

# Suboptimality of Penalized Empirical Risk Minimization in Classification.

Guillaume Lécué

Université Paris 6

COLT 2007, June 13

# Motivation.

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

$n$  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  $f_1, \dots, f_M$ .

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

Estimation :

$\epsilon$ -net :  $f_1, \dots, f_M$  (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  $\mathcal{X} \times \{-1, 1\}$ ,

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

## Model of classification

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

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

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

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

## Model of classification

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

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

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

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

$f : \mathcal{X} \mapsto \{-1, 1\}$  : 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** :  $x \in \mathcal{X} \rightarrow$  label  $y \in \{-1, 1\}$ ?

$f : \mathcal{X} \mapsto \{-1, 1\}$  : 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  $\mathcal{X} \times \{-1, 1\}$ ,

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

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

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

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

**Bayes rule** :  $f^*(x) = \text{Sign}(2\eta(x) - 1)$  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** :  $x \in \mathcal{X} \rightarrow$  label  $y \in \{-1, 1\}$ ?

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

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

**Bayes rule** :  $f^*(x) = \text{Sign}(2\eta(x) - 1)$  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  $\mathcal{X} \times \{-1, 1\}$ ,

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

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

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

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

**Bayes rule** :  $f^*(x) = \text{Sign}(2\eta(x) - 1)$  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  $f^*$ .

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

## Model of classification

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

$$\phi_0(x) = \mathbb{I}_{(x \leq 0)}$$

classical loss or 0 – 1 loss

# Model of classification

$(f : \mathcal{X} \mapsto \mathbb{R}) \rightarrow \text{risk } A_0(f) = \mathbb{E}[\phi_0(Yf(X))]$  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)

$$x \mapsto \log_2(1 + \exp(-x))$$

'Logit-Boosting' loss

$$x \mapsto \exp(-x)$$

exponential Boosting loss

$$x \mapsto (1 - x)^2$$

quadratic loss

$$x \mapsto \max(0, 1 - x)^2$$

2-norm soft margin loss

# Model of classification

$(f : \mathcal{X} \mapsto \mathbb{R}) \rightarrow \text{risk } A_0(f) = \mathbb{E}[\phi_0(Yf(X))]$  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)

$$x \mapsto \log_2(1 + \exp(-x))$$

'Logit-Boosting' loss

$$x \mapsto \exp(-x)$$

exponential Boosting loss

$$x \mapsto (1 - x)^2$$

quadratic loss

$$x \mapsto \max(0, 1 - x)^2$$

2-norm soft margin loss

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

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

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

# Selectors.

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

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

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

# Selectors.

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

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

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

- **penalized Empirical Risk Minimization (pERM)** :

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

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

## Aggregation methods with exponential weights.

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

- Aggregate with Exponential weights (AEW) :

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

$T^{-1}$  : temperature parameter.

## Aggregation methods with exponential weights.

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

- Aggregate with Exponential weights (AEW) :

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

$T^{-1}$  : 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

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

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

## 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  $\bar{f}_n$ ,  $\exists \pi \in \mathcal{P}$ ,  $\forall n \geq 1$

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

## 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  $\forall \pi \in \mathcal{P}$ ,  $\forall n \geq 1$

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

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

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

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

## Aim of Aggregation(2) : Adaptation.

### Definition (Oracle Inequality)

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

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

where  $C \geq 1$ .

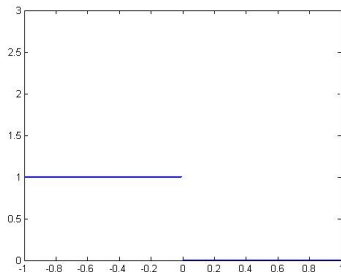
## Continuous scale of loss functions.

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

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

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.

