Support Vector Machines

Additional material, with derivation of dual problem and examples

Author: Ondřej Drbohlav

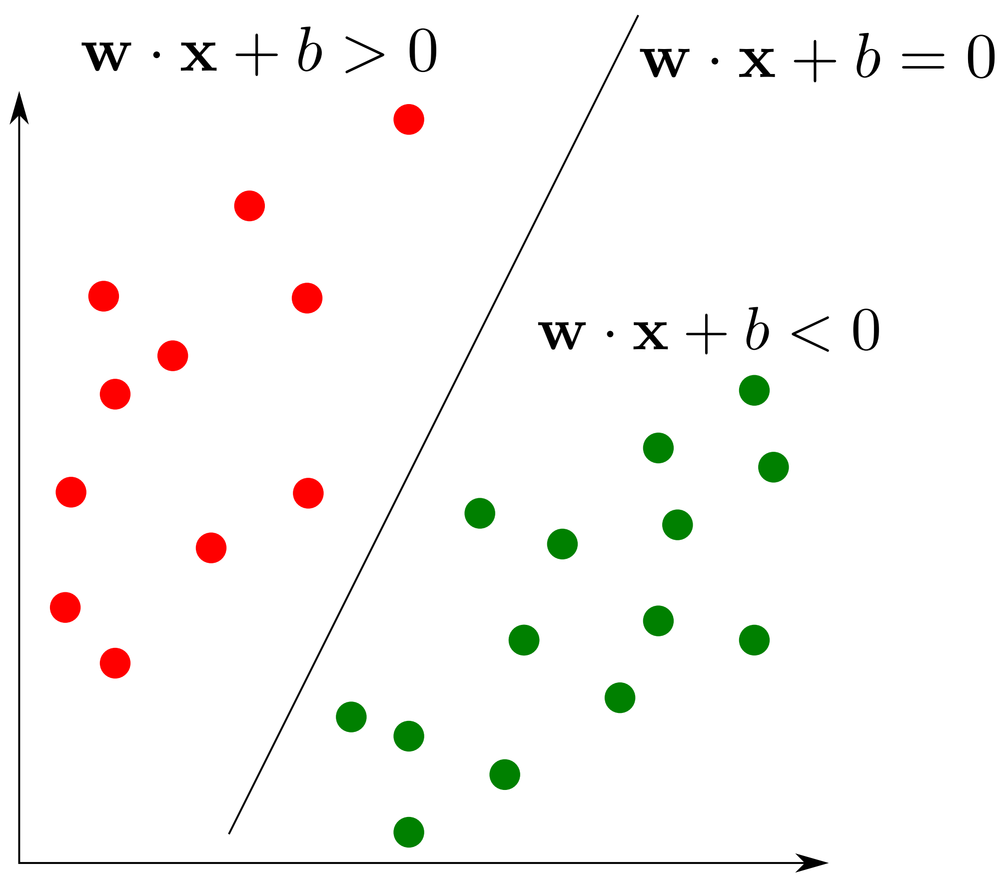
Version 30/Nov/2017

Linear Classifier Revisited (1)

Classification according to signum of an affine function of $\mathbf{x}$:

$$q(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{1}$$

A solution for $\{\mathbf{w}, b\}$ correctly classifying the training set:

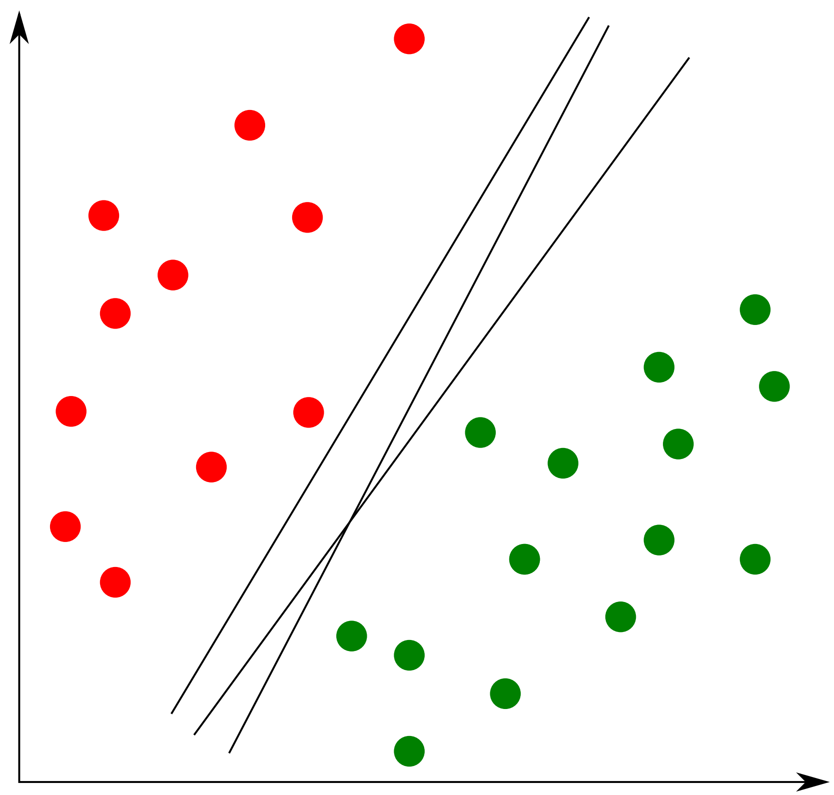


Linear Classifier Revisited (2)

Classification according to signum of an affine function of $\mathbf{x}$:

