Boards online
Exercises for validation
Book for sale! GO

Data (x_i, y_i) $i = 1, \dots, n$
training

$\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$

$f : \mathcal{X} \to \mathcal{Y}$
prediction function

$f_\theta(x) = \theta^\top \varphi(x)$

$\hat{\theta} \in \arg\min \hat{\mathcal{R}}(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

(1) how to get $\hat{\theta}$ => optimization

(2) $\mathcal{R}(f_{\hat{\theta}})$? => statistics

$\frac{\sigma^2 d}{n}$ $\frac{1}{\sqrt{n}} \times C$

(1) Convex surrogates for binary classification

$$\mathcal{Y} = \{-1, 1\} \qquad f : \mathcal{X} \longrightarrow \{-1, 1\}$$

current practice: learn $g : \mathcal{X} \longrightarrow \mathbb{R}$

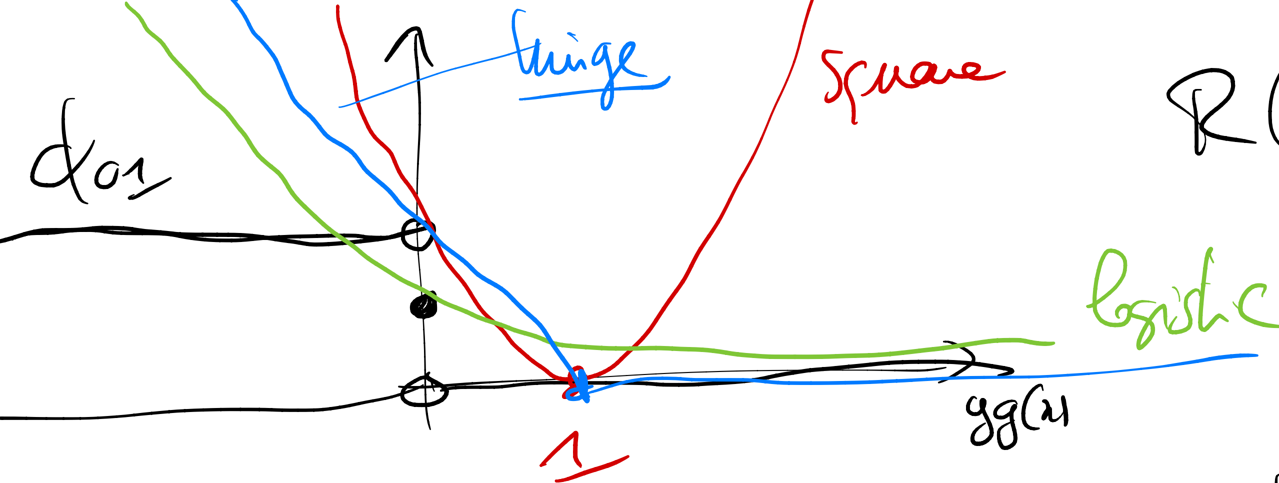
take $f(x) = \text{sign}(g(x)) = \begin{cases} 1 & \text{if } g(x) > 0 \\ -1 & \text{if } g(x) < 0 \\ \text{random in } \{-1, 1\} \end{cases}$

$$\ell(y, f(x)) = 1_{y \neq f(x)}$$

$$= 1_{g(x)=0,\, y \neq f(x)} + 1_{g(x) \neq 0} 1_{y \neq f(x)}$$

$$= \underline{\qquad\qquad} + 1_{g(x) \neq 0} 1_{y g(x) < 0}$$

$$\mathbb{E}[\ell(y, f(x))] = \frac{1}{2}\mathbb{E}(1_{g(x)=0}) + \mathbb{E}(1_{y g(x) < 0}) = \mathbb{E}[\Phi_{01}(y g(x))]$$



$$R(f) = \mathbb{E}\left[\phi_{01}(y\,g(x))\right]$$

convex surrogates: ① square loss $\phi(u) = (u-1)^2$

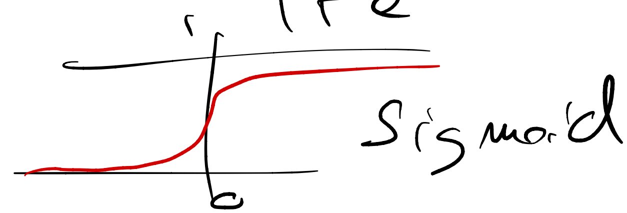
$$\phi(y\,g(x)) = (y\,g(x) - 1)^2$$
$$= (g(x) - y)^2$$

