Soft-Margin Support Vector Machines Part 1

Support Vector Machines

Advantages:

- "Best off-the-shelf classifier" - Andrew Ng
- Nice theoretical bounds
- Allows for nonlinear classification
- Optimization problem for learning parameters is convex

Disadvantages:

- No probabilistic interpretation
- Can be prone to overfitting in the nonlinear case

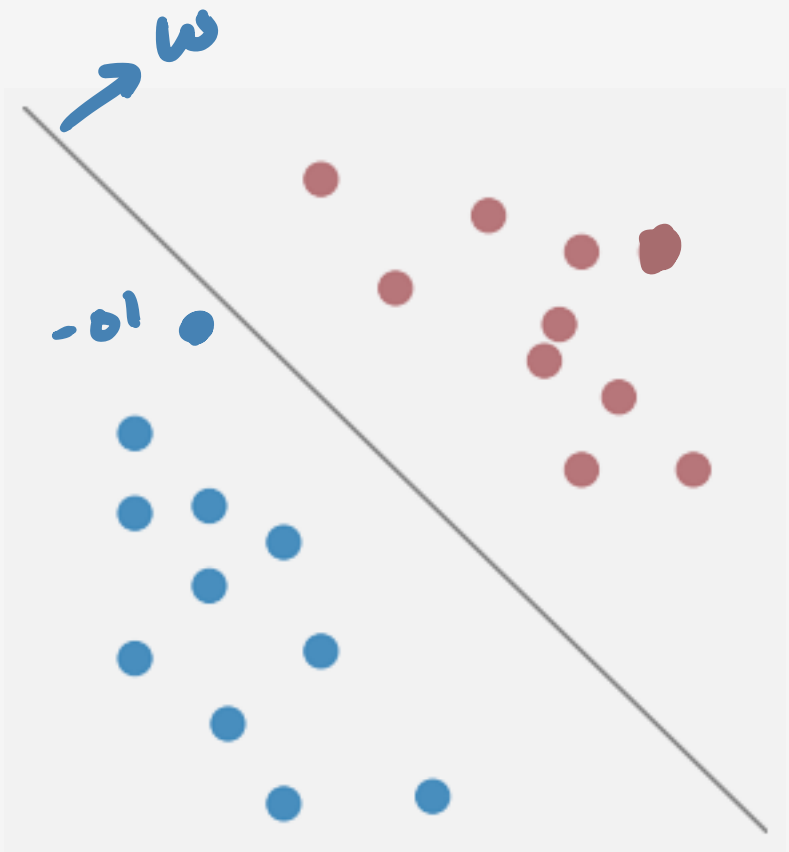
SVM Intuition

Temporary Assumption: Assume the training data is linearly separable

Consider the following training set in 2D

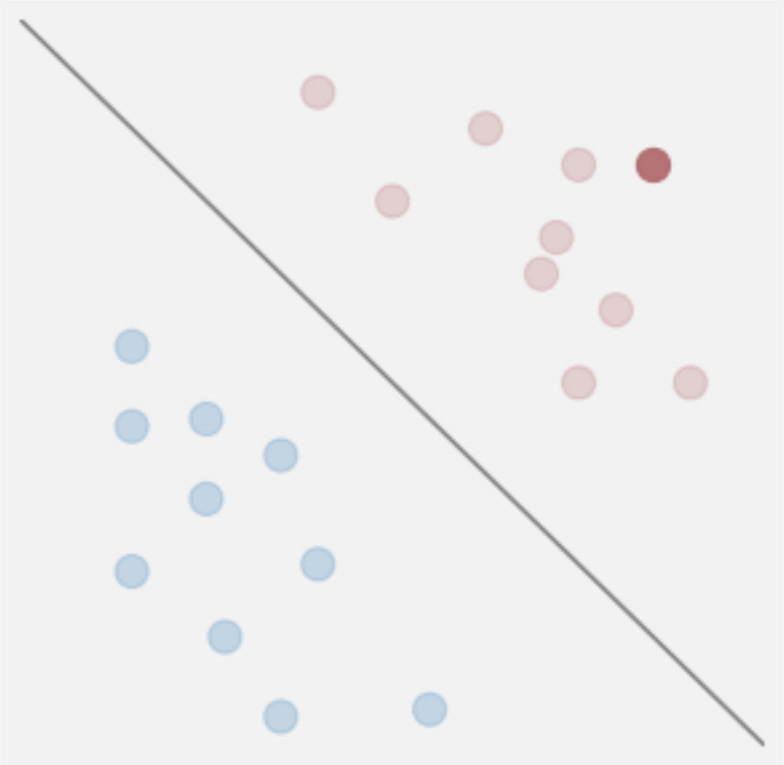
SVM Intuition

Think about the size of $\mathbf{w}^T\mathbf{x}_i + b$ as a measure of confidence in classification of example $\mathbf{x}_i$



