University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 12 - Kernel Methods
Support Vector Machines - Inequality
Constraint Optimization - Kernel SVM

Erik Bekkers

Catch-up:
Probabilistic PCA
Non-linear PCA

Image credit: Kirillm | Getty Images

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 11.3 - Kernel Methods
Support Vector Machines - Maximum Margin Classifier

Erik Bekkers

(Bishop 7.1.0)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Support vector machines

- Kernel method with sparse solutions:
  - prediction for new inputs depend only on kernel function evaluated at a **subset** of the training points

- Applications:
  - Classification
  - Regression
  - novelty detection/anomaly detection

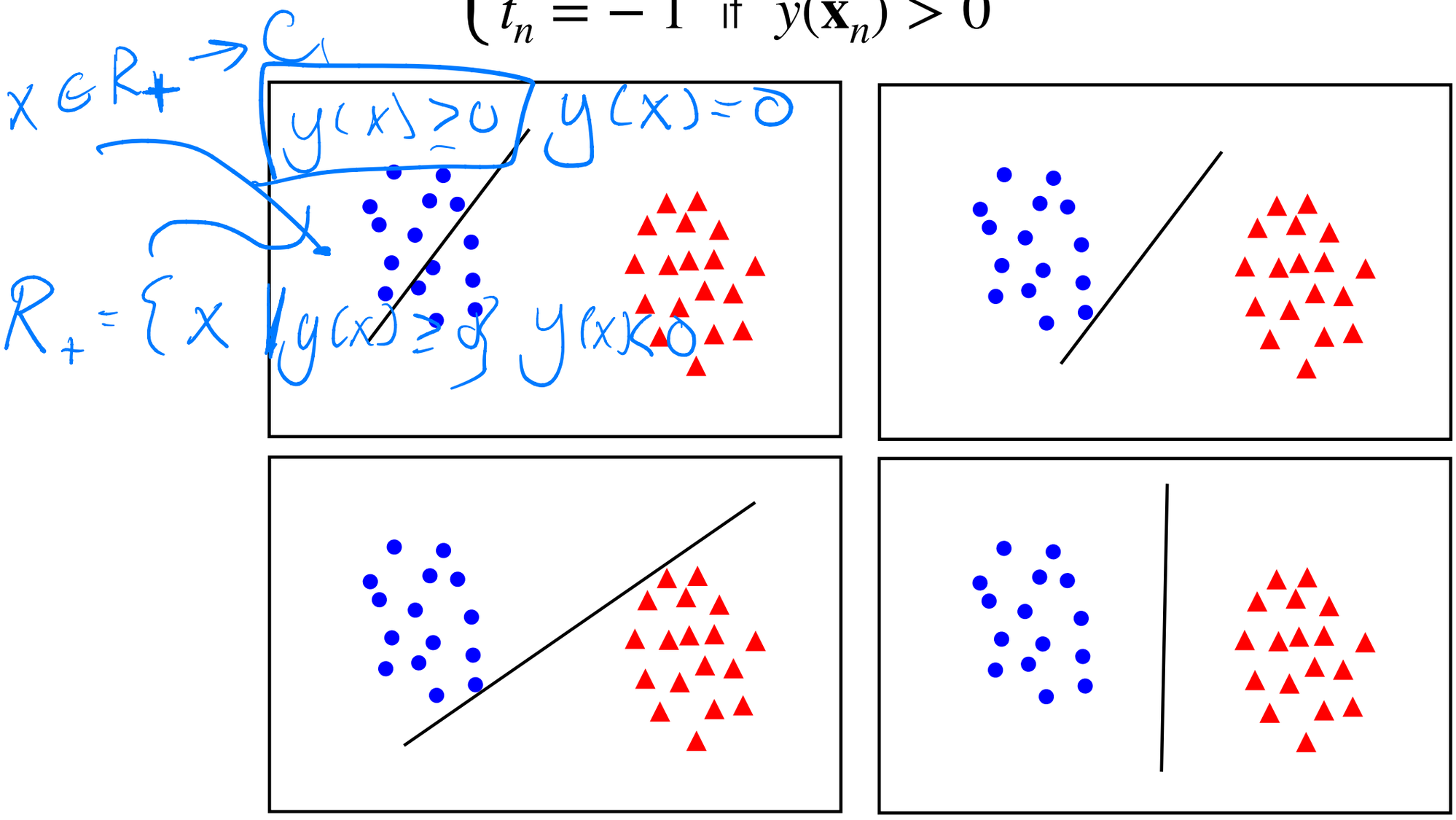
- Convex optimization problem, any local solution is at global optimum!

- No good probabilistic interpretation

- Today: SVM for binary classification -> maximum margin classifier!

Linearly separable dataset

- Linear classifier: $y(\mathbf{x_n}) = \mathbf{w}^t\mathbf{x}_n + b$
- Classification: $\begin{cases} t_n = +1 \text{ if } y(\mathbf{x}_n) > 0 \\ t_n = -1 \text{ if } y(\mathbf{x}_n) > 0 \end{cases}$



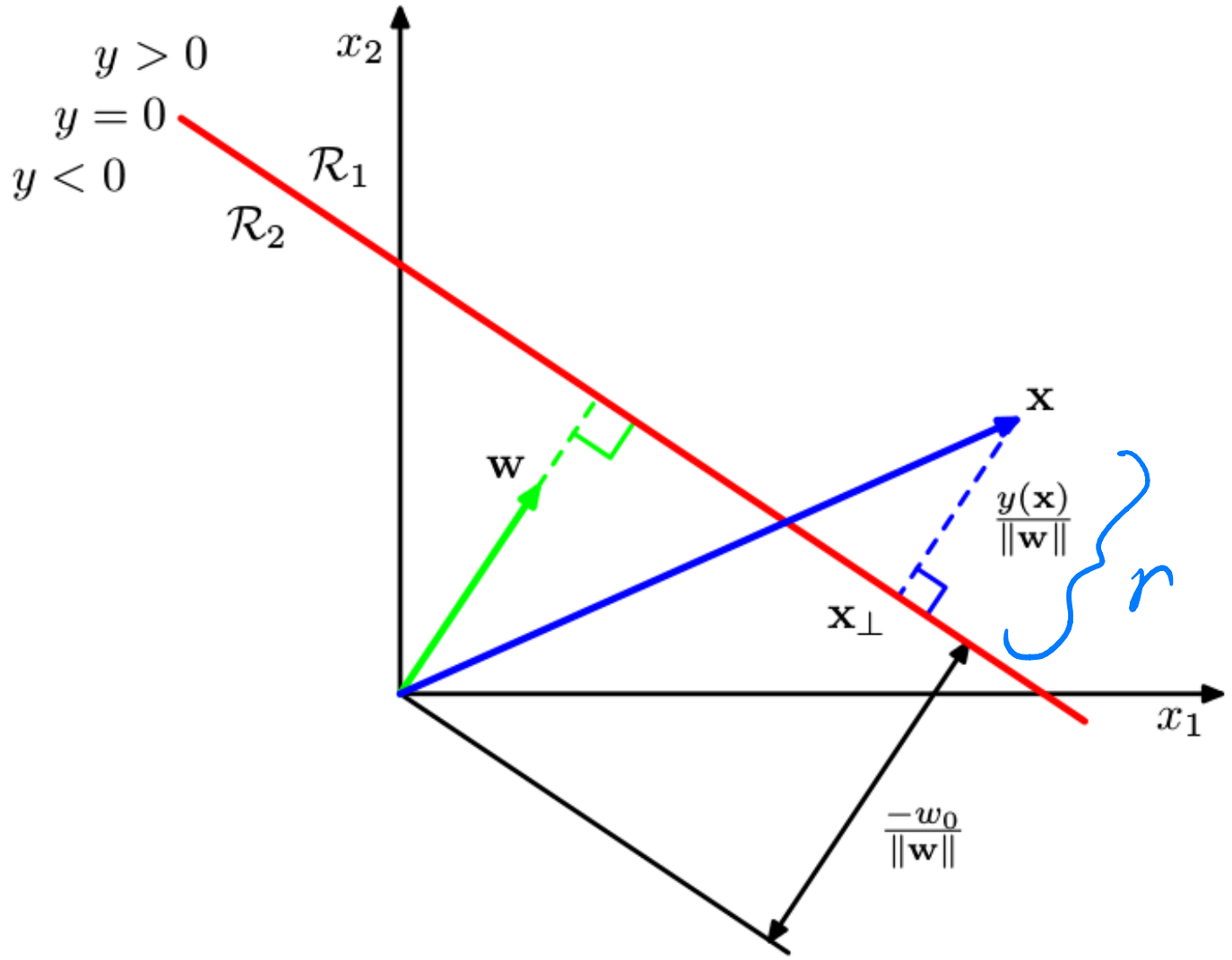
- Maximum Margin: most stable under perturbations of the input

Linearly Separable Dataset

- If $\mathbf{x}'$ lies on decision boundary: $y(\mathbf{x}') = \mathbf{w}^T\mathbf{x}' + b = 0$

- Recall: distance from $\mathbf{x}$ to decision boundary is

$$r = \frac{|y(\mathbf{x}_n)|}{\|\mathbf{w}\|} = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|}$$



