last week: least-squares

$$\ell(y, f(x)) = |y - \varphi(x)^T\theta|^2 \qquad \varphi(x) \in \mathbb{R}^d$$

two results: excess expected risk $\frac{\sigma^2 d}{n}$ (OLS)

model $y = \varphi(x)^T\theta_* + \varepsilon$

$\mathbb{E}\varepsilon = 0$

$\mathbb{E}\varepsilon^2 = \sigma^2$

ridge regression

$$\frac{D}{\sqrt{n}}$$

generic set-up: Expected risk

$$R(f) = \mathbb{E}(\ell(y, f(x))).$$

find bounds on excess (expected) risk: $R(f) - R^*$

Binary classification: $\mathcal{Y} = \{-1, 1\}$ $\neq \mathbb{R}$

⚠ other convention $\{0,1\}$ or $\{1,2\}$
$\Rightarrow$ see chapter 13

f prediction function: $f: \mathcal{X} \to \mathcal{Y} = \{-1,1\}$
(Vapnik - Chervonenkis dimension)

learn $g: \mathcal{X} \to \mathbb{R}$ and define $f(x) = \operatorname{sign}(g(x))$

$$\operatorname{sign}(a) = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \\ ? & \text{if } a = 0 \end{cases}$$

Question: what is the risk of f as a function of g?

$$R(g) = R(\mathrm{sign}\circ g) = \mathbb{E}\left(\mathbb{1}_{\mathrm{sign}(g(x))\neq y}\right)$$

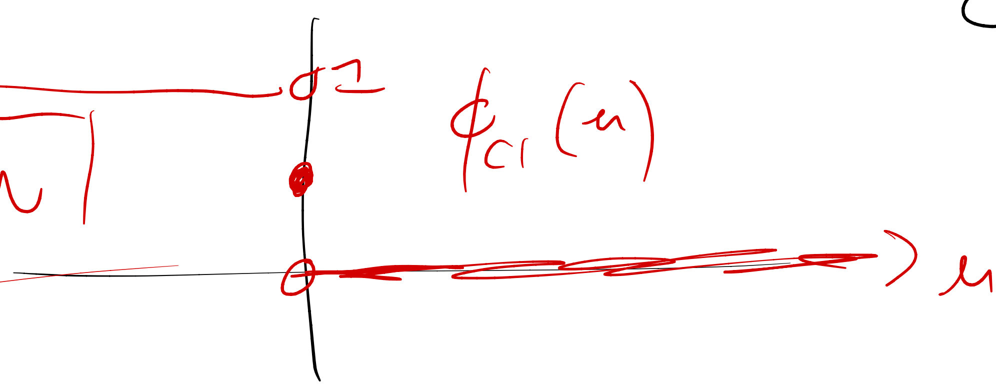
values in $\mathbb{R}$ | values in $\{-1,1\}$ | 0-1 loss

$$= \mathbb{P}\left(\mathrm{sign}(g(x)) \neq y\right)$$

$$= \mathbb{E}\left[\mathbb{1}_{g(x)\neq 0}\,\mathbb{1}_{g(x)\underbrace{y}_{\in\{-1,1\}}<0} + \mathbb{1}_{g(x)=0}\ \text{"random prediction"}\right]$$

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

$$= \Phi_{0-1}(y\,g(x)) \quad \text{with} \quad \Phi_{0-1}(u) = \begin{cases} 1 & \text{if } u<0 \\ 1/2 & \text{if } u=0 \\ 0 & \text{if } u>0 \end{cases}$$



