Aggregated cross-validation

Guillaume Maillard

Université Paris Sud

October 23, 2017

Definition

Let $\mathcal{X}$ be a measure space.

- A *sample* is a finite collection $(x_i, y_i)_{1 \leq i \leq k}$, where $y_i \in \mathcal{Y} = \{0; 1\}$ and $x_i \in \mathcal{X}$.
- A *classifier* is a measurable function $f : \mathcal{X} \to \mathcal{Y}$
- A *learning rule* takes a sample D as input and produces a classifier $G(D) : \mathcal{X} \to \mathcal{Y}$.
- Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a r.v. The *excess risk* of a classifier f, denoted $\ell(f^*, f)$, equals

$$\mathbb{P}(f(X) \neq Y) - \inf_{g \text{ classifieur}} \mathbb{P}(g(X) \neq Y)$$

Cross-validation

Cross-Validation uses data to estimate the risk of a learning rule.

Definition

D_n a sample s.t $|D_n| = n$, $T \subset \{1, ..., n\}$.

$$\mathrm{HO}_T(G) = \frac{1}{|^c T|} |\{j \in^c T : G((x_i, y_i)_{i \in T})(x_j) \neq y_j\}|$$

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

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

Selection through Cross-Validation

Definition

Let $\mathcal{G}$ be a set of learning rules.

- The *hold out classifier* is $\widehat{f}_T^{\text{ho}} = \hat{G}_T^{ho}(D_n^T)$ where $\hat{G}_T^{ho} = \text{argmin}_{G \in \mathcal{G}} \text{HO}_T(G)$.
- A *cross-validated classifier* is defined as $\widehat{f}_{\mathcal{T}}^{CV} = \hat{G}_{\mathcal{T}}^{CV}(D_n)$ where $\hat{G}_{\mathcal{T}}^{CV} = \text{argmin}_{G \in \mathcal{G}} CV_{\mathcal{T}}(G)$

Aggregation and ensemble methods

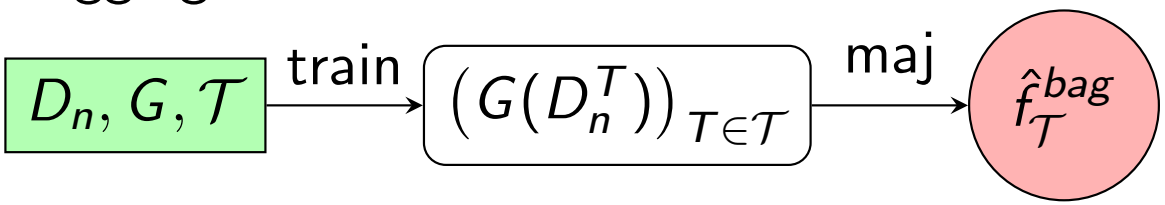
Many classifiers $\Rightarrow$ one prediction

Definition

Let $(f_i)_{i=1..V}$ be classifiers.

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

Bagging:



Aggregated cross-validation

The idea is to aggregate several hold-out classifiers.

- Base method: hold out (not retrained).
- The training set varies
- Compared with CV: selection and averaging switched.
- Compared with bagging: hold out computed from the whole sample.

A precise definition

Definition

$T_1, ..., T_V$ i.i.d $\sim \mathcal{U}(\{T \subset \{1, .., n\} : |T| = n - p\})$.

$$\widehat{f}_V^{\text{ag}} = \text{maj}\left((\widehat{f}_{T_i}^{\text{ho}})_{1 \leq i \leq V}\right)$$

