Aggregation methods: optimality and fast rates.

Guillaume Lecué

Université Paris 6

18 Mai 2007

Motivation.

M prior estimators ('weak' estimators) : $f_1, \ldots, f_M$

n observations : D_n

Motivation.

M prior estimators ('weak' estimators) : $f_1, \ldots, f_M$

n observations : D_n

Aim

Construction of a new estimator which is approximatively as good as the best 'weak' estimator :

$\boxed{\text{Aggregation method}}$ or $\boxed{\text{Aggregate}}$

Examples.

Adaptation :

Observations : D_{m+n}

Estimation : D_m $\longrightarrow$non-adaptive estimators $f_1, \ldots, f_M$.

learning : $D_{(n)}$ $\longrightarrow$aggregate $\tilde{f}_n$ (adaptive).

Examples.

Adaptation :

Observations : D_{m+n}

Estimation : $D_m \longrightarrow$non-adaptive estimators $f_1, \ldots, f_M$.

learning : $D_{(n)} \longrightarrow$aggregate $\tilde{f}_n$ (adaptive).

Estimation :

$\epsilon-$net : $f_1, \ldots, f_M$ (functions)

learning : $D_n \longrightarrow$aggregate $\tilde{f}_n$.

Model.

$(\mathcal{Z}, \mathcal{T})$ a measurable space,

$\mathcal{P}$ the set of all probability measures on $(\mathcal{Z}, \mathcal{T})$,

$F : \mathcal{P} \longmapsto \mathcal{F}$ (example : $F(\pi) = d\pi/d\mu$),

Model.

$(\mathcal{Z}, \mathcal{T})$ a measurable space,

$\mathcal{P}$ the set of all probability measures on $(\mathcal{Z}, \mathcal{T})$,

$F : \mathcal{P} \longmapsto \mathcal{F}$ (example : $F(\pi) = d\pi/d\mu$),

Z : random variable with values in $\mathcal{Z}$,

π probability distribution of Z,

$D_n = (Z_1, \ldots, Z_n)$: n i.i.d. observations of Z.

Model.

$(\mathcal{Z}, \mathcal{T})$ a measurable space,

$\mathcal{P}$ the set of all probability measures on $(\mathcal{Z}, \mathcal{T})$,

$F : \mathcal{P} \longmapsto \mathcal{F}$ (example : $F(\pi) = d\pi/d\mu$),

Z : random variable with values in $\mathcal{Z}$,

π probability distribution of Z,

$D_n = (Z_1, \ldots, Z_n)$: n i.i.d. observations of Z.

Problem : Estimation of $F(\pi)$ from D_n.

Model.

loss : $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$,

Model.

$$\text{loss}: \; Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R},$$

$$\text{risk}: \; A(f) = \mathbb{E}[Q(Z, f)],$$

$$A^* \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}} A(f) = A(f^*), (f^* = F(\pi))$$

Model.

$$\text{loss}: \quad Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R},$$

$$\text{risk}: \quad A(f) = \mathbb{E}[Q(Z, f)],$$

$$A^* \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}} A(f) = A(f^*), (f^* = F(\pi))$$

$$\text{excess risk}: \quad A(f) - A^* \geq 0$$

Model.

$$
\begin{aligned}
\text{loss : } & Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}, \\
\text{risk : } & A(f) = \mathbb{E}[Q(Z, f)], \\
& A^* \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}} A(f) = A(f^*), (f^* = F(\pi)) \\
\text{excess risk : } & A(f) - A^* \geq 0
\end{aligned}
$$

Empirical Risk :

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, f).$$

Regression

$$\text{loss}: \quad Q((x,y),f) = (y - f(x))^2,$$

$$\text{excess risk}: \quad A(f) - A^* = ||f^* - f||^2_{L^2(P^X)} \text{ where } f^*(x) = \mathbb{E}\left[Y|X=x\right].$$

Regression

$$\text{loss}: \quad Q((x,y),f) = (y - f(x))^2,$$
$$\text{excess risk}: \quad A(f) - A^* = ||f^* - f||^2_{L^2(P^X)} \text{ where } f^*(x) = \mathbb{E}[Y|X=x].$$

Density estimation ($f^* = d\pi/d\mu$)

$$\text{KL loss}: \quad Q(z,f) = -\log f(z),$$
$$\textit{excess risk}: \quad A(f) - A^* = K(f^*|f).$$
$$L^2\text{-loss}: \quad Q(z,f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z),$$
$$\textit{excess risk}: \quad A(f) - A^* = ||f^* - f||^2_{L^2(\mu)}.$$

Regression

$$\begin{aligned} \text{loss}: \ & Q((x,y),f) = (y - f(x))^2, \\ \text{excess risk}: \ & A(f) - A^* = ||f^* - f||^2_{L^2(P^X)} \text{ where } f^*(x) = \mathbb{E}\left[Y|X=x\right]. \end{aligned}$$

Density estimation ($f^* = d\pi/d\mu$)

$$\begin{aligned} \text{KL loss}: \ & Q(z,f) = -\log f(z), \\ \textit{excess risk}: \ & A(f) - A^* = K(f^*|f). \\ L^2\text{-loss}: \ & Q(z,f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z), \\ \textit{excess risk}: \ & A(f) - A^* = ||f^* - f||^2_{L^2(\mu)}. \end{aligned}$$

Classification

$$\begin{aligned} \text{loss}: \ & \phi : \mathbb{R} \longmapsto \mathbb{R}, Q((x,y),f) = \phi(yf(x)), y \in \{-1,1\} \\ \phi\text{−risk}: \ & A^\phi(f) = \mathbb{E}[Q((X,Y),f)] = \mathbb{E}[\phi(Yf(X))]. \\ & f^* = f^{\phi *} \text{ s.t. } A^\phi(f^{\phi *}) = \min_{f:\mathcal{X}\longmapsto\mathbb{R}} A^\phi(f). \end{aligned}$$

Selectors.

$\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

- Empirical Risk Minimization (ERM) :(Vapnik, Chervonenkis...)

$$\tilde{f}_n^{ERM} \in \mathrm{Arg}\min_{f \in \mathcal{F}_0} A_n(f).$$

Selectors.

$\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

- Empirical Risk Minimization (ERM) :(Vapnik, Chervonenkis...)

$$\tilde{f}_n^{ERM} \in \mathrm{Arg} \min_{f \in \mathcal{F}_0} A_n(f).$$

- penalized Empirical Risk Minimization (pERM) :

$$\tilde{f}_n^{ERM} \in \mathrm{Arg} \min_{f \in \mathcal{F}_0} [A_n(f) + \mathrm{pen}(f)],$$

where pen is a penalty function. (Barron, Bartlett, Birgé, Boucheron, Koltchinski, Lugosi, Massart,...)

Aggregation methods with exponential weights.

$\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

- Aggregate with Exponential weights (AEW) :

$$\tilde{f}_{n,T}^{AEW} = \sum_{f \in \mathcal{F}_0} w_T^{(n)}(f) f, \text{ where } w_T^{(n)}(f) = \frac{\exp\left(-nTA_n(f)\right)}{\sum_{g \in \mathcal{F}_0} \exp\left(-nTA_n(g)\right)},$$

T^{-1} : temperature parameter.

