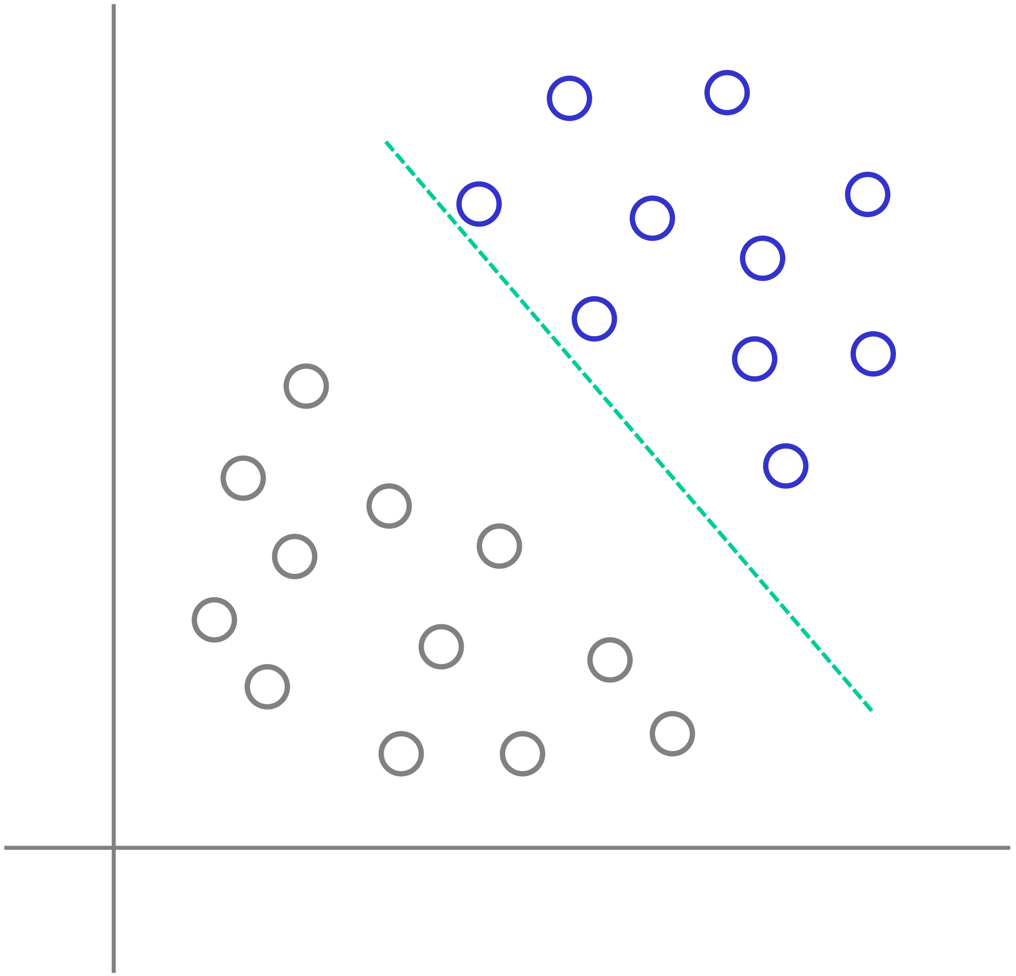
CSE 5526: Introduction to Neural Networks

Support Vector Machines (SVM)

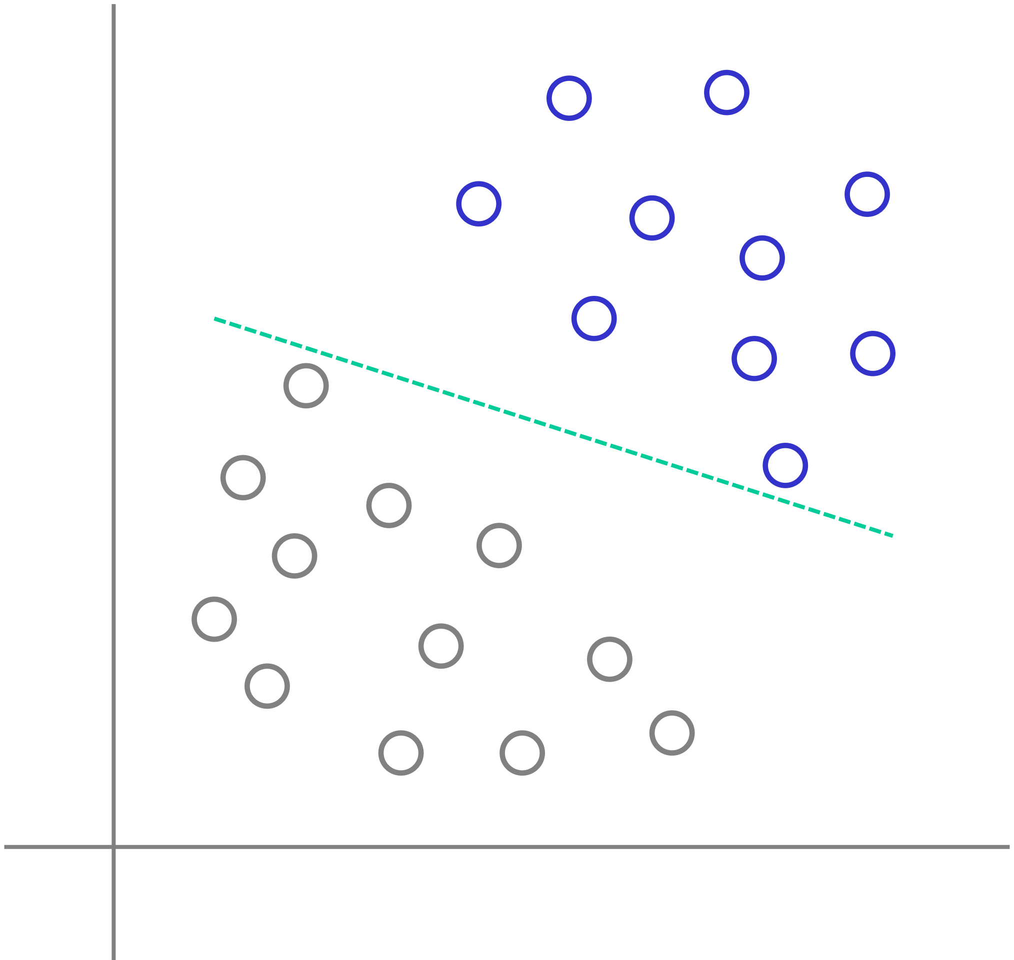
Perceptrons find any separating hyperplane

Depends on initialization and ordering of training points



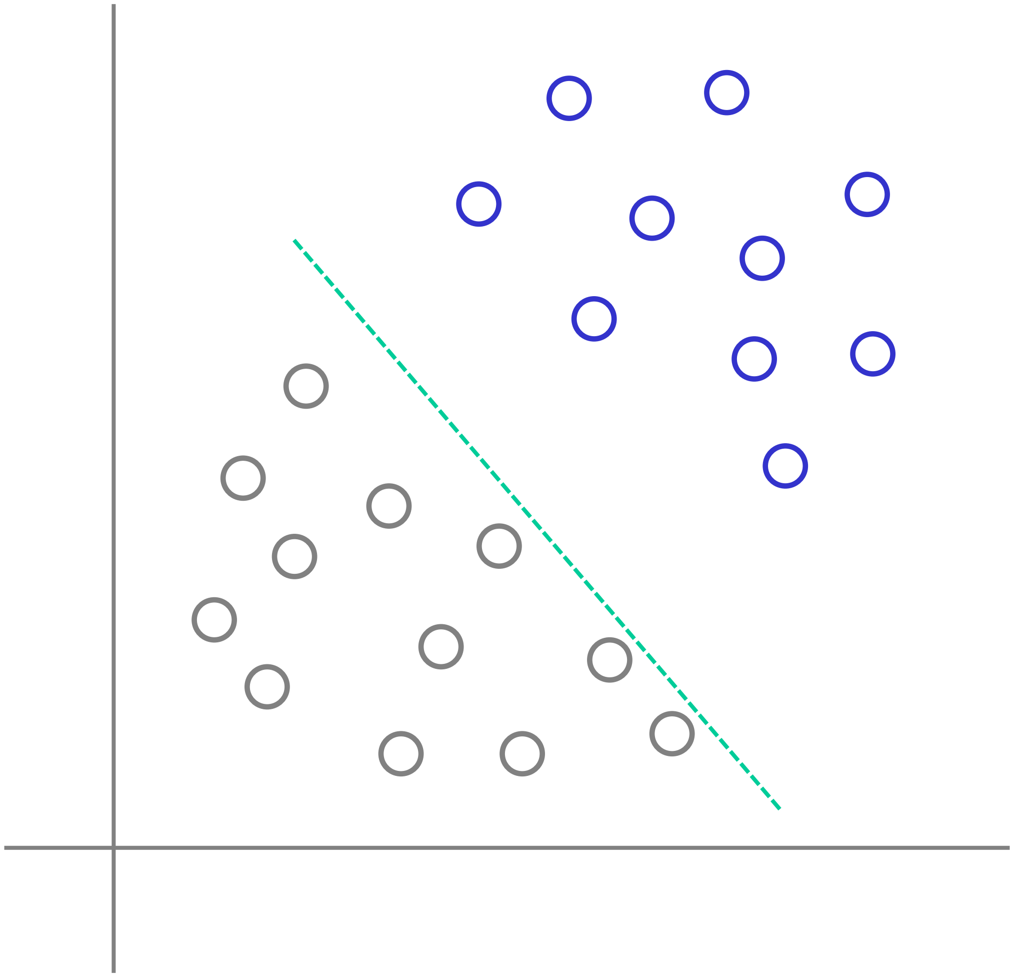
Perceptrons find any separating hyperplane