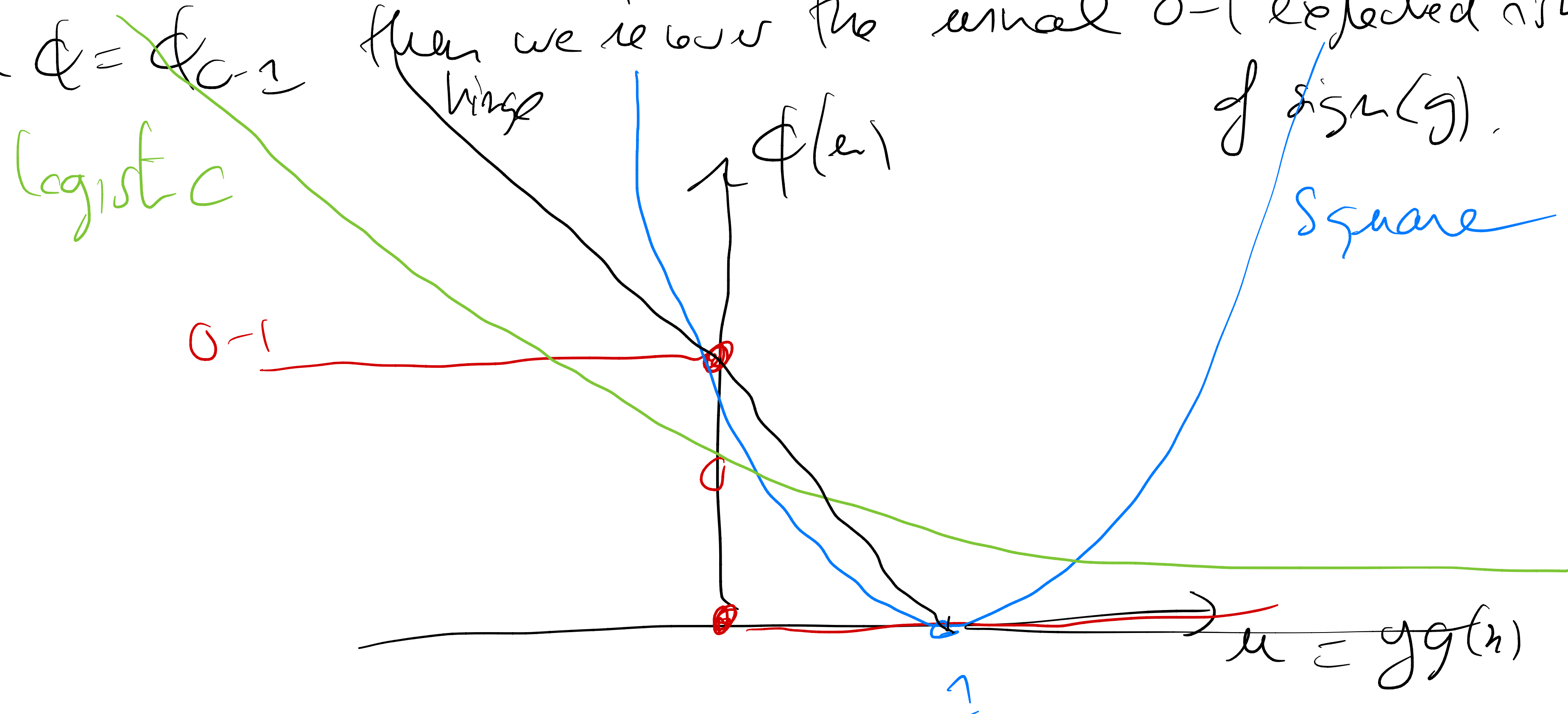
EVALUATION

⚠ avalanche

Convex surrogates : ϕ-risk.

