Fast learning rates for plug-in classifiers under the margin condition

Jean-Yves Audibert[1] Alexandre B. Tsybakov[2]

[1]Certis
ParisTech - Ecole des Ponts, France

[2]LPMA
Université Pierre et Marie Curie, France

Empirical Processes and Asymptotic Statistics, 2007

Outline

1. Motivation
   - A standard statistical learning task
   - Complexity and margin assumptions
2. Contributions
   - A powerful way to exploit the margin assumption for plug-in rules
   - Locally polynomial estimators
   - Plugged locally polynomial estimators

Classification : a standard learning setting (1/2)

- **Training data Z_1^n:** $Z_i = (X_i, Y_i) \quad i = 1, \ldots, n \quad$ i.i.d. $\sim P$
  - $X_i \in \mathbb{R}^d$
  - $Y_i \in \{0; 1\}$
  - $Z_i \in \mathcal{Z} = \mathbb{R}^d \times \{0; 1\}$
- **Classification function:** $\quad f : \mathbb{R}^d \to \{0; 1\}$
- **Risk:** $\quad R(f) = P\big[Y \neq f(X)\big]$

Classification : a standard learning setting (2/2)

- **Regression function:** $\eta : \mathbb{R}^d \to [0;1]$
- **Bayes regression function:** $\eta^* \in \text{argmin}_\eta \mathbb{E}[Y - \eta(X)]^2$
  $\eta^* : x \mapsto \mathbb{E}(Y|X = x) = P(Y = 1|X = x)$
- **Bayes classifier:** $f^* : x \mapsto \mathbf{1}_{\eta^*(x) \geq 1/2} \in \text{argmin}_f R(f)$
- **Model:** $\mathcal{P}$ = the set of proba on $\mathcal{Z}$ in which we assume that P is

The classification task:

Predict as well as f^*. More formally: find a mapping $Z_1^n \mapsto \hat{f}$ such that for any $P \in \mathcal{P}$, we have

$$\mathbb{E}_{Z_1^n} R(\hat{f}) \leq R(f^*) + \textit{small term}$$

A way to solve the classification task: the plug-in rules

Find a mapping $Z_1^n \mapsto \hat{\eta}$ such that $\hat{\eta}$ is close to η^* and use the mapping $Z_1^n \mapsto \hat{f} = \mathbf{1}_{\hat{\eta} \geq 1/2}$. (Recall that $f^* = \mathbf{1}_{\eta^* \geq 1/2}$)

Classification : a standard learning setting (2/2)

- **Regression function:** $\eta : \mathbb{R}^d \to [0;1]$
- **Bayes regression function:** $\eta^* \in \text{argmin}_\eta \mathbb{E}[Y - \eta(X)]^2$
  $\eta^* : x \mapsto \mathbb{E}(Y|X = x) = P(Y = 1|X = x)$
- **Bayes classifier:** $f^* : x \mapsto \mathbf{1}_{\eta^*(x) \geq 1/2} \in \text{argmin}_f R(f)$
- **Model:** $\mathcal{P}$ = the set of proba on $\mathcal{Z}$ in which we assume that P is

The classification task:

Predict as well as f^*. More formally: find a mapping $Z_1^n \mapsto \hat{f}$ such that for any $P \in \mathcal{P}$, we have

$$\mathbb{E}_{Z_1^n} R(\hat{f}) \leq R(f^*) + \ Cn^{-\gamma} \quad \text{for some } \gamma > 0$$

A way to solve the classification task: the plug-in rules

Find a mapping $Z_1^n \mapsto \hat{\eta}$ such that $\hat{\eta}$ is close to η^* and use the mapping $Z_1^n \mapsto \hat{f} = \mathbf{1}_{\hat{\eta} \geq 1/2}$. (Recall that $f^* = \mathbf{1}_{\eta^* \geq 1/2}$)

Specifying the model ... (1/2)

