal{M}_1(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$) such that $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to μ. Consider the constant measure-valued graphons (resp. kernels) $W_n \equiv \mu_n$, $n \in \mathbb{N}$, and $W \equiv \mu$. Then, for every $x, y \in [0,1]$, $W_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$ as $n \to \infty$. As the distance $d_{\square,\mathrm{m}}$ is smooth, we get that $\lim_{n\to\infty} d_{\square,\mathrm{m}}(W_n, W) = 0$. Considering $S = T = [0,1]$ in the cut distance, we deduce that $\lim_{n\to\infty} d_{\mathrm{m}}(\mu_n, \mu) = 0$. □

The next lemma is a partial converse of Lemma 3.3.12, it gives sufficient conditions for $d_{\square,\mathrm{m}}$ to be smooth. Remind the definition of a quasi-convex distance in Definition 3.2.10.

Proposition 3.3.13 ($d_{\square,\mathrm{m}}$ is smooth)**.** *Let d_{m} be distance on $\mathcal{M}_\epsilon(\mathbf{Z})$ with $\epsilon \in \{+, \pm\}$ which is quasi-convex and sequentially continuous w.r.t. the weak topology (on $\mathcal{M}_\epsilon(\mathbf{Z})$). Then, the cut distance $d_{\square,\mathrm{m}}$ is smooth.*

Moreover, for all $U, W \in \mathcal{W}_\epsilon$, and for all measurable $A \subset [0,1]^2$, we have:

$$d_{\mathrm{m}}(U(A;\cdot), W(A;\cdot)) \leq \operatorname*{essup}_{(x,y)\in A} d_{\mathrm{m}}(U(x,y;\cdot), W(x,y;\cdot)). \tag{3.5}$$

To prove Proposition 3.3.13, we first need to prove the following lemma for approximation by $\mathcal{M}$-valued kernels taking finitely many values.

Lemma 3.3.14. *Let $W \in \mathcal{W}_\pm$ and a subset $A \subset [0,1]^2$. There exists a sequence $(W_n)_{n\in\mathbb{N}}$ in $\mathcal{W}_\pm$ such that $(W_n(A;\cdot))_{n\in\mathbb{N}}$ weakly converges to $W(A;\cdot)$ and for all $n \in \mathbb{N}$, W_n is finitely valued and takes its values in $\{W(x,y;\cdot) : (x,y) \in A\}$.*

Proof. By scaling, we may assume that $\|W\|_\infty \leq 1$. Let $(f_k)_{k\in\mathbb{N}}$ be a convergence determining sequence with $f_0 = \mathbb{1}$ and f_k takes values in $[0,1]$. Thus, for all $(x,y) \in [0,1]^2$, $\epsilon \in \{\pm 1\}$ and $k \in \mathbb{N}$, we have $W_\epsilon(x,y;f_k) \in [0,1]$. For all $n \in \mathbb{N}$, let $(C_{n,i})_{1\leq i\leq d_n}$ be a partition of $[0,1]^{2(n+1)}$ into $d_n = n^{2(n+1)}$ hypercubes of edge-length $r_n = 1/n$. Then, for all $n \in \mathbb{N}$ and $i \in [d_n]$, define $B_{n,i} = A \cap (W_+(\cdot;(f_i)_{0\leq i\leq n}, W_-(\cdot;(f_i)_{0\leq i\leq n})^{-1}(C_{n,i})$; thus we get a partition $(B_{n,i})_{1\leq i\leq d_n}$ of A. If $B_{n,i} \neq \emptyset$, fix some $\mu_{n,i} \in \{W(x,y;\cdot) : (x,y) \in B_{n,i}\}$. If $A \neq [0,1]^2$, fix some $\mu_\partial \in \{W(x,y;\cdot) : (x,y) \in [0,1]^2 \setminus A\}$. For $n \in \mathbb{N}$, we define $W_n = \mathbb{1}_{A^c}\ \mu_\partial + \sum_{i=1}^{d_n} \mathbb{1}_{B_{n,i}}\ \mu_{n,i}$, which is finitely valued and takes its values in $\{W(x,y;\cdot) : (x,y) \in A\}$.

Let $k \in \mathbb{N}$ and $\epsilon \in \{\pm\}$. For all $n \geq k$, we have:

$$|W_\epsilon(A;f_k) - (W_n)_\epsilon(A;f_k)| \leq \sum_{i=1}^{d_n} \int_{B_{n,i}} |W_\epsilon(x,y;f_k) - (\mu_{n,i})_\epsilon|\ \mathrm{d}x\mathrm{d}y \leq \frac{1}{n}\cdot$$

As $(f_k)_{k\in\mathbb{N}}$ is convergence determining, this implies that $((W_n)_\epsilon(A;\cdot))_{n\in\mathbb{N}}$ weakly converges to $W_\epsilon(A;\cdot)$ for $\epsilon \in \{\pm\}$. Hence, $(W_n(A;\cdot))_{n\in\mathbb{N}}$ weakly converges to $W(A;\cdot)$. □

Proof of Proposition 3.3.13. As d_{m} is quasi-convex, (3.5) is immediate when U and W take only finitely many values. Now, assume that U and W are arbitrary $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels. Let $\varepsilon > 0$. As d_{m} is sequentially continuous w.r.t. the weak topology, using Lemma 3.3.14, there exist two $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernel U' and W' such that $d_{\mathrm{m}}(U'(A;\cdot), U(A\cdot)) < \varepsilon$ and U' is finitely valued and takes its values in $\{U(x,y;\cdot) : (x,y) \in A\}$, and similarly for W' and W. Thus, we have:

$$d_{\mathrm{m}}(U(A;\cdot), W(A;\cdot)) \leq 2\varepsilon + \operatorname*{essup}_{(x,y)\in A} d_{\mathrm{m}}(U(x,y;\cdot), W(x,y;\cdot)),$$

and this being true for all $\varepsilon > 0$, we get (3.5).

Let $(W_n)_{n\in\mathbb{N}}$ and W be $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels which are uniformly bounded by some constant $C < \infty$ and such that for a.e. $(x,y) \in [0,1]^2$, the sequence $((W_n(x,y;\cdot))_{n\in\mathbb{N}}$ converges to $W(x,y;\cdot)$ for the weak topology, and thus also for d_{m}. Let $\varepsilon > 0$ and $S,T \subset [0,1]$. As d_{m} is quasi-convex and sequentially continuous w.r.t. the weak topology, using Lemma 3.2.11, there exists $\eta > 0$ such that for all $\mu,\nu \in \mathcal{M}_\epsilon(\mathbf{Z})$, we have that $\|\mu - \nu\|_\infty < \eta$ implies $d_{\mathrm{m}}(\mu,\nu) < \varepsilon$. For all $n \in \mathbb{N}$, define the measurable set:

$$A_n = \{(x,y) \in S \times T\ :\ d_{\mathrm{m}}(W_n(x,y;\cdot), W(x,y;\cdot)) < \varepsilon\}.$$

By assumption, we have that $\lim_{n\to\infty} \lambda(A_n) = \lambda(S\times T)$. Let $N \in \mathbb{N}$ be such that for $n \geq N$, we have $\lambda((S\times T)\setminus A_n) < \eta/C$. Let $n \geq N$. Remark that $W_n((S\times T)\setminus A_n;\cdot)$ and $W((S\times T)\setminus A_n;\cdot)$ have total mass at most $C\lambda(A_n^c) < \eta$. Thus, we have that $d_{\mathrm{m}}(W_n(A_n;\cdot), W_n(S\times T;\cdot)) < \varepsilon$ and $d_{\mathrm{m}}(W(A_n;\cdot), W(S\times T;\cdot)) < \varepsilon$. Hence, using (3.5) we get that:

$$\begin{aligned} d_{\mathrm{m}}(W_n(S\times T;\cdot), W(S\times T;\cdot)) &\leq 2\varepsilon + d_{\mathrm{m}}(W_n(A_n;\cdot), W(A_n;\cdot)) \\ &\leq 2\varepsilon + \operatorname*{essup}_{(x,y)\in A_n} d_{\mathrm{m}}(W_n(x,y;\cdot), W(x,y;\cdot)) \\ &\leq 3\varepsilon. \end{aligned}$$

Taking the supremum over $S,T \subset [0,1]$, we get $d_{\square,\mathrm{m}}(W_n,W) \leq 3\varepsilon$. This being true for all $\varepsilon > 0$, we conclude that $(W_n)_{n\in\mathbb{N}}$ converges to W for $d_{\square,\mathrm{m}}$, and thus $d_{\square,\mathrm{m}}$ is smooth. □

3.3.4 The unlabeled cut distance

We can now define the cut distance for unlabeled graphons.

Definition 3.3.15 (The unlabeled cut distance $\delta_{\square,\mathrm{m}}$)**.** *Set $\mathcal{K} \in \{\mathcal{W}_1, \mathcal{W}_+, \mathcal{W}_\pm\}$. Let d be an invariant distance on the kernel space $\mathcal{K}$. The premetric $\delta_\square$ on $\mathcal{K}$, also called the* cut distance*, is defined by:*

$$\delta_\square(U,W) = \inf_{\varphi\in S_{[0,1]}} d(U,W^\varphi) = \inf_{\varphi\in S_{[0,1]}} d\left(U^\varphi, W\right). \tag{3.6}$$

Notice that $\delta_\square$ satisfies the symmetry property (as d is invariant) and the triangular inequality. Hence, $\delta_\square$ induces a distance (that we still denote by $\delta_\square$) on the quotient space $\widetilde{\mathcal{K}}_d = \mathcal{K}/\sim_d$ of kernels in $\mathcal{K}$ associated with the equivalence relation $\sim_d$ defined by $U \sim_d W$ if and only if $\delta_\square(U,W) = 0$.

When the metric $d = d_{\square,\mathrm{m}}$ on $\mathcal{K} = \mathcal{W}_1$ (resp. $\mathcal{W}_+$, resp. $\mathcal{W}_\pm$) derives from a metric d_m on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$, resp. $\mathcal{M}_\pm(\mathbf{Z})$), and is thus invariant thanks to Lemma 3.3.11, we write $\delta_{\square,\mathrm{m}}$ for $\delta_\square$ and $\widetilde{\mathcal{K}}_\mathrm{m}$ for $\widetilde{\mathcal{K}}_{d_{\square,\mathrm{m}}}$. We shall see in Theorem 3.5.5 and Corollary 3.5.6 that under some conditions, different choices of distance d_m, which induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$, lead to the same quotient space, then simply denoted by $\widetilde{\mathcal{W}}_1$, with the same topology.

3.3.5 Weak isomorphism

Similarly to Theorem 8.13 in [Lov12], when the distance d_m is such that $d_{\square,\mathrm{m}}$ is invariant and smooth, we can rewrite the cut distance $\delta_{\square,\mathrm{m}}$ as a minimum instead of an infimum using measure-preserving maps, see the last equality in (3.7).

We introduce a weak isomorphism relation that allows to "un-label" probability-graphons.

Definition 3.3.16 (Weak isomorphism)**.** *We say that two signed measure-valued kernels U and W are* weakly isomorphic *(and we note $U \sim W$) if there exists two measure-preserving maps $\varphi, \psi \in \bar{S}_{[0,1]}$ such that $U^\varphi(x,y;\cdot) = W^\psi(x,y;\cdot)$ for a.e. $x, y \in [0,1]$.*

We denote by $\widetilde{\mathcal{W}}_\pm = \mathcal{W}_\pm/\sim$ (resp. $\widetilde{\mathcal{W}}_1 = \mathcal{W}_1/\sim$) the space of unlabeled signed measure-valued kernels (resp. probability-graphons) i.e. the space of signed measure-valued kernels (resp. probability-graphons) where we identify signed measure-valued kernels (resp. probability-graphons) that are weakly isomorphic.

Notice that $U \sim W$ implies that $\|U\|_\infty = \|W\|_\infty$ (we recall that signed measure-valued kernels are only defined for a.e. $x, y \in [0,1]$ and that $\|W\|_\infty$ in (3.2) is an essup in general). In particular, the notion of uniformly bounded subset defined in (3.4) naturally extends to $\widetilde{\mathcal{W}}_\pm$. The last part of this section is devoted to the proof of the following key result.

Theorem 3.3.17 (Weak isomorphism and $\delta_\square$)**.** *Let d be a distance defined on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) which is invariant and smooth. Then, two kernels are weakly isomorphic, i.e. $U \sim W$, if and only if $U \sim_d W$,* i.e. *$\delta_\square(U,W) = 0$.*

Furthermore, the map $\delta_\square$ is a distance on $\widetilde{\mathcal{W}}_1 = \widetilde{\mathcal{W}}_{1,d}$ (resp. $\widetilde{\mathcal{W}}_+ = \widetilde{\mathcal{W}}_{+,d}$ or $\widetilde{\mathcal{W}}_\pm = \widetilde{\mathcal{W}}_{\pm,d}$).

As a first step in the proof of Theorem 3.3.17, following [Lov12], we give a nice description of $\delta_\square$ using couplings. We say that a measure μ on $[0,1]^2$ is a coupling measure on $[0,1]^2$ (between two copies of $[0,1]$ each equipped with the Lebesgue measure) if the projection maps on each components $\tau, \rho : [0,1]^2 \to [0,1]$ (where $[0,1]^2$ is equipped with the measure μ and $[0,1]$ with the Lebesgue measure λ) are measure-preserving. Thus for every kernel W on $([0,1], \mathcal{B}([0,1]), \lambda)$, the function W^τ is a kernel on the probability space $([0,1]^2, \mathcal{B}([0,1]^2), \mu)$, and similarly for the projection ρ.

Let φ be a given measure-preserving map from $[0,1]$ with the Lebesgue measure to $[0,1]^2$ with a coupling measure μ. For an invariant distance d on $\mathcal{W}_1$ (resp. $\mathcal{W}_\pm$), we define a distance, say d^μ, on kernels on $([0,1]^2, \mathcal{B}([0,1]^2), \mu)$ by:

$$d^\mu(U', W') = d(U'^\varphi, W'^\varphi).$$

It is easy to see that, for U and W kernels on $[0,1]$, we have $d^\mu(U^\tau, W^\tau) = d(U,W)$ as d is invariant and $\tau \circ \varphi$ is a measure-preserving map from $[0,1]$ to itself; and similarly $d^\mu(U^\rho, W^\rho) = d(U,W)$.

A straightforward adaptation of the proof of [Lov12, Theorem 8.13] gives the next result.

Proposition 3.3.18 (Minima in the cut distance $\delta_\square$)**.** *Let d be a distance defined on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) which is invariant and smooth. Then, we have the following alternative formulations for the cut distance $\delta_\square$ on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$):*

$$\begin{aligned}\delta_\square(U,W) &= \inf_{\varphi \in S_{[0,1]}} d(U, W^\varphi) = \inf_{\varphi \in \bar{S}_{[0,1]}} d(U, W^\varphi) \\ &= \inf_{\psi \in S_{[0,1]}} d(U^\psi, W) = \inf_{\psi \in \bar{S}_{[0,1]}} d(U^\psi, W) \\ &= \inf_{\varphi, \psi \in S_{[0,1]}} d(U^\psi, W^\varphi) = \min_{\varphi, \psi \in \bar{S}_{[0,1]}} d(U^\psi, W^\varphi),\end{aligned} \tag{3.7}$$

and

$$\delta_{\square}(U,W) = \min_{\mu} d^{\mu}\left(U^{\tau}, W^{\rho}\right) \tag{3.8}$$

where μ range over all coupling measures on $[0,1]^2$.

Proof of Theorem 3.3.17. We deduce from the last equality in (3.7) that $\delta_{\square}(U,W) = 0$ if and only if there exist measure-preserving maps $\varphi, \psi \in \tilde{S}_{[0,1]}$ such that $U^{\psi}(x,y;\cdot) = W^{\varphi}(x,y;\cdot)$ for a.e. $x, y \in [0,1]$. This gives that the equivalence relations $\sim_d$ and $\sim$ are the same. □

3.3.6 The cut norm for stepfunctions

For a quasi-convex distance d_{m}, the cut distance d_{m} for stepfunctions can be reformulated using a finite combinatorial optimization. For a collection of subsets $\mathcal{P}$, denote by $\sigma(\mathcal{P})$ the σ-field generated by $\mathcal{P}$.

Lemma 3.3.19 (Combinatorial optimization of quasi-convex d_{m} for stepfunctions)**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}$ a convex subset of $\mathcal{M}_{\pm}(\mathbf{Z})$ containing the zero measure. Let $U, W \in \mathcal{W}_{\mathcal{M}}$ be $\mathcal{M}$-valued stepfunctions adapted to the same finite partition $\mathcal{P}$. Then, there exists $S, T \in \sigma(\mathcal{P})$ such that:*

$$d_{\square,\mathrm{m}}(U,W) = d_{\mathrm{m}}(U(S\times T;\cdot), W(S\times T;\cdot)).$$

Proof. Let $\mathcal{P} = \{S_1, \dots, S_k\}$ with $k = |\mathcal{P}|$ the size of the partition $\mathcal{P}$. First, remark that the quantity $d_{\square,\mathrm{m}}(U,W) = d_{\mathrm{m}}(U(S'\times T';\cdot), W(S'\times T';\cdot))$ depends on S' and T' only through the values of $\lambda(S'\cap S_i)$ and $\lambda(T'\cap S_i)$ for $1 \le i \le k$. Thus, the cut distance between U and W can be reformulated as:

$$d_{\square,\mathrm{m}}(U,W) = \sup_{0\le\alpha_i,\beta_i\le\lambda(S_i);\, 1\le i\le k} d_{\mathrm{m}}\left(\sum_{1\le i,j\le k} \alpha_i\beta_j\, \mu_{i,j}(\cdot), \sum_{1\le i,j\le k} \alpha_i\beta_j\, \nu_{i,j}(\cdot)\right),$$

where $\mu_{i,j}$ (resp. $\nu_{i,j}$) is the constant value of $U(x,y;\cdot)$ (resp. $W(x,y;\cdot)$) when $x \in S_i$ and $y \in S_j$. Moreover, when we fix the value of $\beta = (\beta_i)_{1\le i\le k}$, the quantity

$$d_{\mathrm{m}}\left(\sum_{1\le i,j\le k} \alpha_i\beta_j\, \mu_{i,j}(\cdot), \sum_{1\le i,j\le k} \alpha_i\beta_j\, \nu_{i,j}(\cdot)\right)$$

is a quasi-convex function of $\alpha = (\alpha_i)_{1\le i\le k}$, and thus realizes its maximum on the extremal points of the hypercube $\prod_{i=1}^{k}[0,\lambda(S_i)]$, i.e. when α_i equals 0 or $\lambda(S_i)$ for every $1 \le i \le k$. By symmetry, a similar argument holds for β. The cut distance can thus be reformulated as the combinatorial optimization:

$$d_{\square,\mathrm{m}}(U,W) = \max_{\mathrm{I},\mathrm{J}\subset[k]} d_{\mathrm{m}}\left(\sum_{i\in\mathrm{I},j\in\mathrm{J}} \mu_{i,j}(\cdot), \sum_{i\in\mathrm{I},j\in\mathrm{J}} \nu_{i,j}(\cdot)\right).$$

Let $I, J \subset [k]$ that maximizes this combinatorial optimization, and take $S = \cup_{i\in I} S_i$ and $T = \cup_{j\in J} S_j$ to conclude. □

3.3.7 The supremum in S and T in the cut distance $d_{\square,\mathrm{m}}$

In this section, we prove that the supremum in the cut distance $d_{\square,\mathrm{m}}$ is achieved by some subsets $S, T \subset [0,1]$.

For $W \in \mathcal{M}_{\pm}(\mathbf{Z})$ and $f, g : [0,1] \to [0,1]$ measurable, we define the signed measure:

$$W(f\otimes g;\cdot) = \int_{[0,1]^2} W(x,y;\cdot) f(x) g(y)\, \mathrm{d}x\mathrm{d}y.$$

Remark that if we have $W \in \mathcal{W}_{\epsilon}$ with $\epsilon \in \{1, \le 1, +, \pm\}$, then we have $W(f\otimes g;\cdot) \in \mathcal{M}_{\epsilon}(\mathbf{Z})$.

Lemma 3.3.20 (The supremum in the cut distance $d_{\square,\mathrm{m}}$ for quasi-convex distance d_{m})**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}_\epsilon(\mathbf{Z})$ with $\epsilon \in \{+, \pm\}$ that is sequentially continuous w.r.t. the weak topology. Let $U, W \in \mathcal{W}_\epsilon$. Then, there exist measurable subsets $S, T \subset [0,1]$ such that $f = \mathbb{1}_S$ and $g = \mathbb{1}_T$ achieve the supremum in:*

$$\sup_{f,g} d_{\mathrm{m}}\Big(U(f \otimes g; \cdot), W(f \otimes g; \cdot)\Big)$$

where the supremum is taken over measurable functions f, g from $[0,1]$ to itself.

Proof. Define the map $\Psi : (f,g) \mapsto d_{\mathrm{m}}(U(f \otimes g; \cdot), W(f \otimes g; \cdot))$, and denote $C = \sup_{f,g} \Psi(f,g)$, where the supremum is taken over measurable functions f, g from $[0,1]$ to itself. Let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be sequences of measurable functions from $[0,1]$ to itself such that $\lim_{n\to\infty} \Psi(f_n, g_n) = C$. As the unit ball of $L^\infty([0,1], \lambda)$ is compact for the weak-$*$ topology (with primal space $L^1([0,1], \lambda)$), upon taking subsequences, we may assume that $(f_n)_{n\in\mathbb{N}}$ (resp. $(g_n)_{n\in\mathbb{N}}$) weak-$*$ converges to some f (resp. g) which take values in $[0,1]$. Thus, $(f_n \otimes g_n)_{n\in\mathbb{N}}$ weak-$*$ converges to $f \otimes g$ in $L^\infty([0,1]^2, \lambda_2)$. In particular, for every $h \in C_b(\mathbf{Z})$, as $W[h]$ is a real-valued kernel, this implies that $\lim_{n\to\infty} W(f_n \otimes g_n; h) = W(f \otimes g; h)$. This being true for every $h \in C_b(\mathbf{Z})$, we get that the sequence $(W(f_n \otimes g_n; \cdot))_{n\in\mathbb{N}}$ in $\mathcal{M}_\epsilon(\mathbf{Z})$ weakly converges to $W(f \otimes g; \cdot) \in \mathcal{M}_\epsilon(\mathbf{Z})$; and similarly for U. As d_{m} is sequentially continuous w.r.t. the weak topology on $\mathcal{M}_\epsilon(\mathbf{Z})$, we get that $C = \lim_{n\to\infty} \Psi(f_n, g_n) = \Psi(f, g)$.

Now, we show that we can replace the functions f and g by functions that only take the values 0 and 1 (i.e. indicator functions). We first fix g and do this for f. Let X be a random variable uniformly distributed over $[0,1]$, and consider the random function $\mathbb{1}_{X\le f}$. Remark that we have $\mathbb{E}[W(\mathbb{1}_{X\le f} \otimes g; \cdot)] = W(f \otimes g; \cdot)$, and similarly for U. As d_{m} is quasi-convex and sequentially continuous w.r.t. the weak topology, we have:

$$\begin{aligned} C &\ge \sup_{x\in[0,1]} d_{\mathrm{m}}(U(\mathbb{1}_{x\le f} \otimes g; \cdot), W(\mathbb{1}_{x\le f} \otimes g; \cdot)) \\ &\ge d_{\mathrm{m}}\Big(\mathbb{E}[U(\mathbb{1}_{X\le f} \otimes g; \cdot)], \mathbb{E}[W(\mathbb{1}_{X\le f} \otimes g; \cdot)]\Big) \\ &= d_{\mathrm{m}}(U(f \otimes g; \cdot), W(f \otimes g; \cdot)) \\ &= C, \end{aligned}$$

where in the second equality we used the quasi-convex supremum inequality from (3.5) with the $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels $U'(x, y; \cdot) = U(\mathbb{1}_{x\le f} \otimes g; \cdot)$ and $W'(x, y; \cdot) = W(\mathbb{1}_{x\le f} \otimes g; \cdot)$, and $A = [0,1]^2$. All inequalities being equalities, this imposes:

$$C = \sup_{x\in[0,1]} d_{\mathrm{m}}(U(\mathbb{1}_{x\le f} \otimes g; \cdot), W(\mathbb{1}_{x\le f} \otimes g; \cdot)) = \lim_{n\to\infty} d_{\mathrm{m}}(U(\mathbb{1}_{r_n\le f} \otimes g; \cdot), W(\mathbb{1}_{r_n\le f} \otimes g; \cdot)),$$

for some sequence $(x_n)_{n\in\mathbb{N}}$ in $[0,1]$. Upon taking a subsequence, we may assume that the sequence $(x_n)_{n\in\mathbb{N}}$ monotonically converges to some $x \in [0,1]$. In particular, the sequence of functions $(\mathbb{1}_{x_n\le f})_{n\in\mathbb{N}}$ (monotonically) converges to the function $f' = \mathbb{1}_{x\le f}$ (resp. $f' = \mathbb{1}_{x<f}$) if $(x_n)_{n\in\mathbb{N}}$ is non-decreasing (resp. decreasing), and thus also weak-$*$ converges in $L^\infty([0,1], \lambda)$. Using, as in the first part of the proof, the sequential continuity of the function Ψ w.r.t. the weak-$*$ topology on $L^\infty([0,1], \lambda)$, we get that $\Psi(f', g) = d_{\mathrm{m}}(U(f' \otimes g; \cdot), W(f' \otimes g; \cdot)) = C$, that is we can replace f by the indicator function f'. The same argument allows to replace g by an indicator function. □

3.3.8 Examples of distance d_{m}

We consider usual distances and norms on $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$ that induce the weak topology on $\mathcal{M}_+(\mathbf{Z})$. All the distances we consider are quasi-convex, and all the norms we consider are sequentially continuous w.r.t. the weak topology on $\mathcal{M}_\pm(\mathbf{Z})$. Thus their associated cut distances are invariant and smooth by Lemma 3.3.11 and Proposition 3.3.13. Properties for the cut distances associated with those distances and norms are summarized in Corollaries 3.4.14 and 3.5.6.

In this section, we assume that $(\mathbf{Z}, d_0)$ is a Polish metric space, and remind that $\mathcal{B}(\mathbf{Z})$ denotes its Borel σ-field.

The Prohorov distance $d_{\mathcal{P}}$

The Prohorov distance $d_{\mathcal{P}}$ is a complete distance defined on the set of finite measures $\mathcal{M}_+(\mathbf{Z})$ that induces the weak topology (see [Bil99a, Theorem 6.8]). It is defined for $\mu, \nu \in \mathcal{M}_+(\mathbf{Z})$ as:

$$d_{\mathcal{P}}(\mu,\nu) = \inf\{\varepsilon > 0 : \forall A \in \mathcal{B}(\mathbf{Z}),\ \mu(A) \leq \nu(A^\varepsilon) + \varepsilon \quad \text{and} \quad \nu(A) \leq \mu(A^\varepsilon) + \varepsilon\}, \tag{3.9}$$

where $A^\varepsilon = \{x \in \mathbf{Z} : \exists y \in A, d_0(x,y) < \varepsilon\}$. For probability measures, we only need one inequality in (3.9) to define the Prohorov distance; however for positive measures we need both inequalities as two arbitrary positive measures might not have the same total mass. For $d_{\mathrm{m}} = d_{\mathcal{P}}$, we use the subscript $\mathrm{m} = \mathcal{P}$. We now prove that the Prohorov distance is quasi-convex.

Lemma 3.3.21. *The Prohorov distance $d_{\mathcal{P}}$ is quasi-convex on $\mathcal{M}_+(\mathbf{Z})$.*

Proof. Let $\mu_1, \mu_2, \nu_1, \nu_2 \in \mathcal{M}_+(\mathbf{Z})$ and let $\alpha \in [0,1]$. Let $\varepsilon > \max(d_{\mathcal{P}}(\mu_1,\nu_1), d_{\mathcal{P}}(\mu_2,\nu_2))$, then for all $i \in \{1,2\}$ and $B \in \mathcal{B}(\mathbf{Z})$, we have that $\mu_i(B) \leq \nu_i(B^\varepsilon) + \varepsilon$ and $\nu_i(B) \leq \mu_i(B^\varepsilon) + \varepsilon$. Taking a linear combination of those inequalities, we get that for all $B \in \mathcal{B}(\mathbf{Z})$, we have that $\alpha\mu_1(B) + (1-\alpha)\mu_2(B) \leq \alpha\nu_1(B^\varepsilon) + (1-\alpha)\nu_2(B^\varepsilon) + \varepsilon$, and similarly when swapping the role (μ_1,μ_2) and (ν_1,ν_2). Hence, we get that $d_{\mathcal{P}}(\alpha\mu_1 + (1-\alpha)\mu_2, \alpha\nu_1 + (1-\alpha)\nu_2) \leq \varepsilon$, and taking the infimum over ε, we get that $d_{\mathrm{m}}(\alpha\mu_1 + (1-\alpha)\mu_2, \alpha\nu_1 + (1-\alpha)\nu_2) \leq \max(d_{\mathrm{m}}(\mu_1,\nu_1), d_{\mathrm{m}}(\mu_2,\nu_2))$. □

The Kantorovitch-Rubinshtein and Fortet-Mourier norms

The Kantorovitch-Rubinshtein norm $\|\cdot\|_{\mathrm{KR}}$ (sometimes also called the bounded Lipschitz distance) and the Fortet-Mourier norm $\|\cdot\|_{\mathrm{FM}}$ are two norms defined on $\mathcal{M}_\pm(\mathbf{Z})$ that induce the weak topology on $\mathcal{M}_+(\mathbf{Z})$ (see Section 3.2 in [Bog18] for definition and properties of those norms). They are defined for $\mu \in \mathcal{M}_\pm(\mathbf{Z})$ by:

$$\|\mu\|_{\mathrm{KR}} = \sup\left\{\int_{\mathbf{Z}} f \,\mathrm{d}\mu : f \text{ is 1-Lipschitz and } \|f\|_\infty \leq 1\right\},$$
$$\|\mu\|_{\mathrm{FM}} = \sup\left\{\int_{\mathbf{Z}} f \,\mathrm{d}\mu : f \text{ is Lipschitz and } \|f\|_\infty + \mathrm{Lip}(f) \leq 1\right\},$$

where $\|f\|_\infty = \sup_{x\in\mathbf{Z}} |f(x)|$ is the infinite norm and $\mathrm{Lip}(f)$ is the smallest constant $L > 0$ such that f is L-Lipschitz. Those two norms are metrically equivalent, see beginning of Section 3.2 in [Bog18]:

$$\|\mu\|_{\mathrm{FM}} \leq \|\mu\|_{\mathrm{KR}} \leq 2\|\mu\|_{\mathrm{FM}}. \tag{3.10}$$

Note that we have $\|\mu\|_{\mathrm{KR}} \leq \|\mu\|_\infty$, and thus those two norms are sequentially continuous w.r.t. the weak topology on $\mathcal{M}_\pm(\mathbf{Z})$.

An easy adaptation of the proof for Theorem 3.2.2 in [Bog18] gives the following comparison between $d_{\mathcal{P}}$, $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$.

Lemma 3.3.22 (Comparison of $d_{\mathcal{P}}$, $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$)**.** *Let $\mu, \nu \in \mathcal{M}_+(\mathbf{Z})$. Then, we have:*

$$\frac{d_{\mathcal{P}}(\mu,\nu)^2}{1 + d_{\mathcal{P}}(\mu,\nu)} \leq \|\mu - \nu\|_{FM} \leq \|\mu - \nu\|_{KR} \leq \big(2 + \min(\mu(\mathbf{Z}), \nu(\mathbf{Z}))\big)\ d_{\mathcal{P}}(\mu,\nu).$$

In particular, the Prohorov distance $d_{\mathcal{P}}$ is uniformly continuous w.r.t. $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$ on $\mathcal{M}_+(\mathbf{Z})$; and $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$ are uniformly continuous w.r.t. $d_{\mathcal{P}}$ on $\mathcal{M}_{\leq 1}(\mathbf{Z})$.

For the special choice $N_{\mathrm{m}} = \|\cdot\|_{\mathrm{KR}}$ (resp. $N_{\mathrm{m}} = \|\cdot\|_{\mathrm{FM}}$), we use the subscript $\mathrm{m} = \mathrm{KR}$ (resp. $\mathrm{m} = \mathrm{FM}$).

A norm based on a convergence determining sequence

From a convergence determining sequence $\mathcal{F} = (f_k)_{k\in\mathbb{N}}$, where $f_0 = \mathbb{1}$ and $f_k \in C_b(\mathbf{Z})$ takes values in $[0,1]$, we define a norm on $\mathcal{M}_\pm(\mathbf{Z})$ metrizing the weak topology on $\mathcal{M}_+(\mathbf{Z})$, for $\mu \in \mathcal{M}_\pm(\mathbf{Z})$, by:

$$\|\mu\|_{\mathcal{F}} = \sum_{k\in\mathbb{N}} 2^{-k} |\mu(f_k)|. \tag{3.11}$$

Note that we have $\|\mu\|_{\mathcal{F}} \leq 2\|\mu\|_\infty$, and thus $\|\cdot\|_{\mathcal{F}}$ is sequentially continuous w.r.t. the weak topology on $\mathcal{M}_\pm(\mathbf{Z})$. For the special choice $N_{\mathrm{m}} = \|\cdot\|_{\mathcal{F}}$, we use the subscript $\mathrm{m} = \mathcal{F}$.

Even though the norm $\|\cdot\|_{\mathcal{F}}$ is not complete when $\mathbf{Z}$ is not compact (see Lemma 3.3.23 below), the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ and the cut distance $\delta_{\square,\mathcal{F}}$ will turn out to be very useful in Sections 3.6 and 3.7 to link the topology of the cut distance to the homomorphism densities. Recall $d_{\mathcal{F}}$ is the distance derived from the norm $\|\cdot\|_{\mathcal{F}}$.

Lemma 3.3.23 ($d_{\mathcal{F}}$ is not complete in general)**.** *Let $\mathcal{F}$ be a convergence determining sequence. Then, the distance $d_{\mathcal{F}}$ is complete over $\mathcal{M}_1(\mathbf{Z})$ if and only if $\mathcal{M}_1(\mathbf{Z})$ is a compact space,* i.e. *, if and only if $\mathbf{Z}$ is compact.*

Proof. Theorem 3.4 in [Var58] states that $\mathbf{Z}$ is compact if and only if $\mathcal{M}_1(\mathbf{Z})$ is compact. When this is the case, any distance metrizing the weak topology on $\mathcal{M}_1(\mathbf{Z})$ is complete.

Reciprocally, assume that $d_{\mathcal{F}}$ is a complete metric over $\mathcal{M}_1(\mathbf{Z})$ and write $\mathcal{F} = (f_m)_{m\in\mathbb{N}}$. Let $(\mu_n)_{n\in\mathbb{N}}$ be an arbitrary sequence of probability measures from $\mathcal{M}_1(\mathbf{Z})$. For every $m \in \mathbb{N}$, as f_m takes values in $[0,1]$, we have for every $n \in \mathbb{N}$ that $\mu_n(f_m) \in [0,1]$. Hence, using a diagonal extraction, there exists a subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ of the sequence $(\mu_n)_{n\in\mathbb{N}}$ such that for every $m \in \mathbb{N}$, the sequence $(\mu_{n_k}(f_m))_{k\in\mathbb{N}}$ converges, that is, $(\mu_{n_k})_{k\in\mathbb{N}}$ is a Cauchy sequence for the distance $d_{\mathcal{F}}$. As we assumed the distance $d_{\mathcal{F}}$ to be complete, this implies that the sequence $(\mu_n)_{n\in\mathbb{N}}$ has a convergent subsequence. The sequence $(\mu_n)_{n\in\mathbb{N}}$ being arbitrary, we conclude that the space $\mathcal{M}_1(\mathbf{Z})$ is sequentially compact, and thus compact by Remark 3.2.6. □

For $W \in \mathcal{W}_\pm$ and $f \in C_b(\mathbf{Z})$, we denote by $W[f]$ the real-valued kernel defined by:

$$W[f](x,y) = W(x,y;f) = \int_{\mathbf{Z}} f(z)\ W(x,y;\mathrm{d}z). \tag{3.12}$$

We denote by $\|\cdot\|_{\square,\mathbb{R}}$ (resp. $\|\cdot\|^+_{\square,\mathbb{R}}$) the cut norm (resp. one-sided version of the cut norm) for real-valued kernels defined as:

$$\|w\|_{\square,\mathbb{R}} = \sup_{S,T\subset[0,1]} \left|\int_{S\times T} w(x,y)\ \mathrm{d}x\mathrm{d}y\right| \quad \text{and} \quad \|w\|^+_{\square,\mathbb{R}} = \sup_{S,T\subset[0,1]} \int_{S\times T} w(x,y)\ \mathrm{d}x\mathrm{d}y, \tag{3.13}$$

where w is a real-valued kernel w (see [Lov12, Section 8.2], resp. [Lov12, Section 10.3], for definition and properties of those objects).

The following two remarks link the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ of a signed measure-valued kernel W with the cut norm $\|\cdot\|_{\square,\mathbb{R}}$ of the real-valued kernels $W[f]$ for some particular choices of functions $f \in C_b(\mathbf{Z})$. We will reuse those facts in Section 3.6.

Remark 3.3.24 (Link between $\|\cdot\|_{\square,\mathcal{F}}$ and $\|\cdot\|^+_{\square,\mathbb{R}}$)**.** For $\mu \in \mathcal{M}_\pm(\mathbf{Z})$ we have:

$$\|\mu\|_{\mathcal{F}} = \sup_{\varepsilon\in\{\pm1\}^{\mathbb{N}}} \sum_{n\in\mathbb{N}} 2^{-n}\varepsilon_n\mu(f_n) = \sup_{\varepsilon\in\{\pm1\}^{\mathbb{N}}} \mu\left(\sum_{n\in\mathbb{N}} 2^{-n}\varepsilon_n f_n\right), \tag{3.14}$$

with $\varepsilon = (\varepsilon_n)_{n\in\mathbb{N}}$. Hence, for a signed measure-valued kernel $W \in \mathcal{W}_\pm$, we have:

$$\|W\|_{\square,\mathcal{F}} = \sup_{\varepsilon\in\{\pm1\}^{\mathbb{N}}} \sup_{S,T\subset[0,1]} W\left(S\times T; \sum_{n\in\mathbb{N}} 2^{-n}\varepsilon_n f_n\right) = \sup_{\varepsilon\in\{\pm1\}^{\mathbb{N}}} \left\|W\left[\sum_{n\in\mathbb{N}} 2^{-n}\varepsilon_n f_n\right]\right\|^+_{\square,\mathbb{R}}. \tag{3.15}$$

Remark 3.3.25 (Inequality with $\|\cdot\|_{\square,\mathcal{F}}$ and $\|\cdot\|_{\square,\mathbb{R}}$)**.** For a signed measure-valued kernel W, we have:

$$\begin{aligned}
\|W\|_{\square,\mathcal{F}} &= \sup_{S,T\subset[0,1]} \sum_{n=0}^{\infty} 2^{-n} \left|\int_{S\times T} W(x,y,f_n)\ \mathrm{d}x\mathrm{d}y\right| \\
&\leq \sum_{n=0}^{\infty} 2^{-n} \sup_{S,T\subset[0,1]} \left|\int_{S\times T} W(x,y,f_n)\ \mathrm{d}x\mathrm{d}y\right| \\
&= \sum_{n=0}^{\infty} 2^{-n} \|W[f_n]\|_{\square,\mathbb{R}}.
\end{aligned} \tag{3.16}$$

3.4 Tightness and weak regularity

In this section, using a conditional expectation approach as in [Lov12, Chapter 9], we provide approximations of signed measure-valued kernels and probability-graphons by stepfunctions with an explicit bound on the quality of the approximation. This procedure takes into account that signed measure-valued kernels are infinite-dimensional valued.

3.4.1 Approximation by stepfunctions

We start by introducing the partitioning of a signed measure-valued kernel.

Definition 3.4.1 (The stepping operator)**.** *Let $W \in \mathcal{W}_\pm$ be a signed measure-valued kernel and $\mathcal{P} = \{S_1, \cdots, S_k\}$ be a finite partition of $[0,1]$. We define the kernel stepfunction $W_\mathcal{P}$ adapted to the partition $\mathcal{P}$ by averaging W over the partition subsets:*

$$W_\mathcal{P}(x,y;\cdot) = \frac{1}{\lambda(S_i)\lambda(S_j)} W(S_i \times S_j;\cdot) \qquad \text{for } x \in S_i, y \in S_j,$$

when $\lambda(S_i) \neq 0$ and $\lambda(S_j) \neq 0$, and $W_\mathcal{P}(x,y;\cdot) = 0$ the null measure otherwise. We call the map $W \mapsto W_\mathcal{P}$ defined on $\mathcal{W}_\pm$ the stepping operator (associated with the finite partition $\mathcal{P}$).

Since the signed measure-valued kernel are defined up to an a.e. equivalence, the value of $W_\mathcal{P}(x,y;\cdot)$ for $x \in S_i, y \in S_j$ when $\lambda(S_i)\lambda(S_j)$ is unimportant.

Remark 3.4.2 (Link with conditional expectation)**.** The stepfunction $W_\mathcal{P}$ can be viewed as the conditional expectation of W w.r.t. the (finite) sigma-field $\sigma(\mathcal{P} \times \mathcal{P})$ on $[0,1]^2$, where $W : [0,1]^2 \to \mathcal{M}_\pm(\mathbf{Z})$ is seen as a random signed measure in $\mathcal{M}_\pm(\mathbf{Z})$ and the probability measure on $[0,1]^2$ is the Lebesgue measure.

Remark 3.4.3 (Steppings are convex stable)**.** Let $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$ be a convex subset of measures, for instance $\mathcal{M}$ is $\mathcal{M}_1(\mathbf{Z})$, $\mathcal{M}_{\leq 1}(\mathbf{Z})$, $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$. Whenever $W \in \mathcal{W}_\pm$ is a $\mathcal{M}$-valued kernel, then by simple computation its stepping $W_\mathcal{P}$ is also a $\mathcal{M}$-valued kernel.

In the following remark, we give a characterization of refining partitions that generate the Borel σ-field of $[0,1]$.

Remark 3.4.4 (On refining partitions that generates the Borel σ-field)**.** Let $(\mathcal{P}_k)_{k\in\mathbb{N}}$ be a sequence of refining partitions of $[0,1]$. It generates the Borel σ-field of $[0,1]$ (that is, $\{S : S \in \mathcal{P}_k,\, k \in \mathbb{N}\}$ generates the Borel σ-field of $[0,1]$) if and only if $(\mathcal{P}_k)_{k\in\mathbb{N}}$ separates points (that is, for every distinct $x, y \in [0,1]$, there exists $k \in \mathbb{N}$ such that x and y belong to different classes of $\mathcal{P}_k$).

Indeed, assume that $(\mathcal{P}_k)_{k\in\mathbb{N}}$ separates points, and consider the countable family of Borel-measurable functions $\mathcal{F} = \{\mathbb{1}_S : S \in \mathcal{P}_k,\, k \in \mathbb{N}\}$ which separates points. Thus, by [Bog07b, Theorem 6.8.9] (remark that a Polish space is a Souslin space, see [Bog07b, Definition 6.6.1]), the family $\mathcal{F}$ generates the Borel σ-field of $[0,1]$. This implies that the family of Borel sets $\{S : S \in \mathcal{P}_k,\, k \in \mathbb{N}\}$ generates the Borel σ-field of $[0,1]$.

Conversely, assume there exist $x, y \in [0,1]$ which are not separated by $(\mathcal{P}_k)_{k\in\mathbb{N}}$, i.e. for all $k \in \mathbb{N}$, x and y belong to the same class of $\mathcal{P}_k$. This implies that the set $\{x\}$ does not belong to the σ-field generated by $(\mathcal{P}_k)_{k\in\mathbb{N}}$, and thus $(\mathcal{P}_k)_{k\in\mathbb{N}}$ does not generate the Borel σ-field of $[0,1]$.

Recall the definition of the norm $\|\cdot\|_\infty$ on $\mathcal{W}_\pm$ defined in (3.2). The following lemma allows to approximate any signed measure-valued kernel by its steppings.

Lemma 3.4.5 (Approximation using the stepping operator)**.** *Let $W \in \mathcal{W}_\pm$ be a signed measure-valued kernel (which is bounded by definition). Let $(\mathcal{P}_n)_{n\in\mathbb{N}}$ be a refining sequence of finite partitions of $[0,1]$ that generates the Borel σ-field on $[0,1]$. Then, the sequence $(W_{\mathcal{P}_n})_{n\in\mathbb{N}}$ is uniformly bounded by $\|W\|_\infty$, and weakly converges to W almost everywhere (on $[0,1]^2$).*

Proof. Set $W_n = W_{\mathcal{P}_n}$ for $n \in \mathbb{N}$. By definition of the stepping operator, we have for every $n \in \mathbb{N}$ and every $(x,y) \in [0,1]^2$ that the total mass of $W_n(x,y;\cdot)$ is upper bounded by $\|W\|_\infty$.

Recall that for $W \in \mathcal{W}_\pm$ and $f \in C_b(\mathbf{Z})$, the real-valued kernel $W[f]$ is defined by (3.12). First assume that $W \in \mathcal{W}_+$. Let $\mathcal{F} = (f_k)_{k\in\mathbb{N}}$ be a convergence determining sequence, with by convention $f_0 = \mathbb{1}$. For

every $k \in \mathbb{N}$ and $n \in \mathbb{N}$, an immediate computation gives $W_n[f_k] = (W[f_k])_{\mathcal{P}_n}$. For every $k \in \mathbb{N}$, as $W[f_k]$ is a real-valued kernel, we can apply the closed martingale theorem (as $(W[f_k])_{\mathcal{P}_n}$ can be viewed as a conditional expectation, see Remark 3.4.2), and we get that $\lim_{n\to\infty} W_n[f_k] = W[f_k]$ almost everywhere, since $(\mathcal{P}_n)_{n\in\mathbb{N}}$ generates the Borel σ-field. Hence, as the sequence $(f_k)_{k\in\mathbb{N}}$ is convergence determining, the sequence $(W_n)_{n\in\mathbb{N}}$ weakly converges to W almost everywhere.

Now, for $W \in \mathcal{W}_\pm$, write $W = W^+ - W^-$ where $W^+, W^- \in \mathcal{W}_+$ (see Lemma 3.3.3). By linearity of the stepping operator, remark that we have $W_n = (W^+)_{\mathcal{P}_n} - (W^-)_{\mathcal{P}_n}$ for all $n \in \mathbb{N}$. By the first case, we have that the sequence $((W^+)_{\mathcal{P}_n})_{n\in\mathbb{N}}$ weakly converges a.e. to W^+, and similarly for $((W^-)_{\mathcal{P}_n})_{n\in\mathbb{N}}$ and W^-. Hence, the sequence $(W_n)_{n\in\mathbb{N}}$ weakly converges to W almost everywhere. □

We first provide a separability result on the space of probability-graphons.

Proposition 3.4.6 (Separability of $\mathcal{W}_1$ and $\widetilde{\mathcal{W}_1}$)**.** *Let d be a smooth distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$). Then, the space $(\mathcal{W}_1, d)$ (resp. $(\mathcal{W}_+, d)$ or $(\mathcal{W}_\pm, d)$) is separable.*

If furthermore d is invariant (which implies that $\delta_\square$ is a distance), then the space $(\widetilde{\mathcal{W}_1}, \delta_\square)$ (resp. $(\widetilde{\mathcal{W}_+}, \delta_\square)$ or $(\widetilde{\mathcal{W}_\pm}, \delta_\square)$) is separable.

In particular, this proposition can be applied when $d = \delta_{\square,\mathrm{m}}$ and d_m is a quasi-convex distance continuous w.r.t. the weak topology, as then $d_{\square,\mathrm{m}}$ is invariant and smooth (remind Lemma 3.3.11 and Proposition 3.3.13).

Proof. We shall consider the space of probability-graphons $\mathcal{W}_1$, as the proofs for $\mathcal{W}_+$ and $\mathcal{W}_\pm$ are similar. Applying Lemma 3.4.5 with the sequence of dyadic partitions, for every probability-graphon W, we can find a sequence of probability-graphon stepfunctions adapted to finite dyadic partitions and converging to W almost everywhere on $[0,1]^2$.

As the space $\mathbf{Z}$ is separable, the space of probability measures $\mathcal{M}_1(\mathbf{Z})$ is also separable for the weak topology (see [Bil99a, Theorem 6.8]). Let $\mathcal{A} \subset \mathcal{M}_1(\mathbf{Z})$ be an at most countable dense (for the weak topology) subset. Then, for any stepfunction $W \in \mathcal{W}_1$ adapted to a finite dyadic partition, we can approach it everywhere on $[0,1]^2$ by a sequence of $\mathcal{A}$-valued stepfunctions adapted to the same finite dyadic partition.

Hence, for every $W \in \mathcal{W}_1$, there exists a sequence $(W_n)_{n\in\mathbb{N}}$ in the countable set of $\mathcal{A}$-valued stepfunctions adapted to a finite dyadic partition that converges to W almost everywhere on $[0,1]^2$. As d is smooth, we get that this convergence also holds for d. Thus, the space $(\mathcal{W}_1, d)$ is complete.

Remind that by Theorem 3.3.17, when the distance d is invariant and smooth, then the premetric $\delta_\square$ is a distance on $\widetilde{\mathcal{W}_1}$. In that case, convergence for d implies convergence for $\delta_\square$, and thus the space $(\widetilde{\mathcal{W}_1}, \delta_\square)$ is also separable. □

3.4.2 Tightness

Similarly to the case of signed measures (remind Lemma 3.2.8), we introduce a tightness criterion for signed measure-valued kernels that characterizes relative compactness, see Proposition 3.4.8 below. For a signed measure-valued kernel $W \in \mathcal{W}_\pm$, we define the measure $M_W \in \mathcal{M}_+(\mathbf{Z})$ by:

$$M_W(\mathrm{d}z) = |W|([0,1]^2; \mathrm{d}z) = \int_{[0,1]^2} |W|(x, y; \mathrm{d}z)\ \mathrm{d}x\mathrm{d}y, \tag{3.17}$$

where for every $x, y \in [0,1]$, $|W|(x, y; \cdot)$ is the total variation of $W(x, y; \cdot)$ (see Lemma 3.3.3). In particular, if W is a probability-graphon then M_W is a probability measure from $\mathcal{M}_1(\mathbf{Z})$. Notice also that if W and U are weakly isomorphic, then $M_W = M_U$, so that the application $W \mapsto M_W$ can be seen as a map from $\widetilde{\mathcal{W}_1}$ (resp. $\widetilde{\mathcal{W}_\pm}$) to $\mathcal{M}_1(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$).

Definition 3.4.7 (Tightness criterion)**.** *A subset $\mathcal{K} \subset \mathcal{W}_\pm$ (resp. $\mathcal{K} \subset \widetilde{\mathcal{W}_\pm}$) is said to be tight if the subset of measures $\{M_W : W \in \mathcal{K}\} \subset \mathcal{M}_+(\mathbf{Z})$ is tight.*

The following proposition shows the equivalence between a global tightness criterion and a local tightness criterion. Recall that uniformly bounded subsets of $\widetilde{\mathcal{W}_\pm}$ are discussed after Definition 3.3.16. Recall also λ_2 is the Lebesgue measure on $[0,1]^2$.

Proposition 3.4.8 (Alternative tightness criterion)**.** *Let $\mathcal{K} \subset \mathcal{W}_\pm$ (or $\mathcal{K} \subset \widetilde{\mathcal{W}}_\pm$) be a uniformly bounded subset of signed measure-valued kernels. The set $\mathcal{K}$ is tight if and only if for every $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{Z}$, such that for every $W \in \mathcal{K}$ we have:*

$$\lambda_2\Big(\{(x,y) \in [0,1]^2 : |W|(x,y;K^c) \leq \varepsilon\}\Big) > 1 - \varepsilon. \tag{3.18}$$

Proof. As the left hand side of (3.18) is invariant by relabeling, it is enough to do the proof for $\mathcal{W}_\pm$. Let $\mathcal{K} \subset \mathcal{W}_\pm$ be uniformly bounded and set $C = \sup_{W\in\mathcal{K}} \|W\|_\infty < \infty$. Assume that for every $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{Z}$, such that (3.18) holds for every $W \in \mathcal{K}$. Let $1 > \varepsilon > 0$. Thus, there exists a compact subset $K \subset \mathbf{Z}$ such that for every $W \in \mathcal{K}$ there exists a subset $A_W \subset [0,1]^2$ with (Lebesgue) measure at least $1-\varepsilon$, such that for every $(x,y) \in A_W$, we have $|W|(x,y;K^c) \leq \varepsilon$. We have that for all $W \in \mathcal{K}$:

$$M_W(K^c) = \int_{[0,1]^2} |W|(x,y;K^c)\,\mathrm{d}x\mathrm{d}y \leq \|W\|_\infty \lambda_2(A_W^c) + \varepsilon\lambda_2(A_W) \leq (C+1)\varepsilon.$$

Hence, the subset of measures $\{M_W : W \in \mathcal{K}\} \subset \mathcal{M}_+(\mathbf{Z})$ is tight, that is $\mathcal{K}$ is tight.

Conversely, suppose that $\mathcal{K}$ is tight. Let $\varepsilon > 0$. There exists a compact set $K \subset \mathbf{Z}$ such that for every $W \in \mathcal{K}$, we have $M_W(K^c) < \varepsilon^2$. For $W \in \mathcal{K}$, define $A_W = \{(x,y) \in [0,1]^2 : |W|(x,y;K^c) \leq \varepsilon\}$. We have:

$$\varepsilon^2 > M_W(K^c) = \int_{[0,1]^2} |W|(x,y;K^c)\,\mathrm{d}x\mathrm{d}y \geq \varepsilon\lambda_2(A_W^c).$$

Hence, $\lambda_2(A_W) > 1-\varepsilon$, and consequently Equation (3.18) holds. □

We end this section on a continuity result of the map $W \mapsto M_W$.

Lemma 3.4.9 (Regularity of the map $W \mapsto M_W$)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$). Then the map $W \mapsto M_W$ is 1-Lipschitz, and thus continuous, from $(\widetilde{\mathcal{W}}_{1,\mathrm{m}}, \delta_{\square,\mathrm{m}})$ (resp. $(\widetilde{\mathcal{W}}_{+,\mathrm{m}}, \delta_{\square,\mathrm{m}})$) to $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ (resp. $(\mathcal{M}_+(\mathbf{Z}), d_{\mathrm{m}})$).*

Proof. Taking $S = T = [0,1]$ in Definition (3.3) of $d_{\square,\mathrm{m}}$, we get that $d_{\mathrm{m}}(M_U, M_W) \leq d_{\square,\mathrm{m}}(W,U)$. As $M_{U^\varphi} = M_U$ for any measure-preserving map φ thanks to (3.1), we deduce from Definition (3.6) of $\delta_{\square,\mathrm{m}}$ that $d_{\mathrm{m}}(M_U, M_W) \leq \delta_{\square,\mathrm{m}}(U,W)$. □

3.4.3 Weak regularity

We shall consider the following extra regularities of distances on the set of signed measure-valued kernels w.r.t. the stepping operator. For a finite partition $\mathcal{P}$, denote by $|\mathcal{P}|$ the size of the partition $\mathcal{P}$, i.e. the number of sets composing $\mathcal{P}$.

Definition 3.4.10 (Regularities of distances)**.** *Let d be a distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$).*

(i) ***Weak regularity.*** *The distance d is weakly regular if whenever the subset $\mathcal{K}$ of $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) is tight (resp. tight and uniformly bounded), then for every $\varepsilon > 0$, there exists $m \in \mathbb{N}^*$, such that for every kernel $W \in \mathcal{K}$, and for every finite partition $\mathcal{Q}$ of $[0,1]$, there exists a finite partition $\mathcal{P}$ of $[0,1]$ that refines $\mathcal{Q}$ such that:*

$$|\mathcal{P}| \leq m|\mathcal{Q}| \quad \text{and} \quad d(W, W_\mathcal{P}) < \varepsilon.$$

(ii) ***Regularity w.r.t. the stepping operator.*** *The distance d is regular w.r.t. the stepping operator if (resp. for any finite constant $C \geq 0$) there exists a finite constant $C_0 > 0$ such that for every W, U in $\mathcal{W}_1$ (resp. in $\mathcal{W}_+$ or $\mathcal{W}_\pm$, with $\|W\|_\infty \leq C$ and $\|U\|_\infty \leq C$) and every finite partition $\mathcal{P}$ of $[0,1]$, then we have:*

$$d(W, W_\mathcal{P}) \leq C_0\, d(W, U_\mathcal{P}). \tag{3.19}$$

We say that a norm N on $\mathcal{W}_\pm$ is weakly regular (resp. regular w.r.t. the stepping operator) if its associated distance d on $\mathcal{W}_\pm$ is weakly regular (resp. regular w.r.t. the stepping operator).

The weak regularity property is an analogue to the weak regularity lemma for real-valued graphons (see [Lov12, Lemma 9.15]). If a distance d is weakly regular, then for a subset $\mathcal{K} \subset \mathcal{M}_\pm(\mathbf{Z})$ which is tight and uniformly bounded, every $\mathcal{K}$-valued kernel can be approximated by a stepfunction with a uniform bound. The regularity w.r.t. the stepping operator states that the stepping operator gives an almost optimal way to approximate a signed measure-valued kernel using stepfunctions adapted to a given partition.

An example of cut distance regular w.r.t. the stepping operator

Remind the definition of a quasi-convex distance in Definition 3.2.10. We first show that the stepping operator is 1-Lipschitz for the cut distance $d_{\square,\mathrm{m}}$ when the distance d_{m} is quasi-convex.

Lemma 3.4.11 (The stepping operator is 1-Lipschitz)**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}$ a convex subset of $\mathcal{M}_{\pm}(\mathbf{Z})$ containing the zero measure. Then, the stepping operator associated with a given finite partition of $[0,1]$ is 1-Lipschitz on $\mathcal{W}_{\mathcal{M}}$ for the cut distance $d_{\square,\mathrm{m}}$.*

Proof. Let $U, W \in \mathcal{W}_{\mathcal{M}}$ be $\mathcal{M}$-valued kernels, and let $\mathcal{P}$ be a finite measurable partition of $[0,1]$. As $U_{\mathcal{P}}$ and $W_{\mathcal{P}}$ are stepfunctions adapted to the same partition, and as d_{m} is quasi-convex, we can use Lemma 3.3.19 to get for some $S, T \in \sigma(\mathcal{P})$ that:

$$d_{\square,\mathrm{m}}(U_{\mathcal{P}}, W_{\mathcal{P}}) = d_{\mathrm{m}}(U_{\mathcal{P}}(S\times T;\cdot), W_{\mathcal{P}}(S\times T;\cdot)) = d_{\mathrm{m}}(U(S\times T;\cdot), W(S\times T;\cdot)) \leq d_{\square,\mathrm{m}}(U,W),$$

where the second equality comes from the fact that the integrals are equals as $S, T \in \sigma(\mathcal{P})$ and thus the integration is over full steps of the partition. Hence, the stepping operator is 1-Lipschitz on $\mathcal{W}_{\mathcal{M}}$ for the cut distance d_{m}. □

For a quasi-convex distance d_{m}, the cut distance $d_{\square,\mathrm{m}}$ is regular w.r.t. the stepping operator with $C_0 = 2$ in (3.19) (and one can take $C = +\infty$ in Definition 3.4.10 (ii)).

Lemma 3.4.12 ($d_{\square,\mathrm{m}}$ is regular w.r.t. the stepping operator)**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}$ a convex subset of $\mathcal{M}_{\pm}(\mathbf{Z})$ containing the zero measure. Let $W, U \in \mathcal{W}_{\epsilon}$ be $\mathcal{W}_{\mathcal{M}}$-valued kernels, and let $\mathcal{P}$ be a finite partition of $[0,1]$. Then, we have:*

$$d_{\square,\mathrm{m}}(W, W_{\mathcal{P}}) \ \leq \ 2d_{\square,\mathrm{m}}(W, U_{\mathcal{P}}).$$

Proof. The proof is similar to the proof of [Lov12, Lemma 9.12]. As the distance d_{m} is quasi-convex, using Lemma 3.4.11, we get:

$$d_{\square,\mathrm{m}}(W, W_{\mathcal{P}}) \leq \ d_{\square,\mathrm{m}}(W, U_{\mathcal{P}}) + d_{\square,\mathrm{m}}(U_{\mathcal{P}}, W_{\mathcal{P}}) \leq \ 2d_{\square,\mathrm{m}}(W, U_{\mathcal{P}}).$$

This concludes the proof. □

An example of weakly regular cut distance

We have the following general result. Recall Definitions 3.3.10 and 3.4.10 on distances and norms on $\mathcal{W}_{\epsilon}$, with $\varepsilon \in \{+, \pm\}$, being invariant, smooth, weakly regular and regular w.r.t. the stepping operator.

Proposition 3.4.13 (Weak regularity of $d_{\square,\mathrm{m}}$)**.** *Let d_{m} be a quasi-convex distance on $\mathcal{M}_{\epsilon}(\mathbf{Z})$, with $\epsilon \in \{+, \pm\}$, which is sequentially continuous w.r.t. the weak topology. Then, the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_{\epsilon}$ is invariant, smooth, weakly regular and regular w.r.t. the stepping operator.*

Using results from Section 3.3.8, we directly get the following weak regularity of the cut distance $d_{\square,\mathcal{P}}$ and the cut norms $\|\cdot\|_{\square,\mathcal{F}}$, $\|\cdot\|_{\square,\mathrm{KR}}$ and $\|\cdot\|_{\square,\mathrm{FM}}$.

Corollary 3.4.14 (Weak regularity of usual distances and norms)**.** *The cut norms $\|\cdot\|_{\square,\mathcal{F}}$, $\|\cdot\|_{\square,KR}$ and $\|\cdot\|_{\square,FM}$ (resp. the cut distance $d_{\square,\mathcal{P}}$) on $\mathcal{W}_{\pm}$ (resp. $\mathcal{W}_{+}$) are invariant, smooth, weakly regular and regular w.r.t. the stepping operator.*

Proof of Proposition 3.4.13. We deduce from Lemmas 3.3.11 and 3.4.12, Proposition 3.3.13 and that the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_{\epsilon}$ is invariant, smooth and regular w.r.t. the stepping operator. We are left to prove that $d_{\square,\mathrm{m}}$ is weakly regular on $\mathcal{W}_{\epsilon}$. We prove it by considering in the first step the case $\mathbf{Z}$ compact and in a second step the general case $\mathbf{Z}$ Polish.

Step 1. We assume $\mathbf{Z}$ compact. As in the definition of weak regularity, let $\mathcal{K} \subset \mathcal{W}_{\epsilon}$ be a subset of $\mathcal{M}_{\epsilon}(\mathbf{Z})$-valued kernels that is tight and uniformly bounded by some finite constant C. Let $\mathcal{M} \subset \mathcal{M}_{\epsilon}(\mathbf{Z})$ be the subset of elements of $\mathcal{M}_{\epsilon}(\mathbf{Z})$ with total mass at most C; in particular $\mathcal{M}$ is a convex set containing 0 and $\mathcal{K} \subset \mathcal{W}_{\mathcal{M}}$. As $\mathbf{Z}$ is compact, from Remarks 3.2.6 and 3.2.7, we know that the weak topology is metrizable on $\mathcal{M}$ and that $\mathcal{M}$ is compact, and thus sequentially weakly compact. Hence, as d_{m}

is sequentially continuous w.r.t. the weak topology on $\mathcal{M}_\epsilon(\mathbf{Z})$, we have that $(\mathcal{M}, d_{\mathrm{m}})$ is sequentially compact, and thus compact.

Denote by $B(\mu, r) = \{\nu \in \mathcal{M} : d_{\mathrm{m}}(\mu, \nu) < r\}$ the open ball centered at $\mu \in \mathcal{M}$ with radius $r > 0$. Let $\varepsilon > 0$. As $\mathcal{M}$ is compact, there exist $\mu_1, \dots, \mu_n \in \mathcal{M}$, $n \in \mathbb{N}^*$, such that $\mathcal{M} = \cup_{i=1}^n B(\mu_i, \varepsilon)$. For $1 \leq i \leq n$, define $A_i = B(\mu_i, \varepsilon) \setminus \cup_{j<i} B(\mu_j, \varepsilon)$, so that $\{A_1, \dots, A_n\}$ is a finite partition (with possibly some empty sets) of $\mathcal{M}$.

Every $\mathcal{M}$-valued kernel W can be approximated by a $\{\mu_1, \dots, \mu_n\}$-valued kernel U defined for every $(x, y) \in [0, 1]^2$ by $U(x, y; \cdot) = \mu_i$ for i such that $W(x, y; \cdot) \in A_i$. Thus, by construction, we have that for every $(x, y) \in [0, 1]^2$, $d_{\mathrm{m}}(W(x, y; \cdot), U(x, y; \cdot)) < \varepsilon$. Applying the quasi-convex supremum inequality from (3.5) to W and U, we get that:

$$d_{\square,\mathrm{m}}(W, U) \leq \operatorname*{essup}_{(x,y)\in[0,1]^2} d_{\mathrm{m}}(W(x, y; \cdot), U(x, y; \cdot)) \leq \varepsilon.$$

Then, as the stepping operator is 1-Lipschitz for the cut norm, see Lemma 3.4.11, we have for any finite partition $\mathcal{P}$ of $[0, 1]$ that:

$$\begin{aligned} d_{\square,\mathrm{m}}(W, W_\mathcal{P}) &\leq d_{\square,\mathrm{m}}(W, U) + d_{\square,\mathrm{m}}(U, U_\mathcal{P}) + d_{\square,\mathrm{m}}(U_\mathcal{P}, W_\mathcal{P}) \\ &\leq 2\varepsilon + d_{\square,\mathrm{m}}(U, U_\mathcal{P}). \end{aligned} \tag{3.20}$$

Hence, to get the weak regularity property for $\mathcal{M}$-valued kernels, we are left to prove it for the much smaller set of $\mathcal{V}$-valued kernels, where $\mathcal{V}$ is the convex hull of $\{\mu_1, \dots, \mu_n\}$.

As d_{m} is quasi-convex and sequentially continuous w.r.t. the weak topology, using Lemma 3.2.11, there exists $\eta > 0$ such that for all $\mu, \nu \in \mathcal{M}_\epsilon(\mathbf{Z})$, we have that $\|\mu - \nu\|_\infty < \eta$ implies that $d_{\mathrm{m}}(\mu, \nu) \leq \varepsilon$.

As $\mathcal{V}$ is a subset of a vector space with finite dimension n, the norm $\|\cdot\|_\infty$ seen over $\mathcal{V}$ is equivalent to the L_1-norm $\mu = \sum_{i=1}^n \alpha_i \mu_i \mapsto \|\alpha\|_1 = \sum_{i=1}^n |\alpha_i|$. We can now see $\mathcal{V}$-valued kernel as $\mathbb{R}^n$-valued graphon with a cut norm derived from the L_1-norm $\|\cdot\|_1$, and in this case the proof for the weak regularity Lemma 9.9 in [Lov12] can easily be adapted. Hence, we have the weak regularity property for $\mathcal{V}$-valued kernels: there exists $m \in \mathbb{N}^*$, such that for every $\mathcal{V}$-valued kernel U', and for every finite partition $\mathcal{Q}$ of $[0, 1]$ there exists a finite partition $\mathcal{P}$ of $[0, 1]$ that refines $\mathcal{Q}$, and such that $|\mathcal{P}| \leq m|\mathcal{Q}|$ and $\sup_{S,T\subset[0,1]} \|(U' - U'_\mathcal{P})(S \times T; \cdot)\|_\infty < \eta$, and thus $d_{\square,\mathrm{m}}(U', U'_\mathcal{P}) \leq \varepsilon$.

Taking $U' = U$ in (3.20), we get that $d_{\square,\mathrm{m}}(W, W_\mathcal{P}) \leq 3\varepsilon$ and $|\mathcal{P}| \leq m|\mathcal{Q}|$. This concludes the proof of the lemma when $\mathbf{Z}$ is compact.

Step 2. We consider the general case $\mathbf{Z}$ Polish. We now prove that $d_{\square,\mathrm{m}}$ is weakly regular on $\mathcal{W}_\epsilon$. Let $\mathcal{K} \subset \mathcal{W}_\epsilon$ be a subset of $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels that is tight and uniformly bounded, and denote by $C = \sup_{W\in\mathcal{K}} \|W\|_\infty < \infty$.

Let $\varepsilon > 0$. As d_{m} is quasi-convex and sequentially continuous w.r.t. the weak topology, using Lemma 3.2.11, there exists $\eta > 0$ such that for all $\mu, \nu \in \mathcal{M}_\epsilon(\mathbf{Z})$, we have that $\|\mu - \nu\|_\infty < \eta$ implies that $d_{\mathrm{m}}(\mu, \nu) < \varepsilon$. Without loss of generality, we assume that $\eta \leq \varepsilon$. Let $\eta_C = \min(\eta, \eta/C)$.

As $\mathcal{K}$ is tight, using Proposition 3.4.8, there exists a compact set $K \subset \mathbf{Z}$, such that for every $W \in \mathcal{K}$ the subset $A_W = \{(x, y) \in [0, 1]^2 : |W|(x, y; K^c) \leq \eta_C/2\}$ has Lebesgue measure at least $1 - \eta_C/2$. Let $W \in \mathcal{K}$, and define the signed measure-valued kernel U by: $U(x, y; \cdot) = W(x, y; \cdot \cap K)$ for every $(x, y) \in A_W$, and $U(x, y; \cdot) = 0$ otherwise. Let $S, T \subset [0, 1]$. We have:

$$\begin{aligned} \|(W - U)(S \times T; \cdot)\|_\infty &\leq \int_{S\times T} \|W(x, y; \cdot) - U(x, y; \cdot)\|_\infty \, \mathrm{d}x\mathrm{d}y \\ &\leq \int_{A_W\cap(S\times T)} |W|(x, y; K^c) \, \mathrm{d}x\mathrm{d}y + \int_{A_W^c\cap(S\times T)} \|W(x, y; \cdot)\|_\infty \, \mathrm{d}x\mathrm{d}y \\ &\leq \eta_C/2 + C \cdot \eta_C/2 \\ &\leq \eta. \end{aligned}$$

Thus, we have that $d_{\mathrm{m}}(W(S \times T; \cdot), U(S \times T; \cdot)) < \varepsilon$. Since this holds for all $S, T \subset [0, 1]$, we get that $d_{\square,\mathrm{m}}(W, U) \leq \varepsilon$.

Notice that the $\mathcal{M}_\pm(\mathbf{Z})$-valued kernel U is also a $\mathcal{M}_\pm(K)$-valued kernel, where $K \subset \mathbf{Z}$ is a compact set, and that $\|U\|_\infty \leq \|W\|_\infty \leq C$. Further remark that, using Lemma 3.4.11, for every $W \in \mathcal{K}$ and

every finite partition $\mathcal{P}$ of $[0,1]$, we have that:

$$\begin{aligned} d_{\square,\mathrm{m}}(W, W_{\mathcal{P}}) &\leq d_{\square,\mathrm{m}}(W,U) + d_{\square,\mathrm{m}}(U, U_{\mathcal{P}}) + d_{\square,\mathrm{m}}(U_{\mathcal{P}}, W_{\mathcal{P}}) \\ &\leq 2\varepsilon + d_{\square,\mathrm{m}}(U, U_{\mathcal{P}}). \end{aligned}$$

Hence, to get the weak regularity property for $d_{\square,\mathrm{m}}$ on $\mathcal{K}$ (see Definition 3.4.10 (i)), it is enough to prove that $d_{\square,\mathrm{m}}$ restricted to $\mathcal{M}_\epsilon(K)$-valued kernels is weakly regular, which is true by Step 1. As a consequence, we get that $d_{\square,\mathrm{m}}$ on $\mathcal{W}_\epsilon$ is weakly regular. □

3.4.4 A stronger weak regularity lemma for $d_{\square,\mathcal{F}}$

In this subsection, we prove a stronger version of the weak regularity lemma for the special case of the cut distance $d_{\square,\mathcal{F}}$. We shall use this result for the proof of the second sampling Lemma 3.6.12.

Let $\mathcal{F} = (f_n)_{n\in\mathbb{N}}$, with $f_0 = \mathbb{1}$ and f_n takes values in $[0,1]$, be a convergence determining sequence, which is assumed fixed in this section.

Comparison between $\|\cdot\|_{\square,\mathcal{F}}$ and an euclidian norm

To better understand the stepping operator, we introduce a scalar product over signed measure-valued kernels. The link between this scalar product and the norm $\|\cdot\|_{\square,\mathcal{F}}$ is given by Lemma 3.4.15. We define the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ on signed measure-valued kernels for $U, W \in \mathcal{W}_\pm$ by:

$$\langle U, W\rangle_{\mathcal{F}} = \sum_{n\geq 0} 2^{-n}\langle U[f_n], W[f_n]\rangle,$$

where for all n the scalar product taken for $U[f_n]$ and $W[f_n]$ is the usual scalar product in $L^2([0,1]^2, \lambda_2)$ for real-valued kernels:

$$\langle U[f_n], W[f_n]\rangle = \int_{[0,1]^2} U[f_n](x,y)W[f_n](x,y)\,\mathrm{d}x\mathrm{d}y.$$

The scalar product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ induces a norm on $\mathcal{W}_\pm$ which we denote by $\|\cdot\|_{2,\mathcal{F}}$.

Let $\mathcal{P}$ be a finite partition of $[0,1]$. As the stepping operator for measurable real-valued L^2 functions on $[0,1]^2$ is a linear projection, and is idempotent and symmetric, and by definition of the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ for signed measure-valued kernels, we have that the stepping operator for signed measure-valued kernels is linear, idempotent and symmetric for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$. Moreover, the stepping operator is the orthogonal projection for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ onto the space of stepfunctions with steps in $\mathcal{P}$.

Note that for a probability-graphon $W \in \mathcal{W}_1$, we have $\|W\|_{2,\mathcal{F}} \leq \sqrt{2}$ as each f_n takes values in $[0,1]$. The following technical lemma gives a comparison between $\|\cdot\|_{\square,\mathcal{F}}$ and $\|\cdot\|_{2,\mathcal{F}}$.

Lemma 3.4.15 (Comparison between $\|\cdot\|_{\square,\mathcal{F}}$ and $\|\cdot\|_{2,\mathcal{F}}$)**.** *For a signed measure-valued kernel $W \in \mathcal{W}_\pm$, we have $\|W\|_{\square,\mathcal{F}} \leq \sqrt{2}\|W\|_{2,\mathcal{F}}$.*

Proof. Let $S, T \subset [0,1]$ be measurable subsets. Using the Cauchy-Schwarz inequality, we have that $|\langle W[f_n], \mathbb{1}_{S\times T}\rangle|^2 \leq \|W[f_n]\|_2^2 = \langle W[f_n], W[f_n]\rangle$ for every $n \geq 0$. Using this inequality along with Jensen's inequality, we get for every $S, T \subset [0,1]$ that:

$$\begin{aligned} \left(\sum_{n\geq 0} 2^{-n}|W(S\times T, f_n)|\right)^2 &= \left(\sum_{n\geq 0} 2^{-n}|\langle W[f_n], \mathbb{1}_{S\times T}\rangle|\right)^2 \\ &\leq \sum_{n\geq 0} 2^{-n+1}|\langle W[f_n], \mathbb{1}_{S\times T}\rangle|^2 \\ &\leq \sum_{n\geq 0} 2^{-n+1}\langle W[f_n], W[f_n]\rangle \\ &= 2(\|W\|_{2,\mathcal{F}})^2. \end{aligned}$$

Taking the supremum over every measurable subsets $S, T \subset [0,1]$ gives the desired inequality. □

The weak regularity lemma for $\|\cdot\|_{\square,\mathcal{F}}$

The following lemma gives an explicit bound on the approximation of a signed measure-valued kernel, say W, by its steppings $W_{\mathcal{P}}$, with $\mathcal{P}$ a finite partition on $[0,1]$. Its proof is a straightforward adaptation of the proof of the weak regularity lemma for real-valued graphons in [Lov12, Lemma 9.9].

Lemma 3.4.16 (Weak regularity lemma for $\|\cdot\|_{\square,\mathcal{F}}$, simple formulation)**.** *For every signed measure-valued kernel $W \in \mathcal{W}_\pm$ and $k \geq 1$, there exists a finite partition $\mathcal{P}$ of $[0,1]$ such that $|\mathcal{P}| = k$ and:*

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \leq \frac{\sqrt{8}}{\sqrt{\log(k)}}\|W\|_{2,\mathcal{F}}.$$

In particular, if $W \in \mathcal{W}_1$ is a probability-graphon, (as $\|W\|_{2,\mathcal{F}} \leq \sqrt{2}$) we have:

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \leq \frac{4}{\sqrt{\log(k)}}.$$

It is possible in the weak regularity lemma to ask for extra requirements, for instance to start from an already existing partition, or to ask the partition to be balanced, as stated in the following lemma. The proof is a straightforward adaptation of the proof of [Lov12, Lemma 9.15].

Lemma 3.4.17 (Weak regularity lemma for $\|\cdot\|_{\square,\mathcal{F}}$, with extra requirements)**.** *Let $W \in \mathcal{W}_1$ be a probability-graphon, and let $1 \leq m < k$.*

(i) For every partition $\mathcal{Q}$ of $[0,1]$ into m classes, there is a partition $\mathcal{P}$ with k classes refining $\mathcal{Q}$ and such that:

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \leq \frac{4}{\sqrt{\log(k/m)}}.$$

(ii) For every partition $\mathcal{Q}$ of $[0,1]$ into m classes, there is an equipartition (i.e. a finite partition into classes with the same measure) $\mathcal{P}$ of $[0,1]$ into k classes and such that:

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \leq 2\|W - W_{\mathcal{Q}}\|_{\square,\mathcal{F}} + \frac{2m}{k}.$$

3.5 Compactness and completeness of $\mathcal{W}_1$

In Section 3.5.1, we link the tightness criterion for measure-valued kernels with the relative compactness w.r.t. the cut distance $\delta_{\square,\mathrm{m}}$. In Section 3.5.2, we compare the topologies induced by the cut distance $\delta_{\square,\mathrm{m}}$ for different choice of the distance d_{m}, and state that under some conditions on d_{m}, those topologies coincide. In Section 3.5.3, we investigate the completeness of $\mathcal{W}_1$ endowed with the cut distance $\delta_{\square,\mathrm{m}}$ and prove that the space of probability-graphons $\widetilde{\mathcal{W}}_1$ is a Polish space (Theorem 3.5.10), and that it is compact if and only if $\mathbf{Z}$ is compact (Corollary 3.5.13). The technical proofs are postponed to Section 3.8.

3.5.1 Tightness criterion and compactness

Let $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$ be a subset of signed measures on $\mathbf{Z}$. Recall that $\mathcal{W}_\mathcal{M} \subset \mathcal{W}_\pm$ denote the subset of signed measure-valued kernels which are $\mathcal{M}$-valued. In this section, we shall denote by $\widetilde{\mathcal{W}}_\mathcal{M}$ the quotient of $\mathcal{W}_\mathcal{M}$ identifying signed measure-valued kernels that are weakly isomorphic.

Remind from Definition 3.3.15 and Theorem 3.3.17 that for an invariant, smooth and weakly regular distance d on $\mathcal{W}_1$ (resp. $\mathcal{W}_+, \mathcal{W}_\pm$), $\delta_\square$ is defined as $\delta_\square(U,W) = \inf_{\varphi \in S_{[0,1]}} d(U, W^\varphi)$, and is a distance on $\widetilde{\mathcal{W}}_1$ (resp. $\widetilde{\mathcal{W}}_+, \widetilde{\mathcal{W}}_\pm$).

We are now ready to formulate the important following theorem, which relates tightness with compactness and convergence for signed measure-valued kernels. We prove this theorem in Section 3.8.

Theorem 3.5.1 (Compactness theorem for $\widetilde{\mathcal{W}}_1$)**.** *Let d be an invariant, smooth and weakly regular distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_\pm$).*

(i) If a sequence of elements of $\mathcal{W}_1$ or $\widetilde{\mathcal{W}}_1$ (resp. $\mathcal{W}_\pm$ or $\widetilde{\mathcal{W}}_\pm$) is tight (resp. tight and uniformly bounded), then it has a subsequence converging for $\delta_\square$.

(ii) *If $\mathcal{M} \subset \mathcal{M}_1(\mathbf{Z})$ (resp. $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$) is convex and compact (resp. sequentially compact) for the weak topology, then the space $(\widetilde{\mathcal{W}}_{\mathcal{M}}, \delta_\square)$ is convex and compact.*

(iii) *If $\mathbf{Z}$ is compact, then the space $(\widetilde{\mathcal{W}}_1, \delta_\square)$ is compact.*

In particular, this theorem can be applied when $d = d_{\square,\mathrm{m}}$ and d_m is a quasi-convex distance continuous w.r.t. the weak topology, as then $d_{\square,\mathrm{m}}$ is invariant, smooth and weakly regular (remind Proposition 3.4.13).

We deduce from this theorem a characterization of relative compactness for subsets of probability-graphons.

Proposition 3.5.2 (Characterization of relative compactness)**.** *Let d_m be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$) that induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$). Assume that the distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) is (invariant) smooth and weakly regular.*

(i) *If a sequence of elements of $\mathcal{W}_1$ or $\widetilde{\mathcal{W}}_1$ (resp. $\mathcal{W}_+$ or $\widetilde{\mathcal{W}}_+$) is converging for $\delta_{\square,\mathrm{m}}$, then it is tight.*

(ii) *Let $\mathcal{K}$ be a subset of $\widetilde{\mathcal{W}}_1$ (resp. a uniformly bounded subset of $\widetilde{\mathcal{W}}_+$). Then, the set $\mathcal{K}$ is relatively compact for $\delta_{\square,\mathrm{m}}$ if and only if it is tight.*

(iii) *Let $\mathcal{M}$ be a subset of $\mathcal{M}_+(\mathbf{Z})$ which is bounded, convex and closed for the weak topology. Then the set $\widetilde{\mathcal{W}}_{\mathcal{M}}$ is convex and closed in $\widetilde{\mathcal{W}}_+$.*

Remark that convergence for $\delta_{\square,\mathrm{m}}$ does not necessarily imply tightness on $\mathcal{W}_\pm$ or on $\widetilde{\mathcal{W}}_\pm$.

Proof. We consider the case where d_m is a distance on $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$, the case with $\mathcal{M}_{\leq 1}(\mathbf{Z})$ is similar.

We prove Point (i). Let $(W_n)_{n\in\mathbb{N}}$ be a convergent sequence of $\mathcal{W}_+$ (and thus of $\widetilde{\mathcal{W}}_+$) for $\delta_{\square,\mathrm{m}}$. We deduce from the continuity of the map $W \mapsto M_W$, see Lemma 3.4.9, that the sequence $(M_{W_n})_{n\in\mathbb{N}}$ is converging for d_m, and thus is tight as d_m induces the weak topology on $\mathcal{M}_+(\mathbf{Z})$. Then, by definition the sequence $(W_n)_{n\in\mathbb{N}}$ is tight.

