An aggregation procedure in classification

Application to adaptivity

Guillaume Lecué

lecue@ccr.jussieu.fr

Laboratoire de Probabilités et modèles aléatoires

Université Paris 6

Classification: Model

- **Framework:**
  - (X, Y): a random variable $\sim$ probability measure π on $\mathcal{X} \times \{-1, 1\}$.
  - $D_n = (X_i, Y_i)_{i=1,\dots,n}$: a set of n i.i.d. observations of (X, Y).
  - Prediction rule $f : \mathcal{X} \longmapsto \{-1, 1\}$
  - Risk of f: $R(f) = \mathbb{P}(f(X) \neq Y)$.
- **Bayes rule:** f^* minimizes the risk $R(f)$ over all prediction rules,

$$f^*(x) = \mathrm{sign}(2\eta(x) - 1), \eta(x) = \mathbb{P}(Y = 1 | X = x),$$

the Bayes risk: $R^* \stackrel{\text{def}}{=} R(f^*) = \min_f R(f)$.

Classification: Model

- **Classifier:** a procedure, that assigns to observations D_n a prediction rule $\hat{f}_n(\cdot, D_n) : \mathcal{X} \longmapsto \{-1, 1\}$. The excess risk of a classifier $\hat{f}_n$ is the value

$$\mathbb{E}[R(\hat{f}_n) - R^*].$$

- **Rate of convergence:** For a set $\mathcal{P}$ of probability measures on $\mathcal{X} \times \{-1, 1\}$, a classifier $\hat{f}_n$ learns with the rate of convergence $\phi(n)$ over $\mathcal{P}$, if

$$\sup_{\pi \in \mathcal{P}} \mathbb{E}[R(\hat{f}_n) - R^*] \leq C\phi(n), \quad \forall n \geq 1.$$

Classification: results

No classifier can guarantee a rate of convergence that holds for all probability distributions π (Devroye, Györfi and Lugosi, 1996).

Classification: results

No classifier can guarantee a rate of convergence that holds for all probability distributions π (Devroye, Györfi and Lugosi, 1996).

- Complexity assumption on the class of decision sets $\{x \in \mathcal{X} : f^*(x) = 1\} = \{x \in \mathcal{X} : \eta(x) \geq 1/2\}$ ($VC-$dimension, metric entropy).

$$n^{-1/2}, \text{ (up to a logarithm)}$$

Classification: results

No classifier can guarantee a rate of convergence that holds for all probability distributions π (Devroye, Györfi and Lugosi, 1996).

- Complexity assumption on the class of decision sets $\{x \in \mathcal{X} : f^*(x) = 1\} = \{x \in \mathcal{X} : \eta(x) \geq 1/2\}$ ($VC-$dimension, metric entropy).

$$n^{-1/2}, \text{ (up to a logarithm)}$$

- Complexity assumption on the class of conditional probability functions η (smoothness).

$$n^{-\alpha}, 0 < \alpha \leq 1/2, \text{ (up to a logarithm)}$$

Classification: results

No classifier can guarantee a rate of convergence that holds for all probability distributions π (Devroye, Györfi and Lugosi, 1996).

- Complexity assumption on the class of decision sets $\{x \in \mathcal{X} : f^*(x) = 1\} = \{x \in \mathcal{X} : \eta(x) \geq 1/2\}$ ($VC-$dimension, metric entropy).

$$n^{-1/2}, \text{ (up to a logarithm)}$$

- Complexity assumption on the class of conditional probability functions η (smoothness).

$$n^{-\alpha}, 0 < \alpha \leq 1/2, \text{ (up to a logarithm)}$$

No convergence rates faster than $n^{-1/2}$ can be expected if only complexity assumptions are supposed (DGL 96).

Classification: margin assumption

Excess risk is sensitive to the behavior of $\eta(x)$ near the decision boundary $\{x : \eta(x) = 1/2\}$.

Classification: margin assumption

Excess risk is sensitive to the behavior of $\eta(x)$ near the decision boundary $\{x : \eta(x) = 1/2\}$.

Margin assumption MA(α), $0 \le \alpha < +\infty$, (Tsybakov, 2004)

$$\forall t > 0, \quad \mathbb{P}(|2\eta(X) - 1| \le t) \le c_0 t^{\alpha}.$$

Classification: margin assumption

Excess risk is sensitive to the behavior of $\eta(x)$ near the decision boundary $\{x : \eta(x) = 1/2\}$.

Margin assumption MA(α), $0 \leq \alpha < +\infty$, (Tsybakov, 2004)

$$\forall t > 0, \quad \mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^{\alpha}.$$

Under MA(α), we can expect fast rates:

- Tsybakov (2004): $n^{-\frac{\alpha+1}{\alpha+\alpha\rho+2}}$ under MA(α) for massive classes of decision sets (polynomial entropies increasing as $\epsilon^{-\rho}$, $0 < \rho < 1$). Can approach n^{-1} as $\alpha \to +\infty$ and $\rho \to 0$.

Further results on fast rates

- Blanchard, Lugosi and Vayatis (2003)
- Bartlett, Jordan and McAuliffe (2003)
- Blanchard, Bousquet and Massart (2004)
- Nédélec and Massart (2005)
- Scovel and Steinwart (2004, 2005)
- Audibert and Tsybakov (2005)
- Koltchinskii (2005)
- Herbei and Wegkamp (2005)

(non-adaptive)

Classification: adaptivity

If one wants to achieve fast rates, two types of assumptions are needed:

Classification: adaptivity

If one wants to achieve fast rates, two types of assumptions are needed:

- A complexity assumption ($VC-$dimension, entropy)
  $\Longrightarrow \quad \rho \quad$: complexity parameter.