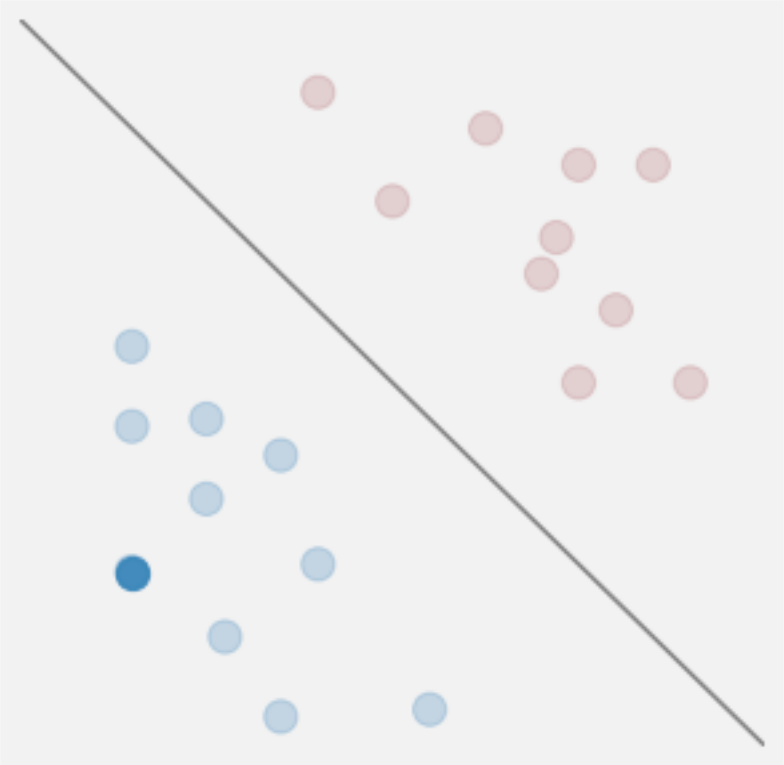
SVM Intuition

Very confident $y_i = 1$ if $\mathbf{w}^T \mathbf{x}_i + b \gg 0$



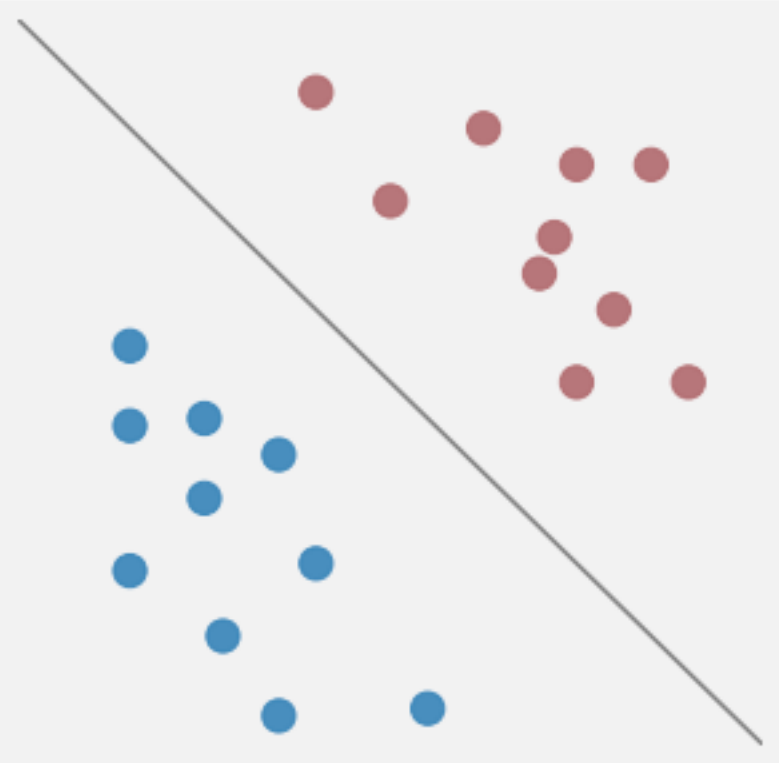
SVM Intuition

Very confident $y_i = -1$ if $\mathbf{w}^T \mathbf{x}_i + b \ll 0$



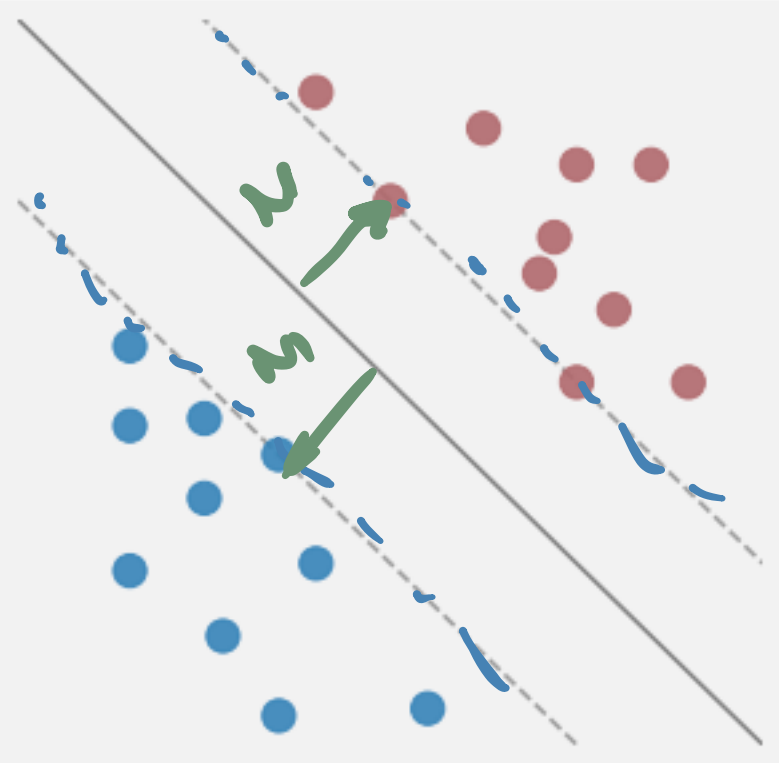
SVM Intuition

Want a DB that makes us very confident in the classification of each training example



SVM Intuition

Do this by maximizing distance from DB to *closest* training examples

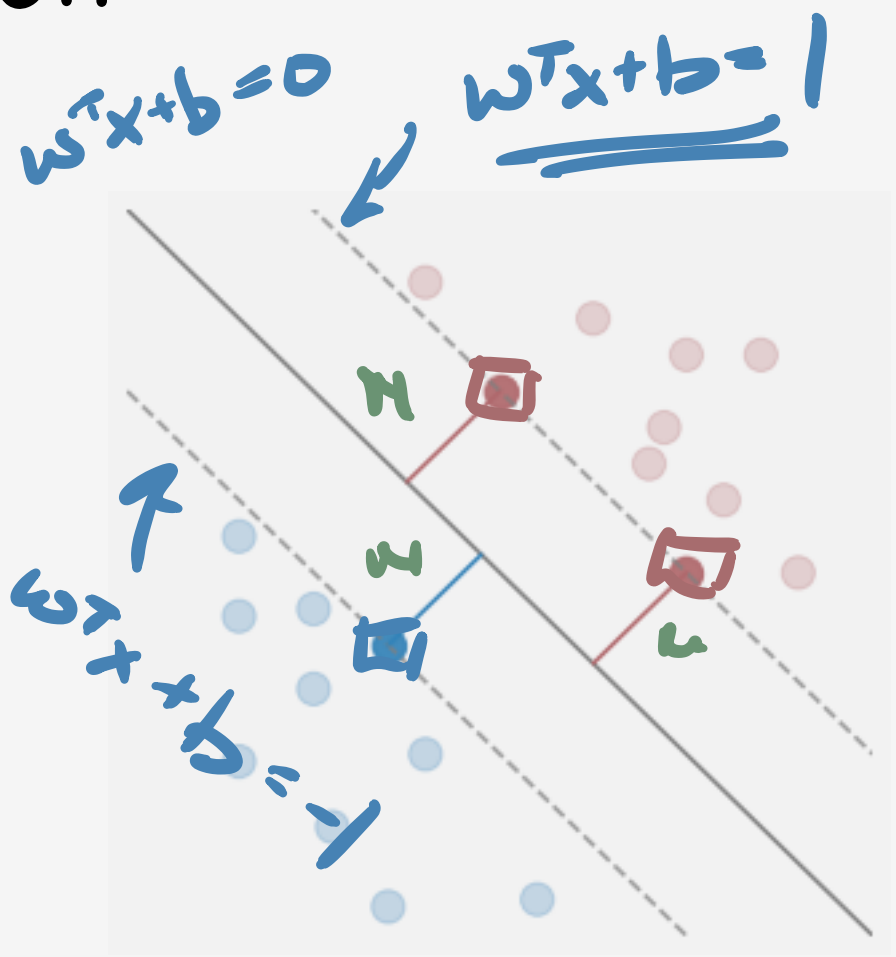


SVM Intuition

Maximizing distance from DB to *closest* training examples

Distance to closest training examples is called the **Margin**

Points closest to the DB are called the **support vectors**



Maximum Margin Classifier

CONSTRAINED OPT

Goal:

$$\min_{\mathbf{w},\, b} \quad \|\mathbf{w}\|_2^2$$

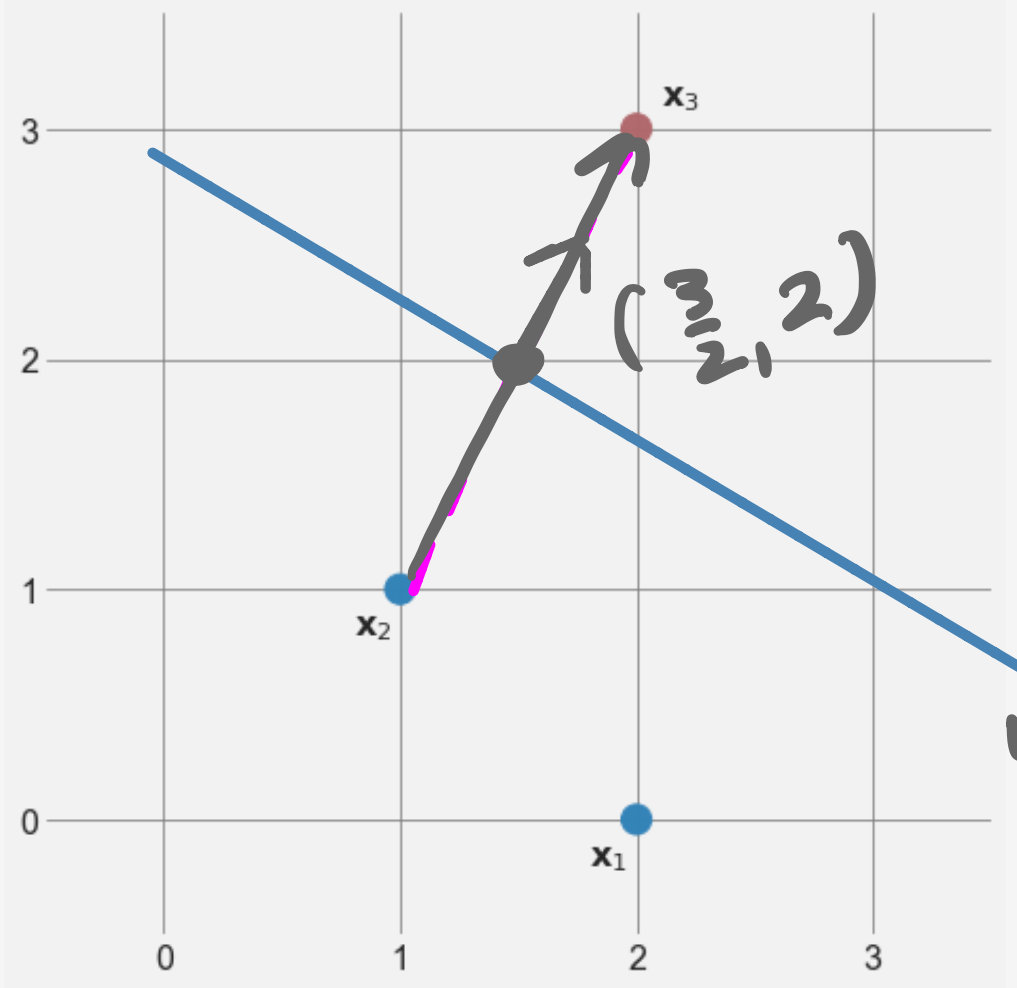
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \text{ for } i = 1, \ldots, m$$

This is a quadratic objective function with linear inequality constraints. Tons of canned software.

$$\|\vec{w}\|_2^2 = w_1^2 + w_2^2 + \cdots + w_p^2$$

Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



$x_3 = (2,3) \quad x_2 = (1,1)$

$w = \begin{bmatrix} 2-1 \\ 3-1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$x_m = \frac{x_2 + x_3}{2} = \frac{(2,3)+(1,1)}{2}$

