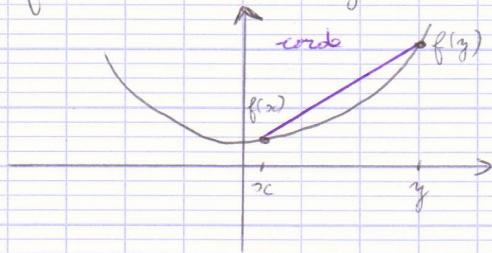


Gradient stochastique, dualité

I / Rappels :

- soit $X = \mathbb{R}^d$ (dimension de l'espace entrée)
soit $f : X \rightarrow \mathbb{R}$
- définition: f est convexe si $\forall x, y$ et $\forall \alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$



f est convexe si elle est dominée par ses cordes

- définition: f est strictement convexe si :

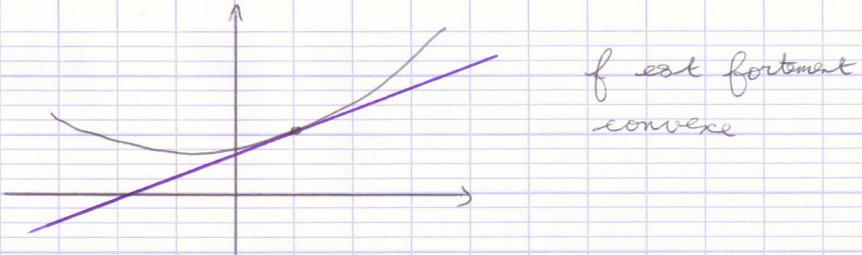
$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)y \quad \forall \alpha \in]0, 1[, \forall x \neq y$$
- propriété: toute fonction f strictement convexe admet au plus un minimiseur
 preuve: supposons que x^* et y^* soient deux minimiseurs et que $x^* \neq y^*$. Par convexité:

$$f\left(\frac{1}{2}x^* + \frac{1}{2}y^*\right) \leq \frac{1}{2}f(x^*) + \frac{1}{2}f(y^*)$$

$$\Leftrightarrow f(z) \leq \frac{1}{2}\min f + \frac{1}{2}\min f = \min f$$
 On obtient une contradiction car cela implique que z entraîne une valeur $f(z)$ inférieure strictement au minimiseur de f .

- définition: f est fortement convexe si $\exists \mu > 0$

$$f(x) - \frac{\mu}{2} \|x\|^2$$
 est convexe



- f fortement convexe $\Rightarrow f$ strictement convexe
En plus, f admet un minimiseur (par stricte convexité)
- propriété: si f est fortement convexe, alors elle admet un unique minimiseur.

II / Algorithme du gradient

- problème = $\min_{x \in X} f(x)$

hypothèse: f est dérivable

- algorithme: on itère:

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

idée: $f(x_k + \delta) \approx f(x_k) + \langle \nabla f(x_k), \delta \rangle$ pour δ petit

si on définit $\delta = -\gamma \nabla f(x_k)$, on obtient:

$$\begin{aligned} f(x_k + \delta) &\approx f(x_k) - \gamma \langle \nabla f(x_k), \nabla f(x_k) \rangle \\ &= f(x_k) - \gamma \|\nabla f(x_k)\|^2 \end{aligned}$$

et on se rend compte que la fonction décroît à chaque itération

- preuve de convergence dans un cas simple:

hypothèses: $f: \mathbb{R} \rightarrow \mathbb{R}$ est deux fois dérivable

$$\forall x, \mu \leq f''(x) \leq L \text{ pour un } \mu > 0$$

remarque: $f'' \geq \mu$ signifie fortement convexe

$f'' \leq L$ signifie que f' est L -lipschitzienne

L'algorithme revient à estimer :

$$x_{k+1} = T(x_k), \text{ où } T(x) = x - \gamma f'(x)$$

On cherche alors un point fixe. En effet, si $x_k \rightarrow x^*$, alors $x^* = T(x^*)$ et on observe :

$$x^* = x^* - \gamma f'(x^*) \Leftrightarrow f'(x^*) = 0 \Leftrightarrow x^* \text{ minimise } f$$

Autrement dit, si on atteint un point fixe en appliquant l'algorithme ce point fixe sera forcément un minimiseur de f .