Aggregation methods with exponential weights.

$\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

- Aggregate with Exponential weights (AEW) :

$$\tilde{f}_{n,T}^{AEW} = \sum_{f \in \mathcal{F}_0} w_T^{(n)}(f) f, \text{ where } w_T^{(n)}(f) = \frac{\exp\left(-nTA_n(f)\right)}{\sum_{g \in \mathcal{F}_0} \exp\left(-nTA_n(g)\right)},$$

T^{-1} : temperature parameter.

- Cumulative Aggregate with Exponential Weights (CAEW) :(Catoni, Yang,...)

$$\tilde{f}_{n,T}^{CAEW} = \frac{1}{n} \sum_{k=1}^{n} \tilde{f}_{k,T}^{AEW}.$$

Aim of Aggregation(1) : Optimal rate of aggregation.

Definition

$\forall \mathcal{F}_0 = \{f_1, \ldots, f_M\} \subseteq \mathcal{F}$, $\exists \tilde{f}_n$ such that $\forall \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{F}_0} \left(A(f) - A^*\right) + C_0 \gamma(n, M).$$

Aim of Aggregation(1) : Optimal rate of aggregation.

Definition

$\forall \mathcal{F}_0 = \{f_1, \ldots, f_M\} \subseteq \mathcal{F}$, $\exists \tilde{f}_n$ such that $\forall \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_0 \gamma(n, M).$$

$\exists \mathcal{F}_0 = \{f_1, \ldots, f_M\}$ such that for any aggregate $\bar{f}_n$, $\exists \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_1 \gamma(n, M).$$

Aim of Aggregation(1) : Optimal rate of aggregation.

Definition

$\forall \mathcal{F}_0 = \{f_1, \ldots, f_M\} \subseteq \mathcal{F}$, $\exists \tilde{f}_n$ such that $\forall \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_0 \gamma(n, M).$$

$\exists \mathcal{F}_0 = \{f_1, \ldots, f_M\}$ such that for any aggregate $\bar{f}_n$, $\exists \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_1 \gamma(n, M).$$

$\gamma(n, M)$ is an optimal rate of aggregation and $\tilde{f}_n$ is an optimal aggregation procedure.

Aim of Aggregation(2) : Adaptation.

Definition (Oracle Inequality)

$\forall \mathcal{F}_0 = \{f_1, \ldots, f_M\} \subseteq \mathcal{F}$, $\exists \tilde{f}_n$ such that $\forall \pi \in \mathcal{P}$, $\forall n \geq 1$

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq C \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_0 \gamma(n, M),$$

where $C \geq 1$.

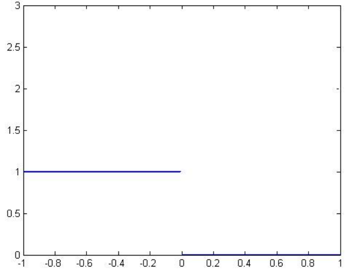
Continuous scale of loss functions.

Classification problem : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1, 1\}, X \in \mathcal{X}$.

$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \leq 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 0$



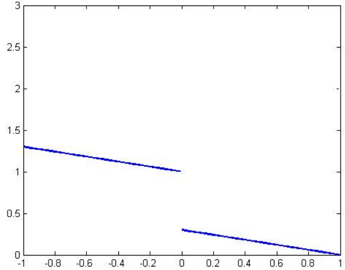
Continuous scale of loss functions.

Classification problem : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1, 1\}, X \in \mathcal{X}.$

$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \leq 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 1/3$



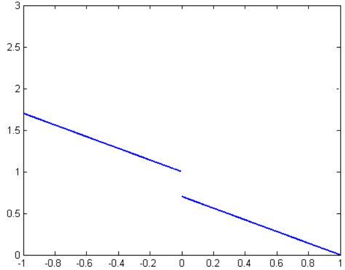
Continuous scale of loss functions.

Classification problem : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1, 1\}, X \in \mathcal{X}.$

$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \leq 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 2/3$



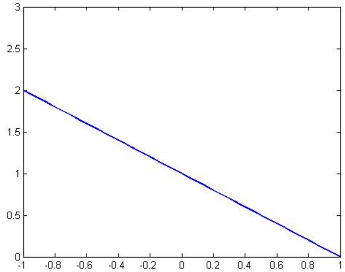
Continuous scale of loss functions.

Classification problem : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1, 1\}, X \in \mathcal{X}$.

$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \leq 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 1$



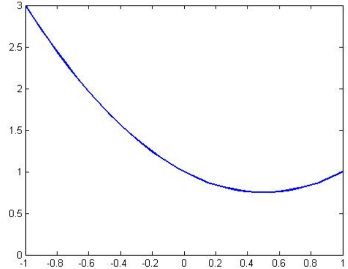
Continuous scale of loss functions.

Classification problem : $A^\phi(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1,1\}, X \in \mathcal{X}$.

$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \le h \le 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \le 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 2$



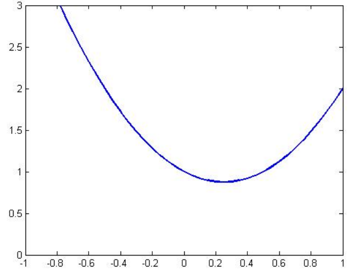
Continuous scale of loss functions.

Classification problem : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))], Y \in \{-1, 1\}, X \in \mathcal{X}$.

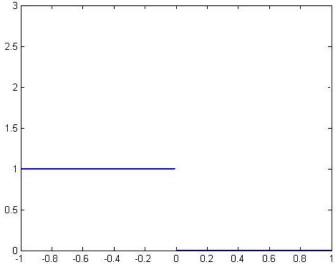
$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

where $\phi_0(z) = \mathbb{I}_{(z \leq 0)}$ is the $0-1$ loss and $\phi_1(z) = \max(0, 1-z)$ is the Hinge loss.

$h = 3$

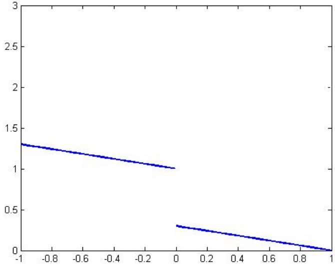


ORA in classification



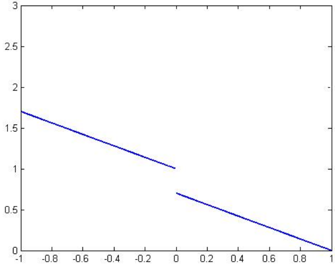
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



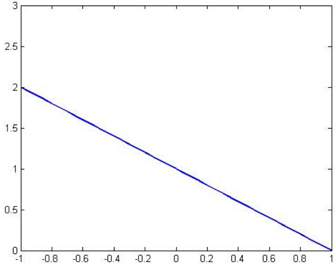
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



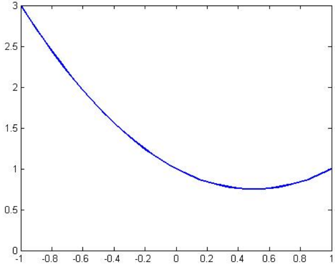
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



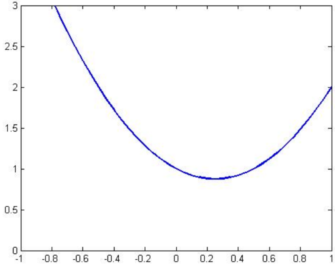
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

2 Questions.

Question 1 : Why is there such a breakdown just after the Hinge loss ?