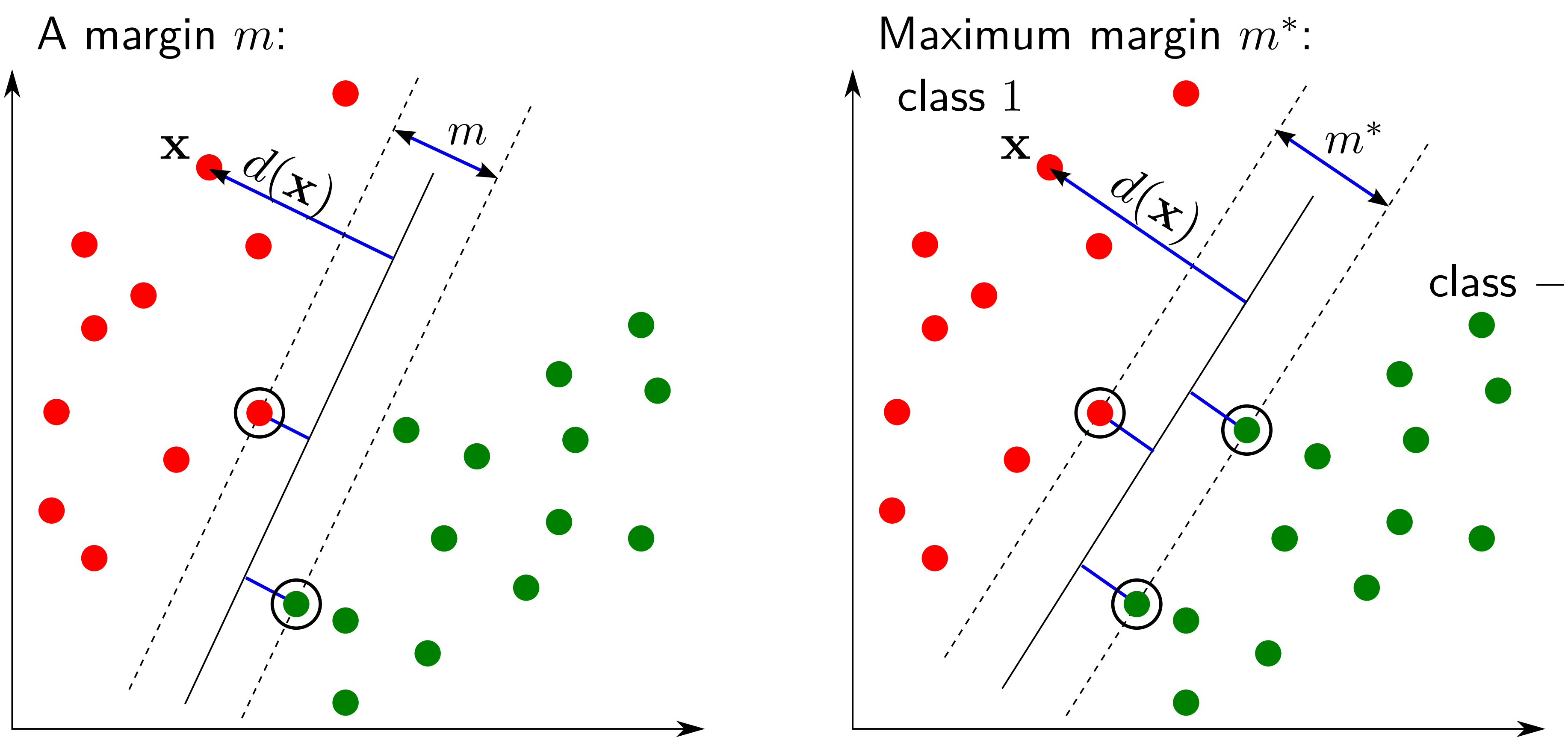
$$q(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \quad (2)$$

But there are many solutions possible. Which one is the best?



Margin, Informal Introduction

- Assume linearly separable data.
- Distance of a point $\mathbf{x}$ to the decision boundary: $d(\mathbf{x})$
- Points closest to the decision boundary are called **support vectors**
- Margin m (our definition): twice the distance to a support vector
- Find the decision boundary maximizing the margin. Vapnik justifies the use of maximum margin from the viewpoint of Structural Risk Minimization.

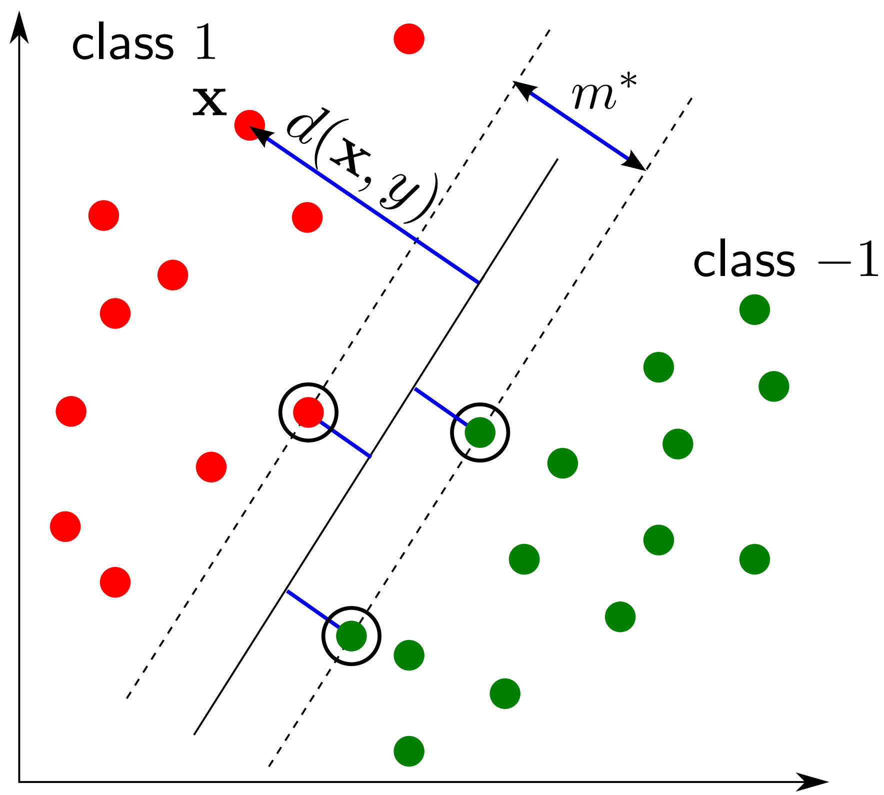


Maximizing Margin, Formulation

- Signed distance of a point $\mathbf{x}$ belonging to class $y \in \{1, -1\}$:

$$d(\mathbf{x}, y) = \frac{y(\mathbf{w} \cdot \mathbf{x} + b)}{\|\mathbf{w}\|} \tag{3}$$

- We require $d(\mathbf{x}, y) > 0$ for all training data (all training points are in their class' half-space). This is equivalent to $y(\mathbf{w} \cdot \mathbf{x} + b) \geq \epsilon > 0$.



Optimization task:

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w}, b} \min_{(\mathbf{x}, y) \in \mathcal{T}} 2d(\mathbf{x}, y)$$

subject to:

$$y(\mathbf{w} \cdot \mathbf{x} + b) \geq \epsilon > 0, \forall (\mathbf{x}, y) \in \mathcal{T} \quad \text{(C)}$$

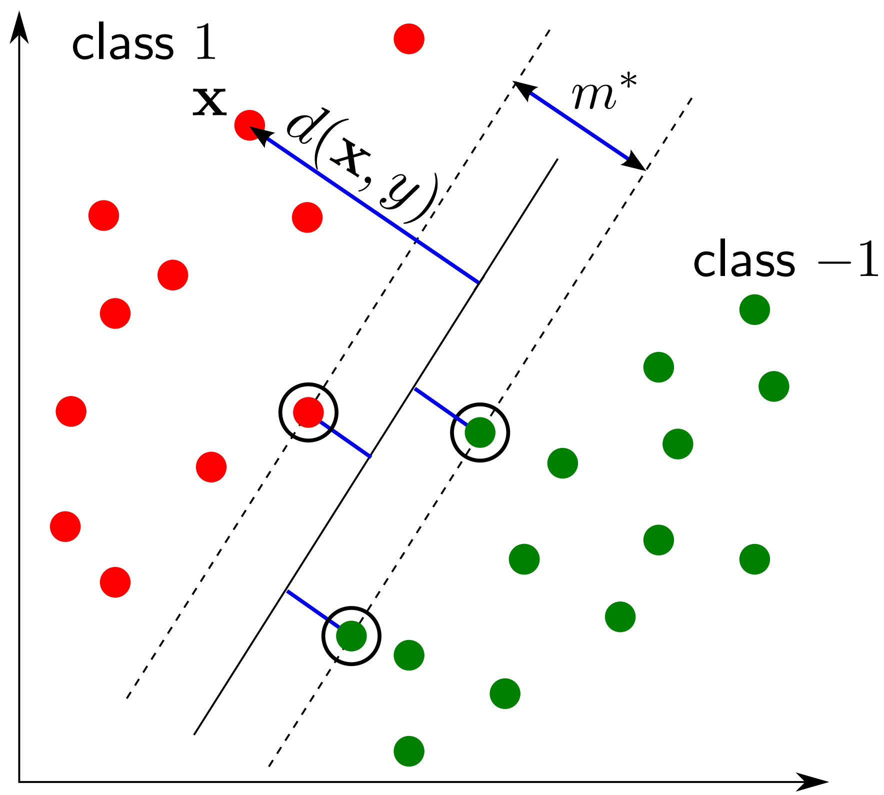
Maximizing Margin, Scale Ambiguity

- There is a scale ambiguity in the parameters $(\mathbf{w}, b)$. Any feasible $(\mathbf{w}, b)$ (that is, satisfying Eq. (C) can be multiplied by a positive constant $k > 0$ to form $(k\mathbf{w}, kb)$, and:

  (i) feasibility does not change, as

$$y(k\mathbf{w}\cdot\mathbf{x}+kb) = ky(\mathbf{w}\cdot\mathbf{x}+b) \geq k\epsilon \Leftrightarrow y(\mathbf{w}\cdot\mathbf{x}+b) \geq \epsilon, \text{ and} \tag{4}$$

  (ii) signed distances do not change, as



$$d(\mathbf{x}, y) = \frac{y(k\mathbf{w}\cdot\mathbf{x}+kb)}{\|k\mathbf{w}\|} = \frac{y(\mathbf{w}\cdot\mathbf{x}+b)}{\|\mathbf{w}\|}. \tag{5}$$

Optimization task:

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x}, y)$$

subject to:

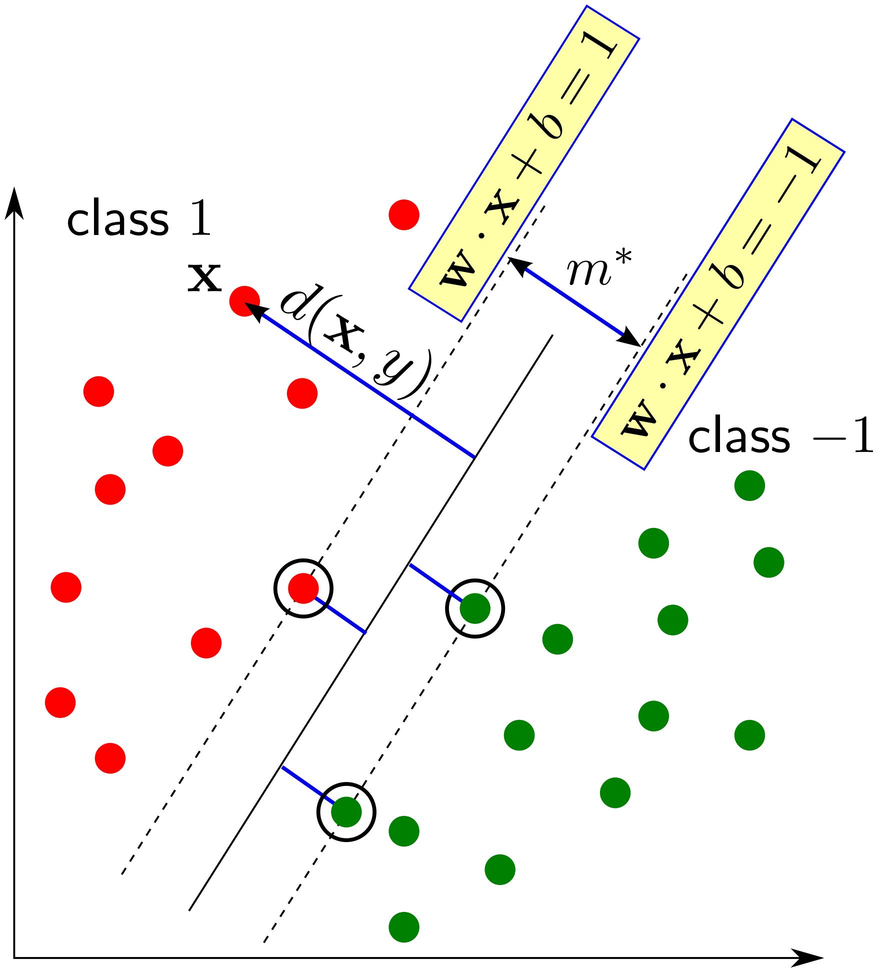
$$y(\mathbf{w}\cdot\mathbf{x}+b) \geq \epsilon > 0, \forall(\mathbf{x}, y) \in \mathcal{T} \quad \text{(C)}$$

Maximizing Margin, Fixing Scale

- Break the scale ambiguity by setting $\epsilon = 1$:

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x}, y)$$

$$\text{subject to: } y(\mathbf{w}\cdot\mathbf{x} + b) \geq 1, \forall(\mathbf{x}, y) \in \mathcal{T} \quad (6)$$



Optimization task (original):

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x}, y)$$

subject to:

$$y(\mathbf{w}\cdot\mathbf{x} + b) \geq \epsilon > 0, \forall(\mathbf{x}, y) \in \mathcal{T} \quad \text{(C)}$$

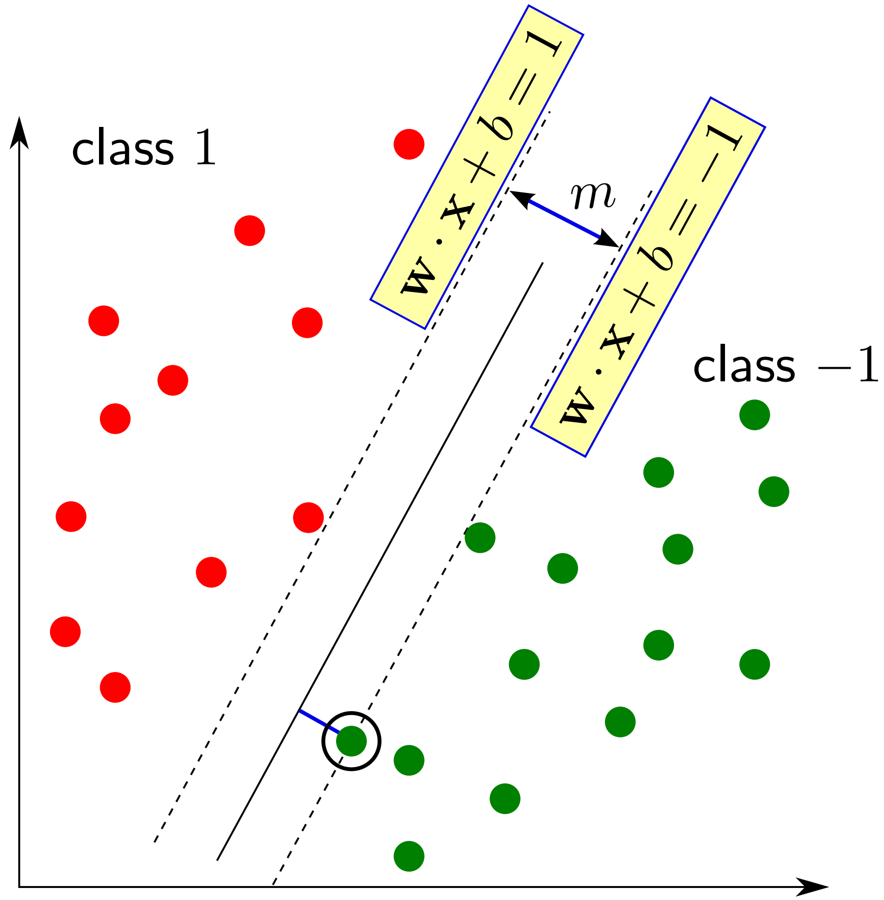
$$d(\mathbf{x}, y) = \frac{y(\mathbf{w}\cdot\mathbf{x} + b)}{\|\mathbf{w}\|}$$

