Approximations stochastiques

Convergence et théorie asymptotique :

Le cadre des approximations stochastiques recouvre un spectre beaucoup plus large que ceux présentés les séances précédentes (gradient stochastique et EM stochastique)

La situation générale est la suivante. On suppose que l'on cherche θ^* tq $h(\theta^*) = \mathbb{E}_{\theta^*}[H(\theta^*, Y)] = 0$ (*)

où H est connue mais la distribution de Y (qui peut dépendre de θ) ne l'est pas.

Le fait est que l'on n'a pas accès directement à la fonction de coût h mais à des observations aléatoires $H(\theta, Y_n)$

ex: • minimiser / maximiser $Q(\theta)$ avec comme observations une version bruitée du gradient de Q : $H(\theta, Y_n) = \nabla Q(\theta) + Y_n$ où $(Y_n)_{n\geq 0}$ sont des réalisations d'un bruit additif.

• le gradient stochastique : $H(\theta, Y_n) = f(\theta, W_{n+1})$ où $W_{n+1} \sim P_W$

Vocabulaire: la fonction h est appelée le champ moyen.

L'idée des approximations stochastiques est de trouver θ^* itérativement en utilisant une mise à jour sous la forme :

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n \qquad (**)$$

où $h(\theta) = \mathbb{E}_\theta[H(\theta, Y)]$ est notre fonction à annuler

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

La raison de ce schéma apparaîtra au fur et à mesure puisque (**) s'écrit en fait $\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n)$ dans bcp de cas ce qui ressemble bien à notre cas du ∇ sto

Pour montrer la convergence de la suite $(\theta_n)_{n\geq 0}$, il est nécessaire de procéder en deux étapes :

① Trouver des conditions générales sur une suite déterministe de $(\eta_n)_{n\geq 0}$ et sur h qui assure la convergence déterministe de $(\theta_n)_{n\geq 0}$

② Montrer que ces conditions sont satisfaites avec probabilité 1 c'est à dire pour presque tout chemin $(\eta_1(\omega), \dots, \eta_n(\omega))$ du processus de bruit.

Le 1er point est développé ci dessous : sous des conditions de stabilité sur $(\theta_n)_n$ et des propriétés sur $(\eta_n)_{n\geq 0}$ (déterministe), nous allons voir que la convergence de la suite est fortement liée à la convergence de l'équation dite du champ moyen : $\frac{d\theta_t}{dt} = h(\theta_t)$

I Convergence ponctuelle dans le cas borné :

On suppose que notre terme de bruit (perturbation) peut s'écrire sous la forme : $\eta_n = e_n + r_n$ où :

- la série $\sum_{i=1}^{n} r_i e_i$ converge
- $r_i \xrightarrow[i\to\infty]{} 0$

On va voir que ces conditions sont imposées par la convergence.

Hypothèses : $\mathcal{O} \subset \mathbb{R}^d$ est un ouvert dont on note $\partial\mathcal{O}$ sa frontière.

(A) h est un champ de vecteurs continu sur $\mathcal{O}$

$\exists V$ une fonction positive (ou nulle) C^1 tq

(i) $\forall x \in \mathcal{O} \quad \langle \nabla V(x), h(x) \rangle \leq 0$

(ii) l'ensemble $\mathcal{S} = \{x : \langle \nabla V(x), h(x) \rangle = 0\}$ satisfait $V(\mathcal{S})$ est d'intérieur vide.

Rq: La fonction V est connue sous le nom de fonction de Lyapunov pour le champ h. C'est une condition très courante est peu restrictive Par exemple, si $h = -\nabla Q$ pour une fonction d'énergie ou de vraisemblance Q, le choix $V = Q$ fonctionne dès que Q est C^1

On va voir que l'hypothèse (A) va permettre d'assurer la convergence vers l'ensemble $\mathcal{S}$. En général, $\mathcal{S}$ coincidera avec $\{x : h(x) = 0\}$; De plus si $|\mathcal{S}| < \infty$, on convergera vers un point de $\mathcal{S}$.

le point (ii) semble compliqué à prouver, cependant le théorème de Sard (géométrie différentielle) peut souvent être utilisé. Il dit que pour une fonction $V \in C^d$ alors $V(\{\nabla V = 0\})$ est d'intérieur vide.