2 Questions.

Question 1 : Why is there such a breakdown just after the Hinge loss ?

$$0 \leq h \leq 1, \sqrt{\frac{\log M}{n}} \quad \longrightarrow \quad \frac{\log M}{n}, h > 1.$$

2 Questions.

Question 1 : Why is there such a breakdown just after the Hinge loss ?

$$0 \leq h \leq 1, \sqrt{\frac{\log M}{n}} \quad \longrightarrow \quad \frac{\log M}{n}, h > 1.$$

Question 2 : Do we really need aggregation procedures with exponential weights to achieve the optimal rates of aggregation ?

2 Questions.

Question 1 : Why is there such a breakdown just after the Hinge loss ?

$$0 \leq h \leq 1, \sqrt{\frac{\log M}{n}} \quad \longrightarrow \quad \frac{\log M}{n}, h > 1.$$

$$\text{ERM} \quad \longrightarrow \quad \text{CAEW}$$

Question 2 : Do we really need aggregation procedures with exponential weights to achieve the optimal rates of aggregation ?

Question 1. Why there is a breakdown at $h = 1$?

Margin assumption for the loss function ϕ :

The probability measure π satisfies the $\phi-$margin assumption $\phi-\mathrm{MA}(\kappa)$, with margin parameter $\kappa \geq 1$ if

$$\mathbb{E}[(\phi(Yf(X)) - \phi(Yf^{\phi*}(X)))^2] \leq c_\phi (A^\phi(f) - A^{\phi*})^{1/\kappa},$$

for any $f : \mathcal{X} \longmapsto \mathbb{R}$.

cf. Mammen and Tsybakov 99 (discriminant analysis) and Tsybakov 04 (classification).

Question 1. Why there is a breakdown at $h = 1$?

Margin assumption for the loss function ϕ :

The probability measure π satisfies the $\phi-$margin assumption $\phi-\mathrm{MA}(\kappa)$, with margin parameter $\kappa \geq 1$ if

$$\mathbb{E}[(\phi(Yf(X)) - \phi(Yf^{\phi*}(X)))^2] \leq c_\phi (A^\phi(f) - A^{\phi*})^{1/\kappa},$$

for any $f : \mathcal{X} \longmapsto \mathbb{R}$.

cf. Mammen and Tsybakov 99 (discriminant analysis) and Tsybakov 04 (classification).

$$\phi_0 - \mathrm{MA}(\kappa) \Longleftrightarrow \mathbb{P}[|2\eta(X) - 1| \leq t] \leq t^\alpha, \forall 0 < t < 1, \alpha = \frac{1}{\kappa - 1}$$

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

Question 1. Why there is a breakdown at $h=1$?

Margin assumption for the loss function ϕ :

The probability measure π satisfies the $\phi-$margin assumption $\phi-\mathrm{MA}(\kappa)$, with margin parameter $\kappa \geq 1$ if

$$\mathbb{E}[(\phi(Yf(X)) - \phi(Yf^{\phi*}(X)))^2] \leq c_\phi (A^\phi(f) - A^{\phi*})^{1/\kappa},$$

for any $f : \mathcal{X} \longmapsto \mathbb{R}$.

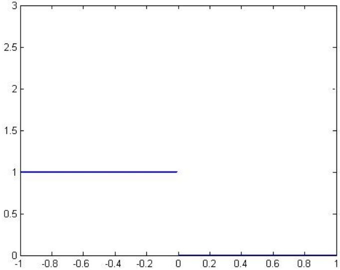
cf. Mammen and Tsybakov 99 (discriminant analysis) and Tsybakov 04 (classification).

$$\phi_0 - \mathrm{MA}(\kappa) \Longleftrightarrow \mathbb{P}[|2\eta(X) - 1| \leq t] \leq t^\alpha, \forall 0 < t < 1, \alpha = \frac{1}{\kappa - 1}$$

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

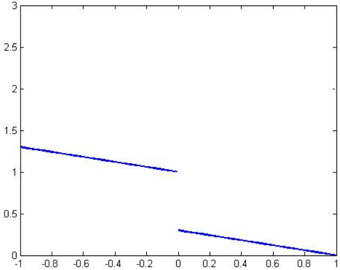
$$(\kappa = 1 \Longleftrightarrow \exists h > 0, |2\eta(X) - 1| \geq h)$$

Question 1. Why there is a breakdown at $h = 1$?



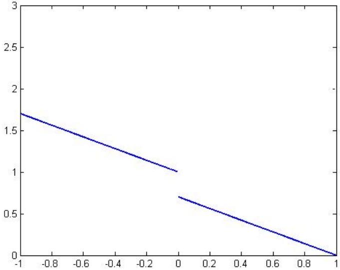
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

Question 1. Why there is a breakdown at $h = 1$?



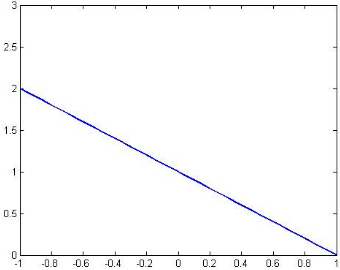
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

Question 1. Why there is a breakdown at $h = 1$?



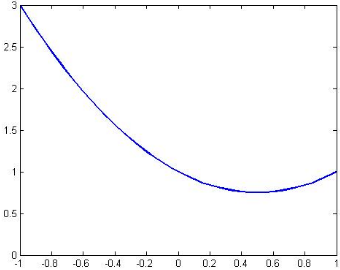
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

Question 1. Why there is a breakdown at $h = 1$?



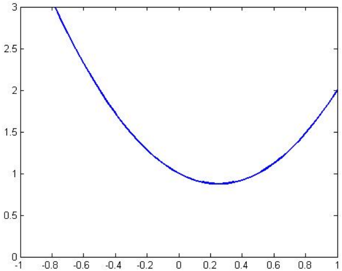
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

Question 1. Why there is a breakdown at $h = 1$?



$\kappa = 1$ for any $h > 1$.

Question 1. Why there is a breakdown at $h = 1$?



$\kappa = 1$ for any $h > 1$.

Question 2 : Do we really need agg. with exp. weights ?

Theorem (suboptimality of selectors)

For any $M \geq 2$, $\phi : \mathbb{R} \longmapsto \mathbb{R}$ s.t. $\phi(-1) \neq \phi(1)$,
$\exists f_1, \ldots, f_M : \mathcal{X} \longmapsto \{-1, 1\}$ s.t. for any selector $\tilde{f}_n$, $\exists \pi$ s.t.

$$\mathbb{E}\left[A^{\phi}(\tilde{f}_n) - A^{\phi *}\right] \geq \min_{j=1,\ldots,M}\left(A^{\phi}(f_j) - A^{\phi *}\right) + C\sqrt{\frac{\log M}{n}}.$$

Question 2 : Do we really need agg. with exp. weights ?