$= \left(\frac{3}{2}, 2\right)$

$w^T x + b = 0$

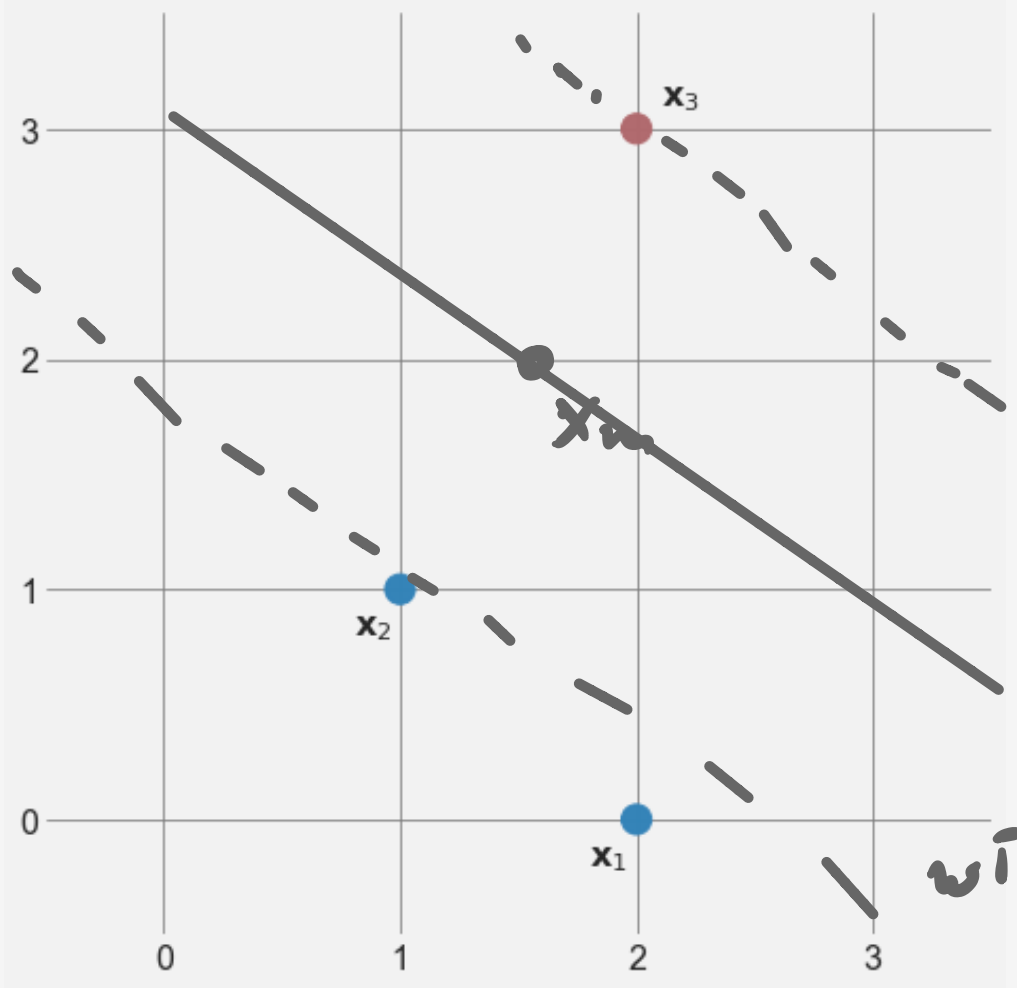
$\frac{3}{2} + 4 + b = 0$

$\Rightarrow$

$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3/2 \\ 2 \end{bmatrix} + b = 0 \Rightarrow b = -\frac{11}{2}$

Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



$w^Tx+b=1$

$w^Tx+b=-1$

guesses: $\vec{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $b = -\frac{11}{2}$

$x_3 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

$w^Tx_3+b =$

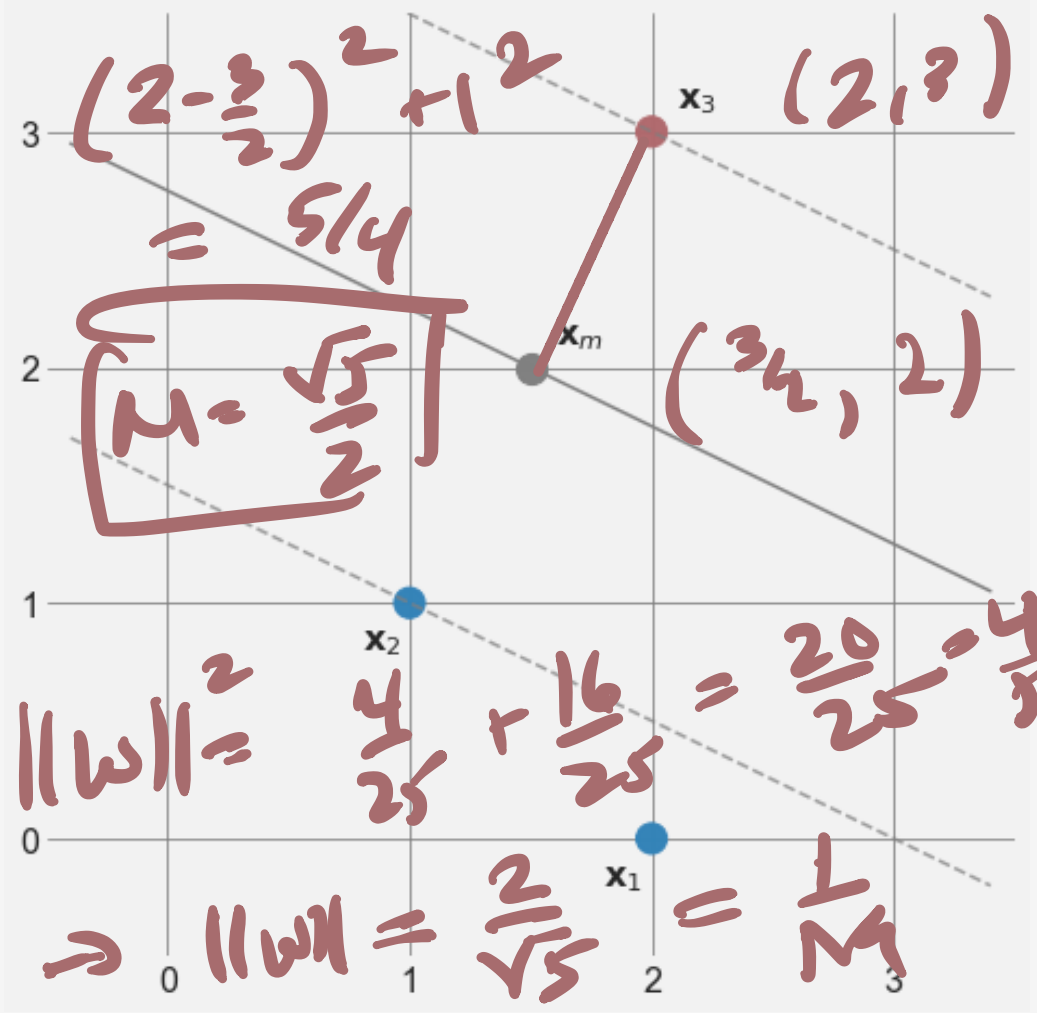
$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} - 11/2$

$= 2 + b - \frac{11}{2} = \frac{5}{2}$

SHOULD BE 1 BUT NOT

Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



$(2-\frac{3}{2})^2 + 1^2 = 5/4$

$M = \frac{\sqrt{5}}{2}$

$(3/2, 2)$

$\|w\|^2 = \frac{4}{25} + \frac{16}{25} = \frac{20}{25} = \frac{4}{5}$

$\Rightarrow \|w\| = \frac{2}{\sqrt{5}} = \frac{1}{M}$

$w \leftarrow \begin{bmatrix} 1/2 \\ 1 \end{bmatrix} / 5/2 \quad b \leftarrow -\frac{11}{2} / 5/2$