- For correct classification:

$$y(\mathbf{x}_n) > 0 \text{ if } t_n = +1$$

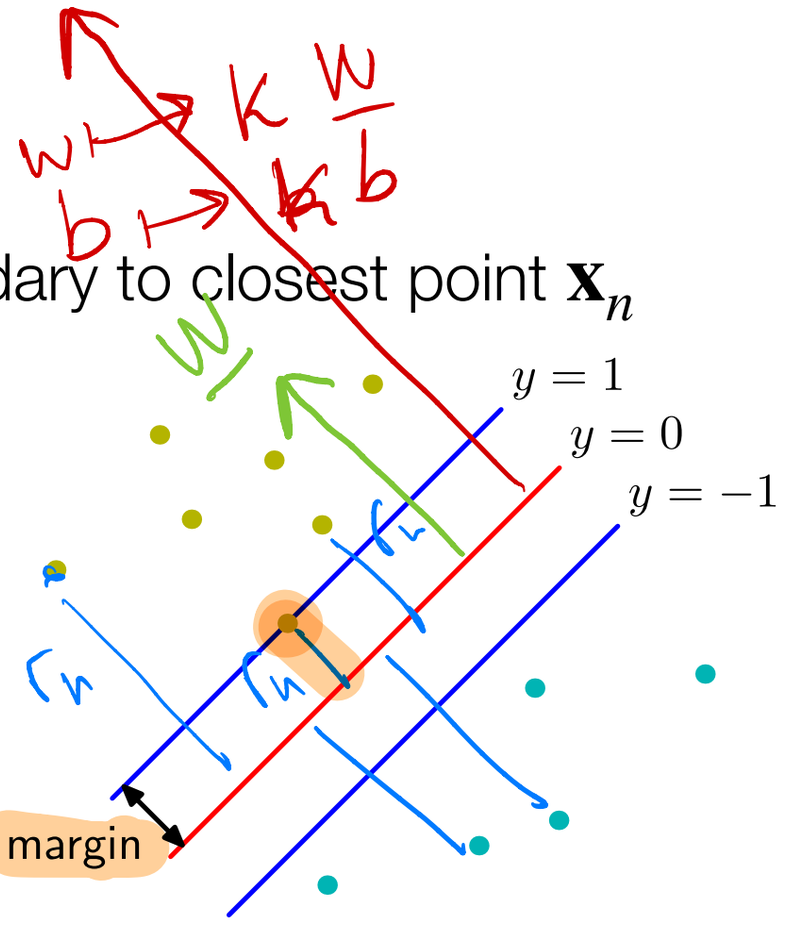
$$y(\mathbf{x}_n) < 0 \text{ if } t_n = -1$$

- So for all $n = 1, \ldots, N$

$$t_n y(\mathbf{x}_n) > 0$$

Maximum Margin Classifier

- Margin: perpendicular distance from decision boundary to closest point $\mathbf{x}_n$

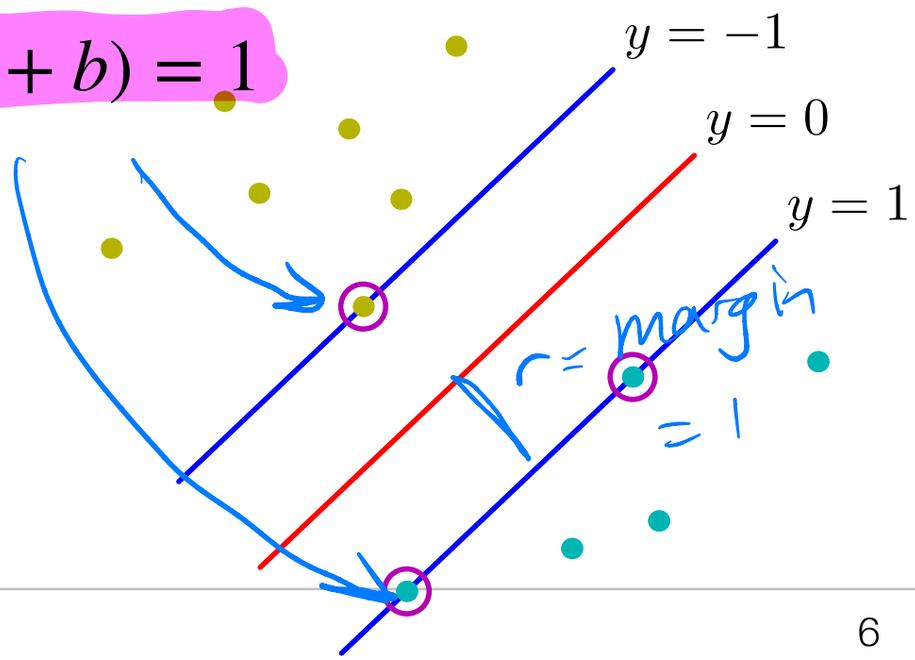


- For all data points distance to decision boundary is

$$r_n = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{\|\mathbf{w}\|}$$

- Margin: $\min_n \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{\|\mathbf{w}\|} = \min_n \frac{t_n(\kappa\mathbf{w}^T\mathbf{x}_n + \kappa b)}{\|\kappa\mathbf{w}\|}$

- For point closest to decision boundary $t_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$



- For all data points: $t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$

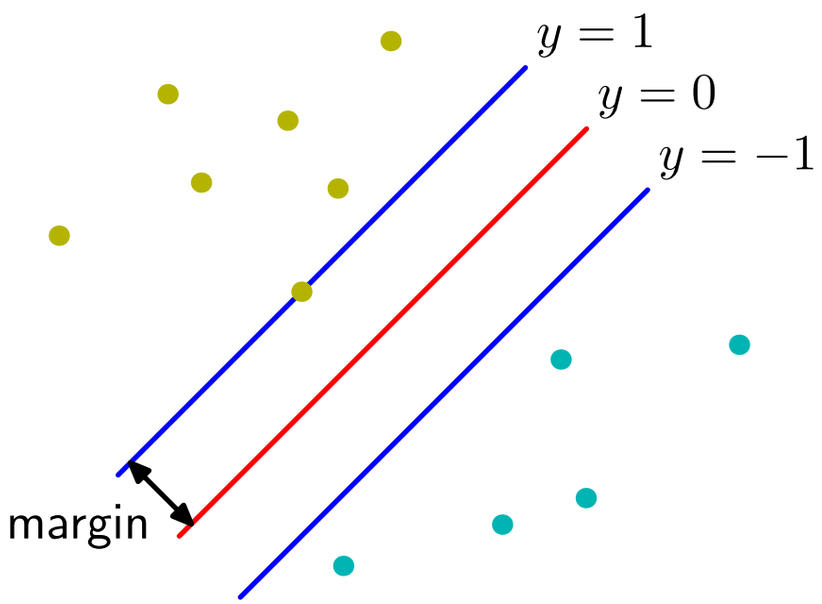
- Maximum margin classifier:

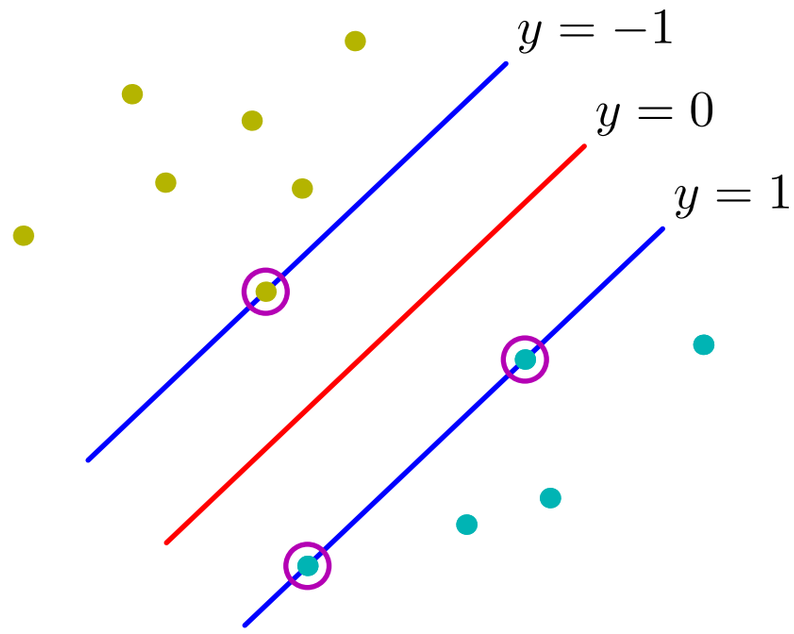
Maximum Margin Classifier

- For all data points distance to decision boundary is

$$r = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{\|\mathbf{w}\|}$$

- For point closest to decision boundary $t_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$
- For all data points: $t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$
- Size of the margin: $\frac{1}{\|\mathbf{w}\|}$