$$R_\phi(g) = \mathbb{E}[\phi(yg(x))]$$

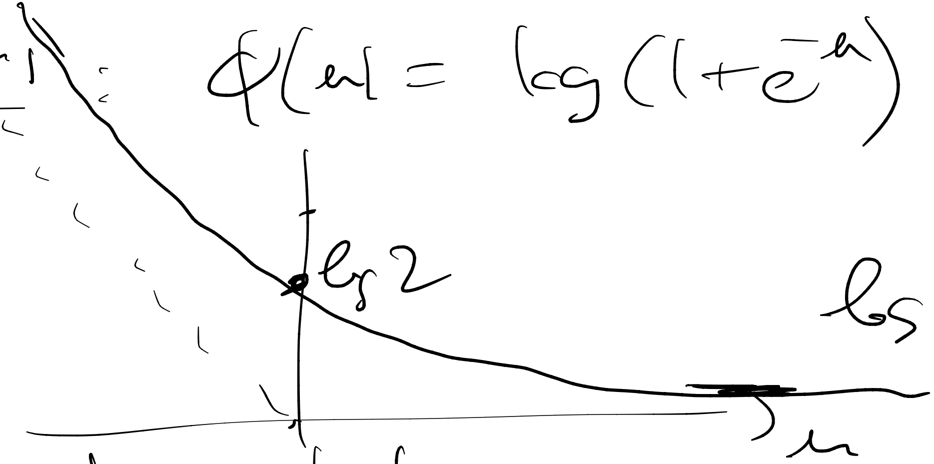
with $\phi = \phi_{0\text{-}1}$ then we recover the usual 0-1 expected risk of $\operatorname{sign}(g)$.



① quadratic loss . $\phi(u) = (u-1)^2$

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

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

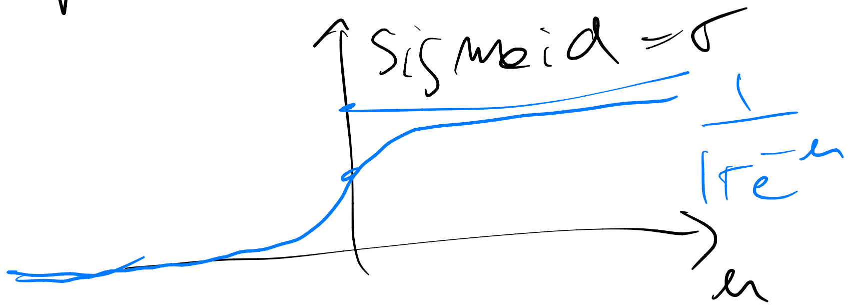


least-squares
$\frac{|y-g(x)|^2}{2\sigma^2} \sim -\log p(y|x)$
y gaussian

probabilistic interpretation:

define a model $p(y=1|x) = \text{sigmoid}(g(x))$

$$= \frac{1}{1+e^{-g(x)}}$$



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

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

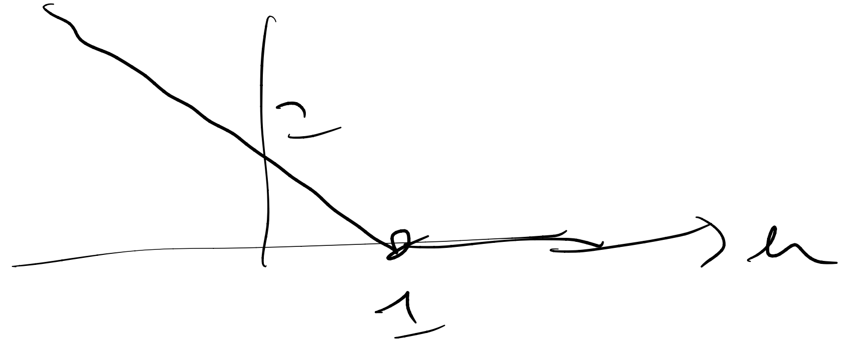
$$p(y|x) = \frac{1}{1+e^{-yg(x)}}$$

$$-\log p(y|x) = \log(1+e^{-yg(x)}) = \text{logistic loss}$$

MAXIMUM LIKELIHOOD / CROSS-ENTROPY

③ hinge loss / Support vector machine (SVM)

$$\varphi(u) = \max\{0, 1-u\} = (1-u)_+$$



convex
non
differentiable

---

if g^* is the minimizer of $R_\phi(g) = \mathbb{E}[\phi(y\, g(x))]$

if $\mathrm{sign}\, g^* = f^*$?