$w \leftarrow \begin{bmatrix} 2/5 \\ 4/5 \end{bmatrix} \quad b = -11/5$

$x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad w^T x_2 + b =$

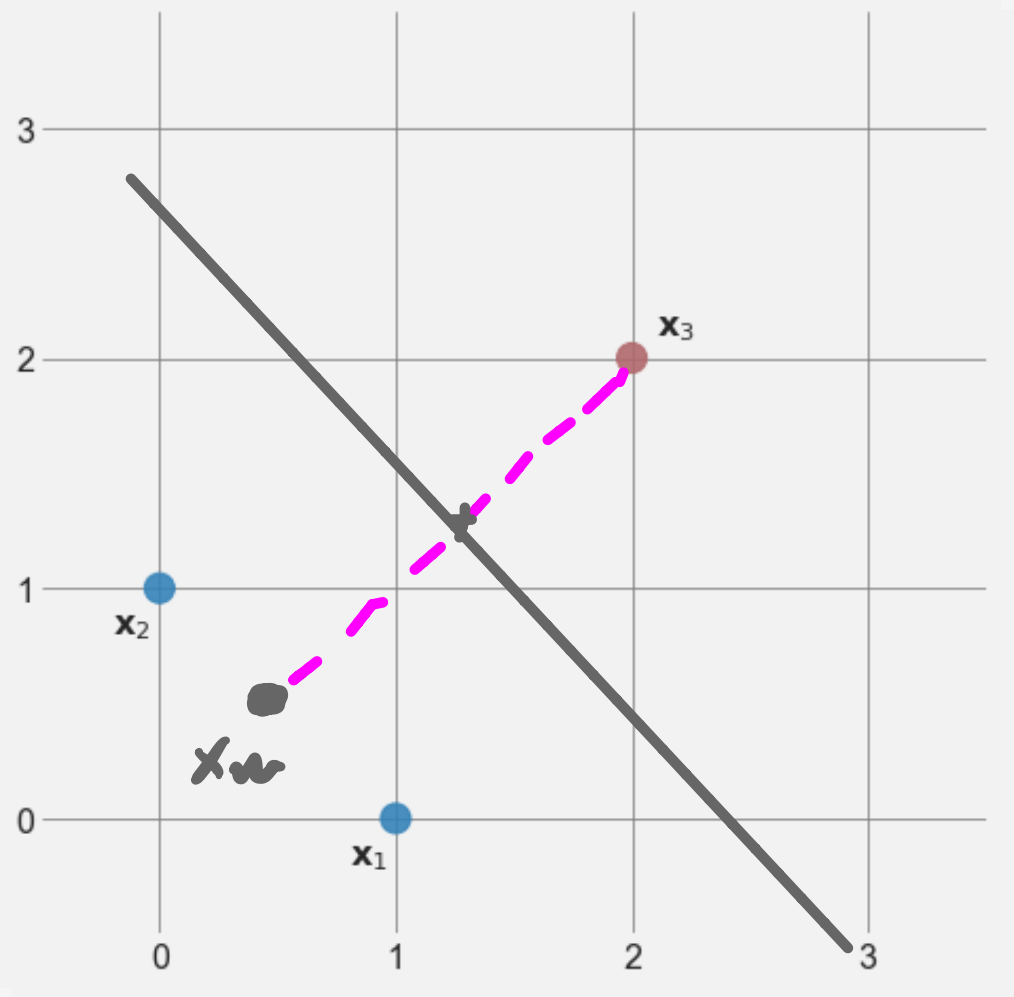
$\begin{bmatrix} 2/5 & 4/5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 11/5$

$= \frac{2}{5} + \frac{4}{5} - \frac{11}{5} = \frac{6}{5} - \frac{11}{5}$

$= -5/5 = -1$ SHOULD BE -1 AND IS!

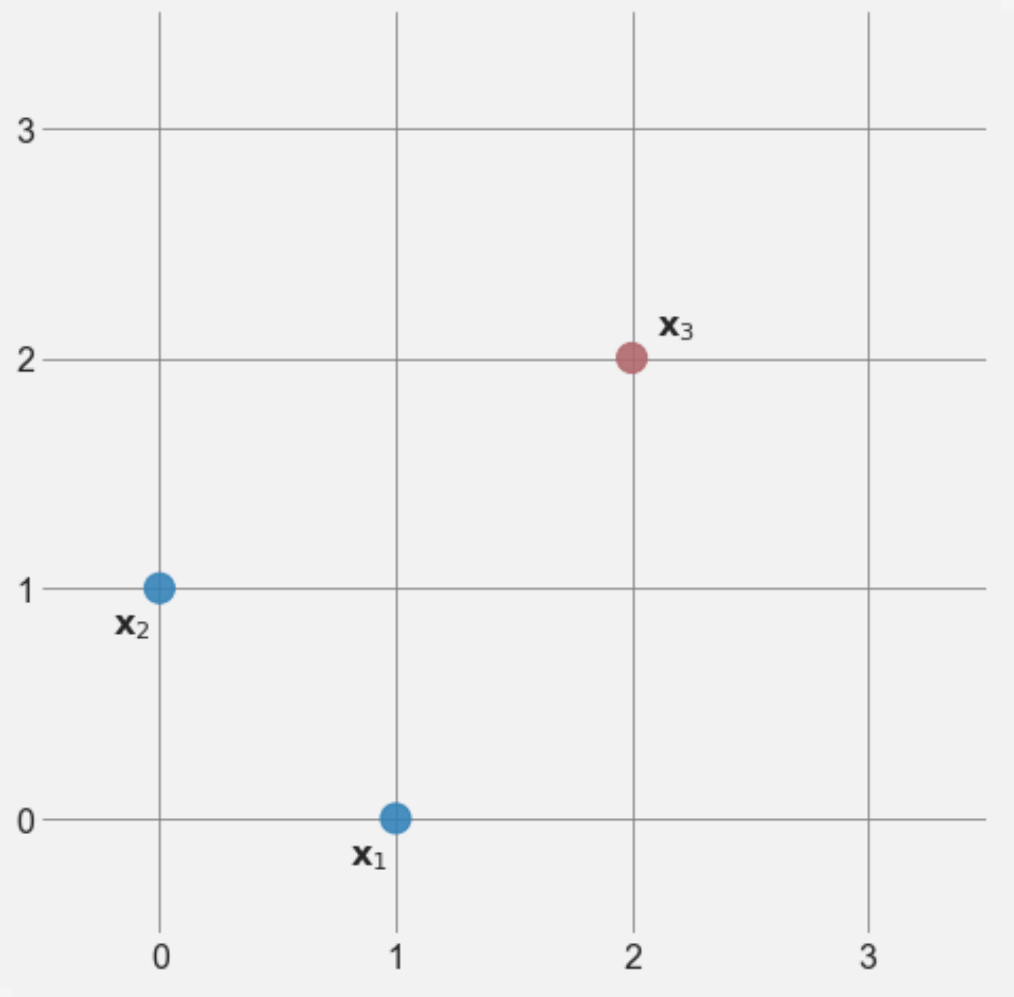
Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



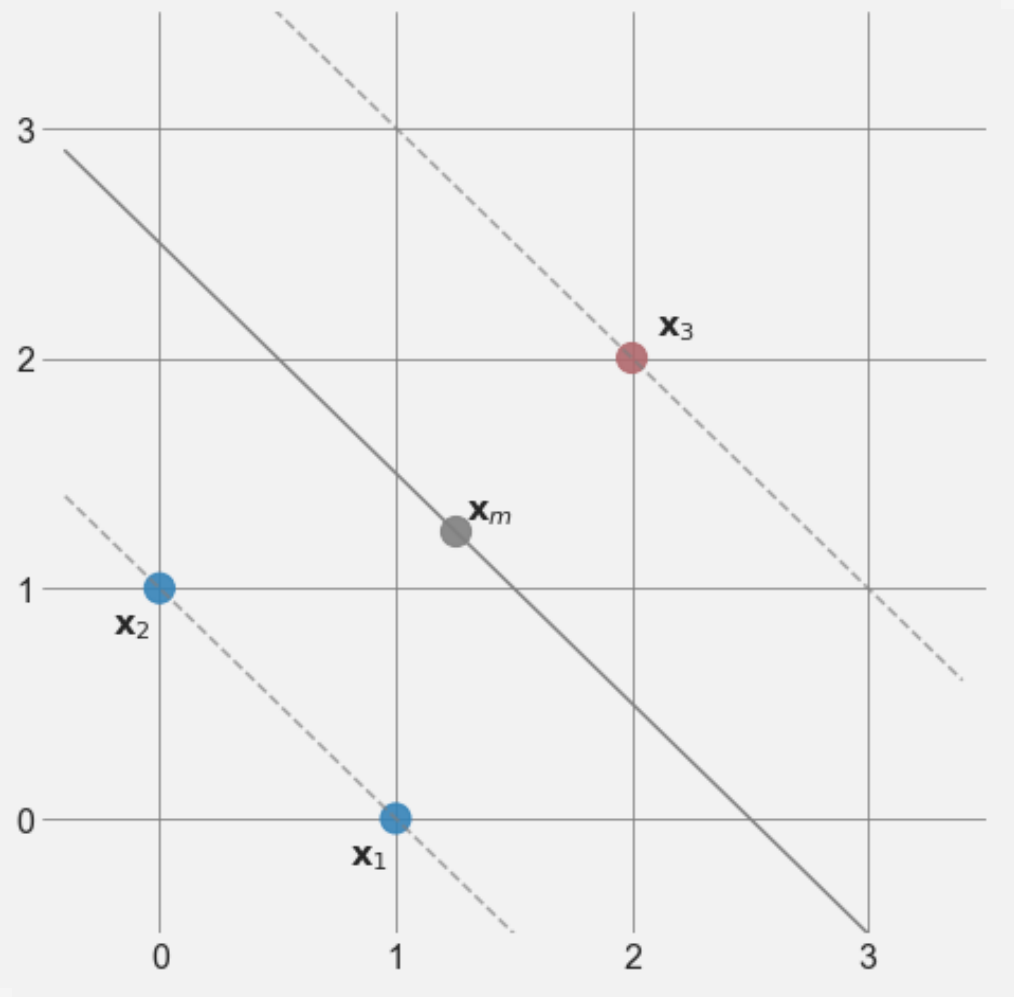
Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