We prove Point (ii). If $\mathcal{K} \subset \widetilde{\mathcal{W}}_+$ is tight and uniformly bounded, then by Theorem 3.5.1 (i) every sequence in $\mathcal{K}$ has a subsequence converging for $\delta_{\square,\mathrm{m}}$, which implies that $\mathcal{K}$ is relatively compact in the metric space $(\widetilde{\mathcal{W}}_+, \delta_{\square,\mathrm{m}})$ (see Remark 3.2.1).

Conversely, assume that $\mathcal{K} \subset \widetilde{\mathcal{W}}_+$ is uniformly bounded and relatively compact for $\delta_{\square,\mathrm{m}}$. Define $\mathcal{M} = \{M_W : W \in \mathcal{K}\} \subset \mathcal{M}_+(\mathbf{Z})$. By Lemma 3.4.9, the mapping $W \mapsto M_W$ is continuous from $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ to $(\mathcal{M}_+(\mathbf{Z}), d_\mathrm{m})$. Hence, as d_m induces the weak topology on $\mathcal{M}_+(\mathbf{Z})$, the set $\mathcal{M}$ is also relatively compact in $\mathcal{M}_+(\mathbf{Z})$ for the weak topology. As the space $\mathbf{Z}$ is Polish, applying Lemma 3.2.8, we get that $\mathcal{M} \subset \mathcal{M}_+(\mathbf{Z})$ is tight, and by Definition 3.4.7, the set $\mathcal{K} \subset \widetilde{\mathcal{W}}_+$ is tight.

We postpone the proof of Point (iii) to Section 3.8 on page 92. □

3.5.2 Equivalence of topologies induced by $\delta_{\square,\mathrm{m}}$

The following lemma allows to show a first result on equivalence of the topologies induced by the cut distance $\delta_{\square,\mathrm{m}}$ for different distances d_m, where the sub-script m is used to distinguish different distances. Its proof is given below. Remind from Theorem 3.3.17 that $d_{\square,\mathrm{m}}$ must be smooth for $\delta_{\square,\mathrm{m}}$ to be a distance.

Lemma 3.5.3 (Comparison of topologies induced by $d_{\square,\mathrm{m}}$ and $\delta_{\square,\mathrm{m}}$)**.** *Let d_m and $d_{\mathrm{m}'}$ be two distances on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that $d_{\mathrm{m}'}$ is uniformly continuous w.r.t. d_m (in particular, d_m induces a finer topology than $d_{\mathrm{m}'}$ on $\mathcal{M}_{\leq 1}(\mathbf{Z})$). Then, we have the following properties.*

(i) *The distance $d_{\square,\mathrm{m}'}$ is uniformly continuous w.r.t. $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$. In particular $d_{\square,\mathrm{m}}$ induces a finer topology than $d_{\square,\mathrm{m}'}$ on $\mathcal{W}_1$.*

(ii) *If the distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is smooth, then the distance $d_{\square,\mathrm{m}'}$ is also smooth and $\delta_{\square,\mathrm{m}'}$ is uniformly continuous w.r.t. $\delta_{\square,\mathrm{m}}$. In particular, $\delta_{\square,\mathrm{m}}$ induces a finer topology than $\delta_{\square,\mathrm{m}'}$ on $\widetilde{\mathcal{W}}_1$.*

(iii) *If the distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is weakly regular, then the distance $d_{\square,\mathrm{m}'}$ is also weakly regular.*

(iv) Assume that the distance $d_{\mathrm{m}'}$ induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$, and that the distance $d_{\square,\mathrm{m}}$ is smooth and weakly regular. In particular, the distance d_{m} also induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$. Then, the distances $\delta_{\square,\mathrm{m}}$ and $\delta_{\square,\mathrm{m}'}$ induce the same topology on $\widetilde{\mathcal{W}}_1$.

We will see some application of Lemma 3.5.3 in Corollary 3.5.6 below.

Remark 3.5.4 (Extension to $\mathcal{W}_\pm$ for topology comparisons)**.** In Lemma 3.5.3 (i)-(iii), one can replace $\mathcal{W}_1$ and $\widetilde{\mathcal{W}}_1$ by $\mathcal{W}_+$ and $\widetilde{\mathcal{W}}_+$ or by $\mathcal{W}_\pm$ and $\widetilde{\mathcal{W}}_\pm$ as soon as the distances d_{m} and $d_{\mathrm{m}'}$ are defined on $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$; in this case comparisons of topologies only apply on uniformly bounded subsets. In Lemma 3.5.3 (iv), one can replace $\widetilde{\mathcal{W}}_1$ by $\widetilde{\mathcal{W}}_{\mathcal{M}}$ with a bounded subset $\mathcal{M} \subset \mathcal{M}_+(\mathbf{Z})$ as soon as the distances d_{m} and $d_{\mathrm{m}'}$ are defined on $\mathcal{M}_+(\mathbf{Z})$.

Proof of Lemma 3.5.3. We prove Point (i). Let $\varepsilon > 0$. As $d_{\mathrm{m}'}$ is uniformly continuous w.r.t. d_{m}, there exists $\eta > 0$ such that for every $\mu, \nu \in \mathcal{M}_{\leq 1}(\mathbf{Z})$, if $d_{\mathrm{m}}(\mu,\nu) < \eta$, then $d_{\mathrm{m}'}(\mu,\nu) < \varepsilon$. Let $U, W \in \mathcal{W}_1$ such that $d_{\square,\mathrm{m}}(U,W) < \eta$. Then, for every subsets $S, T \subset [0,1]$, we have:

$$d_{\mathrm{m}'}\left(U(S\times T;\cdot), W(S\times T;\cdot)\right) < \varepsilon.$$

Thus, $d_{\square,\mathrm{m}'}(U,W) \leq \varepsilon$. Hence, $d_{\square,\mathrm{m}'}$ is uniformly continuous w.r.t. $d_{\square,\mathrm{m}}$.

We prove Point (ii). Assume that $d_{\square,\mathrm{m}}$ is smooth. Let $(W_n)_{n\in\mathbb{N}}$ and W be probability-graphons such that $W_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$ for almost every $x, y \in [0,1]$. Since the cut distance $d_{\square,\mathrm{m}}$ is smooth, we get that $d_{\square,\mathrm{m}}(W_n, W) \to 0$. As $d_{\square,\mathrm{m}'}$ is uniformly continuous (and thus also continuous) w.r.t. $d_{\square,\mathrm{m}}$, we have that $d_{\square,\mathrm{m}'}(W_n, W) \to 0$. Hence, $d_{\square,\mathrm{m}'}$ is smooth.

Furthermore, let $\varepsilon > 0$. Let $\eta > 0$ be such that for every $\mu, \nu \in \mathcal{M}_{\leq 1}(\mathbf{Z})$, $d_{\mathrm{m}}(\mu,\nu) < \eta$ implies $d_{\mathrm{m}'}(\mu,\nu) < \varepsilon$. For every $U, W \in \mathcal{W}_1$ such that $\delta_{\square,\mathrm{m}}(U,W) < \eta$, there exists $\varphi \in S_{[0,1]}$ such that $d_{\square,\mathrm{m}}(U, W^\varphi) < \eta$, which implies that $d_{\square,\mathrm{m}'}(U, W^\varphi) < \varepsilon$, which then implies that $\delta_{\square,\mathrm{m}'}(U,W) < \varepsilon$. That is, $\delta_{\square,\mathrm{m}'}$ is uniformly continuous w.r.t. $\delta_{\square,\mathrm{m}}$.

We prove Point (iii). Assume that $d_{\square,\mathrm{m}}$ is weakly regular. Let $\mathcal{K} \subset \mathcal{W}_1$ be tight. Let $\varepsilon > 0$. As $d_{\square,\mathrm{m}'}$ is uniformly continuous w.r.t. $d_{\square,\mathrm{m}}$, there exists $\eta > 0$ such that for every $U, W \in \mathcal{W}_1$, if $d_{\square,\mathrm{m}}(U,W) < \eta$, then $d_{\square,\mathrm{m}'}(U,W) < \varepsilon$. Since $d_{\square,\mathrm{m}}$ is weakly regular, there exists $m \in \mathbb{N}^*$, such that for every probability-graphon $W \in \mathcal{K}$, and for every finite partition $\mathcal{Q}$ of $[0,1]$, there exists a finite partition $\mathcal{P}$ of $[0,1]$ that refines $\mathcal{Q}$ and such that $|\mathcal{P}| \leq m|\mathcal{Q}|$ and $d_{\square,\mathrm{m}}(W, W_{\mathcal{P}}) < \eta$; and thus we also have $d_{\square,\mathrm{m}'}(W, W_{\mathcal{P}}) < \varepsilon$. Hence, $d_{\square,\mathrm{m}'}$ is weakly regular.

We prove Point (iv). Assume that $d_{\mathrm{m}'}$ induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ and that $d_{\square,\mathrm{m}}$ is smooth and weakly regular. In particular, the topology induced by d_{m} if finer than the topology induced by $d_{\mathrm{m}'}$, i.e. finer than the weak topology. As $d_{\square,\mathrm{m}}$ is smooth, by Lemma 3.3.12, d_{m} is continuous w.r.t. the weak topology (i.e. the weak topology if finer than the topology induced by d_{m}), and thus d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$. By Points (ii) and (iii), we get that $d_{\square,\mathrm{m}'}$ is also smooth and weakly regular. By Point (ii), the distance $\delta_{\square,\mathrm{m}}$ induces a finer topology than $\delta_{\square,\mathrm{m}'}$ on $\widetilde{\mathcal{W}}_1$.

We now prove that the topology of $\delta_{\square,\mathrm{m}'}$ is finer than the topology of $\delta_{\square,\mathrm{m}}$. Let $(W_n)_{n\in\mathbb{N}}$ and W be probability-graphons in $\widetilde{\mathcal{W}}_1$, such that W_n converges to W for $\delta_{\square,\mathrm{m}'}$. By Proposition 3.5.2 (i), we deduce that the sequence $(W_n)_{n\in\mathbb{N}}$ is tight. As $d_{\square,\mathrm{m}}$ is smooth and weakly regular, Theorem 3.5.1 gives that every subsequence $(W_{n_k})_{k\in\mathbb{N}}$ of the sequence $(W_n)_{n\in\mathbb{N}}$ has a further subsequence $(W_{n'_k})_{k\in\mathbb{N}}$ that converges for $\delta_{\square,\mathrm{m}}$ to a limit, say $U \in \widetilde{\mathcal{W}}_1$. Since $\delta_{\square,\mathrm{m}}$ is finer than $\delta_{\square,\mathrm{m}'}$, we deduce that $(W_{n'_k})_{k\in\mathbb{N}}$ converges also to U for $\delta_{\square,\mathrm{m}'}$; but, as a subsequence, it also converges to W for $\delta_{\square,\mathrm{m}'}$. As $\delta_{\square,\mathrm{m}'}$ is a distance on $\mathcal{W}_1$ thanks to Theorem 3.3.17, we get $U = W$. Hence, every subsequence of $(W_n)_{n\in\mathbb{N}}$ has a further subsequence that converges to W for $\delta_{\square,\mathrm{m}}$, therefore the whole sequence itself converges to W for $\delta_{\square,\mathrm{m}}$. Consequently, $\delta_{\square,\mathrm{m}'}$ is finer than $\delta_{\square,\mathrm{m}}$, and thus those two distances induce the same topology on $\widetilde{\mathcal{W}}_1$. □

The following theorem states that the topology induced by $\delta_{\square,\mathrm{m}}$ does not depend on d_{m} under some hypothesis. We prove this theorem in Section 3.8. Recall that under suitable conditions satisfied in the next theorem, the quotient space $\widetilde{\mathcal{W}}_1$ does not depend on the choice of the distance d_{m}, see Theorem 3.3.17.

Theorem 3.5.5 (Equivalence of topologies induced by $\delta_{\square,\mathrm{m}}$ on $\widetilde{\mathcal{W}}_1$)**.** *The topology on the space proba -bility-graphons $\widetilde{\mathcal{W}}_1$ induced by the distance $\delta_{\square,\mathrm{m}}$ does not depend on the choice of the distance d_{m} on*

$\mathcal{M}_{\leq 1}(\mathbf{Z})$, as long as d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ and the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is (invariant) smooth, weakly regular and regular w.r.t. the stepping operator.

Remind from Proposition 3.4.13 that when the distance d_{m} is quasi-convex and continuous w.r.t. the weak topology on $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$, then the cut distance $d_{\square,\mathrm{m}}$ is invariant, smooth, weakly regular and regular w.r.t. the stepping operator. This is in particular the case of $d_{\mathcal{P}}$, $\|\cdot\|_{\mathcal{F}}$, $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathrm{FM}}$.

The next corollary is an immediate consequence of Lemma 3.3.22, Corollary 3.4.14, Lemma 3.5.3 and Theorem 3.5.5. This corollary gathers results comparing the topology induced by the cut distances associated with the distances introduced in Section 3.3.8. It is yet unclear if the distances $d_{\square,\mathcal{F}}$ induces the same topology on the space of labeled probability-graphons $\mathcal{W}_1$ as the one induced by $d_{\square,\mathcal{P}}$, $d_{\square,\mathrm{FM}}$ or $d_{\square,\mathrm{KR}}$.

Corollary 3.5.6 (Topological equivalence of the cut distances associated to $d_{\mathcal{P}}$, $\|\cdot\|_{\mathrm{FM}}$, $\|\cdot\|_{\mathrm{KR}}$ and $\|\cdot\|_{\mathcal{F}}$)**.** *The cut distances $d_{\square,\mathcal{P}}$ on $\mathcal{W}_+$ and $d_{\square,KR}, d_{\square,FM}$ and $d_{\square,\mathcal{F}}$ on $\mathcal{W}_\pm$ are invariant, smooth, weakly regular and regular w.r.t. the stepping operator. Moreover, we have the following comparison between the distances introduced in Section 3.3.8.*

(i) The cut norms $\|\cdot\|_{\square,FM}$ and $\|\cdot\|_{\square,KR}$ (resp. the cut distances $\delta_{\square,FM}$ and $\delta_{\square,KR}$) are metrically equivalent on $\mathcal{W}_\pm$ (resp. $\widetilde{\mathcal{W}}_\pm$).

(ii) The cut distances $\delta_{\square,FM}$, $\delta_{\square,KR}$ and $\delta_{\square,\mathcal{P}}$ (resp. $d_{\square,FM}$, $d_{\square,KR}$ and $d_{\square,\mathcal{P}}$) are uniformly continuous w.r.t. one another, and thus induce the same topology on $\widetilde{\mathcal{W}}_1$ (resp. $\mathcal{W}_1$) and on every uniformly bounded subset of $\widetilde{\mathcal{W}}_+$ (resp. $\mathcal{W}_+$).

(iii) The cut distances $\delta_{\square,FM}$, $\delta_{\square,KR}$, $\delta_{\square,\mathcal{P}}$ and $\delta_{\square,\mathcal{F}}$, for every choice of the convergence determining sequence $\mathcal{F}$, induce the same topology on $\widetilde{\mathcal{W}}_1$.

Proof. The first part of the corollary is a re-statement of Corollary 3.4.14. Point (i) is an immediate consequence of (3.10).

We now prove Point (ii). Thanks to (3.10) and Point (i), it is enough to consider only the Prohorov and the Kantorovitch-Rubinshtein distances. As $d_{\mathcal{P}}$ is uniformly continuous w.r.t. d_{KR} (see Lemma 3.3.22), applying Lemma 3.5.3 (remind Corollary 3.4.14) with Remark 3.5.4 in mind, we get that $\delta_{\square,\mathcal{P}}$ (resp. $d_{\square,\mathcal{P}}$) is uniformly continuous w.r.t. $\delta_{\square,\mathrm{KR}}$ (resp. $d_{\square,\mathrm{KR}}$) on every uniformly bounded subset of $\widetilde{\mathcal{W}}_+$ (resp. $\mathcal{W}_+$) As d_{KR} is also uniformly continuous w.r.t. $d_{\mathcal{P}}$ (see Lemma 3.3.22), applying again Lemma 3.5.3, we have that $\delta_{\square,\mathrm{KR}}$ (resp. $d_{\square,\mathrm{KR}}$) is uniformly continuous w.r.t. $\delta_{\square,\mathcal{P}}$ (resp. $d_{\square,\mathcal{P}}$) on every uniformly bounded subset of $\widetilde{\mathcal{W}}_+$ (resp. $\mathcal{W}_+$).

Point (iii) is an immediate consequence of Corollary 3.4.14 and Theorem 3.5.5, together with Point (ii). □

Remark 3.5.7 (Extension to uniformly bounded subsets of $\widetilde{\mathcal{W}}_+$)**.** In Theorem 3.5.5 and also in Corollary 3.5.6 (iii), one can replace $\widetilde{\mathcal{W}}_1$ by $\widetilde{\mathcal{W}}_{\mathcal{M}}$ with a bounded subset $\mathcal{M} \subset \mathcal{M}_+(\mathbf{Z})$ as soon as the distance d_{m} is defined on $\mathcal{M}_+(\mathbf{Z})$. (One has in mind the case $\mathcal{M} = \mathcal{M}_{\leq 1}(\mathbf{Z})$.) This can be seen by an easy modification in the proof of Theorem 3.5.5. Alternatively, this can be seen using scaling to reduce the case of general $\mathcal{M}$ to the case of $\mathcal{M}_{\leq 1}(\mathbf{Z})$, and then adding a cemetery point (for missing mass of measures) to $\mathbf{Z}$ to further reduce to the case of $\mathcal{M}_1(\mathbf{Z})$.

3.5.3 Completeness

Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ or $\mathcal{M}_+(\mathbf{Z})$. We shall consider a slight modification of the cut distances $d_{\square,\mathrm{m}}$ and $\delta_{\square,\mathrm{m}}$ to achieve completeness. Recall the measure $M_W \in \mathcal{M}_+(\mathbf{Z})$ defined by (3.17) associated to $W \in \mathcal{W}_+$.

Definition 3.5.8 (The cut distances $d^{\mathrm{c}}_{\square,m}$ and $\delta^{\mathrm{c}}_{\square,m}$)**.** *Let d_{m} and d^{c} be two distances on $\mathcal{M}_\epsilon(\mathbf{Z})$ with $\epsilon \in \{\leq 1, +\}$. We define the cut distance $d^{\mathrm{c}}_{\square,m}$ on the space of $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels $\mathcal{W}_\epsilon$ as:*

$$d^{\mathrm{c}}_{\square,m}(U,W) = d_{\square,\mathrm{m}}(U,W) + d^{\mathrm{c}}(M_U, M_W),$$

and the cut (pseudo-)distance $\delta^{\mathrm{c}}_{\square,m}$ on the space of unlabeled $\mathcal{M}_\epsilon(\mathbf{Z})$-valued kernels $\widetilde{\mathcal{W}}_\epsilon$ as:

$$\delta^{\mathrm{c}}_{\square,m}(U,W) = \inf_{\varphi \in S_{[0,1]}} d^{\mathrm{c}}_{\square,m}(U,W^\varphi) = \delta_{\square,\mathrm{m}}(U,W) + d^{\mathrm{c}}(M_U, M_W).$$

Notice that by Lemma 3.3.11 and the definition of M_W, the distance $d^{\mathrm{c}}_{\square,m}$ is invariant.

Lemma 3.5.9 (Topological equivalence of $\delta_{\square,\mathrm{m}}$ and $\delta^{\mathrm{c}}_{\square,m}$)**.** *Let d_{m} and d^{c} be two distances on $\mathcal{M}_\epsilon(\mathbf{Z})$, with $\epsilon \in \{\leq 1, +\}$, such that d^{c} is continuous w.r.t. d_{m} and that $d_{\square,\mathrm{m}}$ is (invariant and) smooth on $\mathcal{W}_\epsilon$. Then, the cut distance $d^{\mathrm{c}}_{\square,m}$ is invariant and smooth and $\delta^{\mathrm{c}}_{\square,m}$ is a distance on $\widetilde{\mathcal{W}}_\epsilon$. Moreover, the distances $d_{\square,\mathrm{m}}$ and $d^{\mathrm{c}}_{\square,m}$ (resp. $\delta_{\square,\mathrm{m}}$ and $\delta^{\mathrm{c}}_{\square,m}$) induce the same topology on the space $\mathcal{W}_\epsilon$ (resp. $\widetilde{\mathcal{W}}_\epsilon$).*

Proof. Let $(W_n)_{n\in\mathbb{N}}$ and W be elements of $\mathcal{W}_{\leq 1}$ such that $(W_n(x,y;\cdot))_{n\in\mathbb{N}}$ weakly converges to $W(x,y;\cdot)$ for almost every $x, y \in [0,1]$. Since the distance $d_{\square,\mathrm{m}}$ is smooth, we have that $\lim_{n\to\infty} d_{\square,\mathrm{m}}(W_n, W) = 0$. Using Lemma 3.4.9 on the continuity of the map $W \mapsto M_W$ and that d^{c} is continuous w.r.t. d_{m}, we obtain that $\lim_{n\to\infty} d^{\mathrm{c}}_{\square,m}(W_n, W) = 0$. This gives that the distance $d^{\mathrm{c}}_{\square,m}$ is smooth. Since we have already seen that $d^{\mathrm{c}}_{\square,m}$ is invariant, we deduce from Theorem 3.3.17 that $\delta^{\mathrm{c}}_{\square,m}$ is a distance on $\widetilde{\mathcal{W}}_1$.

We now prove that the two distances $d_{\square,\mathrm{m}}$ and $\delta^{\mathrm{c}}_{\square,m}$ induce the same topology (which implies that this is also true for $\delta_{\square,\mathrm{m}}$ and $\delta^{\mathrm{c}}_{\square,m}$). As $d_{\square,\mathrm{m}} \leq d^{\mathrm{c}}_{\square,m}$, convergence for $d^{\mathrm{c}}_{\square,m}$ implies convergence for $d_{\square,\mathrm{m}}$. Conversely, let $(W_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{W}_\epsilon$ that converges for $d_{\square,\mathrm{m}}$ to a limit, say $W \in \mathcal{W}_\epsilon$. Using again Lemma 3.4.9 and the continuity of d^{c} w.r.t. d_{m}, we obtain that $\lim_{n\to\infty} d^{\mathrm{c}}(M_{W_n}, M_W) = 0$. This clearly implies that the sequence $(W_n)_{n\in\mathbb{N}}$ converges to W for $d^{\mathrm{c}}_{\square,m}$. Then, the two distances have the same convergent sequences and thus induce the same topology (see Remark 3.2.1). □

Recall $\mathbf{Z}$ is a Polish space. We already proved in Proposition 3.4.6 that the space $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is separable; and we now investigate completeness of this space.

Theorem 3.5.10 ($\widetilde{\mathcal{W}}_1$ is a Polish space)**.** *Let d_{m} and d^{c} be two distances on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that d^{c} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$, d^{c} is complete and continuous w.r.t. d_{m}, and $d_{\square,\mathrm{m}}$ is (invariant) smooth and weakly regular on $\mathcal{W}_1$. Then, the space $(\widetilde{\mathcal{W}}_1, \delta^{\mathrm{c}}_{\square,m})$ is a Polish metric space.*

Note that the assumptions in Theorem 3.5.10 imply that d_{m} also induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$. Indeed, as d^{c} is continuous w.r.t. d_{m}, the topology induced by d_{m} if finer than the topology induced by d^{c}, i.e. finer than the weak topology. As $d_{\square,\mathrm{m}}$ is smooth, by Lemma 3.3.12, d_{m} is continuous w.r.t. the weak topology (i.e. the weak topology if finer than the topology induced by d_{m}), and thus d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$.

Also note that Theorem 3.5.10 can easily be extended to $\mathcal{W}_{\leq 1}$ or the space of unlabeled $\mathcal{M}$-valued kernels $\widetilde{\mathcal{W}}_{\mathcal{M}}$ when $\mathcal{M}$ is a bounded convex closed subset of $\mathcal{M}_+(\mathbf{Z})$.

Proof. From Lemma 3.5.9, we have that $\delta^{\mathrm{c}}_{\square,m}$ is a distance on $\widetilde{\mathcal{W}}_1$ which induces the same topology as $\delta_{\square,\mathrm{m}}$, and from Proposition 3.4.6, we have that $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$, and thus $(\widetilde{\mathcal{W}}_1, \delta^{\mathrm{c}}_{\square,m})$, is separable. To get that this latter space is Polish, we are left to prove that the distance $\delta^{\mathrm{c}}_{\square,m}$ is complete.

Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of probability-graphons that is Cauchy for $\delta^{\mathrm{c}}_{\square,m}$. By definition of the cut distance $\delta^{\mathrm{c}}_{\square,m}$, the sequence of probability measures $(M_{W_n})_{n\in\mathbb{N}}$ is Cauchy in $\mathcal{M}_1(\mathbf{Z})$ for the complete distance d^{c}. Thus, the sequence $(M_{W_n})_{n\in\mathbb{N}}$ is weakly convergent as d^{c} induces the weak topology, which implies that it is tight (see Lemma 3.2.8). Hence, by definition, the sequence of probability-graphons $(W_n)_{n\in\mathbb{N}}$ is tight. By Theorem 3.5.1 (i), there exists a subsequence $(W_{n_k})_{k\in\mathbb{N}}$ that converges for $\delta_{\square,\mathrm{m}}$ to a limit, say $W \in \widetilde{\mathcal{W}}_1$. This subsequence also converges for $\delta^{\mathrm{c}}_{\square,m}$ to W as $\delta_{\square,\mathrm{m}}$ and $\delta^{\mathrm{c}}_{\square,m}$ induce the same topology. Finally, because the sequence $(W_n)_{n\in\mathbb{N}}$ is Cauchy for $\delta^{\mathrm{c}}_{\square,m}$ and has a subsequence converging to W for $\delta^{\mathrm{c}}_{\square,m}$, the whole sequence must also converge to W for $\delta^{\mathrm{c}}_{\square,m}$. Consequently, the distance $\delta^{\mathrm{c}}_{\square,m}$ is complete. □

The following lemma shows that every probability measure in $\mathcal{M}_1(\mathbf{Z})$ can be represented as a constant probability-graphon.

Lemma 3.5.11 ($\mathcal{M}_1(\mathbf{Z})$ seen as a closed subset of $\mathcal{W}_1$)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that $d_{\square,\mathrm{m}}$ is (invariant and) smooth on $\mathcal{W}_1$. Then, the map $\mu \mapsto W_\mu \equiv \mu$ is an injection from $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ to $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ with a closed range and continuous inverse.*

Proof. For any $\mu \in \mathcal{M}_1(\mathbf{Z})$ consider the constant probability-graphon $W_\mu \equiv \mu$, and notice that $M_{W_\mu} = \mu$, that $W_\mu(S \times T; \cdot) = \lambda(S)\lambda(T)\,\mu$ for all measurable $S, T \subset [0,1]$, and that $W_\mu^\varphi = W_\mu$ for any measure-preserving map φ. This readily implies that for $\mu \in \mathcal{M}_1(\mathbf{Z})$ and $W \in \mathcal{W}_1$:

$$\delta_{\square,\mathrm{m}}(W_\mu, W) = d_{\square,\mathrm{m}}(W_\mu, W) = \sup_{S,T\subset[0,1]} d_{\mathrm{m}}(\lambda(S)\lambda(T)\mu, W(S\times T;\cdot)) \geq d_{\mathrm{m}}(\mu, M_W). \tag{3.21}$$

In particular, taking $W = W_\nu$ for $\nu \in \mathcal{M}_1(\mathbf{Z})$ we get that $\delta_{\square,\mathrm{m}}(W_\mu, W_\nu) \geq d_{\mathrm{m}}(\mu,\nu)$. This implies that the map $\mathcal{I} : \mu \mapsto W_\mu \equiv \mu$ is an injection, and its inverse, given by the map $W_\mu \mapsto \mu$, is 1-Lipschitz.

Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{M}_1(\mathbf{Z})$ such that the sequence $(W_{\mu_n})_{n\in\mathbb{N}}$ converges for $\delta_{\square,\mathrm{m}}$ to a limit, say W. We deduce from (3.21) that $(\mu_n)_{n\in\mathbb{N}}$ converges for d_{m} to $\mu = M_W$ and that for all measurable $S, T \subset [0,1]$, $(\lambda(S)\lambda(T)\mu_n)_{n\in\mathbb{N}}$ converges for d_{m} to $W(S\times T;\cdot)$. This implies that $W(S\times T;\cdot) = \lambda(S)\lambda(T)\mu(\cdot)$ for all measurable $S, T \subset [0,1]$, that is, $W = W_\mu$. This implies that the image by $\mathcal{I}$ of any closed subset of $\mathcal{M}_1(\mathbf{Z})$ is a closed subset of $\mathcal{W}_1$, and thus the range of $\mathcal{I}$ is closed. □

Remark 3.5.12 (Extension to isometric representation of $\mathcal{M}_1(\mathbf{Z})$)**.** If the distance d_{m}, in addition to the hypothesis of Lemma 3.5.11, is sub-homogeneous, that is, for all $\mu, \nu \in \mathcal{M}_1(\mathbf{Z})$ we have $d_{\mathrm{m}}(\mu,\nu) = \sup_{r\in[0,1]} d_{\mathrm{m}}(r\mu, r\nu)$ (which is the case if d_{m} is quasi-convex), then we deduce from (3.21) that the map $\mu \mapsto W_\mu \equiv \mu$ is isometric from $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ to $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$.

We now state characterization of compactness and completeness for the space of probability-graphons. Recall $\mathbf{Z}$ is a Polish space.

Corollary 3.5.13 (Characterization of compactness and completeness for $\widetilde{\mathcal{W}}_1$)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$, which induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$, and such that $d_{\square,\mathrm{m}}$ is (invariant) smooth and weakly regular on $\mathcal{W}_1$. We have the following properties.*

(i) $\mathbf{Z}$ *is compact* $\iff$ $(\mathcal{M}_{\leq 1}(\mathbf{Z}), d_{\mathrm{m}})$ *is compact* $\iff$ $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ *is compact.*

(ii) If $(\mathcal{M}_{\leq 1}(\mathbf{Z}), d_{\mathrm{m}})$ *is complete then* $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ *is complete.*

(iii) Assume furthermore that d_{m} is sub-homogeneous (see Remark 3.5.12). If $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ *is complete, then* $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ *is complete.*

Proof. We prove Point (i). rom Remark 3.2.7, we already know that $\mathbf{Z}$ is compact if and only if $\mathcal{M}_{\leq 1}(\mathbf{Z})$ is weakly compact, i.e. compact for d_{m} as d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$.

Now, assume that $(\mathcal{M}_{\leq 1}(\mathbf{Z}), d_{\mathrm{m}})$ is compact. Applying Theorem 3.5.1 (iii), we get that the space $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is also compact.

Conversely, assume that $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is compact. By Lemma 3.4.9, the mapping $W \mapsto M_W$ is continuous from $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ to $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$, and as $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is compact its image through this mapping is also compact. To conclude, it is enough to check that this mapping is surjective. But this is clear as the image of the constant probability-graphon $W_\mu \equiv \mu$ is $M_{W_\mu} = \mu$. Hence, $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ (and thus $(\mathcal{M}_{\leq 1}(\mathbf{Z}), d_{\mathrm{m}})$) is compact.

We prove Point (ii). Assume that $(\mathcal{M}_{\leq 1}(\mathbf{Z}), d_{\mathrm{m}})$ is complete. Thus, we can choose $d^{\mathrm{c}} = d_{\mathrm{m}}$ in Definition 3.5.8, and apply Theorem 3.5.10 to get that $(\widetilde{\mathcal{W}}_1, \delta^{\mathrm{c}}_{\square,m})$ is complete. As $d^{\mathrm{c}} = d_{\mathrm{m}}$, we have $\delta_{\square,\mathrm{m}} \leq \delta^{\mathrm{c}}_{\square,m} \leq 2\delta_{\square,\mathrm{m}}$. Hence, $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is also complete.

We prove Point (iii). Assume that $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is complete. Let $(\mu_n)_{n\in\mathbb{N}}$ be a Cauchy sequence of probability measures in $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$. By Remark 3.5.12, the sequence of constant probability-graphons $(W_{\mu_n})_{n\in\mathbb{N}}$ is also Cauchy for $\delta_{\square,\mathrm{m}}$. As $(\widetilde{\mathcal{W}}_1, \delta_{\square,\mathrm{m}})$ is complete, there exists a probability-graphon $W \in \widetilde{\mathcal{W}}_1$ such that $(W_{\mu_n})_{n\in\mathbb{N}}$ converges to W for the cut distance $\delta_{\square,\mathrm{m}}$. Thanks to Lemma 3.5.11, W is constant equal to some $\mu \in \mathcal{M}_1(\mathbf{Z})$, and $(\mu_n)_{n\in\mathbb{N}}$ converges to μ for d_{m}. Hence, $(\mathcal{M}_1(\mathbf{Z}), d_{\mathrm{m}})$ is complete. □

3.6 Sampling from probability-graphons

Measured-valued graphons allow to define models for generating random weighted graphs that are more general than the models based on real-valued graphons. We prove that the weighted graphs sampled

from probability-graphons are close to their original model for the cut distance $\delta_{\square,\mathcal{F}}$, where $\mathcal{F} = (f_k)_{k\in\mathbb{N}}$ (with $f_0 = \mathbb{1}$) is a convergence determining sequence.

It would have been more natural to work in Sections 3.6 and 3.7 with the Kantorovitch-Rubinshtein norm or the Fortet-Mourier norm that both treats all test functions in a uniform manner. Unfortunately, the supremum in the definition of both of this norms does not behave well regarding the probabilities and expectations of graphs sampled from probability-graphons. We need in our proofs (and in particular that of the First Sampling Lemma 3.6.7 below) to consider simultaneously only a finite number of test functions in order to control the probability of failure for our stochastic bounds.

3.6.1 $\mathcal{M}_1(\mathbf{Z})$-Graphs and weighted graphs

A graph $G = (V, E)$ is composed of a finite set of vertices $V(G) = V$, and a set of edges $E(G) = E$ which is a subset of $V \times V$ avoiding the diagonal. When its set of edges $E(G)$ is symmetric, we say that G is *symmetric* or *non-oriented*. We denote by $v(G) = |V(G)|$ the number of vertices of this graph, and by $e(G) = |E(G)|$ its number of edges.

Definition 3.6.1 ($\mathcal{X}$-graphs)**.** *Let $\mathcal{X}$ be a non-empty set. A* $\mathcal{X}$-graph *is a triplet $G = (V, E, \Phi)$ where (V, E) is a graph and $\Phi : E \to \mathcal{X}$ is a map that associates a decoration $x = \Phi(e) \in \mathcal{X}$ to each edge $e \in E$. When $\mathcal{X} = \mathbf{Z}$, we say that G is a* weighted graph*.*

Furthermore, the graph G is said to be symmetric *if (V, E) is a symmetric graph and if Φ is a symmetric function, i.e. for every edge $(x, y) \in E$, we have $(y, x) \in E$ and $\Phi(x, y) = \Phi(y, x)$.*

Remark 3.6.2 ($\mathcal{M}_1(\mathbf{Z})$-Graphs as probability-graphons)**.** Any labeled $\mathcal{M}_1(\mathbf{Z})$-graph G can be naturally represented as an $\mathcal{M}_1(\mathbf{Z})$-valued graphon, which we denote by W_G, in the following way. Let $G = (V, E, M)$ be a $\mathcal{M}_1(\mathbf{Z})$-graph, with $v(G) = n \in \mathbb{N}^*$. Denote by $V = [n] = \{1, \dots, n\}$ the vertices of G. Consider intervals of length $1/n$: for $1 \leq i \leq n$, let $J_i = \ ((i-1)/n, i/n]$. We then define the $\mathcal{M}_1(\mathbf{Z})$-valued graphon stepfunction W_G associated with the $\mathcal{M}_1(\mathbf{Z})$-graph G by:

$$\forall (i, j) \in E, \quad \forall (x, y) \in J_i \times J_j, \quad W_G(x, y; \mathrm{d}z) = \Phi(i, j)(\mathrm{d}z);$$

and $W_G(x, y; \mathrm{d}z)$ equals the Dirac mass at ∂ otherwise, where ∂ is an element of $\mathbf{Z}$ used as a cemetery point for missing edges in graphs.

In this section, we investigate weighted graphs sampled from probability-graphons. Hence, using the cemetery point argument in the remark above, we only consider complete graphs for the rest of this section.

Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$. If G and H have the same vertex-set, the cut distance between them is defined as the cut distance between their associated graphons:

$$d_{\square,\mathrm{m}}(G, H) = d_{\square,\mathrm{m}}(W_G, W_H).$$

When G and H does not have the same vertex-sets, as the numbering of the vertices in Remark 3.6.2 is arbitrary, we must consider the unlabeled cut distance between them defined as the cut distance between their associated graphons:

$$\delta_{\square,\mathrm{m}}(G, H) = \delta_{\square,\mathrm{m}}(W_G, W_H).$$

Remind that when the distance d_{m} derives from a norm N_{m} on $\mathcal{M}_{\pm}(\mathbf{Z})$, Lemma 3.3.19 applies, and the cut distance $d_{\square,\mathrm{m}}(G, H)$ can be rewritten as a combinatorial optimization over whole steps.

Remark 3.6.3 (Weighted graphs as $\mathcal{M}_1(\mathbf{Z})$-graphs)**.** We will sometimes need to interpret a weighted graph G as a $\mathcal{M}_1(\mathbf{Z})$-graph where a weight x on an edge is replaced by δ_x the Dirac mass laocated at x.

Notation 3.6.4 (The real-weighted graph $G[f]$)**.** For a $\mathcal{M}_1(\mathbf{Z})$-graph (resp. weighted graph) G and a function $f \in C_b(\mathbf{Z})$, we denote by $G[f]$ the real-weighted graph with the same vertex set and edge set as G, and where the edge (i, j) has weight $\Phi_{G[f]}(i, j) = \Phi_G(i, j; f) = \int_{\mathbf{Z}} f(z)\Phi_G(i, j; \mathrm{d}z)$ (resp. $\Phi_{G[f]}(i, j) = f(\Phi_G(i, j))$), where Φ_G is the decoration of the $\mathcal{M}_1(\mathbf{Z})$-graph G.

3.6.2 W-random graphs

Let W be a probability-graphon, and $x = (x_1, \dots, x_n)$, $n \in \mathbb{N}^*$, be a sequence of points from $[0,1]$. We define the $\mathcal{M}_1(\mathbf{Z})$-graph $\mathbb{H}(x, W)$ as the complete graph whose vertex set is $[n] = \{1, \dots, n\}$, and with each edge (i,j) decorated by the probability measure $W(x_i, x_j; \mathrm{d}z)$.

Let H be any $\mathcal{M}_1(\mathbf{Z})$-graph. We can define from H a random weighted (directed) graph $\mathbb{G}(H)$ whose vertex set $V(H)$ and edge set $E(H)$ are the same as H, and with each edge (i,j) having a random weight $\beta_{i,j}$ distributed according to the probability distribution decorating the edge (i,j) in H, all the weights being independent from each other. For the special case where $H = \mathbb{H}(x, W)$, we simply note $\mathbb{G}(x, W) = \mathbb{G}(\mathbb{H}(x, W))$.

An important special case is when the sequence X is chosen at random: $X = (X_i)_{1 \le i \le n}$ where the X_i are independent and uniformly distributed on $[0,1]$. For this special case, we simply note $\mathbb{H}(n, W) = \mathbb{H}(X, W)$ and $\mathbb{G}(n, W) = \mathbb{G}(X, W)$, that are conditionally on $X = x$, distributed respectively as $\mathbb{H}(x, W)$ and $\mathbb{G}(x, W)$. The random graphs $\mathbb{H}(n, W)$ and $\mathbb{G}(n, W)$ are called W-random graphs.

Remark 3.6.5 (The case of symmetric graphons)**.** In the special case where W is a symmetric probability-graphon, the $\mathcal{M}_1(\mathbf{Z})$-graph $\mathbb{H}(x, W)$ is also symmetric. From a symmetric $\mathcal{M}_1(\mathbf{Z})$-graph H, the random weighted graph $\mathbb{G}(H)$ is not necessarily symmetric, but we can define a random symmetric weighted graph $\mathbb{G}^{\mathrm{sym}}(H)$ whose vertex set $V(H)$ and $E(H)$ are the same as H, and with independent weights $\beta_{i,j} = \beta_{j,i}$ on each edge $(i,j) = (j,i)$ distributed according to $\Phi_H(i,j;\cdot)$. For $H = \mathbb{H}(x, W)$ we simply note $\mathbb{G}^{\mathrm{sym}}(x, W)$ and $\mathbb{G}^{\mathrm{sym}}(n, W)$.

For a weighted graph G, and for $1 \le k \le v(G)$, we can define the random weighted graph $\mathbb{G}(k, G)$ as being the sub-graph of G induced by a uniform random subset of k distinct vertices from G. Then, upper bounding by the probability that a uniformly-chosen map $[k] \to V(G)$ is non-injective, we get the following bound on the total variation distance between the graphs obtained from G and its associated graphon W_G:

$$d_{\mathrm{var}}(\mathbb{G}(k, G), \mathbb{G}(k, W_G)) \le \binom{k}{2} \frac{1}{v(G)},$$

where d_{var} is the total variation distance between probability measures.

3.6.3 Estimation of the distance by sampling

The first sampling lemma

In this subsection, we link sampling from graphons with the cut distance. This result is the equivalent of Lemma 10.6 in [Lov12]. The main consequence of the following lemma is that the cut distance $d_{\square,\mathcal{F}}$ between two probability-graphons can be estimated by sampling.

Notation 3.6.6 (The random stepfunction W_X)**.** For a measure-valued kernel W (resp. a real-valued kernel w) and a vector $X = (X_i)_{1 \le i \le k}$ composed of k independent random variables uniformly distributed over $[0,1]$, we denote by $W_X = W_{\mathbb{H}(k,W)}$ (resp. w_X) the random measure-valued (resp. real-valued) stepfunction with k steps of size $1/k$, and where the step (i,j) has value $W(X_i, X_j; \cdot)$ (resp. $w(X_i, X_j)$).

Lemma 3.6.7 (First Sampling Lemma)**.** *Let $\mathcal{F}$ be a convergence determining sequence. Let $k \in \mathbb{N}^*$, and $U, W \in \mathcal{W}_1$ be two probability-graphons, and let X be a random vector uniformly distributed over $[0,1]^k$. Then with probability at least $1 - 4k^{1/4}\,\mathrm{e}^{-\sqrt{k}/10}$, we have:*

$$-\frac{2}{k^{1/4}} \le \|U_X - W_X\|_{\square,\mathcal{F}} - \|U - W\|_{\square,\mathcal{F}} \le \frac{9}{k^{1/4}}.$$

An immediate consequence of Lemma 3.6.7 is that the decorated graphs with probability measures on their edges $\mathbb{H}(k, U)$ and $\mathbb{H}(k, W)$ can be coupled in order that $d_{\square,\mathcal{F}}(\mathbb{H}(k, U), \mathbb{H}(k, W))$ is close to $d_{\square,\mathcal{F}}(U, W)$ with high probability.

To prove the first sampling lemma, we first need to prove the following lemma which states that the cut norm $\|\cdot\|_{\square,\mathcal{F}}$ can be approximated by the maximum of the one-sided cut norm using a finite number of function. Remind from Remark 3.3.24 the definition of the one-sided version of the cut norm $\|\cdot\|^{+}_{\square,\mathbb{R}}$.

Lemma 3.6.8 (Approximation bound with $\|\cdot\|_{\square,\mathcal{F}}$ and $\|\cdot\|^+_{\square,\mathbb{R}}$)**.** *Let $U, W \in \mathcal{W}_1$ and let $N \in \mathbb{N}$. For every $\varepsilon = (\varepsilon_n)_{1\leq n\leq N} \in \{\pm 1\}^N$, define $g_{N,\varepsilon} = \sum_{n=1}^N 2^{-n}\varepsilon_n f_n$. Then, we have:*

$$\|U - W\|_{\square,\mathcal{F}} - 2^{-N} \leq \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)\left[g_{N,\varepsilon}\right]\|^+_{\square,\mathbb{R}} \leq \|U - W\|_{\square,\mathcal{F}}.$$

Proof. First remark that for $n \in \mathbb{N}$, f_n takes values in $[0,1]$, and thus $U[f_n] - W[f_n]$ takes values in $[-1,1]$. Remind that $f_0 = \mathbb{1}$, and thus $U[f_0] - W[f_0] \equiv 0$. Upper bounding integrals by 1 for indices $n > N$, we get:

$$\|U - W\|_{\square,\mathcal{F}} \leq \sup_{S,T\subset[0,1]} \sum_{n=1}^N 2^{-n} \left|\int_{S\times T} (U-W)[f_n](x,y)\ \mathrm{d}x\mathrm{d}y\right| + 2^{-N}.$$

And adding the non-negative terms for $n > N$, we get:

$$\sup_{S,T\subset[0,1]} \sum_{n=1}^N 2^{-n} \left|\int_{S\times T} (U-W)[f_n](x,y)\ \mathrm{d}x\mathrm{d}y\right| \leq \|U - W\|_{\square,\mathcal{F}}.$$

Using the same idea as in (3.14) and (3.15), we get:

$$\sup_{S,T\subset[0,1]} \sum_{n=1}^N 2^{-n} \left|\int_{S\times T} (U-W)[f_n](x,y)\ \mathrm{d}x\mathrm{d}y\right| = \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)\left[g_{N,\varepsilon}\right]\|^+_{\square,\mathbb{R}},$$

which concludes the proof. □

Proof of Lemma 3.6.7. Remark that for $f \in C_b(\mathbf{R})$ and $W \in \mathcal{W}_\pm$, we have $(W_X)[f] = (W[f])_X$, and we thus write $W[f]_X$ without any ambiguity.

Assume that $k \geq 2^4$ (otherwise the lower bound in the lemma is trivial). Set $N = \lceil \log_2(k^{1/4})\rceil$, so that $2^{-1}k^{-1/4} < 2^{-N} \leq k^{-1/4}$. Let $\varepsilon \in \{\pm1\}^N$. Remark that as the f_n take values in $[0,1]$, the real-valued kernels $(U-W)[f_n]$ take values in $[-1,1]$, and thus the real-valued kernel $(U-W)[g_{N,\varepsilon}]$ also take values in $[-1,1]$. Applying Lemma 10.7 in [Lov12] to the real-valued kernel $(U-W)\left[g_{N,\varepsilon}\right]$, we get with probability at least $1 - 2\,\mathrm{e}^{-\sqrt{k}/10}$ that:

$$-\frac{3}{k} \leq \|(U-W)[g_{N,\varepsilon}]_X\|^+_{\square,\mathbb{R}} - \|(U-W)[g_{N,\varepsilon}]\|^+_{\square,\mathbb{R}} \leq \frac{8}{k^{1/4}}, \tag{3.22}$$

where remind that $\|\cdot\|^+_{\square,\mathbb{R}}$ is the one-sided version of the cut norm for real-valued kernels defined in (3.13). Hence, with probability at least $1 - 2^{N+1}\,\mathrm{e}^{-\sqrt{k}/10} \geq 1 - 4k^{1/4}\,\mathrm{e}^{-\sqrt{k}/10}$, we have that the bounds in (3.22) holds for every $\varepsilon \in \{\pm1\}^N$ simultaneously; and when all of this holds, applying Lemma 3.6.8 to U, W and to U_X, W_X, we get:

$$\begin{aligned}
\|U_X - W_X\|_{\square,\mathcal{F}} &\leq \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)[g_{N,\varepsilon}]_X\|^+_{\square,\mathbb{R}} + 2^{-N} \\
&\leq \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)[g_{N,\varepsilon}]\|^+_{\square,\mathbb{R}} + \frac{9}{k^{1/4}} \\
&\leq \|U - W\|_{\square,\mathcal{F}} + \frac{9}{k^{1/4}},
\end{aligned}$$

and similarly:

$$\begin{aligned}
\|U - W\|_{\square,\mathcal{F}} &\leq \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)[g_{N,\varepsilon}]\|^+_{\square,\mathbb{R}} + 2^{-N} \\
&\leq \max_{\varepsilon\in\{\pm1\}^N} \|(U-W)[g_{N,\varepsilon}]_X\|^+_{\square,\mathbb{R}} + \frac{1}{k^{1/4}} + \frac{3}{k} \\
&\leq \|U_X - W_X\|_{\square,\mathcal{F}} + \frac{2}{k^{1/4}}.
\end{aligned}$$

This concludes the proof. □

Approximation with random weighted graphs

As a consequence of the First Sampling Lemma 3.6.7, we get that the cut distance between the sampled graphs $\mathbb{H}(k,U)$ and $\mathbb{H}(k,W)$ (with the proper coupling) is close to the cut distance between the probability-graphons U and W. The following lemma states that if k is large enough, then $\mathbb{G}(k,W)$ is close to $\mathbb{H}(k,W)$ in the cut distance $d_{\square,\mathcal{F}}$, and thus the cut distance between the random weighted graphs $\mathbb{G}(k,U)$ and $\mathbb{G}(k,W)$ is also close to $d_{\square,\mathcal{F}}(U,W)$.

Recall from Section 3.6.2 the definition of the random weighted graph $\mathbb{G}(H)$ when H is a $\mathcal{M}_1(\mathbf{Z})$-graph. Following Remarks 3.6.3 and 3.6.2, we shall see the weighted graph $\mathbb{G}(H)$ as a $\mathcal{M}_1(\mathbf{Z})$-graph or even as a probability-graphon.

Lemma 3.6.9 (Bound in probability for $d_{\square,\mathcal{F}}(\mathbb{G}(H),H)$)**.** *For every $\mathcal{M}_1(\mathbf{Z})$-graph H with k vertices, and for every $\varepsilon \geq 10/\sqrt{k}$, we have:*

$$\mathbb{P}\Big(d_{\square,\mathcal{F}}(\mathbb{G}(H),H) > 2\varepsilon\Big) \leq e^{-\varepsilon^2 k^2}.$$

Remark 3.6.10 (Bound in expectation for $d_{\square,\mathcal{F}}(\mathbb{G}(H),H)$)**.** Remind that $d_{\square,\mathcal{F}}(\mathbb{G}(H),H) \leq 1$. Applying Lemma 3.6.9 with $\varepsilon = 10/\sqrt{k}$, we get the following bound on the expectation of $d_{\square,\mathcal{F}}(\mathbb{G}(H),H)$:

$$\mathbb{E}[d_{\square,\mathcal{F}}(\mathbb{G}(H),H)] \leq \frac{20}{\sqrt{k}} + \mathrm{e}^{-100k} < \frac{21}{\sqrt{k}}.$$

Proof of Lemma 3.6.9. Let H and ε be as in the lemma. Assume that $\varepsilon \leq 1/2$ (otherwise the probability to bound in the lemma is null). To simplify the notations, denote by $G = \mathbb{G}(H)$ through this proof. Define $N = \lceil \log_2(\varepsilon^{-1}) \rceil$, so that $\sum_{n=N+1}^{\infty} 2^{-n} \leq \varepsilon$. Upper bounding by 1 the terms for $n > N$ in (3.16), we get for $U, W \in \mathcal{W}_1$:

$$d_{\square,\mathcal{F}}(U,W) \leq \sum_{n=1}^{N} 2^{-n} \|U[f_n] - W[f_n]\|_{\square,\mathbb{R}} + \ \varepsilon,$$

where remind that $\|\cdot\|_{\square,\mathbb{R}}$ is the cut norm for real-valued kernels defined in (3.13). Using this equation with the graphs G and H, we get:

$$\begin{aligned}\mathbb{P}(d_{\square,\mathcal{F}}(G,H) > 2\varepsilon) &\leq \mathbb{P}\left(\sum_{n=1}^{N} 2^{-n} d_{\square,\mathbb{R}}(G[f_n],H[f_n]) \ > \varepsilon\right) \\ &\leq \sum_{n=1}^{N} \mathbb{P}\left(d_{\square,\mathbb{R}}(G[f_n],H[f_n]) \ > \varepsilon\right), \end{aligned} \tag{3.23}$$

where $d_{\square,\mathbb{R}}$ denotes the cut distance associated to the cut norm $\|\cdot\|_{\square,\mathbb{R}}$ for real-valued graphons and kernels. Remark that for every $n \in \mathbb{N}$, $H[f_n]$ and $G[f_n]$ are real-weighted graphs with weights in $[0,1]$. Thus, by a straightforward adaptation of the proof of [Lov12, Lemma 10.11], we get:

$$\forall n \in [N], \quad \mathbb{P}(d_{\square}(G[f_n],H[f_n]) \ > \varepsilon) \leq 2 \cdot 4^k \, \mathrm{e}^{-2\varepsilon^2 k^2}. \tag{3.24}$$

Combining (3.23) and (3.24), we get for $\varepsilon > 10/\sqrt{k}$:

$$\mathbb{P}(d_{\square,\mathcal{F}}(G,H) > 2\varepsilon) \leq 2N 4^k e^{-2\varepsilon^2 k^2} \leq \mathrm{e}^{-\varepsilon^2 k^2},$$

where the last bound derives from simple calculus. This concludes the proof. □

We can apply the First Sampling Lemma 3.6.7 along with Lemma 3.6.9 to get the following lemma, equivalent of the first sampling lemma for the random weighted graph $\mathbb{G}(k,W)$:

Corollary 3.6.11 (First Sampling Lemma for $\mathbb{G}(k,W)$)**.** *Let $U, W \in \mathcal{W}_1$ be two probability-graphons, and $k \in \mathbb{N}^*$. Then, we can couple the random weighted graphs $\mathbb{G}(k,U)$ and $\mathbb{G}(k,W)$ such that with probability at least $1 - (4k^{1/4}+1)\,\mathrm{e}^{-\sqrt{k}/10}$, we have:*

$$\big|d_{\square,\mathcal{F}}(\mathbb{G}(k,U),\mathbb{G}(k,W)) - d_{\square,\mathcal{F}}(U,W)\big| \leq \frac{13}{k^{1/4}}.$$

Proof. Assume that $k \geq 13^4$ (otherwise the bound in the corollary is trivial). Then, we have with probability at least $1 - 4k^{1/4}\,\mathrm{e}^{-\sqrt{k}/10} - 2\,\mathrm{e}^{-100k} > 1 - (4k^{1/4}+1)\,\mathrm{e}^{-\sqrt{k}/10}$:

$$\begin{aligned}
\left|d_{\square,\mathcal{F}}(\mathbb{G}(k,U),\mathbb{G}(k,W)) - d_{\square,\mathcal{F}}(U,W)\right| &\leq \left|d_{\square,\mathcal{F}}(\mathbb{G}(k,U),\mathbb{G}(k,W)) - d_{\square,\mathcal{F}}(\mathbb{H}(k,U),\mathbb{H}(k,W))\right| \\
&\quad + \left|d_{\square,\mathcal{F}}(\mathbb{H}(k,U),\mathbb{H}(k,W)) - d_{\square,\mathcal{F}}(U,W)\right| \\
&\leq d_{\square,\mathcal{F}}(\mathbb{G}(k,U),\mathbb{H}(k,U)) \\
&\quad + d_{\square,\mathcal{F}}(\mathbb{G}(k,W),\mathbb{H}(k,W)) + \frac{9}{k^{1/4}} \\
&\leq \frac{40}{\sqrt{k}} + \frac{9}{k^{1/4}} \\
&\leq \frac{13}{k^{1/4}},
\end{aligned}$$

where we used the upper bound from the First Sampling Lemma 3.6.7 (which gives the coupling with the same random vector X to define both graphs $U_X = \mathbb{H}(k,U)$ and $W_X = \mathbb{H}(k,W)$) for the second inequality, the upper bound from Lemma 3.6.9 with $\varepsilon = 10/\sqrt{k}$ with both U and W for the third inequality, and that $\frac{1}{\sqrt{k}} \leq \frac{1}{13k^{1/4}}$ for the last inequality. □

3.6.4 The distance between a probability-graphon and its sample

In this section, we present the Second Sampling Lemma, that shows that a sampled $\mathcal{M}_1(\mathbf{Z})$-graph is close to its original probability-graphon with high probability. Note that we use the unlabeled cut distance $\delta_{\square,\mathcal{F}}$ rather than $d_{\square,\mathcal{F}}$ as the sample points are unordered. The bound on the distance is much weaker than the one in the First Sampling Lemma 3.6.7, but nevertheless goes to 0 as the sample size increases.

The proof is a straightforward adaptation of the proof of [Lov12, Lemma 10.16] (replacing the weak regularity lemma and the first sampling lemma by their counterparts for probability-graphons, that is Lemmas 3.4.17 and 3.6.7; the sample concentration theorem for real-valued graphons can easily be adapted to probability-graphons).

Lemma 3.6.12 (Second Sampling Lemma)**.** *Let $\mathcal{F}$ be a convergence determining sequence. Let $W \in \widetilde{\mathcal{W}}_1$ be a probability-graphon and $k \in \mathbb{N}^*$. Then, with probability at least $1 - \exp(-k/(2\ln(k)))$ we have:*

$$\delta_{\square,\mathcal{F}}(\mathbb{H}(k,W),W) \leq \frac{21}{\sqrt{\ln(k)}} \qquad \text{and} \qquad \delta_{\square,\mathcal{F}}(\mathbb{G}(k,W),W) \leq \frac{22}{\sqrt{\ln(k)}}.$$

In the above lemma, the asymmetric random graph $\mathbb{G}(k,W)$ can be replaced by the symmetric random graph $\mathbb{G}^{\mathrm{sym}}(k,W)$ without changing the proof. Similarly, the results in Section 3.6.3 can be reformulated with symmetric random graphs $\mathbb{G}^{\mathrm{sym}}(k,W)$ and $\mathbb{G}^{\mathrm{sym}}(H)$ (but with a slight modification of the proof for Lemma 3.6.9 to symmetrize the random variable $X_{i,j}$ and with the upper bound $e^{-\varepsilon^2 k^2/2}$, see also [Lov12, Lemma 10.11]).

As an immediate consequence of Lemma 3.6.12 and of the Borel-Cantelli lemma, we get the convergence of the sampled subgraphs for the cut distance $\delta_{\square,\mathcal{F}}$.

Theorem 3.6.13 (Convergence of sampled subgraphs)**.** *Let $\mathcal{F}$ be a convergence determining sequence. Let $W \in \widetilde{\mathcal{W}}_1$ be a probability-graphon. Then, a.s. the sequence of sampled subgraphs $(\mathbb{G}(k,W))_{k\in\mathbb{N}^*}$ converges to W for the cut distance $\delta_{\square,\mathcal{F}}$, and thus for any cut distance $\delta_{\square,\mathrm{m}}$ from Theorem 3.5.5.*

3.7 The Counting Lemmas and the link with the topology of probability-graphons

In this section, we introduce the homomorphism densities for probability-graphons, and then we link those to the cut distance $\delta_{\square,\mathcal{F}}$ through the Counting Lemma and the Inverse Counting Lemma. Those results are analogous to the case of real-valued graphons, see [Lov12, Chapter 7] for the definition of homomorphism densities and [Lov12, Chapter 10] for the Counting Lemma and Inverse Counting Lemma.

The main differences with [Lov12] are: the decoration of the edges of the graphs with functions from $C_b(\mathbf{Z})$; the Counting Lemma for the decorations belonging only in the convergence determining sequence $\mathcal{F}$; the more technical proof of the Inverse Counting Lemma. Note that we need to work with $\delta_{\square,\mathcal{F}}$ here as the proof of the Inverse Counting Lemma relies on the second sampling Lemma 3.6.12.

3.7.1 The homomorphism densities

In the case of non-weighted graphs, the homomorphism densities $t(F,G)$ allow to characterize a graph (up to twin-vertices expansion), and also allow to define a topology for real-valued graphons. In the case of weighted graphs and probability-graphons, we need to replace the absence/presence of edges (which is 0-1 valued) by test functions from $C_b(\mathbf{Z})$ decorating each edge.

In this section, we often need to fix the underlying (directed) graph structure $F = (V, E)$ (which may be incomplete) of a $C_b(\mathbf{Z})$-graph and to vary only the $C_b(\mathbf{Z})$-decorating functions $g = (g_e)_{e\in E}$, thus we will write $F^g = (V, E, g)$ for a $C_b(\mathbf{Z})$-graph. Moreover, when there exists a convergence determining sequence $\mathcal{F}$ such that $g_e \in \mathcal{F}$ for every edge $e \in E$, we say that F^g is a $\mathcal{F}$-graph and use the same notation conventions.

Definition 3.7.1 (Homomorphism density)**.** *We define the* homomorphism density *of a $C_b(\mathbf{Z})$-graph F^g in a signed measure-valued kernel $W \in \mathcal{W}_\pm$ as:*

$$t(F^g, W) = M_W^F(g) = \int_{[0,1]^{V(F)}} \prod_{(i,j)\in E(F)} W(x_i, x_j; g_{i,j}) \prod_{i\in V(F)} \mathrm{d}x_i. \tag{3.25}$$

Moreover, M_W^F defines a measure on $\mathbf{Z}^E$ (which we still denote by M_W^F) which is characterized by $M_W^F(\otimes_{e\in E} g_e) = M_W^F(g)$ for $g = (g_e)_{e\in E}$.

Remark 3.7.2 (Invariance under relabeling of homomorphism densities)**.** Let $\varphi : [0,1] \to [0,1]$ be a measure-preserving map. As $\varphi^{\otimes k} : (x_1, \dots, x_k) \mapsto (\varphi(x_1), \dots, \varphi(x_k))$ is a measure-preserving map on $[0,1]^k$, applying the transfer formula (see (3.1)), we get that for every $C_b(\mathbf{Z})$-graph F^g and every signed measure-valued kernel $W \in \mathcal{W}_\pm$, we have $t(F^g, W^\varphi) = t(F^g, W)$. Thus $t(F^g, \cdot)$ can be extending to $\widetilde{\mathcal{W}}_\pm$.

Remark 3.7.3. When $W \in \mathcal{W}_+$ is a measure-valued kernel, and F is the graph with two vertices and one edge, we get that $M_W^F = M_W$ the measure defined in (3.17).

Remark 3.7.4 (Adding missing edges to F)**.** When we work with probability-graphons, we can always assume the graph F to be complete, by adding the missing edges (i, j) and decorating them with the constant function $g_{(i,j)} = \mathbb{1}$.

For a finite weighted graph G, we define the *homomorphism density* of the $C_b(\mathbf{Z})$-graph F^g in G as $t(F^g, G) = t(F^g, W_G)$ (remind from Remark 3.6.2 the definition of W_G), that is:

$$t(F^g, G) = \frac{1}{v(G)^k} \sum_{(x_1,\cdots,x_k)\in V(G)^k} \prod_{(i,j)\in E(F)} g_{(i,j)}(\Phi_G(x_i, x_j)),$$

where $k = v(F)$ and $\Phi_G(x_i, x_j)$ is the weight of the directed edge from x_i to x_j.

3.7.2 The Counting Lemma

The following lemma links the homomorphism densities with the cut distance $\delta_{\square,\mathcal{F}}$ for some convergence determining sequence $\mathcal{F} = (f_n)_{n\in\mathbb{N}}$ (with $f_0 = \mathbb{1}$ and f_n takes values in $[0,1]$). This lemma is a generalization to probability-graphons of the Counting Lemma for real-valued graphons (see Lemmas 10.22 and 10.23 from [Lov12]). Recall that by Remark 3.7.2, $t(F^g, \cdot)$ is defined on $\widetilde{\mathcal{W}}_\pm$.

Lemma 3.7.5 (Counting Lemma)**.** *Let $\mathcal{F} = (f_n)_{n\in\mathbb{N}}$ be a convergence determining sequence (with $f_0 = \mathbb{1}$ and f_n takes values in $[0,1]$). Let F^g be a $\mathcal{F}$-graph, and for every edge $e \in E(F)$, let $n_e \in \mathbb{N}$ be such that $g_e = f_{n_e}$. Then, for every probability-graphons $W, W' \in \widetilde{\mathcal{W}}_1$, we have:*

$$|t(F^g, W) - t(F^g, W')| \leq \left(\sum_{e\in E(F)} 2^{n_e} \right) \delta_{\square,\mathcal{F}}(W, W').$$

Remark 3.7.6 ($W \mapsto t(F^g, W)$ is Lipschitz)**.** The Lipschitz constant given by the lemma is too large to be useful in practical cases. Nevertheless, the homomorphism density function $W \mapsto t(F^g, W)$ is Lipschitz on the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$ equipped with the cut distance $\delta_{\square,\mathcal{F}}$.

Proof of Lemma 3.7.5. To do this proof, we will apply Lemma 10.24 from [Lov12], which applies to graphs F whose edges are decorated with (possibly different) real-valued graphons $w = (w_e : e \in E(F))$, and the associated homomorphism density is defined as

$$t(F, w) = \int_{[0,1]^{V(F)}} \prod_{(i,j)\in E(F)} w_e(x_i, x_j) \prod_{i\in V(F)} \mathrm{d}x_i. \tag{3.26}$$

Remind from (3.12) that for a probability-graphon $W \in \mathcal{W}_1$ and a function $f \in \mathcal{F}$ (which is $[0,1]$-valued by our definition of convergence determining sequences), we have that $W[f]$ is a real-valued graphon. Define the collections of real-valued graphons $w = (W[g_e] : e \in E(F))$ and $w' = (W'[g_e] : e \in E(F))$. Notice from (3.25) and (3.26) that we have $t(F, w) = t(F^g, W)$ and $t(F, w') = t(F^g, W')$. Applying [Lov12, Lemma 10.24] to the graph F and edge-decorations w and w', we get:

$$|t(F^g, W) - t(F^g, W')| = |t(F, w) - t(F, w')| \leq \sum_{e\in E(F)} \|W[g_e] - W'[g_e]\|_{\square,\mathbb{R}},$$

where the norm $\|\cdot\|_{\square,\mathbb{R}}$ in the upper bound is the cut norm for real-valued graphons (see (3.13) for definition of this object). For $e \in E(F)$, by definition of the cut distance $d_{\square,\mathcal{F}}$ and using (3.11), we have:

$$\|W[g_e] - W'[g_e]\|_{\square,\mathbb{R}} \ \leq 2^{n_e}\ d_{\square,\mathcal{F}}(W, W').$$

Hence, combining all those upper bounds, we get the bound in the lemma but with $d_{\square,\mathcal{F}}$ instead of $\delta_{\square,\mathcal{F}}$. Since $t(F^g, \cdot)$ is invariant under relabeling by Remark 3.7.2, taking the infimum other all relabelings allows to replace $d_{\square,\mathcal{F}}$ by $\delta_{\square,\mathcal{F}}$ and to get the bound in the lemma. □

We have just seen that homomorphism densities defined using only functions from $\mathcal{F}$ are Lipschitz. We are going to see that the other homomorphism densities are nevertheless continuous.

Lemma 3.7.7 (Weak Counting Lemma)**.** *Let $\mathcal{F}$ be a convergence determining sequence (with $f_0 = \mathbb{1}$). Let $(W_n)_{n\in\mathbb{N}}$ and W be probability-graphons such that $\lim_{n\to\infty} t(F^g, W_n) = t(F^g, W)$ for all $\mathcal{F}$-graph F^g (which in particular the case if $\lim_{n\to\infty} \delta_{\square,\mathcal{F}}(W_n, W) = 0$ by the Counting Lemma 3.7.5). Then, for every $C_b(\mathbf{Z})$-graph F^g we have:*

$$t(F^g, W_n) \underset{n\to\infty}{\longrightarrow} t(F^g, W).$$

Proof. Let $F = (V, E)$ be some fixed (directed) graph. By assumption, we have for all edge-decorations $g = (g_e)_{e\in E}$ in $\mathcal{F}$ that $\lim_{n\to\infty} M^F_{W_n}(\otimes_{e\in E} g_e) = M^F_W(\otimes_{e\in E} g_e)$ (see Definition 3.7.1). By [EK09, Chapter 3, Proposition 4.6], $\mathcal{F}^{\otimes E}$ is a (countable) convergence determining family on $\mathcal{M}_+(\mathbf{Z}^E)$. Thus, the sequence of measures $(M^F_{W_n})_{n\in\mathbb{N}}$ converges to M^F_W for the weak topology on $\mathcal{M}_+(\mathbf{Z}^E)$. And in particular, for every edge-decoration function $g = (g_e)_{e\in E}$ (here for every $e \in E$, $g_e \in C_b(\mathbf{Z})$ is arbitrary) we have $M^F_{W_n}(\otimes_{e\in E} g_e) = t(F^g, W_n) \to t(F^g, W) = M^F_W(\otimes_{e\in E} g_e)$ as $n \to \infty$. This being true for all choices of the graph F, it concludes the proof. □

3.7.3 The Inverse Counting Lemma

The goal of this subsection is to establish a converse to the Counting Lemma: if two probability-graphons are close in terms of homomorphism densities, then they are close w.r.t. the cut distance $\delta_{\square,\mathcal{F}}$.

Lemma 3.7.8 (Inverse Counting Lemma)**.** *Let $\mathcal{F} = (f_n)_{n\in\mathbb{N}}$ be a convergence determining sequence (with $f_0 = \mathbb{1}$ and f_n takes values in $[0,1]$). Let $U, W \in \widetilde{\mathcal{W}}_1$ be two probability-graphons, and let $k, n_0 \in \mathbb{N}^*$. Assume that we have $|t(F^g, U) - t(F^g, W)| \leq 2^{-k-n_0k^2}$ for every (complete) $C_b(\mathbf{Z})$-graph F^g with k vertices and such that the edge-decoration functions $g = (g_e)_{e\in E(F)}$ are products (without repetition) of the functions $(f_n)_{1\leq n\leq n_0}$ and $(1 - f_n)_{1\leq n\leq n_0}$. Then, we have:*

$$\delta_{\square,\mathcal{F}}(U, W) \leq \frac{44}{\sqrt{\log(k)}} + 2^{-n_0}.$$

To prove Lemma 3.7.8, we first need to prove the special case where the space $\mathbf{Z}$ is finite.

Lemma 3.7.9 (Inverse Counting Lemma, case with finite space $\mathbf{Z}$)**.** *Assume that the space $\mathbf{Z}$ is finite with cardinality n_1, for simplicity say $\mathbf{Z} = [n_1]$. Define the indicator functions $f_n : z \mapsto \mathbb{1}_{\{z=n\}}$ for $n \in [n_1]$, in particular $\mathcal{H} = (f_n)_{1\leq n\leq n_1}$ is a finite convergence determining sequence. Let $U, W \in \widetilde{\mathcal{W}_1}$ be two probability-graphons, and let $k \in \mathbb{N}^*$. Assume that we have $|t(F^g, U) - t(F^g, W)| < 2^{-k-\log_2(n_1)k^2}$ for every (complete) $\mathcal{H}$-graph F^g with k vertices.*

Then, for any (possibly finite) convergence determining sequence $\mathcal{F}$, we have:

$$\delta_{\square,\mathcal{F}}(U, W) \leq \frac{44}{\sqrt{\log(k)}}.$$

Abusing notations, we can identify a weight-value $n \in \mathbf{Z}$ with its indicator functions f_n, and doing this identification for edge-decoration functions, we can identify a $\mathcal{F}$-graph F^g with its corresponding weighted graph. In particular, doing so we get $t(F^g, W) = \mathbb{P}(\mathbb{G}(k, W) = F^g)$ for every $\mathcal{F}$-graph F^g with k vertices. The proof of Lemma 3.7.9 is then a straightforward adaptation of the proof of [Lov12, Lemma 10.31 and Lemma 10.32].

Proof of Lemma 3.7.8. As the functions $(f_n)_{n\in\mathbb{N}}$ take value in $[0, 1]$, for all φ measure-preserving map, for all $S, T \subset [0, 1]$ measurable sets and for all $n \in \mathbb{N}$, we have:

$$\Big|U(S \times T; f_n) - W^{\varphi}(S \times T; f_n)\Big| \leq 1.$$

Using this bound, we get the following bound (remind that $f_0 = \mathbb{1}$):

$$\delta_{\square,\mathcal{F}}(U, W) \leq \inf_{\varphi\in S_{[0,1]}} \sup_{S,T\subset[0,1]} \sum_{n=1}^{n_0} 2^{-n}\Big|U(S \times T; f_n) - W^{\varphi}(S \times T; f_n)\Big| + 2^{-n_0}. \tag{3.27}$$

Hence, for a point $z \in \mathbf{Z}$, the upper bound in (3.27) uses only the information given by $(f_n(z))_{n\in[n_0]}$. In order to discretize the space $[0, 1]^{n_0}$, we replace a point $p = (p_1, \dots, p_{n_0}) \in [0, 1]^{n_0}$ by a random point $(Y_1, \dots, Y_{n_0}) \in \{0, 1\}^{n_0}$ where the Y_i are independent random variables with Bernoulli distribution of parameter p_i. This leads us to replace a $\mathcal{M}_1(\mathbf{Z})$-valued kernel W by the $\mathcal{M}_1(\{0, 1\}^{n_0})$-valued kernel $\tilde{W}$ defined for all $(x, y) \in [0, 1]^2$, and for all $s = (s_1, \dots, s_{n_0}) \in \{0, 1\}^{n_0}$ as:

$$\tilde{W}(x, y; \{s\}) = W(x, y; f^s) \qquad \text{where} \qquad f^s = \prod_{n=1}^{n_0} f_n^{s_n}(1 - f_n)^{1-s_n}.$$

Fix some enumeration $(s^m)_{m\in[2^{n_0}]}$ of the points in $\{0, 1\}^{n_0}$, and define the indicator functions $\tilde{h}_m : s \mapsto \mathbb{1}_{\{s=s^m\}}$ for $m \in [2^{n_0}]$, in particular $\tilde{\mathcal{H}} = (\tilde{h}_m)_{1\leq m\leq 2^{n_0}}$ is a finite convergence determining sequence on $\mathcal{M}_+(\{0, 1\}^{n_0})$. Let $F^{\tilde{g}}$ be a $\tilde{\mathcal{H}}$-graph with vertex set $V(F) = [k]$, and for every edge $e \in E(F)$, let $m_e \in [2^{n_0}]$ be such that $\tilde{g}_e = \tilde{h}_{m_e}$. Define the edge-decoration functions $g = (g_e)_{e\in E(F)}$ for every edge $e \in E(F)$ as $g_e = f^{s^{m_e}}$, then we get:

$$t(F^{\tilde{g}}, \tilde{W}) = \int_{[0,1]^k} \prod_{(i,j)\in E(F)} \tilde{W}(x_i, x_j; \{s^{m_e}\}) \prod_{i=1}^{k} \mathrm{d}x_i = t(F^g, W).$$

Thus, the $\mathcal{M}_1(\{0, 1\}^{n_0})$-valued graphons $\tilde{U}$ and $\tilde{W}$ inherit the bounds on the homomorphism densities: for every $\tilde{\mathcal{H}}$-graph $F^{\tilde{g}}$, we have $|t(F^{\tilde{g}}, \tilde{U}) - t(F^{\tilde{g}}, \tilde{W})| \leq 2^{-k-n_0k^2}$.

Define for all $n \in [n_0]$ the function $\tilde{f}_n : s \mapsto \mathbb{1}_{\{s_n=1\}}$, and let $\tilde{\mathcal{F}}$ be the concatenation of $(\tilde{f}_n)_{n\in[n_0]}$ and $\tilde{\mathcal{H}}$, in particular $\tilde{\mathcal{F}}$ is a finite convergence determining sequence on $\mathcal{M}_+(\{0, 1\}^{n_0})$. Finally, as $\delta_{\square,\tilde{\mathcal{F}}}(\tilde{U}, \tilde{W})$ upper bounds the first term in the upper bound of (3.27), applying Lemma 3.7.9 with the finite space $\mathbf{Z} = \{0, 1\}^{n_0}$ and $n_1 = 2^{n_0}$, the finite convergence determining sequences $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{H}}$, and the $\mathcal{M}_1(\{0, 1\}^{n_0})$-valued graphons $\tilde{U}$ and $\tilde{W}$, we get:

$$\delta_{\square,\mathcal{F}}(U, W) \leq \frac{44}{\sqrt{\ln(k)}} + 2^{-n_0},$$

which concludes the proof. □

3.7.4 Subgraph sampling and the topology of probability-graphons

Thanks to the Weak Counting Lemma 3.7.7 and the Inverse Counting Lemma 3.7.8, we can formulate a new informative characterization of weak isomorphism, i.e. equality in the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$.

Note that the propositions and the theorem in this subsection can in particular be applied to $\delta_{\square,\mathrm{m}}$ when d_{m} is a quasi-convex distance continuous w.r.t. the weak topology, as then $d_{\square,\mathrm{m}}$ is invariant, smooth, weakly regular and regular w.r.t. the stepping operator (remind Proposition 3.4.13).

Proposition 3.7.10 (Characterization of equality for $\delta_{\square,\mathrm{m}}$)**.** *Let $U, W \in \mathcal{W}_1$ be two probability-graphons. The following properties are equivalent:*

(i) $\delta_{\square,\mathrm{m}}(U, W) = 0$ *for some (and hence for every) choice of the distance d_{m} on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is (invariant) smooth.*

(ii) There exist $\varphi, \psi \in \bar{S}_{[0,1]}$ such that $U^\varphi = W^\psi$ almost everywhere on $[0,1]^2$.

(iii) $t(F^g, U) = t(F^g, W)$ *for all $C_b(\mathbf{Z})$-graph F^g.*

(iv) $t(F^g, U) = t(F^g, W)$ *for all $\mathcal{F}$-graph F^g.*

Proof. The equivalence between Properties (i) and (ii) is a consequence of Proposition 3.3.18 on the cut distance. Remark 3.7.2 gives that Property (ii) implies Property (iii). It is clear that Property (iii) implies Property (iv). The Inverse Counting Lemma 3.7.8 with the Weak Counting Lemma 3.7.7 give that Property (iv) implies Property (i) (with $d_{\mathrm{m}} = d_{\mathcal{F}}$). Hence, we have the desired equivalence. □

Thanks to the Weak Counting Lemma 3.7.7 and the Inverse Counting Lemma 3.7.8, we get the following characterization of the topology induced by the cut distance $\delta_{\square,\mathrm{m}}$ on the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$ in terms of homomorphism densities

Theorem 3.7.11 (Characterization of the topology induced by $\delta_{\square,\mathrm{m}}$)**.** *Let $(W_n)_{n\in\mathbb{N}}$ and W be unlabeled probability-graphons from $\widetilde{\mathcal{W}}_1$. The following properties are equivalent:*

(i) $\lim_{n\to\infty} \delta_{\square,\mathrm{m}}(W_n, W) = 0$ *for some (and hence for every) choice of the distance d_{m} on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ and the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is (invariant) smooth, weakly regular and regular w.r.t. the stepping operator.*

(ii) $\lim_{n\to\infty} t(F^g, W_n) = t(F^g, W)$ *for all $C_b(\mathbf{Z})$-graph F^g.*

(iii) $\lim_{n\to\infty} t(F^g, W_n) = t(F^g, W)$ *for all $\mathcal{F}$-graph F^g.*

(iv) For all $k \geq 2$, the sequence of sampled subgraphs $(\mathbb{G}(k, W_n))_{n\in\mathbb{N}}$ converges in distribution to $\mathbb{G}(k, W)$.

In particular, the topology induced by the cut distance $\delta_{\square,\mathcal{F}}$ on the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$ coincides with the topology generated by the homomorphism densities functions $W \mapsto t(F^g, W)$ for all $C_b(\mathbf{Z})$-graph F^g.

Proof. By Theorem 3.5.5, convergence for $\delta_{\square,\mathcal{F}}$ is equivalent to convergence for $\delta_{\square,\mathrm{m}}$ for every choice of the distance d_{m} on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ and the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is (invariant) smooth, weakly regular and regular w.r.t. the stepping operator. Taking $d_{\mathrm{m}} = d_{\mathcal{P}}$, the Weak Counting Lemma 3.7.7 gives that Property (i) implies Property (ii). It is clear that Property (ii) implies Property (iii). The Inverse Counting Lemma 3.7.8 with the Weak Counting Lemma 3.7.7 give that Property (iii) implies Property (i) (with $d_{\mathrm{m}} = d_{\mathcal{F}}$). Notice that when F is the complete graph with k vertices, M_W^F is the joint measure of all the edge-weights of the random graph $\mathbb{G}(k, W)$, and thus characterizes the distribution random graph $\mathbb{G}(k, W)$. Thus (remind Definition 3.7.1), Property (ii) and Property (iv) are equivalent. Hence, we have the desired equivalence. □

Question 3.7.12 (Do the distances $d_{\square,\mathcal{F}}$ all induce the same topology?)**.** *Even though every distance $\delta_{\square,\mathcal{F}}$ generates the same topology on the space of unlabeled probability-graphons $\widetilde{\mathcal{W}}_1$, it is an open question whether or not this is also the case that every distance $d_{\square,\mathcal{F}}$ induces the same topology on the space of labeled probability-graphons $\mathcal{W}_1$.*

The following proposition states that to prove existence of a limit unlabeled probability-graphon it is enough to prove that there exists a convergence determining sequence $\mathcal{F}$ such that for every $\mathcal{F}$-graph F^g the homomorphism densities $t(F^g, \cdot)$ converge.

Proposition 3.7.13 (Existence of a limit unlabeled probability-graphon)**.** *Let d_{m} be a distance on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ such that d_{m} induces the weak topology on $\mathcal{M}_{\leq 1}(\mathbf{Z})$ and the cut distance $d_{\square,\mathrm{m}}$ on $\mathcal{W}_1$ is (invariant) smooth, weakly regular and regular w.r.t. the stepping operator.*

Let $(W_n)_{n\in\mathbb{N}}$ be sequence of unlabeled probability-graphons in $\widetilde{\mathcal{W}}_1$ that is tight. Let $\mathcal{F}$ be a convergence determining sequence such that for every $\mathcal{F}$-graph F^g the sequence $(t(F^g, W_n))_{n\in\mathbb{N}}$ converges. Then, there exists an unlabeled probability-graphon $W \in \widetilde{\mathcal{W}}_1$ such that the sequence $(W_n)_{n\in\mathbb{N}}$ converges to W for $\delta_{\square,\mathrm{m}}$.

Proof. Since the sequence $(W_n)_{n\in\mathbb{N}}$ is tight, by Theorem 3.5.1, there exists a subsequence $(W_{n_k})_{k\in\mathbb{N}}$ of the sequence $(W_n)_{n\in\mathbb{N}}$ that converges to some W for $\delta_{\square,\mathrm{m}}$. By Theorem 3.7.11, we have for every $\mathcal{F}$-graph F^g that $\lim_{k\to\infty} t(F^g, W_{n_k}) = t(F^g, W)$; and as we already know that the sequence $(t(F^g, W_n))_{n\in\mathbb{N}}$ converges, we have that $\lim_{n\to\infty} t(F^g, W_n) = t(F^g, W)$. Hence, by Theorem 3.7.11, we get that the sequence $(W_n)_{n\in\mathbb{N}}$ converges to W for $\delta_{\square,\mathrm{m}}$. □

Remark 3.7.14. For the special case $\mathbf{Z} = \{0, 1\}$, which is compact, we find back that convergence for real-valued graphons is characterized by the convergence of the homomorphism densities. Notice the tightness condition of Proposition 3.7.13 is automatically satisfied as $\mathbf{Z}$ is compact.

