Aggregation methods in classification: optimality and adaptation

Guillaume Lecué

Université Paris 6

ISI 2007, Lisboa. August, 27, 2007

Motivation

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

n observations : D_n

Motivation

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

n observations : D_n

Aim

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

Aggregation method or Aggregate

Examples

Adaptation :

Observations : D_{m+n}

Estimation : $D_m \rightarrow$non-adaptive estimators $f_1, \dots, f_M$.

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

Examples

Adaptation :

Observations : D_{m+n}

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

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

Estimation :

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

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

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$: n i.i.d. observations.

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \to$ label $y \in \{-1, 1\}$?

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \rightarrow$ label $y \in \{-1, 1\}$?

$f : \mathcal{X} \longmapsto \{-1, 1\}$: prediction rule.

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \to$ label $y \in \{-1, 1\}$?

$f : \mathcal{X} \longmapsto \{-1, 1\}$: prediction rule.

Bayes risk : $A_0(f) = \mathbb{P}[f(X) \neq Y]$

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \rightarrow$ label $y \in \{-1, 1\}$?

$f : \mathcal{X} \longmapsto \{-1, 1\}$: prediction rule.

Bayes risk : $A_0(f) = \mathbb{P}[f(X) \neq Y]$

Bayes rule : $f^*(x) = \mathrm{Sign}(2\eta(x) - 1)$ where $\eta(x) = \mathbb{P}[Y = 1 | X = x]$.

$A_0^* \stackrel{\text{def}}{=} \min_f A_0(f) = A_0(f^*)$

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \to$ label $y \in \{-1, 1\}$?

$f : \mathcal{X} \longmapsto \{-1, 1\}$: prediction rule.

Bayes risk : $A_0(f) = \mathbb{P}[f(X) \neq Y]$

Bayes rule : $f^*(x) = \mathrm{Sign}(2\eta(x) - 1)$ where $\eta(x) = \mathbb{P}[Y = 1 | X = x]$.

$A_0^* \stackrel{\text{def}}{=} \min_f A_0(f) = A_0(f^*)$

Prediction$\rightarrow$estimation : estimation of f^*.

Model of classification

$(\mathcal{X}, \mathcal{A})$ a measurable space,

$(X, Y) \sim \pi$ valued in $\mathcal{X} \times \{-1, 1\}$,

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$: n i.i.d. observations.

Problem of prediction : $x \in \mathcal{X} \rightarrow$ label $y \in \{-1, 1\}$?

$f : \mathcal{X} \longmapsto \{-1, 1\}$: prediction rule.

Bayes risk : $A_0(f) = \mathbb{P}[f(X) \neq Y]$

Bayes rule : $f^*(x) = \mathrm{Sign}(2\eta(x) - 1)$ where $\eta(x) = \mathbb{P}[Y = 1 | X = x]$.

$A_0^* \stackrel{\text{def}}{=} \min_f A_0(f) = A_0(f^*)$

Prediction→estimation : estimation of f^*.

excess risk : $A_0(f) - A_0^*$

Model of classification

$(f : \mathcal{X} \longmapsto \mathbb{R}) \to$ risk $A_0(f) = \mathbb{E}[\phi_0(Yf(X))]$ where

$\phi_0(x) = \mathbb{I}_{(x \leq 0)}$ classical loss or $0-1$ loss

Model of classification

$(f : \mathcal{X} \longmapsto \mathbb{R}) \to$ risk $A_0(f) = \mathbb{E}[\phi_0(Yf(X))]$ where

$\phi_0(x) = \mathbb{I}_{(x \leq 0)}$	classical loss or $0-1$ loss
$\phi_1(x) = \max(0, 1-x)$	hinge loss or (SVM loss)
$x \longmapsto \log_2(1+\exp(-x))$	'Logit-Boosting' loss
$x \longmapsto \exp(-x)$	exponential Boosting loss
$x \longmapsto (1-x)^2$	quadratic loss
$x \longmapsto \max(0, 1-x)^2$	2-norm soft margin loss