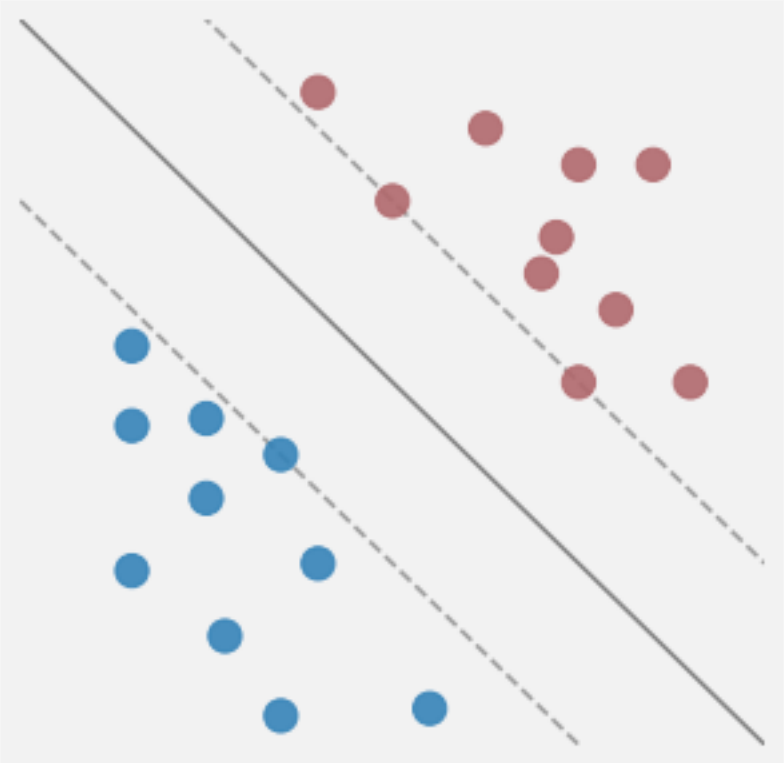
Hard-Margin SVM Example

Geometric Example: Find the Hard-Margin SVM of form $\mathbf{w}^T\mathbf{x} + b = 0$ for the following data