3.8 Proofs of Theorem 3.5.1 and Theorem 3.5.5

We start by proving a lemma that allows to construct a convergent subsequence and its limit kernel for a tight sequence of measure-valued kernels. This lemma is useful for the proofs of both Theorem 3.5.1 and Theorem 3.5.5. For the proof of Theorem 3.5.5, we will also need the convergence to hold simultaneously for two distances $\delta_\square$ and $\delta'_\square$. Remind from Definition 3.4.1 the definition of the stepfunction $W_{\mathcal{P}}$ for a signed measure-valued kernel W and a finite partition $\mathcal{P}$ of $[0, 1]$. For a finite partition $\mathcal{P}$ of $[0, 1]$, define its diameter as the smallest diameter of its sets, i.e. $\mathrm{diam}(\mathcal{P}) = \min_{S\in\mathcal{P}} \mathrm{diam}(S) = \min_{S\in\mathcal{P}} \sup_{x,y\in S} |x - y|$.

Lemma 3.8.1 (Convergence using given approximation partitions)**.** *Let d be an invariant smooth distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$). Let $(W_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) which is tight (resp. uniformly bounded and tight). Further assume that we are given, for every $n, k \in \mathbb{N}$, partitions $\mathcal{P}_{n,k}$ of $[0, 1]$, such that these partitions and the corresponding stepfunctions $W_{n,k} = (W_n)_{\mathcal{P}_{n,k}}$ satisfy the following conditions:*

(i) the partition $\mathcal{P}_{n,k+1}$ is a refinement of $\mathcal{P}_{n,k}$,

(ii) $\mathrm{diam}(\mathcal{P}_{n,k}) \leq 2^{-k}$ and $|\mathcal{P}_{n,k}| = m_k$ depends only on k (and not on n),

(iii) $d(W_n, W_{n,k}) \leq 1/(k + 1)$.

Then, there exists a subsequence $(W_{n_\ell})_{\ell\in\mathbb{N}}$ of the sequence $(W_n)_{n\in\mathbb{N}}$ and a measure-valued kernel $W \in \mathcal{W}_1$ (resp. $W \in \mathcal{W}_+$ or $W \in \mathcal{W}_\pm$) such that $(W_{n_\ell})_{\ell\in\mathbb{N}}$ converges to W for $\delta_\square$.

Moreover, assume that d' is another invariant smooth distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) such that for every $n \in \mathbb{N}$ and $k \in \mathbb{N}$, $W_{n,k}$ also satisfies:

(iv) $d'(W_n, W_{n,k}) \leq 1/(k + 1)$.

Then, there exists a subsequence $(W_{n_\ell})_{\ell\in\mathbb{N}}$ of the sequence $(W_n)_{n\in\mathbb{N}}$ and a measure-valued kernel $W \in \mathcal{W}_1$ (resp. $W \in \mathcal{W}_+$ or $W \in \mathcal{W}_\pm$) such that $(W_{n_\ell})_{\ell\in\mathbb{N}}$ converges to the same measure-valued kernel W simultaneously both for $\delta_\square$ and for $\delta'_\square$, the cut distance associated with d'.

Proof. We adapt here the general scheme from the proof of Theorem 9.23 in [Lov12], but the argument for the convergence of the U_k, defined below, takes into account that measure-valued kernels are infinite-dimensional valued. We set (remind from (3.2) the definition of $\|\cdot\|_\infty$):

$$C = \sup_{n\in\mathbb{N}} \|W_n\|_\infty < +\infty.$$

The proof is divided into four steps.

Step 1: Without loss of generality, the partitions $\mathcal{P}_{n,k}$ are made of intervals. For every $n \in \mathbb{N}$, we can rearrange the points of $[0,1]$ by a measure-preserving map so that the partitions $\mathcal{P}_{n,k}$ are made of intervals, and we replace W_n by its rearranged version.

An argument similar to the next lemma is used in the proof in [Lov12, Proof of Theorem 9.23] without any reference. So, we provide a proof and stress that diameters of the partitions shrinking to zero is an important assumption (see Remark 3.8.3 below).

Lemma 3.8.2 (Kernel rearrangement with interval partitions)**.** *Let $(\mathcal{P}_k)_{k\in\mathbb{N}}$ be a refining sequence of finite partitions of $[0,1]$ whose diameter converges to zero. Then, there exist a measure-preserving map $\varphi \in \bar{S}_{[0,1]}$ and a refining sequence of partitions made of intervals $(\mathcal{Q}_k)_{k\in\mathbb{N}}$ such that for all $k \in \mathbb{N}$, and all set $S \in \mathcal{P}_k$ there exists a set $R \in \mathcal{Q}_k$ such that a.e. $\mathbb{1}_R = \mathbb{1}_{\varphi^{-1}(S)}$.*

In particular, if W is a signed measure-valued kernel, then for $U = W^\varphi$, we have that a.e. $U_{\mathcal{Q}_k} = (W^\varphi)_{\mathcal{Q}_k} = (W_{\mathcal{P}_k})^\varphi$ for all $k \in \mathbb{N}$.

Notice that, according to Remark 3.4.4, the sequence of refining partition $(\mathcal{P}_k)_{k\in\mathbb{N}}$, with a partition diameter converging to 0, separates points and thus generates the Borel σ-field of $[0,1]$.

Proof. Consider the infinite Neveu-Ulam-Harris tree $\mathcal{T}^\infty = \{u_1 \cdots u_k : k \in \mathbb{N}, u_1, \ldots, u_k \in \mathbb{N}^*\}$, where for $k = 0$ the empty word $u = \partial$ is called the root node of the tree; for a node $u = u_1 \cdots u_k \in \mathcal{T}^\infty$, we define its height as $h(u) = k$, and if $k > 0$ we define its parent node as $p(u) = u_1 \cdots u_{k-1}$ and we say that u is a child node of $p(u)$. We order vertices on the tree $\mathcal{T}^\infty$ with the lexicographical (total) order $<_{\text{lex}}$. As a first step, we construct a subtree $\mathcal{T} \subset \mathcal{T}^\infty$ that indexes the sets in the partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$, such that for every $k \in \mathbb{N}$, $\mathcal{P}_k = \{S_u : u \in \mathcal{T}, h(u) = k\}$, and such that if $S_v \subset S_u$ with $S_v \in \mathcal{P}_k$ and $S_u \in \mathcal{P}_{k-1}$, then $p(v) = u$.

Without loss of generality, we may assume that $\mathcal{P}_0 = \{[0,1]\}$, and we label its only set by the empty word ∂, and we set $S_\partial = [0,1]$. Then, suppose we have already labeled the sets from $\mathcal{P}_0, \ldots, \mathcal{P}_k$, and we proceed to label the sets from $\mathcal{P}_{k+1}$. Because the partition $\mathcal{P}_{k+1}$ is a refinement of $\mathcal{P}_k$, we can group the sets of $\mathcal{P}_{k+1}$ by their unique parent set from $\mathcal{P}_k$, i.e. for every $S_u \in \mathcal{P}_k$, let $\mathcal{O}_u = \{S \in \mathcal{P}_{k+1} : S \subset S_u\}$, then $S_u = \cup_{S\in\mathcal{O}_u} S$. For $S_u \in \mathcal{P}_k$, we fix an arbitrary enumeration of $\mathcal{O}_u = \{S^1, \ldots, S^\ell\}$ with $\ell = |\mathcal{O}_u|$, then label the set S^j by uj, and set $S^j = S_{uj}$; remark that the parent node of $w = uj$ is $p(w) = u$, and the height of node w is $h(w) = h(u) + 1 = k + 1$. Hence, we have labeled every set from $\mathcal{P}_{k+1}$. To finish the construction, we set $\mathcal{T} = \{u : \exists k \in \mathbb{N}, \exists S \in \mathcal{P}_k,\ S \text{ has label } u\}$.

We now proceed to construct a measure-preserving map ψ such that the image of every set S_u is a.e. equal to an interval, and such that those intervals are ordered w.r.t. to the order of their labels in $\mathcal{T}$.

Define the map $\sigma : [0,1] \to \mathcal{T}^\mathbb{N}$ by $\sigma(x) = (u^k(x))_{k\in\mathbb{N}} \in \mathcal{T}^\mathbb{N}$ where $u^k(x)$ is the only node of $\mathcal{T}$ with height k such that $x \in S_{u^k(x)}$ (and thus $u^{k+1}(x)$ is a child node of $u^k(x)$). Remark that if $u^{k_0}(x) <_{\text{lex}} u^{k_0}(y)$ for some $k_0 \in \mathbb{N}$, then $u^k(x) <_{\text{lex}} u^k(y)$ for every $k \geq k_0$. We extend naturally the total order $<_{\text{lex}}$ from $\mathcal{T}$ to a the total order on $\mathcal{T}^\mathbb{N}$: for $(u^k)_{k\in\mathbb{N}}, (v^k)_{k\in\mathbb{N}} \in \mathcal{T}^\mathbb{N}$, $(u^k)_{k\in\mathbb{N}} <_{\text{lex}} (v^k)_{k\in\mathbb{N}}$ if $u^{k_0} <_{\text{lex}} v^{k_0}$ where k_0 is the smallest k such that $u^k \neq v^k$.

For every $u \in \mathcal{T}$, define:

$$A^-(u) = \bigcup_{v <_{\text{lex}} u\,:\, h(v)=h(u)} S_v \quad \text{and} \quad A^+(u) = A^-(u) \cup S_u,$$

and then define $C^-(u) = \lambda(A^-(u))$ and $C^+(u) = \lambda(A^+(u))$. Now, define ψ as, for $x \in [0,1]$:

$$\psi(x) = \lambda(A^-(x)) \quad \text{where} \quad A^-(x) = \{y \in [0,1] : \sigma(y) <_{\text{lex}} \sigma(x)\} = \cup_{k\in\mathbb{N}} A^-(u^k(x)).$$

Moreover, as the sequence of partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ has a diameter that converges to zero, and thus separates points, the map σ is injective. Thus, we also have:

$$\psi(x) = \lambda(A^+(x)) \quad \text{where} \quad A^+(x) = \{y \in [0,1] : \sigma(y) \leq_{\text{lex}} \sigma(x)\} = A^-(x) \cup \{x\}.$$

Remark that both $A^-(x)$ and $A^+(x)$ are Borel measurable.

Remark that for every $k \in \mathbb{N}$, we have $A^-(u^k(x)) \subset A^-(x) \subset A^+(x) \subset A^+(u^k(x))$. In particular, for every $u \in \mathcal{T}$, we have $\psi(S_u) \subset [C^-(u), C^+(u)]$; however $\psi(S_u)$ is not necessarily an interval, but we shall see that $\lambda(\psi(S_u)) = C^+(u) - C^-(u)$, i.e. $\psi(S_u)$ is a.e. equal to $[C^-(u), C^+(u)]$. Remark that, as

the sequence of partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ is refining, we get that $[C^-(u), C^+(u)] = \cup_{v\,:\,p(v)=u}[C^-(v), C^+(v)]$ for every $u \in \mathcal{T} \setminus \{\partial\}$.

As the diameter of the partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ converges to zero, we have the following alternative formula for ψ:

$$\psi(x) = \lim_{k\to\infty} C^-(u^k(x)) = \lim_{k\to\infty} C^+(u^k(x)).$$

For every $k \in \mathbb{N}$, the map $x \mapsto C^-(u^k(x))$ is a simple function (constant on each $S \in \mathcal{P}_k$ and takes finitely-many values), and thus ψ is measurable as a limit of measurable maps.

We outline the rest of the proof. We first prove that ψ is measure-preserving. Secondly, we prove that ψ is a.e. bijective and construct its a.e. inverse map φ. Thirdly, we prove that $(\varphi^{-1}(\mathcal{P}_k))_{k\in\mathbb{N}}$ is a refining sequence of partitions. And lastly, we approximate almost everywhere the sequence of partitions $(\varphi^{-1}(\mathcal{P}_k))_{k\in\mathbb{N}}$ by a sequence of refining partitions composed of intervals.

We now prove that ψ is measure preserving. Remark that $\psi(x)$ is a non-decreasing function of $\sigma(x)$ for the total relation order $<_{\text{lex}}$, i.e. $\psi(y) \le \psi(x)$ if and only if $\sigma(y) \le_{\text{lex}} \sigma(x)$. Hence, $\psi^{-1}([0, \psi(x)]) = \{y \in [0,1] : \sigma(y) \le_{\text{lex}} \sigma(x)\}$, and we have:

$$\lambda(\psi^{-1}([0, \psi(x)])) = \lambda(\{y \in [0,1] : \sigma(y) \le_{\text{lex}} \sigma(x)\}) = \psi(x).$$

Thus, to show that ψ is measure preserving we just need to show that $\psi([0,1])$ is dense in $[0,1]$. For every $u \in \mathcal{T}$, as $\psi(S_u) \subset [C^-(u), C^+(u)]$, we know that the interval $[C^-(u), C^+(u)]$ contains at least one point of the form $\psi(x)$. Remark that for all $k \in \mathbb{N}$, we have $[0,1] = \cup_{u\in\mathcal{T}:h(u)=k}[C^-(u), C^+(u)]$. Hence, as $\lambda([C^-(u), C^+(u)]) = \lambda(S_u) \le \text{diam}(\mathcal{P}_{h(u)})$ for every $u \in \mathcal{T}$, and as the diameter of the partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ converges to zero, we know that each interval of positive length contains a point of the form $\psi(x)$ for some $x \in [0,1]$, which implies that $\psi([0,1])$ is indeed dense in $[0,1]$.

We now prove that ψ is a.e. bijective and construct its a.e. inverse map φ. Without loss of generality, assume that there is no set S_u with null measure. Consider two distinct elements $x, y \in [0,1]$ such that $\sigma(x) <_{\text{lex}} \sigma(y)$. Assume that $\psi(x) = \psi(y)$, and let $N \in \mathbb{N}$ be the last index k such that $u^k(x) = u^k(y)$. Then, for every $k > N$, we have $u^k(x) <_{\text{lex}} u^k(y)$, which implies that $\psi(x) \le C^+(u^k(x)) \le C^-(u^k(y)) \le \psi(y)$; and thus $\psi(x) = \psi(y) = C^+(u^k(x)) = C^-(u^k(y))$, which in turn implies that there is no node of $\mathcal{T}$ between $u^k(x)$ and $u^k(y)$. Remark that this situation is analogous to the terminating decimal versus repeating decimal situation. Hence, we proved that there is no node between $u^{N+1}(x)$ and $u^{N+1}(y)$ and that for every $k > N$, $u^{k+1}(x)$ is the right-most child of $u^k(x)$, and $u^{k+1}(y)$ is the left-most child of $u^k(y)$ (i.e. $u^{k+1}(x) = u^k(x)|\mathcal{O}_{u^k(x)}|$ and $u^{k+1}(y) = u^k(y)1$). Remind that the map σ is injective. Putting all of this together, we get that the set $\{(x,y) \in [0,1] : \psi(x) = \psi(y),\, x < y\}$ can be indexed by the nodes of $\mathcal{T}$, and is thus at most countable. Hence, the map ψ is injective on a subset $D \subset [0,1]$ with measure one (indeed $[0,1] \setminus D$ is at most countable), and as ψ is measure preserving, we get that $\psi(D)$ has measure one, and thus ψ is bijective from D to $\psi(D)$, that is, ψ is a.e. bijective. We construct the map φ as the inverse map of ψ for $x \in \psi(D)$ and $\varphi(x) = 0$ for $x \in [0,1] \setminus \psi(D)$. Without loss of generality, we assume that $0 \notin D$. Thus, φ is the a.e. inverse map of ψ, that is, $\varphi \circ \psi(x) = \psi \circ \varphi(x) = x$ for almost every $x \in [0,1]$.

We are left to prove that φ is measurable and measure preserving. As we saw that each point $z \in [0,1]$ as a pre-image $\psi^{-1}(z) = \{x \in [0,1] : \psi(x) = z\}$ at most countable (indeed of cardinal at most 2), thus [Pur66] insures that ψ is bimeasurable (i.e. ψ is (Borel) measurable and for all Borel set $B \subset [0,1]$, $\psi(B)$ is also a Borel set). Let $B \subset]0,1]$ be a Borel set. We have that $\varphi^{-1}(B) = \varphi^{-1}(B \cap D) = \psi(B \cap D)$ is a Borel set, where the first equality uses that $\varphi([0,1]) = D \cup \{0\}$, the second equality uses that ψ is the inverse of φ on D, and lastly we used that ψ is bimeasurable. We also have that $\varphi^{-1}(B \cup \{0\}) = \varphi^{-1}(B) \cup ([0,1] \setminus \psi(D))$ is a Borel set. Moreover, we have:

$$\lambda(\varphi^{-1}(B)) = \lambda(\psi(B \cap D)) = \lambda(\psi^{-1}(\psi(B \cap D))) = \lambda(B \cap D) = \lambda(B),$$

where we used that $\varphi^{-1}(B) = \psi(B \cap D)$ for the first equality, that ψ is measure preserving for the second equality, that ψ is bijective from D to $\psi(D)$ for the third equality, and that D has measure one for the last equality. We also have:

$$\lambda(\varphi^{-1}(B \cup \{0\})) = \lambda(\varphi^{-1}(B)) + \lambda([0,1] \setminus \psi(D)) = \lambda(B) = \lambda(B \cup \{0\}),$$

where we used that $\varphi^{-1}(B) \subset \psi(D)$ and $[0,1] \setminus \psi(D)$ are disjoint sets for the first equality, that $\lambda(\varphi^{-1}(B)) = \lambda(B)$ and that $\psi(D)$ has measure one for the second equality. Hence, the map φ is measurable and measure preserving.

We now prove that $(\varphi^{-1}(\mathcal{P}_k))_{k\in\mathbb{N}}$ is a refining sequence of partitions. For $k \in \mathbb{N}$, as $\mathcal{P}_k$ is a finite partition of $[0,1]$, we have that $\varphi^{-1}(\mathcal{P}_k) = \{\varphi^{-1}(S_u) : u \in \mathcal{T}, h(u) = k\}$ is also a finite partition of $[0,1]$. Moreover, as $(\mathcal{P}_k)_{k\in\mathbb{N}}$ is a refining sequence of partitions, we get that the sequence of partitions $(\varphi^{-1}(\mathcal{P}_k))_{k\in\mathbb{N}}$ is also refining. Remark that the sets $\varphi^{-1}(S_u)$ are not necessarily intervals, they are intervals minus some at most countable sets (this is similar to the unit line minus the Cantor set).

To finish the proof, we are left to construct a refining sequence of partitions made of intervals $(\mathcal{Q}_k)_{k\in\mathbb{N}}$ that agrees almost everywhere with the refining sequence of partitions $(\varphi^{-1}(\mathcal{P}_k))_{k\in\mathbb{N}}$. For $u \in \mathcal{T}$, define $R_u = [C^-(u), C^+(u)[$ (and $R_u = [C^-(u), C^+(u)]$ if u is the unique node such that $v \leq_{\text{lex}} u$ for every $v \in \mathcal{T}$ with $h(v) = h(u)$). As ψ is measure preserving, and as $\psi(S_u) \subset [C^-(u), C^+(u)]$ with $\lambda(S_u) = C^+(u) - C^-(u)$, we get that $\lambda([C^-(u), C^+(u)] \setminus \psi(S_u)) = 0$. As φ is the a.e. inverse map of ψ, we have that a.e. $\mathbb{1}_{\varphi^{-1}(S_u)} = \mathbb{1}_{\psi(S_u)} = \mathbb{1}_{[C^-(u),C^+(u)]} = \mathbb{1}_{R_u}$, i.e. R_u agrees almost everywhere with $\varphi^{-1}(S_u)$. For $k \in \mathbb{N}$, define the finite partition $\mathcal{Q}_k = \{R_u : h(u) = k\}$. Then, by definition of the sets R_u, the sequence of partitions $(\mathcal{Q}_k)_{k\in\mathbb{N}}$ is refining. This concludes the proof. □

Remark 3.8.3 (The shrinking diameter assumption is important)**.** Even if it is not stressed in [Lov12, Proof of Theorem 9.23], the measure preserving map φ (a fortiori an a.e. inversible one) in Lemma 3.8.2 cannot be obtained without any assumption on the refining sequence of partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ (in our case, we assumed that their diameter converges to zero). Indeed consider the sequence of partitions where for every $k \in \mathbb{N}$, $\mathcal{P}_k$ is composed of the sets:

$$S_{k,j} = [j2^{-k-1}, (j+1)2^{-k-1}[\cup [1/2 + j2^{-k-1}, 1/2 + (j+1)2^{-k-1}[, \qquad 0 \leq j < 2^k,$$

i.e. $S_{k,j}$ is the union of two dyadic interval translated by $1/2$, (also add 1 to the set $S_{k,0}$ to get a complete partition). Then, for every $x \in [0, 1/2[$, x and $x + 1/2$ belong to the same set of $\mathcal{P}_k$ for every $k \in \mathbb{N}$; in particular the diameter of the partitions $(\mathcal{P}_k)_{k\in\mathbb{N}}$ does not converge to zero. By contradiction, assume there exist a measure preserving map $\varphi \in \bar{S}_{[0,1]}$ and a sequence of interval partitions $(\mathcal{Q}_k)_{k\in\mathbb{N}}$ such that for all $k \in \mathbb{N}$ and all set $S_{k,j} \in \mathcal{P}_k$ with $0 \leq j < 2^k$, there exists a interval set $I_{k,j} \in \mathcal{Q}_k$ such that a.e. $\mathbb{1}_{I_{k,j}} = \mathbb{1}_{\varphi^{-1}(S_{k,j})}$. In particular, the set $I_{k,j}$ must be an interval of length 2^{-k}. Hence, $\mathcal{Q}_k$ is a dyadic partition with stepsize 2^{-k}, and thus the diameter of the partitions $(\mathcal{Q}_k)_{k\in\mathbb{N}}$ converges to zero. For every $x \in [0, 1/2[$, we get that $\operatorname{diam}(\varphi^{-1}(\{x, x + 1/2\})) \leq \operatorname{diam}(\mathcal{Q}_k) = 2^{-k}$ for all $k \in \mathbb{N}$; this implies that $\varphi^{-1}(\{x, x + 1/2\})$ is a singleton, i.e. either $x \notin \varphi([0,1])$ or $x + 1/2 \notin \varphi([0,1])$. Hence, we have $\lambda([0, 1/2[\cap\varphi([0,1])) = \lambda([1/2, 1[\setminus\varphi([0,1]))$ and $\lambda([0, 1/2[\setminus\varphi([0,1])) = \lambda([1/2, 1[\cap\varphi([0,1]))$. As $\lambda([0, 1/2[) = \lambda([0, 1/2[\cap\varphi([0,1])) + \lambda([0, 1/2[\setminus\varphi([0,1])) = 1/2$ because φ is measure preserving, we get that $\lambda(\varphi([0,1])) = \lambda([0, 1/2[\cap\varphi([0,1])) + \lambda([1/2, 1[\cap\varphi([0,1])) = 1/2$, which contradicts the fact that φ is measure preserving.

Now, for every $n \in \mathbb{N}$, applying Lemma 3.8.2 to $(\mathcal{P}_{n,k})_{k\in\mathbb{N}}$ and W_n, we get a measure-preserving map φ_n and a refining sequence of partitions $(\mathcal{P}'_{n,k})_{k\in\mathbb{N}}$ made of intervals such that for all $k \in \mathbb{N}$, and all set $R \in \mathcal{P}'_{n,k}$ there exists a set $S \in \mathcal{P}_{n,k}$ such that a.e. $\mathbb{1}_R = \mathbb{1}_{\varphi_n^{-1}(S)}$. In particular, for all $k \in \mathbb{N}$, the sequence of partitions $(\mathcal{P}_{n,k})_{k\in\mathbb{N}}$ still satisfy (i)–(ii). Set $W'_n = W_n^{\varphi_n}$ and $W'_{n,k} = W_{n,k}^{\varphi_n}$ so that almost everywhere:

$$W'_{n,k} = \big((W_n)_{\mathcal{P}_{n,k}}\big)^{\varphi_n} = (W_n^{\varphi_n})_{\mathcal{P}'_{n,k}} = (W')_{\mathcal{P}'_{n,k}}.$$

As d and d' are invariant, we have for every $n, k \in \mathbb{N}$ that $d(W_n, W_{n,k}) = d(W'_n, W'_{n,k})$, and similarly for d'. This insures that the signed measure-valued kernels $(W'_n)_{n\in\mathbb{N}}$ and $(W'_{n,k})_{n\in\mathbb{N}}$, $k \in \mathbb{N}$, still satisfy (iii)–(iv). Remind that for a measure-valued kernel W and a measure-preserving map φ, $\delta_{\square,\mathrm{m}}(W, W^\varphi) = 0$. Hence, we can replace the signed measure-valued kernels $(W_n)_{n\in\mathbb{N}}$ and $(W_{n,k})_{n\in\mathbb{N}}$, $k \in \mathbb{N}$, by $(W'_n)_{n\in\mathbb{N}}$ and $(W'_{n,k})_{n\in\mathbb{N}}$, $k \in \mathbb{N}$, and assume that the partitions $\mathcal{P}_{n,k}$ are made of intervals.

Step 2: There exists a subsequence $(W_{n_\ell})_{\ell\in\mathbb{N}}$ such that for every $k \in \mathbb{N}$ and $\epsilon \in \{+,-\}$, the subsequence $(W^\epsilon_{n_\ell,k})_{\ell\in\mathbb{N}}$ weakly converges, as $\ell \to \infty$, almost everywhere to a limit, say U^ϵ_k which is a stepfunction adapted to a partition with m_k elements (some elements might be empty sets).

Fix some $k \in \mathbb{N}$. The stepfunctions $(W_{n,k} = (W_n)_{\mathcal{P}_{n,k}})_{n\in\mathbb{N}}$ all have the same number of steps m_k. For $n \in \mathbb{N}$, denote by $\mathcal{P}_{n,k} = \{S_{n,k,i} : 1 \leq i \leq m_k\}$ the interval partition adapted to $W_{n,k}$ where the intervals are order according to the natural order on $[0,1]$ (note that some intervals might be empty, simply put them at the end). For $n \in \mathbb{N}$ and $1 \leq i \leq m_k$, let $\lambda(S_{n,k,i})$ denote the length of the interval $S_{n,k,i} \in \mathcal{P}_{n,k}$. As the lengths of steps take values in the compact set $[0,1]$, there exists a subsequence of indices $(n_\ell)_{\ell\in\mathbb{N}}$ such that for every $1 \leq i \leq m_k$, there exists $s_{k,i} \in [0,1]$ such that $\lim_{\ell\to\infty} \lambda(S_{n_\ell,k,i}) = s_{k,i}$. Denote by $\mathcal{P}_k = \{S_{k,i} : 1 \leq i \leq m_k\}$ the interval partition composed of m_k intervals where the i-th interval $S_{k,i}$ has length $s_{k,i}$ (note that some intervals might be empty). Up to a diagonal extraction, we can assume that the convergence holds for every $k \in \mathbb{N}$ simultaneously. Remark that for all $n, k \in \mathbb{N}$, the fact that $\mathcal{P}_{n,k+1}$ is a refinement of $\mathcal{P}_{n,k}$ can be simply restated as linear relations on the interval lengths $(\lambda(S_{n,k,i}))_{1\leq i\leq m_k}$ and $(\lambda(S_{n,k+1,i}))_{1\leq i\leq m_{k+1}}$. As linear relations are preserved when taking the limit, we get that the partition $\mathcal{P}_{k+1}$ is a refinement of $\mathcal{P}_k$ for all $k \in \mathbb{N}$. We assume from now on that $(W_n)_{n\in\mathbb{N}}$ and $(W_{n,k})_{n\in\mathbb{N}}$, $k \in \mathbb{N}$, are the corresponding subsequences.

For every $n \in \mathbb{N}$, we decompose $W_n = W_n^+ - W_n^-$ into its positive and negative kernel parts, see Lemma 3.3.3. For $n, k \in \mathbb{N}$ and $\epsilon \in \{+,-\}$, we define $W_{n,k}^\epsilon = (W_n^\epsilon)_{\mathcal{P}_{n,k}}$. In particular, remark that $W_{n,k} = W_{n,k}^+ - W_{n,k}^-$ and for all $\ell \geq k$, that $W_{n,k}^\epsilon = (W_{n,\ell}^\epsilon)_{\mathcal{P}_{n,k}}$. Let $\epsilon \in \{+,-\}$ and $1 \leq i, j \leq m_k$ such that $s_{k,i}s_{k,j} > 0$ be fixed. For every $n \in \mathbb{N}$, we have on $S_{n,k,i} \times S_{n,k,j}$ that $W_{n,k}^\varepsilon = \mu_{n,k}^{i,j,\epsilon} \in \mathcal{M}_+(\mathbf{Z})$ with:

$$\mu_{n,k}^{i,j,\epsilon}(\cdot) = \frac{1}{\lambda(S_{n,k,i})\lambda(S_{n,k,j})} W_n^\epsilon(S_{n,k,i} \times S_{n,k,j}; \cdot).$$

We have that:

$$\|\mu_{n,k}^{i,j,\epsilon}\|_\infty \leq \|W_n\|_\infty \leq C.$$

This gives that the sequence $(\mu_{n,k}^{i,j,\epsilon})_{n\in\mathbb{N}}$ in $\mathcal{M}_\pm(\mathbf{Z})$ is bounded. We now prove it is tight. Let $\varepsilon > 0$. As $\lim_{n\to\infty} \lambda(S_{n,k,\ell}) = s_{k,\ell} > 0$ for $\ell = i, j$, we deduce that there exists $c > 0$ such that for every $n \in \mathbb{N}$ large enough and $\ell = i, j$, we have $\lambda(S_{n,k,\ell}) > c$. Set $\varepsilon' = c^2\varepsilon$. As the sequence $(W_n)_{n\in\mathbb{N}}$ in $\widetilde{\mathcal{W}}_\pm$ is tight, there exists a compact set $K \subset \mathbf{Z}$ such that for every $n \in \mathbb{N}$, $M_{W_n}(K^c) \leq \varepsilon'$. Hence, for every $n \in \mathbb{N}$ large enough, we have:

$$\mu_{n,k}^{i,j,\epsilon}(K^c) \leq \frac{1}{\lambda(S_{n,k,i})\lambda(S_{n,k,j})} M_{W_n}(K^c) \leq \varepsilon.$$

This gives that the sequence $(\mu_{n,k}^{i,j,\epsilon})_{n\in\mathbb{N}}$ in $\mathcal{M}_+(\mathbf{Z})$ is bounded and tight, and thus by Lemma 3.2.8, it has a convergent subsequence. By diagonal extraction, we can assume there is a subsequence $(W_{n_\ell})_{\ell\in\mathbb{N}}$ such that for all $k \in \mathbb{N}$, all $1 \leq i, j \leq m_k$ such that $s_{k,i}s_{k,j} > 0$, and all $\epsilon \in \{+,-\}$, the subsequence $(\mu_{n_\ell,k}^{i,j,\epsilon})_{\ell\in\mathbb{N}}$ weakly converges to a limit say $\mu_k^{i,j,\epsilon}$. Define the stepfunction $U_k^\epsilon \in \mathcal{W}_+$ adapted to the partition $\mathcal{P}_k$ which is equal to $\mu_k^{i,j,\epsilon}$ on $S_{k,i} \times S_{k,j}$ (if $s_{k,i}s_{k,j} = 0$, set $\mu_k^{i,j,\epsilon} = 0$). We have in particular obtained that, for every $k \in \mathbb{N}$, the subsequence $(W_{n_\ell,k}^\epsilon)_{\ell\in\mathbb{N}}$ weakly converges a.e. to U_k^ϵ which is a stepfunction adapted to a partition with m_k elements; this implies that the subsequence $(W_{n_\ell,k})_{\ell\in\mathbb{N}}$ also weakly converges a.e. to $U_k = U_k^+ - U_k^-$. We now assume that $(W_n)_{n\in\mathbb{N}}$ is such a subsequence. With this convention, notice that for all $k, n \in \mathbb{N}$ and $\epsilon \in \{+,-\}$:

$$\|U_k^\epsilon\|_\infty \leq \sup_{n\in\mathbb{N}} \|W_{n,k}^\epsilon\|_\infty \leq \sup_{n\in\mathbb{N}} \|W_n\|_\infty = C < +\infty. \tag{3.28}$$

Step 3: There exists a subsequence of $(U_k)_{k\in\mathbb{N}}$ which weakly converges to a limit $U \in \mathcal{W}_\pm$ almost everywhere on $[0,1]^2$. The proof of this step is postponed to the end. Without loss of generality we still write $(U_k)_{k\in\mathbb{N}}$ for this subsequence.

Step 4: We have $\lim_{n\to\infty} \delta_\square(U, W_n) = \lim_{n\to\infty} \delta'_\square(U, W_n) = 0$. Let $\varepsilon > 0$. As the cut distances d is smooth, we deduce from Step 3, that for k large enough $d(U, U_k) \leq \varepsilon$. By hypothesis (iii) on the sequence $(W_{n,k})_{n\in\mathbb{N}}$, we also have that for k large enough $d(W_n, W_{n,k}) \leq \varepsilon$. For such large k, as by step 2 the sequence $(W_{n,k})_{n\in\mathbb{N}}$ weakly converges almost everywhere to U_k, and again as the cut distances d is smooth, there is a n_0 such that for every $n \geq n_0$, $d(U_k, W_{n,k}) \leq \varepsilon$. Then for all $n \geq n_0$, we have:

$$\begin{aligned}
\delta_\square(U, W_n) &\leq \delta_\square(U, U_k) + \delta_\square(U_k, W_{n,k}) + \delta_\square(W_{n,k}, W_n) \\
&\leq d(U, U_k) + d(U_k, W_{n,k}) + d(W_{n,k}, W_n) \\
&\leq 3\varepsilon.
\end{aligned}$$

This gives that $\lim_{n\to\infty} \delta_\square(W_n, U) = 0$.

If we consider a second distance d' as in the lemma, then similarly $\lim_{n\to\infty} \delta'_\square(W_n, U) = 0$.

Proof of Step 3. Assume that the claim is true for measure-valued kernels. Then, if $(U_k)_{k\in\mathbb{N}}$ is a sequence of signed-measure valued kernels, applying the claim to $(U_k^\epsilon)_{k\in\mathbb{N}}$, for $\epsilon \in \{+,-\}$, we get a measure-valued $U^\epsilon \in \mathcal{W}_+$ such that the sequence $(U_k^\epsilon)_{k\in\mathbb{N}}$ weakly conveges a.e. to U^ϵ. Thus, the sequence $(U_k)_{k\in\mathbb{N}}$ weakly conveges a.e. to $U = U^+ - U^-$.

Hence, we are left to prove the claim for measure-valued kernels. The proof is divided in four steps. The first three steps also work for signed-measure valued kernels, but the last argument of step 3.d. only works for measures.

Step 3.a: The sequence $(U_k)_{k\in\mathbb{N}}$ inherit the tightness property from the sequence $(W_n)_{n\in\mathbb{N}}$. Let $\varepsilon > 0$. Since the sequence $(W_n)_{n\in\mathbb{N}}$ is tight, there exists a compact set $K \subset \mathbf{Z}$ such that for every $n \in \mathbb{N}$, we have $M_{W_n}(K^c) \le \varepsilon$. Remark that:

$$M_{W_{n,k}} = \sum_{1\le i,j\le m_k} \lambda(S_{n,k,i})\lambda(S_{n,k,j})\mu_{n,k}^{i,j} = M_{W_n} \quad \text{and} \quad M_{U_k} = \sum_{1\le i,j\le m_k} s_{k,i}s_{k,j}\mu_k^{i,j}.$$

For all $k \in \mathbb{N}$ and $1 \le i, j \le m_k$, as the sequence $(\mu_{n,k}^{i,j})_{n\in\mathbb{N}}$ weakly converges to $\mu_k^{i,j}$, using [Bog18, Theorem 2.7.4] with the open subset $K^c \subset \mathbf{Z}$, we get that $\mu_k^{i,j}(K^c) \le \liminf_{n\to\infty} \mu_{n,k}^{i,j}(K^c)$. As $\lim_{n\to\infty} \lambda(S_{n,k,i}) = s_{k,i}$ for all $1 \le i \le m_k$, and summing those bounds, we get:

$$M_{U_k}(K^c) \le \liminf_{n\to\infty} M_{W_{n,k}}(K^c) = \liminf_{n\to\infty} M_{W_n}(K^c) \le \varepsilon.$$

Consequently, the sequence $(U_k)_{k\in\mathbb{N}}$ is tight.

Step 3.b: Convergence of the measures $\hat{U}_k$ in $\mathcal{M}_+([0,1]^2 \times \mathbf{Z})$ defined for $k \in \mathbb{N}$ as:

$$\hat{U}_k(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = U_k(x, y; \mathrm{d}z)\, \lambda_2(\mathrm{d}x, \mathrm{d}y). \tag{3.29}$$

Since the sequence $(M_{U_k})_{k\in\mathbb{N}}$ in $\mathcal{M}_+(\mathbf{Z})$ is tight, for all $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{Z}$ such that for every $k \in \mathbb{N}$, $M_{U_k}(K^c) \le \varepsilon$; and thus $\hat{U}_k(\hat{K}^c) = M_{U_k}(K^c) \le \varepsilon$ where $\hat{K} = [0,1]^2 \times K$ is a compact subset of $[0,1]^2 \times \mathbf{Z}$, that is, the sequence $(\hat{U}_k)_{k\in\mathbb{N}}$ in $\mathcal{M}_+([0,1]^2 \times \mathbf{Z})$ is tight. The sequence $(\hat{U}_k)_{k\in\mathbb{N}}$ is also bounded as $\|\hat{U}_k\|_\infty \le \|U_k\|_\infty \le C$ thanks to (3.28). Hence, using Lemma 3.2.8, there exists a subsequence $(\hat{U}_{k_\ell})_{\ell\in\mathbb{N}}$ of the sequence $(\hat{U}_k)_{k\in\mathbb{N}}$ that converges to some measure, say $\hat{U}$, in $\mathcal{M}_+([0,1]^2 \times \mathbf{Z})$. Remark that, when considering the subsequence of indices $(k_\ell)_{\ell\in\mathbb{N}}$, the subsequences $(W_{n,k_\ell})_{\ell\in\mathbb{N}}$, $n \in \mathbb{N}$, still satisfy properties (i)-(iv) of Lemma 3.8.1, and for all $\ell \in \mathbb{N}$, the sequence $(W_{n,k_\ell})_{n\in\mathbb{N}}$ still weakly converges to U_{k_ℓ}. Without loss of generality, we now work with this subsequence and thus write k instead of k_ℓ.

Step 3.c: The measure $\hat{U}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ can be disintegrated w.r.t. $\lambda_2(\mathrm{d}x, \mathrm{d}y)$ giving us an element of $\mathcal{W}_+$. To prove this, we need the following disintegration theorem for measures, see [Kal17, Theorem 1.23] (stated in more the general framework of Borel spaces) which generalizes the disintegration theorem for probability measures [Kal02, Theorem 6.3]. The notation $\mu \sim \nu$ for two measures μ and ν means that $\mu \ll \nu$ and $\nu \ll \mu$, where $\mu \ll \nu$ means that μ is absolutely continuous w.r.t. ν.

Lemma 3.8.4 (Disintegration theorem for measures, [Kal17, Theorem 1.23])**.** *Let ρ be a measure on $S \times T$, where S is a measurable space and T a Polish space. Then there exist a measure $\nu \equiv \rho(\cdot \times T)$ on S and a probability kernel $\mu : S \to \mathcal{M}_1(T)$ such that $\rho = \nu \otimes \mu$ (i.e. $\rho(ds, dt) = \nu(ds)\mu(s; dt)$). Moreover, the measures $\mu_s = \mu(s; \cdot)$ are unique for ν-a.e. $s \in S$.*

Using Lemma 3.8.4 with $S = [0,1]^2$ and $T = \mathbf{Z}$, we get that there exists a probability kernel U' in $\mathcal{W}_1$ such that:

$$\hat{U}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = U'(x, y; \mathrm{d}z)\, \pi(\mathrm{d}x, \mathrm{d}y),$$

where $\pi = \hat{U}(\cdot \times \mathbf{Z})$ is a measure on $[0,1]^2$.

We now need to prove that $\pi \ll \lambda_2$. By contradiction, assume this is false, then there exists a measurable set $A \in \mathcal{B}([0,1]^2)$ such that $\lambda_2(A) = 0$ and $\pi(A) > 0$. As the measure $\int_A U'(x, y; \cdot)\, \pi(\mathrm{d}x, \mathrm{d}y)$ is not null, there exists $f \in C_b(\mathbf{Z})$ such that $\int_A U'(x, y; f)\, \pi(\mathrm{d}x, \mathrm{d}y) \neq 0$. As the sequence $(\hat{U}_k)_{k\in\mathbb{N}}$ weakly converges to $\hat{U}$ in $\mathcal{M}_+([0,1]^2 \times \mathbf{Z})$ by step 3.b, we have that the sequence of measures $\hat{U}_k(\mathrm{d}x, \mathrm{d}y; f) =$

$U_k(x,y;f)\,\lambda_2(\mathrm{d}x,\mathrm{d}y)$ weakly converges as $k\to\infty$ to $\hat{U}(\mathrm{d}x,\mathrm{d}y;f)=U'(x,y;f)\,\pi(\mathrm{d}x,\mathrm{d}y)$ in $\mathcal{M}_+([0,1]^2)$. Moreover, as the maps $x,y\mapsto U_k(x,y;f)$ are uniformly bounded (by $\|f\|_\infty\|U_k\|_\infty\le C\|f\|_\infty$, see (3.28)), they are also uniformly integrable (w.r.t. λ_2), and applying [Bog07a, Corollary 4.7.19] there exist a subsequence $(U_{k_\ell})_{\ell\in\mathbb{N}}$ and a bounded function g_f on $[0,1]^2$ such that for every bounded measurable function $h\in L^\infty([0,1]^2)$, we have:

$$\lim_{\ell\to\infty}\int U_{k_\ell}(x,y;f)h(x,y)\,\lambda_2(\mathrm{d}x,\mathrm{d}y)=\int g_f(x,y)h(x,y)\,\lambda_2(\mathrm{d}x,\mathrm{d}y).$$

In particular, the sequence of measures $(U_{k_\ell}(x,y;f)\,\lambda_2(\mathrm{d}x,\mathrm{d}y))_{\ell\in\mathbb{N}}$ weakly converges to the measure $g_f(x,y)\,\lambda_2(\mathrm{d}x,\mathrm{d}y)$, which imposes the equality between measures:

$$\hat{U}(\mathrm{d}x,\mathrm{d}y,f)=U'(x,y;f)\pi(\mathrm{d}x,\mathrm{d}y)=g_f(x,y)\,\lambda_2(\mathrm{d}x,\mathrm{d}y).$$

Hence, taking $h=\mathbb{1}_A$, we get:

$$\hat{U}(A,f)=\int_A U'(x,y;f)\pi(\mathrm{d}x,\mathrm{d}y)=\int_A g_f(x,y)\lambda_2(\mathrm{d}x,\mathrm{d}y)=0,$$

which yields a contradiction. Consequently, the measure π is absolutely continuous w.r.t. λ_2, with density still denoted by π, and we set λ_2-a.e. on $[0,1]^2$:

$$U(x,y;\mathrm{d}z)=\pi(x,y)\,U'(x,y;\mathrm{d}z)\quad\text{and thus}\quad\hat{U}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z)=U(x,y;\mathrm{d}z)\,\lambda_2(\mathrm{d}x,\mathrm{d}y).\tag{3.30}$$

Step 3.d: The sequence $(U_k)_{k\in\mathbb{N}}$ weakly converges to U almost everywhere on $[0,1]^2$. Recall that by construction, the stepfunction U_k is adapted to the partition $\mathcal{P}_k$ defined in Step 2, and that $\mathcal{P}_{k+1}$ is a refinement of $\mathcal{P}_k$. A closer look at Step 2 yields that for all $\ell\ge k$, since $W_{n,k}=(W_{n,\ell})_{\mathcal{P}_{n,k}}$, we also get:

$$U_k=(U_\ell)_{\mathcal{P}_k}.\tag{3.31}$$

We prove (3.31) for $\ell=k+1$, the other cases follow by induction. As $\mathcal{P}_{n,k+1}$ is a refinement of $\mathcal{P}_{n,k}$, we already know that U_k and $(U_{k+1})_{\mathcal{P}_k}$ are both stepfunctions adapted to the finite partition $\mathcal{P}_k$. Thus, we only need to verify that U_k and $(U_{k+1})_{\mathcal{P}_k}$ take the same value on each step. For every $n\in\mathbb{N}$, the fact that $W_{n,k}=(W_{n,k+1})_{\mathcal{P}_{n,k}}$ implies that for all $1\le i,j\le m_k$ such that $\lambda(S_{n,k,i})\lambda(S_{n,k,j})>0$, we have:

$$\mu_{i,j}^{n,k}=\sum_{i'\in I_i,j'\in I_j}\frac{\lambda(S_{n,k+1,i'})\lambda(S_{n,k+1,j'})}{\lambda(S_{n,k,i})\lambda(S_{n,k,j})}\mu_{i',j'}^{n,k+1},$$

and this equation is preserved when taking the limit $n\to\infty$, which gives us:

$$\mu_{i,j}^{k}=\sum_{i'\in I_i,j'\in I_j}\frac{s_{k+1,i'}s_{k+1,j'}}{s_{k,i}s_{k,j}}\mu_{i',j'}^{k+1},\qquad\text{for all } 1\le i,j\le m_k \text{ such that } s_{k,i}s_{k,j}>0.$$

This proves that the stepfunctions U_k and $(U_{k+1})_{\mathcal{P}_k}$ take the same value on each step $S_{k,i}\times S_{k,j}$ with positive size $s_{k,i}s_{k,j}>0$ (on a step with null size $s_{k,i}s_{k,j}=0$, U_k and $(U_{k+1})_{\mathcal{P}_k}$ are both equal to the null measure). This gives that $U_k=(U_{k+1})_{\mathcal{P}_k}$.

Let $f\in C_b(\mathbf{Z})$ be a bounded continuous function, and X,Y be independent uniform random variables on $[0,1]$. Then (3.31) and (3.28) imply that the sequence $N^f=(N_k^f=U_k(X,Y;f))_{k\in\mathbb{N}}$ is a martingale bounded by $C\|f\|_\infty$ for the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$, where the σ-field $\mathcal{F}_k$ is generated by the events $\{X\in S_{k,i}\}\cap\{Y\in S_{k,j}\}$ for $1\le i,j\le m_k$ and $S_{k,\ell}\in\mathcal{P}_k$. By the martingale convergence theorem, the martingale N^f is almost surely convergent, that is, the sequence $(U_k[f])_{k\in\mathbb{N}}$ converges λ_2-a.e. to a bounded measurable function u_f. Let $g:[0,1]^2\to\mathbb{R}$ be a bounded measurable function. We get:

$$\begin{aligned}\int_{[0,1]^2}g(x,y)\,U(x,y;f)\,\lambda_2(\mathrm{d}x\mathrm{d}y)&=\int_{[0,1]^2\times\mathbf{Z}}g(x,y)\,f(z)\,\hat{U}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z)\\&=\lim_{k\to\infty}\int_{[0,1]^2\times\mathbf{Z}}g(x,y)\,f(z)\,\hat{U}_k(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z)\\&=\lim_{k\to\infty}\mathbb{E}\left[g(X,Y)U_k(X,Y;f)\right]\\&=\int_{[0,1]^2}g(x,y)\,u_f(x,y)\,\lambda_2(\mathrm{d}x\mathrm{d}y),\end{aligned}$$

where we used the definition (3.30) of U for the first equality, that $(\hat{U}_k)_{k\in\mathbb{N}}$ weakly converges to $\hat{U}$ for the second, the definition (3.29) of $\hat{U}_k$ for the third, and the convergence of the martingale N^f for the last. Since g is arbitrary, we deduce that λ_2-a.e. $U(\cdot,\cdot;f) = u_f$ and thus that the sequence $(U_k[f])_{k\in\mathbb{N}}$ converges λ_2-a.e. to $U[f]$. Applying this result for all $f \in \mathcal{F} = (f_m)_{m\in\mathbb{N}}$ a convergence determining sequence (with the convention $f_0 = \mathbb{1}$), we deduce that the sequence $(U_k)_{k\in\mathbb{N}}$ weakly converges to U almost everywhere on $[0,1]^2$. Remind from Section 3.2 that convergence determining sequences exist only for measures and not for signed measures in general, this is why we worked with measures in Step 3. This ends the proof of Step 3, and thus ends the proof of the lemma. □

We are now ready to prove Theorem 3.5.1.

Proof of Theorem 3.5.1. We first prove Point (i) on $\mathcal{W}_\pm$ (the proof on $\mathcal{W}_1$ is similar). Since the distance d is weakly regular and the sequence $(W_n)_{n\in\mathbb{N}}$ is uniformly bounded and tight in $\mathcal{W}_\pm$, we can construct inductively for every $n \in \mathbb{N}$ a sequence $(\mathcal{P}_{n,k})_{k\in\mathbb{N}}$ of partitions of $[0,1]$ such that hypothesis (i)-(iii) of Lemma 3.8.1 are satisfied: $\mathcal{P}_{n,k+1}$ being obtained by applying the weak regularity property (see Definition 3.4.10-(i)) with starting partition $\mathcal{Q}_{n,k} = \mathcal{P}_{n,k} \wedge \mathcal{D}_k$, where $\mathcal{D}_k$ is the dyadic partition with stepsize $2^{-(k+1)}$. (We may assume that the partitions $P_{n,k}$ for all $n \in \mathbb{N}$ have the same size m_k by adding empty sets.) Then as d is also invariant and smooth on $\mathcal{W}_\pm$, the first part of Lemma 3.8.1 directly gives Point (i).

Before proving Point (ii), we first need to prove the following lemma.

Lemma 3.8.5 (Compactness theorem for $\mathcal{W}_\mathcal{M}$)**.** *Let d be an invariant, smooth and weakly regular distance on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$). Let $\mathcal{M}$ be a convex and weakly closed subset of $\mathcal{M}_1(\mathbf{Z})$ (resp. $\mathcal{M}_+(\mathbf{Z})$ or $\mathcal{M}_\pm(\mathbf{Z})$). Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{M}$-valued kernels which is tight and uniformly bounded. Then, $(W_n)_{n\in\mathbb{N}}$ has a subsequence that converges for $\delta_\square$ to some $\mathcal{M}$-valued kernel.*

Proof. First remark that, as $\mathcal{M}$ is convex, the image of $\mathcal{W}_\mathcal{M}$ by the stepping operator $W \mapsto W_\mathcal{P}$, where $\mathcal{P}$ is a finite partition of $[0,1]$, is a subset of $\mathcal{W}_\mathcal{M}$. Hence, a close look at the proof of Lemma 3.8.1 (the partitions are constructed as in the proof of Point (i) from Theorem 3.5.1), and using the notation therein, shows that, up to taking subsequences, one can take the stepping kernels $W_{n,k}$ and U_k in $\mathcal{W}_\mathcal{M}$, such that $(U_k)_{k\in\mathbb{N}}$ weakly converges to U a.e. and the subsequence $(W_{n_\ell})_{\ell\in\mathbb{N}}$ converges to U w.r.t. $\delta_\square$. Since $U_k(x,y;\cdot) \in \mathcal{M}$ weakly converges to $U(x,y;\cdot)$ for almost every $x,y \in [0,1]$ and since $\mathcal{M}$ is weakly closed (and thus sequentially weakly closed), we deduce that $U(x,y;\cdot)$ belongs to $\mathcal{M}$ for almost every $x,y \in [0,1]$. This means that $U \in \mathcal{W}_\mathcal{M}$. □

We prove Point (ii) for $\mathcal{M} \subset \mathcal{M}_\pm(\mathbf{Z})$ (the proof for $\mathcal{M} \subset \mathcal{M}_1(\mathbf{Z})$ is identical). The fact that $\mathcal{W}_\mathcal{M}$ and $\widetilde{\mathcal{W}}_\mathcal{M}$ are convex is clear as $\mathcal{M}$ is convex. Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\widetilde{\mathcal{W}}_\mathcal{M}$. Since $\mathcal{M}$ is convex, we deduce that $(M_{W_n})_{n\in\mathbb{N}}$ is a sequence in $\mathcal{M}$. As $\mathcal{M}$ is sequentially compact for the weak topology, $\mathcal{M}$ is tight and bounded by Lemma 3.2.8, and thus the sequence $(W_n)_{n\in\mathbb{N}}$ is tight and uniformly bounded (remind Definition 3.4.7). Hence, using Lemma 3.8.5, we get that from any sequence in $\widetilde{\mathcal{W}}_\mathcal{M}$, we can extract a subsequence which converges for $\delta_\square$ to an element in $\widetilde{\mathcal{W}}_\mathcal{M}$. This implies that $(\widetilde{\mathcal{W}}_\mathcal{M}, \delta_\square)$ is compact.

Point (iii) is a direct consequence of Point (ii) as if $\mathbf{Z}$ is compact, so is $\mathcal{M}_1(\mathbf{Z})$. □

Proof of Point (iii) from Proposition 3.5.2. We prove Point (iii). The fact that $\mathcal{W}_\mathcal{M}$ and $\widetilde{\mathcal{W}}_\mathcal{M}$ are convex is clear as $\mathcal{M}$ is convex. To prove that $\widetilde{\mathcal{W}}_\mathcal{M}$ is closed, we consider a sequence $(W_n)_{n\in\mathbb{N}}$ in $\widetilde{\mathcal{W}}_\mathcal{M}$ that converges for $\delta_{\square,\mathrm{m}}$ to some $W \in \widetilde{\mathcal{W}}_\pm$. As $(W_n)_{n\in\mathbb{N}}$ is a Cauchy sequence for $\delta_{\square,\mathrm{m}}$, by Lemma 3.4.9, $(M_{W_n})_{n\in\mathbb{N}}$ is a Cauchy sequence for d_m and thus is tight. Hence, $(W_n)_{n\in\mathbb{N}}$ is uniformly bounded and tight. Applying Lemma 3.8.5, there exists a subsequence $(W_{n_k})_{k\in\mathbb{N}}$ of the sequence $(W_n)_{n\in\mathbb{N}}$ which converges for $\delta_{\square,\mathrm{m}}$ to some $\mathcal{M}$-valued kernel $U \in \widetilde{\mathcal{W}}_\mathcal{M}$. But as a subsequence, $(W_{n_k})_{k\in\mathbb{N}}$ must also converge for $\delta_{\square,\mathrm{m}}$ to W. This implies that $W = U$ is a $\mathcal{M}$-valued kernel. □

In order to prove Theorem 3.5.5, we first prove a lemma that allows to construct the partitions needed to use Lemma 3.8.1.

Lemma 3.8.6 (Construction of partitions for two distances)**.** *Let d and d' be two distances on $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) which are invariant, smooth, weakly regular and regular w.r.t. the stepping operator (see Definitions 3.3.10 and 3.4.10). Let $(W_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{W}_1$ (resp. $\mathcal{W}_+$ or $\mathcal{W}_\pm$) which is tight (resp. uniformly bounded and tight). Then, there exists sequences $(\mathcal{P}_{n,k})_{k\in\mathbb{N}}$, $n\in\mathbb{N}$, of partitions of $[0,1]$ such that hypothesis (i)–(iv) of Lemma 3.8.1 are satisfied.*

Proof. We prove the result on $\mathcal{W}_\pm$ (the proof on $\mathcal{W}_1$ and $\mathcal{W}_+$ is similar). To simplify notations, write $d^1 = d$ and $d^2 = d'$. We proceed by induction on $k \in \mathbb{N}\cup\{-1\}$. For every $n\in\mathbb{N}$, set $\mathcal{P}_{n,-1} = \{[0,1]\}$ the trivial partition with size 1. Let $k\in\mathbb{N}$ and assume that we have already constructed partitions $(\mathcal{P}_{n,k-1})_{n\in\mathbb{N}}$ that have the same size m_{k-1}. Now we proceed to construct partitions $(\mathcal{P}_{n,k})_{n\in\mathbb{N}}$ that satisfy hypothesis (i)-(iv).

Set $C = \sup_{n\in\mathbb{N}} \|W_n\|_\infty$, which is finite as the sequence $(W_n)_{n\in\mathbb{N}}$ is uniformly bounded. As d^i, with $i=1,2$, are regular w.r.t. the stepping operator, there exists a finite constant $C_0>0$ such that for every $W,U\in\mathcal{W}_\pm$, with $\|W\|_\infty \le C$ and $\|U\|_\infty \le C$, and U a stepfunction adapted to a finite partition $\mathcal{Q}$:

$$d^i(W, W_{\mathcal{Q}}) \le C_0\, d^i(W,U). \tag{3.32}$$

Set $\varepsilon = 1/C_0(k+1)$. Since d^i, with $i=1,2$, are weakly regular and the sequence $(W_n)_{n\in\mathbb{N}}$ is tight and uniformly bounded, there exists $r_k\in\mathbb{N}^*$, such that for every $n\in\mathbb{N}$, there exists a partition $\mathcal{R}^i_{n,k}$ of $[0,1]$ that refines $\mathcal{Q}_{n,k} = \mathcal{P}_{n,k-1}\wedge\mathcal{D}_k$, where $\mathcal{D}_k$ is the dyadic partition with stepsize 2^{-k}, such that:

$$|\mathcal{R}^i_{n,k}| \le r_k|\mathcal{Q}_{n,k}| \le 2^k r_k|\mathcal{P}_{n,k-1}| \quad\text{and}\quad d^i\Big(W_n, (W_n)_{\mathcal{R}^i_{n,k}}\Big) \le \varepsilon. \tag{3.33}$$

(Indeed, a close look at the proof shows that $\mathcal{P}_{n,k-1}$ refines $\mathcal{D}_{k-1}$ by construction, thus $\mathcal{Q}_{n,k}$ cuts each set of $\mathcal{P}_{n,k-1}$ in at most 2 sets, and we get $|\mathcal{Q}_{n,k}|\le 2|\mathcal{P}_{n,k-1}|$.) Now, let $\mathcal{P}_{n,k}$ be the common refinement of $\mathcal{R}^1_{n,k}$ and $\mathcal{R}^2_{n,k}$; it is a refinement of $\mathcal{P}_{n,k-1}$, has diameter at most 2^{-k} and size:

$$|\mathcal{P}_{n,k}| \le 2^{2k} r_k^2 |\mathcal{P}_{n,k-1}|^2 = 2^{2k} r_k^2 m_{k-1}^2.$$

If necessary, by completing $\mathcal{P}_{n,k}$ with null sets, we may assume that $|\mathcal{P}_{n,k}| = m_k$, where $m_k = 2^{2k} r_k^2 m_{k-1}^2$. As $(W_n)_{\mathcal{R}^i_{n,k}}$ is a stepfunction adapted to the partition $\mathcal{P}_{n,k}$, we deduce from (3.32) and (3.33) that for $i=1,2$ and $n\in\mathbb{N}$:

$$d^i(W_n, (W_n)_{\mathcal{P}_{n,k}}) \le C_0\, d^i\Big(W_n, (W_n)_{\mathcal{R}^i_{n,k}}\Big) \le C_0\,\varepsilon = \frac{1}{k+1}\cdot$$

Hence, for every $n\in\mathbb{N}$, the partition $\mathcal{P}_{n,k}$ satisfies the hypothesis (i)-(iv) of Lemma 3.8.1. Thus, the induction is complete. □

Proof of Theorem 3.5.5. Let d_{m} and $d_{\mathrm{m}'}$ be as in Theorem 3.5.5.

Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of probability-graphons that converges to some $W\in\widetilde{\mathcal{W}}_1$ for $\delta_{\square,\mathrm{m}}$. By Lemma 3.4.9, the sequence of probability measure $(M_{W_n})_{n\in\mathbb{N}}$ converges to M_W for the distance d_{m}. As d_{m} induces the weak topology on $\mathcal{M}_{\le 1}(\mathbf{Z})$, we have that the sequence $(M_{W_n})_{n\in\mathbb{N}}$ is tight, and thus the sequence $(W_n)_{n\in\mathbb{N}}$ is also tight (remind Definition 3.4.7). The sequence $(W_n)_{n\in\mathbb{N}}$ is also uniformly bounded as a sequence in $\widetilde{\mathcal{W}}_1$. Applying Lemma 3.8.6 with the distances $d = d_{\square,\mathrm{m}}$ and $d' = d_{\square,\mathrm{m}'}$, which are invariant, smooth, weakly regular and regular w.r.t. the stepping operator, we get sequences of partitions $(\mathcal{P}_{n,k})_{k\in\mathbb{N}}$, $n\in\mathbb{N}$, that satisfy hypothesis (i)-(iv) of Lemma 3.8.1. We then deduce from the last part of Lemma 3.8.1 that any subsequence of $(W_n)_{n\in\mathbb{N}}$ has a further subsequence which converges to the same limit for both $\delta_{\square,\mathrm{m}}$ and $\delta_{\square,\mathrm{m}'}$, this limit must then be W. This implies that the sequence $(W_n)_{n\in\mathbb{N}}$ converges to W for $\delta_{\square,\mathrm{m}'}$.

The role of d_{m} and $d_{\mathrm{m}'}$ being symmetric, we conclude that the distances $\delta_{\square,\mathrm{m}}$ and $\delta_{\square,\mathrm{m}'}$ induce the same topology on $\widetilde{\mathcal{W}}_1$. □

Acknowledgement

We thank Pierre-André Zitt for some helpful discussion in particular on Lemma 3.3.20.

Index of notations

Measures

- $(\mathbf{Z}, \mathcal{O}_{\mathbf{Z}})$ a topological Polish space
- $\mathcal{B}(\mathbf{Z})$ the Borel σ-field induced by $\mathcal{O}_{\mathbf{Z}}$
- $C_b(\mathbf{Z})$ the set of continuous bounded real-valued functions on $\mathbf{Z}$
- *measure* = positive measure
- $\mathcal{M}_{\pm}(\mathbf{Z})$ the set of signed measures on $\mathbf{Z}$
- $\mathcal{M}_{+}(\mathbf{Z})$ the set of measures on $\mathbf{Z}$
- $\mathcal{M}_1(\mathbf{Z})$ the set of probability measures on $\mathbf{Z}$
- $\mathcal{M}_{\leq 1}(\mathbf{Z})$ the set of sub-probability measures on $\mathbf{Z}$, i.e. measures with total mass at most 1
- μ^+, μ^- the positive and negative parts of μ from its Hahn-Jordan decomposition
- $|\mu| = \mu_+ + \mu_-$ the total variation measure of μ
- $\|\mu\|_\infty = |\mu|(\mathbf{Z})$ the total mass of μ
- d_{m} a distance on either $\mathcal{M}_{\leq 1}(\mathbf{Z})$, $\mathcal{M}_{+}(\mathbf{Z})$ or $\mathcal{M}_{\pm}(\mathbf{Z})$
- N_{m} a norm on $\mathcal{M}_{\pm}(\mathbf{Z})$
- $d_{\mathcal{P}}$ the Prohorov distance
- $\|\cdot\|_{\mathrm{KR}}$ the Kantorovitch-Rubinstein norm
- $\|\cdot\|_{\mathrm{FM}}$ the Fortet-Mourier norm
- $\|\cdot\|_{\mathcal{F}}$ the norm based on a convergence determining sequence $\mathcal{F}$

Relabelings and partitions

- $S_{[0,1]}$ the set of bijective measure-preserving maps from $([0,1], \lambda)$ to itself
- $\bar{S}_{[0,1]}$ the set of measure-preserving maps from $([0,1], \lambda)$ to itself
- $|\mathcal{P}|$ the number of sets in the finite partition $\mathcal{P}$

Kernels and graphons spaces

- $\mathcal{W}_1$ the set of probability-graphons
- $\mathcal{W}_+$ the set of measure-valued kernels
- $\mathcal{W}_{\pm}$ the set of signed measure-valued kernels
- $\mathcal{W}_{\mathcal{M}}$ the set of $\mathcal{M}$-valued kernels with $\mathcal{M} \subset \mathcal{M}_{\pm}(\mathbf{Z})$
- $\widetilde{\mathcal{W}}_1$ the set of unlabeled probability-graphons
- $\widetilde{\mathcal{W}}_+$ the set of unlabeled measure-valued kernels
- $\widetilde{\mathcal{W}}_{\pm}$ the set of unlabeled signed measure-valued kernels
- $\widetilde{\mathcal{W}}_{\mathcal{M}}$ the set of unlabeled $\mathcal{M}$-valued kernels

Kernels and graphons

- W^+ and W^- the positive and negative part of $W \in \mathcal{W}_{\pm}$
- $|W| = W^+ + W^-$
- $W(A;\cdot) = \int_A W(x,y;\cdot)\,\mathrm{d}x\mathrm{d}y$ for $A \subset [0,1]^2$
- $W[f](x,y) = W(x,y;f)$ for $f \in C_b(\mathbf{Z})$
- $W_{\mathcal{P}}$ the stepping of W w.r.t. a partition $\mathcal{P}$
- $\|W\|_\infty := \sup_{x,y\in[0,1]} \|W(x,y;\cdot)\|_\infty$
- $M_W(\mathrm{d}z) = |W|([0,1]^2;\mathrm{d}z)$
- W_G the probability-graphon associated to a $\mathcal{M}_1(\mathbf{Z})$-graph or a weighted graph G
- $\mathbb{H}(k,W)$ the $\mathcal{M}_1(\mathbf{Z})$-graph with k vertices sampled from $W \in \mathcal{W}_1$
- $\mathbb{G}(k,W)$ the $\mathcal{M}_1(\mathbf{Z})$-graph with k vertices sampled from $W \in \mathcal{W}_1$
- F^g a finite graph whose edges are decorated with functions in $C_b(\mathbf{Z})$
- $t(F^g, W) = M_W^F(g)$ the homomorphism density of F^g in W

Distances/norms on graphon spaces

- $d_{\square,\mathrm{m}}$ the cut distance associated to d_{m}
- $N_{\square,\mathrm{m}}$ the cut norm associated to N_{m}
- $\delta_{\square}$ the unlabeled distance associated to an arbitrary distance d
- $\delta_{\square,\mathrm{m}}$ the unlabeled cut distance associated to $d_{\square,\mathrm{m}}$ or $N_{\square,\mathrm{m}}$
- $\|\cdot\|_{\square,\mathbb{R}}$ the cut norm for real-valued kernels
- $\|\cdot\|^+_{\square,\mathbb{R}}$ the positive part of the cut norm for real-valued kernels

Definitions

- *weak isomorphism* of kernels and graphons in Definition 3.3.16 on page 59
- *tightness* for sets of kernels or graphons in Definition 3.4.7 on page 65
- *invariant* and *smooth* for a distance d on graphon spaces in Definition 3.3.10 on page 57
- *weakly regular* and *regular w.r.t. the stepping operator* for a distance d on graphon spaces in Definition 3.4.10 on page 66

3.A Proofs omitted in the main body of the text

In this section, we give all the proofs that were omitted from the main body of the text for being straightforward adaptation of proofs already existing in the literature (mostly from [Lov12], but there is one adaptation from [Bog18]); only the proof of Lemma 3.A.1 is original. Note that this section was not included in [ADW23] for brevity.

3.A.1 Proof from Section 3.3

To prove Proposition 3.3.18, we first need to prove the following lemma.

Lemma 3.A.1. *Any measure-valued kernel W can be approximated by a sequence $(W_n)_{n\in\mathbb{N}}$ of continuous measure-valued kernels for almost everywhere weak convergence (i.e. $W_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$ for almost every $(x,y)\in[0,1]^2$).*

Proof. We are going to prove that the stepfunctions given by Lemma 3.4.5, for the special choice of dyadic partitions, can themselves be approximated by continuous measure-valued kernels.

Let W be a measure-valued kernel, and denote by $C>0$ the constant given by the definition of a measure-valued kernel such that for every $(x,y)\in[0,1]^2$, the total mass of $W(x,y;\cdot)$ is upper bounded by C. Applying Lemma 3.4.5 with the sequence of dyadic partitions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we get a sequence of stepfunctions $(W_n = W_{\mathcal{P}_n})_{n\in\mathbb{N}}$ that are uniformly bounded by C and that converges to W on a subset $D\subset[0,1]^2$ of Lebesgue measure 1. For every $n\in\mathbb{N}$, let $A_n = \{(x,y)\in[0,1]^2 : \exists k\in\mathbb{N},\ x = k2^{-n} \text{ or } y = k2^{-n}\}$ be the dyadic grid of order n of $[0,1]^2$. Remark that A_n is the border set of steps of the stepfunction W_n. Let B_n be the set of points from $[0,1]^2$ that are at (euclidean) distance at most 4^{-n} from A_n. In particular, the Lebesgue measure of B_n is upper bounded by $2\times 2^n\times 2\times 4^{-n} = 4\times 2^{-n}$. We define the continuous measure-valued kernel U_n by $U_n(x,y;\cdot) = W_n(x,y;\cdot)$ for $(x,y)\in[0,1]^2\setminus B_n$, and we define $U_n(x,y;\cdot)$ for $(x,y)\in B_n$ as a convex interpolation of the values from the neighboring steps. In particular, for every $(x,y)\in[0,1]^2$, the total mass of $W(x,y;\cdot)$ is upper bounded by C.

Define the subset $F = \limsup_n B_n = \cap_{k\in\mathbb{N}}\cup_{n\geq k} B_n$. Then, every $(x,y)\in D\setminus F$ belongs to finitely many B_n, hence $U_n(x,y;\cdot) = W_n(x,y;\cdot)$ for n large enough, and thus $U_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$. Denote by λ denotes the Lebesgue measure on $[0,1]^2$, then we have that:

$$\begin{aligned}\lambda(F) &= \lim_{k\to\infty}\lambda(\cup_{n\geq k}B_n)\\ &\leq \lim_{k\to\infty}\sum_{n\geq k}\lambda(B_n)\\ &\leq \lim_{k\to\infty}\sum_{n\geq k}4\times 2^{-n}\\ &= \lim_{k\to\infty}8\times 2^{-k} = 0\end{aligned}$$

Hence, $\lambda(D\setminus F) = 1$, and we have that $U_n(x,y;\cdot)$ weakly converges to $W(x,y;\cdot)$ for almost every $(x,y)\in[0,1]^2$, which complets the proof. □

Proof of Proposition 3.3.18. [This proof is an exact copie of the proof of Theorem 8.13 in [Lov12] to suit a more general setting.]

To simplify the notations, we will denote the cut distances $d = d_{\square,\mathrm{m}}$ and $\delta = \delta_{\square,\mathrm{m}}$ during this proof.

The equality of the first expressions in each line of (3.7) follows from the fact that the set of invertible measure-preserving maps $S_{[0,1]}$ has a group structure. The group structure of $S_{[0,1]}$ also shows that δ

(defined by the first expression of the first line) is symmetric and satisfy the triangular inequality, hence is a premetric.

Let W be a kernel stepfunction, and $\varphi \in \bar{S}_{[0,1]}$ a measure-preserving map. Thus, W^φ is also a stepfunction with the same number of steps, same step sizes, and same step values as W. Hence, there exists a *bijective* measure-preserving map $\psi \in S_{[0,1]}$ (i.e. an *invertible* measure-preserving map) such that $W^\varphi = W^\psi$. This implies that in each line of (3.7) the two expressions are equal when U and W are stepfunctions. The equality in (3.8) follows from a similar argument in this case.

Now, consider arbitrary kernels U and W. Using Lemma 3.4.5, let $(U_n)_{n\in\mathbb{N}}$ and $(W_n)_{n\in\mathbb{N}}$ be sequences of stepfunctions that are uniformly bounded and weakly converging almost everywhere to U and W, respectively. Then, due to the smoothness of d, we have $d(U_n, U) \to 0$ and $d(W_n, W) \to 0$. Using the invariance of d, we have for every measure-preserving map φ that:

$$|d(U_n, W_n^\varphi) - d(U, W^\varphi)| \leq d(U_n, U) + d(W_n^\varphi, W^\varphi) = d(U_n, U) + d(W_n, W) \underset{n\to+\infty}{\longrightarrow} 0.$$

This implies that:

$$\inf_{\varphi\in\bar{S}_{[0,1]}} d(U_n, W_n^\varphi) = \inf_{\psi\in S_{[0,1]}} d(U_n, W_n^\psi) \underset{n\to+\infty}{\longrightarrow} \inf_{\psi\in S_{[0,1]}} d(U, W^\psi) = \delta(U, W);$$

and also that:

$$\inf_{\varphi\in\bar{S}_{[0,1]}} d(U_n, W_n^\varphi) \to \inf_{\varphi\in\bar{S}_{[0,1]}} d(U, W^\varphi),$$

which proves the equality in the first line of (3.7). The equalities in the other lines of (3.7) and in (3.8) follow similarly, but with a min instead of inf in the last line.

We now want to show that:

$$\delta(U, W) = \min_\mu d_\mu(U^\pi, W^\rho),$$

where μ range over all coupling measures on $[0,1]^2$.

To prove that the infimums in the last two expressions of δ are indeed mininums, we begin with (3.8). The space of coupling measures is compact in the weak topology, so it suffices to show that $d_\mu(U^\pi, W^\rho)$, as a function of μ, is lower semicontinuous. This means that if μ_n weakly converges to μ (where $(\mu_n)_{n\in\mathbb{N}}$ and μ are coupling measures), then for every two measure-valued kernels U and W, we have:

$$\liminf_n d_{\mu_n}(U^\pi, W^\rho) \geq d_\mu(U^\pi, W^\rho). \tag{3.34}$$

As a first step, we prove that $\lim_{n\to\infty} d_{\mu_n}(U, W) = d_\mu(U, W)$ for every continuous fonctions $U, W : [0,1]^2 \times [0,1]^2 \to \mathcal{M}_\pm(\mathbf{Z})$. Using Proposition A.6 from [Lov12], as $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to μ, there exist functions f and f_n for all $n \in \mathbb{N}$ which are measure preserving functions from $[0,1]$ (equipped with the Lebesgue measure) to $[0,1]^2$ equipped with the measure μ (resp. μ_n), and such that $(f_n)_{n\in\mathbb{N}}$ converges to f almost everywhere (on $[0,1]$). In particular, we have that $d_{\mu_n}(U, W) = d(U^{f_n}, W^{f_n})$ and $d_\mu(U, W) = d(U^f, W^f)$. Since U and W are continuous, we have that $U^{f_n}(x, y) = U(f_n(x), f_n(y)) \to U(f(x), f(y)) = U^f(x, y)$ as $n \to \infty$ for almost every $(x, y) \in [0,1]^2$. By the smoothness assumption on the distance d, this implies that $d_{\mu_n}(U, W) \to d_\mu(U, W)$ as $n \to \infty$. For the special case $U' = U^\pi$ and $W' = W^\rho$, where U and W are continuous measure-valued kernels, we get (3.34).

Let $U, W : [0,1] \times [0,1] \to \mathcal{M}_\pm(\mathbf{Z})$ be arbitrary measure-valued kernels, and fix any $\varepsilon > 0$. By Lemma 3.A.1, there exist sequences of continuous measure-valued kernels (U_n) and (W_n) such that $U_n \to U$ and $W_n \to W$ as $n \to \infty$ almost everywhere on $[0,1]^2$. By the smoothness assumption on d, we can fix k large enough such that $d(U_k, U) \leq \varepsilon$ and $d(W_k, W) \leq \varepsilon$.

Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be coupling measures on $[0,1]^2$ such that $(\mu_n)_{n\in\mathbb{N}}$ weakly converges to μ. By the special case of continuous kernels proved above, we know that:

$$d_{\mu_n}(U_k^\pi, W_k^\rho) \underset{n\to\infty}{\longrightarrow} d_\mu(U_k^\pi, W_k^\rho),$$

and for n large enough we have $|d_{\mu_n}(U_k^\pi, W_k^\rho) - d_\mu(U_k^\pi, W_k^\rho)| \leq \varepsilon$. Then, we have:

$$\begin{aligned} d_\mu(U^\pi, W^\rho) &\leq d_\mu(U_k^\pi, W_k^\rho) + d_\mu(U_k^\pi, U^\pi) + d_\mu(W_k^\rho, W^\rho) \\ &= d_\mu(U_k^\pi, W_k^\rho) + d(U_k, U) + d(W_k, W) \\ &\leq d_\mu(U_k^\pi, W_k^\rho) + 2\varepsilon. \end{aligned}$$

Moreover, for n large enough we have:

$$\begin{aligned} d_\mu(U_k^\pi, W_k^\rho) &\leq d_{\mu_n}(U_k^\pi, W_k^\rho) + \varepsilon \\ &\leq d_{\mu_n}(U^\pi, W^\rho) + d_{\mu_n}(U_k^\pi, U^\pi) + d_{\mu_n}(W_k^\rho, W^\rho) + \varepsilon \\ &= d_{\mu_n}(U^\pi, W^\rho) + d(U_k, U) + d(W_k, W) + \varepsilon \\ &\leq d_{\mu_n}(U^\pi, W^\rho) + 3\varepsilon. \end{aligned}$$

Combining these inequalities, we get that $d_\mu(U^\pi, W^\rho) \leq d_{\mu_n}(U^\pi, W^\rho) + 5\varepsilon$ if n is large enough. This proves (3.34) and thereby the existence of the minimum in (3.8).

The existence of the minimum in (3.7) follows easily now. Let μ be a coupling measure such that $\delta(U, W) = d_\mu(U^\pi, W^\rho)$. Let σ be a measure-preserving bijection from $[0, 1]$ with the Lebesgue measure into $[0, 1]^2$ with the measure μ. Remind that π and ρ denotes the projection onto the two coordinates. The fact that μ is a coupling measures implies that the compositions $\varphi = \sigma\pi$ and $\psi = \sigma\rho$ are measure-preserving, and

$$d(U^\varphi, W^\psi) = d_\mu(U^\pi, W^\rho) = \delta(U, W).$$

This concludes the proof of the proposition. □

Proof of Lemma 3.3.22. [This proof is a copie of the proof of Theorem 3.2.2.(i) in [Bog18] to extend it to the case of $\mathcal{M}_+(\mathbf{Z})$.]

Remind that the second inequality in the lemma was already stated in (3.10). We start by proving the first inequality in the lemma. Let r be such that $d_{\mathcal{P}}(\mu, \nu) > r > 0$. It follows from the definition of $d_{\mathcal{P}}$ that there exists a closed set $A \subset \mathbf{Z}$ such that:

$$\mu(A) > \nu(A^r) + r \quad (\text{or } \nu(A) > \mu(A^r) + r).$$

Assume without loss of generality that we are in the first case. There is a non-negative continuous function f on $\mathbf{Z}$ such that $|f(x) - f(y)| \leq d(x, y)/r$, $\|f\|_\infty \leq 1$, $f = 1$ on A and $f = 0$ outside A^r; for example, one can take $f(x) = \max(0, 1 - d(x, A)/r)$. Then, we have:

$$(1 + 1/r)\|\mu - \nu\|_{\mathrm{FM}} \geq \int f \ \mathrm{d}(\mu - \nu) \geq \mu(A) - \nu(A^r) \geq r.$$

Therefore, taking the supremum over r, we get the first inequality in the lemma.

We now prove the last inequality in the lemma. Now, let $\varepsilon > 0$ with $\varepsilon < \|\mu - \nu\|_{\mathrm{KR}}$, and let $f \in C_b(\mathbf{Z})$ be a 1-Lipschitz function such that $\|f\|_\infty \leq 1$, and $\int_{\mathbf{Z}} f \ \mathrm{d}(\mu - \nu) > \varepsilon > 0$. Let us consider the distribution functions $\Phi_\mu(t) = \mu(f < t)$ and $\Phi_\nu(t) = \nu(f < t)$. Integrating by parts, and applying Formula 1.2.1 from [Bog18], we obtain:

$$\begin{aligned} \int_{-1}^{1} [\Phi_\nu(t) - \Phi_\mu(t)] \ \mathrm{d}t &= \Phi_\nu(1) - \Phi_\mu(1) + \int_{-1}^{1} t \ \mathrm{d}(\Phi_\mu - \Phi_\nu)(t) \\ &= \nu(\mathbf{Z}) - \mu(\mathbf{Z}) + \int_{\mathbf{Z}} f \ \mathrm{d}(\mu - \nu) \\ &> \nu(\mathbf{Z}) - \mu(\mathbf{Z}) + \varepsilon. \end{aligned} \tag{3.35}$$

Let $r > 0$, and let us find under which condition there exists $\tau \in [-1, 1]$ such that $\Phi_\nu(\tau) > \Phi_\mu(\tau + r) + r$. Let us suppose that $\Phi_\nu(t) \leq \Phi_\mu(t + r) + r$ for all t. Integrating over $[-1, 1]$, we have:

$$\int_{-1}^{1} \Phi_\nu(t) \ \mathrm{d}t \leq \int_{-1+r}^{1+r} \Phi_\mu(t) \ \mathrm{d}t + 2r \leq \int_{-1}^{1} \Phi_\mu(t) \ \mathrm{d}t + (\mu(\mathbf{Z}) + 2)r, \tag{3.36}$$

where in the last inequality we used that $\Phi_\mu(t) = \mu(\mathbf{Z})$ whenever $t > 1$ and $\Phi_\mu(t) = 0$ whenever $t < -1$. In order for (3.35) and (3.36) to yield a contradiction, we need

$$(\mu(\mathbf{Z}) + 2)r \leq \varepsilon + \nu(\mathbf{Z}) - \mu(\mathbf{Z}).$$

By symmetry of μ and ν for distances, without loss of generality we may assume that $\mu(\mathbf{Z}) \leq \nu(\mathbf{Z})$. Hence, in particular the we get the desired contradiction whenever $(\mu(\mathbf{Z}) + 2)r \leq \varepsilon$.

Now, assume that $r = \varepsilon/(\mu(\mathbf{Z}) + 2)$, thus we have $\tau \in [-1, 1]$ such that $\Phi_\nu(\tau) > \Phi_\mu(\tau + r) + r$. Let $B := f^{-1}([-1, \tau])$. Then,

$$B^r \subset f^{-1}([-1, \tau + r])$$

since f is 1-Lipschitz. Hence, we have $\nu(B) > \mu(B^r) + r$, and thus we have $d_{\mathcal{P}}(\mu, \nu) > r$. Taking the supremum over $\varepsilon < \|\mu - \nu\|_{\mathrm{KR}}$, we get the last inequality in the lemma. □

3.A.2 Proof from Section 3.4.4

Before proving Lemma 3.4.16, we first need to prove the following lemma which gives refining approximation partitions.

