

Boards online
Exercises for validation
Book for sale! €

Data
training

$$(x_i, y_i)_{i=1, \dots, n}$$

$$R(\theta) = E[\ell(y, f(x))]$$

$f: X \rightarrow Y$
production
function

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

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

① how to get $\hat{\theta}$ $\Rightarrow$ optimization

② $\hat{R}(\hat{\theta})$ $\Rightarrow$ statistics

$$\frac{\sum \ell}{n} \quad \frac{1}{\sqrt{n}} \times C$$

① convex surrogate for binary classification

$$y = \{-1, 1\} - f: X \rightarrow \{-1, 1\}$$

current practice, learn $g: X \rightarrow \mathbb{R}$

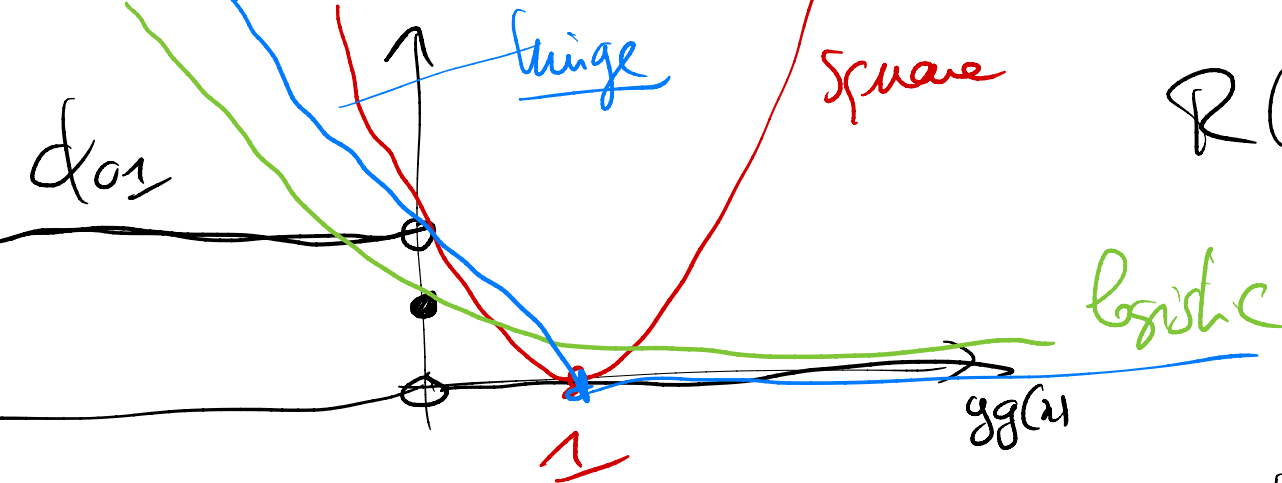
$$\text{take } f(x) = \text{sign}(g(x)) = \begin{cases} 1 & \text{if } g(x) > 0 \\ -1 & \text{if } g(x) < 0 \\ \text{random} & \text{in } \{-1, 1\} \end{cases}$$

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

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

$$= \frac{1}{2} \mathbb{1}_{g(x)=0} + \mathbb{1}_{g(x) \neq 0} \mathbb{1}_{y g(x) < 0}$$

$$\mathbb{E}(\ell(y, f(x))) = \frac{1}{2} \mathbb{E}(\mathbb{1}_{g(x)=0}) + \mathbb{E}(\mathbb{1}_{y g(x) < 0}) = \mathbb{E}[\Phi_{0.5}(y g(x))]$$



$$R(\theta) = E[\phi_{\theta_1}(yg(x))]$$

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

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

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

"log-loss"

convex, smooth

interpreted as maximum likelihood

$p(y=1|x) = \sigma(yg(x))$ where $\sigma(u) = \frac{1}{1 + e^{-u}}$

$$-\log p(y|x) = \phi(yg(x))$$



③ hinge loss: $\phi(u) = (1-u)_+$ SVM

max smooth

Consistency of minimising the Φ -risk $R_{\Phi}(g) = E[\Phi(y f(x))]$

Theorem (McAlleiffe, Bartlett, Tsybakov, Zhang, 2005)

if Φ is convex and $\Phi'(c) < 0$
then there exists a function G s.t

$$G(R_{\Phi}(g) - R_{\Phi}^*) \leq R_{\Phi}(g) - R_{\Phi}^* \quad \text{Excess risk}$$

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

Ex: hinge loss. $G(u) = u \rightarrow$ no loss
logistic / square loss. $G(u) = u^2 \rightarrow$ loss in the rate

Nota: $y = \mathbb{R}$
 $f: X \rightarrow \mathbb{R}$
 $\ell(y, f(x)) = (y - f(x))^2$
 $= \log(1 + e^{-y f(x)})$