Model of classification

$(f : \mathcal{X} \longmapsto \mathbb{R}) \to$ risk $A_0(f) = \mathbb{E}[\phi_0(Yf(X))]$ where

$\phi_0(x) = \mathbb{I}_{(x \leq 0)}$	classical loss or $0-1$ loss
$\phi_1(x) = \max(0, 1-x)$	hinge loss or (SVM loss)
$x \longmapsto \log_2(1+\exp(-x))$	'Logit-Boosting' loss
$x \longmapsto \exp(-x)$	exponential Boosting loss
$x \longmapsto (1-x)^2$	quadratic loss
$x \longmapsto \max(0, 1-x)^2$	2-norm soft margin loss

$\phi-$risk : $A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))]$, $\quad A^{\phi *} \stackrel{\text{def}}{=} \inf_f A(f) = A(f^{\phi *})$,

excess $\phi-$risk : $A^{\phi}(f) - A^{\phi *}$.

empirical $\phi-$risk : $A_n^{\phi}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$.

Selectors

$\phi : \mathbb{R} \longmapsto \mathbb{R}$ a loss, $\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

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

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

Selectors

$\phi : \mathbb{R} \longmapsto \mathbb{R}$ a loss, $\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

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

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

- penalized Empirical Risk Minimization (pERM) :

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

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

Aggregation methods with exponential weights

$\phi : \mathbb{R} \longmapsto \mathbb{R}$ a loss, $\mathcal{F}_0 = \{f_1, \dots, f_M\} \subset \mathcal{F}$ a dictionary.

- Aggregate with Exponential weights (AEW) :

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

T^{-1} : temperature parameter.

Aggregation methods with exponential weights

$\phi : \mathbb{R} \longmapsto \mathbb{R}$ a loss, $\mathcal{F}_0 = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ a dictionary.

- Aggregate with Exponential weights (AEW) :

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

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

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

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

$\exists \mathcal{F}_0 = \{f_1, \dots, f_M\}$ 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 $C \geq 1$.

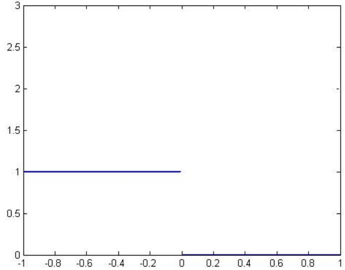
Continuous scale of loss functions

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.

$h = 0$



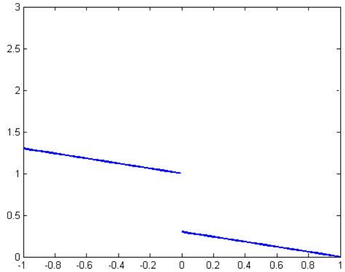
Continuous scale of loss functions

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.

$h = 1/3$



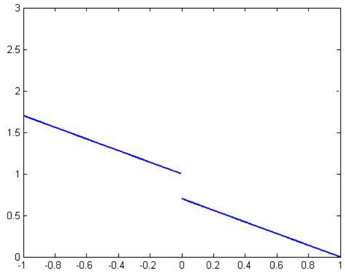
Continuous scale of loss functions

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.

$h = 2/3$



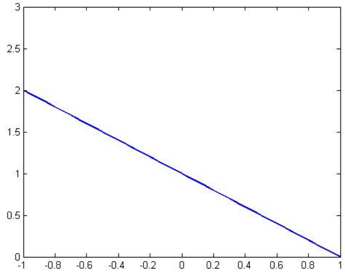
Continuous scale of loss functions

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.

$h = 1$



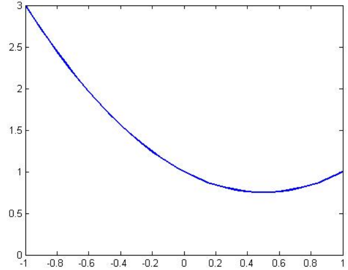
Continuous scale of loss functions

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

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

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.

