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 \(\pi\) on \(\mathcal{X} \times \{-1, 1\}\).
- \(D_n = (X_i, Y_i)_{i=1,...,n}\): a set of \(n\) i.i.d. observations of \((X, Y)\).
- Prediction rule \(f: \mathcal{X} \mapsto \{-1, 1\}\)
- Risk of \(f\): \(R(f) = \mathbb{P}(f(X) \neq Y)\).

**Bayes rule:** \(f^*\) minimizes the risk \(R(f)\) over all prediction rules,

\[
f^*(x) = \text{sign}(2\eta(x) - 1), \quad \eta(x) = \mathbb{P}(Y = 1|X = x),
\]

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

\[\text{An aggregation procedure in classification – p. 2/51}\]
Classification: Model

- **Classifier**: a procedure, that assigns to observations $D_n$ a prediction rule $\hat{f}_n(\cdot, D_n) : \mathcal{X} \mapsto \{-1, 1\}$. 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

$$\sup_{\pi \in \mathcal{P}} \mathbb{E}[R(\hat{f}_n) - R^*] \leq C\phi(n), \quad \forall n \geq 1.$$
No classifier can guarantee a rate of convergence that holds for all probability distributions $\pi$ (Devroye, Györfi and Lugosi, 1996).
Classification: results

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

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

  \[ n^{-1/2}, \text{ (up to a logarithm)} \]
Classification: results

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

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

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

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

  \[ n^{-\alpha}, 0 < \alpha \leq 1/2, \text{ (up to a logarithm)} \]
No classifier can guarantee a rate of convergence that holds for all probability distributions $\pi$ (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 $\eta$ (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 \( \{ x : \eta(x) = 1/2 \} \).
Excess risk is sensitive to the behavior of $\eta(x)$ near the decision boundary $\{x : \eta(x) = 1/2\}$.

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

$$\forall t > 0, \quad \mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^\alpha.$$
Excess risk is sensitive to the behavior of $\eta(x)$ near the decision boundary $\{x : \eta(x) = 1/2\}$.

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

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

Under MA($\alpha$), we can expect fast rates:

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

**Classification: margin assumption**
Further results on fast rates

- Bartlett, Jordan and McAuliffe (2003)
- Blanchard, Bousquet and Massart (2004)
- Nédélec and Massart (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 ($\mathcal{V}C$—dimension, entropy)
  \[ \implies \rho : \text{complexity parameter}. \]
Classification: adaptivity

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

- A complexity assumption \((VC - \text{dimension, entropy})\) \(\implies \rho\) : complexity parameter.

- A margin assumption (behavior of \(\eta\) near the decision boundary) \(\implies \alpha\) : margin parameter.
If one wants to achieve fast rates, two types of assumptions are needed:

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

$\alpha$ and $\rho$ are not known in practice

\[\downarrow\]

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

- A complexity assumption ($\mathcal{VC}$—dimension, entropy) $\implies \rho$ : complexity parameter.

- A margin assumption (behavior of $\eta$ near the decision boundary) $\implies \alpha$ : margin parameter.

$\alpha$ and $\rho$ 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.
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:

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

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

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

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

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

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

The classifier that we propose is: $\tilde{F}_n = \text{sign}(\tilde{f}_n)$
Optimal rate of aggregation

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under $\text{MA}(\alpha)$.
In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under $\text{MA}(\alpha)$.

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

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

In the spirit of Tsybakov (2003), we define optimal rates of model selection aggregation for the classification under $\text{MA}(\alpha)$.

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

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

  \item $\exists \mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any classifier $\bar{f}_n$,

    $\exists \pi \in \text{MA}(\alpha)$, $\forall n \geq 1$

    \[ \mathbb{E} [R(\bar{f}_n) - R^*] \geq \min_{f \in \mathcal{F}} (R(f) - R^*) + C_2 \gamma(n, M, \alpha, \mathcal{F}, \pi). \]
\end{itemize}
Optimal rate of aggregation

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

- $\forall \mathcal{F} = \{f_1, \ldots, f_M\}$, $\exists f_n^*$ such that $\forall \pi \in MA(\alpha)$, $\forall n \geq 1$
  \[ \mathbb{E} \left[ R(f_n^*) - R^* \right] \leq \min_{f \in \mathcal{F}} (R(f) - R^*) + C_1 \gamma(n, M, \alpha, \mathcal{F}, \pi). \]