Lemma 3.A.2 (Refining partition)**.** *Let $U \in \mathcal{W}_\pm$ be a signed measure-valued kernel, and let $\mathcal{Q}$ be a finite partition of $[0, 1]$. Let $\varepsilon > 0$. Then there is a partition $\mathcal{P}$ refining $\mathcal{Q}$ with at most $4|\mathcal{Q}|$ classes such that:*

$$N_{\square,\mathrm{m}}(U_{\mathcal{P}} - U_{\mathcal{Q}}) \;\le\; N_{\square,\mathrm{m}}(U - U_{\mathcal{Q}}) \;\le\; N_{\square,\mathrm{m}}(U_{\mathcal{P}} - U_{\mathcal{Q}}) + \varepsilon.$$

Proof of Lemma 3.A.2. [The proof of this lemma is a straightforward adaptation of the proof of [Lov12, Lemma 9.11.(b)].] The inequality $N_{\square,\mathrm{m}}(U - U_{\mathcal{Q}}) \ge N_{\square,\mathrm{m}}(U_{\mathcal{P}} - U_{\mathcal{Q}})$ follows by Lemma 3.4.11 on the 1-Lipschitzness of the stepping operator. To prove the other direction, let $\varepsilon > 0$, and let S and T be measurable subsets of $[0, 1]$ that almost realizes the supremum in the cut norm with error ε, that is:

$$N_{\square,\mathrm{m}}(U - U_{\mathcal{Q}}) - \varepsilon \le N_{\mathrm{m}}\left(\int_{S\times T} U(x, y; \cdot) - U_{\mathcal{Q}}(x, y; \cdot)\, \mathrm{d}x\mathrm{d}y\right). \tag{3.37}$$

Let $\mathcal{P}$ denote the partition generated by $\mathcal{Q}$, S and T. Clearly, $\mathcal{P}$ has at most $4k$ classes. As S and T are unions of sets from $\mathcal{P}$, we have equality between the integrals:

$$\int_{S\times T} U(x, y; \cdot)\, \mathrm{d}x\mathrm{d}y = \int_{S\times T} U_{\mathcal{P}}(x, y; \cdot)\, \mathrm{d}x\mathrm{d}y. \tag{3.38}$$

Hence, we get:

$$N_{\square,\mathrm{m}}(U - U_{\mathcal{Q}}) - \varepsilon \le N_{\mathrm{m}}\left(\int_{S\times T} U_{\mathcal{P}}(x, y; \cdot) - U_{\mathcal{Q}}(x, y; \cdot)\, \mathrm{d}x\mathrm{d}y\right) \le N_{\square,\mathrm{m}}(U_{\mathcal{P}} - U_{\mathcal{Q}}), \tag{3.39}$$

which completes the proof. □

Let $\mathcal{P}$ be a finite partition of $[0, 1]$. As the stepping operator for measurable real-valued L^2 functions on $[0, 1]^2$ is a linear projection, and is idempotent and symmetric, and by definition of the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ for signed measure-valued kernels, we have that the stepping operator for signed measure-valued kernels is linear, idempotent and symmetric for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$:

$$\langle U_{\mathcal{P}}, W_{\mathcal{P}}\rangle_{\mathcal{F}} = \langle U_{\mathcal{P}}, W\rangle_{\mathcal{F}} = \langle U, W_{\mathcal{P}}\rangle_{\mathcal{F}}. \tag{3.40}$$

The stepping operator is the orthogonal projection for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ onto the space of stepfunctions with steps in $\mathcal{P}$, as shown by the identity:

$$\langle W_{\mathcal{P}}, W - W_{\mathcal{P}}\rangle_{\mathcal{F}} = 0. \tag{3.41}$$

Proof of Lemma 3.4.16. [This proof is a straightforward adaptation of the proof of the weak regularity lemma for real-valued graphons in [Lov12, Lemma 9.9].] Let $W \in \mathcal{W}_\pm$ be a signed measure-valued kernel. Let $\mathcal{Q}$ be a finite partition of $[0, 1]$. Let $\mathcal{P}$ be the finite partition given by Lemma 3.A.2 for W, $\mathcal{Q}$ and $\varepsilon = (1 - 1/\sqrt{2}) \times \|W - W_{\mathcal{Q}}\|_{\square,\mathcal{F}}$. Note that $\|W - W_{\mathcal{Q}}\|_{\square,\mathcal{F}} \le \sqrt{2}\|W_{\mathcal{P}} - W_{\mathcal{Q}}\|_{\square,\mathcal{F}}$. Using Lemma 3.4.15, we get:

$$(\|W - W_{\mathcal{Q}}\|_{\square,\mathcal{F}})^2 \le 2(\|W_{\mathcal{P}} - W_{\mathcal{Q}}\|_{\square,\mathcal{F}})^2 \le 4\|W_{\mathcal{P}} - W_{\mathcal{Q}}\|_{2,\mathcal{F}}^2.$$

As the partition $\mathcal{P}$ is a refinement of the partition $\mathcal{Q}$, we have that $(W_{\mathcal{P}})_{\mathcal{Q}} = W_{\mathcal{Q}}$. Using that the stepping operator is an orthogonal projection for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ (see (3.41)) we get:

$$\|W_{\mathcal{P}} - W_{\mathcal{Q}}\|_{2,\mathcal{F}}^2 = \|W_{\mathcal{P}}\|_{2,\mathcal{F}}^2 - \|W_{\mathcal{Q}}\|_{2,\mathcal{F}}^2 = \|W - W_{\mathcal{Q}}\|_{2,\mathcal{F}}^2 - \|W - W_{\mathcal{P}}\|_{2,\mathcal{F}}^2.$$

Let $\ell \geq 1$. We start with the trivial partition $\mathcal{P}_0 = \{[0,1]\}$. We apply repeatedly Lemma 3.A.2, to get new refining partitions $\mathcal{P}_j$ of size at most 4^j such that:

$$(\|W - W_{\mathcal{P}_{j-1}}\|_{\square,\mathcal{F}})^2 \;\leq\; 4\Big(\|W - W_{\mathcal{P}_{j-1}}\|_{2,\mathcal{F}}^2 - \|W - W_{\mathcal{P}_j}\|_{2,\mathcal{F}}^2\Big),$$

and then summing those inequalities we get:

$$\sum_{j=0}^{\ell-1}(\|W - W_{\mathcal{P}_j}\|_{\square,\mathcal{F}})^2 \;\leq\; 4\Big(\|W - W_{\mathcal{P}_0}\|_{2,\mathcal{F}}^2 - \|W - W_{\mathcal{P}_\ell}\|_{2,\mathcal{F}}^2\Big) \;\leq\; 4\|W - W_{\mathcal{P}_0}\|_{2,\mathcal{F}}^2 \;\leq\; 4\|W\|_{2,\mathcal{F}}^2,$$

where we used in the last inequality that the stepping operator is an orthogonal projection for $\langle\cdot,\cdot\rangle_{\mathcal{F}}$. Hence, there exists $j < \ell$ such that $(\|W - W_{\mathcal{P}_j}\|_{\square,\mathcal{F}})^2 \;\leq\; \frac{4}{\ell}\|W\|_{2,\mathcal{F}}^2$. Setting $\mathcal{P} = \mathcal{P}_j$ we get a finite partition of size at most $4^{\ell-1}$ such that $\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \;\leq\; \frac{\sqrt{4}}{\sqrt{\ell}}\|W\|_{2,\mathcal{F}}$.

Now for $k \geq 1$, let $\ell \geq 1$ be such that $4^{\ell-1} \leq k < 4^\ell$. Thus $\log(k) \leq \ell \log 4 \leq 2\ell$. By the preceding argument, there exists a finite partition $\mathcal{P}$ of size at most $4^{\ell-1} \leq k$ such that:

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \;\leq\; \frac{\sqrt{4}}{\sqrt{\ell}}\|W\|_{2,\mathcal{F}} \;\leq\; \frac{\sqrt{8}}{\sqrt{\log(k)}}\|W\|_{2,\mathcal{F}}.$$

This concludes the proof. □

Proof of Lemma 3.4.17. [The proof is a straightforward adaptation of the proof of [Lov12, Lemma 9.15].] Statement (i) follows by the same argument as Lemma 3.4.16, just starting from the partition $\mathcal{P}_0 = \mathcal{Q}$.

To prove (ii), we partition each class of $\mathcal{Q}$ into classes of measure $1/k$, with at most one exceptional class of size less then $1/k$. Keeping all classes of size $1/k$, let us take the union of the exceptional classes, and repartition it into classes of size $1/k$, to get a partition $\mathcal{P}$.

To analyze this construction, let us also consider the common refinement $\mathcal{R} = \mathcal{P} \wedge \mathcal{Q}$. Then $W_{\mathcal{R}}$ and $W_{\mathcal{P}}$ differ on a set of measure at most $2(m/k)$, and so

$$\|W - W_{\mathcal{P}}\|_{\square,\mathcal{F}} \leq \|W - W_{\mathcal{R}}\|_{\square,\mathcal{F}} + \frac{2m}{k}\cdot$$

Lemma 3.4.12 implies that $\|W - W_{\mathcal{R}}\|_{\square,\mathcal{F}} \leq 2\|W - W_{\mathcal{Q}}\|_{\square,\mathcal{F}}$, which completes the proof. □

3.A.3 Proof from Section 3.6

Concentration of samples

We call a *graph parameter* a function on the set of graphs (simple, weighted or decorated depending on context) with values in $\mathbb{R}$. We say that a graph parameter f is *reasonably smooth* if we have $|f(G) - f(G')| \leq 1$ for every two graphs G and G' defined on the same set of vertices and whose edge set, or the weights/decorations of those edges, may only differ on edges incident with a single vertex. Thus, for every $k \geq 1$, any graph parameter is bounded on the set of graphs with k vertices.

The following proposition gives a concentration bound for reasonably smooth graph parameter evaluated on a W-random graph. The proof is a straightforward adaptation of the proof of [Lov12, Theorem 10.3].

Proposition 3.A.3 (Concentration of samples from probability-graphons)**.** *Let f be a reasonably smooth graph parameter defined on weighted graphs for (i) and on $\mathcal{M}_1(\mathbf{Z})$-graphs for (ii). Let W be a probability-graphon and $k \geq 1$.*

(i) Using the notation $f_0 = \mathbb{E}[f(\mathbb{G}(k,W))]$, we get that for every $t \geq 0$:

$$\mathbb{P}\left(f(\mathbb{G}(k,W)) \geq f_0 + \sqrt{2tk}\right) \leq \mathrm{e}^{-t}.$$

(ii) Using the notation $f_0 = \mathbb{E}[f(\mathbb{H}(k,W))]$, we get that for every $t \geq 0$:

$$\mathbb{P}\left(f(\mathbb{H}(k,W)) \geq f_0 + \sqrt{2tk}\right) \leq \mathrm{e}^{-t}.$$

We can apply the same inequality to the function $-f$ to get a large deviation probability bound from the mean in the other direction. Combining those two bounds, we get a large deviation probability bound from the mean in absolute value.

Proof of Proposition 3.A.3. [This is a copy of the proof of [Lov12, Theorem 10.3] to suit our setting.] For this proof we need a result from the appendix of Lovász's book [Lov12]: the Corollary A.15 of Azuma's inequality. The quantity $f(\mathbb{H}(\{x_1,\dots,x_k\},W))$, as a function of $x_1,\dots,x_k \in [0,1]$, satisfies the conditions of corollary A.15 of Azuma's inequality, and applying this inequality with $n=k$ and $\varepsilon = (2t/k)^{1/2}$, we get the inequality in Point (ii). The proof of the inequality in Point (i) is essentially the same. □

Missing details for the proof of Lemma 3.6.9. In the proof of Proposition 3.A.3, we only skipped the proof of (3.24), which we re-state:

$$\forall n \in [\![1,N]\!], \quad \mathbb{P}(d_{\square}(G[f_n],H[f_n]) \ > \varepsilon) \leq 2\cdot 4^k\, \mathrm{e}^{-2\varepsilon^2 k^2}\,.$$

Remind that for two real-weighted graphs with k vertices G' and H', the cut distance $d_{\square,\mathbb{R}}$ is equivalently defined as (see [Lov12, Chapter 8]):

$$d_{\square,\mathbb{R}}(G',W') = \sup_{I,J\subset[k]} \frac{1}{k^2}\left|\sum_{i\in I, j\in J} \Phi_{G'}(i,j) - \Phi_{H'}(i,j)\right|.$$

In order to control the N terms from the right member of (3.23), we adapt the approach used in the proof of [Lov12, Lemma 10.11]. Remark that for every $n\in\mathbb{N}$, $H[f_n]$ and $G[f_n]$ are real-weighted graphs with weights in $[0,1]$. Fix some $n\in[N]$. For $i,j\in[k]$, define the random variable:

$$X_{i,j} = \Phi_G(i,j;f_n),$$

where we remind that $\Phi_G(i,j;f_n) = f_n(Z_{i,j})$ with $Z_{i,j}$ distributed according to $\Phi_H(i,j;\cdot)$. The random variables $(X_{i,j} : i,j\in[k])$ are independent and take values in $[0,1]$, and we have:

$$\mathbb{E}[X_{i,j}] = \Phi_H(i,j;f_n).$$

Thus, for $S,T\subset[k]$, we have:

$$\sum_{i\in S,j\in T} \Phi_G(i,j;f_n) - \Phi_H(i,j;f_n) = \sum_{i\in S,j\in T} X_{i,j} - \mathbb{E}[X_{i,j}].$$

We say the pair (S,T) is *bad* if $|\sum_{i\in S,j\in T} \Phi_G(i,j;f_n) - \Phi_H(i,j;f_n)| > \varepsilon k^2$. The probability that the pair (S,T) is bad can be estimated using Chernoff-Hoeffding's inequality and the fact that $|S|$ and $|T|$ are bounded by k:

$$\mathbb{P}\left(\left|\sum_{i\in S,j\in T} X_{i,j} - E[X_{i,j}]\right| > \varepsilon k^2\right) \leq 2\exp\left(-2\cdot(\varepsilon k^2)^2\cdot\frac{1}{|S||T|}\right) \leq 2\exp\left(-2\varepsilon^2k^2\right).$$

The total number of pair (S,T) is 4^k, and thus the probability that there is at least one bad pair for the graphs $G[f_n]$ and $H[f_n]$ is upper bounded by $2\cdot 4^k\,\mathrm{e}^{-2\varepsilon^2k^2}$. If there is no bad pair (S,T) for the graphs $G[f_n]$ and $H[f_n]$, then we have $d_{\square}(G[f_n],H[f_n]) \leq \varepsilon$. Hence, we get that (3.24) holds, which conclude the proof. □

Other proofs

Proof of Lemma 3.6.12. [This proof is a straightforward adaptation of the proof of [Lov12, Lemma 10.16], replacing real-valued graphons by probability-graphons.]

We start by upper bounding the expectations of the distances. Define $m = \lceil k^{1/4}\rceil$. We apply the weak regularity Lemma 3.4.17-(i) to the probability-graphon W to get a finite partition $\mathcal{Q}$ of size $q = m/\log(m)^{3/2}$ such that $d_{\square,\mathcal{F}}(W,W_{\mathcal{Q}}) \leq \frac{4}{\sqrt{\ln(q)}}$. Then, we apply Lemma 3.4.17 (ii) to replace $\mathcal{Q}$ by

an equipartition (remind this means that each set as the same measure) $\mathcal{P} = \{V_1, \dots, V_m\}$ of $[0,1]$ into m classes for which we have:

$$d_{\square,\mathcal{F}}(W, W_{\mathcal{P}}) \leq \frac{8}{\sqrt{\ln(q)}} + \frac{2q}{m} \leq \frac{8.5}{\sqrt{\ln(k)}}, \tag{3.42}$$

where the last bound derives from simple calculus. Let $X = (X_1, \dots, X_k)$ be k iid random variables uniformly distributed over $[0,1]$. By the First Sampling Lemma 3.6.7, we have:

$$\left|d_{\square,\mathcal{F}}(\mathbb{H}(X,W), \mathbb{H}(X,W_{\mathcal{P}})) - d_{\square,\mathcal{F}}(W, W_{\mathcal{P}})\right| \leq \frac{9}{k^{1/4}}, \tag{3.43}$$

with probability at least $1 - 4k^{1/4}\,\mathrm{e}^{-\sqrt{k}/10}$. By upper bounding the difference of the two distances by 1 in the exceptional cases, we get:

$$\mathbb{E}\left[\left|d_{\square,\mathcal{F}}(\mathbb{H}(X,W), \mathbb{H}(X,W_{\mathcal{P}})) - d_{\square,\mathcal{F}}(W, W_{\mathcal{P}})\right|\right] \leq \frac{9}{k^{1/4}} + 4k^{1/4}\,\mathrm{e}^{-\sqrt{k}/10} \leq \frac{10}{k^{1/4}} < \frac{1}{\sqrt{\ln(k)}}, \tag{3.44}$$

where the last two bounds derives from simple calculus. Combining the bounds from (3.42) and (3.44), we get:

$$\begin{aligned}\mathbb{E}[d_{\square,\mathcal{F}}(\mathbb{H}(X,W), \mathbb{H}(X,W_{\mathcal{P}}))] &\leq \mathbb{E}\left[\left|d_{\square,\mathcal{F}}(\mathbb{H}(X,W), \mathbb{H}(X,W_{\mathcal{P}})) - d_{\square,\mathcal{F}}(W, W_{\mathcal{P}})\right|\right] \\ &\qquad + d_{\square,\mathcal{F}}(W, W_{\mathcal{P}}) \\ &\leq \frac{9.5}{\sqrt{\ln(k)}}.\end{aligned} \tag{3.45}$$

Hence, combining the bounds from (3.42) and (3.45), we get:

$$\begin{aligned}\mathbb{E}[\delta_{\square,\mathcal{F}}(W, \mathbb{H}(X,W))] &\leq \mathbb{E}\Big[\delta_{\square,\mathcal{F}}(W, W_{\mathcal{P}}) + \delta_{\square,\mathcal{F}}(W_{\mathcal{P}}, \mathbb{H}(X,W_{\mathcal{P}})) + \delta_{\square,\mathcal{F}}(\mathbb{H}(X,W_{\mathcal{P}}), \mathbb{H}(X,W))\Big] \\ &\leq \frac{18}{\sqrt{\ln(k)}} + \mathbb{E}[\delta_{\square,\mathcal{F}}(W_{\mathcal{P}}, \mathbb{H}(X,W_{\mathcal{P}}))].\end{aligned} \tag{3.46}$$

Hence, it is enough to prove that $\delta_{\square,\mathcal{F}}(W_{\mathcal{P}}, \mathbb{H}(X,W_{\mathcal{P}}))$ is small on average. For simplicity, write $H = \mathbb{H}(X, W_{\mathcal{P}})$. The probability-graphons $W_{\mathcal{P}}$ et W_H are almost identical: they are both stepfunctions with m steps, and they take the same $\mathcal{M}_1(\mathbf{Z})$-values on their corresponding steps. Their only difference is that the measure of the i-th step V_i is $1/m$ in $W_{\mathcal{P}}$, whereas the measure of the i-th step of W_H is $|V_i \cap X|/k$, which is close to $1/m$ when k is large enough.

Let us write $|V_i \cap X|/k = 1/m + r_i$, then upper bounding by 1 when the two stepfunctions do not agree, we get $\delta_{\square,\mathcal{F}}(W_{\mathcal{P}}, W_H) \leq \sum_{i=1}^m |r_i|$. We can then easily estimate the expectation of this distance:

$$\mathbb{E}[\delta_{\square,\mathcal{F}}(W_{\mathcal{P}}, W_H)] \leq \sum_{i=1}^m \mathbb{E}\big[|r_i|\big] = m\mathbb{E}\big[|r_1|\big] \leq m\sqrt{\mathbb{E}[r_1^2]} = \sqrt{\frac{m-1}{k}} < \frac{1}{k^{3/8}} < \frac{1}{\sqrt{\ln(k)}}, \tag{3.47}$$

where the last bound derives from simple calculus. And finally, combining (3.46) and (3.47), we get:

$$\mathbb{E}[\delta_{\square}(W, \mathbb{H}(X,W))] \leq \frac{19}{\sqrt{\ln(k)}}. \tag{3.48}$$

We get a similar upper bound on the expectation of the distance $\delta_{\square}(W, \mathbb{G}(X,W))$ by applying Remark 3.6.10:

$$\begin{aligned}\mathbb{E}[\delta_{\square,\mathcal{F}}(W, \mathbb{G}(X,W))] &\leq \mathbb{E}[\delta_{\square,\mathcal{F}}(W, \mathbb{H}(X,W))] + \mathbb{E}[\delta_{\square,\mathcal{F}}(\mathbb{H}(X,W)), \mathbb{G}(X,W))] \\ &\leq \frac{19}{\sqrt{\ln(k)}} + \frac{21}{\sqrt{k}} \\ &< \frac{20}{\sqrt{\ln(k)}},\end{aligned} \tag{3.49}$$

where the last bound derives from simple calculus.

And finally, the lemma is obtained by applying Proposition 3.A.3 for concentration of samples to the reasonably smooth graph parameter $f(G) = \frac{v(G)}{2}\delta_{\square,\mathcal{F}}(G,W)$ with $t = \frac{k}{2\ln(k)}$. Hence, with probability at least $1 - \exp(-k/(2\ln(k)))$, we have that $\delta_{\square,\mathcal{F}}(W, \mathbb{H}(X,W))$ can move away from its expectation of at most $\frac{2}{k}\sqrt{2tk} = \frac{2}{k}\sqrt{\frac{k^2}{\ln(k)}} = \frac{2}{\sqrt{\ln(k)}}$. And similarly for $\delta_{\square,\mathcal{F}}(W, \mathbb{G}(X,W))$. □

3.A.4 Proof from Section 3.7

Proof of Lemma 3.7.9. [This proof is a straightforward adaptation of the proof of [Lov12, Lemma 10.31 and Lemma 10.32], replacing real-valued graphons by probability-graphons.]

Abusing notations, we will identify a weight-value $n \in \mathbf{Z}$ with its indicator functions f_n, and doing this identification for edge-decoration functions, we will identify a $\mathcal{H}$-graph F^g with its corresponding weighted graph. In particular, doing so we get $t(F^g, W) = \mathbb{P}(\mathbb{G}(k, W) = F^g)$ for every $\mathcal{H}$-graph F^g with k vertices, and the assumption bounds in the lemma can be rephrased as:

$$|\mathbb{P}(\mathbb{G}(k, U) = F^g) - \mathbb{P}(\mathbb{G}(k, W) = F^g)| \leq 2^{-k-\log_2(n_0)k^2}.$$

As the space $\mathbf{Z}$ is finite, the random graphs $\mathbb{G}(k, U)$ and $\mathbb{G}(k, W)$ can only take finitely many values, and thus we can easily bound the total variation between them:

$$\begin{aligned} d_{\mathrm{var}}(\mathbb{G}(k, U), \mathbb{G}(k, W)) &= \sum_{F^g} \Big| \mathbb{P}(\mathbb{G}(k, U) = F^g) - \mathbb{P}(\mathbb{G}(k, W) = F^g) \Big| \\ &\leq {n_0}^{k^2} 2^{-k-\log_2(n_0)k^2} \\ &= 2^{-k} \\ &< 1 - 2\exp\left(-\frac{k}{2\ln(k)}\right), \end{aligned}$$

where in the first line, F^g ranges over all weighted graphs with k vertices. Hence, we can couple the two random graphs $\mathbb{G}(k, U)$ and $\mathbb{G}(k, W)$ such that they are equal with probability larger than $2\exp\left(-\frac{k}{2\ln(k)}\right)$.

We can now apply the Second Sampling Lemma 3.6.12 with the convergence determining sequence $\mathcal{F}$ to get that with probability at least $1 - \exp\left(-\frac{k}{2\ln(k)}\right)$, we have:

$$\delta_{\square,\mathcal{F}}(U, \mathbb{G}(k, U)) \leq \frac{22}{\sqrt{\ln(k)}};$$

and the same probabilistic bound hold for W. Hence, it follows that with positive probability those two bounds and the graph equality $\mathbb{G}(k, U) = \mathbb{G}(k, W)$ happen simultaneously, and consequently we get:

$$\delta_{\square,\mathcal{F}}(U, W) \leq \frac{44}{\sqrt{\ln(k)}},$$

which concludes the proof. □

Chapitre 4

Ergodic theorem for branching Markov chains indexed by trees with arbitrary shape

The material for this chapter has been released in [Wei24b] and is currently under review.

Abstract: *We prove an ergodic theorem for Markov chains indexed by the Ulam-Harris-Neveu tree over large subsets with arbitrary shape under two assumptions: with high probability, two vertices in the large subset are far from each other and have their common ancestor close to the root. The assumption on the common ancestor can be replaced by some regularity assumption on the Markov transition kernel. We verify that those assumptions are satisfied for some usual trees. Finally, with Markov-Chain Monte-Carlo considerations in mind, we prove when the underlying Markov chain is stationary and reversible that the Markov chain, that is the line graph, yields minimal variance for the empirical average estimator among trees with a given number of nodes.*

Key words and phrases— Tree indexed Markov chain, ergodic theorem, Bienaymé-Galton-Watson trees, minimal variance

2020 Mathematics Subject Classification— 60J05, 60J80, 60F25

4.1 Introduction

Branching Markov processes, which are a generalization of Markov chains to trees, are useful to describe the evolution and growth of a population. Limit theorems, such as the law of large numbers (sometimes also called ergodic theorem in markovian contexts), are important tools to study properties of a population such as the distribution of traits. The law of large numbers for branching Markov processes has been studied with both discrete and general values [AK98a, AK98b]. To study cellular aging, a more general version of the strong law of large numbers for a wider class of test functions and for non-independent daughter cells was proved in [Guy07], see also [DM10] for an extension to bounded degree Bienaymé-Galton-Watson trees and [Ban19] for an extension to time-varying environnement and trait-dependent offspring distribution. In this article, we present an ergodic theorem for a wide class of test functions as in [Guy07], and for branching Markov processes where reproduction is independent from individual traits but where the genealogical tree of the population can have an arbitrary shape.

A branching Markov process $X = (X_u, u \in T)$ with values in a metric space $\mathcal{X}$ is a random process indexed by a rooted tree T with the Markov property: sibling nodes take independent and identically distributed values that depend only on the value of their parent node. Without loss of generality, we may choose T to be the rooted Ulam-Harris-Neveu tree $\mathbb{T}^\infty = \cup_{n\in\mathbb{N}}(\mathbb{N}^*)^n$. See Definition 4.2.1 below for a complete formal definition.

For simplicity, in this introduction we restrict ourselves to the case where the transition kernel Q of the branching Markov process X is *ergodic* (resp. *uniformly ergodic*), that is for any continuous bounded function f on $\mathcal{X}$, we have for all $x \in \mathcal{X}$ that $\lim_{n\to\infty} |Q^n f(x) - \langle \mu, f\rangle| = 0$ (resp.

$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |Q^n f(x) - \langle \mu, f\rangle| = 0)$ where μ is the unique invariant measure of Q.

For a finite (non-empty) subset $A \subset \mathbb{T}^\infty$ and some function f on $\mathcal{X}$, we define the normalized empirical average:

$$\bar{M}_A(f) = |A|^{-1} \sum_{u\in A} f(X_u).$$

Our goal is to study the asymptotic behavior of the normalized empirical average when the averages are performed on a sequence $(A_n)_{n\in\mathbb{N}}$ of finite subsets of $\mathbb{T}^\infty$ whose size goes to infinity. For instance, the averaging set A_n can be the n-th generation G_n of a tree T, or T_n the tree T up to generation n.

To this end, we need some geometrical assumption on the sequence of finite subsets $(A_n)_{n\in\mathbb{N}}$, which states that vertices are far away from each other with high probability.

Assumption 1 (Geometrical)**.** *Let $A_n \subset \mathbb{T}^\infty$ for $n \in \mathbb{N}$ be finite (non-empty) subsets, and let U_n and V_n be independent and uniformly distributed over A_n. Denoting by d the graph distance on $\mathbb{T}^\infty$, for all $k \in \mathbb{N}$, we have:*

$$\mathbb{P}(d(U_n, V_n) \leq k) = |A_n|^{-2} \sum_{u,v\in A_n} \mathbb{1}_{\{d(u,v)\leq k\}} \underset{n\to\infty}{\longrightarrow} 0.$$

Let us stress that Assumption 1 implies that $\lim_{n\to\infty} |A_n| = \infty$.

We either need to assume that Q is uniformly ergodic, or that Q is ergodic and the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies the following condition stating that the last common ancestor of two vertices is near the root with high probability. Denote by $h(u)$ the height of a vertex u and by $u \wedge v$ the common ancestor of two vertices u and v (see Section 4.2.1).

Assumption 2 (Ancestral)**.** *For all $n \in \mathbb{N}$, let U_n and V_n be independent and uniformly distributed over A_n.*

The sequence of random variables $(h(U_n \wedge V_n))_{n\in\mathbb{N}}$ is tight, that is, for every $\varepsilon > 0$, there exists $k \in \mathbb{N}$ such that $\mathbb{P}(h(U_n \wedge V_n) > k) < \varepsilon$ for n large enough.

Note that Assumptions 1 and 2 are similar to Assumptions 2.(b) and 2.(a), respectively, considered in [Ban19] in the case where A_n is the n-th generation of the tree.

Remark 4.1.1 (Some sufficient conditions for Assumptions 1 and 2, see Section 4.3)**.** Assumption 1 is always satisfied for Cayley and Bethe trees and for bounded degree trees (see Lemma 4.3.1). Assumptions 1 and 2 are satisfied for spherically symmetric trees when $A_n = G_n$ (see Lemma 4.3.4). In Lemma 4.3.5, we prove that Assumptions 1 and 2 are satisfied for super-critical Bienaymé-Galton-Watson trees conditioned on non-extinction when $A_n = G_n$ or T_n.

We can now formulate the ergodic theorem for branching Markov processes on trees with arbitrary shape. In Section 4.2, we prove Theorem 4.2.2, a more general version of this theorem.

Theorem 4.1.2 (Ergodic theorem for Markov processes on trees with arbitrary shape)**.** *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite subsets of $\mathbb{T}^\infty$ that satisfies Assumption 1. Let X be a branching Markov process indexed by $\mathbb{T}^\infty$ with values in $\mathcal{X}$ whose transition kernel Q is ergodic. Assume that either Q is uniformly ergodic or that $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 2. Then, for every continuous bounded function f on $\mathcal{X}$, we have:*

$$\bar{M}_{A_n}(f) = |A_n|^{-1} \sum_{u\in A_n} f(X_u) \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} \langle \mu, f\rangle.$$

Remark 4.1.3. We discuss the main difference between Theorem 4.1.2 and the law of large numbers for branching Markov process found in [Guy07]. The results in [Guy07] apply to Markov processes where daughter nodes can have non-independent distributions when conditioning on their mother, whereas in our case they must be independent. In exchange, our results allow for more flexibility on the shape of the population's genealogical tree: for instance more flexibility on the number of children of each node (including Bienaymé-Galton-Watson trees with unbounded degree), or even allowing the number of children of a node to grow over time (e.g. the degree of the root can increase as $\log n$, i.e. slow condensation). Moreover, in our results, the empirical average can be performed on a wide variety of subsets of the tree, and not only to the n-th generation.

Lastly, motivated by Markov-Chain Monte-Carlo considerations, we study in Section 4.4 the variance of the empirical average estimator $\bar{M}_A(f)$ and its dependence on the shape of A. We perform exact variance computation in the case where the transition kernel Q induces a self-adjoint compact operator on $L^2(\mu)$, which proves the following proposition about non-asymptotic variance comparison.

Proposition 4.1.4 (The line graph has minimal variance)**.** *Let μ be an invariant measure for Q, and assume that the transition kernel Q induces a self-adjoint compact operator on $L^2(\mu)$. Let X be a branching Markov process on $\mathbb{T}^\infty$ with transition kernel Q and initial distribution ν. Let f be a non-constant function in $L^2(\mu)$.*

When $\nu = \mu$, we have that $\mathbb{E}\left[\bar{M}_A(f)\right] = \langle \mu, f\rangle$ for any finite subset $A \subset \mathbb{T}^\infty$, and thus the empirical average estimator has no bias. Moreover, the minimum of the map $A \mapsto \mathrm{Var}(\bar{M}_A(f))$ among subtrees of $\mathbb{T}^\infty$ with a given cardinal n is achieved by the line graph tree (i.e. the Markov chain).

Furthermore, when $n \geq 5$, the line graph is the only subtree of size n achieving this minimum if and only if $f \notin \mathrm{Ker}(Q) \oplus \mathrm{Ker}(Q - I) \oplus \mathrm{Ker}(Q + I)$.

Remark that as Q is a Markov kernel, its spectrum as an $L^2(\mu)$-operator is a subset of $[-1, 1]$. Note that when $f \in \mathrm{Ker}(Q) \oplus \mathrm{Ker}(Q - I)$, then the value of $\mathrm{Var}(\bar{M}_T(f))$ does not depend on the shape of the tree T. Also note that when $f \in \mathrm{Ker}(Q + I)$, then the value of $\mathrm{Var}(\bar{M}_T(f))$ is minimal among subtrees of size n when T has a balanced bipartite 2-coloring, and for $n \geq 5$, the line graph is not the only tree with a balance bipartite 2-coloring.

Hence, if we want to approximate $\langle \mu, f\rangle$, using a branching Markov chain does not improve the rate of convergence compared to a standard Markov chain.

4.2 Main theorem

4.2.1 Notations

Let $\mathbb{T}^\infty = \cup_{n\in\mathbb{N}}(\mathbb{N}^*)^n$ denote the Ulam-Harris-Neveu tree, and denote by ∂ its root, that is the empty word.

Let $u \in \mathbb{T}^\infty$ be a vertex. If u is distinct from the root, we denote by $\mathrm{p}(u)$ its parent vertex. We denote by $h(u)$ its height, i.e. the number of edges separating u from the root ∂. (The height of the root ∂ is zero.) For two vertices $u, v \in \mathbb{T}^\infty$, we denote by $u \wedge v$ the latest common ancestor of u and v, and by $d(u, v)$ the graph-distance between u and v in $\mathbb{T}^\infty$, that is $d(u, v) = h(u) + h(v) - 2h(u \wedge v)$.

Let $X = (X_u, u \in \mathbb{T}^\infty)$ be a stochastic process with values in a metric space $\mathcal{X}$.

Definition 4.2.1 (Markov process)**.** *The stochastic process X is called a* (branching) Markov process *with transition kernel Q and initial distribution ν if for any finite subtree $T \subset \mathbb{T}^\infty$ with $\partial \in T$, we have:*

$$\mathbb{P}(X_u \in \mathrm{d}x_u, u \in T) = \nu(\mathrm{d}x_\partial)\, \Pi_{u\in T\setminus\{\partial\}} Q(x_{\mathrm{p}(u)}; \mathrm{d}x_u).$$

We denote by νQ^n the distribution of a vertex in the n-th generation. For a (Borel) function f, define the function $Qf : x \in \mathcal{X} \mapsto \int f(y)\, Q(x; \mathrm{d}y)$. For a measure μ and a Borel function f on $\mathcal{X}$, we denote $\mu f = \langle \mu, f\rangle = \int_{\mathcal{X}} f \,\mathrm{d}\mu$. Through the rest of this section, we fix ν and Q.

4.2.2 Statement of the main result

Firstly, we need some assumptions on the Borel function f with which we perform the empirical averages.

Assumption 3 (Boundedness and convergence)**.** *Let f be a Borel function on $\mathcal{X}$ such that:*

(i) $\sup_{n\in\mathbb{N}} \nu Q^n(f^2) < \infty$,

(ii) there exists a constant $c_f \in \mathbb{R}$ such that $\lim_{n\to\infty} \nu Q^k((Q^n f - c_f)^2) = 0$ for all $k \in \mathbb{N}$.

Note that if f satisfies Assumption 3, then so does $f - c$ for any $c \in \mathbb{R}$, thus we may assume that $c_f = 0$ when necessary. Also note, using Cauchy-Schwarz and Jensen's inequalities, that Assumption 3-(i) implies that $Q^n f$, $Q^n f^2$, and $Q^k(Q^n f \times Q^m f)$ (with $n, m, k \in \mathbb{N}$) are well-defined and finite ν-almost everywhere and are ν-integrable.

Remark that when Q is ergodic, then Assumption 3 is satisfied by any continuous bounded function f on $\mathcal{X}$, and we get $c_f = \langle \mu, f \rangle$, where μ is the unique invariant measure of Q. Also remark that if F is a subspace of Borel functions on $\mathcal{X}$ that satisfy Assumptions (i)-(vi) in [Guy07], then any function $f \in F$ satisfies Assumption 3 with $c_f = \langle \mu, f \rangle$.

For a finite subset $A \subset \mathbb{T}^\infty$ and a Borel function f, we define the empirical sum $M_A(f) = \sum_{u \in A} f(X_u)$ and the empirical average:

$$\bar{M}_A(f) = |A|^{-1} \sum_{u \in A} f(X_u),$$

where $|A|$ is the cardinal of the set A. Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of finite subsets of $\mathbb{T}^\infty$ on which we perform the empirical averages in the ergodic theorem on trees with arbitrary shape. Remind the geometrical Assumptions 1 and 2.

Contrary to the case of the binary tree considered in [Guy07], Assumption 2 is not always satisfied for a sequence $(A_n)_{n \in \mathbb{N}}$ of finite subsets of $\mathbb{T}^\infty$ with arbitrary shape (e.g. in the case of the line graph, i.e. the Markov chain). Thus, to prove the ergodic theorem for branching Markov chains, as an alternative to the ancestral Assumption 2, we also consider the following conditions on the ergodicity of the transition kernel Q.

Assumption 4 (Stronger ergodicity)**.** *Assume that any of the following conditions holds:*

(i) $\nu = \mu$ *is an invariant measure of* Q*.*

(ii) There is convergence in total variation $\lim_{n \to \infty} \|\nu Q^n - \mu\|_{TV} = 0$ *to some invariant measure* μ *(for* Q*), the function* f *is bounded, and we have* $\lim_{n \to \infty} \mu(Q^n f - c_f)^2 = 0$*, where* c_f *is the same constant as in Assumption 3-(ii).*

(iii) The transition kernel Q *satisfies a uniformly ergodic assumption (with* μ *as its unique invariant measure): there exists a non-negative Borel function* g *on* $\mathcal{X}$ *with* $\sup_{k \in \mathbb{N}} \nu Q^k g^2 < \infty$ *and a sequence of positive numbers* $(a_n)_{n \in \mathbb{N}}$ *that converges to zero, such that for all* $n \in \mathbb{N}$*, we have* $|Q^n f - \langle \mu, f \rangle| \leq a_n \, g$*.*

Remark (using dominated convergence with domination by g) that Assumption 4-(iii) implies Assumption 3-(ii) with $c_f = \langle \mu, f \rangle$. Also remark that when Assumption 3 holds and either Assumption 4-(i) or 4-(ii) holds, then we have $c_f = \langle \mu, f \rangle$ in Assumption 3 (indeed, using Jensen's inequality, we have that $(\langle \mu, f \rangle - c_f)^2 = \limsup_{n \to \infty} (\mu Q^n f - c_f)^2 \leq \limsup_{n \to \infty} \mu (Q^n f - c_f)^2 = 0$).

We can now formulate the ergodic theorem for branching Markov processes on trees with arbitrary shape.

Theorem 4.2.2 (Ergodic theorem for Markov processes on trees with arbitrary shape)**.** *Let* $(A_n)_{n \in \mathbb{N}}$ *be a sequence of finite subsets of* $\mathbb{T}^\infty$ *that satisfies Assumption 1. Let* X *be a branching Markov process on* $\mathbb{T}^\infty$ *with transition kernel* Q *and initial distribution* ν*. Let* f *be a Borel function on* $\mathcal{X}$ *that satisfies Assumption 3. Furthermore assume that either Assumption 2 or 4 holds. Then, we have:*

$$\bar{M}_{A_n}(f) = |A_n|^{-1} \sum_{u \in A_n} f(X_u) \overset{L^2(\nu)}{\underset{n \to \infty}{\longrightarrow}} c_f.$$

In particular, when the transition kernel Q is ergodic and f is a continuous bounded function, then remind that Assumption 3 is satisfied and $c_f = \langle \mu, f \rangle$, where μ is the unique invariant measure of Q. If furthermore Q is uniformly ergodic, then Assumption 4-(iii) holds. Hence, Theorem 4.2.2 implies Theorem 4.1.2.

Proof. Up to replacing f by $f - c_f$, assume that $c_f = 0$. For all $n \in \mathbb{N}$, we have:

$$\mathbb{E}\big[\bar{M}_{A_n}(f)^2\big] = |A_n|^{-2} \sum_{u,v \in A_n} \mathbb{E}[f(X_u) f(X_v)]. \tag{4.1}$$

Remark that for $u, v \in \mathbb{T}^\infty$, we have:

$$\mathbb{E}[f(X_u) f(X_v)] = \nu Q^{h(u \wedge v)} \left(Q^{d(u \wedge v, u)} f \times Q^{d(u \wedge v, v)} f \right). \tag{4.2}$$

Set $C = \sup_{n\in\mathbb{N}} \nu Q^n f^2 < \infty$ which is finite by Assumption 3-(i). Hence, using Cauchy-Schwarz and Jensen's inequalities, for all $k, \ell, m \in \mathbb{N}$, we have:

$$\begin{aligned}|\nu Q^k(Q^\ell f \times Q^m f)| &\leq \left(\nu Q^k(Q^{\min(\ell,m)} f)^2 \times \nu Q^k(Q^{\max(\ell,m)} f)^2\right)^{1/2} \\ &\leq \left(\nu Q^{k+\min(\ell,m)} f^2 \times \nu Q^k(Q^{\max(\ell,m)} f)^2\right)^{1/2} \\ &\leq \sqrt{C} \times \sqrt{\nu Q^k(Q^{\max(\ell,m)} f)^2}. \end{aligned} \tag{4.3}$$

Define the distance $\tilde{d}$ on $\mathbb{T}^\infty$ as:

$$\tilde{d}(u,v) = \max(d(u, u\wedge v), d(v, u\wedge v)) = \max(h(u), h(v)) - h(u\wedge v).$$

Remark that we have $d/2 \leq \tilde{d} \leq d$, thus Assumption 1 is equivalent to: for all $k \in \mathbb{N}$, we have $\lim_{n\to\infty} \mathbb{P}(\tilde{d}(U_n, V_n) \leq k) = 0$. Then, as a consequence of (4.1), (4.2) and (4.3), we get:

$$\begin{aligned}\mathbb{E}\left[\bar{M}_{A_n}(f)^2\right] &\leq \sqrt{C} \times |A_n|^{-2} \sum_{u,v\in A_n} \left(\nu Q^{h(u\wedge v)} \left(Q^{\tilde{d}(u,v)} f\right)^2\right)^{1/2} \\ &\leq \sqrt{C} \times \left(|A_n|^{-2} \sum_{u,v\in A_n} \nu Q^{h(u\wedge v)} \left(Q^{\tilde{d}(u,v)} f\right)^2\right)^{1/2}, \end{aligned}$$

where we used Jensen's inequality in the last inequality. Hence, to conclude the proof it is enough to prove that the following holds:

$$\mathbb{E}\left[\nu Q^{h(U_n\wedge V_n)} \left(Q^{\tilde{d}(U_n,V_n)} f\right)^2\right] = |A_n|^{-2} \sum_{u,v\in A_n} \nu Q^{h(u\wedge v)} \left(Q^{\tilde{d}(u,v)} f\right)^2 \underset{n\to\infty}{\longrightarrow} 0. \tag{4.4}$$

As Assumption 2 (resp. Assumption 4) holds, using Lemma 4.2.3 (resp. Lemma 4.2.4) below, we get that (4.4) holds, which concludes the proof. □

Remind that from Assumption 3, we know that $\lim_{n\to\infty} \nu Q^k(Q^n f)^2 = 0$ for all $k \in \mathbb{N}$, and that (4.4) adds some uniformity in k. In the next lemma, we check that the ancestral Assumption 2, which allows for small values $h(U_n \wedge V_n)$ with high probability, implies (4.4).

Lemma 4.2.3. *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite subsets of $\mathbb{T}^\infty$ that satisfies Assumption 1. Let f be a function on $\mathcal{X}$ that satisfies Assumption 3. Then, Assumption 2 implies* (4.4).

Proof. Without loss of generality, assume that $c_f = 0$. Set $C = \sup_{j\in\mathbb{N}} \nu Q^j f^2 < \infty$ which is finite by Assumption 3-(i). Thus, for $k, m \in \mathbb{N}$, using Jensen's inequality, we get:

$$\nu Q^k(Q^m f)^2 \leq \nu Q^{k+m} f^2 \leq C < \infty. \tag{4.5}$$

Let $\varepsilon > 0$. Using Assumption 2, there exists $K \in \mathbb{N}$ such that for $n \in \mathbb{N}$ large enough, we have $\mathbb{P}(h(U_n \wedge V_n) > K) < \varepsilon$. Using Assumption 3-(ii), let $M \in \mathbb{N}$ be such that for all $m \geq M$ and for all $k \leq K$, we have $\nu Q^k(Q^m f)^2 < \varepsilon$. Using Assumption 1, for n large enough we have $\mathbb{P}(\tilde{d}(U_n, V_n) < M) < \varepsilon$. Hence, using (4.5), for n large enough we get:

$$\mathbb{E}\left[\nu Q^{h(U_n\wedge V_n)} \left(Q^{\tilde{d}(U_n,V_n)} f\right)^2\right] \leq 2C\varepsilon + \max_{k\leq K} \sup_{m\geq M} \nu Q^k(Q^m f)^2 < (1+2C)\varepsilon.$$

This being true for all $\varepsilon > 0$, we get that (4.4) holds, which concludes the proof. □

The following lemma states that the stronger ergodic Assumption 4 implies (4.4); its proof is similar to the proof of Lemma 4.2.3 and is left to the reader.

Lemma 4.2.4. *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite subsets of $\mathbb{T}^\infty$ that satisfies Assumption 1. Let f be a function on $\mathcal{X}$ that satisfies Assumption 3. Then, Assumption 4 implies* (4.4).

4.3 Examples satisfying Assumptions 1 and 2

We now give common examples of trees for which Assumptions 1 and 2 are satisfied.

We denote by T an arbitrary infinite tree rooted at some vertex ∂. In this section, all the trees we consider are locally-finite (i.e. all nodes have finite degree). For $n \in \mathbb{N}$, we denote by G_n the n-th generation of T, that is the set of vertices at distance n from the root, and we denote by $T_n = \cup_{k=0}^{n} G_k$ the tree up to generation n. For a vertex $u \in T$, we denote by $T(u)$ the subtree of T rooted at u and composed of all descendants of u.

4.3.1 Some simple deterministic trees

Firstly, Assumption 1 is always satisfied on trees with bounded vertex degrees, in particular for Cayley and Bethe trees (that is trees where each non-leaf vertex has out-degree D for some $D \geq 1$; except for the root of a Bethe tree which has out-degree $D+1$).

Lemma 4.3.1 (Bounded degree trees)**.** *Let $D \geq 2$, and let T be the infinite rooted complete D-ary tree (that is the tree where each vertex has D children). Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite (non-empty) subsets of T such that $\lim_{n\to\infty} |A_n| = \infty$. Then, the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 1.*

Proof. Let $k \in \mathbb{N}$. For every vertex $u \in T$, the ball $B_T(u,k)$ of radius k and center u has cardinal upper bounded by $c_k = \sum_{j=0}^{k} (D+1)^j$. Hence, we have:

$$\frac{1}{|A_n|^2} \sum_{u,v\in A_n} \mathbb{1}_{\{d(u,v)\leq k\}} = \frac{1}{|A_n|^2} \sum_{u\in A_n} |B_T(u,k)| \leq \frac{1}{|A_n|} c_k.$$

This implies that Assumption 1 is satisfied. □

The following counter-example shows that Assumption 2 is not always satisfied on a bounded degree tree.

Example 4.3.2. Let T be the infinite tree where each vertex has out-degree $D \geq 2$. Set $A_n = T(u_n) \cap G_{2n}$ the n-th descendants of u_n, where $u_n = 1 \cdots 1$ (n times) is the left-most n-th descendant of the root. Then, we have $|A_n| = D^n \to_{n\to\infty} \infty$, and by Lemma 4.3.1, the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 1. However, we have $\mathbb{P}(h(U_n \wedge V_n) \geq n) = 1$ for all $n \in \mathbb{N}$, which implies that the sequence $(A_n)_{n\in\mathbb{N}}$ does not satisfy Assumption 2.

When with high probability, the vertices in A_n are far from the root (e.g. when $A_n = G_n$) and have their common ancestor close from the root, then Assumption 1 is satisfied.

Lemma 4.3.3 (A_n far from the root)**.** *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite subsets of $\mathbb{T}^\infty$. For every $n \in \mathbb{N}$, let U_n and V_n be independent and uniformly distributed over A_n. Assume that for every $k \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{P}(h(U_n) \leq k) = 0$ and that Assumption 2 holds. Then, the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 1.*

Proof. Remind that $d(U_n, V_n) = h(U_n) + h(V_n) - 2h(U_n \wedge V_n)$. Thus, for $k \in \mathbb{N}$ we have:

$$\begin{aligned}\mathbb{P}(d(U_n,V_n) \leq 2k) &\leq 2\,\mathbb{P}(h(U_n) - h(U_n \wedge V_n) \leq k) \\ &\leq 2\,\mathbb{P}(h(U_n \wedge V_n) > k) + 2\,\mathbb{P}(h(U_n) \leq 2k),\end{aligned}$$

where both terms in the upper bound go to zero as $n \to \infty$. Hence, the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 1. □

We say a tree T is spherically symmetric (sometimes also called a generalized Bethe tree) if for all $n \in \mathbb{N}$, every vertex of height n in T has the the same out-degree D_n. When we choose $A_n = G_n$ the n-th generation for all $n \in \mathbb{N}$, Assumptions 1 and 2 are always satisfied on spherically symmetric trees.

Lemma 4.3.4 (Spherically symmetric trees)**.** *Let T be an infinite spherically symmetric tree such that $\lim_{n\to\infty} |G_n| = \infty$. Then, the sequence $(G_n)_{n\in\mathbb{N}}$ satisfies Assumptions 1 and 2.*

Proof. Thanks to Lemma 4.3.3, we only need to prove that Assumption 2 is true. For all $n \in \mathbb{N}$, denote by D_n the out-degree for all vertices of height n. As $\lim_{n\to\infty} |G_n| = \infty$, we have that $D_n > 1$ for infinitely many values of n. Let U_n and V_n be independent random variables uniformly distributed over G_n. Using the Ulam-Harris-Neveu tree notation, write $U_n = U_{(1)} \cdots U_{(n)}$ and $V_n = V_{(1)} \cdots V_{(n)}$, where the random variables $U_{(1)}, \ldots, U_{(n)}$, $V_{(1)}, \ldots, V_{(n)}$ are independent with $U_{(i)}$ and $V_{(i)}$ uniformly distributed over the set $\{1, \ldots, D_{i-1}\}$. Thus, for all $k \in \mathbb{N}$ and $n \geq k$, as $d(U_n, V_n) = 2n - 2h(U_n \wedge V_n)$, we have:

$$\mathbb{P}\Big(h(U_n \wedge V_n) \geq k\Big) = \mathbb{P}\Big(U_{(i)} = V_{(i)}, \forall i \in \{1, \ldots, k\}\Big) = \prod_{i=0}^{k-1} \frac{1}{D_i},$$

where the right hand side goes to 0 as $k \to \infty$. This implies that Assumption 2 is satisfied, and thus concludes the proof. □

4.3.2 Super-critical Bienaymé-Galton-Watson trees

To apply the ergodic Theorem 4.2.2 to a random sequence $(A_n)_{n\in\mathbb{N}}$ of subsets of $\mathbb{T}^\infty$ (independent of the Markov process indexed by $\mathbb{T}^\infty$), we need to verify that a.s. the sequence $(A_n)_{n\in\mathbb{N}}$ satisfy Assumption 1 (and possibly Assumption 2). Note that in this case, Assumptions 1 and 2 should be considered conditionally on A_n, in particular the random vertices U_n and V_n are independent and uniformly distributed over A_n conditionally on A_n.

In this subsection, we consider the case where T is a super-critical Bienaymé-Galton-Watson tree whose offspring distribution $\mathcal{P}$ on $\mathbb{N}$ has finite mean $m > 1$ and finite second moment, and is conditioned on non-extinction. The following lemma states that a.s. both the sequences $(G_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$ satisfy Assumptions 1 and 2.

Lemma 4.3.5 (Super-critical Bienaymé-Galton-Watson trees)**.** *Let T be a super-critical Bienaymé-Galton-Watson tree whose offspring distribution has mean $m > 1$ and finite second moment and is conditioned on non-extinction.*

(i) Let $(\ell_n)_{n\in\mathbb{N}}$ be a sequence of integers such that $0 \leq \ell_n \leq n$ for all $n \in \mathbb{N}$. For every $n \in \mathbb{N}$, let $A_n = \cup_{k=(n-\ell_n)_+}^{n} G_k$ be the subset composed of the last $\ell_n + 1$ generations of T_n. Then, the sequence $(A_n)_{n\in\mathbb{N}}$ a.s. satisfies Assumptions 1 and 2.

(ii) The sequences $(G_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$ a.s. satisfies Assumptions 1 and 2.

In the case of a critical Bienaymé-Galton-Watson tree, it is not possible to condition on non-extinction, but taking $A_n = G_n$ and conditioning on non-extinction at time n, [Ath12a, Theorem 2.1] suggests that Assumption 1 should be satisfied but not Assumption 2.

The proof of Lemma 4.3.5-(i) relies on the case where the sequence $(\ell_n)_{n\in\mathbb{N}}$ is bounded. In this case, we prove a stronger version of [Ath12b, Theorem 2] where individuals can have zero child and the two random individuals are taken from the last $\ell_n + 1$ generations instead of only the last generation.

Before proving Lemma 4.3.5, we need the following lemma stating that the last generations carry most of the weight of T_n.

Lemma 4.3.6 (Last generations carry all the weight)**.** *Let T be a super-critical Bienaymé-Galton-Watson tree whose offspring distribution has mean $m > 1$, and is conditioned on non-extinction. We have:*

$$\forall \ell \in \mathbb{N}, \qquad a.s. \qquad \lim_{n\to\infty} \frac{|\cup_{k=(n-\ell)_+}^{n} G_k|}{|T_n|} = 1 - \frac{1}{m^{\ell+1}}. \tag{4.6}$$

Proof. For all $n \in \mathbb{N}$, we write $Z_n = |G_n|$. Using [AN72, Theorem I.C.12.3], as T is a super-critical Bienaymé-Galton-Watson tree whose offspring distribution $\mathcal{P}$ has a finite mean $m > 1$, and is conditioned on non-extinction, we know that there exists a sequence of positive constants $(C_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} C_n = \infty$ and $\lim_{n\to\infty} C_{n+1}/C_n = m$, and such that a.s. $\lim_{n\to\infty} C_n^{-1} Z_n = W$ where W is a random variable on $\mathbb{R}_+$, and where the event $\{W = 0\} \supset \{\exists n,\, Z_n = 0\}$ has zero probability. Hence, we get that a.s. $C_n^{-1} |\cup_{k=0}^{\ell} G_{n-k}| = C_n^{-1} \sum_{k=0}^{\ell} Z_{n-k}$ converges to $W \sum_{k=0}^{\ell} m^{-k}$ as n goes to infinity. Writing $S_n = \sum_{k=0}^{n} C_k$, as we know that $C_n \sim mC_{n-1}$, we have that $S_{n+1} = S_n + C_{n+1} = mS_n + o(S_n)$, and thus we get that $C_n \sim C_{n+1}/m \sim \frac{m-1}{m} S_n$. As $\lim_{n\to\infty} C_n^{-1} \sum_{k=0}^{n} C_k = \frac{m}{m-1}$, using Stolz-Cesàro theorem, we get that a.s. $C_n^{-1} |T_n| = C_n^{-1} \sum_{k=0}^{n} Z_k$ converges to $W \frac{m}{m-1}$ as n goes to infinity. Consequently, we get (4.6). □

Proof of Lemma 4.3.5. Point (ii) is an immediate consequence of Point (i) taking $\ell_n = 0$ or n.

We now prove Point (i). Using Lemma 4.3.3 with (4.6), it is enough to prove that the sequence $(A_n)_{n\in\mathbb{N}}$ a.s. satisfies Assumption 2. Let U_n and V_n be independent and uniformly distributed over A_n. For $k \in \mathbb{N}$, remark that $h(U_n \wedge V_n) \geq k$ is equivalent to the existence of $u \in G_k$ such that $U_n, V_n \in T(u)$. Note that Assumption 2 for $(A_n)_{n\in\mathbb{N}}$ in the random tree T can be reformulated as:

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(h(U_n \wedge V_n) \geq k \;\Big|\; T\Big) = \lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\exists u \in G_k,\, U_n, V_n \in T(u) \;\Big|\; T\Big) = 0. \tag{4.7}$$

We divide the rest of the proof into two cases: we first consider the case when the sequence $(\ell_n)_{n\in\mathbb{N}}$ is bounded, and then the general case.

Case 1: the sequence $(\ell_n)_{n\in\mathbb{N}}$ is bounded. Following [AN72, Section I.D.12] (and using the survivor/extinct vertices denomination from [AD20, Section 2.2.3]), the vertices of the super-critical Bienaymé-Galton-Watson tree T conditioned on non-extinction can be partitioned into two categories: *survivor* vertices, whose descendants do not suffer extinction, and *extinct* vertices, whose descendants eventually become extinct. The root of T is a survivor vertex due to the conditioning on non-extinction. Denote by T^s the subtree of T composed of survivor vertices. From [AN72, Theorem I.D.12.1], the tree T^s is distributed as a super-critical Bienaymé-Galton-Watson tree whose offspring distribution $\mathcal{P}_0$ on $\mathbb{N}^*$ (hence no extinction) has the same mean $m > 1$ as $\mathcal{P}$ and has finite second moment. Also, from [AN72, Theorem I.D.12.3], we know that if $u \in T$ is an extinct vertex, then its descendant subtree $T(u)$ is distributed as a sub-critical Bienaymé-Galton-Watson tree whose offspring distribution $\mathcal{P}_1$ is explicitly known and has finite second moment.

For all $n > \ell$ and $u \in T$, define $Z_{n,\ell,u} = |(T_n \setminus T_{n-\ell-1}) \cap T(u)|$. Fix some $k \in \mathbb{N}$. Conditionally on T, the unique ancestor of U_n in G_k is $u \in G_k$ with probability $Z_{n,\ell_n,u} / \sum_{v\in G_k} Z_{n,\ell_n,v}$. Define $Z_n = |G_n|$, $G_n^s = G_n \cap T^s$ and $Z_n^s = |G_n^s|$. Then, we have:

$$\mathbb{P}\Big(\exists u \in G_k,\, U_n, V_n \in T(u) \;\Big|\; T\Big) = \frac{\sum_{u\in G_k} Z_{n,\ell_n,u}^2}{\Big(\sum_{u\in G_k} Z_{n,\ell_n,u}\Big)^2} = \frac{\sum_{u\in G_k^s} Z_{n,\ell_n,u}^2}{\Big(\sum_{u\in G_k^s} Z_{n,\ell_n,u}\Big)^2},$$

where the last equality holds a.s. for n large enough as for any extinct vertex $u \in G_k \setminus G_k^s$, we know that a.s. $Z_{n,\ell_n,u} = 0$ for n large enough (remind that the sequence $(\ell_n)_{n\in\mathbb{N}}$ is bounded).

From [AN72, Theorems I.B.6.1 and I.B.6.2], as the offspring distribution $\mathcal{P}$ has finite second moment, we know that $\lim_{n\to\infty} m^{-n} Z_n = W$ a.s. and in L^2 where W is a random variable in $\mathbb{R}_+^*$ with finite second moment. (From [AN72, Theorem I.B.6.2-(iii)], we know that on the non-extinction event, that is when the root is a survivor vertex, a.s. W is positive.) Then, remark that for all survivor vertices $u \in G_k^s$, we have that a.s. $\lim_{n\to\infty} m^{-(n-k)} Z_{n,0,u} = W_u$, where conditionally on G_k^s the random variables $(W_u)_{u\in G_k^s}$ are independent and distributed as W (remind that the root of T is also a survivor vertex). Thus, we get that a.s. for all $\ell \in \mathbb{N}$ and all $u \in G_k^s$, $\lim_{n\to\infty} m^{-(n-k)} Z_{n,\ell,u} = (\sum_{j=0}^{\ell} m^{-j}) W_u$. As the sequence $(\ell_n)_{n\in\mathbb{N}}$ is bounded, this implies that a.s. for all $u \in G_k^s$, $\lim_{n\to\infty} m^{-(n-k)} (\sum_{j=0}^{\ell_n} m^{-j})^{-1} Z_{n,\ell_n,u} = W_u$. Hence, we get:

$$\text{a.s.} \qquad \frac{\sum_{u\in G_k^s} Z_{n,\ell,u}^2}{\Big(\sum_{u\in G_k^s} Z_{n,\ell,u}\Big)^2} \underset{n\to\infty}{\longrightarrow} \frac{\sum_{u\in G_k^s} W_u^2}{\Big(\sum_{u\in G_k^s} W_u\Big)^2}.$$

Combining what we got so far, we get:

$$\text{a.s.} \qquad \lim_{n\to\infty} \mathbb{P}\Big(h(U_n \wedge V_n) \geq k \;\Big|\; T\Big) = \frac{\sum_{u\in G_k^s} W_u^2}{\Big(\sum_{u\in G_k^s} W_u\Big)^2}. \tag{4.8}$$

As the left hand side in (4.8) is a non-increasing function of k, then so is the right hand side in (4.8).

Thus, taking the limit when $k \to \infty$ and then taking the expectation, we get:

$$\mathbb{E}\left[\lim_{k\to\infty}\lim_{n\to\infty}\mathbb{P}\Big(h(U_n \wedge V_n) \geq k \;\Big|\; T\Big)\right] = \mathbb{E}\left[\lim_{k\to\infty}\frac{\sum_{u\in G_k^s} W_u^2}{\Big(\sum_{u\in G_k^s} W_u\Big)^2}\right]$$
$$= \lim_{k\to\infty}\mathbb{E}\left[\frac{\sum_{u\in G_k^s} W_u^2}{\Big(\sum_{u\in G_k^s} W_u\Big)^2}\right], \tag{4.9}$$

where we used the dominated convergence theorem in the last inequality.

Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables distributed as W. For all $n \in \mathbb{N}$, define the random variable $R(n) = \sum_{i=1}^n W_i^2/(\sum_{i=1}^n W_i)^2$. Using the strong law of large numbers, we get that the numerator of $R(n)$ is a.s. equivalent to $n\mathbb{E}[W^2]$, and the denominator of $R(n)$ is a.s. equivalent to $n^2\mathbb{E}[W]^2$. This implies that $R(n)$ is a.s. equivalent to $n^{-1}\mathbb{E}[W^2]/\mathbb{E}[W]^2$, and thus a.s. $\lim_{n\to\infty} R(n) = 0$. Remark that the expectation in the last line of (4.9) can be written as $\mathbb{E}[R(Z_k^s)]$. As we know that a.s. $\lim_{k\to\infty} Z_k^s = \infty$, we get that a.s. $\lim_{k\to\infty} R(Z_k^s) = 0$. Hence, using the dominated convergence theorem (with domination by 1), we get that $\lim_{n\to\infty}\mathbb{E}[R(Z_k^s)] = 0$, which implies that the left hand side in (4.9) is also null. As a consequence, we get that a.s. (4.7) holds, which implies that a.s. the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 2. This concludes the proof of the first case.

Case 2: general case. Let $\varepsilon > 0$, and let $\ell \in \mathbb{N}$ be such that $m^{-\ell-1} < \varepsilon$. Then, using Lemma 4.3.6, we get that a.s. $\mathbb{P}(h(U_n) < n - \ell \,|\, T) < \varepsilon$ for n large enough. Thus, a.s. for n large enough, for all $k \in \mathbb{N}$, we have:

$$\begin{aligned}\mathbb{P}(h(U_n \wedge V_n) \geq k \,|\, T) &\leq \mathbb{P}(h(U_n) < n-\ell \text{ or } h(V_n) < n-\ell \,|\, T)\\ &\quad + \mathbb{P}(h(U_n) \geq n-\ell \text{ and } h(V_n) \geq n-\ell \,|\, T)\\ &\quad\quad \times \mathbb{P}(h(U_n \wedge V_n) \geq k \,|\, \min(h(U_n), h(V_n)) \geq n-\ell,\, T)\\ &\leq 2\varepsilon + \mathbb{P}(h(\bar{U}_n \wedge \bar{V}_n) \geq k \,|\, T),\end{aligned} \tag{4.10}$$

where $\bar{U}_n$ and $\bar{V}_n$ are independent and uniformly distributed over $\cup_{k=(n-\min(\ell_n,\ell))_+}^n G_k$. By the first case, we have that $\lim_{k\to\infty}\limsup_{n\to\infty}\mathbb{P}(h(\bar{U}_n \wedge \bar{V}_n) \geq k \,|\, T) = 0$. Thus, using (4.10), we get that $\lim_{k\to\infty}\limsup_{n\to\infty}\mathbb{P}(h(U_n \wedge V_n) \geq k \,|\, T) \leq 2\varepsilon$. This being true for all $\varepsilon > 0$, we get that a.s. the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies Assumption 2, which concludes the proof. □

4.4 Dependence of the variance on the shape of the tree

In this section, we briefly discuss the variance of the empirical average estimator $\bar{M}_A(f)$ (which estimates $\langle\mu, f\rangle$), that is $\mathbb{E}_\mu\big[|A|^{-1}M_A(f)^2\big]$ for some Borel function f on $\mathcal{X}$, and its dependence on the geometry of the averaging set $A \subset T$. We consider the case where the transition kernel Q induces a self-adjoint compact operator on $L^2(\mu)$, where μ is the unique invariant measure of Q (note that self-adjoint is equivalent to (μ, Q) being reversible). This is in particular the case when the state space $\mathcal{X}$ is finite and (μ, Q) is reversible, or when the operator induced by Q on $L^2(\mu)$ is a symmetric Hilbert-Schmidt operator.

We now prove Proposition 4.1.4, which is a non-asymptotic result that states that among subtrees $T \subset \mathbb{T}^\infty$ of a given finite size, the line graph tree (i.e. the Markov chain) is the one minimizing the variance of the empirical average estimator.

Proof of Proposition 4.1.4. Let $f \in L^2(\mu)$ be some function. As Q induces a self-adjoint compact operator on $L^2(\mu)$, using [Rud96, Theorem 12.29-(d) and Theorem 12.30] (remind that a self-adjoint operator is normal), the spectrum of Q is composed of a (at most) countable number of eigenvalues $(\alpha_k)_{k\in\mathbb{N}}$, and the function $f \in L^2(\mu)$ has a unique expansion $\sum_{k\in\mathbb{N}} f_k$ where we have $Qf_k = \alpha_k f_k$ for all $k \in \mathbb{N}$, and $\langle f_k, f_\ell\rangle_{L^2(\mu)} = 0$ for $k \neq \ell$. As Q is a Markov kernel, we have $\alpha_k \in [-1, 1]$ for all $k \in \mathbb{N}$ (indeed, using Jensen's inequality, we have $\alpha_k^2\langle\mu, f_k^2\rangle = \langle\mu, (Qf_k)^2\rangle \leq \langle\mu, Q(f_k^2)\rangle = \langle\mu, f_k^2\rangle$). Remark that for all $n, m \in \mathbb{N}$, we have:

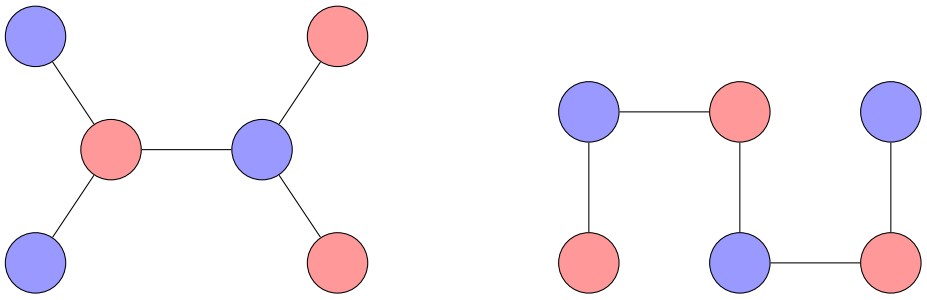
$$\langle\mu, Q^n f \times Q^m f\rangle = \sum_{i,j\in\mathbb{N}}\langle Q^n f_i, Q^m f_j\rangle_{L^2(\mu)} = \sum_{i\in\mathbb{N}}\alpha_i^{n+m}\langle\mu, f_i^2\rangle. \tag{4.11}$$

Using (4.2) with $\nu = \mu$, we get that $\mathbb{E}_\mu[f(X_u)f(X_v)] = \sum_{k\in\mathbb{N}} \alpha_k^{d(u,v)} \langle \mu, f_k^2 \rangle$. In particular, we get that $\mathbb{E}_\mu[M_A(f)^2] = \sum_{k\in\mathbb{N}} \mathbb{E}_\mu[M_A(f_k)^2]$, and thus it is sufficient to prove the result when $f = f_k$ for some $k \in \mathbb{N}$. (Remark that there exists $k \in \mathbb{N}$ such that $f_k \neq 0$ and $\alpha_k \notin \{-1, 0, 1\}$ if and only if $f \notin \mathrm{Ker}(Q) \oplus \mathrm{Ker}(Q - I) \oplus \mathrm{Ker}(Q + I)$.)

Hence, let $k \in \mathbb{N}$ be fixed, and in the rest of the proof, we will write $f = f_k$ and $\alpha = \alpha_k \in [-1, 1]$. Thus, the variance of the empirical average estimator can be written as:

$$|A|^{-1}\, \mathbb{E}_\mu[M_A(f)^2] = |A|^{-1} \langle \mu, f^2 \rangle\, H_A(\alpha), \quad \text{with} \quad H_A(\alpha) = \sum_{u,v\in A} \alpha^{d(u,v)},$$

which involves the Hosoya-Wiener polynomial $H_A(\alpha)$ of A. Note that variance minimization is equivalent to minimization of $H_A(\alpha)$. Also note that we may consider unrooted trees $A = T$, as the definition of $H_T(\alpha)$ is invariant by rerooting the tree. If $\alpha = 0$ or 1, then the value of the Hosoya-Wiener polynomial H_T depends only on the size of T, and thus is the same for every tree T of size n. If $\alpha = -1$, then we have $H_T(-1) = (|B| - |R|)^2$ where $T = B \cup R$ is the bipartite partitioning of vertices in T, that is the value of $H_T(-1)$ is the imbalance between the two bipartite classes of vertices in T (2-coloring of T), and its minimal value is 0 (resp. 1) when n is even (resp. odd), and is achieved by the line graph (but not uniquely for $n \geq 5$, e.g. the double-cherry graph in Figure 4.1a is also a minimizer for $n = 6$). We now assume that $\alpha \in (-1, 1) \setminus \{0\}$, and we are going to prove that in this case the line graph is the unique minimizer of $H_T(\alpha)$. We divide the rest of the proof in two cases depending on the sign of α.



(a) Double-cherry graph (b) Line graph

Figure 4.1 – Comparison of the double-cherry graph and the line graph ($n = 6$); both graphs have an exactly balanced bipartite 2-coloring, and thus satisfy $H_T(-1) = 0$.

Proof for $\alpha \in (0, 1)$. Let $n \in \mathbb{N}^*$. To prove that the line graph minimizes the function $T \mapsto H_T(\alpha)$ among trees of size n, Let u_1 be a leaf of T, and consider the tree T to be rooted at u_1. Let $\ell \in \mathbb{N}^*$ be the first generation with size larger than 1 (which exists as T is not the line graph), and for $i < \ell$, denote by u_{i+1} the only vertex in the i-th generation. Let v be one of the children of the vertex u_ℓ, and denote by T_v the connected component of $T \setminus \{u_\ell\}$ that contains v. We define the tree T' by removing the edge (u_ℓ, v) and by adding the edge (u_1, v) (see Figure 4.2b, where for clarity the other children of u_ℓ have been labeled $v_2, \cdots, v_k$). We now compare the distances $d_{T'}(u, w)$ and $d_T(u, w)$ for $u, w \in T$: if u and w are both in $T \setminus T_v$ or both in T_v, then $d_{T'}(u, w) = d_T(u, w)$; if $w \in T_v$, then $d_{T'}(u_i, w) = d_T(u_{\ell-i+1}, w)$ for $1 \leq i \leq \ell$; and if $u \in T \setminus (T_v \cup \{u_1, \cdots, u_\ell\})$ and $w \in T_v$, then we have:

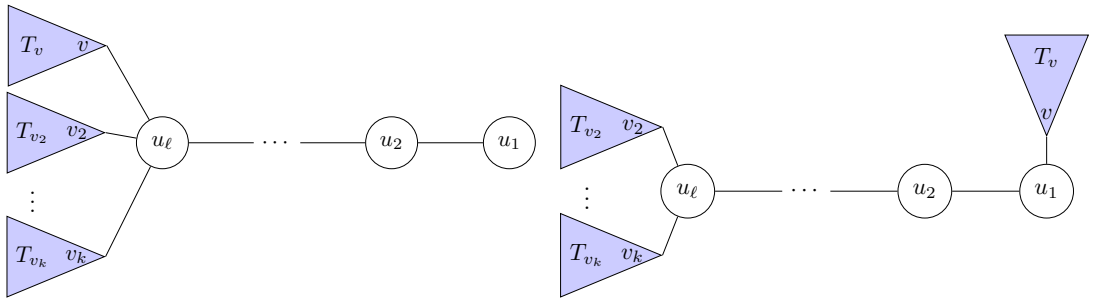
$$d_{T'}(u, w) = d_{T'}(u, u_\ell) + d_{T'}(u_\ell, v) + d_{T'}(w, v) = d_T(u, u_\ell) + \ell + d_T(w, v) > d_T(u, w).$$

Hence, we get:

$$H_T(\alpha) - H_{T'}(\alpha) = 2 \sum_{u\in T_v, w\in T\setminus(T_v\cup\{u_1,\cdots,u_\ell\})} \alpha^{d_T(u,v)} - \alpha^{d_{T'}(u,v)} > 0,$$

and thus T does not minimize $H_T(\alpha)$.

Proof for $\alpha \in (-1, 0)$. We prove the statement by recurrence on the size n of the tree. For $n \in \{1, 2, 3\}$, there exists only one (unrooted) tree of size n, hence the statement is trivial.



(a) The unrooted tree T before modification (b) The unrooted tree T' after modification

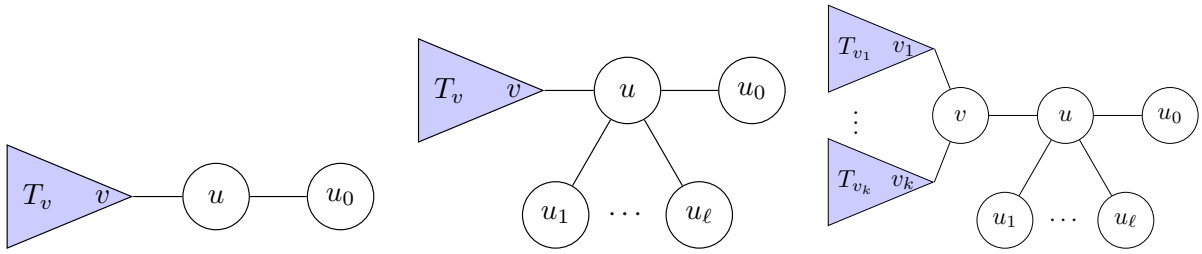
Figure 4.2 – The unrooted trees T and T' for $\alpha \in (0,1)$

Before proving that this property is hereditary, first remark that if T is a tree of size $n \geq 2$ and $u, v \in T$ are two vertices connected by an edge, and we let T_u and T_v be the two rooted subtrees of T obtained by removing the edge (u,v) and rooted at u and v respectively, then we get:

$$\begin{aligned} H_T(\alpha) &= H_{T_u}(\alpha) + H_{T_v}(\alpha) + 2 \sum_{u' \in T_u} \sum_{v' \in T_v} \alpha^{d_T(u',v')} \\ &= H_{T_u}(\alpha) + H_{T_v}(\alpha) + 2\alpha\, C_{T_u}(\alpha) C_{T_v}(\alpha), \end{aligned} \tag{4.12}$$

where for convenience we write $C_{T_v}(\alpha) = \left(\sum_{v' \in T_v} \alpha^{d_T(v,v')}\right)$, and similarly for T_u. We also remark that when T_v is the line graph with k vertices $\{u_1, \cdots, u_k\}$ rooted at the vertex $v = u_j$, then we have $C_{T_v}(\alpha) = \frac{1}{1-\alpha}(1 + \alpha - \alpha^j - \alpha^{k-j+1})$, which is maximal among rooted copies of $\{u_1, \cdots, u_k\}$ only for $j = 1$ or k (remind that $\alpha \in (-1, 0)$). We denote by L_k the line graph with k vertices rooted at one of its extremities.

Now, consider $n > 3$ and assume that for all $k < n$, the line graph is the unique minimizer of $T \mapsto H_T(\alpha)$ among trees of size k. Consider an unrooted tree T that is not the line graph. Let u be a leaf of the tree obtained by removing all the leaves of T (such vertex is sometimes called a non-protected vertex). In particular, the node u has $2 + \ell$ neighbors in T with $\ell \in \mathbb{N}$, which we denote by $u_0, \cdots, u_{\ell+1}$; and at most one of the neighbors of u is not a leaf, say $v = u_{\ell+1}$. (All the neighbors of u are leaves if and only if T is the star graph whose center vertex is u, in which case we still write $v = u_{\ell+1}$.) Denote by T_v the subtree $T \setminus \{u, u_0, \cdots, u_\ell\}$ of T rooted at v. We consider three cases.



(a) Case 1 (b) Case 2 (c) Case 3

Figure 4.3 – The unrooted tree T before modification

Case 1: $\ell = 0$ (see Figure 4.3a). As T is not the line graph and as T_v has size $n-2$, by induction hypothesis, we either have $H_{T_v}(\alpha) > H_{L_{n-2}}(\alpha)$, or T_v is a line graph rooted at a non-extremal vertex (and thus $C_{T_v}(\alpha) < C_{L_{n-2}}(\alpha)$). If $C_{T_v}(\alpha) < C_{L_{n-2}}(\alpha)$, using (4.12) around edge (v,u), as $\alpha(1+\alpha) < 0$ (remind that $\alpha \in (-1,0)$), we get:

$$\begin{aligned} H_T(\alpha) &= H_{T_v}(\alpha) + H_{L_2}(\alpha) + 2\alpha C_{T_v}(\alpha)(1+\alpha) \\ &> H_{L_{n-2}}(\alpha) + H_{L_2}(\alpha) + 2\alpha C_{L_{n-2}}(\alpha)(1+\alpha) \\ &= H_{L_n}(\alpha). \end{aligned}$$

If we have $C_{T_v}(\alpha) \geq C_{L_{n-2}}(\alpha)$ (thus we also have $H_{T_v}(\alpha) > H_{L_{n-2}}(\alpha)$), using (4.12) around edge (u, u_0), as $C_{L_{n-1}}(\alpha) = 1 + \alpha C_{L_{n-2}}(\alpha)$, we get:

$$\begin{aligned} H_T(\alpha) &= H_{T\setminus\{u_0\}}(\alpha) + H_{L_1}(\alpha) + 2\alpha(1 + \alpha C_{T_v}(\alpha)) \\ &> H_{L_{n-1}}(\alpha) + H_{L_1}(\alpha) + 2\alpha(1 + \alpha C_{L_{n-2}}(\alpha)) \\ &= H_{L_n}(\alpha). \end{aligned}$$

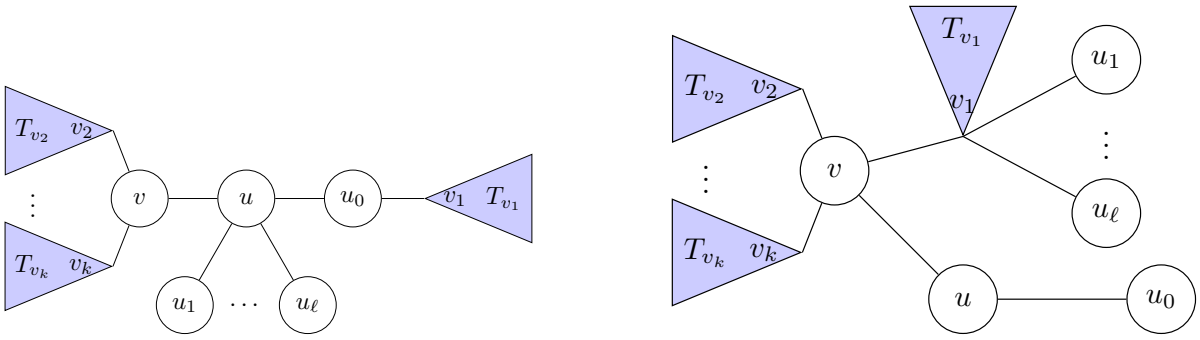
Thus, for all values of $C_{T_v}(\alpha)$, we get that $H_T(\alpha) > H_{L_n}(\alpha)$.

Case 2: $\ell \geq 1$ and $C_{T_v}(\alpha) > 0$. Denote by T_u the subtree of T composed of the $\ell + 2$ vertices $u, u_0, \cdots, u_\ell$ (see Figure 4.3b), and consider the tree T' obtained from T by replacing T_u by a copy of $L_{2+\ell}$. By induction hypothesis, we know that $H_{T_u}(\alpha) \geq H_{L_{2+\ell}}(\alpha)$. Using (4.12) around edge (v, u), as $1 + (\ell+1)\alpha < C_{L_{2+\ell}}(\alpha)$, we get:

$$\begin{aligned} H_T(\alpha) &= H_{T_v}(\alpha) + H_{T_u}(\alpha) + 2\alpha C_{T_v}(\alpha)(1 + (\ell+1)\alpha) \\ &> H_{T_v}(\alpha) + H_{L_{2+\ell}}(\alpha) + 2\alpha C_{T_v}(\alpha) C_{L_{2+\ell}}(\alpha) \\ &= H_{T'}(\alpha), \end{aligned}$$

and thus the tree T does not minimize the Hosoya-Wiener polynomial.

Case 3: $\ell \geq 1$ and $C_{T_v}(\alpha) \leq 0$. As $C_{T_v}(\alpha) \neq 1$, we know that v is not a leaf. Denote by $v_1, \cdots, v_k$ the neighbors of v other than u, and for each $i \in [\![1, k]\!]$, denote by T_{v_i} the subtree of T_v rooted at v_i (see Figure 4.3c). As $C_{T_v}(\alpha) = 1 + \alpha \sum_{i=1}^{k} C_{T_{v_i}}(\alpha) \leq 0$, we have that $\sum_{i=1}^{k} C_{T_{v_i}}(\alpha) \geq -1/\alpha > 0$, and thus there exists $i \in [\![1, k]\!]$ such that $C_{T_{v_i}}(\alpha) > 0$. Without loss of generality, we assume for simplicity that $i = 1$.



(a) Case 3a after moving T_{v_1}

(b) Case 3b after moving $u_1, \cdots, u_\ell$

Figure 4.4 – The unrooted tree T' in Case 3 after modification

Case 3a: $\ell \geq 1$ and $C_{T_v}(\alpha) \leq 0$ and $\sum_{i=2}^{k} C_{T_{v_i}}(\alpha) > 0$. Consider the tree T' obtained from T by replacing edge (v_1, v) by (v_1, u_0), that is grafting T_{v_1} on u_0 (see Figure 4.4a). Remark that going from T to T', the subtrees T_{v_1} and $T \setminus T_{v_1}$ do not change, the distance between v_1 and v_i for $i \in [\![2, k]\!]$ goes from 2 to 4, and that v_1 is still at distance 1 from a leaf of the star graph formed by vertices $\{v, u, u_0, \ldots, u_\ell\}$. Thus, as $\alpha \in (-1, 0)$, we have:

$$H_T(\alpha) - H_{T'}(\alpha) = 2C_{T_{v_1}}(\alpha) \left(\sum_{i=2}^{k} C_{T_{v_i}}(\alpha) \right) (\alpha^2 - \alpha^4) > 0,$$

and thus the tree T does not minimize the Hosoya-Wiener polynomial.

