Le gradient stochastique

I Du déterministe à l'aléatoire:

On considère le problème suivant : $\min_{u \in U} J(u)$ où $\exists j$ tq j soit une fonction dépendant d'une variable aléatoire W définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et à valeurs dans $\mathcal{W}$ muni d'une tribu et $J(u) = \mathbb{E}[j(u, W)]$.

Le problème s'écrit donc trouver $\min_{u \in U} \mathbb{E}[j(u,W)]$ (1)

D'un point de vue purement théorique, on peut calculer, en tout point de u sous certaines conditions de régularité, les valeurs de $J(u)$ et de son gradient $\nabla_u J$. C'est un problème d'intégrale à paramètre. On se ramène donc à un problème déterministe pour lequel plusieurs méthodes de minimisations (gradient, Newton, etc) peuvent être appliquées.

Cependant d'un point de vue pratique, chaque évaluation de $J(u)$ et $\nabla_u J$ nécessite le calcul d'une espérance, ce qui peut être couteux en tps de calcul surtout si W est de grande dimension.

→ Idée du gradient stochastique

l'idée est de faire évoluer simultanément le calcul de $\mathbb{E}$ et la méthode de descente de gradient en s'appuyant sur la méthode de Monte Carlo qui donne une approximation de l'$\mathbb{E}$.

On tire aléatoirement une suite i.i.d de réalisations de W : $(w_1, \dots, w_k, \dots)$ selon sa loi (celle sous laquelle est calculée l'$\mathbb{E}$). On génère une suite $(u_k)_k$ telle que $u_k \to u^*$ solution de (1). La mise à jour de u_k se fait avec une valeur simulée de w_k par un pas de gradient de la fonction j

Il faut bien sûr s'assurer de la convergence de cet algorithme en particulier il faut vérifier que la variation en u est lente pour que la moyennisation en W ait un sens

Voici l'algorithme en pseudo-code :

On se donne u_0

$k \to k+1$: on simule $w^{k+1} \sim \mathbb{P}_W$ et $u_{k+1} = u_k - \varepsilon_k \nabla_{u_k} j(u_k, w_{k+1})$

où $(\varepsilon_k)_k$ suite $\searrow 0$

→ Interprétation intuitive :

L'idée est de profiter des itérations d'optimisation pour effectuer le calcul de l'espérance :

$$u_{k+1} = u_k - \varepsilon_k \nabla_{u_k} j(u_k, w_{k+1})$$

On somme k fois cette formule à partir d'un indice k_0 fixé

$$u_{k+k_0} = u_{k_0} - \sum_{l=0}^{k-1} \varepsilon_{k_0+l} \nabla_u j(u_{k_0+l}, w_{k_0+l+1})$$

Si on sait que $u \mapsto \nabla_u J(u,w)$ est "suffisamment régulière" et que $|u_{k_0+l} - u_{k_0+l'}|$ "petit" (A définir évidemment de manière plus rigoureuse plus tard) alors :

$$u_{k_0+k} \simeq u_{k_0} - \sum_{l=0}^{k-1} \varepsilon_{k_0+l} \nabla_u j(u_{k_0}, w_{k+l+1})$$

$$\simeq u_{k_0} - \left(\sum_{l=0}^{k-1} \varepsilon_{k_0+l}\right) \nabla J(u_{k_0})$$

En utilisant le résultat ci dessous (*)

Cela donne une étape de descente de gradient sur J. Le gradient stochastique utilise les itérations de l'optimisation pour reconstituer l'espérance du gradient du critère (et non l'espérance du critère lui-même).

→ Rq : s'adapte au cas sous contrainte $u \in U^{\text{admissible}}$ convexe fermé $\neq \emptyset$ par projection

Théorème de Césaro en moyenne

(*) Soit $(x_k)_k$ une suite dans un Hilbert X qui converge en moyenne vers μ :

$$\lim_{N \to \infty} \left(\frac{1}{N} \sum_{k=1}^{N} x_k\right) = \mu$$

Soit $(\rho_k)_k$ une suite de réels positifs $\searrow 0$. On suppose que la suite $(\varepsilon_k)_k$ définie par $\varepsilon_k = k(\rho_k - \rho_{k+1}) > 0$ est tq ε_k soit le terme général d'une série divergente. Alors :

$$\lim_{N \to \infty} \frac{\sum_{k=1}^{N} \rho_k x_k}{\sum_{k=1}^{N} \rho_k} = \mu$$

Preuve du théorème de Césaro en moyenne :

On se ramène au cas $\mu = 0$ en translatant x_k par μ.