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

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

$$R(f) - R^* = R(f) - \inf_{g \in \mathcal{F}} R(g) + \inf_{g \in \mathcal{F}} R(g) - R^*$$

stochastic error ≥ 0

approximation error ≥ 0

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

Approximation error

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

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

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

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

$$\begin{aligned} R(\hat{\theta}) - R(\theta_*) &= \mathbb{E} \left[\ell(y, \hat{\theta}(x)) - \ell(y, \theta_*(x)) \right] \\ &\leq \mathbb{E} \left[L \|\hat{\theta}(x) - \theta_*(x)\| \right] \end{aligned}$$

will be
reused
formula

$$= L \mathbb{E} \left[|(\hat{\theta} - \theta_*)^T \phi(x)| \right]$$

with
linear
functions



$$\leq L \cdot \|\hat{\theta} - \theta_*\|_2 \cdot \underbrace{\mathbb{E} \|\phi(x)\|_2}_{\leq R \text{ if } \|\phi(x)\|_2 \leq R}$$

almost
surely

$$\leq LR \|\hat{\theta} - \theta_*\|_2$$

θ_* is a minimizer

$$\inf_{\|\theta\|_2 \leq D} R(\hat{\theta}) - \inf_{\theta \in \mathbb{R}^d} R(\hat{\theta}) \leq \inf_{\|\theta\|_2 \leq D} LR \|\theta - \theta_*\|_2 = LR (\|\theta_*\|_2 - D)_+$$

↑ simple
calculation

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

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - R(g_{\mathcal{F}}) \quad \text{--- minimizer of } R \text{ on } \mathcal{F}$$

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

$$\leq \underbrace{\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)}_{\text{uniform deviation}} + \underbrace{\hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \hat{R}(f)}_{\text{optimization error}} + \underbrace{\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f)}_{\text{uniform deviation}}$$

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

optimization error

$$\frac{GPD}{\sqrt{t}}$$

where

t is # of evaluations of GD

$\Rightarrow$ number of iterations $\Rightarrow O(n^2)$

$$\frac{4GPD}{\sqrt{m}}$$

Hoeffding
McDermid
Rendell

inequality
Finite # of models

$|C| \leq n$ $\rightarrow n \geq 2$

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 $Z_i \in [c, d]$ almost surely

$$P\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z] \geq t\right) \leq \exp\left(-2n\left(\frac{t}{\sigma}\right)^2\right)$$

$\delta'' \quad \left(\sim e^{-\frac{t^2}{2(\sigma^2/n)}} \right)$

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

$$\text{If } \mathbb{P}(c \leq S \leq d) \geq 1 - \delta, \quad \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z] \leq \frac{\sigma}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}}$$

MacDermid inequality, $z_1, \dots, z_n$ iid $\in \mathbb{Z}$

$f: \mathbb{Z}^n \rightarrow \mathbb{R}$ "banded function"

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

$$\mathbb{P}(|f(z_1, \dots, z_n) - \mathbb{E} f(z_1, \dots, z_n)| \geq t) \leq 2e^{-\frac{t^2}{nc^2}}$$

$$\mathbb{P}(|f(z_1, \dots, z_n) - \mathbb{E} f(z_1, \dots, z_n)| \leq c \sqrt{\frac{n}{2}} \sqrt{\log \frac{2}{\delta}})$$

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

Results are $\mathbb{E}T \xrightarrow{\text{if the loss } \in [0, L_\infty)} \text{ with prob } 1 - \delta$

$$T \leq \mathbb{E}T + \frac{L_\infty}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}$$

Application of McDiarmid

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

$z_i = (x_i, y_i) \text{ i.i.d.} \Rightarrow \text{with } C \sim \frac{2L_\infty}{n}$

Exercise of $Z \leq \sqrt{\log \frac{1}{\delta}}$ with prob $(1 - \delta, \forall \delta) \circ$
 $\mathbb{E}Z \leq 2$

