Stochastic gradient

Sylvain Rubenthaler

Function minimisation

We want to find th minimum of

$$Q(w) = \sum_{i=1}^{n} Q_i(w).$$

The simplest solution is a gradient descent :

- start anywhere (w_0)
- when in w_k, compute

$$w_{k+1} = w_k - \eta_k \nabla Q(w) \tag{1}$$

(well chosen decreasing (η_k)).

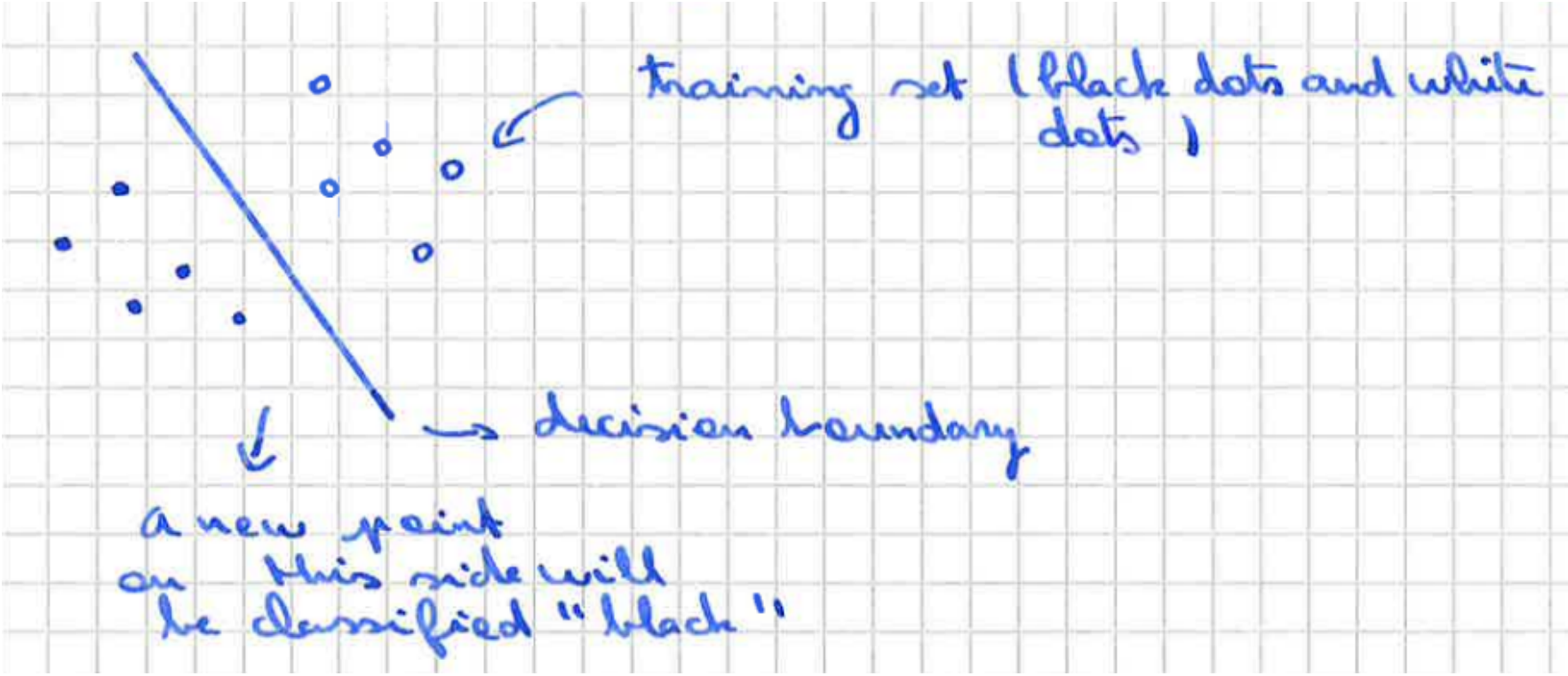
Stochastic gradient

If n is big, the computation of $\nabla Q(w)$ can be too heavy. Stochastic gradient :

- start anywhere in w_0 and choose (η_k) (>0).
- repeat (until you feel it is OK)
  - draw i uniformly in $\{1,2,\ldots,n\}$
  - $w_{k+1} \leftarrow w_k - \eta_k \nabla Q_i(w_k)$