Depends on initialization and ordering of training points



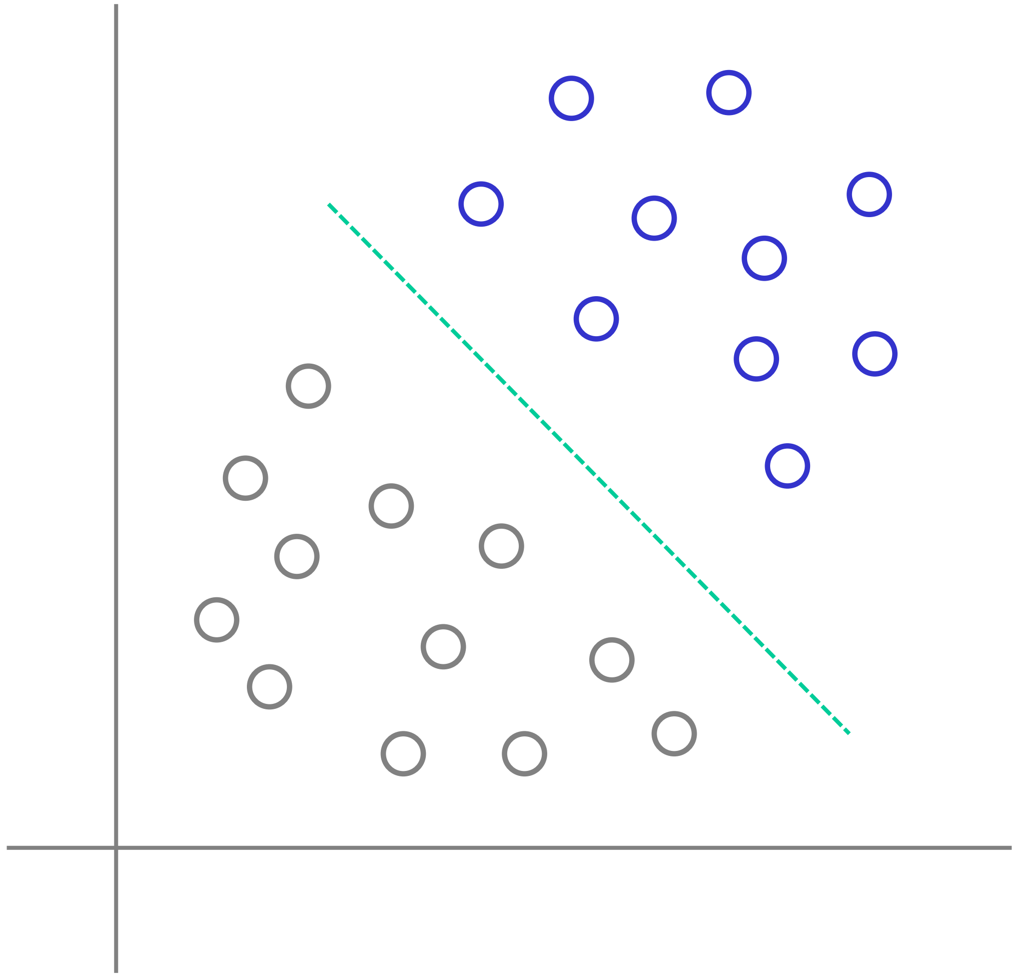
Perceptrons find any separating hyperplane

Depends on initialization and ordering of training points



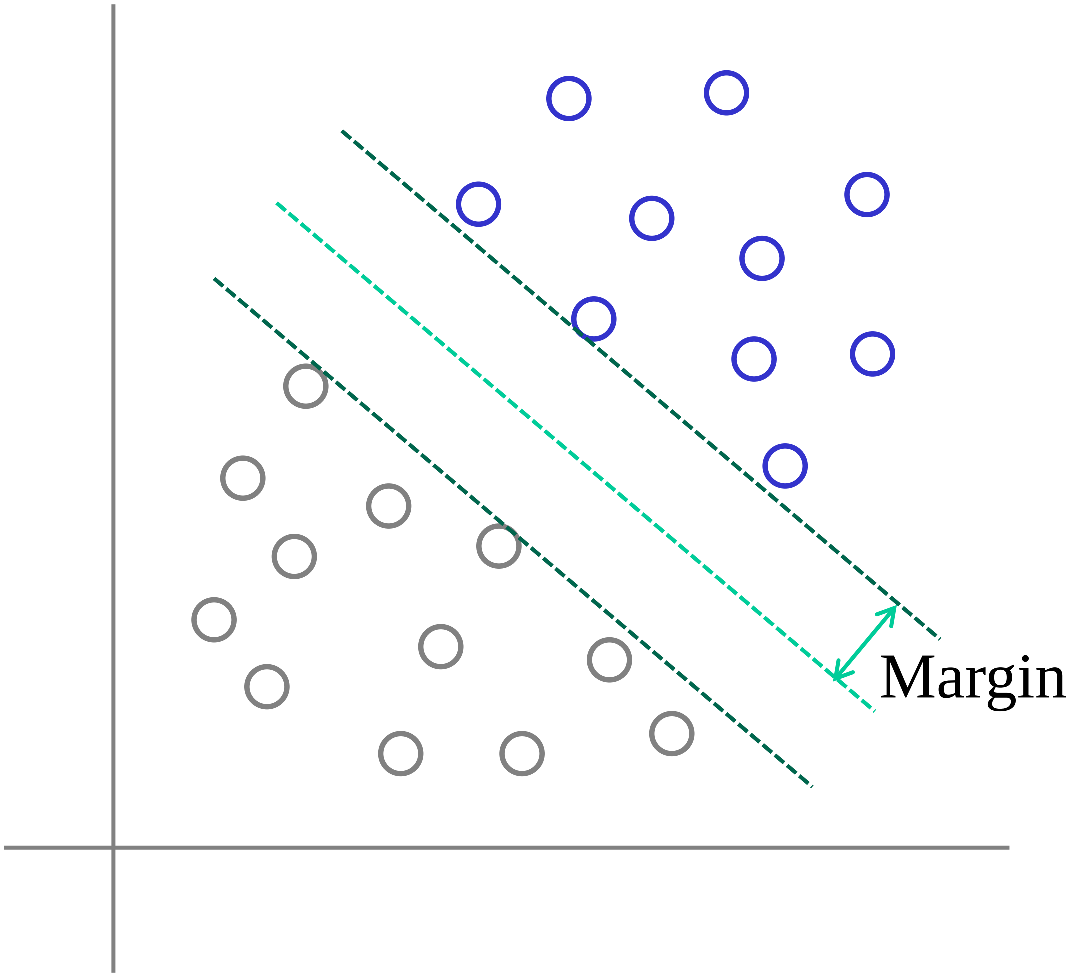
Perceptrons find any separating hyperplane

