

# **Support Vector Machines**

**Intro to ML 466/566  
Winter 2021**

Many of these slides are derived from Seyong  
Kim and Alex Thomo. Thanks!

# **Administrivia**

- As 2 is out
  - Due March 14 (Sunday)
- Proj mid-term report specification is out
- Proj final report specification will be released soon

# The G20, Pittsburgh PA 2009



All photos by Arthur Gretton

<https://www.flickr.com/photos/30686429@N07/3953914015>

# The G20, Pittsburgh PA 2009



# The G20, Pittsburgh PA 2009



# Linear classifier (E.g., Perceptron)

$$h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

Which outputs  $+1$  or  $-1$ .

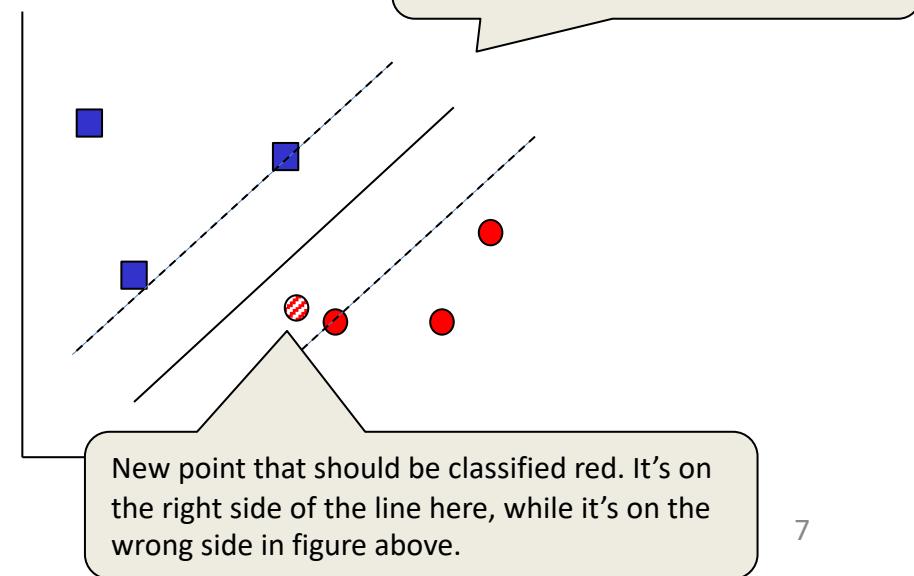
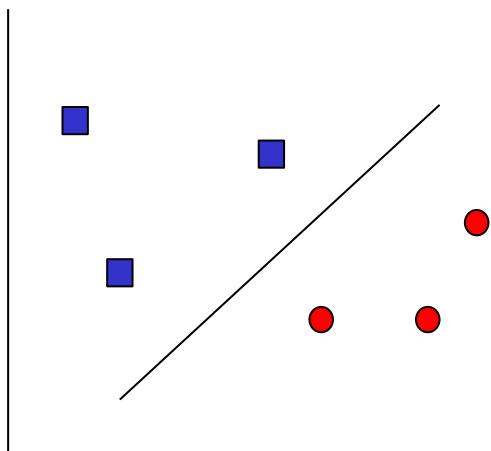
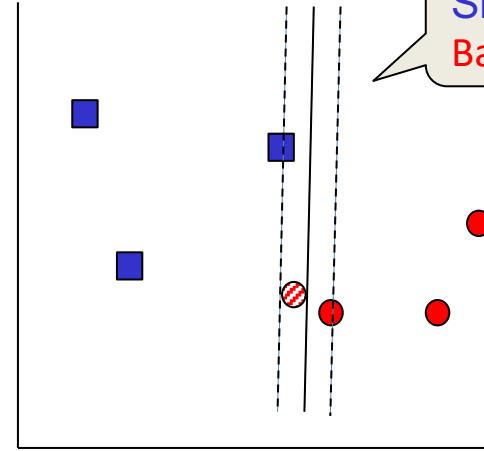
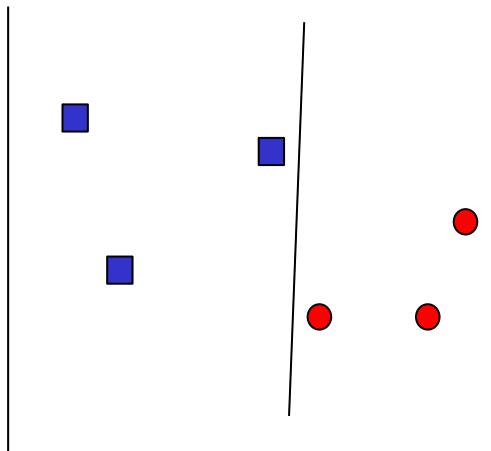
Say:

$+1$  corresponds to blue, and  
 $-1$  to red, or vice versa.

Many lines do the job.

Which one to choose?

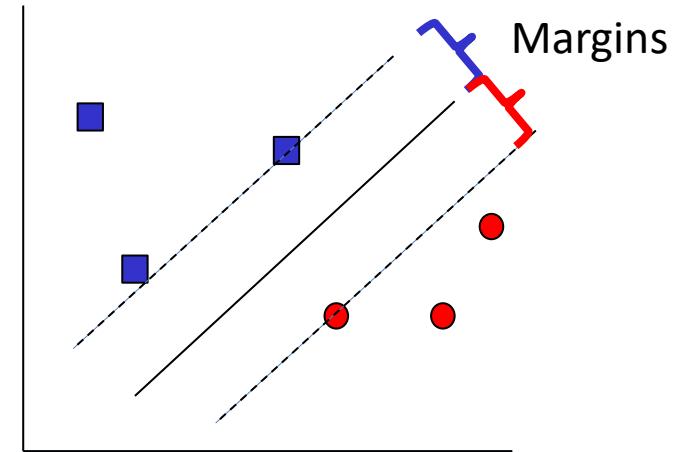
# Margin for different separators



# What is a margin?

Margin is the distance between:

- a point (■, ●)
- and the decision plane



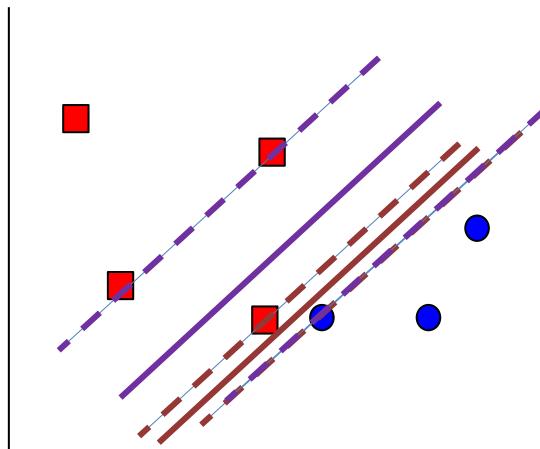
We want to find parameters of a line that:

1. Maximizes the margin for the closest of both point types ■ ●
2. Correctly classify all points

# What about outliers?

We want to find parameters of a line that:

1. Maximizes the margin for both point types
2. Correctly classify all points



There is a trade off to be made!