Maximizing Margin, Final Optimization Formulation (1)

- All points must be outside the strip delineated by the two lines $\mathbf{w} \cdot \mathbf{x} + b = 1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$. The width of this strip is $\frac{2}{\|\mathbf{w}\|}$. It follows that the maximum margin m^* is

$$m^* = \max_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x},y) = \max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$\text{subject to: } y(\mathbf{w}\cdot\mathbf{x}+b) \geq 1, \forall(\mathbf{x},y)\in\mathcal{T} \quad (7)$$



Optimization task (original):

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x},y)$$

subject to:

$$y(\mathbf{w}\cdot\mathbf{x}+b) \geq \epsilon > 0, \forall(\mathbf{x},y)\in\mathcal{T} \quad \text{(C)}$$

$$d(\mathbf{x},y) = \frac{y(\mathbf{w}\cdot\mathbf{x}+b)}{\|\mathbf{w}\|}$$

Maximizing Margin, Final Optimization Formulation (2)

- All points must be outside the strip delineated by the two lines $\mathbf{w} \cdot \mathbf{x} + b = 1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$. The width of this strip is $\frac{2}{\|\mathbf{w}\|}$. It follows that the maximum margin m^* is

$$m^* = \max_{\mathbf{w},b} \min_{(\mathbf{x},y)\in\mathcal{T}} 2d(\mathbf{x},y) = \max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$\text{subject to: } y(\mathbf{w} \cdot \mathbf{x} + b) \geq 1, \forall(\mathbf{x},y) \in \mathcal{T} \tag{8}$$

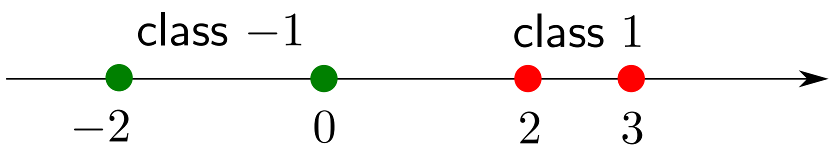
- There holds: $\operatorname{argmax}_{\mathbf{w}} \frac{2}{\|\mathbf{w}\|} = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\| = \operatorname{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$. Therefore, the $(\mathbf{w}^*, b^*)$ maximizing the margin are:

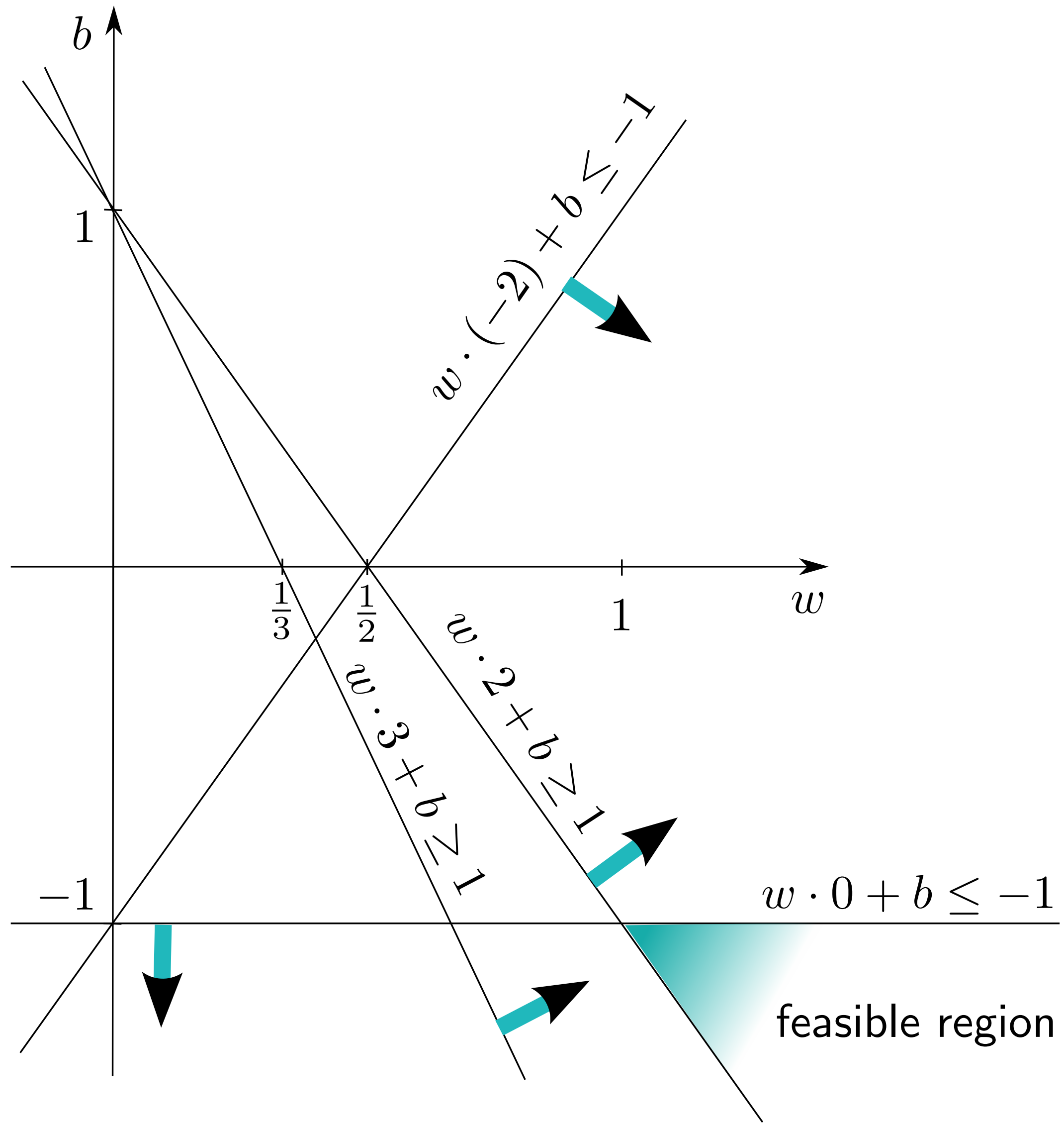
$$(\mathbf{w}^*, b^*) = \operatorname*{argmin}_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to: } y(\mathbf{w} \cdot \mathbf{x} + b) \geq 1, \forall(\mathbf{x},y) \in \mathcal{T} \tag{9}$$

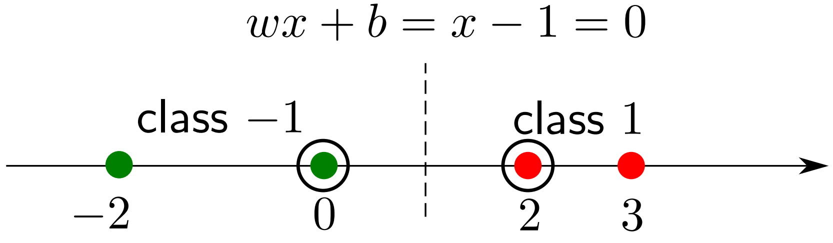
- This is a Quadratic Programming (QP) problem (more generally, it is minimization of a convex function on a convex domain.)