② logistic loss: $\phi(u) = \log(1 + e^{-u})$

convex, smooth

"log-loss" interpretation as maximum likelihood

$p(y=1 \mid x) = \sigma(y\,g(x))$ where $\sigma(u) = \dfrac{1}{1+e^{-u}}$



$$-\log p(y \mid x) = \phi(y\,g(x))$$

③ hinge loss: $\phi(u) = (1-u)_+$ SVM

non smooth

Consistency of minimising the ϕ-risk $R_\phi(g) = \mathbb{E}(\phi(y\, f(x)))$

<u>Theorem</u> (McAuliffe, Bartlett, Jordan, 2006, Zhang, 2005)

if ϕ is cvx and $\phi'(0) < 0$

then there exists a function G s.t

$$G\left(R_{01}(g) - R_{01}^*\right) \leq R_\phi(g) - R_\phi^*$$ Excess ϕ risk

G is increasing, $G(0) = 0$

<u>Ex</u>: hinge loss. $G(u) = u$ → no loss

logistic / square loss. $G(u) = u^2$ → loss in the rate

<u>Now</u>: $y = \mathbb{R}$ $\quad \ell(y, f(x)) = (y - f(x))^2$

$f: \mathcal{X} \to \mathbb{R}$ $\quad = \log(1 + e^{-y f(x)})$

Decomposition of the risk : $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$

$\mathcal{F}$ function space

$\forall f \in \mathcal{F}$

$$\mathcal{R}(f) - \mathcal{R}^* = \underbrace{\mathcal{R}(f) - \inf_{g \in \mathcal{F}} \mathcal{R}(g)}_{\text{estimation error} \geq 0} + \underbrace{\inf_{g \in \mathcal{F}} \mathcal{R}(g) - \mathcal{R}^*}_{\text{approximation error} \geq 0}$$

Assumption : $\mathcal{F} = \{ f_\theta(x) = \theta^\top \varphi(x), \ \|\theta\|_2 \leq D \}$ $\quad \varphi : \mathcal{X} \to \mathbb{R}^d$ feature map

Approximation error

$$\inf_{\|\theta\|_2 \leq D} \mathcal{R}(f_\theta) - \mathcal{R}^* = \underbrace{\inf_{\|\theta\|_2 \leq D} \mathcal{R}(f_\theta) - \overbrace{\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta)}^{\theta_* \text{ is a minimizer}}}_{\text{effect of regularizing}} + \underbrace{\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^*}_{\text{"modelling" error}}$$

if $D \geq \|\theta_*\|_2 \Rightarrow$ equal to 0

if $D < \|\theta_*\|_2 \Rightarrow$?

Assumption: $\ell(y,\cdot)$ is Lipschitz-continuous with constant G $\forall y \in \mathcal{Y}$

$$R(f_\theta) - R(f_{\theta_*}) = \mathbb{E}\big[\ell(y, f_\theta(x)) - \ell(y, f_{\theta_*}(x))\big]$$

will be reused tomorrow

$$\leq \mathbb{E}\big[\, G\,|f_\theta(x) - f_{\theta_*}(x)|\big]$$

$$= G\,\mathbb{E}\big(|(\theta-\theta_*)^\top \varphi(x)|\big)$$

with linear functions

$$\leq G \cdot \|\theta - \theta_*\|_2 \cdot \underbrace{\mathbb{E}\|\varphi(x)\|_2}_{\leq R \text{ if } \|\varphi(x)\|_2 \leq R \text{ almost surely}}$$

$$\leq RG\|\theta - \theta_*\|_2$$

$$\inf_{\|\theta\|_2 \leq D} R(f_\theta) - \overbrace{\inf_{\theta \in \mathbb{R}^d} R(f_\theta)}^{\theta_* \text{ is a minimizer}} \leq \inf_{\|\theta\|_2 \leq D} RG\|\theta - \theta_*\|_2 = RG\big(\|\theta_*\|_2 - D\big)_+$$

simple calculation

Estimation error of $\hat{f} \in \mathcal{F}$ ~ approximate minimizer of $\hat{R}(f)$ on $\mathcal{F}$

$$R(\hat{f}) - \underbrace{\inf_{f \in \mathcal{F}} R(f)}_{} = R(\hat{f}) - R(g_{\mathcal{F}})$$

