Welcome to the Zoo: Fast Rates in Statistical and Online Learning

Tim van Erven

Universiteit Leiden

Leiden, March 23, 2018
Statistical Learning

\[
\left( Y_1, X_1 \right), \ldots, \left( Y_N, X_N \right) \quad \text{independently distributed} \quad \sim P
\]

\[
\hat{f} = f_{\hat{\theta}} \in \mathcal{F} = \{ f_{\theta} \mid \theta \in \Theta \} \quad \text{(estimator inside model)}
\]

Small risk

\[
R(\hat{\theta}) = \mathbb{E}_{(X, Y) \sim P} [\ell(X, Y, \hat{\theta})]
\]

Compared to minimizer \( \theta^* = \arg \min_{\theta \in \Theta} R(\theta) \) of risk

Minimax Rate:
Rate for most difficult possible \( P \)

\[
\min_{\hat{\theta}} \max_P \mathbb{E}[R(\hat{\theta})] - R(\theta^*)
\]
Classification

Given $X \in \mathbb{R}^d$, predict binary label $Y \in \{0, 1\}$

$$\ell(X, Y, \theta) = \begin{cases} 
0 & \text{if } f_\theta(X) = Y, \\
1 & \text{if } f_\theta(X) \neq Y
\end{cases}$$

$$R(\theta) = P(f_\theta(X) \neq Y)$$

Minimax Rate:
For worst-case $P$, learning is slow:

$$\mathbb{E}[R(\hat{\theta})] - R(\theta^*) \asymp \sqrt{\frac{\text{complexity}_N(\Theta)}{N}}$$
But Faster Rates Are Common

- Worst-case distribution: $P(Y = 1 \mid X)$ very close to $\frac{1}{2}$
- But then learning is (almost) useless!

The Margin Condition: [Tsybakov, 2004]

- Common case: $P(Y = 1 \mid X)$ not too close to $\frac{1}{2}$
- Assume $f_{\theta^*}(X) = f_B(X) = \arg \max_y P(Y = y \mid X)$
- Learning can be much faster depending on $\alpha \in [0, \infty]$: 

$$
\mathbb{E}[R(\hat{\theta})] - R(\theta^*) = O\left(\frac{\text{complexity}_N(\Theta)}{N}\right)^{\frac{1+\alpha}{2+\alpha}}
$$
The Margin Condition

\[ P(Y = 1 \mid X) \]

\[
P_x \left( \left| P(Y \mid X) - \frac{1}{2} \right| \leq t \right) \leq c t^\alpha
\]
Fast Rates in Misspecified Regression

Bounded regression: given $X \in \mathbb{R}^d$, predict $Y$, $f_\theta(X) \in [-B, +B]$

$$\ell(X, Y, \theta) = (Y - f_\theta(X))^2, \quad f_B(X) = \mathbb{E}[Y | X]$$

Conclusion: convex $\mathcal{F}$ always safe to get fast rates [Lee et al., 1998].
Fast Rates for Misspecified Density Estimation

Estimate the best density from $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$

$$\ell(Y, \theta) = -\log p_\theta(Y)$$

Assume all densities uniformly bounded: $1/c \leq p_\theta(Y) \leq c$

Non-convex
ERM gets slow rate depending on $P$

Convex
ERM gets fast rate:
$$\tilde{O}\left(\frac{\text{complexity}_N(\mathcal{P})}{N}\right)$$
Fast rates follow from the following supermartingale-like property:

\[
\mathbb{E}_P \left[ \frac{p_\theta}{p_{\theta^*}} \right] \leq 1 \quad \text{for all } p_\theta \in \mathcal{P}. \tag{1}
\]

NB. If \( p \in \mathcal{P} \), then \( p_{\theta^*} = p \), so \( \mathbb{E}_P \left[ \frac{p_\theta}{p_{\theta^*}} \right] = 1 \).

**Lemma ([Li, 1999])**

Convexity of \( \mathcal{P} \) implies (1).