Soit $y_k = \frac{1}{k} \sum_{\ell=1}^{k} x_\ell$. Par hypothèse $y_k \xrightarrow[k\infty]{} \mu = 0$

Comme $\rho_k \to 0$ et $\varepsilon_k = k(\rho_k - \rho_{k+1})$ on obtient par somme télescopique

$$\rho_k = \sum_{\ell=k}^{\infty} \frac{\varepsilon_\ell}{\ell}$$

Soit $N \in \mathbb{N}^*$, $\forall k \leq N$

$$\rho_k \overset{(\cdot)}{=} \rho_N^k + \rho_{N+1} \quad \text{avec} \quad \rho_N^k = \sum_{\ell=k}^{N} \frac{\varepsilon_\ell}{\ell}$$

Avec ces notations :

$$\sum_{k=1}^{N} \varepsilon_k y_k = \sum_{k=1}^{N} \frac{\varepsilon_k}{k} \sum_{\ell=1}^{k} x_\ell = \sum_{\ell=1}^{N} x_\ell \sum_{k=\ell}^{N} \frac{\varepsilon_k}{k} = \sum_{\ell=1}^{N} \rho_N^\ell x_\ell \quad (\cdot\cdot)$$

Ceci est vrai $\forall (x_k)_k$ donc on prend le cas particulier $x_k = 1 \ \forall k$ et on obtient $\sum_{k=1}^{N} \varepsilon_k = \sum_{k=1}^{N} \rho_N^k$. $(\cdot\cdot\cdot)$

On obtient donc :

$$\left\| \frac{\sum_{k=1}^{N} \rho_k x_k}{\sum_{k=1}^{N} \rho_k} \right\| = \frac{\left\| \sum_{k=1}^{N} \rho_N^k x_k + \rho_{N+1} \sum_{k=1}^{N} x_k \right\|}{\sum_{k=1}^{N} \rho_N^k + N\rho_{N+1}} \quad \text{par } (\cdot)$$

$$\leq \frac{\left\| \sum_{k=1}^{N} \rho_N^k x_k \right\|}{\sum_{k=1}^{N} \rho_N^k + N\rho_{N+1}} + \frac{\left\| \rho_{N+1} \sum_{k=1}^{N} x_k \right\|}{\sum_{k=1}^{N} \rho_N^k + N\rho_{N+1}} \quad \text{(inégalité triangulaire)}$$

$$\leq \frac{\left\| \sum_{k=1}^{N} \rho_N^k x_k \right\|}{\sum_{k=1}^{N} \rho_N^k} + \frac{\left\| \rho_{N+1} \sum_{k=1}^{N} x_k \right\|}{N\rho_{N+1}} \quad (\text{car } \rho_k \geq 0 \ \forall k)$$

$$\leq \left\| \frac{\sum_{k=1}^{N} \varepsilon_k y_k}{\sum_{k=1}^{N} \varepsilon_k} \right\| + \underbrace{\frac{1}{N} \left\| \sum_{k=1}^{N} x_k \right\|}_{\to 0 \text{ par hypothèse}} \quad \text{par } (\cdot\cdot) \text{ et } (\cdot\cdot\cdot)$$

Or ε_k est le terme général d'une série divergente donc le théorème de Césaro classique implique que le 1er terme de droite $\xrightarrow[N\infty]{} 0$

Soit $(u_n) \in E$ un R-evec normé, $\lambda_n > 0$ tq $\sum \lambda_n$ diverge. Si $u_n \to \ell$ alors $\sigma_n = \frac{\sum u_k \lambda_k}{\sum \lambda_k} \longrightarrow \ell$

I Le gradient stochastique :

On va ici voir le cadre probabiliste de cet algorithme et l'étude de ses propriétés, en particulier sa convergence.

1) Algorithme :

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et W définie sur Ω à valeurs dans $\mathcal{W}$. On note μ la loi de W. Soit $\mathcal{U}$ un Hilbert et $\mathcal{U}_c$ une partie convexe fermée non vide de $\mathcal{U}$. Soit $j : \mathcal{U} \times \mathcal{W} \longrightarrow \overline{\mathbb{R}}$ et $J = \mathbb{E}_\mu[j(\cdot, W)]$.

On cherche u^* tq $u^* = \underset{u \in \mathcal{U}_c}{\operatorname{argmin}} J(u)$

Sous les hypothèses classiques de convexité de J et de différentiabilité, si on se résoud à calculer J et ∇J sur une trajectoire alors on peut utiliser l'algorithme de descente de gradient qui converge à vitesse $\frac{1}{k}$. (L'existence d'un ensemble $\mathcal{U}^* = \{\underset{u \in \mathcal{U}_c}{\operatorname{argmin}} J(u)\}$ étant supposé $\neq \emptyset$). Ici on essaie d'éviter les calculs des d'espérances.