Theorem (suboptimality of selectors under the margin assumption)

For any $M \geq 2$, $\kappa \geq 1$, $\phi : \mathbb{R} \longmapsto \mathbb{R}$ s.t. $\phi(-1) \neq \phi(1)$, $\exists f_1, \ldots, f_M : \mathcal{X} \longmapsto \{-1, 1\}$ s.t. for any selector $\tilde{f}_n$, $\exists \pi$ satisfying the $\phi_0-\mathrm{MA}(\kappa)$ s.t.

$$\mathbb{E}\left[A^{\phi}(\tilde{f}_n) - A^{\phi*}\right] \geq \min_{j=1,\ldots,M}\left(A^{\phi}(f_j) - A^{\phi*}\right) + C\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}.$$

$$\sqrt{\frac{\log M}{n}} >> \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} >> \frac{\log M}{n}, 1 < \kappa < \infty.$$

Question 2 : Do we really need agg. with exp. weights ?

Suboptimality of Penalized ERM.

For any $M \geq 2$, $\kappa > 1$ and $\phi : \mathbb{R} \longmapsto \mathbb{R}$ s.t. $\phi(-1) \neq \phi(1)$, $\exists f_1, \ldots, f_M : \mathcal{X} \longmapsto \{-1, 1\}$, $\exists \pi$ satisfying the $\phi_0-\mathrm{MA}(\kappa)$ s.t. the pERM aggregate

$$\tilde{f}_n^{pERM} \in \mathrm{Arg} \min_{j=1,\ldots,M} \left(A_n^{\phi}(f_j) + \mathrm{pen}(f_j)\right),$$

where $|\mathrm{pen}(f)| < \frac{1}{6}\sqrt{\frac{\log M}{n}}$, satisfies

$$\mathbb{E}\left[A^{\phi}(\tilde{f}_n^{pERM}) - A^{\phi *}\right] \geq \min_{j=1,\ldots,M} \left(A^{\phi}(f_j) - A^{\phi *}\right) + C\sqrt{\frac{\log M}{n}}$$

if $\sqrt{M \log M} \leq \sqrt{n}/(132 e^3)$, for any integer $n \geq 1$.

Theorem (Exact Oracle Inequalities for the general framework)

$\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subseteq \mathcal{F}$. Assume that π satisfies (MA)$(\kappa, c, \mathcal{F}_0)$ ▸ MH and $|Q(Z,f) - Q(Z,f^*)| \leq K, \forall f \in \mathcal{F}_0$.

$$\mathbb{E}[A(\tilde{f}_n^{ERM}) - A^*] \leq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + 4\gamma(n, M, \kappa, \mathcal{B}),$$

where the residual $\gamma(n, M, \kappa, \mathcal{B})$ equals to

$$\begin{cases} \left(\frac{\mathcal{B}^{\frac{1}{\kappa}} \log M}{\beta_1 n} \right)^{1/2} & \text{if } \mathcal{B} \geq \left(\frac{\log M}{\beta_1 n} \right)^{\frac{\kappa}{2\kappa - 1}} \\ \left(\frac{\log M}{\beta_1 n} \right)^{\frac{\kappa}{2\kappa - 1}} & \text{otherwise,} \end{cases}$$

where $\mathcal{B} = \min_{f \in \mathcal{F}_0} (A(f) - A^*)$ and $\beta_1 > 0$.

Theorem (Exact Oracle Inequalities for the general framework)

$\mathcal{F}_0 = \{f_1, \dots, f_M\} \subseteq \mathcal{F}$. Assume that π satisfies $(\mathrm{MA})(\kappa, c, \mathcal{F}_0)$ MH and $|Q(Z,f) - Q(Z,f^*)| \leq K, \forall f \in \mathcal{F}_0$.
If $Q(z,\cdot)$ is convex for any $z \in \mathcal{Z}$, then

$$\mathbb{E}[A(\tilde{f}_n^{AEW}) - A^*] \leq \min_{f \in \mathcal{F}_0}(A(f) - A^*) + 4\gamma(n, M, \kappa, \mathcal{B}),$$

where the residual $\gamma(n, M, \kappa, \mathcal{B})$ equals to

$$\begin{cases} \left(\dfrac{\mathcal{B}^{\frac{1}{\kappa}} \log M}{\beta_1 n}\right)^{1/2} & \text{if } \mathcal{B} \geq \left(\dfrac{\log M}{\beta_1 n}\right)^{\frac{\kappa}{2\kappa-1}} \\ \left(\dfrac{\log M}{\beta_1 n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{otherwise,} \end{cases}$$

where $\mathcal{B} = \min_{f \in \mathcal{F}_0}(A(f) - A^*)$ and $\beta_1 > 0$.

Oracle Inequality in density estimation

Corollary. In density estimation (Margin parameter $\kappa = 1$.)

Let $f_1, \ldots, f_M : \mathcal{X} \longmapsto [0, B]$. Assume that f^* is bounded by B. For any $\epsilon > 0$, we have

$$\mathbb{E}[||f^* - \tilde{f}_n^{AEW}||^2_{L^2(\mu)}] \leq (1+\epsilon) \min_{j=1,\ldots,M}(||f^* - f_j||^2_{L^2(\mu)}) + \frac{C}{\epsilon}\frac{\log M}{n}.$$

Oracle Inequality in density estimation

Corollary. In density estimation (Margin parameter $\kappa = 1$.)

Let $f_1, \ldots, f_M : \mathcal{X} \longmapsto [0, B]$. Assume that f^* is bounded by B. For any $\epsilon > 0$, we have

$$\mathbb{E}[||f^* - \tilde{f}_n^{AEW}||^2_{L^2(\mu)}] \leq (1+\epsilon) \min_{j=1,\ldots,M}(||f^* - f_j||^2_{L^2(\mu)}) + \frac{C}{\epsilon}\frac{\log M}{n}.$$

Aggregation of wavelet thresholded estimators

$\Downarrow$

adaptive minimax procedure over all Besov Balls $B^s_{p,\infty}$ for $s > 1/p$.

(Chesneau, L. (2007))

Oracle Inequality in the regression framework

Corollary. In bounded regression (Margin parameter $\kappa = 1.$)

Let $f_1, \ldots, f_M : \mathcal{X} \longmapsto [0,1]$. For any $\epsilon > 0$, we have

$$\mathbb{E}[||f^* - \tilde{f}_n^{AEW}||^2_{L^2(P^X)}] \leq (1+\epsilon) \min_{j=1,\ldots,M} (||f^* - f_j||^2_{L^2(P^X)}) + \frac{C}{\epsilon} \frac{\log M}{n}.$$

Oracle Inequality in the regression framework

Corollary. In bounded regression (Margin parameter $\kappa = 1$.)