- ... by the set Σ of regression functions in which we assume that the Bayes regression function is.
  Assumption (CAR): polynomial entropy of Σ for some L_p-distance, i.e.
  $$\mathcal{H}(\varepsilon, \Sigma, L_p) \leq C\varepsilon^{-\rho}, \quad \forall \varepsilon > 0.$$
- ... by the set $\mathcal{F}$ of classification functions in which we assume that the Bayes classification function is.
  Assumption (CAC): polynomial entropy of $\mathcal{F}$ for the pseudo-distance $d_\Delta : (f_1, f_2) \mapsto P\big[f_1(X) \neq f_2(X)\big]$, i.e.
  $$\mathcal{H}(\varepsilon, \mathcal{F}, d_\Delta) \leq C\varepsilon^{-\rho}, \quad \forall \varepsilon > 0.$$

Specifying the model ... (2/2)

- ... by a low noise assumption :

$$\begin{aligned} \text{No noise} \quad &\Leftrightarrow \quad Y = f^*(X) \quad \text{a.s.} \\ &\Leftrightarrow \quad \forall x,\, \eta^*(x) = P(Y = 1 | X = x) \in \{0, 1\} \end{aligned}$$

Margin assumption (MA) : for some $\alpha > 0$*,*

$$P\big(0 < |\eta^*(X) - 1/2| \leq t\big) \leq Ct^{\alpha}, \qquad \forall\, t > 0.$$

Mammen & Tsybakov (1999), Polonik (1995)

Known results about assumption (CAR): $\mathcal{H}(\varepsilon, \Sigma, L_p) \leq C\varepsilon^{-\rho}, \quad \forall \varepsilon > 0$

- well adapted for the study of plug-in rules, since typically

$$\begin{aligned} \text{CAR} \quad &\Rightarrow \quad \text{smoothness of } \eta^* \\ &\Rightarrow \quad \text{good nonparametric estimator } \hat{\eta} \\ &\Rightarrow \quad \text{good plug-in rule } f^{\text{PLUG-IN}} = \mathbf{1}_{\hat{\eta} \geq 1/2} \text{ since} \\ &\qquad\quad \mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq 2\mathbb{E}|\hat{\eta}(X) - \eta^*(X)|. \end{aligned}$$

- Under additional reasonable assumptions:

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq Cn^{-1/(2+\rho)}$$ **Yang (1999)**

- For smooth functions of parameter β, $\rho = d/\beta$.

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq Cn^{-\beta/(2\beta+d)}$$

Known results about assumption (CAC): (1/2)

$\mathcal{H}(\varepsilon, \mathcal{F}, d_\Delta) \leq C\varepsilon^{-\rho}, \quad \forall \varepsilon > 0$

- (CAC) $\Rightarrow$ smoothness of the decision boundary, i.e. the boundary of the set $\{x \in \mathbb{R}^d : f^*(x) = 1\}$
- well adapted for the study of empirical risk minimizer which is based on the concentration of the empirical process

$$f \mapsto r(f) - r(f^*) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \neq f(X_i)} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \neq f^*(X_i)}.$$

Precisely, letting $f^{\mathrm{ERM}} \in \mathrm{argmin}_{f\in\mathcal{F}} r(f)$, when $f^* \in \mathcal{F}$, we have

$$\begin{aligned} R(f^{\mathrm{ERM}}) - R(f^*) &\leq R(f^{\mathrm{ERM}}) - R(f^*) + r(f^*) - r(f^{\mathrm{ERM}}) \\ &\leq \sup_{f\in\mathcal{F}} \left\{R(f) - R(f^*) + r(f^*) - r(f)\right\} \end{aligned}$$

Known results about assumption (CAC): (2/2)

$\mathcal{H}(\varepsilon, \mathcal{F}, d_\Delta) \leq C\varepsilon^{-\rho}, \quad \forall \varepsilon > 0$

- For $0 < \rho < 1$,

$$\mathbb{E}R(\hat{f}) - R(f^*) \leq Cn^{-1/2} \qquad \textbf{Tsybakov (2004)}$$

- For $\rho > 1$,

$$\mathbb{E}R(\hat{f}) - R(f^*) \leq Cn^{-1/(1+\rho)} \qquad \textbf{Tsybakov (2004)}$$