Classification: adaptivity

If one wants to achieve fast rates, two types of assumptions are needed:

- A complexity assumption ($VC-$dimension, entropy) $\Longrightarrow \quad \rho \quad$: complexity parameter.
- A margin assumption (behavior of η near the decision boundary) $\Longrightarrow \quad \alpha \quad$: margin parameter.

Classification: adaptivity

If one wants to achieve fast rates, two types of assumptions are needed:

- A complexity assumption ($VC-$dimension, entropy) $\Longrightarrow \quad \rho \quad$: complexity parameter.
- A margin assumption (behavior of η near the decision boundary) $\Longrightarrow \quad \alpha \quad$: margin parameter.

α and ρ are not known in practice

$$\Downarrow$$

Problem of adaptivity.

Classification: adaptivity

If one wants to achieve fast rates, two types of assumptions are needed:

- A complexity assumption ($VC-$dimension, entropy) $\Longrightarrow \quad \rho \quad$: complexity parameter.
- A margin assumption (behavior of η near the decision boundary) $\Longrightarrow \quad \alpha \quad$: margin parameter.

α and ρ are not known in practice

$$\Downarrow$$

Problem of adaptivity.

Tsybakov (2004); Tsybakov and Van De Geer (2005); Tarigan and Van De Geer (2005); Audibert (2005); Koltchinskii (2005)... Not easy to compute.

Aggregation procedure

Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite set of prediction rules. We define the Aggregation Procedure with Exponential Weights (AEW) by:

$$\boxed{\tilde{f}_n = \sum_{f \in \mathcal{F}} w^{(n)}(f) f,}$$

where, for any prediction rule f, $R_n(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f(X_i) \neq Y_i}$,

$$\boxed{w^{(n)}(f) = \frac{\exp\left(-2nR_n(f)\right)}{\sum_{g \in \mathcal{F}} \exp\left(-2nR_n(g)\right)}, \ \forall f \in \mathcal{F}.}$$

Aggregation procedure

Let $\mathcal{F} = \{f_1, \dots, f_M\}$ be a finite set of prediction rules. We define the Aggregation Procedure with Exponential Weights (AEW) by:

$$\boxed{\tilde{f}_n = \sum_{f \in \mathcal{F}} w^{(n)}(f) f,}$$

where, for any prediction rule f, $R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i}$,

$$\boxed{w^{(n)}(f) = \frac{\exp\left(-2nR_n(f)\right)}{\sum_{g \in \mathcal{F}} \exp\left(-2nR_n(g)\right)}, \ \forall f \in \mathcal{F}.}$$

The classifier that we propose is: $\tilde{F}_n = \mathrm{sign}(\tilde{f}_n)$

Optimal rate of aggregation

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under MA(α).

Optimal rate of aggregation

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under MA(α).

- $\forall \mathcal{F} = \{f_1, \dots, f_M\}$, $\exists f_n^*$ such that $\forall \pi \in MA(\alpha)$, $\forall n \geq 1$

$$\mathbb{E}\left[R(f_n^*) - R^*\right] \leq \min_{f \in \mathcal{F}} \left(R(f) - R^*\right) + C_1 \gamma(n, M, \alpha, \mathcal{F}, \pi).$$

Optimal rate of aggregation

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under MA(α).

- $\forall \mathcal{F} = \{f_1, \dots, f_M\}$, $\exists f_n^*$ such that $\forall \pi \in MA(\alpha)$, $\forall n \geq 1$

$$\mathbb{E}\left[R(f_n^*) - R^*\right] \leq \min_{f \in \mathcal{F}} \left(R(f) - R^*\right) + C_1 \gamma(n, M, \alpha, \mathcal{F}, \pi).$$

- $\exists \mathcal{F} = \{f_1, \dots, f_M\}$ such that for any classifier $\bar{f}_n$, $\exists \pi \in MA(\alpha)$, $\forall n \geq 1$

$$\mathbb{E}\left[R(\bar{f}_n) - R^*\right] \geq \min_{f \in \mathcal{F}} \left(R(f) - R^*\right) + C_2 \gamma(n, M, \alpha, \mathcal{F}, \pi).$$

Optimal rate of aggregation

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under MA(α).

- $\forall \mathcal{F} = \{f_1, \ldots, f_M\}$, $\exists f_n^*$ such that $\forall \pi \in MA(\alpha)$, $\forall n \geq 1$

$$\mathbb{E}\left[R(f_n^*) - R^*\right] \leq \min_{f \in \mathcal{F}} \left(R(f) - R^*\right) + C_1 \gamma(n, M, \alpha, \mathcal{F}, \pi).$$

- $\exists \mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any classifier $\bar{f}_n$, $\exists \pi \in MA(\alpha)$, $\forall n \geq 1$

$$\mathbb{E}\left[R(\bar{f}_n) - R^*\right] \geq \min_{f \in \mathcal{F}} \left(R(f) - R^*\right) + C_2 \gamma(n, M, \alpha, \mathcal{F}, \pi).$$

$\Longrightarrow \quad \gamma(n, M, \alpha, \mathcal{F}, \pi)$: Optimal rate of aggregation.