Algorithme en pseudo-code :

(S)

- On se donne $u_0 \in \mathcal{U}_c$ ainsi qu'une suite de pas $(\varepsilon_k)_{k \geq 0} \geq 0$
- Pour $k = 1$: end

  $w_{k+1} \sim \mu$ (Réalisation $\perp\!\!\!\perp$ de $w_1, \dots, w_k$)

  Calculer $\nabla_u j(u_k, w_{k+1})$

  $u_{k+1} = \operatorname{proj}_{\mathcal{U}_c}[u_k - \varepsilon_k \nabla_u j(u_k, w_{k+1})]$

  fin.

Rq1 : On verra plus loin le critère d'arrêt (end) ; on se doute que ce ne pourra pas être comme pour la descente déterministe dû aux aléats.

Rq2 : Il faut être capable de simuler $(w_k)_k$ la suite i.i.d $\sim \mu$. On verra plus tard (Cours 2-3) quelques méthodes de simulations et pas forcément i.i.d mais avec des garanties quand même.

On va regarder un exemple qui illustre bien ce que l'on fait ~~et~~ présenté en contexte un peu plus général aussi.

<u>Exemple</u>: Soit $W : \Omega \to \mathbb{R}$ L^1. La question est calculer $\mathbb{E}_\mu[W]$;

La manière connue pour approcher $\mathbb{E}[W]$ est d'utiliser la formule de Monte Carlo :

$$\mathbb{E}[W] \simeq \frac{1}{k}\sum_{j=1}^{k} w_j \quad \text{où } w_k \sim \mu \text{ i.i.d.}$$

On note $u_{k+1} = \frac{1}{k+1}\sum_{j=1}^{k+1} w_j$. On sait par la loi forte des grands nombres que $U_k = \frac{1}{k}\sum_{j=1}^{k} W_j \xrightarrow[k\,\infty]{ps} \mathbb{E}_\mu[W]$ quand W_j iid μ.

On va réécrire $(U_k)_k$ sous une autre forme :

$$U_{k+1} = \frac{1}{k+1}\sum_{j=1}^{k} W_j + \frac{W_{k+1}}{k+1} = \frac{1}{k}\sum_{j=1}^{k} W_j - \frac{1}{k+1}\left(\frac{1}{k}\sum_{j=1}^{k} W_j - W_{k+1}\right)$$

$$= U_k - \frac{1}{k+1}\left(U_k - W_{k+1}\right)$$

Soit $\varepsilon_k = \frac{1}{k+1}$ et $j(u,w) = \frac{1}{2}(u-w)^2$; on obtient alors que

$$U_{k+1} = U_k - \varepsilon_k \nabla_u j(U_k, W_{k+1})$$

Or l'espérance d'une variable aléatoire est la valeur minimale du critère de dispersion d'un nuage de points : $\mathbb{E}[W] = \underset{u\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}(u-w)^2$

Donc la méthode d'approximation de l'espérance par somme de Monte Carlo est en fait un gradient stochastique sur la fonctionnelle de dispersion.

On peut avec cet exemple faire quelques remarques :

- On a une idée de la suite (ε_k) dans ce cas précis $\varepsilon_k = \frac{1}{k} \longrightarrow 0$ mais pas trop vite... ($\sum \frac{1}{k} = +\infty$)
- La méthode de Monte Carlo converge ps, ce qui nous laisse espérer de même au moins sous certaines conditions pour le gradient stochastique.
- Il existe aussi des TCL pour les sommes de MC donc peut-être aussi pour le GS.

Rappels de modes de cvce d'une v.a :

- ps : $\forall \varepsilon > 0 \quad \lim_{k\,\infty} \mathbb{P}\left(\sup_{n\geq k} \|U_n - U^*\| > \varepsilon\right) = 0$
- en $\mathbb{P}$: $\forall \varepsilon > 0 \quad \lim_{k\,\infty} \mathbb{P}\left(\|U_k - U^*\| > \varepsilon\right) = 0$ $\Leftarrow$ cvce dominée
- en $\mathbb{E}$: $\lim_{k\,\infty} \mathbb{E}\left[\|U_k - U^*\|\right] = 0$ $\Leftarrow$

2) Résultats de convergence :

On va énoncer et démontrer un thm de cvce en moyenne quadratique car simple à prouver. MAis les conditions sont très restrictives donc on énoncera les thm généraux de convergence pour les Approximations stochastiques qui laisseront voir ce que l'on espère cvce ps et TCL.

