

MACS205 : Méthode de Monte-Carlo

1 Introduction

But du cours : étudier des méthodes aléatoires d'approximation d'intégrales.

Soit (S, \mathcal{S}, μ) un espace mesuré où μ est une mesure positive. Soit $\varphi: S \rightarrow \mathbf{R}$ une fonction intégrable. On cherche à approcher $I(\varphi) = \int \varphi d\mu$.

Plusieurs cas de figure en pratique :

- φ est une fonction continue avec une expression analytique et on arrive à calculer son intégrale,
- l'intégrale de φ est incalculable. Exemples : Gaussienne ou indicatrice d'ensemble S où l'on ne connaît pas de forme analytique.

Les méthodes considérées sont de la forme suivante :

1. choisir/tirer des points X_1, \dots, X_n sur S ,
2. évaluer $\varphi(X_1), \dots, \varphi(X_n)$,
3. trouver une transformation de $(X_1, \varphi(X_1)), \dots, (X_n, \varphi(X_n))$ qui approche $I(\varphi)$.

2 La méthode de Monte-Carlo

$$I(\varphi) = \int \varphi dN = \mathbf{E}_\mu(\varphi)$$

D'après la LFGN, si X_1, \dots, X_n sont i.i.d. de loi μ tel que $\mathbf{E}_\mu |vf| < \infty$ alors $\frac{1}{n} \sum_i \varphi(X_i) \xrightarrow{\text{p.s.}} \mathbf{E}_\mu(\varphi(X_1))$.

Algorithme 1 : Monte-Carlo

Générer X_1, \dots, X_n de façon indépendante sous μ ;

Calculer $\varphi(X_1), \dots, \varphi(X_n)$;

Sorties : $\hat{I}_n(\varphi) = \frac{1}{n} \sum_i \varphi(X_i)$

Prop. Si $\int |\varphi| d\mu < \infty$, $\hat{I}_n(\varphi)$ est non-biaisée et consistante. Si $\int |\varphi|^2 d\mu < \infty$, $\text{Var}(\hat{I}_n(\varphi)) = \frac{1}{n} \text{Var}(\varphi(X_1)) = \frac{1}{n} \sigma^2$ et $\sqrt{n}(\hat{I}_n(\varphi) - I(\varphi)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

On estime σ^2 par $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (\varphi(X_i) - \hat{I}_n(\varphi))^2$.

Prop. Si $\int |\varphi|^2 d\mu < \infty$ alors $\hat{\sigma}^2$ est sans biais et fortement consistant (par la LFGN).

...

Th (Inégalité de Hoeffding). Soit (X_1, \dots, X_n) des v.a i.i.d telles que $\forall i \in [1; n], a \leq X_i \leq b$ p.s. Alors

$$\mathbf{P} \left(\left| \sum_{i=1}^n (X_i - \mathbf{E}(X_i)) \right| > \varepsilon \right) \leq 2e^{-\frac{3\varepsilon^2}{n(b-a)^2}}.$$

...

Concentration

...

Déterministe vs aléatoire

On se place dans le cadre de l'approximation de $\int_{[0;1]^d} \varphi(x) dx$ où $\varphi: [0;1]^d \rightarrow \mathbf{R}$.

Méthode déterministe des sommes de Riemann On se donne n^d points équidistants $(\frac{i_1}{n}, \dots, \frac{i_d}{n})$ où $(i_1, \dots, i_d) \in [1; n]^d$. La méthode des sommes de Riemann est

$$I_n(\varphi) = \frac{1}{n^d} \sum_{(i_1, \dots, i_d) \in [1; n]^d} \varphi\left(\frac{i_1}{n}, \dots, \frac{i_d}{n}\right).$$

Prop. Si $\varphi: [0;1]^d \rightarrow \mathbf{R}$ est L -lipschitzienne alors $|I_n(\varphi) - I(\varphi)| \leq L \frac{\sqrt{d}}{n}$.

...

3 Méthode des variables de contrôle

Présentation

Le contexte est maintenant comme Monte-Carlo avec une variable observée en plus :

- on observe $((X_1, \varphi(X_1), Y_1), \dots, (X_n, \varphi(X_n), Y_n))$ i.i.d,
- on cherche $\mathbf{E}[\varphi(X_1)]$,
- on connaît $\mathbf{E}[Y_1]$.

On peut calculer $\frac{1}{n} \sum_{i=1}^n [\varphi(X_i) - (Y_i - \mathbf{E}[Y_i])]$. C'est un estimateur sans biais de variance

$$\frac{1}{n} \text{Var}(\varphi(X_1) - (Y_1 - \mathbf{E}[Y_1])) = \frac{1}{n} \mathbf{E}[(\varphi(X_1) - (Y_1 - \mathbf{E}[Y_1]))^2] - \frac{1}{n} \mathbf{E}[\varphi(X_1)]^2.$$

Pour un tel estimateur, le but est de réduire ce risque L^2 au maximum en choisissant bien Y_1 . Pour un tel estimateur, on obtient facilement les mêmes résultats que pour Monte-Carlo : forte consistance, normalité asymptotique et estimation consistante de la variance.

