Agrégation d'hold-out

Guillaume Maillard, M.Lerasle, S.Arlot

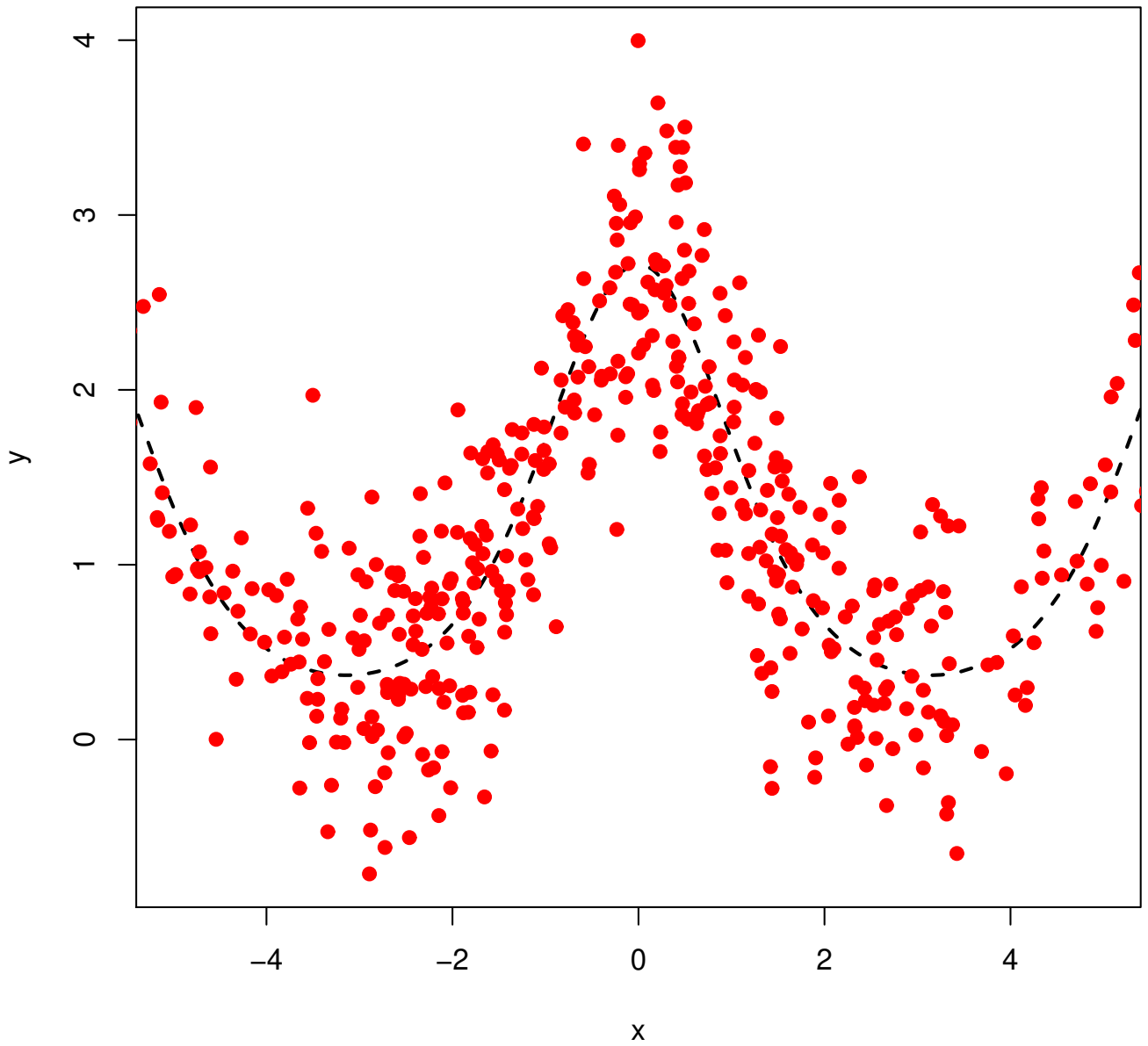
Université Paris Sud

June 20, 2019

Plan

- Presentation of Agghoo
- Theoretical performance
- Simulations in regression
- Classification

1.Presentation of Agghoo: Regression



1.Presentation of Agghoo - prediction

$\mathcal{X}, \mathcal{Y}$ are measurable spaces, P a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\gamma : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a measurable *contrast function*. For ex:

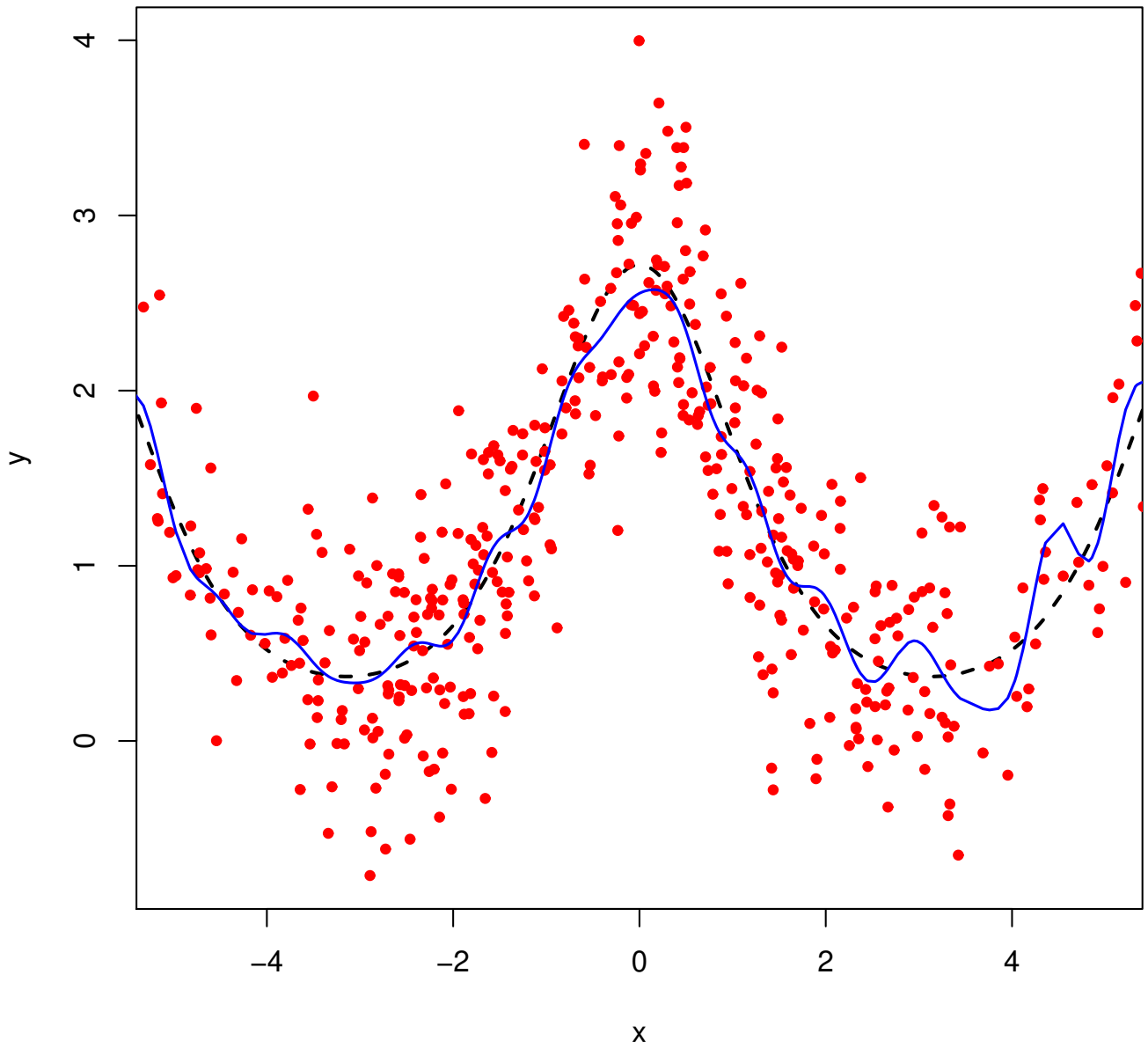
- $\mathcal{Y} = \{-1; 1\}$ and $\gamma(u, y) = \mathbb{I}_{u \neq y}$ in classification
- $\mathcal{Y} = \mathbb{R}$ and $\gamma(u, y) = |y - u|$ in median regression.