H1 la variable aléatoire $j(u,W) : \Omega \to \mathbb{R}$ est mesurable et son espérance existe $\forall u \in \mathcal{U}_c$.

H2 la fonction $\begin{array}{l} u \longmapsto j(u,w) \\ U \longrightarrow \mathbb{R} \end{array}$ est convexe, semi C^0 inférieurement, à valeur dans $\overline{\mathbb{R}}$, différentiable $\forall w \in W$

H3 $\exists\, m > 0$ tq $\forall u \in \mathcal{U}_c$, $\forall w \in W$ $\|\nabla_u j(u,w)\| \leq m$

H4 Le problè $\min\limits_{u \in \mathcal{U}_c} J(u)$ admet un ensemble non vide de solutions $\mathcal{U}^*$ qui vérifie la relation : $\forall u \in \mathcal{U}_c$ $J(u) - J^* \geq c\, \mathrm{dist}_{\mathcal{U}^*}(u)^2$

où $J^* = \min\limits_{u \in \mathcal{U}_c} J(u)$ et $\mathrm{dist}_{\mathcal{U}^*}$ est la fonction distance à $\mathcal{U}^*$

H5 (ε_k) vérifie $\varepsilon_k > 0$ $\forall k$ et $\sum\limits_{k \in \mathbb{N}} \varepsilon_k = +\infty$, $\sum\limits_{k \in \mathbb{N}} \varepsilon_k^2 < \infty$

Théorème de convergence en moyenne quadratique :

Sous les hypothèses (H1 - 5), la suite $(U_k)_k$ de v.a générée par le GS converge en moyenne quadratique vers $\mathcal{U}^*$:

$$\lim_{k\infty} \mathbb{E}\left[\mathrm{dist}_{\mathcal{U}^*}(U_k)^2\right] = 0$$

Quelques remarques avant la preuve :

- convergence L^2 assez faible...
- H_3 très restrictive : par ex ne prend pas en cpt le cas des sommes de N.C
- H_5 : très très courante et se retrouve dans presque ts les thm de cvce des approximations stochastique qui sont des algorithmes s'écrivant sous la forme $\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n$

 où $h(\theta) = \mathbb{E}[H(\theta,W)]$ et la fonction que l'on veut annuler et

 $\eta_n = H(\theta_{n-1}, X_n) - h(\theta_{n-1})$ joue le rôle d'une perturbation aléatoire

Preuve du thm de convergence :

$\mathcal{U}^*$ étant un convexe fermé (car j convexe C'sur convexe fermé), la projection sur $\mathcal{U}^*$ est bien définie. Soit $(u_k)_k$ la suite produite par l'algorithme (GS) avec une réalisation de la suite $(w_k)_k$. Soit $\bar{u}_k = \text{proj}_{\mathcal{U}^*}(u_k)$. On a alors $\text{dist}_{\mathcal{U}^*}(u_k)^2 = \|u_k - \bar{u}_k\|^2$

Soit $d_k = \text{dist}_{\mathcal{U}^*}(u_k)^2$

$d_{k+1} = \|u_{k+1} - \bar{u}_{k+1}\|^2 \leq \|u_{k+1} - \bar{u}_k\|^2$ car d_{k+1} est la distance minimale de u_{k+1} à $\mathcal{U}^*$.

$\leq \|\text{proj}_{\mathcal{U}_c}(u_k - \varepsilon_k \nabla_u j(u_k, w_{k+1})) - \bar{u}_k\|^2$ par déf de u_{k+1}

$\leq \|u_k - \varepsilon_k \nabla_u j(u_k, w_{k+1}) - \bar{u}_k\|^2$ car la projection est contractante

$\leq d_k + \varepsilon_k^2 m^2 - 2\varepsilon_k \langle u_k - \bar{u}_k ; \nabla_u j(u_k, w_{k+1})\rangle$ développem^t du et ∇_j borné

On prend l'espérance conditionnelle sachant la filtration $\mathcal{F}_k$

$\mathbb{E}[d_{k+1} | \mathcal{F}_k] \leq \mathbb{E}[d_k | \mathcal{F}_k] + \varepsilon_k^2 m^2 - 2\varepsilon_k \mathbb{E}[\langle u_k - \bar{u}_k ; \nabla_u j(u_k, w_{k+1})\rangle | \mathcal{F}_k]$

$\leq d_k + \varepsilon_k^2 m^2 - 2\varepsilon_k \langle u_k - \bar{u}_k ; \mathbb{E}[\nabla_u j(u_k, w_{k+1}) | \mathcal{F}_k]\rangle$