**Proof.**

- For arbitrary \( p_\theta \), let \( p_\lambda = (1 - \lambda)p_{\theta^*} + \lambda p_\theta \) and \( h(\lambda) = \mathbb{E}[-\log p_\lambda(Y)] \).

- Convexity: \( h \) is minimized at \( \lambda = 0 \), so \( 0 \leq h'(0) = 1 - \mathbb{E} \left[ \frac{p_\theta}{p_{\theta^*}} \right] \).
Online Learning

For $t = 1, \ldots, T$:

1. Predict parameter vector $\hat{\theta}_t \in \Theta \subset \mathbb{R}^d$

2. Observe outcome $(X_t, Y_t)$ and update $\hat{\theta}_t \to \hat{\theta}_{t+1}$

Goal: achieve small regret

$$\text{Regret}_{\theta^*} = \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{\theta}_t) - \sum_{t=1}^{T} \ell(X_t, Y_t, \theta^*)$$

with respect to the ‘best’ parameters $\theta^* \in \Theta$.

Assume losses bounded and convex in $\theta$, and $\Theta$ convex with bounded diameter.
Online Learning

For \( t = 1, \ldots, T \):

1. Predict parameter vector \( \hat{\theta}_t \in \Theta \subset \mathbb{R}^d \)
2. Observe outcome \((X_t, Y_t)\) and update \( \hat{\theta}_t \rightarrow \hat{\theta}_{t+1} \)

Goal: achieve small \textit{regret}

\[
\text{Regret}^{\theta^*}_T = \sum_{t=1}^T \ell(X_t, Y_t, \hat{\theta}_t) - \sum_{t=1}^T \ell(X_t, Y_t, \theta^*)
\]

with respect to the ‘best’ parameters \( \theta^* \in \Theta \).
Assume losses bounded and convex in \( \theta \), and \( \Theta \) convex with bounded diameter.

\textbf{Minimax Rate:}
Rate for most difficult possible data:

\[
\min \max_{\hat{\theta}_1} \min_{X_1,Y_1} \min_{\hat{\theta}_2} \min_{X_2,Y_2} \cdots \min_{\hat{\theta}_T} \min_{X_T,Y_T} \max_{\theta^* \in \Theta} \text{Regret}^{\theta^*}_T = O(\sqrt{T})
\]
Fast Rates for Exp-concave and Mixable Losses

We can get a much faster $O\left(\frac{d}{\eta} \log T\right)$ rate in the following cases:

**Exp-concavity:**

$$\theta \mapsto e^{-\eta \ell(X_t, Y_t, \theta)}$$ should be concave.

E.g. logistic loss: $\log(1 + e^{-Y_t \theta^\top X_t})$

**Mixability:**

Without knowing $X_t, Y_t$, we can map any probability distribution $\pi$ on $\Theta$ to a prediction $\theta_\pi \in \Theta$ such that

$$e^{-\eta \ell(X_t, Y_t, \theta_\pi)} \geq \int e^{-\eta \ell(X_t, Y_t, \theta)} d\pi(\theta)$$

▶ Intuition: allows being unsure

▶ Exp-concavity is a special case:

$$\theta_\pi = \mathbb{E}_{\pi}[\theta]$$
Fast Rates for Exp-concave and Mixable Losses

We can get a much faster $O\left(\frac{d}{\eta} \log T \right)$ rate in the following cases:

**Exp-concavity:**

$$\theta \mapsto e^{-\eta \ell(X_t, Y_t, \theta)}$$

should be concave.

E.g. logistic loss: $\log(1 + e^{-Y_t \theta^\top X_t})$

**Mixability:**

Without knowing $X_t, Y_t$, we can map any probability distribution $\pi$ on $\Theta$ to a prediction $\theta_\pi \in \Theta$ such that

$$e^{-\eta \ell(X_t, Y_t, \theta_\pi)} \geq \int e^{-\eta \ell(X_t, Y_t, \theta)} d\pi(\theta)$$

- Intuition: allows being unsure
- Exp-concavity is a special case: $\theta_\pi = E_\pi[\theta]$. 
Welcome to the Zoo

How can we understand all these different cases?

