

# Statistical Learning and Applications

Laurent Jacob

September 19, 2014

## Summary of the previous class

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

## Summary of the previous class

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

## Summary of the previous class

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

## Summary of the previous class

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

## Part III

# Supervised learning

- With these ideas in mind, we now turn to concrete examples of statistical learning methods.
- We start with *penalized empirical risk minimization* techniques, which explicitly implements the bias-variance tradeoff.
- We then move to other classical techniques.

- This part is about supervised learning, *i.e.*, inference using **labeled** data (class, real value...).
- If no **labeled** data is available but we want to estimate an assumed hidden structure, we need unsupervised learning (different techniques, next part).
- Although algorithms for unsupervised learning are often different, underlying models are often similar.
- More importantly, the previous bias-variance discussion also applies (we are still doing estimation from data).



# Outline for supervised learning

- 1  $\ell_2$  penalties.
- 2 Ridge regression.
- 3 Fundamentals of constrained optimization.
- 4 Support vector machines.
- 5  $\ell_1$  penalties.
- 6 Cross validation.
- 7 Local methods (nearest neighbors, smoothing).
- 8 Random forests.
- 9 Neural networks.
- 10 Kernel methods.

First five points are related to penalized empirical risk minimization.

# Penalized empirical risk minimization

- 1 **Define nested function sets of increasing complexity.**
- 2 Minimize the empirical risk over each family.
- 3 Choose the solution giving the best generalization performances.

Define a complexity measure  $\Omega$  for functions, and consider the classes

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots,$$

where  $\mathcal{H}_j = \{f, \Omega(f) \leq \mu_j\}$  and  $\mu_1 < \mu_2 < \dots$

## Reminder: structural risk minimization

Define a complexity measure  $\Omega$  for functions, and consider the classes

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots,$$

where  $\mathcal{H}_j = \{f, \Omega(f) \leq \mu_j\}$  and  $\mu_1 < \mu_2 < \dots$

Then (step 2) we can successively solve:

$$\min_{f \in \mathcal{H}_j} \sum_{i=1}^n L(y_i, f(x_i)),$$

i.e., minimize the empirical risk while restricting ourselves to sets of function of increasing complexity.

Note: this results in constrained optimization problems. Solving these problems for different loss functions and function spaces is an active research area.

## Remark: equivalence with a penalized estimator

- We mostly discuss penalized methods:

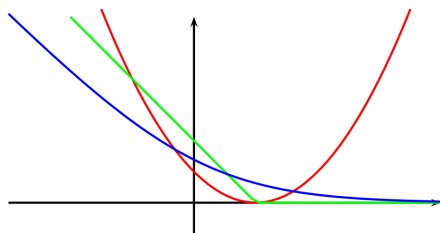
$$\min_f \sum_{i=1}^n L(y_i, f(x_i)) + \lambda \Omega(f)$$

- The first term favors a good fit to the data, the second one favors regularity of  $f$ .
- We will show later that the constrained and penalized forms are often equivalent in some sense (need to introduce some technical tools before that).
- The approach will stay the same: we define an  $\Omega$  which is relevant for our problem and we compare the generalization performances of the functions obtained for decreasing values of  $\lambda$ .

# Usual loss functions

**Regression** :  $y \in \mathbb{R}$

- $\ell_2$  :  $L(y_i, f(x_i)) = (y_i - f(x_i))^2$  (which we used in introduction),
- $\ell_1$  :  $L(y_i, f(x_i)) = |y_i - f(x_i)|$  (robust version, less sensitive to large errors, e.g. median vs mean).

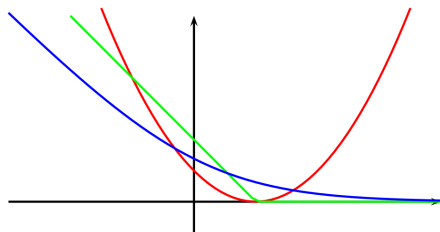


(from J. Mairal's slides)

# Usual loss functions

**Classification** :  $y \in \{0, 1\}$

- 0/1 :  $L(y_i, f(x_i)) = \mathbf{1}_{y_i f(x_i) \geq 0}$ ,
- logistic :  $L(y_i, f(x_i)) = \log(1 + e^{-y_i f(x_i)})$ ,
- hinge :  $L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$ .



(from J. Mairal's slides)

Other problems: ranking, multi-class, survival...

- 0/1 loss counts the number of missclassification, the other ones are convex approximations.
- Convex losses combined with convex penalties lead to convex objectives for which global optima can be found.
- Methods based on convex objectives are also simpler to analyze.
- However, this guarantees by no mean that the convex version of a method performs better than its non-convex counterpart in practice.



- We now present two usual penalties, and analyze their effect on the estimated function.
- We restrict ourselves to linear functions  $f(x) = \theta^\top x$ ,  $\theta \in \mathbb{R}^p$ .

- A very common penalty is the ridge :

$$\Omega(\theta) = \|\theta\|_2^2.$$

- Used in **ridge regression** combined with the  $\ell_2$  loss and **support vector machines** (SVM) combined with the hinge loss.
- Leads to functions with the following type of regularity: two points  $x, x'$  which are close in Euclidean norm have close evaluations by the function since by the Cauchy-Schwarz inequality,

$$|\theta^\top x - \theta^\top x'| \leq \|\theta\|_2 \|x - x'\|_2.$$

- This property can limit the overfit and improve generalization: it makes functions behave similarly over similar, potentially unobserved data.
- Of course if there is no good predictor with this kind of regularity, the risk can be high because of the approximation term.
- We now study more precisely the influence of the ridge penalty in terms of bias-variance tradeoff for the linear model:

$$y = \bar{\theta}^\top x + \varepsilon,$$

where  $\varepsilon$  is a random variable with mean zero and variance  $\sigma^2$ .

# The ridge regression

- We observe  $n$  realizations of the previous linear model, represented by an  $X \in \mathbb{R}^{n,p}$  matrix and a  $Y \in \mathbb{R}^n$  vector.
- Consider the estimator

$$\hat{\theta} = \arg \min_{\theta} (\|Y - X\theta\|^2 + \lambda\|\theta\|^2).$$

- We can show there exists a closed form for this estimator :

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T Y.$$

- **[Exercise]** Show that the bias  $\mathbf{E}[\hat{\theta} - \bar{\theta}]$  of  $\hat{\theta}$  is  $-\lambda(X^T X + \lambda I)^{-1} \bar{\theta}$ .

$$\begin{aligned}\mathbf{E}[\hat{\theta}] &= \mathbf{E}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y}] \\ &= \mathbf{E}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top (\mathbf{X} \bar{\theta} + \varepsilon)] \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \bar{\theta} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{E}[\varepsilon] \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \bar{\theta}\end{aligned}$$

$$\begin{aligned}\mathbf{E}[\hat{\theta} - \bar{\theta}] &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \bar{\theta} - \bar{\theta} \\ &= \left( (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} - \mathbf{I} \right) \bar{\theta} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X} - \lambda \mathbf{I}) \bar{\theta} \\ &= -\lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \bar{\theta}.\end{aligned}$$

We now look at the variance of  $\hat{\theta}$  :

$$\begin{aligned} \text{Var}[\hat{\theta}] &= \text{Var}[(X^T X + \lambda I)^{-1} X^T Y] \\ &= (X^T X + \lambda I)^{-1} X^T \text{Var}[Y] X (X^T X + \lambda I)^{-1} \\ &= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \end{aligned}$$

(reminder : for a deterministic matrix  $A$ ,  $\text{Var}[AX] = A\text{Var}[X]A^T$ ).

