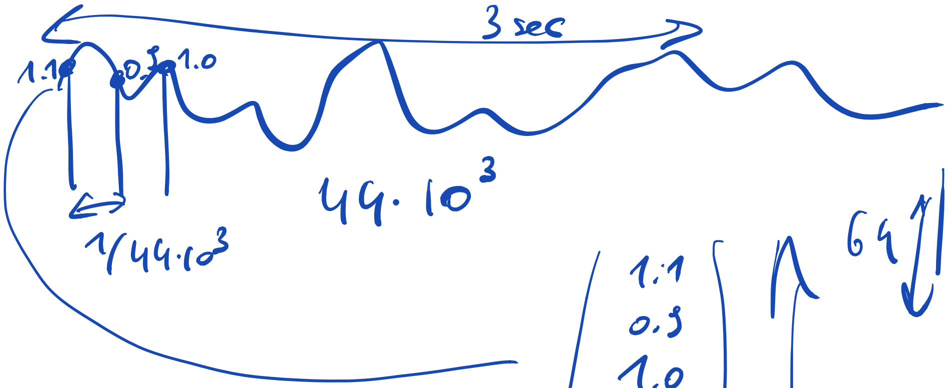
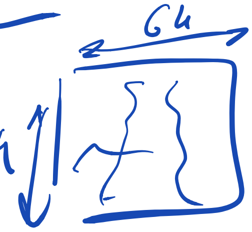




1.1
0.9
1.0

$3 \sec \times 44 \cdot 10^3$

$n \approx 120 \cdot 10^3$

$\underline{x}$

$\hat{y} = \{$ "Chiquitita" "

Song by Abba → Yes
→ No

$h(\underline{x}) = \underline{w}^T \underline{x}$

$h(x) > \varphi \rightarrow \hat{y} =$ "Abba Song"

$< \varphi \rightarrow \hat{y} =$ "No Abba"

true label

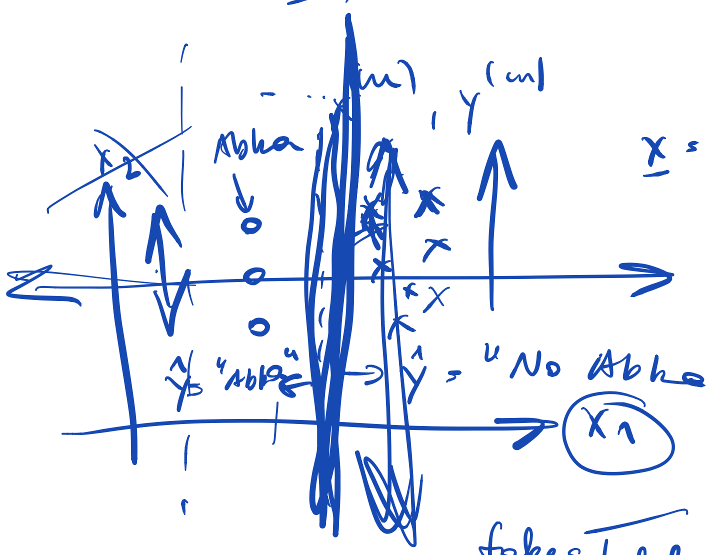
$(\underline{x}^{(1)}, y^{(1)} =$ "Abba"$)$ $(\underline{x}^{(2)}, y^{(2)} =$ "No Abba"$)$