$\longrightarrow$ in both cases, faster than the $n^{-1/(2+\rho)}$ of (CAR), but no true connection with (CAR)

Why assumption (MA) is useful? (1/2)

$$P\big(0 < |\eta^*(X) - 1/2| \leq t\big) \leq Ct^{\alpha}, \qquad \forall\, t > 0.$$

- (MA) $\Rightarrow \forall f, \quad P\big[f(X) \neq f^*(X)\big] \leq C\big[R(f) - R(f^*)\big]^{\frac{\alpha}{1+\alpha}}$
  $\Rightarrow$ the variance of the empirical process

$$f \mapsto r(f^*) - r(f) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{Y_i \neq f^*(X_i)} - \mathbf{1}_{Y_i \neq f(X_i)}\right]$$

can be bounded in terms of excess risk:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[R(f) - R(f^*) + r(f^*) - r(f)\right] &= \mathbb{V}\mathrm{ar}\left[r(f^*) - r(f)\right] \\
&= \tfrac{1}{n}\mathbb{V}\mathrm{ar}\left(\mathbf{1}_{Y \neq f(X)} - \mathbf{1}_{Y \neq f^*(X)}\right) \\
&\leq \frac{\mathbb{E}\left(\mathbf{1}_{Y \neq f(X)} - \mathbf{1}_{Y \neq f^*(X)}\right)^2}{n} \\
&= \frac{P[f(X) \neq f^*(X)]}{n} \\
&\leq \frac{C[R(f) - R(f^*)]^{\frac{\alpha}{1+\alpha}}}{n}
\end{aligned}$$

Why assumption (MA) is useful? (2/2)

$$P\big(0 < |\eta^*(X) - 1/2| \leq t\big) \leq Ct^\alpha, \qquad \forall\, t > 0.$$

- Schematically, for functions f such that $r(f) \leq r(f^*)$ (e.g. $f = f^{\mathrm{ERM}}$ when $f^* \in \mathcal{F}$), with high probability

$$\begin{aligned} R(f) - R(f^*) &\leq R(f) - R(f^*) + r(f^*) - r(f) \\ &\lesssim C\sqrt{\mathbb{V}\mathrm{ar}\,[R(f) - R(f^*) + r(f^*) - r(f)]} \\ &\lesssim C\sqrt{\frac{[R(f) - R(f^*)]^{\frac{\alpha}{1+\alpha}}}{n}}, \end{aligned}$$

  hence $R(f) - R(f^*) \leq Cn^{-\frac{1+\alpha}{2+\alpha}}$.

- We need a complexity assumption if we want nontrivial results on the excess risk: precisely, under assumption (CAC), we have

$$\mathbb{E}R(f^{\mathrm{ERM}}) - R(f^*) \leq Cn^{-\frac{1+\alpha}{2+\alpha+\alpha\rho}}$$

Summary

- Under (CAR): slow rates

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq Cn^{-\gamma_\rho} \quad \text{with} \quad \gamma_\rho < 1/2$$

- Under (CAC): faster slow rates

$$\mathbb{E}R(f^{\text{ERM}}) - R(f^*) \leq Cn^{-\eta_\rho} \quad \text{with} \quad \gamma_\rho < \eta_\rho \leq 1/2$$

- Under (CAC)+(MA): $\mathbb{E}R(f^{\text{ERM}}) - R(f^*) \leq Cn^{-\beta_{\alpha,\rho}}$ with $\eta_\rho < \beta_{\alpha,\rho} \leq 1$, and for high enough α, $\beta_{\alpha,\rho} > 1/2$ (fast rates)

It seems that:

- plug-in rules have only slow rates
- rates cannot be faster than n^{-1}

Summary

- Under (CAR): slow rates

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq Cn^{-\gamma_\rho} \quad \text{with} \quad \gamma_\rho < 1/2$$

- Under (CAC): faster slow rates

$$\mathbb{E}R(f^{\text{ERM}}) - R(f^*) \leq Cn^{-\eta_\rho} \quad \text{with} \quad \gamma_\rho < \eta_\rho \leq 1/2$$