Hinge risk

For any function $f : \mathcal{X} \longmapsto \mathbb{R}$, the hinge risk of f is

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}[(1 - Yf(X))_+].$$

Hinge risk

For any function $f : \mathcal{X} \longmapsto \mathbb{R}$, the hinge risk of f is

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}[(1 - Yf(X))_+].$$

- (Bayes rule) $f^* = \operatorname{Arg\,min}_{f:\mathcal{X}\longmapsto\mathbb{R}} A(f), \quad A^* \stackrel{\text{def}}{=} A(f^*)$,

Hinge risk

For any function $f : \mathcal{X} \longmapsto \mathbb{R}$, the hinge risk of f is

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}[(1 - Yf(X))_+].$$

- (Bayes rule) $f^* = \text{Arg}\min_{f:\mathcal{X}\longmapsto\mathbb{R}} A(f), \quad A^* \stackrel{\text{def}}{=} A(f^*)$,
- Zhang (2004): $R(f) - R^* \leq A(f) - A^*$ for any f with values in $\mathbb{R}$.

Hinge risk

For any function $f : \mathcal{X} \longmapsto \mathbb{R}$, the hinge risk of f is

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}[(1 - Yf(X))_+].$$

- (Bayes rule) $f^* = \mathrm{Arg\,min}_{f:\mathcal{X}\longmapsto\mathbb{R}} A(f), \quad A^* \stackrel{\text{def}}{=} A(f^*)$,
- Zhang (2004): $R(f) - R^* \leq A(f) - A^*$ for any f with values in $\mathbb{R}$.
- $2(R(f) - R^*) = A(f) - A^*$ for any prediction rule $f : \mathcal{X} \longmapsto \{-1, 1\}$.

Optimal rate of aggregation for hinge

Theorem 1 (Oracle inequality). *We assume that π satisfies MA(α). Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a set of prediction rules. The AEW procedure satisfies for any integer $n \geq 1$:*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{F}}(A(f) - A^*) +$$

$$C_1 \left(\sqrt{\frac{(\min_{f \in \mathcal{F}} A(f) - A^*)^{\frac{\alpha}{1+\alpha}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \right),$$

where $C_1 > 0$ is a constant depending only on the constants α and c_0 appearing in the margin assumption.

Optimal rate of aggregation for hinge

Theorem 1 (Oracle inequality). *We assume that π satisfies MA(α). Let $\mathcal{F} = \{f_1, \dots, f_M\}$ be a set of prediction rules. The AEW procedure satisfies for any integer $n \geq 1$:*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{F}}(A(f) - A^*) +$$

$$C_1 \left(\sqrt{\frac{(\min_{f \in \mathcal{F}} A(f) - A^*)^{\frac{\alpha}{1+\alpha}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \right),$$

where $C_1 > 0$ is a constant depending only on the constants α and c_0 appearing in the margin assumption.

Remark: Denote by $\mathcal{C}$ the convex hull of $\mathcal{F}$.

$$\min_{f \in \mathcal{F}} A(f) - A^* = \min_{f \in \mathcal{C}} A(f) - A^*.$$

Optimal rate of aggregation for hinge

Theorem 2 (Lower bound). *There exits $\mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any statistic $\bar{f}_n$ with values in $\mathbb{R}$, there exists a probability measure π MA(α) such that for any n, M satisfying $\log M \leq n$,*

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{F}}(A(f) - A^*) +$$

$$C_2 \left(\sqrt{\frac{(\min_{f\in\mathcal{F}} A(f) - A^*)^{\frac{\alpha}{1+\alpha}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \right),$$

where $C_2 > 0$ is a constant depending only on the constants α and c_0 appearing in the margin assumption MA(α).

Optimal rate of aggregation for hinge

$$\sqrt{\frac{\mathcal{M}(\mathcal{F},\pi)^{\frac{\alpha}{1+\alpha}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\mathcal{M}(\mathcal{F},\pi)=\min_{f\in\mathcal{F}}A(f)-A^*$.

Optimal rate of aggregation for hinge

$$\sqrt{\frac{\mathcal{M}(\mathcal{F},\pi)^{\frac{\alpha}{1+\alpha}}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\mathcal{M}(\mathcal{F},\pi) = \min_{f\in\mathcal{F}} A(f) - A^*$.

- $\mathcal{M}(\mathcal{F},\pi) \preceq \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \Longrightarrow$ rate $\asymp \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$.

Optimal rate of aggregation for hinge

$$\sqrt{\frac{\mathcal{M}(\mathcal{F},\pi)^{\frac{\alpha}{1+\alpha}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\mathcal{M}(\mathcal{F},\pi)=\min_{f\in\mathcal{F}}A(f)-A^*$.

- $\mathcal{M}(\mathcal{F},\pi)\preceq\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \Longrightarrow$ rate $\asymp\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$.
- $\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}\preceq\mathcal{M}(\mathcal{F},\pi)\Rightarrow\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}\preceq$ rate $\preceq\sqrt{\frac{\log M}{n}}$

Optimal rate of aggregation for hinge

$$\sqrt{\frac{\mathcal{M}(\mathcal{F},\pi)^{\frac{\alpha}{1+\alpha}}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\mathcal{M}(\mathcal{F},\pi) = \min_{f\in\mathcal{F}} A(f) - A^*$.

