Topics:-

loss-functions

Math-heavy

practical { - "Kernel" SVMs (Math + intuition)

optmzn (beyond the scope)

[- Naive Bayes]

Simplest

BayesThm

Geom soptmzn

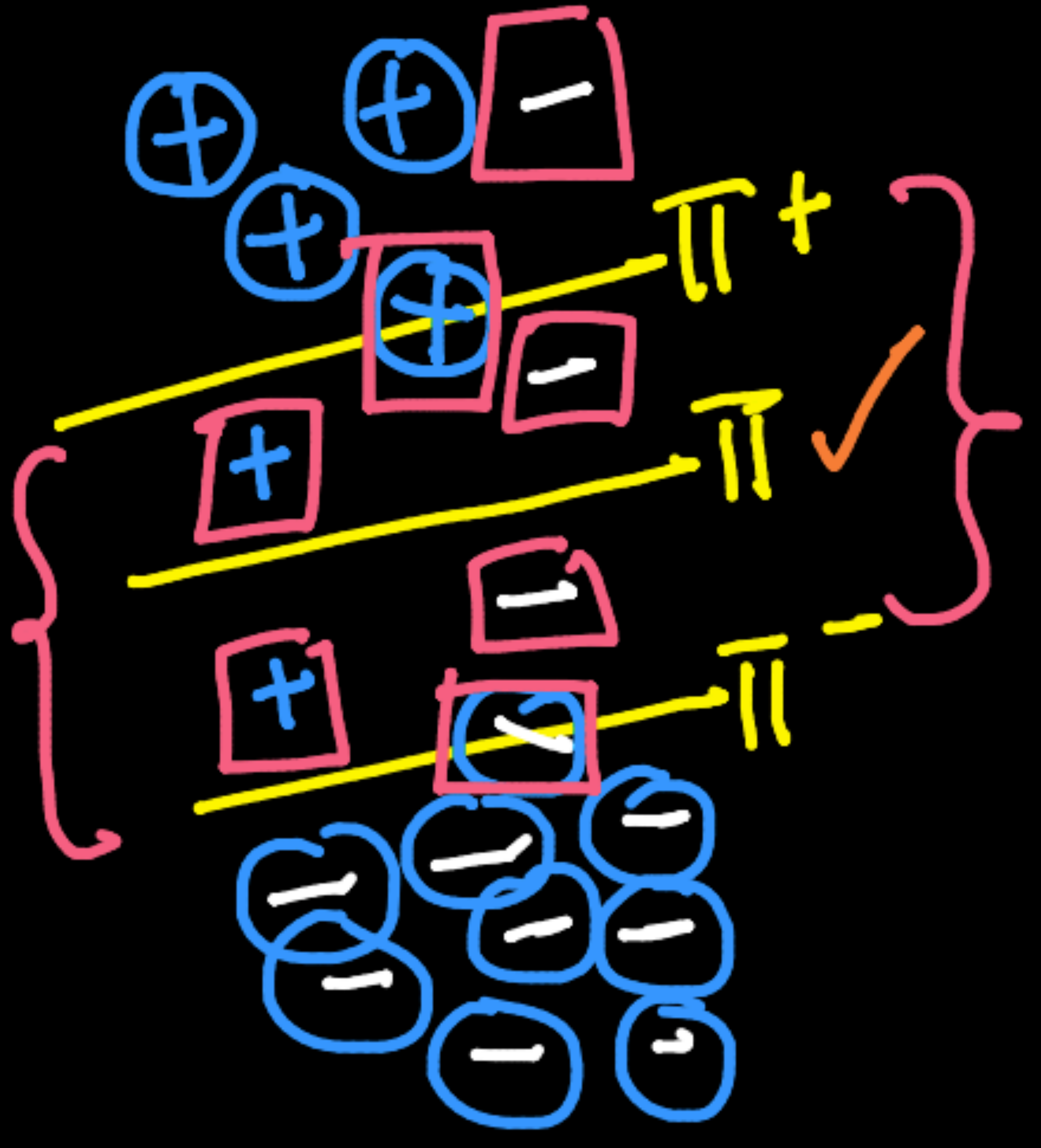
Lr. SVM:

Soft-margin

$$\min_{w,b} \; \underbrace{\frac{\|w\|}{2}}_{\text{inv. of Margin}} + \underbrace{C}_{\text{hyper-param}} \; \underbrace{\frac{1}{n}\sum_{i=1}^{n}\xi_i}_{\text{mean } \xi_i}$$

$$\text{s.t.} \quad \left[\begin{array}{l} y_i(w^T x_i + b) \ge 1 - \xi_i \\ \xi_i \ge 0 \quad \forall i \end{array}\right.$$

Previously:



Def

$\boxed{\text{SV}}$ / $\boxed{x}$: on π^+/π^- or margin / mis class

Loss minimization

Logistic regr $\rightarrow \min$ log-loss + λ reg

Linear reg $\rightarrow$ Min Sq-loss + λ reg

SVM $\rightarrow \min$ C. hinge-loss + reg(L_2)

Soft-margin Linear SVMs:

$$\min_{w,b} \underbrace{\boxed{\frac{\|w\|}{2}}}_{\text{inv. margin}} + \underbrace{\boxed{C}}_{\text{reg-hp}} \underbrace{\boxed{\frac{1}{n} \cdot \sum_{i=1}^{n} \boxed{\xi_i}}}_{\text{error / loss}}$$

10^{10}

hinge-loss (later)

$\frac{1}{2} \cdot L_2 \cdot \text{reg}$

$$\text{s.t} \begin{cases} y_i (w^T x_i + b) \geq 1 - \xi_i & \forall i: 1 \to n \\ \xi_i \geq 0 & \forall i: 1 \to n \end{cases}$$

$C \uparrow \Rightarrow$ more imp to hinge loss $\Rightarrow$ loss is as small as small

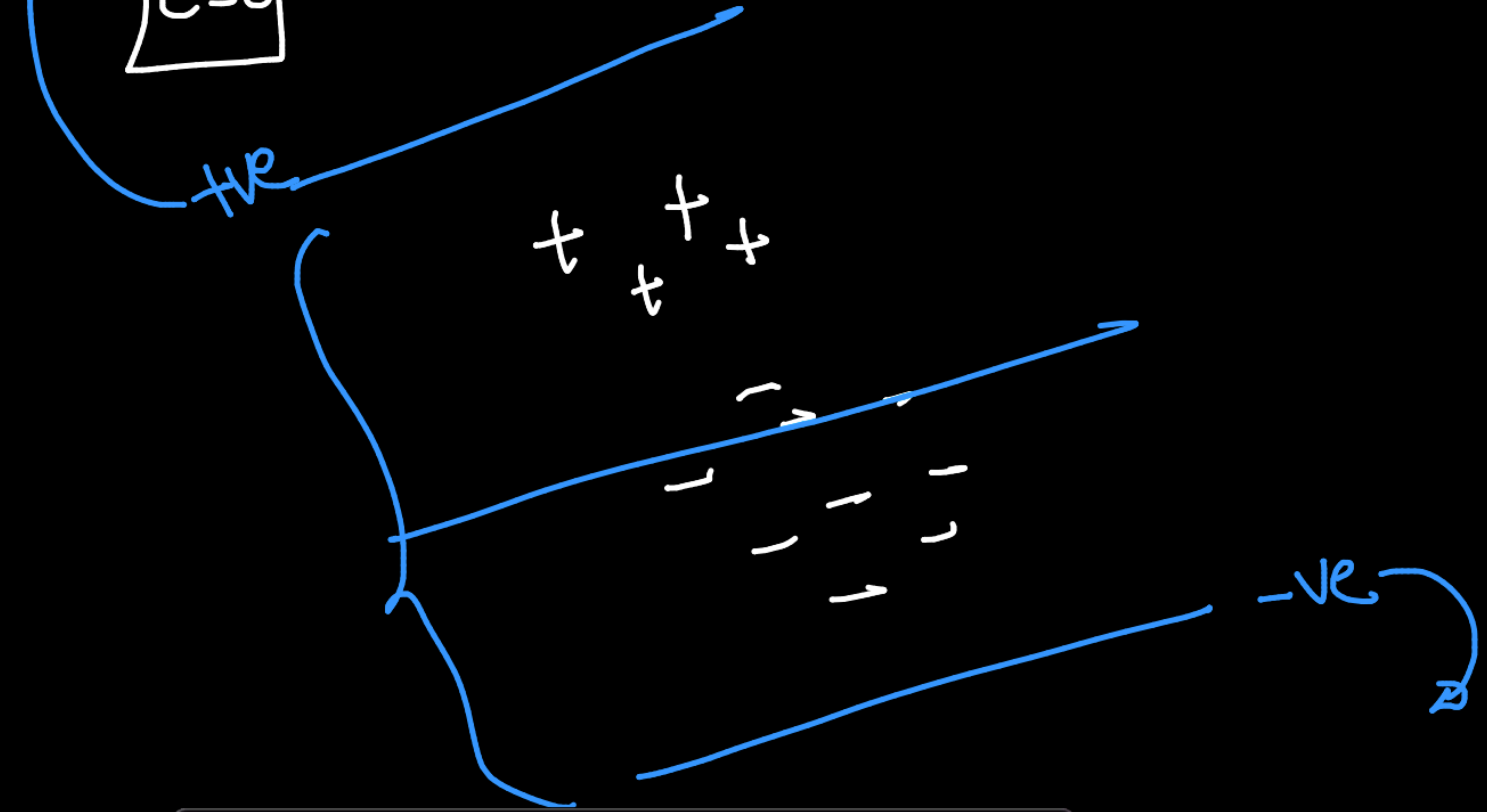
$\Downarrow$

overfit

$\min \|w\|/2$

$C \downarrow \Rightarrow \max \dfrac{2}{\|w\|} \Rightarrow$ perf. on train $\downarrow$ $\Rightarrow$ underfit

