Goal: Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

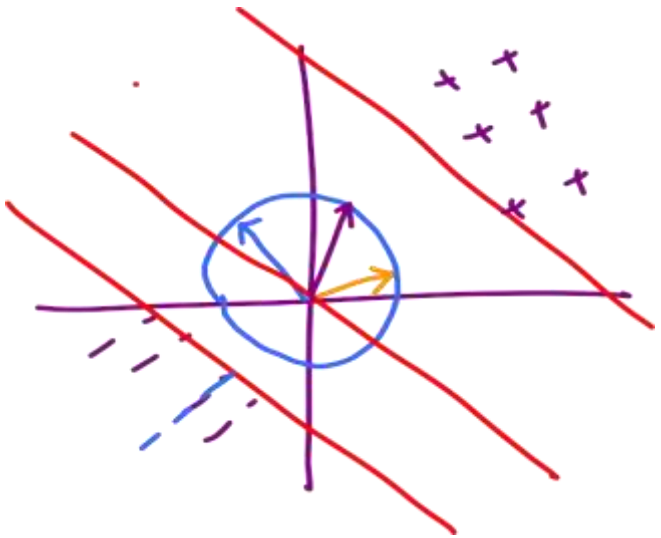
find a w with maximum width s.t all points in dataset D are classified correctly

$$\max_{w, \gamma} \gamma$$

s.t

$$(w^T x_i) y_i \geq \gamma \quad \forall i$$

Issue:
Can scale w arbitrarily.



Possible fix

$$\max_{w,\gamma} \gamma$$

$$(w^T x_i) y_i \geq \gamma$$

$$\|w\|^2 = 1$$

Goal: Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

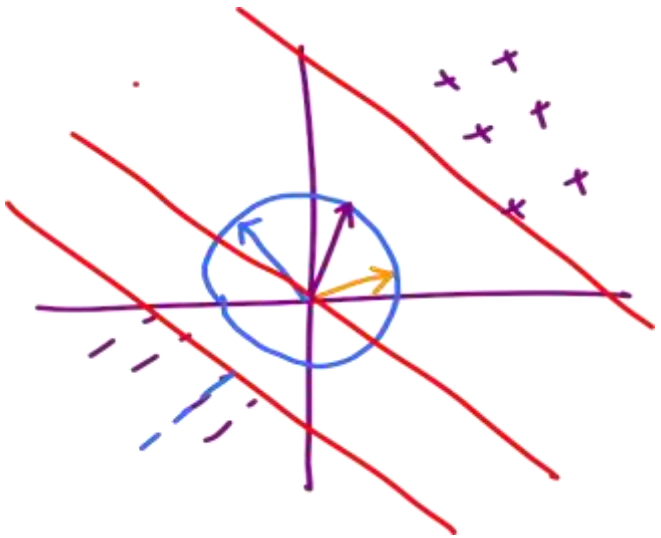
find a w with maximum width s.t all points in dataset D are classified correctly

$$\max_{w, \gamma} \gamma$$

s.t

$$(w^T x_i) y_i \geq \gamma \quad \forall i$$

Issue:
Can scale w arbitrarily.

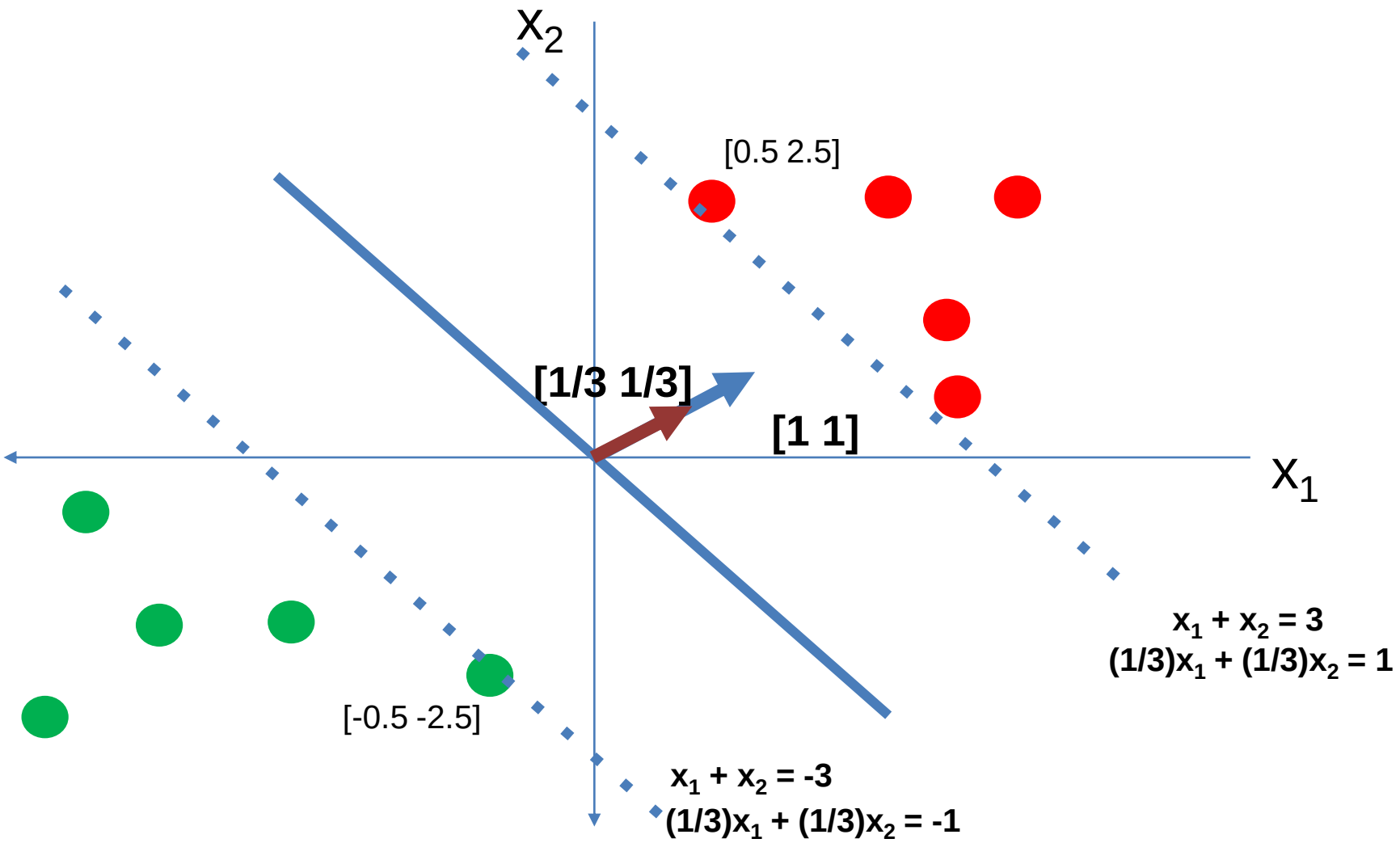


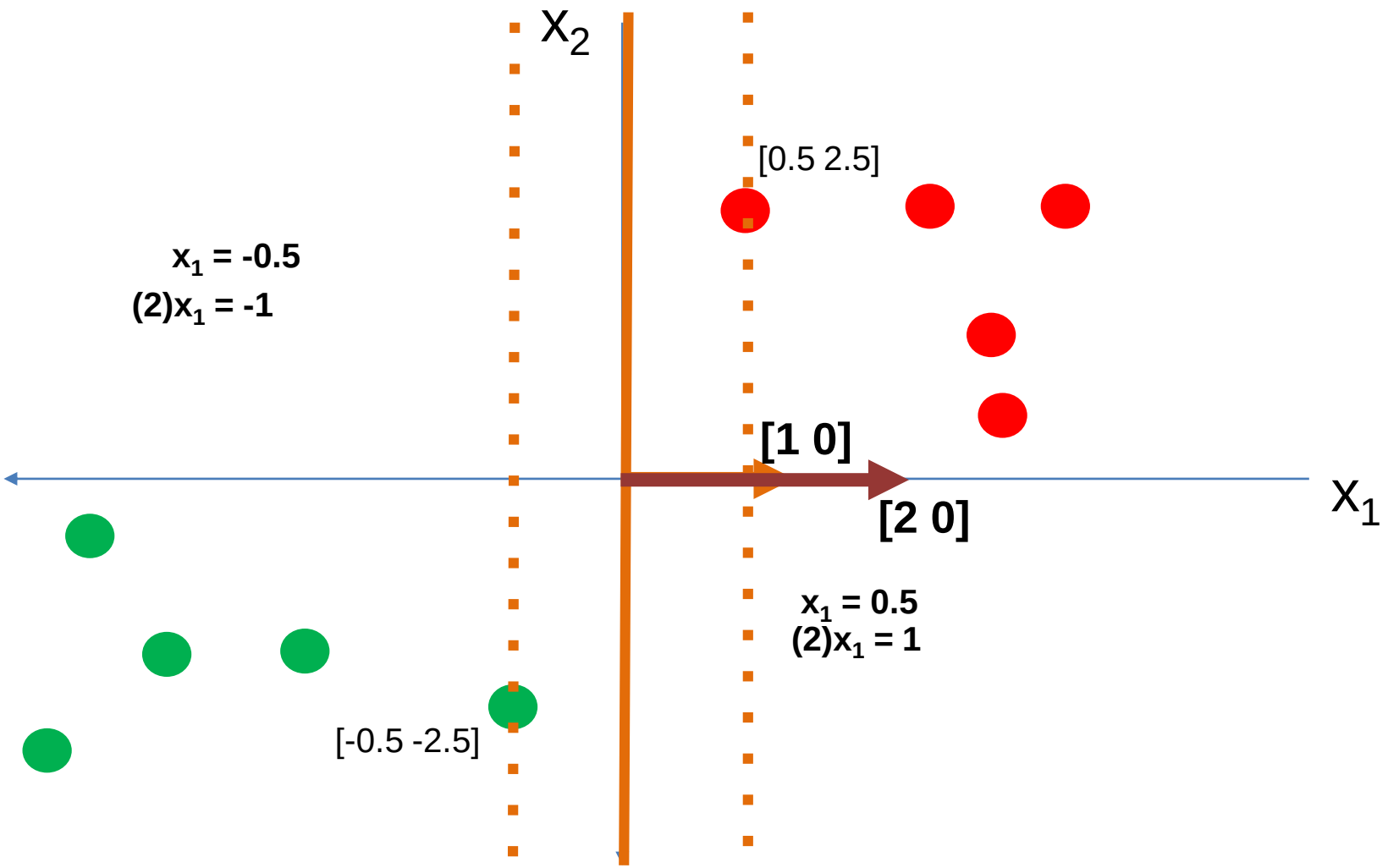
Possible fix

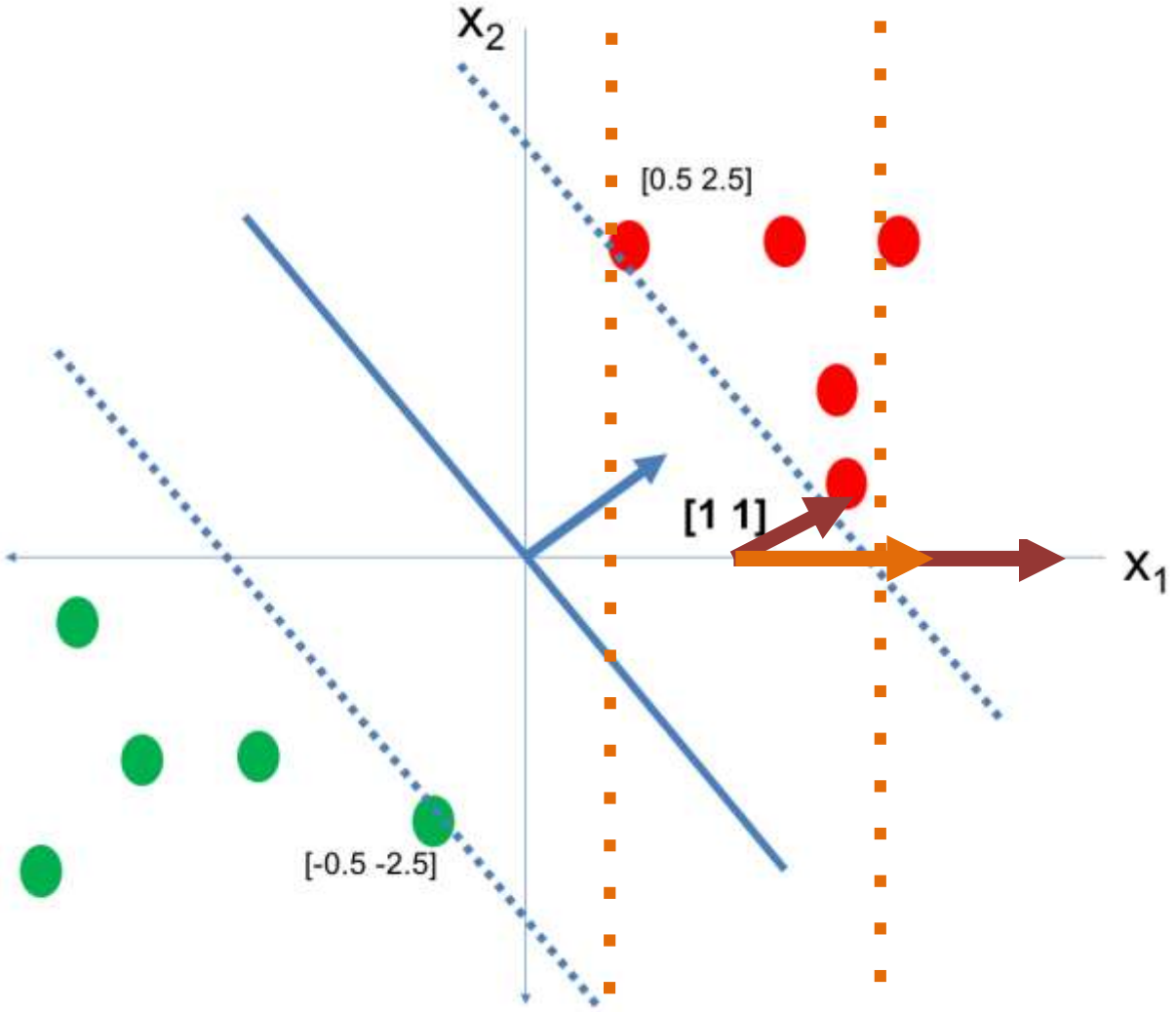
$$\max_{w,\gamma} \gamma$$

$$(w^T x_i) y_i \geq \gamma$$

$$\|w\|^2 = 1$$


x_2
[0.5 2.5]
[1/3 1/3]
[1 1]
x_1
$x_1 + x_2 = 3$
$(1/3)x_1 + (1/3)x_2 = 1$
[-0.5 -2.5]
$x_1 + x_2 = -3$
$(1/3)x_1 + (1/3)x_2 = -1$


