A Proof of Lemma 1: diameter of the cells

We start by recalling some important properties of the Mondrian process, which are exposed in [RT09].

Fact 1 (Consistency, Mondrian slices). Let \( M_\lambda \sim \text{MP}(\lambda, [0, 1]^d) \) be a Mondrian partition, and \( C = \prod_{j=1}^d [a_j, b_j] \subset [0, 1]^d \), be an axis-aligned box (we authorize lower-dimensional boxes when \( a_j = b_j \) for some dimensions \( j \)). Consider the restriction \( M_\lambda | C \) of \( M_\lambda \) on \( C \), i.e. the partition on \( C \) induced by the partition \( M_\lambda \) of \([0, 1]^d\). Then \( M_\lambda | C \sim \text{MP}(\lambda, C) \).

Fact 2 (Dimension 1). For \( d = 1 \), the splits from a Mondrian process \( M_\lambda \sim \text{MP}(\lambda, [0, 1]) \) form a subset of \([0, 1]\), which is distributed as a Poisson point process of intensity \( \lambda dx \).

We will now establish the technical lemma 1. In what follows, \( x \in [0, 1]^d \) is arbitrary, and we let \( A_\lambda(x) \) denote the (random) cell of a Mondrian partition \( M_\lambda \sim \text{MP}(\lambda, [0, 1]^d) \) containing \( x \).

Proof of Lemma 1. Let \( A_\lambda(x) = \prod_{j=1}^d [L^1_j(x), R^1_j(x)] \) denote the (random) cell of a Mondrian partition \( M_\lambda \sim \text{MP}(\lambda, [0, 1]^d) \) containing \( x \in [0, 1]^d \). By definition, the \( \ell_\infty \)-diameter \( D_\lambda(x) \) of \( A_\lambda(x) \) is \( \max_{1 \leq j \leq d} (R^1_j(x) - L^1_j(x)) \). Since the random variables \( R^1_j(x) - L^1_j(x), 1 \leq j \leq d, \) all have the same distribution (by symmetry of the definition of the Mondrian process with respect to the dimension), it suffices to consider \( D^1(x) := R^1(x) - L^1(x) \).

Consider the segment \( I^1 = [0, 1] \times \{(x_j)_{2 \leq j \leq d} \} \simeq [0, 1] \) (through the natural identification) containing \( x = (x_1, 1 \leq j \leq d) \), and denote \( \Phi^1_\lambda(x) \subset [0, 1] \) the restriction of \( M_\lambda \) to \( I^1 \). Note that \( R^1_j(x) \) (resp. \( L^1_j(x) \)) is the lowest element of \( \Phi^1_\lambda(x) \) that is larger than \( x_1 \) (resp. the highest element of \( \Phi^1_\lambda(x) \) that is smaller than \( x_1 \)), and is equal to 1 (resp. 0) if \( \Phi^1_\lambda(x) \cap [x_1, 1] \) (resp. \( \Phi^1_\lambda(x) \cap [0, x_1] \)) is empty. By the facts 1 and 2, \( \Phi^1_\lambda(x) \) is a Poisson point process of intensity \( \lambda \).

Now, note that the characterization of \( L^1(x) \) and \( R^1(x) \) in terms of \( \Phi^1_\lambda(x) \) (a Poisson process on \([0, 1]\)) implies the following: the distribution of \((L^1_j(x), R^1_j(x))\) is the same as that of \((\tilde{L}^1_1(x), \tilde{R}^1_1(x))\), where \( \tilde{\Phi}^1_\lambda(x) \) is a Poisson process on \( \mathbb{R} \) of intensity \( \lambda \), and \( \tilde{L}^1_1(x) = \sup(\tilde{\Phi}^1_\lambda(x)) \cap (-\infty, x_1) \), \( \tilde{R}^1_1(x) = \inf(\tilde{\Phi}^1_\lambda(x)) \cap [x_1, +\infty) \). By the properties of the Poisson point process, this implies that \( \tilde{R}^1_1(x) - x_1 \sim \tilde{L}^1_1(x) \) are independent exponential random variables with parameter \( \lambda \). In particular, \( D^1(x) = \tilde{R}^1_1(x) - x_1 + x_1 - \tilde{L}^1_1(x) \) is stochastically upper bounded by \( E_1 + E_2 \sim \Gamma(2, \lambda) \), so that we have for every \( \delta > 0 \):

\[
\mathbb{P}(D^1(x) \geq \delta) \leq (1 + \lambda \delta) e^{-\lambda \delta} \tag{6}
\]