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

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

$$\leq \sum_{f \in \mathcal{F}} e^{-\frac{2nt^2}{L^2}}$$

$$= |\mathcal{F}| e^{-\frac{2nt^2}{L^2}}$$

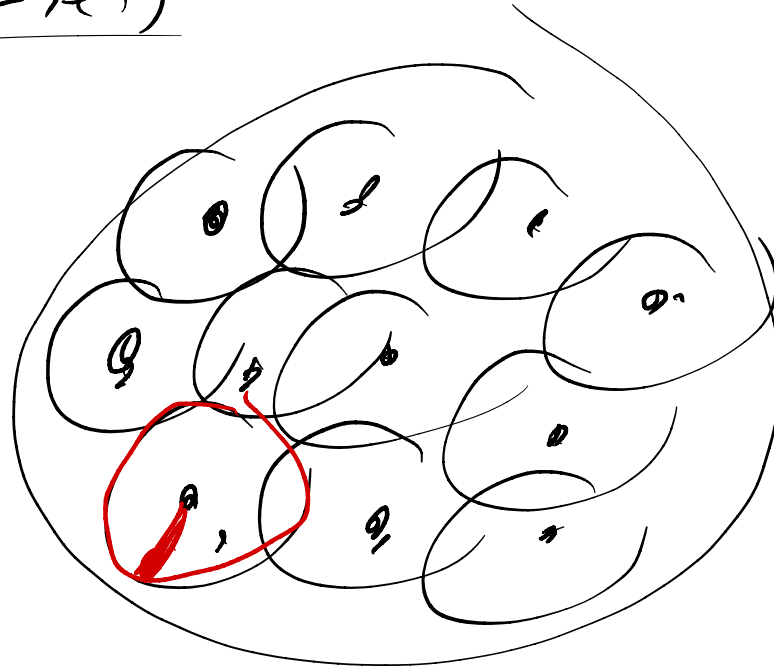
$$\forall \epsilon, \exists n \in \mathbb{N} \text{ s.t. } \mathbb{P}(\hat{R}(f) - R(f) \leq \epsilon) \geq 1 - \delta$$

Hoeffding

with probability $\geq 1 - \delta$

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

Covering numbers



$$\|a\|_2 \in D$$

Very common way $\Rightarrow$ Rademacher

Rademacher averages / complexity
 $\mathcal{F}$ loss function, data $(x_i, y_i)_{i=1 \rightarrow n} = \text{Data}$
 $\mathcal{H} = \{ (x, y) \mapsto \ell(y, f(x)), f \in \mathcal{F} \}$ space of functions
 $z \mapsto h(z)$ for $x \in \mathcal{X} \rightarrow \mathbb{R}$

goal: $\mathbb{E}_{\mathcal{D}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i h(z_i) - \mathbb{E} h(z)$
 $z_i \sim \text{i.i.d.}$

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

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

where $\varepsilon_i = \pm 1$ with prob $\frac{1}{2}$ (Rademacher r.v.).

Symmetrization lemma: $D = \{z_i \quad i=1, \dots, n\}$ iid
independent copy $D' = \{z'_i \quad i=1, \dots, n\}$ iid

