Soft-Margin Support Vector Machines Part 2

Support Vector Machines

Advantages:

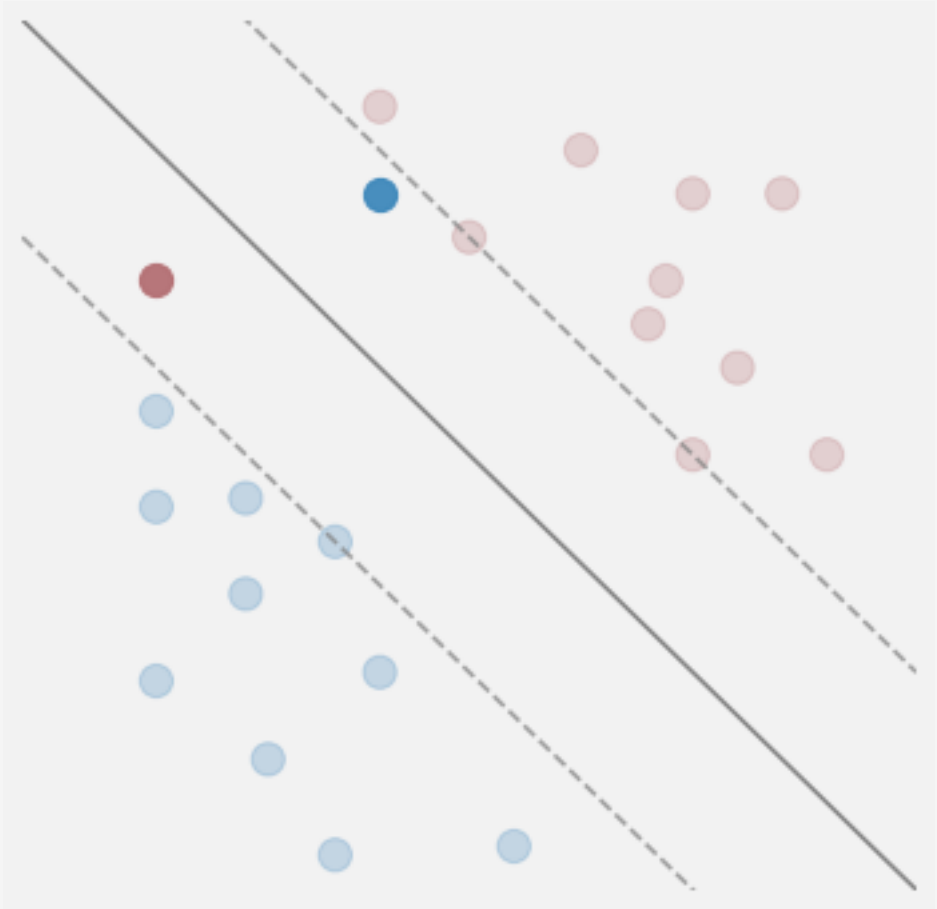
- "Best off-the-shelf classifier" - Andrew Ng
- Nice theoretical bounds
- Allows for nonlinear classification
- Optimization problem for learning parameters is convex

Disadvantages:

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

Soft-Margin SVM

When we left off we were considering the non-separable case



Soft-Margin SVM

How does this change the mathematical landscape?

Objective Function:

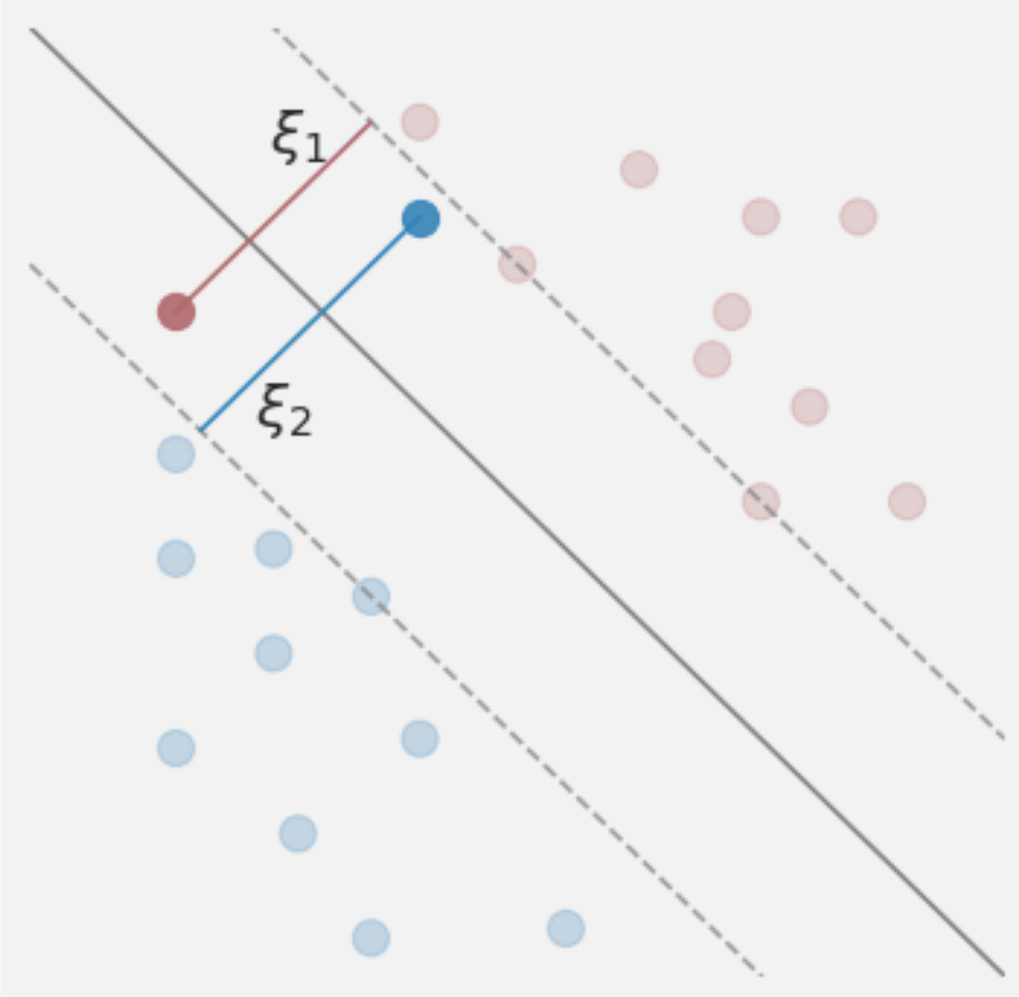
$$\min_{\mathbf{w},\ b,\ \xi} \left(\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i^p \right)$$