$h(\underline{x}^{(1)}) \rightarrow \hat{y}^{(1)}$ . . . $\hat{y}^{(2)}$

predicted label

. . . $\underline{x}^{(m)}, y^{(m)}$

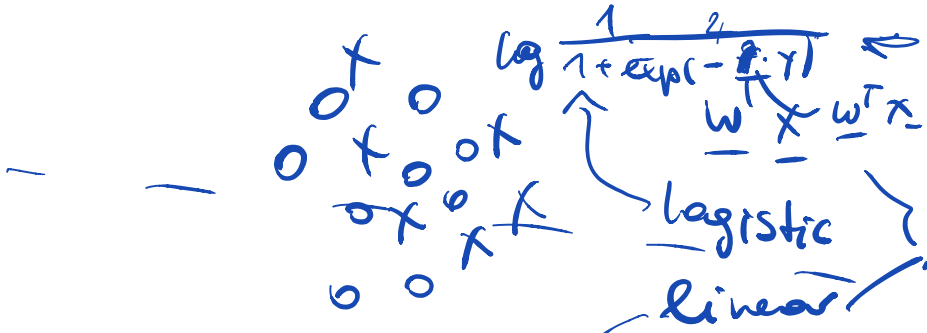
$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



PCA → ignores labels

Fisher discr. analysis

takes labels into account



min Loss

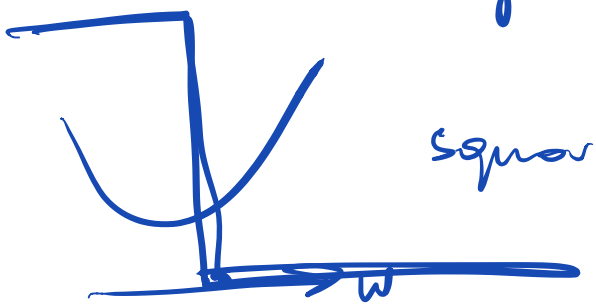
$h(x)$ 0/1 loss

↳ 1 if $\hat{y} \neq y$ $w^T x$

0 else

logistic loss $\rightarrow$ $\log \frac{1}{1+\exp(-w^T x \cdot y)}$ $\{-1, 1\}$

squared loss $(w^T x - y)^2$ $\hat{y}$



Data	Hypothesis Space	Loss-Fct.
$(x^{(1)}, y^{(1)}) \dots (x^{(m)}, y^{(m)})$	$h(x) = w^T x$	$(\hat{y} - y)^2$

Linear Regression

$$\min_{w \in \mathbb{R}} \underbrace{(1/m) \sum (w^T x^{(i)} - y^{(i)})^2}_{\text{training error } E(w)}$$

find number x s.t. $x^2 = 2$
allow adds / mults.

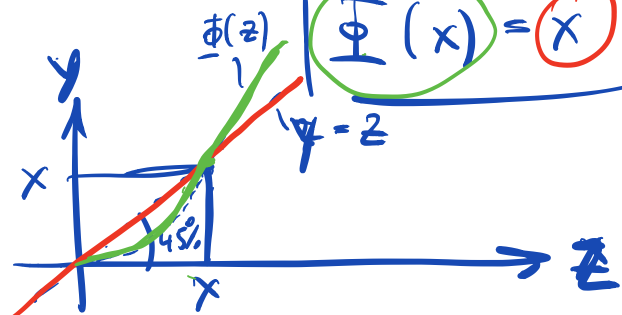
$$x^2 = 2$$

$+x / -2\cdot 10^{-3}$ ~~100~~ $\quad 10^{-3} \cdot x^2 = 2 \cdot 10^{-3}$

$$\underbrace{x + 10^{-3} x^2 - 2\cdot 10^{-3}}_{} = x$$

$$\to \Phi(z) \Rightarrow z + 10^{-3} z^2 - 2\cdot 10^{-3}$$

$$\boxed{\Phi(x) = x}$$ fixed point equation



fixed point iteration

initial guess $\quad x^{(0)} \to x^{(1)} = \Phi(x^{(0)}) \to x^{(2)} = \Phi(x^{(1)})$

