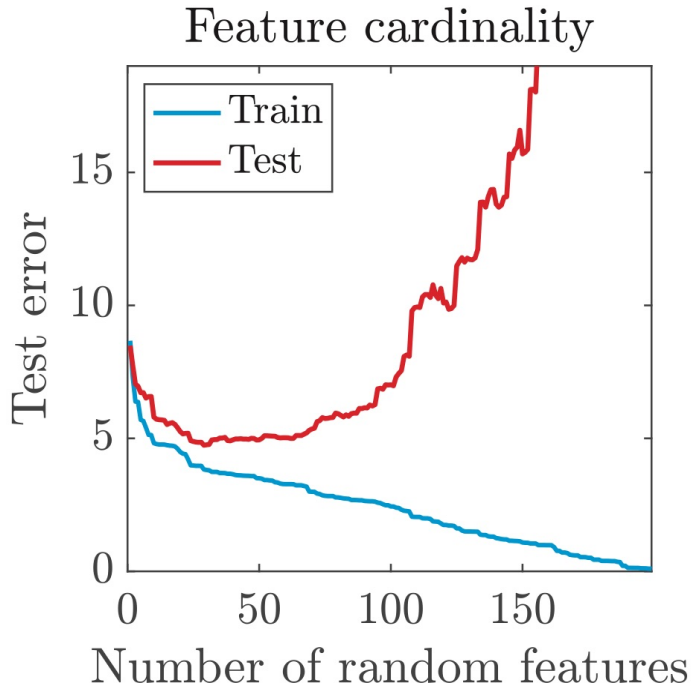
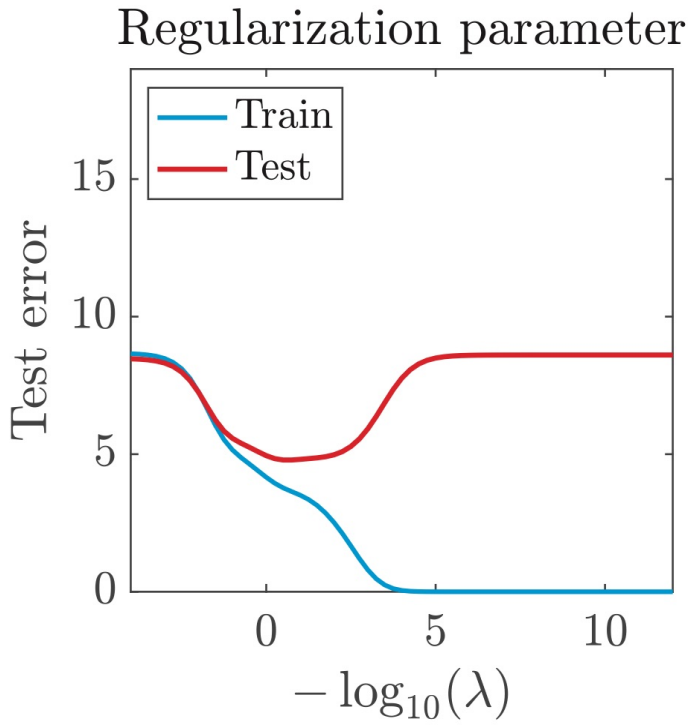