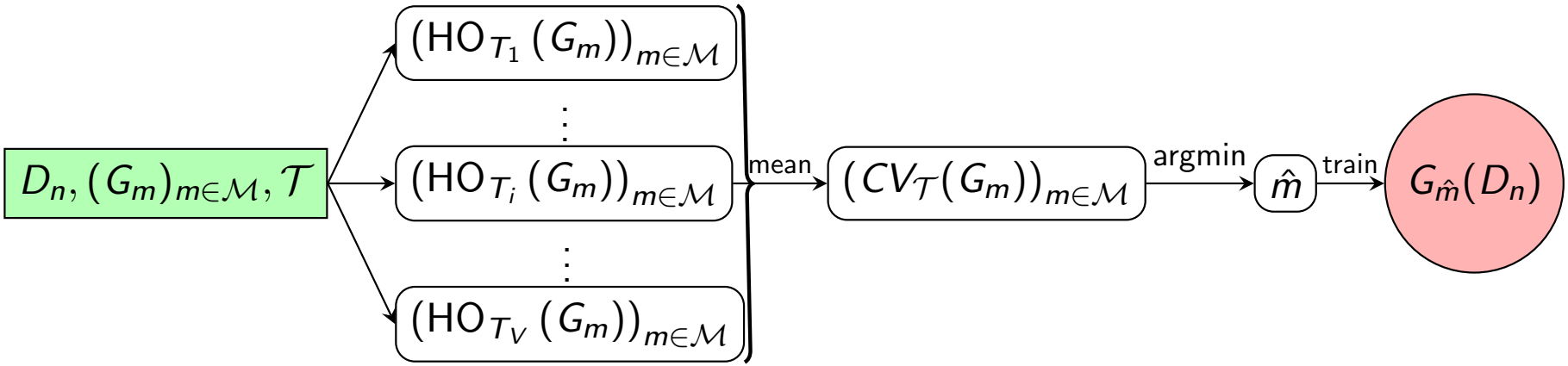
Parameters are V and $\tau = \frac{n-p}{n}$

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

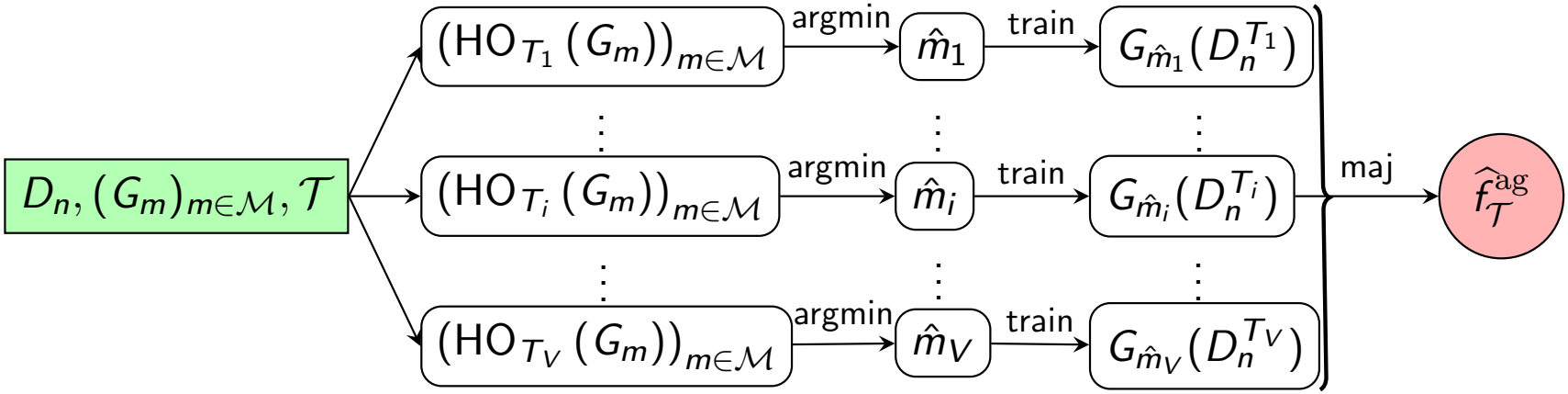
$CV_{\mathcal{T}} \Longrightarrow \widehat{f}_{\mathcal{T}}^{\text{ag}}$.