▶ We made a map…
▶ …but the zoo is huge and the routes are long.

A full map of the zoo
[Van Erven, Grünwald, Mehta, Reid, and Williamson, 2015]
Welcome to the Zoo

How can we understand all these different cases?

- We made a map...
- ...but the zoo is huge and the routes are long.
- The summary: for bounded losses, they are all special cases of (more or less) one central condition.
- Let me give you a tour.

A full map of the zoo [Van Erven, Grünwald, Mehta, Reid, and Williamson, 2015]
The Central Condition

Central Condition
For some $\eta > 0$,
\[
\mathbb{E}_P \left[ e^{-\eta \left( \ell(X, Y, \theta) - \ell(X, Y, \theta^*) \right)} \right] \leq 1 \quad \text{for all } \theta \in \Theta.
\]

- Controls the left tail of $\ell(X, Y, \theta) - \ell(X, Y, \theta^*)$. 

Specialize to Density Estimation

- $\ell(X, Y, \theta) = -\log p_\theta(Y) \leftrightarrow p_\theta(Y) = e^{-\ell(Y, \theta)}$

- For $\eta = 1$, CC specializes to $\mathbb{E}_P \left[ p_\theta(Y) p_{\theta^*}(Y) \right] \leq 1$.

Convex $P$:
\[
\min \pi(\theta) \mathbb{E} \left[ -\log \int p_\theta(Y) d\pi(\theta) \right] = \min \theta \mathbb{E} \left[ -\log p_\theta(Y) \right].
\]

Theorem
For general losses, CC is equivalent to pseudo-probability convexity:
\[
\min \pi(\theta) \mathbb{E} \left[ -\log \int e^{-\eta \ell(X, Y, \theta)} d\pi(\theta) \right] = \min \theta \mathbb{E} \left[ -\log e^{-\eta \ell(X, Y, \theta)} \right].
\]

12 / 26
The Central Condition

Central Condition
For some $\eta > 0$,
\[
\mathbb{E}_{P} \left[ e^{-\eta \left( \ell(X, Y, \theta) - \ell(X, Y, \theta^*) \right)} \right] \leq 1 \quad \text{for all } \theta \in \Theta.
\]

- Controls the left tail of $\ell(X, Y, \theta) - \ell(X, Y, \theta^*)$.

Specialize to Density Estimation

- $\ell(Y, \theta) = -\log p_{\theta}(Y) \leftrightarrow p_{\theta}(Y) = e^{-\ell(Y, \theta)}$

- For $\eta = 1$, CC specializes to $\mathbb{E}_{P} \left[ \frac{p_{\theta}(Y)}{p_{\theta^*}(Y)} \right] \leq 1$.

- Convex $\mathcal{P}$: $\min_{\pi(\theta)} \mathbb{E}[-\log \int p_{\theta}(Y) \, d\pi(\theta)] = \min_{\theta} \mathbb{E}[-\log p_{\theta}(Y)]$. 
The Central Condition

Central Condition
For some $\eta > 0$,
\[
\mathbb{E}_P \left[ e^{-\eta \left( \ell(X, Y, \theta) - \ell(X, Y, \theta^*) \right)} \right] \leq 1 \quad \text{for all } \theta \in \Theta.
\]

- Controls the left tail of $\ell(X, Y, \theta) - \ell(X, Y, \theta^*)$.

Specialize to Density Estimation

- $\ell(Y, \theta) = -\log p_\theta(Y) \leftrightarrow p_\theta(Y) = e^{-\ell(Y, \theta)}$
- For $\eta = 1$, CC specializes to $\mathbb{E}_P \left[ \frac{p_\theta(Y)}{p_{\theta^*}(Y)} \right] \leq 1$.
- Convex $\mathcal{P}$: $\min_{\pi(\theta)} \mathbb{E}[-\log \int p_\theta(Y) \, d\pi(\theta)] = \min_\theta \mathbb{E}[-\log p_\theta(Y)]$.