Definition

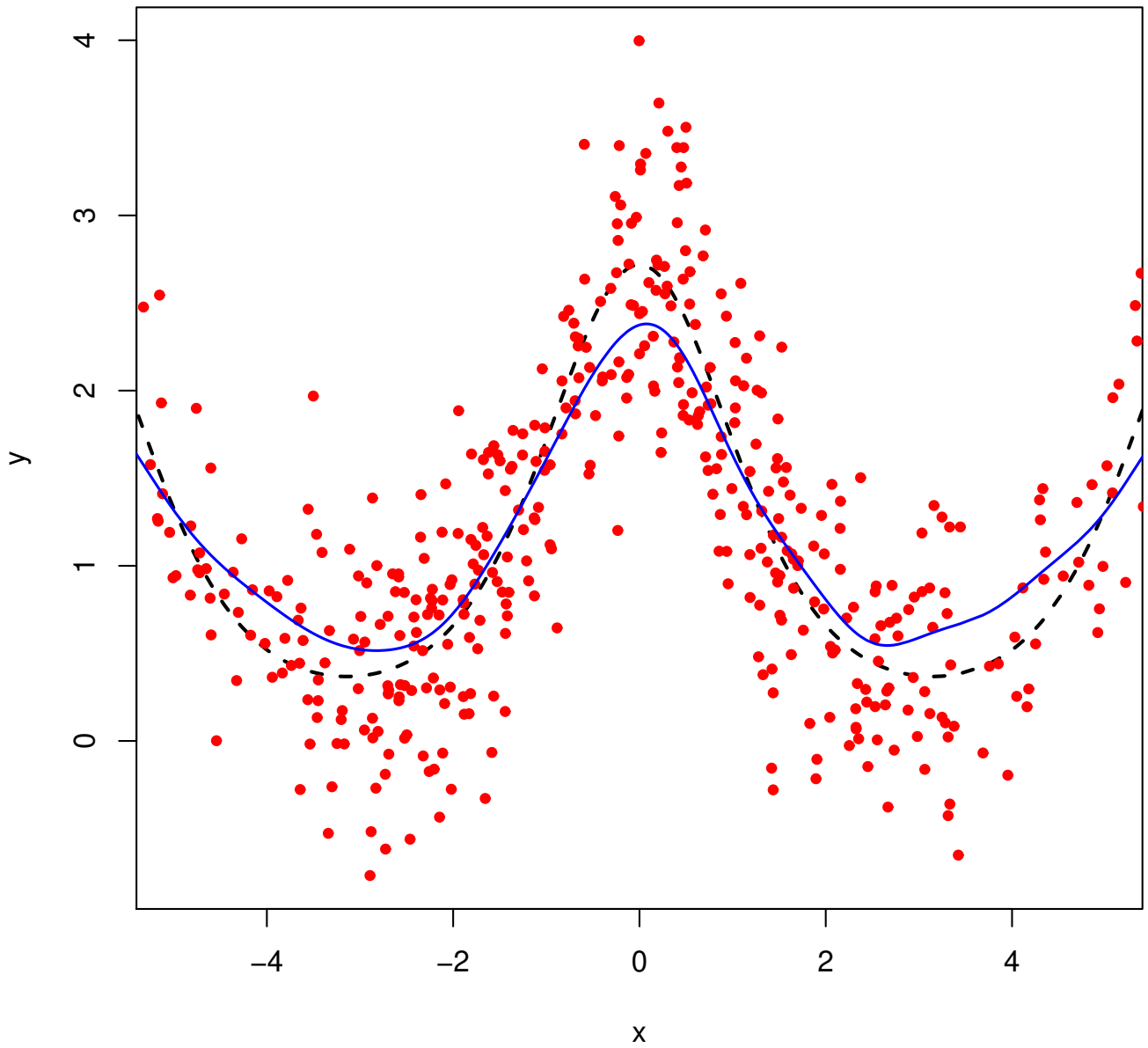
- A *predictor* is a measurable function $t : \mathcal{X} \to \mathcal{Y}$.
- The *excess risk* of a predictor f is defined by

$$\ell(s, t) = \mathbb{E}\left[\gamma(f(X), Y)\right] - \inf_{g \text{ predictor}} \mathbb{E}\left[\gamma(g(X), Y)\right]$$

1.Presentation of Agghoo: example estimator



1.Presentation of Agghoo: example estimator



1.Presentation of Agghoo: notations

Fix an integer n. Denote

Definition

- $D_n = (X_i, Y_i)_{1 \leq i \leq n} \sim P^{\otimes n}$ a *sample* of size n.
- For a sample D_n and $T \subset \{1, \dots, n\}$, $D_n^T = (X_j, Y_j)_{j \in T}$.
- $\mathcal{A} : \cup_{k \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^k \to$ predictors a *learning rule*.

1.Presentation of Agghoo: hyperparameter selection

- Given learning rules $(\mathcal{A}_m)_{m\in\mathcal{M}}$ and a sample D_n, choose parameter m.
- Optimal choice m_* realizes $\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_n))$.

Examples:

- $m = k$ in $k-$NN for classification.
- regularization parameter, $m = \lambda$ (Lasso, Ridge, SVM...)

Definition

Fix a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$. Let $T \subset \{1, \dots, n\}$. For any learning rule $\mathcal{A}$,

$$\mathrm{HO}_T(\mathcal{A}) = \frac{1}{|T^c|} \sum_{j \notin T} \gamma\left(\mathcal{A}(D_n^T)(X_j), Y_j\right).$$

$1 \leq p \leq n$, $\mathcal{T} \subset \{T \subset \{1, ..., n\} : |T| = p\}$

$$CV_{\mathcal{T}}(\mathcal{A}) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \mathrm{HO}_T(\mathcal{A})$$