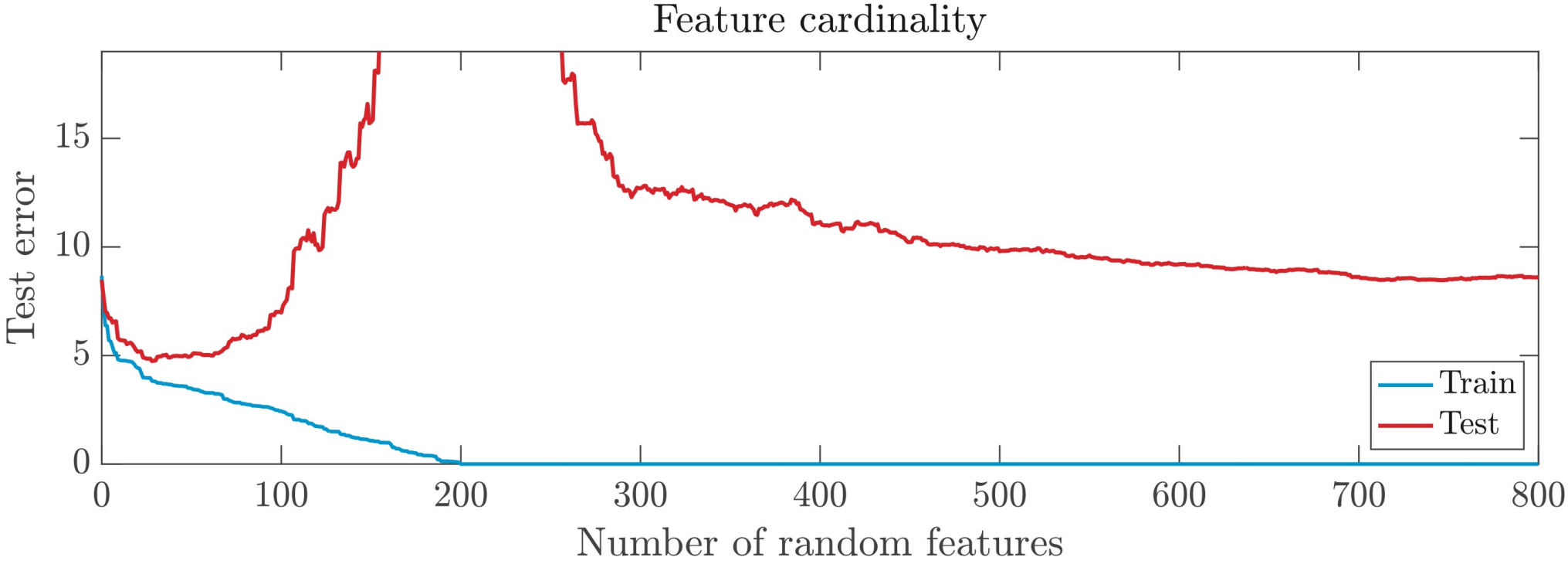
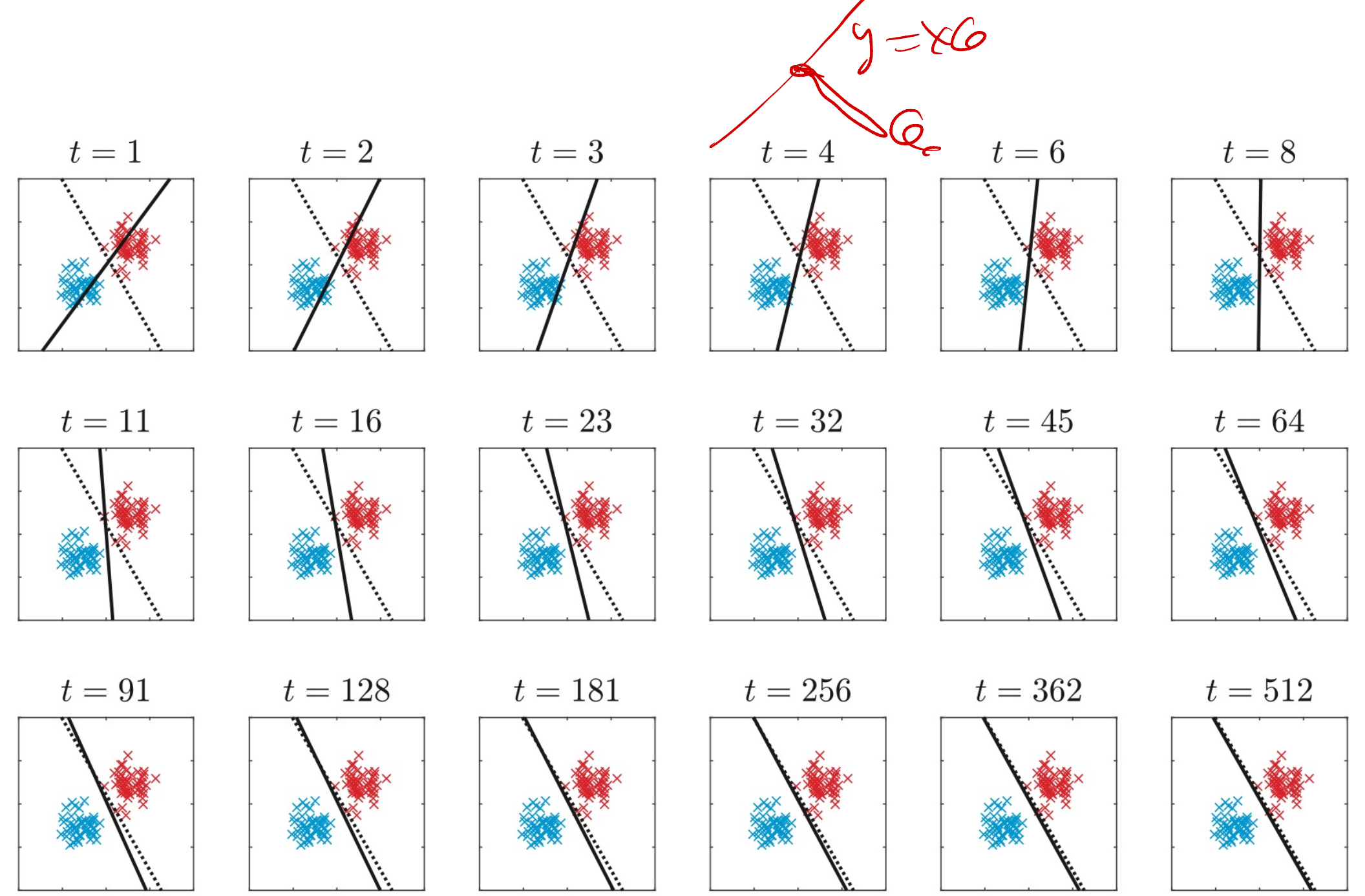