→ minimizer of R on $\mathcal{F}$

$$= R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - \hat{R}(g_{\mathcal{F}}) + \hat{R}(g_{\mathcal{F}}) - R(g_{\mathcal{F}})$$

$$\leq \underbrace{\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)}_{} + \underbrace{\hat{R}(\hat{f}) - \inf_{f} \hat{R}(f)}_{} + \underbrace{\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f)}_{}$$

uniform deviations $\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$

$$\frac{4GRD}{\sqrt{n}}$$

optimization error $\frac{GRD}{\sqrt{t}}$ where t is # of iterations of GD

⇒ n iterations ⇒ $O(n^2)$

Hoeffding
McDiarmid
Rademacher

inequality
Finite # of models

$C^n 5 C \longrightarrow 11^n 12$

Hoeffding's inequality: $z_i \in \mathbb{R}$ iid

$$\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}[z] \sim \frac{1}{\sqrt{n}}$$

Central limit theorem (CLT)

$$\xrightarrow[\text{law}]{} \text{Normal}\left(0, \frac{\sigma^2}{n}\right)$$

$$\sigma = \text{var}(z)$$

if $z_i \in [0, \sigma]$ almost surely

$$\mathbb{P}\left(\frac{1}{n}\sum_i z_i - \mathbb{E}(z) \geq t\right) \leq \exp\left(-2n\left(\frac{t}{\sigma}\right)^2\right)$$

$$\text{"}\delta\text{"} \qquad \left(\sim e^{-\frac{t^2}{2(\sigma^2/n)}}\right)$$

$$t = \frac{\sigma}{\sqrt{2n}}\sqrt{\log\frac{1}{\delta}}$$

$\Leftrightarrow$ with proba $\geq 1-\delta$, $\quad \frac{1}{n}\sum_i z_i - \mathbb{E}z \leq \frac{\sigma}{\sqrt{2n}}\sqrt{\log\frac{1}{\delta}}$

McDiarmid inequality $Z_1, \dots, Z_n$ iid $\in \mathcal{Z}$

$f : \mathcal{Z}^n \to \mathbb{R}$ "bounded variation"

$$\left| f(z_1, \dots, z_n) - f(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n) \right| \leq c$$

$$\mathbb{P}\left(\left| f(Z_1, \dots, Z_n) - \mathbb{E} f(Z_1, \dots, Z_n) \right| \geq t \right) \leq 2 e^{-\frac{2t^2}{nc^2}}$$

$$\Leftrightarrow \left| f(Z_1, \dots, Z_n) - \mathbb{E} f(Z_1, \dots, Z_n) \right| \leq c \sqrt{\frac{n}{2}} \sqrt{\log \frac{2}{\delta}}$$

goal: bound $\sup_{f \in F} \hat{R}(f) - R(f) = T$

Results on $\mathbb{E}T$ $\xrightarrow{\text{if the loss } \in [0, \ell_\infty)}$ with proba $1-\delta$

$$T \le \mathbb{E}T + \frac{\ell_\infty}{\sqrt{n}}\sqrt{\log \tfrac{1}{\delta}}$$

Application of McDiarmid

$$T = \sup_{f \in F} \left(\frac{1}{n} \sum \ell(y_i, f(x_i)) - R(f) \right)$$

$z_i = (x_i, y_i)$ IID $\rightarrow$ with $c \sim \frac{2\ell_\infty}{n}$

Exercise if $Z \le \sqrt{\log 1/\delta}$ with proba $1-\delta$, $\forall \delta > 0$