1.Presentation of Agghoo: CV selection

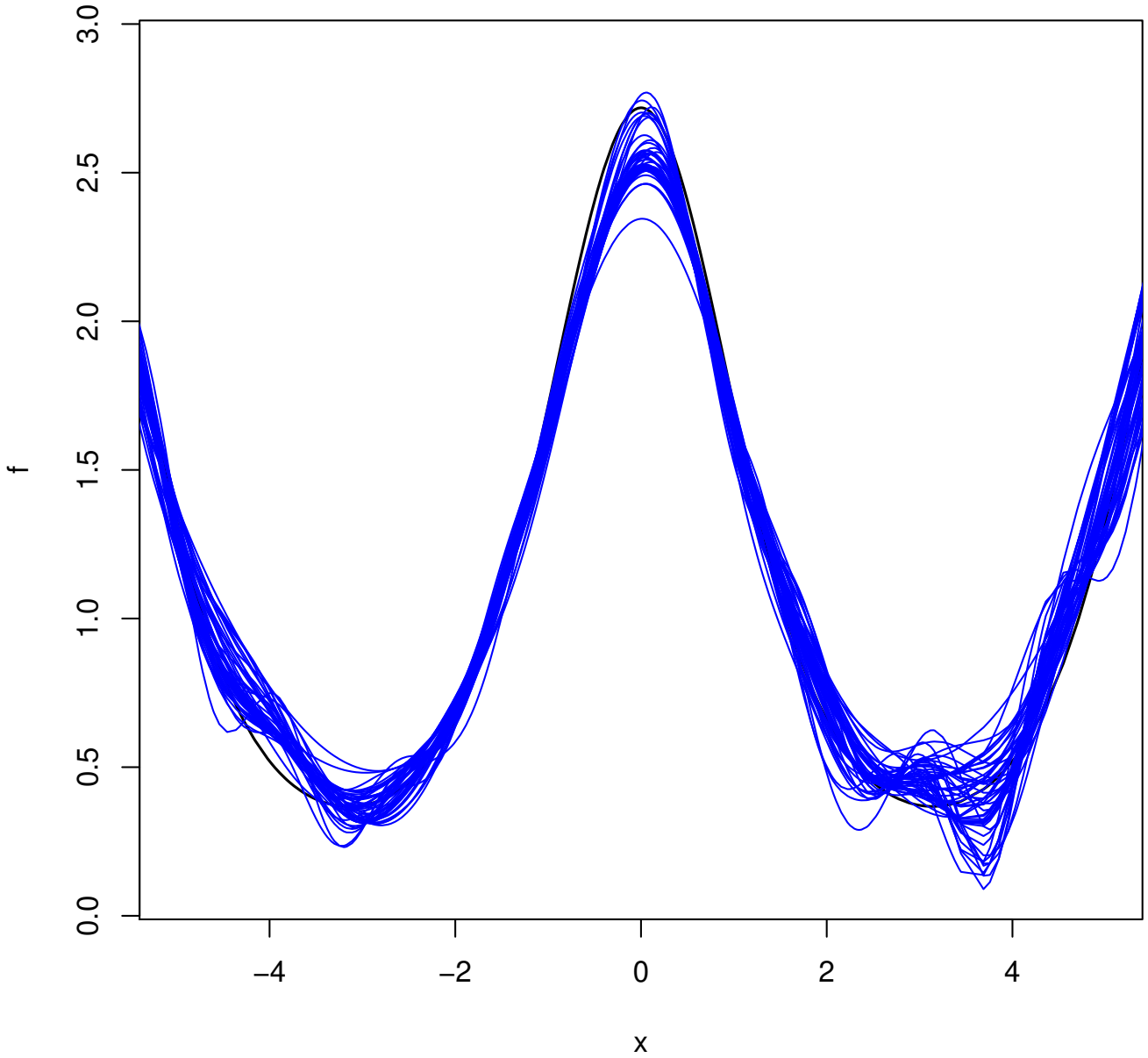
Idea: estimate $\ell(s, \mathcal{A}_m(D))$ and optimize in m.

Definition

Let $\mathcal{M}$ be a set of hyperparameters.

- A *hold out predictor* is $\widehat{f}_T^{\mathrm{ho}} = \mathcal{A}_{\hat{m}_T^{ho}}(D_n^T)$ where $\hat{m}_T^{ho} = \mathrm{argmin}_{m \in \mathcal{M}} \mathrm{HO}_T(\mathcal{A}_m)$.
- A *CV predictor* is defined by $\widehat{f}_{\mathcal{T}}^{CV} = \mathcal{A}_{\hat{m}_{\mathcal{T}}^{CV}}(D_n)$ where $\hat{m}_{\mathcal{T}}^{CV} = \mathrm{argmin}_{m \in \mathcal{M}} CV_{\mathcal{T}}(\mathcal{A}_m)$

1.Presentation of Agghoo: variability of the hold out



Idea: *aggregate* several hold out estimators.

Definition

Assume $\mathcal{Y}$ is convex (regression...).
$T_1, ..., T_V$ i.i.d $\sim \mathcal{U}(\{T \subset \{1, .., n\} : |T| = \lfloor \tau n \rfloor\})$.