Depends on initialization and ordering of training points



But the maximum margin hyperplane generalizes the best to new data

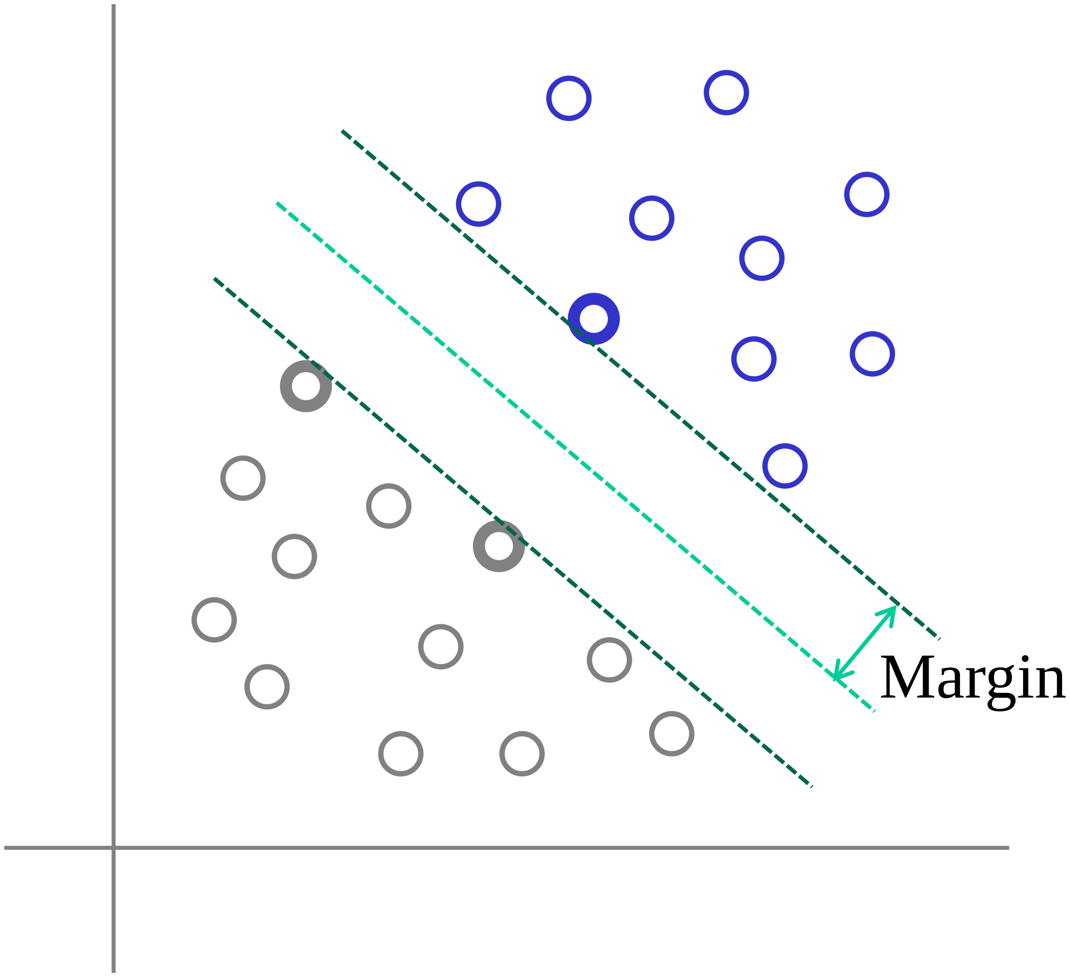
According to computational/statistical learning theory



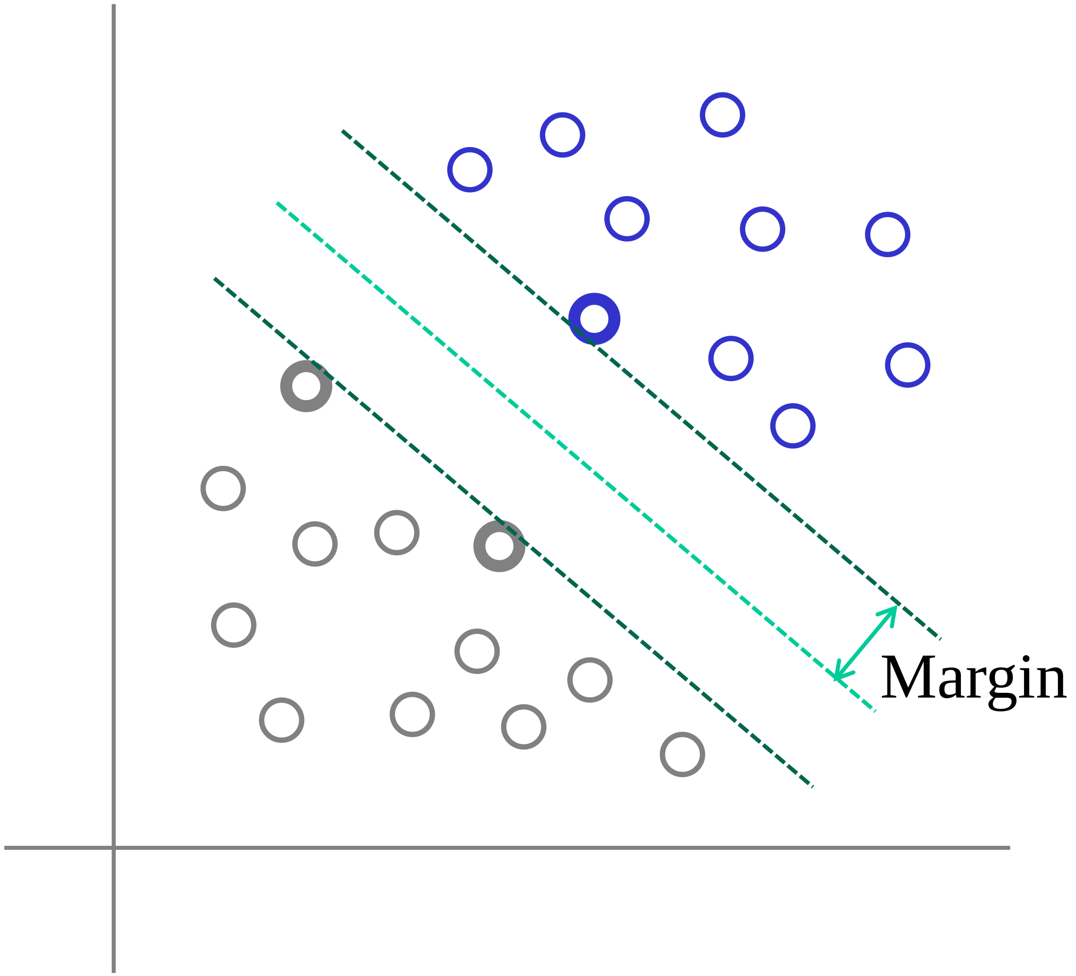
But the maximum margin hyperplane generalizes the best to new data

- According to computational learning theory
- Also known as statistical learning theory
- We won't get into the details of that
- Recall from the perceptron convergence proof
  - We assumed the existance of a best hyperplane $\boldsymbol{w}_0$
  - Which provided the maximum margin α
  - Such that $d_p \boldsymbol{w}_0^T \boldsymbol{x}_p \geq \alpha$ for all training points p
- The SVM actually finds this hyperplane

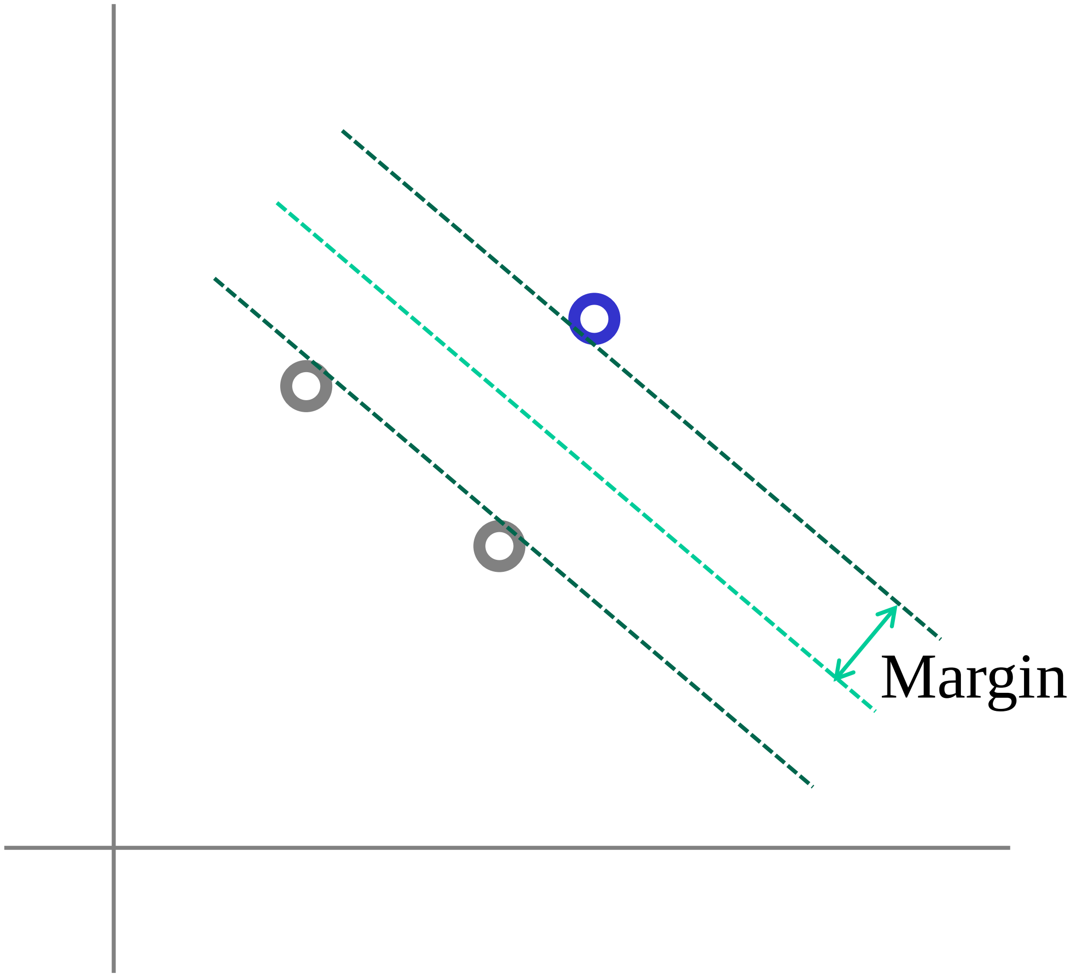
The maximum margin only depends on certain points, the support vectors