$\mathbb{E}Z \le \alpha$

Finite number of models $\mathcal{F} = \{f_1, \dots, f_m\}$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq t\right) = \mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \{\hat{R}(f) - R(f) \geq t\}\right)$$

union bound

$$\leq \sum_{f\in\mathcal{F}} \mathbb{P}(\hat{R}(f) - R(f) \geq t)$$

$$\leq \sum_{f\in\mathcal{F}} e^{-\frac{2mt^2}{\ell_\infty^2}}$$

if $\ell(y, f(x)) \in (0, \ell_\infty)$

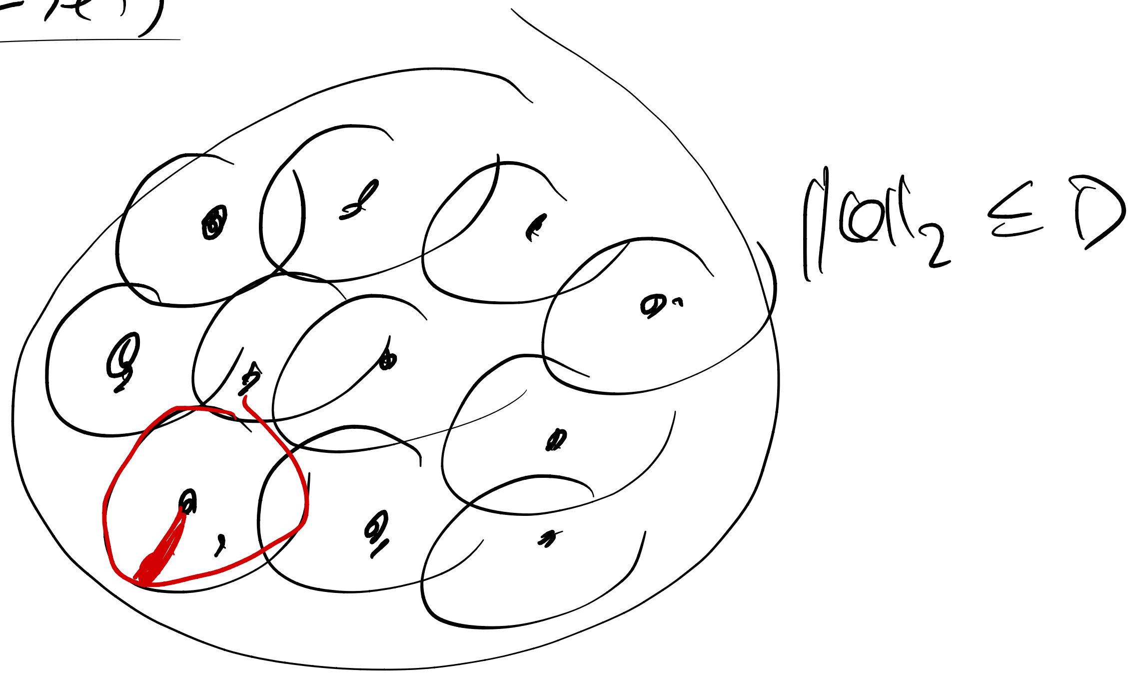
Hoeffding

$$= |\mathcal{F}|\, e^{-\frac{2mt^2}{\ell_\infty^2}}$$

with probability $\geq 1-\delta$

$$\Rightarrow \quad \hat{R}(f) - R(f) \leq \frac{\ell_\infty}{\sqrt{2m}} \sqrt{\log|\mathcal{F}| + \log 1/\delta}$$

Covering numbers



Very common way $\Rightarrow$ Rademacher

Rademacher averages / complexity

$\mathcal{F}$ loss function, data (x_i, y_i) $i = 1 \to n$ = Data D

$\mathcal{H} = \{ (x,y) \mapsto \ell(y, f(x)),\ f \in \mathcal{F} \}$ space of functions from $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

$z \mapsto h(z)$

goal: $\mathbb{E}_D \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i h(z_i) - \mathbb{E} h(z)$

$z_i \sim$ iiD

$\mathbb{E}_D \sup_{h \in \mathcal{H}} \mathbb{E} h(z) - \frac{1}{n} \sum_i h(z_i)$

$$\boxed{R_n(\mathcal{H}) = \mathbb{E}_{D,\varepsilon} \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(z_i) \right\}}$$

where $\varepsilon_i = \begin{matrix} +1 \\ -1 \end{matrix}$ with proba $\frac{1}{2}$ (Rademacher r.v).