Nous allons maintenant donner les conditions de convergence de notre suite $(\theta_n)_n$ que nous allons écrire sous la forme : θ_0 donné

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n \qquad (1)$$

avec $\gamma_n \geq 0$; $\sum_{p=1}^{\infty} \gamma_p = +\infty$ et $\lim_{n \to \infty} \gamma_n = 0$

h est défini sur $\partial C \mathbb{R}^d$, e_n et r_n sont des perturbations et γ_n est appelé suite de gain (scalaire positive ou nulle)

Def: On dit que l'algorithme est A-stable si

θ_n reste dans un sous ensemble compact de ∂

$\lim_{p \to \infty} \sum_{n=1}^{p} \gamma_n e_n$ existe

et $\lim_{n \to \infty} |r_n| = 0$

théorème: On suppose que l'algorithme donné sous la forme (1) est A-stable. Sous l'hypothèse (A), la distance de θ_n à l'ensemble $\mathcal{G}$ noté $d(\theta_n, \mathcal{G})$ converge vers 0. En particulier si $|\mathcal{G}| < \infty$, θ_n tend vers un point de $\mathcal{G}$.

Preuve: Soit $F(\theta) = \langle \nabla V(\theta), h(\theta) \rangle$.

Soit $\theta_n' = \theta_n + \sum_{i=n+1}^{\infty} \gamma_i e_i$

et $\delta_n' = -\sum_{i=n}^{\infty} \gamma_i e_i$

Donc $\theta_{n+1}' = \theta_n + \delta_{n+1} = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n - \delta_{n+1}$

$= \theta_{n+1} - \delta_n + \delta_n - \delta_{n+1} + h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n$

$= \theta_{n-1}' - \gamma_n e_n + h(\theta_{n-1}' + \delta_n) + \gamma_n e_n + \gamma_n r_n$

$= \theta_{n-1}' + h(\theta_{n-1}' + \delta_n) + \gamma_n r_n$

En posant $r_n' = r_n - h(\theta_{n-1}') + h(\theta_{n-1}' + \delta_n)$ on obtient

$\theta_n' = \theta_{n-1}' + \gamma_n h(\theta_{n-1}') + \gamma_n r_n'$.

- De plus comme la série des γ_n en converge, la suite $\delta_n \xrightarrow[n\infty]{} 0$.
Par hypothèse $\theta_n \in K$ compact de S (A stabilité) donc
$\theta_n' \in K_0 \subset S$ compact aussi
Comme h est uniformément continue sur K_0 donc
$$h(\theta_{n-1}') - h(\theta_{n-1}' + \delta_n) \xrightarrow[n\infty]{} 0$$
Enfin $r_n \xrightarrow[n\infty]{} 0$ Donc $r_n' \xrightarrow[n\infty]{} 0$

- la fonction V est C^1 donc en utilisant la formule de Taylor :
$\exists \theta'' \in [\theta_{n-1}' ; \theta_n']$ tq
$$V(\theta_n') = V(\theta_{n-1}') + \gamma_n \langle \nabla V(\theta''), \underbrace{h(\theta_{n-1}') + r_n'}_{\text{car } = \theta_n' - \theta_{n-1}'} \rangle$$
$$= V(\theta_{n-1}') + \gamma_n \langle h(\theta_{n-1}') ; \nabla V(\theta_{n-1}') \rangle + \gamma_n \langle h(\theta_{n-1}') ; \nabla V(\theta'') - \nabla V(\theta_{n-1}') \rangle + \gamma_n \langle \nabla V(\theta'') , r_n' \rangle$$
Soit $r_n'' = \gamma_n \langle h(\theta_{n-1}') ; \nabla V(\theta'') - \nabla V(\theta_{n-1}') \rangle + \langle \nabla V(\theta'') , r_n' \rangle$
alors $|r_n''| \leq \|h\|_{L^\infty(K_0)} \|\nabla V(\theta'') - \nabla V(\theta_{n-1}')\| + \|\nabla V\|_{L^\infty(K_0)} |r_n'|$
or $r_n' \xrightarrow[n\infty]{} 0$ et $\|\nabla V(\theta'') - \nabla V(\theta_{n-1}')\| \xrightarrow[n\infty]{} 0$ car
$\theta'' \in [\theta_n' ; \theta_{n-1}']$ et $|\theta_{n-1}' - \theta_n'| \leq \gamma_n \underbrace{|h(\theta_{n-1}')|}_{\leq C \text{ sur } K_0} + \gamma_n \underbrace{|r_n'|}_{\to 0}$ et $\gamma_n \to 0$
donc $|\theta_{n-1}' - \theta''| \to 0$. Par continuité de ∇V car $V \in C^1$
on a le résultat.

