

Empirical risk minimization

Data $(x_i, y_i) \in X \times Y$ iid from the test distribution

Empirical risk : $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$
 $\hookrightarrow$ loss function $f: X \rightarrow Y$

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

Class 2 : $Y = \mathbb{R}$ $\ell =$ square loss $\hat{f} = \arg \min \hat{R}(f)$
linear model with d -dimensional features
(fixed design) : excess-risk $\sigma^2 d/n$
 $R(\hat{f}) - R_* \sim \sigma^2 d/n$
 $\hookrightarrow$ Bayes risk

Regularization : $\arg \min R(f) + \lambda \Omega(f)$
 $\hookrightarrow \|\beta\|^2$
 $\Rightarrow \frac{\square}{\sqrt{n}}$

Binary classification: $y = \{-1, 1\}$

$$f: X \rightarrow y = \{-1, 1\} \quad \text{goal: } \min_{f \in F} \bar{R}(f)$$

$$\sum y_j (w_j^T x_j + b_j) \in \mathbb{R}$$

"trick":

$$g: X \rightarrow \mathbb{R} \quad \text{and} \quad \underline{f = \text{sign}(g)}$$

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

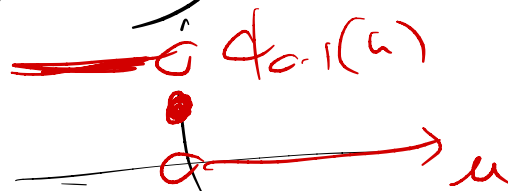
uniformly
random

compute risk of f - 0-1 loss

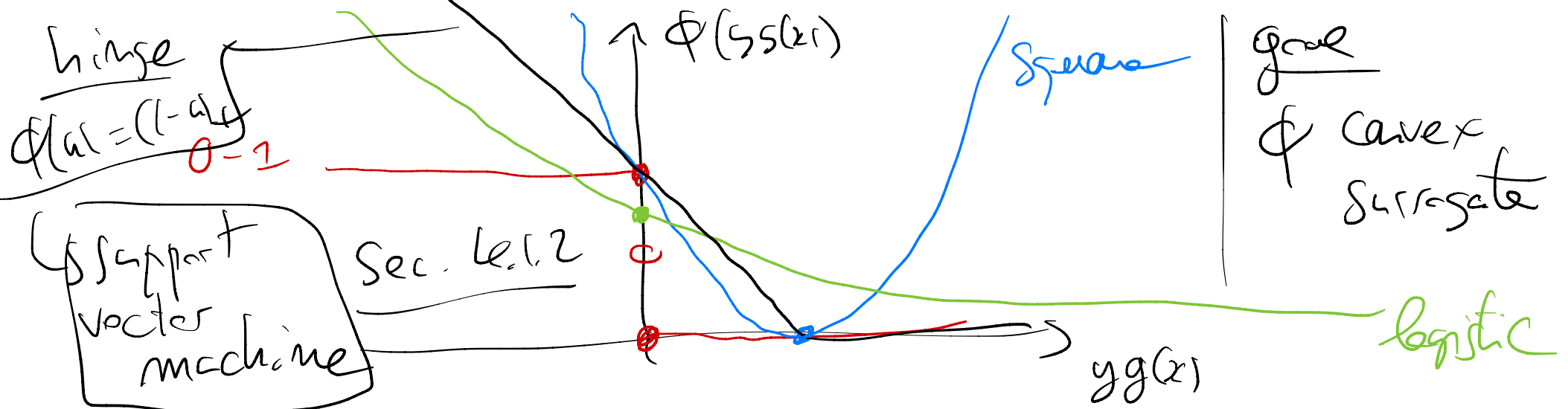
$$R(f) = \mathbb{E}[L(y, f(x))] = \mathbb{P}(y \neq f(x)) = \mathbb{E}[\mathbb{1}_{g(x) \neq 0} \mathbb{1}_{y \neq f(x)}] + \mathbb{E}[\mathbb{1}_{g(x)=0} \mathbb{1}_{y \neq f(x)}]$$

$$= \mathbb{E}[\Phi_{01}(yg(x))]$$

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



$$\mathbb{E}[\mathbb{1}_{g(x)=0}] \leq \frac{1}{2}$$



① square loss

$$\phi(u) = (u-1)_+^2, \quad \mathbb{E}[\phi(yg(x))] = \mathbb{E}[(yg(x)-1)^2] = \mathbb{E}[(g(x)-y)^2]$$

$$f(x) = \text{sign}(g(x))$$

② logistic loss $\Rightarrow$ logistic regression

$$\phi(u) = \log(1 + e^{-u})$$

maximum likelihood

$$\mathbb{E} \phi(yg(x)) = \mathbb{E}[\log(1 + e^{-yg(x)})] = \mathbb{E}[-\log \sigma(yg(x))]$$

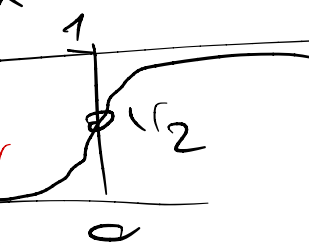
with $\sigma(u) = \frac{1}{1 + e^{-u}}$ sigmoid

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

probabilistic interpretation:

model $P(y|x) = \sigma(yg(x))$

CROSS-ENTROPY LOSS



Classification - calibration ϕ -risk

Question: if $\mathbb{E}[\Phi(yg(x))] - \inf_{\tilde{g}} \mathbb{E}[\Phi(y\tilde{g}(x))]$ small
 is $\mathbb{E}[\Phi_{0-1}(yg(x))] - \inf_{\tilde{g}} \mathbb{E}[\Phi_{0-1}(y\tilde{g}(x))]$ small
 ϕ -risk ϕ -risk



Answer: yes!