Il reste à prouver que il existe bien un point fixe tel que $x_{k+1} = T(x_k)$. On utilise le théorème suivant :

Théorème du point fixe

Si T est une contraction, alors T admet un unique point fixe x^* . (et donc dans le cas $x_k \rightarrow x^*$)

note : T est une contraction si :

$$\forall x, y : |T(x) - T(y)| \leq \alpha |x - y| \text{ pour } 0 < \alpha < 1$$

pour le cas de l'algorithme du gradient, T est une contraction si :

$$|x_{k+1} - x^*| = |T(x_k) - T(x^*)| \leq \alpha |x_k - x^*|$$

montrons le : on prend la dérivée de T :

$$T'(x) = \frac{d(x - \gamma f'(x))}{dx} = 1 - \gamma f''(x)$$

ce qui implique :

$$1 - \gamma L \leq T'(x) \leq 1 - \gamma \mu \text{ par hypothèse}$$

On a une contraction si $\exists p < 1$ tel que :

$$-p \leq T'(x) \leq p \text{ (par intégration ?)}$$

$$T'(x) \leq 1 - \gamma \mu < 1$$

et il faut aussi que : $-1 < 1 - \gamma L$, soit $\gamma < \frac{2}{L}$

III / Algorithme du gradient stochastique (SGD)

- problème = $\min_x \mathbb{E}[f(x, \xi)] = F(x)$
où ξ est une variable aléatoire
- on ne sait pas calculer l'espérance : soit la loi de ξ est inconnue, soit le calcul est trop coûteux.
- on connaît en revanche des réalisations i.i.d de $\xi = \xi_1, \dots, \xi_n$.

- algorithme : on itère :

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k, \xi_{k+1}) \quad \begin{matrix} \text{f} \text{ est } \text{ séquence qui décrivent alors} \\ \text{le temps} \end{matrix}$$

- exemple : minimisation du risque empirique :

$$\rightarrow \min_{w \in X} \frac{1}{n} \sum_{i=1}^n l(h_w(x_i), y_i)$$

avec x_i = feature y_i = label h_w = score l = loss

→ problème avec algorithme du gradient :

$$w_{k+1} = w_k - \gamma \frac{1}{n} \sum_{i=1}^n \frac{\partial l(h_w(x_i), y_i)}{\partial h_w(x_i)} \cdot \nabla_w l_w$$

si n est large, le calcul de la somme est très coûteux.

→ reformulation = $\min_{w \in X} \mathbb{E}[l(h_w(x_\xi), y_\xi)]$

où $\xi \sim \text{Unif}(\{1, 2, \dots, n\})$, $P(\xi=i) = \frac{1}{n} \quad \forall i = 1, \dots, n$

→ on obtient alors l'algorithme SGD pour le risque empirique :

à l'étape k , tirer aléatoirement $\xi_{k+1} \sim \text{U}(\{1, \dots, n\})$

$$w_{k+1} = w_k - \gamma_k \frac{\partial l(h_w(x_{\xi_{k+1}}), y_{\xi_{k+1}})}{\partial h_w(x_{\xi_{k+1}})} \cdot \nabla_w l_w$$