- Under (CAC)+(MA): $\mathbb{E}R(f^{\text{ERM}}) - R(f^*) \leq Cn^{-\beta_{\alpha,\rho}}$ with $\eta_\rho < \beta_{\alpha,\rho} \leq 1$, and for high enough α, $\beta_{\alpha,\rho} > 1/2$ (fast rates)

It seems that:

- plug-in rules have only slow rates FALSE
- rates cannot be faster than n^{-1} FALSE

A key lemma to exploit (CAR) + (MA)

Up to now, we have used for $f^{\text{PLUG-IN}} = \mathbf{1}_{\hat{\eta} \geq 1/2}$

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq 2\mathbb{E}|\hat{\eta}(X) - \eta^*(X)|$$

Lemma 1

Assume that a regression function estimator $\hat{\eta}$ satisfies for some sequence (a_n): for almost all x w.r.t. P_X and any $\delta > 0$,

$$\sup_{P \in \mathcal{P}} P^{\otimes n}\Big(\big|\hat{\eta}(x) - \eta^*(x)\big| \geq \delta\Big) \leq C_1 \exp\big(-C_2 a_n \delta^2\big).$$

If (MA) holds (for all $P \in \mathcal{P}$), we have

$$\sup_{P \in \mathcal{P}} \Big\{\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*)\Big\} \leq C a_n^{-\frac{1+\alpha}{2}}$$

for some constant $C > 0$ depending only on α, C_1 and C_2.

Proof of the key lemma by a peeling device

Consider the sets $A_j \subset \mathbb{R}^d, j = 1, 2, \ldots,$ defined as

$$\begin{aligned} A_0 &\triangleq \{x \in \mathbb{R}^d : 0 < |\eta^*(x) - \tfrac{1}{2}| \leq \delta\}, \\ A_j &\triangleq \{x \in \mathbb{R}^d : 2^{j-1}\delta < |\eta^*(x) - \tfrac{1}{2}| \leq 2^j\delta\}, \qquad \text{for } j \geq 1. \end{aligned}$$

For any $\delta > 0$, we may write

$$\begin{aligned} \mathbb{E}R(\hat{f}) - R(f^*) &= \mathbb{E}\big(|2\eta^*(X) - 1|\mathbf{1}_{\{\hat{f}(X) \neq f^*(X)\}}\big) \\ &= \textstyle\sum_{j=0}^{\infty} \mathbb{E}\big(|2\eta^*(X) - 1|\mathbf{1}_{\{\hat{f}(X) \neq f^*(X)\}}\mathbf{1}_{\{X \in A_j\}}\big) \\ &\leq 2\delta P_X\big(0 < |\eta^*(X) - \tfrac{1}{2}| \leq \delta\big) \\ &\qquad + \textstyle\sum_{j \geq 1} \mathbb{E}\big(|2\eta^*(X) - 1|\mathbf{1}_{\{\hat{f}(X) \neq f^*(X)\}}\mathbf{1}_{\{X \in A_j\}}\big). \end{aligned}$$

On the event $\{\hat{f} \neq f^*\}$ we have $|\eta^* - \frac{1}{2}| \leq |\hat{\eta} - \eta^*|$, hence

$$\begin{aligned} &\mathbb{E}\big(|2\eta^*(X) - 1|\mathbf{1}_{\{\hat{f}(X) \neq f^*(X)\}}\mathbf{1}_{\{X \in A_j\}}\big) \\ &\quad \leq 2^{j+1}\delta\, \mathbb{E}\big[\mathbf{1}_{\{|\hat{\eta}(X) - \eta^*(X)| \geq 2^{j-1}\delta\}}\mathbf{1}_{\{0 < |\eta^*(X) - \frac{1}{2}| \leq 2^j\delta\}}\big] \\ &\quad \leq 2^{j+1}\delta\, E_X\Big[P^{\otimes n}\big(|\hat{\eta}(X) - \eta^*(X)| \geq 2^{j-1}\delta\big)\mathbf{1}_{\{0 < |\eta^*(X) - \frac{1}{2}| \leq 2^j\delta\}}\Big] \\ &\quad \leq C_1 2^{j+1}\delta \exp\Big(-C_2 a_n (2^{j-1}\delta)^2\Big) P_X\Big(0 < |\eta^*(X) - \tfrac{1}{2}| \leq 2^j\delta\Big) \\ &\quad \leq 2C_1 C_0 2^{j(1+\alpha)}\delta^{1+\alpha} \exp\big(-C_2 a_n (2^{j-1}\delta)^2\big) \end{aligned}$$

Now taking $\delta = a_n^{-1/2}$ and using once more (MA), we get the desired result.