Symmetrization lemma: $D = \{ z_i \quad i = 1, \dots, n \}$ iid

independent copy $D' = \{ z'_i \quad \underline{\qquad\qquad} \}$ iid

$$\mathbb{E} \sup_{h \in H} \mathbb{E}[h(z)] - \frac{1}{n} \sum_i h(z_i) = \mathbb{E} \sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[h(z'_i)] - \frac{1}{n} \sum_{i=1}^{n} h(z_i)$$

$$= \mathbb{E} \sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{E}[h(z'_i)(D) - h(z_i)]}_{\mathbb{E}(h(z'_i) - h(z_i) \mid D)}$$

$\sup \mathbb{E} \leq \mathbb{E} \sup$ $\rightarrow$

$$\leq \mathbb{E}\, \mathbb{E}\left[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} [h(z'_i) - h(z_i)) \mid D \right]$$

ε_i independent Rademacher $\rightarrow$

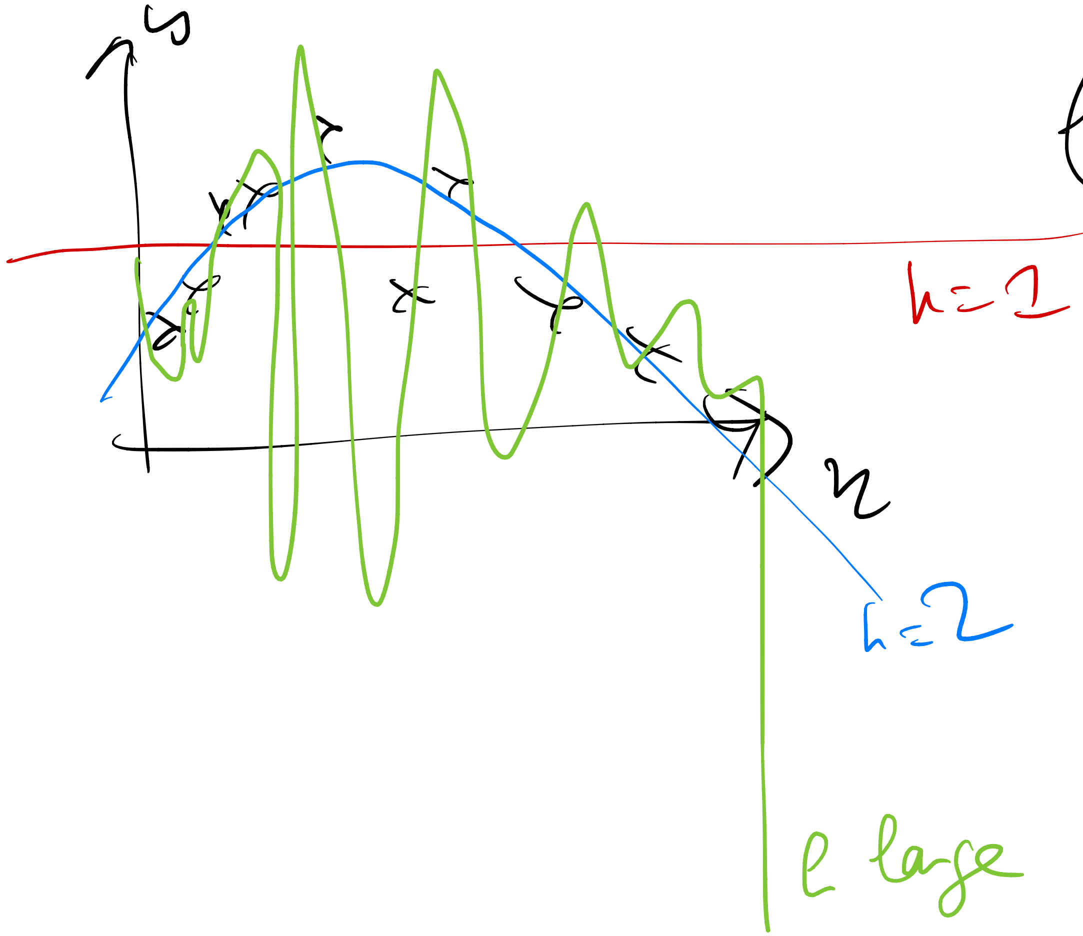
$$= \mathbb{E}_{D, D', \varepsilon} \left[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \{ h(z'_i) - h(z_i) \} \right]$$