① if excess ϕ -risk $= 0$
 then excess 0-1 risk $= 0$

ex. best squares: $\mathbb{E}[\Phi(yg(x))] = \mathbb{E}[(y - g(x))^2]$

$$\mathbb{E}(y|x) = 1 \cdot \mathbb{P}(y=1|x) + (-1) \cdot \mathbb{P}(y=-1|x) \quad \left| \begin{array}{l} g^*(x) = \mathbb{E}(y|x) \\ f^*(x) = \text{sign}(\mathbb{E}(y|x)) \end{array} \right.$$

Bayes classifier: 1 if $\mathbb{P}(y=1|x) > \mathbb{P}(y=-1|x)$
 -1 otherwise

From now on : only consider $f: \mathcal{X} \rightarrow \mathbb{R}$
and losses: $\ell(y, f(x_1))$

$$y = \mathbb{R} \quad \bigwedge \quad y = [-1, 1]$$

square loss

$$\ell(y, f(x_1)) = (y - f(x_1))^2$$

$$\begin{aligned} \ell(y, f(x_1)) \\ = \Phi(y - f(x_1)) \end{aligned}$$

Convex losses

Risk de composition

Data $(x_i, y_i) \mathcal{X} \times \mathcal{Y}$

loss $\ell: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

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

$f: \mathcal{X} \rightarrow \mathbb{R}$

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

ERM: $\hat{f} = \text{empirical risk minimizer}$
 $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{R}(f)$

exact $\rightarrow$
minimizer

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

$$\underbrace{R(\hat{f}) - R_{\star}}_{\substack{\uparrow \\ \text{Bayes risk}}} = \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R_{\star}}_{\substack{\text{approximation} \\ \text{error}}}$$

$\left(\frac{1}{\sqrt{n}} \right)$

Approximation error: $\inf_{f \in \mathcal{F}} R(f) - \inf_{f \text{ measurable}} R(f)$

Model: $\mathcal{F} = \{b_\theta(x), \theta \in \mathbb{C}\}$

$$\inf_{\theta \in \mathbb{C}} R(b_\theta) - \inf_{\theta \in \mathbb{R}^d} R(b_\theta)$$

$$\inf_{\theta \in \mathbb{R}^d} R(b_\theta) - \inf_{f \text{ measurable}} R(f)$$

θ^* regularization / max. error
 θ^* minimizer
 θ^* multipliers

"model error"
model dependent

$$\forall \theta \quad R(b_\theta) - R(b_{\theta^*}) = \mathbb{E}[l(y, b_\theta(x)) - l(y, b_{\theta^*}(x))] \leq \mathbb{E}(|b_\theta(x) - b_{\theta^*}(x)|)$$

Assumption: loss l is Lipschitz-continuous wrt to second variable
 $|l(y, u) - l(y, v)| \leq G|u - v|$

$\inf_{\theta \in \mathbb{C}}$

$$\inf_{\theta \in \mathbb{C}} R(b_\theta) - R(b_{\theta^*}) \leq G \inf_{\theta \in \mathbb{C}} |b_\theta(x) - b_{\theta^*}(x)|$$

$$\mathbb{E}(|b_\theta(x) - b_{\theta^*}(x)|)$$

Estimation error :

$$R(\hat{G}) - \inf_{G \in \mathcal{F}} R(G) = R(\hat{G}) - R(G_F^*) + \underbrace{R(\hat{G}) - \hat{R}(G_F^*)}_{\leq 0} + \underbrace{\hat{R}(G_F^*) - R(G_F^*)}_{\downarrow}$$

ERM
argmin $\hat{R}(G)$
 $G \in \mathcal{F}$

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

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

$\hat{G}$ depends
on the data

$$\leq \sum_{i=1}^n \ell(y_i, G_F^*(x_i)) - \mathbb{E}[\dots]$$

$$\frac{1}{n} \sum_{i=1}^n \ell(y_i, G_F^*(x_i)) - \mathbb{E}[\dots] \sim O\left(\frac{1}{\sqrt{n}}\right)$$

$$\leq \underbrace{\sup_{G \in \mathcal{F}} \{R(G) - \hat{R}(G)\}}_{\text{uniform deviation}} + o(1) + \sup_{G \in \mathcal{F}} \{R(G) - R(G)\}$$

"large"
band

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

in expectation
 $\mathbb{E}(Z) \leq \dots$
in high prob.
with probability
 $1 - \delta, Z \leq \dots + \dots$

Hoeffding's inequality: $Z_i \in [0, 1]$ iid.