- $\exists \mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any classifier $\bar{f}_n$,
  $\exists \pi \in MA(\alpha)$, $\forall n \geq 1$
  \[ \mathbb{E} \left[ R(\bar{f}_n) - R^* \right] \geq \min_{f \in \mathcal{F}} (R(f) - R^*) + C_2 \gamma(n, M, \alpha, \mathcal{F}, \pi). \]

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

For any function \( f : \mathcal{X} \rightarrow \mathbb{R} \), the hinge risk of \( f \) is

\[
A(f) \overset{\text{def}}{=} \mathbb{E}[(1 - Y f(X))_+].
\]
Hinge risk

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

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

(Bayes rule) $f^* = \text{Arg min}_{f : \mathcal{X} \rightarrow \mathbb{R}} A(f)$, \quad $A^* \overset{\text{def}}{=} A(f^*)$, 
Hinge risk

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

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

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

Zhang (2004): $R(f) - R^* \leq A(f) - A^*$ for any $f$ with values in $\mathbb{R}$. 

An aggregation procedure in classification – p. 10/51
For any function $f : \mathcal{X} \mapsto \mathbb{R}$, the hinge risk of $f$ is

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

(Bayes rule) $f^* = \text{Arg min}_{f : \mathcal{X} \mapsto \mathbb{R}} A(f)$, $A^* \overset{\text{def}}{=} A(f^*)$,

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

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

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

where \( C_1 > 0 \) is a constant depending only on the constants \( \alpha \) and \( c_0 \) appearing in the margin assumption.
Theorem 1 (Oracle inequality). We assume that $\pi$ satisfies $\text{MA}(\alpha)$. Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a set of prediction rules. The AEW procedure satisfies for any integer $n \geq 1$:

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

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

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

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

$$
\mathbb{E} \left[ A(\overline{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^*}{n}} \frac{\alpha}{1+\alpha} \log M \right) + \left( \frac{\log M}{n} \right)^{\frac{1+\alpha}{2+\alpha}},
$$

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

Optimal rate of aggregation for hinge
Optimal rate of aggregation for hinge

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

where \( \mathcal{M}(\mathcal{F}, \pi) = \min_{f \in \mathcal{F}} A(f) - A^* \).
Optimal rate of aggregation for hinge

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

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

$$M(\mathcal{F}, \pi) \preccurlyeq \left( \frac{\log M}{n} \right)^{\frac{1+\alpha}{2+\alpha}} \implies \text{rate} \prec \left( \frac{\log M}{n} \right)^{\frac{1+\alpha}{2+\alpha}}.$$
Optimal rate of aggregation for hinge

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

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

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

- \( \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \leq M(\mathcal{F}, \pi) \Rightarrow \left(\frac{\log M}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \leq \text{rate} \leq \sqrt{\frac{\log M}{n}} \).
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 \( \mathcal{M}(\mathcal{F}, \pi) = \min_{f \in \mathcal{F}} A(f) - A^* \).

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

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

- No margin assumption (\( \alpha = 0 \)) or \( \mathcal{M}(\mathcal{F}, \pi) \geq a > 0 \implies \text{rate} \preceq \sqrt{\frac{\log M}{n}}. \)
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 $\Sigma(\beta, L, \mathbb{R}^d)$ ($\beta, L > 0$).

$g : \mathbb{R}^d \mapsto \mathbb{R}$, $[\beta]$-times continuously differentiable.
Hölder class

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

- $g : \mathbb{R}^d \mapsto \mathbb{R}$, $\lceil \beta \rceil$-times continuously differentiable.

- $\forall x, y \in \mathbb{R}^d, |g(y) - g_x(y)| \leq L||x - y||_2^\beta$,

where

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

is the Taylor polynomial of degree $\lceil \beta \rceil$ for $g$ at point $x$. 
Hölder class

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

- $g : \mathbb{R}^d \mapsto \mathbb{R}$, $\lfloor \beta \rfloor$-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:

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

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

- The underlying probability measure $\pi$ satisfies the margin assumption $\text{MA}(\alpha)$.
Define the class of models $\mathcal{P}_{\beta,\alpha}$, $\alpha \geq 0, \beta > 0$, by:

- **The underlying probability measure $\pi$ satisfies the margin assumption MA($\alpha$).**

- **The conditional probability function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$.**
Define the class of models $\mathcal{P}_{\beta, \alpha}$, $\alpha \geq 0$, $\beta > 0$, by:

- The underlying probability measure $\pi$ satisfies the margin assumption $\text{MA}(\alpha)$.

