Least-squares regression -

Estimator

OLS $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$

Model: $y = \Phi \theta_* + \varepsilon$

Regularization

$\frac{\sigma^2 d}{n} \longrightarrow$ Ridge

dimension independent

TODAY:

- Convex surrogates
- Estimation error for ERM

Data $x_i, y_i \quad i = 1, \dots, n$

$x_i \in \mathcal{X}, \quad y_i \in \mathcal{Y}$

loss function : $\ell(y, f(x))$

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

↳ prediction function

empirical risk : $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$ Empirical

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

goal : any $f(x) \in \mathbb{R}$

ok for regression $\ell(y, f(x)) = (y - f(x))^2$

not if $\mathcal{Y} = \{-1, 1\}$

- Convex surrogates for binary classification and 0-1 loss

- going from $g(x) \in \mathbb{R}$ to $f(x) \in \{-1, 1\}$

consider $f(x) = \text{sign}(g(x))$ $= 1$ if $g(x) > 0$

$= -1$ if $g(x) < 0$

( $=$ random if $g(x) = 0$ )

$R(g)$ Risk of $g$ is the risk $R(f)$

$(= R_{01}(g))$ $\quad R(f) = \mathbb{P}(f(x) \neq y) = \mathbb{E}\left[\mathbb{1}_{f(x) \neq y}\right]$

$= \mathbb{P}\left[f(x) \neq y, g(x) > 0\right] + \mathbb{P}\left[f(x) \neq y, g(x) < 0\right]$

$+ \mathbb{P}\left(f(x) \neq y, g(x) = 0\right)$

$R(g)$ Risk of $g$ vs the risk $R(f)$

$(= R_{01}(g))$

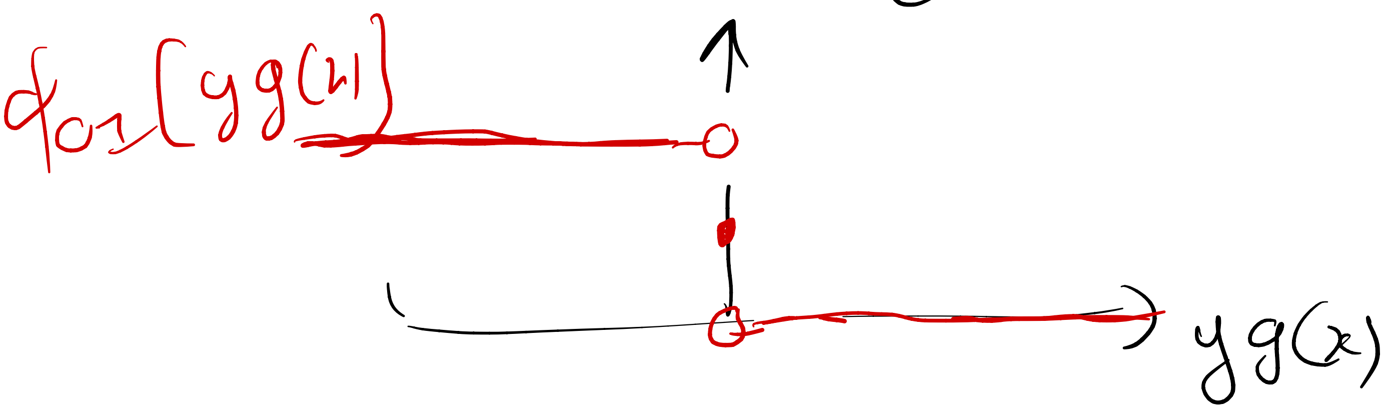
$$R(f) = \mathbb{P}(f(x) \neq y) = \mathbb{E}\left[\mathbb{1}_{f(x)\neq y}\right]$$

$$= \underbrace{\mathbb{P}[f(x) \neq y, g(x) > 0] + \mathbb{P}[f(x) \neq y, g(x) < 0]}_{\mathbb{P}(yg(x) < 0)}$$