- preuve du gradient stochastique :

$$F(x) = \mathbb{E}(f(x, \xi))$$

hypothèses : → F admet un minimiseur x^*

→ F est convexe

→ $\forall x, \mathbb{E}(\|\nabla f(x, \xi)\|^2) \leq c < +\infty$

Démonstration :

$$\begin{aligned}
 \|x_{k+1} - x^*\|^2 &= \|x_k - \gamma_k \nabla f(x_k, \xi_{k+1}) - x^*\|^2 \\
 &= \|x_k - x^* - \gamma_k \nabla f(x_k, \xi_{k+1})\|^2 \\
 &= \|x_k - x^*\|^2 + 2 \langle x_k - x^*, -\gamma_k \nabla f(x_k, \xi_{k+1}) \rangle + \|-\gamma_k \nabla f(x_k, \xi_{k+1})\|^2 \\
 &\text{en utilisant } \|x+y\|^2 = \|x\|^2 + 2 \langle x, y \rangle + \|y\|^2 \\
 &= \|x_k - x^*\|^2 - 2 \gamma_k \langle \nabla f(x_k, \xi_{k+1}), x_k - x^* \rangle + \gamma_k^2 \|\nabla f(x_k, \xi_{k+1})\|^2
 \end{aligned}$$

notons E_k l'espérance conditionnellement à x_k , c'est-à-dire, en intégrant que la variable aléatoire ξ_{k+1} :

$$\begin{aligned}
 E_k(\|x_{k+1} - x^*\|^2) &= \|x_k - x^*\|^2 - 2 \gamma_k \langle E_k(\nabla f(x_k, \xi_{k+1})), x_k - x^* \rangle \\
 &\quad + \gamma_k^2 E_k(\|\nabla f(x_k, \xi_{k+1})\|^2) \leq c
 \end{aligned}$$

remarque : on pose (faute à montrer) : $E(\nabla f(x, \xi)) = \nabla F(x)$
on obtient alors :

$$E_k(\|x_{k+1} - x^*\|^2) \leq \|x_k - x^*\|^2 - 2 \gamma_k \langle \nabla F(x_k), x_k - x^* \rangle + \gamma_k^2 c$$

Or, F est convexe et donc : $\forall y, F(y) \geq f(x_k) + \langle \nabla F(x_k), y - x_k \rangle$

$$F(x^*) \geq F(x_k) + \langle \nabla F(x_k), x^* - x_k \rangle$$

donc :

$$- \langle \nabla F(x_k), x_k - x^* \rangle \leq F(x^*) - F(x_k)$$

$$E_k(\|x_{k+1} - x^*\|^2) \leq \|x_k - x^*\|^2 + 2 \gamma_k (F(x^*) - F(x_k)) + \gamma_k^2 c$$

$$E(\|x_{k+1} - x^*\|^2) \leq E(\|x_k - x^*\|^2) + 2 \gamma_k (F(x^*) - E(F(x_k))) + \gamma_k^2 c$$

(obtenu en utilisant $E(E(X|Y)) = E(X)$)

on backward une période :

$$E(\|x_k - x^*\|^2) \leq E(\|x_{k-1} - x^*\|^2) + 2 \gamma_{k-1} (F(x^*) - E(F(x_{k-1}))) + \gamma_{k-1}^2 c$$

$$E(\|x_{k-1} - x^*\|^2) \leq \dots$$

$$\leq \|x_0 - x^*\|^2 + 2 \gamma_0 \dots$$

On fait une somme télescopique :

$$\begin{aligned}
 0 \leq E(\|x_{k+1} - x^*\|^2) &\leq -2 \sum_{j=0}^k \gamma_j (\mathbb{E}(F(x_j)) - F^*) + \sum_{j=0}^k \gamma_j^2 c \\
 &\quad + \|x_0 - x^*\|^2
 \end{aligned}$$

→

Résultat :

$$\frac{\sum_{j=0}^k \gamma_j E(F(x_j))}{\sum_{j=0}^k \gamma_j} - F^* \leq \frac{\|x_0 - x^*\|^2 + c \sum_{j=0}^k \gamma_j^2}{2 \cdot \sum_{j=0}^k \gamma_j}$$