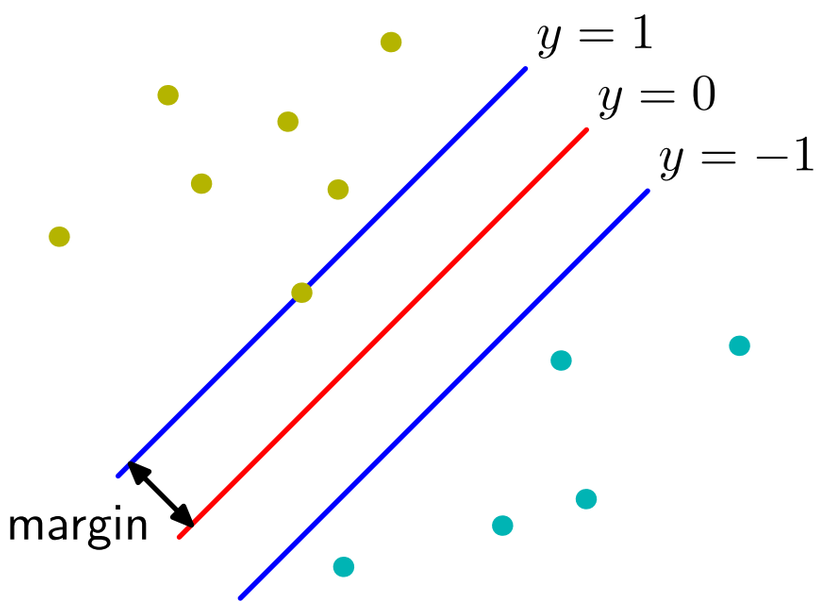


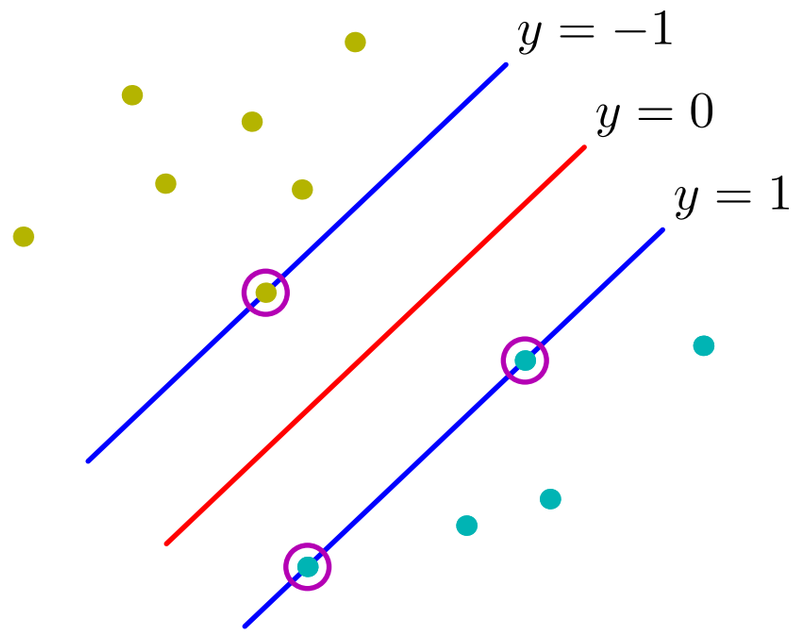
Maximum Margin Classifier

- Size of the margin: $\frac{1}{\|\mathbf{w}\|}$ $\Leftrightarrow$ $\arg\min \frac{1}{2}\|\mathbf{w}\|^2$
- For all data points: $t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$
- **Maximizing the margin**:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- Quadratic programming problem!





University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 11.4 - Kernel Methods
Intermezzo: Constraint Optimization

Erik Bekkers

(Bishop E, 7.1)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Constrained optimization (**equality constraint**)

- Consider a constrained optimization problem of the form

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

- *It is solved via the method of Lagrange multipliers:*
  1. *Define the Lagrangian:* $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$
  2. *Find stationary points of* $L(\mathbf{x}, \lambda)$*:* $\frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}, \lambda) = 0$ *and* $\frac{\partial}{\partial \lambda} L(\mathbf{x}, \lambda) = 0$

1. $\nabla g(\mathbf{x})$ is perpendicular to constraint surface $\mathcal{S}$

   ($\mathcal{S}$ is surface of points $\mathbf{x}$ along which g is constant, i.e., $\mathcal{S} = \{\mathbf{x} \mid g(\mathbf{x}) = c\}$)

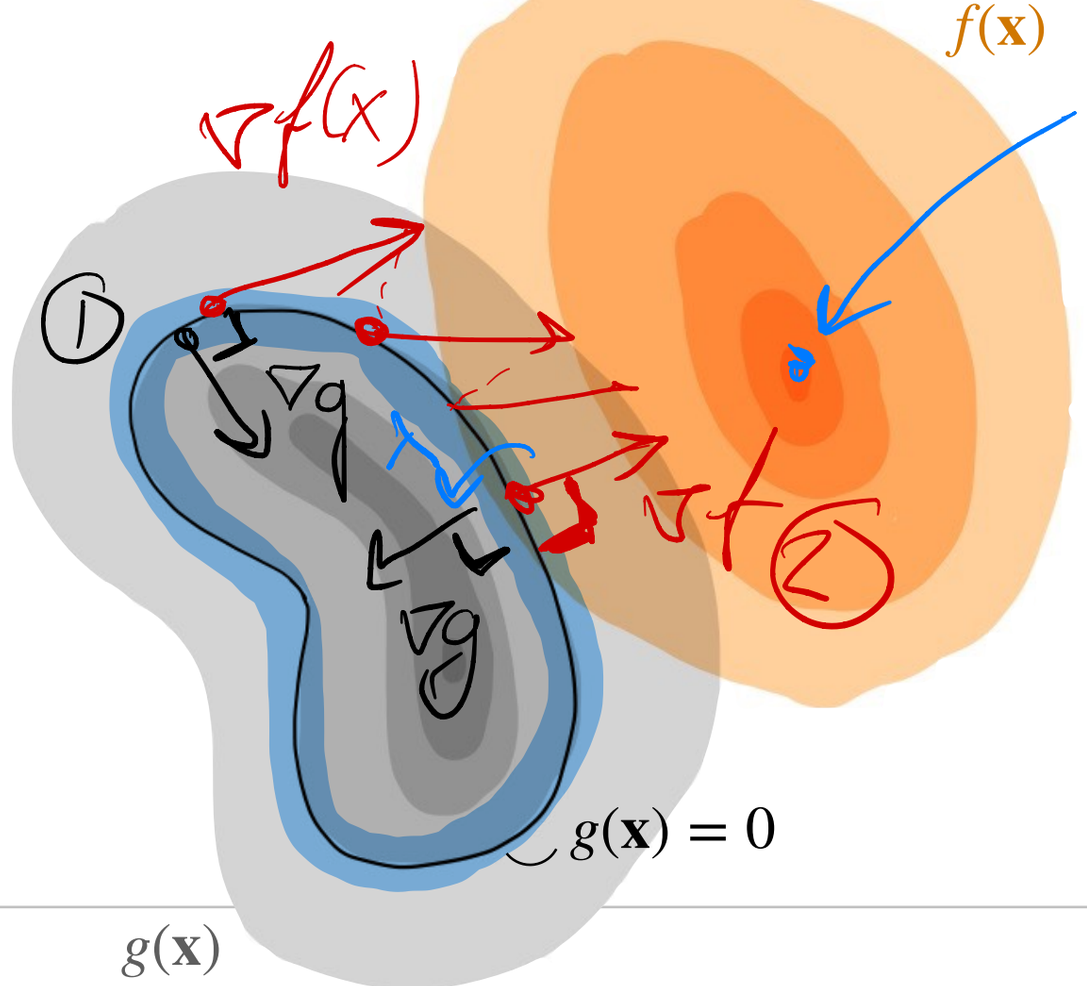
2. At constrained maximum, $\nabla f(\mathbf{x})$ must also be perpendicular to $\mathcal{S}$

- Therefore there exists a $\lambda \in \mathbb{R}$ s.t.

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

$$\Leftrightarrow$$

$$\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0$$



Constrained optimization (**inequality constraint**)

- Consider an inequality constrained optimization problem of the form **Primal problem**

$$\max_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \geq 0$$

- It is solved by the min-max problem (primal-dual optimization):
  - Define Lagrangian: $L(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$
  - Solve $\max_{\mathbf{x}} \min_{\mu} L(\mathbf{x}, \mu)$ subject to **Karush-Kuhn-Tucker conditions**:
    - primal feasibility: $g(\mathbf{x}) \geq 0$
    - dual feasibility: $\mu \geq 0$
    - complementary slackness: $\mu g(\mathbf{x}) = 0$

(* stationarity: $\frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}, \mu) = 0$ and $\frac{\partial}{\partial \mu} L(\mathbf{x}, \mu) = 0$)

- Optimal points satisfy (gradients must be anti-parallel!)