Let $f_1, \ldots, f_M : \mathcal{X} \longmapsto [0,1]$. For any $\epsilon > 0$, we have

$$\mathbb{E}[||f^* - \tilde{f}_n^{AEW}||^2_{L^2(P^X)}] \leq (1+\epsilon) \min_{j=1,\ldots,M}(||f^* - f_j||^2_{L^2(P^X)}) + \frac{C}{\epsilon}\frac{\log M}{n}.$$

Aggregation of wavelet thresholded estimators

$\Downarrow$

adaptive minimax procedure over all Besov Balls $B^s_{p,\infty}$ for $s > 1/p$.

(Chesneau, L. (2007))

Oracle Inequality in classification

Corollary. In classification under the Margin Assumption.

Let $f_1,\ldots,f_M:\mathcal{X}\longmapsto[-1,1]$ and $\kappa\geq 1$. For any π satisfying the margin assumption $\phi_0-\mathrm{MA}(\kappa)$ and any $\epsilon>0$, we have

$$\mathbb{E}[A_1(\tilde{f}_n^{AEW})-A_1^*]\leq(1+\epsilon)\min_{j=1,\ldots,M}(A_1(f_j)-A_1^*)+C\left(\frac{\log M}{\epsilon n}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Oracle Inequality in classification

Corollary. In classification under the Margin Assumption.

Let $f_1, \ldots, f_M : \mathcal{X} \longmapsto [-1, 1]$ and $\kappa \geq 1$. For any π satisfying the margin assumption $\phi_0-\mathrm{MA}(\kappa)$ and any $\epsilon > 0$, we have

$$\mathbb{E}[A_1(\tilde{f}_n^{AEW}) - A_1^*] \leq (1+\epsilon) \min_{j=1,\ldots,M} (A_1(f_j) - A_1^*) + C\left(\frac{\log M}{\epsilon n}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Aggregation of SVM classifiers (or plug-in classifiers)

$\Downarrow$

procedure adaptive simultaneously to the complexity and to the margin parameters.

(Gaïffas, L. (2007))

Single-index model

$(X, Y) \in \mathbb{R}^d \times \mathbb{R}, \qquad Y = g(X) + \sigma(X)\epsilon$

where $\exists \vartheta \in S_+^{d-1} = \left\{ v \in \mathbb{R}^d \mid \|v\|_2 = 1 \text{ et } v_d \geq 0 \right\}$ (*index*) and a function $f : \mathbb{R} \longmapsto \mathbb{R}$ (*link function*) s.t.

$$g(x) = f(\vartheta^\top x).$$

(Gaïffas, L. (2007))

Single-index model

$(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, $\qquad Y = g(X) + \sigma(X)\epsilon$

where $\exists \vartheta \in S_+^{d-1} = \left\{ v \in \mathbb{R}^d \mid \|v\|_2 = 1 \text{ et } v_d \geq 0 \right\}$ (*index*) and a function $f : \mathbb{R} \longmapsto \mathbb{R}$ (*link function*) s.t.

$$g(x) = f(\vartheta^\top x).$$

- $\varepsilon \perp X$ and $\epsilon \sim N(0, 1)$;
- $\sigma_0 \leq \sigma(\cdot) \leq \sigma_1$ (σ_1 known) ;

Aim : estimation of g from n observations $D_n := [(X_i, Y_i); 1 \leq i \leq n]$

Reduction of dimension

Without the assumption of Single-Index and if $g \in H_d(s)$ (Hölder class)

$$\Downarrow$$

$n^{-s/(2s+d)}$ is the minimax rate of convergence.

Reduction of dimension

Without the assumption of Single-Index and if $g \in H_d(s)$ (Hölder class)

$$\Downarrow$$

$n^{-s/(2s+d)}$ is the minimax rate of convergence.

"Open" question 2 of Stone (82)

is $n^{-s/(2s+1)}$ the minimax rate of estimation of the regression function in the Single-Index model, if the link function f belongs to a s-Hölder Class ?

assumption on the Design P^X

Assumption (D)

- P_X is **compactly supported** and $P^X << \lambda_d$
- For any $v \in S_+^{d-1}$, if $\mu := dP_{v^\top X}/Leb$:
  - μ is **continuous** ;
  - $\mathrm{Card}\{z \in \mathrm{Supp}\,\mu \text{ s.t. } \mu(z) = 0\} < \infty$;
  - if $\mu(z) = 0$, then μ is **U-shaped** around z ;
- $\exists \beta \geq 0, \gamma > 0$ s.t. $\forall v \in S_+^{d-1}$ and $\forall$ interval $I \subset \mathrm{Supp}\, P_{v^\top X}$;

$$P_{v^\top X}[I] \geq \gamma |I|^{\beta+1}.$$

Hölder Balls

1-dimensional Hölder Balls $H(s, L)$ (regularity of the link function)

- $H(s, L)$ is made of all functions $f : \mathbb{R} \longmapsto \mathbb{R}$ $\lfloor s \rfloor-$times differentiable satisfying $\forall z_1, z_2 \in \mathbb{R}$

$$|f^{(\lfloor s \rfloor)}(z_1) - f^{(\lfloor s \rfloor)}(z_2)| \leq L|z_1 - z_2|^{s - \lfloor s \rfloor}.$$

- For $Q > 0$, define

$$H^Q(s, L) := H(s, L) \cap \{f \mid \|f\|_\infty := \sup_x |f(x)| \leq Q\}.$$

Upper bound

Theorem

If P_X satisfies assumption (D), we can construct an estimator $\hat{g}$ satisfying for any $s \in [s_{\min}, s_{\max}]$:

$$\sup_{\vartheta \in S_+^{d-1}} \sup_{f \in H^Q(s,L)} E^n \|\hat{g} - g\|_{L^2(P_X)}^2 \leq C n^{-2s/(2s+1)},$$

where $g(\cdot) = f(\vartheta^\top \cdot)$. The constant $C > 0$ depends on $\sigma_1, L, s_{\min}, s_{\max}$ and P_X.

- $\hat{g}$ adapts both to the **index** and the **regularity**.

Lower bound

Theorem

Let $s, L, Q > 0$ and P_X satisfying assumption (D). For any $\vartheta \in S_+^{d-1}$:

$$\inf_{\tilde{g}} \sup_{f \in H(s,L)} E^n \|\tilde{g} - g\|^2_{L^2(P_X)} \geq C' n^{-2s/(2s+1)}$$

where $\inf_{\tilde{g}}$ denotes the infimum over all estimator $\tilde{g}$ constructed on D_n.

- $n^{-2s/(2s+1)}$ is the minimax rate of convergence in the single-index model conjectured by Stone (1982).

Construction of the estimator

We split the sample in two subsamples

- *Training sample :*

$$D_m := [(X_i, Y_i); 1 \leq i \leq m] \text{ (for instance } m = 3n/4)$$

- *Learning sample :*

$$D_{(m)} := [(X_i, Y_i); m+1 \leq i \leq n].$$

weak estimators : local polynomial estimators

Construction of the weak estimators : We **fixe** a parameter

$$\lambda = (v, s) \in \Lambda_n = S_+^{d-1}(\Delta_n) \times [s_{min}, s_{min} + (\log n)^{-1}, \ldots, s_{max}],$$

where $\Delta_n = (n \log n)^{-1/(2s_{min})}$.

weak estimators : local polynomial estimators

Construction of the weak estimators : We **fixe** a parameter

$$\lambda = (v, s) \in \Lambda_n = S_+^{d-1}(\Delta_n) \times [s_{min}, s_{min} + (\log n)^{-1}, \ldots, s_{max}],$$

where $\Delta_n = (n \log n)^{-1/(2s_{min})}$.