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

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

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

ε_i independent Rademacher $\rightarrow$

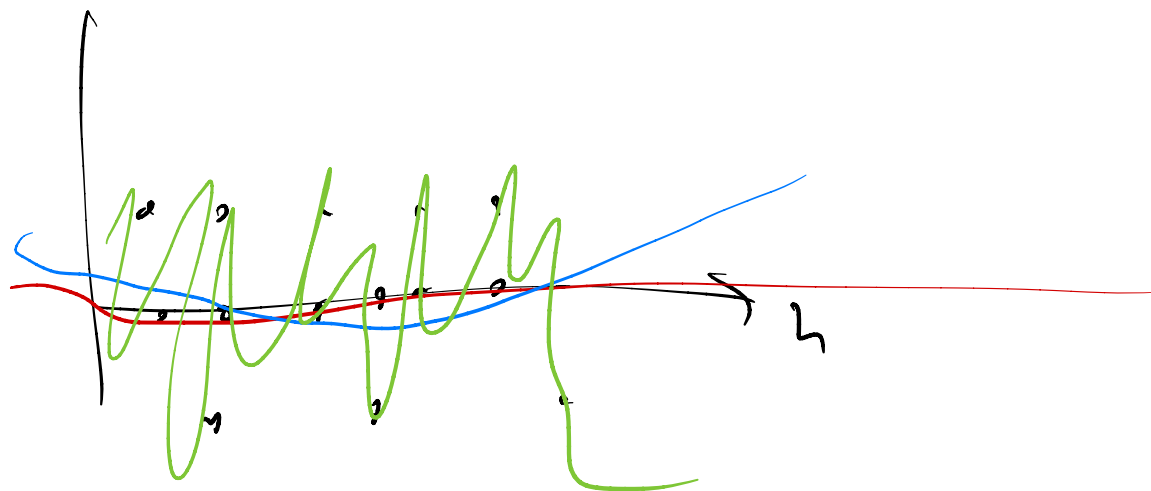
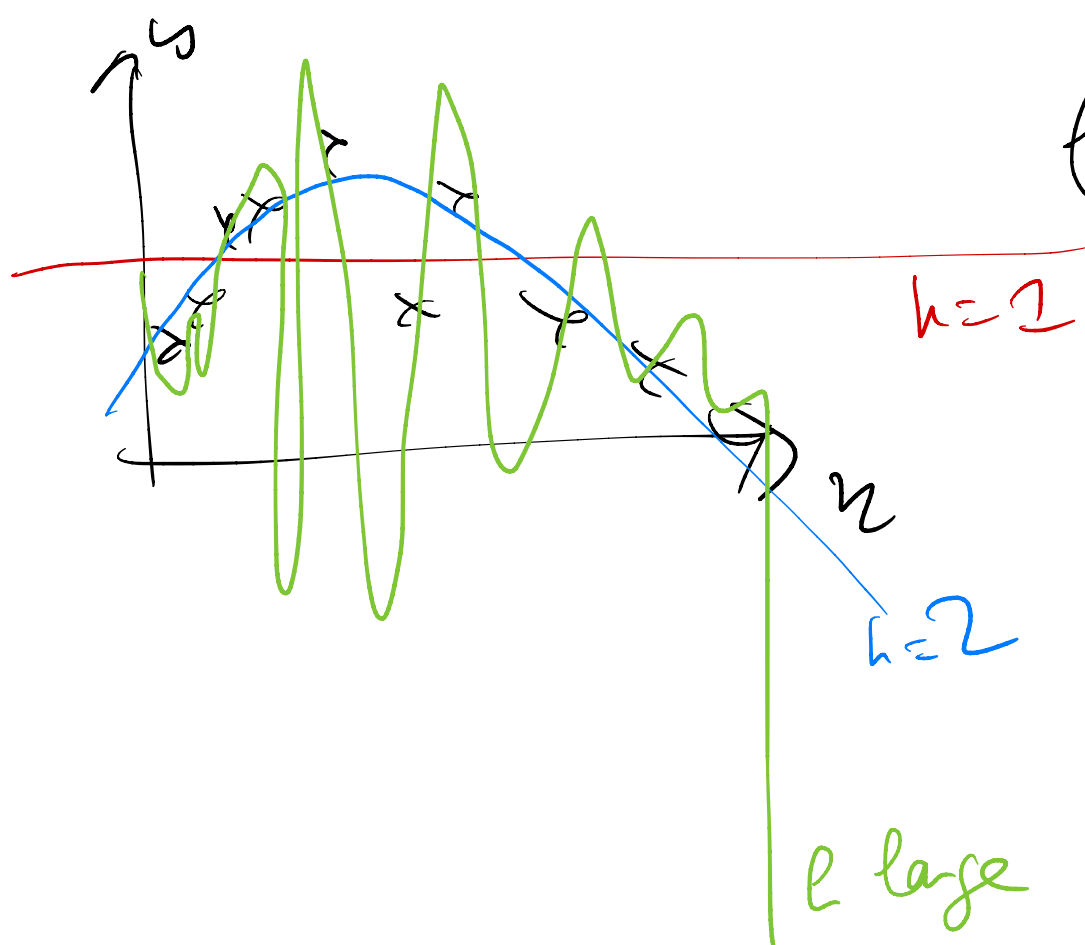
$$= \mathbb{E}_{(D, D')} \mathbb{E}_{\varepsilon} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(h(z'_i) - h(z_i) \right) \right]$$

$$\rightarrow \leq \mathbb{E} \sup_h \frac{1}{n} \sum_{i=1}^n \varepsilon_i h'(z'_i) + \mathbb{E} \sup_h \frac{1}{n} \sum_{i=1}^n -\varepsilon_i h(z_i)$$

$$= 2 R_n(\mathcal{H})$$

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

f degree h polynomials



Contraction principle (Ledoux, Talagrand)

$$\mathbb{E}_\varepsilon \sup_{Q \in \mathcal{Q}} \sum_{i=1}^n \varepsilon_i \varphi_i(Q_i(\omega)) \leq G \mathbb{E}_\varepsilon \sup_{Q \in \mathcal{Q}} \sum_{i=1}^n \varepsilon_i Q_i(\omega)$$

Rademacher

G -Lipschitz
 $f(x) = \mathbb{R} \rightarrow \mathbb{R}$

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

$$\mathcal{R}_n(f) \leq G \mathcal{R}_n(\mathcal{F})$$

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

Applying to linear models . $f_{\Theta}(x) = \Theta^T \varphi(x)$

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

$$J_n(\Theta) = \mathbb{E}_{\mathcal{D}} \sup_{\|\Theta\|_2 \leq D} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot \varphi(x_i) \right)^T \Theta$$

sg Cauchy Schwarz

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

Jensen's inequality

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

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

Estimating error $\leq \frac{DR}{\sqrt{n}} \times G \times 2 \times 2$

+ optim error

summed up twice