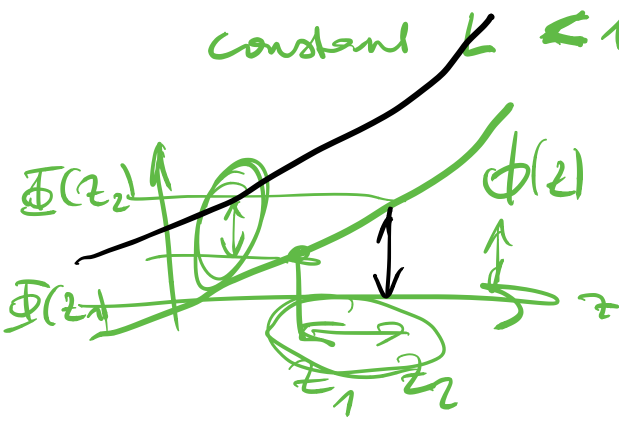
$$\boxed{x^{(k+1)} = \Phi(x^{(k)})}$$

if Φ is Lipschitz ~~constant~~ continuous with

constant $L < 1 \Rightarrow \quad x^{(k)} \to x = \sqrt{2}$

$$|x^{(k)} - x|^2 \leq \text{const.}\, L^k \qquad L^k \approx 10^{-10}$$

$$k \geq \log_L 10^{-10}$$



$$|\Phi(z_2) - \Phi(z_1)| \leq L \cdot |z_1 - z_2|$$

$$\min_{\underline{w}} \; \underline{\mathcal{E}}(w) = \frac{1}{m} \| \underline{y} - \underline{X}\,\underline{w} \|_2^2$$

$\uparrow$ $\begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix}$ $\begin{pmatrix} (\underline{x}^{(1)})^T \\ \vdots \end{pmatrix}$

$$-\alpha \cdot \nabla \underline{\mathcal{E}}(\underline{w}^{(opt)}) = \underline{\phi}$$

$$\underline{w}^{(opt)} - \alpha \nabla \underline{\mathcal{E}}(w^{(opt)}) = w^{(opt)}$$

$$\underline{\Phi}(\underline{w}) \rightarrow \underline{w} - \alpha \nabla \underline{\mathcal{E}}(w) \rightarrow \underline{w}^{(k+1)} = \underline{\Phi}(\underline{w}^{(k)})$$

step-size / learning rate

$$\nabla \underline{\mathcal{E}}(\underline{w}) = -(2/m)\, \underline{X}^T (\underline{y} - \underline{X}\,\underline{w})$$

$$\underline{\Phi}(\underline{w}) = \underline{w} + (2\alpha/m)\left[\underline{X}^T \underline{y} - \underline{X}^T \underline{X} \cdot \underline{w}\right]$$