- $\mathcal{M}(\mathcal{F},\pi) \preceq \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \Longrightarrow$ rate $\asymp \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$.
- $\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \preceq \mathcal{M}(\mathcal{F},\pi) \Rightarrow \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \preceq$ rate $\preceq \sqrt{\frac{\log M}{n}}$
- No margin assumption ($\alpha = 0$) or
  $\mathcal{M}(\mathcal{F},\pi) \geq a > 0 \Longrightarrow$ rate $\asymp \sqrt{\frac{\log M}{n}}$.

Construction of Classifiers

Hölder class

The d-dimensional Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ ($\beta, L > 0$).

Hölder class

The d-dimensional Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ $(\beta, L > 0)$.

- $g : \mathbb{R}^d \longmapsto \mathbb{R}$, $\lfloor \beta \rfloor$-times continuously differentiable.

Hölder class

The d-dimensional Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ ($\beta, L > 0$).

- $g : \mathbb{R}^d \longmapsto \mathbb{R}$, $\lfloor \beta \rfloor$-times continuously differentiable.
- $\forall x, y \in \mathbb{R}^d, |g(y) - g_x(y)| \leq L||x - y||_2^\beta,$

where

$$g_x(y) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(y-x)^s}{s!} D^s g(x)$$

is the Taylor polynomial of degree $\lfloor \beta \rfloor$ for g at point x.

Hölder class

The d-dimensional Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ ($\beta, L > 0$).

- $g : \mathbb{R}^d \longmapsto \mathbb{R}$, $\lfloor \beta \rfloor$-times continuously differentiable.
- $\forall x, y \in \mathbb{R}^d, |g(y) - g_x(y)| \leq L||x - y||_2^\beta,$

where

$$g_x(y) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(y - x)^s}{s!} D^s g(x)$$

is the Taylor polynomial of degree $\lfloor \beta \rfloor$ for g at point x.

- $\epsilon-$entropy of the Hölder class:

$$\log(\mathcal{N}(\Sigma(\beta, L, \mathbb{R}^d), \epsilon, L^\infty([0, 1]^d))) \leq A\epsilon^{-d/\beta}, \forall \epsilon > 0.$$

Aggregation over a sieve

Define the class of models $\mathcal{P}_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

Aggregation over a sieve

Define the class of models $\mathcal{P}_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**

Aggregation over a sieve

Define the class of models $\mathcal{P}_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**
- **The conditional probability function** $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$.

Aggregation over a sieve

Define the class of models $\mathcal{P}_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**
- **The conditional probability function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$.**
- **The marginal distribution of X is supported on $[0,1]^d$ and has a Lebesgue density upper bounded by a constant.**

Aggregation over a sieve

$\Sigma_\epsilon(\beta)$: ϵ-net of $\Sigma(\beta, L, \mathbb{R}^d)$ for the $L^\infty-$norm on $[0,1]^d$, such that:

$$\log \mathrm{Card}(\Sigma_\epsilon(\beta)) \leq A\epsilon^{-d/\beta}.$$

Aggregation over a sieve

$\Sigma_\epsilon(\beta)$: ϵ-net of $\Sigma(\beta, L, \mathbb{R}^d)$ for the $L^\infty-$norm on $[0,1]^d$, such that:

$$\log \mathrm{Card}(\Sigma_\epsilon(\beta)) \leq A\epsilon^{-d/\beta}.$$

- $$\tilde{f}_n^{(\epsilon,\beta)} = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta,$$ where $f_\eta(x) = \mathrm{Sign}\left(\eta(x) - 1/2\right).$

Aggregation over a sieve

$\Sigma_\epsilon(\beta)$: ϵ-net of $\Sigma(\beta, L, \mathbb{R}^d)$ for the $L^\infty-$norm on $[0,1]^d$, such that:

$$\log \mathrm{Card}(\Sigma_\epsilon(\beta)) \leq A\epsilon^{-d/\beta}.$$

- $$\tilde{f}_n^{(\epsilon,\beta)} = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta, \text{ where } f_\eta(x) = \mathrm{Sign}\left(\eta(x) - 1/2\right).$$

- We chose the step of the ϵ-net by a trade-off:

$$\epsilon_n = n^{-\frac{\beta}{\beta(\alpha+2)+d}}.$$

Aggregation over a sieve

Theorem 3: *Let* $\alpha \geq 0$ *and* $\beta > 0$*. Let* $a_1 > 0$ *be an absolute constant, we consider* $\epsilon_n = a_1 n^{-\frac{\beta}{\beta(\alpha+2)+d}}$*, then, the sign of the aggregate* $\tilde{f}_n^{(\epsilon_n,\beta)}$ *satisfies, for any* $\pi \in \mathcal{P}_{\beta,\alpha}$ *and integer* $n > 0$*,*

$$\mathbb{E}_\pi \left[R(\mathrm{Sign}(\tilde{f}_n^{(\epsilon_n,\beta)})) - R^* \right] \leq C_3(\alpha, \beta, d) n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}},$$

where $C_3(\alpha, \beta, d) > 0$*.*

Audibert and Tsybakov (2005) have shown the optimality, in a minimax sense, of this rate.

Problem of Adaptivity

Construction of the classifier $\mathrm{Sign}(\tilde{f}_n^{(\epsilon_n,\beta)})$ needs to know the parameters α and β which are not available in practice.

$$\Downarrow$$

Problem of adaptivity with respect to α and β.