- 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

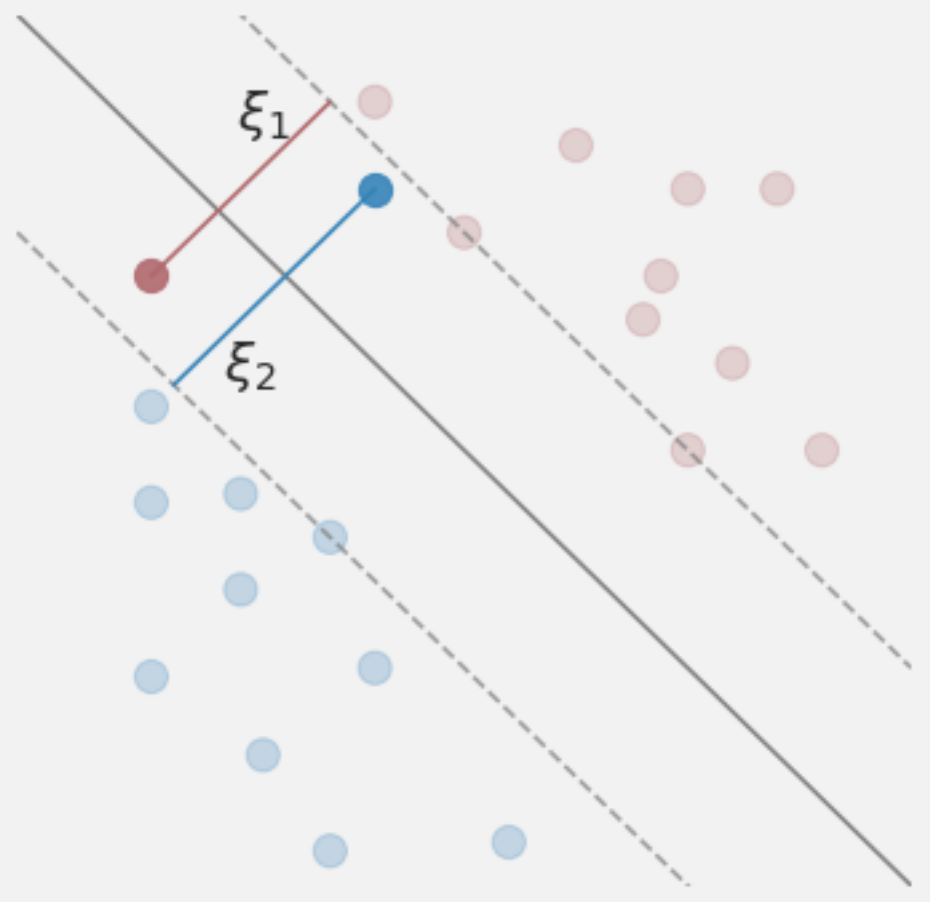
How does this change the mathematical landscape?

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

Primal Optimization Problem:

PRIMAL VARIABLES

$\vec{w}, b, \xi_i$

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

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

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

- 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

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

$\vec{w}$ FEASIBLE $\Leftrightarrow$ $g_i(w) \leq 0$

$\vec{w}$ NOT FEASIBLE $\Leftrightarrow$ $g_i(w) > 0$ FOR SOME i

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** ($\alpha_i \geq 0$)

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?