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

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

with probability $1 - \delta$, $\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z) \leq \sqrt{\frac{\ln(1/\delta)}{2n}}$

$\left(\frac{1}{\sqrt{n}}\right)$

Chebyshev: $\text{var}(Z_i) = \sigma^2$

$$\mathbb{E}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z)\right|^2\right) = \mathbb{E}\left(\left|\frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}Z)\right|^2\right)$$
$$= \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{E}\left((Z_i - \mathbb{E}Z)^2\right)}{n^2} = \frac{\sigma^2}{n}$$

σ^2

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

Markov inequality

Uniformly distributed : band $\sup_{\theta} R(\theta) - \hat{R}(\theta)$

① finite set $F = \{b_1, \dots, b_n\}$

$$\begin{aligned} \text{finite set } \mathcal{F} &= \{b_1, \dots, b_m\} \\ \mathbb{P}\left(\sup_{b \in \mathcal{F}} R(b) - \hat{R}(b) > t\right) &= \mathbb{P}\left(\bigcup_{b \in \mathcal{F}} \{R(b) - \hat{R}(b) > t\}\right) \\ &\leq \sum_{b \in \mathcal{F}} \mathbb{P}(R(b) - \hat{R}(b) > t) \end{aligned}$$

$\rightarrow \leq \sum_{f \in F} \mathbb{P}(R(f) - \bar{R}(f) > t)$
 $-2nt^2$

$$\rightarrow \sum_{f \in F} e^{-2nt^2} = |F| e^{-2nt^2} = f$$

bsi $\in [c]$
+ the effding

$$\Rightarrow \frac{1}{2n} \log \frac{|S|}{f} = t$$

$\Rightarrow$ with probability $1 - \delta$

proceeding 1-8

$$\sup_{b \in F} R(b) - \hat{R}(b) \leq \sqrt{\frac{1}{2n}} \left(\sqrt{\log |F|} + \sqrt{\log \frac{1}{\delta}} \right).$$

Rademacher complexity: $\sup_{f \in \mathcal{F}} \mathbb{E} |f - \bar{\mathbb{E}}(f)|$

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

with $f \in \mathcal{F} = \left\{ (z, y) \mapsto \mathbb{E}[y, f(x)] \right\}$

$z \mapsto h(z)$

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

$\mathcal{D}$: data $z_1, \dots, z_n$

$\varepsilon \in \{-1, 1\}^n$ vector of independent Rademacher random variables

$\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$

Symmetrization lemma

$$\mathbb{E}_{\mathcal{D}} \left(\sup_{h \in \mathcal{H}} \mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \leq 2 R_n(\mathcal{H})$$

$\neq \sup_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}(\dots)$

NB: for results in high probability
use McDiarmid inequality

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} \left(\mathbb{E}[h(z_1)] - \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \right] = \mathbb{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}[h(z'_i) | \mathcal{D}]}_{\mathbb{E}[h(z_1)]} - \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \right]$$

$\mathcal{D}'$ are independent copy of $\mathcal{D}$

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

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

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

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

ε_i Rademacher

$$\mathbb{E}(A) = \mathbb{E}(\mathbb{E}(A | \mathcal{D})) \quad \rightarrow$$

$$\sup(A+B) \leq \sup A + \sup B \quad \rightarrow \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(z'_i) \right] + \mathbb{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(z_i) \right]$$

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

Contraction principle: $f(x) = \int (|y|+1) \ell(y, f(x))$, $f \in \mathcal{F}$
 if ℓ is G -Lipschitz w.r.t to the second variable

$$R_n(f) \leq G R_n(\mathcal{F})$$

$$\Rightarrow \mathbb{E} \left(\sup_{f \in \mathcal{F}} |R(f)| \right) \leq 2G R_n(\mathcal{F})$$

Linear model: $f_\theta(x) = \theta^T \phi(x)$ Assume $\|\phi(x)\| \leq R$ a.s. $\mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right)$

$$R(\{f_\theta, \|\theta\|_2 \leq D\}) = \mathbb{E} \sup_{\|\theta\|_2 \leq D} \left| \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \phi(x_i) \right)^T \theta \right|$$

Cauchy Schwarz

$$= \mathbb{E} \left[D \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \phi(x_i) \right\|_2 \right]$$

Jensen $\rightarrow$

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

$\leq RD/\sqrt{n}$ $\leq R^2$

Summary for linear models

$$\hat{\theta} = \inf_{\|\theta\|_2 \leq D} \hat{R}(\theta)$$

$$\text{for } \theta_0(x) = \varphi(x)^T \theta$$

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

estimation error

G = Lipschitz constant of the θ_0

R = radius of data $\|\varphi(x)\|_2 \leq R$

D = bound on predictors

Next week : optimization error
 $t = \# \text{ iter of } g(s)(n)$

$$\frac{GRD}{\sqrt{t}}$$