Idea: We aggregate classifiers $\tilde{f}_n^{(\epsilon,\beta)}$, for different values of (ϵ, β) lying in a finite grid.

Aggregation of Aggregate-Classifiers

Adaptivity

We use a split of the sample to construct our adaptive classifier:

Adaptivity

We use a split of the sample to construct our adaptive classifier:

- $l = \left\lceil \frac{n}{\log n} \right\rceil$ and $m = n - l$.

Adaptivity

We use a split of the sample to construct our adaptive classifier:

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

$$\downarrow$$

Construction of the class of aggregate-classifiers

$$\mathcal{F} = \left\{ \mathrm{Sign}(\tilde{f}_m^{(\epsilon_m^k, \beta_p)}) : \begin{array}{l} \epsilon_m^k = m^{-k/\Delta} : k \in \{1, \ldots, \lfloor \Delta/2 \rfloor\} \\ \beta_p = p/\Delta : p \in \{1, \ldots, \lceil \Delta \rceil^2\} \end{array} \right\},$$

where $\Delta_n = \log n$.

Adaptivity

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

$$\downarrow$$

Construction of the weights:

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

$$F \in \mathcal{F} = \left\{ \mathrm{Sign}(\tilde{f}_m^{(\epsilon_m^k, \beta_p)}) : \begin{array}{l} \epsilon_m^k = m^{-k/\Delta} : k \in \{1, \ldots, \lfloor \Delta/2 \rfloor\} \\ \beta_p = p/\Delta : p \in \{1, \ldots, \lceil \Delta \rceil^2\} \end{array} \right\},$$

Adaptivity

The classifier that we propose is $\tilde{F}_n^{adp} = \mathrm{Sign}(\tilde{f}_n^{adp})$, where:

Adaptivity

The classifier that we propose is $\tilde{F}_n^{adp} = \mathrm{Sign}(\tilde{f}_n^{adp})$, where:

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

Adaptivity

The classifier that we propose is $\tilde{F}_n^{adp} = \mathrm{Sign}(\tilde{f}_n^{adp})$, where:

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

and

$$\mathcal{F} = \left\{ \mathrm{Sign}(\tilde{f}_m^{(\epsilon_m^k, \beta_p)}) : \begin{array}{l} \epsilon_m^k = m^{-k/\Delta} : k \in \{1, \dots, \lfloor \Delta/2 \rfloor\} \\ \beta_p = p/\Delta : p \in \{1, \dots, \lceil \Delta \rceil^2\} \end{array} \right\},$$

$\tilde{f}_n^{(\epsilon,\beta)} = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta$ is the aggregate over the minimal sieve $\Sigma_\epsilon(\beta)$ over $\Sigma(\beta, L, \mathbb{R}^d)$ for the $L^\infty-$norm.

Adaptivity

Theorem 4. *Let K be a compact subset of $(0,+\infty)\times(0,+\infty)$. There exists a constant $C_4>0$ depending only on K and d such that for any integer $n\geq 1$, any $(\alpha,\beta)\in K$ and any $\pi\in\mathcal{P}_{\beta,\alpha}$, we have,*

$$\mathbb{E}_\pi\left[R(\tilde{F}_n^{adp})-R^*\right]\leq C_4 n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}}.$$

Recall: $n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}}$ is an optimal rate of convergence for the model $\mathcal{P}_{\beta,\alpha}$.

Adaptivity

Theorem 4. *Let K be a compact subset of $(0,+\infty)\times(0,+\infty)$. There exists a constant $C_4>0$ depending only on K and d such that for any integer $n\geq 1$, any $(\alpha,\beta)\in K$ and any $\pi\in\mathcal{P}_{\beta,\alpha}$, we have,*

$$\mathbb{E}_\pi\left[R(\tilde{F}_n^{adp})-R^*\right]\leq C_4 n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}}.$$

Recall: $n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}}$ is an optimal rate of convergence for the model $\mathcal{P}_{\beta,\alpha}$.

Problem: The aggregate $\tilde{f}_n^{(\epsilon,\beta)}$ are not realizable in practice.

Adaptive SVM

Adaptivity

Aggregation of $L1-$SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

Adaptivity

Aggregation of $L1-$SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

Adaptivity

Aggregation of $L1-$SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

$$\Downarrow$$

Problem: simultaneous adaptation to the margin α and to geometry exponent γ.

Adaptivity

Aggregation of $L1-$SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

$$\Downarrow$$

Problem: simultaneous adaptation to the margin α and to geometry exponent γ.

We use our aggregation procedure to construct adaptive classifiers both to the margin and to geometry.

Adaptivity

Aggregation of $L1-$SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

$$\Downarrow$$

Problem: simultaneous adaptation to the margin α and to geometry exponent γ.

We use our aggregation procedure to construct adaptive classifiers both to the margin and to geometry.

We aggregate classifiers for different values of α and γ in a finite grid, thus giving an adaptive version of the result of Scovel and Steinwart (2004).

Conclusion

The Aggregation procedure with Exponential Weights:

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.
- achieves optimal rates of aggregation under the margin assumption.

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.
- achieves optimal rates of aggregation under the margin assumption.
- can be used to achieve simultaneous adaptation to the margin and to complexity with fast rates.