Theorem

For general losses, CC is equivalent to pseudo-probability convexity:
\[
\min_{\pi(\theta)} \mathbb{E}[-\log \int e^{-\eta \ell(X, Y, \theta)} \, d\pi(\theta)] = \min_\theta \mathbb{E}[-\log e^{-\eta \ell(X, Y, \theta)}]
\]
Understanding Online Learning Conditions

Mixability

Without knowing $X_t, Y_t$, we can map any probability distribution $\pi$ on $\Theta$ to a prediction $\theta_\pi \in \Theta$ such that

$$e^{-\eta \ell(X_t, Y_t, \theta_\pi)} \geq \int e^{-\eta \ell(X_t, Y_t, \theta)} d\pi(\theta)$$

$$\ell(X_t, Y_t, \theta_\pi) \leq -\frac{1}{\eta} \log \int e^{-\eta \ell(X_t, Y_t, \theta)} d\pi(\theta)$$

Stochastic Mixability

Without knowing $P$, we can map any probability distribution $\pi$ on $\Theta$ to a prediction $\theta_\pi \in \Theta$ such that

$$\mathbb{E}_P[\ell(X, Y, \theta_\pi)] \leq \mathbb{E}_P \left[ -\frac{1}{\eta} \log \int e^{-\eta \ell(X, Y, \theta)} d\pi(\theta) \right]$$
Understanding Online Learning Conditions

Mixability

Without knowing \( X_t, Y_t \), we can map any probability distribution \( \pi \) on \( \Theta \) to a prediction \( \theta_\pi \in \Theta \) such that

\[
e^{-\eta \ell(X_t, Y_t, \theta_\pi)} \geq \int e^{-\eta \ell(X_t, Y_t, \theta)} \, d\pi(\theta)
\]

\[
\ell(X_t, Y_t, \theta_\pi) \leq -\frac{1}{\eta} \log \int e^{-\eta \ell(X_t, Y_t, \theta)} \, d\pi(\theta)
\]

Stochastic Mixability

Without knowing \( P \), we can map any probability distribution \( \pi \) on \( \Theta \) to a prediction \( \theta_\pi \in \Theta \) such that

\[
\mathbb{E}_P[\ell(X, Y, \theta_\pi)] \leq \mathbb{E}_P \left[ -\frac{1}{\eta} \log \int e^{-\eta \ell(X, Y, \theta)} \, d\pi(\theta) \right]
\]

Theorem

*Stochastic mixability implies the central condition, and under technical conditions the reverse also holds.*
The Technical Conditions...

\[ S_\pi(P, \theta) = \mathbb{E}_{(X,Y) \sim P, \theta' \sim \pi} \left[ \exp(\eta(\ell(X,Y,\theta) - \ell(X,Y,\theta'))) \right] \]

**Theorem (Detailed)**

Stochastic mixability w.r.t. all \( P \in \mathcal{P} \) implies the central condition for all \( P \in \mathcal{P} \) if, for all \( \pi \),

\[
\sup_{P \in \mathcal{P}} \inf_{\theta \in \Theta} S_\pi(P, \theta) \leq 1 \implies \inf_{\theta \in \Theta} \sup_{P \in \mathcal{P}} S_\pi(P, \theta) \leq 1. \tag{\ast}
\]

**Sufficient Conditions for \( \ast \):**

1. \( \ell(X,Y,\theta) \) continuous in \( (X,Y) \) in Polish space
2. \( \ell(X,Y,\theta) \) or \( \exp(\eta(\ell(X,Y,\theta)}) \) convex in \( \theta \)
3. \( \mathcal{P} \) closed, convex and tight in weak topology
4. \( \xi_\theta(X,Y) = \mathbb{E}_{\theta' \sim \pi} \left[ \exp(\eta(\ell(X,Y,\theta) - \ell(X,Y,\theta'))) \right] \) uniformly integrable over \( \theta \in \Theta, P \in \mathcal{P} \).
Understanding Regression