A general picture of regression function estimators **Györfi, Kohler, Krzyzak and Walk (2004)**

- Algorithms by local averaging: to estimate $\eta^*(x) = \mathbb{E}(Y|X = x)$, you average the Y_i's of the X_i close to x

$$\hat{\eta}(x) = \sum_{i=1}^n W_i(x) Y_i \qquad \text{with } \sum_{i=1}^n W_i(x) = 1$$

  - kernel estimators: $W_i(x) = \frac{K(\|X_i - x\|)}{\sum_{j=1}^n K(\|X_j - x\|)}$, with, for instance, $K(u) = e^{-\frac{u^2}{2h^2}}$ and $h > 0$.
  - k-nearest neighbors, algorithms by partitioning the space,...

  Key tool: concentration of sums of i.i.d. random variables (typically Hoeffding and Bernstein's inequalities)

- Algorithms by minimization of the empirical risk: neural networks, support vector machines, AdaBoost, ...
  Key tool: the concentration of empirical processes

Definition

Definition of locally polynomial estimator

For $h > 0$, $x \in \mathbb{R}^d$, for an integer $\ell \geq 0$ and a function $K : \mathbb{R}^d \to \mathbb{R}_+$, denote by $\hat{\theta}_x$ a polynomial on $\mathbb{R}^d$ of degree ℓ which minimizes

$$\sum_{i=1}^{n} \left[Y_i - \hat{\theta}_x (X_i - x) \right]^2 K \left(\frac{X_i - x}{h} \right). \tag{1}$$

The **locally polynomial estimator** $\hat{\eta}^{LP}(x)$ **of order** ℓ, of the value $\eta^*(x)$ of the regression function at point x is defined by:

$$\hat{\eta}^{LP}(x) \triangleq \hat{\theta}_x(0)$$

if $\hat{\theta}_x$ is the unique minimizer of (1) and $\hat{\eta}^{LP}(x) \triangleq 0$ otherwise.

Model A

Let $\beta > 0$ and $L > 0$ be given. We consider the class $\mathcal{P}_A$ of probability distributions P such that

- (MA) is satisfied
- $\eta^* : \mathbb{R}^d \to [0; 1]$ is β-times continuously differentiable and all its partial derivatives of order β is bounded by L
- P_X admits a density w.r.t. the Lebesgue measure on $[0; 1]^d$ bounded away from zero and infinity

Key results for model A (1/2)

Theorem 1

A slight modification $\hat{\eta}$ of the LP estimator satisfies for $h = n^{-\frac{1}{2\beta+d}}$, any $\delta > 0$ and almost all x w.r.t. P_X:

$$\sup_{P \in \mathcal{P}_A} P^{\otimes n}\Big(\big|\hat{\eta}(x) - \eta^*(x)\big| \geq \delta\Big) \leq C_1 \exp\big(-C_2 n^{\frac{2\beta}{2\beta+d}} \delta^2\big).$$

- LP estimators do not require smoothness assumption on the density of P_X. **(Stone 1980)**

Key results for model A (2/2)

Corollary of Lemma 1 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(f^{\text{PLUG-IN}}) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

- $\alpha\beta > d/2 \Rightarrow$ fast rates
- $\alpha\beta > d \Rightarrow$ super-fast rates
- minimax optimal for $\alpha\beta \leq d$ since any classifier $\hat{f}$ (plug-in or not) satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(\hat{f}) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Key results for model A (2/2)

Corollary of Lemma 1 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(f^{\text{PLUG-IN}}) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