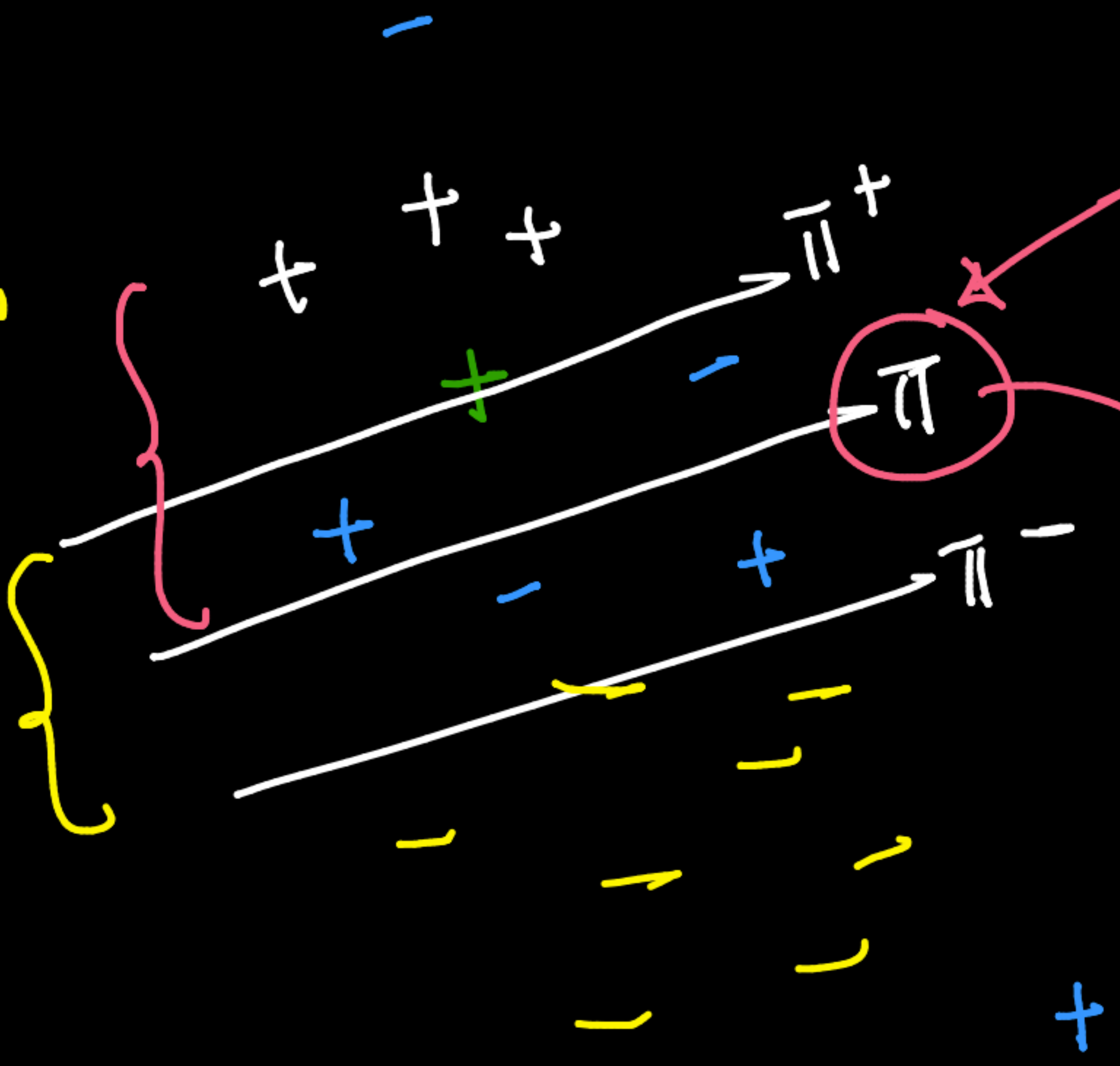
$\boxed{C=0}$



hard-margin SVM

↓

perfectly linearly-sep data $\Rightarrow$ no error



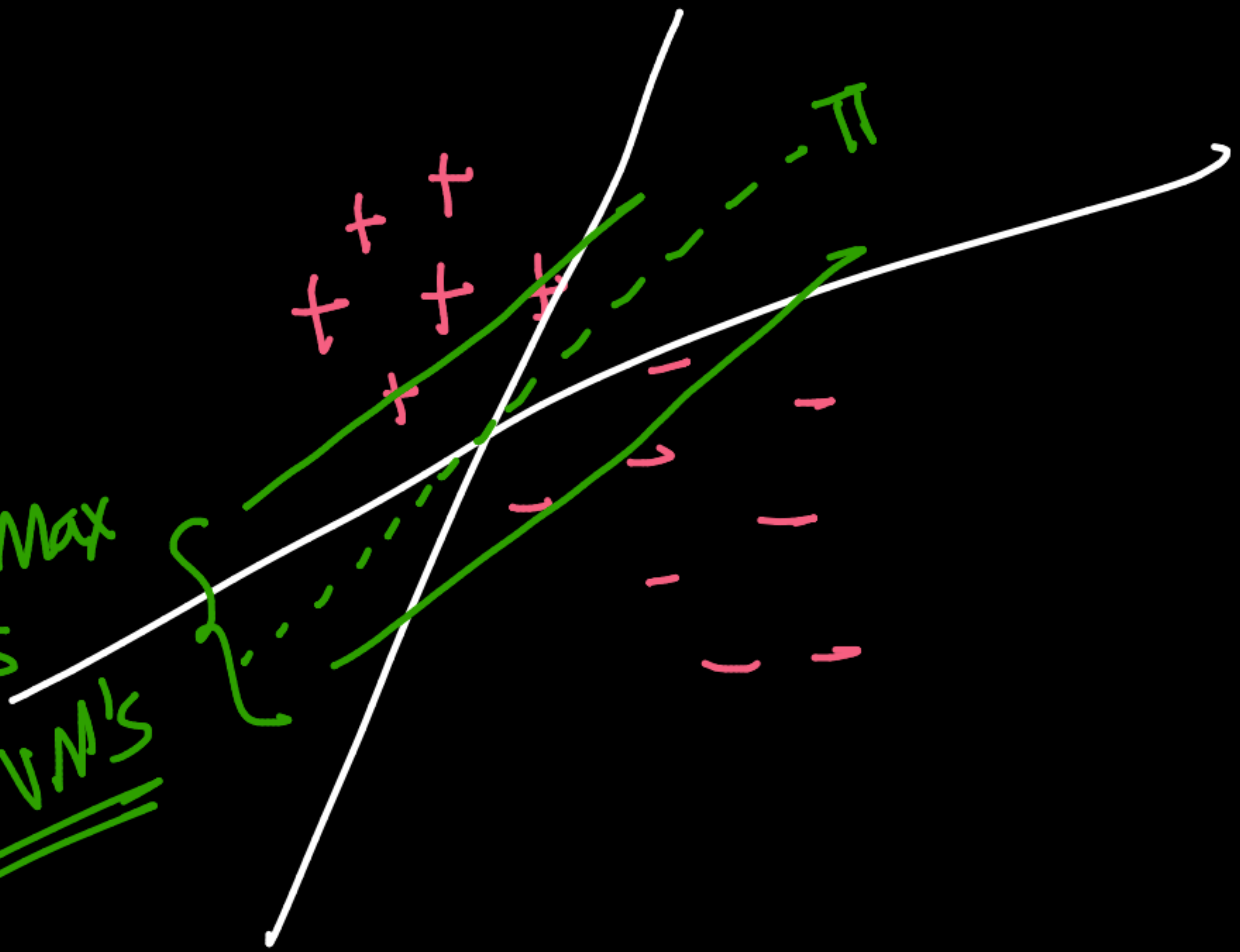
linear-SVM

(ideal)

$w^{*T}x + b^* = 0$

solve (optimzn)

$x_q \rightarrow w^{*T}x_q + b^*$



Margin -Max models

SVM's

Constrained optmzn → lagrange multipliers

[PCA / SVM ...]