Case 3b: $\ell \geq 1$ and $C_{T_v}(\alpha) \leq 0$ and $\sum_{i=2}^{k} C_{T_{v_i}}(\alpha) \leq 0$. In particular, we get that $C_{T_{v_1}}(\alpha) \geq -1/\alpha \geq 1$. Consider the tree T' obtained from T by replacing edge (u_i, u) by (u_i, v_1) for all $i \in [\![1, \ell]\!]$, that is grafting leaves $u_1, \ldots, u_\ell$ to v_1 (see Figure 4.4b). Remark that going from T to T', the subtree $T \setminus \{u_1, \cdots, u_\ell\}$ does not change, and for for $i, j \in [\![1, \ell]\!]$, the distance between u_i and u_j (resp. v) does not change, and the distance between u_i and v_1 (resp. u) goes from 3 to 1 (resp. from 1 to 3). Thus, as

$\alpha \in (-1, 0)$, we have:

$$\begin{aligned} H_T(\alpha) - H_{T'}(\alpha) &= 2\ell C_{T_{v_1}}(\alpha)(\alpha^3 - \alpha) + 2\ell(1+\alpha)(\alpha - \alpha^3) \\ &\geq 2\ell(\alpha^3 - \alpha) + 2\ell(1+\alpha)(\alpha - \alpha^3) \\ &\geq 2\ell(\alpha^2 - \alpha^4) > 0, \end{aligned}$$

and thus the tree T does not minimize the Hosoya-Wiener polynomial. □

Chapitre 5

Asymptotic properties of the maximum likelihood estimator for Hidden Markov Models indexed by binary trees

The material for this chapter has been released in [Wei24a].

Abstract: *We consider hidden Markov models indexed by a binary tree where the hidden state space is a general metric space. We study the maximum likelihood estimator (MLE) of the model parameters based only on the observed variables. In both stationary and non-stationary regimes, we prove strong consistency and asymptotic normality of the MLE under standard assumptions. Those standard assumptions imply uniform exponential memorylessness properties of the initial distribution conditional on the observations. The proofs rely on ergodic theorems for Markov chain indexed by trees with neighborhood-dependent functions.*

Key words and phrases— Hidden Markov tree (HMT), hidden Markov model (HMM), branching process, maximum likelihood estimator (MLE), asymptotic normality, consistency, geometric ergodicity

2020 Mathematics Subject Classification— 62M05, 62F12, 60J80, 60J85

5.1 Introduction

In this article, we consider a generalization of the hidden Markov chain/model (HMM) where the process is indexed by a binary tree, which we call hidden Markov tree (HMT). The HMT is composed of a hidden process and an observed process. The hidden process is a branching Markov process, that is, a random process $X = (X_u, u \in T)$ with values in a metric space $\mathcal{X}$ indexed by a rooted tree T with the Markov property: sibling nodes take independent and identically distributed values that depend only on the value of their parent node. Note that the hidden process is sometimes called latent process in the literature. Conditionally on the hidden process X, the observed process $Y = (Y_u, u \in T)$, with values in another metric space $\mathcal{Y}$, is composed of independent random variables Y_u which only depends on X_u for all $u \in T$. See Definitions 5.2.1 and 5.2.2 below for a complete formal definitions. In this article, we consider the case where the tree T is the (deterministic) complete infinite rooted binary tree, that is, each vertex has exactly two children. See Figure 5.1 for a graphical representation of the dependance between the variables composing the HMT process (X, Y) indexed by T.

5.1.1 Literature review

HMMs were first introduced by Baum and Petrie in [BP66] and were popularized by Rabiner's tutorial [Rab89]. Since then, HMMs have been used in a wide variety of applications such as speech recognition [YD15], bioinformatics [Kos01], finance [ME14], and time-series analysis [ZM09]; see also [BFA22] for a more global reference on HMMs applications.

HMTs were first introduced in [CNB98] to account for the multi-scale dependency of wavelet coefficients in statistical signal processing with applications in wavelet-based image processing [RCBK00, CB01,

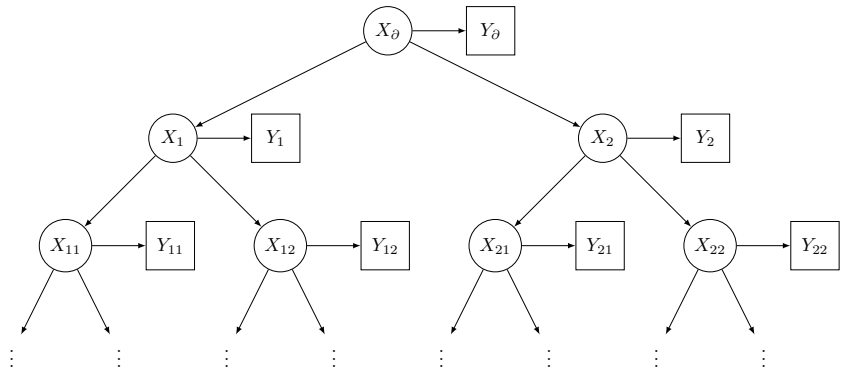


Figure 5.1 – Graph of dependance for variables of a HMT process indexed by the complete infinite rooted binary tree T. The observed variables are represented inside square, while the hidden variables are represented inside circles.

DWB08, SH17]. After that, HMTs have been used in several application contexts such as natural language processing [GOB13, KDM13], flood mapping [XJS18], medical imaging [MBY$^+$12, HYG17, HBSLLB$^+$17], plant growth modeling [DGCC05], and bioinformatics [OCB$^+$09, BWX13, NSK20].

In practice, maximum likelihood estimation for HMMs often relies on iterative numerical methods to approximate the maximum likelihood estimator (MLE). Those methods are often based on the expectation-maximization algorithm which is an algorithm for models with missing data and was popularized by Dempster et al. [DLR77] in a celebrated article. For HMMs with finite hidden state space, the first presentation of a complete expectation-maximization strategy is due to Baum et al. [BPSW70], and is the well-known "forward-backward" or Baum-Welch algorithm. For more details on the expectation-maximization and "forward-backward" algorithms and their stochastic approximations, see [CMR05, Chapters 10 and 11]. In the HMT case, the "forward-backward" algorithm must be replaced by the "upward-downward" algorithm developed in [CNB98]. See also [DGG04] for alternative "upward-downward" recursive formulae that can handle underflow issues implicitly.

The statistical properties of the MLE for the HMM were first studied in [BP66] which proved consistency and asymptotic normality in the case where both the hidden and the observed processes can only take finitely many values. Those results were then successively extended in a series of articles [Ler92, BRR98, JP99, LGM00, DM01]. An extension of all those results for HMMs with autoregression (that is, when conditionally on the hidden Markov chain, the observed process is an inhomogeneous s-order Markov chain for some $s \in \mathbb{N}$) was later developed in [DMR04], which proved, using weaker assumptions, strong consistency and asymptotic normality of the MLE for auto-regressive HMMs with compact hidden state space and with possibly non-stationary regime. The methods used in [DMR04] relies on expressing the log-likelihood as an additive function of an extended Markov chain with infinite past thanks to stationarity and using geometric ergodicity of this extended chain (extension to non-stationary regime is then made separately). The method of [DMR04] was adapted in [KS19] under similar assumptions to allow the transition densities of the hidden process to be zero valued. Since the article [DMR04], the strong consistency of the MLE was proved under weaker assumptions in [GL06, DMOVH11, DRS16], but no generalization has been made for the asymptotic normality of the MLE.

In this article, we will adapt the proof method of [DMR04] to the HMT case. We shall also refer to the monograph [CMR05] which exposes in details the theory of HMMs, and in particular to its Chapter 12 which covers the strong consistency and asymptotic normality of the MLE, under the same assumptions used in [DMR04], for HMMs where the hidden state space is a general metric space.

To adapt the proof method of [DMR04] to the HMT case, we will need almost sure (a.s.) and L^2 ergodic convergence results for branching Markov chains under geometric ergodicity of the transition kernel as in [Guy07, Wei24b]. Indeed, we will need variants of those results for neighborhood-dependent functions (that is, the function associated to each vertex u depends on variables X_v for vertices v in the neighborhood of u) which we develop in Section 5.2.4 and Section 5.A.

5.1.2 New contribution

In this article, we consider the case where the distribution of the HMT is parametrized by some vector θ, that is, the transition kernel Q_θ between the hidden variables and the transition kernel G_θ from hidden variables to observed variables both depend on θ. As an example, if the hidden state space $\mathcal{X}$ is finite and Y_u conditioned on X_u is a Gaussian random variable for each $u \in T$, then θ could parametrized the transition matrix of the hidden process and the mean and variances of the Gaussian distribution associated to each hidden state values. Our goal is to estimate the true parameter $\theta^\star$ of the HMT process among a compact set of possible parameters $\Theta \subset \mathbb{R}^d$, for some integer d, using only the knowledge of the observed process Y over n generations of the tree. Note that as our assumptions will imply uniform exponential memorylessness properties for the initial distribution, we cannot try estimate the initial distribution. Denote ∂ the root of the tree T. Thus, we assume that the distribution of the hidden root variable X_∂ is some unknown measure ζ which does not depend on θ. Denote by $\mathbb{P}_{\theta^\star,\zeta}$ the probability distribution of the HMT under the true parameter $\theta^\star$ when the initial unknown distribution of X_∂ is ζ. When ζ is the unique invariant measure of Q_θ (i.e. in the stationary case), we write $\mathbb{P}_{\theta^\star}$ instead of $\mathbb{P}_{\theta^\star,\zeta}$.

To estimate the true parameter $\theta^\star$ of the HMT, we will use the maximum likelihood estimator (MLE). We will work with the likelihood conditioned on the hidden state of the root vertex X_∂. The reason to do this is that the computation of the stationary distribution of the joint process (X, Y), and thus also the true likelihood, is intractable in typical applications. Note that the idea of conditioning on the initial hidden state was already used in [DMR04] for HMMs with the same motivation, and conditioning on initial observations in time series goes back at least to [MW43]. Remind that T denote the (deterministic) complete infinite rooted binary tree. Denote T_n the tree T up to and including the n-th generation. Hence, for any value $x \in \mathcal{X}$, we denote by $\ell_{n,x}(\theta)$ the log-likelihood under the parameter θ of the observed process $(Y_u, u \in T_n)$ until the n-th generation of the tree T conditionally on $X_\partial = x$ (see (5.7) on page 125 for exact definition). Then, for any value $x \in \mathcal{X}$, we define the MLE $\hat{\theta}_{n,x}$ as the maximizer of $\ell_{n,x}$ over Θ (see (5.33) on page 135 for exact definition).

Our goal is to study the asymptotic properties of the MLE. We prove the strong consistency and the asymptotic normality of the MLE in the stationary case in Sections 5.3 and 5.4, respectively. We then extend those results to the non-stationary case in Section 5.5. In our results, the hidden state space $\mathcal{X}$ and the observed state space $\mathcal{Y}$ are both general metric spaces. We prove our results under the same assumptions used in [DMR04] and in [CMR05, Chapter 12] for HMMs with L^1 and L^2 integrability assumptions replaced by L^2 and L^4 integrability assumptions, respectively, to accommodate the stronger assumptions needed in ergodic theorems for branching Markov chains. See Remark 5.1.6 below for a discussion on the main differences between the HMM case as in [DMR04, CMR05] and the HMT case we develop in this article.

We first state that strong consistency of the MLE holds under standard assumptions for HMMs. Following [DMR04], we assume a fully dominated model, that is, the transition kernels Q_θ and G_θ admits densities q_θ and g_θ w.r.t. to common measures λ and μ, respectively (see Assumption 6). We also assume (see Assumption 7) :

$$0 < \sigma^- \leq \inf_{x,x' \in \mathcal{X}} q_\theta(x, x') \leq \sup_{x,x' \in \mathcal{X}} q_\theta(x, x') \leq \sigma^+ < \infty. \tag{5.1}$$

This assumption is rather strong as it imposes a full connection for the hidden space, see [KS19] for an extension of the method in [DMR04] for HMMs where q_θ is allowed to be zero valued. Nevertheless, this assumption implies the uniform exponential memorylessness properties with mixing rate $\rho := 1 - \sigma^-/\sigma^+$ of the initial distribution conditional on the observations $(Y_u, u \in T_n)$. The other assumptions are more standard regularity assumptions for the densities q_θ and g_θ (see Assumptions 6-10), and identifiability of the model. We can now state the strong consistency of the MLE under those assumptions, see Theorems 5.3.11 and 5.5.1 for the precise statements in the stationary and non-stationary case, respectively.

Theorem 5.1.1 (Strong consistency of the MLE)**.** *Under those assumptions of fully dominated model with density satisfying* (5.1) *and other more standard regularity assumptions, and under the assumption that the model is identifiable, for any $x \in \mathcal{X}$, the MLE $\hat{\theta}_{n,x}$ is strongly consistent, that is, the sequence $(\hat{\theta}_{n,x})_{n\in\mathbb{N}}$ converges $\mathbb{P}_{\theta^\star,\zeta}$-almost surely to the true parameter $\theta^\star \in \Theta$.*

To prove asymptotic normality of the MLE, in addition to the assumptions used in Theorem 5.1.1, we need existence and regularity assumptions for the gradient and the Hessian of the transition densities

q_θ and g_θ (see Assumptions 11-13). Denote by $\mathcal{I}(\theta^\star)$ the limiting Fisher information matrix of the model (see (5.54) on page 144 for precise definition). The proof of asymptotic normality in the non-stationary case is an extension of the stationary case. The proof of asymptotic normality in the stationary case follows from a standard argument for asymptotic normality of the MLE that relies on Theorem 5.1.1 and Theorems 5.1.2 and 5.1.3 below.

The following theorem, which we only prove in the stationary case, states that the normalized score $|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star)$ has asymptotic normal fluctuations with covariance matrix $\mathcal{I}(\theta^\star)$, see Theorem 5.4.3 for the precise statement. Note that the extra assumption in Theorem 5.1.2 (not present in the case of HMMs) that $\rho < 1/\sqrt{2}$ for the mixing rate ρ of the HMT process comes from the approximation bounds used in the proof of this theorem. See Remark 5.1.5 below for a discussion on this condition on ρ.

Theorem 5.1.2 (Asymptotic normality of the normalized score)**.** *Under the assumptions from Theorem 5.1.1 and existence and regularity assumptions for the gradient and the Hessian of the transition densities (see Assumptions 11-13), and under the assumption that $\rho < 1/\sqrt{2}$ for the mixing rate ρ of the HMT process, in the stationary case we have:*

$$|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)) \quad \textit{under } \mathbb{P}_{\theta^\star}.$$

The following theorem states the locally uniform convergence $\mathbb{P}_{\theta^\star,\zeta}$-a.s. of the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta)$ towards the Fisher information matrix $\mathcal{I}(\theta^\star)$, see Theorems 5.4.6 and 5.5.2 for the precise statements in the stationary and non-stationary case, respectively. Note that in this theorem we need the stronger assumption $\rho < 1/2$ for the mixing rate ρ of the HMT process as we use more restrictive approximation bounds in the proof of this theorem than the ones used in the proof of Theorem 5.1.2.

Theorem 5.1.3 (Convergence of the normalized observed information)**.** *Under the assumptions from Theorem 5.1.2 on the HMT model, and under the assumption that $\rho < 1/2$ for the mixing rate ρ of the HMT process, for all $x \in \mathcal{X}$, we have:*

$$\lim_{\delta\to 0}\lim_{n\to\infty} \sup_{\theta\in\Theta\,:\,\|\theta-\theta^\star\|\le\delta} \left\| -|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta) - \mathcal{I}(\theta^\star)\right\| = 0 \quad \mathbb{P}_{\theta^\star,\zeta}\textit{-a.s.}$$

In particular, combining Theorems 5.1.1 and 5.1.3, we get that the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\hat\theta_{n,x})$ at the MLE $\hat\theta_{n,x}$ is a strongly consistent estimator of the Fisher information matrix $\mathcal{I}(\theta^\star)$.

As announced above, following a standard argument for asymptotic normality of the MLE, Theorems 5.1.1, 5.1.2 and 5.1.3 imply the following theorem which states that the MLE has asymptotic normal fluctuations with covariance matrix $\mathcal{I}(\theta^\star)^{-1}$. See Theorems 5.4.7 and 5.5.5 for the precise statements in the stationary and non-stationary case, respectively.

Theorem 5.1.4 (Asymptotic normality of the MLE)**.** *Under the assumptions from Theorem 5.1.2 on the HMT model, that $\theta^\star$ is an interior point of Θ, and the Fisher information matrix $\mathcal{I}(\theta^\star)$ is non-singular, and under the assumption that $\rho < 1/2$ for the mixing rate ρ of the HMT process, we have the following convergence in distribution:*

$$|T_n|^{1/2}\big(\hat\theta_n - \theta^\star\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1}) \quad \textit{under } \mathbb{P}_{\theta^\star,\zeta},$$

where $\mathcal{N}(0, M)$ denotes the centered Gaussian distribution with covariance matrix M.

Note that the standard argument used in the proof of Theorem 5.1.4 implies that we have the following joint convergence in distribution:

$$\Big(|T_n|^{1/2}\big(\hat\theta_n - \theta^\star\big),\, |T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star)\Big) \xrightarrow[n\to\infty]{(d)} \big(\mathcal{I}(\theta^\star)^{-1/2}\,G,\, \mathcal{I}(\theta^\star)^{1/2}\,G\big) \quad \text{under } \mathbb{P}_{\theta^\star},$$

where G is Gaussian random variable distributed as $\mathcal{N}(0, I_d)$ with I_d the identity matrix of dimension $d \times d$, and $\mathcal{I}(\theta^\star)^{1/2}$ is a root matrix of $\mathcal{I}(\theta^\star)$.

The following remark is a discussion on the condition on the mixing rate ρ of the HMT process (X, Y) that appear in Theorems 5.1.4, 5.1.2 and 5.1.3.

Remark 5.1.5 (On the condition on the mixing rate ρ)**.** Note that in central limit theorems for branching Markov chains, three regimes with different asymptotic behaviors (and different normalization terms) for $\rho < 1/\sqrt{2}$, $\rho = 1/\sqrt{2}$ and $\rho > 1/\sqrt{2}$ were observed in [BPD22a], corresponding to a competition between the ergodic mixing rate ρ and the branching rate 2 in T, see also [Ath69, BPDG14, BPD22b]. However, the condition on ρ disappears when we consider martingale increments in the central limit theorem for branching Markov chains, see [Guy07, BDSG09, DM10].

In our case, the condition $\rho < 1/\sqrt{2}$ on the mixing rate ρ that appears in Theorem 5.1.2 is due to the coupling bounds and the grouping of terms used in the proof of Lemma 5.4.2 (the upper bounds at the end of the proof only add a constant multiplicative factor). It is an open question whether or not some convergence would still hold in Theorem 5.1.2 with $\rho \geq 1/\sqrt{2}$ even with a possibly stronger normalization term and a possibly non Gaussian limit. Nevertheless, note that the proof of Theorem 5.1.2 relies on decomposing the score $\nabla_\theta \ell_{n,x}(\theta)$ as a sum of martingale increments, which could indicate that convergence is possible for $\rho \geq 1/\sqrt{2}$.

Moreover, the stronger condition $\rho < 1/2$ on the mixing rate ρ that appears in Theorem 5.1.3, and thus in Theorem 5.1.4, is due to the coupling bounds from Lemma 5.4.16 and the grouping of terms used in the proof of Lemma 5.4.17 (the upper bounds in the rest of the proof only add a constant multiplicative factor). It is an open question whether or not convergence would still hold in Theorem 5.1.3 and in Theorem 5.1.4 with $\rho \geq 1/2$ even with a possibly stronger normalization term and a possibly non Gaussian limit in Theorem 5.1.4. Also note that the condition $\rho < 1/2$ is used when proving that Theorem 5.1.4 extends to the non-stationary case to construct a coupling between a stationary HMT process and a non-stationary HMT process, see Lemma 5.5.3.

In the following remark, we discuss the main differences between the HMM case as in [DMR04, CMR05] and the HMT case we develop in this article.

Remark 5.1.6 (On main differences with the HMM case)**.** In both HMM and HMT cases, the study of the log-likelihood is based on decomposing it as a sum of increments, and then extending the "past" seen by each variable. However, while the extended "past" only spreads backwards in the HMM case, the extended "past" in the HMT case is a subtree that also spreads laterally due to the different topologies between the line $\mathbb{Z}$ and the binary tree, see Figure 5.3 on page 126 for an illustration. See also Sections 5.2.4 and 5.3.1 for the definition of those "past" and extended "past". Moreover, due to the enumeration of vertices in the tree in a breadth-first-search manner, those extended "past" do not have the same "shapes" for all vertices, see Section 5.2.4. Also note that the infinite expanded "past" of a vertex relies on a random infinite "backward spine" of left / right ancestors (see Figure 5.4 on page 128), which adds extra randomness to the "shape" of the "past".

Furthermore, contrary to the HMM case, the lateral spreading of each vertex's "past" in the HMT case implies that log-likelihood increments with infinite extended "pasts" do not form a branching Markov process. For this reason, we need to work with log-likelihood increments whose "past" is trimmed to a fixed common subtree height, and only expand to infinite "past" in the limit. To prove convergence for sums of log-likelihood increments with trimmed "pasts" which have different shapes, we need to develop new ergodic theorems for branching Markov chains and neighborhood-dependent functions, see Section 5.2.4 and Section 5.A.

In the proof of asymptotic normality of the normalized score, the score is decomposed as a sum of martingale increments which is no longer stationary in the HMT case due to the "pasts" of vertices having different shapes. Thus, to apply the central limit theorem for martingales, we first need to verify convergence for the quadratic variations of the martingale increment sequences and Lindeberg's condition. Moreover, the computation of the approximation bounds for the increments used to decompose the score and the observed information are more involved and impose conditions on the value of the mixing rate ρ, as already discussed in Remark 5.1.5. This also implies that the proof scheme for convergence of the observed information matrix needs to be modified as we cannot have almost sure convergence for all the increments simultaneously, and we must rely on L^2 convergence instead.

Lastly, as discussed in Section 5.1.1, the results for HMMs in [DMR04] allowed for autoregression (remind, that is, when conditionally on the hidden Markov chain, the observed process is an inhomogeneous s-order Markov chain for some $s \in \mathbb{N}$). Our results for HMTs are stated for processes without autogression. However, as our approach adapts the proof scheme of [DMR04], note that with straightforward modifications of our proofs, we could allow for autoregression in HMT processes.

5.1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 5.2, we define the notations used in this article, HMT processes and the log-likelihood for the HMT. For the stationary case, we prove the strong consistency of the MLE in Section 5.3, and its asymptotic normality in Section 5.4. In Section 5.5, we extend those results to the non-stationary case. In Section 5.A, we develop the ergodic theorems for branching Markov chains with neighborhood-dependent functions needed in the proofs of the asymptotic properties of the MLE.

5.2 Definition of HMT and notations

In this section, we first define the notations we use for the infinite complete binary tree T. We then define branching Markov chains and hidden Markov models (HMMs) indexed by the binary tree T, which we will simply call Hidden Markov Tree (HMT) models. We continue with the basic assumptions we need to define the log-likelihood for the HMT. Lastly, we present the ergodic theorems for branching Markov chains and neighborhood-dependent functions needed in this article, whose proofs can be found in Section 5.A.

5.2.1 Notations for trees

Let $T = \cup_{n\in\mathbb{N}}\{0,1\}^n$ denote the infinite complete plane rooted binary tree, that is the plane rooted tree where each vertex u has exactly two children $u0$ and $u1$. Denote by ∂ the root vertex of T (which is the unique point in $\{0,1\}^0$). If u is distinct from the root, we denote by $\mathrm{p}(u)$ its parent vertex. We denote by $h(u)$ its height, i.e. the number of edges separating u from the root ∂. (The height of the root ∂ is zero.) In particular, for $k \leq h(u)$, note that $\mathrm{p}^k(u)$ denotes the k-th ancestor of u. For two vertices $u, v \in T$, we denote by $u \wedge v$ the most recent common ancestor of u and v, and by $d(u,v)$ the graph-distance between u and v in T, that is $d(u,v) = h(u) + h(v) - 2h(u \wedge v)$. For all $n \in \mathbb{N}$, denote by G_n the n-th generation of the tree, that is vertices that are at distance n from the root, and denote by $T_n = \cup_{0\leq k\leq n} G_k$ the tree up to generation n. For a vertex $u \in T$, we denote by $T(u)$ the subtree of T composed of descendants of u, and for all $k \in \mathbb{N}$, we denote by $T(u,k) = T(u) \cap T_{h(u)+k}$ the subtree of $T(u)$ composed of descendants of u at distance up to k from u. We will use the convention that for a subtree T_{sub} of T, we write T^*_{sub} for the subtree T_{sub} without its root vertex, for instance, $T^*_n = T_n \setminus \{\partial\}$ and $T(u)^* = T(u) \setminus \{u\}$. For a finite subset $A \subset T$, we denote by $|A|$ its cardinal.

We will sometimes use Neveu's notation, which we define recursively: a vertex $u \in T$ with height $h(u) = n$ can be represented as a sequence $(u_{(j)})_{1\leq j\leq n}$ where u is the $u_{(n)}$-th child of $\mathrm{p}(u)$ and $\mathrm{p}(u)$ can be represented by $(u_{(j)})_{1\leq j\leq n-1}$; and the representation of the root ∂ is the empty sequence. Note that Neveu's notation can also be interpreted as encoding the path from the root ∂ to the vertex u: starting from the root $u_0 = \partial$, at each generation j we go from u_j to its $u_{(j+1)}$ child which we denote by u_{j+1}, and at generation n we get $u_n = u$.

For simplicity, we will write $u_{(k:n)} = (u_{(j)})_{k\leq j\leq n}$ and $u_{(k:n)} = (u_{(j)})_{k\leq j\leq n}$ for path sequences where $k, n \in \mathbb{Z}$ with $k < n$.

As T is a plane rooted tree, we can order its vertices in a breadth-first-search manner, that is, the total order relation $\leq$ on T is defined for all $u, v \in T$ as $u \leq v$ if $h(u) < h(v)$ or $h(u) = h(v)$ and $u \leq_{\mathrm{lex}} v$ (where $\leq_{\mathrm{lex}}$ is the lexicographical order on T). Moreover, we denote by $v < u$ if $u \leq v$ and $v \neq u$.

5.2.2 Definition of HMT processes

For a sequence $(x_u, u \in T)$, for simplicity, we will write $x_A = (x_u, u \in A)$ for all subsets $A \subset T$. For a metric space $\mathcal{X}$, we will always assume it is equipped with its Borel σ-field $\mathcal{B}(\mathcal{X})$.

For a measure μ on a metric space $\mathcal{X}$ and an integrable function $f \in L^1(\mu)$, we write $\mu(f) = \int_{\mathcal{X}} f \,\mathrm{d}\mu$. For two probability measures μ_1, μ_2 on a metric space $\mathcal{X}$, we denote the total variation norm between them by $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_{A\subset\mathcal{X}} |\mu_1(A) - \mu_2(A)|$ (where A ranges over all measurable subsets of $\mathcal{X}$). We also remind the identities $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2}\sup_{f:|f|\leq 1} |\mu_1(f) - \mu_2(f)| = \sup_{f:0\leq f\leq 1} |\mu_1(f) - \mu_2(f)|$ (where f is a measurable function). Note that $\|\mu_1 - \mu_2\|_{\mathrm{TV}}$ takes value in $[0,1]$.

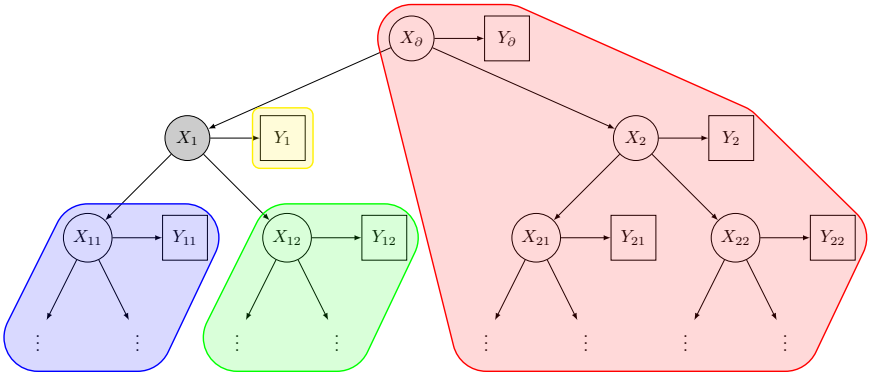


Figure 5.2 – Illustration of the Markov property for the HMT. Conditioning on X_1 (in grey) implies that the HMT process (X, Y) becomes independent between the four connected components of variable-dependence tree from Figure 5.1 where the vertex X_1 is removed, that is, Y_1, $(X_{T(11)}, Y_{T(11)})$, $(X_{T(12)}, Y_{T(12)})$ and $(X_{T\setminus T(1)}, Y_{T\setminus T(1)})$ (respectively in yellow, blue, green and red) are independent conditionally on X_1.

Denote by $X = (X_u, u \in T)$ the hidden (stochastic) process with values in a metric space $\mathcal{X}$, and by $Y = (Y_u, u \in T)$ the observed (stochastic) process with values in a metric space $\mathcal{Y}$. We assume that the hidden process X is a branching Markov process.

Definition 5.2.1 (Branching Markov process)**.** *The stochastic process X is called a* (branching) Markov process *with transition (probability) kernel Q on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and initial (probability) distribution ν on $\mathcal{X}$ if for all $n \in \mathbb{N}$, we have:*

$$\mathbb{P}(X_{T_n} \in \mathrm{d}x_{T_n}) = \nu(\mathrm{d}x_\partial)\, \Pi_{u \in T_n^*} Q(x_{\mathrm{p}(u)}; \mathrm{d}x_u).$$

We can now define the Hidden Markov Tree process.

Definition 5.2.2 (Hidden Markov Tree process)**.** *The stochastic process $(X, Y) = ((X_u, Y_u), u \in T)$ is called a* Hidden Markov Tree (HMT) *with parameter (Q, G, ν) if:*

(i) the hidden process $X = (X_u, u \in T)$ is a branching Markov process with transition kernel Q and initial distribution ν,

(ii) the observed process $Y = (Y_u, u \in T)$ conditioned on the hidden process X is composed of independent variables whose laws factorize using the transition (probability) kernel G on $(\mathcal{X}, \mathcal{B}(\mathcal{Y}))$, that is for all $n \in \mathbb{N}$, we have:

$$\mathbb{P}(Y_{T_n} = y_{T_n} \,|\, X_{T_n} = x_{T_n}) = \Pi_{u \in T_n} G(x_u; \mathrm{d}y_u).$$

Note that the definitions of branching Markov chains and HMT processes also work for non-plane rooted trees.

In particular, note that if $(X, Y) = ((X_u, Y_u), u \in T)$ is a HMT process, then the joint process $((X_u, Y_u), u \in T)$ is also a branching Markov chain (but the observed process Y is not necessarily Markov). The following fact, which we shall reuse later, illustrates the Markov property of the HMT process (X, Y). For any $k \in \mathbb{N}^*$, any $u \in T$ with height at least k, and any subset $A \subset T$, we have :

$$\mathcal{L}(X_u \,|\, Y_A, X_{\mathrm{p}^k(u)}) = \mathcal{L}(X_u \,|\, Y_{A \cap T(\mathrm{p}^{k-1}(u))}, X_{\mathrm{p}^k(u)}) \tag{5.2}$$

where $\mathcal{L}(R \,|\, S)$ denotes the distribution of a random variable R conditionally on another random variable S.

We say that a branching Markov process X is *stationary* if all its variables are identically distributed (that is, for all $u \in T$, X_u has the same distribution as X_∂), or equivalently if its initial distribution ν is invariant for its transition kernel Q (that is, $\nu Q = \nu$). Moreover, if (X, Y) is a HMT process and the hidden process X is stationary, then the joint process (X, Y) is also stationary.

5.2.3 Basic assumptions and definition of the log-likelihood

In this article, we consider the case where the kernels in the definition of the HMT $(Q_\theta, G_\theta, \nu_\theta)$ are parametrized by some vector θ that we want to estimate using only the knowledge of the observed process $(Y_u, u \in T_n)$ up to generation n. We denote by Θ the set of all possible vector parameters θ, which we assume to be a subset of $\mathbb{R}^d$ for some integer d. And we denote by $\theta^\star$ the true parameter of the HMT.

Through this article, with the exception of Section 5.5, we assume that the hidden process X is stationary.

Assumption 5 (Stationarity)**.** *The hidden process $(X_u, u \in T)$ is stationary, (and thus the joint process $((X_u, Y_u), u \in T)$ is also stationary).*

We denote by $\mathbb{P}_\theta$ the probability distribution under the parameter θ of the stationary joint process (X, Y), and by $\mathbb{E}_\theta$ the corresponding expectation.

To prove asymptotic properties of the MLE for the HMT, we will use assumptions similar to the HMM case in [CMR05, Chapter 12] and [DMR04]. We first assume that the HMT model is fully dominated. For two probability measures λ, μ on the same space, we write $\lambda \ll \mu$ to denote that λ is absolutely continuous w.r.t. to μ.

Assumption 6 (Fully dominated model, [CMR05, Assumption 12.0.1])**.**

(i) There exists a probability measure λ on $\mathcal{X}$ such that for every $x \in \mathcal{X}$ and every $\theta \in \Theta$, $Q_\theta(x, \cdot) \ll \lambda$, with density $q_\theta(x, \cdot)$. That is, $Q_\theta(x; A) = \int_A q_\theta(x, x')\, \lambda(\mathrm{d}x')$ for all $A \in \mathcal{B}(\mathcal{X})$. Moreover, the density function $q_\theta(\cdot, \cdot)$ is $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$ measurable.

(ii) There exists a σ-finite measure μ on $\mathcal{Y}$ such that for every $x \in \mathcal{X}$ and every $\theta \in \Theta$, $G_\theta(x, \cdot) \ll \mu$, with density $g_\theta(x, \cdot)$. That is, $G_\theta(x; B) = \int_B g_\theta(x, y)\, \mu(\mathrm{d}y)$ for all $B \in \mathcal{B}(\mathcal{Y})$. Moreover, the density function $g_\theta(\cdot, \cdot)$ is $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$ measurable.

In addition to Assumption 6, we use the following regularity assumptions on the density functions q_θ and g_θ.

Assumption 7 (Regularity, [CMR05, Assumption 12.2.1])**.**

(i) The transition density q_θ is bounded: there exist $\sigma^-, \sigma^+ \in (0, +\infty)$ such that $\forall x, x' \in \mathcal{X},\ \forall \theta \in \Theta$, $0 < \sigma^- \leq q_\theta(x, x') \leq \sigma^+ < +\infty$.

(ii) For every $y \in \mathcal{Y}$, the function $\theta \mapsto \int_{\mathcal{X}} g_\theta(x, y)\, \lambda(\mathrm{d}x)$ is bounded away from 0 and ∞ uniformly on Θ, that is, $\sup_{\theta \in \Theta} \int_{\mathcal{X}} g_\theta(x, y)\, \lambda(\mathrm{d}x) < \infty$ and $\inf_{\theta \in \Theta} \int_{\mathcal{X}} g_\theta(x, y)\, \lambda(\mathrm{d}x) > 0$.

(iii) For every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $g_\theta(x, y) > 0$.

We will denote by $\rho = 1 - \sigma^-/\sigma^+ \in (0, 1)$ the *mixing rate* of the hidden process X.

Note that Assumption 7-(iii) is due to our choice of conditioning on $X_\partial = x$ for some $x \in \mathcal{X}$ in the log-likelihood we analyse, we discuss how to get rid of this assumption after the definition of the log-likelihood at the end of this subsection.

As λ is a probability measure and q_θ is the density of a probability kernel, Assumption 7-(i) implies that $\sigma^- \leq 1 \leq \sigma^+$. Moreover, Assumption 7-(i) also implies the following result.

Lemma 5.2.3. *Assume that Assumptions 6-(i) and 7-(i) hold. Then, for all $\theta \in \Theta$, the transition kernel Q_θ has a unique invariant measure π_θ and is uniformly geometrically ergodic, that is:*

$$\forall x \in \mathcal{X}, \forall \theta \in \Theta, \forall n \geq 0, \quad \|Q_\theta^n(x, \cdot) - \pi_\theta\|_{TV} \leq (1 - \sigma^-)^n \leq \rho^n.$$

Note that due to structure of the HMT, Lemma 5.2.3 extends naturally to the transition kernel of the joint process (X, Y) with the same mixing rate ρ. Moreover, note that Assumption 7-(i) also implies that $\pi_\theta \ll \lambda$ with density $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\lambda}$ taking value in $[\sigma^-, \sigma^+]$.

Lemma 5.2.3 is due to Assumption 7-(i) implying the Doeblin condition:

$$\forall \theta \in \Theta, \forall x \in \mathcal{X} \qquad \sigma^- \lambda(\cdot) \leq Q_\theta(x; \cdot). \tag{5.3}$$

As we will reuse Doeblin conditions later, before proving Lemma 5.2.3, we give a quick summary on results for the Doeblin condition. For a transition kernel K on a metric space $\mathcal{X}$ (to itself), we define its *Dobrushin coefficient* $\delta(K)$ as:

$$\delta(K) = \sup_{x, x' \in \mathcal{X}} \|K(x; \cdot) - K(x'; \cdot)\|_{\mathrm{TV}}. \tag{5.4}$$

The Dobrushin coefficient gives the following coupling bound in the total variation norm. (Note that the definition of the total variation norm $\|\cdot\|_{\mathrm{TV}}$ used in [CMR05, Chapter 4] differs by a factor 2 from ours, see [CMR05, Lemma 4.3.5].)

Lemma 5.2.4 ([CMR05, Lemma 4.3.8])**.** *Let μ_1, μ_2 be two probability measures on a metric space $\mathcal{X}$, and let K be a transition kernel on $\mathcal{X}$. Then, we have:*

$$\|(\mu_1 - \mu_2)K\|_{TV} \leq \delta(K)\|\mu_1 - \mu_2\|_{TV} \leq \delta(K).$$

Moreover, the Dobrushin coefficient is sub-multiplicative.

Lemma 5.2.5 ([CMR05, Proposition 4.3.10])**.** *The Dobrushin coefficient is sub-multiplicative. That is, if K, R are two transition kernels on a metric space $\mathcal{X}$, then we have $\delta(KR) \leq \delta(K)\delta(R)$.*

We know define the Doeblin condition.

Definition 5.2.6 (Doeblin condition, [CMR05, Definition 4.3.12])**.** *We say that a transition kernel K on a metric space $\mathcal{X}$ satisfies a* Doeblin condition *if there exist $\varepsilon > 0$ and a probability measure ν on $\mathcal{X}$ such that for all $x \in \mathcal{X}$ and measurable subset $A \subset \mathcal{X}$, we have:*

$$K(x; A) \geq \varepsilon\nu(A).$$

The Doeblin condition gives an upper bound on the Dobrushin coefficient.

Lemma 5.2.7 ([CMR05, Lemma 4.3.13])**.** *Let K be transition kernel (on a metric space $\mathcal{X}$) that satisfies a Doeblin condition with (ε, ν). Then, we have $\delta(K) \leq 1 - \varepsilon$.*

Lastly, the Doeblin condition implies the existence of a unique invariant probability measure, as well as uniform geometric ergodicity.

Lemma 5.2.8 ([CMR05, Theorem 4.3.16])**.** *Let K be a transition kernel on a metric space $\mathcal{X}$ that satisfies a Doeblin condition with (ε, ν). Then, K admits a unique invariant probability measure π. Moreover, for any probability measure ζ on $\mathcal{X}$, we have for all $n \in \mathbb{N}$:*

$$\|\zeta K^n - \pi\|_{TV} \leq (1 - \varepsilon)^n \|\zeta - \pi\|_{TV}.$$

Lemma 5.2.3 then follows immediately from Lemma 5.2.8 and the uniform Doeblin condition (5.3).

Remark 5.2.9 (More properties of the transition kernel from the Doeblin condition)**.** For a transition kernel K on a metric space $\mathcal{X}$, the Doeblin condition also implies that $\mathcal{X}$ is an (accessible) 1-small set. In particular, we get that K satisfies some extra classical properties (that we will not use here): K is positive (i.e. irreducible and admits a unique invariant probability measure), strongly aperiodic and Harris recurrent (see [DMPS18, Chapter 9 and 10] for definitions and details).

We will use the letter p to denote (possibly conditional) probability density. For instance, for any $\theta \in \Theta$, $y_{T_n} \in \mathcal{Y}^{T_n}$ and $x_\partial \in \mathcal{X}$, we denote by:

$$p_\theta(y_{T_n} \,|\, X_\partial = x_\partial) = g_\theta(x_\partial, y_\partial) \int_{\mathcal{X}^{T_n^*}} \prod_{v \in T_n^*} q_\theta(x_{\mathrm{p}(v)}, x_v) g_\theta(x_v, y_v) \lambda(\mathrm{d}x_v), \tag{5.5}$$

the conditional density w.r.t. $\mu^{\otimes T_n}$ under the parameter θ of $Y_{T_n} = y_{T_n}$ conditionally on $X_\partial = x_\partial$. Note that Assumption 7 guarantees that $p_\theta(y_{T_n} \,|\, X_\partial = x_\partial)$ is positive for all $y_{T_n} \in \mathcal{Y}^{T_n}$ and $x_\partial \in \mathcal{X}$.

We are now ready to define the log-likelihood. As discussed in Section 5.1.2, we will analyze the log-likelihood of the observed process $(Y_u, u \in T_n)$ up to generation n conditioned on the hidden value of the root $X_\partial = x$ for some $x \in \mathcal{X}$. Thus, for any $x \in \mathcal{X}$, we define the log-likelihood function as:

$$\ell_{n,x}(\theta; y_{T_n}) := \log\big(p_\theta(y_{T_n} \,|\, X_\partial = x)\big). \tag{5.6}$$

We then define the log-likelihood that we will analyze as the following random variable

$$\ell_{n,x}(\theta) := \ell_{n,x}(\theta; Y_{T_n}). \tag{5.7}$$

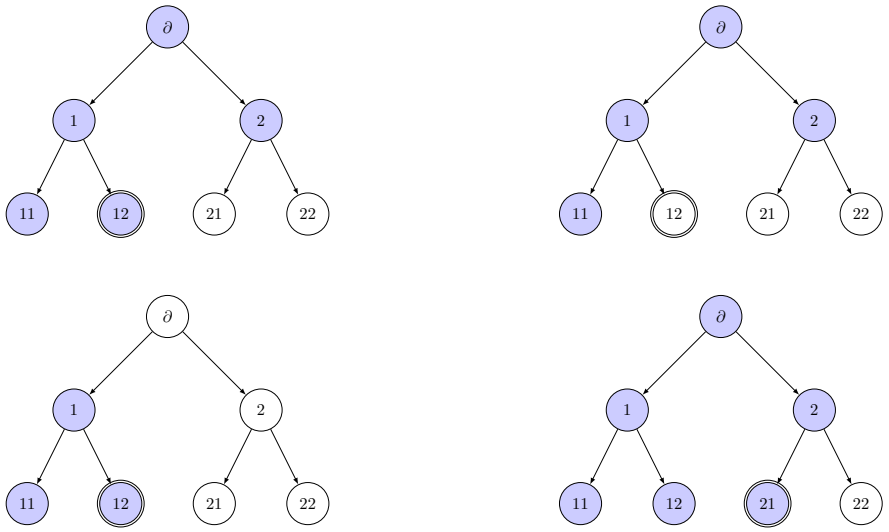


Figure 5.3 – Illustration of the subtrees $\Delta(u,k)$ and $\Delta^*(u,k)$ where vertices in $\Delta(u,k)$ and $\Delta^*(u,k)$ are in blue and the vertex u is inside a double circle. From left to right and top to bottom, using Neveu's notation, we have $\Delta(12) = \Delta(12,2)$, $\Delta^*(12) = \Delta^*(12,2)$, $\Delta(12,1)$ and $\Delta(21) = \Delta(21,2)$. Note that $\Delta(12)$ and $\Delta(21)$ have a different number of vertices, while vertices 12 and 21 are in the same generation.

For simplicity, we will write $\ell_{n,x}(\theta)$ instead of $\ell_{n,x}(\theta; Y_{T_n})$ making the dependence on the observed variables $(Y_u, u \in T_n)$ implicit. We will keep this convention for all quantities considered in this article, and only make the dependence explicit when necessary. The MLE is then the maximizer over Θ of the log-likelihood $\ell_{n,x}$; we post-pone the precise definition of the MLE to when we will first use it in Theorem 5.3.11.

Remark 5.2.10 (On Assumption 7-(iii))**.** Note that Assumption 7-(iii) is due to our choice of conditioning on $X_\partial = x$ for some $x \in \mathcal{X}$ in the log-likelihood $\ell_{n,x}(\theta)$ we analyse. Indeed, without Assumption 7-(iii), there could be a non-zero probability under $\mathbb{P}_{\theta^\star}$ that $g_{\theta^\star}(x, Y_\partial) = 0$ for some $x \in \mathcal{X}$, implying $\ell_{n,x}(\theta^\star) = -\infty$, and thus preventing the MLE to converge to $\theta^\star$. Several modifications of the log-likelihood $\ell_{n,x}(\theta)$ can be considered to get rid of Assumption 7-(iii). A first option would be to replace $p_\theta(y_{T_n} \,|\, X_\partial = x)$ by $p_\theta(y_{T_n^*} \,|\, X_\partial = x)$ in (5.6). A second option would be to extend the tree T and the HMT (X,Y) to add a parent vertex $\mathrm{p}(\partial)$ for the root vertex ∂ (see Section 5.3.1), and then replace $p_\theta(y_{T_n} \,|\, X_\partial = x)$ by $p_\theta(y_{T_n} \,|\, X_{\mathrm{p}(\partial)} = x)$ in (5.6).

5.2.4 Ergodic theorems with neighborhood-dependent functions

For all $u \in T$ and $0 \le k \le h(u)$, define the subtrees of T:

$$\Delta^*(u,k) = \{v \in T(\mathrm{p}^k(u)) \,:\, v < u\},$$

and $\Delta(u,k) = \Delta^*(u,k) \cup \{u\}$. In particular, note that when $k = h(u)$, we have that $\Delta^*(u) := \Delta^*(u, h(u)) = \{v \in T \,:\, v < u\}$, and we also write $\Delta(u) := \Delta(u, h(u))$. See Figure 5.3 for an illustration of those subtrees. The subtree $\Delta^*(u)$ represents the past of the vertex u.

For the ergodic convergence results needed in this article, we will need to consider different functions for each vertex $u \in T$ depending on the "shape" of the subtree $\Delta(u,k)$ for some common $k \in \mathbb{N}$. For $k \in \mathbb{N}$ and vertices $u, v \in T$ both with height at least k, we say that $\Delta(u,k)$ and $\Delta(v,k)$ *have the same shape* if they are equal up to translation, that is, if they are isomorphic as (finite) rooted plane trees. For $k \in \mathbb{N}$ and any vertex $u \in T$ with $h(u) \ge k$, there exists a unique $v_u \in G_k$ such that $\Delta(u,k)$ and $\Delta(v_u)$ have the same shape, and we thus define the *shape* of $\Delta(u,k)$ as:

$$\mathcal{S}h\big(\Delta(u,k)\big) = \Delta(v_u). \tag{5.8}$$

Note that as $|\Delta(v)|$ is different for each $v \in G_k$, thus the shape of $\Delta(u,k)$ is characterized by its size. For

any $k \in \mathbb{N}$, we define the (finite) set $\mathcal{N}_k$ of possible shapes for $\Delta(u,k)$ with $u \in T$ as:

$$\mathcal{N}_k = \{\Delta(v) \,:\, v \in G_k\}. \tag{5.9}$$

For any $k \in \mathbb{N}$, we define a collection of *neighborhood-shape-dependent* functions as a collection of functions $(f_{\mathcal{S}} : \mathcal{Z}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ where $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}, \mathcal{X}\times\mathcal{Y}\}$. For such a collection of functions, we will simply write $f_{\Delta(u,k)}$ instead of $f_{\mathcal{S}h(\Delta(u,k))}$. And we will also write $f_{\Delta(u,k)}(Y_{\Delta(u,k)})$ for the evaluation of $f_{\Delta(u,k)}$ on $Y_{\Delta(u,k)}$. Note that indexing such a collection of functions with G_k or with $\mathcal{N}_k$ is equivalent in light of (5.9).

We prove the following ergodic convergence lemma for neighborhood-shape-dependent functions. The proof of this lemma relies on the theorems in Section 5.A. Note that if U_n is uniformly distributed over G_n with $n \geq k$, then $\mathcal{S}h(\Delta(u,k))$ is uniformly distributed over $\mathcal{N}_k$.

Lemma 5.2.11 (Ergodic theorem for neighborhood-dependent functions)**.** *Assume that Assumptions 5–7 hold. Let $k \geq 0$. Let $(f_{\mathcal{S}} : \mathcal{Y}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(\mathbb{P}_{\theta^\star})$. Then, we have:*

$$\lim_{n\to\infty} \frac{1}{|T_n|} \sum_{u\in T_n\setminus T_{k-1}} f_{\Delta(u,k)}(Y_{\Delta(u,k)}) = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}\left[f_{\Delta(U_k)}(Y_{\Delta(U_k)})\right] \quad \mathbb{P}_{\theta^\star}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^\star}), \tag{5.10}$$

with the convention $T_{-1} = \emptyset$, and where U_k is uniformly distributed over G_k and independent of the process X, and $\mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}$ denotes the joint expectation over U_k and X (under the true parameter $\theta^\star$).

Moreover, there exist finite constants $C < \infty$ and $\alpha \in (0,1)$ such that:

$$\forall n \geq k, \quad \mathbb{E}\left[\left(\frac{1}{|T_n|} \sum_{u\in T_n\setminus T_{k-1}} f_{\Delta(u,k)}(Y_{\Delta(u,k)}) - \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}\left[f_{\Delta(U_k)}(Y_{\Delta(U_k)})\right]\right)^2\right] \leq C\alpha^n. \tag{5.11}$$

Remark that in the left hand side of (5.10) the subtrees $\Delta(u,k)$ are deterministic, while the subtree $\Delta(U_k)$ is a random function of U_k.

Proof. Using Lemma 5.2.3, remind that under Assumptions 5–7, the branching Markov process (X,Y) is stationary and its transition kernel has a unique invariant probability and is uniformly geometrically ergodic. Hence, the lemma follows immediately from applying the ergodic Theorems 5.A.2 and 5.A.4 for neighborhood-shape-dependent functions from the appendix. □

As T is a plane rooted tree, we can enumerate its vertices as a sequence $(v_j)_{j\in\mathbb{N}}$ in a breadth-first-search manner, that is, which is increasing for $<$ (note that $u_0 = \partial$). Note that if V_n is uniformly distributed over $A_n := \{v_j \,:\, |T_{k-1}| < j \leq n\} = \Delta(v_n) \setminus T_{k-1}$, then the distribution of $\mathcal{S}h(\Delta(V_n,k))$ converges to the uniform distribution over $\mathcal{N}_k$ as $n \to \infty$. We will also need the following variant of Lemma 5.2.11 where $T_n \setminus T_{k-1}$ is replaced by A_n.

Lemma 5.2.12 (Another ergodic theorem for neighborhood-dependent functions)**.** *Assume that Assumptions 5–7 hold. Let $k \geq 0$. Let $(f_{\mathcal{S}} : \mathcal{Y}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(\mathbb{P}_{\theta^\star})$. Let $(v_j)_{j\in\mathbb{N}}$ be the sequence enumerating the vertices in T in a breadth-first-search manner. For all $n > |T_{k-1}|$, define $A_n = \Delta(v_n) \setminus T_{k-1}$. Then, we have:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{u\in A_n} f_{\Delta(u,k)}(Y_{\Delta(u,k)}) = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}\left[f_{\Delta(U_k)}(Y_{\Delta(U_k)})\right] \quad \text{in } L^2(\mathbb{P}_{\theta^\star}),$$

where U_k is uniformly distributed over G_k and independent of the process X, and $\mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}$ denotes the joint expectation over U_k and X (under the true parameter $\theta^\star$).

Proof. Using Lemma 5.2.3, remind that under Assumptions 5–7, the branching Markov process (X,Y) is stationary and its transition kernel has a unique invariant probability and is uniformly geometrically ergodic. Hence, the lemma follows immediately from applying the ergodic Theorem 5.A.2 for neighborhood-shape-dependent functions from the appendix. □

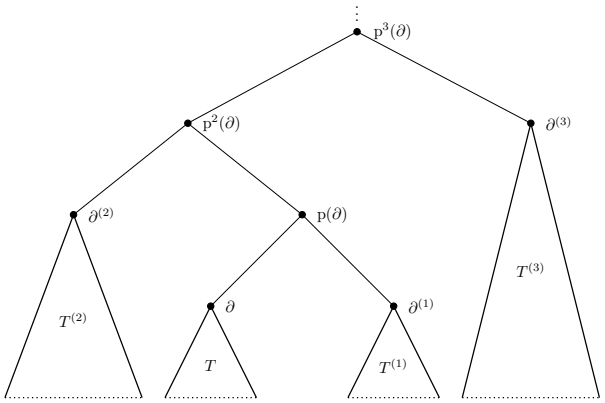


Figure 5.4 – Illustration of the construction of the extended random plane rooted tree T^∞ (which is rooted at ∂) for $\mathcal{U}_{(-1)} = 0$ and $\mathcal{U}_{(-2)} = 1$.

5.3 Strong consistency of the MLE

In this section, we first define the extended tree T^∞ to get an infinite past horizon and rewrite the log-likelihood as a sum of increments. Then, we construct the log-likelihood increments with infinite past, which allows to define the contrast function. We prove properties for this contrast function. Finally, we prove the strong consistency of the MLE.

5.3.1 Decomposition of the log-likelihood into increments

The extended tree T^∞ to get an infinite past horizon

Remind that the subtree $\Delta^*(u) = \Delta^*(u, h(u))$ represents the past of the vertex u.

To get an infinite past horizon, we will consider an extended version of the tree T. Thus, we are going to define a random (countable) plane rooted tree T^∞ that contains T as a subtree and is also rooted at ∂ the root vertex of T, and where each vertex (including ∂) has exactly one parent node and two children nodes. To construct T^∞, we start from T and add a line $L = \{u_{-j} : j \in \mathbb{N}^*\}$ of ancestors for ∂ (that is, $u_{-j} = \mathrm{p}^j(\partial)$ for $j \in \mathbb{N}$, where $u_0 = \partial$), and then for all $j \in \mathbb{N}^*$, we graft on u_{-j} a copy $T^{(j)}$ of T (that is, u_{-j} is the parent of the root vertex $\partial^{(j)}$ of $T^{(j)}$). We extend the height function h from T to T^∞ as follows: for all $j \in \mathbb{N}^*$, we set $h(u_{-j}) = -j$ and for all $u \in T^{(j)}$, we define $h(u)$ as $-j$ plus the number of edges separating u from u_{-j}. For $u, v \in T^\infty$, denote by $u \wedge v$ their most recent common ancestor, and by $d(u, v) = h(u) + h(v) - 2h(u \wedge v)$ the graph distance between u and v. The definition of the subtrees $T^\infty(u)$ and $T^\infty(u, k)$ then naturally extend to T^∞.

Thus, we have constructed the deterministic non-plane version of the tree T^∞, and we are left to define the random plane embedding of T^∞. That is, for each vertex $u \in T^\infty$, we have to define a possibly random ordering of its children. As T is a plane rooted tree, note that if $u \in T$ or $u \in T^{(j)}$ for some $j \in \mathbb{N}^*$, then its children are already order deterministically. Let $\mathcal{U} = (\mathcal{U}_{(j)})_{-\infty<j\leq 0}$ be a sequence of independent random variables with Bernoulli distribution of parameter $1/2$, and which is independent of the HMT process (X, Y). For all $j \in \mathbb{N}$, we order the children of u_{-j-1}, that is u_{-j} and $\delta^{(j+1)}$ (the root vertex of $T^{(j)}$), as follows: u_{-j} is the left child of u_{-j-1} if $\mathcal{U}_{(-j)} = 0$, and is the right child otherwise. Hence, we have constructed the random plane rooted tree T^∞. (Note that $\mathcal{U}$ can be seen as the random shape of the backward spine of ∂.) See Figure 5.4 for an illustration of the extended random plane rooted tree T^∞. We denote by $\mathbb{P}_\mathcal{U}$ the distribution of the random sequence $\mathcal{U}$, and by $\mathbb{E}_\mathcal{U}$ the corresponding expectation.

Note that the random plane embedding of T^∞ allows to use Neveu's notation to represent the random path between any vertex in the plane tree T^∞ and one of its descendants as a random sequence $\mathcal{U}_{(k:n)}$ (which depends on $\mathcal{U}$) for some $k, n \in \mathbb{Z}$ with $k < n$. The random breadth-first-search order relation $\leq := \leq_\mathcal{U}$ can then be naturally extend from T to T^∞ using the random plane embedding of T^∞ (which depends on $\mathcal{U}$): we have $u \leq v$ for $u, v \in T^\infty$ if either $h(u) < h(v)$, or $h(u) = h(v)$ and $U_{(k:n)} \leq_{\mathrm{lex}} V_{(k:n)}$ where $U_{(k:n)}$ (resp. $V_{(k:n)}$) is Neveu's notation for the random path (which depends on $\mathcal{U}$) from $u \wedge v$ to

u (resp. v) with $k = h(u \wedge v) + 1$ and $n = h(u)$.

Thanks to the stationarity assumption, for all $k \in \mathbb{N}$, the HMT process (X, Y) can be defined on the (rooted) tree $T^\infty(\mathrm{p}^k(\partial))$, and thus by Kolmogorov's extension theorem, the HMT process (X, Y) can be defined on the whole tree T^∞. In particular, note that the stationarity assumption implies that the distribution of the HMT process (X, Y) is invariant by translation on T^∞, that is, is the same (up to translation) on T and on $T^\infty(u)$ for any $u \in T^\infty$. Note that the extended process does not depend on $\mathcal{U}$. Thus, we will now assume that the HMT process (X, Y) is defined on the whole tree T^∞.

For all $u \in T^\infty$ and $k \in \mathbb{N}$, define the subtrees (which are measurable functions of $\mathcal{U}$):

$$\Delta^*_{\mathcal{U}}(u, k) = \{v \in T^\infty(\mathrm{p}^k(u)) \,:\, v <_{\mathcal{U}} u\},$$

and $\Delta_{\mathcal{U}}(u, k) = \Delta^*_{\mathcal{U}}(u, k) \cup \{u\}$. For simplicity, we will write instead $\Delta^*(u, k)$ and $\Delta(u, k)$, making the dependence on the random variable $\mathcal{U}$ implicit, and only make the dependence explicit when necessary. The following fact illustrates that the shape and size of the $\Delta(u, k)$ do indeed depend on the value of $\mathcal{U}$: for $u = \partial$ and $k = 1$, note that $\Delta(\partial, 1)$ contains two vertices if $\mathcal{U}_{(0)} = 0$, and contains three vertices if $\mathcal{U}_{(0)} = 1$. Remark 5.3.1 below, which we shall reuse later, further illustrates the randomness of the set $\Delta(u, k)$. However, for $u \in T$ and $k \leq h(u)$, we have that $\Delta(u, k) = \Delta(u, k)$ and $\Delta^*(u, k) = \Delta^*(u, k)$ are deterministic. Also note that we have the following inclusions:

$$T^\infty(u, k-1) \subset \Delta(u, k) \subset T^\infty(u, k), \tag{5.12}$$

where remind that the subtrees $T^\infty(u, k-1)$ and $T^\infty(u, k)$ are deterministic.

Remark 5.3.1. For a vertex $u = u_{(1:n)}$ in T with $h(u) = n \geq k$, note that $\Delta(u, k)$, up to re-rooting (i.e. up to translation), can be identified with $\Delta(\partial, k)$ conditioned on $\mathcal{U}_{(-k+1:0)} = u_{(n-k+1:n)}$. In particular, when U_n is a random vertex uniformly distributed over G_n for $n \geq k$, we get the following equality between the distribution of the shapes (that is, when the subtrees are seen up to translation / re-rooting) for the subtrees $\Delta(\partial, k)$, $\Delta(U_n, k)$ and $\Delta(U_k)$:

$$\mathcal{S}h\big(\Delta(\partial, k)\big) \overset{\mathcal{L}}{=} \mathcal{S}h(\Delta(U_n, k)) \overset{\mathcal{L}}{=} \Delta(U_k). \tag{5.13}$$

Moreover, if $(f_{\mathcal{S}} : \mathcal{Y}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$ is a collection of neighborhood-shape-dependent Borel functions that are in $L^2(\mathbb{P}_{\theta^\star})$ (as in Lemmas 5.2.11 and 5.2.12), then we have:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[f_{\Delta(\partial,k)}(Y_{\Delta(\partial,k)})\big] = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^\star}\big[f_{\Delta(U_k)}(Y_{\Delta(U_k)})\big], \tag{5.14}$$

where $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}$ is the expectation corresponding to $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$.

The log-likelihood as a sum of increments

For any (possibly random) subtree Δ of T^∞ with root vertex w, note that we have:

$$p_\theta(y_\Delta \,|\, X_w = x) = g_\theta(x, y_\partial) \int_{\mathcal{X}^{|\Delta|-1}} \prod_{v \in \Delta \setminus \{w\}} q_\theta(x_{\mathrm{p}(v)}, x_v) g_\theta(x_v, y_v) \lambda(\mathrm{d}x_v). \tag{5.15}$$

We will use the convention $p_\theta(Y_\Delta \,|\, X_w = x) = 1$ whenever $\Delta = \emptyset$ and w is any vertex in T^∞. For all $u \in T$, $k \in \mathbb{N}$, $x \in \mathcal{X}$ and $\theta \in \Theta$, using the conditional probabilities formula, define:

$$\begin{aligned} \mathrm{H}_{u,k,x}(\theta) &= \int_{\mathcal{X}} g_\theta(x_u, Y_u)\, \mathbb{P}_\theta(X_u \in \mathrm{d}x_u \,|\, Y_{\Delta^*(u,k)}, X_{\mathrm{p}^k(u)} = x) \\ &= \frac{p_\theta(Y_{\Delta(u,k)} \,|\, X_{\mathrm{p}^k(u)} = x)}{p_\theta(Y_{\Delta^*(u,k)} \,|\, X_{\mathrm{p}^k(u)} = x)}. \end{aligned} \tag{5.16}$$

We then define the log-likelihood contribution of node $u \in T$ with past over $k \in \mathbb{N}$ generation as:

$$\mathrm{h}_{u,k,x}(\theta) = \log\big(\mathrm{H}_{u,k,x}(\theta)\big). \tag{5.17}$$

Note that $\mathrm{h}_{u,k,x}(\theta)$ (resp. $\mathrm{H}_{u,k,x}(\theta)$) is a random variable as a function of $Y_{\Delta(u,k)}$ with an implicit dependence on $\mathcal{U}$ through $\Delta(u, k)$, and that $\mathrm{h}_{u,k,x}(\theta)$ (resp. $\mathrm{H}_{u,k,x}(\theta)$) does not depend on $\mathcal{U}$ is $k \leq h(u)$.

Hence, using (5.6), (5.7), (5.16) and (5.17) and a telescopic sum argument, the log-likelihood of the observed variables Y_{T_n} can be rewritten as the sum of the log-likelihood contributions defined in (5.17):

$$\ell_{n,x}(\theta) = \sum_{u \in T_n} \mathrm{h}_{u,h(u),x}(\theta). \tag{5.18}$$

5.3.2 Construction of the log-likelihood increments with infinite past

In this subsection, we construct the log-likelihood increment functions with infinite past.

The following lemma states that, as the HMT is uniformly geometrically ergodic, the tree forgets exponentially fast its starting state. Recall the mixing ratio $\rho = 1 - \sigma^-/\sigma^+ \in (0,1)$ is defined just after Assumption 7.

Lemma 5.3.2 (Exponential forgetting of the initial state)**.** *Assume that Assumptions 6 and 7 hold. We have for all $u \in T$, $\theta \in \Theta$, $n \in \mathbb{N}$ and $y_{T_n} \in \mathcal{Y}^{T_n}$, and all initials distributions ν and ν' on $\mathcal{X}$, that:*

$$\left\| \int_{\mathcal{X}} \mathbb{P}_\theta\Big(X_u \in \cdot \;\Big|\; Y_{T_n} = y_{T_n}, X_\partial = x\Big)[\nu(\mathrm{d}x) - \nu'(\mathrm{d}x)] \right\|_{\mathrm{TV}} \leq \rho^{h(u)}. \tag{5.19}$$

For simplicity, Lemma 5.3.2 is stated with ∂ as the initial vertex, but note that the results still holds when replacing ∂ and T_n by v and $T(v,n)$ for any $v \in T^\infty$. We shall reuse this fact later.

Proof. Fix some $u \in T$, $\theta \in \Theta$, an integer n and observables $y_{T_n} \in \mathcal{Y}^{T_n}$. Denote by $u_0, \cdots, u_k$ with $k = h(u)$ the vertices on the path from ∂ to u. The proof relies on the fact that conditionally on $Y_{T_n} = y_{T_n}$, the sequence $(X_{u_j})_{0 \leq j \leq k}$ is an inhomogeneous Markov chain where for $1 \leq j \leq k$, the (forward smoothing) transition kernel F_j from $X_{u_{j-1}}$ to X_{u_j} is defined if $j \leq n$ as:

$$\begin{aligned}
\mathrm{F}_j[y_{T(u_j,n-j)}](x_{u_{j-1}}; f) &= \mathbb{E}_\theta\big[f(X_{u_j}) \;\big|\; Y_{T(u_j,n-j)} = y_{T(u_j,n-j)}, X_{u_{j-1}} = x_{u_{j-1}}\big] \\
&= \mathbb{E}_\theta\big[f(X_{u_j}) \;\big|\; Y_{T_n} = y_{T_n}, X_{u_{j-1}} = x_{u_{j-1}}\big] \\
&= \frac{\int_{\mathcal{X}} f(x_{u_j}) p_\theta(y_{T(u_j,n-j)} \,|\, X_{u_j} = x_{u_j}) q_\theta(x_{u_{j-1}}, x_{u_j})\, \lambda(\mathrm{d}x_{u_j})}{\int_{\mathcal{X}} p_\theta(y_{T(u_j,n-j)} \,|\, X_{u_j} = x_{u_j}) q_\theta(x_{u_{j-1}}, x_{u_j})\, \lambda(\mathrm{d}x_{u_j})},
\end{aligned}$$

for any $x_{u_{j-1}} \in \mathcal{X}$ and any bounded Borel function f on $\mathcal{X}$ (note that in the second equality, we used the Markov property of the HMT process, see (5.2)); and is defined as $F_j = Q$ for $j > n$. (Note that Assumption 7-(ii) is only used to insure that $p_\theta(y_{T(u_j,n-j)} \,|\, X_{u_j} = x_{u_j})$ is positive, and thus the denominator in the last equality is also positive.)

Note that for all $1 \leq j \leq k \wedge n$, using Assumption 7-(i), the transition kernel F_j satisfies the following Doeblin condition:

$$\frac{\sigma^-}{\sigma^+} \nu_j[y_{T(u_j,n-j)}](f) \leq \mathrm{F}_j[y_{T(u_j,n-j)}](x; f),$$

where for any bounded Borel function f on $\mathcal{X}$, we have:

$$\begin{aligned}
\nu_j[y_{T(u_j,n-j)}](f) &= \mathbb{E}_\theta[f(X_{u_j}) \,|\, Y_{T(u_j,n-j)} = y_{T(u_j,n-j)}] \\
&= \frac{\int_{\mathcal{X}} f(x_{u_j}) p_\theta(y_{T(u_j,n-j)} \,|\, X_{u_j} = x_{u_j})\, \lambda(\mathrm{d}x_{u_j})}{\int_{\mathcal{X}} p_\theta(y_{T(u_j,n-j)} \,|\, X_{u_j} = x_{u_j})\, \lambda(\mathrm{d}x_{u_j})}.
\end{aligned}$$

Note that the difference between the definitions of F_j and ν_j is that the term $q_\theta(x_{\mathrm{p}(u_j)}, x_{u_j})$ has disappear from both the numerator and the denominator of ν_j. Remark that (5.3) also implies the Doeblin condition $\sigma^- \lambda(\cdot) \leq Q(x, \cdot)$ for the transition kernel Q. Thus, Lemma 5.2.7 shows that the Dobrushin coefficient of each transition kernel F_j for $1 \leq j \leq k$ is upper bounded by $\rho = 1 - \sigma^-/\sigma^+$. Therefore, as the Dobrushin coefficient is sub-multiplicative (see Lemma 5.2.5), applying Lemma 5.2.4, we get that (5.19) holds. This concludes the proof. □

To construct the limit of the functions $\mathrm{h}_{u,k,x}(\theta)$ we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when $k \to \infty$. For this lemma, we need the following assumption on the density function g_θ that strengthens Assumption 7-(ii). Remind that $\mathbb{P}_\theta$ denotes the stationary probability distribution under the parameter $\theta \in \Theta$ of the HMT process (X, Y), and by $\mathbb{E}_\theta$ the corresponding expectation.

Assumption 8 (L^1 regularity, [CMR05, Assumption 12.3.1])**.** *Assume that we have:*

(i) $b^+ := 1 \wedge \sup_\theta \sup_{x,y} g_\theta(x,y) < \infty$.

(ii) $\mathbb{E}_{\theta^\star} |\log b^-(Y_\partial)| < \infty$, *where* $b^-(y) := \inf_\theta \int_{\mathcal{X}} g_\theta(x,y) \lambda(\mathrm{d}x)$.

Note that $b^-(y) > 0$ for all $y \in \mathcal{Y}$ by Assumption 6-(ii).

Lemma 5.3.3 (Uniform bounds for $\mathrm{h}_{u,k,x}(\theta)$)**.** *Assume that Assumptions 6–7 and 8-(ii) hold. For all vertices $u \in T$ and all integers $k, k' \in \mathbb{N}^*$, the following assertions hold true:*

$$\sup_{\theta\in\Theta}\sup_{x,x'\in\mathcal{X}} |\mathrm{h}_{u,k,x}(\theta) - \mathrm{h}_{u,k',x'}(\theta)| \leq \frac{\rho^{(k\wedge k')-1}}{1-\rho}, \tag{5.20}$$

$$\sup_{\theta\in\Theta}\sup_{k\in\mathbb{N}^*}\sup_{x\in\mathcal{X}} |\mathrm{h}_{u,k,x}(\theta)| \leq \log b^+ \vee |\log(\sigma^- b^-(Y_u))|. \tag{5.21}$$

Proof. [The proof is a straightforward adaptation of the proof of [CMR05, Lemma 12.3.2] with the use of Lemma 5.3.2 for the coupling.] Remind the definition of $\mathrm{H}_{u,k,x}(\theta)$ in (5.16). Let $k' \geq k \geq 1$, and write $w = \mathrm{p}^k(u)$, $w' = \mathrm{p}^{k'}(u)$. Then, write:

$$\begin{aligned}\mathrm{H}_{u,k,x}(\theta) = \int_{\mathcal{X}\times\mathcal{X}} \left[\int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{\mathrm{p}(u)}, x_u)\lambda(\mathrm{d}x_u)\right] \\ \times \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, Y_{\Delta^*(u,k)}, X_w = x_w) \times \delta_x(\mathrm{d}x_w),\end{aligned} \tag{5.22}$$

and using the Markov property at X_w, write:

$$\begin{aligned}\mathrm{H}_{u,k',x'}(\theta) = \int_{\mathcal{X}\times\mathcal{X}} \left[\int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{\mathrm{p}(u)}, x_u)\lambda(\mathrm{d}x_u)\right] \\ \times \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, Y_{\Delta^*(u,k)}, X_w = x_w) \\ \times \mathbb{P}_\theta(X_w \in \mathrm{d}x_w \,|\, Y_{\Delta^*(u,k')}, X_{w'} = x').\end{aligned} \tag{5.23}$$

Applying Lemma 5.3.2, we get (note that the integrands in (5.22) and (5.23) are non-negative):

$$\begin{aligned}|\mathrm{H}_{u,k,x}(\theta) - \mathrm{H}_{u,k',x'}(\theta)| &\leq \rho^{k-1} \sup_{x_{\mathrm{p}(u)}\in\mathcal{X}} \int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{\mathrm{p}(u)}, x_u)\lambda(\mathrm{d}x_u) \\ &\leq \rho^{k-1}\sigma^+ \int_{\mathcal{X}} g_\theta(x_u, Y_u)\lambda(\mathrm{d}x_u).\end{aligned} \tag{5.24}$$

The integral in (5.22) can be lower bounded giving us:

$$\mathrm{H}_{u,k,x}(\theta) \geq \sigma^- \int_{\mathcal{X}} g_\theta(x_u, Y_u)\lambda(\mathrm{d}x_u), \tag{5.25}$$

where the right hand side is positive by Assumption 7-(ii); and similarly for (5.23). Combining (5.24) with (5.25), and with the inequality $|\log x - \log y| \leq |x-y|/(x\wedge y)$, we get the first assertion of the lemma:

$$|\mathrm{h}_{u,k,x}(\theta) - \mathrm{h}_{u,k',x'}(\theta)| \leq \frac{\sigma^+}{\sigma^-}\rho^{k-1} = \frac{\rho^{k-1}}{1-\rho}.$$

Combining (5.16) and (5.25), we get that $\sigma^- b^-(Y_u) \leq \mathrm{H}_{u,k,x}(\theta) \leq b^+$ (remind that $b^-(Y_u) > 0$ by Assumption 7-(ii)), which yields the second assertion of the lemma. □

We are now ready to construct the limit of the functions $\mathrm{h}_{u,k,x}(\theta)$ and state some properties of this limit. Note that this result is stated for every $u \in T$, but we will only need it for $u = \partial$. Remind that we are in the stationary case, and that the HMT process (X, Y) is defined on T^∞.

Proposition 5.3.4 (Properties of the limit function $\mathrm{h}_{u,\infty}(\theta)$)**.** *Assume that Assumptions 5–8 hold. For every $u \in T$ and $\theta \in \Theta$, there exists $\mathrm{h}_{u,\infty}(\theta) \in L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ such that for all $x \in \mathcal{X}$, the sequence $(\mathrm{h}_{u,k,x}(\theta))_{k\in\mathbb{N}}$ converges $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ to $\mathrm{h}_{u,\infty}(\theta)$.*

Furthermore, this convergence is uniform over $\theta \in \Theta$ and $x \in \mathcal{X}$, that is, we have $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ that:

$$\lim_{k\to\infty}\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}} |\mathrm{h}_{u,k,x}(\theta) - \mathrm{h}_{u,\infty}(\theta)| = 0.$$

The limit function $\mathrm{h}_{u,\infty}(\theta)$ can be interpreted as $\log p_\theta(Y_u \,|\, Y_{\Delta^*(u,\infty)})$, where $\Delta^*(u,\infty) = \{v \in T^\infty : v <_\mathcal{U} u\}$ is a random subset of vertices. Note that $\mathrm{h}_{u,\infty}(\theta)$ is a function of the random set of variables $(Y_v, v \in \Delta(u,\infty))$, where $\Delta(u,\infty) = \Delta^*(u,\infty) \cup \{u\}$, and thus implicitly depend on $\mathcal{U}$ trough $\Delta(u,\infty)$.

Proof. Fix some $u \in T$. Note that (5.20) shows that the sequence $(\mathrm{h}_{u,k,x}(\theta))_{k\in\mathbb{N}}$ is Cauchy uniformly in θ and x, and thus has $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-almost surely a limit when $k \to \infty$ which does not depend on x; we denote this limit by $\mathrm{h}_{u,\infty}(\theta)$. Furthermore, we get from (5.21) that $(\mathrm{h}_{u,k,x}(\theta))_{k\in\mathbb{N}}$ is uniformly bounded in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$, and thus $\mathrm{h}_{u,\infty}(\theta)$ is in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ and the convergence also holds in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$. Finally, as the bound in (5.20) is uniform in θ and x, we get that the convergence holds uniform over θ and x both $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-almost surely and in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$. □

5.3.3 Properties of the contrast function

As the functions $\mathrm{h}_{\partial,\infty}(\theta)$ are in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ under the assumptions used in Proposition 5.3.4, we can now define the *contrast function* ℓ (which is deterministic) as:

$$\ell(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,\infty}(\theta)\big], \tag{5.26}$$

where remind $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}$ is the expectation corresponding to $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$.

We prove under the following L^2 regularity assumption the convergence of the normalized log-likelihood to the contrast function. Remind that $b^-(y) = \inf_\theta \int_{\mathcal{X}} g_\theta(x,y)\lambda(\mathrm{d}x)$. Also remind that $\mathbb{P}_\theta$ denotes the stationary probability distribution under the parameter $\theta \in \Theta$ of the HMT process (X,Y), and by $\mathbb{E}_\theta$ the corresponding expectation.

Assumption 9 (L^2 regularity)**.** *Assume that* $\mathbb{E}_{\theta^\star}\big[(\log b^-(Y_\partial))^2\big] < \infty$.

Remind that the log-likelihood $\ell_{n,x}$ is defined in (5.7) on page 125.

Proposition 5.3.5 (Ergodic convergence for the stationary log-likelihood)**.** *Assume that Assumptions 5–9 hold. Then, for all $x \in \mathcal{X}$, the normalized log-likelihood $|T_n|^{-1}\ell_{n,x}(\theta)$ converges $\mathbb{P}_{\theta^\star}$-a.s. to the contrast function $\ell(\theta)$ as $n \to \infty$.*

Proof. Let $\theta \in \Theta$ be some parameter. Fix some $k \in \mathbb{N}^*$ and $x \in \mathcal{X}$. Remind from (5.18) that $\ell_{n,x}(\theta) = \sum_{u\in T_n} \mathrm{h}_{u,h(u),x}(\theta)$. Applying (5.20) for each vertex $u \in T_n \setminus T_{k-1}$, we get:

$$\frac{1}{|T_n|}\left|\ell_{n,x}(\theta) - \sum_{u\in T_n\setminus T_{k-1}} \mathrm{h}_{u,k,x}(\theta)\right| \le \frac{\rho^{k-1}}{1-\rho} + \frac{1}{|T_n|}\sum_{u\in T_{k-1}} |\mathrm{h}_{u,h(u),x}(\theta)|. \tag{5.27}$$

Note that by (5.21), we have that $|\mathrm{h}_{u,h(u),x}(\theta)| < \infty$ $\mathbb{P}_{\theta^\star}$-a.s. for all $u \in T \setminus \{\partial\}$. For $u = \partial$, we have $\mathrm{h}_{\partial,0,x}(\theta) = \log g_\theta(x, Y_\partial)$ which is finite $\mathbb{P}_{\theta^\star}$-a.s. by Assumption 7-(iii).

For a vertex u in $T \setminus T_{k-1}$, let $v_u \in G_k$ be the unique vertex that satisfies (5.8) (on page 126), then we have:

$$\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \mathrm{h}_{v_u,k,x}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{5.28}$$

Moreover, using (5.21) together with Assumption 9, we get for every $u \in T \setminus T_{k-1}$ that the random variable $\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$. Hence, applying Lemma 5.2.11 to the collection of neighborhood-shape-dependent functions $(\mathrm{h}_{v,k,x}(\theta; Y_{\Delta(v)} = \cdot))_{v\in G_k}$ (remind that indexing functions with G_k or with $\mathcal{N}_k$ is equivalent by (5.9)), and using (5.28) and (5.14) (in Remark 5.3.1), we get:

$$|T_n|^{-1}\sum_{u\in T_n\setminus T_{k-1}} \mathrm{h}_{u,k,x}(\theta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,k,x}(\theta)\big] \qquad \mathbb{P}_{\theta^\star}\text{-a.s. (and in } L^2(\mathbb{P}_{\theta^\star})). \tag{5.29}$$

Using (5.20) with Proposition 5.3.4, we get:

$$\big|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,k,x}(\theta)\big] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,\infty}(\theta)\big]\big| \le \frac{\rho^{k-1}}{1-\rho}.$$

Thus, combining this bound with (5.27) and (5.29), we get $\mathbb{P}_{\theta^\star}$-a.s. that:

$$\limsup_{n\to\infty}\Big||T_n|^{-1}\ell_{n,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,\infty}(\theta)\big]\Big| \le 2\,\frac{\rho^{k-1}}{1-\rho}.$$

As the left hand side does not depend on k, letting $k \to \infty$, we get that $|T_n|^{-1}\ell_{n,x}(\theta)$ converges $\mathbb{P}_{\theta^\star}$-a.s. to $\ell(\theta)$ as $n \to \infty$. This concludes the proof. □

We are going to prove that this convergence holds uniformly in θ. First, we need to prove that the contrast function is continuous and has a unique global maximum at $\theta^\star$. In order to get those results, we need a natural continuity assumption on the transition functions.

Assumption 10 (Continuity, [CMR05, Assumption 12.3.5])**.** *For all $(x, x') \in \mathcal{X} \times \mathcal{X}$ and $y \in \mathcal{Y}$, the functions $\theta \mapsto q_\theta(x', x)$ and $\theta \mapsto g_\theta(x, y)$ defined on $\Theta \subset \mathbb{R}^d$ are continuous.*

We denote by $\|\cdot\|$ the euclidean norm on $\mathbb{R}^d$.

Proposition 5.3.6 (ℓ is continuous)**.** *Assume that Assumptions 5–8 and 10 hold. Then, for any $n \in \mathbb{N}$ and $x \in \mathcal{X}$, the log-likelihood function $\theta \mapsto \ell_{n,x}(\theta)$ is $\mathbb{P}_{\theta^\star}$-a.s. continuous on Θ.*

Moreover, for any $\theta \in \Theta$, we have:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta' \in \Theta : \|\theta - \theta'\| \le \delta} \left|\mathrm{h}_{\partial,\infty}(\theta') - \mathrm{h}_{\partial,\infty}(\theta)\right|\right] \to 0 \quad \text{as } \delta \to 0,$$

and the contrast function $\theta \mapsto \ell(\theta)$ is continuous on Θ.

Proof. This proof is a straightforward adaptation from the proof of [CMR05, Proposition 12.3.6].