$h = 0$



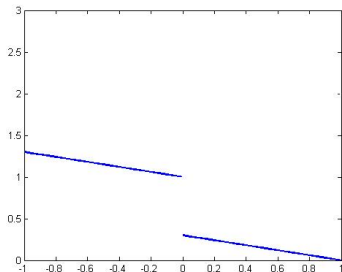
## Continuous scale of loss functions.

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

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

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

$$h = 1/3$$



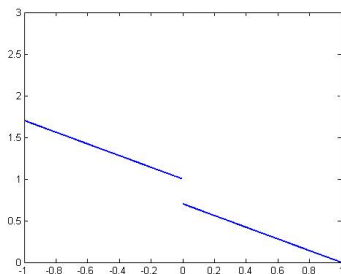
## Continuous scale of loss functions.

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

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

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

$$h = 2/3$$



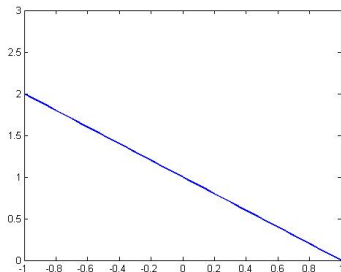
## Continuous scale of loss functions.

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

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

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.

$h = 1$



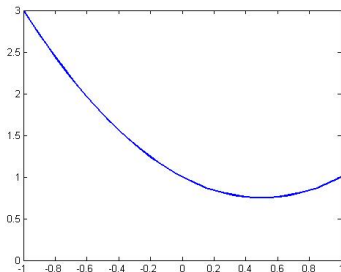
## Continuous scale of loss functions.

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

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

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.

$h = 2$



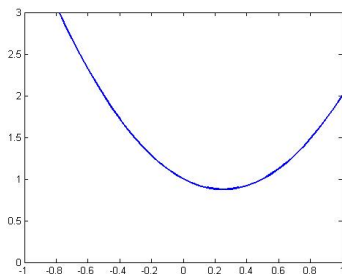
## Continuous scale of loss functions.

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

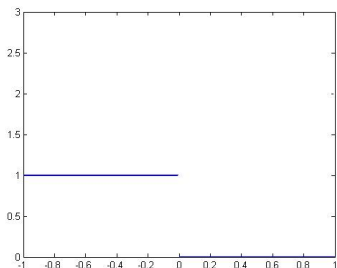
$$\phi(x) = \phi_h(x) = \begin{cases} (1-h)\phi_0(x) + h\phi_1(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \quad \forall x \in \mathbb{R}$$

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.

$h = 3$

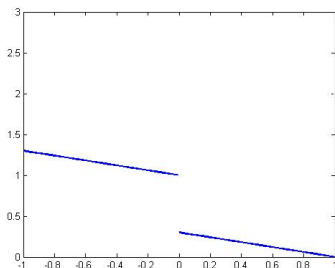


## ORA in classification



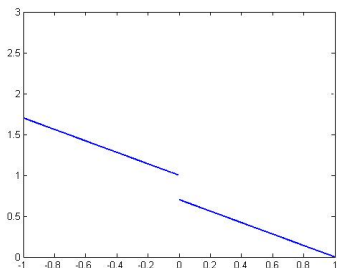
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

## ORA in classification



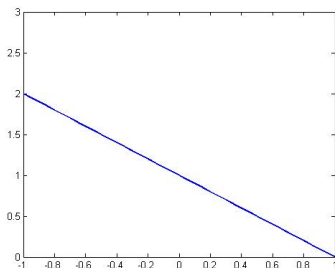
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

## ORA in classification



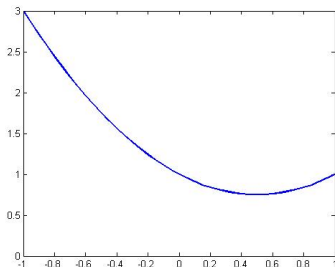
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

## ORA in classification



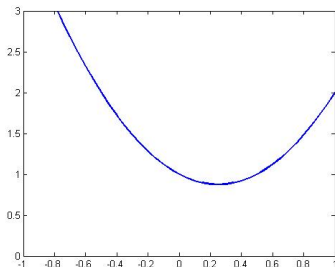
Loss function	$0 \leq h < 1$	$h = 1$	$h > 1$
Optimal rate of aggregation (ORA)	$\sqrt{\frac{\log M}{n}}$	$\sqrt{\frac{\log M}{n}}$	$\frac{\log M}{n}$
Optimal aggregation procedure	ERM	ERM, AEW, CAEW	CAEW

## 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 ?

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

## 2 Questions.

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

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

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 ?

