Notes de la Séance 6

OTOFA Samuel & SAID Meriem & SEYE El hadji babou

27 Septembre 2023

Table des matières

1 Statistique inférentielle asymptotique

1.1 Modes de convergence

Il y a quatre modes de convergences différents.
On a $X_1, \ldots, X_n$ une suite de variables aléatoires réelles (v.a.r.) définies sur Ω. On va définir la limite $\lim_{n\to+\infty} X_n$. Or, comme il y a plusieurs types de convergences, on note plutôt :

1.1.1 Convergence presque sûre ou presque partout

$X_n \to_{a.e.} X \sim P$, $a.e.$ = almost everywhere,
$\equiv X_n \to_{p.s.} X \sim P$, $p.s.$ = presque sûr,
$\equiv X_n \to_{a.s.} X \sim P$, $a.s.$ = almost sure.
ssi $\exists N \subset \Omega$, tq $P(N) = 0, \forall \omega \in \Omega \setminus N, \lim_{n\to+\infty} X_n(\omega) = X(\omega)$.
$\equiv \forall \varepsilon > 0, \lim_{n\to+\infty} P(\{\omega \mid \omega \in \Omega \text{ et } \exists m \geq n\} \|X_m(\omega) - X(\omega)\| > \varepsilon)$

Un estimateur θ_n de θ est fortement consistant de θ si et seulement si la suite de var $(\theta_n)n \in \mathbb{N}^*$ converge presque sûrement vers θ.

La convergence presque sure se généralise à des valeurs aléatoires à valeurs dans un espace topologique muni de sa structure borélienne. Il est même possible de généraliser cette notion de convergence à des fonctions mesurables sur un espace mesuré, on parle alors de convergence presque partout.

1.1.2 Convergence en probabilité

Soit $(X_n)_n$ avec n $\in \mathbb{N}$ une suite de variables aléatoires et $\mathbf{X}$ une variable aléatoire. on dit que $(X_n)_n$ converge en probabilité vers X lorsque :

$$\text{ssi } \forall \varepsilon > 0, \lim_{n\to+\infty} P(\|X_n - X\| \geq \varepsilon) = 0$$

Et on le note comme suit :

$$X_n \to_p X \sim P$$

Un estimateur θ_n de θ est faiblement consistant de θ si et seulement si la suite de var $(\theta_n)n \in \mathbb{N}^*$ converge en probabilité vers θ.

i.e. son biais tend vers 0 quand n tend vers l'infini.

1.1.3 Convergence en loi

$X_1, \ldots, X_n$ v.a.r. définies sur Ω. On Dit que (X_n) tend vers $X \sim P$ en distribution, écrit $X_n \to_d X$, ssi pour tout $x \in \mathbb{R}$ tel que F_x la fonction de répartition de X soit continue en x, alors $\lim_{n\to+\infty} F_{n,X}(x) = F_X(x)$ où $F_{n,X}$ est la fonction de répartition de X_n.

Aussi, elle diffère des 2 lois précédentes, car on ne regarde pas événement par événement, mais on regarde directement les fonctions de répartitions.

Si $X_n \to_{a.e.} X \sim P$ alors $X_n \to_p X \sim P$.
Si $X_n \to_p X \sim P$ alors $X_n \to_d X \sim P$.

La convergence presque sûre est plus forte que la convergence en probabilité. Cela veut dire que s'il y a convergence presque sûre, alors il y a convergence en probabilité (mais la réciproque n'est pas vraie). Il y a la même relation entre convergence en probabilité et convergence en loi.

1.2 Applications des modes de convergence

1.2.1 Loi forte des grands nombres

$X_1, X_2, \ldots$ variables aléatoires i.i.d.

(The SLLN). Une condition nécessaire et suffisante pour l'existence d'une constante c telle que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c$$

est que $\mathbb{E}[X_1] < \infty$, dans ce cas $c = \mathbb{E}[X_1]$.
On peut réécrire cette condition comme $X1$ a une moyenne ou $X1$ a une espérance $< \infty$.