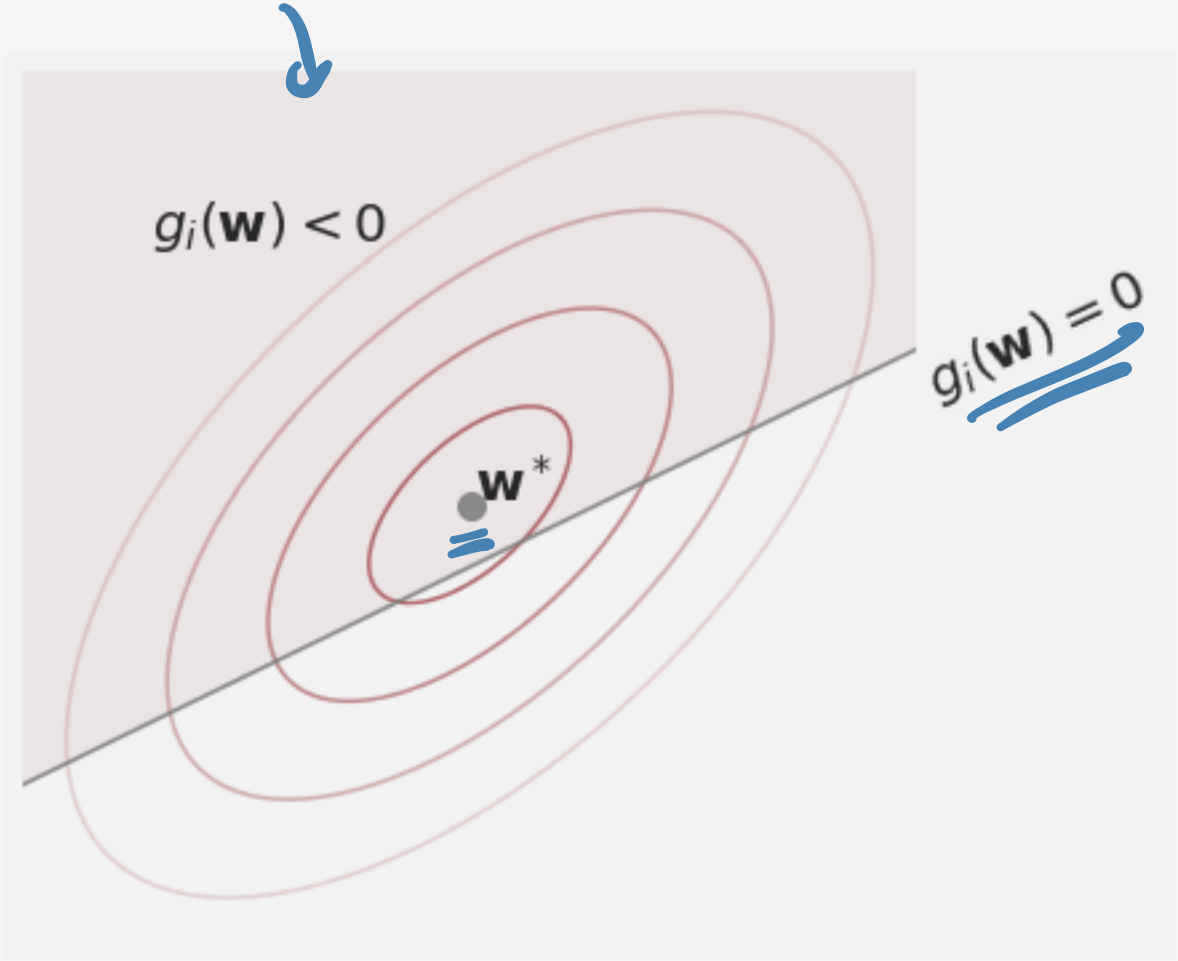
$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