$h = 2$



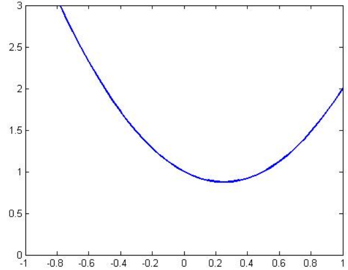
Continuous scale of loss functions

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

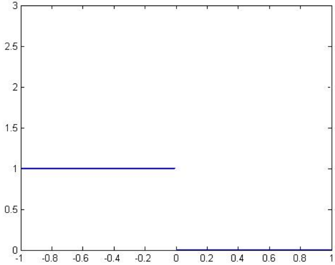
$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \le h \le 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}$$

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.

$h = 3$

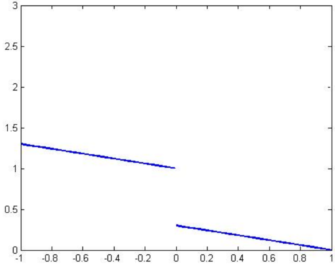


ORA in classification



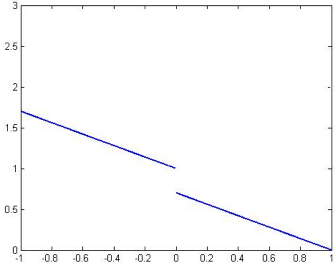
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



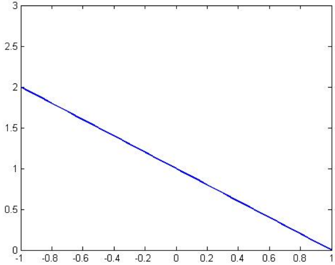
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



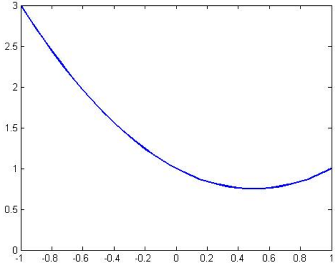
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

ORA in classification



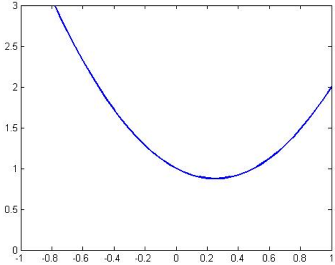
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

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

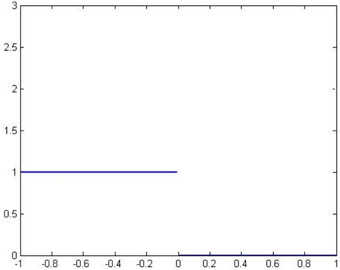
cf. Mammen and Tsybakov 99 (discriminant analysis) and Tsybakov 04 (classification).

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

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

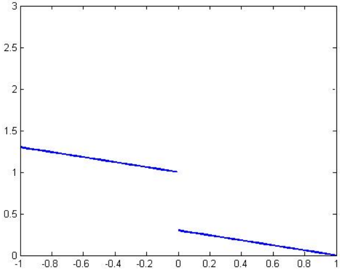
$$(\kappa = 1 \Longleftrightarrow \exists h > 0, |2\eta(X) - 1| \geq h)$$

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



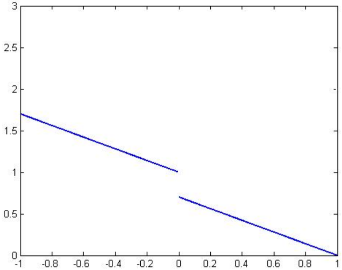
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

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



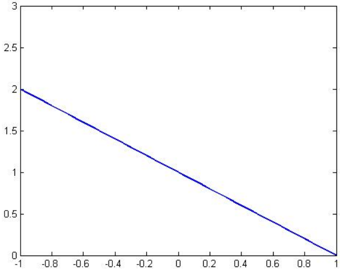
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