Donc $V(\theta_n') = V(\theta_{n-1}') + \gamma_n f(\theta_{n-1}') + \gamma_n r_n''$
avec $r_n'' \xrightarrow[n\infty]{} 0$

Soit maintenant $\mathcal{P}$ un voisinage de $\mathcal{G} \cap K_0$. On rappelle que

$\mathcal{G} = \{F=0\}$. et $F(\theta) \leq 0 \quad \forall \theta \in K_0$ par (A)

Par continuité de F sur K_0 :

si $\theta'_{n-1} \in K_0 \setminus \mathcal{P}$ alors $F(\theta'_{n-1}) \leq -2\varepsilon$

si $\theta'_{n-1} \notin K_0 \setminus \mathcal{P}$ ie $\theta'_{n-1} \in \mathcal{P}$ alors $F(\theta'_{n-1}) \leq C_1$ entant que fonction C^0 sur un compact.

Donc pour n assez grand : $|r_n| \leq \varepsilon$ donc $\gamma_n |u_n| \leq \varepsilon \gamma_n$

$$(\Delta) \quad V(\theta'_n) \leq V(\theta'_{n-1}) - \gamma_n \varepsilon \mathbb{1}_{\{\theta'_{n-1} \notin \mathcal{P}\}} + \gamma_n C \mathbb{1}_{(\theta'_{n-1} \in \mathcal{P})}$$

où $C = C_1 + \varepsilon$.

↳ qui prend en compte $\gamma_n r_n$ de le cas $\in \mathcal{P}$

On va terminer la preuve en démontrant que $V(\theta'_n) \to V(\mathcal{G})$ et ensuite $\theta'_n \to \mathcal{G}$

- Soit $A_\alpha = \{x \in \mathbb{R} : d(x, V(\mathcal{G} \cap K_0)) < \alpha\}$

comme $\mathring{\mathcal{G}} = \emptyset$ A_α est une union d'intervalles de longueur 2α

Comme K_0 compact (ie borné), A_α est donc une union finie d'intervalles de longueur au moins 2α. (on peut avoir $\frac{1}{n} \to 0$ donc plus gd)

Soit α fixé, petit. On définit $\mathcal{P} = V^{-1}(A_\alpha)$

Dans ce cas, (Δ) définit une suite $u_n = V(\theta'_n)$ tq

$$u_n \leq u_{n-1} - \gamma_n \varepsilon \mathbb{1}_{A_\alpha^c}(u_{n-1}) + \gamma_n C \mathbb{1}_{A_\alpha}(u_{n-1})$$

A chaque fois que $u_{n-1} \notin A_\alpha$ $u_n \leq u_{n-1} - \gamma_n \varepsilon$

Puisque $\sum \gamma_i = +\infty$ u_n ne peut pas être une infinité de fois en dehors de A_α. (car $u_n \in V(\theta'_n) \subset V(K_0)$ compact)

Donc à partir d'un certain rang $u_n \in A_\alpha$ ie $d(u_n, A_\alpha) = 0$

- $\mathcal{S} \cap K_0$ est défini comme $\mathcal{S} = F^{-1}(0)$ donc $\mathcal{S}$ fermé donc $\mathcal{S} \cap K_0$ fermé dans un compact donc compact