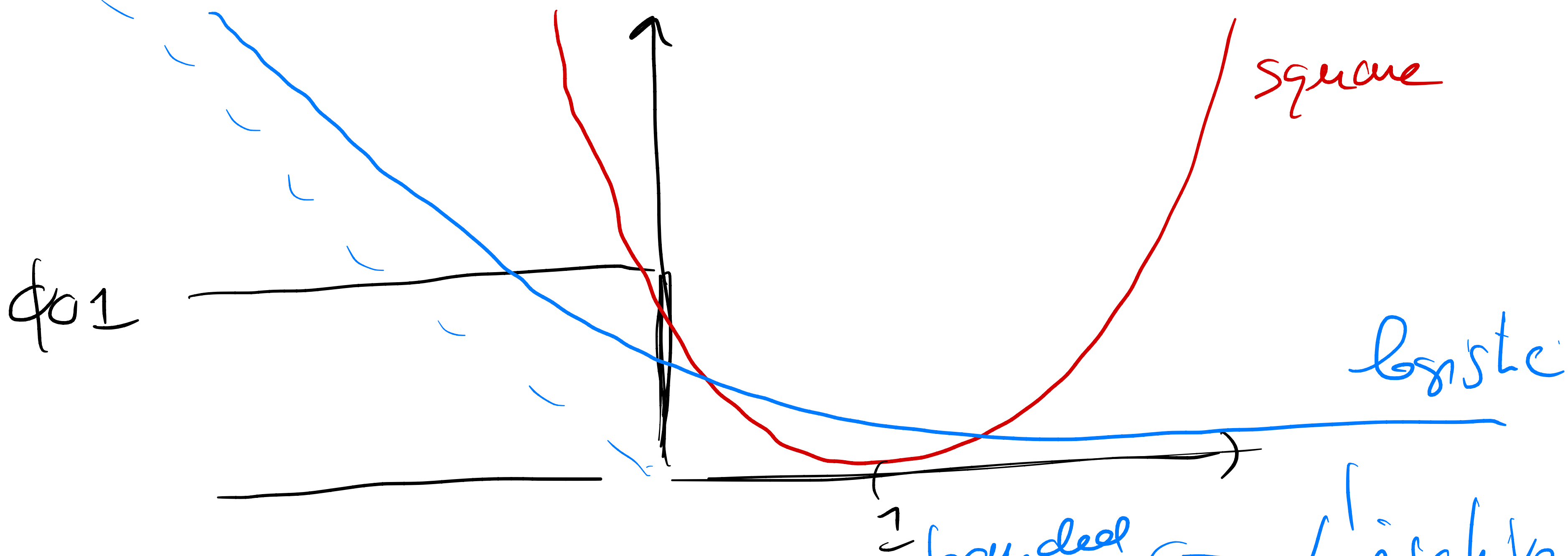
$$+ \underbrace{\mathbb{P}(f(x) \neq y, g(x) = 0)}_{\frac{1}{2}\mathbb{E}\left[\mathbb{1}_{g(x)=0}\right]}$$

$$R(g) = \mathbb{E}\left[\mathbb{1}_{yg(x)<0} + \frac{1}{2}\mathbb{1}_{yg(x)=0}\right]$$

$$= \mathbb{E}\left[\phi_{01}(yg(x))\right]$$




$$\phi_{01}(u) = \begin{cases} 0 & \text{if } u > 0 \\ 1 & \text{if } u < 0 \\ \frac{1}{2} & \text{if } u = 0 \end{cases}$$