Comparison with cross-validation

Cross-validation:



Agghoo:



Comparison with subbagging

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

Subbagging $\underbrace{\underbrace{\xi_{i_1}, \ldots, \xi_{i_k}}_{T'}, \underbrace{\xi_{i_{k+1}}, \ldots, \xi_{i_l}}_{T'^c}}_{T}, \underbrace{\xi_{i_{l+1}}, \ldots, \xi_n}_{T^c}$ $D_n^{T^c}$ is unused

Agghoo $\underbrace{\xi_{i_1}, \ldots, \xi_{i_k}, \xi_{i_{k+1}}, \ldots, \xi_{i_l}}_{T}, \underbrace{\xi_{i_{l+1}}, \ldots, \xi_n}_{T^c}$ $D_n^{T^c}$ is used

Theoretical analysis: selection

Benchmark:

$$\inf_{G \in \mathcal{G}} \ell(f_*, G(D_n))$$

called *oracle*.
Hold out: oracle on D_n^T:

$$\inf_{G \in \mathcal{G}} \ell(f_*, G(D_n^T))$$

oracle inequality = bounding the risk by an affine function of the oracle.

Oracle performance: feasible?

Theorem

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

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

(margin hypothesis). Then for D_n *i.i.d* $\sim (X, Y)$,

$$\mathbb{E}[\ell(f_*, \widehat{f}_T^{\text{ho}})] \leq 1.5\mathbb{E}[\inf_{G \in \mathcal{G}} \ell(f_*, G(D_n^T))] + \frac{14.5c^{\frac{1}{\beta+2}} \log(e|\mathcal{G}|)}{p^{\frac{\beta+1}{\beta+2}}}$$

where $n - p = |T|$

Theoretical analysis: aggregation

Proposition

Let $(f_i)_{i=1\dots V}$ be classifiers, and $f^{maj} = maj(f_1, \dots, f_V)$. We obtain

$$\ell(f_*, f^{maj}) \leq \frac{2}{V} \sum_{i=1}^{V} \ell(f_*, f_i)$$

In particular if the classifiers f_i are equal in distribution,

$$\mathbb{E}[\ell(f_*, f^{maj})] \leq 2\mathbb{E}[\ell(f_*, f_1)]$$

Application to aggregated cross-validation

An oracle inequality for $\widehat{f}_V^{\mathrm{ag}}$:

$$\mathbb{E}[\ell(f_*, \widehat{f}_V^{\mathrm{ag}})] \leq 3\mathbb{E}[\inf_{G \in \mathcal{G}} \ell(f_*, G(D_n^T))] + \frac{29c^{\frac{1}{\beta+2}} \log(e|\mathcal{G}|)}{p^{\frac{\beta+1}{\beta+2}}}$$

Relevance:

- The oracle is *adaptative*.
- It is *simultaneously minimax* for well chosen $\mathcal{G}$.
- So is the hold-out if $p^{\frac{\beta+1}{\beta+2}} = o(\mathit{oracle})$.

Adaptation: an example

Definition

Let $\mathcal{P}_{\gamma,L,\beta,c}$ be the set of r.v (X, Y) such that:

- $\eta(x) = \mathbb{P}[Y = 1|X = x]$ is γ-Hölder with constant L.
- $$\forall h \in [0;1], \mathbb{P}(|2\eta(X) - 1| \leq h) \leq ch^{\beta}$$
- $\text{supp}(X) \subset [0;1]^d$.

The following result is due to Audibert et Tsybakov:

Theorem

The minimax risk over $\mathcal{P}_{\gamma,L,\beta,c}$ *is of order* $n^{-\frac{\gamma(1+\beta)}{\gamma(2+\beta)+d}}$.

Assessment of theoretical approach

- No such guarantee for cross-validation.
- Optimal up to a constant.
- Importance of aggregation not shown
- Why not hold-out?
- Choice of parameters?