Remark

Consider $\mathcal{P}_1$ the model made of all underlying probability measure on $[0,1]^d \times \{-1,1\}$ such that:

- $\pi^X = \lambda_d$ (Lebesgue probability measures on $[0,1]^d$).
- π satisfies MA(∞) $\Leftrightarrow |2\eta(X) - 1| \geq h$ a.s.. Assume $h = 1$ $\Leftrightarrow Y = f^*(X) = \eta(X) \Leftrightarrow R^* = 0$.

Theorem 1. *For any classifier $\bar{f}_n$ constructed from a sample of size n, we have*

$$\sup_{\pi \in \mathcal{P}_1} \mathbb{E}[R(\bar{f}_n) - R^*] \geq \frac{1}{8e}$$

Corollaries for excess risk

Using Zhang's inequality, we obtain:

Corollaries for excess risk

Using Zhang's inequality, we obtain:

• $\forall \mathcal{F} = \{f_1, \ldots, f_M\}$, the (AEW) procedure satisfies for any π satisfying MA(α), $\forall n \geq 1, a > 0$

$$\mathbb{E}\left[R(\tilde{F}_n) - R^*\right] \leq 2(1+a)\min_{f\in\mathcal{F}}(R(f) - R^*) + C_1(a)\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

Corollaries for excess risk

Using Zhang's inequality, we obtain:

•$\forall \mathcal{F} = \{f_1, \dots, f_M\}$, the (AEW) procedure satisfies for any π satisfying MA(α), $\forall n \geq 1, a > 0$

$$\mathbb{E}\left[R(\tilde{F}_n) - R^*\right] \leq 2(1+a)\min_{f \in \mathcal{F}}(R(f) - R^*) + C_1(a)\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

•$\exists \mathcal{F} = \{f_1, \dots, f_M\}$ such that for any classifier $\bar{f}_n$, $\exists \pi$ satisfying MA(α), $\forall n \geq 1, a > 0$

$$\mathbb{E}\left[R(\bar{f}_n) - R^*\right] \geq 2(1+a)\min_{f \in \mathcal{F}}(R(f) - R^*) + C_2(a)\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

Corollaries for excess risk

Using Zhang's inequality, we obtain:

•$\forall \mathcal{F} = \{f_1, \dots, f_M\}$, the (AEW) procedure satisfies for any π satisfying MA(α), $\forall n \geq 1, a > 0$

$$\mathbb{E}\left[R(\tilde{F}_n) - R^*\right] \leq 2(1+a)\min_{f\in\mathcal{F}}(R(f) - R^*) + C_1(a)\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

•$\exists \mathcal{F} = \{f_1, \dots, f_M\}$ such that for any classifier $\bar{f}_n$, $\exists \pi$ satisfying MA(α), $\forall n \geq 1, a > 0$

$$\mathbb{E}\left[R(\bar{f}_n) - R^*\right] \geq 2(1+a)\min_{f\in\mathcal{F}}(R(f) - R^*) + C_2(a)\left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

$\Rightarrow \quad \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$: Almost an optimal rate of aggregation.

Classification: adaptivity problem

We propose a procedure which is:

Classification: adaptivity problem

We propose a procedure which is:

- **Easily computable.**

Classification: adaptivity problem

We propose a procedure which is:

- **Easily computable.**
- **Provides classifiers simultaneously adaptive to the margin and to complexity.**

Classification: adaptivity problem

We propose a procedure which is:

- **Easily computable.**
- **Provides classifiers simultaneously adaptive to the margin and to complexity.**
- **Achieves optimal rates of aggregation under the margin assumption.**

Aggregation of Plug-in Classifiers

Adaptivity

Define the class of models $\mathcal{P}'_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

Adaptivity

Define the class of models $\mathcal{P}'_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**

Adaptivity

Define the class of models $\mathcal{P}'_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**
- **The a conditional probability function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$.**

Adaptivity

Define the class of models $\mathcal{P}'_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure π satisfies the margin assumption MA(α).**
- **The a conditional probability function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$.**
- **The marginal distribution of X is compactly supported and has a Lebesgue density lower bounded and upper bounded by two constants.**

Adaptivity

Theorem 5 (Audibert and Tsybakov (2005)): *Let* $\alpha \geq 0, \beta > 0$. *The excess risk of the plug-in classifier* $\hat{f}_n^{(\beta)} = 2\mathbb{1}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$ *satisfies*

$$\sup_{\pi \in \mathcal{P}'_{\beta,\alpha}} \mathbb{E}\left[R(\hat{f}_n^{(\beta)}) - R^*\right] \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}},$$

where $\hat{\eta}_n^{(\beta)}(\cdot)$ *is the locally polynomial estimator of* $\eta(\cdot)$ *of order* $\lfloor \beta \rfloor$ *with bandwidth* $h = n^{-\frac{1}{2\beta+d}}$.

Adaptivity

Theorem 5 (Audibert and Tsybakov (2005)): *Let $\alpha \geq 0, \beta > 0$. The excess risk of the plug-in classifier $\hat{f}_n^{(\beta)} = 2\mathbb{1}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$ satisfies*

$$\sup_{\pi \in \mathcal{P}'_{\beta,\alpha}} \mathbb{E}\left[R(\hat{f}_n^{(\beta)}) - R^*\right] \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}},$$

where $\hat{\eta}_n^{(\beta)}(\cdot)$ is the locally polynomial estimator of $\eta(\cdot)$ of order $\lfloor \beta \rfloor$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$.