bounded gradient $\Leftarrow$ Lipschitz continuous

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

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

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

probabilistic model : $\mathbb{P}(y=1|x) = \sigma(g(x))$

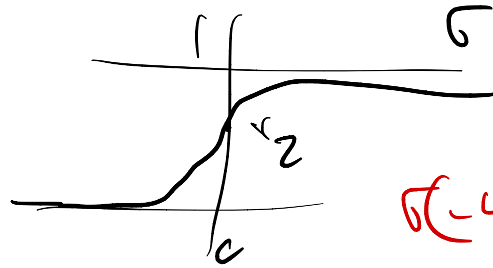
└ sigmoid

$$\mathbb{P}(y|x) = \sigma(yg(x))$$

$$\sigma(u) = \frac{1}{1+e^{-u}}$$

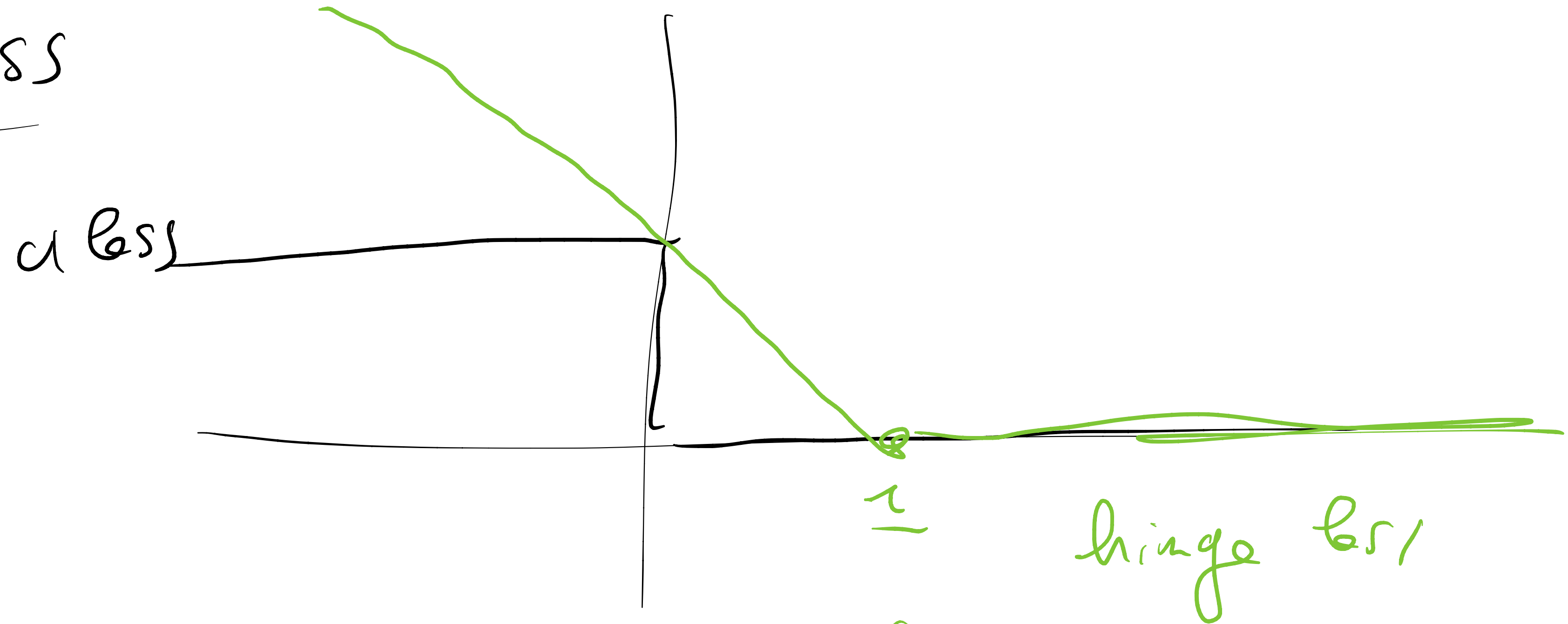
maximum likelihood

max $\log \mathbb{P}(y|x)$




$\sigma(-u) = 1-\sigma(u)$

# Hinge loss




$(1-u)_+$

link with support vector machine (SVM)