Comme j est différentiable de gradient en u borné :

$$\mathbb{E}[\nabla_u j(u_k, w_{k+1}) | \mathcal{F}_k] = \nabla_u \mathbb{E}[j(u_k, w_{k+1}) | \mathcal{F}_k] = \nabla_u J(u_k)$$

D'où $\mathbb{E}[d_{k+1} | \mathcal{F}_k] \leq d_k + \varepsilon_k^2 m^2 - 2\varepsilon_k \langle u_k - \bar{u}_k ; \nabla_u J(u_k)\rangle$

$\leq d_k + \varepsilon_k^2 m^2 - 2\varepsilon_k (J(u_k) - J^*)$ par les propriétés de convexité

$\leq d_k(1 - 2\varepsilon_k c) + \varepsilon_k^2 m^2$ par H4

En reprenant $\mathbb{E}$ on obtient,

$$\mathbb{E}[d_{k+1}] \leq (1 - 2\varepsilon_k c)\,\mathbb{E}[d_k] + \varepsilon_k^2 m^2$$

On montre alors par récurrence que $\forall k$ suffisamment grand : $\forall n \in \mathbb{N}^*$

$$\mathbb{E}[d_{k+n+1}] \leq \underbrace{\left[\prod_{l=0}^{n} (1 - 2\varepsilon_{k+l} c)\right]}_{\to 0 \ (1)} \mathbb{E}[d_k] + \underbrace{\left(\sum_{l=0}^{n} \varepsilon_{k+l}^2\right)}_{\to 0 \text{ car } \sum_{k \in \mathbb{N}} \varepsilon_k^2 < \infty} m^2$$

CQFD

(1) car soit k_0 tq $\forall l \geq k_0$ $0 < (1 - 2\varepsilon_l c) \leq 1$ existe car $\varepsilon_k \to 0$

donc $P_k = \prod_{l=1}^{k} (1 - 2\varepsilon_k c) = C \prod_{l=k_0}^{k} (1 - 2\varepsilon_l c) = C\, \gamma_k$. $\gamma_k > 0$ ↘

donc converge. De plus $\log \gamma_k = \sum_{l=k_0}^{k} \log(1 - 2\varepsilon_l c) \leq -2c \underbrace{\sum_{l=k_0}^{k} \varepsilon_l}_{\to +\infty}$

$\Rightarrow \gamma_k \to 0$

Théorème (Vitesse de convergence)

Sous les mêmes hypothèses que précédemment, on choisit $\varepsilon_k = \dfrac{1}{ck + \frac{m^2}{cd^0}}$

avec $d_0 = \text{dist}_{U^*}(u_0)^2$ on obtient la borne de la vitesse de convergence

$$\mathbb{E}[\text{dist}_{U^*}(u_k)^2] \leq \frac{1}{\frac{c^2}{m^2}k + \frac{1}{d_0}} \quad \forall k \in \mathbb{N}$$

Preuve : On repart de l'inégalité suivante :

$$\mathbb{E}[d_{k+1}] \leq (1 - 2\varepsilon_k c)\,\mathbb{E}[d_k] + \varepsilon_k^2 m^2$$

Si $\varepsilon_k = \dfrac{\gamma}{\alpha k + \beta}$ alors on montre (par récurrence) que si $\alpha = \dfrac{c^2}{m^2}$, $\beta = \dfrac{1}{d_0}$

et $\gamma = \dfrac{c}{m^2}$: $(\alpha k + \beta)\,\mathbb{E}[d_k] \leq 1$ cqfd

3) Considérations pratiques de l'algorithme (GS) :

- Critère d'arrêt : * Ne peut être, comme pour la DG, basé sur $\|u_{k+1} - u_k\|$ car du fait que $\varepsilon_k \to 0$ et $|\nabla j| \leq m$ cette différence $\to 0$ par construction.
  - * pas non plus sur ∇j qui n'est pas informatif sur ∇J.
  - * une approximation de $\mathbb{E}[\nabla j]$ peut être utilisée
  - * le plus souvent un nb fixé par l'utilisateur.
- suite ε_k : en théorie pour avoir la plus grande vitesse assymptotique il faut $\varepsilon_k = \frac{1}{k}$. Mais en pratique (et pour le TCL aussi) $\frac{1}{\varepsilon^\alpha}$, $\alpha \in]\frac{1}{2},$ est parfois meilleur pour avoir un poids assez important de ∇j % à la précédente valeur. De plus, on peut aussi avoir besoin de ce qu'on appelle un tps de chauffe pour retirer les 1ères valeurs aberrantes.

On va maintenant regarder un cadre un peu plus général qui nous donnera des théorèmes de convergence plus forts (on attend p.s.) et un TCL.