Ex: least squares $=$ $R_\phi(g) = \mathbb{E}\left((y - g(x))^2\right)$

$$g^*(x) \overset{?}{=} \mathbb{E}[y|x]$$

$f^*(x) = 1$ if $p(y=1|x) > \frac{1}{2}$

$= -1$ if $p(y=1|x) < \frac{1}{2}$

$$\mathrm{sign}(g^*(x)) = \mathrm{sign}(\mathbb{E}(y|x))$$

$$\mathbb{E}(y|x) = 1 \cdot p(y=1|x) - 1 \cdot (1 - p(y=1|x)) = 2p(y=1|x) - 1$$

$$= f^*$$

All bounds will be of the form: $R_{\phi}(S) - R_{\phi}^{*} \le \frac{\text{[illegible]}}{\sqrt{n}}$

$$R_{\phi_{0-1}}(S) - R^{*}_{\phi_{0-1}} \le H\left[R_{\phi}(S) - R_{\phi}^{*}\right]$$

calibration function

$H(u) = \sqrt{u}$ for quadratic, logistic

$H(u) = u$ for hinge

$$R(f) = \mathbb{E}\left[\ell(y, f(x))\right] = \mathbb{E}\,(y - f(x))^2 \quad \text{regression}$$

$$= \mathbb{E}\,\phi(y f(x)) \quad \text{classification}$$

$$f : X \to \mathbb{R}$$

Training data: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ iid from distribution $p(x,y)$, $i = 1, \dots, n$

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

Expected risk $R(f) = \mathbb{E}\left(\ell(y, f(x))\right)$.

Empirical risk minimization: $\hat{f} \approx \arg\min \hat{R}(f)$

$f \in \mathcal{F}$ function class

Decomposition

$$R(\hat{f}) - R^* = \underbrace{R(\hat{f}) - \inf_{g \in \mathcal{F}} R(g)}_{\text{Estimation error}} + \underbrace{\inf_{g \in \mathcal{F}} R(g) - R^*}_{\substack{\text{approximation error} \\ \text{decreasing in} \\ \text{model "size"}}}$$

Approximation error : $\mathcal{F} = \{ f_\theta , \theta \in \Theta \}$

$\Theta \in \mathbb{R}^d$

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

neural network

$$\inf_{\theta \in \Theta} R(f_\theta) - R^* = \underbrace{\inf_{\theta \in \Theta} R(f_\theta) - \inf_{\theta' \in \mathbb{R}^d} R(f_{\theta'})} + \underbrace{\inf_{\theta' \in \mathbb{R}^d} R(f_{\theta'}) - R^*}_{\text{"incompressible" approx. err}}$$

Assumption: $\inf_{\theta' \in \mathbb{R}^d} R(f_{\theta'}) = R(f_{\theta_*})$ $\boxed{\neq f^*}$

$$R(f_\theta) - R(f_{\theta_*}) = \mathbb{E}\left(\ell(y, f_\theta(x)) - \ell(y, f_{\theta_*}(x)) \right)$$
$$\leq \mathbb{E}\left(G \,| f_\theta(x) - f_{\theta_*}(x) | \right)$$

Assumptions: loss is G-Lipschitz continuous wrt the second variable

$$\inf_{\theta\in\Theta} R(f_\theta) - R^* = \underbrace{\inf_{\theta\in\Theta} R(f_\theta) - \inf_{\theta'\in\mathbb{R}^d} R(f_{\theta'})}_{} + \underbrace{\inf_{\theta'\in\mathbb{R}^d} R(f_{\theta'}) - R^*}_{}$$

"incompressible"