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

ERM  $\longrightarrow$  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  $\phi$  :

The probability measure  $\pi$  satisfies the  $\phi$ -margin assumption  $\phi$ -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  $f : \mathcal{X} \mapsto \mathbb{R}$ .

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  $\phi$  :

The probability measure  $\pi$  satisfies the  $\phi$ -margin assumption  $\phi$ -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  $f : \mathcal{X} \mapsto \mathbb{R}$ .

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

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

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

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

Margin assumption for the loss function  $\phi$  :

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

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

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

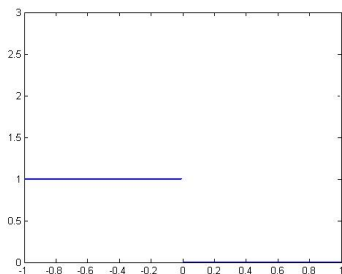
cf. Mammen and Tsybakov 99 (discriminant analysis) and Tsybakov 04 (classification).

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

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

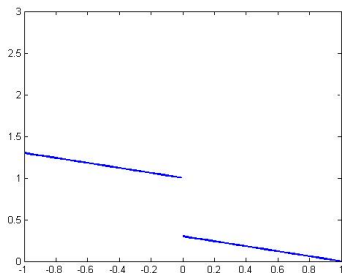
$$(\kappa = 1 \iff \exists h > 0, |2\eta(X) - 1| \geq h)$$

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



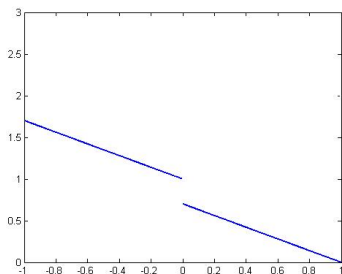
$$\kappa = +\infty \text{ for any } 0 \leq h \leq 1.$$

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



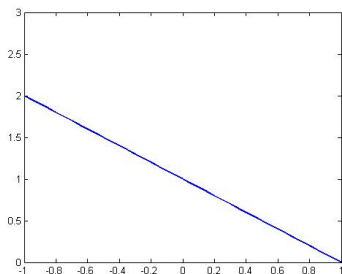
$$\kappa = +\infty \text{ for any } 0 \leq h \leq 1.$$

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



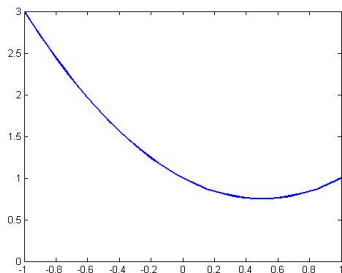
$\kappa = +\infty$  for any  $0 \leq h \leq 1$ .

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



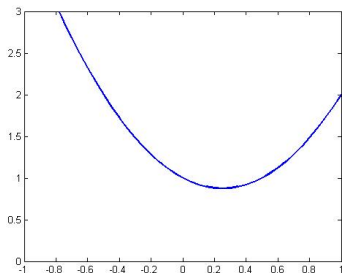
$\kappa = +\infty$  for any  $0 \leq h \leq 1$ .

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

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

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

## Theorem (suboptimality of selectors under the margin assumption)

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

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

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

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

### Suboptimality of Penalized ERM.

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

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

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

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

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

## Conclusion of optimality

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

## Conclusion of 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.