X2
[0.5 2.5]
x1 = -0.5
(2)x1 = -1
[1 0]
[2 0]
X1
x1 = 0.5
(2)x1 = 1
[-0.5 -2.5]



Notice what happens to the lengths of the w as we adjust it to have margin 1

OBSERVATIONS

-> Once a direction is fixed,
the width between the margin lines
is fixed

-> If the width is large, then the w that
achieves margin 1 in that direction
has smaller length

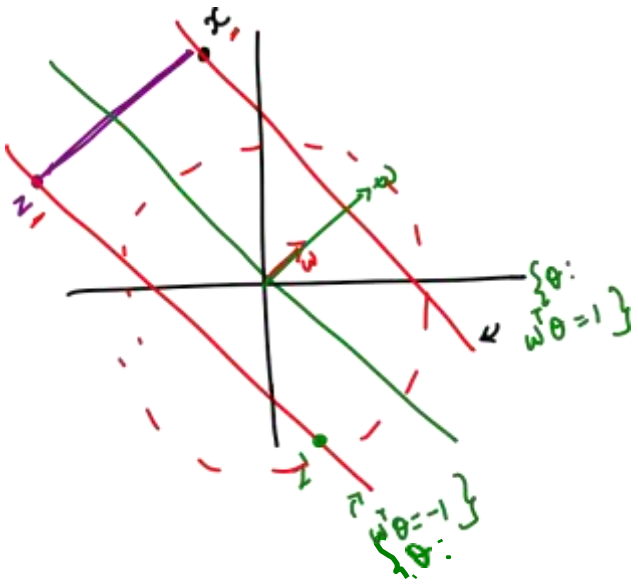
-> If the width is small, then the w that
achieves margin 1 in that direction
has larger length

-> In general, width(w) seems to be
inversely proportional to length(w)

$$\max_{w} \quad \text{width}(w)$$

$$\text{s.t} \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

What is width(w) ?



$$\min_{z} \quad \frac{1}{2}\|x - z\|^2$$

$$\text{s.t} \quad w^T x = 1, \quad w^T z = -1$$

[Exercise]

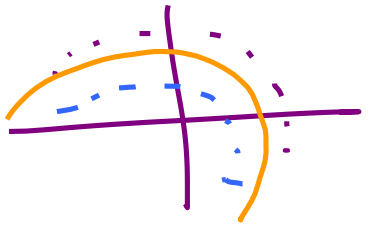
$$\text{width}(w) = \frac{2}{\|w\|^2}$$

$$\max_{w} \quad \frac{2}{\|w\|^2}$$

$$s.t \quad \forall i \quad (w^T x_i) y_i \geq 1$$

$$\min_{w} \quad \frac{1}{2} \|w\|^2$$

$$s.t \quad \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption
- Non-linear structure?

DETOUR

$$\min_{w} \; f(w)$$
$$g(w) \leq 0$$

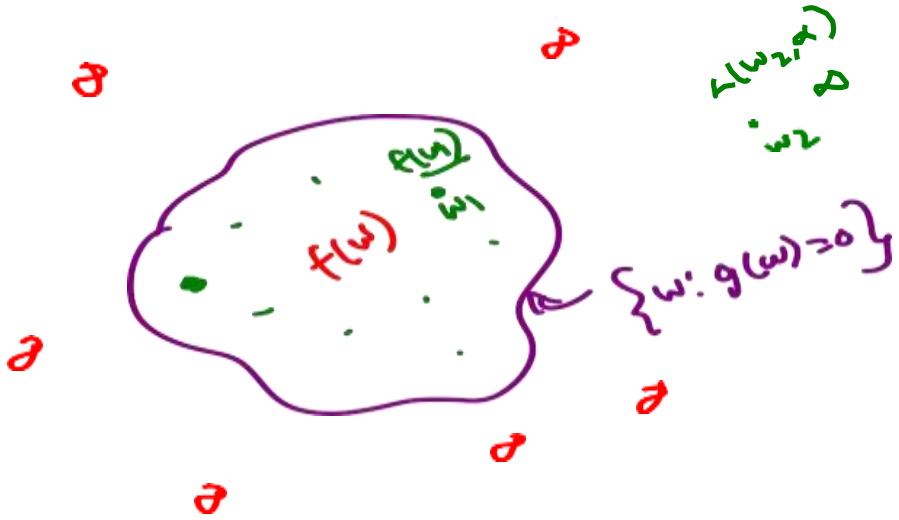
$$\mathcal{L}(w, \alpha) = f(w) + \alpha \cdot g(w)$$

Fix some w.

Consider

$$\max_{\alpha \geq 0} \; L(w, \alpha)$$

$$= \max_{\alpha \geq 0} \; f(w) + \alpha\, g(w)$$



$$\begin{cases} \infty & \text{if } g(w) > 0 \\ f(w) & \text{if } g(w) \leq 0 \end{cases}$$

$$\min_{w} \Big[\underbrace{\max_{\alpha \geq 0} \overset{\theta(w)}{\overline{\mathcal{L}(w,\alpha)}}}_{} \Big] \overset{\text{equivalent}}{\equiv} \left.\begin{array}{l} \min_{w} \; f(w) \\ \text{s.t. } g(w) \leq 0. \end{array}\right\}$$

- Can we swap min and max?

mi
w

Multiple constraints

→ Same idea

$$\min_{w} f(w) \quad \text{s.t} \quad g_i(w) \le 0 \quad \forall i = 1 \dots k$$

$$= \min_{w} \left[\max_{\{\alpha_1, \dots, \alpha_k \ge 0\}} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) \right] \right]$$

$$\Longleftarrow \text{Strong duality for convex } f, g_i$$

$$\max_{\alpha_1, \dots, \alpha_k \ge 0} \min_{w} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\min_{w} \quad \underbrace{\frac{1}{2}\|w\|^2}_{f(w)}$$

$$\text{s.t} \quad \underbrace{(w^T x_i) y_i \geq 1}_{1-(w^T x_i) y_i \leq 0} \quad \forall i$$

$$g_i(w) = 1 - (w^T x_i) y_i$$

$$\mathcal{L}(w, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

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

$$\min_{w} \left[\max_{\alpha \geq 0} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

where $\alpha \geq 0$ means $\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0$

$$\Downarrow\Downarrow\Downarrow$$

$$\max_{\alpha \geq 0} \left[\min_{w} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Fix $\alpha \geq 0$

$$\min_{w} \left[\frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - w^T x_i y_i\right) \right]$$

Grad w.r.t w

$$w^* + \sum_{i=1}^{n} -\alpha_i x_i y_i = 0$$

$$w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

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

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

α_i: Fixed Choice

In matrix notation

$$w^* = X Y \alpha$$

$$X = \begin{bmatrix} 1 & 1 & & 1 \\ x_1 & x_2 & \cdots & x_n \\ 1 & 1 & & 1 \end{bmatrix}_{d \times n} \quad Y = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{bmatrix}_{n \times n} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

Substituting $\overset{\text{Soln}}{=}$ back in the objective.

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - w^T x_i) y_i)$$

$$= \frac{1}{2} \underline{w^T w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i (w^T x_i) y_i$$

$$\frac{1}{2}(xy\alpha)^T(xy\alpha) + \alpha^T 1 - \sum_{i=1}^{n}(xy\alpha)^T x_i\, y_i\, \alpha_i$$

$\alpha^T 1 = \begin{bmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{bmatrix}\begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix}$

On Simplification [please do this]

$$\alpha^T 1 - \frac{1}{2}(xy\alpha)^T(xy\alpha)$$

DUAL PROBLEM

$$\max_{\alpha \geq 0} \quad \alpha^T 1 - \frac{1}{2}\alpha^T y^T x^T x\, y\, \alpha$$

$x^T x$ → n x n.

Can be KERNELIZED!

Solving in α instead of d.

$\alpha \geq 0$ ← easy constraints

Revisiting The Lagrangian

$$\min_{w} \left[\max_{\alpha \geq 0} f(w) + \alpha g(w) \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} f(w) + \alpha g(w) \right]$$

PRIMAL DUAL

w^* α^*

$$\max_{\alpha \geq 0} f(w^*) + \alpha g(w^*) = \min_{w} f(w) + \alpha^* g(w)$$

$$f(w^*) = f(w) + \alpha^* g(w')$$

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \alpha^* g(w^*) \geq 0$$

But we know $\alpha^* g(w^*) \leq 0$

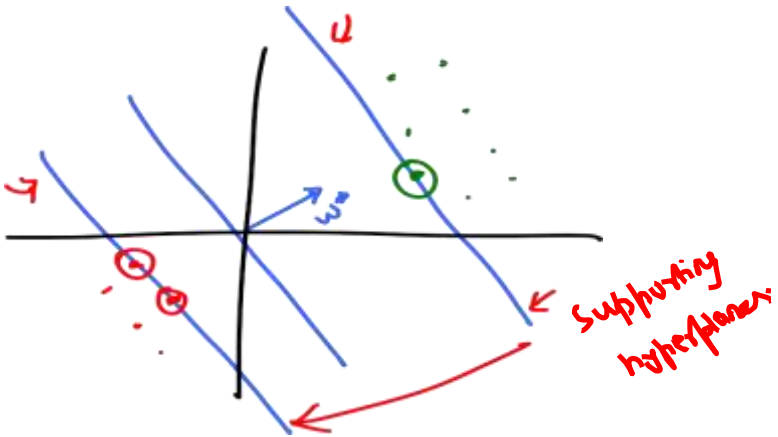
$$\Rightarrow \alpha^* g(w^*) = 0 \rightarrow$$ COMPLEMENTARY SLACKNESS

For multiple constraints

$\alpha_i^* \, g_i(\omega^*) = 0 \quad \forall i$

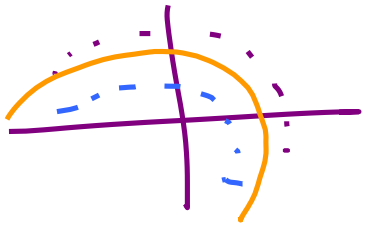
For our problem

$$\alpha_i^* \left(1 - (w^T x_i) y_i\right) = 0 \quad \forall i$$



$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

- Non-linear structure?

So far

Support Vector Machines
Primal Problem – Margin Maximization
Dual Problem
- Kernel Version

Now

- What if there are **outliers** in the problem?

Idea (to deal with outliers):

Fix any w. w classifies some points correct and some incorrectly. Let the incorrect points pay "bribe" to get to the correct side.

Modified formulation

$C > 0$ [hyper parameter]

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

$$\rightarrow \quad (w^T x_i)\, y_i + \xi_i \geq 1 \quad \leftarrow \forall i$$

$$\rightarrow \quad \xi_i \geq 0 \quad \leftarrow \forall i$$

If $C = 0$ $\Rightarrow$ Bribes don't cost $\Rightarrow$ $w = 0$ is solution

$C \rightarrow \infty$ $\Rightarrow$ Bribes are too costly $\Rightarrow$ Linear separable case.

$$L(w, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i - \xi_i\right) + \sum_{i=1}^{n} \beta_i (-\xi_i)$$

Dual:

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0}} \min_{w} L(w, \xi, \alpha, \beta)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

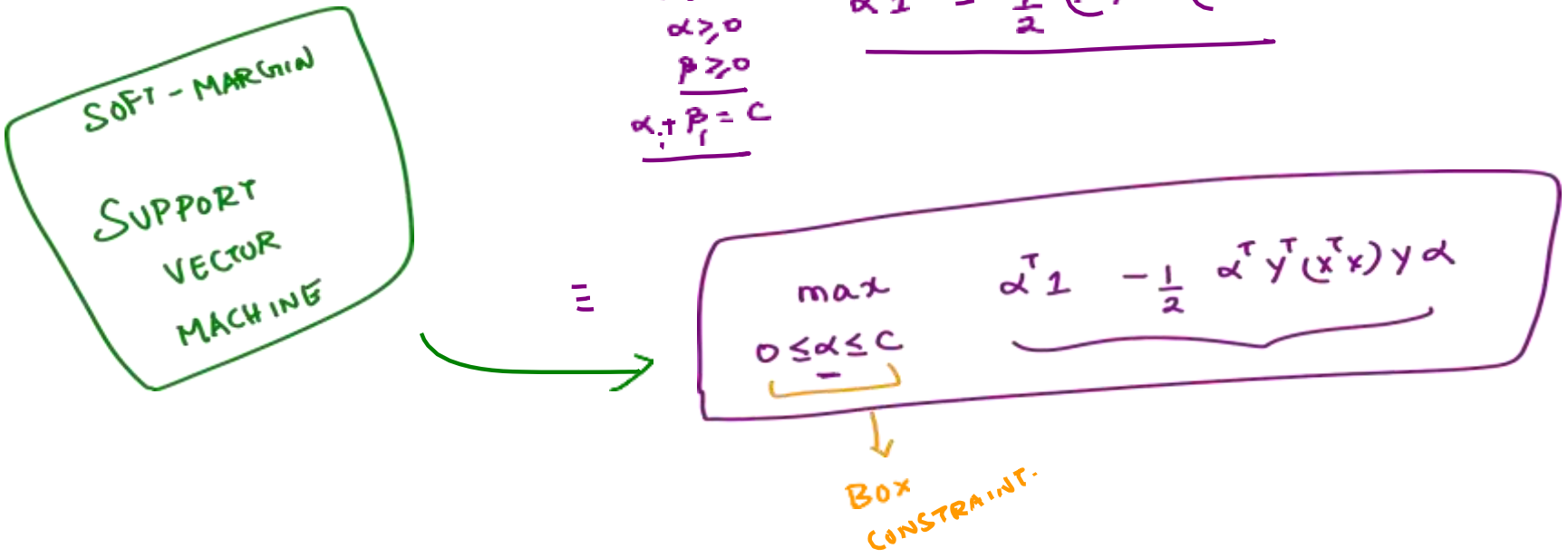
$$w^* = X Y \alpha$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \boxed{C - \alpha_i - \beta_i = 0}$$

$$\boxed{\alpha_i + \beta_i = C} \quad \forall i$$

Substitute $w = X Y \alpha$ in the original objective

$$\frac{1}{2}(XY\alpha)^T(XY\alpha) + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\xi_i + \alpha^T 1 - (XY\alpha)^T(XY\alpha)$$


SOFT - MARGIN
SUPPORT VECTOR MACHINE
max
$\alpha \geq 0$
$\beta \geq 0$
$\alpha_i + \beta_i = C$
$\alpha^T 1 - \frac{1}{2} (Xy\alpha)^T (Xy\alpha)$
$\equiv$
max
$0 \leq \alpha \leq C$
$\alpha^T 1 - \frac{1}{2} \alpha^T y^T (x^T x) y \alpha$
BOX CONSTRAINT.