(with equality if \( \delta \leq x_1 \wedge (1 - x_1) \)), and \( \mathbb{E}[D^1_1(x)]^2 \leq \mathbb{E}(E_1^2) + \mathbb{E}(E_2^2) = \frac{4}{\pi} \). Finally, the bound (2) for the diameter \( D_\lambda(x) = \sqrt{\sum_{j=1}^d D^2_j(x)} \) follows from the observation that \( \mathbb{P}(D_\lambda(x) \geq \delta) \leq \mathbb{P}(\exists : D^1_j(x) \geq \frac{\delta}{\sqrt{2}}) \leq d \mathbb{P}(D^1_1(x) \geq \frac{\delta}{\sqrt{2}}) \) and inequality (6); the bound (3) is obtained by noting that \( \mathbb{E}[D_\lambda(x)]^2 \leq d \mathbb{E}[D^1_1(x)]^2 \leq \frac{4d}{\pi} \).

B Proof of Lemma 2: number of splits

Proof. Let \( A \subset \mathbb{R}^d \) be an arbitrary box, and let \( K^A_\lambda \) denote the number of splits performed by \( M_\lambda \sim \text{MP}(\lambda, A) \). As shown in the proof of Proposition 3 in [BLG+16], since the time until a leaf \( \phi \) is split follows an exponential distribution of rate \( |A_\phi| \leq |A| \) (independently of the other leaves), the number of leaves \( K_{\phi} + 1 \geq K_{\phi} \) at time \( t \) is dominated by the number of individuals in a Yule process with rate \( |A| \), which gives the first estimate

\[
\mathbb{E}(K^A_\lambda) \leq \exp(\Lambda |A|) \tag{7}
\]
This bound can be refined to the correct order of magnitude in λ in the following way. Consider the covering C of A by a regular grid of ⌈λ⌉d boxes obtained by dividing each coordinate of A in ⌈λ⌉. Since each split of A induces a split in at least one box C ∈ C (i.e. a split in the restriction MλC of MλA to C), and since MλC ∼ MP(λ, C) by Fact 1,
\[ \mathbb{E}(K^A_λ) \leq \sum_{C \in \mathcal{C}} \mathbb{E}(K^C_λ) \leq \lceil λ \rceil d \exp \left( \frac{\lceil |A| \rceil}{|A|} \right) \leq (λ + 1)^d \exp(|A|) \] (8)
where in the inequality (*) we applied the bound (7) to every cell C ∈ C (and the fact that |C| = |A|/⌈λ⌉). The bound of Lemma 2 follows by taking A = [0, 1]^d in (8).

**C  Proof of Proposition 1: original Mondrian Forests are inconsistent**

In this appendix, we show that Mondrian Forests with fixed lifetime λ are inconsistent, as stated in Proposition 1. We establish that this is true both for the variant based on the full domain [0, 1]^d, and for the original Mondrian Forests algorithm [LRT14] that restricts to the range of training data.

**C.1 Reduction to the full domain**

First, we begin by showing that, asymptotically, there is little difference between Mondrian trees constructed on the full domain and those restricted to the range of the training data. This is due to the fact that, as the sample size n grows large, the training data will span the whole domain, as well as every cell contained in it.

**Lemma 3.** Assume the distribution µ of X satisfies: µ(A) ≥ α vol(A) for every measurable A ⊆ [0, 1]^d, for some α ∈ (0, 1]. Fix λ > 0. For every n ≥ 1, there exists a couple (Mλ, Mλ\text{range}(n)) such that Mλ ∼ MP(λ, [0, 1]^d), Mλ\text{range}(n) is a Mondrian partition with parameter λ restricted to the range defined by the data points X1, ..., Xn, and P(Mλ = Mλ\text{range}(n)) → 1 as n → ∞.

**Proof of Lemma 3.** Let Mλ ∼ MP(λ, [0, 1]^d) be sampled by the procedure SampleMondrian (Algorithm 1). We will define explicitly each Mλ\text{range}(n) so that they have the desired distribution, and agree with Mλ on an event of high probability.