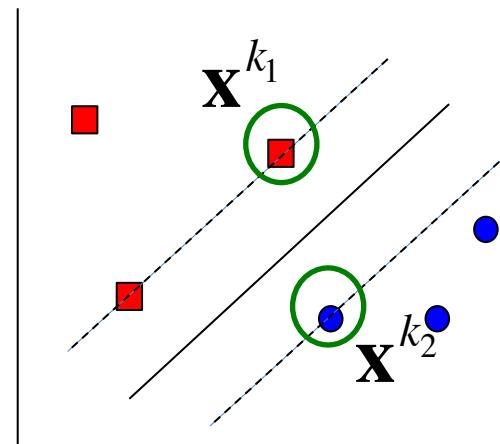
# Support Vectors

We want to find parameters of a line that:

1. **Maximizes the margin for the closest of both point types**
2. Correctly classify all points

Because of point 1, we will always have (at least) one point on each of the +1 and -1 lines

Closest points are called “support vectors”.



# Scale Invariance

$$h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

- We rescale  $\mathbf{w}$  and  $b$  (without changing the line) such that:

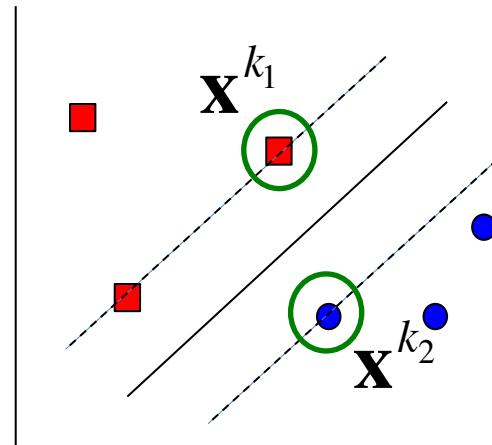
$$\mathbf{w} \cdot \mathbf{x}^{k_1} + b = 1$$

for the closest point(s) to the line on the +1 side, and

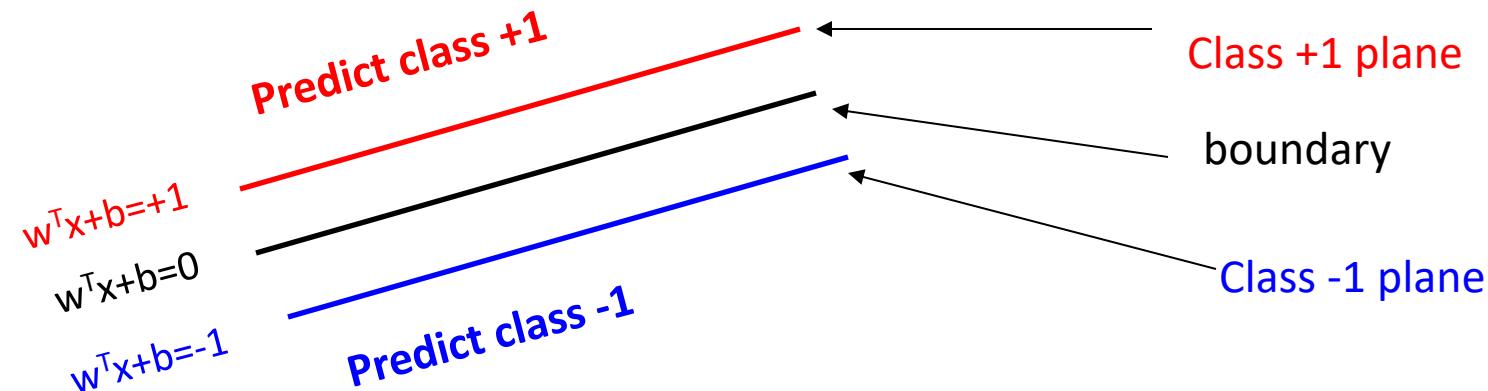
$$\mathbf{w} \cdot \mathbf{x}^{k_2} + b = -1$$

for the closest point(s) to the line on the -1 side.

Closest points are called “support vectors”.



# Specifying a max margin classifier



Classify as +1      if       $w^T x + b \geq 1$

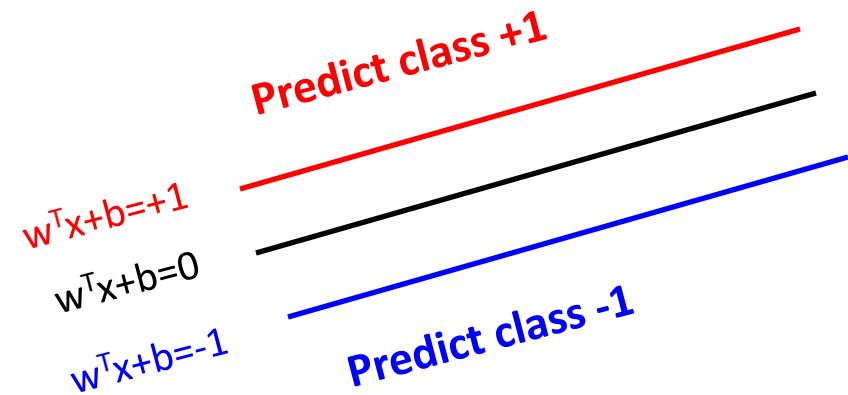
Classify as -1      if       $w^T x + b \leq -1$

Undefined      if       $-1 < w^T x + b < 1$

Learn one  $w$  and  $b$   
to satisfy these  
equations

In practice, at test time we would still give +1, -1 classes to the undefined points based on the sign. At train time, we are setting up the problem so no training points fall into this intermediate zone.

# Specifying a max margin classifier



Is the linear separation assumption realistic?