HARD-MARGIN SVM

PRIMAL

$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$s.t \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

$1 - w^T x_i y_i \leq 0$

DUAL

$$\max_{\alpha \geq 0} \quad \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$$\alpha_i^* (1 - w^{*T} x_i y_i) = 0 \quad \forall i$$

$$w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i$$

SOFT-MARGIN SVM

PRIMAL ✓

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

$$s.t \quad \begin{cases} (w^T x_i) y_i + \xi_i \geq 1 & \forall i = 1, \dots n \\ \xi_i \geq 0 & \forall i = 1, \dots n \end{cases}$$

α, β

DUAL ✓

$$\max_{\substack{\alpha + \beta = C \\ \alpha \geq 0 \\ \beta \geq 0}} \quad \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$0 \leq \alpha \leq C$

- Let $(\underline{w}^*, \underline{\xi}^*)$ be the primal optimal solution
- Let $(\underline{\alpha}^*, \underline{\beta}^*)$ be the dual optimal solution

COMPLEMENTARY SLACKNESS

$$\alpha_i^* \left(1 - (w^{*T} x_i) y_i - \xi_i^* \right) = 0 \quad \forall i$$

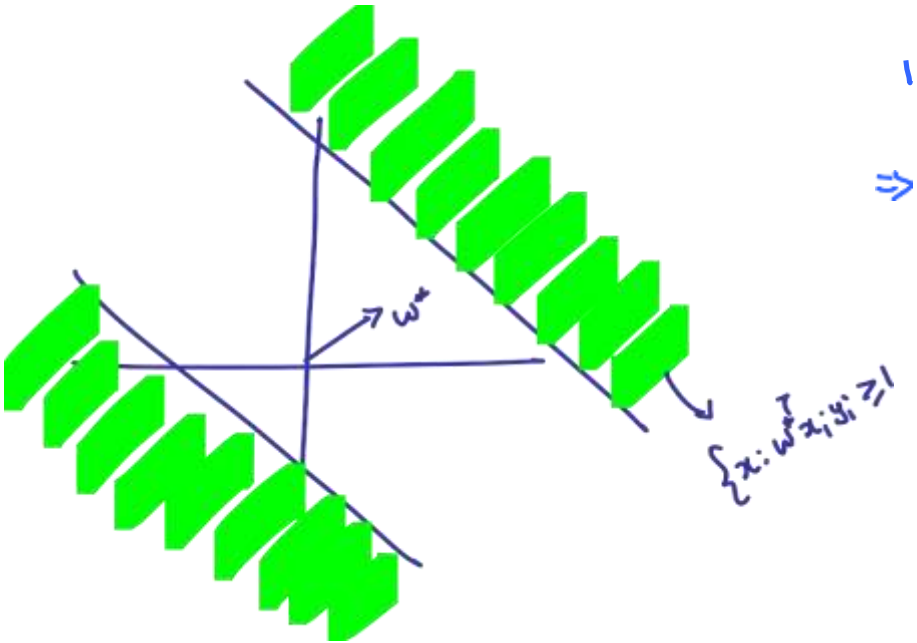
$$\beta_i^* \xi_i^* = 0 \quad \forall i$$

$\alpha_i^* + \beta_i^* = C \quad \forall i$

↳ (A)

Various cases possible

Case 1: $\alpha_i^* = 0$



$\overset{(A)}{\Rightarrow} \quad \beta_i^* = C \quad \overset{CS}{\Rightarrow} \quad \xi_i^* = 0$

$1 - (w^{*T} x_i) y_i - \underbrace{\xi_i^*}_{=0} \leq 0$ [Primal feasibility]

$\Rightarrow \quad 1 - (w^{*T} x_i) y_i \leq 0$

$\Rightarrow \quad w^{*T} x_i y_i \geq 1$

$\Rightarrow$ w^* classifies (x_i, y_i) correctly.

Case 2: $0 < \alpha_i^* < C$ $\overset{\text{KKT}}{\Rightarrow}$ $0 < \beta_i^* < C$ $\overset{\text{CS}}{\Rightarrow}$ $\xi_i^* = 0$

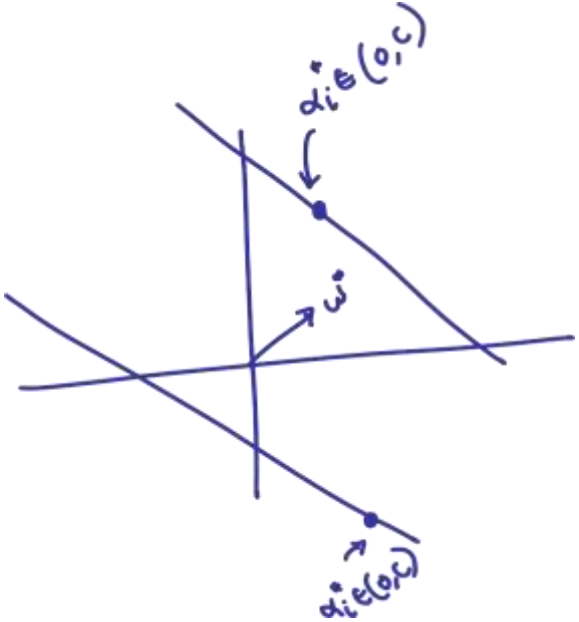
$\Downarrow$ CS

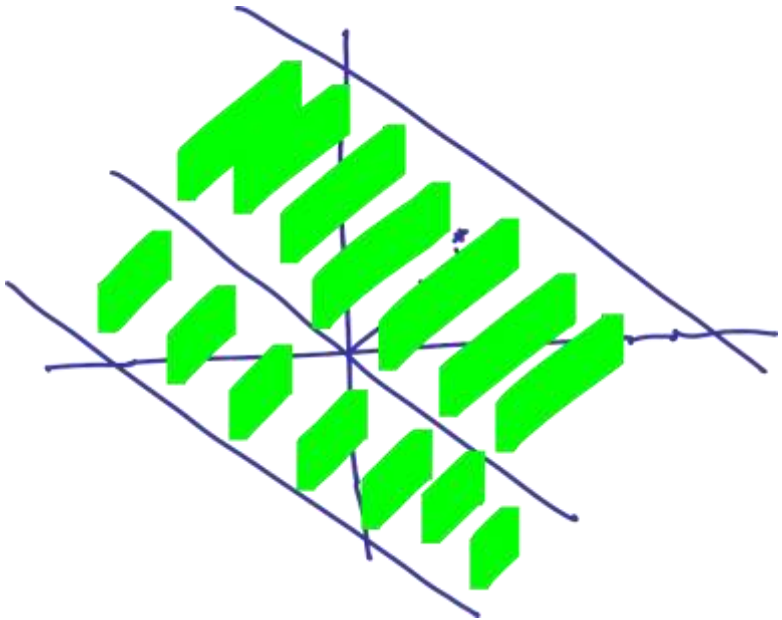
$1 - (w^{*T} x_i) y_i - \xi_i^* = 0$

$\Downarrow$

$(w^{*T} x_i) y_i = 1$

$\Rightarrow$ (x_i, y_i) lies on the supporting hyperplane.





Case 3: $\alpha_i^* = C \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$

$\Downarrow$ CS

$1 - w^{*T} x_i y_i - \xi_i^* = 0$

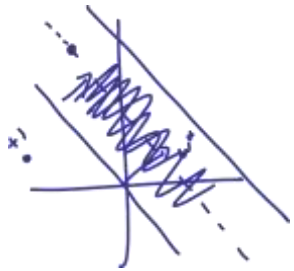
$\xi_i^* = 1 - w^{*T} x_i y_i \geq 0$

$\Rightarrow \boxed{w^{*T} x_i y_i \leq 1}$

Let's see this from P.O.V of data

CASE 1

$w^{*T} x_i y_i < 1$



$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0$$

$$w^{*T} x_i y_i \geq 1 - \xi_i^*$$

$$\xi_i^* \geq 1 - w^{*T} x_i y_i$$

$\Rightarrow \xi_i^* > 0 \Rightarrow \beta_i^* = 0 \Rightarrow \alpha_i^* = C$

$$\alpha_i^* (1 - w^{*T} x_i y_i - \xi_i^*) = 0$$

$$\beta_i^* \xi_i^* = 0$$

CASE 2: $w^{*T} x_i y_i = 1$

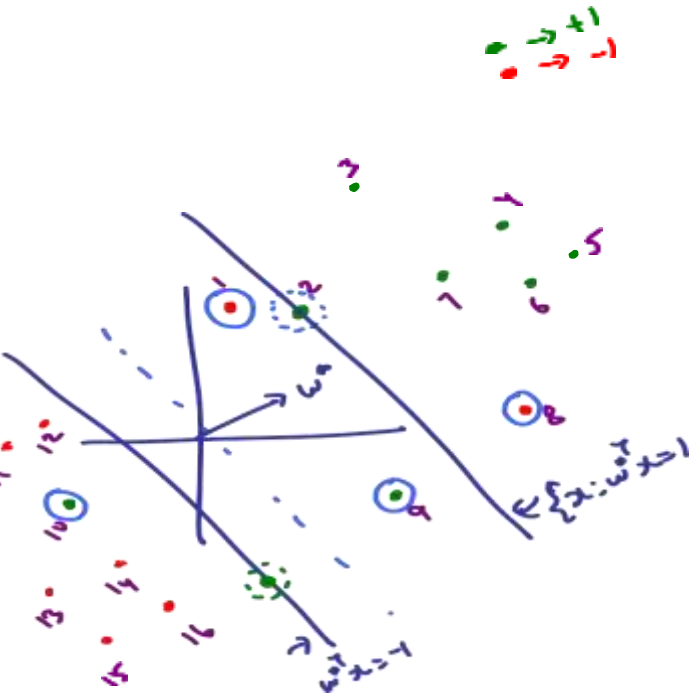
$$\xi_i^* \geq 1 - w^{*T} x_i y_i$$

$$\Rightarrow \quad \xi_i^* \geq 0 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$$

CASE 3 $w^{*T} x_i y_i > 1$

$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0 \quad \text{[Primal feasibility]}$$

$$\Rightarrow \quad 1 - w^{*T} x_i y_i - \xi_i^* < 0 \quad \overset{\text{C.S}}{\Rightarrow} \quad \alpha_i^* = 0$$



SUMMARY

$\alpha_i^* = 0 \quad \Rightarrow \quad w^{*T} x_i y_i \geq 1$

$0 < \alpha_i^* < C \quad \Rightarrow \quad w^{*T} x_i y_i = 1$

$\alpha_i^* = C \quad \Rightarrow \quad w^{*T} x_i y_i \leq 1$

✓ $w^{*T} x_i y_i < 1 \quad \Rightarrow \quad \alpha_i^* = C$

→ $w^{*T} x_i y_i = 1 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$

→ $w^{*T} x_i y_i > 1 \quad \Rightarrow \quad \alpha_i^* = 0.$

Binary classification

- GENERATIVE
  - Naive Bayes
  - G.D.A
- DISCRIMINATIVE
  - k-NN
  - Decision trees
  - Perceptron
  - Support-vector-machines

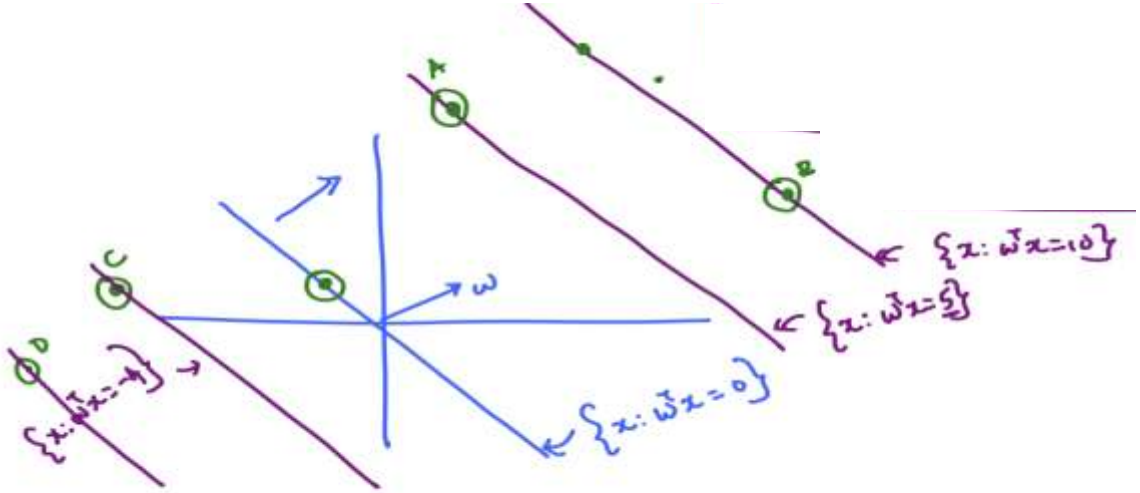
Doesn't "really" model $P(y|x)$

→ Just finds $f: \mathbb{R}^d \to \{\pm 1\}$

- Can we model $P(y=+1/x)$ differently?

Start with a simple model

Given $x \in \mathbb{R}^d$ $\quad z = w^T x \quad$ $w \in \mathbb{R}^d$.



$$P(y=+1 \mid x) = g(w^T x)$$

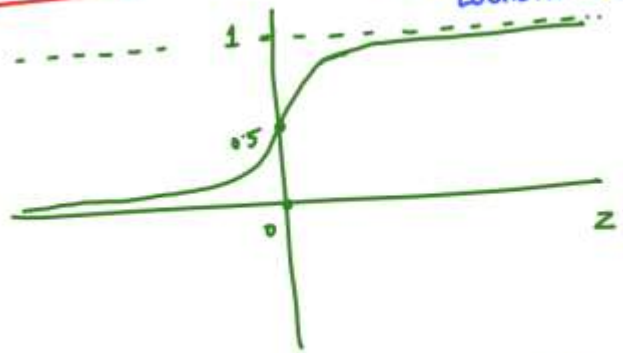
- $g(z) \in [0,1]$
- $g(z) \to 1$ as $z \to \infty$
- $g(z) \to 0$ as $z \to -\infty$
- $g(z) = 0.5$ if $z = 0$.

LINK FUNCTION

ONE POPULAR CHOICE

SIGMOID FUNCTION
LOGISTIC FUNCTION.