Feature cardinality
Train
Test
Test error
0
5
10
15
0
50
100
150
Number of random features


Regularization parameter
Train
Test
Test error
0
5
10
15
0
5
10
$-\log_{10}(\lambda)$


Feature cardinality
Train
Test
Test error
0
5
10
15
0
100
200
300
400
500
600
700
800
Number of random features


$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 6$
$t = 8$
$t = 11$
$t = 16$
$t = 23$
$t = 32$
$t = 45$
$t = 64$
$t = 91$
$t = 128$
$t = 181$
$t = 256$
$t = 362$
$t = 512$

Exam : Wednesday Dec 18 9-11.30 am Room 6/8/3

Two sides of A4 allowed

Mix of simple and deeper exercises

TODAY : . Summary / Review of each chapter

, Chapter 12 : overparameterized models

"implicit bias of GD"

Chapter 2 : Decision theory for ML

$X \times Y$ — X: input space; Y: output space

$f : X \to Y$ prediction function

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

Expected risk / Test error / generalization error

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

↑ testing data

⚠ if f is random, $R(f)$ is random

$= \mathbb{E}_{\text{training data}} R(\hat{f})$

Bayes predictor / target function: $f^* \in \arg\min R(f)$

Bayes error: $R(f^*) = \min_{f: \mathcal{X} \to \mathcal{Y}} R(f)$ $\overset{?}{=} 0$

$$f^*(x) = \arg\min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) \mid x]$$

two cases: classification: $\mathcal{Y} = \{1, \dots, k\}$ $\ell(y, \bar{y}) = \mathbb{1}_{y \neq \bar{y}}$

$$f^*(x) = \arg\max_z \mathbb{P}(y = z \mid x)$$

regression: $\mathcal{Y} = \mathbb{R}^z$ $\ell(y, \bar{y}) = |y - \bar{y}|^2$

$$f^*(x) = \mathbb{E}[y \mid x]$$

Chapter 3 : least-squares : $\ell(y, f(x)) = (y - f(x))^2$

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

$\varphi : \mathcal{X} \to \mathbb{R}^d$

linear model in θ

$\theta \mapsto R(f_\theta)$ | are convex in Θ

$\theta \mapsto \hat{R}(f_\theta)$ empirical risk |

$$L \triangleq \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^\top \varphi(x_i))^2 = \frac{1}{n} \| y - \Phi \theta \|_2^2$$

$y \in \mathbb{R}^n$: label response | vector

$\Phi \in \mathbb{R}^{n \times d}$: design data matrix

OLS estimator : $\hat{\theta} = \arg\min \hat{R}(f_\theta)$

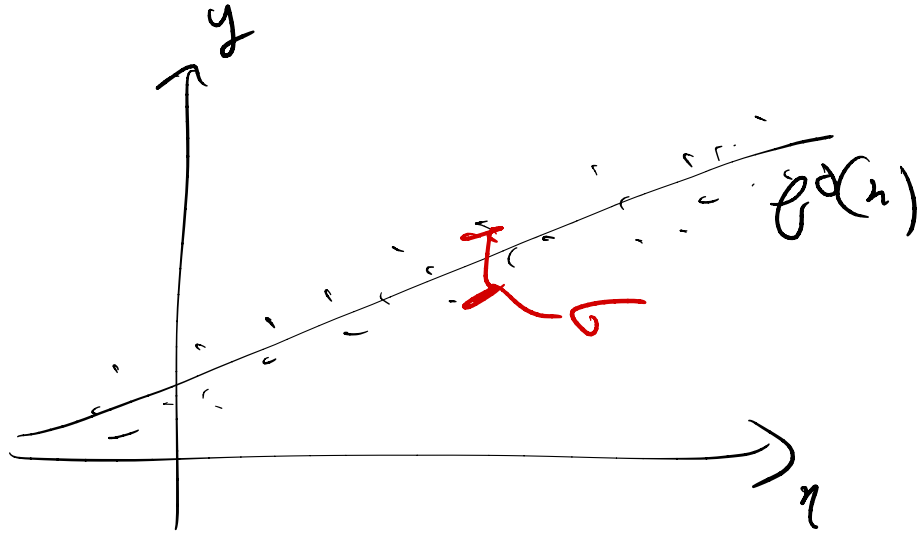
Normal equations / Fixed design

Model : $y_i = \varphi(x_i)^T \theta_* + \varepsilon_i$ with $\mathbb{E}\varepsilon_i = 0$

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

$$\mathbb{E}\left(\text{Excess risk}\right) = \frac{\sigma^2 d}{n}$$

only valid for $d < n$



⚠ Not regularized

Ridge regression : $\min_{\theta} \hat{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2$