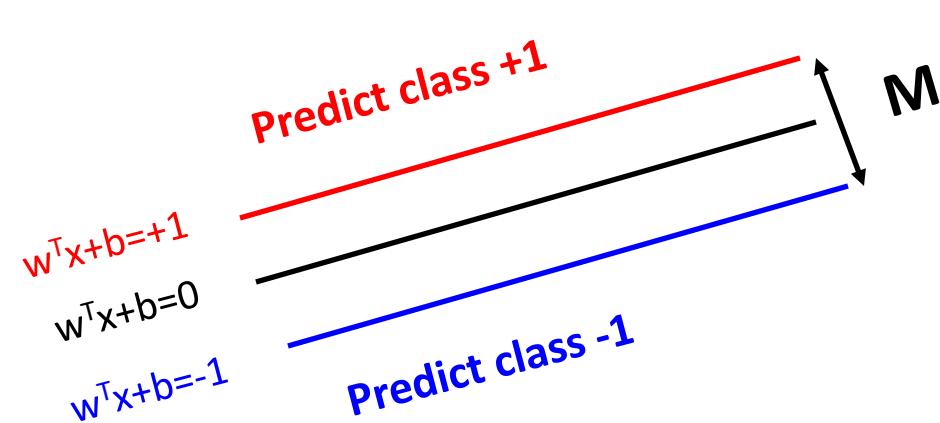
We will deal with this shortly, but lets assume it is for now

Classify as +1 if  $w^T x + b \geq 1$

Classify as -1 if  $w^T x + b \leq -1$

Undefined if  $-1 < w^T x + b < 1$

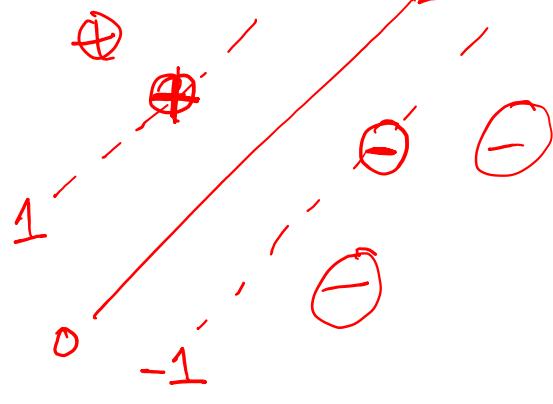
# Maximizing the margin



Classify as +1 if $w^T x + b \geq 1$
Classify as -1 if $w^T x + b \leq -1$
Undefined if $-1 < w^T x + b < 1$

- Lets define the width of the margin by  $M$
- How can we encode our goal of maximizing  $M$  in terms of our parameters ( $w$  and  $b$ )?
- Lets start with a few observations

$$h(x) = \text{sign}(w^T x + b)$$



$$w^T x + b = 1$$

$x^+$  is on +1 plane

$x^-$  is on -1 plane

$w$  is orthogonal to  
+1 plane

distance from  $x^+$  to  $x^-$  is  $\|w\|$

start at  $x^+$  walk  $\|w\|$  steps

we get to  $x^-$

$$x^+ = \|w\| w + x^-$$

$$w^T x^+ + b = 1$$

$$\rightarrow w^T \bar{x}^- + b = -1$$

$$\rightarrow x^+ = \lambda w + \bar{x}^-$$

$$|x^+ - \bar{x}^-| = M \quad (\text{margin})$$

Let's find  $\lambda$

$$w^T x^+ + b = 1$$

$$w^T (\lambda w + \bar{x}^-) + b = 1$$

$$w^T \lambda w + \underbrace{w^T \bar{x}^- + b}_{-1} = 1$$

$$w^T \lambda w = 2$$

$$\lambda = \frac{2}{w^T w}$$

Let's find M

$$M = |x^+ - x^-| = |\lambda w|$$

$$= \lambda |w|$$

$$= \underline{\lambda} \sqrt{w^T w}$$

$$= \frac{2}{w^T w} \sqrt{w^T w}$$

$$= \cancel{2 \sqrt{w^T w}}$$

we want w that

① correctly classifies all pts.