Recall that $\mathrm{h}_{\partial,\infty}(\theta)$ is the limit of $\mathrm{h}_{\partial,k,x}(\theta)$ as $k \to \infty$. We first prove that, for every $x \in \mathcal{X}$ and $k \ge 0$, $\mathrm{h}_{\partial,k,x}(\theta)$ is a continuous function of θ, and then use this to show continuity of the limit. Recall from (5.16) the second equality defining $\mathrm{H}_{u,k,x}(\theta)$, which we remind for convenience for any $u \in T$ and $x \in \mathcal{X}$:

$$\mathrm{H}_{u,k,x}(\theta) = \frac{p_\theta(Y_{\Delta(u,k)} \,|\, X_{\mathrm{p}^k(u)} = x)}{p_\theta(Y_{\Delta^*(u,k)} \,|\, X_{\mathrm{p}^k(u)} = x)}.$$

Recall from (5.15) the definition of $p_\theta(Y_\Delta \,|\, X_{\mathrm{p}^k(u)} = x)$ where here the possibly random subtree Δ is either $\Delta(u,k)$ or $\Delta^*(u,k)$. First note that the integrand in (5.15) is by assumption continuous w.r.t. θ and upper bounded by $(1 \vee \sigma^+ b^+)^{|\Delta|}$. Thus, dominated convergence shows that $p_\theta(Y_\Delta \,|\, X_{\mathrm{p}^k(u)} = x)$ is continuous w.r.t. to θ (remind that λ, defined in Assumption 6, is finite). Moreover, note that $p_\theta(Y_{\Delta^*(u,k)} \,|\, X_{\mathrm{p}^k(u)} = x)$ is lower bounded by $\prod_{v \in \Delta^*(u,k) \setminus \{\mathrm{p}^k(u)\}} \sigma^- b^-(Y_v)$ which is positive $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. (by Assumption 7). Thus, $\mathrm{H}_{u,k,x}(\theta)$ and $\mathrm{h}_{u,k,x}(\theta) = \log \mathrm{H}_{u,k,x}(\theta)$ (remind (5.17)) are continuous w.r.t. θ $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. as well. Hence, using (5.6), for all $n \in \mathbb{N}$ and $x \in \mathcal{X}$, we get that $\ell_{n,x}(\theta)$ is also continuous w.r.t. θ $\mathbb{P}_{\theta^\star}$-a.s.

Remind from Proposition 5.3.4 that $(\mathrm{h}_{u,k,x}(\theta))_{k \in \mathbb{N}}$ converges to $\mathrm{h}_{u,\infty}(\theta)$ uniformly in θ $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. Thus, the function $\theta \mapsto \mathrm{h}_{u,\infty}(\theta)$ is continuous $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. Using the uniform bound (5.21), Assumption 8-(ii) and dominated convergence, we obtain the first part of the proposition.

We deduce the second part from the first one, as:

$$\begin{aligned}\sup_{\theta' \in \Theta : \|\theta' - \theta\| \le \delta} |\ell(\theta') - \ell(\theta)| &= \sup_{\theta' \in \Theta : \|\theta' - \theta\| \le \delta} \left|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\mathrm{h}_{\partial,\infty}(\theta') - \mathrm{h}_{\partial,\infty}(\theta)\right]\right| \\ &\le \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta' \in \Theta : \|\theta - \theta'\| \le \delta} |\mathrm{h}_{\partial,\infty}(\theta') - \mathrm{h}_{\partial,\infty}(\theta)|\right].\end{aligned}$$

This concludes the proof. □

We are now ready to state and prove that the convergence to the contrast function ℓ holds uniformly in θ.

Proposition 5.3.7 (Uniform convergence to ℓ)**.** *Assume that Assumptions 5–10 hold and Θ is compact. Then, we have:*

$$\lim_{n \to \infty} \sup_{\theta \in \Theta} \left| |T_n|^{-1} \ell_{n,x}(\theta) - \ell(\theta) \right| = 0 \quad \mathbb{P}_{\theta^\star}\text{-a.s.}$$

Proof. [We mimic the proof of [CMR05, Proposition 12.3.7].] As Θ is compact, it is sufficient to prove that for every $\theta \in \Theta$:

$$\limsup_{\delta \to 0} \limsup_{n \to \infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \le \delta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \right| = 0 \quad \mathbb{P}_{\theta^\star}\text{-a.s.} \tag{5.30}$$

As this claim is not proven in the proof of [CMR05, Proposition 12.3.7], we give a short proof. Indeed, assume that (5.30) holds for all $\theta \in \Theta$. Let $\varepsilon > 0$. By Proposition 5.3.6, the function ℓ is continuous, and thus uniformly continuous as Θ is compact. In particular, there exists $\delta > 0$ such that for all $\theta, \theta' \in \Theta$, we have that $\|\theta - \theta'\| \leq \delta$ implies $|\ell(\theta) - \ell(\theta')| \leq \varepsilon$. For every $\theta \in \Theta$, let $\delta_\theta \in (0, \delta)$ be such that $\limsup_{n\to\infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta_\theta} \Big| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \Big| < \varepsilon$. As $\cup_{\theta \in \Theta} \{\theta' : \|\theta' - \theta\| \leq \delta_\theta\}$ is an open cover of Θ and as Θ is compact, there exists a finite subset $\{\theta_j : 1 \leq j \leq m\}$ of Θ with $m \geq 1$ such that $\Theta = \cup_{j=1}^m \{\theta' : \|\theta' - \theta_j\| \leq \delta_{\theta_j}\}$. Note that for n large enough, for all $1 \leq j \leq m$, we have that $\sup_{\theta' \in \Theta : \|\theta' - \theta_j\| \leq \delta_{\theta_j}} \Big| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta_j) \Big| < \varepsilon$. Thus, for n large enough, we have:

$$\sup_{\theta \in \Theta} \Big| |T_n|^{-1} \ell_{n,x}(\theta) - \ell(\theta) \Big| \leq \varepsilon + \max_{1 \leq j \leq m} \sup_{\theta' \in \Theta : \|\theta' - \theta_j\| \leq \delta_{\theta_j}} \Big| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta_j) \Big| \leq 2\varepsilon.$$

This being true for all $\varepsilon > 0$, we get that the statement in the proposition holds.

We now prove (5.30). Fix some $\theta \in \Theta$. By Proposition 5.3.5, remind that $\lim_{n\to\infty} |T_n|^{-1} \ell_n(\theta) = \ell(\theta)$ $\mathbb{P}_{\theta^\star}$-a.s. Using this fact, we get:

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \Big| \\ &\qquad = \limsup_{n\to\infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| |T_n|^{-1} \ell_{n,x}(\theta') - |T_n|^{-1} \ell_{n,x}(\theta) \Big|. \end{aligned} \tag{5.31}$$

Using (5.27), for any $k \geq 1$, we get that (5.31) is $\mathbb{P}_{\theta^\star}$-a.s. bounded by:

$$\begin{aligned} 2 \limsup_{n\to\infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} & |T_n|^{-1} \Big| \ell_{n,x}(\theta') - \sum_{u \in T_n \setminus T_{k-1}} \mathrm{h}_{u,k,x}(\theta') \Big| \\ & + \limsup_{n\to\infty} |T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| \mathrm{h}_{u,k,x}(\theta') - \mathrm{h}_{u,k,x}(\theta) \Big| \\ & \leq 2 \frac{\rho^{k-1}}{1-\rho} + \limsup_{n\to\infty} |T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| \mathrm{h}_{u,k,x}(\theta') - \mathrm{h}_{u,k,x}(\theta) \Big| \\ & = 2 \frac{\rho^{k-1}}{1-\rho} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| \mathrm{h}_{\partial,k,x}(\theta') - \mathrm{h}_{\partial,k,x}(\theta) \Big| \right] \\ & \leq 4 \frac{\rho^{k-1}}{1-\rho} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} \Big| \mathrm{h}_{\partial,\infty}(\theta') - \mathrm{h}_{\partial,\infty}(\theta) \Big| \right], \end{aligned} \tag{5.32}$$

where we used Lemma 5.2.11 for ergodic convergence (with $L^2(\mathbb{P}_{\theta^\star})$ boundedness given by (5.21)) in the equality, and we used (5.20) (with Proposition 5.3.4) in the second inequality. Then, letting $k \to \infty$ in the upper bound of (5.32) (note that (5.31) does not depend on k), and then letting $\delta \to 0$ with Proposition 5.3.6, we get that (5.30) holds. This concludes the proof. □

Remark 5.3.8 (Uniform convergence for the log-likelihood with general initial condition)**.** Let ν be a probability distribution on $\mathcal{X}$ such that $\sup_\theta |\int g_\theta(x, Y_\partial) \nu(\mathrm{d}x)|$ is finite $\mathbb{P}_{\theta^\star}$-a.s. The uniform convergence of $|T_n|^{-1} \ell_{n,x}(\theta)$ to $\ell(\theta)$ still holds when modifying the definition of the log-likelihood $\ell_{n,x}(\theta)$ of the HMT to replace the Dirac mass δ_x by ν for the distribution of the root hidden variable X_∂. When ν is the stationary distribution π_θ associated to q_θ, uniform convergence holds without this extra regularity assumption by conditioning on the state of the root's parent $X_{\mathrm{p}(\partial)}$ instead (which allows to replace $\mathrm{h}_{\partial,0,x}(\theta) = g_\theta(x, Y_\partial)$ in (5.27) by $\mathrm{h}_{\partial,1,\nu}(\theta) := \log \int_{\mathcal{X}} \mathrm{H}_{\partial,1,x}(\theta)\, \nu(\mathrm{d}x)$ for which $\sup_\theta |\mathrm{h}_{\partial,1,\nu}(\theta)|$ is finite by an immediate adaptation of (5.21)).

5.3.4 Identifiability and strong consistency

In this subsection, we prove the strong consistency of the MLE. We must first study the identifiability of the parameter of the HMT model. We start with a definition of equivalent parameters.

Definition 5.3.9 (Equivalent parameters)**.** *We say that two parameters $\theta, \theta' \in \Theta$ are* equivalent *if they define the same distribution for the process $(Y_u, u \in T)$, i.e. $\mathbb{P}_\theta(Y \in \cdot) = \mathbb{P}_{\theta'}(Y \in \cdot)$.*

Note that by Kolmogorov's extension theorem, θ and θ' are equivalent if and only if they define the same law on every finite tree T_n, i.e. for $(Y_u, u \in T_n)$.

The following proposition characterizes global maxima of the contrast function ℓ.

Proposition 5.3.10 (Global maxima of the contrast function ℓ)**.** *Assume that Assumptions 5–9 hold. Then a parameter $\theta \in \Theta$ is a global maximum of ℓ if and only if θ is equivalent to $\theta^\star$.*

We get as an immediate corollary that $\theta^\star$ is a global maximum of ℓ.

The proof of Proposition 5.3.10, which is postponed to the end of this section, is an adaptation of the proof of [CMR05, Theorem 12.4.2]. This adaptation comes from the difference of topology between the tree and the line.

Remind that the log-likelihood function $\theta \mapsto \ell_{n,x}(\theta)$ is continuous $\mathbb{P}_{\theta^\star}$-a.s. under Assumptions 5-8 and 10. Thus, when we further assume that Θ is compact, we get that the argmax set $\operatorname{argmax}_{\theta\in\Theta} \ell_{n,x}(\theta)$ is non-empty. The maximum likelihood estimator (MLE) is then defined as the maximizer over Θ of the log-likelihood $\ell_{n,x}$, that is as the following random variable (which depends on Y_{T_n}):

$$\hat{\theta}_{n,x} = \hat{\theta}_{n,x}(Y_{T_n}) \in \operatorname{argmax}_{\theta\in\Theta} \ell_{n,x}(\theta). \tag{5.33}$$

Note that the argmax set in (5.33) is not necessarily unique, in which case we select one parameter θ from the argmax set in a measurable manner (which is possible, see [BS96, Proposition 7.33]).

We are now ready to prove the following theorem that states the strong consistency of the MLE for the HMT model in the stationary case.

Theorem 5.3.11 (Strong consistency of the MLE)**.** *Assume that Assumptions 5–10 hold, the contrast function ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 5.3.10) and Θ is compact. Then, for any $x \in \mathcal{X}$, the MLE $\hat{\theta}_{n,x}$ (defined in* (5.33)*) converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to the true parameter $\theta^\star \in \Theta$, i.e. the MLE is strongly consistent.*

Proof. [The proof is a straightforward adaptation of an argument for HMMs in [CMR05, Section 12.1], which itself adapts an argument that goes back to [Wal49].]

By definition of $\hat{\theta}_n$, we have that $\ell_{n,x}(\hat{\theta}_n) \geq \ell_{n,x}(\theta)$ for every $\theta \in \Theta$. As the contrast function ℓ has a unique maximum located at $\theta^\star$, we have that $\ell(\theta^\star) \geq \ell(\theta)$ for every $\theta \in \Theta$, and in particular, $\ell(\theta^\star) \geq \ell(\hat{\theta}_n)$ for every $n \in \mathbb{N}$. Combining those two bounds, we get that:

$$\begin{aligned}
0 &\leq \ell(\theta^\star) - \ell(\hat{\theta}_n) \\
&\leq \ell(\theta^\star) - |T_n|^{-1}\ell_{n,x}(\theta^\star) + |T_n|^{-1}\ell_{n,x}(\theta^\star) - |T_n|^{-1}\ell_{n,x}(\hat{\theta}_n) + |T_n|^{-1}\ell_{n,x}(\hat{\theta}_n) - \ell(\hat{\theta}_n) \\
&\leq 2\sup_{\theta\in\Theta}\Big|\ell(\theta) - |T_n|^{-1}\ell_{n,x}(\theta)\Big|,
\end{aligned}$$

where the upper bound in the last line goes to zero $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ by Proposition 5.3.7 as Θ is compact. Hence, we get that $\ell(\hat{\theta}_n) \to \ell(\theta^\star)$ $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$. Consequently, as ℓ is continuous (by Proposition 5.3.6) and has a unique global maximum located at $\theta^\star$, and as Θ is compact, we get that $\hat{\theta}_n$ converges $\mathbb{P}_{\theta^\star}$-a.s. to $\theta^\star$ as $n \to \infty$. □

We now prove Proposition 5.3.10.

Proof of Proposition 5.3.10. Remind that $\mathrm{h}_{u,k,x}(\theta)$ is defined in (5.17). By definition of $\ell(\theta)$ (see (5.26)) and using the $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ convergence of $(\mathrm{h}_{\partial,k,x}(\theta))_{k\in\mathbb{N}}$ to $\mathrm{h}_{\partial,\infty}(\theta)$ (remind Proposition 5.3.4), we have:

$$\begin{aligned}
\ell(\theta^\star) - \ell(\theta) &= \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,\infty}(\theta^\star) - \mathrm{h}_{\partial,\infty}(\theta)\big] \\
&= \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^\star}\big[\lim_{k\to\infty}\big(\mathrm{h}_{\partial,k,x}(\theta^\star) - \mathrm{h}_{\partial,k,x}(\theta)\big)\big] \\
&= \lim_{k\to\infty} \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{\partial,k,x}(\theta^\star) - \mathrm{h}_{\partial,k,x}(\theta)\big].
\end{aligned}$$

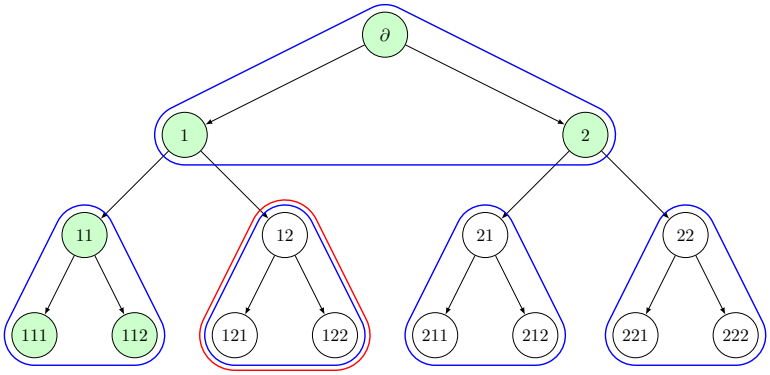


Figure 5.5 – Illustration of the "past" subtree $\Delta^*(T(u,m),k)$ of the block subtree $T(u,m)$ for $m=1$, $u=12$ and $k=1$. The block subtrees are circled with blue lines, and the block subtree $T(12,1)$ is circled a second time with a red line. The vertices in green are those in $\Delta^*(T(12,1),1)$. Note the difference with $\Delta(u',k')$, e.g. vertex 111 is in $\Delta^*(T(12,1),1)$ but not in $\Delta^*(12,2)$, and vertex 21 is in $\Delta^*(121,3)$ but not in $\Delta^*(T(12,1),1)$.

Remind that $\mathrm{H}_{u,k,x}(\theta)$ is defined in (5.16). Then, write:

$$\mathbb{E}_{\theta^\star}[\mathrm{h}_{\partial,k,x}(\theta^\star)-\mathrm{h}_{\partial,k,x}(\theta)]=\mathbb{E}_{\mathcal{U}}\left[\mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}\left[\log\frac{\mathrm{H}_{\partial,k,x}(\theta^\star)}{\mathrm{H}_{\partial,k,x}(\theta)}\,\middle|\,Y_{\Delta^*(\partial,k)},X_{\mathrm{p}^k(\partial)}=x\right]\right]\right], \tag{5.34}$$

where the inner expectation is on Y_∂ conditionally on $X_{\mathrm{p}^k(\partial)}=x$ and $Y_{\Delta^*(\partial,k)}$ (and thus also implicitly on $\mathcal{U}$ as $\Delta^*(\partial,k)=\Delta^*_{\mathcal{U}}(\partial,k)$). Recalling from (5.16) that $\mathrm{H}_{\partial,k,x}(\theta)$ is the conditional density of Y_∂ given $Y_{\Delta^*(\partial,k)}$ and $X_{\mathrm{p}^k(\partial)}=x$, we see that the inner (conditional) expectation in the right hand side is a Kullback-Leibler divergence and thus is non-negative. Hence, the two outer expectations and the limit $\ell(\theta^\star)-\ell(\theta)$ as $k\to\infty$ are non-negative as well, and thus $\theta^\star$ is a global maximum of ℓ.

Remark that if θ is equivalent to $\theta^\star$, then as the process $(Y_u,u\in T)$ is stationary and has same law under both parameters, the roles of $\theta^\star$ and θ can be swapped in the argument above, and thus we get $\ell(\theta)=\ell(\theta^\star)$. Hence, any θ equivalent to $\theta^\star$ is a global maximum of ℓ.

We now turn to prove that any global maximum $\theta\in\Theta$ of ℓ is equivalent to $\theta^\star$.

Remind that we use the letter p to denote (possibly conditional) densities of random variables, e.g. $p_\theta(Y_u\,|\,Y_{\Delta^*(u,k)},X_{\mathrm{p}^k(u)}=x)$ denotes the *conditional density* (w.r.t. the measure μ defined in Assumption 6-(i)) under the parameter θ of Y_u conditionally on $Y_{\Delta^*(u,k)}$ and $X_{\mathrm{p}^k(u)}=x$. Note that $\mathbb{P}_\theta(Y_u\in\cdot\,|\,Y_{\Delta^*(u,k)},X_{\mathrm{p}^k(u)}=x)$ denotes the *distribution* under the parameter θ of Y_u conditionally on $Y_{\Delta^*(u,k)}$ and $X_{\mathrm{p}^k(u)}=x$.

We first need a variant of the convergence in Proposition 5.3.4 where instead of considering one vertex u as in $\mathrm{h}_{u,k,x}(\theta)$ we consider a whole subtree $T^\infty(u,m)$ for any $m\geq 1$ (this can be seen as a convergence by block). To this end, we need to define an analogue of the breadth-first-search order relation $<$ on T^∞ for subtree blocks of the form $T^\infty(u,m)$. Let $m\geq 1$ be fixed. For $u,v\in T^\infty$ with $h(u)\equiv h(v) \mod m+1$, we write $T^\infty(u,m)<T^\infty(v,m)$ if $u<v$ (informally, "$T^\infty(u,m)$ is above or on the left of $T^\infty(v,m)$"). Note that the modulo congruence is there to insure the collection of block subtrees $T^\infty(u,m)$ with $h(u)\equiv h(\partial) \mod m+1$ form a partition (i.e. a cover with non-overlapping subsets) of T^∞ (this still holds for any other class of congruence $\mod m+1$). Also note that in this congruence we have $m+1$ and not m, because any subtree $T^\infty(u,m)$ (e.g. $T_m=T^\infty(\partial,m)$) spans over $m+1$ different generations (remind that $h(\partial)=0$). We can then define the analogue of the subset $\Delta^*(u,k)$ for subtree blocks, that is, for all $u\in T^\infty$ and $k\in\mathbb{N}$, we define:

$$\Delta^*(T(u,m),k)=\bigcup\big\{T(v,m):v\in\Delta^*(u,k(m+1))\text{ such that }h(v)\equiv h(u) \mod m+1\big\}.$$

See Figure 5.5 for an illustration of the "past" subtree $\Delta^*(T(u,m),k)$ of the block subtree $T(u,m)$. (Informally, "the subset $\Delta^*(T(u,m),k)$ is the union of the subtree blocks (with height m) above and on the left of $T(u,m)$ up to k block generations". Note that we will not need to understand in details the

geometry of the subset $\Delta^*(T(u,m),k)$, we only need to remember that all its vertices are upstream of the edge $(\mathrm{p}(u),u)$, and we will then use the Markov property.) Remind that $T^\infty(\partial,m)=T_m$. Then, a straightforward adaptation of Lemma 5.3.3, and Propositions 5.3.4 and 5.3.5 to a decomposition of the log-likelihood into non-overlapping subtrees of height m instead of single vertices (see Section 5.C for detailed proofs of those adaptations) give us for all $\theta\in\Theta$, $x\in\mathcal{X}$ and $m\in\mathbb{N}^*$:

$$\lim_{k\to\infty}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log p_\theta(Y_{T_m}\,|\,Y_{\Delta^*(T_m,k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)\right]=|T_m|\,\ell(\theta). \tag{5.35}$$

Let $\mathcal{U}^+=(\mathcal{U}_{(j)})_{1\le j<\infty}$ be a sequence of independent random variables with Bernoulli distribution of parameter $1/2$ (note that $\mathcal{U}^+$ can be seen as a random forward spine), which is independent of $\mathcal{U}$ and of the HMT process (X,Y). For all $n\in\mathbb{N}^*$, define the random vertex U_n as the unique vertex in G_n whose path from ∂ is encoded by $\mathcal{U}_{(1:n)}$ in Neveu's notation. For all $n\in\mathbb{N}$, define the deterministic vertex $U_{-n}=\mathrm{p}^n(\partial)$. Note that $\partial=U_0$ and that U_{n-1} is the parent vertex of U_n for all $n\in\mathbb{Z}$. Moreover, using a similar argument as in Remark 5.3.1, note that for any $m,k\in\mathbb{N}$, the sequence of random shapes $(\mathcal{S}h(\Delta(T^\infty(U_n,m),k)))_{n\in\mathbb{Z}}$ is stationary.

Now, pick $\theta\in\Theta$ such that $\ell(\theta)=\ell(\theta^\star)$. Thus for any positive integer $n<m$, we have:

$$\begin{aligned}
0&=|T_m|\,(\ell(\theta^\star)-\ell(\theta))\\
&=\lim_{k\to\infty}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log\frac{p_{\theta^\star}(Y_{T_m}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}{p_\theta(Y_{T_m}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}\right]\\
&=\lim_{k\to\infty}\left\{\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log\frac{p_{\theta^\star}(Y_{T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}{p_\theta(Y_{T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}\right]\right.\\
&\qquad\left.+\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log\frac{p_{\theta^\star}(Y_{T_m\setminus T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)\cup T(U_{m-n},n)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}{p_\theta(Y_{T_m\setminus T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)\cup T(U_{m-n},n)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}\right]\right\}\\
&\ge\limsup_{k\to\infty}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log\frac{p_{\theta^\star}(Y_{T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}{p_\theta(Y_{T(U_{m-n},n)}\,|\,Y_{\Delta^*(T(\partial,m),k)},X_{\mathrm{p}^{k(m+1)}(\partial)}=x)}\right]\\
&=\limsup_{k\to\infty}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\log\frac{p_{\theta^\star}(Y_{T_n}\,|\,Y_{\Delta^*(T(U_{-m+n},m),k)},X_{\mathrm{p}^{k(m+1)}(U_{-m+n})}=x)}{p_\theta(Y_{T_n}\,|\,Y_{\Delta^*(T(U_{-m+n},m),k)},X_{\mathrm{p}^{k(m+1)}(U_{-m+n})}=x)}\right],
\end{aligned}\tag{5.36}$$

where the inequality follows by noting that the second term is non-negative as an expectation of a (conditional) Kullback-Leibler divergence (using an argument similar as for (5.34) above), and the last equality follows by using stationarity of the HMT process (X,Y), of the spinal process $(U_n)_{n\in\mathbb{Z}}$, and of the shape process $(\mathcal{S}h(\Delta(T^\infty(U_n,m),k)))_{n\in\mathbb{Z}}$. Note that the term in the lower bound is also non-negative as an expectation of a (conditional) Kullback-Leibler divergence.

Let $n\in\mathbb{N}$ be fixed. Now, we define for all $\theta\in\Theta$ and $m,k\in\mathbb{N}^*$:

$$\begin{aligned}
&W_{m,k}(\theta)=\log p_\theta(Y_{T_n}\,|\,Y_{\Delta^*(T(U_{-m},m+n),k)},X_{\mathrm{p}^{k(m+n+1)}(U_{-m})}=x),\\
\text{and}\qquad&W(\theta)=\log p_\theta(Y_{T_n}).
\end{aligned}\tag{5.37}$$

Note that $\log p_\theta(Y_{T_n})$ is well defined using an integral expression similar to (5.5) along Assumptions 6 and 7 and the comment on π_θ after Lemma 5.2.3. From (5.36), we deduce that (where m in (5.37) and (5.38) below corresponds to $m-n$ in (5.36)):

$$\forall m\in\mathbb{N}^*,\qquad\lim_{k\to\infty}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[W_{m,k}(\theta^\star)-W_{m,k}(\theta)\right]=0. \tag{5.38}$$

Hence, we have managed to insert a gap between the variables $(Y_v,v\in T_n)$ whose density we examine and the variables $(Y_v,v\in\Delta^*(T(U_{-m},m+n),k))$ and $X_{\mathrm{p}^{k(m+n+1)}(U_{-m})}$ that appear in the conditioning. Remark that the following fact illustrates the gap between the variables: if $u\in T_n$ and $v\in\Delta^*(T(U_{-m},m+n),k)$, then the most recent common ancestor $u\wedge v$ of u and v has height $h(u\wedge v)<-m$, that is $u\wedge v$ is an ancestor of U_{-m}. See Figure 5.6 for a graphical illustration of this gap.

The idea is now to let this gap tend to infinity to show that in the limit the conditioning has no effect. Our next goal is thus to prove that:

$$\lim_{m\to\infty}\sup_{k\in\mathbb{N}}\left|\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[W_{m,k}(\theta^\star)-W_{m,k}(\theta)\right]-\mathbb{E}_{\theta^\star}\left[W(\theta^\star)-W(\theta)\right]\right|=0. \tag{5.39}$$

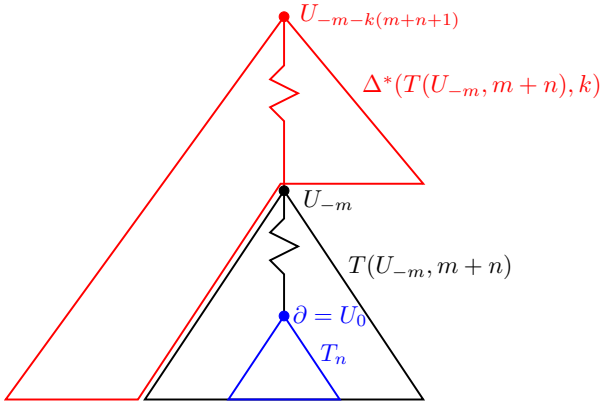


Figure 5.6 – Illustration of the gap in (5.38) between the variables $(Y_v, v \in T_n)$ (bottom triangle in blue) and the variables $(Y_v, v \in \Delta^*(T(U_{-m}, m+n), k))$ and $X_{\mathrm{p}^{k(m++n1)}(U_{-m})}$ that appear in the conditioning (top partial triangle in red). Note that the two groups of variables are separated by the path from $U_{-m-1} = \mathrm{p}(U_{-m})$ to $\partial = U_0$, which is of length $m+1$.

Combining (5.39) with (5.38), it is clear that if $\theta \in \Theta$ is such that $\ell(\theta) = \ell(\theta^\star)$, then we have that $\mathbb{E}_{\theta^\star}[\log[p_{\theta^\star}(Y_{T_n})/p_\theta(Y_{T_n})]] = 0$, that is, the Kullback-Leibler divergence between the $|T_n|$-dimensional densities $p_{\theta^\star}(Y_{T_n})$ and $p_\theta(Y_{T_n})$ is null. This implies, by the information inequality, that these densities coincide except on a set with $\mu^{\otimes|T_n|}$-measure zero, so that the T_n-marginal laws of $\mathbb{P}_{\theta^\star}$ and $\mathbb{P}_\theta$ agree. Because n was arbitrary, we find that $\theta^\star$ and θ are equivalent.

What remains to do to complete the proof is thus to prove (5.39). Remind the definition of $W_{m,k}(\theta)$ and $W(\theta)$ in (5.37). Obviously, it is enough to prove that for all $\theta \in \Theta$, we have:

$$\lim_{m\to\infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{k\in\mathbb{N}} |W_{m,k}(\theta) - W(\theta)|\right] = 0. \tag{5.40}$$

Let $\theta \in \Theta$ be fixed. To prove that (5.40) holds for θ, we write (remind the discussion above on the gap between variables):

$$\begin{aligned}\exp(W_{m,k}(\theta)) &= p_\theta(Y_{T_n} \,|\, Y_{\Delta^*(T(U_{-m},m+n),k)}, X_{\mathrm{p}^{k(m+n+1)}(U_{-m})} = x)\\ &= \int_{\mathcal{X}\times\mathcal{X}} p_\theta(Y_{T_n} \,|\, X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)})\, Q_\theta^{m-1}(x_{U_{-m}}; \mathrm{d}x_{\mathrm{p}(\partial)})\\ &\qquad \times \mathbb{P}_\theta(X_{U_{-m}} \in \mathrm{d}x_{U_{-m}} \,|\, Y_{\Delta^*(T(U_{-m},m+n),k)}, X_{\mathrm{p}^{k(m+n+1)}(U_{-m})} = x),\end{aligned}$$

and

$$\exp(W(\theta)) = p_\theta(Y_{T_n}) = \int_{\mathcal{X}\times\mathcal{X}} p_\theta(Y_{T_n} \,|\, X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)})\, Q_\theta^{m-1}(x_{U_{-m}}; \mathrm{d}x_{\mathrm{p}(\partial)})\pi_\theta(\mathrm{d}x_{U_{-m}}),$$

where remind from Lemma 5.2.3 that π_θ is the stationary distribution of the process $(X_u, u \in T^\infty)$ with transition kernel Q_θ (that is, under the distribution $\mathbb{P}_\theta$). Note that we have the upper bound (remind that b^+ is defined in Assumption 8):

$$p_\theta(Y_{T_n} \,|\, X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)}) = \int_{\mathcal{X}^{T_n}} \prod_{u\in T_n} q_\theta(x_{\mathrm{p}(u)}, x_u) g_\theta(x_u, Y_u)\, \lambda(\mathrm{d}x_u) \le (b^+)^{|T_n|}. \tag{5.41}$$

Thus, as Assumptions 6 and 7 hold, applying the uniform geometric bound from Lemma 5.2.3 to the Markov chain $(X_{U_j})_{j\in\mathbb{Z}}$ with transition kernel Q_θ, we obtain $\mathbb{P}_{\theta^\star}$-a.s. :

$$\sup_{k\in\mathbb{N}} \left| p_\theta(Y_{T_n} \,|\, Y_{\Delta^*(T(U_{-m},m+n),k)}, X_{\mathrm{p}^{k(m+n+1)}(U_{-m})} = x) - p_\theta(Y_{T_n}) \right| \le (b^+)^{|T_n|}(1-\sigma^-)^{m-1}. \tag{5.42}$$

Moreover, as we have the lower bound:

$$p_\theta(Y_{T_n} \,|\, X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)}) = \int_{\mathcal{X}^{T_n}} \prod_{u \in T_n} q_\theta(x_{\mathrm{p}(u)}, x_u) g_\theta(x_u, Y_u) \lambda(\mathrm{d}x_u)$$
$$\geq \prod_{u \in T_n} \sigma^- b^-(Y_u), \tag{5.43}$$

this implies that $p_\theta(Y_{T_n} \,|\, Y_{\mathbb{A}^*(T(U_{-m},m+n),k)}, X_{\mathrm{p}^{k(m+n+1)}(U_{-m})} = x)$ and $p_\theta(Y_{T_n})$ both obey the same lower bound. This lower bound combined with the observation that $b^-(Y_u) > 0$ for all $u \in T_n$ (which follows from Assumption 7-(ii)), and the bound $|\log(x) - \log(y)| \leq |x - y|/x \wedge y$, (5.42) shows that:

$$\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}\text{-a.s.} \qquad \lim_{m \to \infty} \sup_{k \in \mathbb{N}} |W_{m,k}(\theta) - W(\theta)| = 0.$$

Using the bounds (5.41) and (5.43) with Assumptions 7 and 8-(ii), we get:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{m \in \mathbb{N}^*} \sup_{k \in \mathbb{N}} |W_{m,k}(\theta)| \right] < \infty.$$

Hence, as this expectation is finite, (5.40) follows from dominated convergence. This concludes the proof. □

5.4 Asymptotic normality of the MLE

In this section, we prove that the MLE for the HMT has asymptotic normal fluctuations. We keep the assumptions used in Section 5.3. This section is divided in two parts: we first prove the asymptotic normality of the score, and then we prove a strong law of large numbers for the observed information. Together with the strong consistency, those two results imply the asymptotic normality of the MLE.

We will need the following assumption for existence and regularity of the gradient and Hessian of the transition kernels. Remind that $\mathbb{P}_\theta$ denotes the stationary probability distribution under the parameter $\theta \in \Theta$ of the HMT process (X, Y), and by $\mathbb{E}_\theta$ the corresponding expectation. Also remind that the measures λ and μ are defined in Assumption 6. We denote by ∇_θ and ∇^2_θ, respectively, the gradient and Hessian operator w.r.t. the parameter $\theta \in \Theta$. With a slight abuse of notations, we denote by $\|\cdot\|$ the euclidean norm on either $\mathbb{R}^d$ or $\mathbb{R}^{d \times d}$.

Assumption 11 (Regularity of the gradient, [CMR05, Assumption 12.5.1])**.** *There exists an open (for the trace topology on $\Theta \subset \mathbb{R}^d$) neighborhood $\mathcal{O} = \{\theta \in \Theta : \|\theta - \theta^\star\| < \delta_0\}$ of $\theta^\star$ such that the following hold.*

(i) For all $(x, x') \in \mathcal{X} \times \mathcal{X}$ and all $y \in \mathcal{Y}$, the functions $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(x, y)$ are twice continuously differentiable on $\mathcal{O}$.

(ii) We have:

$$\sup_{\theta \in \mathcal{O}} \sup_{x,x'} \|\nabla_\theta \log q_\theta(x, x')\| < \infty, \qquad \text{and} \qquad \sup_{\theta \in \mathcal{O}} \sup_{x,x'} \|\nabla^2_\theta \log q_\theta(x, x')\| < \infty.$$

(iii) We have:

$$\mathbb{E}_{\theta^\star} \left[\sup_{\theta \in \mathcal{O}} \sup_{x} \|\nabla_\theta \log g_\theta(x, Y_\partial)\|^2 \right] < \infty, \qquad \text{and} \qquad \mathbb{E}_{\theta^\star} \left[\sup_{\theta \in \mathcal{O}} \sup_{x} \|\nabla^2_\theta \log g_\theta(x, Y_\partial)\| \right] < \infty.$$

(iv) For μ-almost all $y \in \mathcal{Y}$, there exists a function $f_y : \mathcal{X} \to \mathbb{R}_+$ in $L^1(\lambda)$ such that we have $\sup_{\theta \in \mathcal{O}} g_\theta(x, y) \leq f_y(x)$.

(v) For λ-almost all $x \in \mathcal{X}$, there exist functions $f^1_x : \mathcal{X} \to \mathbb{R}_+$ and $f^2_x : \mathcal{X} \to \mathbb{R}_+$ in $L^1(\mu)$ such that $\sup_{\theta \in \mathcal{O}} \|\nabla_\theta g_\theta(x, y)\| \leq f^1_x(y)$ and $\sup_{\theta \in \mathcal{O}} \|\nabla^2_\theta g_\theta(x, y)\| \leq f^2_x(y)$.

These assumptions insures that the log-likelihood $\ell_{n,x}$ is twice continuously differentiable, and that the *score function* $\nabla_\theta \ell_{n,x}(\theta)$ and the *observed information* $-\nabla^2_\theta \ell_{n,x}(\theta)$ exist and are in $L^2(\mathbb{P}_{\theta^\star})$ and $L^1(\mathbb{P}_{\theta^\star})$, respectively.

5.4.1 Asymptotic normality of the score

In this subsection, we prove the asymptotic normality of the score under the true parameter $\theta^\star$. Note that the score function can be written for all $n \in \mathbb{N}$ and $x \in \mathcal{X}$ as:

$$\nabla_\theta \ell_{n,x}(\theta) = \sum_{u \in T_n} \nabla_\theta \log \left[\int g_\theta(X_u, Y_u)\, \mathbb{P}_\theta(X_u \in \mathrm{d}x_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x) \right],$$

and $\nabla_\theta \ell_{n,x}(\theta)$ is implicitly a function of Y_{T_n}.

Decomposition of the score as a sum of increments

Remind that for $u \in T$, the subtrees $\Delta^*(u,k)$ and $\Delta(u,k)$ are defined in Section 5.2.4 for $k \le h(u)$ (with $\Delta^*(u) = \Delta^*(u,h(u))$ and $\Delta(u) = \Delta(u,h(u))$) and the random subtrees $\Delta^*(u,k)$ and $\Delta(u,k)$ are defined in Section 5.3.1 for $k > h(u)$. Also remind that we use the letter p to denote (possibly conditional) probability density, and in particular remind that $p_\theta(Y_\Delta \,|\, X_\partial = x_\partial)$ for any subtree $\Delta \subset T$ with root ∂ is defined in (5.15) in Section 5.3.1 (with the convention $p_\theta(Y_\emptyset \,|\, X_\partial = x_\partial) = 1$). Using (5.16) and (5.17) in Section 5.3.1, note that for any $u \in T$ and $x \in \mathcal{X}$, we have:

$$\mathrm{h}_{u,h(u),x}(\theta) = \log p_\theta(Y_{\Delta(u)} \,|\, X_\partial = x) - \log p_\theta(Y_{\Delta^*(u)} \,|\, X_\partial = x). \tag{5.44}$$

Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 11 (note that this result is also known as *Fisher identity*, see [CMR05, Proposition 10.1.6]), we get:

$$\begin{aligned} &\nabla_\theta \log p_\theta(Y_{\Delta(u)} \,|\, X_\partial = x) \\ &\qquad = \nabla_\theta \log g_\theta(x, Y_\partial) + \mathbb{E}_\theta \left[\sum_{v \in \Delta(u) \setminus \{\partial\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,\middle|\, Y_{\Delta(u)}, X_\partial = x \right], \end{aligned} \tag{5.45}$$

where

$$\phi_\theta(x', x, y) = \nabla_\theta \log\big[q_\theta(x', x) g_\theta(x, y) \big]. \tag{5.46}$$

Note that under Assumption 11, $\|\phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)\|$ is upper bounded by a square integrable function of Y_v (which does not depend on θ), and $\phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)$ is thus integrable conditionally on $Y_{\Delta(u)}$ and $X_\partial = x$. Also note that $\nabla_\theta \log g_\theta(x, Y_\partial)$ is $\mathbb{P}_{\theta^\star}$-a.s. finite by Assumption 11-(iii).

Combining those two equations with (5.18) in Section 5.3.1, we can express the score function as:

$$\begin{aligned} \nabla_\theta \ell_{n,x}(\theta) = \nabla_\theta \log g_\theta(x, Y_\partial) &+ \sum_{u \in T_n^*} \mathbb{E}_\theta \left[\sum_{\Delta(u) \setminus \{\partial\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\Delta(u)}, X_\partial = x \right] \\ &- \sum_{u \in T_n^*} \mathbb{E}_\theta \left[\sum_{\Delta^*(u) \setminus \{\partial\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\Delta^*(u)}, X_\partial = x \right]. \end{aligned}$$

We want to express the score function $\nabla_\theta \ell_{n,x}(\theta)$ as a sum of increments (conditional scores) in order to apply a convergence result for the normalized score. To this end, define for every $u \in T$, $k \in \mathbb{N}$ and $x \in \mathcal{X}$, the function $\dot{\mathrm{h}}_{u,k,x}(\theta)$ by $\dot{\mathrm{h}}_{u,0,x}(\theta) = \nabla_\theta \log g_\theta(x, Y_u)$ if $k = 0$, and otherwise by:

$$\begin{aligned} \dot{\mathrm{h}}_{u,k,x}(\theta) = \mathbb{E}_\theta &\left[\sum_{v \in \mathbb{\Delta}(u,k) \setminus \{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x \right] \\ &- \mathbb{E}_\theta \left[\sum_{v \in \mathbb{\Delta}^*(u,k) \setminus \{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x \right]. \end{aligned}$$

Note that $\dot{\mathrm{h}}_{u,k,x}(\theta)$ is well defined as $\mathbb{\Delta}(u,k)$ is finite and as $\phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)$ is integrable conditionally on $Y_{\mathbb{\Delta}^*(u,k)}$ and $X_{\mathrm{p}^k(u)} = x$ under Assumption 11 (see the comment after (5.46)). Also note that

$\dot{\mathrm{h}}_{u,k,x}(\theta)$ is the gradient w.r.t. θ of $\mathrm{h}_{u,k,x}(\theta)$ defined in (5.17) (see (5.44) and (5.45) for the case $k = h(u)$). Furthermore, note that $\dot{\mathrm{h}}_{u,k,x}(\theta)$ is a function of $Y_{\mathbb{\Delta}(u,k)}$ with an implicit dependence on $\mathcal{U}$ through $\mathbb{\Delta}(u,k)$, and that $\dot{\mathrm{h}}_{u,k,x}(\theta)$ does not depend on $\mathcal{U}$ is $k \leq h(u)$.

Using the increment functions $\dot{\mathrm{h}}_{u,k,x}(\theta)$, we can rewrite the score function as:

$$\nabla_\theta \ell_{n,x}(\theta) = \sum_{u \in T_n} \dot{\mathrm{h}}_{u,h(u),x}(\theta). \tag{5.47}$$

Construction of score increments with infinite past

Our goal is to let $k \to \infty$ as before to get a limit function $\dot{\mathrm{h}}_{u,\infty}$. We now proceed to construct $\dot{\mathrm{h}}_{u,\infty}$. First, we rewrite $\dot{\mathrm{h}}_{u,k,x}(\theta)$ (which is in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ by Assumption 11), as:

$$\begin{aligned}\dot{\mathrm{h}}_{u,k,x}(\theta) = \mathbb{E}_\theta[\phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u) \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] \\ + \sum_{v \in \mathbb{\Delta}^*(u,k) \setminus \{\mathrm{p}^k(u)\}} \Big(\mathbb{E}_\theta[\phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] \\ - \mathbb{E}_\theta[\phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x] \Big).\end{aligned} \tag{5.48}$$

We will need the following lemma that states a coupling bound that works "backwards in time", or rather along the path between a vertex v and the newly observed vertex u. Remind from (5.12) on page 129 that $\mathbb{\Delta}(u,k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$.

Lemma 5.4.1 (Total variation bound "backwards in time")**.** *Assume that Assumptions 6–7 hold. Let $k \in \mathbb{N}^*$, $x \in \mathcal{X}$ and $u \in T$, and let $v \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$. Then, we have:*

$$\left\| \mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) - \mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x) \right\|_{TV} \leq \rho^{d(u,v)-1}.$$

The proof of Lemma 5.4.1, which is postponed to Section 5.B, relies on a "backward in time" bound from $\mathrm{p}(u)$ to $u \wedge v$ and then a "forward in time" bound from $u \wedge v$ to v, and using the initial distributions $\mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \cdot \,|\, Y_\Delta, X_{\mathrm{p}^k(u)} = x)$ with Δ equal to $\mathbb{\Delta}(u,k)$ and $\mathbb{\Delta}^*(u,k)$, respectively. Note that this proof is similar to the proofs for Lemma 5.3.2 and [CMR05, Proposition 12.5.4].

The following lemma gives an L^2-bound on the difference between $\dot{\mathrm{h}}_{u,k,x}(\theta)$ and $\dot{\mathrm{h}}_{u,k',x'}(\theta)$ with a geometric decay. As we will reuse this result later with different functions, we state a more general version.

Note that the condition $\rho < 1/\sqrt{2}$ on the mixing rate ρ of the HMT process (X, Y) is due to the coupling bounds and the grouping of terms used in the proof of Lemma 5.4.2 (the upper bounds at the end of the proof only add a constant multiplicative factor). See the discussion in Remark 5.1.5.

Lemma 5.4.2. *Assume that Assumptions 5–8. Further assume that $\rho < 1/\sqrt{2}$.*

Let Θ_0 be a closed ball in Θ, and let ψ be a Borel function from $\Theta_0 \times \mathcal{X}^2 \times \mathcal{Y}$ to $\mathbb{R}^d$ for some $d \in \mathbb{N}$ such that for all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\theta \mapsto \psi(\theta, x, x', y) = \psi_\theta(x, x', y)$ is a continuous function on Θ_0. Furthermore, assume that there exists $b \in [1, +\infty)$ such that:

$$\mathbb{E}_{\theta^\star}\left[\sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\psi_\theta(x, x', Y_\partial)\|^b \right] < \infty.$$

Let $\xi_{u,k,x}(\theta)$ be defined as in (5.48) (with $\dot{\mathrm{h}}_{u,k,x}(\theta)$ and ϕ_θ replaced by $\xi_{u,k,x}(\theta)$ and ψ_θ, respectively), and note that it is in $L^b(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$. Then, there exists a finite constant $C < \infty$ such that for all $u \in T$ and $k' \geq k \geq 1$, we have:

$$\begin{aligned}\left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\xi_{u,k,x}(\theta) - \xi_{u,k',x'}(\theta)\|^b \right] \right)^{1/b} \\ \leq C \left(\mathbb{E}_{\theta^\star}\left[\sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\psi_\theta(x, x', Y_\partial)\|^b \right] \right)^{1/b} k \left(\max(\rho, 2\rho^2) \right)^{k/2}.\end{aligned}$$

As a consequence of Lemma 5.4.2, for all $u \in T$ and $x \in \mathcal{X}$, the sequence of function $(\xi_{u,k,x}(\theta))_{n\in\mathbb{N}}$ converges in $L^b(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ to some limit function $\xi_{u,\infty}(\theta)$ which does not depend on x. Moreover, the bound in Lemma 5.4.2 still holds when $\xi_{u,k',x'}(\theta)$ is replaced by $\xi_{u,\infty}(\theta)$.

For the particular choice of $\psi_\theta = \phi_\theta$, under Assumptions-5-8 and 11, for all $u \in T$, we denote by $\dot{\mathrm{h}}_{u,\infty}(\theta)$ the limit function of the sequence $(\dot{\mathrm{h}}_{u,k,x}(\theta))_{n\in\mathbb{N}}$ (for all $x \in \mathcal{X}$) which is in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$.

As an immediate corollary of Lemma 5.4.2, there exists a finite constant $C' < \infty$ such that for all $x \in \mathcal{X}$, $u \in T^*$ and $k \geq 1$, we have that $(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b])^{1/b} \leq \mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b])^{1/b} + C' < \infty$, (note that by stationarity, $\mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b])^{1/b} = \mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{v,1,x}(\theta)\|^b])^{1/b}$ for any other $v \in T^*$). Hence, we get:

$$\sup_{\theta\in\mathcal{O}} \sup_{u\in T} \sup_{k\geq 1} \left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{x\in\mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b \right] \right)^{1/b} < \infty. \tag{5.49}$$

Proof. We mimic the scheme of the proof of [CMR05, Lemma 12.5.3].

Let $u \in T$ and $k' \geq k \geq 1$ be fixed. The idea of the proof is to match, for each vertex index v of the sums expressing $\xi_{u,k,x}(\theta)$ and $\xi_{u,\infty}(\theta)$, pairs of terms that are close. To be more precise, we match:

1. For v close to u,
$$\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x]$$
and
$$\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'],$$
and similarly for the corresponding terms with $\Delta(u,k)$ and $\Delta(u,k')$ replaced by $\Delta^*(u,k)$ and $\Delta^*(u,k')$, respectively;
2. For v far from u,
$$\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x]$$
and
$$\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x],$$
and similarly for the corresponding terms with k and x replaced by k' and x', respectively.

Remind from (5.12) on page 129 that $\mathbb{\Delta}(u,k) \subset T^\infty(\mathrm{p}^k(u), k)$ and that the subtree $\mathbb{\Delta}(u,k)$ is random while the subtree $T^\infty(\mathrm{p}^k(u), k)$ is deterministic. Let $(x, x') \in \mathcal{X} \times \mathcal{X}$ and let $v \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(u)\}$, which implies that $\mathrm{p}(v) \in \Delta(u,k)$.

We start with the first kind of matches. Using the Markov property (remind (5.2)), we have:

$$\begin{aligned}
&\Big\| \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'] \Big\| \\
&\qquad = \Big\| \int_{\mathcal{X}^3} \psi_\theta(x_{\mathrm{p}(v)}, x_v, Y_v)\, \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}(u,k)\cap T(v)}, X_{\mathrm{p}(v)} = x_{\mathrm{p}(v)}) \\
&\qquad\qquad \times \mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x_{\mathrm{p}^k(u)}) \\
&\qquad\qquad \times \big[\delta_x(\mathrm{d}x_{\mathrm{p}^k(u)}) - \mathbb{P}_\theta(X_{\mathrm{p}^k(u)} \in \mathrm{d}x_{\mathrm{p}^k(u)} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x')\big] \Big\| \\
&\qquad \leq 2 \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\|\, \rho^{d(v,\mathrm{p}^k(u))-1},
\end{aligned} \tag{5.50}$$

where the inequality is obtained using Lemma 5.3.2 (note that $d(\mathrm{p}(v), \mathrm{p}^k(u)) = d(v, \mathrm{p}^k(u)) - 1$). Note that this upper bound is a.s. finite as $\sup_{x_1,x_2\in\mathcal{X}} \|\phi_\theta(x_1, x_2, Y_v)\|$ is in $L^b(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ by assumption (remind that the HMT process (X, Y) is stationary by Assumption 5). For $v \neq u$, note that this bound remains valid if $\mathbb{\Delta}(u,k)$ and $\mathbb{\Delta}(u,k')$ are replaced by $\mathbb{\Delta}^*(u,k)$ and $\mathbb{\Delta}^*(u,k')$, respectively. Obviously, this bound is small if v is far away from $\mathrm{p}^k(u)$ (remind that k is fixed).

We now give a bound for the second kind of matches. Assume that $v \neq u$. If v is not an ancestor of

u (then $d(u,v) = d(u,\mathrm{p}(v)) + 1$), using the Markov property (remind (5.2)) and Lemma 5.4.1, we get:

$$\begin{aligned}
&\big\|\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)},\ X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x]\big\| \\
&= \Big\| \int_{\mathcal{X}^3} \psi_\theta(x_{\mathrm{p}(v)}, x_v, Y_v)\, \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}(u,k)\cap T(v)}, X_{\mathrm{p}(v)} = x_{\mathrm{p}(v)}) \\
&\qquad \times \big[\mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) - \mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x)\big] \Big\| \\
&\le 2 \sup_{\theta\in\Theta_0} \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|\, \rho^{d(u,v)-2}.
\end{aligned}$$

If v is an ancestor of u (then $d(u,\mathrm{p}(v)) = d(u,v) + 1$), using the Markov property (remind (5.2)) and Lemma 5.4.1, we get:

$$\begin{aligned}
&\big\|\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)},\ X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x]\big\| \\
&\quad = \Big\| \int_{\mathcal{X}^3} \psi_\theta(x_{\mathrm{p}(v)}, x_v, Y_v)\, \mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)\setminus T(v)}, X_v = x_v) \\
&\qquad\qquad \times \big[\mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) - \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x)\big] \Big\| \\
&\quad \le 2 \sup_{\theta\in\Theta_0} \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|\, \rho^{d(u,v)-1}.
\end{aligned}$$

In both cases, we get:

$$\begin{aligned}
&\big\|\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)},\ X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x]\big\| \\
&\qquad\qquad \le 2 \sup_{\theta\in\Theta_0} \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|\, \rho^{d(u,v)-2}. \quad (5.51)
\end{aligned}$$

Note that the same bound remain valid for the corresponding terms with k and x replaced by k' and x', respectively, and with $v \in \Delta(u,k') \setminus \{\mathrm{p}^{k'}(u)\}$ instead of $v \in \Delta(u,k) \setminus \{\mathrm{p}^k(u)\}$. This bound is small if v is far away from u.

Remind from (5.12) on page 129 that $\mathbb{\Delta}(u,k) \subset T^\infty(\mathrm{p}^k(u),k)$ and as $k' \ge k$ note that:

$$\mathbb{\Delta}(u,k') \setminus \mathbb{\Delta}(u,k) \subset T^\infty(\mathrm{p}^{k'}(u),k') \setminus T^\infty(\mathrm{p}^k(u),k).$$

For a vertex $v \in T^\infty(\mathrm{p}^k(u),k) \setminus \{\mathrm{p}^k(u)\}$ (note that $u \wedge v \in T^\infty(\mathrm{p}^k(u))$), note that the term $\rho^{d(v,\mathrm{p}^k(u))-1}$ is smaller than $\rho^{d(u,v)-2}$ whenever $d(v,\mathrm{p}^k(u)) > d(u,v) - 1$, that is when $d(u\wedge v, \mathrm{p}^k(u)) \ge d(u\wedge v, u)$, that is when $d(u\wedge v, u) \le k/2$.

Combining those facts with the bounds (5.50) and (5.51), and using Minkowski's inequality for the L^b-norm, we find that $(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\sup_{\theta\in\Theta_0} \sup_{x,x'\in\mathcal{X}} \|\xi_{u,k,x}(\theta) - \xi_{u,k',x'}(\theta)\|^b]\ )^{1/b}$ is upper bounded by:

$$4 \sum_{v\in T^\infty(\mathrm{p}^{\lfloor k/2\rfloor}(u),\lfloor k/2\rfloor)} \rho^{d(v,\mathrm{p}^k(u))-1} + 4 \sum_{v\in T^\infty(\mathrm{p}^{k'}(u),k')\setminus T^\infty(\mathrm{p}^{\lfloor k/2\rfloor}(u),\lfloor k/2\rfloor)} \rho^{d(u,v)-2}, \quad (5.52)$$

up to the factor $\big(\mathbb{E}_{\theta^\star}\big[\sup_{\theta\in\Theta_0} \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|^b\big]\big)^{1/b}$ (remind that the process $(Y_u, u \in T^\infty)$ is stationary under Assumption 5). Denote by A_1 and A_2 respectively the first and second terms in (5.52). We are going to reindex those sums by $j := d(u, u\wedge v)$ and $q := d(u\wedge v, v)$ with $q \le j$. Note that if $q > 0$, then the first vertex after $u \wedge v = \mathrm{p}^j(u)$ on the path from u to v cannot be $\mathrm{p}^{j-1}(u)$ and must be the other children of $\mathrm{p}^j(u)$. Thus, there are 2^{q-1} choices of v with the same coding (j,q) with $0 < q \le j$. Hence, we get:

$$A_1 = 4 \sum_{j=0}^{\lfloor k/2\rfloor} \rho^{k-j-1} \left(1 + \sum_{q=1}^{j} 2^{q-1}\rho^q\right) \quad \text{and} \quad A_2 = 4 \sum_{j=\lfloor k/2\rfloor+1}^{k'} \rho^{j-2} \left(1 + \sum_{q=1}^{j} 2^{q-1}\rho^q\right).$$

Remark that there exists a finite constant $C < \infty$ (which depends on the value of ρ) such that for all $j \in \mathbb{N}^*$ we have $(1 + \sum_{q=1}^{j} 2^{q-1}\rho^q) \le C\max(j, (2\rho)^j)$ and $\sum_{q=j}^{\infty} q\rho^q \le C\rho^j$. Hence, there exists a finite constant $C' < \infty$ (which depends only on the value of ρ) such that (remind that $\rho < 1/\sqrt{2}$):

$$A_1 \le C' \left(k\rho^{k/2} + (2\rho^2)^{k/2}\right) \quad \text{and} \quad A_2 \le C' \left(\rho^{k/2} + (2\rho^2)^{k/2}\right). \quad (5.53)$$

Combining (5.52) and (5.53), we get that the bound in the lemma holds. This concludes the proof of the lemma. □

Asymptotic normality of the score

Define the limiting Fisher information as:

$$\mathcal{I}(\theta^\star) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\Big[\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)^t\Big], \tag{5.54}$$

where we see $\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)$ as a column vector.

For the asymptotic normality of the score, we need the following extra regularity assumption of the gradient.

Assumption 12 (L^4 gradient regularity)**.** *In addition to Assumption 11, we have:*

$$\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\mathcal{O}}\sup_{x\in\mathcal{X}} \|\nabla_\theta \log g_\theta(x, Y_\partial)\|^4\right] < \infty.$$

We are now ready to prove the following theorem stating the asymptotic normality of the normalized score towards a centered Gaussian random variable whose variance is the limiting Fisher information. Note that the condition $\rho < 1/\sqrt{2}$ on the mixing rate ρ of the HMT process (X, Y) comes from the use of Lemma 5.4.2 in the proof of this theorem. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Theorem 5.4.3 (Asymptotic normality of the normalized score)**.** *Assume that Assumptions 5–8 and 11–12 hold with $\theta^\star \in \Theta$ given. Further assume that $\rho < 1/\sqrt{2}$. Then, for all $x \in \mathcal{X}$, we have:*

$$|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)) \quad \textit{under } \mathbb{P}_{\theta^\star},$$

where $\mathcal{N}(0, M)$ denotes the centered Gaussian distribution with covariance matrix M, and $\mathcal{I}(\theta^\star)$ is the limiting Fisher information defined in (5.54)*.*

Proof. **Step 1: Approximation of the score by the stationary score.**

Remind from Lemma 5.2.3 that $\pi_{\theta^\star}$ denotes the invariant distribution for the hidden process X associated with $Q_{\theta^\star}$. Define the stationary score $\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta)$ as:

$$\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta) := \int_{\mathcal{X}} \nabla_\theta \ell_{n,x}(\theta)\, \pi_{\theta^\star}(\mathrm{d}x).$$

First, for all $x, x' \in \mathcal{X}$ and $\theta \in \mathcal{O}$, write:

$$\nabla_\theta \ell_{n,x}(\theta) - \nabla_\theta \ell_{n,x'}(\theta) = \sum_{u\in T_n^*} \Phi(\theta; u, x, x'),$$

where:

$$\Phi(\theta; u, x, x') = \mathbb{E}_\theta[\phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u) \,|\, Y_{T_n}, X_\partial = x] - \mathbb{E}_\theta[\phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u) \,|\, Y_{T_n}, X_\partial = x'].$$

Using Minkowski's inequality and the upper bound (5.50) from the proof of Lemma 5.4.2, we get:

$$\begin{aligned}
\left(\mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}} \frac{1}{|T_n|}\,\|\nabla_\theta \ell_{n,x}(\theta) - \nabla_\theta \ell_{n,x'}(\theta)\|^2\right]\right)^{1/2} & \\
&\le \frac{1}{|T_n|^{1/2}} \sum_{u\in T_n^*} \left(\mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}} \|\Phi(\theta; u, x, x')\|^2\right]\right)^{1/2} \\
&\le 2\left(\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\Theta_0}\sup_{x,x'\in\mathcal{X}} \|\phi_\theta(x, x', Y_\partial)\|^2\right]\right)^{1/2} \frac{1}{|T_n|^{1/2}} \sum_{k=1}^{n} 2^k \rho^{k-1} \\
&\le C \max(n2^{-n}, (2\rho^2)^{n/2}),
\end{aligned}$$

where $C < \infty$ is some finite constant. Thus (remind that $\rho < 1/\sqrt{2}$), for any $x \in \mathcal{X}$, we have:

$$\lim_{n\to\infty} \frac{1}{|T_n|^{1/2}} \Big(\nabla_\theta \ell_{n,x}(\theta^\star) - \nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta^\star) \Big) = 0 \qquad \text{in } L^2(\mathbb{P}_{\theta^\star}). \tag{5.55}$$

In particular, to prove asymptotic normality for the score $\nabla_\theta \ell_{n,x}(\theta^\star)$ for any $x \in \mathcal{X}$, it is enough to prove asymptotic normality for the stationary score $\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta^\star)$ (see for instance [Bil99b, Theorem 3.1]).

For any $u \in T$ and $k \in \mathbb{N}$ and $\theta \in \mathcal{O}$, define:

$$\dot{\mathrm{h}}_{u,k,\pi_{\theta^\star}}(\theta) := \int_{\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta)\, \pi_{\theta^\star}(\mathrm{d}x). \tag{5.56}$$

In particular, note that, as the bound in Lemma 5.4.2 is uniform in $x \in \mathcal{X}$, this bound still holds with $\dot{\mathrm{h}}_{u,k,x}(\theta)$ replaced by $\dot{\mathrm{h}}_{u,k,\pi_{\theta^\star}}(\theta)$. Using (5.47), note that we have:

$$\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta) = \sum_{u\in T_n} \dot{\mathrm{h}}_{u,h(u),\pi_{\theta^\star}}(\theta).$$

Moreover, remark that for $\theta = \theta^\star$ and for any $u \in T$ and $k \in \mathbb{N}^*$, we have:

$$\begin{aligned}
\dot{\mathrm{h}}_{u,k,\pi_{\theta^\star}}(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u)}, X_u, Y_u) \,|\, Y_{\mathbb{\Delta}(u,k)}] \\
+ \sum_{v\in \mathbb{\Delta}^*(u,k)\setminus\{\mathrm{p}^k(u)\}} \Big(\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(u,k)}] \\
- \mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}^*(u,k)}] \Big).
\end{aligned} \tag{5.57}$$

Step 2: The stationary score is a sum of martingale increments.

As T is a plane rooted tree, we can enumerate its vertices in a breadth-first-search manner, that is, as a sequence $(u_j)_{j\in\mathbb{N}}$ which is increasing for $<$. (Note that $u_0 = \delta$.) Remind that $\Delta(u_{j-1}) = \Delta^*(u_j)$ for all $j \geq 1$. Define the filtration $\mathcal{F}$ by $\mathcal{F}_j = \sigma(Y_v : v \in T,\, v \leq u_j) = \sigma(Y_{\Delta(u_j)})$ for all $j \in \mathbb{N}$, and note that $\mathcal{F}_j \subset \sigma(Y_T)$. Let $j \in \mathbb{N}^*$, $1 \leq k \leq h(u_j)$, $x \in \mathcal{X}$ and $v \in Y_{\Delta^*(u_j,k)}$. Note that we have:

$$\mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\Delta(u_j,k)}] \;\Big|\; \mathcal{F}_{j-1} \Big] \;=\; \mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\Delta^*(u_j,k)}].$$

Also note that Assumption 11 (on page 139) implies that:

$$\begin{aligned}
&\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, X_{\mathrm{p}(u_j)}] \\
&\qquad = \int_{\mathcal{X}\times\mathcal{Y}} \nabla_\theta \log[q_\theta(X_{\mathrm{p}(u_j)}, x) g_\theta(x,y)]\, q_\theta(X_{\mathrm{p}(u_j)}, x) g_\theta(x,y)\, \lambda(\mathrm{d}x)\mu(\mathrm{d}y) \\
&\qquad = \int_{\mathcal{X}\times\mathcal{Y}} \nabla_\theta [q_\theta(X_{\mathrm{p}(u_j)}, x) g_\theta(x,y)]\, \lambda(\mathrm{d}x)\mu(\mathrm{d}y) \\
&\qquad = \nabla_\theta \left[\int_{\mathcal{X}\times\mathcal{Y}} q_\theta(X_{\mathrm{p}(u_j)}, x) g_\theta(x,y)\, \lambda(\mathrm{d}x)\mu(\mathrm{d}y) \right] \\
&\qquad = 0.
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
&\mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta(u_j,k)}] \;\Big|\; \mathcal{F}_{j-1} \Big] \\
&\qquad = \mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta^*(u_j,k)}] \\
&\qquad = \mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta^*(u_j,k)}, X_{\mathrm{p}(u_j)}] \;\Big|\; Y_{\mathbb{\Delta}^*(u_j,k)} \Big] \\
&\qquad = \mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, X_{\mathrm{p}(u_j)}] \;\Big|\; Y_{\Delta^*(u_j,k)} \Big] \\
&\qquad = 0,
\end{aligned}$$

where we used the Markov property for the inner expectation in the third equality. Moreover, it is immediate that $\dot{\mathrm{h}}_{u_j,k,\pi_{\theta^\star}}(\theta^\star)$ is $\mathcal{F}_j$-measurable for all $j \in \mathbb{N}^*$ and $1 \leq k \leq h(j)$. Hence, we get that the sequence $\left(\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)\right)_{j\in\mathbb{N}^*}$ is a $\mathbb{P}_{\theta^\star}$-martingale increment sequence adapted to the filtration $\mathcal{F} = (\mathcal{F}_j)_{j\in\mathbb{N}}$ in $L^2(\mathbb{P}_{\theta^\star})$ (thanks to Assumption 11). We are going to apply a central limit theorem for martingales (see [Duf11, Corollary 2.1.10]). For all $n \in \mathbb{N}$, define $M_n = \sum_{j=0}^{n} \dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)$. Note that $M_0 = \dot{\mathrm{h}}_{\partial,0,\pi_{\theta^\star}}(\theta^\star) = \int_{\mathcal{X}} \nabla_\theta \log g_{\theta^\star}(x, Y_\partial)\, \pi_{\theta^\star}(\mathrm{d}x)$ is in $L^2(\mathbb{P}_{\theta^\star})$ by Assumption 11. Hence, the sequence $(M_n)_{n\in\mathbb{N}}$ is a $\mathbb{P}_{\theta^\star}$-martingale sequence adapted to the filtration $\mathcal{F} = (\mathcal{F}_j)_{j\in\mathbb{N}}$ in $L^2(\mathbb{P}_{\theta^\star})$, and whose quadratic variation is:

$$\langle M\rangle_n = \sum_{j=1}^{n} \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)^t \;\middle|\; \mathcal{F}_{j-1}\right],$$

where, as in (5.54), we see $\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)$ as a column vector. Note that for all $n \in \mathbb{N}$, M_n and $\langle M\rangle_n$ do not depend on $\mathcal{U}$.

Step 3: Convergence of the quadratic variation. Before applying the central limit theorem for martingales, we first need to prove that $\lim_{n\to\infty} n^{-1}\langle M\rangle_n = \mathcal{I}(\theta^\star)$ in $\mathbb{P}_{\theta^\star}$-probability. Indeed, we will prove that this convergence holds in $L^2(\mathbb{P}_{\theta^\star})$. Let $k \in \mathbb{N}^*$ and $x \in \mathcal{X}$. Note that for $u_j \in T \setminus T_{k-1}$ is equivalent to $j \geq |T_{k-1}|$ (remind that $u_0 = \partial$). Using (5.48) along with Assumption 12, we get that the random variable $\sup_{x\in\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\sup_{x\in\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})^t$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T \setminus T_{k-1}$. Thus, using (5.56) and Lemma 5.4.2 (remind that $\rho < 1/\sqrt{2}$) for the first moment ($b = 2$), there exists a finite constant $C > 0$ and $\alpha \in (0, 1)$ such that we have (remind (5.49)):

$$\mathbb{E}_{\theta^\star}\left[\left\|n^{-1}\langle M\rangle_n - \frac{1}{n}\sum_{j=|T_{k-1}|}^{n} \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{u_j,k,x}(\theta^\star)\dot{\mathrm{h}}_{u_j,k,x}(\theta^\star)^t \;\middle|\; \mathcal{F}_{j-1}\right]\right\|\right] \leq C\alpha^k + \frac{C|T_{k-1}|}{n}, \tag{5.58}$$

where remind that $\|\cdot\|$ denotes the euclidean norm for $d \times d$ matrices (or any other norm as they are all equivalent in finite dimension). To prove that the second term inside the expectation in the left hand side of (5.58) converges in $L^2(\mathbb{P}_{\theta^\star})$ as $n \to \infty$, we are going to apply the ergodic convergence Lemma 5.2.12 where the averages are done on the vertex subset $\{u_j : |T_{k-1}| \leq j \leq n\}$. Note that this lemma is stated for scalar-valued functions, but we can apply it individually for each of the matrix coefficients to get the equivalent for matrix-valued functions.

For all $u \in T \setminus T_{k-1}$, define the function:

$$\Psi_{u,k,x} : y_{\Delta^*(u,k)} \in \mathcal{Y}^{\Delta^*(u,k)} \mapsto \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{u,k,x}(\theta^\star; Y_{\Delta(u,k)})\dot{\mathrm{h}}_{u,k,x}(\theta^\star; Y_{\Delta(u,k)})^t \;\middle|\; Y_{\Delta^*(u,k)} = y_{\Delta^*(u,k)}\right].$$

For a vertex u in $T \setminus T_{k-1}$, let $v_u \in G_k$ be the unique vertex that satisfies the shape equality constraint (5.8) (on page 126), then we have the equality between functions:

$$\Psi_{u,k,x} = \Psi_{v_u,k,x}. \tag{5.59}$$

Moreover, using (5.48) along with Assumption 12, we get that $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\Psi_{u,k,x}(Y_{\Delta^*(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T \setminus T_{k-1}$. Hence, applying Lemma 5.2.12 to the collection of neighborhood-shape-dependent functions $(\Psi_{v,k,x})_{v\in G_k}$ (remind that indexing functions with G_k or with $\mathcal{N}_k$ is equivalent by (5.9)), and using (5.59) and (5.14) (in Remark 5.3.1), we get that the second term inside the expectation in the left hand side of (5.58) converges in $L^2(\mathbb{P}_{\theta^\star})$ to $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right]$ as $n \to \infty$. Using Lemma 5.4.2, we have that $\lim_{k\to\infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right] = \mathcal{I}(\theta^\star)$. Combining those facts with (5.58), we get that $\lim_{n\to\infty} n^{-1}\langle M\rangle_n = \mathcal{I}(\theta^\star)$ in $L^2(\mathbb{P}_{\theta^\star})$.

Step 4: Lindeberg's condition holds. We now need to verify that Lindeberg's condition holds (see [Duf11, Corollary 2.1.10]), that is, to prove for all $\varepsilon > 0$ that $\lim_{n\to\infty} F_n(\varepsilon\sqrt{n}) = 0$ in $\mathbb{P}_{\theta^\star}$-probability where for all $n \in \mathbb{N}^*$ and $A \in \mathbb{R}_+$:

$$F_n(A) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{\theta^\star}\left[\left\|\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)\right\|^2 \mathbb{1}_{\{\|\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)\|\geq A\}} \;\middle|\; \mathcal{F}_{j-1}\right]. \tag{5.60}$$

Remind that by Assumption 12 and Lemma 5.4.2 (remind that $\rho < 1/\sqrt{2}$) for the fourth moment ($b = 4$), we have:

$$C := \sup_{u\in T}\sup_{k\in\mathbb{N}^*} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \left\|\dot{\mathrm{h}}_{u,k,x}(\theta^\star)\right\|^4\right] < \infty.$$

Using Cauchy-Schwarz inequality and Markov inequality, we get:

$$\mathbb{E}_{\theta^\star}[F_n(A)] \le \frac{1}{n}\sum_{j=1}^{n} \frac{\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \left\|\dot{\mathrm{h}}_{u_j,h(u_j),x}(\theta^\star)\right\|^4\right]}{A^2} \le \frac{C}{A^2}.$$

Let $\varepsilon > 0$. Then, setting $A_n = \varepsilon\sqrt{n}$ for all $n \in \mathbb{N}^*$, we get that $\lim_{n\to\infty} F_n(\varepsilon\sqrt{n}) = 0$ in $L^1(\mathbb{P}_{\theta^\star})$, and thus in $\mathbb{P}_{\theta^\star}$-probability. Hence, we get that Lindeberg's condition holds.

Step 5: Applying the central limit theorem for martingales. Hence, we can apply the central limit theorem for martingales (see [Duf11, Corollary 2.1.10]), which gives us that $\lim_{n\to\infty} n^{-1}M_n = 0$ $\mathbb{P}_{\theta^\star}$-a.s. and that the sequence $(n^{-1/2}M_n)_{n\in\mathbb{N}^*}$ converges in $\mathbb{P}_{\theta^\star}$-distribution towards a centered Gaussian distribution $\mathcal{N}(0, \mathcal{I}(\theta^\star))$ whose covariance matrix is $\mathcal{I}(\theta^\star)$. In particular, using (5.55), we get that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} |T_n|^{-1}\nabla_\theta \ell_{n,x}(\theta^\star) = 0$ and that:

$$|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)) \quad \text{under } \mathbb{P}_{\theta^\star}.$$

This concludes the proof of the theorem. □

5.4.2 Law of large number for the normalized observed information

In this subsection, we prove that for all possibly random sequence $(\theta_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}\theta_n = \theta^\star$ $\mathbb{P}_{\theta^\star}$-a.s., then the normalized observed information $-n^{-1}\nabla^2_\theta \ell_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ which is defined in (5.54).

Remind the definition of the log-likelihood $\ell_{n,x}(\theta)$ in (5.7) on page 125. We start by decomposing the Hessian of the log-likelihood $\ell_{n,x}(\theta)$ as a sum of increment indexed by the tree T. Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 11 (note that this result is also known as *Louis missing information principle*, see [CMR05, Proposition 10.1.6]), we get for all $\theta \in \mathcal{O}$ and $x \in \mathcal{X}$:

$$\begin{aligned}\nabla^2_\theta \ell_{n,x}(\theta) = \nabla^2_\theta \log(g_\theta(X_\partial, Y_\partial)) &+ \mathbb{E}_\theta\left[\sum_{u\in T_n^*} \varphi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)\,\middle|\, Y_{T_n}, X_\partial = x\right] \\ &+ \mathrm{Var}_\theta\left[\sum_{u\in T_n^*} \phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)\,\middle|\, Y_{T_n}, X_\partial = x\right],\end{aligned}$$

where remind that ϕ_θ is defined in (5.46) on page 140, and φ_θ is defined as:

$$\varphi_\theta(x', x, y) = \nabla^2_\theta \log(q_\theta(x', x) g_\theta(x, y)). \tag{5.61}$$

Note that similarly to the case of ϕ_θ, the random variale $\varphi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)$ is integrable conditionally on $Y_{\Delta(u)}$ and $X_\partial = x$ (see the discussion after (5.46)). Also note that $\nabla_\theta \log g_\theta(x, Y_\partial)$ is $\mathbb{P}_{\theta^\star}$-a.s. finite by Assumption 11-(iii).

For all $u \in T$, $k \in \mathbb{N}^*$ and $x \in \mathcal{X}$, we define:

$$\begin{aligned}\Lambda_{u,k,x}(\theta) = \mathbb{E}_\theta&\left[\sum_{v\in\mathbb{\Delta}(u,k)\setminus\{\mathrm{p}^k(u)\}} \varphi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)\,\middle|\, Y_{\mathbb{\Delta}(u,k)}, X_\partial = x_\partial\right] \\ &- \mathbb{E}_\theta\left[\sum_{v\in\mathbb{\Delta}^*(u,k)\setminus\{\mathrm{p}^k(u)\}} \varphi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)\,\middle|\, Y_{\mathbb{\Delta}^*(u,k)}, X_\partial = x_\partial\right],\end{aligned} \tag{5.62}$$

and:

$$\Gamma_{u,k,x}(\theta) = \mathrm{Var}_\theta\left[\sum_{v\in\mathbb{\Delta}(u,k)\setminus\{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)\,\middle|\, Y_{\mathbb{\Delta}(u,k)}, X_\partial = x_\partial\right] - \mathrm{Var}_\theta\left[\sum_{v\in\mathbb{\Delta}^*(u,k)\setminus\{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v)\,\middle|\, Y_{\mathbb{\Delta}^*(u,k)}, X_\partial = x_\partial\right], \tag{5.63}$$

where Var_θ (resp. Cov_θ) denotes the (possibly conditional) variance (resp. covariance) corresponding to $\mathbb{P}_\mathcal{U}\otimes\mathbb{P}_\theta$. Note that $\Lambda_{u,k,x}(\theta)$ and $\Gamma_{u,k,x}(\theta)$ are random variables which depend on $Y_{\mathbb{\Delta}(u,k)}$ with an implicit dependence on $\mathcal{U}$, and that they do not depend on $\mathcal{U}$ if $k\le h(u)$.

Then, using telescopic sums involving the quantities defined in (5.62) and (5.63), the Hessian of the log-likelihood $\ell_{n,x}(\theta)$ can be rewritten for all $\theta\in\mathcal{O}$ and $x\in\mathcal{X}$ as:

$$\nabla^2_\theta \ell_{n,x}(\theta) = \nabla^2_\theta \log(g_\theta(X_\partial, Y_\partial)) + \sum_{u\in T_n^*} \Lambda_{u,h(u),x}(\theta) + \sum_{u\in T_n^*} \Gamma_{u,h(u),x}(\theta). \tag{5.64}$$

To prove the convergence of the two sums in the right hand side of (5.64), and thus get the convergence of the normalized observed information $-n^{-1}\nabla^2_\theta\ell_{n,x}(\theta)$, we will need the following L^2 regularity assumption on the Hessian of the transition kernel g_θ of the HMT.

Assumption 13 (L^2 Hessian regularity)**.** *In addition to Assumption 11, assume that we have:*

$$\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\mathcal{O}}\sup_x \|\nabla^2_\theta \log g_\theta(x, Y_\partial)\|^2\right] < \infty.$$

Propositions 5.4.4 and 5.4.5 below (whose proofs are given in Sections 5.4.2 and 5.4.2, respectively) state that $\Lambda_{u,k,x}(\theta)$ and $\Gamma_{u,k,x}(\theta)$ both have limits $\mathbb{P}_{\theta^\star}$-a.s. and in $L^2(\mathbb{P}_{\theta^\star})$ when $k\to\infty$. Denote those limits by $\Lambda_{u,\infty}(\theta)$ and $\Gamma_{u,\infty}(\theta)$, respectively. Furthermore, Propositions 5.4.4 and 5.4.5 also state that the two sums in the right hand side of (5.64) converge to $\mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]$ and $\mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$, respectively, with some uniformity in θ near $\theta^\star$.

We start with the proposition for the terms $\Lambda_{u,k,x}(\theta)$. Note that the condition $\rho<1/\sqrt{2}$ on the mixing rate ρ of the HMT process (X,Y) is due to the use of Lemma 5.4.2 in the proof of Proposition 5.4.4. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Proposition 5.4.4 (Convergence for averages of $\Lambda_{u,k,x}(\theta)$)**.** *Assume that Assumptions 5–8, 10-11 and 13 hold. Assume that $\rho<1/\sqrt{2}$. Then, for each $\theta\in\mathcal{O}$, we have that $\Lambda_{u,k,x}(\theta)$ converges in $L^2(\mathbb{P}_\mathcal{U}\otimes\mathbb{P}_{\theta^\star})$ to some limit $\Lambda_{u,\infty}(\theta)$ (that does not depend on x) as $k\to\infty$. Moreover, we have:*

$$\lim_{n\to\infty}\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Lambda_{u,h(u),x}(\theta^\star) - \mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}\left[\Lambda_{\partial,\infty}(\theta^\star)\right]\right|\right] = 0. \tag{5.65}$$

Furthermore, the function $\theta\mapsto\mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]$ is continuous on $\mathcal{O}$, and for all $x\in\mathcal{X}$ and $\theta\in\mathcal{O}$, we have:

$$\lim_{\delta\to0}\lim_{n\to\infty}\sup_{\theta'\in\mathcal{O}:\|\theta'-\theta\|\le\delta}\left|\,|T_n|^{-1}\sum_{u\in T_n^*}\Lambda_{u,h(u),x}(\theta') - \mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]\right| = 0, \quad \mathbb{P}_{\theta^\star}\text{-a.s.}$$

The following proposition is the equivalent of Proposition 5.4.4 for the terms $\Gamma_{u,k,x}(\theta)$. Note that the condition $\rho<1/2$ on the mixing rate ρ of the HMT process (X,Y) is due to the use of Lemma 5.4.17 in the proof of Proposition 5.4.5. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Proposition 5.4.5 (Convergence for the averages of $\Gamma_{u,k,x}(\theta)$)**.** *Assume that Assumptions 5–8 and 10-12 hold. Assume that $\rho<1/2$. Then, for each $\theta\in\mathcal{O}$, we have that $\Gamma_{u,k,x}(\theta)$ converges in $L^2(\mathbb{P}_\mathcal{U}\otimes\mathbb{P}_{\theta^\star})$ to some limit $\Gamma_{u,\infty}(\theta)$ (that does not depend on x) as $k\to\infty$. Moreover, we have:*

$$\lim_{n\to\infty}\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta^\star) - \mathbb{E}_\mathcal{U}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,\infty}(\theta^\star)\right]\right|\right] = 0. \tag{5.66}$$

Furthermore, the function $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$ *is continuous on* $\mathcal{O}$*, and for all* $x \in \mathcal{X}$ *and* $\theta \in \mathcal{O}$*, we have:*

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{\theta' \in \mathcal{O} : \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)] \right| = 0, \quad \mathbb{P}_{\theta^\star}\text{-}a.s.$$

With Propositions 5.4.4 and 5.4.5, we are now ready to prove the following theorem which states that the normalized observed information $-|T_n|^{-1}\nabla_\theta^2 \ell_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star}$-a.s. locally uniformly to the limiting Fisher information $\mathcal{I}(\theta^\star)$ (which is defined in (5.54)). Note that the condition $\rho < 1/2$ on the mixing rate ρ of the HMT process (X, Y) is inherited from Proposition 5.4.5. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Theorem 5.4.6 (Convergence of the normalized observed information)**.** *Assume that Assumptions 5–8 and 10–13 hold. Assume that* $\rho < 1/2$*. Assume that* Θ *is compact. Then, for all* $x \in \mathcal{X}$*, we have:*

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{\theta \in \mathcal{O} \,:\, \|\theta - \theta^\star\| \leq \delta} \left\| -|T_n|^{-1}\nabla_\theta^2 \ell_{n,x}(\theta) - \mathcal{I}(\theta^\star) \right\| = 0 \quad \mathbb{P}_{\theta^\star}\text{-}a.s. \tag{5.67}$$

As an immediate corollary, for any possibly random sequence $(\theta_n)_{n \in \mathbb{N}}$ such that $\lim_{n\to\infty} \theta_n = \theta^\star$ $\mathbb{P}_{\theta^\star}$-a.s. and any $x \in \mathcal{X}$, we get that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} -|T_n|^{-1}\nabla_\theta^2 \ell_{n,x}(\theta_n) = \mathcal{I}(\theta^\star)$. In particular, choosing $\theta_n = \hat{\theta}_{n,x}$ for all $n \in \mathbb{N}$ (remind that the MLE $\hat{\theta}_{n,x}$ is defined in (5.33) on page 135), and combining Theorems 5.3.11 and 5.4.6, we get that the normalized observed information $-|T_n|^{-1}\nabla_\theta^2 \ell_{n,x}(\hat{\theta}_{n,x})$ at the MLE $\hat{\theta}_{n,x}$ is a strongly consistent estimator of the Fisher information matrix $\mathcal{I}(\theta^\star)$.