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

$\alpha_i = 0$



Convex Opt. with Inequality Constraints

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

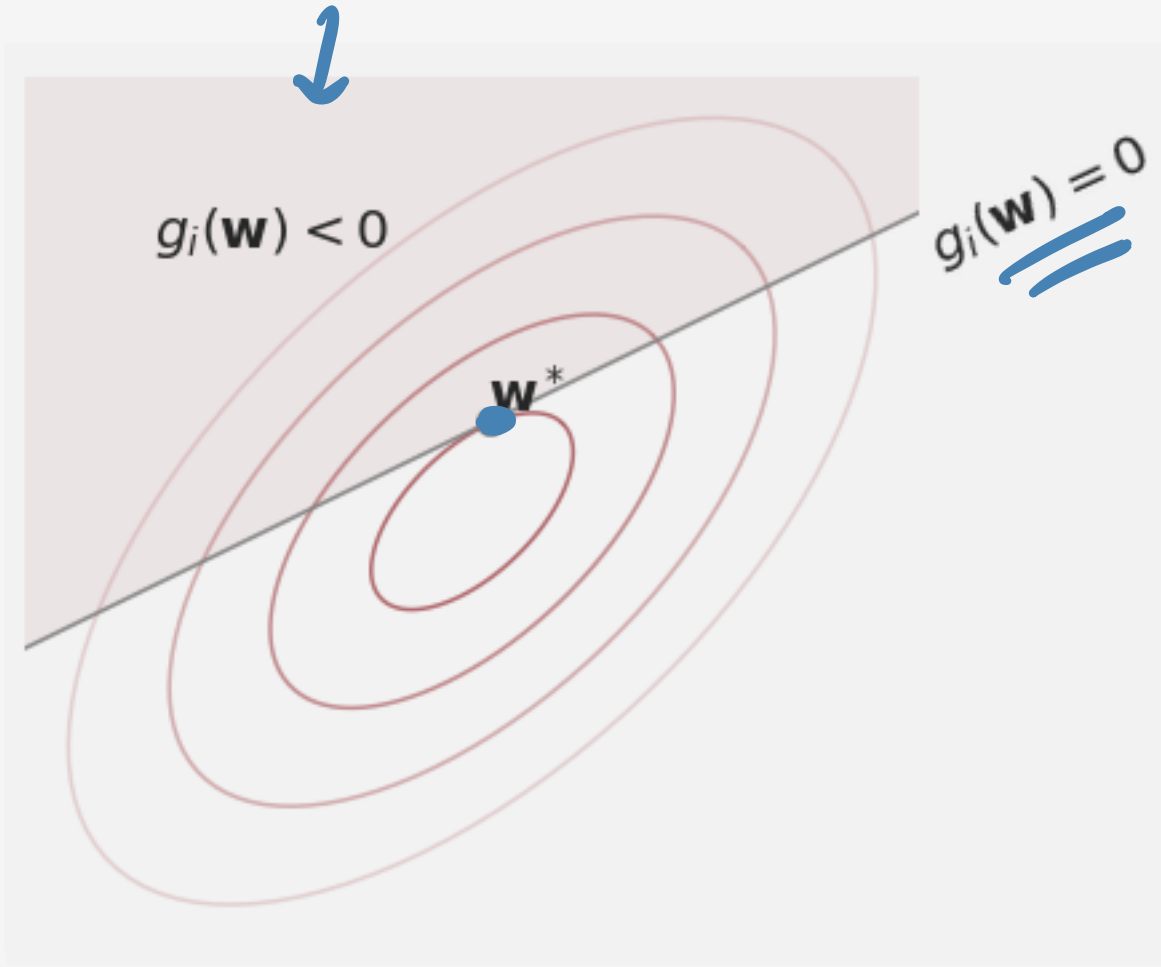
$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

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

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

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

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

Under certain circumstances, we can switch the order of the max and min

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

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

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

Under certain circumstances, we can switch the order of the max and min

Convex Opt. with Inequality Constraints

The Lagrangian formulation is

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

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

Now, let's make some progress by solving in the inner minimization problem first. How?

$\nabla_w L(w, \alpha) = 0 \Rightarrow$ solve

Convex Opt. with Inequality Constraints

The Lagrangian formulation is

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

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

Now, let's make some progress by solving in the inner minimization problem first. How?

We can find the optimal w by setting $\nabla_{\mathbf{w}} L(\mathbf{w}, \alpha) = 0$ and solving for w

Let's define a new function to represent what's left after the minimization

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

Convex Opt. with Inequality Constraints

What's left after the dust settles is called the **Dual Problem**

DUAL VARIABLES

$$\begin{aligned} \max_{\alpha} \quad & h(\alpha) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ for } i = 1, \ldots, n \end{aligned}$$

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

Unsurprisingly, the solution to the Dual problem is found when $\nabla_{\alpha} L(\mathbf{w}, \alpha) = 0$. So we have:

Primal Problem

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

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \alpha) = 0$$

Dual Problem

$$\begin{aligned} \max_{\alpha} \quad & h(\alpha) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ for } i = 1, \ldots, n \end{aligned}$$

$$\nabla_{\alpha} L(\mathbf{w}, \alpha) = 0$$

There's just one thing missing, which we need to tie the two problems together

Convex Opt. with Inequality Constraints

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

Primal Problem

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

Dual Problem

$$\max_{\alpha} \quad h(\alpha)$$
$$\text{s.t.} \quad \alpha_i \geq 0 \text{ for } i = 1, \ldots, n$$

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \alpha) = 0 \qquad \alpha_i g_i(\mathbf{w}) = 0 \text{ for } i = 1, \ldots, n \qquad \nabla_{\alpha} L(\mathbf{w}, \alpha) = 0$$

The unique solution to both problems occurs when at $\mathbf{w}^\star$ and $\alpha^\star$ where all of the constraints are satisfied and the objective functions are minimized/maximized.

Def: The set of constraints for the two problems are called the **KKT Conditions**

The Dual Problem for SVM

Recall that the Soft-Margin SVM optimization problem is

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

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

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

The Dual Problem for SVM

Recall that the Soft-Margin SVM optimization problem is

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

$$\text{s.t.} \quad 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0 \quad \text{for } i = 1, \ldots, m$$

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

Let α_i and β_i be the Lagrange multipliers. Then the associated Lagrangian is

$$L = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i$$

The Dual Problem for SVM

$$L = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i$$

Get to the Dual Problem by taking derivatives wrt primal variables and setting equal to zero

$$0 = \frac{\partial L}{\partial w_k} = w_k - \sum_{i=1}^{m}\alpha_i y_i x_{ik} \Rightarrow \vec{w} = \sum_{i=1}^{m}\alpha_i y_i \vec{x}_i$$

$$0 = \frac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i(-y_i) \Rightarrow \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$0 = \frac{\partial L}{\partial \xi_k} = C - \alpha_i - \beta_i \Rightarrow \alpha_i + \beta_i = C \quad i = 1, \ldots, m$$