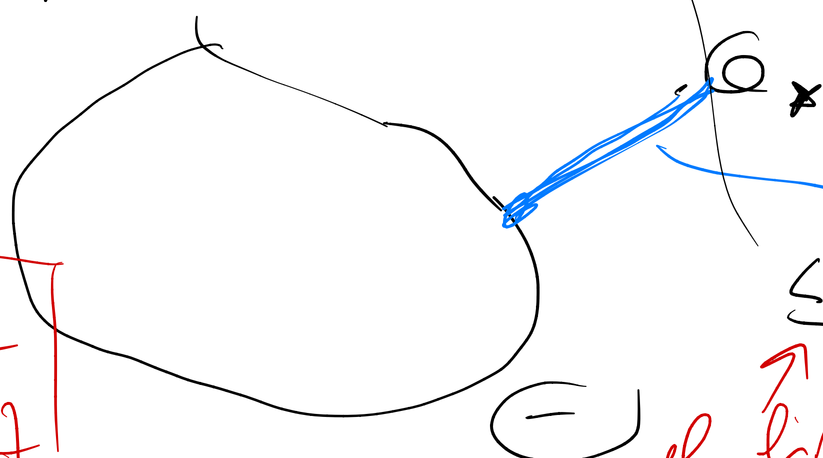
approx. err

Assumption: $\inf_{\theta'\in\mathbb{R}^d} R(f_{\theta'}) = R(f_{\theta_*})$ $\boxed{\exists\, \theta^*}$

$$\inf_{\theta\in\Theta} R(f_\theta) - R(f_{\theta_*}) = \mathbb{E}\left[\ell(y, f_\theta(x)) - \ell(y, f_{\theta_*}(x))\right]$$

$$\le \inf_{\theta\in\Theta} \mathbb{E}\left(G\,|f_\theta(x) - f_{\theta_*}(x)|\right)$$

Assumption: loss is G-Lipschitz continuous wrt to second variable



distance between f_{θ_*} and f_θ, $\theta\in\Theta$

$$\le \mathbb{E}\left(G\,\varphi(x)^\top(\theta - \theta_*)\right)$$

if linear model

$$\le G\,\mathbb{E}\|\varphi(x)\|\,\|\theta-\theta_*\|_2$$

Estimation error:

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - R(f_{\mathcal{F}}) \quad \text{if optimal risk attained}$$

$$= R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - \hat{R}(f_{\mathcal{F}}) + \underbrace{\hat{R}(f_{\mathcal{F}}) - R(f_{\mathcal{F}})}_{\substack{\text{LLN} \to 0 \\ O\left(\frac{1}{\sqrt{n}}\right)}}$$

$$\hat{R}(\hat{f}) = \frac{1}{n} \sum_{i} \ell(y_i, \hat{f}(x_i))$$

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

$$\leq \text{optimization error} + \underbrace{\sup_{f} R(f) - \hat{R}(f) + \sup_{f} \hat{R}(f) - R(f)}_{\text{uniform deviations}} \quad D$$

GOAL: $\mathbb{E} D \leq \ldots$, or $D \leq \ldots(\delta)$ with proba. $1-\delta$

Concentration inequalities : Hoeffding & McDiarmid

Hoeffding's inequality : $Z_i \in [c, c]$ independent

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Z_i)$$

$$= \frac{1}{n}\sum_{i} (Z_i - \mathbb{E}Z_i) \qquad \sim \frac{1}{\sqrt{n}}$$

Central limit theorem

$$\boxed{\mathbb{P}\left(\frac{1}{n}\sum_{i} Z_i - \frac{1}{n}\sum \mathbb{E}Z_i \geq t\right) \leq e^{-2nt^2}} = \delta$$

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

$\Longleftrightarrow$

with probability greater than $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}Z_i < \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Uniform deviation: $\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)$, $\boxed{\mathcal{F} \text{ finite}}$

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

union bound

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

$\ell(y, f(x)) \in [0, \ell_\infty)$
$\forall f \in \mathcal{F}$

hoeffding

$$\le \sum_{f\in\mathcal{F}} e^{-2nt^2/\ell_\infty^2}$$

$$= \underbrace{|\mathcal{F}|\, e^{-2nt^2/\ell_\infty^2}}_{=\delta}$$

This implies that with proba $1-\delta$

$$\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \le \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log\frac{1}{\delta} + \log|\mathcal{F}|}$$

$\Leftrightarrow t = \sqrt{\frac{\ell_\infty^2}{2n} \log\frac{|\mathcal{F}|}{\delta}}$

$\Rightarrow$ extra cost: $\sqrt{\frac{\log|\mathcal{F}|}{n}}$

Rademacher average / complexity:

$\mathcal{F}$ set of functions from $\mathcal{X} \to \mathbb{R}$

loss function Lipschitz continuous (wrt to 2^{nd} arg.)