First, consider the event Ωn that all splits of Mλ occur inside the range defined by the feature points among X1, ..., Xn that belong to the cell to be split. We will show that P(Ωn) → 1 as n → ∞. Since the tree Mλ is grown independently of (X1, ..., Xn), we may reason conditionally on Mλ, and (X1, ..., Xn) remains distributed as µ\text{range}n. Note that Ωn is equivalent to the following: no leaf cell of Mλ contain no points among X1, ..., Xn. We can now write, denoting Ωn the complementary of Ωn,
\[ P(Ω^c_n | Mλ) = P(∃φ \in \mathcal{L}(Mλ) : A_φ \cap \{X_1, \ldots, X_n\} = \emptyset) \]
\[ \leq \sum_{φ \in \mathcal{L}(Mλ)} P(A_φ \cap \{X_1, \ldots, X_n\} = \emptyset) \]
\[ = \sum_{φ \in \mathcal{L}(Mλ)} (1 - μ(A_φ))^n \]
\[ \leq \sum_{φ \in \mathcal{L}(Mλ)} (1 - α \text{vol}(A_φ))^n \]
\[ \rightarrow_0 0 \quad \text{a.s.} \quad (9) \]
where equation (9) used the hypothesis μ ≥ α vol, and the convergence (10) is almost sure with respect to Mλ, since a.s. vol(A_φ) > 0 for every φ ∈ L(Mλ). By the dominated convergence theorem (since each random variable P(Ω^c_n | Mλ), n ≥ 1, is dominated by 1), we have P(Ωn) = E[P(Ω^c_n | Mλ)] → 0 as n → ∞.

For every n ≥ 1, we define Mλ\text{range}(n) as follows: on Ω^c_n, we let Mλ\text{range}(n) be a random Mondrian partition of lifetime λ, on the range defined by the data points X1, ..., Xn. On Ωn, we take
$M^{\text{range}(n)}_{\lambda}$ to be a pruning of $M_{\lambda}$. Specifically, for $\eta \in \mathcal{N}(M_{\lambda})$, denote $E_{\eta} = E_{A_{\eta}} \sim \text{Exp}(|A_{\eta}|)$ the exponential random variables drawn during the construction of $M_{\lambda}$ (see Algorithm 1). Now, set $E_{\eta}^{\text{range}(n)} := \frac{|A_{\eta}|}{|A_{\eta}^{\text{range}(n)}|} E_{\eta} \sim \text{Exp}(|A_{\eta}^{\text{range}(n)}|)$, and $r_{\eta}^{\text{range}(n)} := \sum_{\eta'} E_{\eta'}^{\text{range}(n)}$, where the sum spans over the (strict) ancestors $\eta' \in \mathcal{N}(M_{\lambda})$ of $\eta$. Finally, we define $M^{\text{range}(n)}_{\lambda}$ on $\Omega_{n}$ to be equal to the pruning of $M_{\lambda}$ obtained by keeping only the nodes $\mathcal{N}$ such that $r_{\eta}^{\text{range}(n)} \leq \lambda$. By construction, $M^{\text{range}(n)}_{\lambda}$ has the distribution of a Mondrian process of parameter $\lambda$ restricted to the range of the data $X_{1}, \ldots, X_{n}$.

It remains to show that $\mathbb{P}(M^{\text{range}(n)}_{\lambda} = M_{\lambda}) \to 1$. Since we already proved that $\mathbb{P}(\Omega_{n}) \to 1$, it suffices to show that $\mathbb{P}(M^{\text{range}(n)}_{\lambda} = M_{\lambda} \mid \Omega_{n}) \to 1$.

Second, consider the random variable $\Delta_{n} = \sup_{\phi \in \mathcal{L}(M_{\lambda})} \frac{|A_{\phi}|}{|A_{\phi}^{\text{range}(n)}|} - 1 \geq 0$. By the same argument as above, but replacing the boxes $A_{\phi}$ ($\phi \in \mathcal{L}(M_{\lambda})$) by interior cubes of size $\varepsilon$ around the edges of the cells $A_{\eta}$ ($\eta \in \mathcal{N}(M_{\lambda})$), we see that $\Delta_{n} \to 0$ in probability as $n \to \infty$. Since a.s. $\tau_{\phi} < \lambda$ and $r_{\phi}^{\text{range}(n)} \leq (1 + \Delta_{n}) \tau_{\phi}$ for every $\phi \in \mathcal{L}(M_{\lambda})$, we have $\mathbb{P}(M^{\text{range}(n)}_{\lambda} = M_{\lambda} \mid \Omega_{n}) \to 1$, which concludes the proof.