(making constraints part of objective)

↓

Gradient Descent

$$\underbrace{\min_{w,b} \frac{\|w\|}{2} + \boxed{C} \sum_{i=1}^{n} \xi_i + \lambda_1 \boxed{\text{Constr 1}} + \lambda_2 \boxed{\text{Constr2}}}_{\mathcal{L}(w,b)}$$

$\downarrow$

Gradient-Descent / Specialized algo (SMO) - later

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Convexity → [SVM - optmzn]

one minima / global minima

We dont care

$\boxed{\text{DL}}$ is non-convex

↓

hacks / algo . . .

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:40

$$\sum_{i=1}^{n} \xi_i$$

SVM:- +1 & -1

y_i's

logistic reg: +1 & 0

+1 & -1

$$\sum_{i=1}^{n} \log(1+\exp(-y_i(w^T x_i + b)))$$

exercise

$\sigma(w^T x_i + b)$

+1 & 0: $\rightarrow \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$

11 / 12

binary Classfn

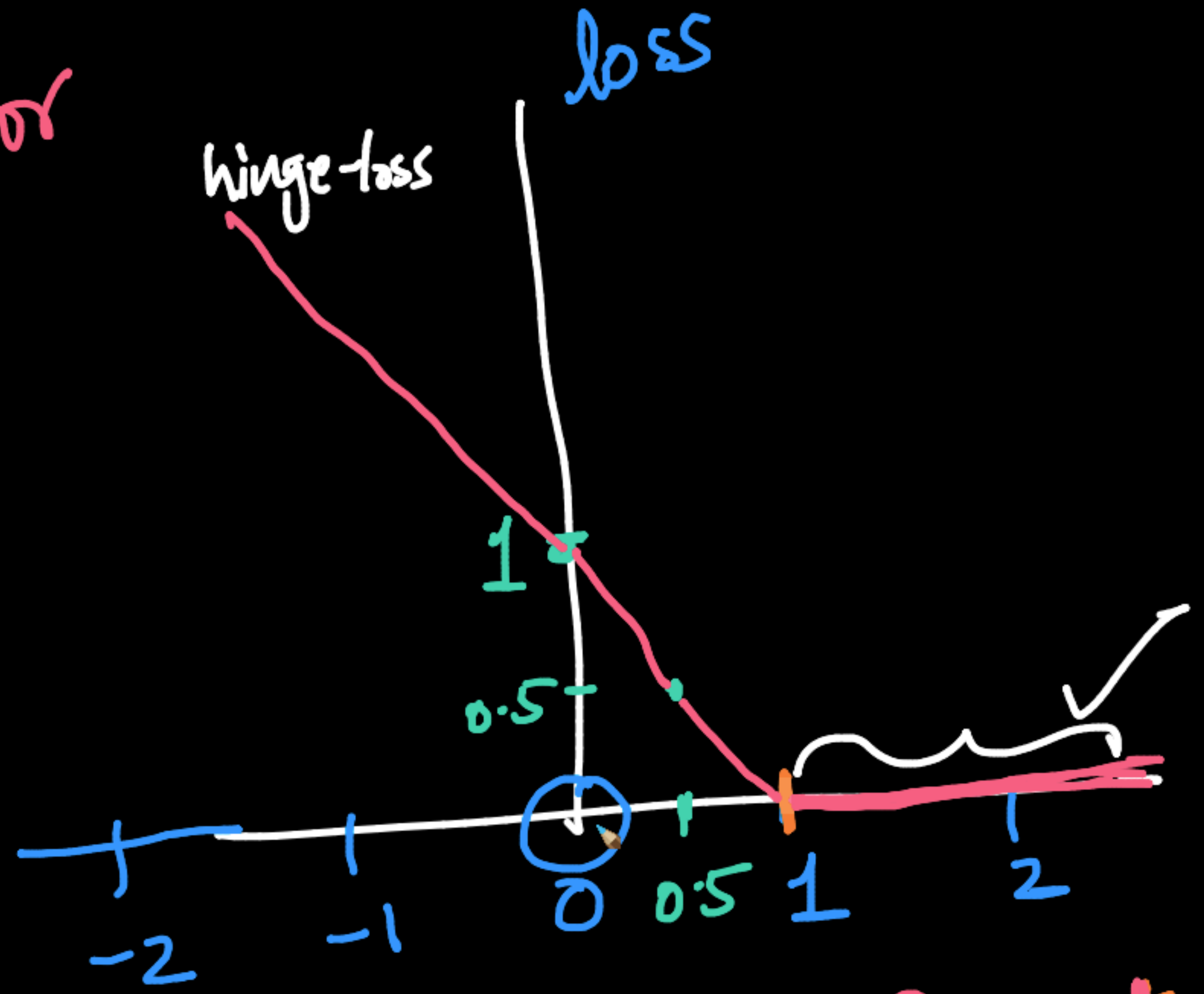
$f(x_i) = \text{model}$

binary-classification :→ $y_i \in \{+1, -1\}$

$y_i \underbrace{f(x_i)}_{(w^T x_i + b)} = Z_i$

defined

ξ_i : loss | error



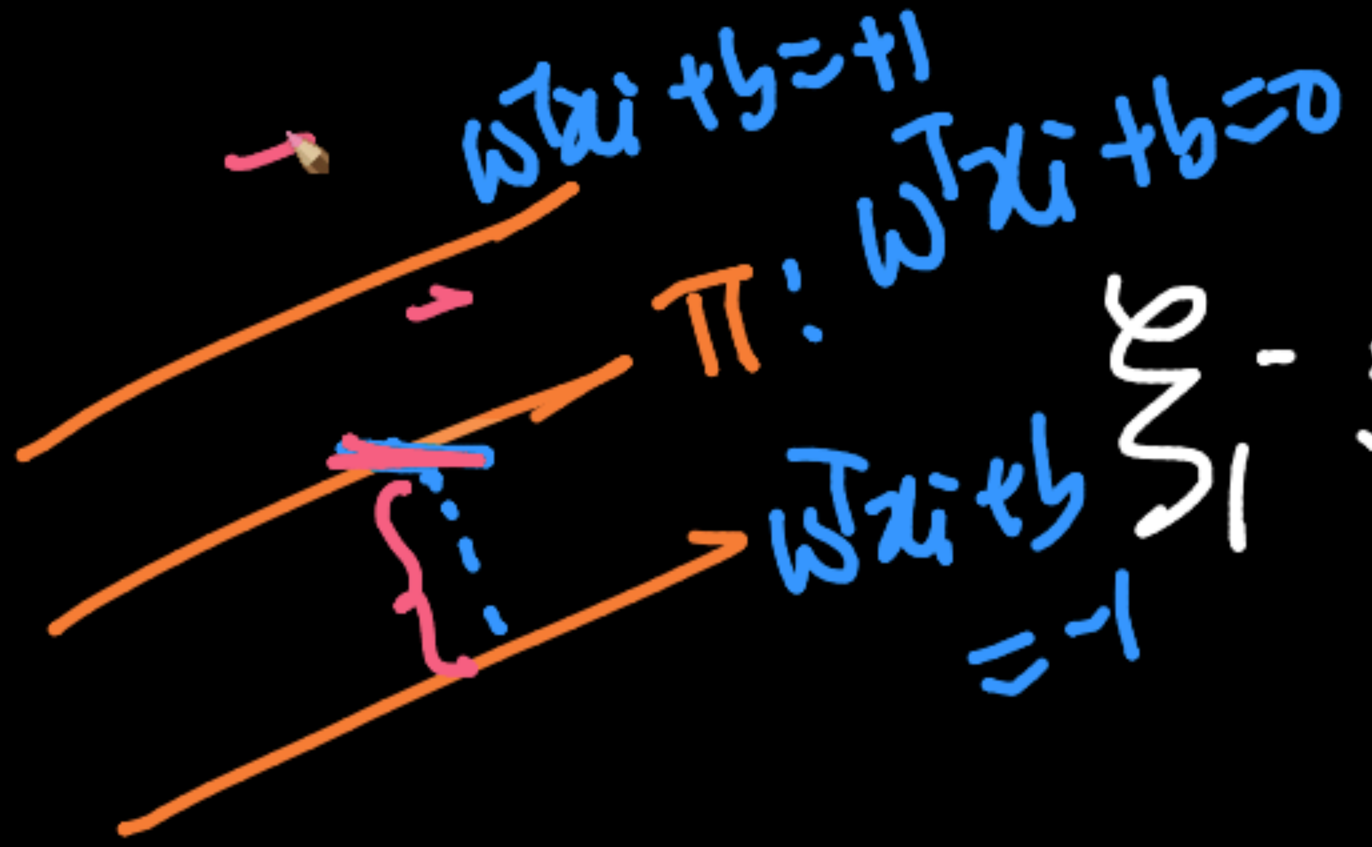
SVM-loss

= hinge-loss

+ve

$y_i f(x_i) = z_i$

$\xi_i =$



ξ_i :

hinge loss:

$$\underbrace{y_i \overbrace{(w^T x_i + b)}^{-ve}}_{z_i} \geq 1 \Rightarrow \xi_i = 0$$