$\Rightarrow$ dimension independent rate in $\frac{R}{\sqrt{n}}$

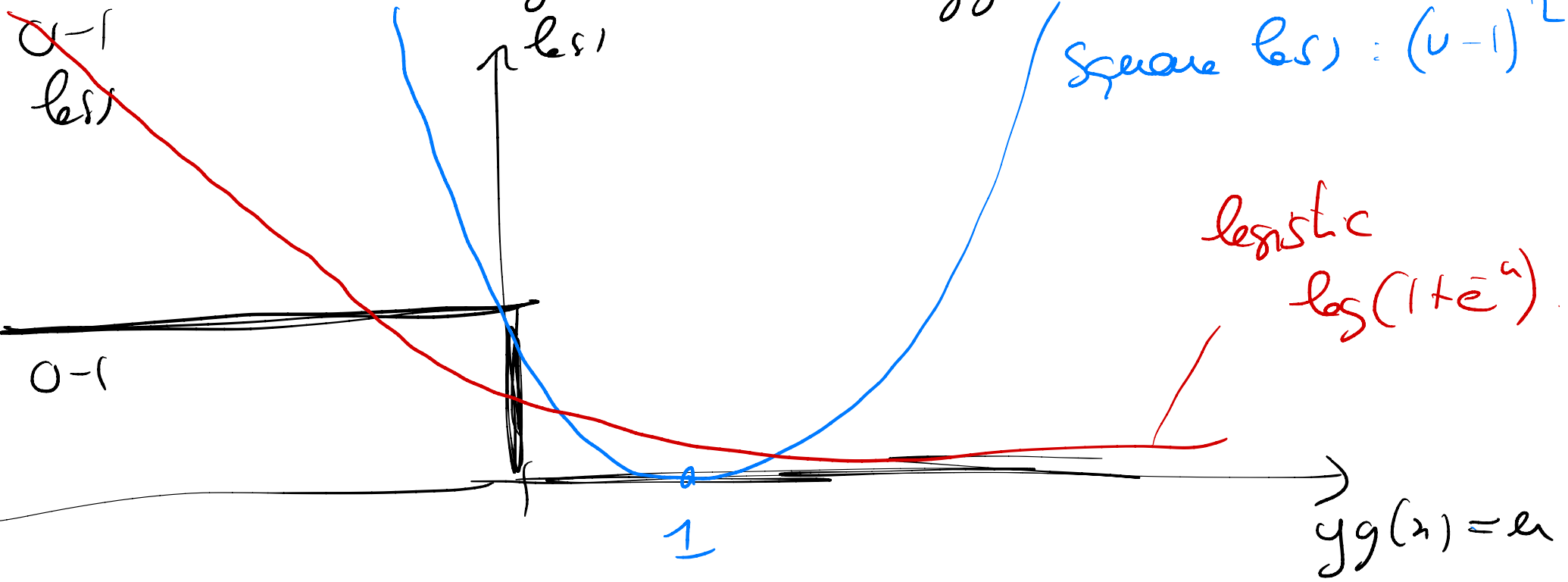
$$\|\varphi(x)\|_2 \leq R$$

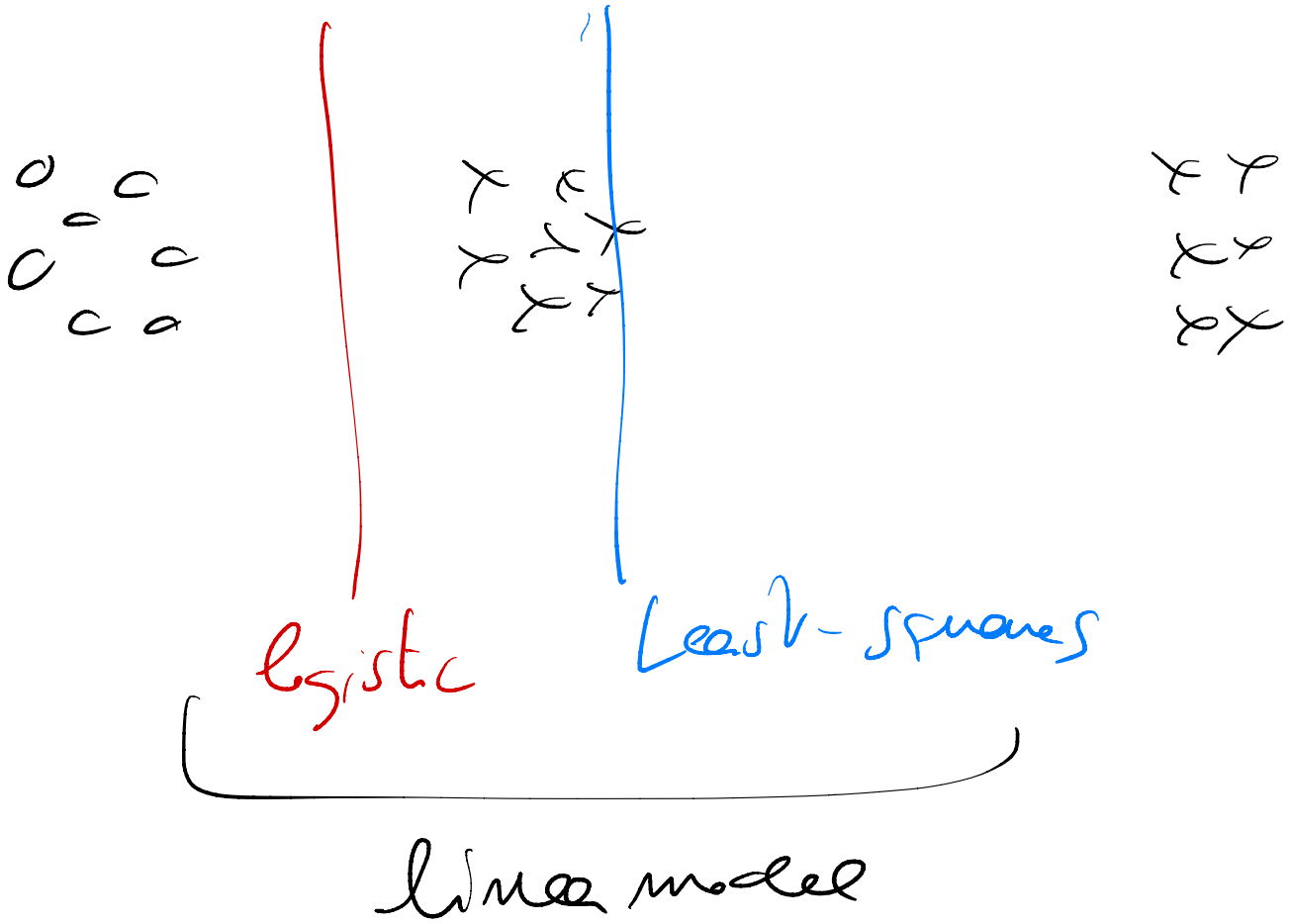
Chapter 4 : Empirical risk minimization

① Convex surrogates for classification $\mathcal{Y} = \{-1, 1\}$.

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

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




logistic
Least-squares
linear model

Decomposition: approximation and estimation error

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{R}(f)$$

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

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

single f: $\sim \frac{1}{\sqrt{n}}$

m functions $\sim \frac{\sqrt{\log m}}{\sqrt{n}}$

Rademacher | complexity | average

$$\underbrace{R}_{R}(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \xi_i f(x_i)$$