Pour faciliter la notation on supposera maintenant $\mathbf{E}Y_1 = 0$.

Rem.

- $Y = 0 \rightarrow$ Monte-Carlo,
- $Y = \frac{-\varphi \circ L(X) - \varphi(X)}{2} \rightarrow$ variables antithétiques.

Rem. La méthode des variables de contrôle (VC) est plus performante que MC si $\text{Var}(\varphi(X_1) - Y_1) \leq \text{Var}(\varphi(X_1))$
 $(\frac{1}{n} \sum \varphi(X_i) + \frac{\varphi \circ L(X_i) - \varphi(X_i)}{2} = \frac{1}{2n} \sum \frac{\varphi(X_i) + \varphi \circ L(X_i)}{2})$.

Afin de prévenir d'une mauvaise variable de contrôle, on définit l'estimateur $\forall \beta \in \mathbf{R}, \hat{\mu}_n(\beta) = \frac{1}{n} \sum_{i=1}^n (\varphi(X_i) - \beta Y_i)$, à utiliser si $\text{Var}(\varphi(X_1) - \beta Y_1) \leq \text{Var}(\varphi(X_1))$.

$\rightarrow \beta^* = \arg \min_{\beta} \text{Var}(\varphi(X_1) - \beta Y_1), \min_{\beta} \text{Var}(\varphi - \beta Y) \leq \text{Var}(\varphi)$.

Soit f_1, \dots, f_m une collection de fonctions dont on connaît les intégrales. Supposons $\forall L \in [1; m], \int f_L d\lambda = 0$. Alors VC donne $\frac{1}{n} \sum_{i=1}^n [\varphi(u_i) - \sum_{j=1}^m \beta_j f_j(u_i)]$.

Ex. (f_L) polynômes, (f_L) base de Fourier ou (f_L) indicatrices.

Propriétés asymptotiques

Soient $((X_i, Y_i))_i$ une suite de v.a i.i.d à valeurs dans $S \times \mathbf{R}^m$. On définit l'estimateur de $\mathbf{E}[\varphi(X_1)]$ par $\forall \beta \in \mathbf{R}^m, \hat{\mu}_n(\beta) = \frac{1}{n} \sum_{i=1}^n (\varphi(X_i) - \beta^T Y_i)$.

Comme dans l'intro, on suppose $\mathbf{E}Y_1 = \begin{pmatrix} \mathbf{E}[Y_{1,1}] \\ \vdots \\ \mathbf{E}[Y_{1,m}] \end{pmatrix} = 0$.

$\{\mu_n(\beta), \beta \in \mathbf{R}^m\}$ est une collection d'estimateurs sans biais. Trouvons l'élément de variance minimale :

$$\begin{aligned} \beta^* &= \arg \min_{\beta} \frac{1}{n} \text{Var}(\varphi - \beta^T T) \\ &= \arg \min_{\beta} \text{Var}(\varphi - \beta^T T) \\ &= \arg \min_{\beta} \mathbf{E}[(\varphi - \beta^T Y)^2] - \mathbf{E}[\varphi]^2 \\ &= \arg \min_{\beta} \mathbf{E}[(\varphi - \beta^T Y)^2] \end{aligned}$$

Si $\mathbf{E}[Y_1 Y_1^T]$ est inversible, les équations normales / du premier ordre admettent une unique solution :

$$\beta^* = \mathbf{E}[Y_1 Y_1^T]^{-1} \mathbf{E}[Y_1 \varphi(X_1)]$$

Il faut utiliser $\hat{\mu}_n(\beta^*)$, mais β^* est inconnue.

Idée : estimer β^* sur les données $\rightarrow \hat{\beta}$, et utiliser $\hat{\mu}_n(\hat{\beta})$, qui a la même variance asymptotique que $\hat{\mu}_n(\beta^*)$.

Si $\frac{1}{n} \sum_{i=1}^n Y_i Y_i^T$ est inversible :

$$\hat{\beta} = \arg \min_{\beta \in \mathbf{R}^m} \frac{1}{n} \sum_{i=1}^n ([\varphi(X_i) - \beta^T Y_i] - \hat{\mu}_n(\beta))^2$$

estimateur classique de la covariance

Ce choix ne va pas entraîner de changement à l'asymptotique mais pratique il procure de meilleurs performances. Donc $\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n ((\varphi(X_i) - \bar{\varphi}) - \beta^T (Y_i - \bar{Y}))^2$.