Proof. Using (5.64) and Propositions 5.4.4 and 5.4.5, we get that (5.67) holds with $\mathcal{I}(\theta^\star)$ replaced by $-\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta^\star) + \Gamma_{\partial,\infty}(\theta^\star)]$. Thus, it remains to prove that this latter quantity is equal to $\mathcal{I}(\theta^\star)$.

Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 11 (note that this result is also known as *Fisher information matrix identity*, see [RS18, p.21] or [CMR05, p.355]), we get for all $\theta \in \mathcal{O}$ and $x \in \mathcal{X}$:

$$|T_n|^{-1}\, \mathbb{E}_\theta\big[\nabla_\theta \ell_{n,x}(\theta) \nabla_\theta \ell_{n,x}(\theta)^t \;\big|\; X_\partial = x\big] = -|T_n|^{-1}\, \mathbb{E}_\theta\big[\nabla_\theta^2 \ell_{n,x}(\theta) \;\big|\; X_\partial = x\big].$$

Setting $\theta = \theta^\star$ and taking the expectation over X_∂, we get:

$$|T_n|^{-1}\, \mathbb{E}_{\theta^\star}\left[\nabla_\theta \ell_{n,X_\partial}(\theta^\star) \nabla_\theta \ell_{n,X_\partial}(\theta^\star)^t\right] = -|T_n|^{-1}\, \mathbb{E}_{\theta^\star}\left[\nabla_\theta^2 \ell_{n,X_\partial}(\theta^\star)\right]. \tag{5.68}$$

Using (5.64) on page 148, Propositions 5.4.4 and 5.4.5 give us that the right hand side of (5.68) converges as $n \to \infty$ to $-\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta^\star) + \Gamma_{\partial,\infty}(\theta^\star)]$.

Remind that using (5.48) along with Assumption 12, we get that $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})^t$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T \setminus T_{k-1}$. Thus, using Lemma 5.4.2 for the first moment ($b = 1$), there exists a finite constant $C > 0$ and $\alpha \in (0, 1)$ such that for any $k \in \mathbb{N}^*$ and $x \in \mathcal{X}$, we have:

$$\mathbb{E}_{\theta^\star}\left[\frac{1}{|T_n|}\left\|\nabla_\theta \ell_{n,X_\partial}(\theta^\star)\nabla_\theta \ell_{n,X_\partial}(\theta^\star)^t - \sum_{u \in T_n \setminus T_{k-1}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star)\dot{\mathrm{h}}_{u,k,x}(\theta^\star)^t\right\|\right] \leq C\alpha^k + \frac{C|T_{k-1}|}{|T_n|}, \tag{5.69}$$

where remind that we see $\dot{\mathrm{h}}_{u,k,x}(\theta^\star)$ as a column vector. Then, using an ergodic convergence argument similar to the one used in Step 3 in the proof of Theorem 5.4.3, we get:

$$\lim_{n\to\infty} \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star)\dot{\mathrm{h}}_{u,k,x}(\theta^\star)^t = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right] \qquad \text{in } L^2(\mathbb{P}_{\theta^\star}).$$

Using Lemma 5.4.2, we have that $\lim_{k\to\infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right] = \mathcal{I}(\theta^\star)$. Combining those facts with (5.69), we get that the left hand side in (5.68) converges to $\mathcal{I}(\theta^\star)$ as $n \to \infty$.

Hence, we get $\mathcal{I}(\theta^\star) = -\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta^\star) + \Gamma_{\partial,\infty}(\theta^\star)]$, which concludes the proof. □

Using Theorems 5.3.11, 5.4.3 and 5.4.6, we can prove the following theorem which states that the MLE has asymptotic normal fluctuations with covariance matrix $\mathcal{I}(\theta^\star)^{-1}$ where the Fisher information matrix $\mathcal{I}(\theta^\star)$ is defined in (5.54) on page 144. Recall that the contrast function ℓ is defined in (5.26) on page 132, that the MLE $\hat{\theta}_{n,x}$ is defined in (5.33) on page 135, and that the mixing rate ρ of the HMT process (X, Y) is defined after Assumption 7 on page 124.

Note that the condition $\rho < 1/2$ on the mixing rate ρ of the HMT process (X, Y) is inherited from Theorem 5.4.6, and thus from Proposition 5.4.5. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Theorem 5.4.7 (Asymptotic normality of the MLE)**.** *Assume that Assumptions 5–13 hold. Assume that $\rho < 1/2$. Further assume that the contrast function ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 5.3.10) and that Θ is compact, $\theta^\star$ is an interior point of Θ, and the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ (which is defined in* (5.54)*) is non-singular. Then, for all $x \in \mathcal{X}$, we have:*

$$|T_n|^{1/2}\big(\hat{\theta}_{n,x} - \theta^\star\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1}) \quad \text{under } \mathbb{P}_{\theta^\star},$$

where $\mathcal{N}(0, M)$ denotes the centered Gaussian distribution with covariance matrix M.

Proof. The proof is a standard argument and is similar to the proof of [BRR98, Theorem 1]. Remind that the gradient of $\ell_{n,x}$ vanishes at the MLE $\hat{\theta}_{n,x}$ by definition. Thus, using a Taylor expansion for $\nabla_\theta \ell_{n,x}$ around $\theta^\star$, we get:

$$0 = \nabla_\theta \ell_{n,x}(\hat{\theta}_{n,x}) = \nabla_\theta \ell_{n,x}(\theta^\star) + \left(\int_0^1 \nabla_\theta^2 \ell_{n,x}(\theta^\star + t(\hat{\theta}_{n,x} - \theta^\star))\,\mathrm{d}t\right)(\hat{\theta}_{n,x} - \theta^\star),$$

where we see $\hat{\theta}_{n,x}$ and $\theta^\star$ as column vectors. As $\mathcal{I}(\theta^\star)$ is non-singular (indeed definite positive), remark that Theorems 5.3.11 and 5.4.6 insure that $\mathbb{P}_{\theta^\star}$-a.s. for n large enough the integrand in the integral of the above formula is non-singular (indeed definite positive) for all values of t, and thus the matrix-valued integral is non-singular. Thus, from the above equation, we obtain $\mathbb{P}_{\theta^\star}$-a.s. for n large enough:

$$|T_n|^{1/2}\big(\hat{\theta}_{n,x} - \theta^\star\big) = \left(-|T_n|^{-1}\int_0^1 \nabla_\theta^2 \ell_{n,x}(\theta^\star + t(\hat{\theta}_{n,x} - \theta^\star))\,\mathrm{d}t\right)^{-1} |T_n|^{-1/2}\nabla_\theta \ell_{n,x}(\theta^\star).$$

As by Theorem 5.3.11, we have that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} \hat{\theta}_{n,x} = \theta^\star$, using Theorem 5.4.6, we get that the first factor in the right hand side $\mathbb{P}_{\theta^\star}$-a.s. converges to $\mathcal{I}(\theta^\star)$ as $n \to \infty$. Using Theorem 5.4.3, we get that the second factor in the right hand side converges $\mathbb{P}_{\theta^\star}$-weakly as $n \to \infty$ to the Gaussian random distribution $\mathcal{N}(0, \mathcal{I}(\theta^\star))$. Hence, using Cramér-Slutsky's theorem, we get that $|T_n|^{1/2}\big(\hat{\theta}_{n,x} - \theta^\star\big)$ converges $\mathbb{P}_{\theta^\star}$-weakly as $n \to \infty$ to the Gaussian random distribution $\mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1})$. This concludes the proof. □

Proof of Proposition 5.4.4

We are going to prove a version of Proposition 5.4.4 where the functions φ_θ used in (5.62) to define $\Lambda_{u,k,x}(\theta)$ are replaced by scalar-valued functions, still denoted by φ_θ, under more general assumptions. The extension to matrix-valued functions is then straightforward by applying the result coordinate-wise.

Let Θ_0 be a compact subset of Θ, Let Θ_0 be a closed ball in Θ, and let $\varphi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \to \mathbb{R}$ be a Borel function such that for all $x', x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\theta \mapsto \varphi(\theta, x', x, y) = \varphi_\theta(x', x, y)$ is a continuous function on Θ_0, and such that:

$$\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\Theta_0}\sup_{x,x'\in\mathcal{X}} |\varphi_\theta(x, x', Y_\partial)|^2\right] < \infty.$$

Let $\Lambda_{u,k,x}(\theta)$ be defined as in (5.62) on page 147 and note that it is in $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$.

The proof of Proposition 5.4.4 is decomposed into several lemmas.

We start with the following lemma, stating a uniform $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ approximation bound on the quantities $\Lambda_{u,k,x}(\theta)$, and the existence of a limit function $\Lambda_{u,\infty}(\theta)$ which does not depend on x. This lemma is an immediate consequence of Lemma 5.4.2 (remind that $\rho < 1/\sqrt{2}$ under the assumptions of Proposition 5.4.4) for the second moment ($b = 2$) with $\psi_\theta = \varphi_\theta$, see also the discussion after Lemma 5.4.2 for the existence of the limit function.

Lemma 5.4.8. *Under the assumptions of Proposition 5.4.4, there exist finite constants $C < \infty$ and $\alpha \in (0,1)$ such that for all $u \in T$ there exists some function $\Lambda_{u,\infty}(\theta)$ in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ such that for all $k \in \mathbb{N}^*$, we have:*

$$\left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta \in \Theta_0} \sup_{x, \in \mathcal{X}} |\Lambda_{u,k,x}(\theta) - \Lambda_{u,\infty}(\theta)|^2\right]\right)^{1/2} \leq C\alpha^k.$$

In particular, for all $u \in T$, $\theta \in \Theta_0$ and $x \in \mathcal{X}$, the sequence $(\Lambda_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$ converges in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ to $\Lambda_{u,\infty}(\theta)$ which does not depend on x.

The following lemma gives an exponential bound on the $L^2(\mathbb{P}_{\theta^\star})$ norm uniformly in $x \in \mathcal{X}$ for the the average of the quantities $\Lambda_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$.

Lemma 5.4.9. *Under the assumptions of Proposition 5.4.4, for all $x \in \mathcal{X}$ and $\theta \in \Theta_0$, there exist finite constants $C < \infty$ and $\alpha \in (0,1)$ such that for all $n \in \mathbb{N}^*$ we have:*

$$\mathbb{E}_{\theta^\star}\left[\sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\Lambda_{\partial,\infty}(\theta)\right] \right|^2\right]^{1/2} \leq C\alpha^n.$$

Proof. Let $x' \in \mathcal{X}$ and $\theta \in \Theta_0$. Using Minkowski's inequality and Jensen's inequality, for all $n, k \in \mathbb{N}^*$, we get:

$$\begin{aligned}
&\mathbb{E}_{\theta^\star}\left[\sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\Lambda_{\partial,\infty}(\theta)\right] \right|^2\right]^{1/2} \\
&\qquad \leq \mathbb{E}_{\theta^\star}\left[\sup_{x,x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_{k-1}^*} \Lambda_{u,h(u),x}(\theta) \right|^2\right]^{1/2} \\
&\qquad\quad + \mathbb{E}_{\theta^\star}\left[\sup_{x,x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Lambda_{u,h(u),x}(\theta) - \Lambda_{u,k,x'}(\theta) \right|^2\right]^{1/2} \\
&\qquad\quad + \mathbb{E}_{\theta^\star}\left[\left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Lambda_{u,k,x'}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\Lambda_{\partial,k,x'}(\theta)\right] \right|^2\right]^{1/2} \\
&\qquad\quad + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[|\Lambda_{\partial,k,x'}(\theta) - \Lambda_{\partial,\infty}(\theta)|^2\right]^{1/2}.
\end{aligned} \tag{5.70}$$

Using Lemma 5.4.8 together with (5.49) on page 142 (which, remind, are both immediate consequences of Lemma 5.4.2), there exists a finite constant $C < \infty$ and $\beta \in (0,1)$ such that the first term in the right hand side of (5.70) is upper bounded by $C2^{-(n-k)}$ (note that $\frac{|T_{k-1}|}{|T_n|} \leq 2^{-(n-k)}$), and the second and fourth terms in the right hand side of (5.70) are both upper bounded by $C\beta^{k/2}$.

We now give an upper bound for the second term in the right hand side of (5.70). For a vertex u in $T \setminus T_{k-1}$, let $v_u \in G_k$ be the unique vertex that satisfies the shape equality constraint (5.8) (on page 126), then we have:

$$\Lambda_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \Lambda_{v_u,k,x'}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{5.71}$$

Moreover, using the definition of $\Lambda_{u,k,x}(\theta)$ in (5.62) together with the assumption on φ_θ in Proposition 5.4.4, we get that the random variable $\Lambda_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T \setminus T_{k-1}$. Thus, we can apply Lemma 5.2.11 (see in particular (5.11)) to the collection of neighborhood-shape-dependent functions $(\Lambda_{v_u,k,x'}(\theta; Y_{\Delta(v)} = \cdot))_{v \in G_k}$ (remind that indexing functions with G_k or with $\mathcal{N}_k$ is equivalent by (5.9)). Using (5.11) in Lemma 5.2.11 together with (5.28) and (5.14) in Remark 5.3.1, we get that there exist $\gamma \in (0,1)$ and a finite constant $C' < \infty$ (note that they both do not depend on k

and n) such that for all $n, k \in \mathbb{N}^*$ with $n \geq k$, the second term in the right hand side of (5.70) is upper bounded by $C'\gamma^{n-k}$.

Hence, taking $k = \lceil n/2 \rceil$, we get that the left hand side of (5.70) is upper bounded by $2C\beta^{n/4} + C'\alpha^{n/2} + C2^{-n/2+1}$, and thus decays at exponential rate as desired. This concludes the proof. □

Lemma 5.4.9 implies as a corollary the convergence $\mathbb{P}_{\theta^\star}$-a.s. and in $L^2(\mathbb{P}_{\theta^\star})$ uniformly in $x \in \mathcal{X}$ for the the sum of the quantities $\Lambda_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$.

Corollary 5.4.10. *Under the assumptions of Proposition 5.4.4, for all $x \in \mathcal{X}$ and $\theta \in \Theta_0$, we have:*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u\in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\Lambda_{\partial,\infty}(\theta)\big] \right| = 0 \quad \mathbb{P}_{\theta^\star}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^\star}).$$

Proof. The convergence in $L^2(\mathbb{P}_{\theta^\star})$ follows immediately from Lemma 5.4.9. Moreover, using again Lemma 5.4.9, we have:

$$\sum_{n\in\mathbb{N}^*} \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u\in T_n^*} \Lambda_{u,\infty}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)] \right|^2 \right] < \infty.$$

Hence, Borel-Cantelli lemma and Markov's inequality imply that the convergence in the lemma also holds $\mathbb{P}_{\theta^\star}$-a.s. □

The following lemma gives some continuity properties of the function $\theta \mapsto \Lambda_{\partial,k,x}(\theta)$.

Lemma 5.4.11. *Under the assumptions of Proposition 5.4.4, for all $x \in \mathcal{X}$ and $k \in \mathbb{N}$, the random function $\theta \mapsto \Lambda_{\partial,k,x}(\theta)$ is $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. continuous on Θ_0. Moreover, for all $\theta \in \Theta_0$, we have:*

$$\lim_{\delta\to 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} |\Lambda_{\partial,k,x}(\theta') - \Lambda_{\partial,k,x}(\theta)|^2 \right] = 0.$$

Proof. We mimic the proof of [DMR04, Lemma 14].

For all $v \in T^\infty$, define the random variable $\|\varphi^v\|_\infty = \sup_{\theta'\in\Theta_0} \sup_{x,x'\in\mathcal{X}} |\varphi_{\theta'}(x', x, Y_v)|$. Remind that under the assumptions of Proposition 5.4.4, the HMT process (X, Y) is stationary and the random variable $\|\varphi^\partial\|_\infty$ is in $L^2(\mathbb{P}_{\theta^\star})$. Thus, for all $v \in T^\infty$, the random variable $\|\varphi^v\|_\infty$ is in $L^2(\mathbb{P}_{\theta^\star})$. Remind from (5.12) on page 129 that $\mathbb{\Delta}(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$. Then, note that we have:

$$\sup_{\theta\in\Theta_0} |\Lambda_{\partial,k,x}(\theta)| \leq 2 \sum_{v\in T^\infty(\mathrm{p}^k(\partial),k)} \|\varphi^v\|_\infty,$$

where the upper bound is a random variable in $L^2(\mathbb{P}_{\theta^\star})$ (and thus in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$) which depends only on $Y_{T^\infty(\mathrm{p}^k(u),k)}$ but not on $\mathcal{U}$. Hence, to prove the lemma, it suffices to prove that for all $v \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(\partial)\}$ we have:

$$\lim_{\delta\to 0} \sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} \Big| \mathbb{E}_{\theta'}[\varphi_{\theta'}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] - \mathbb{E}_{\theta}[\varphi_{\theta}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] \Big| = 0, \qquad \mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}\text{-a.s.}$$

Denote $x_{\mathrm{p}^k(\partial)} = x$, and write:

$$\mathbb{E}_{\theta}[\varphi_{\theta}(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbb{\Delta}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] = \frac{\int_{\mathcal{X}^{\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}} \varphi_{\theta}(x_{\mathrm{p}(v)}, x_v, Y_v) \prod_{w\in\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} q_\theta(x_{\mathrm{p}(w)}, x_w) g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w)}{\int_{\mathcal{X}^{\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}} \prod_{w\in\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} q_\theta(x_{\mathrm{p}(w)}, x_w) g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w)}. \quad (5.72)$$

Using Assumptions 6-8 (which are part of the assumptions in Proposition 5.4.4), we know that the integrand in the numerator of the right hand side of (5.72) is continuous w.r.t. θ and is upper bounded

by the random variable $\|\varphi^v\|_\infty(\sigma^+b^+)^{|T^\infty(\mathrm{p}^k(u),k)|-1}$ (remind that $\sigma^+ \geq 1$ and $b^+ \geq 1$). And similarly, the denominator is continuous w.r.t. θ, and, using Assumption 7-(ii), is lower bounded by the random variable:

$$\prod_{w\in\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} \sigma^- \inf_{\theta'\in\Theta} \int g_{\theta'}(x_w, Y_w)\lambda(\mathrm{d}x_w) > 0.$$

Hence, using dominated convergence, we conclude that $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. the left hand side of (5.72) is continuous w.r.t. θ. This concludes the proof. □

As a corollary of Lemma 5.4.11, we get that the function $\theta \mapsto \Lambda_{\partial,\infty}(\theta)$ is continuous in $L^2(\mathbb{P}_{\theta^\star})$.

Corollary 5.4.12. *Under the assumptions of Proposition 5.4.4, for all $\theta \in \Theta_0$, we have:*

$$\lim_{\delta\to 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} |\Lambda_{\partial,\infty}(\theta') - \Lambda_{\partial,\infty}(\theta)|^2\right] = 0.$$

In particular, the function $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]$ is continuous on Θ_0.

Proof. Using Minkowski's inequality and Lemma 5.4.8, there exist a finite constant $C < \infty$ and $\beta \in (0,1)$ such that for all $x \in \mathcal{X}$ and $k \in \mathbb{N}^*$, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} |\Lambda_{\partial,\infty}(\theta') - \Lambda_{\partial,\infty}(\theta)|^2\right]^{1/2} \\ \leq 2C\beta^{k/2} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} |\Lambda_{\partial,k,x}(\theta') - \Lambda_{\partial,k,x}(\theta)|^2\right]^{1/2}. \quad (5.73)\end{aligned}$$

Using Lemma 5.4.11, we get:

$$\limsup_{\delta\to 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta} |\Lambda_{\partial,\infty}(\theta') - \Lambda_{\partial,\infty}(\theta)|^2\right]^{1/2} \leq 2C\beta^{k/2},$$

and taking $k \to \infty$, the upper bound vanishes. This concludes the proof. □

We now prove a locally uniform law of large numbers for the quantities $\Lambda_{u,k,x}(\theta)$.

Lemma 5.4.13. *Under the assumptions of Proposition 5.4.4, for all $x \in \mathcal{X}$, we have:*

$$\lim_{\delta\to 0}\lim_{n\to\infty} \sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta} \left|\frac{1}{|T_n|}\sum_{u\in T_n^*} \Lambda_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]\right| = 0, \quad \mathbb{P}_{\theta^\star}\text{-}a.s.$$

Proof. First, write:

$$\begin{aligned}\sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta} \left|\frac{1}{|T_n|}\sum_{u\in T_n^*} \Lambda_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]\right| \\ \leq \frac{1}{|T_n|}\sum_{u\in T_n^*} \sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta} \left|\Lambda_{u,h(u),x}(\theta') - \Lambda_{u,h(u),x}(\theta)\right| \\ + \left|\frac{1}{|T_n|}\sum_{u\in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]\right|. \quad (5.74)\end{aligned}$$

Then, we use the exact same argument as in the proofs of Lemma 5.4.9 and Corollary 5.4.10 where for all $u \in T^*$, the random variable $\Lambda_{u,h(u),x}(\theta)$ is replaced by the random variable:

$$\sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta} \left|\Lambda_{u,h(u),x}(\theta') - \Lambda_{u,h(u),x}(\theta)\right|,$$

which are in $L^2(\mathbb{P}_{\theta^\star})$ using the assumptions of Proposition 5.4.4. This gives us that the first term in the upper bound of (5.74) converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{\theta' : \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial,\infty}(\theta') - \Lambda_{\partial,\infty}(\theta)| \right],$$

which, by Corollary 5.4.12, vanishes when $\delta \to 0$. Corollary 5.4.10 implies that the second term in the upper bound of (5.74) vanishes $\mathbb{P}_{\theta^\star}$-a.s. when $n \to \infty$. This concludes the proof. □

Combining the previous lemmas in this subsection, we are now ready to prove Proposition 5.4.4.

Proof of Proposition 5.4.4. By Lemma 5.4.8, for all $u \in T$, we have that $(\Lambda_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$ is a Cauchy sequence uniformly w.r.t. $\theta \in \Theta_0$ in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ that converges to some limit $\Lambda_{u,\infty}(\theta)$ (that does not depend on x). By Corollary 5.4.10, we have that $\mathbb{P}_{\theta^\star}$-a.s. the convergence for the the average of the quantities $\Lambda_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$ holds uniformly in $x \in \mathcal{X}$, that is, (5.65) in Proposition 5.4.4 holds. By Corollary 5.4.12, we have that the function $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]$ is continuous on Θ_0. Finally, the last part of the proposition is given by Lemma 5.4.13. □

Proof of Proposition 5.4.5

Similarly to what we have done for Proposition 5.4.4, we are going to prove a version of Proposition 5.4.5 where the functions ϕ_θ used in (5.63) to define $\Gamma_{u,k,x}(\theta)$ are replaced by scalar-valued functions, still denoted by φ_θ, under more general assumptions. The extension to matrix-valued functions is then straightforward by applying the result coordinate-wise.

Let Θ_0 be a compact subset of Θ, Let Θ_0 be a closed ball in Θ, and let $\phi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \to \mathbb{R}$ be a Borel function such that for all $x', x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\theta \mapsto \phi(\theta, x', x, y) = \phi_\theta(x', x, y)$ is a continuous function on Θ_0, and such that:

$$\mathbb{E}_{\theta^\star} \left[\sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\phi_\theta(x, x', Y_\partial)|^4 \right] < \infty.$$

Let $\Gamma_{u,k,x}(\theta)$ be defined as in (5.63) on page 148 and note that it is in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$.

The proof of Proposition 5.4.4 can be straightforwardly adapted to Proposition 5.4.5 except for Lemma 5.4.8. Thus, for brevity, we only present the adaptation of Lemma 5.4.8 to the terms $\Gamma_{u,k,x}(\theta)$. (The details of the adaptation for the rest of the proof of Proposition 5.4.4 to the terms $\Gamma_{u,k,x}(\theta)$ can be found in Section 5.D.)

We start with two lemmas giving coupling bounds that will be used to control the covariance terms that appear in the definition of the terms $\Gamma_{u,k,x}(\theta)$. The following lemma is a variant of Lemma 5.3.2 for two vertices of T.

Lemma 5.4.14 (Forward coupling bound for two vertices)**.** *Assume that Assumptions 6 and 7 hold. Then, for all $u, v \in T$, all $y_{T_n} \in \mathcal{Y}^{|T_n|}$ (and $n \in \mathbb{N}$) and all initials distributions ν and ν' on $\mathcal{X}$, we have:*

$$\left\| \int_{\mathcal{X}} \mathbb{P}_\theta\Big(X_u \in \cdot, X_v \in \cdot \;\Big|\; Y_{T_n} = y_{T_n}, X_\partial = x\Big)[\nu(\mathrm{d}x) - \nu'(\mathrm{d}x)] \right\|_{\mathrm{TV}} \leq 2\,\rho^{\min(h(u),h(v))}.$$

For simplicity, Lemma 5.4.14 is stated with ∂ as the initial vertex, but note that the results still holds when replacing ∂ and T_n by v and $T(v, n)$ for any $v \in T^\infty$. We shall reuse this fact later.

Proof. We are going to construct a coupling that achieves this minimum. Denote by $((X'_w, Y'_w), w \in T)$ the process started from the distribution ν' (and similarly without the $'$). Remark that we only need to define the (joint) coupling for the variables X_w and X'_w for w on the paths between the vertices ∂, u and v.

Lemma 5.3.2 applied to the vertex $u \wedge v$ gives us a coupling for the variables X_w and X'_w for w on the path between ∂ and $u \wedge v$ with successful coupling probability upper bounded by $1 - \rho^{h(u \wedge v)}$. On this successful coupling event before or on $u \wedge v$, the two processes are still defined to be equal after the fork on both branches leading to u and v.

On the complementary event (no successful coupling before or on $u \wedge v$), we get two new distributions $\nu_{u\wedge v}$ and $\nu'_{u\wedge v}$ for the variables $X_{u\wedge v}$ and $X'_{u\wedge v}$, respectively. Note that conditioned on the value of $X_{u\wedge v}$, the two branches leading to u and v are independent. Thus, applying Lemma 5.3.2 to u (resp. v) with the initial distributions $\nu_{u\wedge v}$ and $\nu'_{u\wedge v}$, we construct a coupling of the processes X and X' on the branch from $u \wedge v$ to u (resp. v) with successful coupling probability upper bounded by $1 - \rho^{h(u)-h(u\wedge v)}$ (resp. $1 - \rho^{h(v)-h(u\wedge v)}$).

Hence, the probability that we do not get a successful coupling on at least one of the two variables X_u and X_v is upper bounded by $\rho^{h(u\wedge v)}(\rho^{h(u)-h(u\wedge v)} + \rho^{h(v)-h(u\wedge v)}) \leq 2\rho^{\min(h(u),h(v))}$. This concludes the proof. □

The following lemma is a variant of Lemma 5.4.1 giving a "backward in time" coupling bound for two vertices of T.

Lemma 5.4.15 (Backward coupling bound for two vertices)**.** *Assume that Assumptions 6–7 hold. Let $k \in \mathbb{N}^*$, $x \in \mathcal{X}$ and $u \in T$, and let $v, w \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$. Then, we have:*

$$\left\|\mathbb{P}_\theta(X_v \in \cdot, X_w \in \cdot \,|\, Y_{\Delta(u,k)}, X_{\mathrm{p}^k(u)} = x) - \mathbb{P}_\theta(X_v \in \cdot, X_w \in \cdot \,|\, Y_{\Delta^*(u,k)}, X_{\mathrm{p}^k(u)} = x)\right\|_{TV} \leq 2\,\rho^{\min(d(u,v),d(u,w))-1}.$$

Proof. The idea of the proof is similar to that of Lemma 5.4.14. We explicitly construct a coupling with coupling failure probability upper bounded by $2\,\rho^{\min(d(u,v),d(u,w))-1}$. Denote by w_0 the vertex on the path between v and w that is the closest to u, and note that we have $w_0 \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$. Note that w_0 is on the path from $\mathrm{p}(u)$ to v, and thus $d(\mathrm{p}(u), v) = d(\mathrm{p}(u), w_0) + d(w_0, v)$, and similarly when replacing v by w. On the path from $\mathrm{p}(u)$ to w_0, we use the coupling provided by the "backward in time" coupling bound from Lemma 5.4.1 with successful probability $1 - \rho^{d(\mathrm{p}(u),w_0)}$. On the path from w_0 to v and the other path from w_0 to w, which are independent using the Markov property, we use two independent couplings that are constructed using a similar coupling argument as in Lemma 5.4.1 with $\mathrm{p}(u)$ replaced by w_0. Those independent couplings have successful probabilities $1 - \rho^{d(w_0,v)}$ and $1 - \rho^{d(w_0,w)}$, respectively. Note that the coupling we have constructed has a coupling failure probability upper bounded by $2\,\rho^{\min(d(u,v),d(u,w))-1}$. □

For brevity, for all $u \in T$ we will denote $\phi_{\theta,u} = \phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)$ and $\|\phi_u\|_\infty = \sup_{\theta\in\Theta_0} \sup_{x,x'\in\mathcal{X}} |\phi_\theta(x', x, Y_u)|$. The following lemma gives several upper bounds on the covariance terms that appear in the definition of the terms $\Gamma_{u,k,x}(\theta)$. Remind from (5.12) on page 129 that $\Delta(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$.

Note that this lemma is stated under the assumptions of Proposition 5.4.5, but here we do not need the assumption that $\rho < 1/2$ for the mixing rate ρ of the HMT process (X, Y).

Lemma 5.4.16. *Under the assumptions of Proposition 5.4.5 (without the need for the assumption that $\rho < 1/2$), for all $x, x' \in \mathcal{X}$, $\theta \in \Theta_0$, $k' \geq k > 0$ and $u \in T$, and for all $v, w \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(u)\}$, we have:*

$$|\mathrm{Cov}_\theta[\ \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\Delta(u,k)}, X_{\mathrm{p}^k(u)} = x]| \leq 2\,\|\phi_v\|_\infty \|\phi_w\|_\infty\, \rho^{d(v,w)-2}, \tag{5.75}$$

and,

$$|\mathrm{Cov}_\theta[\ \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\Delta(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathrm{Cov}_\theta[\ \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\Delta(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']| \leq 8\,\|\phi_v\|_\infty \|\phi_w\|_\infty\, \rho^{\min(d(\mathrm{p}^k(u),v),d(\mathrm{p}^k(u),w))-2}. \tag{5.76}$$

Moreover, if $v, w \in \Delta^*(u, k)$, then we have:

$$|\mathrm{Cov}_\theta[\ \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\Delta(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathrm{Cov}_\theta[\ \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\Delta^*(u,k)}, X_{\mathrm{p}^k(u)} = x]| \leq 8\,\|\phi_v\|_\infty \|\phi_w\|_\infty\, \rho^{\min(d(u,v),d(u,w))-2}. \tag{5.77}$$

Proof. We start by proving (5.75), that is, the first inequality in the lemma. Let A_1, A_2, B_1, B_2 be measurable subsets of $\mathcal{X}$, and we write $A = A_1 \times A_2$ and $B = B_1 \times B_2$. If one of the two vertices v and

w is an ancestor of the other, say w is an ancestor of v (which implies that w is also an ancestor of $\mathrm{p}(v)$), then using the Markov property of the HMT process (X, Y), we get:

$$\begin{aligned}
&\big|\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A, X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\
&\qquad -\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\big| \\
&\qquad\qquad = \mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\
&\qquad\qquad\qquad \times \Big| \int_{\mathcal{X}^2} \mathbb{1}_{\{(x_{\mathrm{p}(v)}, x_v) \in A\}} \, \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}(v)} = x_{\mathrm{p}(v)}) \\
&\qquad\qquad\qquad\qquad \times [\mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_w \in B_2, X_{\mathrm{p}^k(u)} = x) \\
&\qquad\qquad\qquad\qquad\qquad -\mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)]\Big| \\
&\qquad\qquad \le \rho^{d(\mathrm{p}(v),w)} \\
&\qquad\qquad = \rho^{d(v,w)-1},
\end{aligned}$$

where the inequality follows using the same argument as in the proof of the "backward in time" coupling Lemma 5.4.1 (with the role of $\mathrm{p}(u)$ replaced by w and using the initial distributions $\mathbb{P}((X_{\mathrm{p}(w)}, X_w) \in \cdot \,|\, Y_{\mathbb{\Delta}(u,k)}, X_w \in B_2, X_{\mathrm{p}^k(u)} = x)$ with $B' = B$ and $\mathcal{X}$ respectively).

Otherwise, we have that both $\mathrm{p}(v)$ and $\mathrm{p}(w)$ are on the path between v and w, and similarly to the first case, we get:

$$\begin{aligned}
&\big|\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A, X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\
&\qquad -\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\big| \\
&\qquad\qquad = \mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\
&\qquad\qquad\qquad \times \Big| \int_{\mathcal{X}^2} \mathbb{1}_{\{(x_{\mathrm{p}(v)}, x_v) \in A\}} \, \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}(v)} = x_{\mathrm{p}(v)}) \\
&\qquad\qquad\qquad\qquad \times [\mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}(w)} \in B_1, X_{\mathrm{p}^k(u)} = x) \\
&\qquad\qquad\qquad\qquad\qquad -\mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)]\Big| \\
&\qquad\qquad \le \rho^{d(\mathrm{p}(v),\mathrm{p}(w))} \\
&\qquad\qquad = \rho^{d(v,w)-2}.
\end{aligned}$$

Thus, in both case, we get:

$$\begin{aligned}
&\big|\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A, X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\
&\qquad -\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\big| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \le \rho^{d(v,w)-2}.
\end{aligned}$$

This gives that (5.75) holds. (Note that the functions $\phi_{\theta,v}$ and $\phi_{\theta,w}$ can take positive, null or negative values.)

For (5.76), that is, the second inequality in the lemma, use the decomposition:

$$\begin{aligned}
&|\mathrm{Cov}_\theta[\, \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathrm{Cov}_\theta[\, \phi_{\theta,v}, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']| \\
&\qquad \le |\mathbb{E}_\theta[\, \phi_{\theta,v}\, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\, \phi_{\theta,v}\, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']| \\
&\qquad\quad + |\mathbb{E}_\theta[\, \phi_{\theta,v} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\, \phi_{\theta,v} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']| \\
&\qquad\qquad\qquad \times |\mathbb{E}_\theta[\, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x]| \\
&\qquad\quad + |\mathbb{E}_\theta[\, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\, \phi_{\theta,w} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']| \\
&\qquad\qquad\qquad \times |\mathbb{E}_\theta[\, \phi_{\theta,v} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']|,
\end{aligned}$$

and then use the joint coupling Lemma 5.4.14 with $v' = \mathrm{p}(v)$ and $w' = \mathrm{p}(w)$ for the first term in the upper bound, and use the coupling Lemma 5.3.2 for the other two terms with $\mathrm{p}(v)$ and $\mathrm{p}(w)$, respectively.

For (5.77), that is, the third inequality in the lemma, use a similar decomposition as for (5.76), and then use the "backward in time" coupling for two vertices from Lemma 5.4.15 for the first term in the upper bound, and use the "backward in time" coupling Lemma 5.4.1 for the other two terms. This gives an upper bound of $8\,\|\phi_v\|_\infty\|\phi_w\|_\infty\,\rho^m$ with $m = \min\{d(\mathrm{p}(u), w_0) \,:\, w_0 \in \{\mathrm{p}(v), v, \mathrm{p}(w), w\}\}$. Noting $m \le \min(d(u,v), d(u,w)) - 2$, we get that (5.77) holds. This concludes the proof of the lemma. □

We are now ready to prove the following lemma which is the adaptation of Lemma 5.4.8 to the terms $\Gamma_{u,k,x}(\theta)$, and which gives us a uniform $L^2(\mathbb{P}_{\theta^\star})$ approximation bound.

Note that the condition $\rho < 1/2$ on the mixing rate ρ of the HMT process (X, Y) is due to the coupling bounds from Lemma 5.4.16 and the grouping of terms used in the proof of Lemma 5.4.17 (the upper bounds at the end of the proof only add a constant multiplicative factor). See the discussion in Remark 5.1.5.

Lemma 5.4.17. *Under the assumptions of Proposition 5.4.5, there exists a positive constant $C < \infty$ such that for all $u \in T$ and $0 < k \le k'$, we have:*

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\Theta_0}\sup_{x,x'\in\mathcal{X}} |\Gamma_{u,k,x}(\theta) - \Gamma_{u,k',x'}(\theta)|^2\right]^{1/2} \le C\,\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\Theta_0}\sup_{x,x'\in\mathcal{X}} |\varphi_\theta(x,x',Y_\partial)|^4\right]^{1/2} k^2\,(2\rho)^{k/3}.$$

Proof. Let u, k and k' be as in the lemma. Similarly to the proof of Lemma 5.4.2, we use the bounds from Lemma 5.4.16 and Minkowski's inequality to bound the left hand side of the inequality in the lemma. For a finite subset $I \subset T^\infty$, we write $S_I = \sum_{v\in I}\phi_{\theta,v}$ (the dependence on θ is implicit). Similarly to the proof of [DMR04, Lemma 17], the difference $\Gamma_{u,k,x}(\theta) - \Gamma_{u,k',x'}(\theta)$ may be rewritten as $A+2B+C+D+2E+2F$, where all those terms are random variables which depend on $Y_{\mathbb{\Delta}(u,k')}$ and implicitly on $\mathcal{U}$, and are define as:

$$\begin{aligned}
A &= \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{\mathrm{p}^k(u)} = x] \\
&\quad - \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'] + \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}^*(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'], \\
B &= \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k)}, \phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] \\
&\quad - \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k)}, \phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'], \\
C &= \mathrm{Var}_\theta[\phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathrm{Var}_\theta[\phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'], \\
D &= \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'] \\
&\quad - \mathrm{Var}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}^*(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'], \\
E &= \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)}, S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'] \\
&\quad - \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)}, S_{\mathbb{\Delta}^*(u,k)} \,|\, Y_{\mathbb{\Delta}^*(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'], \\
F &= \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)}, \phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'] \\
&\quad - \mathrm{Cov}_\theta[S_{\mathbb{\Delta}^*(u,k')\setminus\mathbb{\Delta}^*(u,k)}, \phi_{\theta,u} \,|\, Y_{\mathbb{\Delta}^*(u,k')}, X_{\mathrm{p}^{k'}(u)} = x'].
\end{aligned}$$

Using Minkowski's inequality, we will upper bound each of those six terms separately. First remark using Cauchy-Schwarz inequality, the stationarity of the process $((X_u, Y_u), u \in T^\infty)$ and the assumptions in the proposition, that we have $\mathbb{E}_{\theta^\star}[\|\phi_v\|_\infty^2\|\phi_w\|_\infty^2] \le \mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4] < \infty$ for all $v, w \in T^\infty$.

Remind from (5.12) on page 129 that $\Delta(u,k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$.

Upper bound for A: Applying the three inequalities in Lemma 5.4.16 and Minkowski's inequality, we get that $\mathbb{E}_{\theta^\star}[|A|^2]^{1/2}$ is upper bounded (up to the factor $\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4]$) by:

$$\begin{aligned}
&2\sum_{v,w\in T^\infty(\mathrm{p}^k(u),k)} \left(2\times 8\rho^{\min(d(v,u),d(w,u))-2} \wedge 2\times 8\rho^{\min(d(v,\mathrm{p}^k(u)),d(w,\mathrm{p}^k(u)))-2} \wedge 4\times 2\rho^{d(v,w)-2}\right) \\
&\le \frac{32}{\rho^2}\sum_{v,w\in T^\infty(\mathrm{p}^k(u),k)} \left(\rho^{\min(d(v,u),d(w,u))} \wedge \rho^{\min(d(v,\mathrm{p}^k(u)),d(w,\mathrm{p}^k(u)))} \wedge \rho^{d(v,w)}\right). \qquad (5.78)
\end{aligned}$$

Note that the value of this sum does not depend on the choice of $u \in T$.

For all $j \in \mathbb{N}$, denote $u_j = \mathrm{p}^j(u)$. We will divide the sum in the upper bound of (5.78) according to four cases: $v, w \in T^\infty(u_k, k) \setminus T^\infty(u_{\lfloor k/3 \rfloor}, \lfloor k/3 \rfloor)$, or $v, w \in T^\infty(u_{\lfloor 2k/3 \rfloor}, \lfloor 2k/3 \rfloor)$, or $v \in T^\infty(u_k, k) \setminus T^\infty(u_{\lfloor 2k/3 \rfloor}, \lfloor 2k/3 \rfloor)$ and $w \in T^\infty(u_{\lfloor k/3 \rfloor}, \lfloor k/3 \rfloor)$ or similarly exchanging the roles of v and w. Note that those conditions are non-exclusive and we will count some vertices several times, but this is not a problem.

Let $i, j \in \mathbb{N}$ be such that $u \wedge v = u_i$ and $u \wedge w = u_j$, and let $a, b \in \mathbb{N}$ be such that $a = d(u_i, v)$ and $b = d(u_j, w)$. Note that for v, w in the first case, either $\min(d(v, u), d(w, u))$ or $d(v, w)$ is large, and thus using elementary computation we upper bound the sum for v, w in the first case by:

$$\begin{aligned}
&\sum_{i=\lfloor k/3 \rfloor}^{k} \sum_{j=\lfloor k/3 \rfloor}^{k} \sum_{a=0}^{i} \sum_{b=0}^{j} 2^{a+b} \big(\rho^{\min(i+a, j+b)} \wedge \rho^{a+b+|j-i|}\big) \\
&\leq 2 \sum_{i=\lfloor k/3 \rfloor}^{k} \sum_{j=i}^{k} \sum_{a=0}^{i} \sum_{b=0}^{j} 2^{a+b} \big(\rho^{i+\min(a,b)} \wedge \rho^{a+b}\big) \\
&\leq \frac{8(1-\rho)}{(1-2\rho)^3}\, k\, (2\rho)^{\lfloor k/3 \rfloor + 1}.
\end{aligned}$$

Note that for v, w in the second case, either $\min(d(v, \mathrm{p}^k(u)), d(w, \mathrm{p}^k(u))$ or $d(v, w)$ is large, and thus using elementary computation we upper bound the sum for v, w in the second case by:

$$\sum_{i=0}^{\lfloor 2k/3 \rfloor} \sum_{j=0}^{\lfloor 2k/3 \rfloor} \sum_{a=0}^{i} \sum_{b=0}^{j} 2^{a+b} \big(\rho^{\min(k-i+a, k-j+b)} \wedge \rho^{a+b+|j-i|}\big) \leq \frac{8(1-\rho)}{(1-2\rho)^3}\, k\, (2\rho)^{\lfloor k/3 \rfloor + 1}.$$

Note that for v, w in the third and fourth case, either $\min(d(v, \mathrm{p}^k(u)), d(w, \mathrm{p}^k(u))$ or $d(v, w)$ is large, and thus we upper bound the sum for v, w in the third and fourth case by:

$$2 \sum_{i=0}^{\lfloor k/3 \rfloor} \sum_{j=\lfloor 2k/3 \rfloor}^{k} \sum_{a=0}^{i} \sum_{b=0}^{j} 2^{a+b} \rho^{a+b+|j-i|} \leq \frac{2}{1-2\rho}\, k^2\, \rho^{k/3-1}.$$

Putting those three upper bounds together, we get that $\mathbb{E}_{\theta^\star}[|A|^2]^{1/2}$ is upper bounded by an expression as in the lemma.

Upper bound for B: Using the first and second inequalities in Lemma 5.4.16 and Minkowski's inequality, we get that $\mathbb{E}_{\theta^\star}[|B|^2]^{1/2}$ is upper bounded (up to the factor $\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4]$) by:

$$8 \sum_{v \in \mathbb{\Delta}^*(u,k)} (\rho^{d(v,u)-2} \wedge \rho^{d(v, \mathrm{p}^k(u))-2}) \leq C\, k\, \big(\max(\rho, 2\rho^2)\big)^{k/2},$$

where we used the same computation as in the proof of Lemma 5.4.2 and $C < \infty$ is some finite constant (which depends only on ρ).

Upper bound for C: Using the second equation in Lemma 5.4.16, we get that $\mathbb{E}_{\theta^\star}[|C|^2]^{1/2} \leq 8\rho^{k-2}\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^2]$.

Upper bound for D: Using the first and third equation in Lemma 5.4.16, Minkowski's inequality and elementary computation, we get that $\mathbb{E}_{\theta^\star}[|D|^2]^{1/2}$ is upper bounded (up to the factor $\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4]$) by:

$$\sum_{v,w \in T^\infty(u_{k'}, k') \setminus T^\infty(u_k, k)} (2 \times 2\rho^{d(v,w)-2} \wedge 8\rho^{\min(d(v,u), d(w,u))-2}) \leq \frac{96}{\rho^2(1-2\rho)^4}\, k\, (2\rho)^k.$$

Upper bound for E: Using the first and third equation in Lemma 5.4.16, Minkowski's inequality and elementary computation, we get that $\mathbb{E}_{\theta^\star}[|E|^2]^{1/2}$ is upper bounded (up to the factor $\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4]$) by:

$$\begin{aligned}
&\sum_{v \in T^\infty(u_{k'}, k') \setminus T^\infty(u_k, k)} \sum_{w \in T^\infty(u_k, k)} (2 \times 2\rho^{d(v,w)-2} \wedge 8\rho^{\min(d(v,u), d(w,u))-2}) \\
&\qquad\qquad \leq \frac{64}{\rho^2(1-\rho)(1-2\rho)^2}\, k^2\, (2\rho)^{\lfloor k/2 \rfloor}.
\end{aligned}$$

Upper bound for F: Using the first equation in Lemma 5.4.16 and Minkowski's inequality, we get that $\mathbb{E}_{\theta^\star}[|F|^2]^{1/2}$ is upper bounded (up to the factor $\mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4]$) by:

$$2 \times 2 \sum_{v \in T^\infty(u_{k'}, k') \setminus T^\infty(u_k, k)} \rho^{d(v,u)-2} \leq \frac{4}{1-2\rho} \frac{\rho^{k-1}}{1-\rho}.$$

Hence, as the $L^2(\mathbb{P}_{\theta^\star})$ norm for the six terms A, B, C, D, E and F are all upper bounded by expressions as in the lemma, we get that the upper bound in the lemma holds. This concludes the proof. □

As annonce at the beginning of this subsection, the rest of the proof of Proposition 5.4.5 closely follows the lines of the proof of Proposition 5.4.4.

5.5 Extension to the non-stationary case

In Sections 5.3 and 5.4, the stationarity assumption of the process $(Y_u : u \in T)$ played a crucial role. In this section, we extend the strong consistency and the asymptotic normality of the MLE for the HMT to the case where this process is not stationary.

Hence, we assume that the HMT process $(X', Y') = ((X'_u, Y'_u) : u \in T)$ has the same transition kernel $Q_{\theta^\star}$ and $G_{\theta^\star}$ that are parametrized by some $\theta^\star \in \Theta$ as before, and the hidden variable X'_∂ of the root vertex ∂ has distribution ζ. This initial distribution ζ is unknown to us, may depend on $\theta^\star$, and in general is different from the invariant distribution $\pi_{\theta^\star}$. As before, we will denote by $(X, Y) = ((X_u, Y_u) : u \in T)$ a stationary process distributed according to the same parameter $\theta^\star$. Note that, in this section, we will use the convention that objects with an added $'$ symbol are related to the non-stationary process (X', Y'), while those without the $'$ symbol are their counterpart for the stationary process (X, Y). Also note that due to the non-stationarity assumption, in this section, we will only consider the HMT process on the original tree $T^\infty = T(\partial)$.

For the non-stationary process (X', Y'), similarly to the stationary case in (5.7) on page 125, define its log-likelihood for all $n \in \mathbb{N}$ and $x \in \mathcal{X}$ as:

$$\ell'_{n,x}(\theta) := \ell_{n,x}(\theta; Y'_{T_n}), \tag{5.79}$$

where $\ell_{n,x}(\theta; \cdot)$ is defined in (5.6) on page 125. Moreover, when Assumptions 5-8 and 10 hold and Θ is compact, similarly to the stationary case in (5.33) on page 135, we define the MLE $\hat{\theta}'_{n,x}$ for the non-stationary process as:

$$\hat{\theta}'_{n,x} = \hat{\theta}'_{n,x}(Y'_{T_n}) \in \operatorname{argmax}_{\theta \in \Theta} \ell'_{n,x}(\theta). \tag{5.80}$$

Denote by $\mathbb{P}_{\theta^\star,\zeta}$ the probability distribution of the non-stationary HMT process (X', Y'), and by $\mathbb{E}_{\theta^\star,\zeta}$ the corresponding expectation.

We can now prove the strong consistency of the MLE for a non-stationary HMT process.

Theorem 5.5.1 (Strong consistency of the MLE, non-stationary case)**.** *Assume that Assumptions 6–10 hold. the contrast function ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 5.3.10) and Θ is compact. Then, the MLE is strongly consistent, that is, for all initial distributions ζ and all $x \in \mathcal{X}$, the MLE $\hat{\theta}'_{n,x}$ converges $\mathbb{P}_{\theta^\star,\zeta}$-a.s. as $n \to \infty$ to the true parameter $\theta^\star \in \Theta$.*

Proof. We start by proving that for any $n \in \mathbb{N}^*$, the distribution of the non-stationary HMT process (X', Y') on T^* is absolutely continuous w.r.t. the distribution of the stationary HMT process (X, Y) on T^*, that is:

$$\mathbb{P}_{\theta^\star,\zeta}(X'_{T^*} \in \cdot, Y'_{T^*} \in \cdot) \ll \mathbb{P}_{\theta^\star,\pi_{\theta^\star}}(X'_{T^*} \in \cdot, Y'_{T^*} \in \cdot) = \mathbb{P}_{\theta^\star}(X_{T^*} \in \cdot, Y_{T^*} \in \cdot). \tag{5.81}$$

Remind that Assumption 7-(i) implies that $\pi_{\theta^\star} \ll \lambda$ with density $\frac{\mathrm{d}\pi_{\theta^\star}}{\mathrm{d}\lambda}$ taking value in $[\sigma^-, \sigma^+]$. Denote by u_1 and u_2 the two children vertices of ∂. Using Assumption 7, for any non-negative measurable

function f from $\mathcal{X}^2$ to $\mathbb{R}_+$, we get:

$$\begin{aligned}\int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2})\, \mathbb{P}_{\theta^\star,\zeta}(X'_{u_1} \in \mathrm{d}x_{u_1}, X'_{u_2} \in \mathrm{d}x_{u_2}) \\ &= \int_{\mathcal{X}^3} f(x_{u_1}, x_{u_2})\, q_{\theta^\star}(x_\partial, x_{u_1}) q_{\theta^\star}(x_\partial, x_{u_2})\, \lambda(\mathrm{d}x_{u_1})\lambda(\mathrm{d}x_{u_2})\zeta(\mathrm{d}x_\partial) \\ &\le \left(\frac{\sigma^+}{\sigma^-}\right)^2 \int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2})\, \pi_{\theta^\star}(\mathrm{d}x_{u_1})\pi_{\theta^\star}(\mathrm{d}x_{u_2}) \\ &= \left(\frac{\sigma^+}{\sigma^-}\right)^2 \int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2})\, \mathbb{P}_{\theta^\star}(X_{u_1} \in \mathrm{d}x_{u_1}, X_{u_2} \in \mathrm{d}x_{u_2}).\end{aligned}$$

In particular, for any measurable subset A of $\mathcal{X}^{T^*} \times \mathcal{Y}^{T^*}$, we can choose f to be define as:

$$\begin{aligned} f(x_{u_1}, x_{u_2}) &= \mathbb{E}_{\theta^\star,\zeta}[\mathbb{1}_A(X'_{T^*}, Y'_{T^*}) \,|\, X'_{u_1} = x_{u_1}, X'_{u_2} = x_{u_2}] \\ &= \mathbb{E}_{\theta^\star}[\mathbb{1}_A(X_{T^*}, Y_{T^*}) \,|\, X_{u_1} = x_{u_1}, X_{u_2} = x_{u_2}] \end{aligned}$$

Hence, we get that (5.81) holds.

Using (5.81), we get that Proposition 5.3.7 also holds $\mathbb{P}_{\theta^\star,\zeta}$-a.s. with $\ell_{n,x}(\theta)$ replaced by $\ell'_{n,x}(\theta)$, that is, in the non-stationary case. Thus, the proof of Theorem 5.3.11 can be immediately adapted to the non-stationary case (note that Propositions 5.3.6 and 5.3.10 state properties of the contrast function ℓ, which is the same in the stationary and non-stationary cases). This concludes the proof of the Theorem. □

Using a similar argument as for Theorem 5.5.1, we get that in the non-stationary case, the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell'_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star,\zeta}$-a.s. locally uniformly to the limiting Fisher information $\mathcal{I}(\theta^\star)$ (which is defined in (5.54)). Note that the condition $\rho < 1/2$ on the mixing rate ρ of the HMT process (X, Y) is inherited from Theorem 5.4.6 in the stationary case. See the discussion in Remark 5.1.5 for comments on this condition on ρ.

Theorem 5.5.2 (Convergence of the normalized observed information, non-stationary case)**.** *Assume that Assumptions 6–8 and 10–13 hold. Assume that $\rho < 1/2$. Assume that Θ is compact. Then, for all initial distributions ζ and all $x \in \mathcal{X}$, we have:*

$$\lim_{\delta\to 0} \lim_{n\to\infty} \sup_{\theta\in\mathcal{O}\,:\,\|\theta-\theta^\star\|\le\delta} \left\| -|T_n|^{-1}\nabla^2_\theta \ell'_{n,x}(\theta) - \mathcal{I}(\theta^\star) \right\| = 0 \quad \mathbb{P}_{\theta^\star,\zeta}\text{-a.s.}$$

In particular, combining Theorems 5.5.1 and 5.5.2, we get that the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\hat{\theta}_{n,x})$ at the MLE $\hat{\theta}_{n,x}$ is a strongly consistent estimator of the Fisher information matrix $\mathcal{I}(\theta^\star)$.

Before proving the asymptotic normality of the MLE in the non-stationary case, we start with the following lemma which present a coupling construction for the two processes (X, Y) and (X', Y').

Lemma 5.5.3 (Coupling construction of two HMTs)**.** *Assume that Assumptions 6 and 7 hold. Further assume that $\sigma^- \ge 1/2$. Then, it is possible to construct the two processes (X, Y) and (X', Y') on a common probability space such that there exists an a.s. finite random time N, which we call the* coupling time*, such that $(X_u, Y_u) = (X'_u, Y'_u)$ for all $u \in T_n$ with $n \ge N$.*

We will denote by $\mathbb{P}_{\theta^\star\bowtie\zeta}$ the probability distribution that realizes this coupling.

Note that $\rho \le 1/2$ implies that $\sigma^- \ge \sigma^+/2 \ge 1/2$ (see Assumption 7).

Proof. We first construct the coupling only for the process X and X'. We define the coupling construction inductively on the height of the tree. For the root vertex, we use an independent coupling construction for X_∂ and X'_∂, which are distributed according to $\pi_{\theta^\star}$ and ζ respectively (note that it is also possible to use a perfect coupling with probability error $\|\pi_{\theta^\star} - \zeta\|_{\mathrm{TV}}$). Then, if the coupling has been constructed up to generation $n \in \mathbb{N}$, using the Markov property, we proceed to construct independently the coupling for each vertices in generation $n+1$. Let $u \in G_{n+1}$. If the variables were already coupled for the parent vertex $\mathrm{p}(u)$, that is $X_{\mathrm{p}(u)} = X'_{\mathrm{p}(u)}$, then we choose the new value $X_u = X'_u$ according to the transition kernel $Q_{\theta^\star}$. Otherwise, if $X_{\mathrm{p}(u)} \ne X'_{\mathrm{p}(u)}$, then using the uniform geometric ergodicity (remind Assumption 7)

of the transition kernel $Q_{\theta^\star}$, we know that $\sup_{x,x'\in\mathcal{X}} \|Q_{\theta^\star}(x;\cdot) - Q_{\theta^\star}(x';\cdot)\|_{\mathrm{TV}} \leq 1-\sigma^-$, and thus we can construct a coupling of X_u and X'_u conditionally on $X_{\mathrm{p}(u)} \neq X'_{\mathrm{p}(u)}$ with exact matching probability at least $1-\sigma^-$. We have constructing the matching for u, and thus for the whole generation $n+1$. Using Kolmogorov extension theorem, there exists a coupling measure for the whole tree T whose finite dimensional marginals are the ones given above.

Remark that the joint process (X,Y) satisfies a uniform geometric ergodicity bound with the same constant $1-\sigma^-$. Thus, the construction above can be extended to the joint process (X,Y). Denote by $\mathbb{P}_{\theta^\star\bowtie\zeta}$ the probability distribution of the coupling we have constructed for the joint process (X,Y).

Define the random coupling time $N = \inf\{n\in\mathbb{N} \,|\, \forall u\in G_n,\, X_u = X'_u\}$, which is the first generation for which the exact coupling occurs for all vertices (and $N=\infty$ is this never happens). We are left to prove that $\mathbb{P}_{\theta^\star\bowtie\zeta}$-a.s. $N<\infty$. We say that a vertex u is a special vertex if $X_u \neq X'_u$. Note that if u is not special, then all its descendants are also not special. Also note that special vertices form a Bienaymé-Galton-Watson tree whose (homogeneous) offspring distribution takes the values: 0 with probability $(\sigma^-)^2$; 1 with probability $2\sigma^-(1-\sigma^-)$; and 2 with probability $(1-\sigma^-)^2$. The average of this offspring distribution is $2(1-\sigma^-)$. Hence, the number of special vertices if finite $\mathbb{P}_{\theta^\star\bowtie\zeta}$-a.s., that is, N is finite $\mathbb{P}_{\theta^\star\bowtie\zeta}$-a.s., if and only if $2(1-\sigma^-)\leq 1$, that is, $\sigma^-\geq 1/2$. This concludes the proof. □

Remind that the log-likelihood function $\ell_{n,x}$ (resp. $\ell'_{n,x}$), which is a random function depending on Y_{T_n} from the stationary HMT process (resp. on Y'_{T_n} from the non-stationary HMT process), is defined in (5.7) on page 125 (resp. just before (5.79) on page 159). For all $\theta\in\Theta$, define:

$$\begin{aligned} D_{n,x}(\theta) &= \ell'_{n,x}(\theta) - \ell_{n,x}(\theta) \\ &= \sum_{u\in T_n} \log p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x), \end{aligned}$$

where remind that p_θ denotes possibly conditional density (see (5.5) on page 125).

Remind that when Assumptions 5-8 and 10 hold and Θ is compact, for all $x\in\mathcal{X}$, the MLE $\hat\theta_{n,x}$ (resp. $\hat\theta'_{n,x}$) is a random variable which depends on Y_{T_n} from the stationary HMT process (resp. on Y'_{T_n} from the non-stationary HMT process) and is defined in (5.33) on page 135 (resp. in (5.80) on page 159).

To prove that $\lim_{n\to\infty} |T_n|^{1/2}(\hat\theta'_{n,x} - \hat\theta_{n,x}) = 0$ $\mathbb{P}_{\theta^\star\bowtie\zeta}$-a.s., and thus that $\hat\theta'_{n,x}$ and $\hat\theta_{n,x}$ are asymptotically normal with the same covariance matrix (remind Theorem 5.4.7), we must first prove that the function $\theta\mapsto D_{n,x}(\theta)$ satisfies some kind of continuity property. Note that we proved in the proof of Theorem 5.5.1 that Proposition 5.3.7 holds both in the stationary and the non-stationary cases, and thus we have:

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \Big| |T_n|^{-1} D_{n,x}(\theta) \Big| = 0 \quad \mathbb{P}_{\theta^\star\bowtie\zeta}\text{-a.s.}$$

However, we need some kind of continuity property without the normalizing term $||T_n|^{-1}$, which is given by the following lemma.

Lemma 5.5.4. *Assume that Assumptions 6–10 hold. Further assume that $\rho<1/2$. Then, for all initial distributions ζ and all $x\in\mathcal{X}$, we have:*

$$\lim_{n\to\infty} |D_{n,x}(\hat\theta'_{n,x}) - D_{n,x}(\hat\theta_{n,x})| = 0, \quad \mathbb{P}_{\theta^\star\bowtie\zeta}\text{-a.s.}$$

Proof. [The proof is a straightforward adaptation of the proof of [DMR04, Lemmas 11 and 12].]

Let N be the random time provided by Lemma 5.5.3. We first prove that $\mathbb{P}_{\theta^\star\bowtie\zeta}$-a.s., we have:

$$\sum_{u\in T\setminus T_N} \sup_{\theta\in\Theta} |\log p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x)| < \infty. \tag{5.82}$$

Note that for all $u\in T_n$ and v an ancestor of u (distinct of u), we have:

$$\begin{aligned} &p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x) \\ &\quad = \int_{\mathcal{X}^3} g_\theta(x_u, Y_u) q_\theta(x_{\mathrm{p}(u)}, x_u)\, \lambda(\mathrm{d}x_u) \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, X_v = x_v, Y_{\Delta^*(u,h(u)-h(v))}) \\ &\qquad\qquad \times \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x), \end{aligned}$$

and similarly for $p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x)$. Using the fact that for $v \in T$ with height $h(v) \geq N$, we have $Y_v = Y'_v$, and using Lemma 5.3.2, we have for all $u \in T$ with height $h(u) > N$:

$$\begin{aligned}|p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x)| \\ \leq 2\,\rho^{h(u)-N-1}\sigma^+ \int g_\theta(x, Y_u)\,\lambda(\mathrm{d}x).\end{aligned}$$

Thus, using a similar argument as in the proof of Lemma 5.3.3, we get:

$$|\log p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x)| \leq \frac{\rho^{h(u)-N-1}}{1-\rho}.$$

Hence, the sum in (5.82) is $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. upper bounded by a constant times $\sum_{k=N+1}^{\infty} 2^k \rho^k \leq (2\rho)^{N+1}/(1-2\rho)$, and is thus finite $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s.

Let $\varepsilon > 0$. Using (5.82), there exists a random integer N_ε which $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. is finite and satifies:

$$\sum_{u \in T \setminus T_{N_\varepsilon}} \sup_{\theta \in \Theta} |\log p_\theta(Y'_u \,|\, Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u \,|\, Y_{\Delta^*(u,h(u))}, X_\partial = x)| \leq \varepsilon.$$

Thus, $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s., for all $n \geq N_\varepsilon$, we have:

$$|D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x})| \leq 2\varepsilon + |\ell'_{N_\varepsilon,x}(\hat{\theta}'_{n,x}) - \ell'_{N_\varepsilon,x}(\hat{\theta}_{n,x})| + |\ell_{N_\varepsilon,x}(\hat{\theta}'_{n,x}) - \ell_{N_\varepsilon,x}(\hat{\theta}_{n,x})|.$$

Note that under the given assumptions, the functions $\theta \mapsto \ell'_{N_\varepsilon,x}(\theta)$ and $\theta \mapsto \ell_{N_\varepsilon,x}(\theta)$ are continuous $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. (see the proof of Proposition 5.3.6). Hence, the proof is complete upon observing that $\hat{\theta}'_{n,x}$ and $\hat{\theta}_{n,x}$ both converge $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. to $\theta^\star$ (see Theorem 5.5.1), and that ε was arbitrary. □

We can now prove the asymptotic normality of the MLE $\hat{\theta}'_{n,x}$ in the non-stationary case. Remind that the contrast function ℓ is defined in (5.26) on page 132.

Theorem 5.5.5 (Asymptotic normality of the MLE, non-stationary case)**.** *Assume that Assumptions 6–13 hold. Assume that $\rho < 1/2$. Further assume that the contrast function ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 5.3.10) and that Θ is compact, $\theta^\star$ is an interior point of Θ, and the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ (which is defined in* (5.54)*) is non-singular. Then, for all initial distributions ζ and for all $x \in \mathcal{X}$, we have:*

$$|T_n|^{1/2}\big(\hat{\theta}'_{n,x} - \theta^\star\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1}) \quad \textit{under } \mathbb{P}_{\theta^\star,\zeta},$$

where $\mathcal{N}(0, M)$ denotes the centered Gaussian distribution with covariance matrix M.

Proof. [The proof is a straightforward adaptation of the proof of [DMR04, Theorem 6].]

Define $\varepsilon_n = |T_n|^{1/2}(\hat{\theta}_{n,x} - \hat{\theta}'_{n,x})$ for all $n \in \mathbb{N}$, and remark that it is sufficient to prove that $\lim_{n\to\infty} \varepsilon_n = 0$ $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. Since $\hat{\theta}'_{n,x}$ is the maximizer of the function $\theta \mapsto \ell'_{n,x}(\theta)$, we have that $\ell'_{n,x}(\hat{\theta}'_{n,x}) \geq \ell'_{n,x}(\hat{\theta}_{n,x})$. Thus, using a Taylor expansion of $\ell_{n,x}$ around its maximizer $\hat{\theta}_{n,x}$ (for which we have $\nabla_\theta \ell_{n,x}(\hat{\theta}_{n,x}) = 0$), we get that there exists $t_n \in [0,1]$ such that:

$$\begin{aligned}D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x}) &\geq \ell_{n,x}(\hat{\theta}_{n,x}) - \ell_{n,x}(\hat{\theta}'_{n,x}) \\ &= -\frac{1}{2}|T_n|^{-1}\varepsilon_n^t \nabla^2_\theta \ell_{n,x}(t_n\hat{\theta}'_{n,x} + (1-t_n)\hat{\theta}_{n,x})\varepsilon_n.\end{aligned}$$

Note that we have $\lim_{n\to\infty} t_n\hat{\theta}'_{n,x} + (1-t_n)\hat{\theta}_{n,x} = \theta^\star$ $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. by Theorem 5.5.1. Thus, applying Theorem 5.4.6, we have:

$$\lim_{n\to\infty} -|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(t_n\hat{\theta}'_{n,x} + (1-t_n)\hat{\theta}_{n,x}) = \mathcal{I}(\theta^\star), \quad \mathbb{P}_{\theta^\star \bowtie \zeta}\text{-a.s.}$$

As $\mathcal{I}(\theta^\star)$ is positive definite, there exits $M > 0$ such that on a set with $\mathbb{P}_{\theta^\star \bowtie \zeta}$-probability one and for n sufficiently large, we have:

$$D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x}) \geq M|\varepsilon_n|^2.$$

Then, the proof is complete by applying Lemma 5.5.4. □

5.A Ergodic theorem for Markov processes indexed by trees with neighborhood-dependent functions

In this appendix, we prove generalization of the ergodic theorems in [Guy07] and in [Wei24b], which give a.s. and L^2 convergences for branching Markov chains, to allow for neighborhood-dependent functions. Those ergodic theorems are used to prove the ergodic convergence lemmas in Section 5.2.4 which are used in the main body of this article. Remind that we need those generalization as in the study of asymptotic property of the MLE for the HMT relies on the study of the likelihood contribution functions $\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)})$ (defined in (5.17) on page 129) which are neighborhood-dependent.

Remind from Section 5.2 that if (X, Y) is a HMT process, then the joint process $((X_u, Y_u), u \in T)$ is a branching Markov chain. Thus, it is enough to prove those ergodic theorems for branching Markov chains instead of HMT processes.

Let Q be a transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ where $\mathcal{X}$ is a metric space. We assume that Q has a unique invariant probability distribution π and is uniformly geometrically ergodic, that is, there exists $\rho \in (0,1)$ and a finite positive constant C such that for all $x \in \mathcal{X}$, we have $\|Q^n(x;\cdot) - \pi\|_{\mathrm{TV}} \leq C\rho^n$. Remind from Lemma 5.2.3 that this covers the case $Q = Q_\theta$ for any $\theta \in \Theta$. Let $X = (X_u, u \in T)$ be a branching Markov chain with transition kernel Q and initial distribution π. Denote by $\mathbb{P}_Q$ the probability distribution of the process X, and by $\mathbb{E}_Q$ the corresponding expectation.

In this section, for a probability measure ν on $\mathcal{X}$, a transition kernel Q on $(\mathcal{X}, \mathcal{B}(\mathcal{Z}))$ and a Borel integrable function f on $\mathcal{B}(\mathcal{Z})$ where $\mathcal{Z} = \mathcal{X}^A$ for some finite subset $A \subset T$, we will write νQ for the image probability measure $(\nu Q)(\cdot) = \int_{\mathcal{X}} Q(x;\cdot)\,\nu(\mathrm{d}x)$, and Qf for the Borel function $(Qf)(x) = \int_{\mathcal{Z}} f(z)\,Q(x;\mathrm{d}z)$. For a probability measure ν on $\mathcal{X}$ and a Borel integrable function f on $\mathcal{X}$, we will write $\langle \nu, f\rangle = \nu f = \int_{\mathcal{X}} f\,\mathrm{d}\nu$.

We will need the following lemma which states geometric convergence bounds for functions in $L^2(\pi)$.

Lemma 5.A.1 (Convergence bounds when Q is uniformly geometrically ergodic)**.** *Assume that the transition kernel Q has a unique invariant measure π, and that Q is uniformly geometrically ergodic. Then, there exists finite positive constants $\alpha \in (0,1)$ and $M < \infty$ such that for all functions $f \in L^2(\pi)$, we have:*

$$\forall n \in \mathbb{N}, \qquad \sup_{k\in\mathbb{N}} \pi Q^k (Q^n f - \langle \pi, f\rangle)^2 = \pi(Q^n f - \langle \pi, f\rangle)^2 \leq M\alpha^{2n} \|f - \langle \pi, f\rangle\|^2_{L^2(\pi)}.$$

In particular, the function f satisfies $\sup_{n\in\mathbb{N}} \pi Q^n f^2 < \infty$.

Note, using Cauchy-Schwarz and Jensen's inequalities, that $\sup_{n\in\mathbb{N}} \pi Q^n f^2 < \infty$ implies that $Q^n f$, $Q^n f^2$, and $Q^k(Q^n f \times Q^m f)$ (with $n, m, k \in \mathbb{N}$) are well-defined and finite π-almost everywhere and are π-integrable.

Proof. Using [DMPS18, Proposition 22.3.5 and Definition 22.3.1], we get that there exists finite positive constants $\alpha \in (0,1)$ and $M < \infty$ such that for all functions $f \in L^2(\pi)$, we have $\pi(Q^n f - \langle \pi, f\rangle)^2 = \|Q^n(f - \langle \pi, f\rangle)\|^2_{L^2(\pi)} \leq M\alpha^{2n}\|f - \langle \pi, f\rangle\|^2_{L^2(\pi)}$ for all $n \in \mathbb{N}$. In particular, we get that $\sup_{n\in\mathbb{N}} \pi Q^n f^2 < \infty$. □

Let $k \in \mathbb{N}$ be fixed. Remind from Section 5.2.4 the definitions of the subtrees $\Delta(u,k)$, of their shapes $\mathcal{S}h\big(\Delta(u,k)\big)$ (defined in (5.8)), and of the finite set of possible shapes $\mathcal{N}_k$ (defined in (5.9)). For simplicity, in this appendix we will write $\mathcal{S}_u$ instead of $\mathcal{S}h\big(\Delta(u,k)\big)$. Also remind from Section 5.2.4 the definition of a collection of neighborhood-shape-dependent functions $(f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$. Remind that for such a collection of functions, we simply write $f_{\Delta(u,k)}$ or $f_{\mathcal{S}_u}$ instead of $f_{\mathcal{S}h(\Delta(u,k))}$. And also remind that we write $f_{\mathcal{S}_u}(X_{\Delta(u,k)})$ for the evaluation of $f_{\mathcal{S}_u} = f_{\Delta(u,k)}$ on $X_{\Delta(u,k)}$. Note that up to translation, we may identify $\mathcal{X}^{\mathcal{S}}$ and $\mathcal{X}^{\Delta(u,k)}$ for any $u \in T \setminus T_{k-1}$ such that $\mathcal{S}_u = \mathcal{S}$.

Remind that any subset $A \subset T$, we denote by X_A the gathered variables $(X_v : v \in A)$. For a collection of neighborhood-shape-dependent functions $\mathrm{f} = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$, define the empirical average of f over a finite subset $A \subset T \setminus T_{k-1}$ as:

$$\bar{M}_A(\mathrm{f}) = |A|^{-1} \sum_{u\in A} f_{\mathcal{S}_u}(X_{\Delta(u,k)}). \tag{5.83}$$

For a neighborhood shape $\mathcal{S} \in \mathcal{N}_k$, let $u \in G_k$ be the unique vertex such that $\mathcal{S}_u = \mathcal{S}$, and define the transition kernel $Q^{\mathcal{S}}$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}^{\mathcal{S}}))$ for any $x \in \mathcal{X}$ and any Borel function f on $\mathcal{X}^{\mathcal{S}} = \mathcal{X}^{\Delta(u)}$ which is in $L^1(X_{\Delta(u)}) = L^1(X_{\mathcal{S}})$ by:

$$Q^{\mathcal{S}} f(x) = \mathbb{E}_Q\Big[f(X_{\Delta(u)}) \;\Big|\; X_\partial = x\Big]. \tag{5.84}$$

That is, from the value $x \in \mathcal{X}$ of the root vertex v in $\mathcal{S}$, the transition kernel $Q^{\mathcal{S}}$ returns the distribution of the Markov process X on $\mathcal{S}$ with transition kernel Q conditioned on the value X_v of the vertex v being x. Note that (5.84) also extends to any vertex $u \in T \setminus T_{k-1}$ such that $\mathcal{S}_u = \mathcal{S}$, which gives us:

$$Q^{\mathcal{S}_u} f(x) = \mathbb{E}_Q\Big[f(X_{\Delta(u,k)}) \;\Big|\; X_{\mathrm{p}^k(u)} = x\Big]. \tag{5.85}$$

Moreover, using Jensen's inequality, note that if $f \in L^2(X_{\mathcal{S}})$, then $Q^{\mathcal{S}} f$ is in $L^2(\pi) = L^2(X_\partial)$.

Remind that as T is a plane rooted tree, we can enumerate its vertices as a sequence $(v_j)_{j\in\mathbb{N}}$ in a breadth-first-search manner, that is, which is increasing for $<$ (note that $u_0 = \partial$). Also remind that, for $n \geq |T_{k-1}|$, if V_n is uniformly distributed over $A_n := \{v_j : |T_{k-1}| < j \leq n\} = \Delta(v_n) \setminus T_{k-1}$, then the distribution of $\mathcal{S}_{V_n}$ converges to the uniform distribution over $\mathcal{N}_k$ as $n \to \infty$.

We are now ready to state the ergodic theorem with neighborhood-shape-dependent functions for branching Markov chains indexed by the infinite complete binary tree T. Remind that $\bar{M}_{A_n}(\mathrm{f})$ is defined in (5.83).

Theorem 5.A.2 (Ergodic theorem with neighborhood-dependent functions)**.** *Let $k \in \mathbb{N}$ be fixed. Let $(v_j)_{j\in\mathbb{N}}$ be the sequence enumerating the vertices in T in a breadth-first-search manner. For all $n > |T_{k-1}|$, define $A_n = \Delta(v_n) \setminus T_{k-1}$.*

Let Q be a transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ which is uniformly geometrically ergodic and has a unique invariant probability measure π. Let $X = (X_u, u \in T)$ be a branching Markov chain with transition kernel Q and initial distribution π.

Let $\mathrm{f} = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(X)$. Then, we have:

$$\bar{M}_{A_n}(\mathrm{f}) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{U_k} \otimes \mathbb{E}_Q\big[f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})\big] \qquad \text{in } L^2(\pi) = L^2(X), \tag{5.86}$$

where U_k is uniformly distributed over G_k and independent of the process X, and $\mathbb{E}_{U_k} \otimes \mathbb{E}_Q$ denotes the joint expectation over U_k and X.

As an immediate corollary, we get that the result still holds if A_n is replaced by $T_n \setminus T_{k-1}$.

Remark 5.A.3 (More general assumptions)**.** Note that without changing the proof, we could replace the subtrees $\Delta(u,k)$ by general subtrees (i.e. a connected subsets) $\mathcal{O}_u$ of the the k-neighborhood $B_T(u,k) := \{v \in T : d(u,v) \leq k\}$ of u such that $\mathcal{O}_u$ contains u. In that case, we must assume that the distribution of the shape (i.e. when seen up to translation) $\mathcal{S}h(\mathcal{O}_{V_n})$ converges to some limit distribution. Also note that we could allow more general choices as in [Wei24b] for the averaging sets $(A_n)_{n\in\mathbb{N}}$, for the tree T, and for the transition kernel Q and initial distribution of the branching Markov chain X.

Proof. **First case: we only have constant functions $f_{\mathcal{S}} \equiv c(\mathcal{S})$ for all $\mathcal{S} \in \mathcal{N}_k$.** Then, for every $n \in \mathbb{N}$ we have:

$$\bar{M}_{A_n}(\mathrm{f}) = \sum_{\mathcal{S}\in\mathcal{N}_k} c(\mathcal{S})\, \frac{|u \in A_n : \mathcal{S}_u = \mathcal{S}|}{|A_n|}, \tag{5.87}$$

where the right hand side converges in distribution (and thus in L^2) as $n \to \infty$ to $\mathbb{E}_{U_k}\big[c(\mathcal{S}_{U_k})\big]$ (remind that the distribution of $\mathcal{S}_{V_n}$ converges to the uniform distribution on G_k when $n \to \infty$). This concludes the proof in this first case.

General case: We adapt the proof of [Wei24b, Theorem 2.2] to the case of neighborhood-shape-dependent functions.

Using the first case and the linearity in f of the empirical averages, and replacing $f_{\mathcal{S}}$ by $f_{\mathcal{S}} - \langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle$, we may assume that $\langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle = 0$ for all $\mathcal{S} \in \mathcal{N}_k$. For all $n \in \mathbb{N}$, we have:

$$\mathbb{E}_Q\Big[\bar{M}_{A_n}(\mathrm{f})^2\Big] = \frac{1}{|A_n|^2} \sum_{u,v\in A_n} \mathbb{E}_Q\Big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)})\Big]. \tag{5.88}$$

Using Lemma 5.A.1, as the transition kernel Q is uniformly geometrically ergodic and has unique invariant probability measure π, and as the function $Q^{\mathcal{S}} f_{\mathcal{S}}$ for $\mathcal{S} \in \mathcal{N}_k$ are all in $L^2(\pi)$ (see the comment just after (5.85)), we have that $C_{\mathcal{S}} := \sup_{n\in\mathbb{N}} \pi Q^n (Q^{\mathcal{S}} f_{\mathcal{S}})^2 < \infty$ for all $\mathcal{S} \in \mathcal{N}_k$. Define $C := \max_{\mathcal{S}\in\mathcal{N}_k} C_{\mathcal{S}} < \infty$ (remind that $\mathcal{N}_k$ is finite). Thus, for $u \in T$, we have:

$$\mathbb{E}_Q\Big[f_{\mathcal{S}_u}(X_{\Delta(u,k)})^2\Big] = \pi Q^{(h(u)-k)_+} (Q^{\mathcal{S}_u} f_{\mathcal{S}_u})^2 \le C < \infty.$$

Hence, for all $u, v \in T$, using Cauchy-Schwarz inequality, we have:

$$\mathbb{E}_Q\Big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)})\Big] \le \Big(\mathbb{E}_Q\Big[f_{\mathcal{S}_u}(X_{\Delta(u,k)})^2\Big]\mathbb{E}_Q\Big[f_{\mathcal{S}_v}(X_{\Delta(v,k)})^2\Big]\Big)^{1/2} \le C < \infty. \tag{5.89}$$

Let $u, v \in T$ such that $d(u,v) > 2k$, which implies that $\Delta(u,k) \cap \Delta(v,k) = \emptyset$. Without loss of generality, assume that $h(u) \ge h(v)$. Then, we have that $h(u \wedge v) < h(u) - k$. Denote by v_0 the last ancestor of u in $\Delta(v,k) \cup \{u \wedge v\}$. Remark that $u \wedge v$ is an ancestor of v_0. Then, we have:

$$\begin{aligned}
&\mathbb{E}_Q\Big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)})\Big] \\
&\qquad = \mathbb{E}_Q\Big[\mathbb{E}_Q\big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \,\big|\, X_{\Delta(v,k)},\, X_{u\wedge v}\big] f_{\mathcal{S}_v}(X_{\Delta(v,k)})\Big] \\
&\qquad \le \Big(\mathbb{E}_Q\Big[\mathbb{E}_Q\big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \,\big|\, X_{\Delta(v,k)},\, X_{u\wedge v}\big]^2\Big]\mathbb{E}_Q\Big[f_{\mathcal{S}_v}(X_{\Delta(v,k)})^2\Big]\Big)^{1/2} \\
&\qquad \le C^{1/2}\Big(\mathbb{E}_Q\Big[\mathbb{E}_Q\big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \,\big|\, X_{v_0}\big]^2\Big]\Big)^{1/2} \\
&\qquad \le C^{1/2}\Big(\mathbb{E}_Q\Big[\mathbb{E}_Q\big[f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \,\big|\, X_{u\wedge v}\big]^2\Big]\Big)^{1/2} \\
&\qquad = C^{1/2}\Big(\pi Q^{h(u\wedge v)}\Big(Q^{d(u\wedge v,u)-k} Q^{\mathcal{S}_u} f_{\mathcal{S}_u}\Big)^2\Big)^{1/2} \\
&\qquad \le C^{1/2}\Big(\max_{\mathcal{S}\in\mathcal{N}_k}\ \pi Q^{h(u\wedge v)}\Big(Q^{\tilde{d}(u,v)-k} Q^{\mathcal{S}} f_{\mathcal{S}}\Big)^2\Big)^{1/2},
\end{aligned} \tag{5.90}$$

where we used Cauchy-Schwarz inequality in the first inequality, we used (5.89) and the Markov property of the process X in the second inequality, and we used Jensen's inequality in the third inequality. Remark that $\tilde{d}(u,v) = \max(d(u\wedge v, u), d(u\wedge v, v))$ is a distance on T that satisfies $d/2 \le \tilde{d} \le d$.