② maximizes  $\frac{2}{\sqrt{w^T w}} \Rightarrow \min \frac{1}{2} w^T w$

$$\min w^T w / 2$$

Subject to

$$w^T x_i + b \geq 1 \quad y_i \in \{+1\}$$

$$w^T x_i + b \leq -1 \quad y_i \in \{-1\}$$

$$y_i (w^T x_i + b) \geq 1 \quad y_i \in \{+1\}$$

$$y_i (w^T x_i + b) \geq 1 \quad y_i \in \{-1\}$$

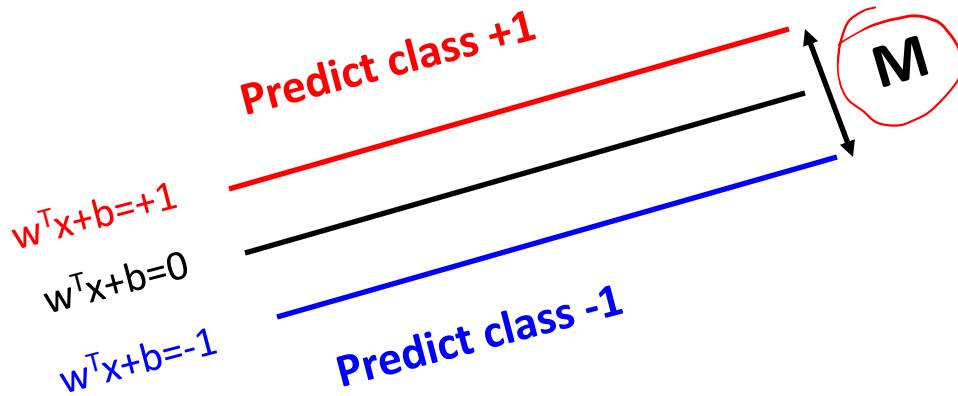
---

Subject to

$$y_i (w^T x_i + b) \geq 1 \quad \forall (x_i, y_i)$$



# Maximizing the margin



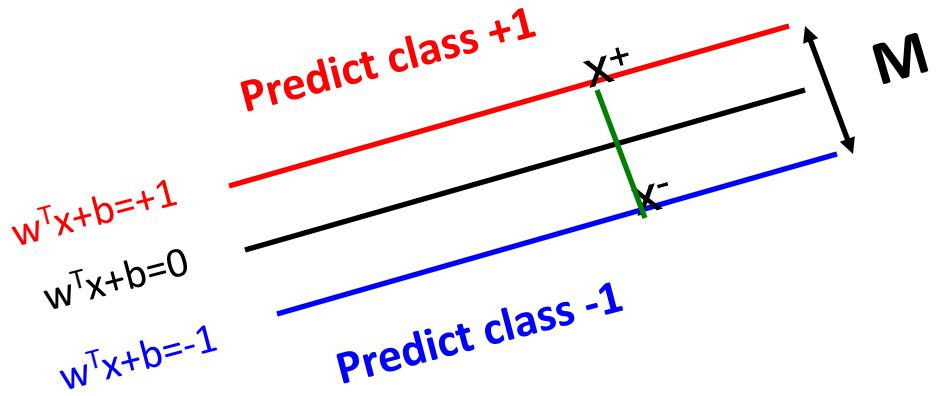
Classify as +1 if  $w^T x + b \geq 1$   
Classify as -1 if  $w^T x + b \leq -1$   
Undefined if  $-1 < w^T x + b < 1$

- Observation 1: the vector  $w$  is orthogonal to the +1 plane
- Why?

Let  $u$  and  $v$  be two points on the +1 plane, then for the vector defined by  $u$  and  $v$  we have  $w^T(u-v) = 0$

Corollary: the vector  $w$  is orthogonal to the -1 plane

# Maximizing the margin



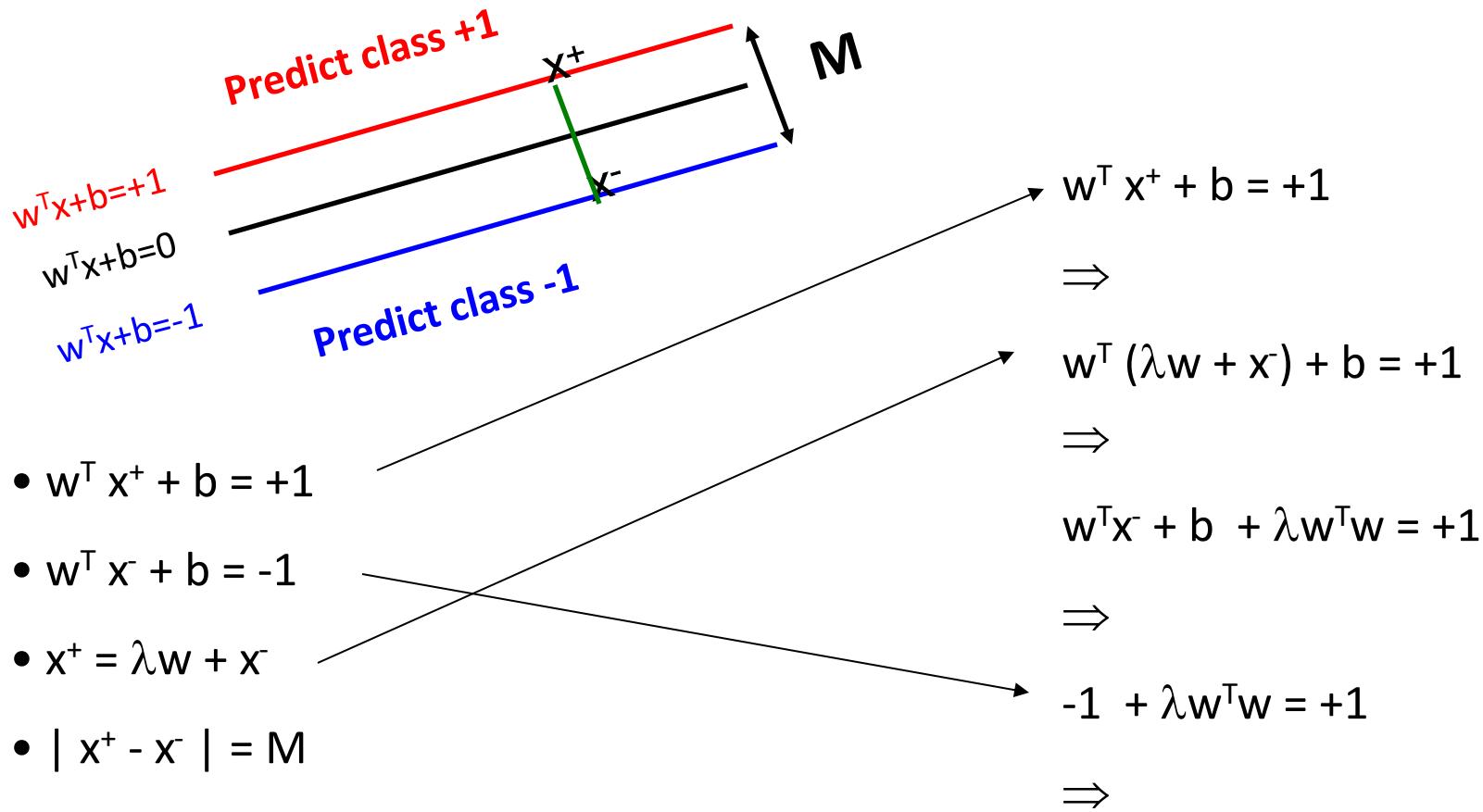
Classify as +1 if  $w^T x + b \geq 1$   
Classify as -1 if  $w^T x + b \leq -1$   
Undefined if  $-1 < w^T x + b < 1$