The Dual Problem for SVM

$$L = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i$$

Get to the Dual Problem by taking derivatives wrt primal variables and setting equal to zero

$$0 = \frac{\partial L}{\partial w_k} = w_k - \sum_{i=1}^{m}\alpha_i y_i x_{ik} \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$$

$$0 = \frac{\partial L}{\partial b} = -\sum_{i=1}^{m}\alpha_i y_i \quad \Rightarrow \quad \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$0 = \frac{\partial L}{\partial \xi_k} = C - \alpha_k - \beta_k \quad \Rightarrow \quad \alpha_k + \beta_k = C \text{ for } k = 1, \ldots, m$$

Eliminating the Primal Variables

$$\max_{\alpha} \left(\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1-\xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i \right)$$

Plug in the primal variables in terms of the dual variables obtained from the minimization

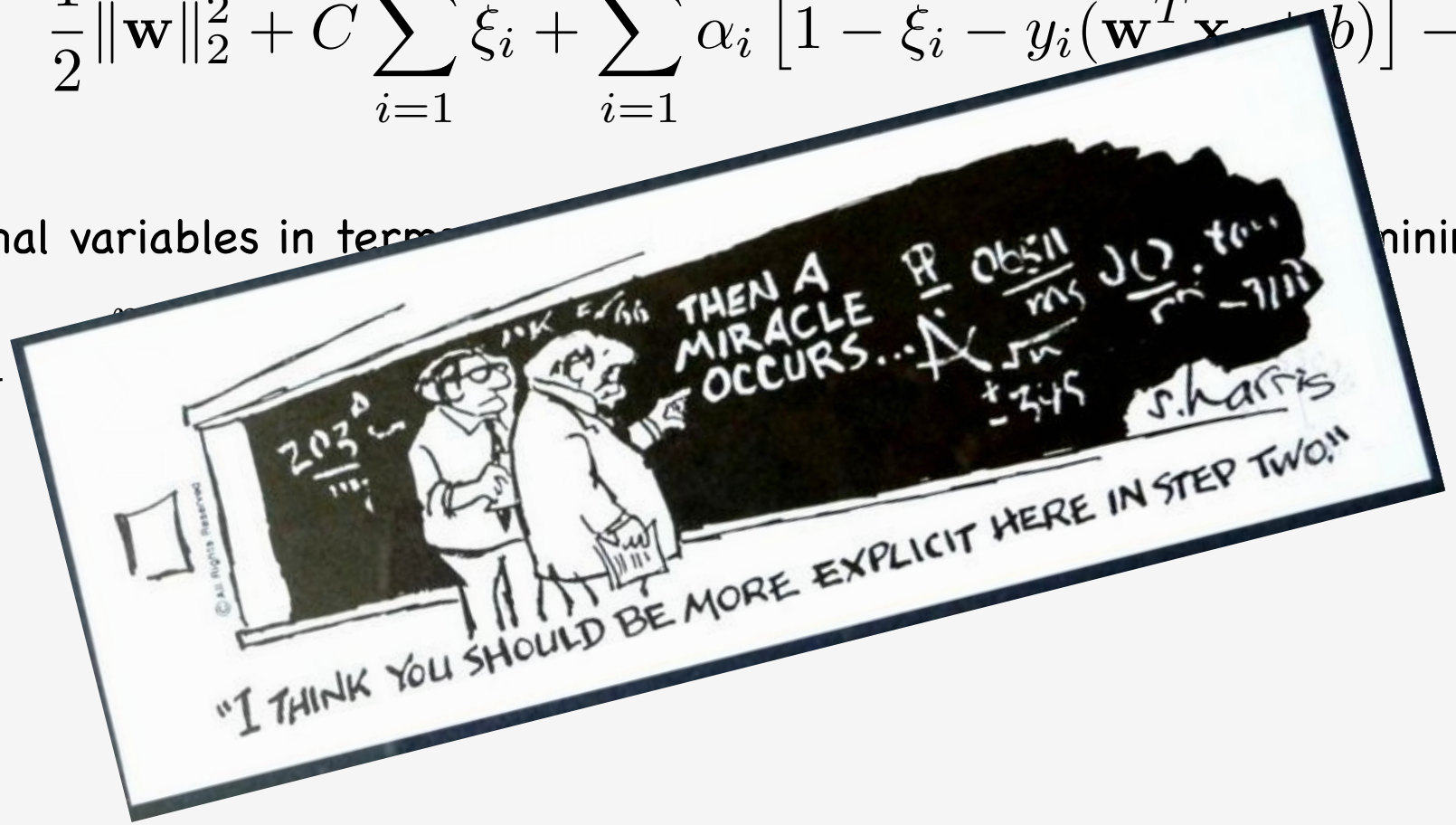
$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{m}\alpha_i y_i = 0 \qquad \alpha_i + \beta_i = C$$

Eliminating the Primal Variables

$$\max_{\alpha} \left(\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i \right)$$

Plug in the primal variables in term… minimization

$\mathbf{w}$ =



Eliminating the Primal Variables

$$\max_{\alpha}\left(\min_{\mathbf{w},b,\xi}\quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1-\xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] - \sum_{i=1}^{m}\beta_i\xi_i\right)$$