Let U_n and V_n be uniformly distributed over A_n, and independent of each other and of the branching Markov chain X, and denote by $\mathbb{P}_{U_n,V_n}$ their joint probability distribution. Using again Lemma 5.A.1, there exist finite constants $M < \infty$ and $\alpha \in (0,1)$ such that for all $\mathcal{S} \in \mathcal{N}$ we have

$$\forall m \in \mathbb{N}, \qquad \sup_{j\in\mathbb{N}} \pi Q^j \big(Q^m Q^{\mathcal{S}} f_{\mathcal{S}}\big)^2 \le M^2 \alpha^{2m}. \tag{5.91}$$

Hence, combining (5.88), (5.89), (5.90) and (5.91), we get for any $K \ge k$:

$$\begin{aligned}
\mathbb{E}_Q\Big[\bar{M}_{A_n}(\mathrm{f})^2\Big] &\le C\,\mathbb{P}_{U_n,V_n}(\tilde{d}(U_n,V_n) \le 2K) \\
&\quad + C^{1/2}\,|A_n|^{-2} \sum_{u,v\in A_n:\tilde{d}(u,v)>2K} \Big(\max_{\mathcal{S}\in\mathcal{N}_k}\ \pi Q^{h(u\wedge v)}\Big(Q^{\tilde{d}(u,v)-K} Q^{\mathcal{S}} f_{\mathcal{S}}\Big)^2\Big)^{1/2} \\
&\le C\,\mathbb{P}_{U_n,V_n}(d(U_n,V_n) \le 4K) \\
&\quad + C^{1/2}\,|A_n|^{-2} \sum_{u,v\in A_n:\tilde{d}(u,v)>2K} M\alpha^{\tilde{d}(u,v)-K} \qquad (5.92) \\
&\le C\,\mathbb{P}_{U_n,V_n}(d(U_n,V_n) \le 4K) + C^{1/2} M \alpha^K. \qquad (5.93)
\end{aligned}$$

Let $\varepsilon > 0$. Let $K \ge k$ be such that $C^{1/2} M \alpha^K < \varepsilon$. Using [Wei24b, Lemma 3.1], we get that the first term in the right hand side of (5.93) goes to zero as $n \to \infty$. Thus, for n large enough, the right hand side of (5.93) is upper bounded by 2ε. This being true for all $\varepsilon > 0$, we get that $\lim_{n\to\infty} \mathbb{E}_Q\big[\bar{M}_{A_n}(\mathrm{f})^2\big] = 0$. This concludes the proof. □

We now state and prove a strong law of large numbers for branching Markov chains indexed by the infinite complete binary tree T and with neighborhood-shape-dependent functions. This result uses the same assumptions as in Theorem 5.A.2.

Theorem 5.A.4 (Strong law of larger numbers with neighborhood-dependent function)**.** *Let Q be a transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ which is uniformly geometrically ergodic and has a unique invariant probability measure π. Let $X = (X_u, u \in T)$ be a branching Markov chain with transition kernel Q and initial distribution π.*

Let $k \in \mathbb{N}$ be fixed. Let U_k be uniformly distributed over G_k and independent of the process X, and let $\mathbb{E}_{U_k} \otimes \mathbb{E}_Q$ denote the joint expectation over U_k and X. Let $\mathrm{f} = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(X)$.

Then, we have:

$$\text{a.s.} \qquad \lim_{n\to\infty} \bar{M}_{G_n}(\mathrm{f}) = \lim_{n\to\infty} \bar{M}_{T_n \setminus T_{k-1}}(\mathrm{f}) = \mathbb{E}_{U_k} \otimes \mathbb{E}_Q\big[f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})\big].$$

Moreover, there exist finite constants $C_0 < \infty$ and $\beta \in (0,1)$ such that:

$$\forall n \geq k, \qquad \mathbb{E}_Q\left[\left(\bar{M}_{G_n}(\mathrm{f}) - \mathbb{E}_{U_k} \otimes \mathbb{E}_Q\big[f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})\big]\right)^2\right] \leq C_0 \beta^n. \tag{5.94}$$

Proof. After using (5.92), the proof is an easy adaptation of the proof of [Guy07, Theorem 14].

The case of $\bar{M}_{T_n \setminus T_{k-1}}(\mathrm{f})$ follows directly from the case of $\bar{M}_{G_n}(\mathrm{f})$ as:

$$|\bar{M}_{T_n}(\mathrm{f})| \leq \sum_{j=k}^{n} \frac{|G_j|}{|T_n \setminus T_{k-1}|} \bar{M}_{G_j}(\mathrm{f}).$$

Thus, it it enough to treat the case of $\bar{M}_{G_n}(\mathrm{f})$. In the case where all functions $f_{\mathcal{S}}$ for $\mathcal{S} \in \mathcal{N}_k$ are constant, writing $\bar{M}_{G_n}(\mathrm{f})$ as in (5.87), and using the convergence in distribution of $(\mathcal{S}_{U_n})_{n\in\mathbb{N}}$ the uniform distribution over $\mathcal{N}_k$, we get that the sought convergence holds a.s. for $\bar{M}_{G_n}(\mathrm{f})$. Thus, without loss of generality, we assume that $\langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle = 0$ for all $\mathcal{S} \in \mathcal{N}_k$.

Remark that it is enough to prove that $\sum_{n\geq k} \mathbb{E}[\bar{M}_{G_n}(\mathrm{f})^2] < \infty$, as then we can immediately conclude using Borel-Cantelli lemma with Markov's inequality. Thus, for $n \geq k$, using (5.92) with $n' = |T_n|$ (such that $A_{n'} = G_n$) and $K = k$, we get:

$$\begin{aligned}
\mathbb{E}_Q\big[\bar{M}_{G_n}(\mathrm{f})^2\big] &\leq C\, 2^{-(n-2k)} + C^{1/2} M\, 2^{-2n} \sum_{u,v \in G_n : d(u,v) > 2k} \alpha^{\tilde{d}(u,v)-k} \\
&= C\, 2^{-(n-2k)} + C^{1/2} M \sum_{j=k}^{n} 2^{-(n-j)-\mathbb{1}_{\{j>0\}}} \alpha^{j-k} \\
&\leq C\, 2^{-(n-2k)} + C^{1/2} M\, 2^{-(n-k)} \sum_{j=k}^{n} 2^{j-k} \alpha^{j-k} \\
&\leq C\, 2^{-(n-2k)} + C^{1/2} M\, 2^{-(n-k)} C' \max(n+1, (2\alpha)^{n-k}) \\
&\leq C\, 2^{-(n-2k)} + C^{1/2} M C'\, \max((n+1)2^{-(n-k)}, \alpha^{n-k}),
\end{aligned}$$

where C' is a constant whose value only depends on the value of 2α. Hence, there exist finite constants $C_0 < \infty$ and $\beta \in (0,1)$ such that (5.94) holds. In particular, we get that $\sum_{n\geq k} \mathbb{E}[\bar{M}_{G_n}(\mathrm{f})^2] < \infty$. This concludes the proof of the theorem. □

5.B Proof of the "backward" coupling Lemma 5.4.1

We now prove Lemma 5.4.1.

Proof of Lemma 5.4.1. The proof relies on a "backward in time" bound from u to $u \wedge v$, and then a "forward in time" bound from $u \wedge v$ to v. We divide the proof in two cases: first when v is an ancestor of u, and then the general case.

For all $j \leq k$, define the vertex $U_j = \mathrm{p}^j(u)$ which is random for $j > h(u)$ (in which case, it depends on $\mathcal{U}$). Write $x_{U_k} = x$. For all $j \in \{1, \cdots, k\}$, define the random set (which depends on $\mathcal{U}$):

$$\Delta^-(u,k,j) = (\Delta^*(u,k) \setminus \{U_k\}) \cap (T^\infty(U_k) \setminus T^\infty(U_{j-1})),$$

and in particular remark that $U_0 = u \notin \Delta^-(u,k,j)$ and $U_k = \mathrm{p}^k(u) \notin \Delta^-(u,k,j)$.

Case 1: v is an ancestor of u. We mimic the proof of [CMR05, Proposition 12.5.4]. The proof of the first case relies on the observation that conditioned on $X_{\mathrm{p}^k(u)}$ and $Y_{\Delta(u,k)}$, the backward ancestral process X from $U_0 = u$ to $U_k = \mathrm{p}^k(u)$ is a non-homogeneous Markov chain satisfying a uniform mixing condition. The fact that $(X_{U_j})_{0 \leq j \leq k}$ is a Markov chain comes from the Markov property of the HMT (X, Y) (remind the discussion around (5.2) on page 123) which gives for all $j \in \{1, \cdots, k\}$:

$$\begin{aligned}\mathcal{L}(X_{U_j} \,|\, Y_{\Delta(u,k)}, X_{U_k}, X_{T^\infty(U_{j-1})}) &= \mathcal{L}(X_{U_j} \,|\, Y_{\Delta(u,k)}, X_{U_k}, X_{U_{j-1}}) \\ &= \mathcal{L}(X_{U_j} \,|\, Y_{\Delta^-(u,k,j)}, X_{U_k}, X_{U_{j-1}}). \end{aligned} \tag{5.95}$$

For all integers $j \in \{1, \cdots, k\}$, the backward transition kernel (which depends on $\mathcal{U}$) from $X_{U_{j-1}}$ to X_{U_j} is defined as:

$$\mathrm{B}_{x_{U_k},j}[y_{\Delta(u,k)}](x_{U_{j-1}}; f) = \mathbb{E}_\theta\big[f(X_{U_j}) \;\big|\; Y_{\Delta(u,k)} = y_{\Delta(u,k)}, X_{U_k} = x_{U_k}, X_{U_{j-1}} = x_{U_{j-1}}\big],$$

for any $x_{U_{j-1}} \in \mathcal{X}$ and any bounded Borel function f on $\mathcal{X}$. By the Markov property (see (5.95)), note that $\mathrm{B}_{x_{U_k},j}[y_{\Delta(u,k)}](x_{U_{j-1}}, f)$ only depends on $y_{\Delta^-(u,k,j)}$ instead of $y_{\Delta(u,k)}$, that is:

$$\begin{aligned}&\mathrm{B}_{x_{U_k},j}[y_{\Delta(u,k)}](x_{U_{j-1}}; f) \\ &\qquad = \mathbb{E}_\theta\big[f(X_{U_j}) \;\big|\; Y_{\Delta^-(u,k,j)} = y_{\Delta^-(u,k,j)}, X_{U_k} = x_{U_k}, X_{U_{j-1}} = x_{U_{j-1}}\big] \\ &\qquad = \frac{\int_{\mathcal{X}} f(x_{U_j})\, p_\theta(y_{\Delta^-(u,k,j)}, x_{U_j} \,|\, X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}})\, \lambda(\mathrm{d}x_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\Delta^-(u,k,j)}, x_{U_j} \,|\, X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}})\, \lambda(\mathrm{d}x_{U_j})}, \end{aligned} \tag{5.96}$$

where:

$$\begin{aligned}&p_\theta(y_{\Delta^-(u,k,j)}, x_{U_j} \,|\, X_{U_k} = x_{U_k}) = \\ &\qquad \int_{\mathcal{X}^{\Delta^-(u,k,j)}} \prod_{w \in \Delta^-(u,k,j)} q_\theta(x_{\mathrm{p}(w)}, x_w) g_\theta(x_w, y_w) \prod_{w \in \Delta^-(u,k,j) \setminus \{U_j\}} \lambda(\mathrm{d}x_w).\end{aligned}$$

To simplify notations, we will keep the dependence on $y_{\Delta(u,k)}$ for all indices j. Note that the integral in the denominator in the right hand side of (5.96) is lower bounded by:

$$\prod_{w \in \Delta^-(u,k,j)} \sigma^- \int_{\mathcal{X}} g_\theta(x_w, y_w) \lambda(\mathrm{d}x_w),$$

and is thus positive $\mathbb{P}_\theta$-a.s. under Assumption 7.

Using Assumption 7, we get that those backward transition kernels satisfy the following Doeblin condition (remind Definition 5.2.6):

$$\frac{\sigma^-}{\sigma^+} \nu_{x_{U_k},j}[y_{\Delta(u,k)}](f) \leq \mathrm{B}_{x_{U_k},j}[y_{\Delta(u,k)}](x_{U_{j-1}}; f),$$

where for any bounded Borel function f on $\mathcal{X}$, we have:

$$\begin{aligned}\nu_{x_{U_k},j}[y_{\Delta(u,k)}](f) &= \mathbb{E}_\theta\big[f(X_{U_j}) \;\big|\; Y_{\Delta^-(u,k,j)} = y_{\Delta^-(u,k,j)}, X_{U_k} = x_{U_k}\big] \\ &= \frac{\int_{\mathcal{X}} f(x_{U_j})\, p_\theta(y_{\Delta^-(u,k,j)}, x_{U_j} \,|\, X_{U_k} = x_{U_k})\, \lambda(\mathrm{d}x_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\Delta^-(u,k,j)}, x_{U_j} \,|\, X_{U_k} = x_{U_k})\, \lambda(\mathrm{d}x_{U_j})}, \end{aligned}$$

where note that the only difference with the definition of $\mathrm{B}_{x_{U_k},j}[y_{\Delta(u,k)}](x_{U_{j-1}}; f)$ is that the term $q_\theta(x_{U_j}, x_{U_{j-1}})$ has disappeared in both the numerator and the denominator of $\nu_{x_{U_k},j}[y_{\Delta(u,k)}](f)$. Thus, Lemma 5.2.7 shows that the Dobrushin coefficient $\delta(B_{x_{U_k},j})$ (defined in (5.4)) of the backward transition kernel $B_{x_{U_k},j}$ is upper bounded by $\rho = 1 - \sigma^-/\sigma^+$.

Note that the Markov property in (5.95) (with $j = 1$) also gives us:

$$\begin{aligned}\mathcal{L}(X_{U_1} \,|\, Y_{\mathbb{\Delta}(u,k)}, X_{U_k}, X_{U_0}) &= \mathcal{L}(X_{U_1} \,|\, Y_{\mathbb{\Delta}^-(u,k,j)}, X_{U_k}, X_{U_0}) \\ &= \mathcal{L}(X_{U_1} \,|\, Y_{\mathbb{\Delta}^*(u,k)}, X_{U_k}, X_{U_0}). \end{aligned} \tag{5.97}$$

Finally, if we write:

$$\begin{aligned}&\mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}(u,k)} = y_{\mathbb{\Delta}(u,k)}, X_{U_k} = x_{U_k}) \\ &\qquad = \int \mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}^-(u,k,1)} = y_{\mathbb{\Delta}^-(u,k,1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \\ &\qquad\qquad \times \mathbb{P}_\theta(X_{U_1} \in \mathrm{d}x_{U_1} \,|\, Y_{\mathbb{\Delta}(u,k)} = y_{\mathbb{\Delta}(u,k)}, X_{U_k} = x_{U_k}),\end{aligned} \tag{5.98}$$

and we also write (using (5.97)):

$$\begin{aligned}&\mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}^*(u,k)} = y_{\mathbb{\Delta}^*(u,k)}, X_{U_k} = x_{U_k}) \\ &\qquad = \int \mathbb{P}_\theta(X_v \in \cdot \,|\, Y_{\mathbb{\Delta}^-(u,k,1)} = y_{\mathbb{\Delta}^-(u,k,1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \\ &\qquad\qquad \times \mathbb{P}_\theta(X_{U_1} \in \mathrm{d}x_{U_1} \,|\, Y_{\mathbb{\Delta}^*(u,k)} = y_{\mathbb{\Delta}^*(u,k)}, X_{U_k} = x_{U_k}),\end{aligned} \tag{5.99}$$

then the two distributions (for X_v) on the left hand sides of those displayed equations can be considered as obtained through running $d(u,v) - 1$ iterations of the backward ancestral conditional Markov chain described above, using two different initial conditions. Therefore, as the Dobrushin coefficient is sub-multiplicative (remind Lemma 5.2.5), we get that those two probability distribution differ by at most $2\,\rho^{d(u,v)-1}$ in total variation. This concludes the proof of the first case.

Case 2: general case. The proof of the second case relies on the observation that conditioned on $X_{\mathrm{p}^k(u)}$ and $Y_{\mathbb{\Delta}(u,k)}$, if we consider the process X backward from u to $u \wedge v$ (remind that $v \in \mathbb{\Delta}^*(u,k)$) and then forward from $u \wedge v$ to v, we get a non-homogeneous Markov chain satisfying uniform mixing rate ρ. Note that as $v \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$, we have that $u \wedge v \in \{U_1, \cdots, U_k\}$. Using the first case, it only remains to check those observations for the forward segment, which were already proved in the proof of Lemma 5.3.2.

Hence, if we use the same decomposition as in (5.98) and (5.98), which corresponds to run $d(u,v) - 1$ iterations of the backward-forward conditional chain described above ($d(u, u \wedge v) - 1$ backward iterations and $d(u \wedge v, v)$ forward iterations), we get as in the first case that those two probability distribution differ by at most $2\,\rho^{d(u,v)-1}$ in total variation. This concludes the proof of the lemma. □

5.C Proof of (5.35) (used in the proof of Proposition 5.3.10)

Let $m \in \mathbb{N}^*$ be fixed through this section.

For ease of read, we restate some notation definitions used only in the proof of Proposition 5.3.10. For $u, v \in T^\infty$ with $h(u) \equiv h(v) \mod m+1$, we write $T(u,m) < T(v,m)$ if $u < v$. Moreover for $u, v \in T^\infty$, we write $u < T(v,m)$ if $h(u) < h(v)$ or $h(u) \leq h(v) + m$ and for all $w \in T(v,m)$ with $h(w) = h(u)$ (note that such w must exist), we have $u < w$. Informally u is "above or on the left of $T(v,m)$". For all $u \in T$, $k \in \mathbb{N}$, define the random subtrees which depend on $\mathcal{U}$:

$$\mathbb{\Delta}^*(T(u,m),k) = \bigcup \big\{ T(v,m) : v \in \mathbb{\Delta}^*(u, k(m+1)) \text{ such that } h(v) \equiv h(u) \mod m+1 \big\},$$

and $\mathbb{\Delta}(T(u,m),k) = \mathbb{\Delta}^*(T(u,m),k) \cup T(u,m)$. When $h(u) \geq k(m+1)$, then those subtrees do not depend on $\mathcal{U}$, and we write $\Delta^*(T(u,m),k) = \mathbb{\Delta}^*(T(u,m),k)$ and $\Delta(T(u,m),k) = \mathbb{\Delta}(T(u,m),k)$ to indicate it. See Figure 5.5 on page 136 for an illustration of the "past" subtree $\mathbb{\Delta}^*(T(u,m),k)$ of the block subtree $T(u,m)$.

The goal of this section is to prove (5.35) for all $\theta \in \Theta$ and $x \in \mathcal{X}$, which we restate here for ease of read:

$$\lim_{k \to \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\log p_\theta(Y_{T_m} \,|\, Y_{\mathbb{\Delta}^*(T_m,k)}, X_{\mathrm{p}^{k(m+1)}(\partial)} = x) \right] = |T_m|\, \ell(\theta).$$

5.C.1 Decomposition of the log-likelihood into subtree increments

Following (5.16), for all $u \in T$, $k \in \mathbb{N}$, $x \in \mathcal{X}$ and $\theta \in \Theta$, using the conditional probabilities formula, define:

$$\begin{aligned} \mathrm{H}_{T(u,m),k,x}(\theta) &= \frac{p_\theta(Y_{\mathbb{\Delta}(T(u,m),k)} \,|\, X_{\mathrm{p}^k(u)} = x)}{p_\theta(Y_{\mathbb{\Delta}^*(T(u,m),k)} \,|\, X_{\mathrm{p}^k(u)} = x)} \\ &= \int p_\theta(Y_{T(u,m)} \,|\, X_u = x_u)\, \mathbb{P}_\theta(X_u \in \mathrm{d}x_u \,|\, Y_{\mathbb{\Delta}^*(T(u,m),k)}, X_{\mathrm{p}^{(k-1)(m+1)}(u)} = x), \end{aligned} \tag{5.100}$$

where:

$$p_\theta(Y_{T(u,m)} \,|\, X_u = x_u) = \int_{\mathcal{X}^{|T(u,m)|}} g_\theta(x_u, Y_u) \prod_{w \in T(u,m)\setminus\{u\}} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w)\, \lambda(\mathrm{d}x_w).$$

We then define the log-likelihood contribution of the subtree $T(u,m)$ with past over $k \in \mathbb{N}$ subtree generations (that is, $k(m+1)$ (node) generations) as:

$$\mathrm{h}_{T(u,m),k,x}(\theta) = \log \mathrm{H}_{T(u,m),k,x}(\theta)$$

For all $n \in \mathbb{N}^*$, we decompose the tree $T_{n(m+1)-1}$ into subtrees of height m (such as T_m), and we order those subtrees according to $<$. Hence, using (5.6), (5.7) and (5.100) and a telescopic sum argument, the log-likelihood of the observed variables $Y_{T_{n(m+1)-1}}$ can be rewritten as:

$$\ell_{n(m+1)-1,x}(\theta) = \sum_{k=0}^{n-1} \sum_{u \in G_{k(m+1)}} \mathrm{h}_{T(u,m),k,x}(\theta). \tag{5.101}$$

5.C.2 Construction of the log-likelihood increments with infinite past for subtree blocks

In this subsection, we construct the log-likelihood increments with infinite past for subtrees.

To construct the limit of the functions $\mathrm{h}_{T(u,m),k,x}(\theta)$ we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when $k \to \infty$.

Lemma 5.C.1 (Uniform bounds for $\mathrm{h}_{T(u,m),k,x}(\theta)$)**.** *Assume that Assumptions 6–7 and 8-(ii) hold. For all vertices $u \in T$ and all integers $k, k' \in \mathbb{N}^*$, the following assertions hold true:*

$$\sup_{\theta\in\Theta} \sup_{x,x'\in\mathcal{X}} |\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),k',x'}(\theta)| \le \frac{\rho^{(k\wedge k')(m+1)-1}}{(1-\rho)^{|T_m|}}, \tag{5.102}$$

$$\sup_{\theta\in\Theta} \sup_{k\in\mathbb{N}^*} \sup_{x\in\mathcal{X}} |\mathrm{h}_{T(u,m),k,x}(\theta)| \le \big(|T_m| \log b^+\big) \vee \left| \sum_{w\in T(u,m)} \log(\sigma^- b^-(Y_w)) \right|. \tag{5.103}$$

Proof. [The proof is a straightforward adaptation of the proof of [CMR05, Lemma 12.3.2] with the use of Lemma 5.3.2 for the coupling.] Let $k' \ge k \ge 1$, and write $v = \mathrm{p}^{k(m+1)}(u)$, $v' = \mathrm{p}^{k'(m+1)}(u)$. Then, write:

$$\begin{aligned} \mathrm{H}_{T(u,m),k,x}(\theta) = \int_{\mathcal{X}^2} \left[\int_{\mathcal{X}^{T(u,m)}} \prod_{w\in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w) \lambda(\mathrm{d}x_w) \right] \\ \times \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, Y_{\mathbb{\Delta}^*(T(u,m),k)}, X_v = x_v) \times \delta_x(\mathrm{d}x_v), \end{aligned} \tag{5.104}$$

and using the Markov property at X_v, write:

$$\begin{aligned} \mathrm{H}_{T(u,m),k',x'}(\theta) = \int_{\mathcal{X}^2} \left[\int_{\mathcal{X}^{T(u,m)}} \prod_{w\in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w) \lambda(\mathrm{d}x_w) \right] \\ \times \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, Y_{\mathbb{\Delta}^*(T(u,m),k)}, X_v = x_v) \\ \times \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbb{\Delta}^*(T(u,m),k')\setminus\mathbb{\Delta}(T(u,m),k)}, X_{v'} = x'). \end{aligned} \tag{5.105}$$

Applying Lemma 5.3.2, we get (note that the integrands in (5.104) and (5.105) are non-negative):

$$\begin{aligned}|\mathrm{H}_{T(u,m),k,x}(\theta) - \mathrm{H}_{T(u,m),k',x'}(\theta)| \\ &\leq \rho^{k(m+1)-1} \sup_{x_{\mathrm{p}(u)} \in \mathcal{X}} \int \prod_{w \in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w) \lambda(\mathrm{d}x_w) \\ &\leq \rho^{k(m+1)-1} (\sigma^+)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w). \end{aligned} \tag{5.106}$$

The integral in (5.104) can be lower bounded giving us:

$$\mathrm{H}_{u,k,x}(\theta) \geq (\sigma^-)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w), \tag{5.107}$$

where the right hand side is positive by Assumption 7-(ii); and similarly for (5.105). Combining (5.106) with (5.107), and with the inequality $|\log x - \log y| \leq |x-y|/(x \wedge y)$, we get the first assertion of the lemma:

$$|\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),k',x'}(\theta)| \leq \left(\frac{\sigma^+}{\sigma^-}\right)^{|T_m|} \rho^{k(m+1)-1} = \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Combining (5.100) and (5.107), we get:

$$\prod_{w \in T(u,m)} \sigma^- b^-(Y_w) \leq \mathrm{H}_{T(u,m),k,x}(\theta) \leq (b^+)^{|T_m|},$$

which yields the second assertion of the lemma (remind that $b^-(Y_w) > 0$ for all $w \in T^\infty$ by Assumption 7-(ii)). □

We are now ready to construct the limit of the functions $\mathrm{h}_{T(u,m),k,x}(\theta)$ and state some properties of this limit. Note that this result is stated for every $u \in T$, but we will only need it for $u = \partial$. Remind that we are in the stationary case, and that the HMT process (X, Y) is defined on T^∞.

Proposition 5.C.2 (Properties of the limit function $\mathrm{h}_{T(u,m),\infty}(\theta)$)**.** *Assume that Assumptions 5–8 hold. For every $u \in T$ and $\theta \in \Theta$, there exists $\mathrm{h}_{T(u,m),\infty}(\theta) \in L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ such that for all $x \in \mathcal{X}$, the sequence $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k\in\mathbb{N}}$ converges $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ to $\mathrm{h}_{T(u,m),\infty}(\theta)$.*

Furthermore, this convergence is uniform over $\theta \in \Theta$ and $x \in \mathcal{X}$, that is, we have $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ that:

$$\lim_{k\to\infty} \sup_{\theta\in\Theta} \sup_{x\in\mathcal{X}} |\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),\infty}(\theta)| = 0.$$

The limit function $\mathrm{h}_{T(u,m),\infty}(\theta)$ can be interpreted as $\log p_\theta(Y_{T(u,m)} \,|\, Y_{\Delta^*(T(u,m),\infty)})$, where the random subset of vertices $\Delta^*(T(u,m),\infty)$ is defined as $\Delta^*(T(u,m),\infty) = \{v \in T^\infty : v <_\mathcal{U} T(u,m)\}$. Note that $\mathrm{h}_{T(u,m),\infty}(\theta)$ is a function of the random set of variables $(Y_v, v \in \Delta(T(u,m),\infty))$, where we define $\Delta(T(u,m),\infty) = \Delta^*(T(u,m),\infty) \cup T(u,m)$, and thus implicitly depend on $\mathcal{U}$ trough $\Delta(T(u,m),\infty)$.

Proof. Fix some $u \in T$. Note that (5.102) shows that the sequence $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k\in\mathbb{N}}$ is Cauchy uniformly in θ and x, and thus has $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-almost surely a limit when $k \to \infty$ which does not depend on x; we denote this limit by $\mathrm{h}_{T(u,m),\infty}(\theta)$. Furthermore, we get from (5.103) that $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k\in\mathbb{N}}$ is uniformly bounded in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$, and thus $\mathrm{h}_{T(u,m),\infty}(\theta)$ is in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ and the convergence also holds in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$. Finally, as the bound in (5.102) is uniform in θ and x, we get that the convergence holds uniform over θ and x both $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-almost surely and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$. □

5.C.3 Properties of the contrast function

As the functions $\mathrm{h}_{T(u,m),\infty}(\theta)$ are in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ under the assumptions used in Proposition 5.C.2, we can now define the *contrast function* $\ell^{(m)}$ (which is deterministic) for block subtree of height m as:

$$\ell^{(m)}(\theta) = \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,\infty}(\theta)\big],$$

where remind $\mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^\star}$ is the expectation corresponding to $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$. We prove under the L^2 regularity assumption Assumption 9 the convergence of the normalized log-likelihood to this contrast function.

Proposition 5.C.3 (Ergodic convergence for the log-likelihood)**.** *Assume that Assumptions 5–9 hold. Then, for all $x \in \mathcal{X}$, the normalized log-likelihood $|T_{n(m+1)-1}|^{-1}\ell_{n(m+1)-1,x}(\theta)$ converges $\mathbb{P}_{\theta^\star}$-a.s. to the contrast function $\ell^{(m)}(\theta)$ as $n \to \infty$.*

$$\lim_{n\to\infty} \frac{|T_m|}{|T_{n(m+1)-1}|}\, \ell_{n(m+1)-1,x}(\theta) = \ell^{(m)}(\theta) \qquad \mathbb{P}_{\theta^\star}\text{-a.s.} \tag{5.108}$$

In particular, we get that $\ell^{(m)}(\theta) = |T_m|\ell(\theta)$.

Proof. Let $\theta \in \Theta$ be some parameter. Fix some $k \in \mathbb{N}^*$ and $x \in \mathcal{X}$. Remind (5.101). Applying (5.102) for each vertex $u \in G_{j(m+1)}$ with $j \in \{k, \cdots, n-1\}$, we get:

$$\frac{|T_m|}{|T_{n(m+1)-1}|}\left|\ell_{n(m+1)-1,x}(\theta) - \sum_{j=k}^{n-1}\sum_{u\in G_{j(m+1)}} \mathrm{h}_{T(u,m),k,x}(\theta)\right| \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}} + \frac{|T_m|}{|T_{n(m+1)-1}|}\sum_{j=0}^{k-1}\sum_{u\in G_{j(m+1)}} |\mathrm{h}_{T(u,m),j,x}(\theta)|. \tag{5.109}$$

Note that by (5.103), we have that $|\mathrm{h}_{T(u,m),j,x}(\theta)| < \infty$ $\mathbb{P}_{\theta^\star}$-a.s. for all $j \in \mathbb{N}^*$ and $u \in G_{j(m+1)}$. For $u = \partial$, we have $\mathrm{h}_{T_m,0,x}(\theta) = \log p_\theta(Y_{T_m} \,|\, X_\partial = x)$ which is finite $\mathbb{P}_{\theta^\star}$-a.s. by Assumption 7-(iii).

The definition of the *shape* from Section 5.2.4 can straightforwardly be adapted to the (deterministic) subtrees $\Delta(T(u,m),k)$ for vertices $u \in G_{j(m+1)}$ with $j \geq k$, where u is seen as a distinguished vertex of $\Delta(T(u,m),k)$. Following (5.8) (on page 126) in the initial vertex-by-vertex decomposition setting, for a vertex $u \in G_{j(m+1)}$ with $j \geq k$, let $v_u \in G_{k(m+1)}$ be the unique vertex $G_{k(m+1)}$ such that $\Delta(T(u,m),k)$ and $\Delta(T(v_u,m),k)$ have the same shape. Then, we have:

$$\mathrm{h}_{T(u,m),k,x}\big(\theta; Y_{\Delta(T(u,m),k)} = y_{\Delta(T(u,m),k)}\big) = \mathrm{h}_{T(v_u,m),k,x}\big(\theta; Y_{\Delta(T(v_u,m),k)} = y_{\Delta(T(u,m),k))}\big). \tag{5.110}$$

Moreover, using (5.103) together with Assumption 9, we get for every $u \in G_{j(m+1)}$ with $j \geq k$ that the random variable $\mathrm{h}_{T(u,m),k,x}(\theta; Y_{\Delta(T(u,m),k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$. Hence, applying a straightforward modification of Lemma 5.2.11 for subtree blocks $T(u,m)$ to the collection of neighborhood-shape-dependent functions $(\mathrm{h}_{T(v,m),k,x}(\theta; Y_{\Delta(T(v,m))} = \cdot))_{v\in G_{k(m+1)}}$ (remind that indexing functions with $G_{k(m+1)}$ or with the set of possible shapes is equivalent by (5.9)), and using (5.110) and (5.14) (in Remark 5.3.1), we get:

$$\frac{|T_m|}{|T_{n(m+1)-1}|}\sum_{j=k}^{n-1}\sum_{u\in G_{j(m+1)}} \mathrm{h}_{T(u,m),k,x}(\theta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,k,x}(\theta)\big] \qquad \mathbb{P}_{\theta^\star}\text{-a.s.} \tag{5.111}$$

Using (5.102) with Proposition 5.C.2, we get:

$$\big|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,k,x}(\theta)\big] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,\infty}(\theta)\big]\big| \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Thus, combining this bound with (5.109) and (5.111), we get $\mathbb{P}_{\theta^\star}$-a.s. that:

$$\limsup_{n\to\infty}\left|\frac{|T_m|}{|T_{n(m+1)-1}|}\ell_{n(m+1)-1,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,\infty}(\theta)\big]\right| \leq 2\,\frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

As the left hand side does not depend on k, letting $k \to \infty$, we get that (5.108) in the lemma holds. Lastly, as the limit must be the same as in Proposition 5.3.5, we get that $\ell^{(m)}(\theta) = |T_m|\ell(\theta)$. This concludes the proof. □

We are now ready to close this section by proving that (5.35) holds.

Proposition 5.C.4. *Assume that Assumptions 5–9 hold. Then,* (5.35) *holds for all $\theta \in \Theta$ and $x \in \mathcal{X}$.*

Proof. Applying Proposition 5.C.2, we get that the left hand side of (5.35) is equal to $\ell^{(m)}(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[h_{T(\partial,m),\infty}(\theta)]$, which is equal to $|T_m|\ell(\theta)$ by Proposition 5.C.3. □

5.D Details of the proof of Proposition 5.4.5

Remind that the proof of Proposition 5.4.4 can be straightforwardly adapted to Proposition 5.4.5 except for Lemma 5.4.8. In Section 5.4.2, for brevity, we have only presented the adaptation of Lemma 5.4.8 to the terms $\Gamma_{u,k,x}(\theta)$. In this appendix, we present all the details of the adaptation of the rest of the proof of Proposition 5.4.4 to the terms $\Gamma_{u,k,x}(\theta)$.

The following lemma gives an exponential bound on the $L^2(\mathbb{P}_{\theta^\star})$ norm uniformly in $x \in \mathcal{X}$ for the the average of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$.

Lemma 5.D.1. *Under the assumptions of Proposition 5.4.5, for all $x \in \mathcal{X}$ and $\theta \in \Theta_0$, there exist finite constants $C < \infty$ and $\alpha \in (0,1)$ such that for all $n \in \mathbb{N}^*$ we have:*

$$\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,\infty}(\theta)\right]\right|^2\right]^{1/2} \leq C\alpha^n. \tag{5.112}$$

Proof. Let $x' \in \mathcal{X}$ and $\theta \in \Theta_0$. Using Minkowski's inequality and Jensen's inequality, for all $n, k \in \mathbb{N}^*$, we get:

$$\begin{aligned}
&\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,\infty}(\theta)\right]\right|^2\right]^{1/2} \\
&\qquad\leq \mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_{k-1}^*}\Gamma_{u,h(u),x}(\theta)\right|^2\right]^{1/2} \\
&\qquad\quad+ \mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Gamma_{u,h(u),x}(\theta) - \Gamma_{u,k,x'}(\theta)\right|^2\right]^{1/2} \\
&\qquad\quad+ \mathbb{E}_{\theta^\star}\left[\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Gamma_{u,k,x'}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,k,x'}(\theta)\right]\right|^2\right]^{1/2} \\
&\qquad\quad+ \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[|\Gamma_{\partial,k,x'}(\theta) - \Gamma_{\partial,\infty}(\theta)|^2\right]^{1/2}.
\end{aligned} \tag{5.113}$$

Using Lemma 5.4.17 together with (5.49) on page 142 (which, remind, are both immediate consequences of Lemma 5.4.2), there exists a finite constant $C < \infty$ and $\beta \in (0,1)$ such that the first term in the right hand side of (5.113) is upper bounded by $C2^{-(n-k)}$ (note that $\frac{|T_{k-1}|}{|T_n|} \leq 2^{-(n-k)}$), and the second and fourth terms in the right hand side of (5.113) are both upper bounded by $C\beta^{k/2}$.

We now give an upper bound for the second term in the right hand side of (5.113). For a vertex u in $T \setminus T_{k-1}$, let $v_u \in G_k$ be the unique vertex that satisfies the shape equality constraint (5.8) (on page 126), then we have:

$$\Gamma_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \Gamma_{v_u,k,x'}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{5.114}$$

Moreover, using the definition of $\Gamma_{u,k,x}(\theta)$ in (5.63) together with the assumption on ϕ_θ in Proposition 5.4.5, we get that the random variable $\Gamma_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T\setminus T_{k-1}$. Thus, we can apply Lemma 5.2.11 (see in particular (5.11)) to the collection of neighborhood-shape-dependent functions $(\Gamma_{v_u,k,x'}(\theta; Y_{\Delta(v)} = \cdot))_{v\in G_k}$ (remind that indexing functions with G_k or with $\mathcal{N}_k$ is equivalent by (5.9)). Using (5.11) in Lemma 5.2.11 together with (5.28) and (5.14) in Remark 5.3.1, we get that there exist $\gamma \in (0,1)$ and a finite constant $C' < \infty$ (note that they both do not depend on k and n) such that for all $n, k \in \mathbb{N}^*$ with $n \geq k$, the second term in the right hand side of (5.113) is upper bounded by $C'\gamma^{n-k}$.

Hence, taking $k = \lceil n/2 \rceil$, we get that the left hand side of (5.113) is upper bounded by $2C\beta^{n/4} + C'\alpha^{n/2} + C2^{-n/2+1}$, and thus decays at exponential rate as desired. This concludes the proof. □

Lemma 5.D.1 implies as a corollary the convergence $\mathbb{P}_{\theta^\star}$-a.s. and in $L^2(\mathbb{P}_{\theta^\star})$ uniformly in $x \in \mathcal{X}$ for the the sum of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$.

Corollary 5.D.2. *Under the assumptions of Proposition 5.4.5, for all $x \in \mathcal{X}$ and $\theta \in \Theta_0$, we have:*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u\in T_n^*} \Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\Gamma_{\partial,\infty}(\theta)\big] \right| = 0 \quad \mathbb{P}_{\theta^\star}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^\star}).$$

Proof. The convergence in $L^2(\mathbb{P}_{\theta^\star})$ follows immediately from Lemma 5.D.1. Moreover, using again Lemma 5.D.1, we have:

$$\sum_{n\in\mathbb{N}^*} \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u\in T_n^*} \Gamma_{u,\infty}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)] \right|^2 \right] < \infty.$$

Hence, Borel-Cantelli lemma and Markov's inequality imply that the convergence in the lemma also holds $\mathbb{P}_{\theta^\star}$-a.s. □

The following lemma gives some continuity properties of the function $\theta \mapsto \Gamma_{\partial,k,x}(\theta)$.

Lemma 5.D.3. *Under the assumptions of Proposition 5.4.5, for all $x \in \mathcal{X}$ and $k \in \mathbb{N}$, the random function $\theta \mapsto \Gamma_{\partial,k,x}(\theta)$ is $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. continuous on Θ_0. Moreover, for all $\theta \in \Theta_0$, we have:*

$$\lim_{\delta\to 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta} |\Gamma_{\partial,k,x}(\theta') - \Gamma_{\partial,k,x}(\theta)|^2 \right] = 0.$$

Proof. We mimic the proof of [DMR04, Lemma 14].

For all $v \in T^\infty$, define the random variable $\|\phi^v\|_\infty = \sup_{\theta'\in\Theta_0} \sup_{x,x'\in\mathcal{X}} |\phi_{\theta'}(x', x, Y_v)|$. Remind that under the assumptions of Proposition 5.4.5, the HMT process (X, Y) is stationary and the random variable $\|\phi^\partial\|_\infty$ is in $L^4(\mathbb{P}_{\theta^\star})$. Thus, for all $v \in T^\infty$, the random variable $\|\phi^v\|_\infty$ is in $L^4(\mathbb{P}_{\theta^\star})$. Remind from (5.12) on page 129 that $\Delta(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$. Then, note that we have:

$$\sup_{\theta\in\Theta_0} |\Gamma_{\partial,k,x}(\theta)| \le 4 \left(\sum_{v\in T^\infty(\mathrm{p}^k(\partial),k)} \|\phi^v\|_\infty \right)^2,$$

where the upper bound is a random variable in $L^2(\mathbb{P}_{\theta^\star})$ (and thus in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$) which depends on $Y_{T^\infty(\mathrm{p}^k(u),k)}$ but not on $\mathcal{U}$. Hence, to prove the lemma, it suffices to prove that for all $v_1, v_2 \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(\partial)\}$ and $\epsilon \in \{0, 1\}$, we have $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. :

$$\lim_{\delta\to 0} \sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta} \Big| \mathbb{E}_{\theta'}[\phi_{\theta'}^{(2,\epsilon)}(X_{\mathrm{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathrm{p}(v_2)}, X_{v_2}, Y_{v_2}) \,|\, Y_{\Delta(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] - \mathbb{E}_{\theta}[\phi_{\theta}^{(2,\epsilon)}(X_{\mathrm{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathrm{p}(v_2)}, X_{v_2}, Y_{v_2}) \,|\, Y_{\Delta(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] \Big| = 0,$$

where:

$$\phi_{\theta'}^{(2,\epsilon)}(X_{\mathrm{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathrm{p}(v_2)}, X_{v_2}, Y_{v_2}) := \phi_{\theta'}(X_{\mathrm{p}(v_1)}, X_{v_1}, Y_{v_1}) \phi_{\theta'}(X_{\mathrm{p}(v_2)}, X_{v_2}, Y_{v_2})^\epsilon.$$

Denote $x_{\mathrm{p}^k(\partial)} = x$, and write:

$$\mathbb{E}_{\theta}[\phi_{\theta'}^{(2,\epsilon)}(X_{\mathrm{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathrm{p}(v_2)}, X_{v_2}, Y_{v_2}) \,|\, Y_{\Delta(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] = \frac{\int_{\mathcal{X}^{|\Delta(\partial,k)|-1}} \phi_{\theta'}^{(2,\epsilon)}(x_{\mathrm{p}(v_1)}, x_{v_1}, Y_{v_1}, x_{\mathrm{p}(v_2)}, x_{v_2}, Y_{v_2}) \, \Psi(\mathrm{d}x_{\Delta(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}})}{\int_{\mathcal{X}^{\Delta(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}} 1 \, \Psi(\mathrm{d}x_{\Delta(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}})}. \quad (5.115)$$

where:

$$\Psi(\mathrm{d}x_{\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}) := \prod_{w\in\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} q_\theta(x_{\mathrm{p}(w)},x_w)g_\theta(x_w,Y_w)\lambda(\mathrm{d}x_w).$$

Using Assumptions 6-8 (which are part of the assumptions in Proposition 5.4.5), we know that the integrand in the numerator of the right hand side of (5.115) is continuous w.r.t. θ and is upper bounded by the random variable $\|\phi^{v_1}\|_\infty(\|\phi^{v_2}\|_\infty)^\epsilon(\sigma^+b^+)^{|T^\infty(\mathrm{p}^k(u),k)|-1}$ (remind that $\sigma^+\geq 1$ and $b^+\geq 1$). And similarly, the denominator is continuous w.r.t. θ, and, using Assumption 7-(ii), is lower bounded by the random variable:

$$\prod_{w\in\mathbb{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} \sigma^- \inf_{\theta'\in\Theta}\int g_{\theta'}(x_w,Y_w)\lambda(\mathrm{d}x_w) > 0.$$

Hence, using dominated convergence, we conclude that $\mathbb{P}_{\mathcal{U}}\otimes\mathbb{P}_{\theta^\star}$-a.s. the left hand side of (5.115) is continuous w.r.t. θ. This concludes the proof. □

As a corollary of Lemma 5.D.3, we get that the function $\theta\mapsto\Gamma_{\partial,\infty}(\theta)$ is continuous in $L^2(\mathbb{P}_{\theta^\star})$.

Corollary 5.D.4. *Under the assumptions of Proposition 5.4.5, for all $\theta\in\Theta_0$, we have:*

$$\lim_{\delta\to 0}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta}|\Gamma_{\partial,\infty}(\theta')-\Gamma_{\partial,\infty}(\theta)|^2\right]=0.$$

In particular, the function $\theta\mapsto\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$ is continuous on Θ_0.

Proof. Using Minkowski's inequality and Lemma 5.4.17, there exist a finite constant $C<\infty$ and $\beta\in(0,1)$ such that for all $x\in\mathcal{X}$ and $k\in\mathbb{N}^*$, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}&\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta}|\Gamma_{\partial,\infty}(\theta')-\Gamma_{\partial,\infty}(\theta)|^2\right]^{1/2}\\&\leq 2C\beta^{k/2}+\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta}|\Gamma_{\partial,k,x}(\theta')-\Gamma_{\partial,k,x}(\theta)|^2\right]^{1/2}. \quad (5.116)\end{aligned}$$

Using Lemma 5.D.3, we get:

$$\limsup_{\delta\to 0}\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\leq\delta}|\Gamma_{\partial,\infty}(\theta')-\Gamma_{\partial,\infty}(\theta)|^2\right]^{1/2}\leq 2C\beta^{k/2},$$

and taking $k\to\infty$, the upper bound vanishes. This concludes the proof. □

We now prove a locally uniform law of large numbers for the quantities $\Gamma_{u,k,x}(\theta)$.

Lemma 5.D.5. *Under the assumptions of Proposition 5.4.5, for all $x\in\mathcal{X}$, we have:*

$$\lim_{\delta\to 0}\lim_{n\to\infty}\sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta')-\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]\right|=0,\quad \mathbb{P}_{\theta^\star}\text{-a.s.}$$

Proof. First, write:

$$\begin{aligned}\sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta}&\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta')-\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]\right|\\&\leq\frac{1}{|T_n|}\sum_{u\in T_n^*}\sup_{\theta'\in\Theta_0\,:\,\|\theta'-\theta\|\leq\delta}\left|\Gamma_{u,h(u),x}(\theta')-\Gamma_{u,h(u),x}(\theta)\right|\\&\quad+\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta)-\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]\right|. \quad (5.117)\end{aligned}$$

Then, we use the exact same argument as in the proofs of Lemma 5.D.1 and Corollary 5.D.2 where for all $u \in T^*$, the random variable $\Gamma_{u,k,x}(\theta)$ is replaced by the random variable:

$$\sup_{\theta' \in \Theta_0 \,:\, \|\theta' - \theta\| \leq \delta} \left| \Gamma_{u,h(u),x}(\theta') - \Gamma_{u,h(u),x}(\theta) \right|,$$

which are in $L^2(\mathbb{P}_{\theta^\star})$ using the assumptions of Proposition 5.4.5. This gives us that the first term in the upper bound of (5.117) converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[\sup_{\theta' : \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial,\infty}(\theta') - \Gamma_{\partial,\infty}(\theta)| \right],$$

which, by Corollary 5.D.4, vanishes when $\delta \to 0$. Corollary 5.D.2 implies that the second term in the upper bound of (5.117) vanishes $\mathbb{P}_{\theta^\star}$-a.s. when $n \to \infty$. This concludes the proof. □

Combining the previous lemmas in this appendix and Lemma 5.4.17, we are now ready to prove Proposition 5.4.5.

Proof of Proposition 5.4.5. By Lemma 5.4.17, for all $u \in T$, we have that $(\Gamma_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$ is a Cauchy sequence uniformly w.r.t. $\theta \in \Theta_0$ in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ that converges to some limit $\Gamma_{u,\infty}(\theta)$ (that does not depend on x). By Corollary 5.D.2, we have that $\mathbb{P}_{\theta^\star}$-a.s. the convergence for the the average of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ over $u \in T_n^*$ holds uniformly in $x \in \mathcal{X}$, that is, (5.66) in Proposition 5.4.5 holds. By Corollary 5.D.4, we have that the function $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$ is continuous on Θ_0. Finally, the last part of the proposition is given by Lemma 5.D.5. □

Bibliographie

[Abb18] Emmanuel Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[AD20] Romain Abraham and Jean-François Delmas. An introduction to Galton-Watson trees and their local limits, September 2020, arXiv:1506.05571, [math].

[ADW23] Romain Abraham, Jean-François Delmas, and Julien Weibel. Probability-graphons: Limits of large dense weighted graphs, December 2023, arXiv:2312.15935, [cs, math].

[ACC13] Edo M. Airoldi, Thiago B. Costa, and Stanley H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[AAN04] Réka Albert, István Albert, and Gary L. Nakarado. Structural vulnerability of the North American power grid. *Physical Review E*, 69(2):025103, February 2004.

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002.

[Ald81] David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, December 1981.

[Ald97] John Aldrich. R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3), September 1997.

[AN22] Giacomo Aletti and Giovanni Naldi. Opinion dynamics on graphon: The piecewise constant case. *Applied Mathematics Letters*, 133:108227, November 2022.

[AMGM18] Lázaro Alonso, Jose A. Méndez-Bermúdez, A González-Meléndrez, and Yamir Moreno. Weighted random-geometric and random-rectangular graphs: Spectral and eigenfunction properties of the adjacency matrix. *Journal of Complex Networks*, 6(5):753–766, October 2018.

[ADL13] Hamed Amini, Moez Draief, and Marc Lelarge. Flooding in Weighted Sparse Random Graphs. *SIAM Journal on Discrete Mathematics*, 27(1):1–26, January 2013.

[AL15] Hamed Amini and Marc Lelarge. The diameter of weighted random graphs. *The Annals of Applied Probability*, 25(3), June 2015, 1112.6330.

[AWC99] Carolyn J. Anderson, Stanley Wasserman, and Bradley Crouch. A p* primer: Logit models for social networks. *Social Networks*, 21(1):37–66, January 1999.

[ANS24] Eleanor Archer, Asaf Nachmias, and Matan Shalev. The GHP Scaling Limit of Uniform Spanning Trees in High Dimensions. *Communications in Mathematical Physics*, 405(3):73, March 2024.

[Ath12a] Krishna B. Athreya. Coalescence in Critical and Subcritical Galton-Watson Branching Processes. *Journal of Applied Probability*, 49(3):627–638, September 2012.

[Ath12b] Krishna B. Athreya. Coalescence in the recent past in rapidly growing populations. *Stochastic Processes and their Applications*, 122(11):3757–3766, November 2012.

[AK98a] Krishna B. Athreya and Hye-Jeong Kang. Some limit theorems for positive recurrent branching Markov chains: I. *Advances in Applied Probability*, 30(3):693–710, September 1998.

[AK98b] Krishna B. Athreya and Hye-Jeong Kang. Some limit theorems for positive recurrent branching Markov chains: II. *Advances in Applied Probability*, 30(3):711–722, September 1998.

[AN72] Krishna B. Athreya and Peter E. Ney. *Branching Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1972.

[Ath69] Krishna Balasundaram Athreya. Limit theorems for multitype continuous time markov branching processes: I. The case of an eigenvector linear functional. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 12(4):320–332, 1969.

[ADL22] Konstantin Avrachenkov, Maximilien Dreveton, and Lasse Leskelä. Community recovery in non-binary and temporal stochastic block models, August 2022, arXiv:2008.04790, [cs, math, stat].

[AD23] Nathalie Ayi and Nastassia Pouradier Duteil. Graph Limit for Interacting Particle Systems on Weighted Random Graphs, July 2023, arXiv:2307.12801, [math].

[AD24] Nathalie Ayi and Nastassia Pouradier Duteil. Large-population limits of non-exchangeable particle systems, January 2024, arXiv:2401.07748, [math].

[Bac11] Nicolas Bacaër. *A Short History of Mathematical Population Dynamics*. Springer London, London, 2011.

[BS22] Ágnes Backhausz and Balázs Szegedy. Action convergence of operators and graphs. *Canadian Journal of Mathematics*, 74(1):72–121, February 2022.

[Ban19] Vincent Bansaye. Ancestral Lineages and Limit Theorems for Branching Markov Chains in Varying Environment. *Journal of Theoretical Probability*, 32(1):249–281, March 2019.

[BA99] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[BAJ00] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1-4):69–77, June 2000.

[BBB+23] Jnaneshwar Baslingker, Shankar Bhamidi, Nicolas Broutin, Sanchayan Sen, and Xuan Wang. Scaling limits and universality: Critical percolation on weighted graphs converging to an l^3 graphon, March 2023, arXiv:2303.10082, [math].

[BP66] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966.

[BPSW70] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171, February 1970.

[BK24] Erhan Bayraktar and Donghan Kim. Concentration of measure for graphon particle system. *Advances in Applied Probability*, pages 1–28, January 2024.

[BWZ23] Erhan Bayraktar, Ruoyu Wu, and Xin Zhang. Propagation of Chaos of Forward–Backward Stochastic Differential Equations with Graphon Interactions. *Applied Mathematics & Optimization*, 88(1):25, August 2023.

[BC78] Edward A. Bender and E. Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, May 1978.

[BS01] Itai Benjamini and Oded Schramm. Recurrence of Distributional Limits of Finite Planar Graphs. *Electronic Journal of Probability*, 6(none), January 2001.

[BDSG09] Bernard Bercu, Benoîte De Saporta, and Anne Gégout-Petit. Asymptotic Analysis for Bifurcating AutoRegressive Processes via a Martingale Approach. *Electronic Journal of Probability*, 14(none), January 2009.

[BCM09] Marc Bernot, Vicent Caselles, and Jean-Michel Morel. *Optimal Transportation Networks: Models and Theory*. Number 1955 in Lecture Notes in Mathematics. Springer, Berlin, 2009.

[BS96] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Optimization and Neural Computation Series. Athena Scientific, Belmont, Mass, 1996.

[BVDHH10] Shankar Bhamidi, Remco Van Der Hofstad, and Gerard Hooghiemstra. First passage percolation on random graphs with finite mean degrees. *The Annals of Applied Probability*, 20(5), October 2010.

[BRR98] Peter J. Bickel, Ya'acov Ritov, and Tobias Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4), August 1998.

[Bie45] Irenée Jules Bienaymé. De la loi de multiplication et de la durée des familles. *Société philomathique de Paris Extraits*, 5:37–39, 1845.

[BWX13] Jacob Biesinger, Yuanfeng Wang, and Xiaohui Xie. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics*, 14(S5):S4, April 2013.

[Bil99a] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, New York, second edition edition, 1999.

[Bil99b] Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Probability and Statistics Section. Wiley, New York, 2nd ed edition, 1999.

[BPD22a] S. Valère Bitseki Penda and Jean-François Delmas. Central limit theorem for bifurcating Markov chains under pointwise ergodic conditions. *The Annals of Applied Probability*, 32(5), October 2022.

[BPD22b] S. Valère Bitseki Penda and Jean-François Delmas. Central Limit Theorem for Kernel Estimator of Invariant Density in Bifurcating Markov Chains Models. *Journal of Theoretical Probability*, October 2022.

[BPDG14] S. Valère Bitseki Penda, Hacène Djellout, and Arnaud Guillin. Deviation inequalities, moderate deviations and some limit theorems for bifurcating Markov chains with application. *The Annals of Applied Probability*, 24(1), February 2014.

[BRZ95] Pavel M. Bleher, Jean Ruiz, and Valentin A. Zagrebnov. On the purity of the limiting gibbs state for the Ising model on the Bethe lattice. *Journal of Statistical Physics*, 79(1-2):473–482, April 1995.

[Bog07a] Vladimir I. Bogachev. *Measure Theory Vol. 1*, volume 1. Springer, Berlin ; New York, 2007.

[Bog07b] Vladimir I. Bogachev. *Measure Theory Vol. 2*, volume 2. Springer, Berlin ; New York, 2007.

[Bog18] Vladimir I. Bogachev. *Weak Convergence of Measures*. Number volume 234 in Mathematical Surveys and Monographs. American Mathematical Society, Providence, Rhode Island, 2018.

[Bol80] Béla Bollobás. A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. *European Journal of Combinatorics*, 1(4):311–316, December 1980.

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, August 2007.

[BJR10] Béla Bollobás, Svante Janson, and Oliver Riordan. The Cut Metric, Random Graphs, and Branching Processes. *Journal of Statistical Physics*, 140(2):289–335, July 2010.

[BR09] Béla Bollobás and Oliver Riordan. Metrics for sparse graphs. In Sophie Huczynska, James D. Mitchell, and Colva M. Roney-Dougal, editors, *Surveys in Combinatorics 2009*, pages 211–288. Cambridge University Press, 1 edition, July 2009.

[BR11] Béla Bollobás and Oliver Riordan. Sparse graphs: Metrics and random models. *Random Structures & Algorithms*, 39(1):1–38, August 2011.

[Bop87] Ravi B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (Sfcs 1987)*, pages 280–285, Los Angeles, CA, USA, October 1987. IEEE.

[BC17] Christian Borgs and Jennifer Chayes. Graphons: A Nonparametric Method to Model, Estimate, and Design Algorithms for Massive Networks. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 665–672, Cambridge Massachusetts USA, June 2017. ACM.

[BCCZ19] Christian Borgs, Jennifer Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062, May 2019.

[BCL+12] Christian Borgs, Jennifer Chayes, László Lovász, Vera Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Annals of Mathematics*, 176(1):151–219, July 2012.

[BCL+06a] Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, Balázs Szegedy, and Katalin Vesztergombi. Graph limits and parameter testing. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, pages 261–270, Seattle WA USA, May 2006. ACM.

[BCL+06b] Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi. Counting Graph Homomorphisms. In Martin Klazar, Jan Kratochvíl, Martin Loebl, Jiří Matoušek, Pavel Valtr, and Robin Thomas, editors, *Topics in Discrete Mathematics*, volume 26, pages 315–371. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[BCSZ18] Christian Borgs, Jennifer Chayes, Adam Smith, and Ilias Zadik. Revealing Network Structure, Confidentially: Improved Rates for Node-Private Graphon Estimation. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 533–543, Paris, October 2018. IEEE.

[BCCH18] Christian Borgs, Jennifer T. Chayes, Henry Cohn, and Nina Holden. Sparse Exchangeable Graphs and Their Limits via Graphon Processes. *Journal of Machine Learning Research*, 18(210):1–71, 2018.

[BCCL19] Christian Borgs, Jennifer T. Chayes, Henry Cohn, and László Miklós Lovász. Identifiability for Graphexes and the Weak Kernel Metric. In Imre Bárány, Gyula O. H. Katona, and Attila Sali, editors, *Building Bridges II*, volume 28, pages 29–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 2019.

[BCCV19] Christian Borgs, Jennifer T. Chayes, Henry Cohn, and Victor Veitch. Sampling perspectives on sparse exchangeable graphs. *The Annals of Probability*, 47(5), September 2019.

[BCCZ18] Christian Borgs, Jennifer T. Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence II: LD convergence, quotients and right convergence. *The Annals of Probability*, 46(1), January 2018.

[BCDS21] Christian Borgs, Jennifer T. Chayes, Souvik Dhara, and Subhabrata Sen. Limits of sparse configuration models and beyond: Graphexes and MultiGraphexes. *The Annals of Probability*, 49(6), November 2021.

[BCL+08] Christian Borgs, Jennifer T. Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, December 2008.

[BCS15] Christian Borgs, Jennifer T. Chayes, and Adam Smith. Private Graphon Estimation for Sparse Graphs, June 2015, arXiv:1506.06162, [cs, math, stat].

[BS02] Stefan Bornholdt and Hans Georg Schuster, editors. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley, 1 edition, November 2002.

[BFA22] Nizar Bouguila, Wentao Fan, and Manar Amayri, editors. *Hidden Markov Models and Applications*. Unsupervised and Semi-Supervised Learning. Springer International Publishing, Cham, 2022.

[BGJM11] Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, 2011.

[BCLS87] Thang N. Bui, Soma Chaudhuri, F. Tom Leighton, and Michael Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, June 1987.

[CH21] Peter E. Caines and Minyi Huang. Graphon Mean Field Games and Their Equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, January 2021.

[CMR05] Olivier Cappé, Éric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer New York, New York, NY, 2005.

[CF17] François Caron and Emily B. Fox. Sparse Graphs Using Exchangeable Random Measures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1295–1366, November 2017.

[CL95] Kung-Sik Chan and Johannes Ledolter. Monte Carlo EM Estimation for Time Series Models Involving Counts. *Journal of the American Statistical Association*, 90(429):242–252, March 1995.

[Cha15] Sourav Chatterjee. Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1), February 2015.

[CB01] Hyeokho Choi and Richard G. Baraniuk. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Transactions on Image Processing*, 10(9):1309–1321, September 2001.

[CNB98] Matt S. Crouse, Robert D. Nowak, and Richard G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, April 1998.

[DJS95] Piet De Jong and Neil Shephard. The simulation smoother for time series models. *Biometrika*, 82(2):339–350, 1995.

[DDZ22] Jean-François Delmas, Dylan Dronnier, and Pierre-André Zitt. An infinite-dimensional metapopulation SIS model. *Journal of Differential Equations*, 313:1–53, March 2022.

[DDZ23] Jean-François Delmas, Dylan Dronnier, and Pierre-André Zitt. Optimal vaccination: Various (counter) intuitive examples. *Journal of Mathematical Biology*, 86(2):26, February 2023.

[DM10] Jean-François Delmas and Laurence Marsalle. Detection of cellular aging in a Galton–Watson process. *Stochastic Processes and their Applications*, 120(12):2495–2519, December 2010.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):1–22, September 1977.

[DJ08] Petri Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni. Serie VII*, pages 33–61, 2008.

[DGKR15] Peter Diao, Dominique Guillot, Apoorva Khare, and Bala Rajaratnam. Differential calculus on graphon space. *Journal of Combinatorial Theory, Series A*, 133:183–227, July 2015.

[DM01] Randal Douc and Catherine Matias. Asymptotics of the Maximum Likelihood Estimator for General Hidden Markov Models. *Bernoulli*, 7(3):381, June 2001, 3318493.

[DMOVH11] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1), February 2011.

[DMPS18] Randal Douc, Éric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2018.

[DMR04] Randal Douc, Éric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5), October 2004.

[DRS16] Randal Douc, François Roueff, and Tepmony Sim. The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains. *The Annals of Applied Probability*, 26(4), August 2016.

[DM10] Moez Draief and Laurent Massoulié. *Epidemics and Rumours in Complex Networks*. Number 369 in London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge ; New York, 2010.

[DWB08] Marco F. Duarte, Michael B. Wakin, and Richard G. Baraniuk. Wavelet-domain compressive signal reconstruction using a Hidden Markov Tree model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5137–5140, Las Vegas, NV, USA, March 2008. IEEE.

[DF64] Lester Dubins and David Freedman. Measurable sets of measures. *Pacific Journal of Mathematics*, 14(4):1211–1222, December 1964.

[Duf11] Marie Duflo. *Random Iterative Models*. Springer, Berlin, 2011.

[DGG04] Jean-Baptiste Durand, Paulo Gonçalves, and Yann Guédon. Computational Methods for Hidden Markov Tree Models—An Application to Wavelet Trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, September 2004.

[DGCC05] Jean-Baptiste Durand, Yann Guédon, Yves Caraglio, and Evelyne Costes. Analysis of the plant architecture via tree-structured statistical models: The hidden Markov tree models. *New Phytologist*, 166(3):813–825, June 2005.

[DK97] James Durbin and Siem Jan Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, September 1997.

[Dur06] Rick Durrett. *Random Graph Dynamics*. Cambridge University Press, 1 edition, October 2006.

[DF89] Martin E. Dyer and Alan M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, December 1989.

[Eng89] Ryszard Engelking. *General topology*. Number 6 in Sigma series in pure mathematics. Heldermann, Berlin, rev. and completed ed edition, 1989.

[ER59] Paul Erdős and Alfréd Rényi. On random graphs. I. *Publicationes Mathematicae Debrecen*, 6(3-4):290–297, 1959.

[ER60] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[ER61a] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist.*, 38:343–347, 1961.

[ER61b] Paul Erdős and Alfréd Rényi. On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.*, 12:261–267, 1961.

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[Eul36] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1736.

[EKPS00] William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the Ising model. *The Annals of Applied Probability*, 10(2), May 2000.

[FOSU16] Victor Falgas-Ravry, Kelly O'Connell, Johanna Strömberg, and Andrew Uzzell. Multicolour containers and the entropy of decorated graph limits, July 2016, arXiv:1607.08152, [math].

[FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. *ACM SIGCOMM Computer Communication Review*, 29(4):251–262, October 1999.

[Fis92] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 11–44. Springer New York, New York, NY, 1992.

[Fis22] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, January 1922.

[FS86] Ove Frank and David Strauss. Markov Graphs. *Journal of the American Statistical Association*, 81(395):832–842, September 1986.

[FLS06] Michael Freedman, László Lovász, and Alexander Schrijver. Reflection positivity, rank connectivity, and homomorphism of graphs. *Journal of the American Mathematical Society*, 20(1):37–51, April 2006.

[FK99] Alan Frieze and Ravi Kannan. Quick Approximation to Matrices and Applications. *Combinatorica*, 19(2):175–220, February 1999.

[FK15] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 1 edition, October 2015.

[GW74] Francis Galton and H. W. Watson. On the probability of the extinction of families. *J. Anthropol. Inst.*, 4:138–144, 1874.

[GLZ15] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6), December 2015.

[GM21] Chao Gao and Zongming Ma. Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing. *Statistical Science*, 36(1), February 2021.

[Gar09] Diego Garlaschelli. The weighted random graph model. *New Journal of Physics*, 11(7):073005, July 2009, 0902.0897.

[GL06] Valentine Genon-Catalot and Catherine Laredo. Leroux's method for general hidden Markov models. *Stochastic Processes and their Applications*, 116(2):222–243, February 2006.

[Gil59] Edgar N. Gilbert. Random Graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, December 1959.

[GOB13] Édouard Grave, Guillaume Obozinski, and Francis Bach. Hidden Markov tree models for semantic class induction. In *CoNLL*, volume Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 94–103, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[Guy07] Julien Guyon. Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. *The Annals of Applied Probability*, 17(5-6):1538–1569, October 2007, 0710.5434.

[HJV07] Patricia Haccou, Peter Jagers, and Vladimir A. Vatutin. *Branching Processes: Variation, Growth, and Extinction of Populations*. Number 5 in Cambridge Studies in Adaptive Dynamics. Cambridge Univ. Press, Cambridge, digitally printed version edition, 2007.

[HBSLLB+17] Houda Hanzouli-Ben Salah, Jerome Lapuyade-Lahorgue, Julien Bert, Didier Benoit, Philippe Lambin, Angela Van Baardwijk, Emmanuel Monfrini, Wojciech Pieczynski, Dimitris Visvikis, and Mathieu Hatt. A framework based on hidden Markov trees for multimodal image co-segmentation. *Medical Physics*, 44(11):5835–5848, November 2017.

[Har63] Theodore Edward Harris. *The Theory of Branching Processes*, volume 119 of *Die Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin; Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

[HLS14] Hamed Hatami, László Lovász, and Balázs Szegedy. Limits of locally–globally convergent graph sequences. *Geometric and Functional Analysis*, 24(1):269–296, February 2014.

[HLM12] Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. Community Detection in the Labelled Stochastic Block Model. *arXiv:1209.2910 [physics]*, September 2012, 1209.2910.

[HNT18] Jan Hladký, Asaf Nachmias, and Tuan Tran. The local limit of the uniform spanning tree on dense graphs. *Journal of Statistical Physics*, 173(3-4):502–545, November 2018, 1711.09788.

[HV23] Jan Hladký and Gopal Viswanathan. Random minimum spanning tree and dense graph limits, October 2023, arXiv:2310.11705, [math].

[HLL83] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, June 1983.

[HL81] Paul W. Holland and Samuel Leinhardt. An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50, March 1981.

[Hoo79] D. N. Hoover. Relations on probability spaces and arrays of random variables, 1979.

[HYG17] Kai Hu, Wei Yang, and Xieping Gao. Microcalcification diagnosis in digital mammography using extreme learning machine based on hidden Markov tree model of dual-tree complex wavelet transform. *Expert Systems with Applications*, 86:135–144, November 2017.

[HWYZ23] Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute*, September 2023.

[HS96] A. Humphreys and Stephen Simpson. Separable Banach space theory needs strong set existence axioms. *Transactions of the American Mathematical Society*, 348(10):4231–4255, 1996.

[HG13] Thomas R. Hurd and James P. Gleeson. On Watts' cascade model with random link weights. *Journal of Complex Networks*, 1(1):25–43, June 2013.

[Jan13] Svante Janson. Graphons, cut norm and distance, couplings and rearrangements. *New York Journal of Mathematics*, 4, March 2013, 1009.2376.

[Jan16] Svante Janson. Graphons and cut metric on sigma-finite measure spaces, August 2016, arXiv:1608.01833, [math].

[Jan22] Svante Janson. On convergence for graphexes. *European Journal of Combinatorics*, 104:103549, August 2022.

[JŁR00] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley, New York, 2000.

[JP99] Jens Ledet Jensen and Niels Væver Petersen. Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics*, 27(2), April 1999.

[Jia18] Xu Jiaming. Rates of Convergence of Spectral Methods for Graphon Estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5433–5442. PMLR, 2018.

[JL15] Varun Jog and Po-Ling Loh. Recovering communities in weighted stochastic block models. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1308–1315, Monticello, IL, September 2015. IEEE.

[Kal02] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[Kal17] Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, Cham, 2017.

[KS19] Hiroyuki Kasahara and Katsumi Shimotsu. Asymptotic properties of the maximum likelihood estimator in regime switching econometric models. *Journal of Econometrics*, 208(2):442–467, February 2019.

[Ken75] David G. Kendall. The Genealogy of Genealogy Branching Processes before (and after) 1873. *Bulletin of the London Mathematical Society*, 7(3):225–253, November 1975.

[KBV21] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. On the Universality of Graph Neural Networks on Large Random Graphs. *Advances in Neural Information Processing Systems*, 34:6960–6971, 2021, 2105.13099.

[KV23] Nicolas Keriven and Samuel Vaiter. What functions can Graph Neural Networks compute on random graphs? The role of Positional Encoding. In *Advances in Neural Information Processing Systems*, volume 36, pages 11823–11849. Curran Associates, Inc., 2023.

[KS66] Harry Kesten and Bernt P. Stigum. A Limit Theorem for Multidimensional Galton-Watson Processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, October 1966.

[KSC98] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *The Review of Economic Studies*, 65(3):361–393, July 1998.

[KA15] Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*, volume 19 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY, 2015.

[KMS17] István Z. Kiss, Joel C. Miller, and Péter L. Simon. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*, volume 46 of *Interdisciplinary Applied Mathematics*. Springer International Publishing, Cham, 2017.

[KTV17] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1), February 2017.

[KV19] Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174(3-4):1033–1090, August 2019.

[KR11] István Kolossváry and Balázs Ráth. Multigraph limits and exchangeability. *Acta Mathematica Hungarica*, 130(1-2):1–34, January 2011.

[KL20] Júlia Komjáthy and Bas Lodewijks. Explosion in weighted hyperbolic random graphs and geometric inhomogeneous random graphs. *Stochastic Processes and their Applications*, 130(3):1309–1367, March 2020.

[KDM13] Shuhei Kondo, Kevin Duh, and Yuji Matsumoto. Hidden markov tree model for word alignment. In *WMT*, volume Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 503–511, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[Kos01] Timo Koski. *Hidden Markov Models for Bioinformatics*. Number 2 in Computational Biology. Kluwer academic publ, Dordrecht [etc.], 2001.

[KLS19] Dávid Kunszenti-Kovács, László Lovász, and Balázs Szegedy. Measures on the square as sparse graph limits. *Journal of Combinatorial Theory, Series B*, 138:1–40, September 2019.

[KLS22] Dávid Kunszenti-Kovács, László Lovász, and Balázs Szegedy. Multigraph limits, unbounded kernels, and Banach space decorated graphs. *Journal of Functional Analysis*, 282(2):109284, January 2022.

[LS22] Daniel Lacker and Agathe Soret. A Label-State Formulation of Stochastic Graphon Games and Approximate Equilibria on Large Networks. *Mathematics of Operations Research*, page moor.2022.1329, November 2022.

[LGM00] François Le Gland and Laurent Mevel. Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models. *Mathematics of Control, Signals, and Systems*, 13(1):63–93, February 2000.

[LMX13] Marc Lelarge, Laurent Massoulie, and Jiaming Xu. Reconstruction in the labeled stochastic block model. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5, Sevilla, Spain, September 2013. IEEE.

[LMX15] Marc Lelarge, Laurent Massoulie, and Jiaming Xu. Reconstruction in the Labelled Stochastic Block Model. *IEEE Transactions on Network Science and Engineering*, 2(4):152–163, October 2015.

[LW06] Alberto Leon-Garcia and Indra Widjaja. *Communication Networks: Fundamental Concepts and Key Architectures*. McGraw-Hill Series in Computer Science. McGraw-Hill, Boston, 2. ed., internat. ed edition, 2006.

[Ler92] Brian G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143, February 1992.

[Lew09] Theodore Gyle Lewis. *Network Science: Theory and Practice*. Wiley, Hoboken, NJ, 2009.

[LMD+08] Ariel B. Lindner, Richard Madden, Alice Demarez, Eric J. Stewart, and François Taddei. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proceedings of the National Academy of Sciences*, 105(8):3076–3081, February 2008.

[LCL13] Chuanhai Liu, Raymond Carroll, and Faming Liang. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley, Hoboken, N.J., 2013.

[Lov67] László Lovász. Operations with structures. *Acta Mathematica Academiae Scientiarum Hungaricae*, 18(3-4):321–328, September 1967.

[Lov12] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Society, Colloquium Publications, 2012.

[LS06] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, November 2006.

[LS07] László Lovász and Balázs Szegedy. Szemerédi's Lemma for the Analyst. *GAFA Geometric And Functional Analysis*, 17(1):252–270, April 2007.

[LS10] László Lovász and Balázs Szegedy. Limits of compact decorated graphs, October 2010, arXiv:1010.5155, [math].

[LG24] Yuetian Luo and Chao Gao. Computational Lower Bounds for Graphon Estimation via Low-degree Polynomials, May 2024, arXiv:2308.15728, [cs, math, stat].

[LKR13] Dean Lusher, Johan Koskinen, and Garry Robins, editors. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Number 32 in Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge, 2013.

[MK23] Nabeela Majid and Rizwan Hasan Khan. Protein aggregation: Consequences, mechanism, characterization and inhibitory strategies. *International Journal of Biological Macromolecules*, 242:125123, July 2023.

[MBY+12] Mahender K. Makhijani, Niranjan Balu, Kiyofumi Yamada, Chun Yuan, and Krishna S. Nayak. Accelerated 3D MERGE carotid imaging using compressed sensing with a hidden markov tree model. *Journal of Magnetic Resonance Imaging*, 36(5):1194–1202, November 2012.

[ME14] Rogemar S. Mamon and Robert J. Elliott, editors. *Hidden Markov Models in Finance: Further Developments and Applications, Volume II*, volume 209 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA, 2014.

[MW43] Henry B. Mann and Abraham Wald. On the Statistical Treatment of Linear Stochastic Difference Equations. *Econometrica*, 11(3/4):173, July 1943, 1905674.

[Mar06] Andreï Andreïvitch Markov. Extension of the law of large numbers to dependent quantities (in Russian). *Izvestiia Fiz.-Matem. Obsch. Kazan Univ., (2nd Ser.)*, 15:135–156, 1906.

[Mar04] Andreï Andreïvitch Markov. The extension of the law of large numbers onto quantities depending on each other. In *Probability and Statistics. Russian Papers*, Selected and Translated by Oscar Sheynin, pages 143–158. NG Verlag, Berlin, 2004.

[MT10] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Univ. Press, Cambridge, 2. ed., reprint edition, 2010.

[NSK20] So Nakashima, Yuki Sughiyama, and Tetsuya J Kobayashi. Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics*, 36(9):2829–2838, May 2020.

[NODM20] Jaroslav Nešetřil and Patrice Ossona De Mendez. A Unified Approach to Structural Limits and Limits of Graphs with Bounded Tree-Depth. *Memoirs of the American Mathematical Society*, 263(1272):0–0, January 2020.

[Nev86] Jacques Neveu. Arbres et processus de Galton-Watson. *Annales de l'I.H.P. Probabilités et statistiques*, 22:199–207, 1986.

[New03] Mark E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, January 2003.

[OCB+09] Victor Olariu, Daniel Coca, Stephen A. Billings, Peter Tonge, Paul Gokhale, Peter W. Andrews, and Visakan Kadirkamanathan. Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics*, 25(21):2824–2830, November 2009.

[PN04] Juyong Park and Mark E. J. Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):066117, December 2004.

[Pen03] Mathew Penrose. *Random Geometric Graphs*. Number 5 in Oxford Studies in Probability. Oxford university press, Oxford, 2003.

[Pur66] Robert Purves. Bimeasurable functions. *Fund. Math.*, 58:149–157, 1966.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[RB05] Bin Ran and David E. Boyce. *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*. Springer-Verlag, Berlin, 2005.

[RS18] Dieter Rasch and Dieter Schott. *Mathematical Statistics*. John Wiley & Sons, Hoboken, 2018.

[RPKL07] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, May 2007.

[RCBK00] Justin Romberg, Hyeokho Choi, Richard G. Baraniuk, and Nicholas Kingsbury. Multiscale classification using complex wavelets and hidden Markov tree models. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, pages 371–374 vol.2, Vancouver, BC, Canada, 2000. IEEE.

[Rud96] Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Boston, Mass., 2. ed., [nachdr.] edition, 1996.

[SH17] Hamid Reza Shahdoosti and Seyede Mahya Hazavei. Image denoising in dual contourlet domain using hidden Markov tree models. *Digital Signal Processing*, 67:17–29, August 2017.

[SAL18] Farhad Shahnia, Ali Arefi, and Gerard Ledwich. *Electric Distribution Network Planning*. Power Systems. Springer, Singapore, 2018.

[Str86] David Strauss. On a General Class of Models for Interaction. *SIAM Review*, 28(4):513–527, December 1986.

[SKK+23] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023.

[TMB10] Jens Tyedmers, Axel Mogk, and Bernd Bukau. Cellular strategies for controlling protein aggregation. *Nature Reviews Molecular Cell Biology*, 11(11):777–788, November 2010.

[van16] Remco van der Hofstad. *Random Graphs and Complex Networks*. Cambridge University Press, 1 edition, November 2016.

[van24] Remco van der Hofstad. *Random Graphs and Complex Networks. Volume 2*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2024.

[Var58] Veeravalli S. Varadarajan. Weak convergence of measures on separable metric spaces. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):15–22, 1958.

[VR15] Victor Veitch and Daniel M. Roy. The Class of Random Graphs Arising from Exchangeable Random Measures, December 2015, arXiv:1512.03099, [cs, math, stat].

[VR19] Victor Veitch and Daniel M. Roy. Sampling and estimation for (sparse) exchangeable graphs. *The Annals of Statistics*, 47(6), December 2019.

[Wal49] Abraham Wald. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, December 1949.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[Wei24a] Julien Weibel. Asymptotic properties of the maximum likelihood estimator for Hidden Markov Models indexed by binary trees, September 2024, arXiv:2409.06295, [math, stat].

[Wei24b] Julien Weibel. Ergodic theorem for branching Markov chains indexed by trees with arbitrary shape, March 2024, arXiv:2403.16505, [math].

[WHF+23] Jun-Hao Wen, Xiang-Hong He, Ze-Sen Feng, Dong-Yi Li, Ji-Xin Tang, and Hua-Feng Liu. Cellular Protein Aggregates: Formation, Biological Effects, and Ways of Elimination. *International Journal of Molecular Sciences*, 24(10):8593, May 2023.

[WO13] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation, September 2013, arXiv:1309.5936, [math, stat].

[XJS18] Miao Xie, Zhe Jiang, and Arpan Man Sainju. Geographical Hidden Markov Tree for Flood Extent Mapping. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2545–2554, London United Kingdom, July 2018. ACM.

[XML14] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 903–920, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[XJL20] Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, 48(1), February 2020.

[YD15] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. Springer London, London, 2015.

[YP16] Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[ZM09] Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall/CRC, 0 edition, April 2009.

Julien WEIBEL

Graphons de probabilités, limites de graphes pondérés aléatoires et chaînes de Markov branchantes cachées

Résumé :
Les graphes sont des objets mathématiques qui servent à modéliser tout type de réseaux, comme les réseaux électriques, les réseaux de communications et les réseaux sociaux. Formellement un graphe est composé d'un ensemble de sommets et d'un ensemble d'arêtes reliant des paires de sommets. Les sommets représentent par exemple des individus, tandis que les arêtes représentent les interactions entre ces individus. Dans le cas d'un graphe pondéré, chaque arête possède un poids ou une décoration pouvant modéliser une distance, une intensité d'interaction, une résistance. La modélisation de réseaux réels fait souvent intervenir de grands graphes qui ont un grand nombre de sommets et d'arêtes.
La première partie de cette thèse est consacrée à l'introduction et à l'étude des propriétés des objets limites des grands graphes pondérés : les graphons de probabilités. Ces objets sont une généralisation des graphons introduits et étudiés par Lovász et ses co-auteurs dans le cas des graphes sans poids sur les arêtes. À partir d'une distance induisant la topologie faible sur les mesures, nous définissons une distance de coupe sur les graphons de probabilités. Nous exhibons un critère de tension pour les graphons de probabilités lié à la compacité relative dans la distance de coupe. Enfin, nous prouvons que cette topologie coïncide avec la topologie induite par la convergence en distribution des sous-graphes échantillonnés.
Dans la deuxième partie de cette thèse, nous nous intéressons aux modèles markoviens cachés indexés par des arbres. Nous montrons la consistance forte et la normalité asymptotique de l'estimateur de maximum de vraisemblance pour ces modèles sous des hypothèses standards. Nous montrons un théorème ergodique pour des chaînes de Markov branchantes indexés par des arbres avec des formes générales. Enfin, nous montrons que pour une chaîne stationnaire et réversible, le graphe ligne est la forme d'arbre induisant une variance minimale pour l'estimateur de moyenne empirique parmi les arbres avec un nombre donné de sommets.

Mots clés : graphes pondérés aléatoires, réseaux stochastiques, graphons de probabilités, modèles markoviens cachés indexés par des arbres, chaînes de Markov branchantes, processus de branchement

Probability-graphons, limits of weighted random graphs and hidden branching Markov chains

Abstract :
Graphs are mathematical objects used to model all kinds of networks, such as electrical networks, communication networks, and social networks. Formally, a graph consists of a set of vertices and a set of edges connecting pairs of vertices. The vertices represent, for example, individuals, while the edges represent the interactions between these individuals. In the case of a weighted graph, each edge has a weight or a decoration that can model a distance, an interaction intensity, or a resistance. Modeling real-world networks often involves large graphs with a large number of vertices and edges.
The first part of this thesis is dedicated to introducing and studying the properties of the limit objects of large weighted graphs : probability-graphons. These objects are a generalization of graphons introduced and studied by Lovász and his co-authors in the case of unweighted graphs. Starting from a distance that induces the weak topology on measures, we define a cut distance on probability-graphons. We exhibit a tightness criterion for probability-graphons related to relative compactness in the cut distance. Finally, we prove that this topology coincides with the topology induced by the convergence in distribution of the sampled subgraphs.
In the second part of this thesis, we focus on hidden Markov models indexed by trees. We show the strong consistency and asymptotic normality of the maximum likelihood estimator for these models under standard assumptions. We prove an ergodic theorem for branching Markov chains indexed by trees with general shapes. Finally, we show that for a stationary and reversible chain, the line graph is the tree shape that induces the minimal variance for the empirical mean estimator among trees with a given number of vertices.

Keywords : random weighted graphs, stochastic networks, probability-graphons, hidden Markov models indexed by trees, branching Markov chains, branching processes

INSTITUT ORLÉANS · TOURS Mathématiques & Physique Théorique DENIS POISSON

Institut Denis Poisson
Rue de Chartres B.P. 6759
45067 ORLEANS CEDEX 2

cnrs