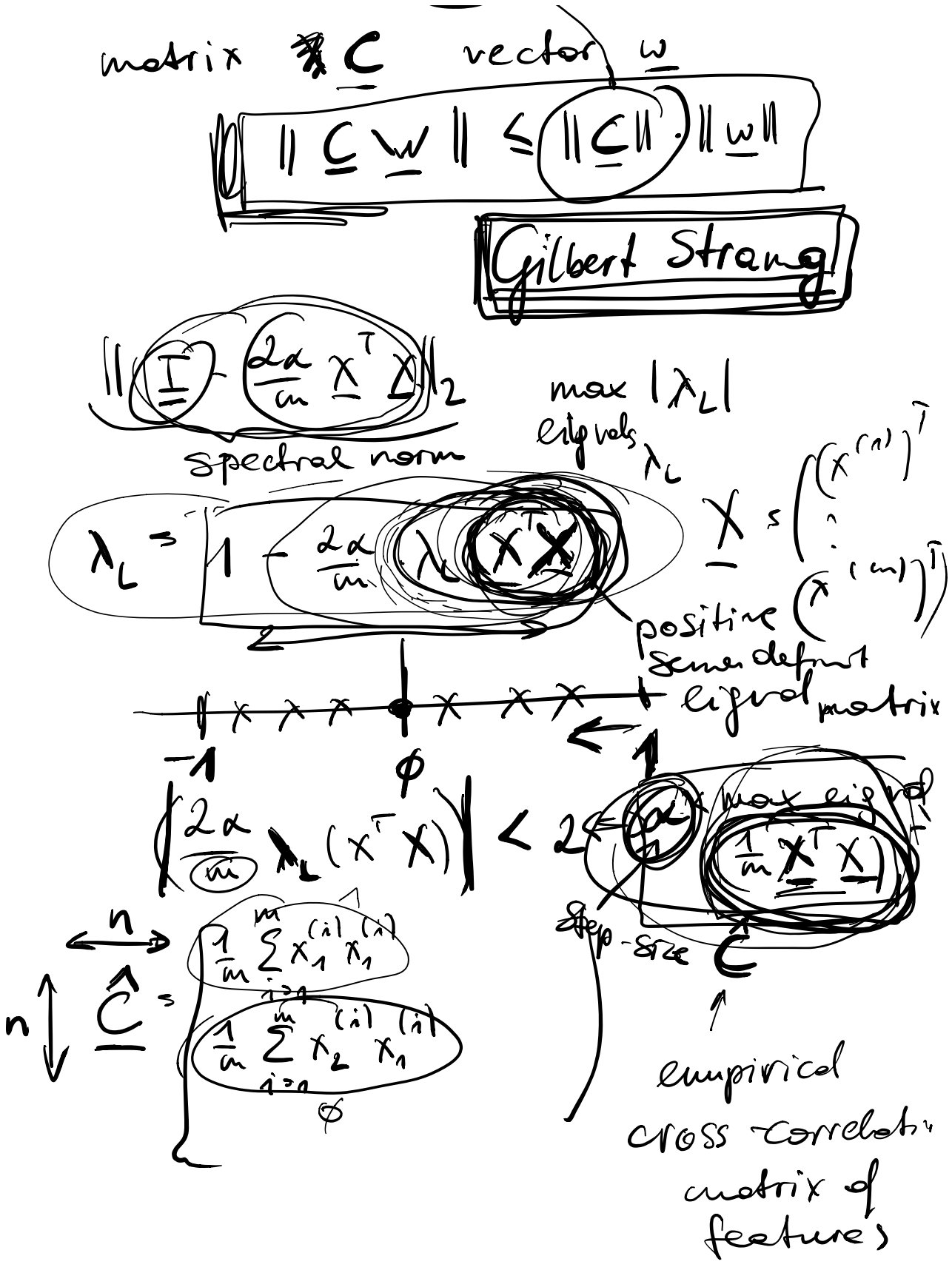
$$\underline{I} \cdot \underline{w} = \left[\underline{I} - \frac{2\alpha}{m} \underline{X}^T \underline{X}\right] \underline{w} + (2\alpha/m)\, \underline{X}^T \underline{y}$$

doesn't depend on w

$$\underline{\Phi}(\underline{w}^{(1)}) - \underline{\Phi}(\underline{w}^{(2)}) =$$

$$\left\| \left(\underline{I} - \frac{2\alpha}{m} \underline{X}^T \underline{X}\right) (\underline{w}^{(1)} - \underline{w}^{(2)}) \right\|$$

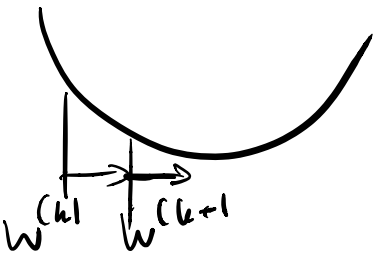
$$\leq L \, \| \underline{w}^{(1)} - \underline{w}^{(2)} \|$$


matrix C
vector w
$\|Cw\| \le \|C\| \cdot \|w\|$
Gilbert Strang
$\left\| I - \frac{2\alpha}{m} X^T X \right\|_2$
spectral norm
$\max |\lambda_L|$
eigvals λ_L
$\lambda_L = 1 - \frac{2\alpha}{m} \lambda_L (X^T X)$
$X = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{pmatrix}$
positive semi-definite
eigval matrix
-1
0
1
$\left| \frac{2\alpha}{m} \lambda_L (X^T X) \right| < 2$
α
max eigval
$\frac{1}{m} X^T X$
step-size
α
n
n
$\hat{C} = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^{m} x_1^{(i)} x_1^{(i)} & \\ \frac{1}{m} \sum_{i=1}^{m} x_2^{(i)} x_1^{(i)} & \phi \end{pmatrix}$
empirical cross-correlation matrix of features