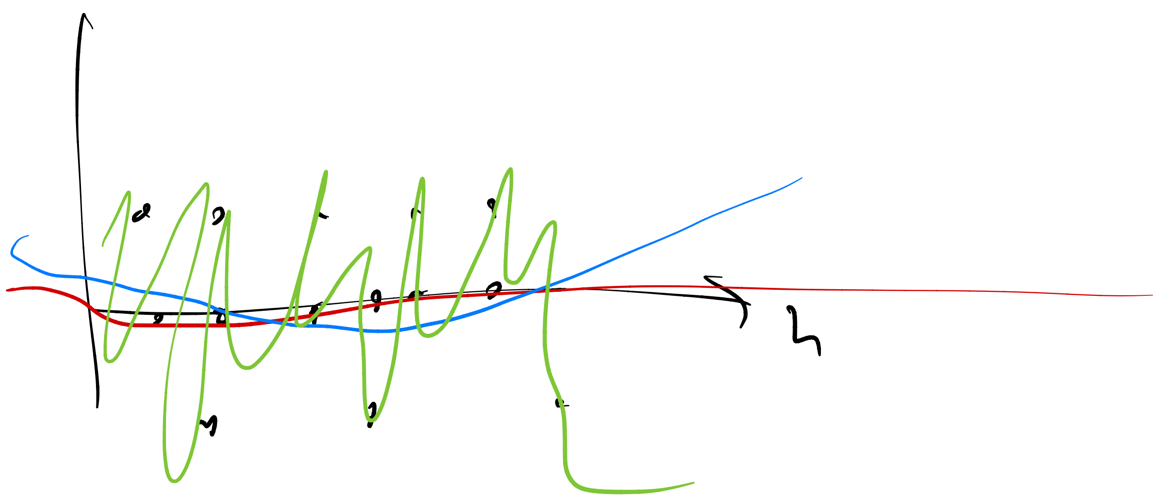
$\sup(A+B) \leq \sup A + \sup B$ $\rightarrow$

$$\leq \mathbb{E} \sup_h \frac{1}{n} \sum_i \varepsilon_i h'(z'_i) + \mathbb{E} \sup \frac{1}{n} \sum_i -\varepsilon_i h(z_i)$$

$$= 2 R_n(H)$$

f degree k polynomials





Contraction principle (Ledoux, Talagrand)

$$\mathbb{E}_\varepsilon \sup_{\theta \in \Theta} \sum_{i=1}^{n} \varepsilon_i \varphi_i(a_i(\theta)) \leq G\, \mathbb{E}_\varepsilon \sup_{\theta \in \Theta} \sum_{i=1}^{n} \varepsilon_i a_i(\theta)$$

Rademacher

G-Lipschitz $f(\cdot) : \mathbb{R} \to \mathbb{R}$

our case

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell(y_i, f(x_i)) \leq G\, \mathbb{E}_\varepsilon \sup_{f} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)$$

G-Lipschitz loss

$$\mathcal{R}_n(\mathcal{F}) \leq G \mathcal{R}_n(\mathcal{F}) \qquad \leq \frac{GRD}{\sqrt{n}}$$

Application to linear models

$f_\theta(x) = \theta^T \varphi(x)$

$\|\theta\|_2 \leq D \quad \|\varphi(x)\|_2 \leq R$

$$R_n(\mathcal{F}) = \mathbb{E}_{\varepsilon, \mathcal{D}} \sup_{\|\theta\|_2 \leq D} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \varphi(x_i) \right)^T \theta$$

by Cauchy Schwarz

$$= \mathbb{E}_{\varepsilon, \mathcal{D}}\, D \cdot \left\| \frac{1}{n} \sum_{i} \varepsilon_i \varphi(x_i) \right\|_2$$

Jensen's inequality

$$\leq D \left(\mathbb{E}_{\varepsilon, \mathcal{D}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \varphi(x_i) \right\|_2^2 \right)^{1/2}$$

$$= \frac{D}{n} \left(\mathbb{E}_{\mathcal{D}} \sum_{i=1}^{n} \underbrace{\|\varphi(x_i)\|^2}_{\leq R^2} \right)^{1/2} \leq \frac{DR}{\sqrt{n}}$$

Estimation error $\leq \underbrace{\dfrac{DR}{\sqrt{n}}}_{\text{opt im error}} \times G \times 2 \times 2$

symmetrize twice uniform deviation