The maximum margin only depends on certain points, the support vectors



The maximum margin only depends on certain points, the support vectors



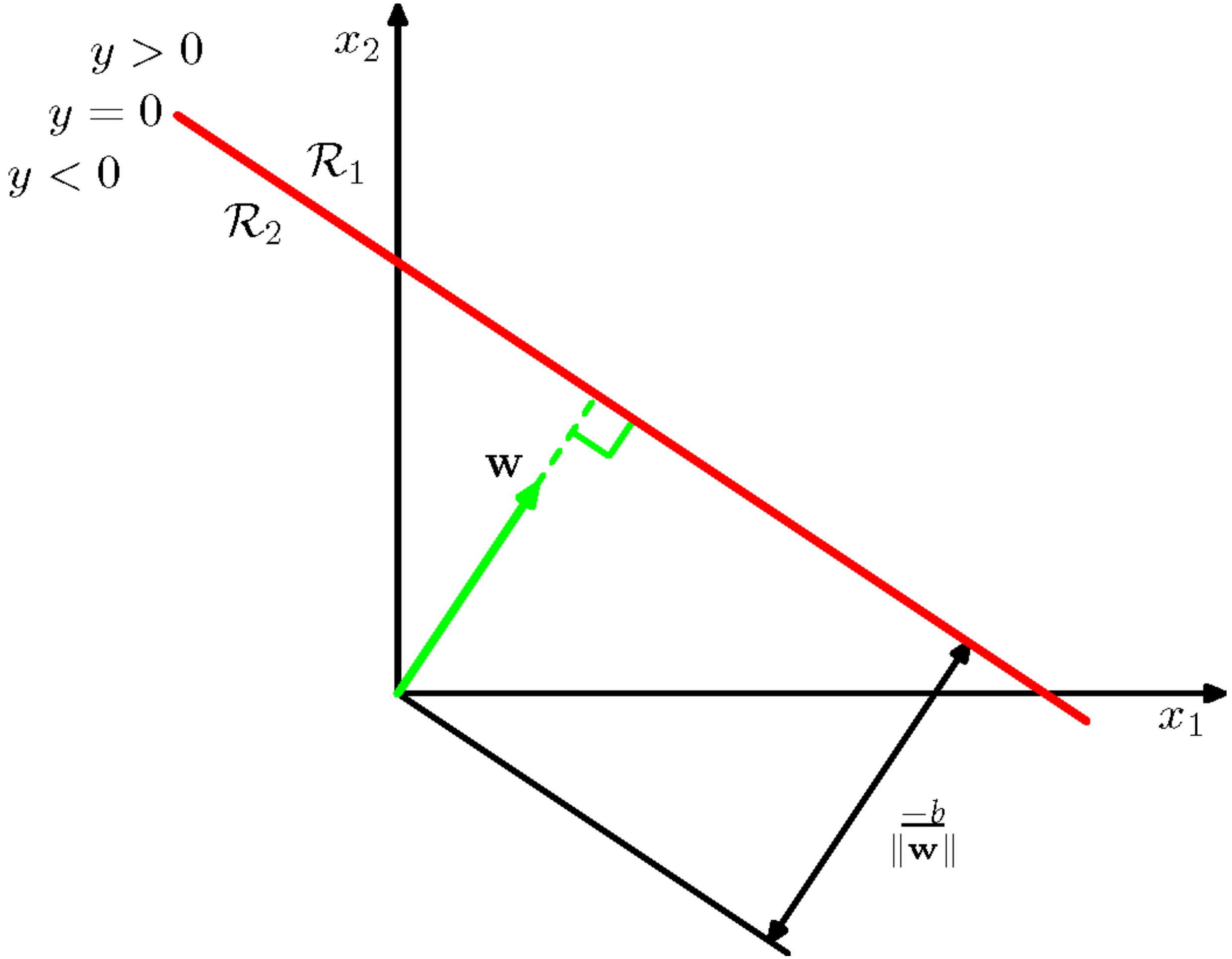
Maximum margin problem

- Given a set of data from two classes $\{\boldsymbol{x}_p, d_p\}$
  - $x_p \in \mathbb{R}^D$ and $d_p \in \{-1,1\}$
  - Assume the classes are linearly separable for now
- Find the hyperplane that separates them
  - with maximum margin
- Equation of general linear discriminant function

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$$

- Find $\boldsymbol{w}$ and b that give maximum margin
- How can we quantify margin?

w is perpendicular to the hyperplane, *b* defines its distance from the origin



$\boldsymbol{w}$ is perpendicular to the hyperplane

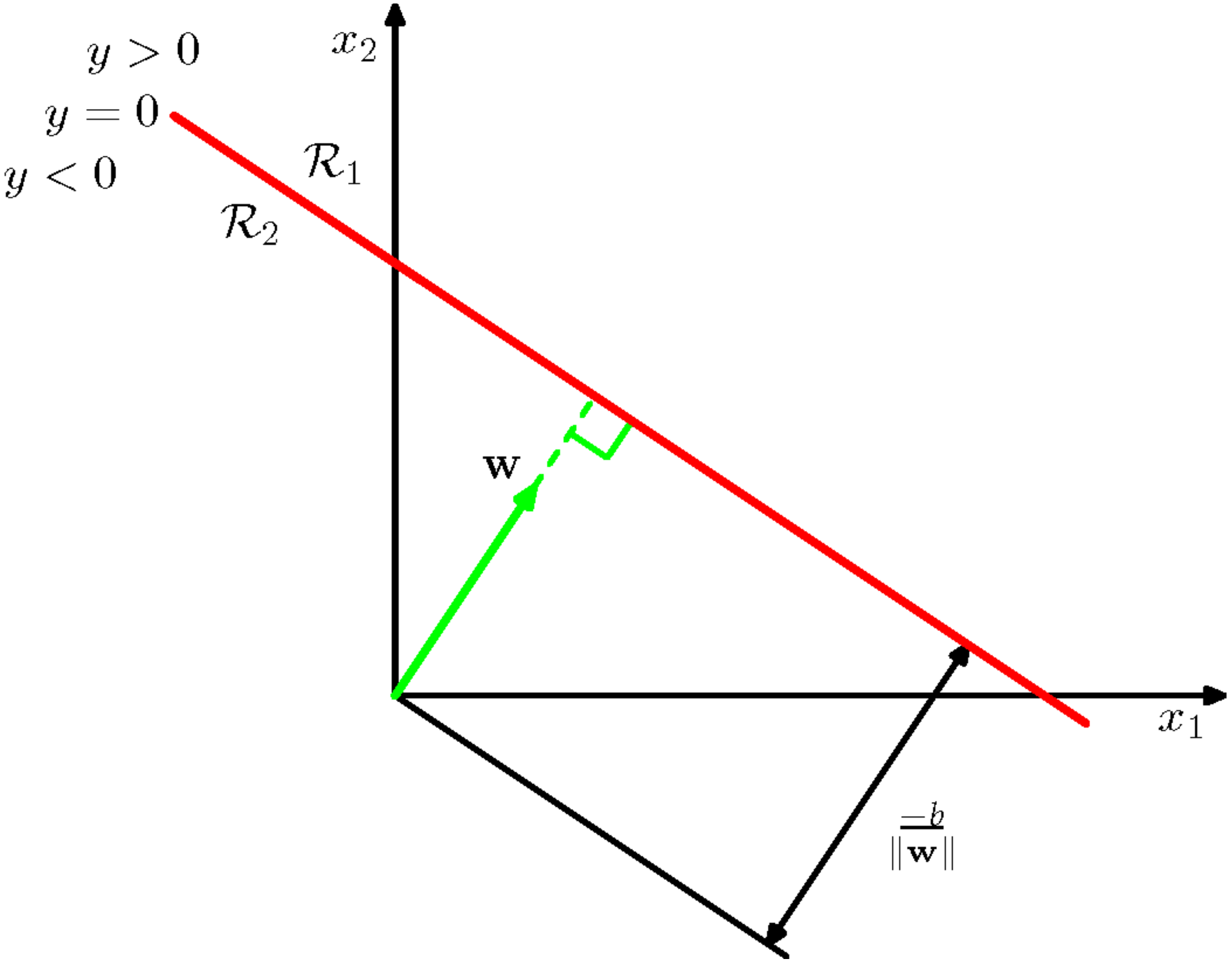
- Consider two points on the hyperplane $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$
- Then $y(\boldsymbol{x}_A) = y(\boldsymbol{x}_B) = 0$ by definition
- So $0 = y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B) = \boldsymbol{w}^T(\boldsymbol{x}_A - \boldsymbol{x}_B)$
- $\boldsymbol{x}_A - \boldsymbol{x}_B$ is a vector pointing along the hyperplane
- So $\boldsymbol{w}$ is perpendicular to the hyperplane