## Bias (1/3)

- The bias  $-\lambda(X^T X + \lambda I)^{-1} \bar{\theta}$  increases with  $\lambda$  and tends to  $-\bar{\theta}$ .
- Remark:  $\hat{\theta} \rightarrow 0$  when  $\lambda \rightarrow \infty$ , so the limit bias is the one incurred by estimating  $\bar{\theta}$  by 0.
- If  $\lambda = 0$  (unpenalized linear regression), the bias is zero.
- The amplitude of the bias also depends on the norm of  $\bar{\theta}$ : if the  $\bar{\theta}$  which generated the data has a small norm, the bias/approximation error incurred by restricting ourselves to small norm estimators is smaller.



## Bias (2/3)

- A little more precisely, the squared norm of the bias is ( $\lambda \neq 0$ ) :

$$\begin{aligned}\| -\lambda(X^T X + \lambda I)^{-1} \bar{\theta} \|^2 &= \| (\lambda^{-1} X^T X + I)^{-1} \bar{\theta} \|^2 \\ &= \| U \Sigma U^T \bar{\theta} \|^2 = \| \Sigma U^T \bar{\theta} \|^2,\end{aligned}$$

where  $U \Sigma U^T$  is the spectral decomposition of  $(\lambda^{-1} X^T X + I)^{-1}$ .

- The eigenvalues of  $(\lambda^{-1} X^T X + I)^{-1}$  are **[Exercise]** :

## Bias (2/3)

- A little more precisely, the squared norm of the bias is ( $\lambda \neq 0$ ) :

$$\begin{aligned}\| -\lambda(X^T X + \lambda I)^{-1} \bar{\theta} \|^2 &= \| (\lambda^{-1} X^T X + I)^{-1} \bar{\theta} \|^2 \\ &= \| U \Sigma U^T \bar{\theta} \|^2 = \| \Sigma U^T \bar{\theta} \|^2,\end{aligned}$$

where  $U \Sigma U^T$  is the spectral decomposition of  $(\lambda^{-1} X^T X + I)^{-1}$ .

- The eigenvalues of  $(\lambda^{-1} X^T X + I)^{-1}$  are **[Exercise]** :

$$\Sigma = \text{Diag} (\lambda^{-1} e_i^2 + 1)^{-1} = \text{Diag} \left( \frac{\lambda}{e_i^2 + \lambda} \right),$$

where the  $e_i$  are the eigenvalues of  $X$ .

## Bias (3/3)

$$\| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta} \|^2 = \left\| \text{Diag} \left( \frac{\lambda}{e_i^2 + \lambda} \right) U^\top \bar{\theta} \right\|^2,$$

- Provides the shape of the convergence towards maximum bias as  $\lambda$  increases.
- If  $n/p$  is small,  $X^\top X$  has small eigenvalues  $e_i^2$ , and the bias is larger (even more if  $\theta$  is aligned with eigenvectors corresponding to small eigenvalues).
- Statistical interpretation: the bias is larger if the vector to be estimated lies in a direction of low empirical variance of the  $X$ .

## Variance

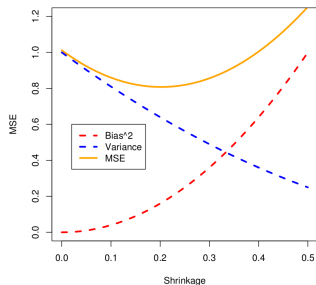
- Total variance

$$\begin{aligned}\text{tr Var}[\hat{\theta}] &= \text{tr} \left( \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \right) \\ &= \sigma^2 \text{tr} \left( (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-2} \mathbf{X}^\top \mathbf{X} \right) \\ &= \sigma^2 \sum_i \frac{e_i^2}{(e_i^2 + \lambda)^2}.\end{aligned}$$

tends to 0 as  $\lambda$  increases, and to the variance  $\sigma^2 \text{tr}(\mathbf{X}^\top \mathbf{X})^{-1}$  of unpenalized linear regression as  $\lambda \rightarrow 0$  (if  $\mathbf{X}^\top \mathbf{X}$  is invertible).

- Here again if  $n/p$  is small,  $\mathbf{X}^\top \mathbf{X}$  has small eigenvalues  $e_i^2$  and the variance for  $\lambda = 0$  increases.

# Ridge regression: bias and variance



(from J. Taylor's slides)

- Illustrates the phenomenon we discussed abstractly on a particular estimator.
- In practice for a dataset, some  $\lambda$  tradeoffs yield smaller risks than others.
- $\lambda$  can be chosen by hold out or cross validation (later in this class).

- We can also justify ridge regression from a numerical point of view: the  $\lambda I$  term decreases the condition number of the  $X^T X$  matrix, which can otherwise get very small eigenvalues.
- A poorly conditioned  $X^T X$  leads to results which are very sensitive to small variations in the data.

- Historically, this motivated the introduction of ridge regression by Hoerl et Kennard (1970):

*We were charging \$90/day for our time, but had to charge \$450/hour for computer time [...], we found that we had both encountered the same phenomenon, one that had caused some embarrassment with clients. We found that multiple linear regression coefficients computed using least squares didn't always make sense when put into the context of the process generating the data. The coefficients tended to be too large in absolute value, some would even have the wrong sign, and they could be unstable with very small changes in the data.*

- Tikhonov (1943) and Philips (1962) already introduced Hilbert norm penalties to improve the conditioning of integral equation solutions.

# Fundamentals of constrained optimization (from JP Vert)



## Setting

- We consider an equality and inequality constrained optimization problem over a variable  $x \in \mathcal{X}$ :

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && h_i(x) = 0, \quad i = 1, \dots, m, \\ & && g_j(x) \leq 0, \quad j = 1, \dots, r, \end{aligned}$$

making **no assumption** of  $f$ ,  $g$  and  $h$ .