Momentum

$$\underbrace{w^{(k+1)}}_{\text{new guess for optimal weights}} = \underbrace{w^{(k)}}_{\text{previous guess}} - \alpha \nabla \mathcal{E}(w^{(k)}) + \underbrace{\beta (w^{(k)} - w^{(k-1)})}_{\text{momentum}}$$

gradient descent



Deep Learning Book

Stochastic Gradient Descent

$$w^{(k+1)} = w^{(k)} - \alpha \nabla \mathcal{E}(w^{(k)}) + \beta (w^{(k)} - w^{(k-1)})$$

$$\mathcal{E}(w) = \frac{1}{m} \| y - Xw \|_2^2$$

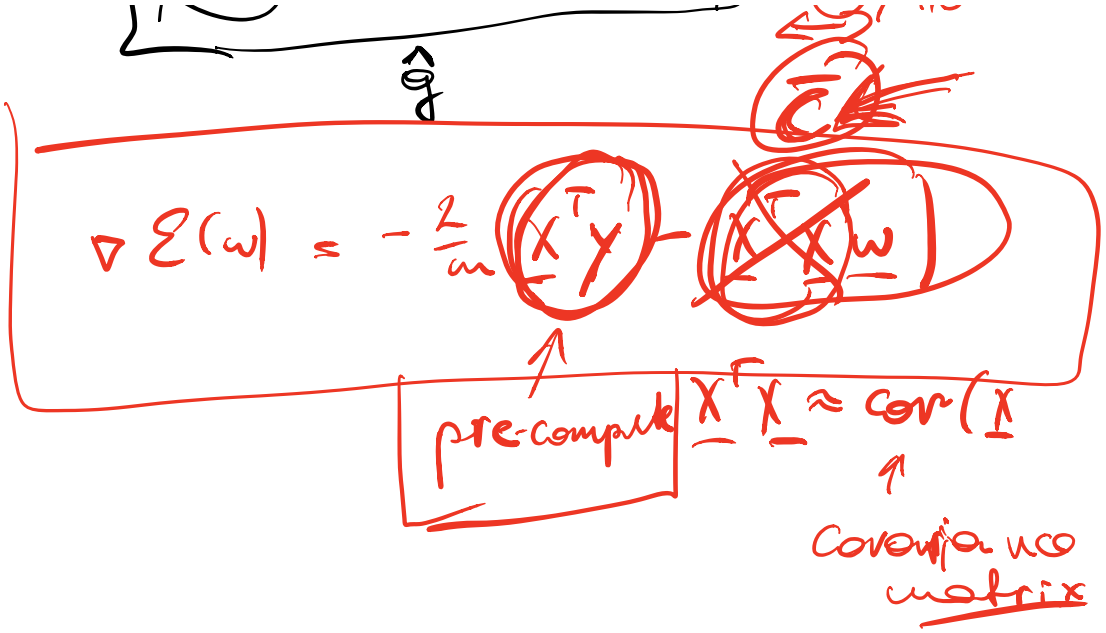
$$\nabla \mathcal{E}(w) = -\frac{2}{m} X^T (y - Xw)$$

Squared loss

$$-\frac{2}{m} \sum_{i=1}^{m} x^{(i)} (y^{(i)} - w^T x^{(i)}) \quad \text{with} \quad X^T X = \sum_{i=1}^{m}$$

≈ randomly select subset $\mathcal{B} \subseteq \{1, \dots, m\}$

$$\left(\sum_{i \in \mathcal{B}} x^{(i)} (y^{(i)} - w^T x^{(i)}) \right) \quad |\mathcal{B}| / m$$


$\nabla \mathcal{E}(w) = -\frac{2}{m}(X^T Y - X^T X w)$
pre-compute
$X^T X \approx cov(X$
covariance matrix