data

ξ Rademacher $\in\{-1,1\}$

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\hat{R}(f) - \underbrace{R(f)}_{R(f)}| \le 4G\, R(\mathcal{F})$$

G: Lipschitz constant of the loss

ex: if $\mathcal{F} = \{x \mapsto \varphi(x)^T\theta,\ \|\theta\|_2 \le D\}$ and $\|\varphi(x)\|_2 \le \underbrace{R}_{\text{a.s}}$ — R

$\|\theta\|_2 \le D$ → 1

$\|\varphi(x)\|_2$ → ∞

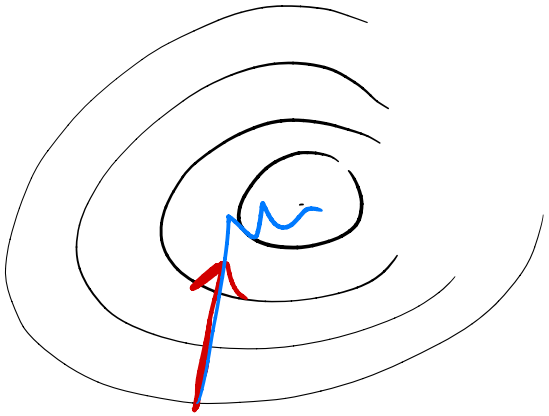
$$R(\mathcal{F}) \le \frac{RD}{\sqrt{n}} \quad \sqrt{\log d}$$

Chapter 5 : Optimization

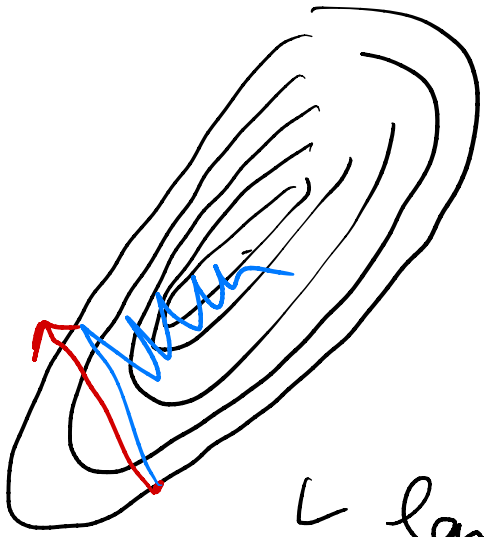
① Gradient descent on smooth functions / convex

all F''s have eigenvalues in $[\mu, L]$

$\mu \geq 0$



$\frac{L}{\mu}$ close to 1



$\frac{L}{\mu}$ large

each has number

$$F(\theta_{t+1}) - F(\theta_*) \leq \left(1 - \frac{\mu}{L}\right)^t \left(F(\theta_1) - F(\theta_*)\right)$$