- 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

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

\[ \log \text{Card}(\Sigma_\epsilon(\beta)) \leq A \epsilon^{-d/\beta}. \]
Aggregation over a sieve

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

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

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

An aggregation procedure in classification – p. 17/51
Aggregation over a sieve

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

\[ \log \text{Card}(\Sigma_\epsilon(\beta)) \leq A\epsilon^{-d/\beta}. \]

\[ \tilde{f}_{n, \epsilon}(\epsilon, \beta) = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta, \text{ where } f_\eta(x) = \text{Sign}(\eta(x) - 1/2). \]

We chose the step of the \( \epsilon \)-net by a trade-off:

\[ \epsilon_n = n^{-\frac{\beta}{\beta(\alpha+2)+d}}. \]
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 $n > 0$,
\[
\mathbb{E}_\pi \left[ R(\text{Sign}(\tilde{f}_n^{(\epsilon_n, \beta)})) - R^* \right] \leq C_3(\alpha, \beta, d) n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}},
\]
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 $\text{Sign}(\tilde{f}_n(\epsilon_n, \beta))$ needs to know the parameters $\alpha$ and $\beta$ which are not available in practice.

⇓

Problem of adaptivity with respect to $\alpha$ and $\beta$.

Idea: We aggregate classifiers $\tilde{f}_n(\epsilon, \beta)$, for different values of $(\epsilon, \beta)$ 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:

\[ l = \left\lceil \frac{n}{\log n} \right\rceil \text{ and } m = n - l. \]
Adaptivity

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

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

\[ \downarrow \]

Construction of the class of aggregate-classifiers

\[ \mathcal{F} = \left\{ \text{Sign}(f_m^{(\epsilon_m^k, \beta_p)}) : \begin{array}{l}
\epsilon_m^k = m^{-k/\Delta} : k \in \{1, \ldots, \lfloor \Delta/2 \rfloor \}
\beta_p = p/\Delta : p \in \{1, \ldots, \lceil \Delta \rceil^2 \}
\end{array} \right\}, \]

where \( \Delta_n = \log n \).
Adaptivity

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

\[ \downarrow \]

Construction of the weights:

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

\[ F \in \mathcal{F} = \left\{ \text{Sign} (\tilde{f}_{m}^{(\epsilon_m, \beta_p)}) : \epsilon_m = m^{-k/\Delta} : k \in \{1, \ldots, \lfloor \Delta/2 \rfloor \}, \beta_p = p/\Delta : p \in \{1, \ldots, \lceil \Delta \rceil^2 \} \right\}, \]
Adaptivity

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

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

$$\tilde{f}_n^{adp} = \sum_{F \in \mathcal{F}} w[l](F) F,$$
The classifier that we propose is $\tilde{F}_n^{adp} = \text{Sign}(\tilde{f}_n^{adp})$, where:

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

and

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

$\tilde{f}_n^{(\epsilon, \beta)} = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta$ 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 $\pi \in P_{\beta, \alpha}$, 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 $P_{\beta, \alpha}$.

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

Aggregation of $L_1$–SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).
Adaptivity

Aggregation of $L_1$–SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$. 
Adaptivity

Aggregation of $L_1$–SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$.

/problem: simultaneous adaptation to the margin $\alpha$ and to geometry exponent $\gamma$.\"
Adaptivity

Aggregation of $L^1$-SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$.

\[ \downarrow \]

Problem: simultaneous adaptation to the margin $\alpha$ and to geometry exponent $\gamma$.

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

Aggregation of $L_1$–SVM classifiers under margin assumption and geometric noise assumption of Scovel and Steinwart (2004).

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$.

\[\Downarrow\]

Problem: simultaneous adaptation to the margin $\alpha$ and to geometry exponent $\gamma$.

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

We aggregate classifiers for different values of $\alpha$ and $\gamma$ 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 $[0, 1]^d$).
- $\pi$ satisfies $\text{MA}(\infty) \iff |2\eta(X) - 1| \geq h$ a.s.. Assume $h = 1$ $\iff Y = f^*(X) = \eta(X) \iff R^* = 0$.

**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:
Using Zhang’s inequality, we obtain:

- \( \forall \mathcal{F} = \{ f_1, \ldots, f_M \} \), the (AEW) procedure satisfies for any \( \pi \) satisfying \( \text{MA}(\alpha) \), \( \forall n \geq 1, a > 0 \)

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

Using Zhang’s inequality, we obtain:

\( \forall \mathcal{F} = \{f_1, \ldots, f_M\} \), the (AEW) procedure satisfies for any \( \pi \) satisfying \( \text{MA}(\alpha) \), \( \forall n \geq 1, a > 0 \)

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

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

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

Using Zhang’s inequality, we obtain:

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

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

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

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

\[ \Rightarrow \left( \frac{\log M}{n} \right)^{\frac{1+\alpha}{2+\alpha}} : \text{Almost an optimal rate of aggregation.} \]
We propose a procedure which is:
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.
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 $\pi$ satisfies the margin assumption $\text{MA}(\alpha)$.
Adaptivity

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

- The underlying probability measure $\pi$ satisfies the margin assumption $\text{MA}(\alpha)$.