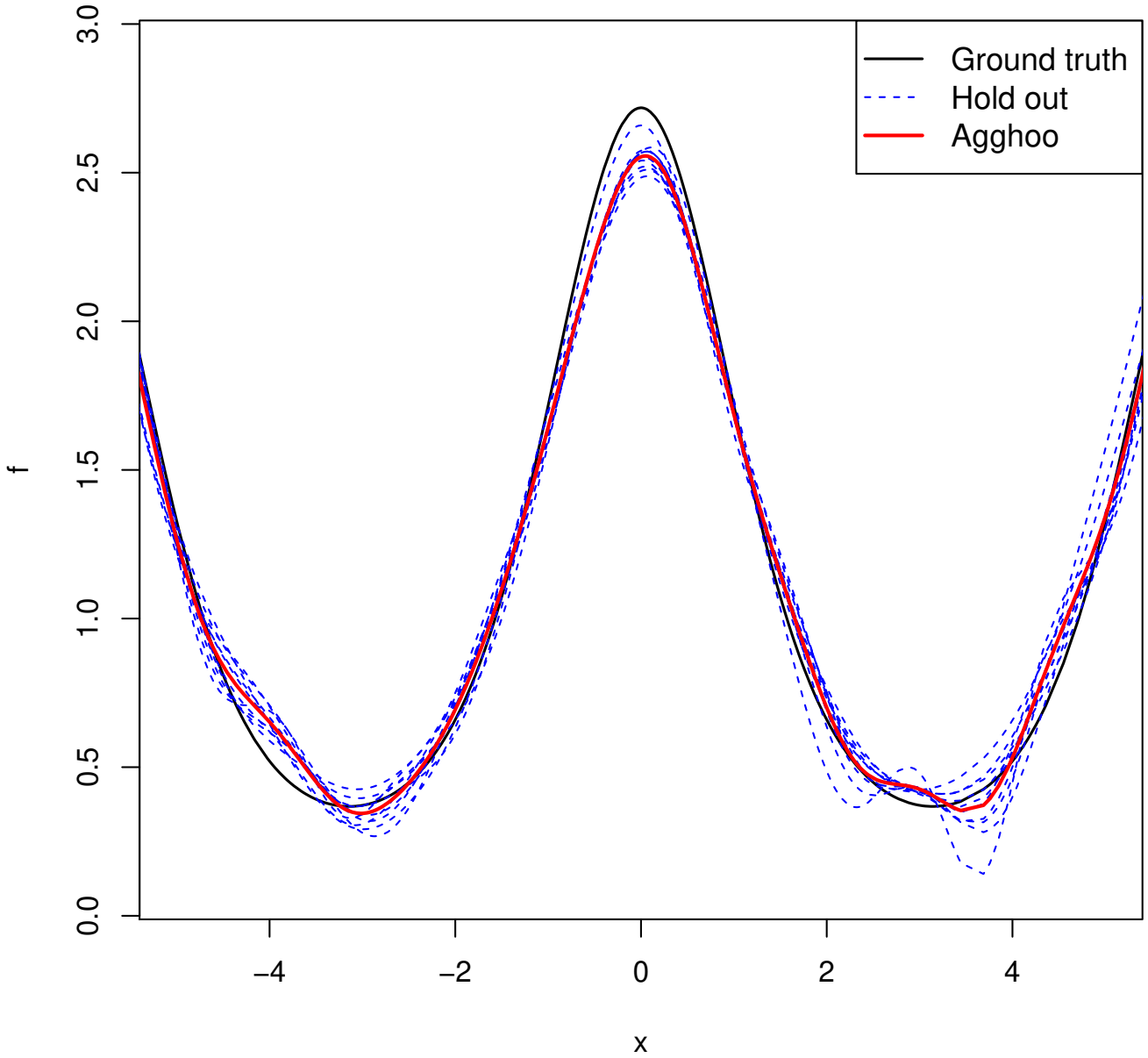
$$\widehat{f}^{\mathrm{ag}}_{\tau,V} = \frac{1}{V} \sum_{i=1}^{V} \widehat{f}^{\mathrm{ho}}_{T_i}$$

Parameters are V and τ.

V-fold CV $\Longrightarrow$ V-fold aggregation.