- Let us denote by  $f^*$  the optimal value of the decision function under the constraints, i.e.,  $f^* = f(x^*)$  if the minimum is reached at a global minimum  $x^*$ .

## Lagrangian

The **Lagrangian** of this problem is the function  $L : \mathcal{X} \times \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}$  defined by:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \mu_j g_j(x).$$

## Lagrangian dual function

The **Lagrange dual function**  $q : \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}$  is:

$$\begin{aligned} q(\lambda, \mu) &= \inf_{x \in \mathcal{X}} L(x, \lambda, \mu) \\ &= \inf_{x \in \mathcal{X}} \left( f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \mu_j g_j(x) \right). \end{aligned}$$

- $q$  is **concave in**  $(\lambda, \mu)$ , even if the original problem is not convex.
- The dual function yields lower bounds on the optimal value  $f^*$  of the original problem when  $\mu$  is nonnegative:

$$q(\lambda, \mu) \leq f^* , \quad \forall \lambda \in \mathbb{R}^m, \forall \mu \in \mathbb{R}^r, \mu \geq 0 .$$

- For each  $x$ , the function  $(\lambda, \mu) \mapsto L(x, \lambda, \mu)$  is linear, and therefore both convex and concave in  $(\lambda, \mu)$ . The pointwise minimum of concave functions is concave, therefore  $q$  is **concave**.
- Let  $\bar{x}$  be any feasible point, i.e.,  $h(\bar{x}) = 0$  and  $g(\bar{x}) \leq 0$ . Then we have, for any  $\lambda$  and  $\mu \geq 0$ :

$$\sum_{i=1}^m \lambda_i h_i(\bar{x}) + \sum_{i=1}^r \mu_i g_i(\bar{x}) \leq 0 ,$$

$$\implies L(\bar{x}, \lambda, \mu) = f(\bar{x}) + \sum_{i=1}^m \lambda_i h_i(\bar{x}) + \sum_{i=1}^r \mu_i g_i(\bar{x}) \leq f(\bar{x}) ,$$

$$\implies q(\lambda, \mu) = \inf_x L(x, \lambda, \mu) \leq L(\bar{x}, \lambda, \mu) \leq f(\bar{x}) , \quad \forall \bar{x} . \quad \square$$

## Definition

For the (primal) problem:

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && h(x) = 0, \quad g(x) \leq 0, \end{aligned}$$

the **Lagrange dual problem** is:

$$\begin{aligned} & \text{maximize} && q(\lambda, \mu) \\ & \text{subject to} && \mu \geq 0, \end{aligned}$$

where  $q$  is the (concave) Lagrange dual function and  $\lambda$  and  $\mu$  are the Lagrange multipliers associated to the constraints  $h(x) = 0$  and  $g(x) \leq 0$ .

- Let  $d^*$  the optimal value of the Lagrange dual problem. Each  $q(\lambda, \mu)$  is a lower bound for  $f^*$  and by definition  $d^*$  is the best lower bound that is obtained. The following **weak duality inequality** therefore **always hold**:

$$d^* \leq f^* .$$

- This inequality holds when  $d^*$  or  $f^*$  are infinite. The difference  $d^* - f^*$  is called the **optimal duality gap** of the original problem.

- We say that **strong duality** holds if the optimal duality gap is zero, i.e.:

$$d^* = f^* .$$

- If strong duality holds, then the best lower bound that can be obtained from the Lagrange dual function is **tight**.
- Strong duality does **not hold** for general nonlinear problems.
- It usually holds for **convex problems**.
- Conditions that ensure strong duality for convex problems are called **constraint qualification**.

Strong duality holds for a **convex** problem:

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && g_j(x) \leq 0, \quad j = 1, \dots, r, \\ & && Ax = b, \end{aligned}$$

if it is **strictly feasible**, i.e., there exists at least one **feasible point** that satisfies:

$$g_j(x) < 0, \quad j = 1, \dots, r, \quad Ax = b.$$



Suppose that strong duality holds,  $x^*$  is primal optimal,  $(\lambda^*, \mu^*)$  is dual optimal. Then we have:

$$\begin{aligned} f(x^*) &= q(\lambda^*, \mu^*) \\ &= \inf_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{i=1}^m \lambda_i^* h_i(x) + \sum_{j=1}^r \mu_j^* g_j(x) \right\} \\ &\leq f(x^*) + \sum_{i=1}^m \lambda_i^* h_i(x^*) + \sum_{j=1}^r \mu_j^* g_j(x^*) \\ &\leq f(x^*) \end{aligned}$$

Hence both inequalities are in fact **equalities**.

The second equality shows that:

$$\mu_j g_j(x^*) = 0, \quad j = 1, \dots, r.$$

This property is called **complementary slackness**:  
**the  $i$ th optimal Lagrange multiplier is zero unless the  $i$ th constraint is active at the optimum.**

# Penalized vs constrained empirical risk minimization

# Equivalence with a penalized estimator

In some cases, the constrained problem

$$\min_{\Omega(f) \leq \mu} \sum_{i=1}^n L(y_i, f(x_i)), \quad (1)$$

is equivalent in some sense to the penalized problem

$$\min_f \sum_{i=1}^n L(y_i, f(x_i)) + \lambda \Omega(f). \quad (2)$$

Any solution of (1) is a solution of (2) for some  $\lambda$  depending of  $\mu$ , and vice-versa.

- The latter problem is sometimes easier to solve in practice.
- We will see later that estimators obtained by maximizing the **posterior likelihood** of some probabilistic models have this form.

**Example:**  $L$  and  $\Omega$  convex,  $f \in \mathbb{R}^p$ . We assume there exists  $f$  such that  $\Omega(f) < \mu$ . We note  $L(f) \triangleq \sum_{i=1}^n L(y_i, f(x_i))$  and

$$f_\mu \in \arg \min_{\Omega(f) \leq \mu} L(f),$$

$$f_\lambda \in \arg \min_f L(f) + \lambda \Omega(f).$$

## Remark: equivalence with a penalized estimator

We first show that  $f_\lambda$  is a solution of the constrained problem for some  $\mu$   
**[Exercise]:**

We first show that  $f_\lambda$  is a solution of the constrained problem for some  $\mu$   
**[Exercise]:**

- $f_\lambda$  verifies the constraint of the constrained problem for  $\mu = \Omega(f_\lambda)$ .
- If there exists  $f'$  such that  $L(f') < L(f_\lambda)$  and  $\Omega(f') \leq \mu = \Omega(f_\lambda)$ , then  $L(f') + \lambda\Omega(f') < L(f_\lambda) + \lambda\Omega(f_\lambda)$  which contradicts the optimality of  $f_\lambda$  for the penalized problem.

## Remark: equivalence with a penalized estimator

- This first part does not use convexity.
- We can therefore say in general that the regularization path of the penalized problem is included in the one of the constrained problem.