---

0-1 loss binary classification

$$R(f) = \mathbb{E}\,\phi_{01}(y\,g(x)) = \mathbb{E}\,\phi(y\,g(x))$$

values in $\{-1,1\}$ values in $\mathbb{R}$

0-1 loss binary classification

$$R(f) = \mathbb{E}\,\phi_{01}(y g(x)) = \mathbb{E}\,\phi(y g(x))$$

values in $\{-1,1\}$ — values in $\mathbb{R}$

(1) <u>Classification calibration</u> = if $g^*(x)$ which minimizes the expected $\phi$-risk $\mathbb{E}\left(\phi(y g(x))\right)$

Assumption: $\phi$ cvx / $\phi'(0) < 0$

then $\text{sign}(g^*(x))$ is a Bayes predictor

if square loss $g^*(x) = \mathbb{E}(y|x)$
$\text{sign}(\mathbb{E}(y|x))$ is the Bayes predictor

(2) There is a calibration function

$$R_a(g) - R_a^* \leq H\left[R_\phi(g) - R_\phi^*\right]$$

( $H(u) = u$ hinge loss), $\sqrt{u}$ for square / log

## Risk decomposition = $\mathcal{F}$ family of functions : $\mathcal{X} \to \mathbb{R}$

ERM $\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min}\ \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

↳ square
↳ logistic
↳ hinge

in most cases : $\mathcal{F} = \{ f_\theta, \theta \in \Theta \}$

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

$\mathbb{R}^4\ 15 \to \mathbb{R}^4\ 39$

Approximation error $= \mathcal{F} = \{f_\theta;\ \theta \in \Theta \subset \mathbb{R}^d\}$

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

$$\inf_{\theta\in\Theta} R(f_\theta) - R^* = \underbrace{\inf_{\theta\in\Theta} R(f_\theta) - \inf_{\theta\in\mathbb{R}^d} R(f_\theta)}_{\text{effect of restricting } \theta\in\Theta} + \underbrace{\inf_{\theta\in\mathbb{R}^d} R(f_\theta) - R^*}_{\text{incompressible approx error}}$$

ex: $\Theta = \{\theta,\ \|\theta\|_2 \le D\}$

Example: linear model with G-Lipschitz-continuous loss

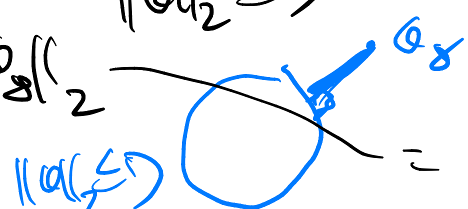
Model $f_\theta(x) = \theta^T\varphi(x)$, with optimal $f_\theta = f_{\theta^*}$

Bounding $\inf_{\|\theta\|_2\le D} R(f_\theta) - R(f_{\theta_*}) = \inf_{\|\theta\|_2\le D} \mathbb{E}\left[\ell(y, f_\theta(x)) - \ell(y, f_{\theta_*}(x))\right]$

$$\le \inf_{\|\theta\|_2\le D} \mathbb{E}\big[G\,|\underbrace{f_\theta(x) - f_{\theta^*}(x)}_{(\theta-\theta_*)^T\varphi(x)}|\big]$$

Cauchy schwarz

$$\le \inf_{\|\theta\|_2\le D} G\,\mathbb{E}(\|\varphi(x)\|_2)\cdot\|\theta-\theta_*\|_2 = G\,\mathbb{E}\|\varphi(x)\|_2\,(\|\theta_*\|_2 - D)_+$$



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

$$= \underbrace{R(\hat{f}) - \hat{R}(\hat{f})}_{} + \underbrace{\hat{R}(\hat{f}) - \hat{R}(g_{\mathcal{F}})}_{\le \varepsilon \text{ optimization error}} + \underbrace{\hat{R}(g_{\mathcal{F}}) - R(g_{\mathcal{F}})}_{}$$

$\frac{1}{n}\sum_i Z_i \to \mathbb{E}Z$

iid

$\frac{1}{n}\sum \underline{\ell(y_i, \hat{f}(x_i))} - R(\cdot)$

$\frac{1}{n}\sum_i \ell(y_i, g_{\mathcal{F}}(x_i)) - \mathbb{E}$

$$\le \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) + \varepsilon + \sup_f \hat{R}(f) - R(f)$$

uniform deviation

$$\le 2 \sup_f \left| R(f) - \hat{R}(f) \right| + \varepsilon$$

goal: control uniform deviations: $\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)$

Two types of guarantees: (1) $\mathbb{E}\left[\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)\right] \leq$ ——

(2) with probability $1-\delta$, $\sup_{f} R(f) - \hat{R}(f) \leq \square(\delta)$.