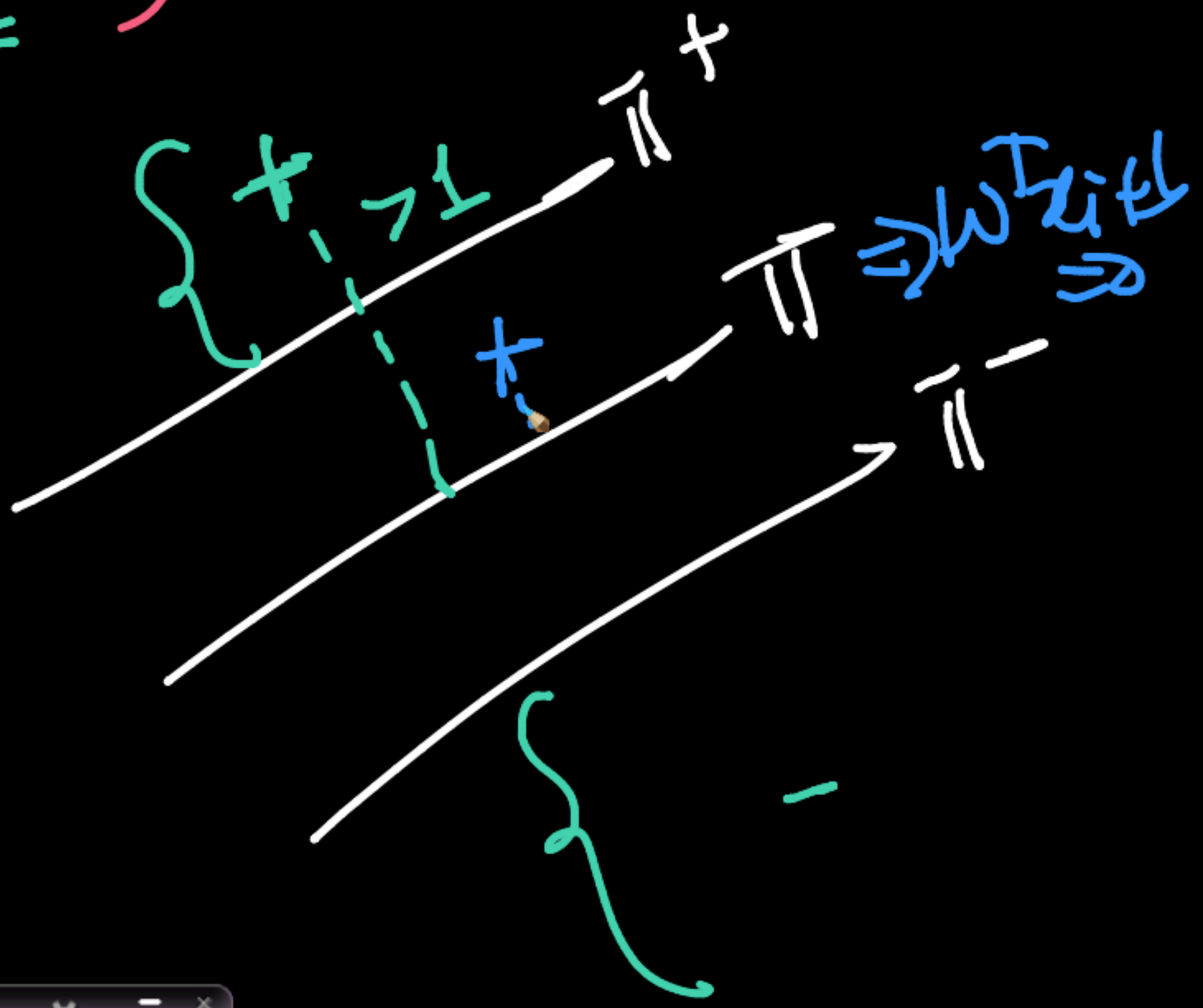
$$\xi_i = 1 - z_i$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i$$

$$\min \ \xi_i = 1 - \underbrace{y_i (w^T x_i + b)}_{z_i} \text{ if}$$

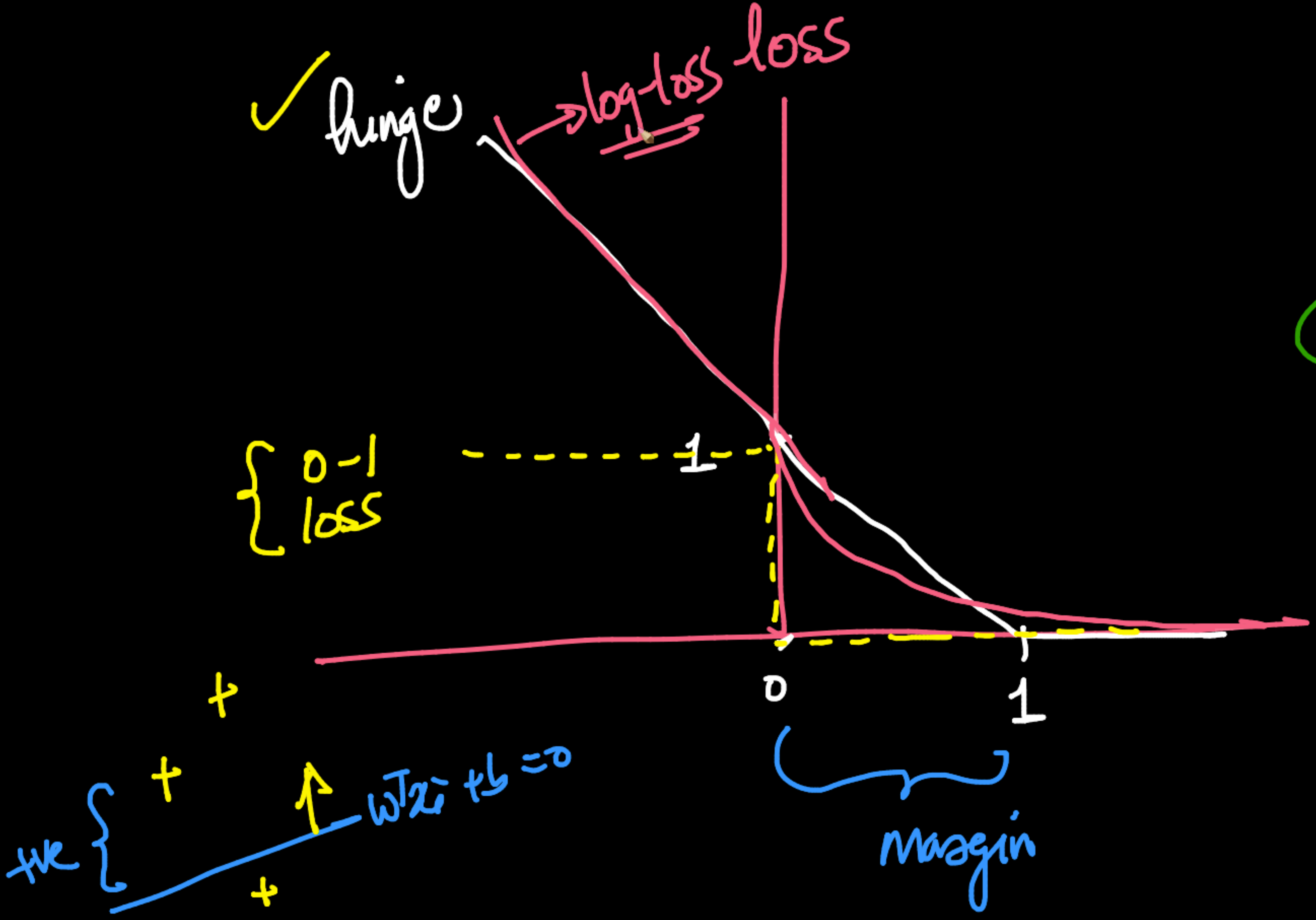
$$\xi_i = 0 \text{ or } 1 - z_i$$

$$\xi_i = \max(0, 1 - z_i)$$



$$1 - z_i = 1 - \underbrace{(y_i(w^T x_i + b))}_{>1}$$

$1 - z_i \rightarrow (-ve)$



0-1 loss:

logistic reg:

$\{+1, -1\} = y_i$

loss: $\log(1 + \exp(\underbrace{-y_i(w^T x_i + b)}_{z_i}))$

$\underline{y_i f(x_i)} = z_i$

>0 : correctly classified

<0 : incorrectly classified

log(1+exp(-x)) - Google Search | Loss functions for classification

google.com/search?q=log(1%2Bexp(-x))&rlz=1C5CHFA_enIN958IN958&oq=&aqs=chrome.3.69i59i450l8.4295406j0j7&sourceid=chrome&ie=UTF-8

Google

log(1+exp(-x))

All · Shopping · Images · News · Videos · More · Tools

About 7,73,00,00,000 results (0.65 seconds)

Graph for log(1+exp(-x))

$y_i f(x_i)$

More info

https://cran.r-project.org › Rmpfr › log1mexp-note PDF

Accurately Computing log(1 - exp(.)) – Assessed by Rmpfr

by M Mächler · Cited by 1 — l1e() computes the relative error of three different ways to compute **log(1 − exp(−a))** for positive a (instead of computing **log(1 − exp(x))** for negative **x**). R> ...