## Remark: equivalence with a penalized estimator

We now show that  $f_\mu$  is a solution of the penalized problem for some  $\lambda$ :

- Let  $\mathcal{L}(f, \lambda) \triangleq L(f) + \lambda(\Omega(f) - \mu)$  be the **Lagrangian** of the constrained problem (1).

## Remark: equivalence with a penalized estimator

We now show that  $f_{\mu}$  is a solution of the penalized problem for some  $\lambda$ :

- Let  $\mathcal{L}(f, \lambda) \triangleq L(f) + \lambda(\Omega(f) - \mu)$  be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is  $q(\lambda) \triangleq \min_f \mathcal{L}(f, \lambda)$ .

## Remark: equivalence with a penalized estimator

We now show that  $f_\mu$  is a solution of the penalized problem for some  $\lambda$ :

- Let  $\mathcal{L}(f, \lambda) \triangleq L(f) + \lambda(\Omega(f) - \mu)$  be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is  $q(\lambda) \triangleq \min_f \mathcal{L}(f, \lambda)$ .
- We note that here,

$$\min_f \mathcal{L}(f, \lambda) = \mathcal{L}(f_\lambda, \lambda). \quad (3)$$

## Remark: equivalence with a penalized estimator

We now show that  $f_\mu$  is a solution of the penalized problem for some  $\lambda$ :

- Let  $\mathcal{L}(f, \lambda) \triangleq L(f) + \lambda(\Omega(f) - \mu)$  be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is  $q(\lambda) \triangleq \min_f \mathcal{L}(f, \lambda)$ .
- We note that here,

$$\min_f \mathcal{L}(f, \lambda) = \mathcal{L}(f_\lambda, \lambda). \quad (3)$$

- The dual always provides a lower bound to the primal solution:  
 $\forall \lambda \geq 0, \min_{\Omega(f) \leq \mu} L(f) \geq q(\lambda) = \min_f \mathcal{L}(f, \lambda)$ .

## Remark: equivalence with a penalized estimator

We now show that  $f_\mu$  is a solution of the penalized problem for some  $\lambda$ :

- Let  $\mathcal{L}(f, \lambda) \triangleq L(f) + \lambda(\Omega(f) - \mu)$  be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is  $q(\lambda) \triangleq \min_f \mathcal{L}(f, \lambda)$ .
- We note that here,

$$\min_f \mathcal{L}(f, \lambda) = \mathcal{L}(f_\lambda, \lambda). \quad (3)$$

- The dual always provides a lower bound to the primal solution:  
 $\forall \lambda \geq 0, \min_{\Omega(f) \leq \mu} L(f) \geq q(\lambda) = \min_f \mathcal{L}(f, \lambda)$ .
- Here by **strong duality** (obtained through Slater's conditions: convex problem and strictly feasible primal), we have

$$\min_{\Omega(f) \leq \mu} L(f) \stackrel{\text{s.d.}}{=} \max_{\lambda \geq 0} \min_f \mathcal{L}(f, \lambda) \stackrel{(3)}{=} \max_{\lambda \geq 0} (L(f_\lambda) + \lambda(\Omega(f_\lambda) - \mu))$$

## Remark: equivalence with a penalized estimator

$$\min_{\Omega(f) \leq \mu} L(f) = \max_{\lambda \geq 0} (L(f_\lambda) + \lambda(\Omega(f_\lambda) - \mu))$$

- In addition, by Slater's conditions again, there exists  $\lambda^*$  such that

$$L(f_{\lambda^*}) + \lambda^*(\Omega(f_{\lambda^*}) - \mu) = \min_{\Omega(f) \leq \mu} L(f) = L(f_\mu).$$

- By *complementary slackness*, it is necessary that  $\lambda^*(\Omega(f_{\lambda^*}) - \mu) = 0$ , which implies that  $L(f_\mu) = L(f_{\lambda^*})$  and:

- either  $\lambda^* = 0$  and  $L(f_\mu) + 0\Omega(f_\mu) = L(f_{\lambda^*}) + 0\Omega(f_{\lambda^*})$ ,
- or  $\Omega(f_{\lambda^*}) = \mu$  and

$$L(f_\mu) + \lambda^*\Omega(f_\mu) = L(f_{\lambda^*}) + \lambda^* \underbrace{\Omega(f_\mu)}_{\leq \mu = \Omega(f_{\lambda^*})} \leq L(f_{\lambda^*}) + \lambda^*\Omega(f_{\lambda^*}).$$

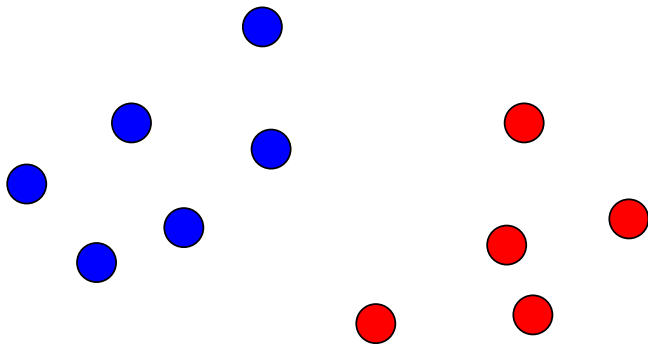
- In both cases,  $f_\mu$  is a solution of the problem penalized by  $\lambda^*$ .

# Support vector machines

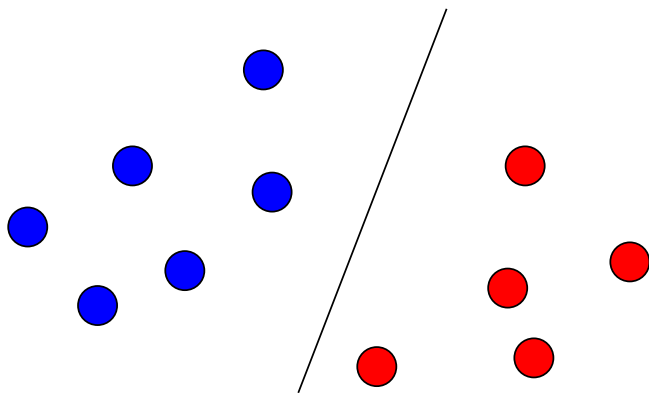
- We now present Support Vector Machines, a classical statistical learning algorithm.
- Fits into the penalized/constrained empirical risk minimization framework.
- We choose the historical *large margin* presentation.