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



MODEL: LOGISTIC REGRESSION

Data: $\{(x_1, y_1) \cdots (x_n, y_n)\}$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$

Max. Likelihood

$$L(w, \text{Data}) = \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{y_i} \left(1 - g(w^T x_i)\right)^{(1-y_i)}$$

$$\log L(w, \text{Data}) = \sum_{i=1}^{n} y_i \log\left(g(w^T x_i)\right) + (1-y_i) \log\left(1 - g(w^T x_i)\right)$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-w^T x_i}} \right) + (1-y_i) \log\left(\frac{e^{-w^T x_i} \times 1}{1+e^{-w^T x_i}} \right) \right]$$

$$= \sum_{i=1}^{n} \left[\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - y_i(-w^T x_i) \right]$$

$$= \sum_{i=1}^{n} \left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right) \right]$$

- No closed form solution

- Gradient ascent

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

$$= \sum_{i=1}^{n} x_i \left(y_i - \left(1 - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) \right)$$

$$= \sum_{i=1}^{n} x_i \left(y_i - \frac{1}{1+e^{-w^T x_i}} \right)$$

($x_i \in \mathbb{R}^d$, $y_i \in \{0,1\}$, $g(w^T x_i)$)

$$w_{t+1} = w_t + \eta_t \nabla \log L(w_t)$$

REGULARIZED VERSION

$$\min_w \sum_{i=1}^{n} (1-y_i) w^T x_i + \log\left(1+e^{-w^T x_i}\right) + \frac{\lambda}{2}\|w\|^2$$

KERNEL VERSION

- Can argue $w = \sum_{i=1}^{n} \alpha_i x_i$

[Formal Theorem

Representer Theorem]

Exercise: Derive the kernel version of logistic regression

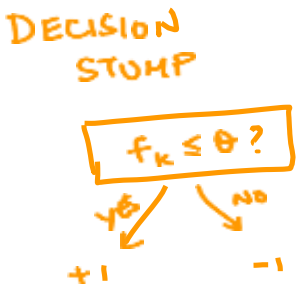
META CLASSIFIERS (OR)

ENSEMBLE CLASSIFIERS.

WEAK CLASSIFIERS [better than random]

STRONG CLASSIFIERS

Weak classifiers



high bias, low variance

Overfit decision tree

…

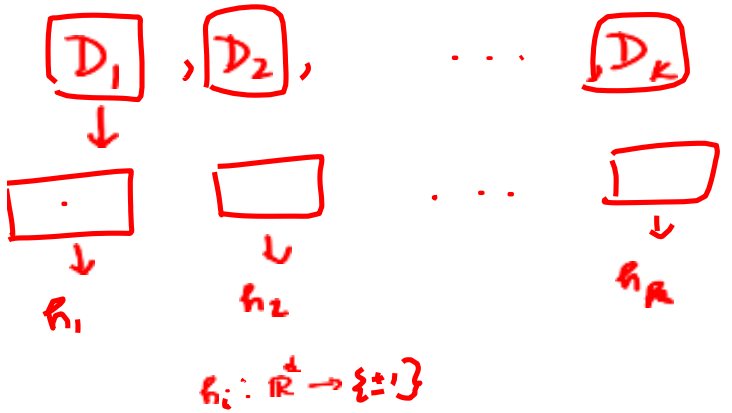
……

low bias, high variance

$x_1, x_2, \ldots x_n \sim N(\mu, 1)$

$\hat{\mu}_1 = x_1 \qquad \hat{\mu}_2 = x_2 \qquad \ldots \qquad \hat{\mu}_n = x_n \qquad \hat{\mu}_{ML} = \frac{1}{n}\sum x_i$

Overfit decision trees



$h^*(x) = \text{majority}(h_1(x), \ldots, h_K(x))$

BAGGING - Bootstrap Aggregation

$D = \{(x_1, y_1) \cdots (x_n, y_n)\}$

- Create datasets $D_1, \cdots, D_k$ from D by
  - Sampling with replacement.
- Run weak classifier on $D_1, \cdots, D_k$ to get $h_1, \cdots, h_k$
- Aggregate $h_1, \cdots, h_k$ using majority.

Chance that a point appears in a dataset

$1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{1}{n}\right)$

$1 - \left(1 - \frac{1}{n}\right)^n$

$1 - \frac{1}{e}$ (as $n \to \infty$)

$\simeq 66\%$

FEATURE BAGGING -> Bag the features in addition to data points

Feature bagged decision trees -> RANDOM FOREST

BOOTSTRAP - Sampling with Replacement ?

AGGREGATION - Majority. ?

BOOSTING

[Freund & Schapire. 1995 Godel Prize]

↑

ADA-BOOST

Distribution D over (x, y) — $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$

↳ unknown but fixed.

$x_1, \ldots, x_n$ are iid from D.

$h : \mathbb{R}^d \to \{\pm 1\}$

x $\quad$ y.

Measure performance using

$$P_{(x,y)\sim D}\left(h(x) \neq y\right)$$

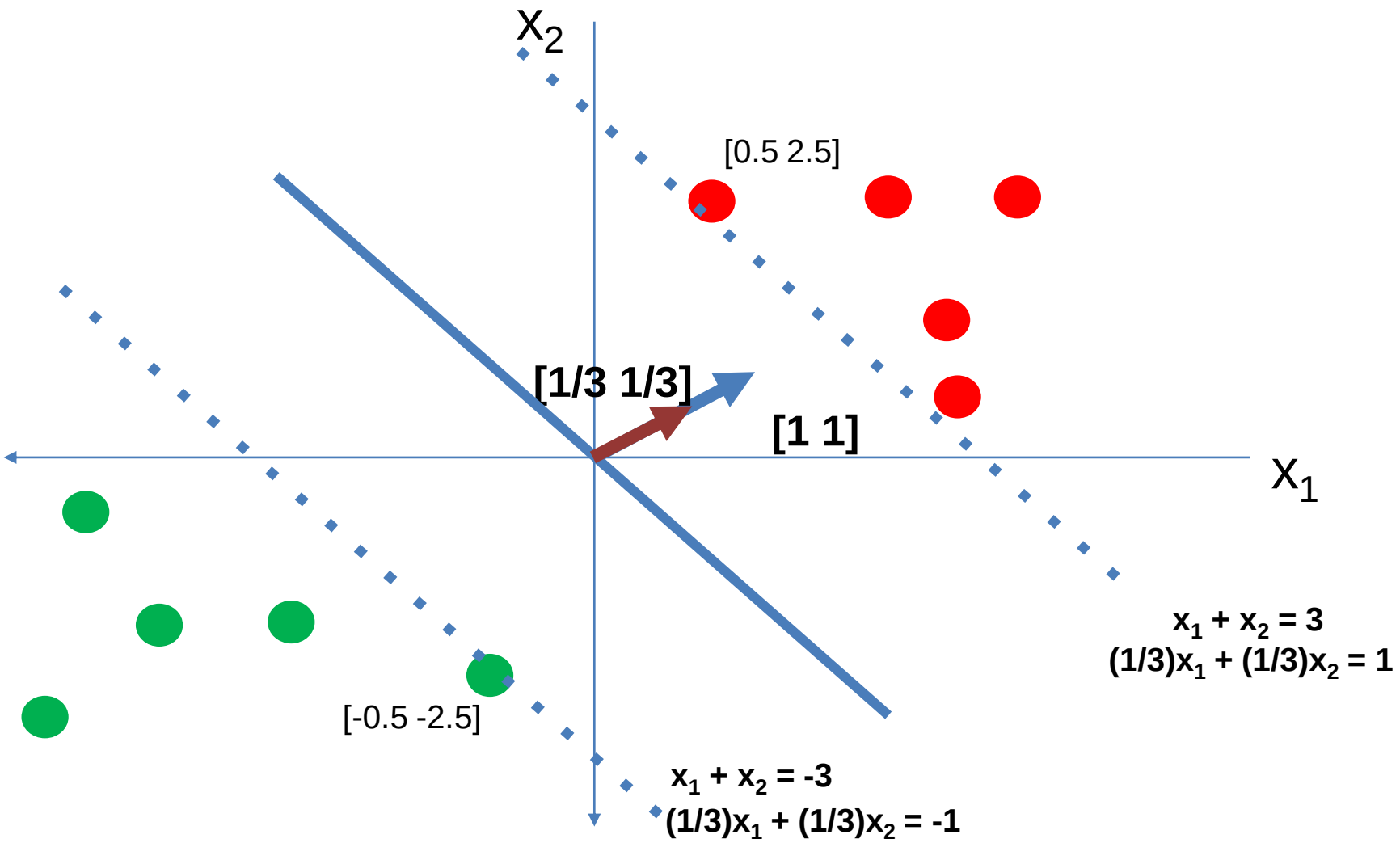
Misclassification probability.

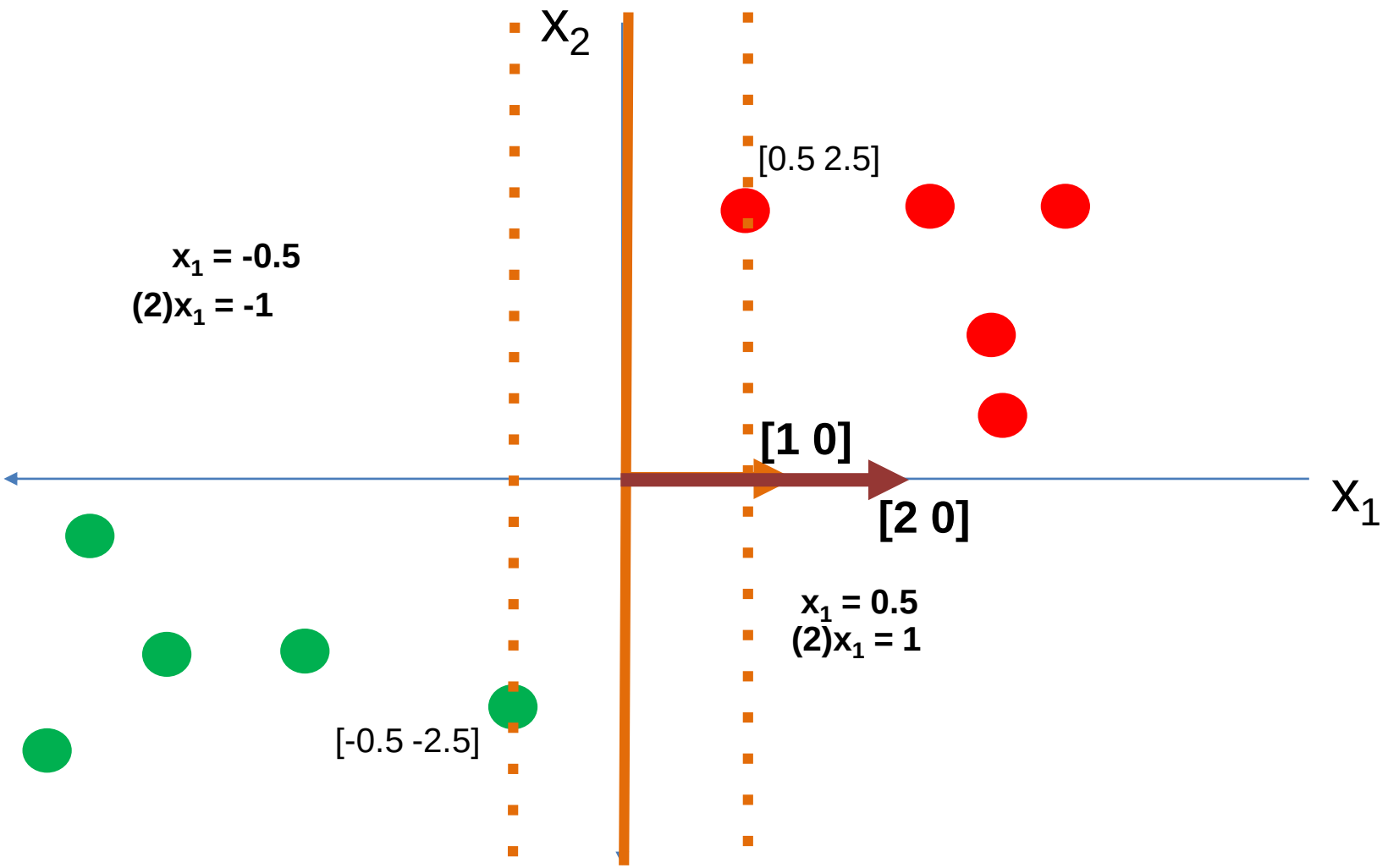
A weak / Strong learner is one which outputs a classifier h for which

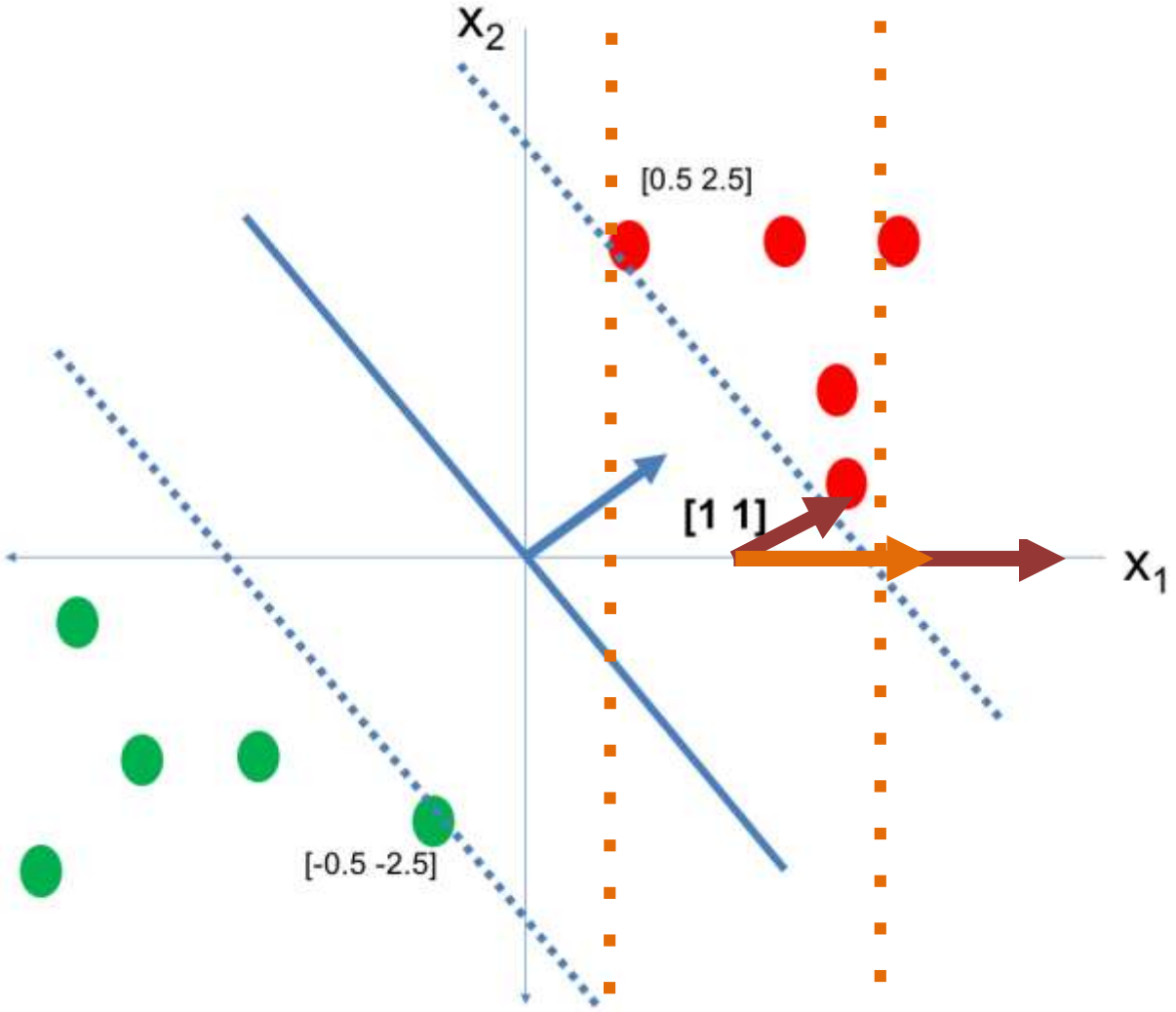
$$P_{x,y\sim D}\left(h(x) = y\right) \geq \frac{1}{2} + \frac{1-\epsilon}{2}$$