Remark: Audibert and Tsybakov (2005) show that the rate $n^{-\frac{\beta(\alpha+1)}{2\beta+d}}$ is optimal over $\mathcal{P}'_{\beta,\alpha}$, if $\alpha\beta \leq d$. Fast rate: Can achieve $1/n$.

Adaptivity

Theorem 5 (Audibert and Tsybakov (2005)): *Let* $\alpha \geq 0, \beta > 0$*. The excess risk of the plug-in classifier* $\hat{f}_n^{(\beta)} = 2\mathbb{I}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$ *satisfies*

$$\sup_{\pi \in \mathcal{P}'_{\beta,\alpha}} \mathbb{E}\left[R(\hat{f}_n^{(\beta)}) - R^*\right] \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}},$$

where $\hat{\eta}_n^{(\beta)}(\cdot)$ *is the locally polynomial estimator of* $\eta(\cdot)$ *of order* $\lfloor \beta \rfloor$ *with bandwidth* $h = n^{-\frac{1}{2\beta+d}}$*.*

Remark: Audibert and Tsybakov (2005) show that the rate $n^{-\frac{\beta(\alpha+1)}{2\beta+d}}$ is optimal over $\mathcal{P}'_{\beta,\alpha}$, if $\alpha\beta \leq d$. Fast rate: Can achieve $1/n$.

Idea:We aggregate the classifiers $\hat{f}_n^{(\beta)}$ for different values of β lying in a finite grid.

Adaptivity

We use a split of the sample to construct our adaptive classifier:

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

$$\downarrow$$

Construction of the class of plug-in classifiers

$$\mathcal{F} = \left\{ \hat{f}_m^{(\beta_k)} : \beta_k = \frac{kd}{\Delta - 2k}, k \in \{1, \dots, \lfloor \Delta/2 \rfloor\} \right\},$$

where $\Delta = \log n$.

Adaptivity

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

$$\downarrow$$

Construction of the weights:

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

$$f \in \mathcal{F} = \left\{ \hat{f}_m^{(\beta_k)} : \beta_k = \frac{kd}{\Delta - 2k}, k \in \{1, \ldots, \lfloor \Delta/2 \rfloor\} \right\},$$

where $\Delta = \log n$.

Adaptivity

The classifier that we propose is $\tilde{F}_n^{adp} = \mathrm{sign}(\tilde{f}_n^{adp})$, where:

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

and

$$\mathcal{F} = \left\{ \hat{f}_m^{(\beta_k)} : \beta_k = \frac{kd}{\Delta - 2k}, k \in \{1, \dots, \lfloor \Delta/2 \rfloor\} \right\}, \Delta = \log n,$$

$\hat{f}_n^{(\beta)} = 2\mathbb{1}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$ and $\hat{\eta}_n^{(\beta)}$ is the locally polynomial estimator of $\eta(\cdot)$ of order $\lfloor \beta \rfloor$ with bandwidth $h = n^{-\frac{1}{2\beta + d}}$.

Adaptivity

Theorem 6. *Let K be a compact subset of $[0,+\infty) \times (0,+\infty)$. There exists a constant $C_3 > 0$ depending only on K and d such that for any integer $n \geq 1$, any $(\alpha,\beta) \in K$, such that $d > \alpha\beta$, and any $\pi \in \mathcal{P}'_{\beta,\alpha}$, we have,*

$$\mathbb{E}_\pi \left[R(\tilde{F}_n^{adp}) - R^* \right] \leq C_3 n^{-\frac{\beta(\alpha+1)}{2\beta+d}}.$$

Adaptivity

Theorem 6. *Let* K *be a compact subset of* $[0, +\infty) \times (0, +\infty)$*. There exists a constant* $C_3 > 0$ *depending only on* K *and* d *such that for any integer* $n \geq 1$*, any* $(\alpha, \beta) \in K$*, such that* $d > \alpha\beta$*, and any* $\pi \in \mathcal{P}'_{\beta,\alpha}$*, we have,*

$$\mathbb{E}_\pi \left[R(\tilde{F}_n^{adp}) - R^* \right] \leq C_3 n^{-\frac{\beta(\alpha+1)}{2\beta+d}}.$$

Recall: $n^{-\frac{\beta(\alpha+1)}{2\beta+d}}$ is an optimal rate of convergence for the model $\mathcal{P}'_{\beta,\alpha}$.

Adaptive SVM

Kernels and RKHS

kernel: A symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that for all integer $n \geq 1$ and all $x_1, \ldots, x_n \in \mathcal{X}$, the matrix

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

$\Leftrightarrow$ there exists a Hilbert space H (feature space) and a feature map $\phi : \mathcal{X} \mapsto H$ with

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

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

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

Kernels and RKHS

RKHS: For a kernel k, the reproducing kernel Hilbert space (RKHS) 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 dot product:

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

The RKHS is a feature space of k with feature map $\phi : \mathcal{X} \mapsto H, \phi(x) = k(x, .)$.

Kernels and RKHS

The RKHS of the gaussian kernel, denoted by H_σ, is

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

If a gaussian kernel is considered on a compact subset $\mathcal{X} \subset \mathbb{R}^d$, then its RKHS is dense in $\mathcal{C}(\mathcal{X}, \mathbb{R})$.