b defines the hyerplane's distance from the origin

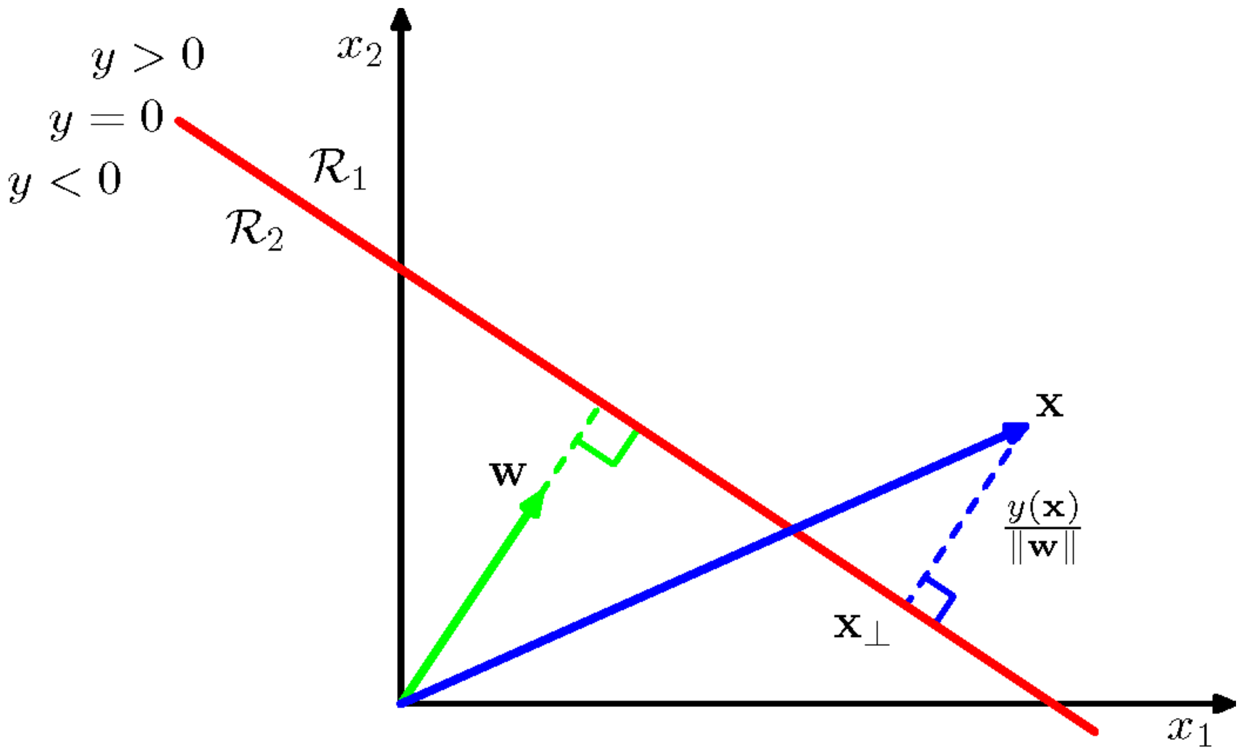
- Consider a general point $\boldsymbol{x}$
- Its distance to the origin is $D = \frac{\boldsymbol{w}^T\boldsymbol{x}}{||\boldsymbol{w}||}$
- If $\boldsymbol{x}$ is on the hyperplane, then $y(\boldsymbol{x}) = 0$
  - So $\boldsymbol{w}^T\boldsymbol{x} = -b$
- So the distance from the hyperplane to the origin is

$$D = -\frac{b}{||\boldsymbol{w}||}$$

w is perpendicular to the hyperplane, *b* defines its distance from the origin



The distance from point $\boldsymbol{x}$ to the hyperplane is $y(\boldsymbol{x})/||\boldsymbol{w}||$



The distance from point $\boldsymbol{x}$ to the hyperplane is $y(\boldsymbol{x})/||\boldsymbol{w}||$

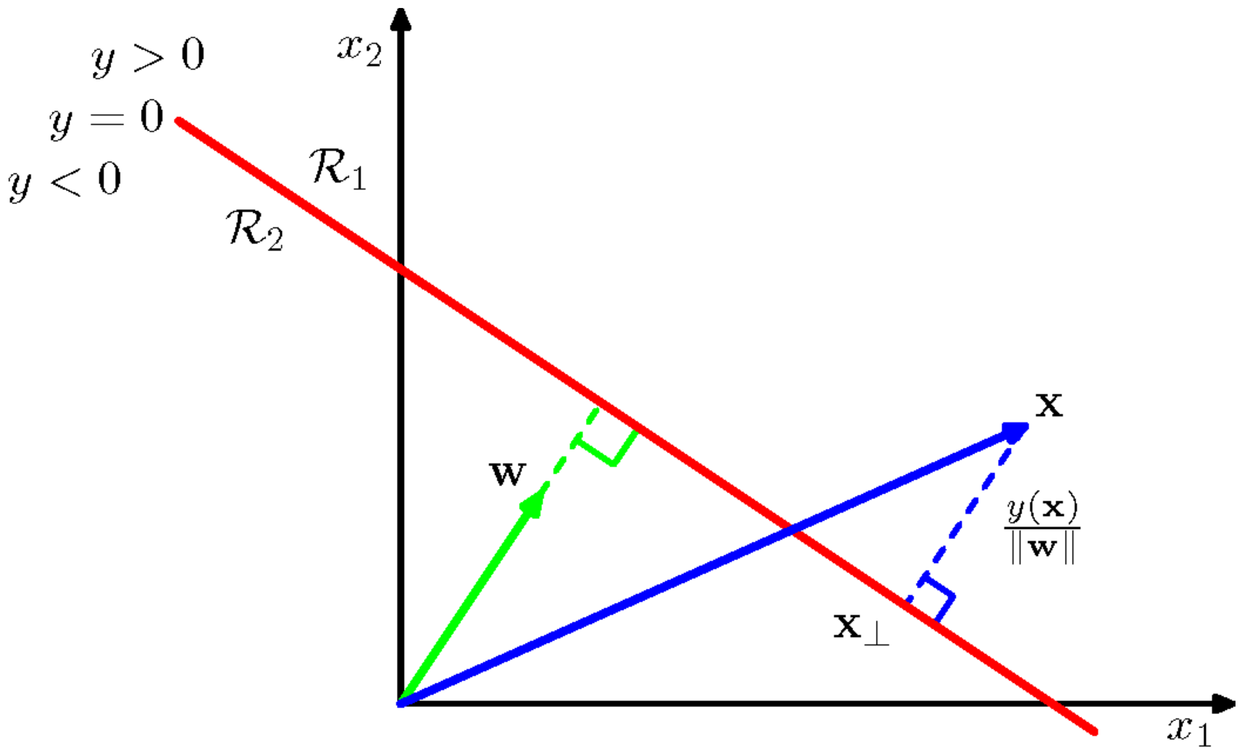
- Consider a point $\boldsymbol{x}$ and its projection onto the hyperplane $\boldsymbol{x}_\perp$ so that $\boldsymbol{x} = \boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$
- We want to find r, the distance to the hyperplane
- Multiply both sides by $\boldsymbol{w}^T$ and add b

$$\boldsymbol{w}^T\boldsymbol{x} + b = \boldsymbol{w}^T\boldsymbol{x}_\perp + b + r\frac{||\boldsymbol{w}||^2}{||\boldsymbol{w}||}$$