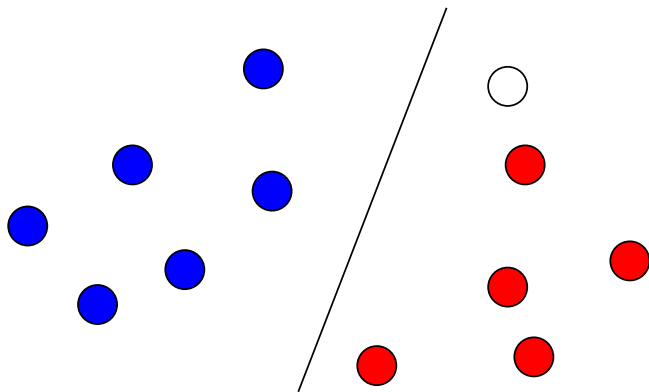
# Linear classifier



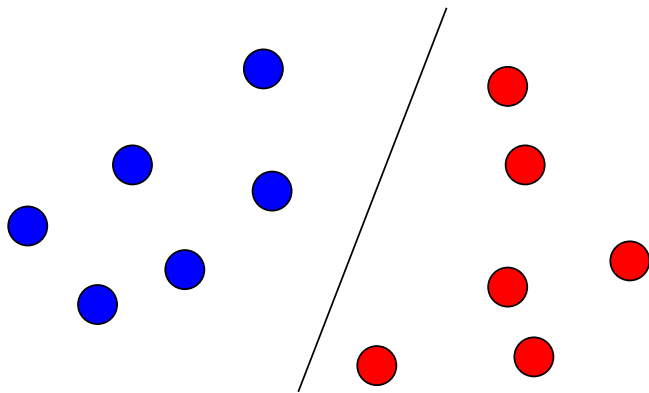
# Linear classifier



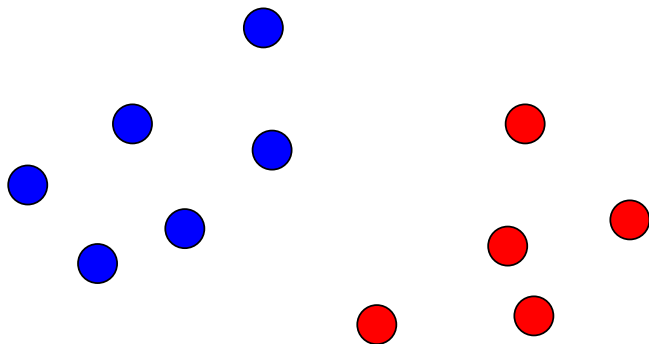
# Linear classifier



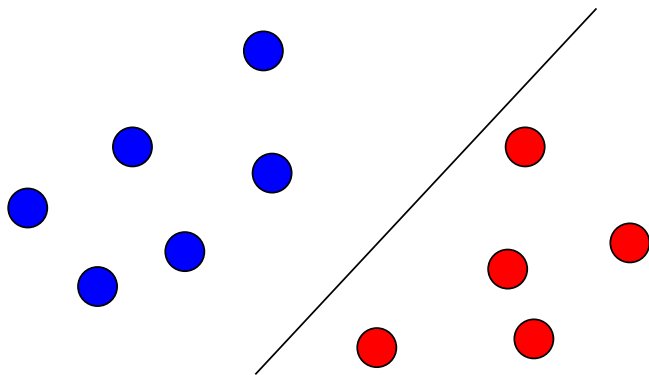
# Linear classifier



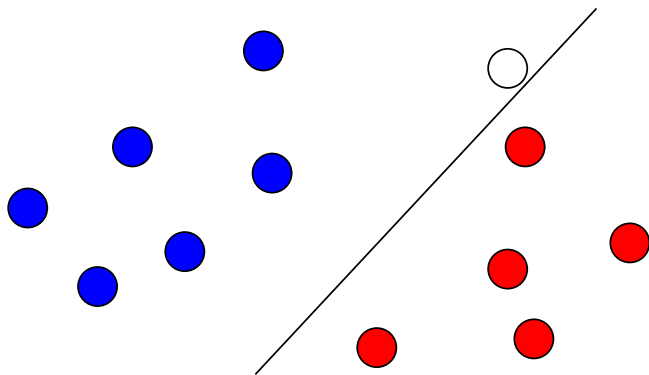
# Linear classifier



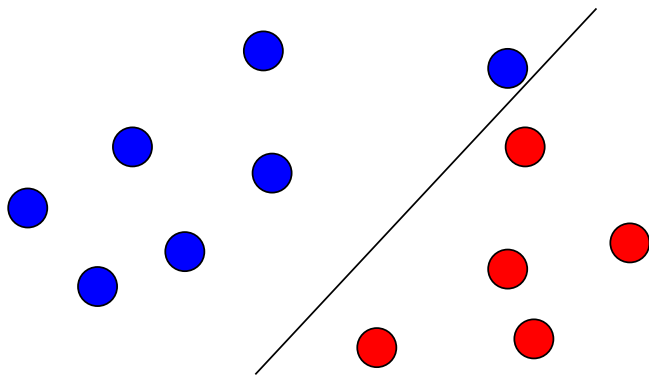
# Linear classifier



# Linear classifier

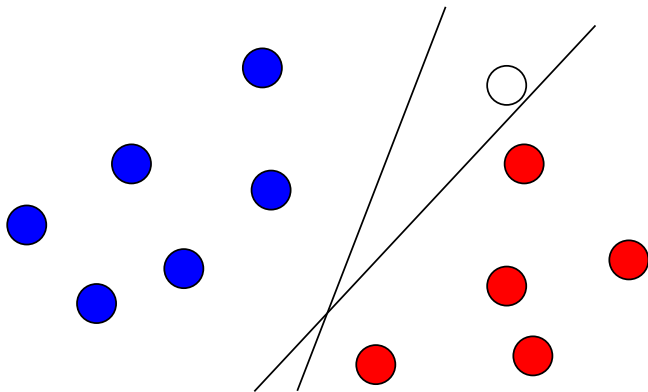


# Linear classifier

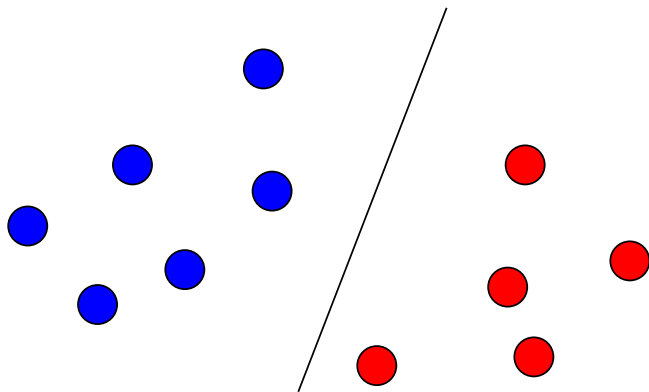




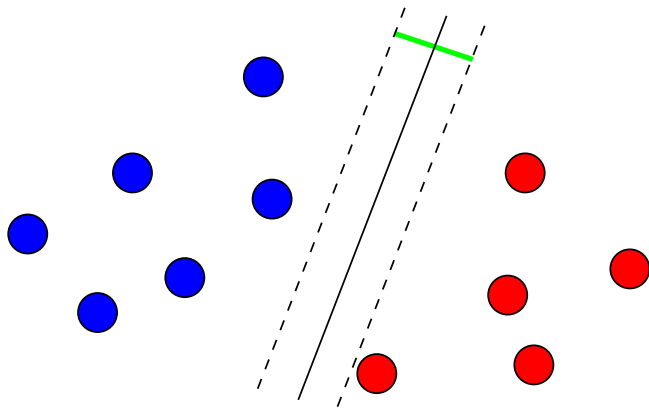
Which one is better?