Description of the simulation

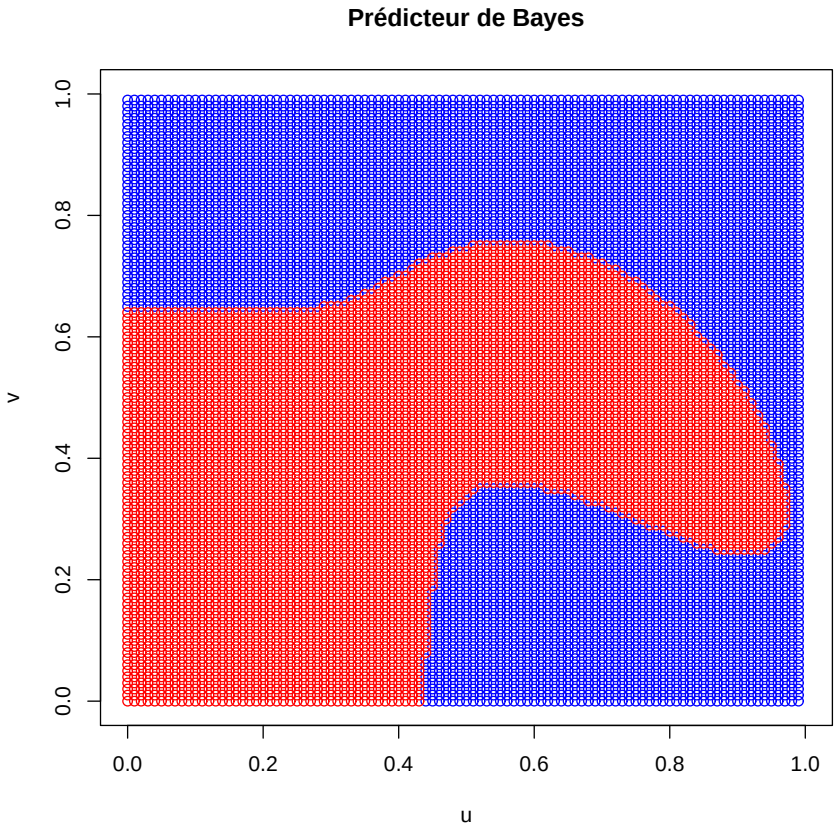
- monte carlo aggregated cross validation is compared to monte carlo cross-validation and the oracle.
- $\mathcal{G} =$ k-NN for $1 \leq k \leq n$, k odd.
- i.i.d sample, $n = 500$, with distribution