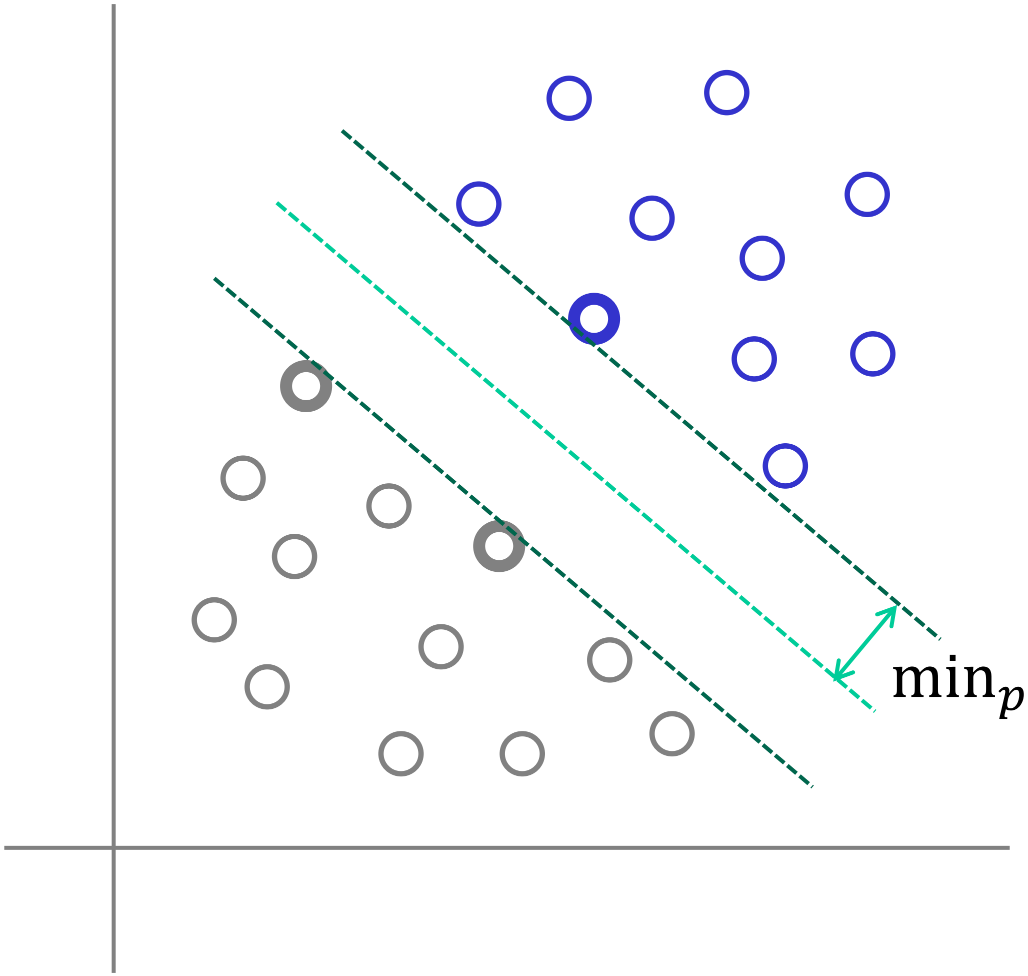
$$y(\boldsymbol{x}) = \cancel{y(\boldsymbol{x}_\perp)} + r||\boldsymbol{w}||$$

$$r = \frac{y(\boldsymbol{x})}{||\boldsymbol{w}||}$$

The distance from point $\boldsymbol{x}$ to the hyperplane is $y(\boldsymbol{x})/\|\boldsymbol{w}\|$



The maximum margin hyperplane is farthest from all of the data points



$\min_p d_p y(\boldsymbol{x}_p)/\|\boldsymbol{w}\|$

The maximum margin hyperplane is farthest from all of the data points

- The margin is defined as

$$\alpha = \min_p d_p \frac{y(\boldsymbol{x}_p)}{||\boldsymbol{w}||}$$

$$= \frac{1}{||\boldsymbol{w}||} \min_p d_p y(\boldsymbol{x}_p)$$

- We want to find $\boldsymbol{w}$ and b that maximize the margin

$$\text{argmax}_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \min_p d_p y(\boldsymbol{x}_p)$$

- Solving this problem is hard as it is written

We are free to choose a rescaling of $\boldsymbol{w}$

- If we replace $\boldsymbol{w}$ by $a\boldsymbol{w}$ and b with ab
- Then the margin is unchanged

$$\min_p d_p \frac{a\boldsymbol{w}^T\boldsymbol{x}_p + ab}{a||\boldsymbol{w}||} = \min_p d_p \frac{\boldsymbol{w}^T\boldsymbol{x}_p + b}{||\boldsymbol{w}||}$$

- So choose a such that $\min_p d_p\left(a\boldsymbol{w}^T\boldsymbol{x}_p + ab\right) = 1$
- Which means that for all points

$$d_p\left(\boldsymbol{w}^T\boldsymbol{x}_p + b\right) \geq 1$$

Maximum margin constrained optimization problem

- Then the maximum margin optimization becomes

$$\operatorname{argmax}_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \min_p d_p y(x_p)$$

$$= \operatorname{argmax}_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||}$$

- With the constraints $d_p(\boldsymbol{w}^T \boldsymbol{x}_p + b) \geq 1$

Maximum margin constrained optimization problem

- Which is equivalent to

$$\text{argmin}_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2 \;\; \text{subject to } d_p\left(\boldsymbol{w}^T \boldsymbol{x}_p + b\right) \geq 1$$

- This is a well studied type of problem
  - A quadratic program with linear inequality constraints

Detour: Lagrange multipliers solve constrained optimization problems

- Want to maximize a function $f(x_1, x_2)$
- Subject to the equality constraint $g(x_1, x_2) = 0$
- Could solve $g(x_1, x_2) = 0$ for x_1 in terms of x_2
  - But that is hard to do in general (i.e., on computers)
- Or could use Lagrange multipliers
  - Which are easier to use in general (i.e., on computers)

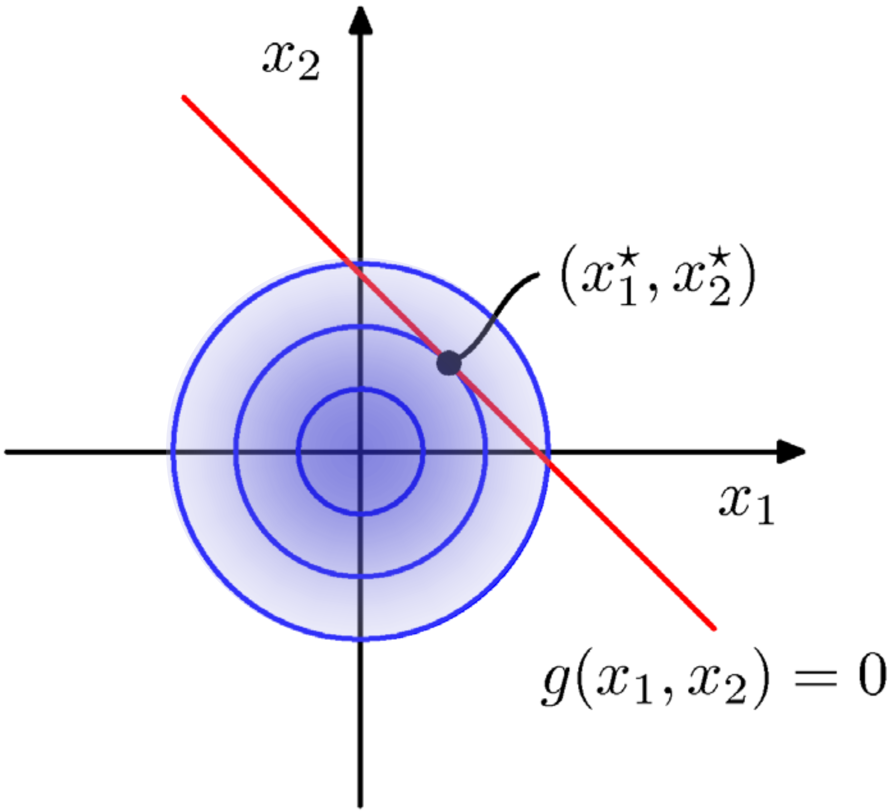
Lagrange multipliers with general $\boldsymbol{x}$