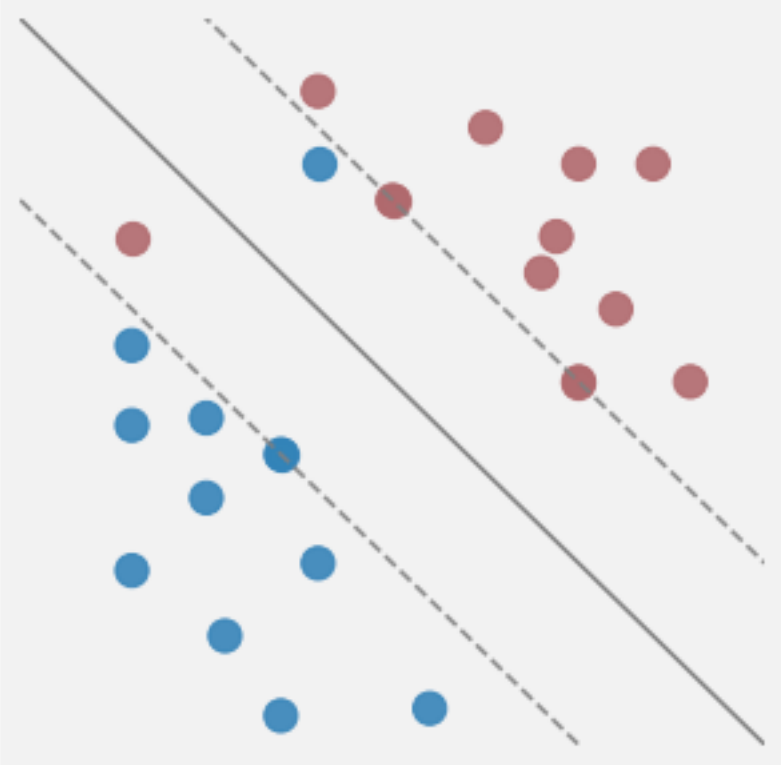
Soft-Margin SVM

OK, so far we've considered the linearly separable case



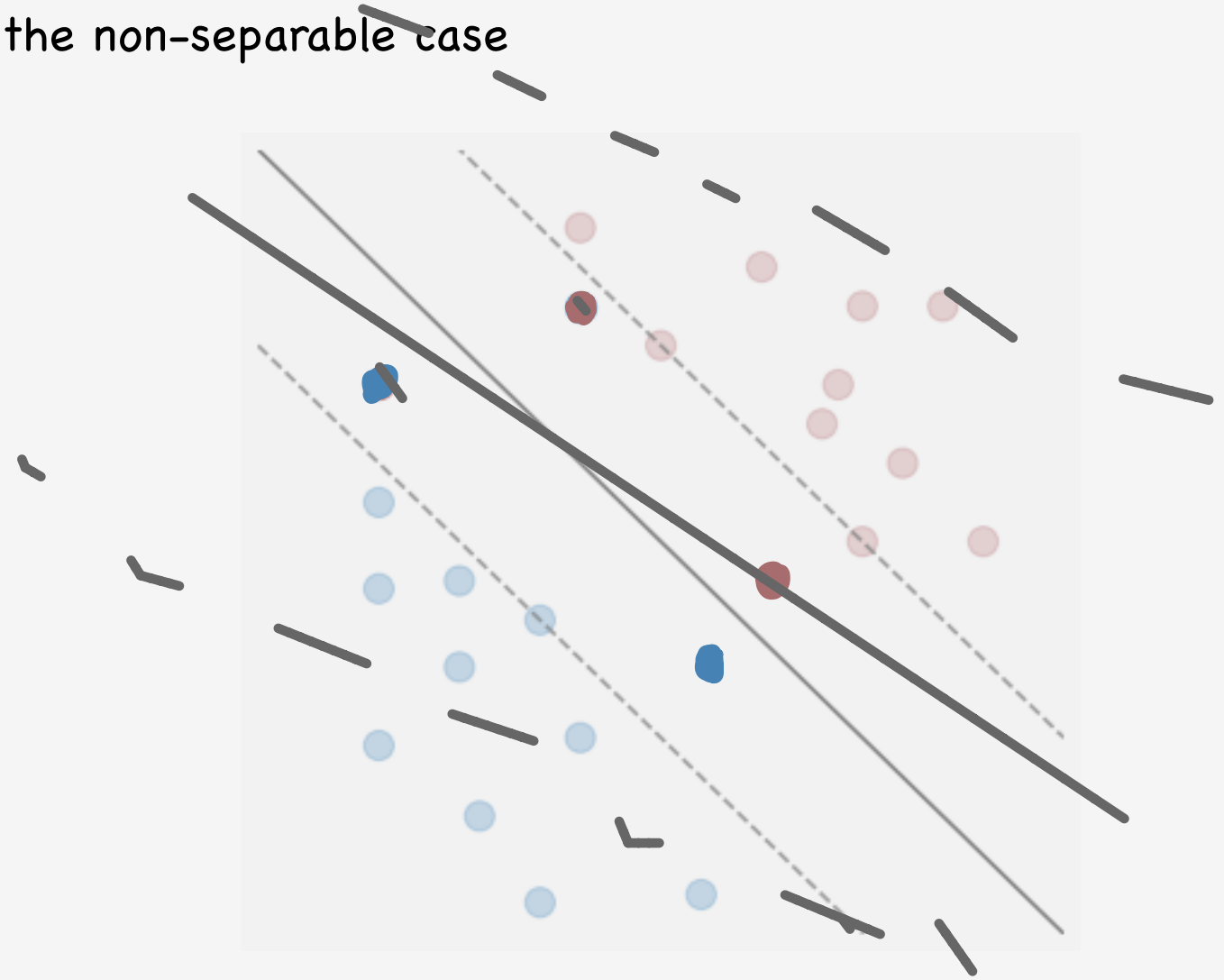
Soft-Margin SVM

Now we consider the non-separable case



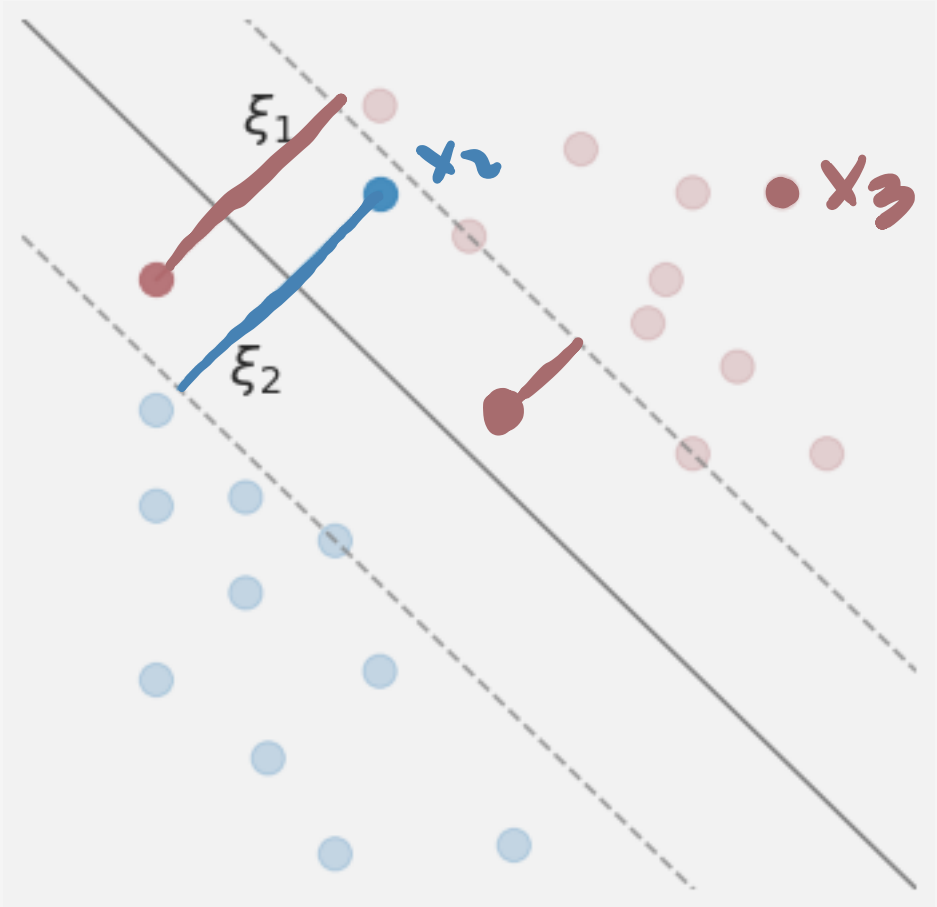
Soft-Margin SVM

Now we consider the non-separable case



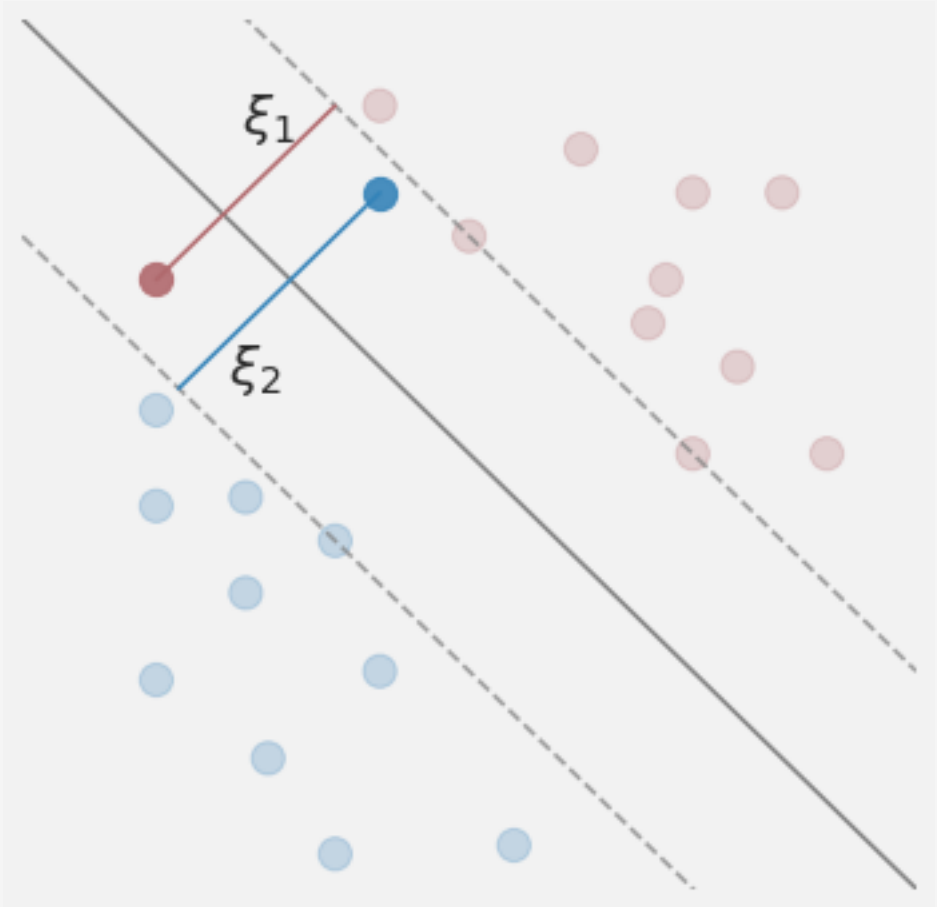
Soft-Margin SVM

Introduce nonnegative slack variable ξ_i associated with each training example



Soft-Margin SVM

Allow some $\xi_i > 0$ by hope that most are $\xi_i = 0$



Soft-Margin SVM

How does this change the mathematical landscape?

Objective Function:

$$\min_{\mathbf{w},\ b,\ \xi} \left(\frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{m} \xi_i^p \right)$$

- C is a tuning parameter that balances
  - Maximizing the margin
  - Classifying training examples correctly

Soft-Margin SVM

How does this change the mathematical landscape?

Objective Function:

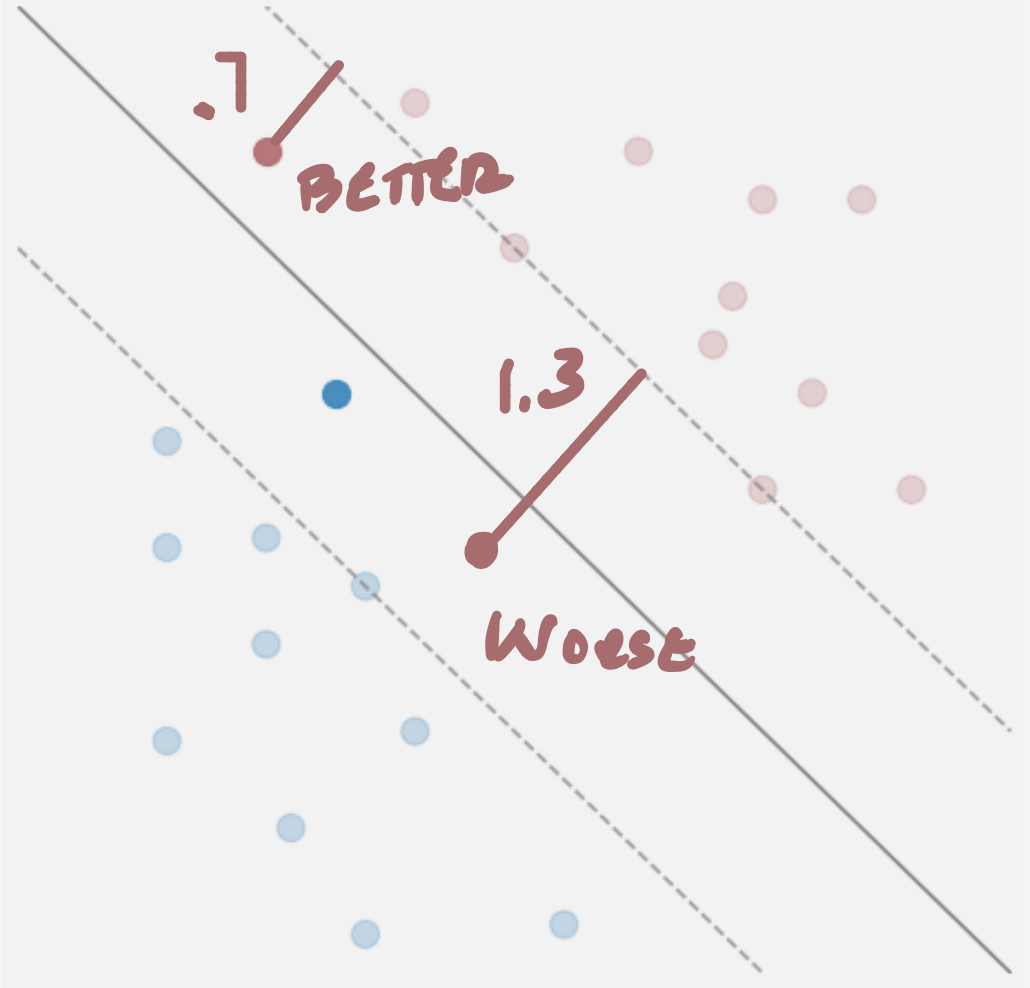
$$\min_{\mathbf{w},\ b,\ \xi} \left(\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i^p \right)$$

- C is a tuning parameter that balances

  Maximizing the margin

  Classifying training examples correctly

- Think of penalty term as regularization



Soft-Margin SVM

How does this change the mathematical landscape?