- Observation 1: the vector  $w$  is orthogonal to the +1 and -1 planes
- Observation 2: if  $x^+$  is a point on the +1 plane and  $x^-$  is the closest point to  $x^+$  on the -1 plane then

$$x^+ = \lambda w + x^-$$

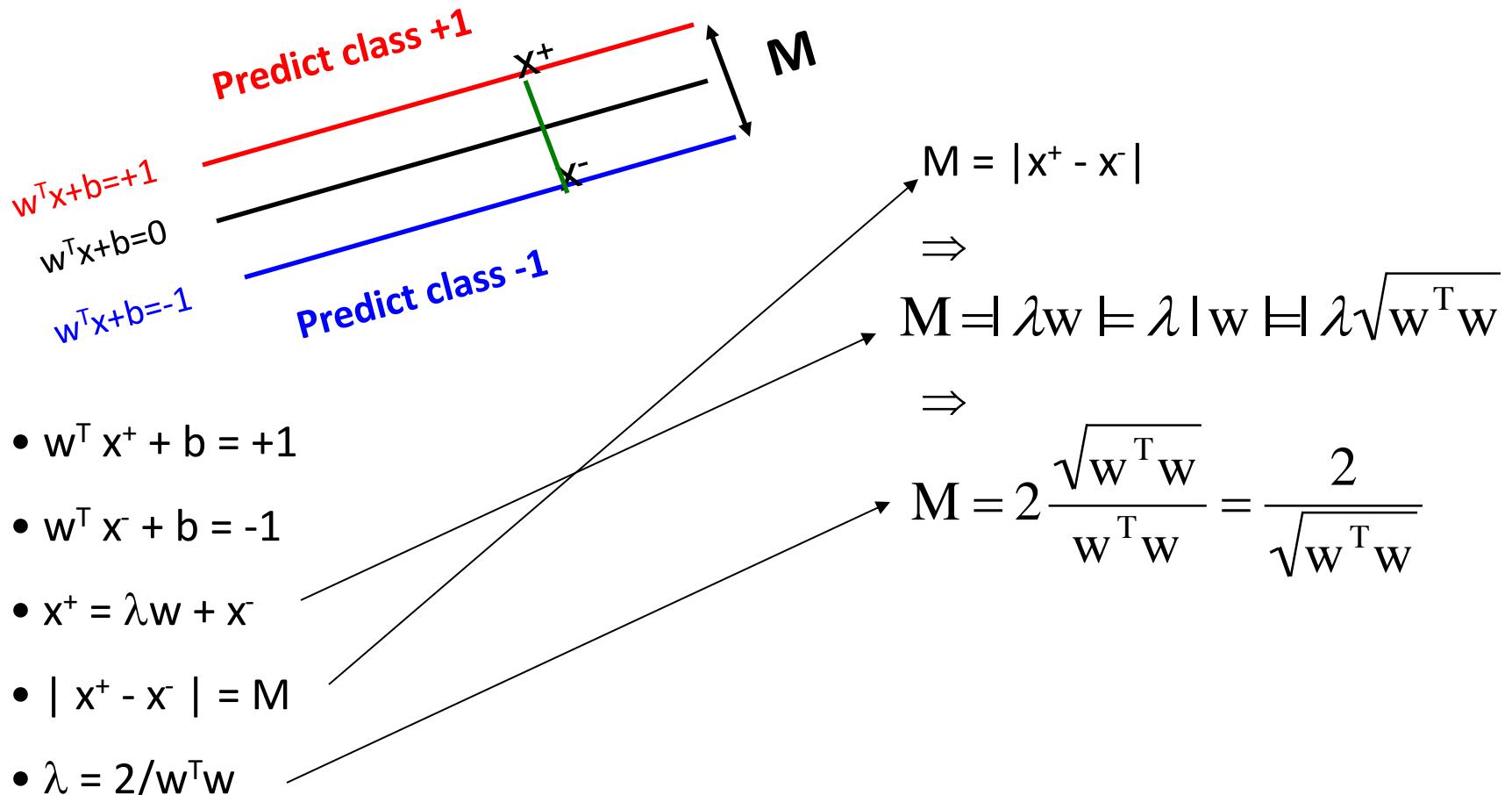
Since  $w$  is orthogonal to both planes we need to ‘travel’ some distance along  $w$  to get from  $x^+$  to  $x^-$

# Putting it together



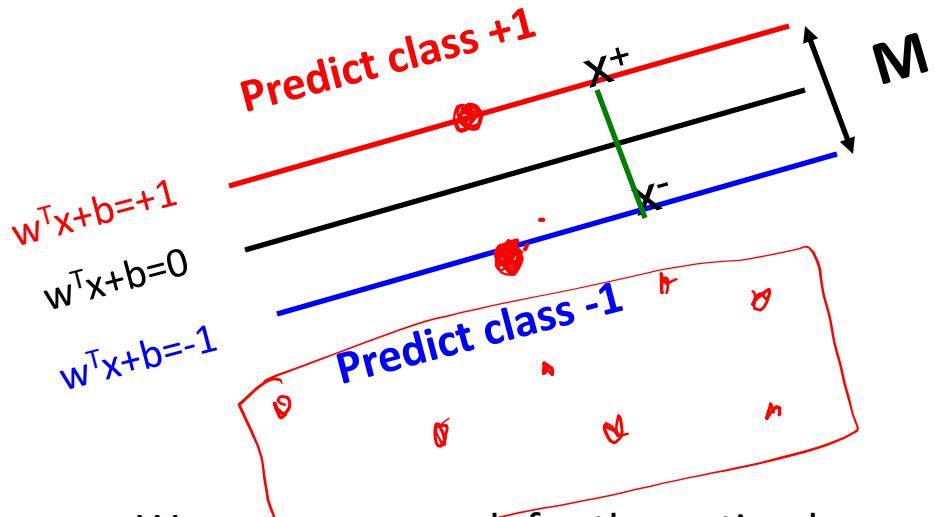
We can now define  $M$  in terms of  $w$  and  $b$

# Putting it together



We can now define  $M$  in terms of  $w$  and  $b$

# Finding the optimal parameters



$$M = \frac{2}{\sqrt{w^T w}}$$

We can now search for the optimal parameters by finding a solution that:

1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes  $w^T w$ )

Several optimization methods can be used:  
Gradient descent, simulated annealing, EM etc.

# Aside: Quadratic programming (QP)

Quadratic programming solves optimization problems of the following form:

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \dots \leq b_1$$

$$\vdots \quad \vdots \quad \vdots$$

$$a_{n1}u_1 + a_{n2}u_2 + \dots \leq b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \dots = b_{n+1}$$

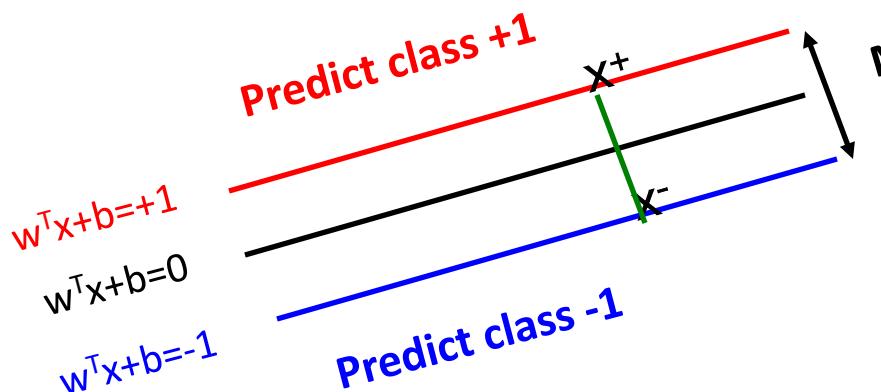
$$\vdots \quad \vdots \quad \vdots$$

$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \dots = b_{n+k}$$

Quadratic term

When a problem can be specified as a QP problem we can use solvers that are better than gradient descent or simulated annealing