$\gamma > 0$

for any unknown but fixed distribution D.


x_2
[0.5 2.5]
[1/3 1/3]
[1 1]
x_1
$x_1 + x_2 = 3$
$(1/3)x_1 + (1/3)x_2 = 1$
[-0.5 -2.5]
$x_1 + x_2 = -3$
$(1/3)x_1 + (1/3)x_2 = -1$


X2
[0.5 2.5]
x1 = -0.5
(2)x1 = -1
[1 0]
[2 0]
X1
x1 = 0.5
(2)x1 = 1
[-0.5 -2.5]



Notice what happens to the lengths of the w as we adjust it to have margin 1

OBSERVATIONS

-> Once a direction is fixed,
the width between the margin lines
is fixed

-> If the width is large, then the w that
achieves margin 1 in that direction
has smaller length

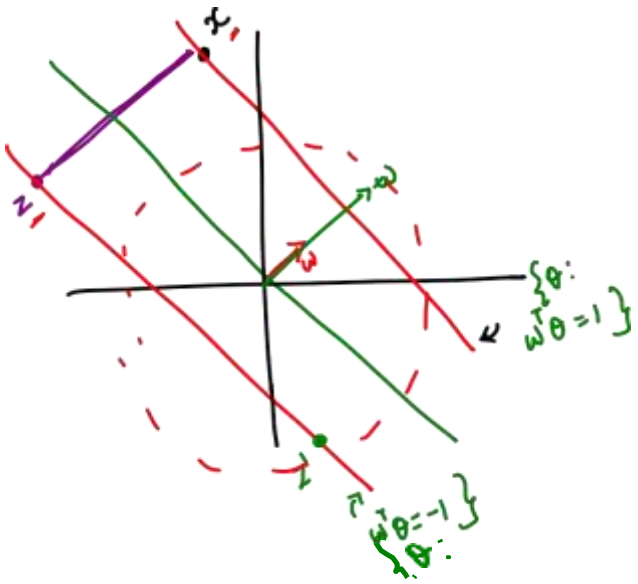
-> If the width is small, then the w that
achieves margin 1 in that direction
has larger length

-> In general, width(w) seems to be
inversely proportional to length(w)

$$\max_{w} \quad \text{width}(w)$$

$$\text{s.t} \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

What is width(w) ?



$$\min_{z} \quad \frac{1}{2}\|x - z\|^2$$

$$\text{s.t} \quad w^T x = 1, \quad w^T z = -1$$

[Exercise]

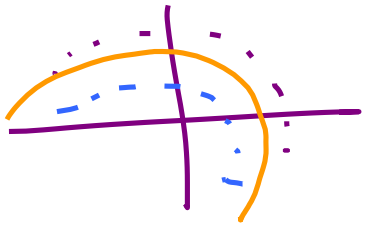
$$\text{width}(w) = \frac{2}{\|w\|^2}$$

$$\max_{w} \quad \frac{2}{\|w\|^2}$$

$$s.t \quad \forall i \quad (w^T x_i) y_i \geq 1$$

$$\min_{w} \quad \frac{1}{2} \|w\|^2$$

$$s.t \quad \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption
- Non-linear structure?

DETOUR

$$\min_{w} \; f(w)$$
$$g(w) \leq 0$$

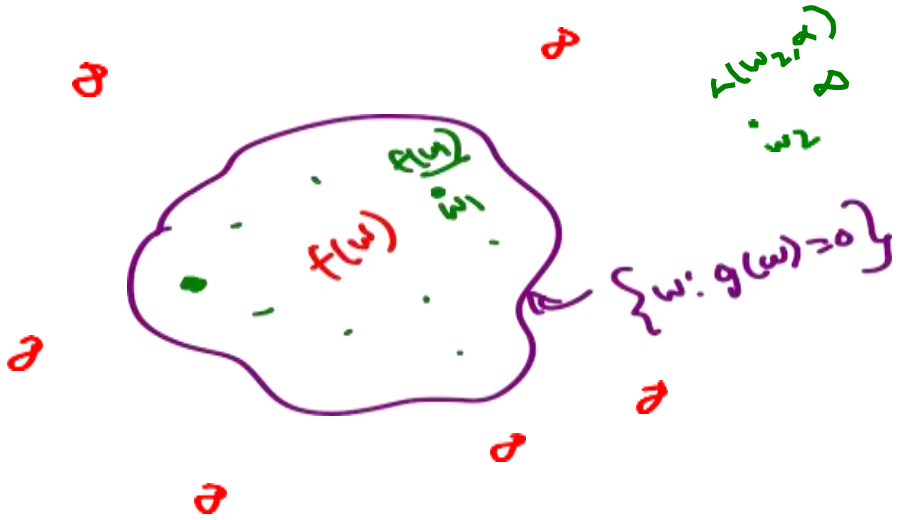
$$\mathcal{L}(w, \alpha) = f(w) + \alpha \cdot g(w)$$

Fix some w.

Consider

$$\max_{\alpha \geq 0} \; L(w, \alpha)$$

$$= \max_{\alpha \geq 0} \; f(w) + \alpha\, g(w)$$



$$\begin{cases} \infty & \text{if } g(w) > 0 \\ f(w) & \text{if } g(w) \leq 0 \end{cases}$$

$$\min_{w} \underbrace{\left[\max_{\alpha \geq 0} \underbrace{\mathcal{L}(w, \alpha)}_{B(w)} \right]}$$

$\equiv$ equivalent

$$\min_{w} \; f(w)$$

$$\text{s.t} \; \; g(w) \leq 0.$$

- Can we swap min and max?

mi
w

Multiple constraints

→ Same idea

$$\min_{w} f(w) \quad \text{s.t} \quad g_i(w) \le 0 \quad \forall i = 1 \dots k$$

$$= \min_{w} \left[\max_{\{\alpha_1, \dots, \alpha_k \ge 0\}} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) \right] \right]$$

$$\Longleftarrow \text{Strong duality for convex } f, g_i$$

$$\max_{\alpha_1, \dots, \alpha_k \ge 0} \min_{w} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\min_{w} \quad \underbrace{\frac{1}{2}\|w\|^2}_{f(w)}$$

$$\text{s.t} \quad \underbrace{(w^T x_i) y_i \geq 1}_{1-(w^T x_i) y_i \leq 0} \quad \forall i$$

$$g_i(w) = 1 - (w^T x_i) y_i$$

$$\mathcal{L}(w, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

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

$$\min_{w} \left[\max_{\alpha \geq 0} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

where $\alpha \geq 0$ means $\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0$

$$\Downarrow\Downarrow\Downarrow$$

$$\max_{\alpha \geq 0} \left[\min_{w} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Fix $\alpha \geq 0$

$$\min_{w} \left[\frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - w^T x_i y_i\right) \right]$$

Grad w.r.t w

$$w^* + \sum_{i=1}^{n} -\alpha_i x_i y_i = 0$$

$$w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

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

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

α_i: Fixed Choice

In matrix notation

$$w^* = X Y \alpha$$

$$X = \begin{bmatrix} 1 & 1 & & 1 \\ x_1 & x_2 & \cdots & x_n \\ 1 & 1 & & 1 \end{bmatrix}_{d \times n} \quad Y = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{bmatrix}_{n \times n} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

Substituting $\overset{\text{Soln}}{=}$ back in the objective.

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - w^T x_i) y_i)$$

$$= \frac{1}{2} \underline{w^T w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i (w^T x_i) y_i$$

$$\frac{1}{2}(xy\alpha)^T(xy\alpha) + \alpha^T 1 - \sum_{i=1}^{n}(xy\alpha)^T x_i\, y_i\, \alpha_i$$

$\alpha^T 1 = \begin{bmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{bmatrix}\begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix}$

On Simplification [please do this]

$$\alpha^T 1 - \frac{1}{2}(xy\alpha)^T(xy\alpha)$$

DUAL PROBLEM

$$\max_{\alpha \geq 0} \quad \alpha^T 1 - \frac{1}{2}\alpha^T y^T x^T x\, y\, \alpha$$

$x^T x$ → n x n.

Can be KERNELIZED!

Solving in α instead of d.

$\alpha \geq 0$ ← easy constraints

Revisiting The Lagrangian

$$\min_{w} \left[\max_{\alpha \geq 0} f(w) + \alpha g(w) \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} f(w) + \alpha g(w) \right]$$

PRIMAL DUAL

w^* α^*

$$\max_{\alpha \geq 0} f(w^*) + \alpha g(w^*) = \min_{w} f(w) + \alpha^* g(w)$$

$$f(w^*) = f(w) + \alpha^* g(w')$$

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \alpha^* g(w^*) \geq 0$$

But we know $\alpha^* g(w^*) \leq 0$

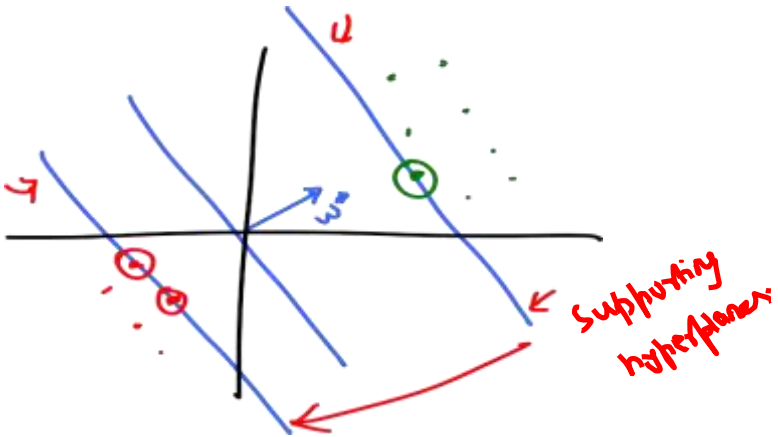
$$\Rightarrow \alpha^* g(w^*) = 0 \rightarrow$$ COMPLEMENTARY SLACKNESS

For multiple constraints

$\alpha_i^* \, g_i(\omega^*) = 0 \quad \forall i$

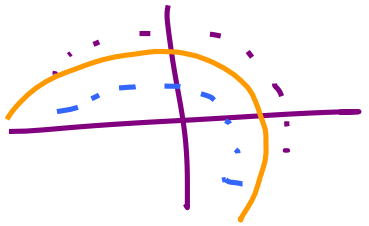
For our problem

$$\alpha_i^* \left(1 - (w^T x_i) y_i\right) = 0 \quad \forall i$$



$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

- Non-linear structure?

So far

Support Vector Machines
Primal Problem – Margin Maximization
Dual Problem
- Kernel Version

Now

- What if there are **outliers** in the problem?

Idea (to deal with outliers):

Fix any w. w classifies some points correct and some incorrectly. Let the incorrect points pay "bribe" to get to the correct side.

Modified formulation

$C > 0$ [hyper parameter]

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

$$\rightarrow \quad (w^T x_i)\, y_i + \xi_i \geq 1 \quad \leftarrow \forall i$$

$$\rightarrow \quad \xi_i \geq 0 \quad \leftarrow \forall i$$

If $C = 0$ $\Rightarrow$ Bribes don't cost $\Rightarrow$ $w = 0$ is solution

$C \rightarrow \infty$ $\Rightarrow$ Bribes are too costly $\Rightarrow$ Linear separable case.

$$L(w, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i - \xi_i\right) + \sum_{i=1}^{n} \beta_i (-\xi_i)$$

Dual:

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0}} \min_{w} L(w, \xi, \alpha, \beta)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

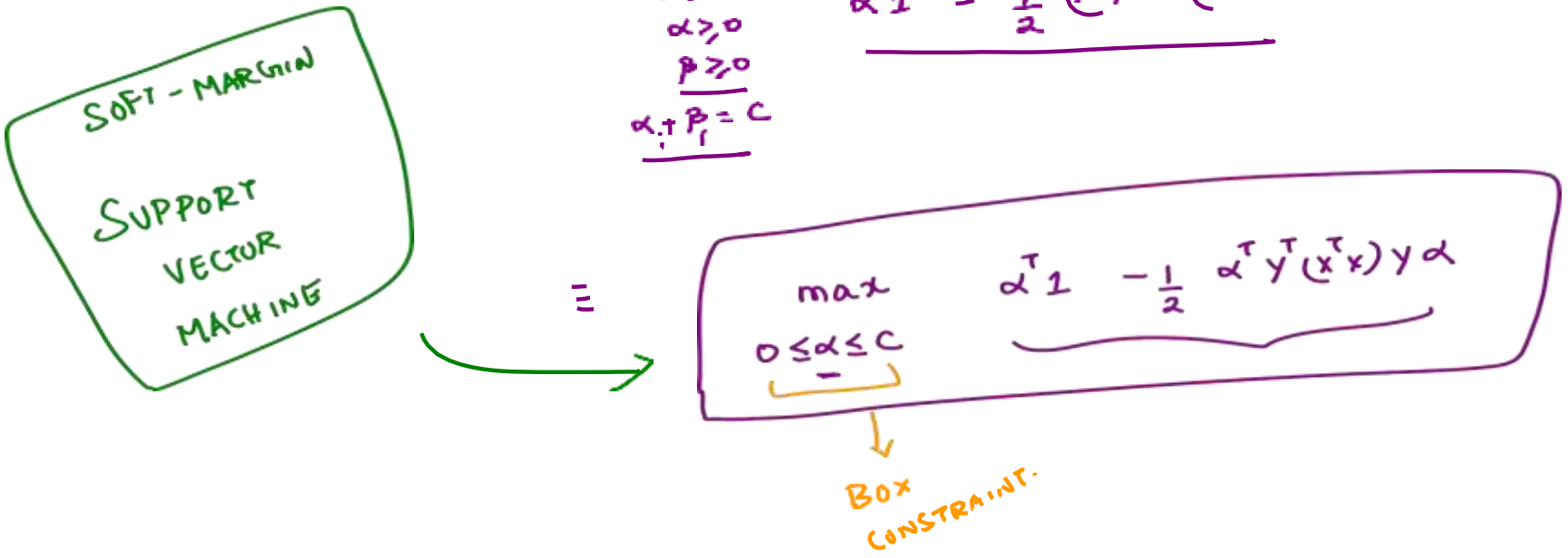
$$w^* = X Y \alpha$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \boxed{C - \alpha_i - \beta_i = 0}$$

$$\boxed{\alpha_i + \beta_i = C} \quad \forall i$$

Substitute $w = X Y \alpha$ in the original objective

$$\frac{1}{2}(XY\alpha)^T(XY\alpha) + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\xi_i + \alpha^T 1 - (XY\alpha)^T(XY\alpha)$$


SOFT - MARGIN
SUPPORT VECTOR MACHINE
max
$\alpha \geq 0$
$\beta \geq 0$
$\alpha_i + \beta_i = C$
$\alpha^T 1 - \frac{1}{2} (Xy\alpha)^T (Xy\alpha)$
$\equiv$
max
$0 \leq \alpha \leq C$
$\alpha^T 1 - \frac{1}{2} \alpha^T y^T (x^T x) y \alpha$
BOX CONSTRAINT.