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



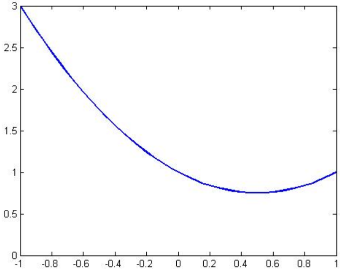
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

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



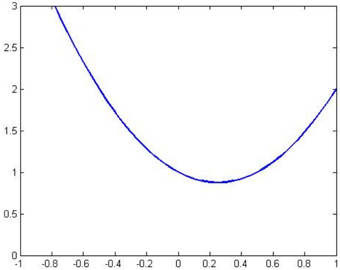
$\kappa = +\infty$ for any $0 \leq h \leq 1$.

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

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

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 $n \geq 1$.

Conclusion on optimality

- The margin parameter characterizes the quality of aggregation and estimation in a given model.

Conclusion on optimality

- The margin parameter characterizes the quality of aggregation and estimation in a given model.

- We need convex aggregates to achieve the optimal rate of aggregation for convex losses.

Exact Oracle Inequality

$\mathcal{F}_0 = \{f_1, \ldots, f_M\}$, $\phi : \mathbb{R} \longmapsto \mathbb{R}$ bounded, $\kappa \geq 1$. Assume that π satisfies ϕ−MA(κ).

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

where the residual term is

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

where $\mathcal{B} = \min_{f \in \mathcal{F}_0}\left(A^\phi(f) - A^{\phi *}\right)$.

Exact Oracle Inequality

$\mathcal{F}_0 = \{f_1, \ldots, f_M\}$, $\phi : \mathbb{R} \longmapsto \mathbb{R}$ bounded, $\kappa \geq 1$. Assume that π satisfies $\phi-$MA(κ). If ϕ is convex then,

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

where the residual term is

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

where $\mathcal{B} = \min_{f \in \mathcal{F}_0} \left(A^{\phi}(f) - A^{\phi *}\right)$.

Idea of SVM

Observations : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$.
$\mathcal{X}$ small dimension $\Rightarrow$ linear separation unlikely.

Idea of SVM

Observations : $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$.
$\mathcal{X}$ small dimension $\Rightarrow$ linear separation unlikely.

Transfer function : $\phi : \mathcal{X} \longmapsto \mathcal{H}$ (dim∞).

Idea of SVM

Observations : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$.
$\mathcal{X}$ small dimension $\Rightarrow$ linear separation unlikely.

Transfer function : $\phi : \mathcal{X} \longmapsto \mathcal{H}$ (dim∞).

New observations : $(\phi(X_1), Y_1), \ldots, (\phi(X_n), Y_n)$.

Idea of SVM

Observations : $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$.
$\mathcal{X}$ small dimension $\Rightarrow$ linear separation unlikely.

Transfer function : $\phi : \mathcal{X} \longmapsto \mathcal{H}$ (dim∞).

New observations : $(\phi(X_1), Y_1), \dots, (\phi(X_n), Y_n)$.

$$\Downarrow$$

Best linear separation in $\mathcal{H}$.

Kernel

Kernel : Symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ s.t. $\forall n \geq 1$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, the matrix

$$(k(x_i, x_j))_{1 \leq i,j \leq n} \quad \text{is semi-definite positive.}$$

$\Leftrightarrow$ there exists a Hilbert space $\mathcal{H}$ and a transfer function $\phi : \mathcal{X} \mapsto \mathcal{H}$ s.t.

$$k(x, x') = \langle \phi(x), \phi(x') \rangle, \quad \forall x, x' \in \mathcal{X}.$$