Review of concentration inequality

(1) Markov inequality $\mathbb{P}(Z > t) \leq \boxed{\dfrac{\mathbb{E}[Z]}{t} = \delta}$ Solve

if $Z \geq 0$

$\Leftrightarrow$ with probability at least $1-\delta$, $Z \leq \dfrac{\mathbb{E}[Z]}{\delta}$

(2) Hoeffding's inequality. if $Z_1, \dots, Z_n \in [0,1]$ and independent

then $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Z_i] > t\right) \leq e^{-2nt^2} = \delta$

with proba at least $1-\delta$, $\frac{1}{n}\sum_i Z_i - \frac{1}{n}\sum_i \mathbb{E}(Z_i) \leq \sqrt{\log\frac{1}{\delta}\,\frac{1}{2n}}$

Concentration inequality for $R(f) - \hat{R}(f)$ for fixed $f$

of loss function $\in [0, \ell_\infty]$, $Z_i := -\ell(y_i, f(x_i))$

Hoeffding
$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E} Z_i - \frac{1}{n}\sum_{i=1}^{n} Z_i}_{R(f) - \hat{R}(f)} \le \ell_\infty \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$$

" $\cancel{\Rightarrow}$ " with probability $1-\delta$ $\quad Z = \boxed{\sup_{f \in F} R(f) - \hat{R}(f)} \le$ ___

$\Rightarrow$ look at Section 4.4.1 on MacDiarmid inequality

with probability at least $1-\delta$, $\quad Z \le \mathbb{E} Z + \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}}$

## How to bound uniform deviations?

- finite class of models
- general case

### ① Finite class of models

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

union bound

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

$$\leq \sum_{f\in\mathcal{F}} e^{-2nt^2/\ell_\infty^2} = |\mathcal{F}|\, e^{-2nt^2/\ell_\infty^2} \overset{"}{=} \delta$$

$$t=\frac{\ell_\infty}{\sqrt{2n}}\left(\sqrt{\log\frac{1}{\delta}+\log|\mathcal{F}|}\right)$$

with proba $\geq 1-\delta$, $\sup_{f\in\mathcal{F}} R(f)-\hat{R}(f)\leq$

Rademacher complexity: $\mathcal{F}$ family of functions: $\mathcal{X} \to \mathbb{R}$

$\mathcal{H} = \{ (x,y) \mapsto \ell(y, f(x)), f \in \mathcal{F} \}$.

family of functions from $\mathcal{X} \times \mathbb{R} \to \mathbb{R}$.

$$\sup_f R(f) - \hat{R}(f) = \sup_{h \in \mathcal{H}} \mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^{n} \underbrace{h(z_i)}_{\ell(y_i, f(x_i))}$$

goal: bound the expectation of

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

Rademacher complexity of $\mathcal{H}$

$z_1, \dots, z_n$

$\varepsilon \in \{-1, 1\}^n$ with independent uniform values

$\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$ Rademacher r.v

Symmetrization : $z_1, \dots, z_n$ i.i.d. = data $D$

independent copy $z'_1, \dots, z'_n$ i.i.d $\to$ data $D'$

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

because $D$ & $D'$ are indep.

$$= \mathbb{E}\Big[\sup_{h \in H} \frac{1}{n}\sum_i \mathbb{E}[h(z'_i) - h(z_i) \,|\, D]\Big]$$

because

$$\mathbb{E}[h(z_i) \,|\, D] = h(z_i)$$

$$= \mathbb{E}\Big[\sup_h \mathbb{E}\Big(\frac{1}{n}\sum_i \big(h(z'_i) - h(z_i)\big) \,\Big|\, D\Big)\Big]$$