\section*{C.2 A simple example for fixed lifetime and range}

In order to establish Proposition 1, it remains to provide a simple counter-example that proves the inconsistency of the Mondrian Forest algorithm for a fixed range and lifetime.

\textbf{Proof.} Fix $\lambda > 0$, and let $\varepsilon \in (0, \frac{1}{2})$ to be specified later. Let $X$ be uniformly distributed on $[0, 1]$; we set $Y = 1$ if $|X - \frac{1}{2}| \leq \varepsilon$, and 0 otherwise. Clearly, we have $L^{*} = 0$.

Denote $g_{\lambda,n}^{(K)}$ the classifier described in Algorithm 4 with $\lambda_{n} = \lambda$, trained on the dataset $((X_{1}, Y_{1}), \ldots, (X_{n}, Y_{n}))$, and denote $\hat{g}_{\lambda,n}^{(K)}$ the corresponding estimate of the conditional probability $\eta$. Also, let $M_{\lambda} \sim \text{MP}(\lambda, [0, 1]^{d})$ and denote $A_{\lambda}(x) \subset [0, 1]$ the cell of $x \in [0, 1]$, as well as

$$N_{\lambda,n}(x) := \sum_{i=1}^{n} 1_{\{X_{i} \in A_{\lambda}(x)\}}, \quad \hat{\eta}_{\lambda,n}(x) := \frac{1}{N_{\lambda,n}(x)} \sum_{i=1}^{n} Y_{i} \cdot 1_{\{X_{i} \in A_{\lambda}(x)\}}$$

(with $\hat{\eta}_{\lambda,n}(x) = 0$ if $N_{\lambda,n}(x) = 0$) and $g_{\lambda,n}(x) := 1_{\{\hat{\eta}_{\lambda,n}(x) \geq \frac{1}{2}\}}$. For each $x \in \left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$, we have

$$\mathbb{P}(\hat{g}_{\lambda,n}^{(K)}(x) = 1) = \mathbb{P}(\hat{\eta}_{\lambda,n}^{(K)}(x) \geq 1/2) \leq 2 \mathbb{E}[\hat{\eta}_{\lambda,n}^{(K)}(x)] = 2 \mathbb{E}[\hat{\eta}_{\lambda,n}(x)]$$

by Markov’s inequality and the fact the $K$ trees in the forest have the same distribution as $M_{\lambda}$. Now, conditionally on $A_{\lambda}(x)$ and on $N_{\lambda,n}(x) = N \geq 1$, the points among $X_{1}, \ldots, X_{n}$ that fall in $A_{\lambda}(x)$ are $N$ i.i.d. points drawn uniformly in the interval $A_{\lambda}(x)$, and $\hat{\eta}_{\lambda,n}(x)$ is just the fraction of those points that satisfy $|X_{i} - \frac{1}{2}| \leq \varepsilon$. In particular,

$$\mathbb{E}[\hat{\eta}_{\lambda,n}(x) \mid A_{\lambda}(x), N_{\lambda,n}(x) = N] = \frac{|A_{\lambda}(x) \cap [1/2 - \varepsilon, 1/2 + \varepsilon]|}{|A_{\lambda}(x)|} \leq \frac{2\varepsilon}{|A_{\lambda}(x)|}$$

so that

$$\mathbb{P}(\hat{g}_{\lambda,n}^{(K)}(x) = 1) \leq 2\varepsilon \mathbb{E}[|A_{\lambda}(x)|^{-1}].$$

(11)