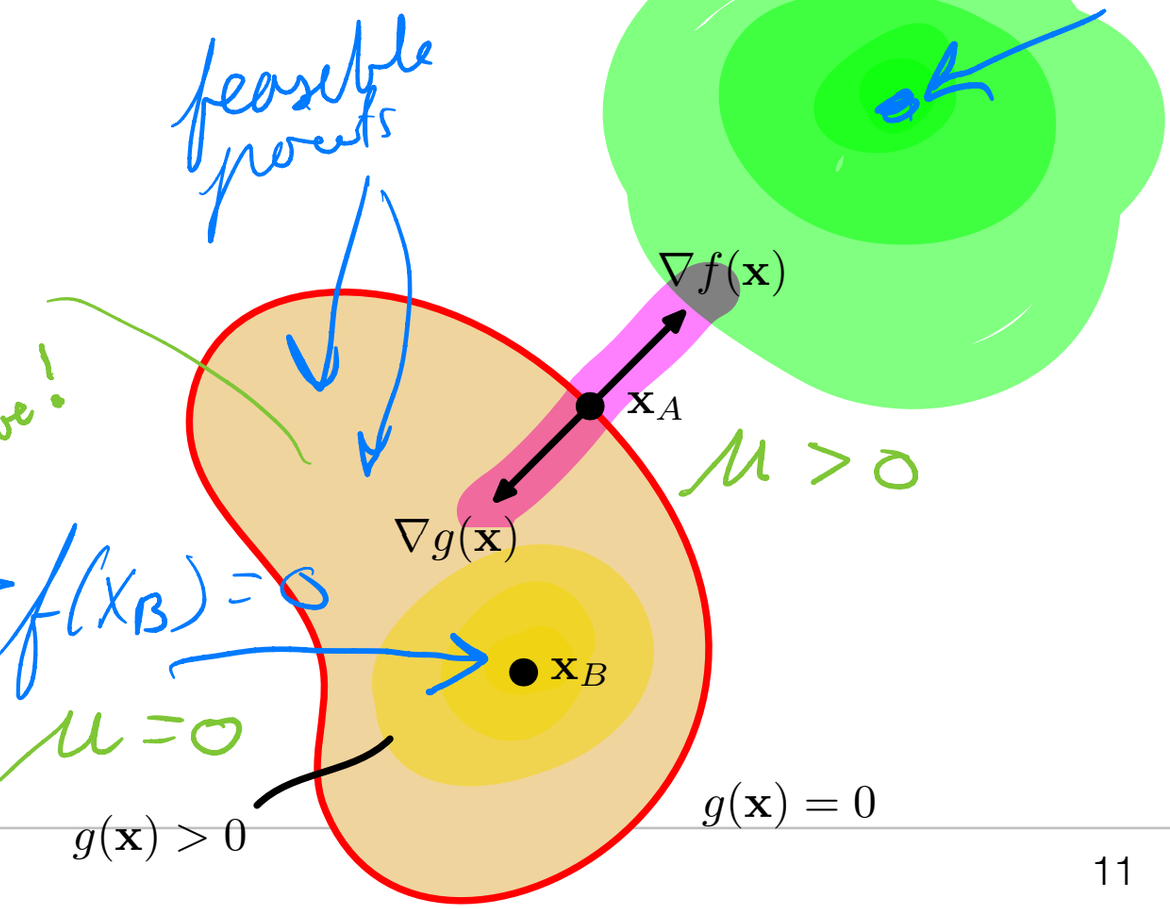
$$\nabla f(\mathbf{x}) + \mu \nabla g(\mathbf{x}) = 0 \quad \text{with } \mu \geq 0$$

- Follows same reasoning as in equality constraint case, however, now there are two types of solutions:
  - Stationary point in region $g(\mathbf{x}) \geq 0$: (*inactive constraint*)

$$\nabla f(\mathbf{x}) = 0, \qquad \mu = 0$$

  - Stationary point on boundary $g(\mathbf{x}) = 0$: (*active constraint!*)

$$\nabla f(\mathbf{x}) = -\mu \nabla g(\mathbf{x}), \qquad \mu > 0$$



Constrained optimization (**inequality constraint**)

- Consider an inequality constrained optimization problem of the form **Primal problem**

$$\max_{\mathbf{x}} \; f(\mathbf{x}) \qquad \text{subject to} \qquad g(\mathbf{x}) \geq 0$$

- Solved by $\max_{\mathbf{x}} \min_{\mu} L(\mathbf{x}, \mu)$ subject to **Karush-Kuhn-Tucker conditions**

1. Define **Dual Lagrangian** (optimize w.r.t. primal variables $\mathbf{x}$ for fixed dual variable μ)

$$\tilde{L}(\mu) = \max_{\mathbf{x}} L(\mathbf{x}, \mu) \qquad \text{with} \qquad \text{Primal Lagrangian } L(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$$

2. Obtain dual Langrangian $\tilde{L}(\mu)$ analytically:
   - Use stationarity condition $\nabla_{\mathbf{x}} L = 0$ to eliminate $\mathbf{x}$ from L
   - This gives $\tilde{L}$ which now only depends on μ
   - This is an upper bound for (1) as function of μ

3. Minimize **Duality gap** (solve **Dual Problem**):
   - For every $\mathbf{x}'$ satisfying $g(\mathbf{x}') \geq 0$ we have ← feasible point

$$f(\mathbf{x}') \leq L(\mathbf{x}', \mu) \leq \tilde{L}(\mu)$$

   - It follows (weak duality):

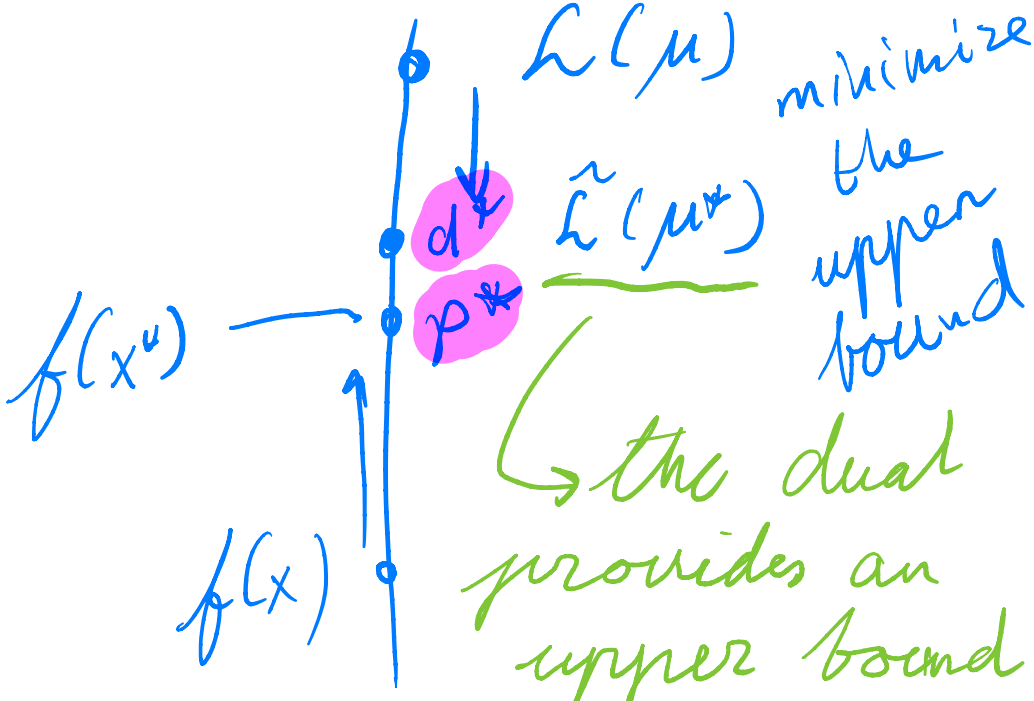
$$\mathbf{p}^* = \max_{\mathbf{x},\, g(\mathbf{x}) \geq 0} f(\mathbf{x}) \leq \min_{\mu} \tilde{L}(\mu) = \mathbf{d}^*$$

- **feasible point x'**: $g(x') \geq 0$
- we also have $\mu \geq 0$

$$\mathcal{L}(x', \mu) = f(x') + \underbrace{\mu \cdot g(x')}_{\geq 0}$$

$$\Rightarrow f(x') \leq \mathcal{L}(x', \mu) \leq \tilde{\mathcal{L}}(\mu)$$

- $\tilde{\mathcal{L}}(\mu) = \max_x \mathcal{L}(x, \mu)$



Constrained optimization (**inequality constraint**)

- Consider an inequality constrained optimization problem of the form **Primal problem**

$$\max_{\mathbf{x}} \; f(\mathbf{x}) \qquad \text{subject to} \qquad g(\mathbf{x}) \geq 0$$

- Solved by $\max_{\mathbf{x}} \min_{\mu} L(\mathbf{x}, \mu)$ subject to **Karush-Kuhn-Tucker conditions**

- For almost all convex problems:
  - Strong duality $\mathbf{p}^* = \mathbf{d}^*$
  - So if we have solved the dual problem, we have solved the primal problem!

- So solve the dual problem (find the lowest upper bound): **Dual problem**

$$\min_{\mu} \tilde{L}(\mu) \qquad \text{subject to} \qquad \mu \geq 0$$

- Recipe:
  1. Define Lagrangian $L(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$
  2. Compute dual Lagrangian $\tilde{L}(\mu)$
  3. Solve dual problem $\mu^* = \operatorname{argmin}_{\mu} \tilde{L}(\mu)$ subject to $\mu \geq 0$
  4. Maximize primal Lagrangian $\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x}} L(\mathbf{x}, \mu^*)$

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 11.5 - Kernel Methods
Support Vector Machines - Kernel SVM

Erik Bekkers

(Bishop 7.1.0)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Maximum Margin Classifier

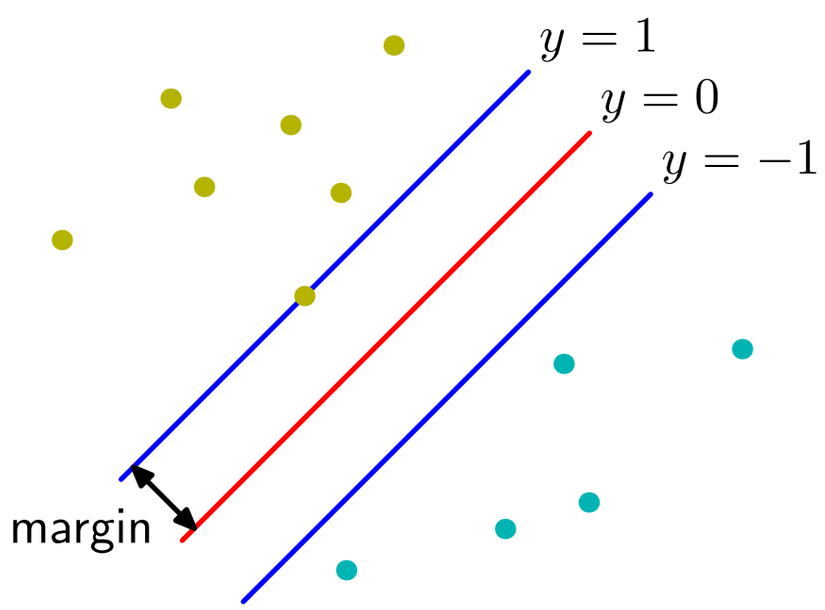
- Maximizing the margin:

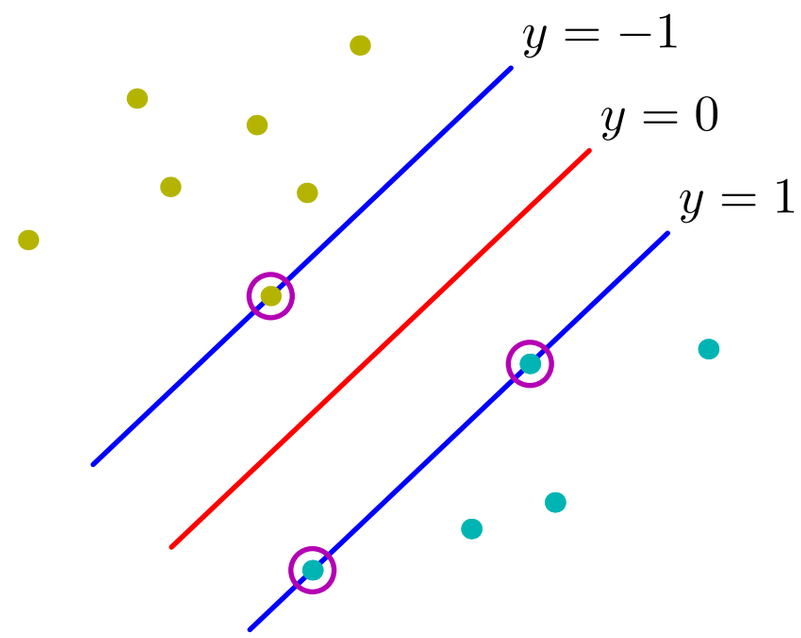
$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- We decided to "calibrate" $\mathbf{w}$ s.t. for the nearest point $t_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$

- Then the size of the margin is given by $\frac{1}{\|\mathbf{w}\|}$

- And for all data points we have $t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$





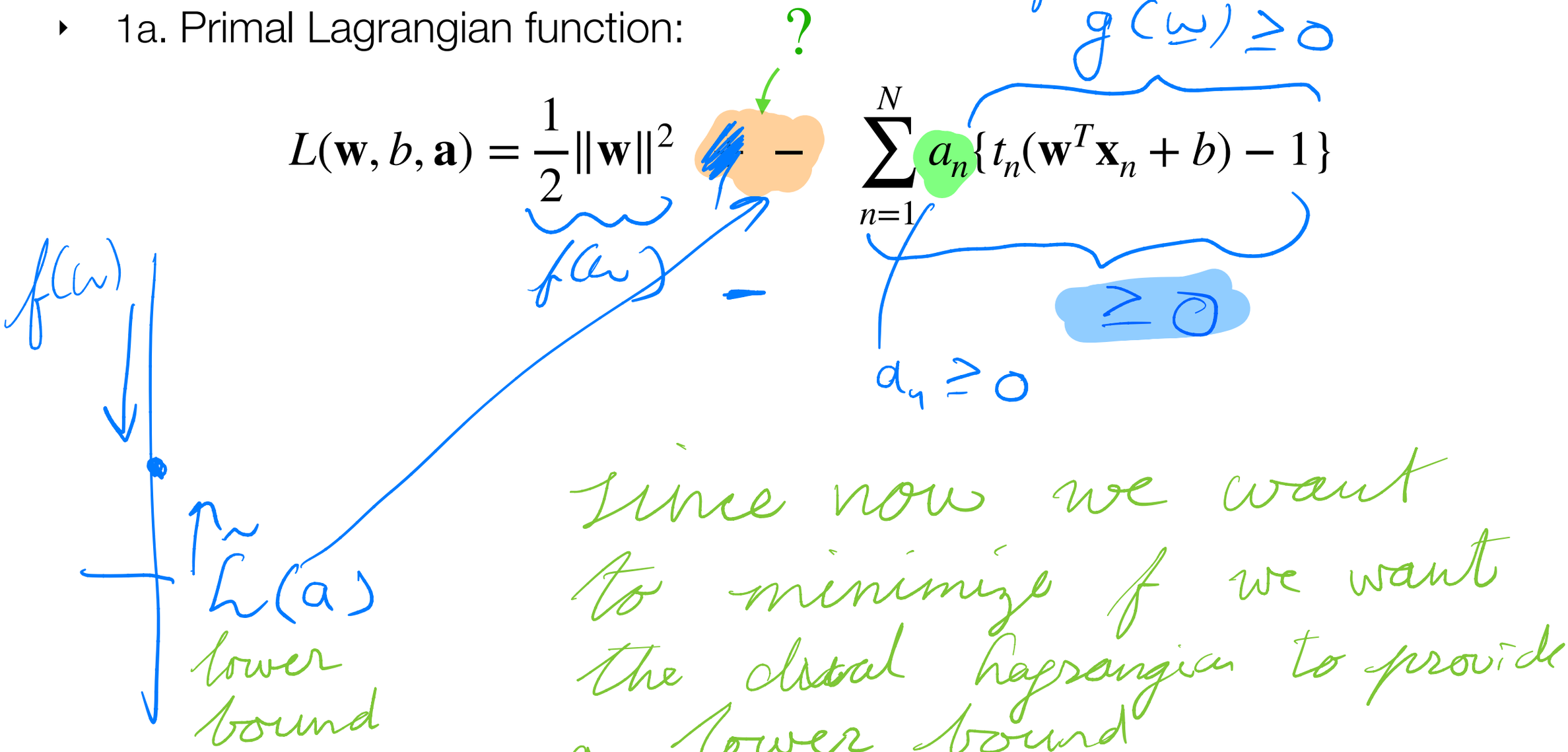
Maximum Margin Classifier

- Maximizing the margin:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- 1a. Primal Lagrangian function:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$$



Maximum Margin Classifier

- Maximizing the margin:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- 1a. Primal Lagrangian function:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$$

- 1b. With KKT conditions:

(primal feasibility)	$t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \geq 0$	for	$n = 1,\ldots,N$
(dual feasibility)	$a_n \geq 0$	for	$n = 1,\ldots,N$
(complimentary slackness)	$a_n(t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1) = 0$	for	$n = 1,\ldots,N$

$3 \cdot N$ constraints

3 per data point

Maximum Margin Classifier

- Maximizing the margin:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- 1a. Primal Lagrangian function:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$$

- 1b. With KKT conditions:

(primal feasibility) $t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \geq 0$ for $n = 1,\ldots,N$

(dual feasibility) $a_n \geq 0$ for $n = 1,\ldots,N$

(complimentary slackness) $a_n(t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1) = 0$ for $n = 1,\ldots,N$

- 2. Dual Lagrangian obtained via (stationarity conditions) $\frac{\partial L}{\partial \mathbf{w}} = 0, \quad \frac{\partial L}{\partial b} = 0$

$$\tilde{L}(\mathbf{a}) = \min_{\mathbf{x}.b} L(\mathbf{x}, b, \mathbf{a})$$

Maximum Margin Classifier

- Maximizing the margin:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } N \text{ constraints } t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

- 1a. Primal Lagrangian function:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$$

- 1b. With KKT conditions:

(primal feasibility) $t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \geq 0$ for $n = 1,\ldots,N$

(dual feasibility) $a_n \geq 0$ for $n = 1,\ldots,N$

(complimentary slackness) $a_n(t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1) = 0$ for $n = 1,\ldots,N$

- 2. Dual Lagrangian obtained via (stationarity conditions) $\frac{\partial L}{\partial \mathbf{w}} = 0, \quad \frac{\partial L}{\partial b} = 0$

$$\tilde{L}(\mathbf{a}) = \min_{\mathbf{x},b} L(\mathbf{x}, b, \mathbf{a})$$

- 3. **Solution**: $\mathbf{a}^* = \arg\max_{\mathbf{a}} \tilde{L}(\mathbf{a})$ → $\mathbf{w}^*, b^* = \arg\min_{\mathbf{w},b} L(\mathbf{w}, b, \mathbf{a}^*)$

Maximum Margin Classifier

- Primal Lagrangian function:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$$

with Langrange multipliers: $a_n \geq 0$ for $n = 1, \ldots, N$

- First step towards dual Langrangian: obtain stationarity conditions

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w}^T - \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T = 0 \quad \rightarrow \quad \boxed{\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n}$$

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \boxed{\sum_{n=1}^{N} a_n t_n = 0}$$

- Eliminate $\mathbf{w}$ and b from L then gives the dual representation!

$\tilde{L}(\mathbf{a}) = \ldots$

Maximum Margin Classifier

- Stationarity conditions:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w}^T - \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n$$

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0$$

- Eliminate $\mathbf{w}$ and b from L then gives the dual representation!