"$\sup \mathbb{E} \le \mathbb{E} \sup$"

$$\le \mathbb{E}\Big(\mathbb{E}\Big(\sup_h \frac{1}{n}\sum_i \big(h(z'_i) - h(z_i)\big) \,\Big|\, D\Big)\Big)$$

$$= \mathbb{E}_{\substack{data \\ \varepsilon}}\Big[\sup_h \frac{1}{n}\sum_i \varepsilon_i \big(h(z'_i) - h(z_i)\big)\Big]$$

because $h(z'_i) - h(z_i)$ "symmetric"

$$\le \mathbb{E}_{\substack{data \\ \varepsilon}}\Big[\sup_h \frac{1}{n}\sum_i \varepsilon_i h(z'_i) + \sup_h \frac{1}{n}\sum_i (-\varepsilon_i) h(z_i)\Big]$$

$$= 2 R_n(H) = 2\, \mathbb{E}_{data, \varepsilon}\Big(\sup_h \frac{1}{n}\sum_i \varepsilon_i h(z_i)\Big)$$

$$\mathbb{E}[\text{estimation error}] \boxed{\leq} 4\, R_n(\mathcal{H}) = 4\,\mathbb{E}_{\text{data},\varepsilon}\Big[\sup_h \frac{1}{n}\sum_i \varepsilon_i\, h(z_i)\Big]$$

$$= 4\,\mathbb{E}_{\text{data},\varepsilon}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \varepsilon_i\, \ell(y_i, f(x_i))\Big]$$

<u>Contraction principle</u>: if the loss fct is $G$-Lipschitz continuous

then $\boxed{R_n(\mathcal{H}) \leq G\cdot R_n(\mathcal{F})}$ $= G\,\mathbb{E}_{\text{data},\varepsilon}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \varepsilon_i f(x_i)\Big]$

<u>Example</u>: $f_\theta(x) = \theta^\top\varphi(x)$, $\|\theta\|_2 \leq D$

Cauchy-Schwarz linear in $\theta$

$$R_n(\mathcal{F}) = \mathbb{E}_{\text{data},\varepsilon} \sup_{\|\theta\|_2\leq D}\Big(\frac{1}{n}\sum_i \varepsilon_i\varphi(x_i)\Big)^\top\theta = \mathbb{E}_{\text{data},\varepsilon}\, D\,\Big\|\frac{1}{n}\sum_i \varepsilon_i\varphi(x_i)\Big\|_2$$

$\mathbb{E}\|A\|_2 \leq \sqrt{\mathbb{E}\|A\|_2^2}$

Jensen's inequality

$$\leq D\sqrt{\mathbb{E}_{\text{data},\varepsilon}\Big\|\frac{1}{n}\sum_i \varepsilon_i\varphi(x_i)\Big\|_2^2} = D\sqrt{\frac{1}{n^2}\sum_i \mathbb{E}\,\underbrace{\varepsilon_i^2}_{1}\,\|\varphi(x_i)\|_2^2}$$

$$= \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}\|\varphi(x)\|^2}$$

# Summary for linear models

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

$\Theta = \{ \theta, \|\theta\|_2 \leq D \}$.

Empirical risk minimizer

$\theta_* = \arg\min_{\theta} \mathcal{R}(f_\theta)$.

$\|\varphi(x)\|_2 \leq R$

Approximation error $\leq GR(\|\theta_*\|_2 - D)_+$

+

Estimation error $\leq 4G \dfrac{RD}{\sqrt{n}}$

$$D = \|\theta_*\|_2 \Rightarrow \underbrace{\frac{4GR\|\theta_*\|_2}{\sqrt{n}}}_{\text{excess risk compared to } \mathcal{R}(f_{\theta_*})}$$

MAGIC 3 : t iterations of GD

optim error : $GRD/\sqrt{t}$