# SVM as a QP problem



$$M = \frac{2}{\sqrt{w^T w}}$$

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \dots \leq b_1$$

$$\vdots \quad \vdots \quad \vdots$$

$$a_{n1}u_1 + a_{n2}u_2 + \dots \leq b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \dots = b_{n+1}$$

$$\vdots \quad \vdots \quad \vdots$$

$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \dots = b_{n+k}$$

Min  $(w^T w)/2$

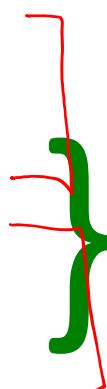
subject to the following inequality constraints:

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$



A total of n constraints if we have n input samples

# SVM as a QP problem: a simplification

$$\text{Min } (w^T w)/2$$

subject to the following inequality constraints:

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$

$$\text{Min } (w^T w)/2$$

subject to the following inequality constraints:

For all  $x$  in class + 1

$$y(w^T x + b) \geq 1$$

For all  $x$  in class - 1

$$y(w^T x + b) \geq 1$$

The same constraint!!  
So much easier to handle!

# Example

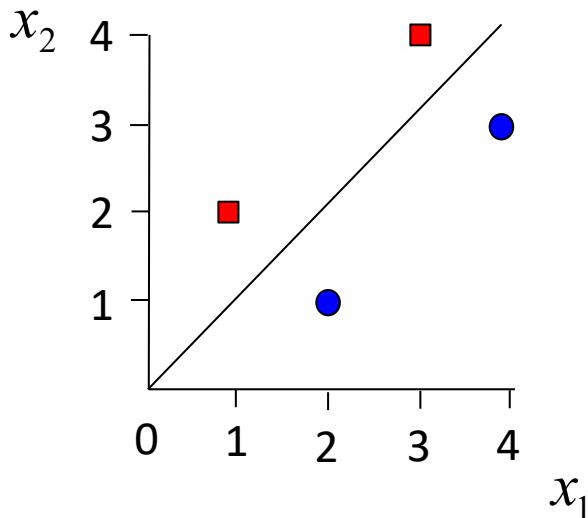
Training tuples:

$$([1, 2], +1)$$

$$([2, 1], -1)$$

$$([3, 4], +1)$$

$$([4, 3], -1)$$



$$\min_{w_1, w_2} \frac{1}{2} (w_1^2 + w_2^2)$$

subject to

$$(+1)(w_1 + 2w_2) \geq 1$$

$$(-1)(2w_1 + w_2) \geq 1$$

$$(+1)(3w_1 + 4w_2) \geq 1$$

$$(-1)(4w_1 + 3w_2) \geq 1$$

For this example, solution is easy to see:

$b = 0$  and  $w = [-1, +1]$ , i.e. the line is:

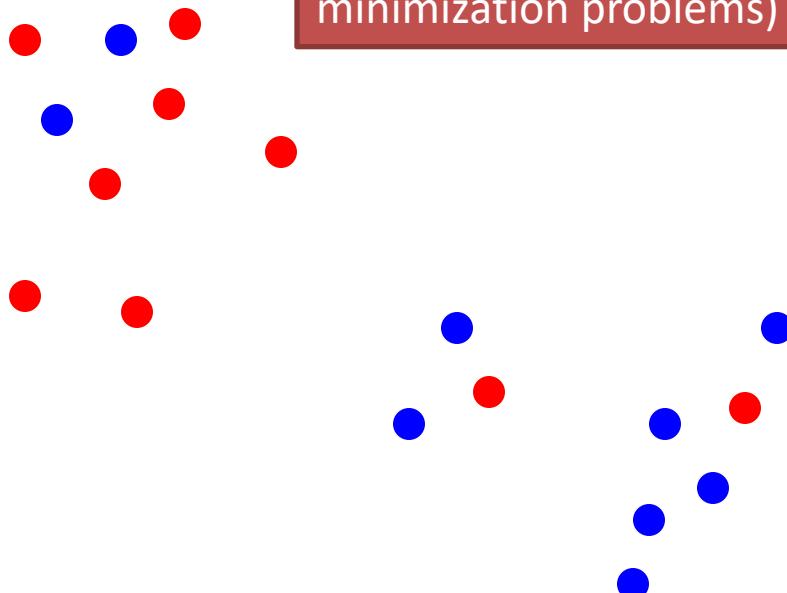
$$-x_1 + x_2 = 0$$

All conditions are satisfied.

$$M = \frac{2}{\|w\|} = \frac{2}{\sqrt{1+1}} = \frac{2}{\sqrt{2}} = \sqrt{2}$$

# Non linearly separable case

- So far we assumed that a linear plane can perfectly separate the points
- But this is not usually the case
  - noise, outliers



Hard to solve (two minimization problems)

How can we convert this to a QP problem?

- Minimize training errors?

$$\min \underline{\underline{w^T w}}$$

$$\min \underline{\underline{\# \text{errors}}}$$

- Penalize training errors:

$$\min w^T w + C * (\# \text{errors})$$

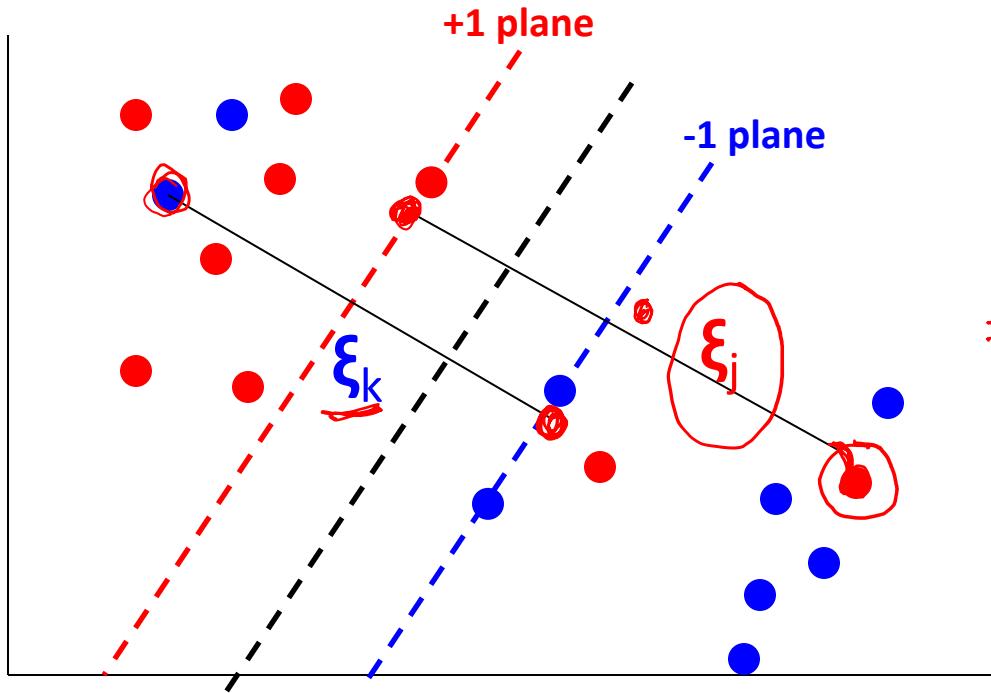
$$\overbrace{\hspace{10em}}^{\text{hyper param}}$$

Hard to encode in a QP problem

# Non linearly separable case

- Instead of minimizing the number of misclassified points we can minimize the *distance* between these points and their correct plane

$\xi$  slack var.



The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \xi_i$$

subject to the following inequality constraints:

For all  $x_i$  in class +1 or class -1

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

Wait. Are we missing something?

# Final optimization for non linearly separable case

The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \xi_i$$

subject to the following inequality constraints:

For all  $x_i$  in class +1  
or class -1

$$y^*(w^T x + b) \geq 1 - \xi_i$$

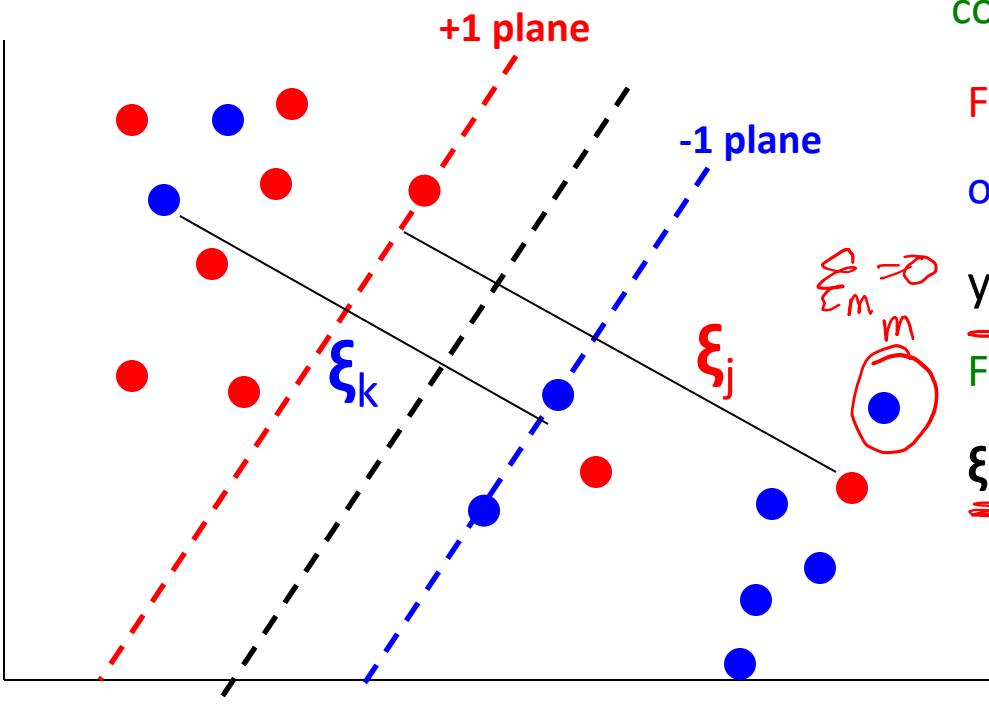
For all i

$$\xi_i \geq 0$$

n constraints

Another n constraints

Slack variables ( $\xi$ )



# Where we are

Two optimization problems: For the separable and non separable cases

$$\min_w \frac{w^T w}{2}$$

For all  $x$  in class +1

$$w^T x + b \geq 1$$

For all  $x$  in class -1

$$w^T x + b \leq -1$$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \xi_i$$

For all  $x_i$  in class +1

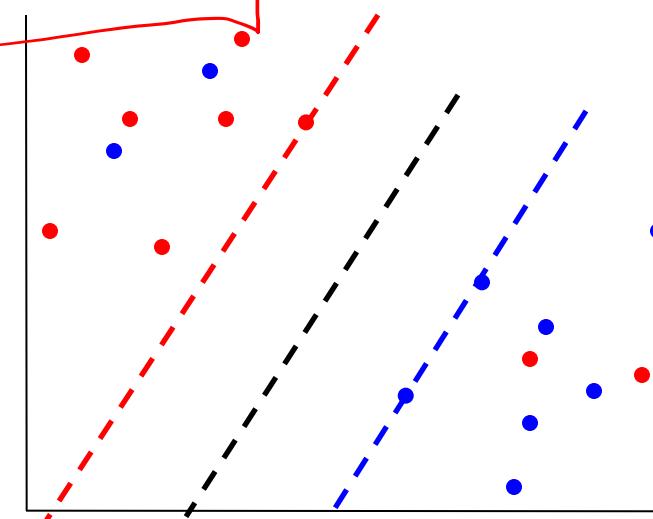
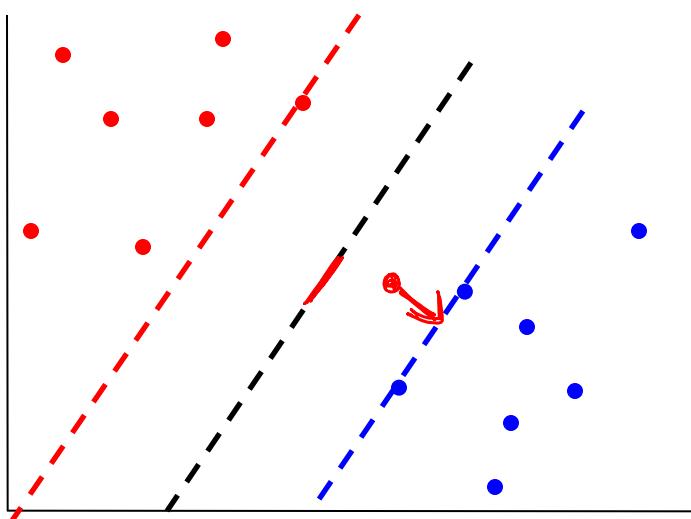
$$w^T x_i + b \geq 1 - \xi_i$$