- In general, we can write

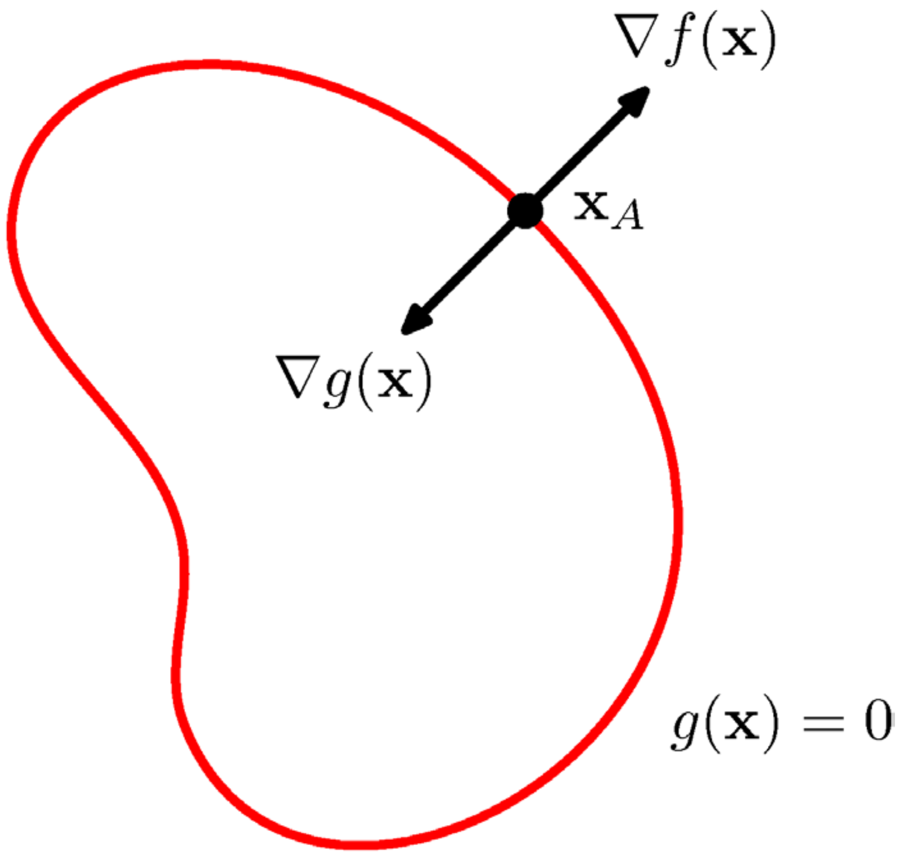
$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) \text{ subject to } g(\boldsymbol{x}) = 0$$

- Constraint $g(\boldsymbol{x}) = 0$ defines a $D - 1$ dimensional surface for D dimensional $\boldsymbol{x}$

Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = x_1 + x_2 - 1 = 0$



Gradients of g and f are orthogonal to surface at solution point



Gradients of g and f are orthogonal to surface at maximum of f

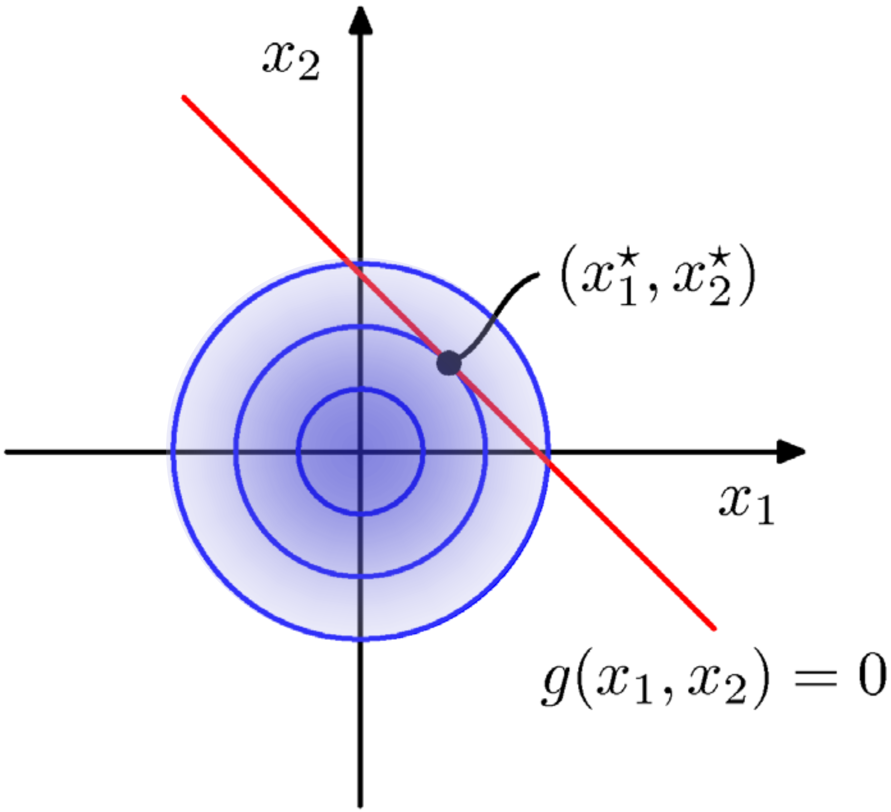
- For g because on all points on the surface $g(\boldsymbol{x}) = 0$
- For f because if it wasn't, you could move along the surface in the direction of the gradient to find a better maximum
- Thus ∇f and ∇g are (anti-)parallel
- And there must exist a scalar λ such that
$$\nabla f + \lambda \nabla g = 0$$

The Lagrangian function captures the constraints on x and on the gradients

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

- Setting gradient of L with respect to $\boldsymbol{x}$ to 0 gives
$$\nabla f + \lambda \nabla g = 0$$
- Setting partial of L with respect to λ to 0 gives
$$g(\boldsymbol{x}) = 0$$
- Thus stationary points of L solve the constrained optimization problem

Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = x_1 + x_2 - 1 = 0$



Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = x_1 + x_2 - 1 = 0$

- So the Lagrangian function is

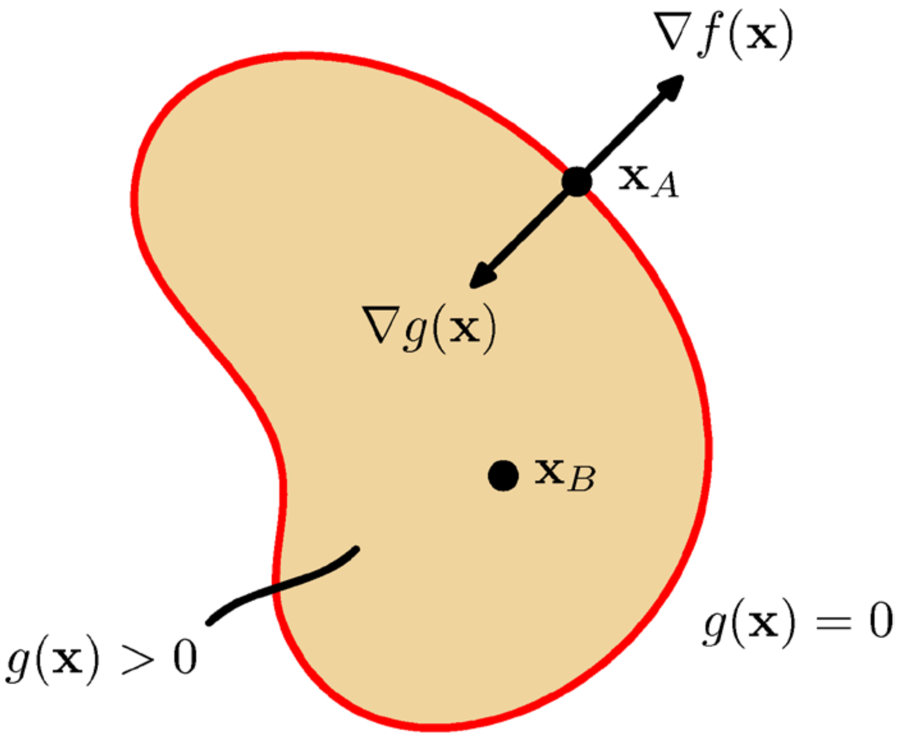
$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$
$$= 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

- The conditions for L to be stationary are

$$\partial L / \partial x_1 = -2x_1 + \lambda = 0$$
$$\partial L / \partial x_2 = -2x_2 + \lambda = 0$$
$$\partial L / \partial \lambda \;\; = x_1 + x_2 - 1 = 0$$

- Can solve to find $\lambda = 1, x_1 = x_2 = \frac{1}{2}$

Lagrange multipliers can also be used with inequality constraints $g(\boldsymbol{x}) \geq 0$



Lagrange multipliers can also be used with inequality constraints $g(\boldsymbol{x}) \geq 0$

- Now two kinds of solutions:
- If $g(\boldsymbol{x}) > 0$, then the solution only depends on $f(\boldsymbol{x})$
  - Inside constraint surface with $\nabla f = 0$
  - Stationary point of $L(\boldsymbol{x}, \lambda)$ with $\lambda = 0$
  - Constraint $g(\boldsymbol{x})$ is said to be inactive
- If $g(x) = 0$, then same as before (with equality constraint)
  - On boundary of constraint surface with ∇f pointing out
  - Stationary point of $L(\boldsymbol{x}, \lambda)$ with $\lambda > 0$
  - Constraint $g(\boldsymbol{x})$ is said to be active