Kernel

Kernel : Symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ s.t. $\forall n \geq 1$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, the matrix

$$(k(x_i, x_j))_{1 \leq i,j \leq n} \quad \text{is semi-definite positive.}$$

$\Leftrightarrow$ there exists a Hilbert space $\mathcal{H}$ and a transfer function $\phi : \mathcal{X} \mapsto \mathcal{H}$ s.t.

$$k(x, x') = \langle \phi(x), \phi(x') \rangle, \quad \forall x, x' \in \mathcal{X}.$$

Gaussian kernel : For all $\sigma > 0$ (σ is called the window),

$$k(x, x') = \exp\left(-\sigma^2 ||x - x'||_2^2\right), \quad x, x' \in \mathbb{R}^d.$$

RKHS

RKHS : k is a kernel. The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ associated to k is the completion of the pre-hilbert space

$$\left\{\sum_{i=1}^{n}\alpha_i k(x_i,.) : n\in\mathbb{N},\alpha_1,\ldots,\alpha_n\in\mathbb{R},x_1,\ldots,x_n\in\mathcal{X}\right\},$$

endowed with the inner product :

$$\left\langle\sum_{i=1}^{n}\alpha_i k(x_i,.),\sum_{j=1}^{m}\beta_i k(y_i,.)\right\rangle=\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j k(x_i,x_j).$$

RKHS

RKHS : k is a kernel. The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ associated to k is the completion of the pre-hilbert space

$$\left\{\sum_{i=1}^{n}\alpha_i k(x_i,.) : n\in\mathbb{N}, \alpha_1,\ldots,\alpha_n\in\mathbb{R}, x_1,\ldots,x_n\in\mathcal{X}\right\},$$

endowed with the inner product :

$$\left\langle\sum_{i=1}^{n}\alpha_i k(x_i,.),\sum_{j=1}^{m}\beta_i k(y_i,.)\right\rangle=\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j k(x_i,x_j).$$

Transfer function : $\phi:\mathcal{X}\mapsto\mathcal{H}, \phi(x)=k(x,.)$

Examples

- For the gaussian kernel

$$\mathcal{H}_\sigma = \left\{ f \in \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(w)|^2 \exp\left(\frac{\sigma^2 w^2}{2}\right) dw < \infty \right\}.$$

If $\mathcal{X} \subset \mathbb{R}^d$ is a compact subset then, $\mathcal{H}_\sigma$ is dense in $\mathcal{C}(\mathcal{X}, \mathbb{R})$.

Examples

- For the gaussian kernel

$$\mathcal{H}_\sigma = \left\{ f \in \mathcal{C}(\mathbb{R}^d, \mathbb{R}) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(w)|^2 \exp(\frac{\sigma^2 w^2}{2}) dw < \infty \right\}.$$

If $\mathcal{X} \subset \mathbb{R}^d$ is a compact subset then, $\mathcal{H}_\sigma$ is dense in $\mathcal{C}(\mathcal{X}, \mathbb{R})$.

- For $\quad k(x, x') = \min(x, x'), \quad \forall x, x' \in [0, 1]$

$$\mathcal{H} = \left\{ f \in \mathcal{C}([0,1], \mathbb{R}) \text{ a.e. differentiable}, f' \in L^2([0,1]), f(0) = 0 \right\}.$$

SVM Estimators

k : kernel on $\mathcal{X}$, $\mathcal{H}_k$: RKHS of k, $\lambda > 0$.

SVM Estimators

k : kernel on $\mathcal{X}$, $\mathcal{H}_k$: RKHS of k, $\lambda > 0$.

SVM Estimator :

$$\hat{f}_n^{\lambda} = \mathrm{Arg}\min_{f \in \mathcal{H}_k} \left(A_n^{\phi_1}(f) + \lambda ||f||^2_{\mathcal{H}_k} \right),$$

fo $A_n^{\phi_1}(f) = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+$

λ : regularization parameter.

SVM Estimators

k : kernel on $\mathcal{X}$, $\mathcal{H}_k$: RKHS of k, $\lambda > 0$.