Remarque :
Il existe aussi la loi faible des grands nombres qui repose sur une convergence en probabilité.
Dans certains cas de figure, la loi faible s'applique, mais pas la forte.
La moyenne empirique est un estimateur fortement convergent de l'espérance selon cette loi.

1.2.2 Théorème de la limite centrale

(Multivariée) Soit $X_1, \ldots, X_n$ i.i.d. des k-vecteurs aléatoires avec $\Sigma = \mathrm{Var}(X_1)$ fini.
Alors

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \to_d N_k(0, \Sigma).$$

où en multipliant par $\frac{n}{n}$ on obtient $\sqrt{n}(\hat{\mu}_n - \mu)$ et où $\mathbb{E}[X_1]$ peut-être réécrit μ.

Pour n grand, $\hat{\mu}_n - \mu \sim$ approx $\mathcal{N}\left(0, \frac{\Sigma}{n}\right)$
Pour n grand, $\hat{\mu}_n \sim$ approx $\mathcal{N}\left(\mu, \frac{\Sigma}{n}\right)$

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \\
&= \frac{n}{n} * \frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \\
&= \frac{\sqrt{n}\sqrt{n}}{n\sqrt{n}}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \\
&= \sqrt{n} * \frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_1]) \\
&= \sqrt{n}(\hat{\mu}_n - \mu)
\end{aligned}$$

1.2.3 Transformations continues

Théorème 1.10.
Soit $X, X_1, X_2, \ldots$ des k-vecteurs aléatoires définis sur un espace de probabilité
et g une fonction de $\left(\mathcal{R}^k\right)$ en $\left(\mathcal{R}^l\right)$.
Supposons que g est continue.
Alors
(i) $X_n \to_{\text{a.s.}} X$ implique $g(X_n) \to_{\text{a.s.}} g(X)$;
(ii) $X_n \to_p X$ implique $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implique $g(X_n) \to_d g(X)$.

Exemple pour (i) :
$X_1, \ldots, X_n$ i.i.d. sur P avec $\mu = \mathbb{E}[P]$.
$\hat{\mu}_n(X_1, \ldots, X_n) \to_{a.e.} \mu$
$g : x \mapsto x^2$
$g(\hat{\mu}_n) \to_{a.s.} g(\mu)$
$\hat{\mu}_n^2 \to_{a.s.} \mu^2$

1.2.4 Théorème de Slutsky

Théorème 1.11 (Théorème de Slutsky).
Soit $X, X_1, X_2, \ldots, Y_1, Y_2, \ldots$ des variables aléatoires sur un espace de probabilité.
Supposons que $X_n \to_d X$ et $Y_n \to_p c$, où c est une constante réelle.
Alors
(i) $X_n + Y_n \to_d X + c$

(ii) $Y_n X_n \to_d cX$
(iii) $X_n/Y_n \to_d X/c$ si $c \neq 0$.

1.2.5 Méthode delta

La méthode delta nous permet d'étudier les fluctuations de $g(X_n)$ autour de $g(c)$.

Soit $X_1, X_2, \ldots, Y$ des vecteurs aléatoires de dimension k tels que :

$$a_n(X_n - c) \to_{d.} Y$$

pour c appartenant à $\mathbb{R}^k$ et (a_n) une suite de nombres positifs tendant vers $+\infty$.