HARD-MARGIN SVM

PRIMAL

$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$s.t \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

$1 - w^T x_i y_i \leq 0$

DUAL

$$\max_{\alpha \geq 0} \quad \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$$\alpha_i^* (1 - w^{*T} x_i y_i) = 0 \quad \forall i$$

$$w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i$$

SOFT-MARGIN SVM

PRIMAL ✓

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

$$s.t \quad \begin{cases} (w^T x_i) y_i + \xi_i \geq 1 & \forall i = 1, \dots n \\ \xi_i \geq 0 & \forall i = 1, \dots n \end{cases}$$

α, β

DUAL ✓

$$\max_{\substack{\alpha + \beta = C \\ \alpha \geq 0 \\ \beta \geq 0}} \quad \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$0 \leq \alpha \leq C$

- Let $(\underline{w}^*, \underline{\xi}^*)$ be the primal optimal solution
- Let $(\underline{\alpha}^*, \underline{\beta}^*)$ be the dual optimal solution

COMPLEMENTARY SLACKNESS

$$\alpha_i^* \left(1 - (w^{*T} x_i) y_i - \xi_i^* \right) = 0 \quad \forall i$$

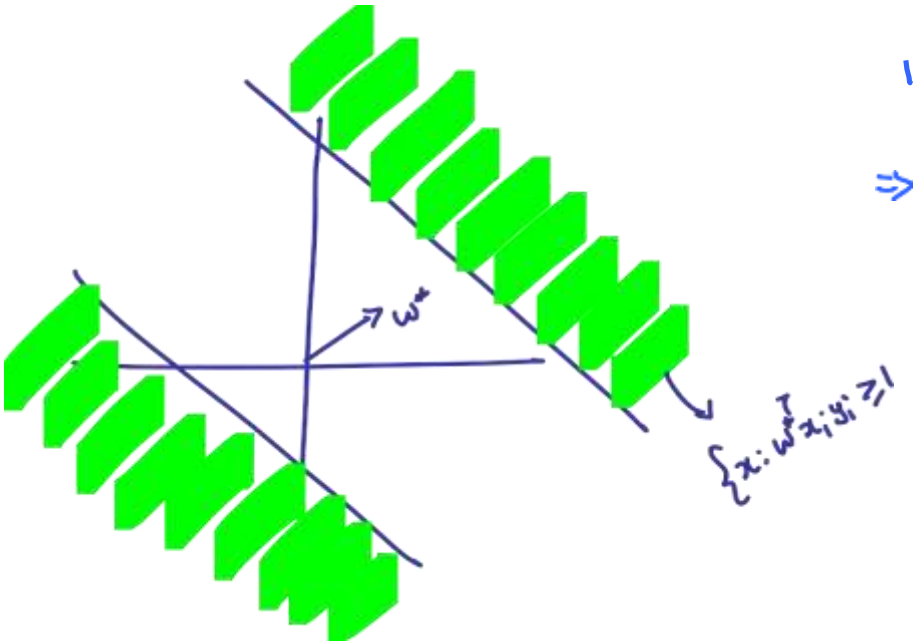
$$\beta_i^* \xi_i^* = 0 \quad \forall i$$

$\alpha_i^* + \beta_i^* = C \quad \forall i$

↳ (A)

Various cases possible

Case 1: $\alpha_i^* = 0$



$\overset{(A)}{\Rightarrow} \quad \beta_i^* = C \quad \overset{CS}{\Rightarrow} \quad \xi_i^* = 0$

$1 - (w^{*T} x_i) y_i - \underbrace{\xi_i^*}_{=0} \leq 0$ [Primal feasibility]

$\Rightarrow \quad 1 - (w^{*T} x_i) y_i \leq 0$

$\Rightarrow \quad w^{*T} x_i y_i \geq 1$

$\Rightarrow$ w^* classifies (x_i, y_i) correctly.

Case 2: $0 < \alpha_i^* < C$ $\overset{\text{KKT}}{\Rightarrow}$ $0 < \beta_i^* < C$ $\overset{\text{CS}}{\Rightarrow}$ $\xi_i^* = 0$

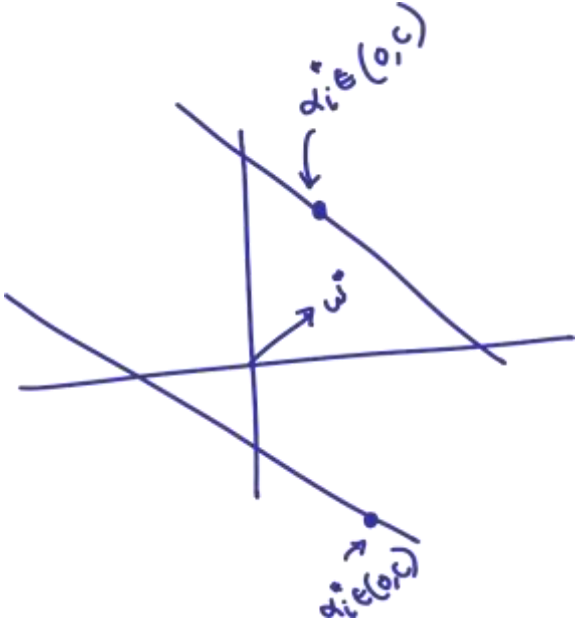
$\Downarrow$ CS

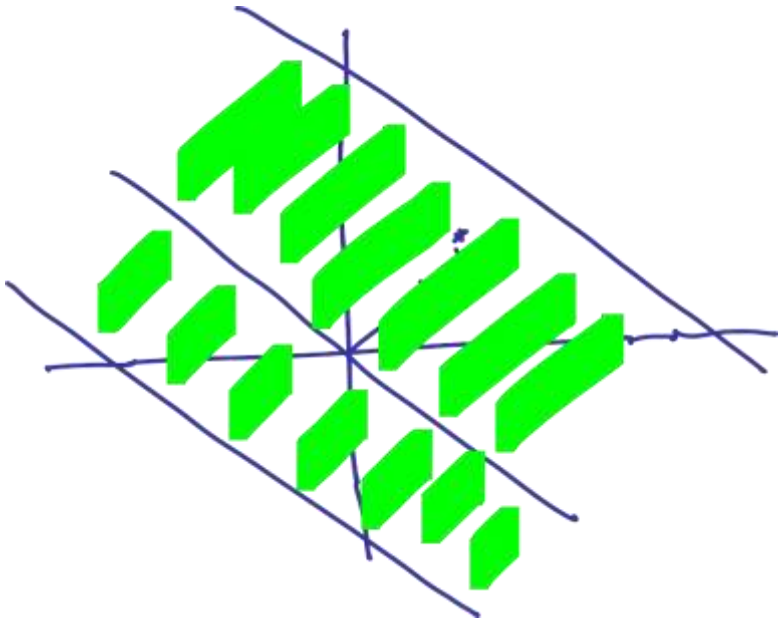
$1 - (w^{*T} x_i) y_i - \xi_i^* = 0$

$\Downarrow$

$(w^{*T} x_i) y_i = 1$

$\Rightarrow$ (x_i, y_i) lies on the supporting hyperplane.





Case 3: $\alpha_i^* = C \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$

$\Downarrow$ CS

$1 - w^{*T} x_i y_i - \xi_i^* = 0$

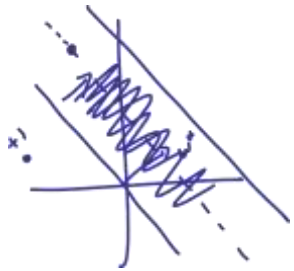
$\xi_i^* = 1 - w^{*T} x_i y_i \geq 0$

$\Rightarrow \boxed{w^{*T} x_i y_i \leq 1}$

Let's see this from P.O.V of data

CASE 1

$w^{*T} x_i y_i < 1$



$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0$$

$$w^{*T} x_i y_i \geq 1 - \xi_i^*$$

$$\xi_i^* \geq 1 - w^{*T} x_i y_i$$

$\Rightarrow \xi_i^* > 0 \Rightarrow \beta_i^* = 0 \Rightarrow \alpha_i^* = C$

$$\alpha_i^* (1 - w^{*T} x_i y_i - \xi_i^*) = 0$$

$$\beta_i^* \xi_i^* = 0$$

CASE 2: $w^{*T} x_i y_i = 1$

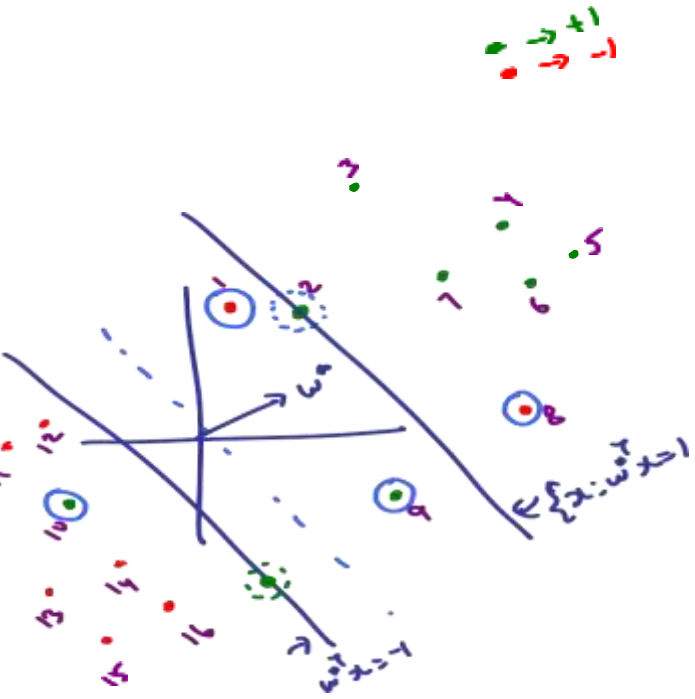
$$\xi_i^* \geq 1 - w^{*T} x_i y_i$$

$$\Rightarrow \quad \xi_i^* \geq 0 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$$

CASE 3 $w^{*T} x_i y_i > 1$

$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0 \quad \text{[Primal feasibility]}$$

$$\Rightarrow \quad 1 - w^{*T} x_i y_i - \xi_i^* < 0 \quad \overset{\text{C.S}}{\Rightarrow} \quad \alpha_i^* = 0$$



SUMMARY

$\alpha_i^* = 0 \quad \Rightarrow \quad w^{*T} x_i y_i \geq 1$

$0 < \alpha_i^* < C \quad \Rightarrow \quad w^{*T} x_i y_i = 1$

$\alpha_i^* = C \quad \Rightarrow \quad w^{*T} x_i y_i \leq 1$

✓ $w^{*T} x_i y_i < 1 \quad \Rightarrow \quad \alpha_i^* = C$

→ $w^{*T} x_i y_i = 1 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$

→ $w^{*T} x_i y_i > 1 \quad \Rightarrow \quad \alpha_i^* = 0.$

Binary classification

- GENERATIVE
  - Naive Bayes
  - G.D.A
- DISCRIMINATIVE
  - k-NN
  - Decision trees
  - Perceptron
  - Support-vector-machines

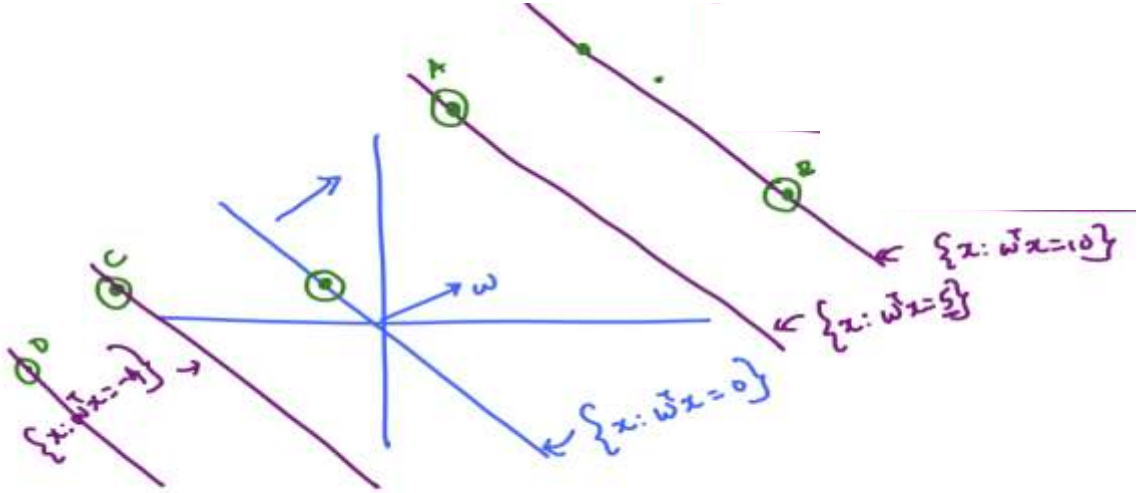
Doesn't "really" model $P(y|x)$

→ Just finds $f: \mathbb{R}^d \to \{\pm 1\}$

- Can we model $P(y=+1/x)$ differently?