Posons $\bar{x}_k = \frac{\sum_{j=0}^k \gamma_j x_j}{\sum_{j=0}^k \gamma_j}$ par Jensen: $F(\bar{x}_k) \leq \frac{\sum_{j=0}^k \gamma_j F(x_j)}{\sum_{j=0}^k \gamma_j}$

et finalement :

$$E(F(\bar{x}_k)) \leq F^* + \frac{\|x_0 - x^*\|^2 + c \sum_{j=0}^k \gamma_j^2}{2 \sum_{j=0}^k \gamma_j}$$

si γ constant : $\frac{\|x_0 - x^*\|^2 + c \cdot k \cdot \gamma^2}{2 \cdot k \cdot \gamma} \underset{k \rightarrow \infty}{\approx} \frac{c \gamma}{2}$

la convergence la plus rapide (meilleure vitesse de convergence) est obtenue pour :

$$\gamma_k = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$

IV / Dualité en optimisation

problème : $\min_x f(x)$ s.c $Ax = b, g_1(x) \leq 0, \dots, g_m(x) \leq 0$
 g_1, \dots, g_m convexes

dans ce cours, on ne considère que les contraintes d'égalité :

$$\min_x f(x) \text{ s.c } Ax = b$$

$$p = \inf_x f(x) \text{ s.c } Ax = b$$

$$\text{Lagrangien} = \mathcal{L}(x, \lambda) = f(x) + \langle \lambda, Ax - b \rangle$$

$$\in \mathbb{R}^d \quad \in \mathbb{R}^m$$

"problème primal"

"valeur primaire"

$$\Phi(\lambda) = \inf_{x \in \mathbb{R}^d} L(x, \lambda)$$

"fonction dual"

$$d = \sup_{\lambda \in \mathbb{R}^m} \Phi(\lambda)$$

"valeur dual"

$$\max_{\lambda} \Phi(\lambda)$$

"problème dual"

• propriété :

$$\begin{cases} d = \sup_{\lambda} \inf_x L(x, \lambda) & \text{vrai par définition} \\ \varphi = \inf_x \sup_{\lambda} L(x, \lambda) \end{cases}$$

En effet : $\sup_{\lambda} L(x, \lambda) = \sup_{\lambda} (f(x) + \langle \lambda, Ax - b \rangle)$
 $= f(x) + \sup_{\lambda} \langle \lambda, Ax - b \rangle$
 $= \begin{cases} f(x) & \text{si } Ax = b \\ +\infty & \text{sinon} \end{cases}$

On conclut que $\inf_x (\sup_{\lambda} L(x, \lambda))$ sera forcément atteint sur un point où $Ax = b$ (car sinon $L(x, \lambda) = +\infty$). Donc :

$$\inf_x (\sup_{\lambda} L(x, \lambda)) = \inf_{x: Ax = b} \sup_{\lambda} L(x, \lambda) = \inf_{x: Ax = b} f(x) = p$$

• résultat (dualité faible) : $d \leq p$

preuve : $\inf_x L(x, \lambda) \leq L(x, \lambda) \leq \sup_{\lambda} L(x, \lambda)$

on utilise le fait que $F(\lambda) \leq p \Rightarrow \sup_{\lambda} F(\lambda) \leq p$ et donc :

$$\sup_{\lambda} \inf_x L(x, \lambda) \leq \inf_x \sup_{\lambda} L(x, \lambda) \quad \text{vrai } \forall x$$

$$= d \qquad \qquad \qquad = p$$

• définition : (x, λ) est un point zelle du Lagrangien si :

$$L(x, \lambda) = \sup_{\lambda'} L(x, \lambda') \rightarrow \text{si } x \text{ fixé, je prends } L \text{ comme une fonction de } \lambda \text{ seulement. } \lambda \text{ maximise } L(x, \cdot)$$

$$L(x, \lambda) = \inf_x L(x, \lambda)$$

point zelle : si je bouge en direction de λ , j'augmente L (car λ = maximum)
 si je bouge en direction de x , j'augmente L

- Théorème: (x, λ) est un point selle de L si et seulement si:
 - x est solution primaire
 - λ est solution duale
 - $\rho = d$
 autrement dit : si on trouve un point selle du lagrangien, on a résolu le problème d'optimisation (sc = multiplicateur de lagrange).