9 pages

https://math.stackexchange.com › questions › any-simpl...

Any simplification of $\log_e{(1+e^x)}? - Math Stack Exchange

22-Jun-2017 — can you explain me how **log(1+e^x)** becomes **x+log(1+e^(−x))** please? – Sam. Jun 22, 2017 at 7:48. 1.

2 answers · Top answer: You can use the Ta

Approximate $\log(1-e^x)$ where $x<0 - Ma

17 / 20

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:05

log(1+exp(-x)) - Google Search | Loss functions for classification

google.com/search?q=log(1%2Bexp(-x))&rlz=1C5CHFA_enIN958IN958&oq=&aqs=chrome.3.69i59i450l8.4295406j0j7&sourceid=chrome&ie=UTF-8

Google

log(1+exp(-x))

All · Shopping · Images · News · Videos · More · Tools

About 7,73,00,00,000 results (0.65 seconds)

Graph for log(1+exp(-x))

More info

https://cran.r-project.org › Rmpfr › log1mexp-note PDF

[Accurately Computing log(1 - exp(.)) – Assessed by Rmpfr]

by M Mächler · Cited by 1 — l1e() computes the relative error of three different ways to compute **log(1 − exp(−a))** for positive a (instead of computing **log(1 − exp(x))** for negative **x**). R> ...

9 pages

https://math.stackexchange.com › questions › any-simpl...

[Any simplification of $\log_e{(1+e^x)}? - Math Stack Exchange]

22-Jun-2017 — can you explain me how **log(1+e^x)** becomes **x+log(1+e^(−x))** please? – Sam. Jun 22, 2017 at 7:48. 1.

2 answers · Top answer: You can use the Ta

Approximate $\log(1-e^x)$ where $x<0 - Ma

18 / 21

real-world:

Lr- soft-margin C

SVM: - min hinge loss + reg (L_2)

margin

log-reg: - min log-loss + λ l_2-reg

Similar performance

SUGGESTION: on your own (Math) ✓

19 / 22

Chrome File Edit View History Bookmarks Profiles Tab Window Help

10:08

22 / 22

Then, why so much hype on SVM 90's

$\boxed{\text{Kernel-}}$SVM

model:

$w^T x + b$

Model

$$\min_{w,\, b} \frac{1}{2}\|w\| + \boxed{C}\, \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

$$\text{s.t} \quad y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i : 1 \to n$$

$$\xi_i \geq 0 \quad \forall i : 1 \to n$$

Primal

Primal-dual Equivalence

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:28

dual:

α_i's

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t} \quad 0 \le \alpha_i \le C \quad \forall i : 1 \rightarrow n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

1. $x_i \rightarrow \boxed{\alpha_i}$

* 2. x_i's occur only in the form of $x_i^T x_j$

3. $f(x_q) = \underbrace{\sum_{i=1}^{n} \alpha_i y_i x_i^T}_{w} x_q + b$

α_i's & x_i's

24 / 27

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:25

4

$\alpha_i = 0$ for non support vectors $\rightarrow$ $1 - z_i \le 0$

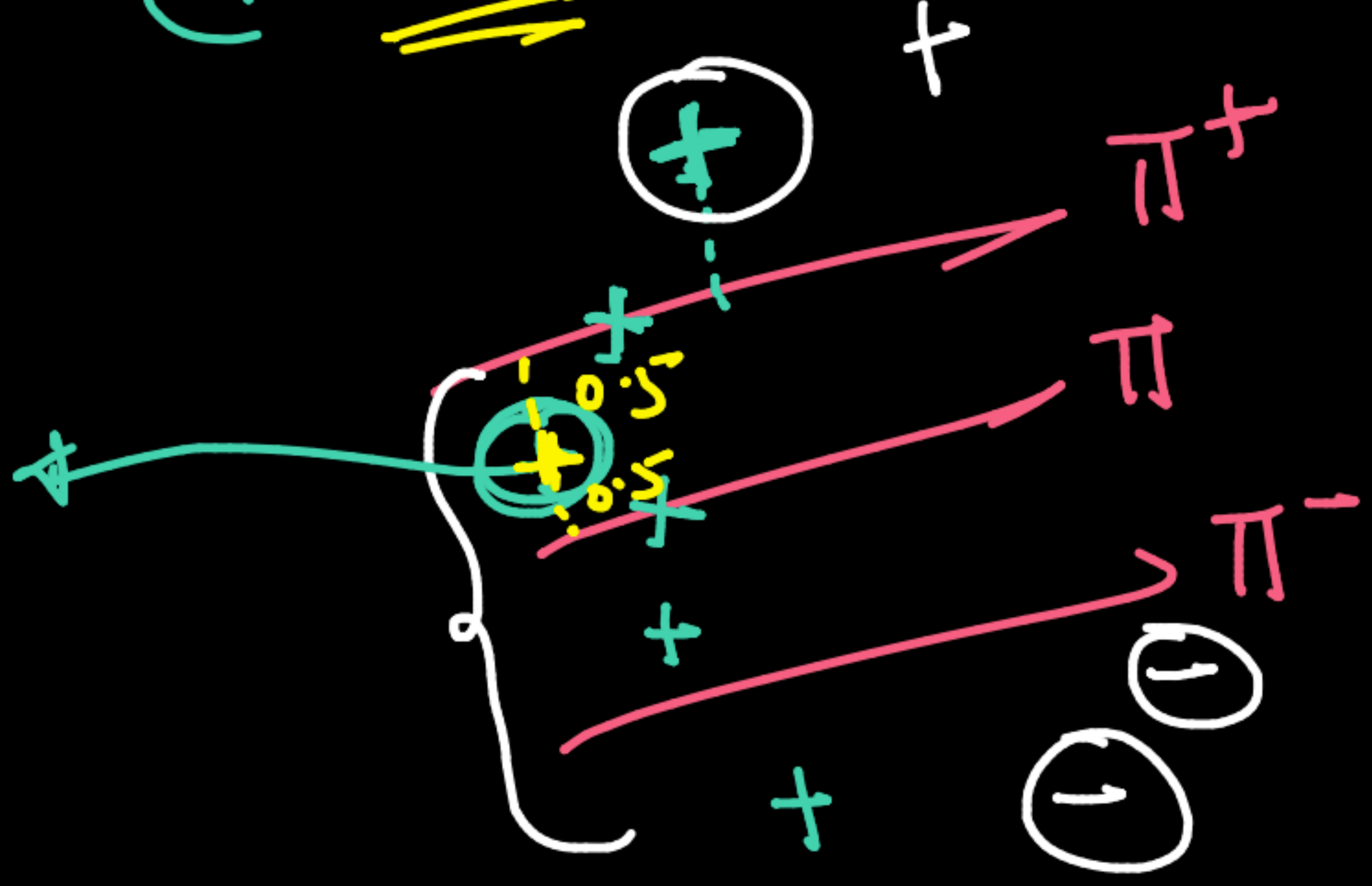
$\alpha_i > 0$ for SV

$z_i = y_i(w^T x_i + b)$

$(1 - z_i) > 0$

$$1 - \underbrace{y_i(w^T x_i + b)}_{0.5} = 0.5$$

$\xi_i = 0.5$



25 / 27

$$\hat{y}_q = f(x_q) = \sum_{i=1}^{n} \alpha_i y_i x_i^T \boxed{x_q} + b \rightarrow \alpha_i\text{'s} \;\&\; x_i\text{'s}$$

10,000 pts

α_i:

- $= 0$ for non SVs (9900)
- > 0 for SVs (100)

@ runtime

$\alpha_i > 0$ (SVs)

let 100 SVs

y_i, x_i s.t $\alpha_i > 0$ (SVs)

need only $\{\alpha_i, y_i, x_i$ for $\boxed{SVs}\}$

$$f(x_q) = \hat{y}_q = \sum_{i=1}^{n} \alpha_i y_i x_i^T \cdot x_q$$

$\checkmark$

dual → solved $\qquad \overbrace{\alpha_{10}, \alpha_{12}, \alpha_{30}} > 0$

all remaining $\alpha_i = 0$

$$f(x_q) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x_q + \cancel{b}$$

$$\hat{y}_q = \text{scalar} = \left[\alpha_{10} y_{10} \overbrace{x_{10}^T} x_q + \alpha_{12} y_{12} \overbrace{x_{12}^T} x_q + \alpha_{30} y_{30} \underbrace{x_{30}^T} x_q \right]$$