- Primal: $L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\}$, with $a_n \geq 0$ for $n = 1, \ldots, N$

- Dual: $\tilde{L}(\mathbf{a}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T\sum_{n=1}^{N} a_n t_n \mathbf{x}_n - b\sum_{n=1}^{N} a_n t_n + \sum_{n=1}^{N} a_n$

$$= -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} a_n$$

$$= \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

with $a_n \geq 0$ for $n = 1, \ldots, N$

and with $\sum_{n=1}^{N} a_n t_n = 0$

Maximum Margin Classifier

now we get $\underline{a}^*$

- The dual representation of the maximum margin, where we maximize w.r.t. $\mathbf{a}$:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

with constraints: $a_n \geq 0$ for $n = 1, \ldots, N$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$\underline{\phi}_n^T \underline{\phi}_m$

kernel implicitly defines basis

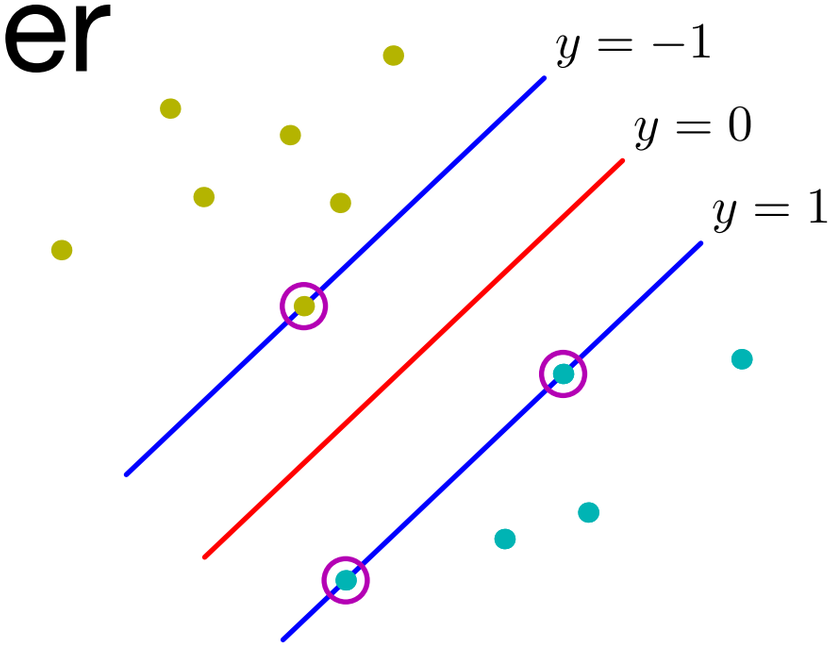
- Apply the **KERNEL TRICK**: replace $\mathbf{x}_n^T \mathbf{x}_m$ with $k(\mathbf{x}_n, \mathbf{x}_m)$

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

$= \exp\left(\frac{1}{2\sigma^2} \|x_n - x_m\|^2\right)$

- Advantage: can now learn complex nonlinear decision boundaries!

Maximum Margin Classifier



- Prediction of class for datapoint $\mathbf{x}_n$:

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

- Use $\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n$ so that

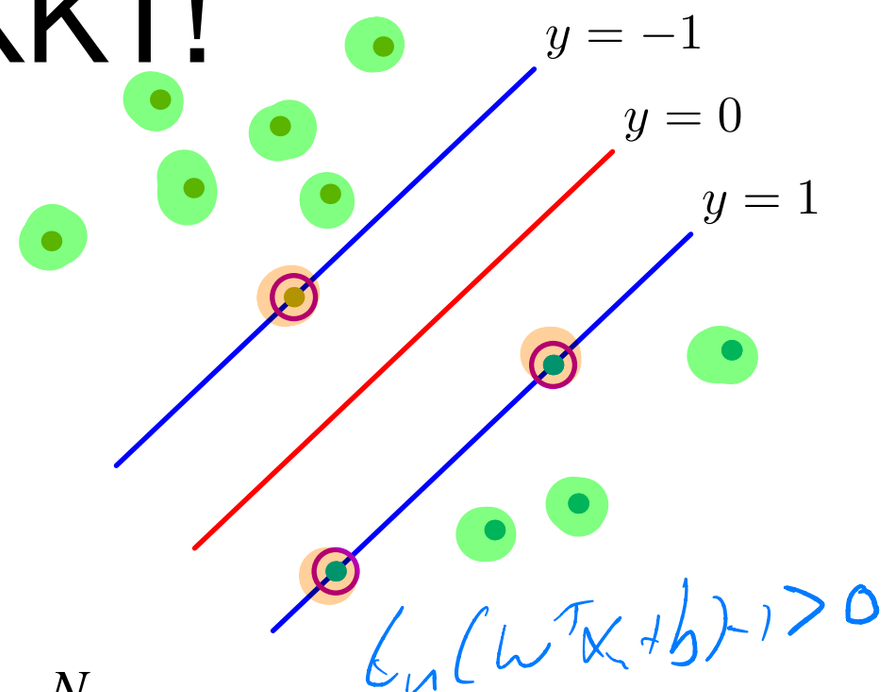
$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T \mathbf{x} + b \quad \rightarrow \quad y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$$

kernel trick!

still running over many points

to how many a_n do we know are zero? → KKT

Sparse solutions due to KKT!



$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$$

- Remember the KKT conditions:

(primal feasibility)	$t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \geq 0$	for	$n = 1,\ldots,N$
(dual feasibility)	$a_n \geq 0$	for	$n = 1,\ldots,N$
(complimentary slackness)	$a_n(t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1) = 0$	for	$n = 1,\ldots,N$

$> 0 \longrightarrow a_n = 0$

- Support vectors lie on maximum margin hyperplanes

$a_n > 0 \quad \rightarrow \quad t_n\, y(\mathbf{x}_n) = 1$ (support vectors)

$a_n = 0 \quad \leftarrow \quad t_n\, y(\mathbf{x}_n) > 1$ (all other points)

SVM: Solution for bias b

- Prediction of class for datapoint $\mathbf{x}$:

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T \mathbf{x} + b \quad \rightarrow \quad y(\mathbf{x}) = \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}) + b$$

- Find b by using that $t_n y_n(\mathbf{x}) = 1$ if $\mathbf{x}_n$ lies on the margin boundary!
  ($\mathbf{x}_n$ is a support vector)

- Then $t_n \left(\sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) + b \right) = 1$

multiply both sides with t_n. Note $t_n \cdot t_n = 1$

$$\sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) + b = t_n$$

$$\rightarrow \quad b = t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n)$$

- More stable to average over all support vectors (depending on optimizer, $\mathbf{a}_n$ may not be perfect)

$$\rightarrow \quad b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) \right)$$

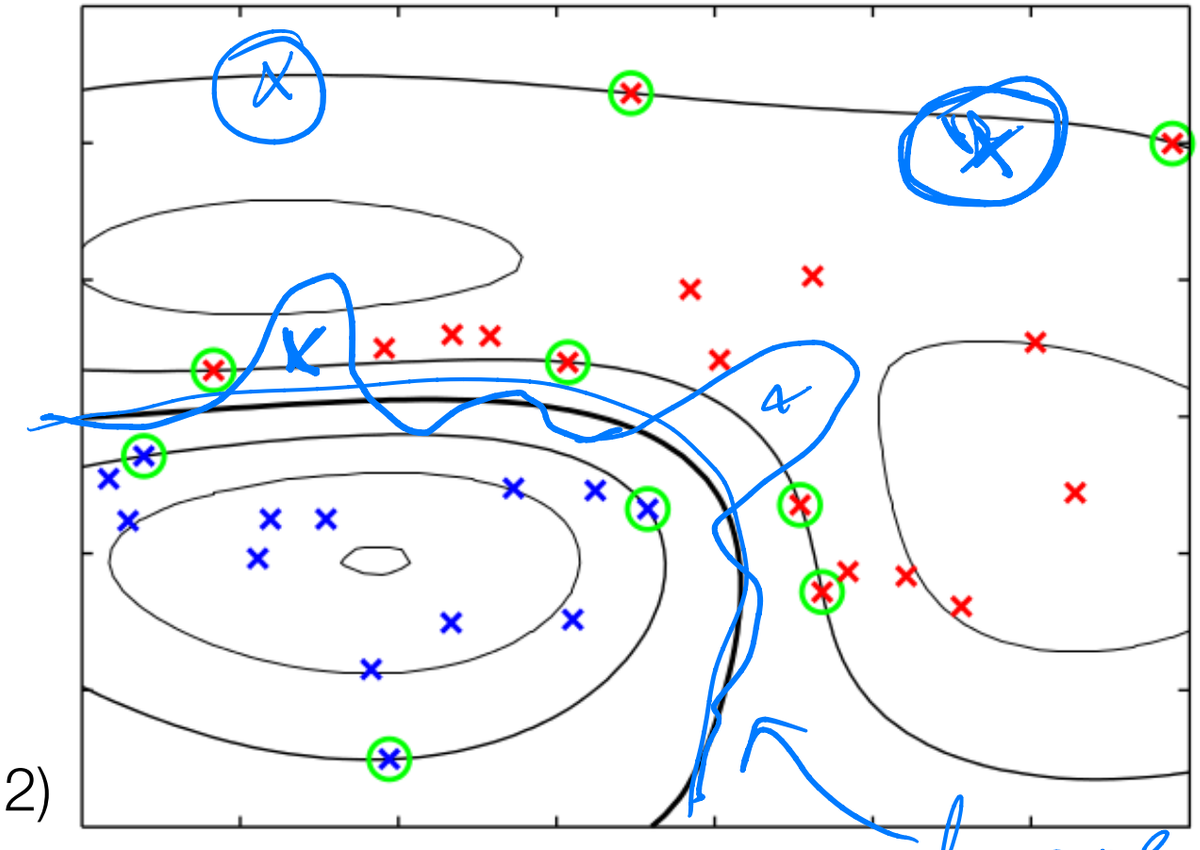
Maximum Margin Classifier

implicitly we have $\phi(x_n) \in \mathbb{R}^{\infty}$

- Maximum Margin Classifier with Gaussian Kernel

$$y(\mathbf{x}) = \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}) + b, \quad \text{with} \quad k(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right)$$

- Dataset is not linearly separable
- With Gaussian kernel one can separate the data perfectly!