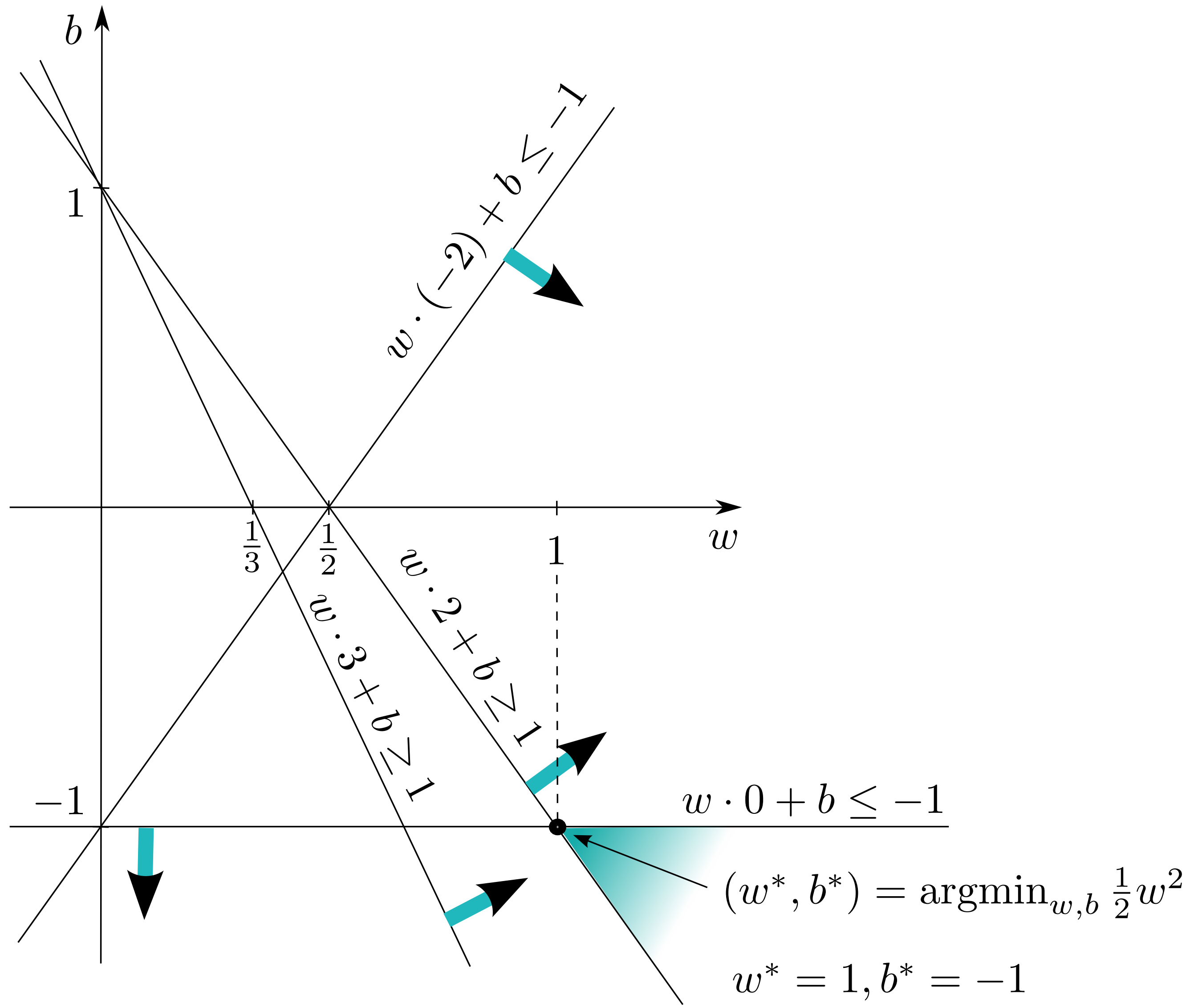
SVM, Example (1D)





SVM, Example (1D), Result





SVM, Primal Problem

The derived optimization problem for $\mathbf{w}$ and b is

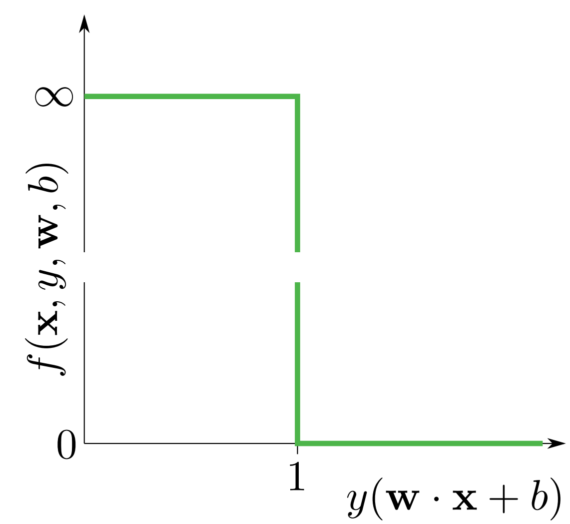
$$(\mathbf{w}^*, b^*) = \operatorname*{argmin}_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to: } y(\mathbf{w}\cdot\mathbf{x}+b) \geq 1, \forall(\mathbf{x},y)\in\mathcal{T} \tag{10}$$

It is called *primal* problem. We will also soon derive the *dual* problem. For now, note that the above optimization task can be equivalently regarded as solving an unconstrained problem (this observation will become handy when deriving the dual problem):

$$(\mathbf{w}^*, b^*) = \operatorname*{argmin}_{(\mathbf{w},b)} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{(\mathbf{x},y)\in\mathcal{T}} f(\mathbf{x},y,\mathbf{w},b) \right\}, \text{ where} \tag{11}$$

$$f(\mathbf{x},y,\mathbf{w},b) = \begin{cases} 0 & \text{if } y(\mathbf{w}\cdot\mathbf{x}+b)\geq 1, \\ \infty, & \text{otherwise} \end{cases} \tag{12}$$



Note that $f(\mathbf{x},y,\mathbf{w},b)$ for a given $(\mathbf{x},y)$ is a convex function of $\mathbf{w}, b$.