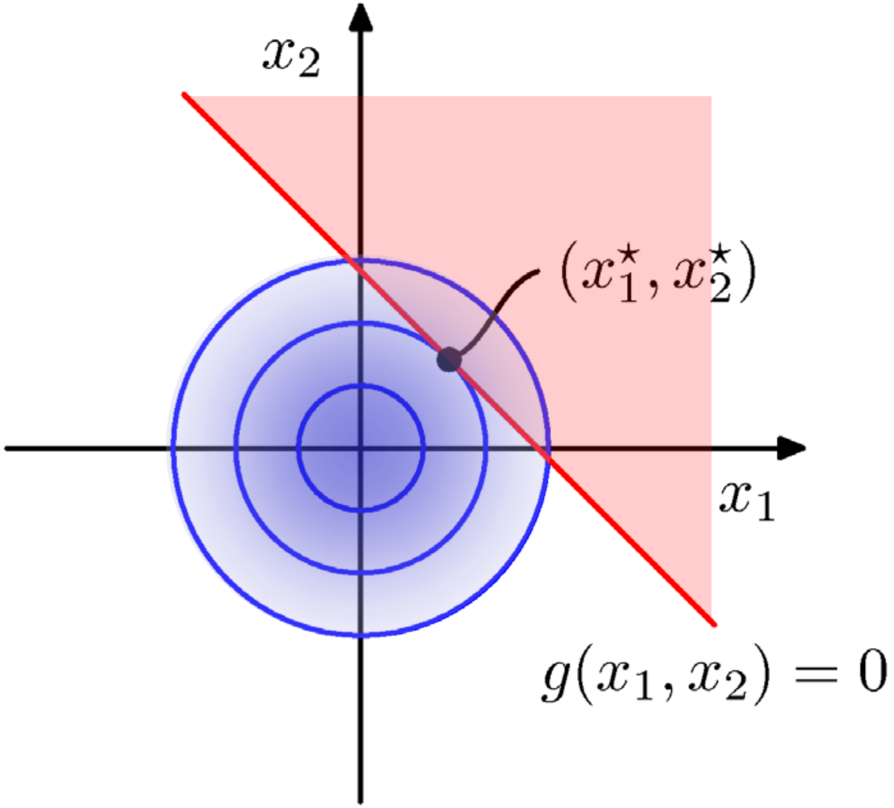
Lagrange multipliers can also be used with inequality constraints $g(\boldsymbol{x}) \geq 0$

- In either case, $\lambda g(\boldsymbol{x}) = 0$
- Thus maximizing $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) \geq 0$ is obtained by optimizing $L(\boldsymbol{x}, \lambda)$ WRT $\boldsymbol{x}$ and λ subject to

$$g(\boldsymbol{x}) \geq 0$$
$$\lambda \geq 0$$
$$\lambda g(\boldsymbol{x}) = 0$$

- These are known as the Karush-Kuhn-Tucker (KKT) conditions

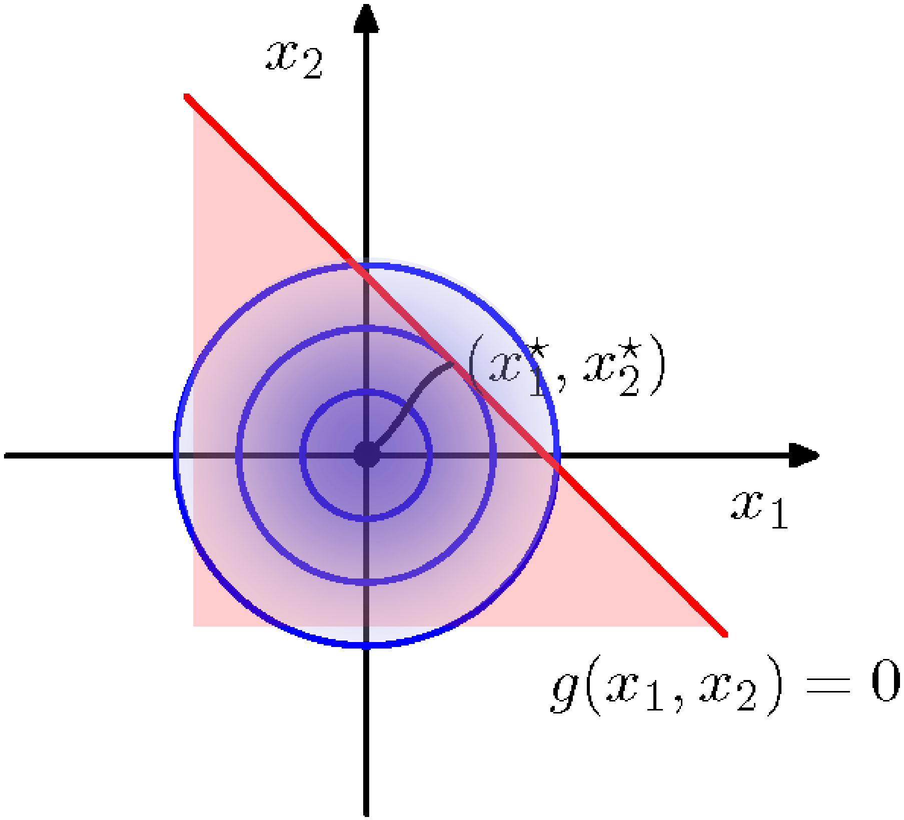
Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = x_1 + x_2 - 1 \geq 0$



Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = x_1 + x_2 - 1 \geq 0$

- So the Lagrangian function is
$$L(\boldsymbol{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$
- The conditions for L to be stationary are
$$\partial L / \partial x_1 = -2x_1 + \lambda = 0$$
$$\partial L / \partial x_2 = -2x_2 + \lambda = 0$$
$$\partial L / \partial \lambda = x_1 + x_2 - 1 = 0$$
- Can solve to find $\lambda = 1$, $x_1 = x_2 = \frac{1}{2}$
  - Which still satisfies KKT conditions

Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = -x_1 - x_2 + 1 \geq 0$



Example: Maximize $f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2$ subject to $g(\boldsymbol{x}) = -x_1 - x_2 + 1 \geq 0$

- So the Lagrangian function is
$$L(\boldsymbol{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(-x_1 - x_2 + 1)$$
- The conditions for L to be stationary are
$$\partial L/\partial x_1 = -2x_1 - \lambda = 0$$
$$\partial L/\partial x_2 = -2x_2 - \lambda = 0$$
$$\partial L/\partial \lambda \ \ = x_1 + x_2 - 1 = 0$$
- Can solve to find $\lambda = -1$
  - which does not satisfy KKT condition $\lambda \geq 0$
- Instead use unconstrained solution $x_1 = x_2 = 0$
  - which does satisfy KKT conditions

Multiple constraints each get their own Lagrange multiplier

- Maximize $f(\boldsymbol{x})$ subject to $g_i(\boldsymbol{x}) = 0$ and $h_j(\boldsymbol{x}) \geq 0$
- Leads to the Lagrangian function

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \mu_j h_j(\boldsymbol{x})$$

- Still solve for $\nabla L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0$
- Trickier in general to figure out which $h_j(\boldsymbol{x})$ constraints should be active

Minimizing $f(\boldsymbol{x})$ with an inequality constraint requires a slightly different Lagrangian

- Minimize WRT $\boldsymbol{x}$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$$

- Still subject to

$$g(\boldsymbol{x}) \geq 0$$

Summary of Lagrange multipliers with multiple inequality constraints

- Goal: maximize $f(\boldsymbol{x})$ subject to $g_i(\boldsymbol{x}) \geq 0$
- Write down Lagrangian function

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x})$$

- Find points where $\nabla L(\boldsymbol{x}, \lambda) = 0$
- Keep points that satisfy constraints $g_i(\boldsymbol{x}) \geq 0$
- Figure out which KKT conditions should be active
  - Don't need to try all 2^I combinations for SVMs
  - Because $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ form a "quadratic program"

Back to SVMs: Maximum margin solution is a fixed point of the Lagrangian function

- Recall, the maximum margin hyperplane is

$$\text{argmin}_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \;\; \text{subject to } d_p\left(\boldsymbol{w}^T\boldsymbol{x}_p + b\right) \geq 1$$

  - Minimization of a quadratic function subject to multiple linear inequality constraints
- Will use Lagrange multipliers, a_p, to write Lagrangian function

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_p a_p(d_p\left(\boldsymbol{w}^T\boldsymbol{x}_p + b\right) - 1)$$

- Note that $\boldsymbol{x}_p$ and d_p are fixed for the optimization