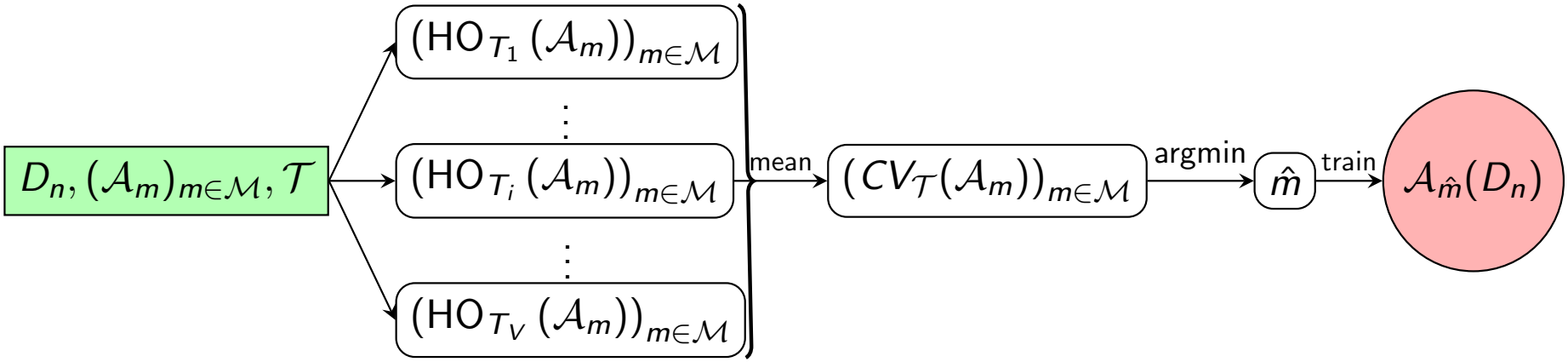
$CV_{\mathcal{T}} \Longrightarrow \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$.

1.Presentation of Agghoo: an example

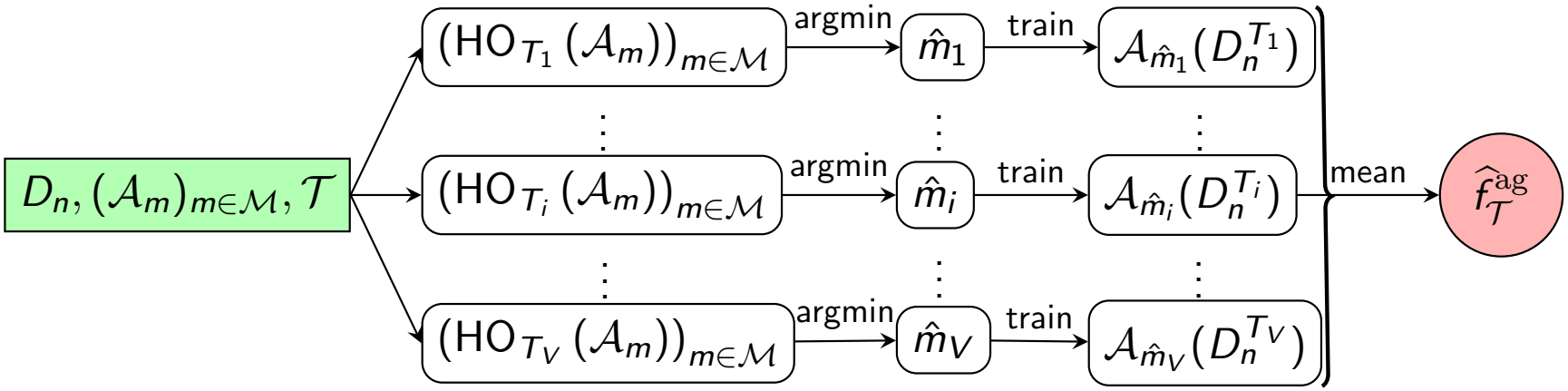


1.Presentation of Agghoo: comparison with cross-validation

Cross-validation:



Agghoo:



1.Presentation of Agghoo: comparison with subagging

Let $\xi_i = (X_i, Y_i)$. For $T \in \mathcal{T}$,

Subbagging $\underbrace{\underbrace{\xi_{i_1}, \dots, \xi_{i_k}}_{T' \text{ train}}, \underbrace{\xi_{i_{k+1}}, \dots, \xi_{i_l}}_{T \setminus T' \text{ val}}}_{T}, \underbrace{\xi_{i_{l+1}}, \dots, \xi_{i_n}}_{T^c}$ D^{T^c} is unused

Agghoo $\underbrace{\xi_{i_1}, \dots, \xi_{i_k}, \xi_{i_{k+1}}, \dots, \xi_{i_l}}_{T \text{ train}}, \underbrace{\xi_{i_{l+1}}, \dots, \xi_{i_n}}_{T^c \text{ val}}$ D^{T^c} is used

2.Theoretical performance: comparison with hold out

Agghoo is safe:

- $\widehat{f}^{\mathrm{ag}}_{\tau,1} = \widehat{f}^{\mathrm{ho}}_{T_1}$ is just the hold out.
- $\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})\right] \leq \mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})\right]$ if γ is convex.
- For any $k \geq 1$, $\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})^k\right] \leq \mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})^k\right]$ if γ is convex.

Oracle inequalities for the hold-out imply oracle inequalities for Agghoo.

Aim is to show that

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})\right] \leq C_n \mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_n^T))\right] + r_n$$

where ideally,

- $C_n \to 1$, r_n negligible.
- C_n, r_n only depend on $(\mathcal{A}_m)_{m\in\mathcal{M}}$ through $|\mathcal{M}|$(generality).
- Few assumptions are made on P.