$$\leq \frac{L\|\theta_0 - \theta_*\|_2^2}{t}$$

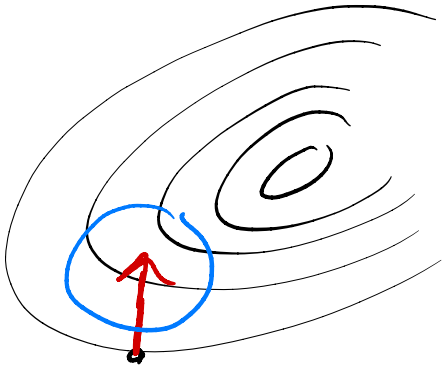
if γ step-size $= \frac{1}{L}$

② Stochastic Gradient Descent

$$\theta_t = \theta_{t-2} - \gamma g_t$$

$$\boxed{\mathbb{E}[g_t | \text{past}] = F'(\theta_{t-1})}$$

$$\|g_t\|_2^2 \le B^2 \quad \text{untriased}$$



$$\mathbb{E}\left(F(\theta_{t+1}) - F(\theta_t)\right) \le \frac{\|\theta_0 - \theta_*\|^2}{\gamma t} + \gamma B^2$$

$$\gamma \propto \frac{1}{\sqrt{t}}$$

$$O\left(\frac{1}{\sqrt{t}}\right)$$

F = training error (multiple passes)

F = testing error (single pass) $t = n$

Bound on test error

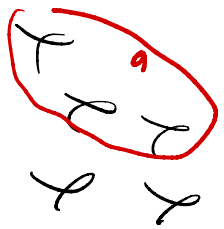
$= \mathbb{E}\,\ell(y, f_{\theta_0}(x))$

Chapter: local averagings : k-nearest neighbors

β^* = Lipschitz-continuous

Excess risk $\leq O\left(\frac{1}{n^{1/d+2}}\right)$

Curse of dimensionality



No optimality

chap 7 : kernel methods

① linear model $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$

$\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space

② Representer theorem

$$\min \hat{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|^2$$

$\theta \in$ Empirical risk

can be found as $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$

$$\boxed{\alpha \in \mathbb{R}^n}$$

$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ kernel function

Pb: need to compute kernel matrix $k(x_i, x_j)$

(3) kernels $k(x,x') = \varphi(x-x')$ $\quad x \in \mathbb{R}^q$

$= \exp\left(-\|x-x'\|_2^2\right)$ Gaussian

$= \exp\left(-\|x-x'\|_2\right)$ Abel exponential

positive definite

$\hookrightarrow \exists \varphi : \mathcal{X} \to \mathcal{H} \quad k(x,x') = \langle \varphi(x), \varphi(x') \rangle$

Function space = space of smooth functions with $s > \frac{d}{2}$ derivatives.

$f(x) = \langle \Theta, \varphi(x) \rangle_{\mathcal{H}}$ $\quad \|\Theta\|_{\mathcal{H}}$ finite

(4) kernels = $\frac{1}{\sqrt{n}}$ of well-specified eq

$f^*(x) = \langle \Theta_*, \varphi(x) \rangle$

Chap 8 : sparse method linear model $f_\theta(x) = \varphi(x)^T \theta$

$\min_\theta \hat{R}(f_\theta)$ — ℓ_0 penalty/constraint

$\|\theta\|_0 = \#$ of non zero components

$\sigma^2 k \log d$ for exact-sparse

n $k = \|\theta_*\|_0$

ℓ_1-norm $= \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$

Convex optimization

$\frac{k \log d}{n} \longrightarrow k \sqrt{\frac{\log d}{n}}$

$$\mathbb{E}\left(\left\| \frac{1}{n} \sum \varepsilon_i \varphi(x_i) \right\|_\infty \right) \sim \sqrt{\log d}$$

Chap 9 : Neural net

$$f_\theta(x) = \sum_{j=1}^{m} a_j \, \sigma(w_j^T x + b_j)$$

$\sigma(u) = u_+$ ReLU

Optimisation : non cvx.

Estimation err $\leq \frac{\|f\|_2}{\sqrt{n}}$ independent of m !!!