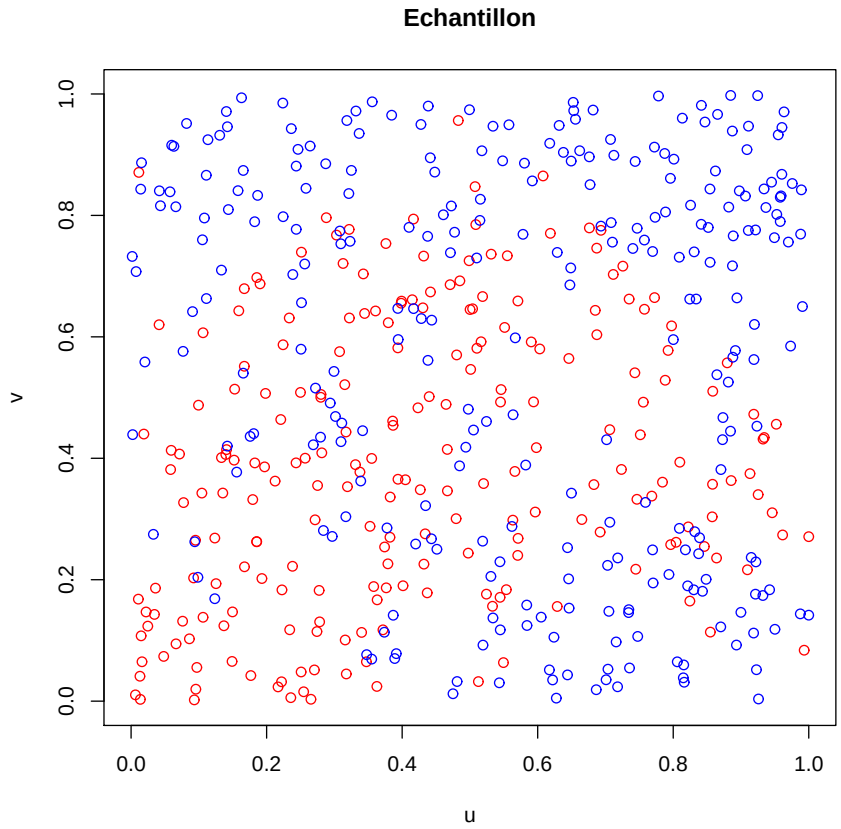
$$X \sim \mathcal{U}([0;1]^2)$$

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

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

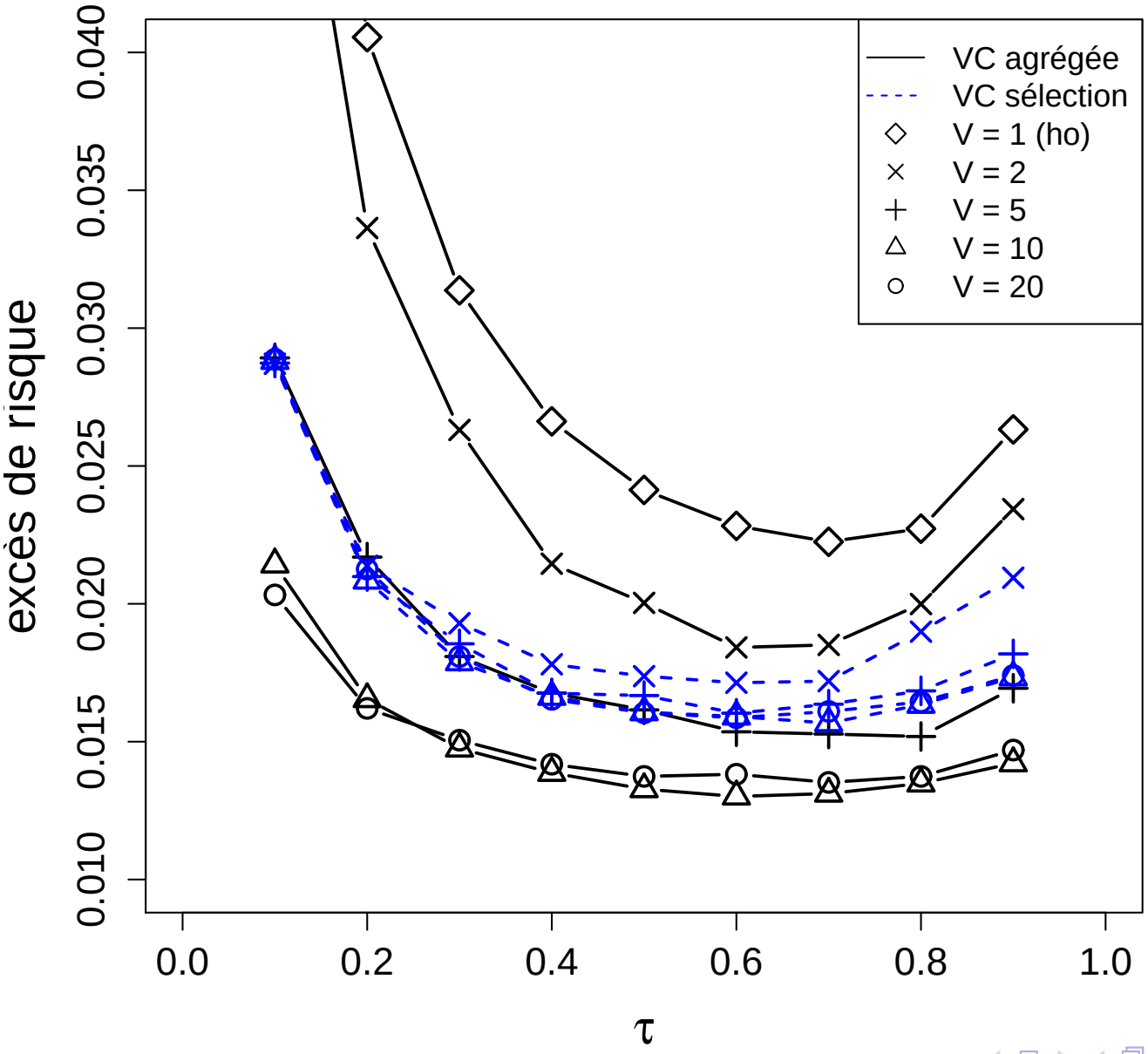
Picture of the distribution





The Bayes risk is approximately 0.2418

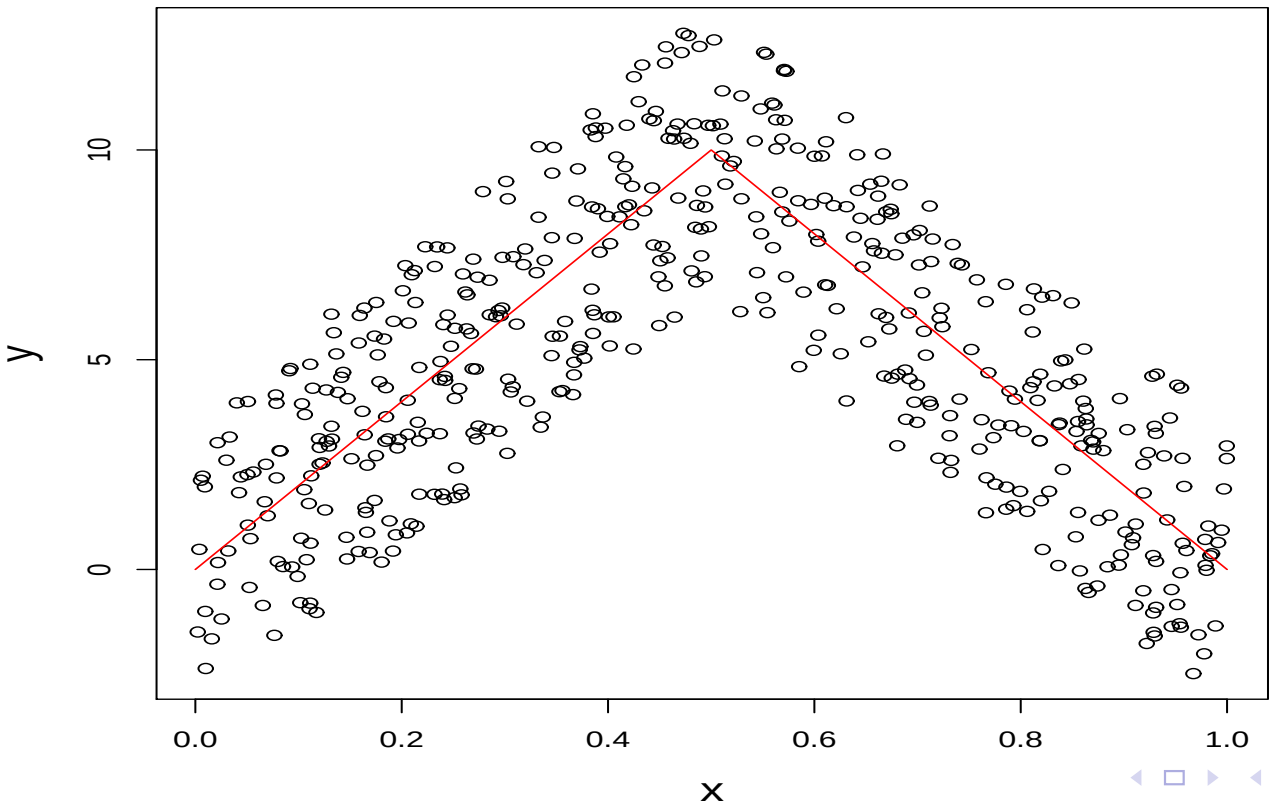