- We work with the data projected in the direction v :

$$D_m(v) = [(v^\top X_i, Y_i); 1 \leq i \leq m];$$

weak estimators : local polynomial estimators

Construction of the weak estimators : We **fixe** a parameter

$$\lambda = (v, s) \in \Lambda_n = S_+^{d-1}(\Delta_n) \times [s_{min}, s_{min} + (\log n)^{-1}, \dots, s_{max}],$$

where $\Delta_n = (n \log n)^{-1/(2s_{min})}$.

- We work with the data projected in the direction v :

$$D_m(v) = [(v^\top X_i, Y_i); 1 \leq i \leq m];$$

- Construction of a 1-dimensional polynomial estimator $f^{(\lambda)}(\cdot)$ with the data $D_m(v)$ for the regularity s LPE.

weak estimators : local polynomial estimators

Construction of the weak estimators : We **fixe** a parameter

$$\lambda = (v, s) \in \Lambda_n = S_+^{d-1}(\Delta_n) \times [s_{min}, s_{min} + (\log n)^{-1}, \dots, s_{max}],$$

where $\Delta_n = (n \log n)^{-1/(2s_{min})}$.

- We work with the data projected in the direction v :

$$D_m(v) = [(v^\top X_i, Y_i); 1 \leq i \leq m];$$

- Construction of a 1-dimensional polynomial estimator $f^{(\lambda)}(\cdot)$ with the data $D_m(v)$ for the regularity s LPE.
- $\bar{g}^{(\lambda)}(x) := \tau_Q\big(\bar{f}^{(\lambda)}(v^\top x)\big)$ where $\tau_Q(g) := \max(-Q, \min(Q, g))$.

We do this for any $\lambda \in \Lambda_n$! ReCo

Aggregation of the weak estimators

Adaptation to the regularity and the index by aggregation. Once we have the dictionary $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda_n\}$ of weak estimators.

- Empirical risk of $\bar{g}$:

$$A_{(m)}(\bar{g}) := \sum_{i=m+1}^{n} (Y_i - \bar{g}(X_i))^2. \tag{1}$$

- For a *temperature* $T^{-1} > 0$, we put a Gibbs measure ▸ Gibbs on $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda_n\}$:

$$w(\bar{g}) := \frac{\exp\big(-TA_{(m)}(\bar{g})\big)}{\sum_{\lambda\in\Lambda_n} \exp\big(-TA_{(m)}(\bar{g}^{(\lambda)})\big)}. \tag{2}$$

- the final estimator is the AEW aggregate of local polynomial estimators :

$$\hat{g} := \sum_{\lambda\in\Lambda_n} w(\bar{g}^{(\lambda)})\bar{g}^{(\lambda)}.$$

Remarks on the temperature parameter

- if T^{-1} large $\Rightarrow$exponential weights are close to the uniform weights.
- if T^{-1} small $\Rightarrow$all the weights equal zero except one : $\hat{g}$ is the ERM.

Remarks on the temperature parameter

- if T^{-1} large $\Rightarrow$exponential weights are close to the uniform weights.
- if T^{-1} small $\Rightarrow$all the weights equal zero except one : $\hat{g}$ is the ERM.

T is a parameter of a trade-off between an aggregate with uniform weights and the ERM.

Remarks on the temperature parameter

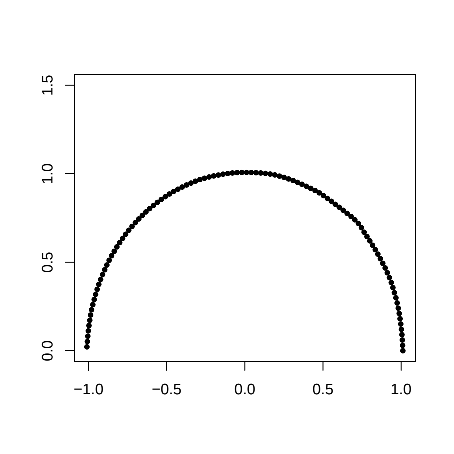
- if T^{-1} large $\Rightarrow$exponential weights are close to the uniform weights.
- if T^{-1} small $\Rightarrow$all the weights equal zero except one : $\hat{g}$ is the ERM.

T is a parameter of a trade-off between an aggregate with uniform weights and the ERM.

Quality of the aggregate depends on the choice of T.

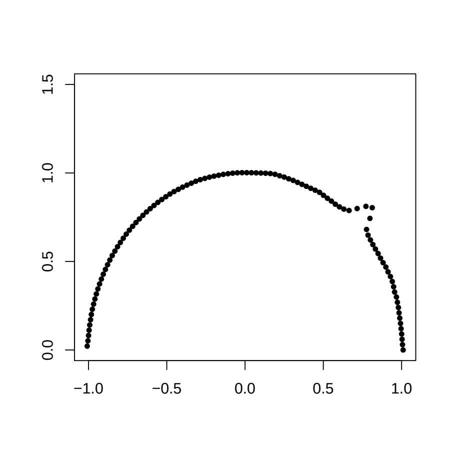
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda_n\}$ on $\bar{S}^1_+(\Delta_n)$ for $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ and $T = 0.05, 0.2, 0.5, 10$



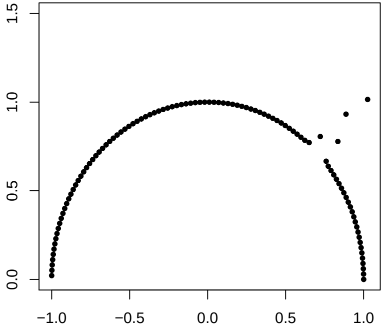
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda_n\}$ on $\bar{S}^1_+(\Delta_n)$ for $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ and $T = 0.05, 0.2, 0.5, 10$



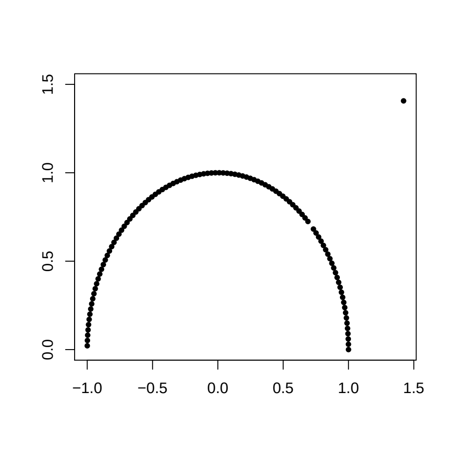
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda_n\}$ on $\bar{S}^1_+(\Delta_n)$ for $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ and $T = 0.05, 0.2, 0.5, 10$