Bounded regression: given $X \in \mathbb{R}^d$, predict $Y, f_\theta(X) \in [-B, +B]$

$$\ell(X, Y, \theta) = (Y - f_\theta(X))^2$$

Proposition

For convex $\mathcal{F}$ parametrized by $\theta = f_\theta$, the squared loss is exp-concave with $\eta \propto 1/B^2$.

exp-concavity $\rightarrow$ mixability $\rightarrow$ stochastic mixability $\rightarrow$ central condition
Another Way to See the Central Condition

Abbreviate $\Delta_\theta = \ell(X, Y, \theta) - \ell(X, Y, \theta^*)$. Then

$$\mathbb{E}[\Delta_\theta] = R(\theta) - R(\theta^*)$$

Central Condition

$$\mathbb{E}[\text{e}^{-\eta \Delta_\theta}] \leq 1$$

$(B, 1)$-Bernstein Condition

The closer $R(\theta)$ to $R(\theta^*)$, the smaller the variance:

$$\mathbb{E}[\Delta_\theta^2] \leq B \mathbb{E}[\Delta_\theta]$$
Another Way to See the Central Condition

Abbreviate $\Delta_{\theta} = \ell(X, Y, \theta) - \ell(X, Y, \theta^*)$. Then

$$\mathbb{E}[\Delta_{\theta}] = R(\theta) - R(\theta^*)$$

Central Condition

$$\mathbb{E}[e^{-\eta\Delta_{\theta}}] \leq 1$$

$(B, 1)$-Bernstein Condition

The closer $R(\theta)$ to $R(\theta^*)$, the smaller the variance:

$$\mathbb{E}[\Delta_{\theta}^2] \leq B \mathbb{E}[\Delta_{\theta}]$$

Proposition

*For bounded losses, CC and $(B, 1)$-Bernstein are equivalent for $B \propto 1/\eta$.*

Proof.

By $e^{-z} \approx 1 - z + \frac{1}{2} z^2$ applied to $z = \eta \Delta_{\theta}$.  

\[ \]
Understanding Classification

$p(y | x) = \begin{cases} 0 & x < a \\ \frac{1}{2} & a \leq x < b \\ 1 & x \geq b \end{cases}$

Easy: $\alpha = \infty$

Moderate: $\alpha = 1$

Hard: $\alpha = 0$

$P_X(|P(Y | X) - \frac{1}{2}| \leq t) \leq ct^\alpha$ (\(\alpha\)-margin)

Lemma (Tsybakov)

If $f_B \in \mathcal{F}$. Then, for 0/1-loss, \(\alpha\)-margin is equivalent to the $(B, \beta)$-Bernstein condition:

$$\mathbb{E}[\Delta_\theta^2] \leq B \mathbb{E}[\Delta_\theta]^{\beta}$$

with $\beta = \frac{\alpha}{1+\alpha} \in [0, 1]$ and some $B \geq 0$. 

Intermediate Rates

Abbreviate \( \Delta_\theta = \ell(X, Y, \theta) - \ell(X, Y, \theta^*) \)

**Generalized Central Condition**
For all \( \epsilon \geq 0 \)

\[
\mathbb{E}[e^{-\eta_\epsilon \Delta_\theta}] \leq e^{\eta_\epsilon \epsilon}
\]

**(B, \beta)**-Bernstein Condition
For some \( B \geq 0, \beta \in [0, 1] \):

\[
\mathbb{E}[\Delta_\theta^2] \leq B \mathbb{E}[\Delta_\theta]^\beta
\]

**Theorem**

*For bounded losses, generalized CC and **(B, \beta)**-Bernstein are equivalent for \( \eta_\epsilon \propto \epsilon^{1-\beta}/B \).*
Online Learning: Prediction with Expert Advice

Prediction with Expert Advice

- Interpret the components of $X_t \in [0, 1]^d$ as predictions of $d$ experts, who are predicting $Y_t \in \{0, 1\}$.
- Our choice $P_\theta$ is a probability distribution on these $d$ experts
- $\ell(X_t, Y_t, \theta) = |Y_t - \mathbb{E}_{P_\theta(i)}[X_{t,i}]| = \mathbb{E}_{P_\theta(i)}[|Y_t - X_{t,i}|]$
Online Learning: Prediction with Expert Advice