Objective Function:

$$\min_{\mathbf{w},\ b,\ \xi} \quad \left(\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i^p \right)$$

- Exponent p controls how bad *wrongness* of point scales
- We'll choose p=1 but other values are popular as well

Soft-Margin SVM

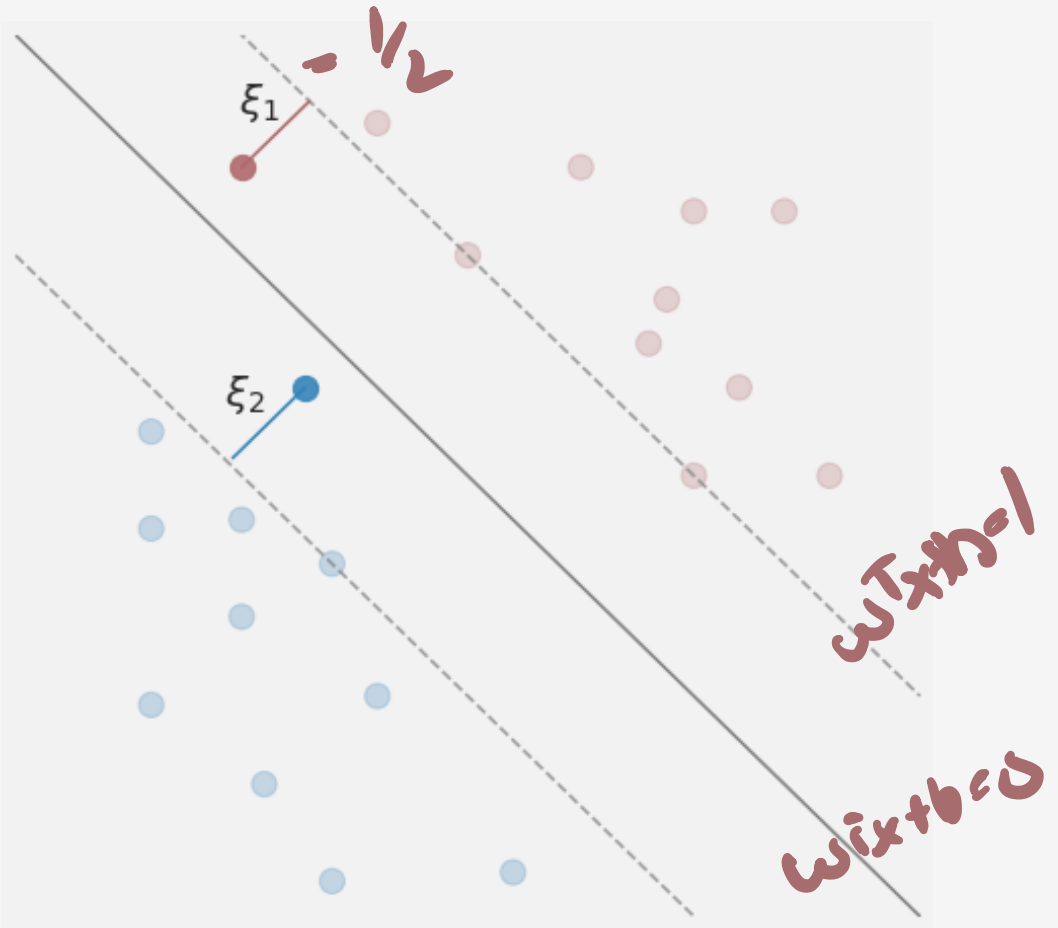
How does this change the mathematical landscape?

Constraints:

$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ becomes $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$

Intuition:

- $\xi_i = 0$: at least one margin on **correct** side of DB
- $\xi_i = 1/2$: one half margin on **correct** side of DB
- $\xi_i = 1$: **on** the DB
- $\xi_i = 2$: one margin on **wrong** side of DB



Soft-Margin SVM

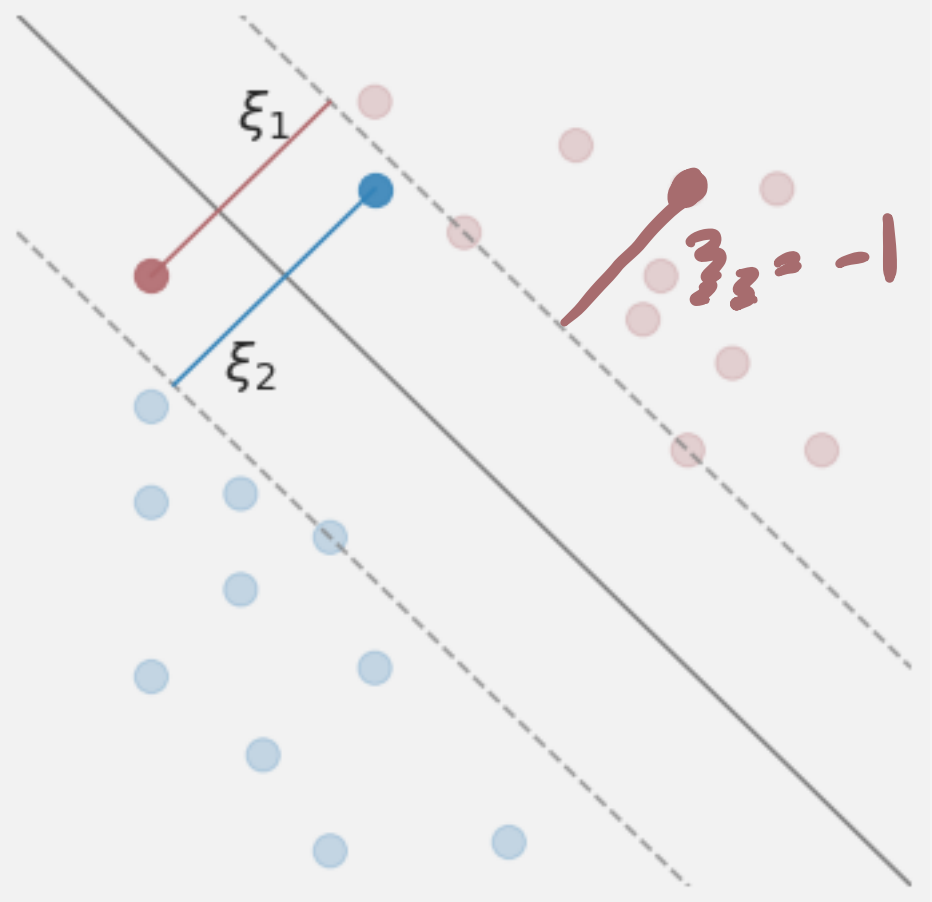
How does this change the mathematical landscape?

Constraints:

$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ becomes $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$

Intuition:

- $\xi_i = 0$: at least one margin on **correct** side of DB
- $\xi_i = 1/2$: one half margin on **correct** side of DB
- $\xi_i = 1$: **on** the DB
- $\xi_i = 2$: one margin on **wrong** side of DB



Soft-Margin SVM

Primal Optimization Problem:

$$\min_{\mathbf{w},\ b,\ \xi} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{for } i = 1, \ldots, m$$

$$\text{s.t.} \quad \xi_i \geq 0 \quad \text{for } i = 1, \ldots, m$$

- We can solve this problem via canned quadratic program solvers
- But it turns out, there are better things to do when we get to the nonlinear case …

Convex Opt. with Inequality Constraints

Consider the following general problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 \text{ for } i = 1, \ldots, n \end{aligned}$$

Define a modified objective function called the **Lagrangian**:

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

The parameters α_i associated with each constraint are called the **Lagrange Multipliers**

Notice that:

$$\max_{\alpha} L(\mathbf{w}, \alpha) = \begin{cases} f(\mathbf{w}) & \text{if } \mathbf{w} \text{ is feasible} \\ \infty & \text{if } \mathbf{w} \text{ is not feasible} \end{cases}$$