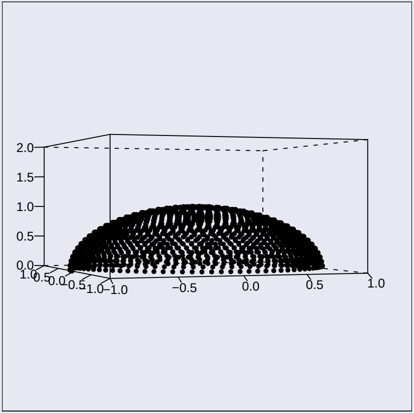
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda_n\}$ on $\bar{S}^1_+(\Delta_n)$ for $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ and $T = 0.05, 0.2, 0.5, 10$



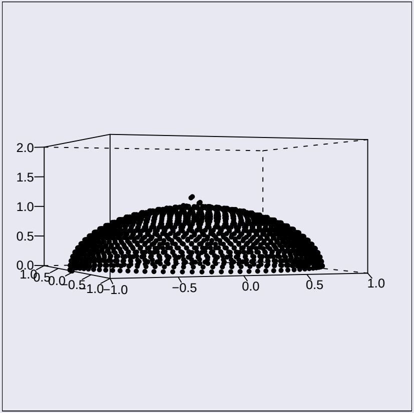
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda\}$ on $\bar{S}_+^2(\Delta_n)$ for $\vartheta = (0,0,1)$ and $T = 0.05, 0.3, 0.5, 10$



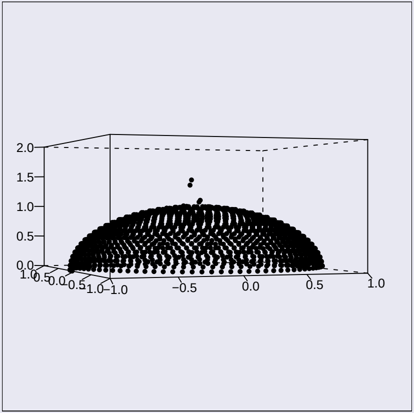
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda\}$ on $\bar{S}_+^2(\Delta_n)$ for $\vartheta = (0,0,1)$ and $T = 0.05, 0.3, 0.5, 10$



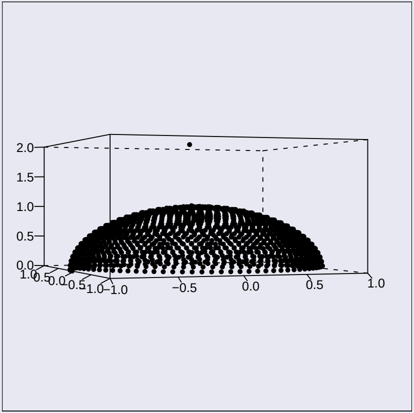
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda\}$ on $\bar{S}^2_+(\Delta_n)$ for $\vartheta = (0,0,1)$ and $T = 0.05, 0.3, 0.5, 10$



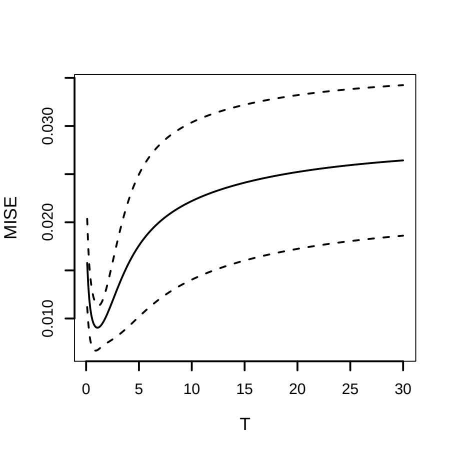
Aggregation phenomenon depending on the temperature

Weights $\{w(\bar{g}^{(\lambda)}) : \lambda \in \Lambda\}$ on $\bar{S}^2_+(\Delta_n)$ for $\vartheta = (0,0,1)$ and $T = 0.05, 0.3, 0.5, 10$



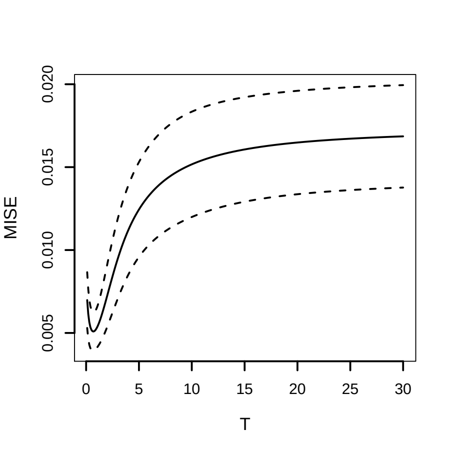
For 100 simulations

MISE against T for $f = \texttt{hardsine}$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$ and $n = 200$



For 100 simulations

MISE against T for $f = \texttt{hardsine}$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$ and $n = 400$



For 100 simulations

MISE against T ($f = \texttt{hardsine}$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$)

n \ T	0.1	0.5	0.7	1.0	1.5	2.0	ERM	aggCVT
100	0.029	0.021	0.019	0.018	**0.017**	0.018	0.037	0.020
	(.011)	(.008)	(.008)	(.007)	(.008)	(.009)	(.022)	(.008)
200	0.016	0.010	0.010	**0.009**	**0.009**	0.010	0.026	0.010
	(.005)	(.003)	(.003)	(.002)	(.002)	(.003)	(0.008)	(.003)
400	0.007	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.003)	(.001)

MISE against T ($f = \texttt{hardsine}$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$)

n \ T	0.1	0.5	0.7	1.0	1.5	2.0	ERM	aggCVT
100	0.038	0.027	0.021	0.019	**0.017**	**0.017**	0.038	0.020
	(.016)	(.010)	(.009)	(.008)	(.007)	(.007)	(.025)	(.010)
200	0.019	0.013	**0.012**	**0.012**	0.013	0.014	0.031	0.013
	(.014)	(.009)	(.010)	(.011)	(.012)	(.012)	(.016)	(.010)
400	0.009	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.004)	(.001)

Some perspectives.

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.
- To construct of aggregation methods in a framework without empirical risk (pointwise estimation, L^p-risk for $p \neq 2, \ldots$).

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.
- To construct of aggregation methods in a framework without empirical risk (pointwise estimation, L^p-risk for $p \neq 2, \ldots$).
- To explore the quality of some randomized aggregates for non-convex losses :

$$\tilde{f}_n^{Rand} = f_j, \text{ with probability } w_T(f_j).$$

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.
- To construct of aggregation methods in a framework without empirical risk (pointwise estimation, L^p-risk for $p \neq 2, \ldots$).
- To explore the quality of some randomized aggregates for non-convex losses :

$$\tilde{f}_n^{Rand} = f_j, \text{ with probability } w_T(f_j).$$

- To explore the quality of aggregation of the ERM over the convex class of the dictionary :

$$\tilde{f}_n \in \mathrm{Arg}\min_{f \in \mathcal{C}} A_n(f), \text{ where } \mathcal{C} = \{\sum_{j=1}^{M} \lambda_j f_j ; \lambda_j \geq 0, \sum \lambda_j = 1\}.$$

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.
- To construct of aggregation methods in a framework without empirical risk (pointwise estimation, L^p-risk for $p \neq 2, ...$).
- To explore the quality of some randomized aggregates for non-convex losses :

$$\tilde{f}_n^{Rand} = f_j, \text{ with probability } w_T(f_j).$$

- To explore the quality of aggregation of the ERM over the convex class of the dictionary :

$$\tilde{f}_n \in \mathrm{Arg}\min_{f \in \mathcal{C}} A_n(f), \text{ where } \mathcal{C} = \{\sum_{j=1}^{M} \lambda_j f_j; \lambda_j \geq 0, \sum \lambda_j = 1\}.$$

- To construct of sparse aggregate for models selection.