Now, recall that $M_{\lambda}$ is a partition of $[0, 1]$ into subintervals whose endpoints form a Poisson point process of intensity $\lambda$ (Fact 2). In particular, a direct derivation shows that $\mathbb{E}[|A_{\lambda}(x)|^{-1}] \leq F(\lambda) := \lambda + 4\varepsilon e^{-\lambda/4} < +\infty$. Choosing $\varepsilon := \frac{1}{4} \wedge \frac{1}{4F(\lambda)}$ and using Equation (11), we get $\mathbb{P}(\hat{g}_{\lambda,n}^{(K)}(x) = 1) \leq \frac{1}{2}$. Finally, integrating over $X$, we get for each $n \geq 1$:

$$L(g_{n}^{(K)}) \geq \int_{1/2 - \varepsilon}^{1/2 + \varepsilon} \mathbb{P}(\hat{g}_{\lambda,n}^{(K)}(x) = 0) dx \geq \varepsilon > 0,$$

so that $L(g_{n}^{(K)})$ is bounded away from 0, as announced.
D Proof of Theorem 1: consistency for Mondrian forests

D.1 Some general consistency results

Let us recall two general consistency results that will be used in the proof. First, the consistency of Mondrian forests can be deduced from that of the individual trees, using Proposition 2.

**Proposition 2** (Proposition 1 in [BDL08]). If a sequence $(\hat{g}_n)_{n \geq 1}$ of randomized classifiers is consistent, then for each $K \geq 1$, the averaged classifier $\hat{g}_n^{(K)}$ is consistent.

Then, to establish the consistency of individual trees, we use the following consistency theorem for partitioning classifiers.

**Proposition 3** ([DGL96], Theorem 6.1). Consider a sequence of randomized tree classifiers $(\hat{g}_n(\cdot, Z))$, grown independently of the labels $Y_1, \ldots, Y_n$. For $x \in [0, 1]^d$, denote $A_n(x) = A_n(x, Z)$ the cell containing $x$, $\text{diam} A_n(x)$ its diameter, and $N_n(x) = N_n(x, Z)$ the number of input vectors among $X_1, \ldots, X_n$ that fall in $A_n(x)$. Assume that, if $X$ is drawn from the distribution $\mu$:

1. $\text{diam} A_n(x) \to 0$ in probability, as $n \to \infty$,
2. $N_n(X) \to \infty$ in probability, as $n \to \infty$.

Then, the tree classifier $\hat{g}_n$ is consistent.

D.2 Universal consistency

We will need Lemma 4 which states that the number of training observations in the cell of a point tends to infinity with $n$, if the number of splits is controlled.

**Lemma 4.** Assume that the total number of splits $K_{\lambda_n}$, performed by the Mondrian tree partition $M_{\lambda_n}$ satisfies $\mathbb{E}(K_{\lambda_n})/n \to 0$. Then, $N_n(X) \to \infty$ in probability.

**Proof.** The proof extends a result in [BDL08] to a random number of splits. We fix $n \geq 1$, and reason conditionally on $M_{\lambda_n}$, which is by construction independent of $\mathcal{D}_n$ and $X$. Note that the number of leaves is $|\mathcal{L}(M_{\lambda_n})| = K_{\lambda_n} + 1$, and let $(A_\phi)_{\phi \in \mathcal{L}(M_{\lambda_n})}$ be the corresponding cells. For $\phi \in \mathcal{L}(M_{\lambda_n})$ we define $N_\phi$ to be the number of points (with repetition) among $X_1, \ldots, X_n$ that fall in the cell $A_\phi$. Since $X_1, \ldots, X_n, X$ are i.i.d., so that the joint distribution of $(X_1, \ldots, X_n, X)$ is invariant under permutation of the $n + 1$ points, conditionally on the set $S = \{X_1, \ldots, X_n, X\}$ (and on $M_{\lambda_n}$) the probability that $X$ falls in the cell $A_\phi$ is $N_\phi / n + 1$. Therefore, for each $t > 0$,

$$
\mathbb{P}(N_n(X) \leq t) = \mathbb{E}\left\{\mathbb{P}(N_n(X) \leq t \mid S, M_{\lambda_n})\right\}
= \mathbb{E}\left\{\sum_{\phi \in \mathcal{L}(M_{\lambda_n}) : N_\phi \leq t} \frac{N_\phi}{n + 1}\right\}
\leq \mathbb{E}\left\{\frac{t|\mathcal{L}(M_{\lambda_n})|}{n + 1}\right\}
= \frac{t(\mathbb{E}(K_{\lambda_n}) + 1)}{n + 1},
$$