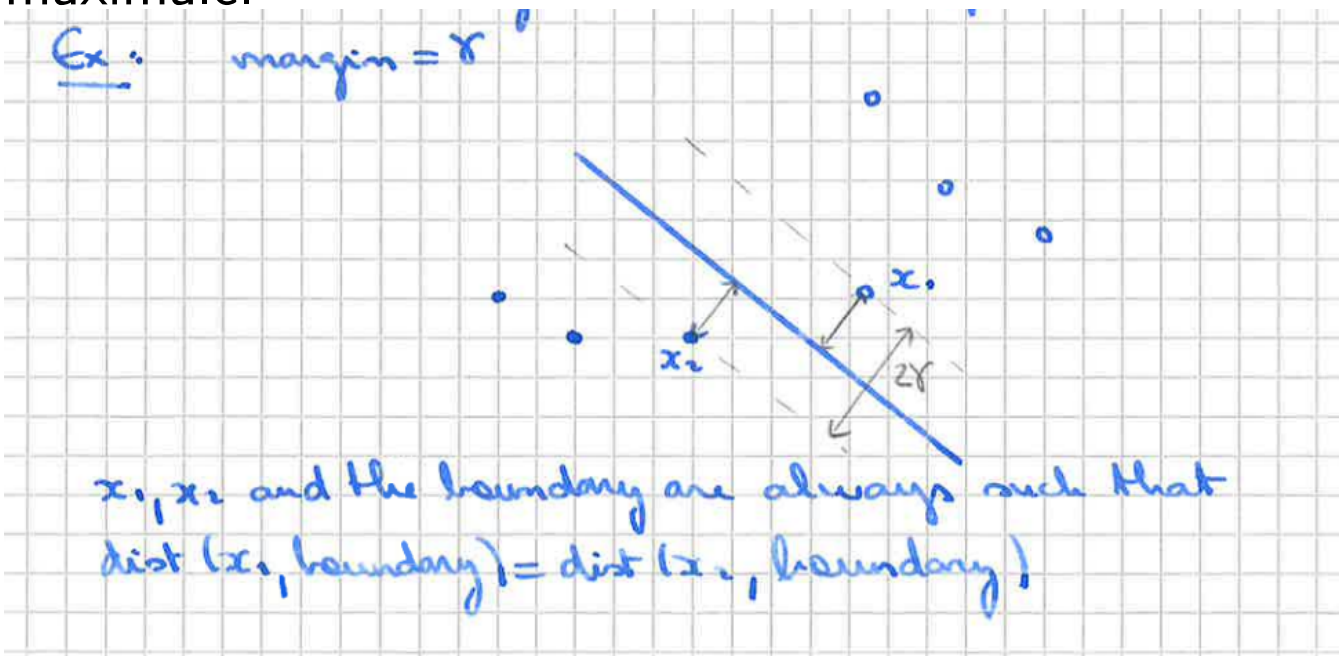
Exemple : SVM

Support Vector Machine



Exemple : SVM

On cherche l'hyperplan séparant les données et avec une marge maximale.



Les petits vecteurs orthogonaux à l'hyperplan et partant de x_1, x_2 sont les *vecteurs de support* (ils « supportent » l'hyperplan). C'est ce qui donne son nom à l'algorithme.

Exemple : SVM

Nous avons un ensemble d'apprentissage
$\{(x_n, t_n)_{1\leq n\leq N}, x_n \in \mathbb{R}^d, t_n \in \{+1; -1\}\}$. (Les t_n sont les étiquettes.)
On fixe : $w^T x + b = \pm 1$ pour les points support et alors (calcul),
$\gamma = 1/\|w\|$.
Nous cherchons (w et b, paramètres de l'hyperplan)

$$\begin{cases} \widehat{w} = \text{argmin}_w \, w^T w \\ \text{sous la contrainte } : t_n(w^T x_n + b) \geq 1, \forall n. \end{cases}$$

On relâche un peu la contrainte et on cherche le (w, b) minimisant :

$$J(w) = \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(0, 1 - t_n(b + w^T x_n)).$$

Exemple : SVM

- Le C est un paramètre de pénalisation à bien choisir (par cross-validation, par exemple).
- Le w est le vecteur orthogonal à l'hyperplan.
- La fonction max peut être remplacée par quelque chose de différentiable.

On voit l'intérêt de la descente de gradient stochastique.