Comme $V C^\circ$: $V(\mathcal{S} \cap K_0)$ compact donc fermé et $V(\mathcal{S})$ est d'intérieur vide donc par des arguments topologiques

$u_n \xrightarrow[n \to \infty]{}$ un élément de $V(\mathcal{S})$

càd $V(\theta'_{n-1}) \xrightarrow[n \to \infty]{} \{V(\mathcal{S})\}$

- Cette convergence de $V(\theta'_{n-1})$ implique que $\forall \varepsilon > 0 \quad \exists n(\varepsilon)$ tq

$\forall p > n \geq n(\varepsilon) \quad |V(\theta'_n) - V(\theta'_p)| \leq \varepsilon$

Donc par (Δ) en itérant en θ'_n et θ'_p

$$V(\theta'_{p-1}) - V(\theta'_p) \geq \gamma_p \varepsilon \mathbb{1}_{\{\theta'_{p-1} \notin \mathcal{P}\}} - \gamma_p C \mathbb{1}_{\{\theta'_{p-1} \in \mathcal{P}\}}$$

$$\Rightarrow V(\theta'_n) - V(\theta'_p) \geq \varepsilon \sum_{k=n+1}^{p} \gamma_k \mathbb{1}_{\{\theta'_{k-1} \notin \mathcal{P}\}} - C \sum_{k=n+1}^{p} \gamma_p \mathbb{1}_{\{\theta'_{k-1} \in \mathcal{P}\}}$$

D'où $\varepsilon \sum_{k=n+1}^{p} \gamma_k \mathbb{1}_{\{\theta'_{k-1} \notin \mathcal{P}\}} - C \sum_{k=n+1}^{p} \gamma_p \mathbb{1}_{\{\theta'_{k-1} \in \mathcal{P}\}} \leq \varepsilon$ (*)

Soit p tq $p > n \geq n(\varepsilon)$ et $\varepsilon(\gamma_{n+1} + \gamma_{n+2} + \dots + \gamma_p) > \varepsilon$
(possible car la série diverge)

Donc $\exists k \in \{n, \dots, p\}$ tq $\theta'_k \in \mathcal{P}$

Puisque $|\theta'_k - \theta'_n|$ peut être arbitrairement petit par (*) de l'ordre de $\frac{\varepsilon}{\varepsilon}$ par (*) (on n'a plus que les termes $\notin \mathcal{P}$ donc que le 1er membre $\leq \varepsilon$)

$d(\theta'_n, \mathcal{P}) \leq \theta(\frac{\varepsilon}{\varepsilon})$. Puisque ε arbitraire, on peut choisir $\varepsilon = \varepsilon^2$ et donc $d(\theta'_n, \mathcal{P}) \xrightarrow[n \to \infty]{} 0$

Comme le voisinage $\mathcal{P}$ de $\mathcal{S}$ est aussi arbitraire, on obtient

$$d(\theta'_n, \mathcal{S}) \xrightarrow[n \to \infty]{} 0$$

$$\Rightarrow \lim_{n \to \infty} d(\theta_n, \mathcal{S}) = 0.$$

Remarque : La condition d'un algorithme A stable est l'hypothèse de base de la convergence de $(\theta_n)_{n\geq 0}$. Le problème vient du fait de devoir supposer que θ_n reste dans un compact.

On peut montrer qu'en utilisant une suite de compacts croissants et en supposant que $(\theta_n)_n$ ne tend pas vers l'infini, on revient à la notion de A stabilité. Je vous renvoie vers le poly de Bernard Delyon (Stochastic approximation with decreasing gain : convergence and asymptotic theory, (2000)) Cela utilise de manière plus fine la fonction de Lyapunov.

La condition devient :

(B) : $h \in C^0(\mathcal{O})$; $\exists V$ positive (ou nulle) C^1, $\exists K \subset \mathcal{O}$ compact tq : (i) $V(x) \to +\infty$ si $x \to \partial\mathcal{O}$ ou $|x| \to +\infty$

(ii) $\langle \nabla V(x), h(x) \rangle < 0$ si $x \notin K$.