Plug in the primal variables in terms of the dual variables obtained from the minimization

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{m}\alpha_i y_i = 0 \qquad \alpha_i + \beta_i = C$$

And we obtain the following **Dual Objective Function**

$$\max_{\alpha,\beta}\ h(\alpha,\beta) = \max_{\alpha,\beta}\left(\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{m}\alpha_i\alpha_k y_i y_k \mathbf{x}_i^T\mathbf{x}_k\right)$$

The KKT Conditions for SVM

Primal and Dual Feasibility:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \alpha_i \geq 0, \quad \beta_i \geq 0$$

From maximizing Lagrangian wrt to primal variables:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \alpha_i + \beta_i = C$$

Complementary Slackness Conditions:

$$\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] = 0 \qquad \beta_i \xi_i = 0$$

The KKT Conditions for SVM

We can actually combine a few of these to get some simpler constraints. In particular:

$$\alpha_i + \beta_i = C, \quad \alpha_i \geq 0, \quad \beta_i \geq 0 \quad \text{for } i = 1, \ldots, m$$

From this we can derive the simpler constraint

$$0 \leq \alpha_i \leq C \quad \text{for } i = 1, \ldots, m$$

Proof:

$\alpha_i = C - \beta_i \qquad \beta_i \geq 0 \Rightarrow \alpha_i \leq C$

$\alpha_i \geq 0$

Better Complementary Slackness

$$\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right] = 0$$

Three possible cases for a training example:

- $\mathbf{x}_i$ on correct side of support vector boundary
- $\mathbf{x}_i$ on the support vector boundary
- $\mathbf{x}_i$ on wrong side of support vector boundary

Better Complementary Slackness

$$\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right] = 0$$

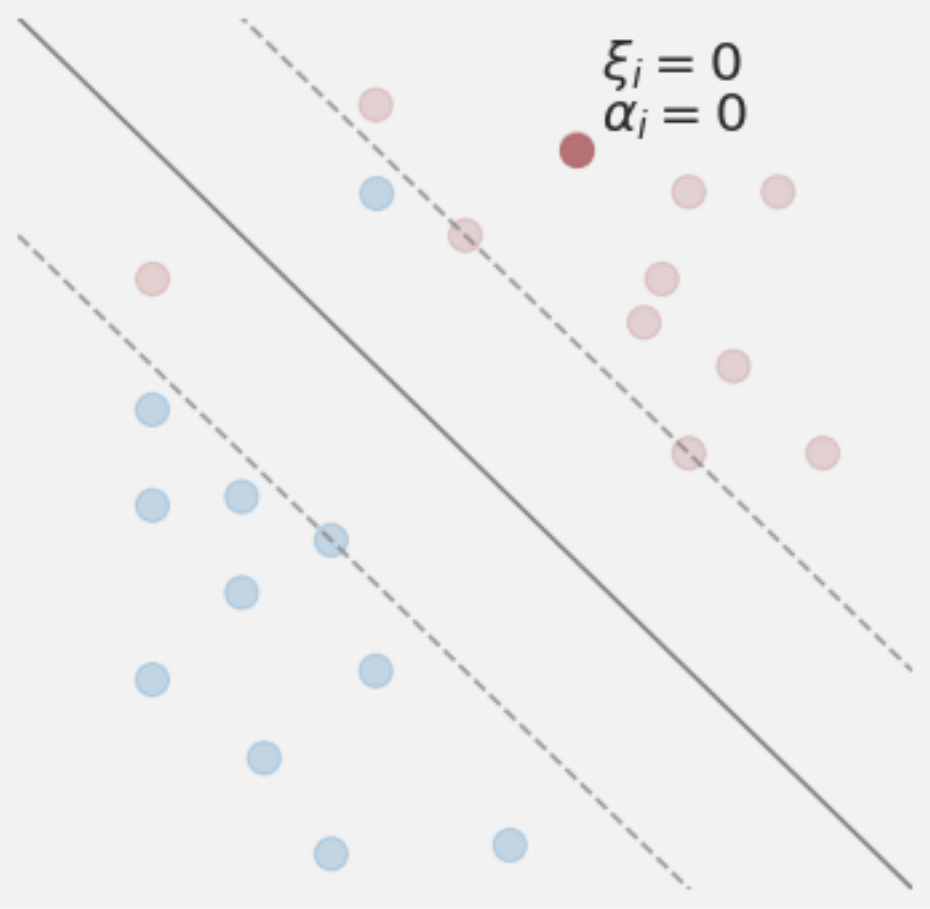
Three possible cases for a training example:

- $\mathbf{x}_i$ on correct side of support vector boundary

$$\xi_i = 0 \quad \Rightarrow \quad \alpha_i \left[1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right] = 0$$

Inactive constraint implies that

$$\alpha_i = 0 \quad \text{and} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$$



Better Complementary Slackness

$$\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] = 0$$

Three possible cases for a training example:

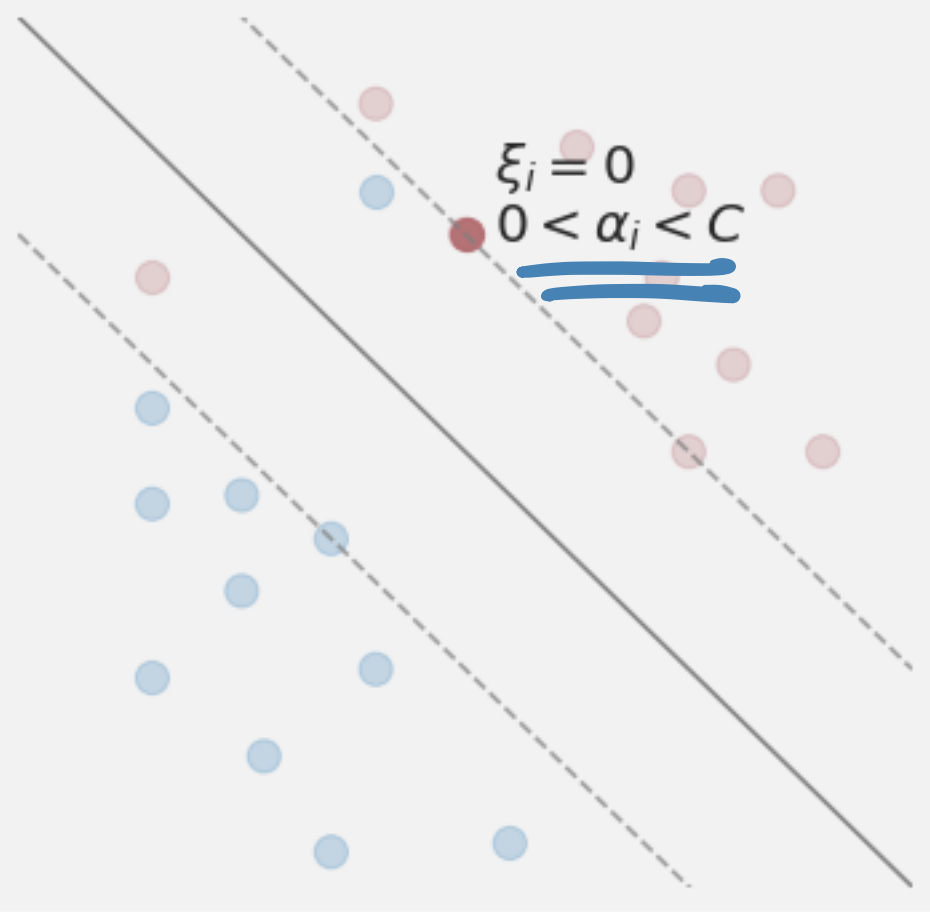
- $\mathbf{x}_i$ on the support vector boundary

$$\xi_i = 0 \quad \Rightarrow \quad \alpha_i \left[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right] = 0$$

Active constraints implies that

$$\beta_i > 0 \Rightarrow \quad 0 < \alpha_i < C \quad \text{and}$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$$



Better Complementary Slackness

$$\alpha_i \left[1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right] = 0$$

Three possible cases for a training example:

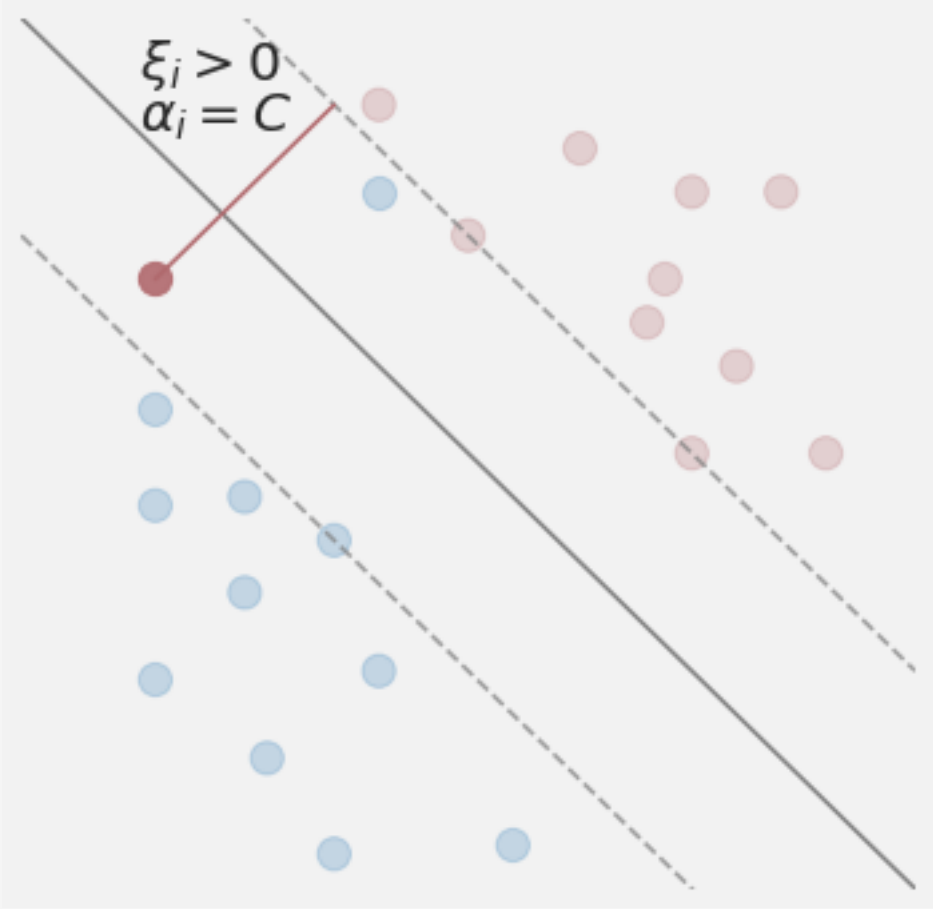
- $\mathbf{x}_i$ on wrong side of support vector boundary