2.Theoretical performance: known results on the hold out

- Generality in classification. Margin adaptivity (G.Blanchard, P.Massart, 2006)
- Least squares regression with bounded Y and linear models (F.Navarro and A.Saumard, 2017)
- Abstract theorems - how do you check the assumptions?
- Specific settings (ex: regressograms, S.Arlot, 2008)

See 2010 survey by S.Arlot and A.Celisse for more references.

A new setting: penalization hyperparameters.

2.Theoretical performance: an RKHS setting

For any sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$, let

$$\mathcal{A}_\lambda(D) = \underset{t \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n_t} c(t(X_i), Y_i) + \lambda \|t\|_{\mathcal{H}}^2 .$$

Penalized empirical risk minimization where

- The model is an RKHS $\mathcal{H}$
- The penalty is $\lambda \|\cdot\|_{\mathcal{H}}^2$
- c is convex and Lipschitz continuous in its first argument

Examples: $c(u, y) = (1 - uy)_+$ (SVM), $c(u, y) = (|u - y| - \varepsilon)_+$ (SVMR).

2.Theoretical performance: an oracle inequality

Now assume that:

- The kernel K is bounded.
- $c(u, y) = \gamma(u, y) = |u - y|$

Theorem

Assume that for a.e. $x \in \mathcal{X}$, for all $u \in \mathbb{R}$,

$$\left|\mathbb{P}(Y \leq u | X = x) - \mathbb{P}(Y \leq s(x) | X = x)\right| \geq \frac{|u - s(x)|}{\delta} \mathbb{I}_{|u - s(x)| \leq \frac{\delta}{4}}.$$

Let $\Lambda \subset \mathbb{R}_+$ and $\lambda_m = \min \Lambda$. Then for all $\theta \in (0; 1]$,

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})] \leq (1+\theta)\mathbb{E}\Big[\min_{\lambda \in \Lambda} \ell(s, \mathcal{A}_\lambda(D_{n_t}))\Big] + \frac{12\delta}{n_v} + \frac{C_3}{\lambda_m}\left[\frac{\log^2 n_v}{\theta^3 n_v^2} + \frac{\log^{\frac{3}{2}} |\Lambda|}{\theta n_v \sqrt{n_t}}\right]$$

where $n_t = \lfloor \tau n \rfloor$ and $n_v = n - n_t$.

From theory to practice

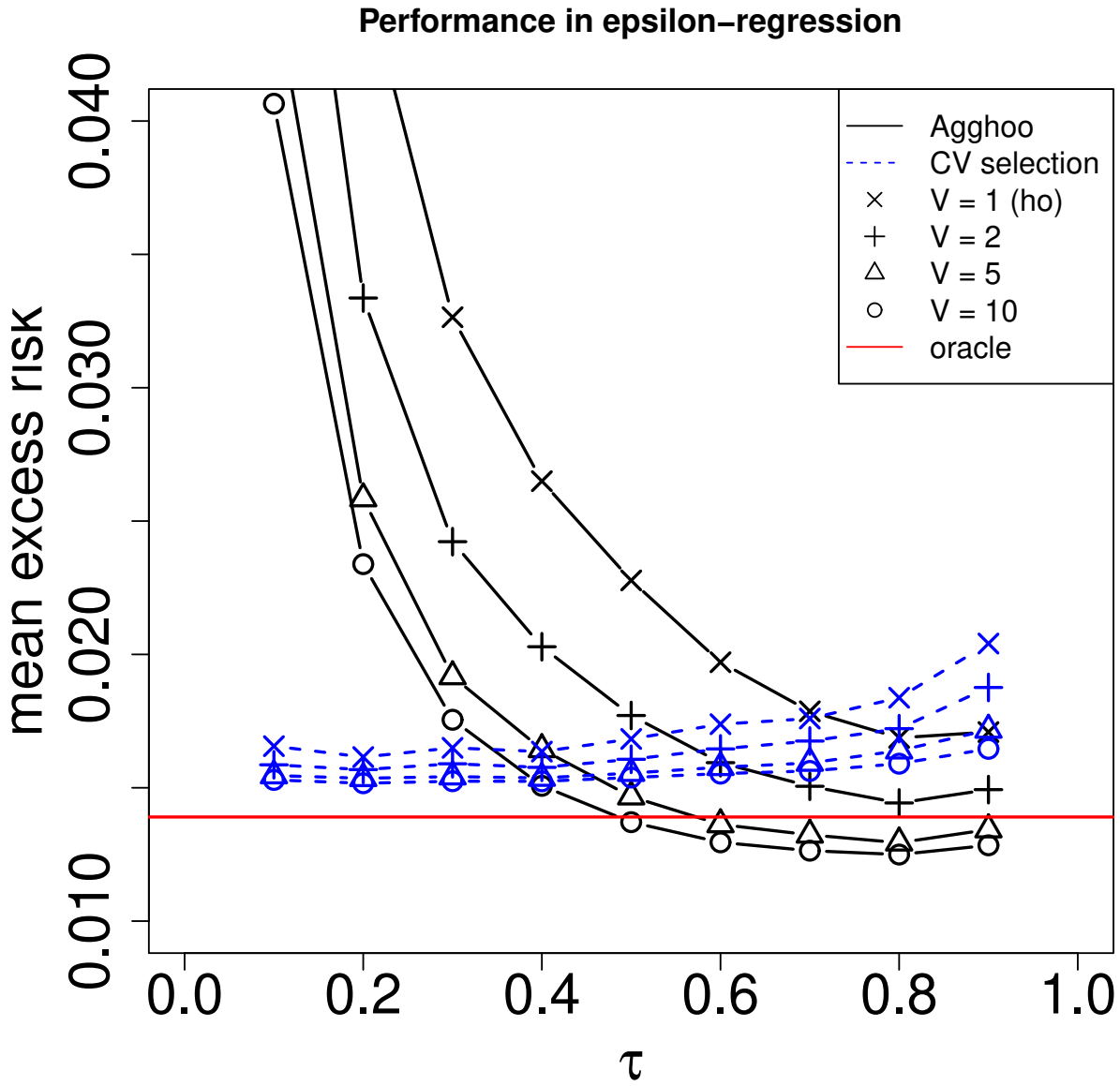
Questions not answered by theory:

- What is the effect of aggregation?
- Is Agghoo competitive with cross-validation?
- How should τ, V be chosen?

3. Simulations in regression: description

- Kernel rules $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ with Gaussian kernel $k(x,y) = e^{-2(x-y)^2}$, cost $c(u,y) = (|u-y| - 0.25)_+$ and grid $\Lambda = \left\{ \frac{2^{j-1}}{500\tau n} | 1 \leq j \leq 17 \right\}$
- $Y = e^{\cos(X)} + \xi$ where $X \sim \mathcal{N}(0,\pi)$, $\xi \sim \mathcal{N}\left(0, \frac{1}{2}\right)$.
- I.i.d sample of size $n = 500$.
- Risk evaluation using $\gamma(u,y) = |u-y|$.

3. Simulations in regression: Results



4.Classification: Majhoo

What can be done if $\mathcal{Y}$ is not convex? $\mathcal{Y} = \{0; 1\}$, $\gamma(u, y) = \mathbb{I}_{u \neq y}$.

Definition

Majority vote: Let $(f_i)_{i=1..V}$ be predictors.

$$\text{maj}\,(f_1, \ldots, f_V) = x \to \underset{y \in \{0;1\}}{\text{argmax}} \left|\{i : f_i(x) = y\}\right|$$

Majhoo: $\widehat{f}^{\text{ho}}_{T_1}, \ldots, \widehat{f}^{\text{ho}}_{T_v} \to \text{maj}\left(\widehat{f}^{\text{ho}}_{T_1}, \ldots, \widehat{f}^{\text{ho}}_{T_v}\right)$.

Comparison with hold-out: a factor 2 is lost:

$$\forall k \geq 1, \mathbb{E}\left[\ell(s, \widehat{f}_{\mathcal{T}}^{\mathrm{maj}})^k\right] \leq 2^k \mathbb{E}\left[\ell(s, \widehat{f}_{T_1}^{\mathrm{ho}})^k\right].$$

Using Massart (2007) yields for example:

Theorem

Let $\eta(x) = \mathbb{P}[Y = 1 | X = x]$ *Suppose that*

$$\forall h \in [0; 1], \mathbb{P}[|2\eta(X) - 1| \leq h] \leq c h^{\beta}$$

(margin hypothesis). Then

$$\mathbb{E}[\ell(s, \widehat{f}_{\tau,V}^{\mathrm{ag}})] \leq 3\mathbb{E}\left[\inf_{m \in \mathcal{M}} \ell(s, \mathcal{A}(D_n^{T_1}))\right] + \frac{29 c^{\frac{1}{\beta+2}} \log(e|\mathcal{M}|)}{(1-\tau)^{\frac{\beta+1}{\beta+2}} n^{\frac{\beta+1}{\beta+2}}}$$

4.Classification: simulation description

- Binary classification with $0-1$ loss $\gamma(u,y) = \mathbb{I}_{u \neq y}$.
- k-NN rules for $k \in \{k | 1 \leq k \leq n, k \text{ odd.}\}$
- I.i.d sample, $n = 500$, with distribution

$$X \sim \mathcal{U}([0;1]^2)$$

$$\mathbb{P}(Y = 1|X) = \sigma\left(\frac{g(X) - b}{\lambda}\right) \quad \text{where} \quad \sigma(u) = \frac{1}{1+e^{-u}},$$

$$g(u,v) = e^{-(u^2+v)^3} + u^2 + v^2$$

4.Classification: simulation results