Excess risks as a function of τ



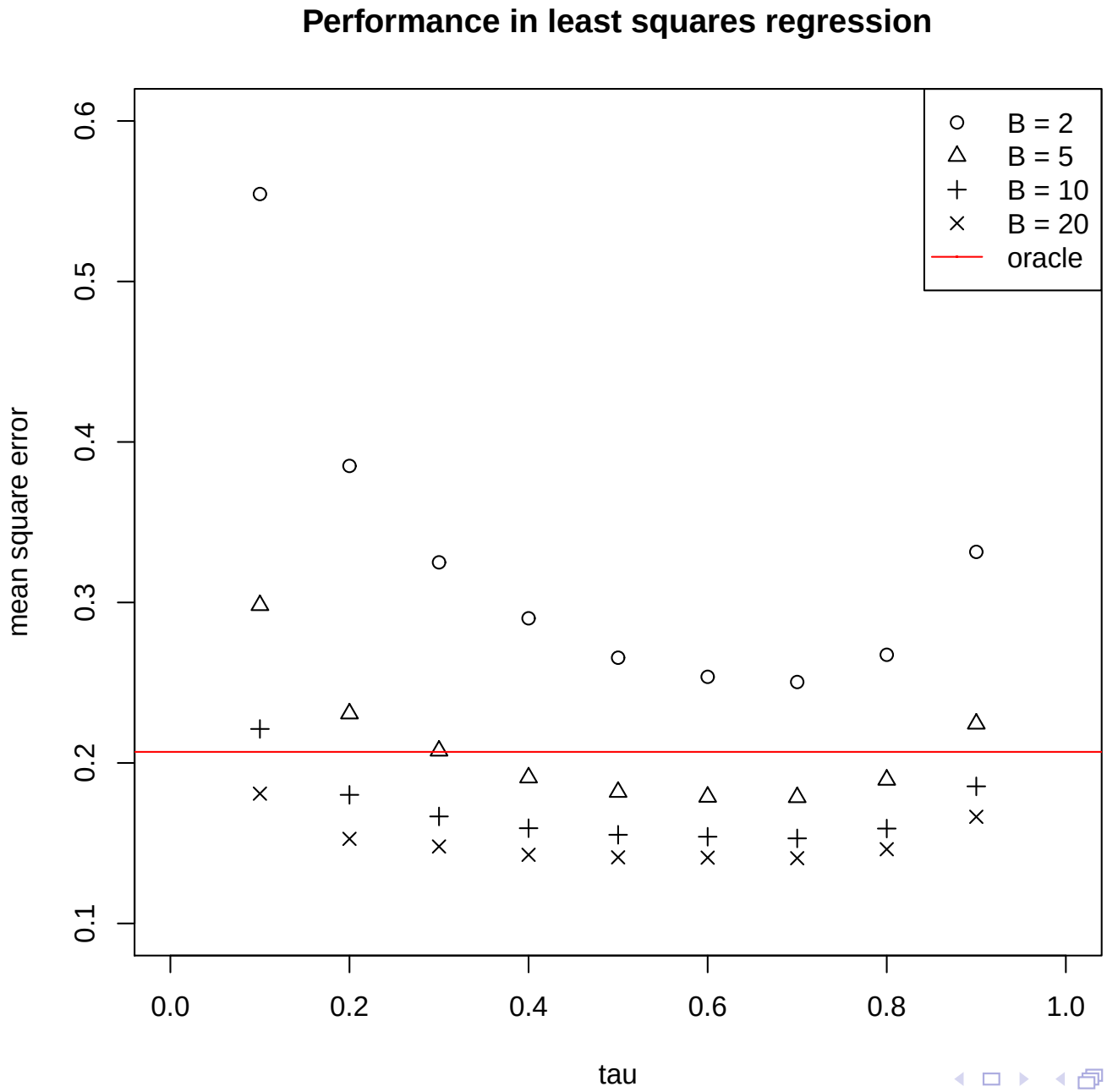
An example in regression

Least squares, $\mathcal{G}$ contains regressograms with step $h \in \{\frac{1}{k} : 1 \leq k \leq n\}$

- $X \sim \mathcal{U}([0;1])$
- $Y \sim t(X) + \varepsilon$ where $\varepsilon \sim \mathcal{U}([-2\sqrt{3}, 2\sqrt{3}])$ indt of X
- $t(x) = 10(1 - 2|x - 0.5|)$



Performance in least-squares regression



Conclusion

- Aggregation of hold-out estimators.
- An alternative to cross-validation in classification and regression.
- Satisfies an oracle inequality in classification
- Aggregation useful in practice, not in theory
- Can beat the oracle!
- More theory to understand aggregation