$$\xi_i > 0 \quad \Rightarrow \quad \beta_i = 0 \quad \Rightarrow \quad \alpha_i = C$$

Plugging into the complementary slackness condition:

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

So we have: $\alpha_i = C \;\text{ and }\; y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq 1$



The Dual Problem for SVM

After all of that work, we have the final **Dual Problem**

$$\max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \quad \text{for } i = 1, \ldots, m$$

$$\text{s.t.} \quad \sum_{i=1}^{m} y_i \alpha_i = 0$$

With simplified KKT conditions:

$$\begin{aligned} \alpha_i = 0 &\quad \Leftrightarrow \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) > 1 \\ \alpha_i = C &\quad \Leftrightarrow \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) < 1 \\ 0 < \alpha_i < C &\quad \Leftrightarrow \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 \end{aligned}$$

Solution Techniques: Again, quadratic programming solutions. More advanced: SMO Algorithm.

Neato Consequences of the Dual Problem

One of the KKT conditions:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

$\mathbf{w}$ depends on training ex's with nonzero α_i

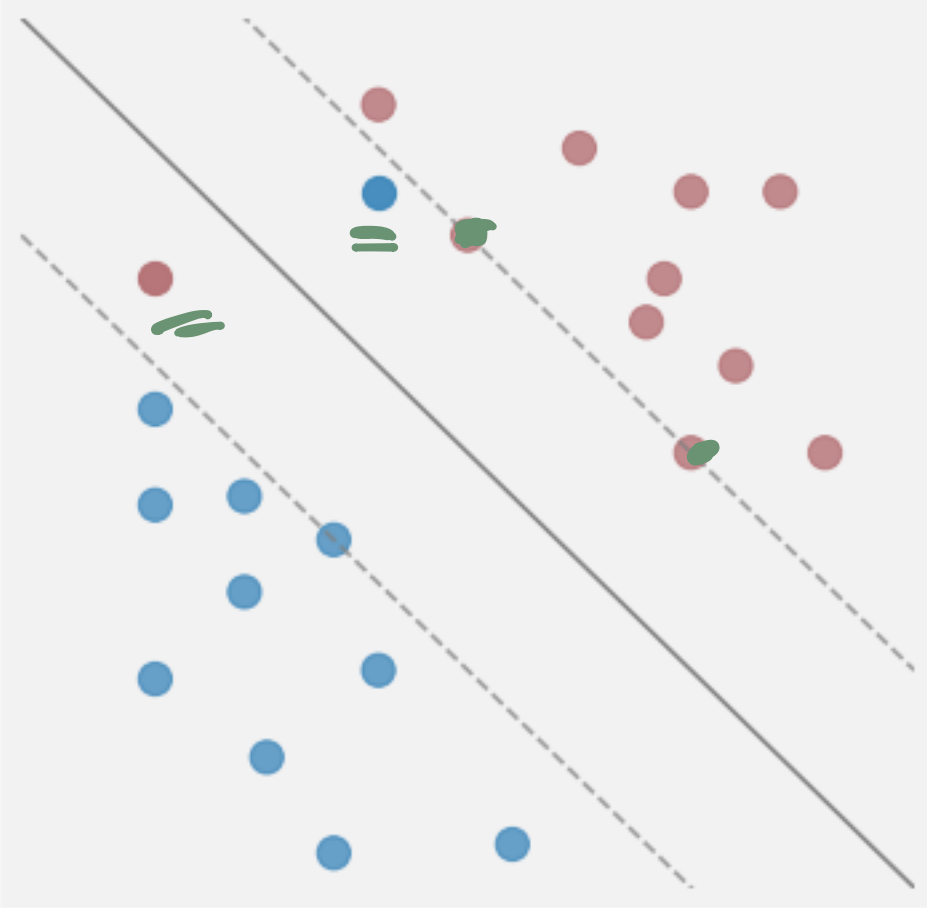
These are the support vectors, but now two types

Inbound Support Vectors:

$$0 < \alpha_i < C \;\Leftrightarrow\; y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$$

Outbound Support Vectors:

$$\alpha_i = C \;\Leftrightarrow\; y_i(\mathbf{w}^T \mathbf{x}_i + b) < 1$$



Neato Consequences of the Dual Problem

The Dual objective function

$$\max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right)$$

Looking Ahead:

- The Dual objective function only depends on **dot products** of training examples
- This will save our collective butts when we get to non-linear SVMs

If-Time Bonus: SMO Algorithm