Alors pour toute fonction $g : \mathbb{R}^k \to \mathbb{R}$ différentiable en c, on a :

$$a_n[g(X_n) - g(c)] \to_{d.} [\nabla g(c)]^T Y$$

2 Optimisation

2.1 Analyse d'algorithmes d'optimisation

2.1.1 Descente de gradient à pas fixe

$$Input : x_0 \in \mathbb{R},\ f : \mathbb{R} \to \mathbb{R}$$
$$iteration : x_{k+1} = x_k + \gamma \nabla f(x_k)$$

2.1.2 Descente de gradient : 'Backtracking' avec la règle d'Armijo

Input: $x_0 \in \mathbf{R}^n, f : \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1.$

$$\text{Iteration : } \ x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$
$$\alpha_k = \beta^{m_k} s$$

m_k plus petit entier positif tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma \beta^{m_k} s \nabla f(x_k)^T \nabla f(x_k)$$

2.1.3 Présence de contraintes : Descente de gradient projété

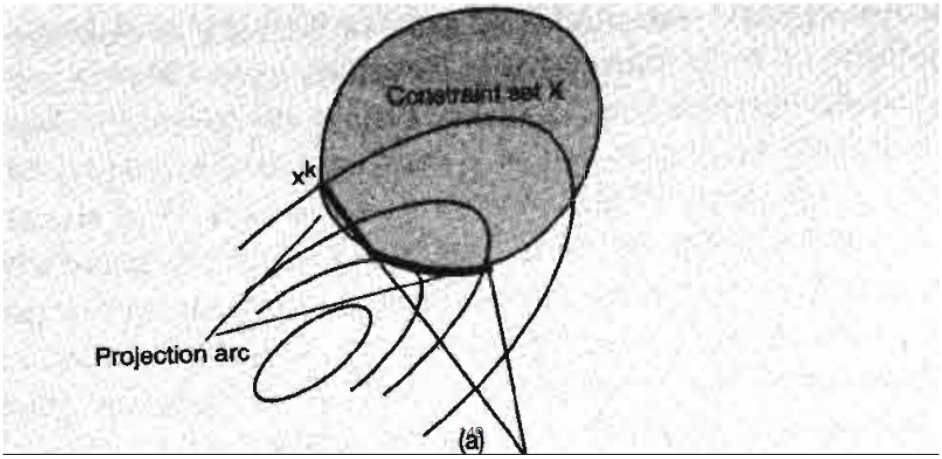
'Backtracking' avec la règle d'Armijo le long de l'arc de projection

Input: $x_0 \in \mathbf{R}^n, f : U \subset \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1.$
U convexe, fermé, non-vide

Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

$p_k(r) = [x_k - r \nabla f(x_k)]_U$ et m_k plus petit entier m tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma \nabla f(x_k)^T (x_k - x_{k+1})$$



2.1.4 Optimisation stochastique

Contexte : minimisation du risque empirique pour une fonction de coût "séparable par point de donnée"

$$R_n(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Descente de gradient stochastique

$$w_1 \in \mathbb{R}^d \text{ given}$$

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k)$$

i_k is chosen *randomly* from $\{1, \ldots, n\}$ and α_k is a positive stepsize

Exemple de garantie de convergence

(cf. Bottou, Curtis et Nocedal (2018) Optimisation Methods for Large-Scale Machine Learning)

Si

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$$

Assumption 4.1 (**Lipschitz-continuous objective gradients**). *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and the gradient function of F, namely, $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$, is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leq L\|w - \overline{w}\|_2 \quad \textit{for all} \quad \{w, \overline{w}\} \subset \mathbb{R}^d.$$

plus des conditions de régularité pas très contraignantes

Alors

$$\liminf_{k \to \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0$$

152

3 Normes de vecteurs

A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a *vector norm* if it has the following properties:

1. $\|\mathbf{x}\| \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar $\alpha \in \mathbb{R}$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Si Q est une matrice orthogonale, $\|Qx\|_2 = \|x\|_2$

155

4 Normes de matrices

A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that has the following properties:

- $\|A\| \geq 0$ for any $A \in \mathbb{R}^{m\times n}$, and $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$ for any $m \times n$ matrix A and scalar α
- $\|A + B\| \leq \|A\| + \|B\|$ for any $m \times n$ matrices A and B

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \qquad \|A\|_2 = \sigma_1$$

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}. \qquad \|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$