Start with a simple model

Given $x \in \mathbb{R}^d$ $\quad z = w^T x \quad$ $w \in \mathbb{R}^d$.



$$P(y=+1 \mid x) = g(w^T x)$$

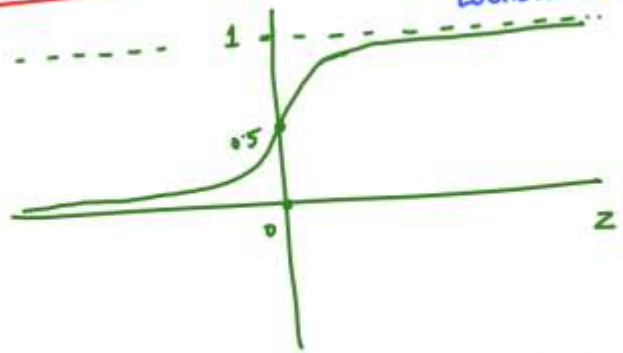
- $g(z) \in [0,1]$
- $g(z) \to 1$ as $z \to \infty$
- $g(z) \to 0$ as $z \to -\infty$
- $g(z) = 0.5$ if $z = 0$.

LINK FUNCTION

ONE POPULAR CHOICE

SIGMOID FUNCTION
LOGISTIC FUNCTION.

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



MODEL: LOGISTIC REGRESSION

Data: $\{(x_1, y_1) \cdots (x_n, y_n)\}$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$

Max. Likelihood

$$L(w, \text{Data}) = \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{y_i} \left(1 - g(w^T x_i)\right)^{(1-y_i)}$$

$$\log L(w, \text{Data}) = \sum_{i=1}^{n} y_i \log\left(g(w^T x_i)\right) + (1-y_i) \log\left(1 - g(w^T x_i)\right)$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-w^T x_i}} \right) + (1-y_i) \log\left(\frac{e^{-w^T x_i} \times 1}{1+e^{-w^T x_i}} \right) \right]$$

$$= \sum_{i=1}^{n} \left[\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - y_i(-w^T x_i) \right]$$

$$= \sum_{i=1}^{n} \left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right) \right]$$

- No closed form solution

- Gradient ascent

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

$$= \sum_{i=1}^{n} x_i \left(y_i - \left(1 - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) \right)$$

$$= \sum_{i=1}^{n} x_i \left(y_i - \frac{1}{1+e^{-w^T x_i}} \right)$$

($x_i \in \mathbb{R}^d$, $y_i \in \{0,1\}$, $g(w^T x_i)$)

$$w_{t+1} = w_t + \eta_t \nabla \log L(w_t)$$

REGULARIZED VERSION

$$\min_w \sum_{i=1}^{n} (1-y_i) w^T x_i + \log\left(1+e^{-w^T x_i}\right) + \frac{\lambda}{2}\|w\|^2$$

KERNEL VERSION

- Can argue $w = \sum_{i=1}^{n} \alpha_i x_i$

[Formal Theorem

Representer Theorem]

Exercise: Derive the kernel version of logistic regression

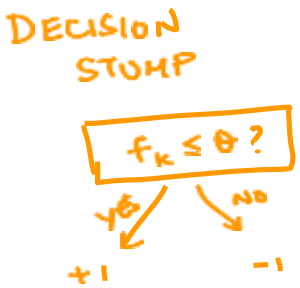
META CLASSIFIERS (OR)

ENSEMBLE CLASSIFIERS.

WEAK CLASSIFIERS [better than random]

STRONG CLASSIFIERS

Weak classifiers



high bias, low variance

Overfit decision tree

…

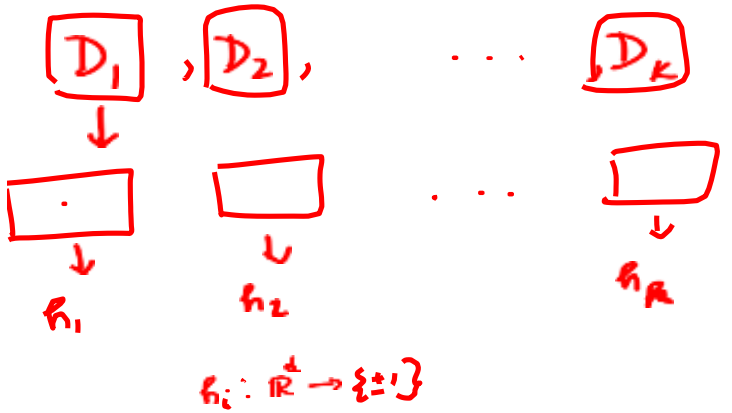
……

low bias, high variance

$x_1, x_2, \ldots x_n \sim N(\mu, 1)$

$\hat{\mu}_1 = x_1 \qquad \hat{\mu}_2 = x_2 \qquad \ldots \qquad \hat{\mu}_n = x_n \qquad \hat{\mu}_{ML} = \frac{1}{n}\sum x_i$

Overfit decision trees



$h^*(x) = \text{majority}(h_1(x), \ldots, h_K(x))$

BAGGING - Bootstrap Aggregation

$D = \{(x_1, y_1) \cdots (x_n, y_n)\}$

- Create datasets $D_1, \cdots, D_k$ from D by
  - Sampling with replacement.
- Run weak classifier on $D_1, \cdots, D_k$ to get $h_1, \cdots, h_k$
- Aggregate $h_1, \cdots, h_k$ using majority.

Chance that a point appears in a dataset

$1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{1}{n}\right)$

$1 - \left(1 - \frac{1}{n}\right)^n$

$1 - \frac{1}{e}$ (as $n \to \infty$)

$\simeq 66\%$

FEATURE BAGGING -> Bag the features in addition to data points

Feature bagged decision trees -> RANDOM FOREST

BOOTSTRAP - Sampling with Replacement ?

AGGREGATION - Majority. ?

BOOSTING

[Freund & Schapire. 1995 Godel Prize]

↑

ADA-BOOST

Distribution D over (x, y) — $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$

↳ unknown but fixed.

$x_1, \ldots, x_n$ are iid from D.

$h : \mathbb{R}^d \to \{\pm 1\}$

x $\quad$ y.

Measure performance using

$$P_{(x,y)\sim D}\left(h(x) \neq y\right)$$

Misclassification probability.

A weak / Strong learner is one which outputs a classifier h for which

$$P_{x,y\sim D}\left(h(x) = y\right) \geq \frac{1}{2} + \frac{1-\epsilon}{2}$$

$\gamma > 0$

for any unknown but fixed distribution D.

BOOSTING

Weak learner → Strong learner.

$\{(x_1, y_1), \ldots (x_n, y_n)\}$ → Weak learner → $h_t : \mathbb{R}^d \to \{\pm 1\}$

$p_1, \ldots p_n$

$\sum_i p_i = 1$

$$\sum_{i=1}^{n} p_i \, \mathbb{1}(h(x_i) \neq y_i) \leq \frac{1}{2} - \gamma$$

$\gamma > 0$.

Strong learner classifies all data points correctly.

BOOSTING [ADABOOST]

$S = \{(x_1, y_1) \ldots (x_n, y_n)\}$

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

$y_i \in \{\pm 1\}$.

Initialize $D_1(i) = \frac{1}{n} \quad \forall i$

D_1: iteration

for $t = 1, \ldots, T$

h_t = [Input S, D_t to the weak learner to get h_t]

$$D_{t+1}(i) = \frac{D_t(i) \cdot e^{-\alpha_t y_i h_t(x_i)}}{\sum_{j=1}^{n} D_t(j)\, e^{-\alpha_t y_j h_t(x_j)}} \quad \forall i$$

$\alpha_t > 0 \quad \forall t.$

end.

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

$$C(x) = \text{sign}(H(x))$$

Analysis of Boosting

$$D_{t+1}(i) = \frac{D_t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

→ update rule.

$$Z_t = \sum_{j=1}^{n} D_t(j)\, e^{-\alpha_t y_j h_t(x_j)}$$

$$D_{t+1}(i) = \left(\frac{D_{t-1}(i)\, e^{-\alpha_{t-1} y_i h_{t-1}(x_i)}}{Z_{t-1}} \right) \cdot \frac{e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

$$D_{t+1}(i) = \frac{\frac{1}{n} e^{-\sum_{k=1}^{t} \alpha_k y_i h_k(x_i)}}{\prod_{k=1}^{t} Z_k}$$

$$D_{T+1}(i) = \frac{\frac{1}{n} e^{-y_i H(x_i)}}{\prod_{k=1}^{T} Z_k} \longrightarrow \sum_{k=1}^{T} \alpha_k h_k(x_i)$$

$$\underbrace{\sum_{i=1}^{n} D_{t+1}(i)}_{=1} = \frac{\sum_{i=1}^{n} \frac{1}{n} e^{-y_i H(x_i)}}{\prod_{l=1}^{T} Z_l}$$

$$\prod_{l=1}^{T} Z_l = \underbrace{\frac{1}{n} \sum_{i=1}^{n} e^{-y_i H(x_i)}}_{} \quad \text{— (A)}$$

Recall, $C(x) = \text{Sign}(H(x))$

$$\mathbb{1}(C(x_i) \neq y_i) = \mathbb{1}(H(x_i)\, y_i < 0) \leq e^{-H(x_i) y_i} \quad \text{— (B)}$$

$$\Rightarrow \quad \prod_{t=1}^{T} z_t \quad = \quad \frac{1}{n} \sum_{i=1}^{n} e^{-y_i H(x_i)} \qquad \text{(from A)}$$

$$\geq \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(H(x_i)\, y_i < 0 \right) \qquad \text{(from B)}$$

$$\geq \quad \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mathbb{1}\left(C(x) \neq y_i \right)}_{\text{error}(C)}$$

$$\boxed{\text{error}(C) \quad \leq \quad \prod_{t=1}^{T} z_t} \quad \leq \quad \square$$

$$\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} \left[\sum_{j=1}^{n} D_t(j)\, e^{-\alpha_t y_j h_t(x_j)} \right]$$

$$= \prod_{t=1}^{T} \left[\sum_{j=1}^{n} e^{-\alpha_t} D_t(j)\, \mathbb{1}\left(y_j = h_t(x_j)\right) + \sum_{j=1}^{n} e^{\alpha_t} D_t(j)\, \mathbb{1}\left(y_j \neq h_t(x_j)\right) \right]$$

$$= \prod_{t=1}^{T} e^{-\alpha_t}\left(1 - \text{error}(h_t)\right) + e^{\alpha_t}\left(\text{error}(h_t)\right)$$

Choose α_t s.t.

$e^{-\alpha_t}(1-\text{error}(h_t)) + e^{\alpha_t}(\text{error}(h_t))$ is as small as possible.

$$\alpha_t^* = \ln\sqrt{\frac{1-\text{err}(h_t)}{\text{err}(h_t)}}$$

$$= \prod_{t=1}^{T} 2\sqrt{\text{error}(h_t)(1-\text{error}(h_t))}$$

$$\leq \prod_{t=1}^{T} 2\sqrt{\left(\tfrac{1}{2}-\gamma\right)\left(\tfrac{1}{2}+\gamma\right)} \quad [\text{show this}]$$

$$= \prod_{t=1}^{T} 2\sqrt{\frac{1-4\gamma^2}{4}}$$

$e^{-\alpha_t}(1-\theta) + e^{\alpha_t}\theta$

$(1-\theta)e^{-\alpha_t}(-1) + e^{\alpha_t}\theta = 0$

$-1 + \theta e^{-\alpha_t} + \theta e^{\alpha_t} = 0$

$$= \prod_{t=1}^{T} \sqrt{1-4\gamma^2}$$

$$\leq \prod_{t=1}^{T} \left(e^{-4\gamma^2}\right)^{1/2} = \prod_{t=1}^{T} e^{-2\gamma^2}$$

$$= e^{-2T\gamma^2} \quad \text{--- (2)}$$

$$\frac{1}{n}\sum_{i} \mathbb{1}(c(x_i) \neq y_i)$$

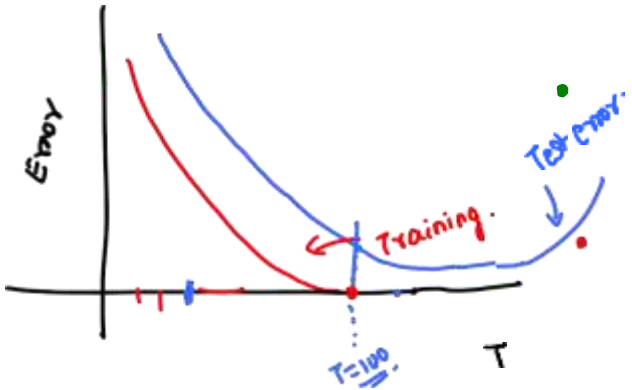
$$\text{error}(c) \leq \prod_{t=1}^{T} z_t \leq e^{-2T\gamma^2} \leq \frac{1}{2n}$$

$$T \geq \frac{1}{2\gamma^2} \ln(2n)$$

$$\Rightarrow \text{error}(c) = 0$$

Strong Learner

Boosting - Summary



- Every bag depends on previous bag's performance
- Next bag "focusses" on previous bag's mistakes
- Can show training error goes to 0 as number of rounds increases

Cannot run in parallel – hence sometimes bagging is preferred in practice

Usually weak learners are decision stumps -> high bias, low variance
Boosting reduces bias without affecting variance a lot

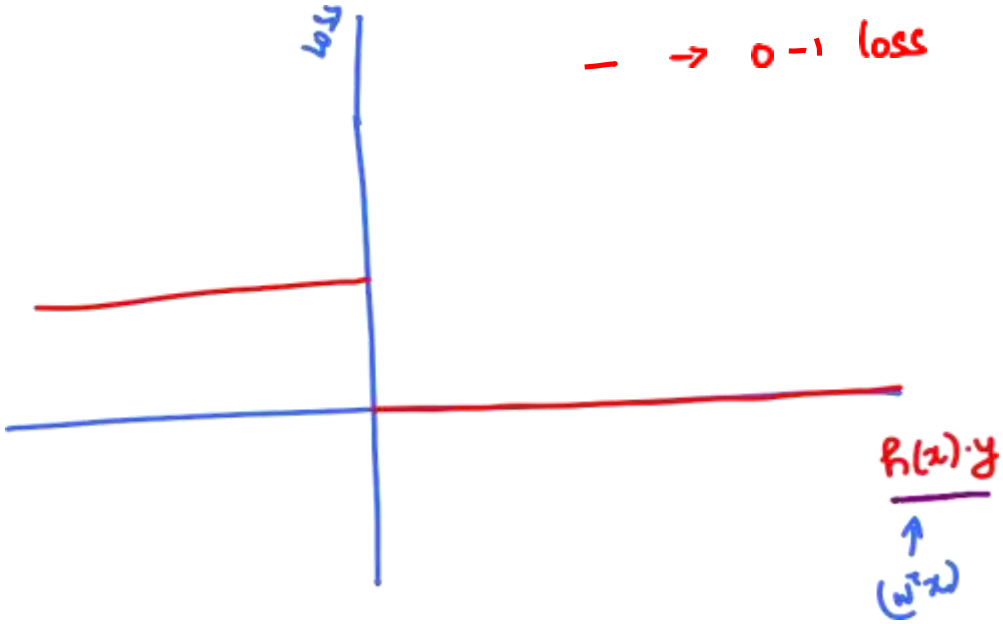
Why are there so many binary classification algorithms?