Multi-class classfn using SVM

$\downarrow$

one vs rest

SVR → (later)

↳ basic formulation

Model

retrain → Some-hacks

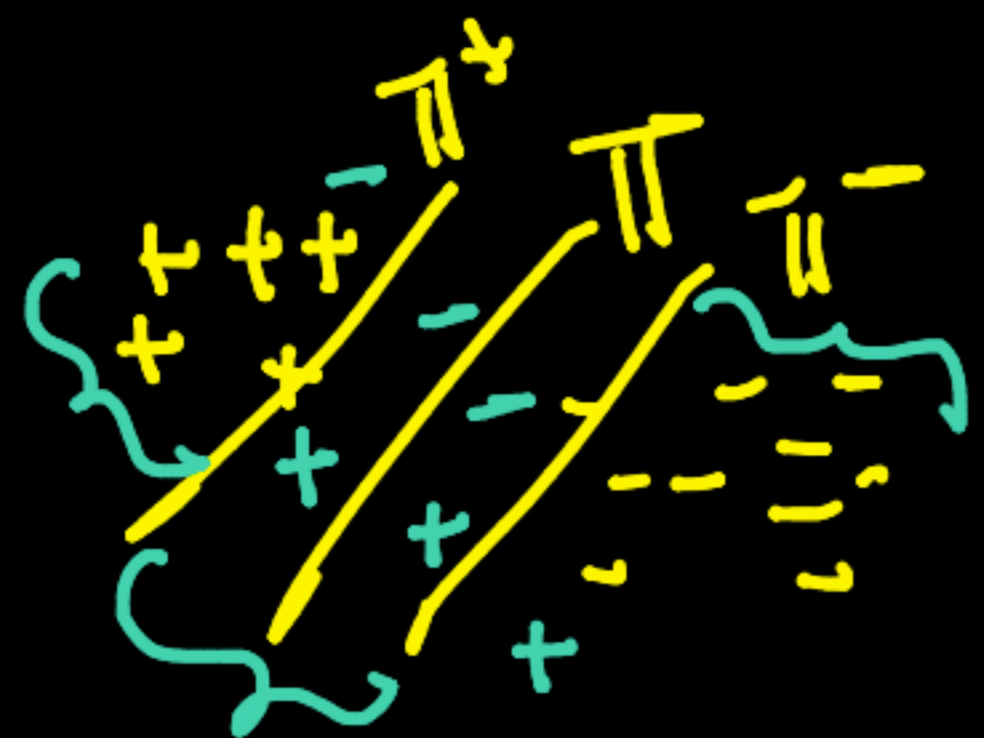
using whole data

Kernel trick

10:39

Recap: (SVMs)

Hard-Margin → (margin-max) → L_2-reg..

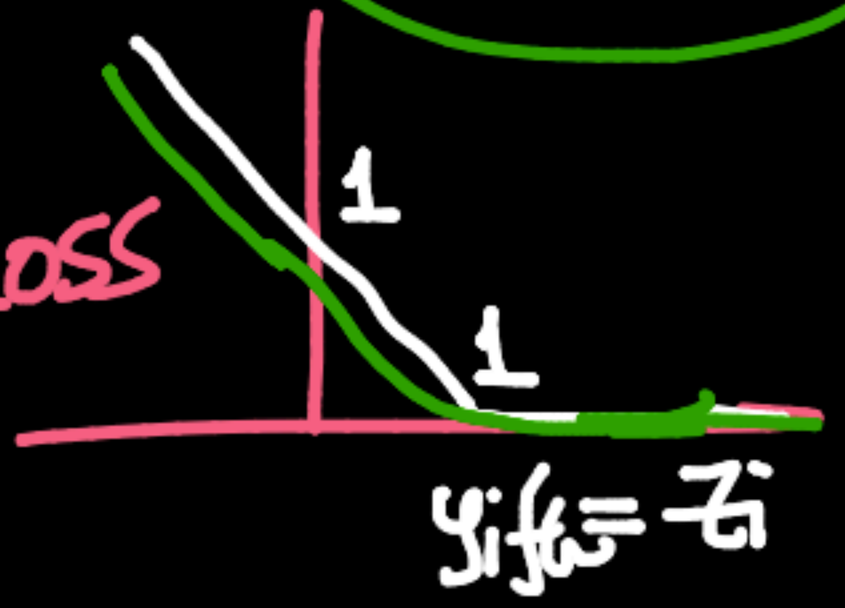


linearly sep-data

Soft-margin → ξ_i : loss/error → optmzn problem

LR-SVM

Loss-minimization → hinge-loss



Primal-Dual → primal = Dual

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:56

T

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \cancel{x_i^T x_j} \; K(x_i, x_j)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C \quad \forall i$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

① $x_i \rightarrow \alpha_i$

$\hat{y}_q$ $\quad K(x_i, x_q)$

② $f(x_q) = \sum_{i=1}^{n} \alpha_i y_i \cancel{x_i^T x_q} + b$

③ SVs: $\rightarrow \alpha_i > 0$, non SVs : $\alpha_i = 0$

④ $x_i^T x_j$

$x^T x^q$

33 / 38

$x_i^T x_j \rightarrow$ similarity between $\overbrace{x_i \,\&\, x_j}$

$\rightarrow$ cosine-sim(x_i, x_j)

$\downarrow$

$K(x_i, x_j)$

biz rules

1000 molecules

Chemical/bio properties

sim(C_1, C_2)

{

return val;

}

x_i

Sim(x_i, x_j)

logistic regr?

Cust 1 X

Cust 2 X

Ques

n - Customers → do NOT have $x_i \in \mathbb{R}^d$

dual-form of SVM

sim (c_i, c_j)

{

Logistic reg?

x_i's need

↓

optmzn problem

x_i's x_j's

$\boxed{Q}$

$\underline{\text{sim}} \propto \frac{1}{\underline{\text{dist}}}$

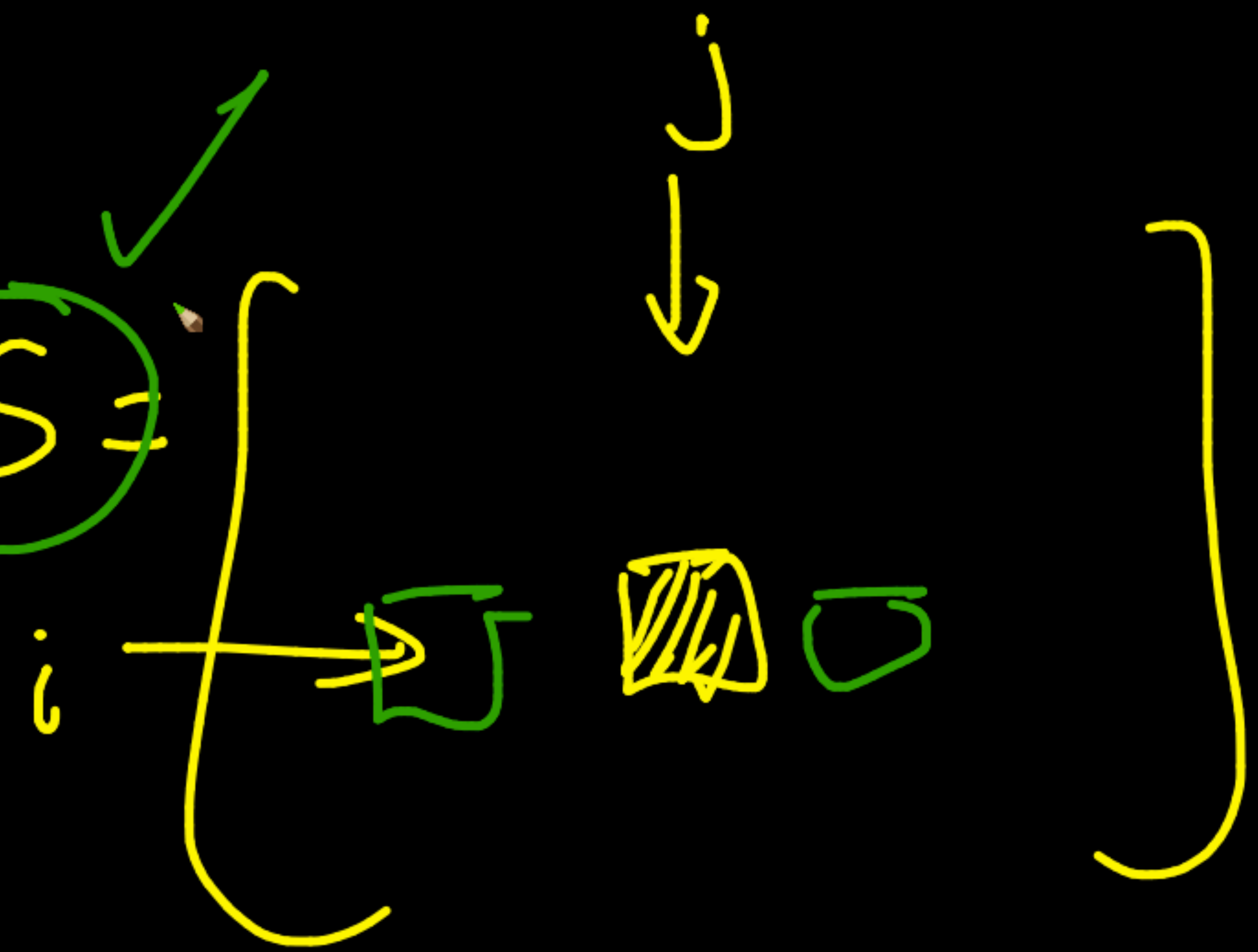
$\underline{\text{K-NN}}$

$Sim(C_1, C_2)$

$\{$

$\}$

$S =$



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:19

Kernel SVM

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{K(x_i, x_j)}$$

$$\text{s.t} \quad 0 \le \alpha_i \le C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$x_i^T x_j$