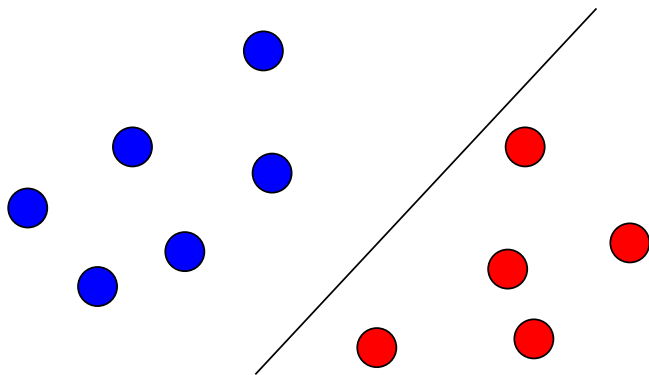
# The margin of a linear classifier



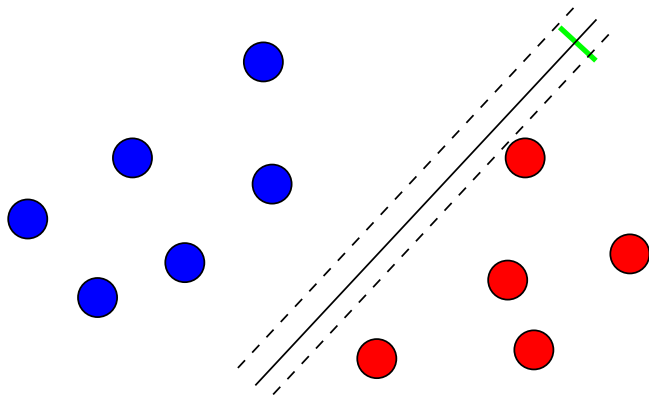
# The margin of a linear classifier



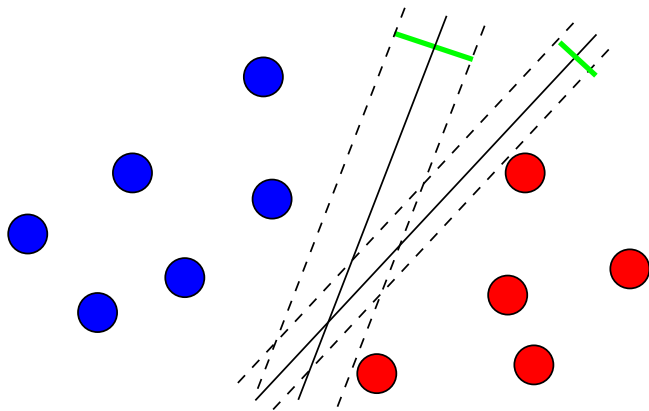
# The margin of a linear classifier



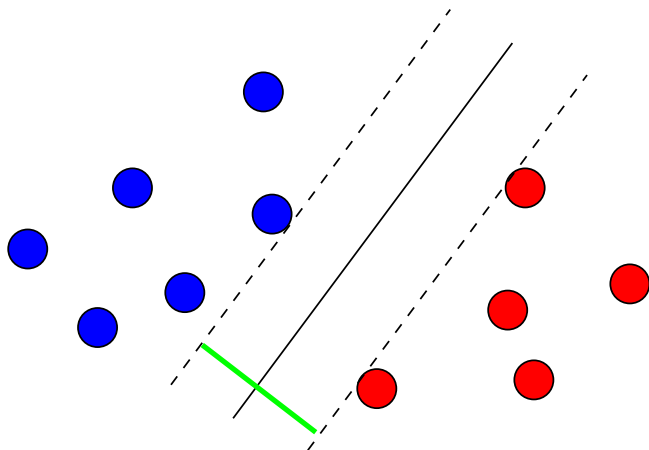
# The margin of a linear classifier



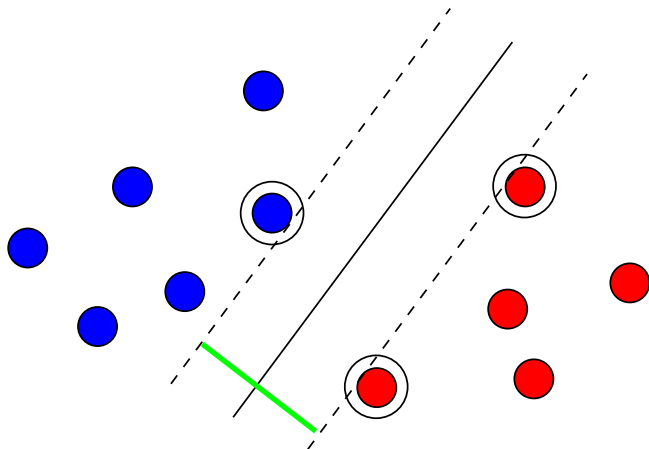
# The margin of a linear classifier



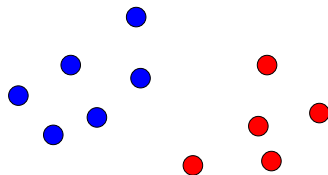
# Largest margin classifier (*hard-margin SVM*)



# Support vectors







- The **training set** is a finite set of  $n$  data/class pairs:

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\},$$

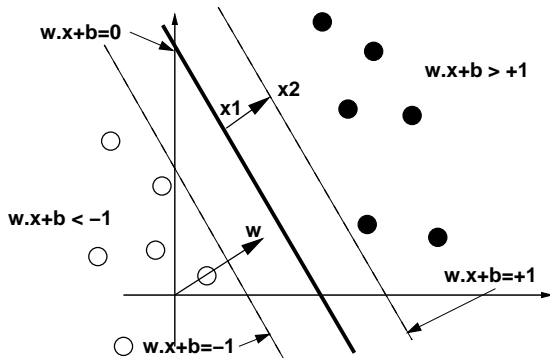
where  $x_i \in \mathbb{R}^p$  and  $y_i \in \{-1, 1\}$ .