Prediction with Expert Advice

- Interpret the components of $\mathbf{X}_t \in [0, 1]^d$ as predictions of $d$ experts, who are predicting $Y_t \in \{0, 1\}$.
- Our choice $P_\theta$ is a probability distribution on these $d$ experts.
- $\ell(\mathbf{X}_t, Y_t, \theta) = |Y_t - \mathbb{E}_{P_\theta(i)}[X_{t,i}]| = \mathbb{E}_{P_\theta(i)}[|Y_t - X_{t,i}|]$.

Suppose i.i.d. expert losses...

- Suppose $|Y_t - X_{t,i}|$ are i.i.d. with mean $\mu_i = \mathbb{E}_{\mathbf{X}_t, Y_t}[|Y_t - X_{t,i}|]$.
- Let $i^* = \text{arg min}_i \mu_i$.

Proposition ([Koolen, Grünwald, and van Erven, 2016])

Then the $(B, 1)$-Bernstein condition is satisfied with

$$B = \min_{i \neq i^*} \frac{\mathbb{E}_{Y_t, X_{t,i}}[\left( |Y_t - X_{t,i}| - |Y_t - X_{t,i^*}| \right)^2]}{\mu_i - \mu_{i^*}}$$
Achieving Fast Rates in Prediction with Expert Advice

Theorem ([Koolen, Grünwald, and van Erven, 2016])

*If the $(B, \beta)$-Bernstein condition is satisfied for prediction with expert advice, then the Squint algorithm [Koolen and van Erven, 2015] achieves (pseudo)-regret*

\[
\mathbb{E}[\text{Regret}^*_{T}] = O\left( (B \log d)^{\frac{1}{2-\beta}} T^{\frac{1-\beta}{2-\beta}} \right)
\]

\[
\text{Regret}^*_{T} = O\left( (B \log d - \log \delta)^{\frac{1}{2-\beta}} T^{\frac{1-\beta}{2-\beta}} \right) \quad \text{w.p. } \geq 1 - \delta
\]

w.r.t. \( i^* = \arg \min_i \mu_i \).
Bernstein Condition for General Online Learning

Linearizing Losses

In online learning it is common to perform linear approximations of the loss:

\[
\tilde{\ell}(X_t, Y_t, \theta) = \ell(X_t, Y_t, \theta_t) + (\theta - \theta_t)^T \nabla_\theta \ell(X_t, Y_t, \theta_t),
\]

which overestimates the regret.
Bernstein Condition for General Online Learning

Linearizing Losses

In online learning it is common to perform linear approximations of the loss:

\[ \tilde{\ell}(X_t, Y_t, \theta) = \ell(X_t, Y_t, \theta_t) + (\theta - \theta_t)^T \nabla_\theta \ell(X_t, Y_t, \theta_t), \]

which overestimates the regret.

Hinge Loss

- Suppose \((X_t, Y_t)\) are i.i.d., and let \(\theta, X_t\) in the \(d\)-dimensional unit ball
- Hinge loss: \(\ell(X_t, Y_t, \theta) = \max\{Y_t - \theta^T X_t, 0\}\)

**Theorem ([Koolen, Grünwald, and van Erven, 2016])**

*Then the \((B, 1)\)-Bernstein condition is satisfied for \(\tilde{\ell}\) with*

\[ B = \frac{2\lambda_{\text{max}}(\mathbb{E}[XX^\top])}{\| \mathbb{E}[YX] \|} \]
Theorem ([Koolen, Grünwald, and van Erven, 2016])