- $\alpha\beta > d/2 \Rightarrow$ fast rates
- $\alpha\beta > d \Rightarrow$ super-fast rates
- minimax optimal for $\alpha\beta \leq d$ since any classifier $\hat{f}$ (plug-in or not) satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(\hat{f}) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Key results for model A (2/2)

Corollary of Lemma 1 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(f^{\text{PLUG-IN}}) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

- $\alpha\beta > d/2 \Rightarrow$ fast rates
- $\alpha\beta > d \Rightarrow$ super-fast rates
- minimax optimal for $\alpha\beta \leq d$ since any classifier $\hat{f}$ (plug-in or not) satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(\hat{f}) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Key results for model A (2/2)

Corollary of Lemma 1 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(f^{\text{PLUG-IN}}) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

- $\alpha\beta > d/2 \Rightarrow$ fast rates
- $\alpha\beta > d \Rightarrow$ super-fast rates *since* $\mathcal{P}_A$ *limited*
- minimax optimal for $\alpha\beta \leq d$ since any classifier $\hat{f}$ (plug-in or not) satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(\hat{f}) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Key results for model A (2/2)

Corollary of Lemma 1 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(f^{\text{PLUG-IN}}) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

- $\alpha\beta > d/2 \Rightarrow$ fast rates
- $\alpha\beta > d \Rightarrow$ super-fast rates *since* $\mathcal{P}_A$ *limited*
- minimax optimal for $\alpha\beta \leq d$ since any classifier $\hat{f}$ (plug-in or not) satisfies

$$\sup_{P \in \mathcal{P}_A} \left\{ \mathbb{E} R(\hat{f}) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Exponential convergence rates under the strong margin assumption

- (MA) with $\alpha = \infty \Leftrightarrow P_X\big(0 < |\eta^*(X) - 1/2| \leq t_0\big) = 0.$

Lemma 2

The plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ we have

$$\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*) \leq P\big(|\hat{\eta}(X) - \eta^*(X)| > t_0\big).$$

Corollary of Lemma 2 and Theorem 1

The excess risk of the plug-in classifier $f^{\text{PLUG-IN}} = \mathbf{1}_{\{\hat{\eta} \geq \frac{1}{2}\}}$ with appropriate bandwidth h (independent of n) satisfies

$$\sup_{P \in \mathcal{P}_A} \Big\{\mathbb{E}R(f^{\text{PLUG-IN}}) - R(f^*)\Big\} \leq C_1 \exp(-C_2 n).$$

Conclusion

- At the level of the statistical learning community
  - Plug-in rules can achieve fast rates, super-fast rates and even exponential rates
- At the level of the empirical process community
  - The statistical learning community is interested in fast rates
  - These rates are achievable under a low noise assumption
  - Comparison lemmas are very powerful tools

$$f \mapsto r(f) - r(f^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{Y_i \neq f(X_i)} - \mathbf{1}_{Y_i \neq f^*(X_i)}\right]$$

$$\downarrow$$

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \eta(X_i))^p - (Y_i - \eta^*(X_i))^p\right]$$

Conclusion

- At the level of the statistical learning community
  - Plug-in rules can achieve fast rates, super-fast rates and even exponential rates
- At the level of the empirical process community
  - The statistical learning community is interested in fast rates
  - These rates are achievable under a low noise assumption
  - Comparison lemmas are very powerful tools

$$f \mapsto r(f) - r(f^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{Y_i \neq f(X_i)} - \mathbf{1}_{Y_i \neq f^*(X_i)}\right]$$

$$\downarrow$$

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \eta(X_i))^p - (Y_i - \eta^*(X_i))^p\right]$$

Conclusion

- At the level of the statistical learning community
  - Plug-in rules can achieve fast rates, super-fast rates and even exponential rates
- At the level of the empirical process community
  - The statistical learning community is interested in fast rates
  - These rates are achievable under a low noise assumption
  - Comparison lemmas are very powerful tools

$$f \mapsto r(f) - r(f^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{Y_i \neq f(X_i)} - \mathbf{1}_{Y_i \neq f^*(X_i)}\right]$$

$$\downarrow$$

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \eta(X_i))^p - (Y_i - \eta^*(X_i))^p\right]$$