(Bishop 7.2)

large σ means smooth boundaries.

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 11.6 - Kernel Methods
Support Vector Machines - Soft Margin Classifier

Erik Bekkers

(Bishop 7.1.1)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

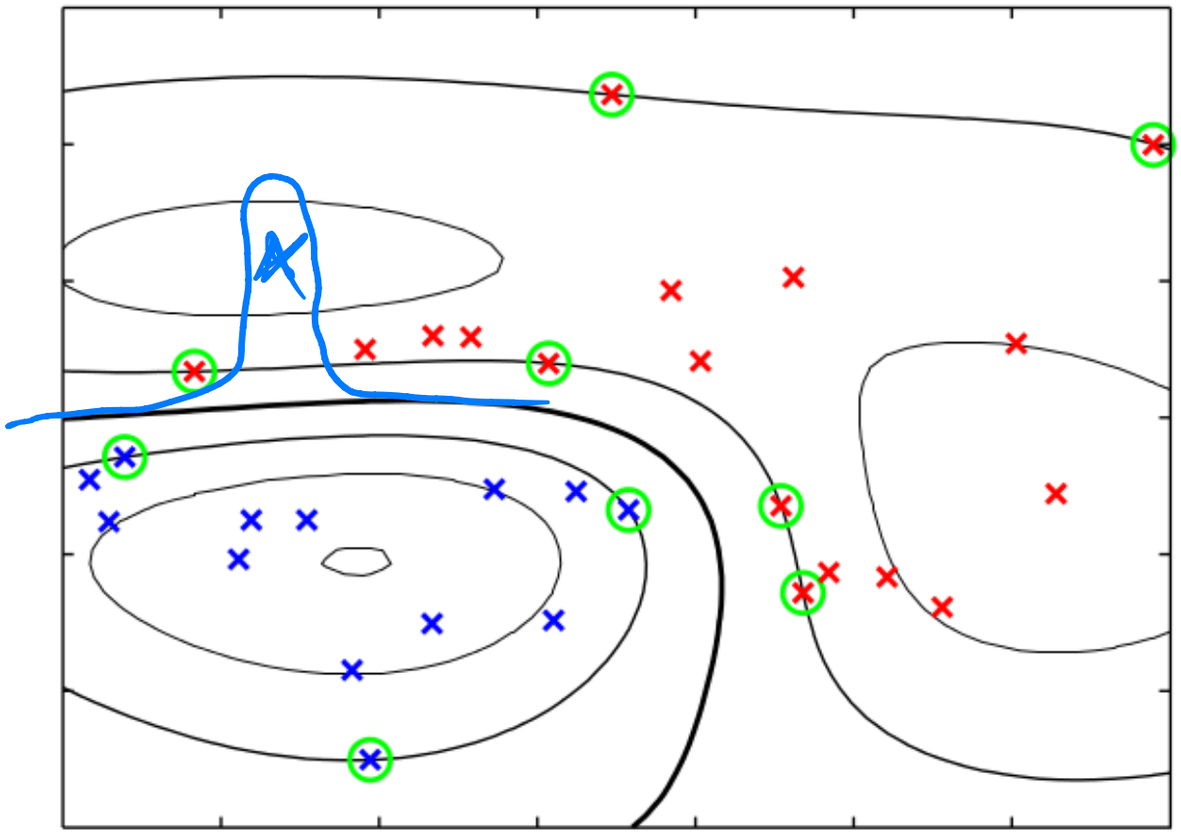
Maximum Margin Classifiers

- So far we have assumed the data points are perfectly separable with a linear decision boundary, or with a nonlinear decision boundary by using a nonlinear kernel.

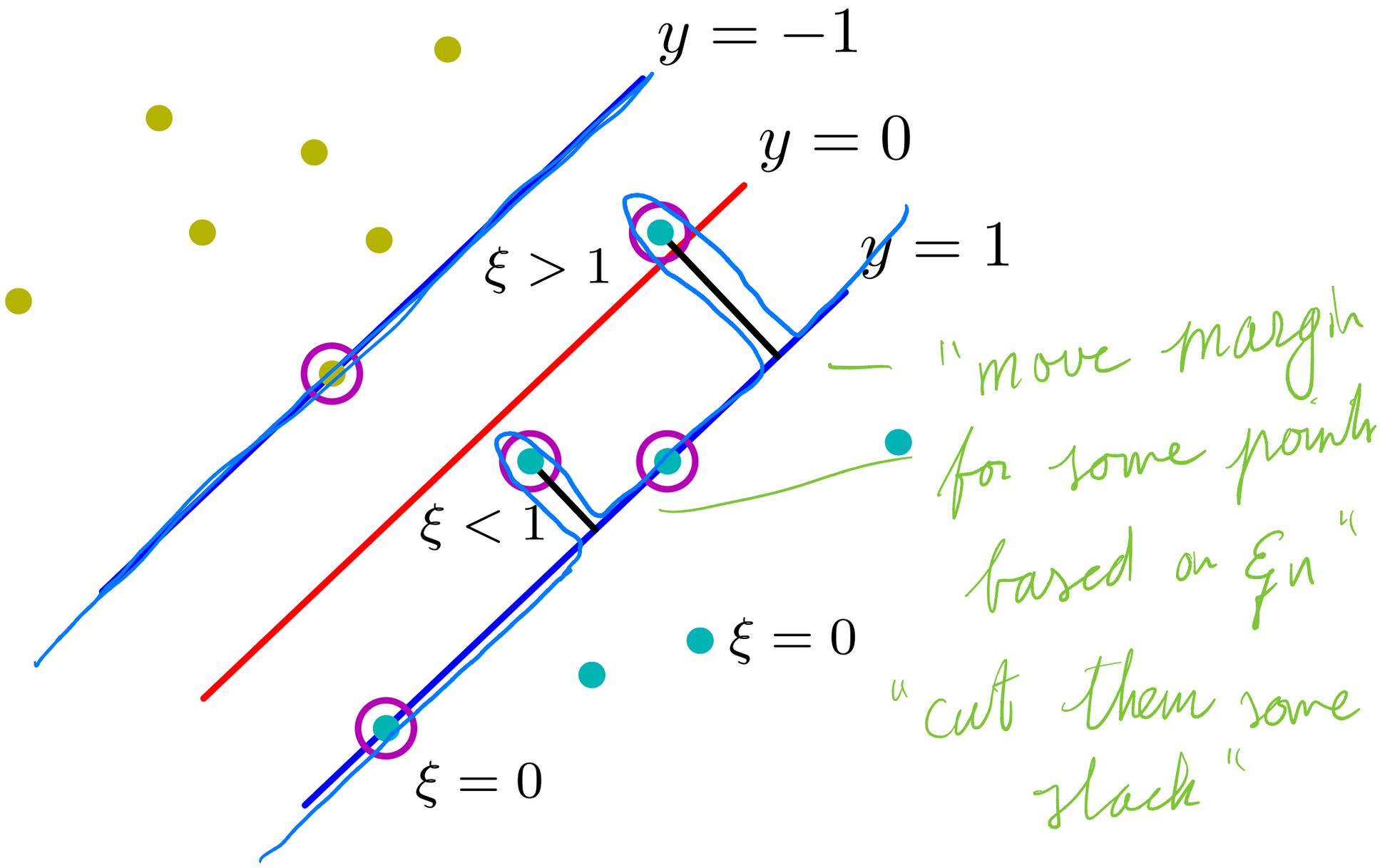
- Sometimes the class conditional distributions have overlap!

- We need to modify the Maximum Margin classifier to allow for some training points to be misclassified.

- Datapoints are allowed to be on the “wrong” side of the margin boundary, but they have to pay a penalty distance to the margin boundary.



Soft Margin Classifiers



- Allows datapoints to lie on the wrong side of the margin boundary.

- Those datapoints pay a penalty proportional to the distance to the margin boundary.

Maximum Margin Classifiers: Soft Margins!