- preuve: soit (x, λ) un point selle :

$$\rho = \inf_{x^*} \sup_{\lambda^*} L(x^*, \lambda^*) \leq \sup_{\lambda^*} L(x, \lambda^*)$$

$$= L(x, \lambda) \quad (\text{définition point-selle})$$

$$= \inf_{x^*} L(x^*, \lambda)$$

$$\mathbb{E}(\lambda^*)$$

$$\Leftrightarrow \sup_{\lambda^*} \inf_{x^*} L(x^*, \lambda^*) = d$$

$$\text{donc } \rho \leq d \quad \left. \begin{array}{l} d = p \\ \text{par dualité faible, on a aussi } d \leq \rho \end{array} \right\} d = p$$

Comme $d = p$, les inégalités doivent être des égalités.

Donc: $\mathbb{E}(\lambda) = \inf_{x^*} L(x^*, \lambda)$ $\mathbb{E}(\lambda) = \sup_{\lambda^*} \mathbb{E}(\lambda^*)$ et λ

minimise la fonction duale : on a prouvé que λ est solution duale. On suit la même logique pour le premier point.

- utilité:

① Si le problème dual est plus facile à résoudre. Une fois qu'on a λ , la solution duale, on peut trouver une solution primaire en minimisant $L(x, \lambda)$.

② méthodes primales-duales (sc = ADMM)
trouver un point-selle.

II / Application : SVM

- problème : $\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i x_i^\top w) + \frac{\alpha}{2} \|w\|_2^2$

- Hinge loss : $l(u) = \max(0, 1 - u)$

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \quad (n \times d) \quad D(Y) = \begin{pmatrix} y_1 & \stackrel{\pm 1}{=} & 0 \\ \ddots & & \\ 0 & & y_m \end{pmatrix} \quad (n \times n) \quad Xw = \begin{pmatrix} x_1^\top w \\ \vdots \\ x_n^\top w \end{pmatrix}$$

$$D(Y)Xw = D(Y) \begin{pmatrix} x_1^\top w \\ \vdots \\ x_n^\top w \end{pmatrix} = \begin{pmatrix} y_1 x_1^\top w \\ \vdots \\ y_m x_n^\top w \end{pmatrix}$$

- posons $g \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n l(z_i)$ alors $g(D(Y)Xw) = \frac{1}{n} \sum_{i=1}^n l(y_i x_i^\top w)$

- réformulation du problème SVM :

$$\min_{w \in \mathbb{R}^d} g(D(Y)Xw) + \frac{\alpha}{2} \|w\|_2^2$$

$$\Leftrightarrow \min_{\substack{w \in \mathbb{R}^d \\ z \in \mathbb{R}^m}} g(z) + \frac{\alpha}{2} \|w\|_2^2 \quad \text{s.c. } z = D(Y)Xw$$

$$\Leftrightarrow \min_{(w, z)} g(z) + \frac{\alpha}{2} \|w\|_2^2$$

$D(Y)Xw - z = 0$ rappel : contrainte de formulation de forme $Ax = b \Rightarrow [D(Y)X, -I] \begin{pmatrix} w \\ z \end{pmatrix} = 0$

$$\Leftrightarrow \min_{w, z: A \begin{pmatrix} w \\ z \end{pmatrix} = 0} g(z) + \frac{\alpha}{2} \|w\|_2^2$$

$$\mathcal{L}(w, z, \lambda) = g(z) + \frac{\alpha}{2} \|w\|_2^2 + \langle \lambda, D(Y)Xw - z \rangle$$

(rappel : $\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, Ax - b \rangle$)