The Dual Formulation (1)

Start with just discussed primal formulation. Let $\mathcal{T} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$ be the training set. We want to solve

$$(\mathbf{w}^*, b^*) = \underset{(\mathbf{w},b)}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{N} f(\mathbf{x}_i, y_i, \mathbf{w}, b) \right\}, \text{ where}$$

$$f(\mathbf{x}_i, y_i, \mathbf{w}, b) = \begin{cases} 0 & \text{if } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1. \\ \infty, & \text{otherwise} \end{cases} \tag{13}$$

This is the same as (α_i's are non-negative multipliers):

$$(\mathbf{w}^*, b^*) = \underset{\mathbf{w},b}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0 \\ i \in \{1,..,N\}}} \left(-\sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \right) \right\}. \tag{14}$$

because

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1 \;\Rightarrow\; \max_{\alpha_i}(-\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]) = 0 \text{ for } \alpha_i = 0\,, \tag{15}$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) < 1 \;\Rightarrow\; \max_{\alpha_i}(-\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]) = \infty \text{ for } \alpha_i = \infty\,, \tag{16}$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 \;\Rightarrow\; \max_{\alpha_i}(-\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]) = 0 \text{ for any } \alpha_i \geq 0\,. \tag{17}$$

The Dual Formulation (2)

This is in turn the same as

$$(\mathbf{w}^*, b^*) = \operatorname*{argmin}_{\mathbf{w},b} \max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0 \\ i \in \{1,..,N\}}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1] \right\} . \tag{18}$$

There holds, in full generality, that $\max_p \min_q f(p,q) \leq \min_q \max_p f(p,q)$. For our case,

$$\min_{\mathbf{w},b} \max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0 \\ i \in \{1,..,N\}}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1] \right\} \geq$$

$$\geq \max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0 \\ i \in \{1,..,N\}}} \min_{\mathbf{w},b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1] \right\} \tag{19}$$

This is the essence of converting the primal problem to the dual one. And, our case is even better: strong duality holds, and the two terms are equal (duality gap is zero). Denote the inner term by $L(\mathbf{w}, b, \alpha)$ (corresponds to what's commonly known as the Lagrangian):

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1] \tag{20}$$

The Dual Formulation (3)

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \tag{21}$$

We want to find $\operatorname{argmax}_{\alpha \geq 0} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$. First, for fixed α, find $\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0 \;\Rightarrow\; \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{22}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{23}$$

Put this to Lagrangian:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = \tag{24}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i\right) \cdot \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i b + \sum_{i=1}^{N} \alpha_i \tag{25}$$

$$= -\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{26}$$

The Dual Formulation, Result and Insights

The dual optimization problem:

$$\alpha = \underset{\alpha}{\operatorname{argmax}} \left(\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) \right) = \underset{\alpha}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right\} \tag{27}$$

$$\text{subject to: } \sum_i \alpha_i y_i = 0; \quad \alpha_i \geq 0, \quad \forall i \in \{1, 2, ..., N\} \tag{28}$$

- Number of optimization variables α_i's is N (the number of training data). But at the solution, all α_i's but those of support vectors are zero.
- Once the dual problem is solved, the primal variables can be computed as

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \qquad \text{only support vectors } (\alpha_i > 0) \text{ contribute} \tag{29}$$

$$y^S[\mathbf{w} \cdot \mathbf{x}^S + b] = 1 \text{ for any support vector } (\mathbf{x}^S, y^S) \;\Rightarrow\; b = y^S - \mathbf{w} \cdot \mathbf{x}^S \tag{30}$$

- The discriminant function $\mathbf{w} \cdot \mathbf{x} + b$ thus takes the form ($\mathcal{P}$ are indices of all support vectors):

$$\mathbf{w} \cdot \mathbf{x} + b = \sum_{i \in \mathcal{P}} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + \underbrace{y^S - \sum_{i \in \mathcal{P}} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^S)}_{\text{constant, independent of } \mathbf{x}} \tag{31}$$

- Both the dual classification problem and the discriminant function involve data points **only** in the form of **dot products**.

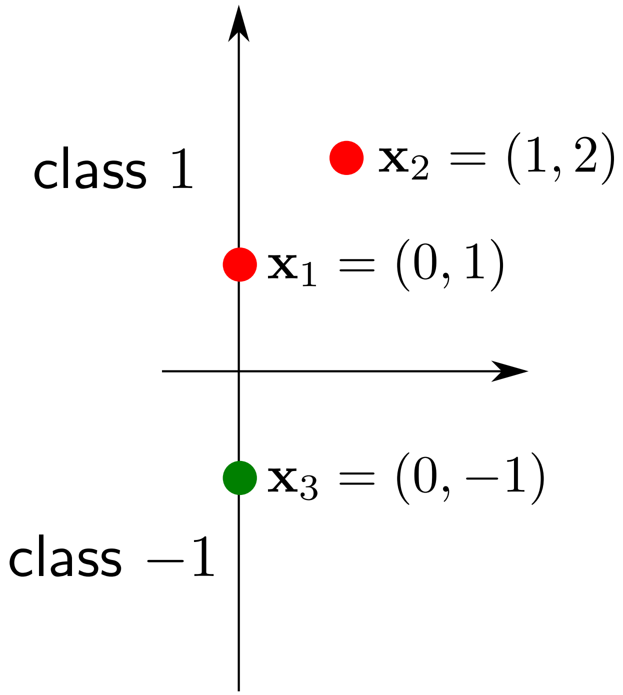
The Dual Problem, Example (1)

Consider the 3 points as below

Objective: maximize

$$\alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}^T \begin{bmatrix} y_1 y_1 \mathbf{x}_1 \cdot \mathbf{x_1} & y_1 y_2 \mathbf{x}_1 \cdot \mathbf{x_2} & y_1 y_3 \mathbf{x}_1 \cdot \mathbf{x_3} \\ y_2 y_1 \mathbf{x}_2 \cdot \mathbf{x_1} & y_2 y_2 \mathbf{x}_2 \cdot \mathbf{x_2} & y_2 y_3 \mathbf{x}_2 \cdot \mathbf{x_3} \\ y_3 y_1 \mathbf{x}_3 \cdot \mathbf{x_1} & y_3 y_2 \mathbf{x}_3 \cdot \mathbf{x_2} & y_3 y_3 \mathbf{x}_3 \cdot \mathbf{x_3} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

subject to: $\alpha_1, \alpha_2, \alpha_3 \geq 0$; $\alpha_1 + \alpha_2 - \alpha_3 = 0$



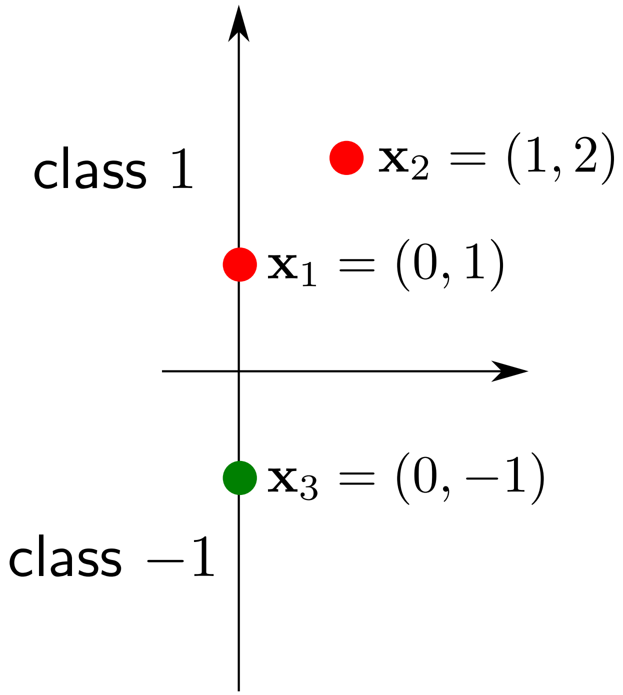
The Dual Problem, Example (2)