Approximation err : (universality
feature learning
$f_\theta(x) = g(w^T x)$

$\ell^4 \mathbb{Z}$

$\ell^4 \mathbb{Z}^2$

Implicit Bias of gradient descent

for "overparameterized" models

Training data (x_i, y_i), $i = 1, \dots, n$ $\quad f_\theta(x)$ model

overparam $\Rightarrow \exists \theta$ s.t f_θ leads to perfect predictions on the training data

$$\hat{R}(f_{\hat{\theta}}) = 0$$

Using uniform deviations: $\quad R(f_{\hat{\theta}}) \leq 0 + \frac{C}{\sqrt{n}}$

$\forall$ all Empirical risk minimizers

Set of minimizers $\quad \hat{R}(f_\theta) = 0$

If f_θ is linear and $\hat{R}$ is cvx $\Rightarrow$ convex set



Goal = least-squares $F(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$

$y \in \mathbb{R}^n$ $X \in \mathbb{R}^{n \times d}$ data matrix

Assumptions = $\exists \theta$ s.t $X\theta = y$

strange assumption: XX^T is invertible $\in \mathbb{R}^{n \times n}$

$(XX^T)_{ij} = x_i^T x_j \quad i, j \in \{1, \dots, n\}$

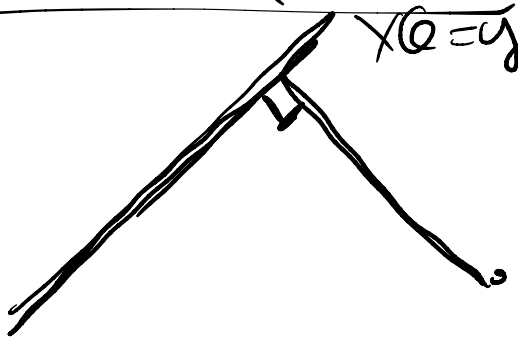
(gram matrix
kernel

$n \leq d$

$\neq$ $X^TX \in \mathbb{R}^{d \times d} \Rightarrow \frac{1}{n} X^TX$ empirical covariance matrix

Normal equations: $F'(\theta) = 0$

$X^T(X\theta - y) = 0$



$\Leftrightarrow XX^T(X\theta - y) = 0 \Leftarrow \boxed{X\theta = y}$

Proof of exponential convergence of GD

$$X\theta_t = X\theta_{t-1} - \gamma X F'(\theta_{t-1}) = X\theta_{t-1} - \gamma \frac{XX^T(X\theta_{t-1} - y)}{n}$$

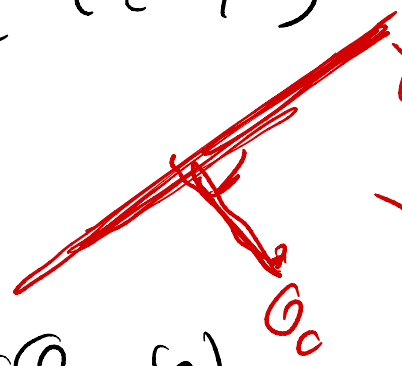
$$X\theta_t - y = X\theta_{t-1} - y - \gamma \frac{XX^T}{n}(X\theta_{t-1} - y)$$

$$= \left(I - \gamma \frac{XX^T}{n}\right)(X\theta_{t-1} - y)$$

$$(XX^T)(XX^T)$$
$$X(X^TX)X^T$$

where $F'(\eta_2) = 0$

NB: $\theta_t - \eta_\delta = \left[I - \gamma \frac{X^TX}{n}\right]\left[\theta_{t-1} - \eta_\delta\right]$



$X\theta_t \to y$

$\theta_t \to ??$

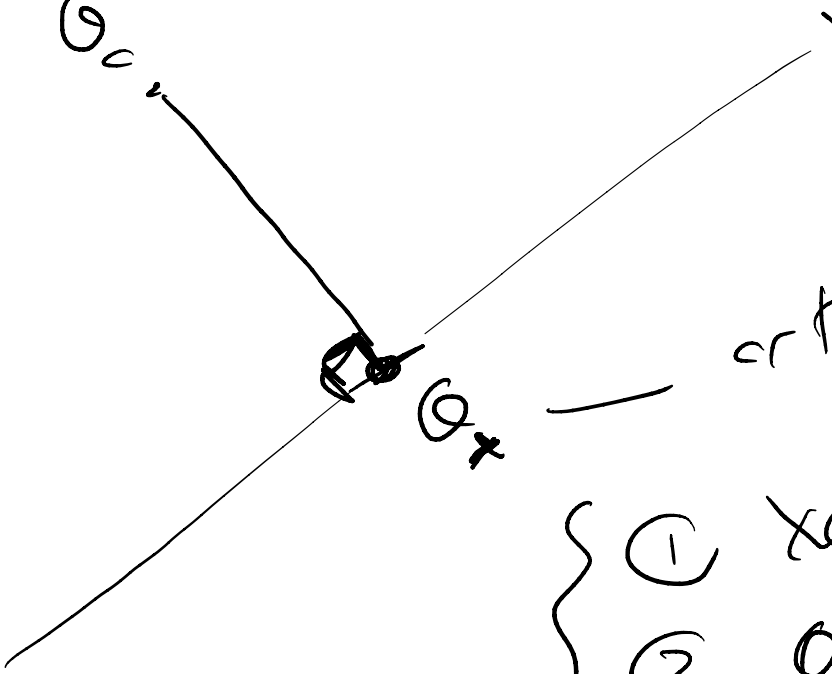
if $\lambda_{min}\left(\frac{XX^T}{n}\right) \geq \lambda > 0$

$$X\theta_t - y = \left(I - \gamma \frac{XX^T}{n}\right)^t (X\theta_0 - y)$$

$$\|X\theta_t - y\|_2 \leq (1-\lambda)^t \|X\theta_0 - y\|_2$$

exponential cycle

Characterizing the orthogonal projection of θ_c on $\{X\theta = y\}$



$$\begin{cases} (1)\ X\theta_* = y \\ (2)\ \theta_c - \theta_* \perp \{X\theta = y\} \end{cases}$$

$$\min_{\theta} \frac{1}{2}\|\theta - \theta_c\|^2 \ s.t.\ y = X\theta = \max_{\alpha \in \mathbb{R}^n} \min_{\theta} \frac{1}{2}\|\theta - \theta_c\|^2 + \alpha^T(y - X\theta)$$