which tends to 0 as $n \to \infty$ by assumption. \hfill \square

**Proof of Theorem 1.** To prove the consistency of Mondrian forest with a lifetime sequence, we show that the two assumptions of Proposition 3 are satisfied, which proves Theorem 1 since our algorithm performs splits independently of the labels $Y_1, \ldots, Y_n$. First, Lemma 1 ensures that, if $\lambda_n \to \infty$, $D_{\lambda_n}(x) = \text{diam} A_{\lambda_n}(x) \to 0$ in probability for every $x \in [0, 1]^d$. In particular, for every $\delta > 0$, $\mathbb{P}(\text{diam} A_{\lambda_n}(X) \geq \delta) = \int_{[0, 1]^d} \mathbb{P}(\text{diam} A_{\lambda_n}(x) \geq \delta) \mu(dx) \to 0$ as $n \to \infty$ by the dominated convergence theorem. This establishes the first condition.

For the second condition, Lemma 2 implies that $\mathbb{E}(K_{\lambda_n})/n \leq e^{d(\lambda_n + 1)^d}/n \to 0$ by hypothesis. By Lemma 4, this establishes the second condition of Lemma 3, which concludes the proof. \hfill \square
E  Proof of Theorem 2: Minimax rates for Mondrian forests in regression

In this section, we demonstrate how the properties about Mondrian trees established in Lemmas 1 and 2 imply minimax rates over the class of Lipschitz regression function, in arbitrary dimension $d$. We consider the following regression problem

$$Y = f(X) + \epsilon,$$

where $X$ is a $[0,1]^d$-valued random variable, $\epsilon$ is a real-valued random variable such that $\mathbb{E}(\epsilon | X) = 0$ and $\text{Var}(\epsilon | X) \leq \sigma^2 < \infty$ a.s., and $f : [0,1]^d \to \mathbb{R}$ is $L$-Lipschitz. We assume to be given $n$ i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, distributed as $(X, Y)$. We draw $K$ i.i.d. Mondrian tree partitions $M^{(1)}_{\lambda_n}, \ldots, M^{(K)}_{\lambda_n}$, distributed as $\text{MP}(\lambda_n, [0,1]^d)$. For all $k = 1, \ldots, K$, we let $\tilde{f}_n^{(k)}(x)$ be the $k$th Mondrian tree estimate at $x$, that is the average\footnote{With the convention that if no training point $X_i$, $1 \leq i \leq n$, falls in $A_{\lambda_n}(x)$, then $\tilde{f}_n(x) := 0$.} of the labels $Y_i$ such that $X_i$ belongs to the cell containing $x$ in the partition $M_{\lambda_n}^{(k)}$. Finally, the Mondrian forest estimate at $x$ is given by

$$\tilde{f}_n = \frac{1}{K} \sum_{k=1}^K \tilde{f}_n^{(k)} : [0,1]^d \to \mathbb{R}.$$

**Proposition 4.** The quadratic risk $R(\tilde{f}_n) = \mathbb{E}(f(X) - f(X))^2$ of $\tilde{f}_n$ is upper bounded as follows:

$$R(\tilde{f}_n) \leq \frac{4dL^2}{\lambda^2_n} + \frac{1 + e^d(1 + \lambda_n)^d}{n} (2\sigma^2 + 9\|f\|_{\infty}).$$  \hspace{1cm}(13)

In particular, the choice $\lambda_n = n^{1/(d+2)}$ yields a risk rate $R(\tilde{f}_n) = O(n^{-2/(d+2)})$.

**Proof.** First, by the convexity of the function $y \mapsto (y - f(x))^2$ for any $x \in [0,1]^d$, we have $R(\hat{f}_n) \leq \frac{1}{K} \sum_{k=1}^K R(\hat{f}_n^{(k)}) = R(\hat{f}_n^{(1)})$ since the random trees classifiers have the same distribution. Hence, it suffices to prove the risk bound (13) for a single tree; in the following, we assume that $K = 1$, and consider the random estimator $\hat{f}_n$ associated to a tree partition $M_{\lambda_n} \sim \text{MP}(\lambda_n, [0,1]^d)$.