SVM Estimator :

$$\hat{f}_n^{\lambda} = \mathrm{Arg}\min_{f \in \mathcal{H}_k} \left(A_n^{\phi_1}(f) + \lambda ||f||^2_{\mathcal{H}_k} \right),$$

fo $A_n^{\phi_1}(f) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i f(X_i))_+$

λ : regularization parameter.

SVM classifier :

$$\hat{F}_n^{\lambda}(x) = \mathrm{sign}(\hat{f}_n^{\lambda}(x)).$$

SVM Estimators

$$\hat{f}_n^{\lambda}(x) = \sum_{i=1}^{n} \hat{C}_i k(X_i, x), \forall x \in \mathcal{X},$$

where $\hat{C}_1, \ldots, \hat{C}_n$ are solutions of

$$\max_{0 \leq 2\lambda C_i Y_i \leq n^{-1}} \left\{ 2 \sum_{i=1}^{n} C_i Y_i - \sum_{i,j=1}^{n} C_i C_j k(X_i, X_j) \right\}.$$

Convergence rates for SVM

(GNA) Geometric noise assumption. *(Steinwart and Scovel)* $\mathcal{X} \subseteq \mathbb{R}^d$, $\exists C_1 > 0$ *and* $\gamma > 0$ *s.t.*

$$\mathbb{E}\left[|2\eta(X) - 1| \exp\left(-\frac{\tau(X)^2}{t}\right)\right] \leq C_1 t^{\frac{\gamma d}{2}}, \quad \forall t > 0.$$

$$\tau(x) = \begin{cases} d(x, G_0 \cup G_1), & \text{if } x \in G_{-1}, \\ d(x, G_0 \cup G_{-1}), & \text{if } x \in G_1, \\ 0, & \text{otherwise}, \end{cases} \quad \text{for all } x \in \mathcal{X},$$

$G_0 = \{x \in \mathcal{X} : \eta(x) = 1/2\}$,
$G_1 = \{x \in \mathcal{X} : \eta(x) > 1/2\}$,
$G_{-1} = \{x \in \mathcal{X} : \eta(x) < 1/2\}$.

Convergence rates for SVM

Théorème (Steinwart and Scovel (2005)) :

i) $\mathcal{X}$: unit ball of $\mathbb{R}^d$, k : gaussian kernel.

Convergence rates for SVM

Théorème (Steinwart and Scovel (2005)) :

i) $\mathcal{X}$: unit ball of $\mathbb{R}^d$, k : gaussian kernel.

ii) π satisfies ϕ_0-MA$((\alpha+1)/\alpha)$ and GNA(γ).

Convergence rates for SVM

Théorème (Steinwart and Scovel (2005)) :

i) $\mathcal{X}$: unit ball of $\mathbb{R}^d$, k : gaussian kernel.

ii) π satisfies ϕ_0-MA$((\alpha+1)/\alpha)$ and GNA(γ).

For the regularization parameter

$$\lambda_n^{\alpha,\gamma} = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} & \text{if } \gamma \leq \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2(\gamma+1)(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}} & \text{otherwise,} \end{cases}$$

and $\sigma_n^{\alpha,\gamma} = (\lambda_n^{\alpha,\gamma})^{-\frac{1}{(\gamma+1)d_0}}$ for window,

Convergence rates for SVM

Théorème (Steinwart and Scovel (2005)) :

i) $\mathcal{X}$: unit ball of $\mathbb{R}^d$, k : gaussian kernel.

ii) π satisfies ϕ_0−MA$((\alpha+1)/\alpha)$ and GNA(γ).