Lagrangian

$$\Rightarrow \min_{\theta} \frac{1}{2}\|\theta - \theta_c\|^2 + \alpha^T(y - X\theta)$$

$$\frac{\partial}{\partial \theta} \Rightarrow \theta - \theta_c - X^T\alpha = 0$$

$$\boxed{\theta - \theta_c = X^T\alpha} \quad (1)$$

Dual problem $= \max_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|X^T\alpha\|_2^2 + \alpha^T(y - X(\theta_c + X^T\alpha))$

$= \alpha^T(y - X\theta_c) - \frac{1}{2}\|X^T\alpha\|_2^2$

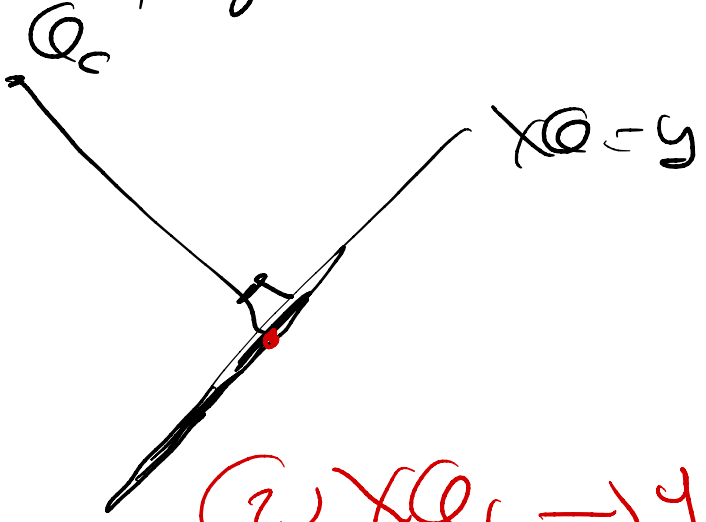
$\partial/\partial \alpha \Rightarrow \boxed{y - X\theta_c - XX^T\alpha = 0}$

$\Rightarrow \alpha = (XX^T)^{-1}(y - X\theta_c)$ (2)

Summary $\Theta = \Pi_{y=X\theta} (\theta_c)$

$$\Leftrightarrow \begin{cases} \theta - \theta_c = X^T \alpha \\ y - X\theta_c - XX^T\alpha = 0 \end{cases} \Leftrightarrow y - X(\underbrace{\theta_c + X^T\alpha}_{\theta = c}) = 0$$

$$\Leftrightarrow \begin{cases} \theta = \theta_c + X^T\alpha = \theta_c + \sum_{i=1}^{n} \alpha_i x_i \in \mathbb{R}^d \\ y = X\theta \end{cases}$$



① For GD

$$\theta_t = \theta_{t-1} - \gamma X^T(X\theta_{t-1} - y)$$
$$\theta_1 = \theta_c - \gamma X^T(X\theta_c - y)$$

$\Rightarrow \theta_t - \theta_c$ is of the form $X^T\alpha_t$

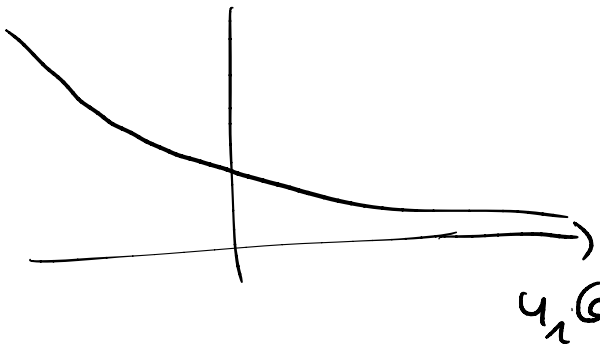
$$X\theta_t - X\theta_c = XX^T\alpha_t$$
$$\alpha_t = (XX^T)^{-1}(X\theta_t - X\theta_c) \to (XX^T)^{-1}(y - X\theta_c)$$

② $X\theta_t \to y$

Logistic regression: $y_i \in \{-1, 1\}$

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \theta^T x_i})$$

minimum value of F?



$\inf F(\theta) = c$

θ_t diverges but $\frac{\theta_t}{\|\theta_t\|_2}$ converges to u is the direction of the

SVM classifier

Support vector machine