For all  $x_i$  in class -1

$$w^T x_i + b \leq -1 + \xi_i$$

For all  $i$

$$\xi_i \geq 0$$



# Where we are

Two optimization problems: For the separable and non separable cases

$$\min_w \frac{w^T w}{2}$$

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \xi_i$$

For all  $x_i$  in class + 1

$$w^T x + b \geq 1 - \xi_i$$

For all  $x_i$  in class - 1

$$w^T x + b \leq -1 + \xi_i$$

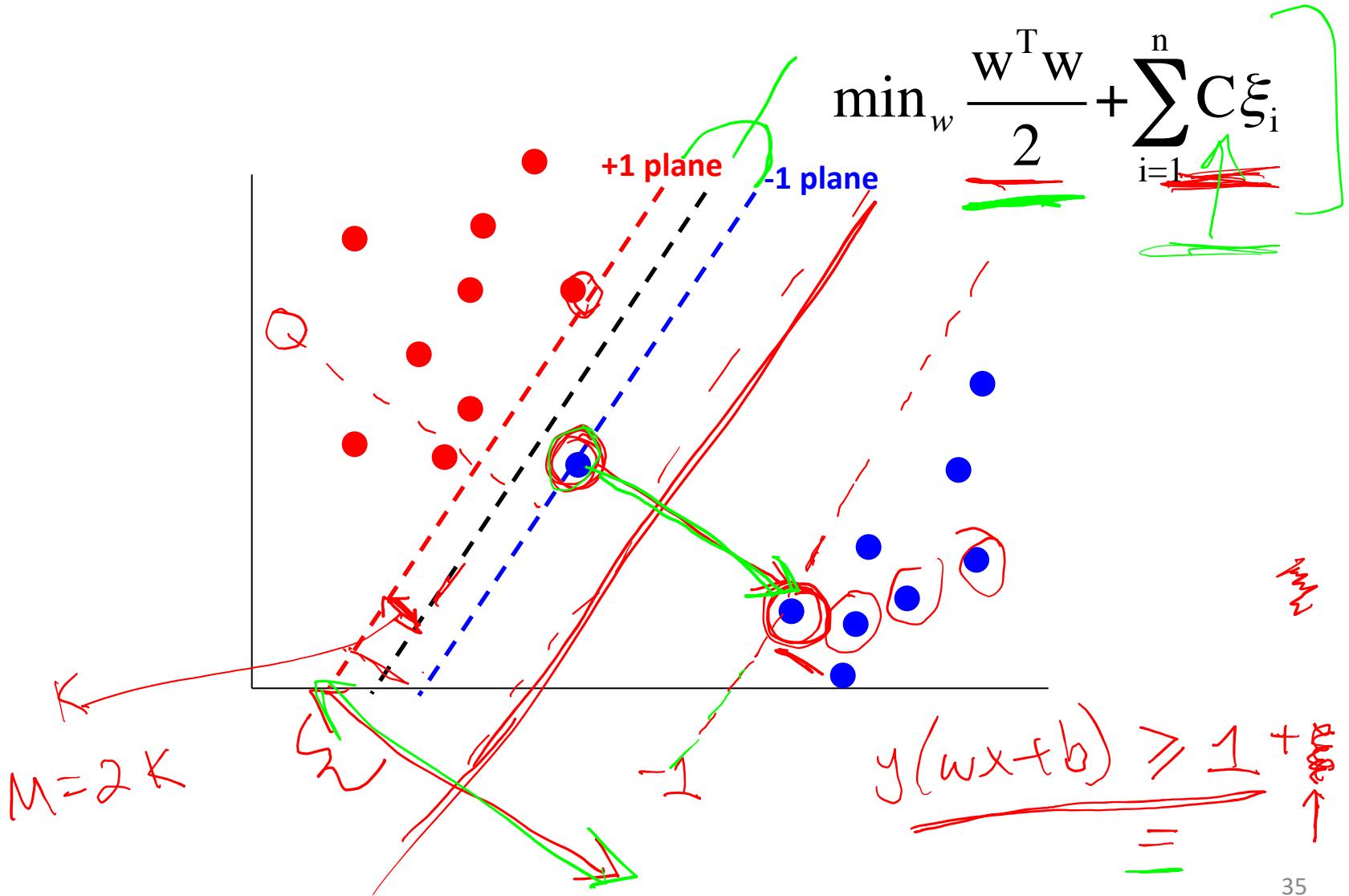
For all  $i$

$$\xi_i \geq 0$$

- Instead of solving these QPs directly we will solve a **dual** formulation of the SVM optimization problem
- One reason for switching to this type of representation (the dual formulation) is that it allows us to use a neat trick that will make our lives easier (and the run time faster)

**Can you think of a case where the data is linearly separable, but a slack variable will be non-zero?**

Can you think of a case where the data is linearly separable, but a slack variable will be non-zero?



# **Some example code**

[https://colab.research.google.com/drive/1QRxaGk\\_YZaKeO0Q3S42NKdi4g2dGuTD1?usp=sharing](https://colab.research.google.com/drive/1QRxaGk_YZaKeO0Q3S42NKdi4g2dGuTD1?usp=sharing)

# SVM Resources

- Nice tutorials
  - <http://cs229.stanford.edu/notes/cs229-notes3.pdf>
  - [https://www.youtube.com/watch?v=\\_PwhiWxHK8o](https://www.youtube.com/watch?v=_PwhiWxHK8o)
- Visualize the Kernel Trick
  - <https://www.youtube.com/watch?v=3liCbRZPrZA>
- The ~~textbook~~ has a nice section on SVMs