Some perspectives.

- To introduce a margin parameter in the problem of prediction of individual sequences.
- To construct of aggregation methods in a framework without empirical risk (pointwise estimation, L^p-risk for $p \neq 2, ...$).
- To explore the quality of some randomized aggregates for non-convex losses :

$$\tilde{f}_n^{Rand} = f_j, \text{ with probability } w_T(f_j).$$

- To explore the quality of aggregation of the ERM over the convex class of the dictionary :

$$\tilde{f}_n \in \operatorname{Arg}\min_{f \in \mathcal{C}} A_n(f), \text{ where } \mathcal{C} = \{\sum_{j=1}^{M} \lambda_j f_j ; \lambda_j \geq 0, \sum \lambda_j = 1\}.$$

- To construct of sparse aggregate for models selection.
- To find the optimal Temperature parameter.

Margin assumption in the general framework

Margin Assumption :

The probability measure π satisfies the margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$, where $\kappa \geq 1, c > 0$ and $\mathcal{F}_0 \subset \mathcal{F}$ if

$$\mathbb{E}[(Q(Z,f) - Q(Z,f^*))^2] \leq c(A(f) - A^*)^{1/\kappa},$$

for any function $f \in \mathcal{F}_0$.

Where does this Gibbs measure come from ?

- The weights $w = (w_\lambda)_{\lambda\in\Lambda} := (w(\bar{g}^{(\lambda)}))_{\lambda\in\Lambda}$ are solution of

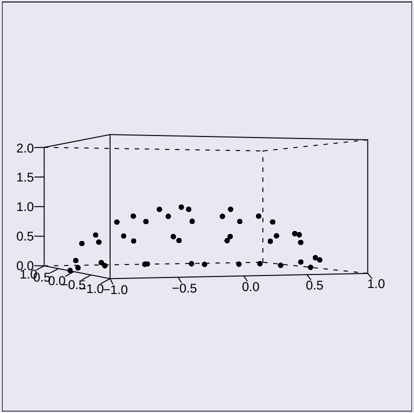
$$\min\Big(\tilde{R}_{(m)}(\theta) + \frac{1}{T}\sum_{\lambda\in\Lambda}\theta_\lambda\log\theta_\lambda \;\Big|\; (\theta_\lambda)\in\mathcal{C}\Big),$$

where $\tilde{R}_m(\theta) := \sum_{\lambda\in\Lambda}\theta_\lambda R_{(m)}(\bar{g}^{(\lambda)})$ and

$$\mathcal{C} := \Big\{(\theta_\lambda)_{\lambda\in\Lambda} \text{ s.t. } \theta_\lambda \geq 0 \text{ and } \sum_{\lambda\in\Lambda}\theta_\lambda = 1\Big\}.$$

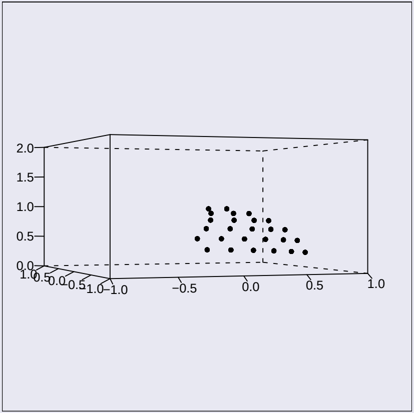
Reduction of complexity : "preselection" of the weak estimators.

iterative Construction of the dictionary : step 1, 2, 3, 4 ($\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$)



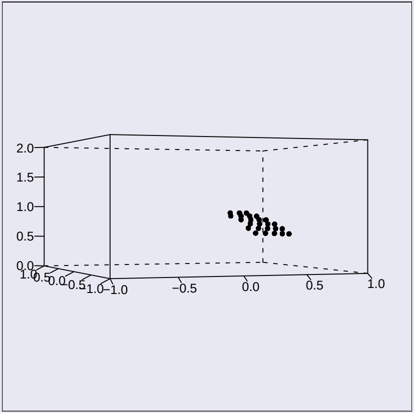
Reduction of complexity : “preselection” of the weak estimators.

iterative Construction of the dictionary : step 1, 2, 3, 4 ($\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$)



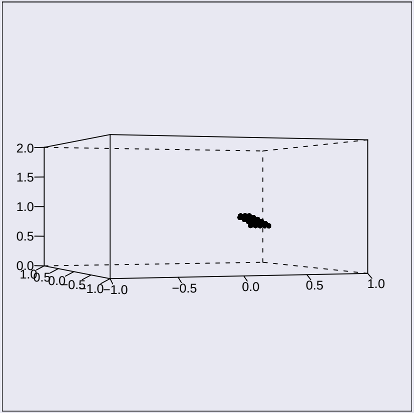
Reduction of complexity : “preselection” of the weak estimators.

iterative Construction of the dictionary : step 1, 2, 3, 4 ($\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$)



Reduction of complexity : "preselection" of the weak estimators.

iterative Construction of the dictionary : step 1, 2, 3, 4 ($\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$)



weak estimators : local polynomial estimators

Construction of the weak estimators : the parameter $\lambda = (v, s) \in \Lambda_n = S_+^{d-1}(\Delta_n) \times [s_{min}, s_{min} + (\log n)^{-1}, \dots, s_{max}]$ is **fixed**

- We work with the data projected in the direction v :

$$D_m(v) = [(Z_i, Y_i); 1 \leq i \leq m] \text{ where } Z_i := v^\top X_i;$$

- If $h > 0$ (*window*), define $\bar{P}_{(z,h)} \in \mathrm{Pol}_r$ minimizing

$$\sum_{i=1}^{m} (Y_i - P(Z_i - z))^2 \mathbf{1}_{Z_i \in [z-h, z+h]},$$

with the window at the point z given by

$$H_m(z) := \operatorname*{argmin}_{h>0} \left\{ L h^s \geq \frac{\sigma_1}{(m \bar{P}_Z[z-h, z+h])^{1/2}} \right\}$$

where $\bar{P}_Z(A) := \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{Z_i \in A}$

- Set $\bar{f}(z) := \bar{P}_{(z, H_m(z))}(z)$, and for $Q > 0$ fixed and all $x \in \mathbb{R}^d$,

$$\bar{g}^{(\lambda)}(x) := \tau_Q\big(\bar{f}^{(\lambda)}(v^\top x)\big) \text{ where } \tau_Q(g) := \max(-Q, \min(Q, g)).$$

We do this for any $\lambda \in \Lambda_n$! ReCo