If the \((B, \beta)\)-Bernstein condition is satisfied for \(\tilde{\ell}\) in general online learning, then the MetaGrad algorithm [Van Erven and Koolen, 2016] achieves (pseudo)-regret

\[
\mathbb{E}[\text{Regret}_{T}^{\theta^*}] = O\left((Bd \log T)^{\frac{1}{2-\beta}} T^{\frac{1-\beta}{2-\beta}}\right)
\]

\[
\mathbb{E}[\text{Regret}_{T}^{\theta^*}] = O\left((Bd \log T - \log \delta)^{\frac{1}{2-\beta}} T^{\frac{1-\beta}{2-\beta}}\right) \quad \text{w.p. } \geq 1 - \delta
\]

w.r.t. \(\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}[\ell(X, Y, \theta)]\).
Achieving Fast Rates in Statistical Learning

- Simplest example: prior $\pi$ on countable model $\Theta = \{\theta_1, \theta_2, \ldots\}$.
- Penalized ERM $\hat{\theta}$ minimizes

$$
\sum_{i=1}^{N} \ell(X_i, Y_i, \theta) + \lambda \log \frac{1}{\pi(\theta)}
$$
Achieving Fast Rates in Statistical Learning

- Simplest example: prior $\pi$ on countable model $\Theta = \{\theta_1, \theta_2, \ldots\}$.
- Penalized ERM $\hat{\theta}$ minimizes

$$\sum_{i=1}^{N} \ell(X_i, Y_i, \theta) + \lambda \log \frac{1}{\pi(\theta)}$$

Proposition (Bernstein Condition Rate)

Under $(B, \beta)$-Bernstein condition, bounded loss, $\lambda = \left( \frac{N}{B \log \frac{1}{\pi(\theta^*)}} \right)^{\frac{1-\beta}{2-\beta}}$ achieves

$$R(\hat{\theta}) - R(\theta^*) = O \left( \frac{B \log \frac{1}{\delta \pi(\theta^*)}}{N} \right)^{\frac{1}{2-\beta}} \quad \text{w.p.} \geq 1 - \delta.$$
Achieving Fast Rates in Statistical Learning

▶ Simplest example: prior $\pi$ on countable model $\Theta = \{\theta_1, \theta_2, \ldots\}$.

▶ Penalized ERM $\hat{\theta}$ minimizes

$$\sum_{i=1}^{N} \ell(X_i, Y_i, \theta) + \lambda \log \frac{1}{\pi(\theta)}$$

Proposition (Bernstein Condition Rate)

Under $(B, \beta)$-Bernstein condition, bounded loss, $\lambda = \left(\frac{N}{B \log \frac{1}{\pi(\theta^*)}}\right)^{\frac{1-\beta}{2-\beta}}$ achieves

$$R(\hat{\theta}) - R(\theta^*) = O \left( \frac{B \log \frac{1}{\delta \pi(\theta^*)}}{N} \right)^{\frac{1}{2-\beta}} \quad \text{w.p.} \geq 1 - \delta.$$ 

▶ Simple approach: estimate $\lambda$ using cross-validation

▶ Or sophisticated approaches:
  ▶ Slope heuristic (Birgé, Massart)
  ▶ Lepski’s method
  ▶ Safe Bayes (Grünwald)
Summary

Conditions for fast rates *all the same or closely related*:

- **Central Condition**: density estimation
- Pseudo-probability convexity: convex set of pseudo-probabilities
- Stochastic mixability (stronger): bounded squared loss (convex model)
- Bernstein Condition: classification
- Bernstein for Online Learning: gap in prediction with expert advice, hinge loss
Summary

Conditions for fast rates *all the same or closely related:*

- **Central Condition**: density estimation
- Pseudo-probability convexity: convex set of pseudo-probabilities
- Stochastic mixability (stronger): bounded squared loss (convex model)
- Bernstein Condition: classification
- Bernstein for Online Learning: gap in prediction with expert advice, hinge loss

Achieving these *fast rates:*

- In statistical learning: use **cross-validation** to select regularization parameter
- In online learning: **Squint** (experts), **MetaGrad** (general online learning)
Van Erven, Grünwald, Mehta, Reid, Williamson. **Fast Rates in Statistical and Online Learning.** Journal of Machine Learning Research, 2015. (Special issue dedicated to the memory of Alexey Chervonenkis.)