For the regularization parameter

$$\lambda_n^{\alpha,\gamma} = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} & \text{if } \gamma \leq \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2(\gamma+1)(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}} & \text{otherwise,} \end{cases}$$

and $\sigma_n^{\alpha,\gamma} = (\lambda_n^{\alpha,\gamma})^{-\frac{1}{(\gamma+1)d_0}}$ for window, the SVM estimator satisfies $(\forall \epsilon > 0)$

$$\mathbb{E}\left[A_0(\hat{F}_n^{(\sigma_n^{\alpha,\gamma},\lambda_n^{\alpha,\gamma})}) - A_0^*\right] \leq C_\epsilon \begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\epsilon} & \text{if } \gamma \leq \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2\gamma(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}+\epsilon} & \text{otherwise,} \end{cases}$$

Problem of adaptation

This estimator depends on α (margin) and γ (geometric margin)

Problem of adaptation

This estimator depends on α (margin) and γ (geometric margin)

$$\Downarrow$$

Problem : simultaneous adaptation to α and γ

Problem of adaptation

This estimator depends on α (margin) and γ (geometric margin)

$\Downarrow$

Problem : simultaneous adaptation to α and γ

$\Downarrow$

Aggregation methods

Adaptation Problem

Split the observations in two parts :

- $l = \left\lceil \frac{n}{\log n} \right\rceil$ et $m = n - l$.
- $D_m^1 = ((X_1, Y_1), \ldots, (X_m, Y_m))$ (training sample)

$$\downarrow$$

Construction of SVM estimators

$$\mathcal{F} = \left\{ \hat{F}_m^{(\sigma_k, \lambda_l)} : \begin{array}{ll} \sigma_k = m^{\frac{k}{2\Delta d_0}}, & k = 1, \ldots, 2\lfloor \Delta \rfloor \\ \lambda_l = m^{-(\frac{1}{2} + \frac{l}{\Delta})}, & l = 1, \ldots, \lfloor \Delta/2 \rfloor \end{array} \right\}$$

for $\Delta = \log n$.

- $D_l^2 = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n))$ (learning sample).

$$\downarrow$$

Construction of the weights :

$$w^{[l]}(F) = \frac{\exp\left(\sum_{i=m+1}^{n} Y_i F(X_i)\right)}{\sum_{\bar{F}\in\mathcal{F}} \exp\left(\sum_{i=m+1}^{n} Y_i \bar{F}(X_i)\right)}, \forall F \in \mathcal{F}.$$

- $D_l^2 = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n))$ (learning sample).

$$\downarrow$$

Construction of the weights :

$$w^{[l]}(F) = \frac{\exp\left(\sum_{i=m+1}^{n} Y_i F(X_i)\right)}{\sum_{\bar{F}\in\mathcal{F}} \exp\left(\sum_{i=m+1}^{n} Y_i \bar{F}(X_i)\right)}, \forall F \in \mathcal{F}.$$

- $\tilde{F}_n^{adp} = \mathrm{sign}(\tilde{f}_n^{adp})$ where

$$\tilde{f}_n^{adp} = \sum_{F\in\mathcal{F}} w^{[l]}(F)F,$$

Theorem (adaptive SVM)

Two compact subsets $K \subset \mathcal{U} = \{(\alpha,\gamma) \in (0,+\infty)^2 : \gamma > \frac{\alpha+2}{2\alpha}\}$ and $K' \subset \mathcal{U}' = \{(\alpha,\gamma) \in (0,+\infty)^2 : \gamma \leq \frac{\alpha+2}{2\alpha}\}$. We have $(\forall \epsilon > 0)$,

$$\sup_{\pi \in \mathcal{P}_{\alpha,\gamma}} \mathbb{E}\left[A_0(\tilde{F}_n^{adp}) - A_0^*\right] \leq C_\epsilon \begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\epsilon} & \text{if } (\alpha,\gamma) \in K', \\ n^{-\frac{2\gamma(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}+\epsilon} & \text{if } (\alpha,\gamma) \in K, \end{cases}$$

$\forall (\alpha,\gamma) \in K \cup K'$ and $\mathcal{P}_{\alpha,\gamma}$ is the set of probility measure on $\mathcal{X} \times \{-1,1\}$ satisfying MA(α) and GNA(γ).