Consider the 3 points as below

Objective: maximize

$$\alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}^T \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

subject to: $\alpha_1, \alpha_2, \alpha_3 \geq 0; \;\; \alpha_1 + \alpha_2 - \alpha_3 = 0$

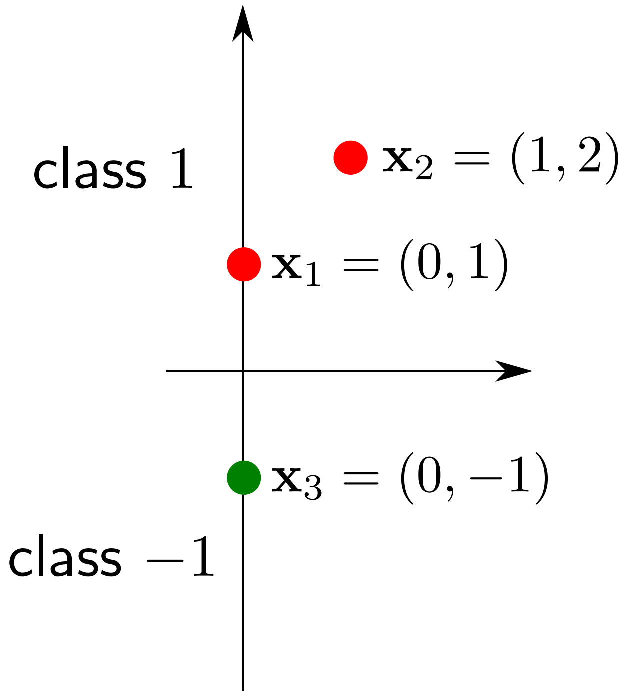


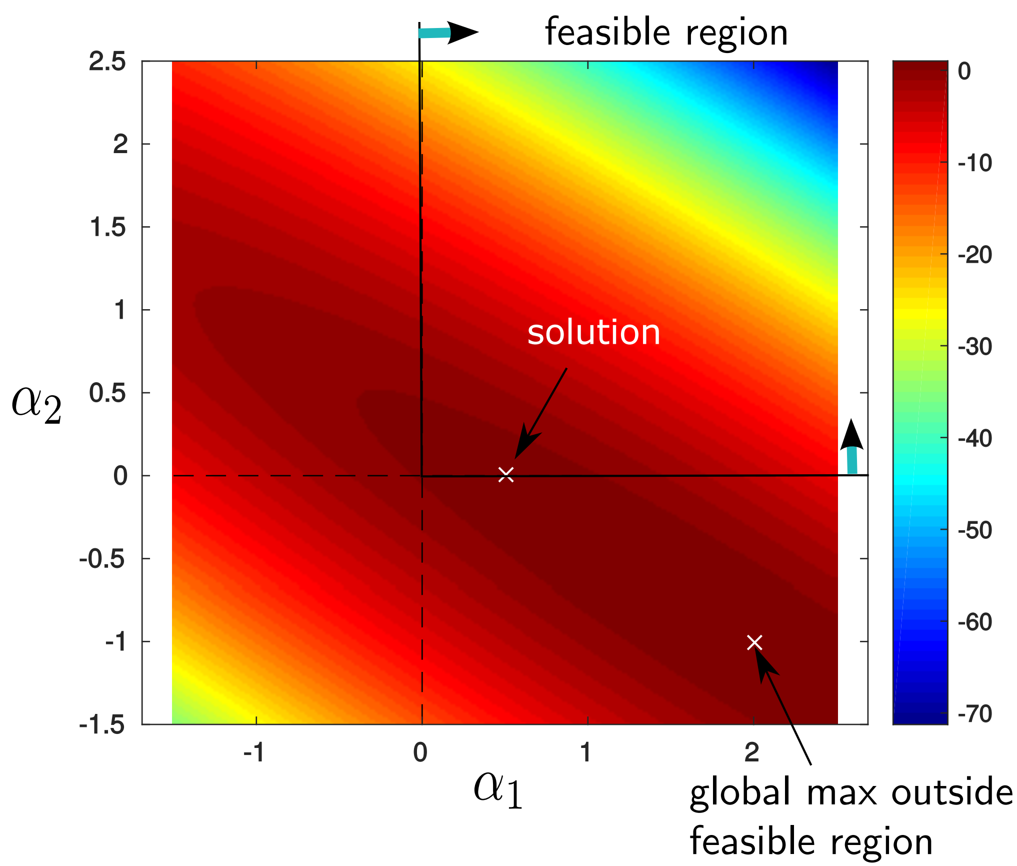
The Dual Problem, Example (3)

Substitute $\alpha_3 = \alpha_1 + \alpha_2$ and search for solution as a problem in α_1, α_2. After some straightforward computation, the original problem turns to:

$$\text{maximize } 2(\alpha_1 + \alpha_2) - \frac{1}{2} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}^T \begin{bmatrix} 4 & 6 \\ 6 & 10 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

subject to: $\alpha_1, \alpha_2 \geq 0$. **Solution**: $(\alpha_1, \alpha_2) = (\frac{1}{2}, 0)$, $\alpha_3 = \frac{1}{2} + 0 = \frac{1}{2}$.