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

$y \in \{\pm 1\}$

$$\min_{w} \sum_{i=1}^{n} \mathbb{1}(w^T x \cdot y < 0)$$

NP-Hard problem



Linear Regression for classification

$$\min_h \sum_{i=1}^{n} (h(x) - y_i)^2$$

Final classifier: $\text{sign}(h(x))$

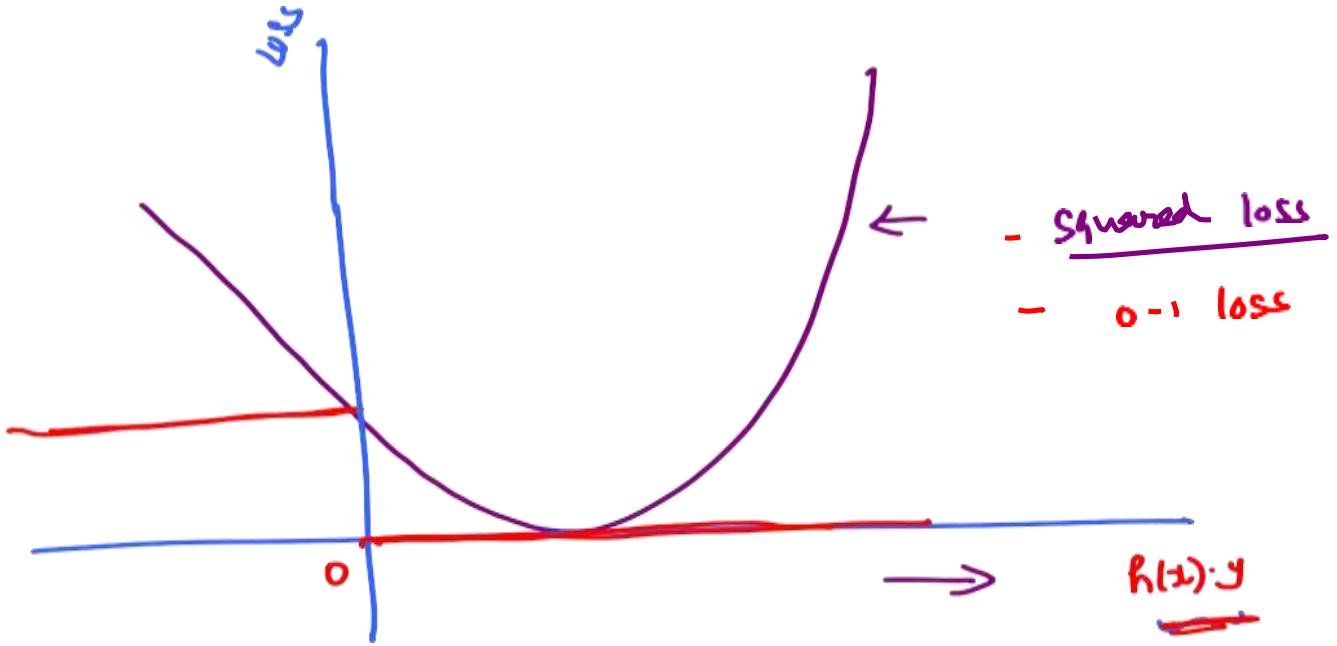
Loss per point for h

$$(h(x) - y)^2 = \begin{cases} (h(x) - 1)^2 & \text{if } y = 1 \\ (h(x) + 1)^2 & \text{if } y = -1 \end{cases}$$

If $y = 1$: $(h(x))^2 + 1 - 2h(x) = (h(x))^2 + 1 - 2h(x)y$

$y = -1$: $(h(x))^2 + 1 + 2h(x) = (h(x))^2 + 1 - 2h(x) \cdot y$

$2 = (h(x) \cdot y - 1)^2$


loss
0
$h(x) \cdot y$
- squared loss
- 0-1 loss

SUPPORT VECTOR MACHINES

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

$$w^T x_i y_i + \xi_i \geq 1 \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

Equivalently

$$\xi_i \geq 1 - w^T x_i y_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

Equivalently

$$\xi_i \geq \max(1 - w^T x_i y_i, 0)$$

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

$$\xi_i \geq \max(1 - w^T x_i y_i, 0)$$

Equivalently,

$$\min_{w,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\max(1 - w^T x_i y_i, 0)$$

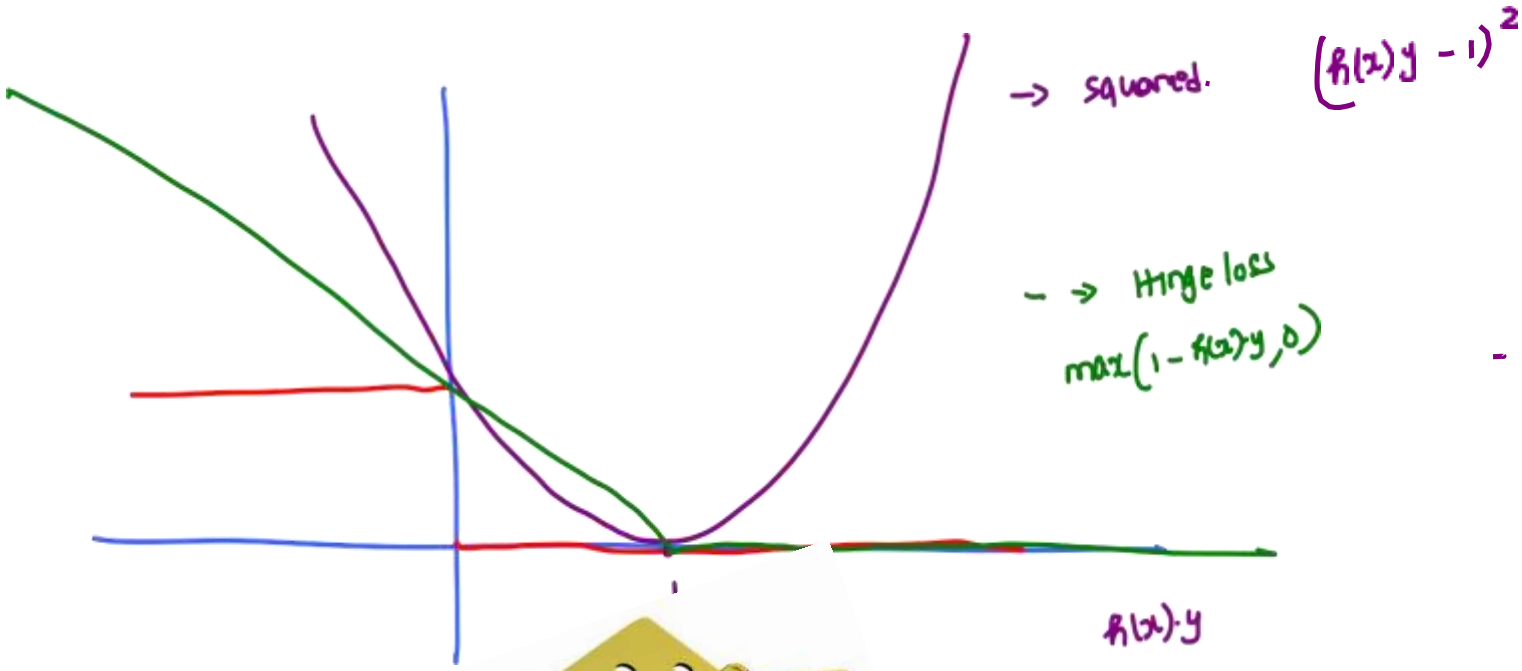
Model Regularization term

Data Loss term

So what exactly is the loss?

$$\ell(w, (x,y)) = \max(1 - w^T x y, 0)$$

HINGE LOSS



Hinge

Courtesy: Google images

Logistic regression

$$\max_{w} \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{z_i} \left(1 - g(w^T x_i)\right)^{(1-z_i)}$$

$z_i = 1$ if $y_i = 1$

$z_i = 0$ if $y_i = -1$

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

$$\max_{w} \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{z_i} \left(1 - g(w^T x_i)\right)^{(1-z_i)}$$

$$\max_{w} \sum_{i=1}^{n} z_i \log\left(g(w^T x_i)\right) + (1-z_i) \log\left(1 - g(w^T x_i)\right)$$

$$= \min_{w} \sum_{i=1}^{n} \left[-z_i \log\left(g(w^T x_i)\right) + (z_i - 1) \log\left(1 - g(w^T x_i)\right) \right]$$

Loss for a single point when $z_i = 1$ $(y_i = 1)$

$$-\log(g(w^T x_i)) = -\log\left(\frac{1}{1+e^{-w^T x_i}}\right)$$

$$= \log\left(1+e^{-w^T x_i}\right) = \log\left(1+e^{-y_i w^T x_i}\right)$$

Loss for a single point when $z_i = 0$ $(y_i = -1)$

$$= -\log(1-g(w^T x_i))$$

$$= -\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}\right)$$

$$= -\log\left(1-\frac{1}{1+e^{-w^T x_i}}\right)$$

$$= \log\left(1+e^{w^T x_i}\right) = \log\left(1+e^{-y_i w^T x_i}\right)$$

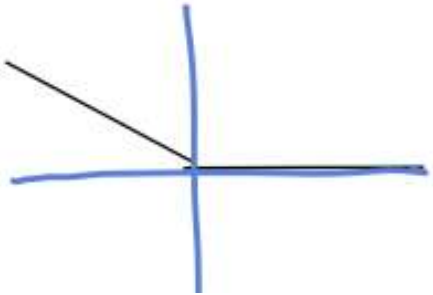

$\mathbb{1}(h(x)\cdot y < 0)$
$\max(1 - h(x)\cdot y, 0)$
$(h(x)\cdot y - 1)^2$
Logistic loss
$\log(1 + e^{-h(x)\cdot y})$
$h(x)\cdot y$

Perceptron

$$W_{t+1} = W_t + x_t y_t$$

Consider the hinge loss

$$\text{loss}(w, (x, y)) = \max(0, 1 - w^T x y)$$



$$\text{loss}(w,(x,y)) = \max(0,\ 1 - \vec{w}^T x y)$$

$$\frac{\partial \text{loss}}{\partial w} = \begin{cases} -x \cdot y \text{ if } (w^T x) y < 0 \\ = 0 \text{ if } (w^T x) y > 0 \\ = [-1, 0]\, x \cdot y \text{ if } (w^T x) y = 0 \end{cases}$$

sub-gradient

↳ choose $-xy$ if mistake.

When mistake

$$w_{t+1} = w_t - \eta_t(-x_t y_t)$$

$\eta_t = 1$

- Perceptron can be viewed as S.G.D with Hinge loss with step size $= 1$

BOOSTING

Recall

$$\prod_{t=1}^{T} Z_t = \frac{1}{n} \sum_{i=1}^{n} e^{-\left(\sum_{t=1}^{T} \alpha_t h_t(x_i)\right) y_i}$$

$$:= \frac{1}{n} \sum_{i=1}^{n} e^{-H_T(x_i)\, y_i}$$

at round T, we so far have

$$\sum_{t=1}^{T-1} \alpha_t h_t(x)$$

We need to add one more classifier in round T to get

$$\sum_{i=1}^{T-1} \alpha_t h_t(x) + \alpha_T h_T(x)$$

greedy choice to minimize

$$\frac{1}{n} \sum_{i=1}^{n} e^{\left[\sum_{i=1}^{T-1} \alpha_t h_t(x_i) + \alpha_T h_T(x)\right] y_i}$$

Define $\text{loss}(h, x, y) = e^{-y\,g(x)}$

- At every round, one chooses α_t to minimize Z_t.
- Thus AdaBoost can be viewed as "greedy / co-ordinate descent" on exponential loss

What about Peceptron and Boosting?

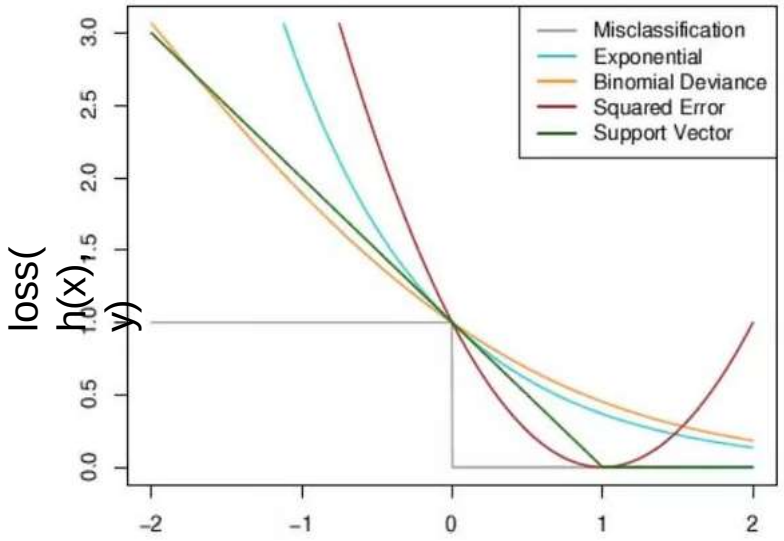
- one can argue perceptron update rule is equivalent to doing a SSGD on hinge loss with stepsize = 1

  SSGD → Stochastic sub-gradient descent.

$$w_{t+1} = w_t + x_i y_i$$

- Boosting (Adaboost) can be viewed as "greedy / co-ordinate wise" descent of exponential loss $e^{-f(x) \cdot y}$

SUMMARY



0-1
Exponential
Logistic
Squared
Hinge

-> The 0-1 loss is NP-hard to optimize even for linear classifiers

-> Different algorithms get around this by using a “surrogate” loss function

-> Surrogates are usually ***“convex” surrogates*** that are easy to optimize