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

<u>Goal</u>: bound

$$\sup_{f \in \mathcal{F}} R(\ell) - \hat{R}(\ell) = \sup_{f \in \mathcal{F}} \mathbb{E}\Big(\ell(y, f(x)) - \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))\Big)$$

$$= \boxed{\sup_{h \in \mathcal{H}} \mathbb{E}(h(z)) - \frac{1}{n} \sum_{i=1}^{n} h(z_i)}$$

$$z = (x, y)$$

<u>Def</u>: Rademacher average

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

data

where $\varepsilon_i \in \{-1, 1\}$ are independent and zero mean, $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$

"Rademacher" random variables

Goal $\mathbb{E}(\text{Uniform deviation}) \le 2G\, R_n(\mathcal{F})$

Symmetrization lemma: $\mathbb{E}\Big(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} h(z_i) - \mathbb{E}h(z)\Big) \le 2R_n(\mathcal{H})$

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

$$= \mathbb{E}\sup_{h\in\mathcal{H}} \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(h(z_i) - h(z_i')\big)\,\Big|\,D\Big)$$

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

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

Create an independent copy of data $D' = (z_1', \dots, z_n')$ with the same distribution as D

$\mathbb{E}(z) = \mathbb{E}(z_i) = \mathbb{E}(z_i' \mid D)$.

$\mathbb{E}(h(z)) = \mathbb{E}(h(z_i)) = \mathbb{E}(h(z_i') \mid D)$.

$$= \mathbb{E}_{D,D',\varepsilon}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_i \varepsilon_i\big(h(z_i) - h(z_i')\big)\Big]$$

by symmetry

$$\le \mathbb{E}_{D,D',\varepsilon}\sup\Big(\frac{1}{n}\sum_i \varepsilon_i h(z_i)\Big) + \mathbb{E}_{D,D',\varepsilon}\sup\Big(\frac{1}{n}\sum_i \varepsilon_i h(z_i')\Big) = 2R_n(\mathcal{H})$$

$$R_n(\mathcal{H}) = \mathbb{E}_{D,\varepsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(z_i) = \mathbb{E}_{D,\varepsilon} \sup_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \ell(y_i, f(x_i))$$

data

$$\le G \cdot \mathbb{E}_{D,\varepsilon} \sup_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)$$

G is Lipschitz constant of the loss)

$$= G\, R_n(F)$$

(Ledoux + Talagrand)

CONTRACTION PRINCIPLE

$\Rightarrow \mathbb{E}$(uniform deviation for F) $\le 2G\, R_n(F)$

$$\mathbb{E}_{\mathcal{D},\varepsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)$$ for linear predictor

$$\mathcal{F} = \{ x \mapsto \theta^T \varphi(x),\ \|\theta\|_2 \le D \}$$

$\|\varphi(x)\|_2 \le R$ almost surely

$\varphi(x) \in \mathbb{R}^d$

$\mathcal{R}$, R, R

risk, rademacher, bound on the data

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

$$\le D \sqrt{\mathbb{E} \left\| \frac{1}{n}\sum_i \varepsilon_i \varphi(x_i) \right\|_2^2} = D \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\, \cancel{\varepsilon_i^2}\, \underbrace{\|\varphi(x_i)\|^2}_{\le R^2}}$$

$$\boxed{\le DR/\sqrt{n}}$$

Jensen's inequality $\mathbb{E}\|\cdot\|_2 \le \sqrt{\mathbb{E}\|\cdot\|_2^2}$

Put everything together

$\hat{\theta}$ = minimizer of $\hat{R}(f_\theta)$ on $\|\theta\|_2 \leq D$

(linear model

$$\mathbb{E}\, R(f_{\hat{\theta}}) - \inf_{\|\theta\|_2 \leq D} R(f_\theta) \leq \frac{4GRD}{\sqrt{n}}$$

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