Convex Opt. with Inequality Constraints

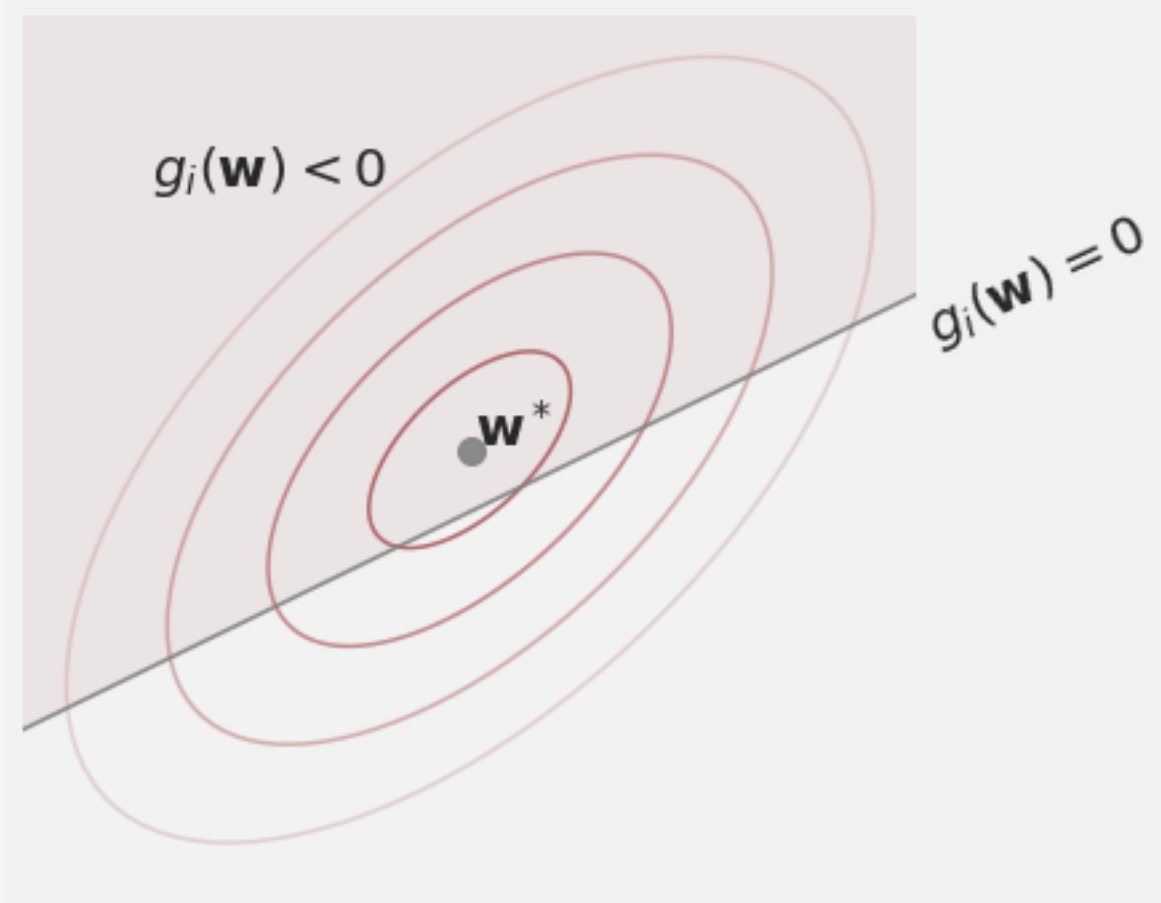
Active and Inactive constraints: What do the Lagrange Multipliers actually do?

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

Suppose the $\mathbf{w}^\star$ that minimizes $f(\mathbf{w})$ is feasible

We say that the constraint $g_i(\mathbf{w}) \leq 0$ is **inactive**

We can set the associated α_i to zero



Convex Opt. with Inequality Constraints

Active and Inactive constraints: What do the Lagrange Multipliers actually do?

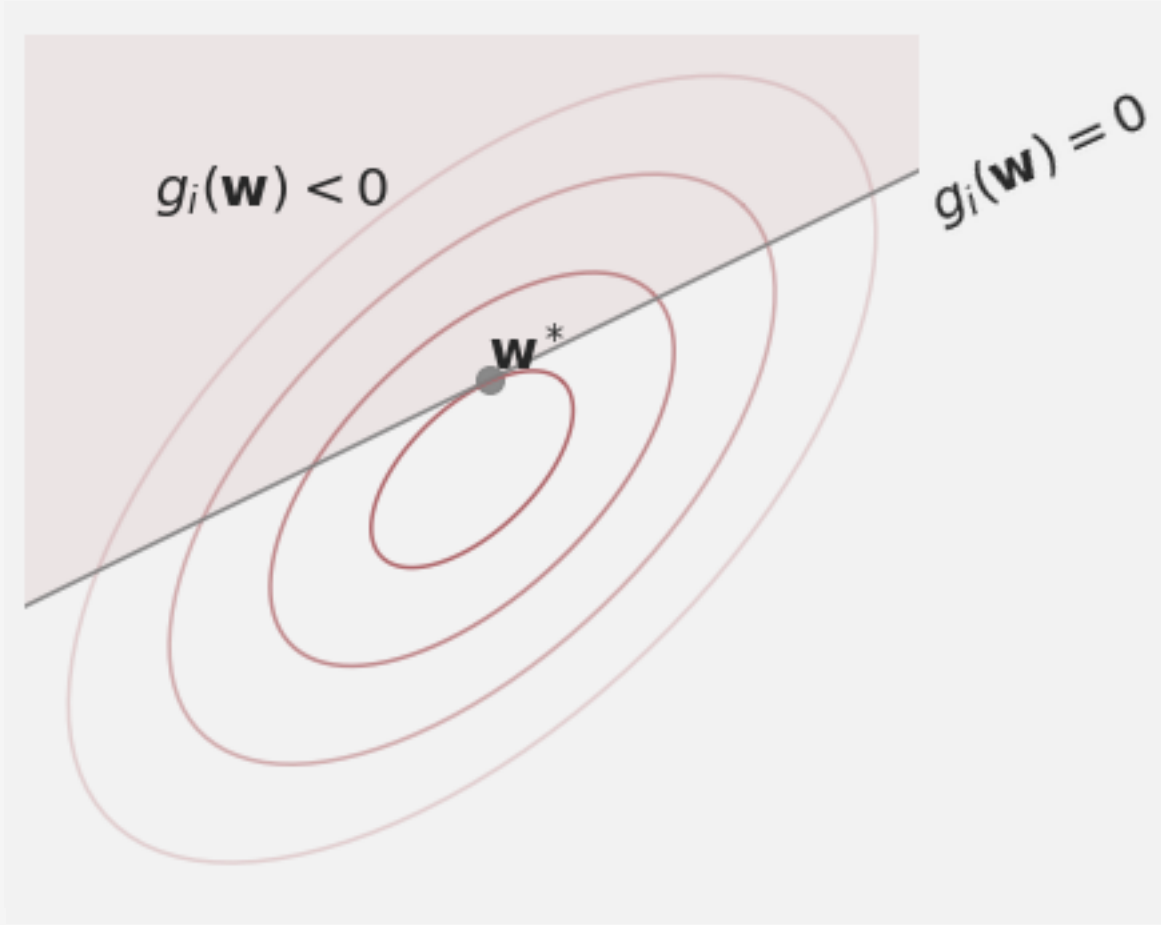
$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

Suppose the $\mathbf{w}^\star$ that minimizes $f(\mathbf{w})$ is infeasible

We say that the constraint $g_i(\mathbf{w}) < 0$ is **active**

Then the Lagrange multiplier satisfies $\alpha_i > 0$

Solution to constrained opt. problem changes



Convex Opt. with Inequality Constraints

These two ideas together give us a new constraint:

The **Complementary Slackness Condition**:

$$\alpha_i g_i(\mathbf{w}^\star) = 0 \text{ for } i = 1, \ldots, n$$

- If the minimizer is on an inequality boundary ($g_i(\mathbf{w}^\star) = 0$) then $\alpha_i > 0$
- If the minimizer is not on an inequality boundary ($g_i(\mathbf{w}^\star) < 0$) then $\alpha_i = 0$

Convex Opt. with Inequality Constraints

The original **Primal problem** was

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 \text{ for } i = 1, \ldots, n \end{aligned}$$

and the Lagrangian formulation becomes

$$\min_{\mathbf{w}} \quad \max_{\alpha} \left[f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w}) \right]$$

$$\text{s.t.} \quad \alpha_i \geq 0 \text{ for } i = 1, \ldots, n$$

$$\text{s.t.} \quad \alpha_i g_i(\mathbf{w}) \geq 0 \text{ for } i = 1, \ldots, n$$

Convex Opt. with Inequality Constraints

Under certain conditions we can swap the order of the min and max, to get the **Dual problem**

$$\max_{\alpha} \quad \min_{\mathbf{w}} \left[f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w}) \right]$$

$$\text{s.t.} \quad \alpha_i \geq 0 \text{ for } i = 1, \ldots, n$$

$$\text{s.t.} \quad \alpha_i g_i(\mathbf{w}) \geq 0 \text{ for } i = 1, \ldots, n$$

The nice thing about the Dual problem is that we can minimize wrt to w first

This will make amazing things happen when we apply this theory to SVMs