- Introduce slack variables: $\xi_n \geq 0$ for $n = 1, \ldots, N$
- If on the correct side of the margin: $\xi_n = 0$
- If on the wrong side of the margin: $\xi_n = |t_n - y(\mathbf{x}_n)|$

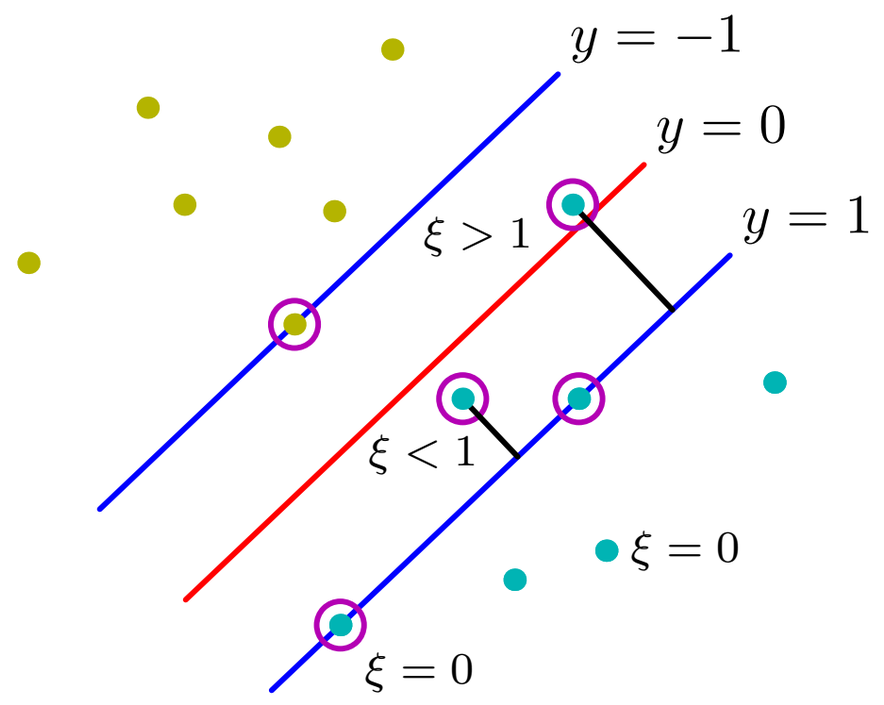
slack penalty proportional to distance to margin

- Previously: hard constraints/hard margin

$$t_n y(\mathbf{x}_n) \geq 1, \quad n = 1, \ldots, N$$

- Now: Soft constraint/soft margin

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n, \quad n = 1, \ldots, N$$



Maximum Margin Classifiers: Soft Margins!

- Goal: maximize margin, give a penalty to points that lie on the wrong side of the boundary!

- We minimize $\arg\min_{\mathbf{w},b,\xi_n} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$

subject to constraints $t_n y(\mathbf{x}_n) \geq 1 - \xi_n$, for $n = 1,\ldots,N$

$\xi_n \geq 0$, for $n = 1,\ldots,N$

- Corresponding Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

- Lagrange multipliers

$\forall n = 1,\ldots,N$: $a_n \geq 0, \quad \mu_n \geq 0$

1 lagrangian
2 KKT
3 solve for stationarity w.r.t primals
4 obtain dual $\tilde{L}(\underline{a}, \underline{\mu})$
5 Solve dual problem

Maximum Margin Classifiers: Soft Margins!

- Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

- Lagrange multipliers. $a_n \geq 0, \quad \mu_n \geq 0$

- KKT conditions:

dual feasibility

$$a_n \geq 0 \qquad \mu_n \geq 0$$

primal

$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0 \qquad \xi_n \geq 0$$

compl. slackness

$$a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} = 0 \qquad \mu_n\xi_n = 0$$

$n = 1, \ldots, N$

- How many KKT conditions?

$6N$

Maximum Margin Classifiers: Soft Margins!

- Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

3

- Minimize L w.r.t. primal variables $\mathbf{w}, b, \xi_n$ and use the KKT conditions to eliminate $\mathbf{w}, b, \xi_n$ from Lagrangian to obtain dual formulation!

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w}^T - \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n$$

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = C - a_n - \mu_n = 0 \quad \rightarrow \quad a_n = C - \mu_n$$

new constraint

4

- Use this to eliminate $\mathbf{w}, b, \xi_n$, dual Lagrangian:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

Maximum Margin Classifiers: Soft Margins!

- Minimization of primal variables gave these conditions:

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n \,, \qquad \sum_{n=1}^{N} a_n t_n = 0 \,, \qquad a_n = C - \mu_n$$

so $a_n \leq C$

- KKT conditions:

$$a_n \geq 0 \qquad \mu_n \geq 0$$
$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0 \qquad \xi_n \geq 0$$
$$a_n \{ t_n y(\mathbf{x}_n) - 1 + \xi_n \} = 0 \qquad \mu_n \xi_n = 0$$

- Dual lagrangian:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

- Remaining constraints:

Box constraints:

$$0 \leq a_n \leq C$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$a_n \leq C - \mu_n$

$\mu_n \geq 0$

Maximum Margin Classifiers: Soft Margins!

- Dual problem: Maximize w.r.t an

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

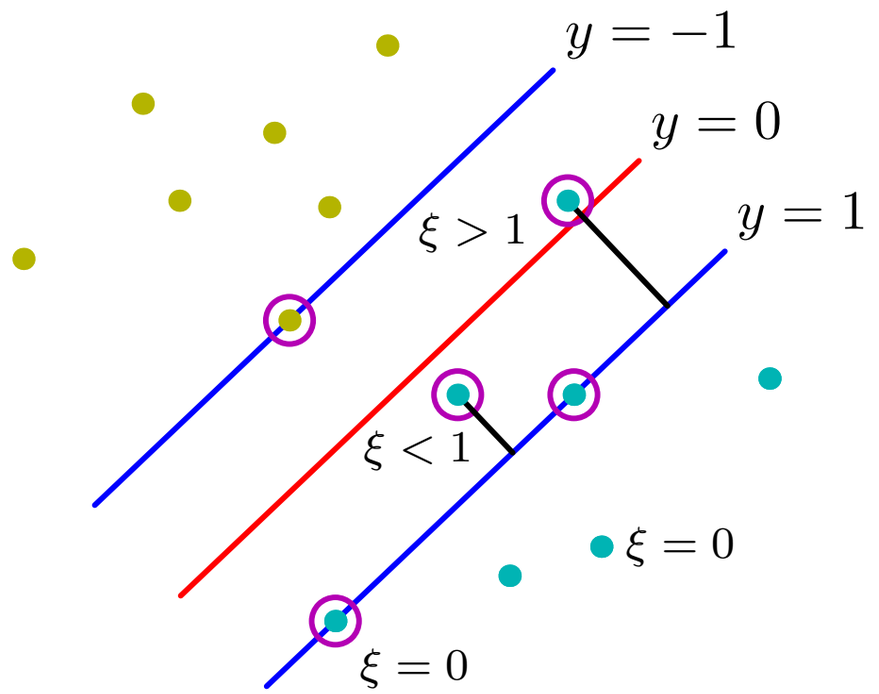
subject to $0 \leq a_n \leq C\,, \qquad \sum_{n=1}^{N} a_n t_n = 0$

- Kernel trick:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

- Prediction

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$$



Again KKT tells us what the option for a_n are

Maximum Margin Classifiers: Soft Margins!

- Prediction: $y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$ (with $0 \leq a_n \leq C$ and $\sum_{n=1}^{N} a_n t_n = 0$)

3 types of SV's

1 $a_n = 0$

2 $0 < a_n < C$ } $a_n > 0$

3 $a_n = C$

- Remember $a_n \geq 0$ $\mu_n \geq 0$

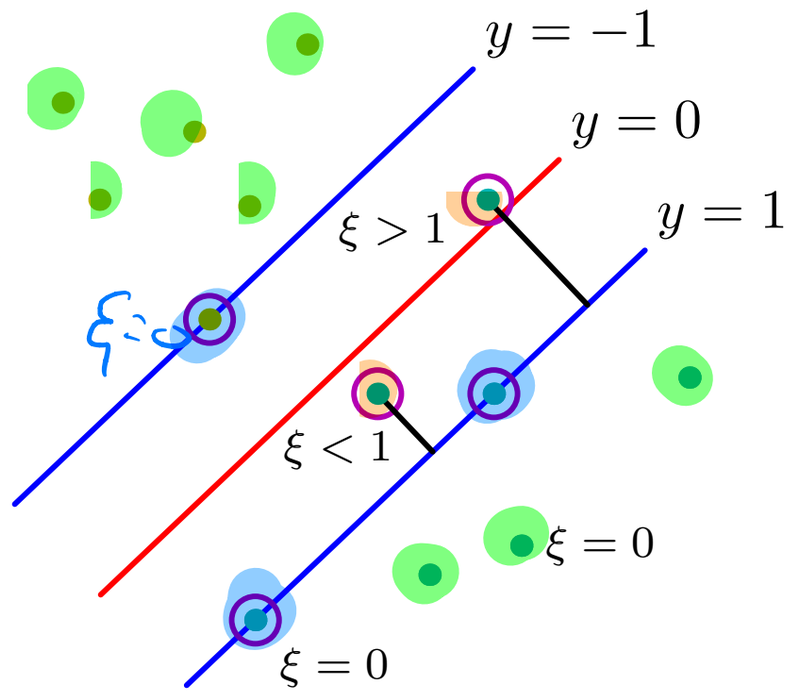
  $t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0$ $\xi_n \geq 0$

  $a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} = 0$ $\mu_n \xi_n = 0$

- **Support vectors** (If $a_n > 0$ then $t_n y(\mathbf{x}_n) = 1 - \xi_n$):

  $a_n = C - \mu_n$

  - If also $a_n < C$ then $\mu_n > 0$ so $\xi_n = 0$:
    - Points on margin
  - If $\xi_n \geq 0$ then $\mu_n = 0$ and thus $a_n = C$:
    - Correctly classified but within margin: $\xi_n \leq 1$
    - Misclassified $\xi_n > 1$



Maximum Margin Classifiers: Soft Margins!

- Goal: maximize margin, give penalty to points that lie on the wrong side of the boundary!

- We minimize $\arg\min_{\mathbf{w},b,\xi_n} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$

$$\text{subject to} \quad t_n y(\mathbf{x}_n) \geq 1 - \xi_n \,, \quad \text{for} \quad n = 1,\ldots,N$$
$$\xi_n \geq 0 \,, \quad \text{for} \quad n = 1,\ldots,N$$

- What happens in the limit: $C \to \infty$

– Hard margin (small margin)

- What happens when $C \to 0$

– possibly as margin

– every point become a support vector