- We assume (for the moment) that the data are **linearly separable**, i.e., that there exists  $(w, b) \in \mathbb{R}^p \times \mathbb{R}$  such that:

$$\begin{cases} w \cdot x_i + b > 0 & \text{if } y_i = 1, \\ w \cdot x_i + b < 0 & \text{if } y_i = -1. \end{cases}$$

# How to find the largest separating hyperplane?

For a given linear classifier  $f(x) = w \cdot x + b$  consider the "tube" defined by the values  $-1$  and  $+1$  of the decision function:



The margin is  $2/\|w\|$

Indeed, the points  $x_1$  and  $x_2$  satisfy:

$$\begin{cases} w \cdot x_1 + b = 0, \\ w \cdot x_2 + b = 1. \end{cases}$$

By subtracting we get  $w \cdot (x_2 - x_1) = 1$ , and therefore:

$$\gamma = 2\|x_2 - x_1\| = \frac{2}{\|w\|}.$$

All training points should be on the correct side of the dotted line

For positive examples ( $y_i = 1$ ) this means:

$$w \cdot x_i + b \geq 1.$$

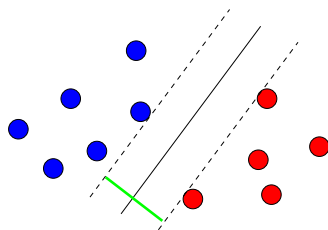
For negative examples ( $y_i = -1$ ) this means:

$$w \cdot x_i + b \leq -1.$$

Both cases are summarized by:

$$\forall i = 1, \dots, n, \quad y_i (w \cdot x_i + b) \geq 1.$$

# Finding the optimal hyperplane



Find  $(w, b)$  which minimize:

$$\|w\|^2$$

under the constraints:

$$\forall i = 1, \dots, n, \quad y_i (w \cdot x_i + b) - 1 \geq 0.$$

*This is a classical quadratic program on  $\mathbb{R}^{p+1}$ .*

In order to minimize:

$$\frac{1}{2} \|w\|_2^2$$

under the constraints:

$$\forall i = 1, \dots, n, \quad y_i (w \cdot x_i + b) - 1 \geq 0,$$

we introduce **one dual variable**  $\alpha_i$  **for each constraint**, i.e., for each **training point**. The Lagrangian is:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i (w \cdot x_i + b) - 1).$$

- $L(w, b, \alpha)$  is convex quadratic in  $w$ . It is minimized for:

$$\nabla_w L = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad \implies \quad \mathbf{w} = \sum_{i=1}^n \alpha_i \mathbf{y}_i \mathbf{x}_i.$$

- $L(w, b, \alpha)$  is affine in  $b$ . Its minimum is  $-\infty$  except if:

$$\nabla_b L = \sum_{i=1}^n \alpha_i \mathbf{y}_i = \mathbf{0}.$$

- We therefore obtain the **Lagrange dual function**:

$$q(\alpha) = \inf_{w \in \mathbb{R}^p, b \in \mathbb{R}} L(w, b, \alpha)$$
$$= \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j x_i \cdot x_j & \text{if } \sum_{i=1}^n \alpha_i y_i = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

- The dual problem is:

$$\begin{aligned} & \text{maximize} && q(\alpha) \\ & \text{subject to} && \alpha \geq 0. \end{aligned}$$



Find  $\alpha^* \in \mathbb{R}^n$  which maximizes

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i \cdot x_j,$$

under the (simple) constraints  $\alpha_i \geq 0$  (for  $i = 1, \dots, n$ ), and

$$\sum_{i=1}^n \alpha_i y_i = 0.$$

- This is a quadratic program on  $\mathbb{R}^n$ , with "box constraints".  $\alpha^*$  can be found efficiently using dedicated optimization softwares.
- This dual shows how SVM are an instance of **kernel methods** (later in this class).

# Recovering the optimal hyperplane

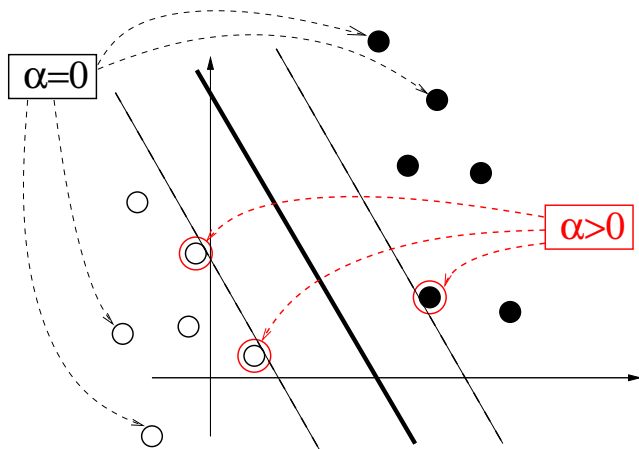
Once  $\alpha^*$  is found, we recover  $(w^*, b^*)$  corresponding to the optimal hyperplane.  $w^*$  is given by:

$$w^* = \sum_{i=1}^n \alpha_i x_i,$$

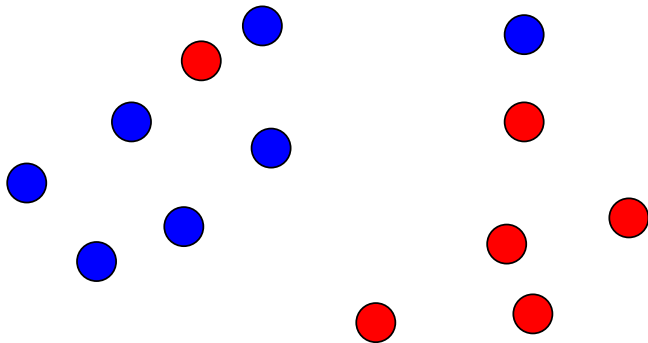
and the **decision function** is therefore:

$$\begin{aligned} f^*(x) &= w^* \cdot x + b^* \\ &= \sum_{i=1}^n \alpha_i x_i \cdot x + b^* . \end{aligned} \tag{4}$$