Et le théorème :

théorème : On suppose la condition (B) vérifiée. On suppose qu'il existe un compact $K_0 \subset \mathcal{O}$ tq

$\theta_n \in K_0$ infiniment souvent

On suppose que $\forall n \in \mathbb{N}$

$\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i \mathbb{1}_{\{V(\theta_{i-1}) \leq n\}}$ existe et $\lim_{n\to\infty} |r_n| \mathbb{1}_{\{V(\theta_{i-1}) \leq n\}} = 0$

Alors l'algorithme est A-stable.

Dans le cas de bruit aléatoire : cas particulier des algo de Robbins-Monro :

On s'est pour l'instant concentrés sur le cas déterministe ; maintenant on regarde un cas plus général où le bruit est aléatoire et donc la convergence sera soit ps soit en proba (on aura ps !)

Pour comprendre et parce qu'ils reflètent bien le cas général, on se concentre sur les algorithmes dits de Robbins-Monro qui s'écrivent

$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n)$ où $Y_n \sim \mathbb{P}_{\theta_{n-1}}$ et tq

$\mathbb{P}(Y_n \in A \mid Y_{n-1}, Y_{n-2}, \dots, \theta_0) = \mathbb{P}_{\theta_{n-1}}(Y_n \in A)$

L'algorithme cherche la solution de $\mathbb{E}_\theta[H(\theta, Y)] = 0 = h(\theta)$

De nouveau on réécrit cet équation en faisant apparaître h:

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \varepsilon_n \qquad (RM)$$

avec $\varepsilon_n = H(\theta_{n-1}, Y_n) - h(\theta_{n-1})$

On pose les hypothèses suivantes :

(R0): $\sum \gamma_n = +\infty \qquad \sum \gamma_n^2 < \infty$

(R1): h C^0 ; $\mathcal{S} = \{\theta : h(\theta) = 0\}$ est fini et $\exists$ V continuement différentiable tq

$$\begin{cases} \lim_{x \to \partial\Theta} V(x) = \infty & \text{non nécessaire si } (\theta_n) \text{ reste dans un compact.} \\ \langle \nabla V(\theta), h(\theta) \rangle \leq 0 \\ \{\langle \nabla V(\theta), h(\theta) \rangle = 0\} = \mathcal{S} \end{cases}$$

(R2): $\forall K$ compact de Θ $\quad \sup_{\theta \in K} \mathbb{E}_\theta[|H(\theta, Y)|^2] < \infty$

Alors

Théorème : sous les hypothèses (R0),(R1) et (R2), l'algorithme (RM) converge vers θ^* tq $h(\theta^*) = 0$ avec probabilité 1.

Preuve : On regarde le cas compact. $\theta_n \in K_0$ $\forall n \geq 0$

Soit $X_n = \gamma_n \varepsilon_n$ et $S_n = \sum_{k=1}^{n} X_k$

$$\begin{aligned} \mathbb{E}_\theta[|X_n|^2 \mid \mathcal{F}_{n-1}] &= \gamma_n^2 \, \mathbb{E}_\theta[|\varepsilon_n|^2 \mid \mathcal{F}_{n-1}] \\ &\leq \gamma_n^2 \sup_{\theta \in K_0} \mathbb{E}_\theta[|H(\theta, Y) - h(\theta)|^2 \mid \mathcal{F}_{n-1}] \\ &\leq \gamma_n^2 \, C_{K_0} \quad \text{car } h \ C^0 \text{ et } H \text{ aussi } \star (R_2) \end{aligned}$$

Donc $\mathbb{E}_\theta[|X_n|^2] \leq \gamma_n C_{K_0}$

Comme $\sum \gamma_n^2 < \infty$ la suite S_n converge ps (thm de Proba)

Et on retombe sur les conditions du précédent théorème ($\varepsilon_n = 0$)

Maintenant : étape suivante : (Y_n) chaîne de Markov !!! La suite à la séance 10 où on aura revu et manipulé les CM.