Since the splits of the tree partition $M_{\lambda_n}$ are performed independently of the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we can write the following bias-variance decomposition of the risk for purely random forests, first noticed by [Gen12]:

$$R(\hat{f}_n) = \mathbb{E}(f(X) - \hat{f}_{\lambda_n}(X))^2 + \mathbb{E}(\hat{f}_{\lambda_n}(X) - \hat{f}_{\lambda_n}(X))^2,$$  \hspace{1cm}(14)

where we denoted $\hat{f}_{\lambda_n}(x) := \mathbb{E}(f(X)|X \in A_{\lambda_n}(x))$ (which only depends on the random partition $M_{\lambda_n}$) for every $x$ in the support of $\mu$. The first term of the sum, the *bias*, measures how close $\hat{f}$ is to its best approximation $\hat{f}_n$ that is constant on the leaves of $M_{\lambda_n}$ (on average over $M_{\lambda_n}$). The second term (the *variance*) measures how well the expected value $\hat{f}_n(x) = \mathbb{E}(Y|X \in A_{\lambda_n}(x))$ (i.e. the optimal label on the leaf $A_{\lambda_n}(x)$) is estimated by the empirical average $\hat{f}_n(x)$, averaged over the sample $\mathcal{S}_n$ and the partition $M_{\lambda_n}$.

The bias term is bounded as follows: for each $x \in [0,1]^d$ in the support of $\mu$, we have

$$|f(x) - \hat{f}_n(x)| = \frac{1}{\mu(A_{\lambda_n}(x))} \left| \int_{A_{\lambda_n}(x)} (f(x) - f(z)) \mu(dz) \right| \leq \sup_{z \in A_{\lambda_n}(x)} |f(x) - f(z)| \leq L \sup_{z \in A_{\lambda_n}(x)} \|x - z\|_2 \leq LD_{\lambda_n}(x),$$  \hspace{1cm}(15)

where $D_{\lambda_n}(x)$ is the $\ell^2$-diameter of $A_{\lambda_n}(x)$; note that inequality (15) used the assumption that $f$ is $L$-Lipschitz. By Lemma 1, this implies

$$\mathbb{E}(f(x) - \hat{f}_n(x))^2 \leq L^2 \mathbb{E}[D_{\lambda_n}(x)²] \leq \frac{4dL^2}{\lambda^2_n}. \hspace{1cm}(16)$$
Integrating the bound (16) with respect to \( \mu \) yields the following bound on the integrated bias:

\[
\mathbb{E}(f(X) - \tilde{f}_n(X))^2 \leq \frac{4dL^2}{\lambda_n^2}.
\] (17)

In order to bound the variance term, we use the following fact ([AG14], Proposition 2): if \( U \) is a random tree partition of the unit cube in \( k + 1 \) cells (with \( k \in \mathbb{N} \) deterministic) formed independently of the training data \( \mathcal{D}_n \), we have

\[
\mathbb{E}(\tilde{f}_U(X) - \hat{f}_U(X))^2 \leq \frac{k + 1}{n} \left( 2\sigma^2 + 9\|f\|_{\infty} \right).
\] (18)

For every \( k \in \mathbb{N} \), applying the upper bound (18) to the random partition \( M_\lambda n \sim \text{MP}(\lambda_n; [0, 1]^d) \) conditionally on the event \( \{K_\lambda n = k\} \) (where \( K_\lambda n \) denotes the number of splits performed by \( M_\lambda n \)), and summing over \( k \), we get

\[
\mathbb{E}(\tilde{f}_\lambda n(X) - \hat{f}_\lambda n(X))^2 \leq \sum_{k=0}^{\infty} \mathbb{P}(K_\lambda n = k) \mathbb{E}(\tilde{f}_\lambda n(X) - \hat{f}_\lambda n(X))^2 | K_\lambda n = k)
\]

\[
\leq \sum_{k=0}^{\infty} \mathbb{P}(K_\lambda n = k) \frac{k + 1}{n} \left( 2\sigma^2 + 9\|f\|_{\infty} \right)
\]

\[
= \frac{1 + \mathbb{E}(K_\lambda n)}{n} \left( 2\sigma^2 + 9\|f\|_{\infty} \right).
\]

Then, applying Lemma 2 gives an upper bound of the variance term:

\[
\mathbb{E}(\tilde{f}_\lambda n(X) - \hat{f}_\lambda n(X))^2 \leq \frac{1 + e^d(1 + \lambda_n)^d}{n} \left( 2\sigma^2 + 9\|f\|_{\infty} \right).
\] (19)

Combining the bounds (17) and (19) with the decomposition (14) yields the desired bound (13). \( \Box \)

References