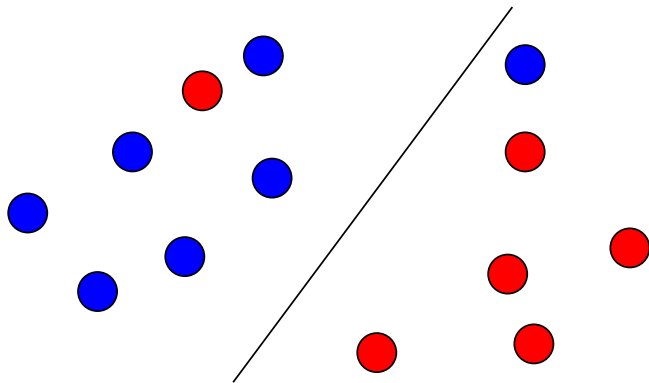
# Interpretation: support vectors



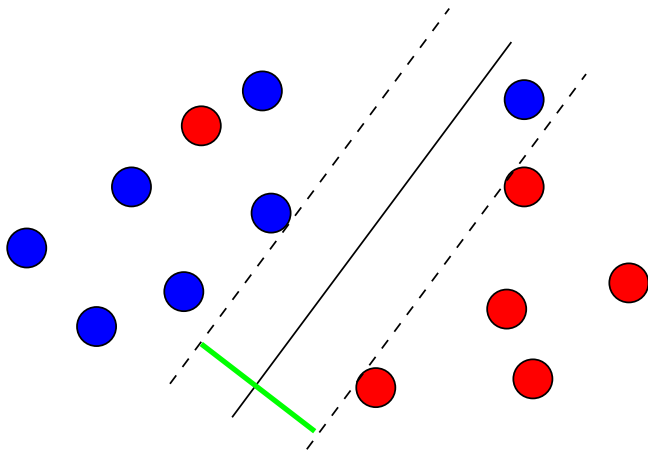
# What if data are not linearly separable?



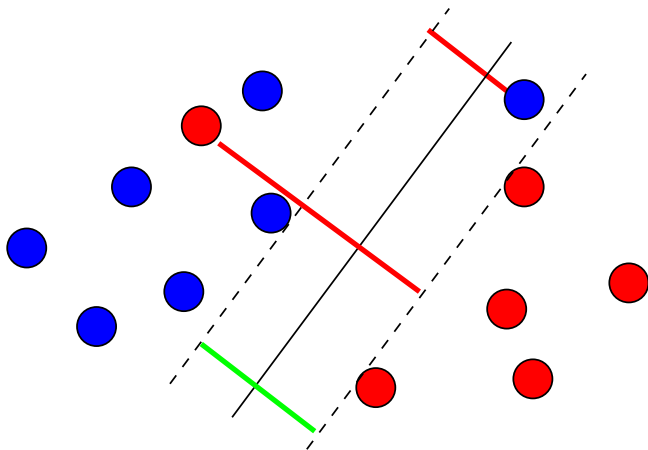
# What if data are not linearly separable?



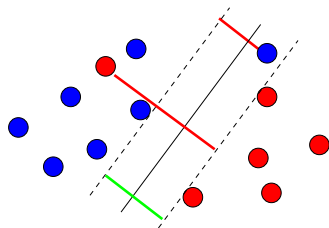
# What if data are not linearly separable?



# What if data are not linearly separable?



# Relaxing the separation constraints

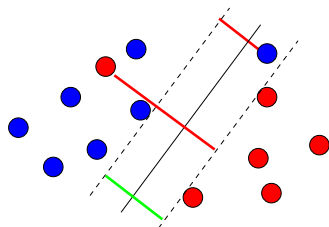


The problem is not feasible anymore. We need to **relax the separation constraints**:

$$\forall i = 1, \dots, n, \quad y_i (w \cdot x_i + b) \geq 1 - \xi_i.$$

The  $\xi_i$  are called **slack variables**.





- Allowing a larger slack makes a larger margin possible.
- One way to control the trade-off is to integrate the slack variables as a cost in the objective function:

$$\begin{cases} \min_{w,b,\xi} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \forall i = 1, \dots, n, & y_i (w \cdot x_i + b) \geq 1 - \xi_i \\ \forall i = 1, \dots, n, & \xi_i \geq 0 \end{cases}$$

- $C \in \mathbb{R}^+$  controls the trade-off.

## Dual formulation of soft-margin SVM (*exercice*)

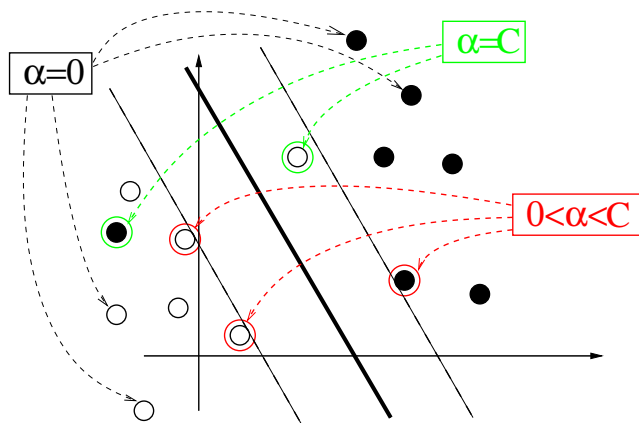
Maximize

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i \cdot x_j,$$

under the constraints:

$$\begin{cases} 0 \leq \alpha_i \leq \mathbf{C}, & \text{for } i = 1, \dots, n \\ \sum_{i=1}^n \alpha_i y_i = 0. \end{cases}$$

# Interpretation: bounded and unbounded support vectors



- Relaxed problem

$$\begin{cases} \min_{w, \xi} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \forall i = 1, \dots, n, & y_i (w \cdot x_i + b) \geq 1 - \xi_i \\ \forall i = 1, \dots, n, & \xi_i \geq 0 \end{cases}$$

is equivalent to

$$\min_w \|w\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i (w \cdot x_i + b))$$

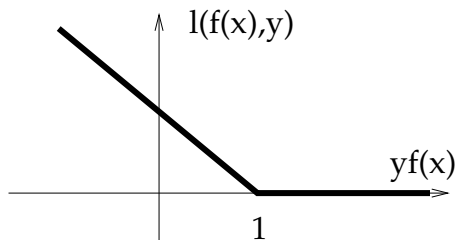
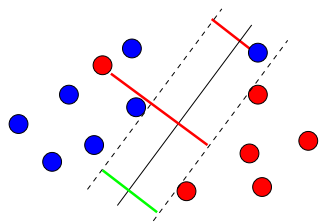
- Note: this is a useful trick to turn piecewise linear objective into linear objective with linear constraints.

# Soft-margin SVM and hinge loss

$$\min_{w,b} \left\{ \sum_{i=1}^n \ell_{\text{hinge}}(w \cdot x_i + b, y_i) + \lambda \|w\|_2^2 \right\},$$

for  $\lambda = 1/C$  and the hinge loss function:

$$\ell_{\text{hinge}}(u, y) = \max(1 - yu, 0) = \begin{cases} 0 & \text{if } yu \geq 1, \\ 1 - yu & \text{otherwise.} \end{cases}$$



- We started from a different perspective (maximize margin) and showed retrospectively that the problem we solved could be thought of as penalized empirical risk minimization.
- Yields another interpretation for  $\ell_2$  regularization of linear functions.
- In practice controlling this trade-off makes sense even if the classes are linearly separable (as discussed during the first class).