Code

Math fn:

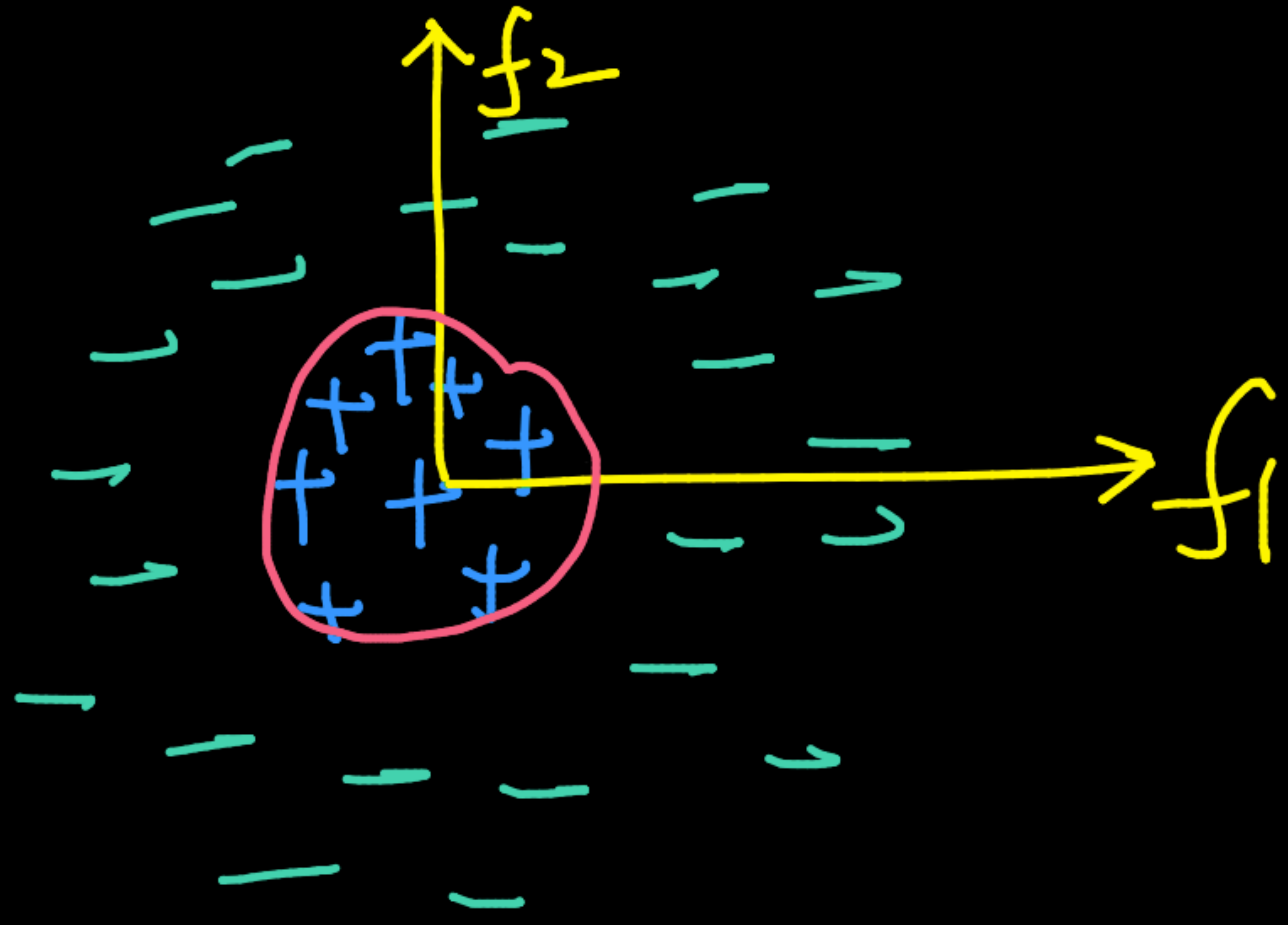
$\boxed{x_i^T x_j}$

38 / 47

logistic reg

f_1^2 , f_2^2

(polynomial features)



Polynomial Kernel

$$K(x_1, x_2) = (x_1^T x_2 + c)^m$$

Quadratic-Kernel $= (1 + x_1^T x_2)^2 = K_Q(x_1, x_2)$

what if we dont know the shape

↓

RBF (magic)

(most-popular)

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Kernelization

d-dim → d'dim

$d' >> d$

Some-properties

$$K_Q(x_1, x_2) = (1 + x_1^T x_2)^2$$

Π_{6d}

Π_{2d}

2d → 6d

$$\begin{cases} x_1 = [x_{11}, x_{12}] \\ x_2 = [x_{21}, x_{22}] \end{cases}$$

42 / 43

$O(n^2)$

reg:

Kernel-Trick

implicitly

$d \longrightarrow d'$

$\prod d'$

$\prod d$

$d' >> d$

$$f(x_q) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x_q + b$$

SVM → Kernel-trick ✓

X

Kernel log-reg X

Kernel-DL-models

log-reg + poly-features + reg

vs

SVM + ~~poly-kernel~~ RBF + reg

Most-widely

RBF → radial basis fn!

d → d' : ∞-many dim ...

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:27

Quad kernel

$$K_Q(x_1, x_2) = (1 + x_1^T \cdot x_2)^2 = \left(1 + [x_{11}, x_{12}]\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}\right)^2$$

$$\boxed{2\text{-}D}$$

$$2d \rightarrow \begin{cases} x_1 = (x_{11}, x_{12})^T \\ x_2 = (x_{21}, x_{22})^T \end{cases}$$

$$= \{1 + (x_{11}x_{21} + x_{12}x_{22})\}^2$$

$$= 1 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21} + 2x_{12}x_{22} + 2x_{11}x_{21}x_{12}x_{22}$$

46 / 52

$$\left(1 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21} + 2x_{12}x_{22} + 2x_{11}x_{21}x_{12}x_{22}\right)$$

$$\| \; x_1'^T x_2'$$

$$x_1' = [1, x_{11}^2, x_{12}^2, \sqrt{2}x_{11}, \sqrt{2}x_{12}, \sqrt{2}x_{11}x_{12}] \; : \text{6 dim}$$

$$x_2' = [1, x_{21}^2, x_{22}^2, \sqrt{2}x_{21}, \sqrt{2}x_{22}, \sqrt{2}x_{21}x_{22}] \rightarrow \text{6 dim}$$

$$\left(x_1'^T x_2'\right)$$

$$K_Q : \quad 2d \xrightarrow{\text{implicitly}} 6d$$

Computationally

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:21

Computation

Log-reg: → 6-dim space

explicitly

SVM → implicitly

2d

49 / 49

logistic-reg

Primal-SVM → explicitly create the 6-dim π^6

=

Dual-SVM → $x_i^T x_j$ → $K(x_i, x_j)$

non linear model

linear model

$d \rightarrow \boxed{d'}$ ✓

mimic

$d' >> d$

Computationally

RBF : d' : ∞-large ✓

K-NN (later)

Non-linear models

without creating
d'-dim explicitly

Chrome File Edit View History Bookmarks Profiles Tab Window Help

52 / 52

Why NOT (Kernels)

→ time-complexity ←→ - - - #SVS

designing "special purpose" Kernels (Bio-inf ...)