- The a conditional probability function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$. 
Define the class of models $\mathcal{P}_{\beta,\alpha}^{\prime}$, $\alpha \geq 0, \beta > 0$, by:

- The underlying probability measure $\pi$ satisfies the margin assumption $\text{MA}(\alpha)$.
- The 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 $h = n^{-\frac{1}{2\beta+d}}$. 

An aggregation procedure in classification – p. 33/51
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 $1/n$. 
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
\[
\sup_{\pi \in \mathcal{P}_{\beta,\alpha}} \mathbb{E} \left[ R(\hat{f}_n^{(\beta)}) - R^* \right] \leq C n^{-\beta(1+\alpha)/(2\beta+d)},
\]
where $\hat{\eta}_n^{(\beta)}(\cdot)$ is the locally polynomial estimator of $\eta(\cdot)$ of order $\lfloor \beta \rfloor$ with bandwidth $h = n^{-1/(2\beta+d)}$.

Remark: Audibert and Tsybakov (2005) show that the rate $n^{-\beta(\alpha+1)/(2\beta+d)}$ 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 $\beta$ 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 \( m = n - l \).
- \( D_m^1 = ((X_1, Y_1), \ldots, (X_m, Y_m)) \) (training sample)

\[ \downarrow \]

Construction of the class of plug-in classifiers

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

where \( \Delta = \log n \).
Adaptivity

\[ D_i^2 = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)) \text{ (validation sample)}. \]

\[ \downarrow \]

Construction of the weights:

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

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

where \( \Delta = \log n. \)
Adaptivity

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

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

and

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

$$\hat{f}_n^{(\beta)} = 2 \mathbb{I}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$$

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,

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

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

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

$$(k(x_i, x_j))_{1 \leq i, j \leq n}$$

is positive semi-definite.

$\iff$ 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 $\sigma > 0$ ($\sigma$ 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

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

equipped with the dot product:

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

The RKHS is a feature space of \( k \) with feature map

\[ \phi : \mathcal{X} \mapsto H, \phi(x) = k(x,.) \]
The RKHS of the gaussian kernel, denoted by $H_\sigma$, is

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

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

An aggregation procedure in classification – p. 41/5
**SVM**

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

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

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

and $\lambda$ is a free parameter, called regularity parameter.

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

An aggregation procedure in classification – p. 42/51
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.
Scovel and Steinwart (2004) introduced the following assumption:

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

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

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

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

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

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

and $\sigma_{\alpha,\gamma}^n = (\lambda_{\alpha,\gamma}^n)^{-\frac{1}{(\gamma+1)d_0}}$, satisfies

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

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

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$. 
Adaptivity

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$.

⇓

Problem: simultaneous adaptation to the margin $\alpha$ and to geometry exponent $\gamma$. 
Adaptivity

These classifiers depend on the margin parameter $\alpha$ and the geometric noise parameter $\gamma$.

$\Downarrow$

Problem: simultaneous adaptation to the margin $\alpha$ and to geometry exponent $\gamma$.

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:

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

\[ D_m^1 = ((X_1, Y_1), \ldots, (X_m, Y_m)) \text{ (training sample)} \]

\[ \downarrow \]

Construction of the class of SVM classifiers

\[ \mathcal{F} = \left\{ \hat{F}_m(\sigma_k, \lambda_l) : \sigma_k = m^{k/2\Delta d_0}, \lambda_l = m^{-(1/2+l/\Delta)} \right\} \]

\[ k \in \{1, \ldots, 2\lceil \Delta \rceil \}, l \in \{1, \ldots, \lceil \Delta/2 \rceil \} \text{, } \Delta = \log n. \]
Adaptivity

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

↓

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 $\tilde{F}_n^{\text{adp}} = \text{sign}(\tilde{f}_n^{\text{adp}})$, where:

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

and

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

$$k \in \{1, \ldots, 2\lfloor \Delta \rfloor \}, l \in \{1, \ldots, \lfloor \Delta/2 \rfloor \}, \quad \Delta = \log n. \right.$$  

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

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

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

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

The Aggregation procedure with Exponential Weights:
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.