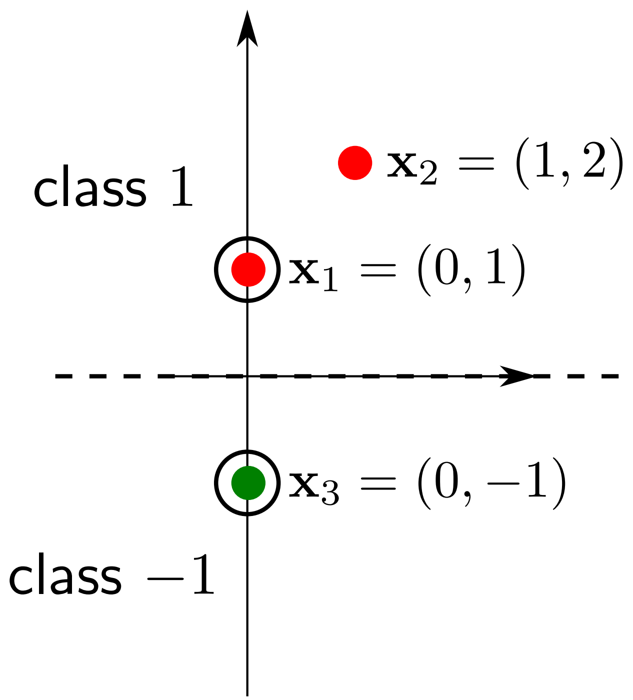
The Dual Problem, Example, Result

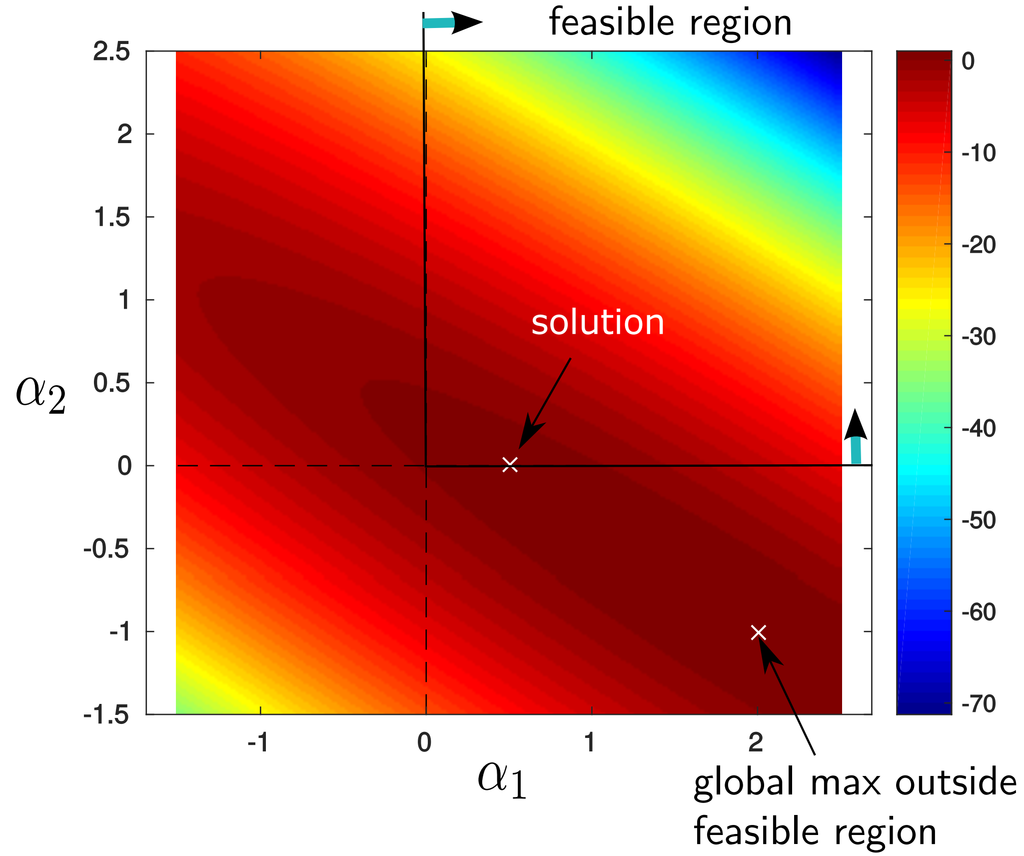
Result: $(\alpha_1, \alpha_2, \alpha_3) = (\frac{1}{2}, 0, \frac{1}{2})$. The support vectors are $\mathbf{x}_1$ and $\mathbf{x}_3$ because their $\alpha_i > 0$.

Vector $\mathbf{w} = \sum_{i=\{1,3\}} \alpha_i y_i \mathbf{x}_i = \frac{1}{2}(0,1) - \frac{1}{2}(0,-1) = (0,1)$.

Offset $b = y^S - \mathbf{w}\mathbf{x}^S = 1 - \mathbf{w}\mathbf{x}_1 = -1 - \mathbf{w}\mathbf{x}_3 = 0$.

Decision boundary $(0,1)^T \cdot \mathbf{x} = 0\,.$

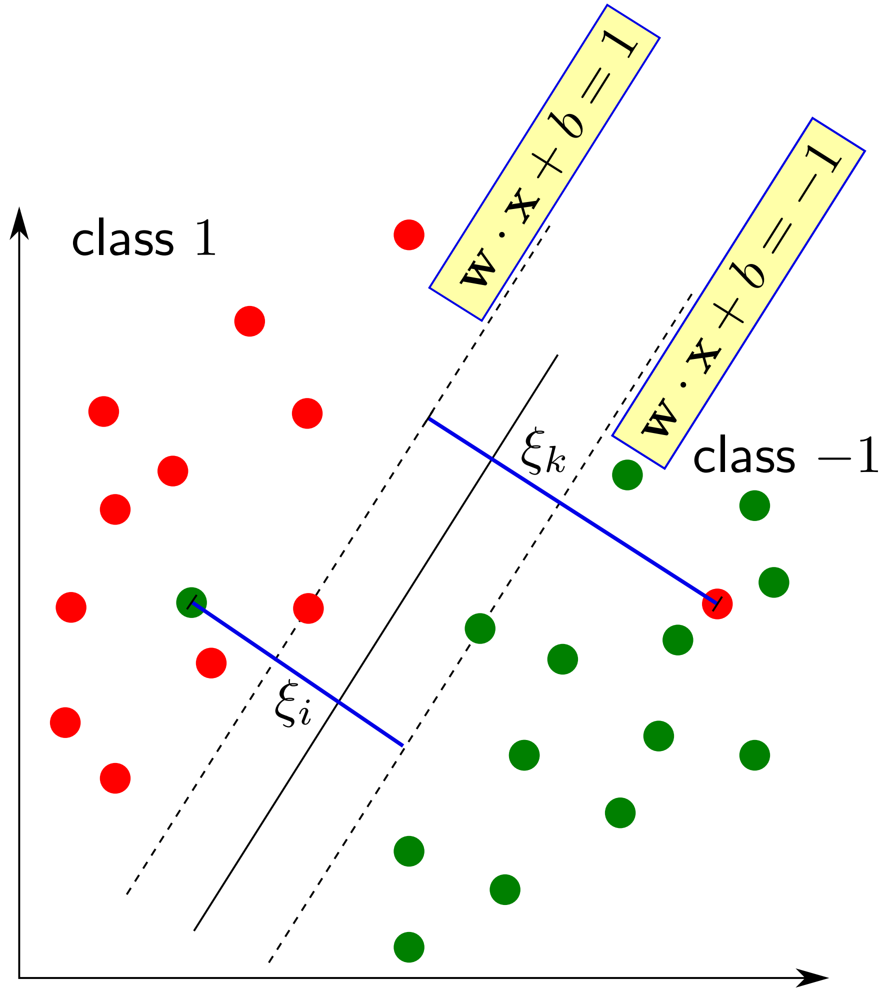




Soft Margin SVM

If the data are not linearly separable, *slack variables* ξ_i need to be introduced.

- Position and size of margin is implied by $\mathbf{w}$ and b, as before.
- If a point $(\mathbf{x}, y)$ fulfills the condition $y(\mathbf{w} \cdot \mathbf{x} + b) \geq 1$ then no penalty is paid.
- Otherwise, the condition is relaxed to $y(\mathbf{w} \cdot \mathbf{x} + b) \geq 1 - \xi$ and penalty $C \cdot \xi$ is paid ($C > 0$)



Optimization problem:

$$(\mathbf{w}^*, b^*) = \underset{(\mathbf{w},b)}{\text{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \quad (32)$$

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad (33)$$

$$\xi_i \geq 0, \quad (34)$$

$$\forall i = 1, ..., N$$

Soft Margin SVM

The primal problem

$$(\mathbf{w}^*, b^*) = \underset{(\mathbf{w},b)}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{subject to: } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i,\ \ \forall i = 1, ..., N \tag{35}$$

$$\xi_i \geq 0,\ \ \forall i = 1, ..., N \tag{36}$$

The dual problem:

$$\alpha = \underset{\alpha}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right\} \tag{37}$$

$$\text{subject to: } \sum_i \alpha_i y_i = 0 \tag{38}$$

$$0 \leq \alpha_i \leq C,\ \ \forall i \in \{1, 2, ..., N\} \tag{39}$$