- fonction duale :

$$\begin{aligned} \Phi(\lambda) &= \inf_{(w, z)} \mathcal{L}(w, z, \lambda) \\ &= \inf_w \inf_z [g(z) + \frac{\alpha}{2} \|w\|_2^2 + \langle \lambda, D(Y)Xw \rangle - \langle \lambda, z \rangle] \\ &= \inf_w \inf_z [(g(z) - \langle \lambda, z \rangle) + (\frac{\alpha}{2} \|w\|_2^2 + \langle \lambda, D(Y)Xw \rangle)] \\ &= \inf_z (g(z) - \langle \lambda, z \rangle) + \inf_w \left[\frac{\alpha}{2} \|w\|_2^2 + \langle \lambda, D(Y)Xw \rangle \right] \end{aligned}$$

- on admet le résultat de A :

$$A^* = \begin{cases} -\sum_{i=1}^m \lambda_i & \text{if } \forall i, \lambda_i \in \left[-\frac{1}{n}, 0\right] \\ -\infty & \text{otherwise} \end{cases}$$

- calculons maintenant le terme B :

$$B = \inf_w \frac{\alpha}{2} \|w\|_2^2 + \langle (D(Y)X)^T \lambda, w \rangle$$

$$\text{Solution: } 0 = \alpha w + (DX)^\top \lambda \rightarrow w^* = -\frac{1}{\alpha} (D(Y)X)^\top \lambda$$

on substitue par retour dans la fonction B :

$$= \frac{\alpha}{2} \left\| -\frac{(D(Y)X^\top)\lambda}{\alpha} \right\|_2^2 + \left\langle (D(Y)X)^\top \lambda, -\frac{(D(Y)X)^\top \lambda}{\alpha} \right\rangle$$

$$= -\frac{1}{2x} \|\mathcal{D}(y)x\|^2_2$$

- on conclut la solution globale suivante :

$$\mathbb{E}(\lambda) = \begin{cases} -\sum_{i=1}^n \lambda_i - \frac{1}{2\alpha} \| (\mathbf{D}(Y)\mathbf{X})^\top \boldsymbol{\lambda} \|_2^2 & \text{si } \lambda \in [-\frac{1}{n}, 0] \\ -\infty & \text{sinon} \end{cases}$$

- problème dual = maximiser $\Xi(\lambda)$ équivaut à :

$$\min_{\lambda} \sum_{i=1}^n \lambda_i + \frac{1}{2\alpha} \|(\mathcal{D}(Y)X)^T \lambda\|_2^2 \quad \text{avec } \lambda \text{ tel que } \nu_i, \lambda_i \in [-\frac{1}{n}, 0]$$

- commentaire : si $n < d$, utiliser la dualité sera moins coûteux en calcul

$$\begin{array}{ccccccc} x & & \times & \times & x & & \\ & & \diagup & & \diagdown & & \\ x & x & & 0 & 0 & & \\ & & & x & & & \\ x & & & 0 & 0 & & \\ & & & x & & & \\ \xrightarrow{\quad\quad\quad} & & 0 & 0 & & & \\ & & x & x & & & \\ & & & x & x & & \\ \end{array} \longrightarrow \varphi = R^d \rightarrow R^{d'}, d' \gg d$$

- autre intérêt de la dualité :

$$\|(\mathbf{D}(Y)\mathbf{X})^\top \lambda\|_2^2 = \lambda^\top \mathbf{D}(Y) \mathbf{X} \mathbf{X}^\top \mathbf{D}(Y)^\top \lambda = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j \frac{(x_i^\top x_j)}{\text{produit scalaire}}$$

→ méthodes de noyaux = kernel SVM

→ méthodes à noyaux : kernel SVM

- $$\bullet \text{problem dual} = \min_{\lambda \in \mathbb{R}^m} \sum_{i=1}^m \lambda_i + \sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j y_i y_j \varphi(x_i)^T \varphi(x_j) \quad \forall i, \lambda_i \in [-\frac{1}{n}, 0]$$

seul élément qui intervient dans le problème dual : $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$

il suffit de connaître le noyau $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$