90's early 00's

→ RBF-Kernel SVM → similar to KNN

(∞-many)

→ GBDT/RF → in practice

REVISIT

→ DL

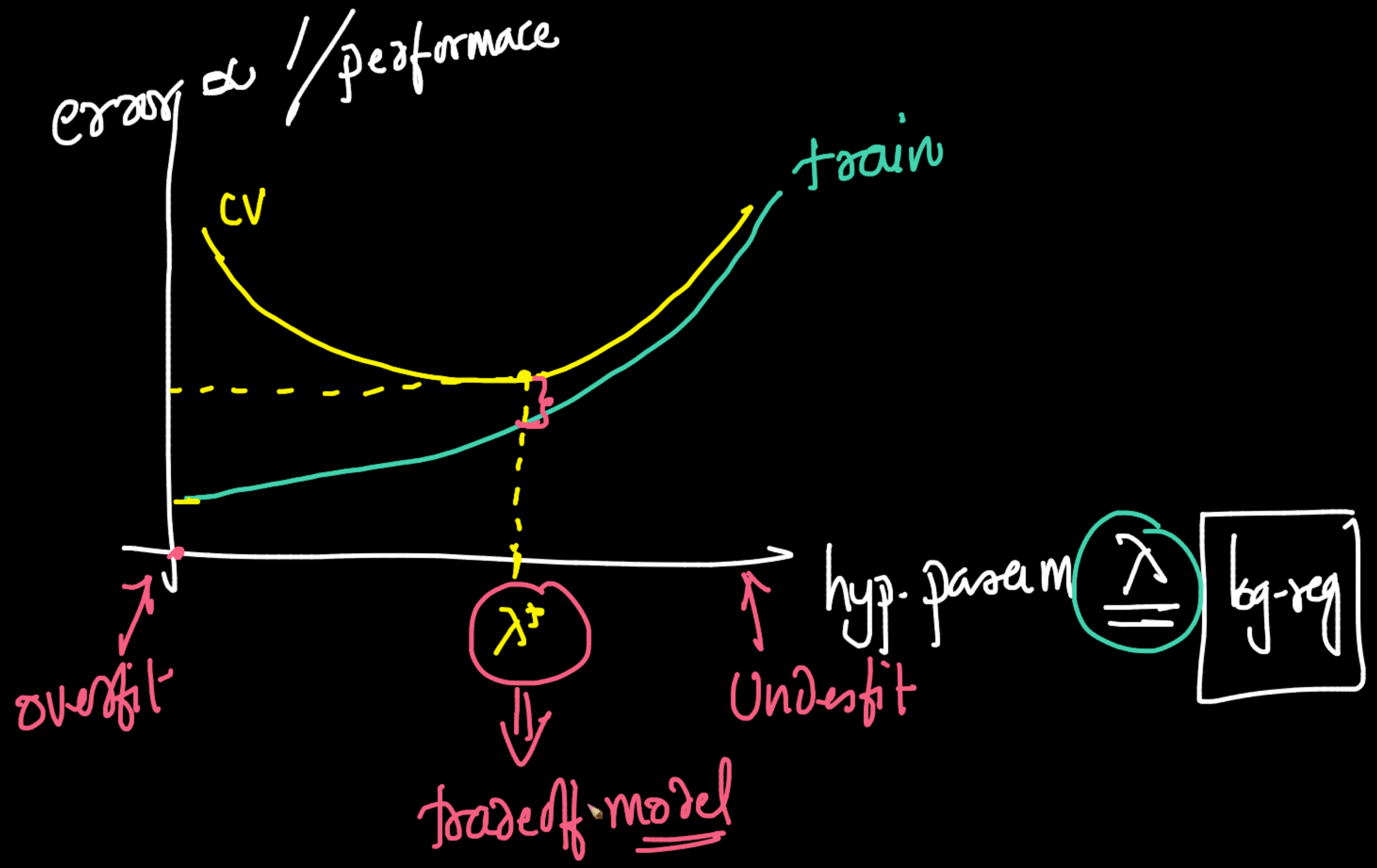
RF/GBDT/log-reg

feature-engineers

SVM

kernel-design

Q & A


error ∝ 1/performance
CV
train
hyp. param
λ
log-reg
overfit
λ*
underfit
tradeoff model

log(1+exp(-x)) - Google Search | Loss functions for classification

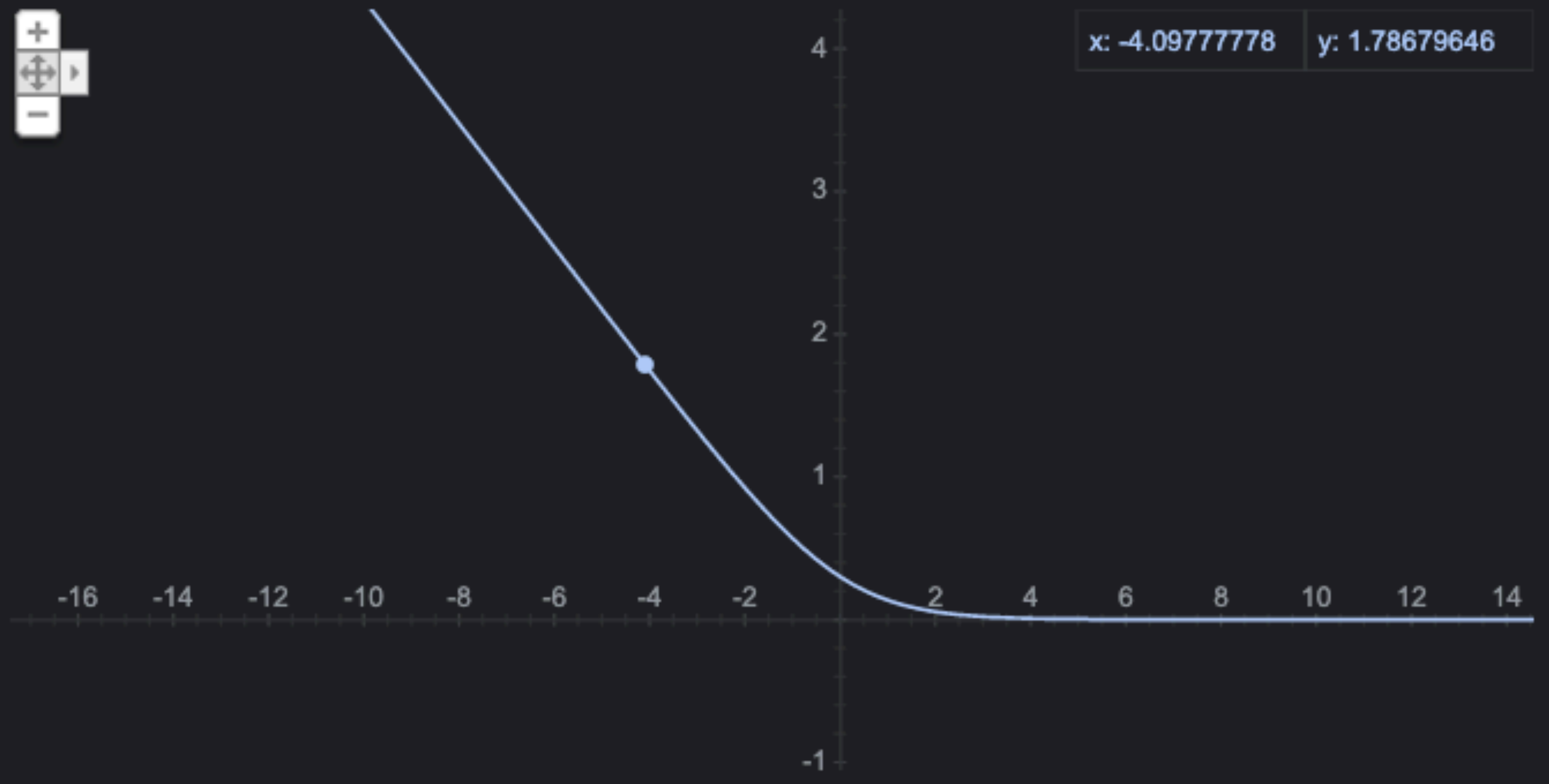
google.com/search?q=log(1%2Bexp(-x))&rlz=1C5CHFA_enIN958IN958&oq=&aqs=chrome.3.69i59i450l8.4295406j0j7&sourceid=chrome&ie=UTF-8

log(1+exp(-x))

All · Shopping · Images · News · Videos · More · Tools

About 7,73,00,00,000 results (0.65 seconds)

Graph for log(1+exp(-x))



More info

https://cran.r-project.org › Rmpfr › log1mexp-note PDF

[Accurately Computing log(1 - exp(.)) – Assessed by Rmpfr]

by M Mächler · Cited by 1 — l1e() computes the relative error of three different ways to compute **log(1 − exp(−a))** for positive a (instead of computing **log(1 − exp(x))** for negative **x**). R> ...
9 pages

https://math.stackexchange.com › questions

[Any simplification of $\log_e{(1+e^x)}? - Math Stack Exchange]

22-Jun-2017 — can you explain me how **log(1+e^x)** becomes **x+log(1+e^(−x))** please? – Sam. Jun 22, 2017 at 7:48. 1.
2 answers · Top answer: You can use the Taylor series of log: log(1+ex)=x+log(1+e−x)=x+e−x…

[Approximate $\log(1-e^x)$ where $x<0 - Math Stack Exchange] 14 Apr 2016

56 / 56