SVM

k: a kernel over $\mathcal{X}$ (an abstract space). H_k: the RKHS associated to k. $D_n = ((X_i, Y_i)_{1 \le i \le n})$: n observations, with $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$. Let $\lambda > 0$. The Support Vector Machine (SVM) estimator is

$$\hat{f}_n^{\lambda} = \operatorname{Arg} \min_{f \in H_k} \left(A_n(f) + \lambda ||f||_{H_k}^2 \right),$$

empirical Hinge-risk of f: $A_n(f) = \frac{1}{n} \sum_{i=1}^n (1 - Y_i f(X_i))_+$

and λ is a free parameter, called regularity parameter.

SVM classifier: $\hat{F}_n^{\lambda}(x) = \operatorname{sign}(\hat{f}_n^{\lambda})$.

SVM

Using the standard development related to SVM (cf. Schölkopf and Smola (2002)), we may write

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

where $\hat{C}_1, \ldots, \hat{C}_n$ are solutions of the following maximization problem

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

that can be obtained using a standard quadratic programming software.

Rates for SVM

Scovel and Steinwart (2004) introduced the following assumption:

(GNA) Geometric noise assumption. *There exists* $C_1 > 0$ *and* $\gamma > 0$ *such that*

$$\mathbb{E}\left[|2\eta(X) - 1| \exp\left(-\frac{\tau(X)^2}{t}\right)\right] \le C_1 t^{\frac{\gamma d_0}{2}}, \quad \forall t > 0.$$

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

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

Rates for SVM

Theorem 7(Steinwart and Scovel (2005)): *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$. Assume that π satisfies MA(α) and GNA(γ). The SVM classifier for the gaussian kernel with regularization parameter and width:*

$$\lambda_n^{\alpha,\gamma} = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} & \textit{if } \gamma \leq \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2(\gamma+1)(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}} & \textit{otherwise,} \end{cases} \quad \textit{and } \sigma_n^{\alpha,\gamma} = (\lambda_n^{\alpha,\gamma})^{-\frac{1}{(\gamma+1)d_0}},$$

satisfies

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

for all $\epsilon > 0$ and $C = C(\alpha, \gamma, \epsilon)$.

Adaptivity

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

Adaptivity

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

$$\Downarrow$$

Problem: simultaneous adaptation to the margin α and to geometry exponent γ.

Adaptivity

These classifiers depend on the margin parameter α and the geometric noise parameter γ.

$$\Downarrow$$

Problem: simultaneous adaptation to the margin α and to geometry exponent γ.

We use our aggregation procedure to construct adaptive classifiers both to the margin and to geometry.

Adaptivity

We use a split of the sample to construct our adaptive classifier:

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

$$\downarrow$$

Construction of the class of SVM classifiers

$$\mathcal{F} = \left\{ \hat{F}_m^{(\sigma_k, \lambda_l)} : \sigma_k = m^{k/2\Delta d_0}, \lambda_l = m^{-(1/2 + l/\Delta)}, \right.$$
$$\left. k \in \{1, \ldots, 2\lfloor \Delta \rfloor\}, l \in \{1, \ldots, \lfloor \Delta/2 \rfloor\} \right\}, \quad \Delta = \log n.$$

Adaptivity

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

$$\downarrow$$

Construction of the weights:

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

Adaptivity

The classifier that we propose is $\tilde{F}_n^{adp} = \mathrm{sign}(\tilde{f}_n^{adp})$, where:

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

and

$$\mathcal{F} = \left\{ \hat{F}_m^{(\sigma_k, \lambda_l)} : \sigma_k = m^{k/2\Delta d_0}, \lambda_l = m^{-(1/2 + l/\Delta)}, \right.$$
$$\left. k \in \{1, \dots, 2\lfloor \Delta \rfloor\}, l \in \{1, \dots, \lfloor \Delta/2 \rfloor\} \right\}, \quad \Delta = \log n.$$

$\hat{F}_m^{(\sigma,\lambda)} = \mathrm{sign}(\hat{f}_m^{(\sigma,\lambda)})$ where

$$\hat{f}_m^{(\sigma,\lambda)} = \mathrm{Arg}\min_{f \in H_\sigma} \left(A_m(f) + \lambda ||f||_{H_\sigma}^2 \right).$$

Adaptivity

Theorem 8. *Let K be a compact subset of $\mathcal{U} = \{(\alpha,\gamma) \in (0,+\infty)^2 : \gamma > \frac{\alpha+2}{2\alpha}\}$ and K' a compact subset of $\mathcal{U}' = \{(\alpha,\gamma) \in (0,+\infty)^2 : \gamma \leq \frac{\alpha+2}{2\alpha}\}$. Then the aggregate $\tilde{F}_n^{adp}$ satisfies*

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

for all $(\alpha,\gamma) \in K \cup K'$ and $\epsilon > 0$, where $C > 0$ depends only on ϵ, K, K', a and b_0, and $\mathcal{P}_{\alpha,\gamma}$ is the set of all probability measure on $\mathcal{X} \times \{-1,1\}$ satisfying MA(α) and GNA(γ).

Conclusion

The Aggregation procedure with Exponential Weights:

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.
- achieves optimal rates of aggregation under the margin assumption.

Conclusion

The Aggregation procedure with Exponential Weights:

- is easily implementable.
- achieves optimal rates of aggregation under the margin assumption.
- can be used to achieve simultaneous adaptation to the margin and to complexity with fast rates.