Notons

$$Z_{n,m} = \begin{pmatrix} Y_{11} - \bar{Y}_1 & \cdots & Y_{1m} - \bar{Y}_m \\ \vdots & & \vdots \\ Y_{n1} - \bar{Y}_1 & \cdots & Y_{nm} - \bar{Y}_m \end{pmatrix} \in \mathbf{R}^{n \times m}, \quad Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im} \end{pmatrix} \in \mathbf{R}^m$$

(Y_i est la covariable du problème de régression). On a $\bar{Y}_k = \frac{1}{n} \sum_{i=1}^n Y_{ik}$.

Notons également $\Psi_n = \begin{pmatrix} \varphi(X_1) - \bar{\varphi} \\ \vdots \\ \varphi(X_n) - \bar{\varphi} \end{pmatrix}$. Alors $\hat{\beta} = \arg \min_{\beta} \|\Psi_n - Z_{n,m}\beta\|^2$.

Le théorème de projection nous donne une unique solution qui, si $Z_{n,m}^\top Z_{n,m}$ est inversible, vérifie :

$$(Z_{n,m}^\top Z_{n,m})\beta = Z_{n,m}^\top \Psi_n$$

$$\hat{\beta} = (Z_{n,m}^\top Z_{n,m})^{-1} Z_{n,m}^\top \Psi_n$$

Prop (asymptotique de $\hat{\mu}_n(\beta)$). Supposons que $\mathbf{E}|\varphi(X_1)| < \infty$, $\forall k \in \llbracket 1; m \rrbracket$, $\mathbf{E}|\varphi(X_1)Y_{1k}| < \infty$ et $\mathbf{E}[Y_1 Y_1^\top]$ existe et est inversible. Alors $\hat{\mu}_n(\hat{\beta}) \xrightarrow{p.s.} \mathbf{E}[\varphi(X_1)]$. Si de plus $\mathbf{E}|\varphi(X_1)|^2 < \infty$, alors $\sqrt{n}(\hat{\mu}_n(\hat{\beta}) - \mathbf{E}[\varphi(X_1)]) \rightarrow \mathcal{N}(0, \sigma_m^2)$ avec $\sigma_m^2 = \text{Var}(\varphi(X_1) - \beta^{*\top} Y_1)$.

Rem. $\hat{\beta}$ n'a pas d'effet en l'asymptotique (c'est comme si on connaissait β^*).

Rem. D'autres estimateurs de β^* peuvent être légitimes sous condition d'inversibilité :

$$\hat{\beta} = \begin{cases} \left(\frac{1}{n} \sum Y_i Y_i^\top \right)^{-1} \frac{1}{n} \sum Y_i \varphi(X_i) \\ \left(\frac{1}{n} \sum (Y_i - \bar{Y})(Y_i - \bar{Y})^\top \right)^{-1} \frac{1}{n} \sum Y_i \varphi(X_i) \end{cases}$$

Lorsque $\mathbf{E}[Y_1 Y_1^\top]$ est connu, $\hat{\beta} = \mathbf{E}[Y_1 Y_1^\top]^{-1} \frac{1}{n} \sum_{i=1}^n (Y_i(\varphi(X_i) - \bar{\varphi}))$.

...

Temps de calcul

Soit F une c.d.f sur \mathbf{R} et $\varphi: \mathbf{R} \rightarrow \mathbf{R}$. On veut calculer $\mathbf{E}_F[\varphi]$.

Le nombre d'échantillons n'est pas fixé par le problème initial. Il est donc à déterminer par rapport à la précision souhaitée et le temps de calcul dont on dispose.

Données massives \rightarrow la problématique du temps de calcul est redevenue essentielle aujourd'hui.

Mesure du temps \rightarrow par simulation, en terme d'opérations élémentaires.

Règles du temps de calcul (peuvent changer selon le problème) :

- générer $X_1 \rightarrow 1$ opération élémentaire,
- générer $Y_{1,k}$ pour chaque $k \rightarrow 1$ opération élémentaire,
- évaluer $\varphi(X_1) \rightarrow 1$ opération élémentaire.

MC	nombre d'opérations élémentaires
X_1, \dots, X_n	n
$\varphi(X_1), \dots, \varphi(X_n)$	n
$\frac{1}{n} \sum_{i=1}^n \varphi(X_i)$	$\sim n$
	$O(n)$
VC	nombre d'opérations élémentaires
X_1, \dots, X_n	n
$\varphi(X_1), \dots, \varphi(X_n)$	n
Y_1, \dots, Y_n	mn
$\frac{1}{n} \sum_{i=1}^n (\varphi(X_i) - \beta^\top Y_i)$	mn
	$O(mn)$
$\hat{\beta}$	nombre d'opérations élémentaires
...	...
...	...
...
	$O(m^3 + m^2 n)$

4 Échantillonnage d'importance

Présentation

On se place dans le cadre de l'approximation de $I(\varphi) = \int \varphi(x) dx$, où φ est intégrable.

Def. Si $g: \mathbf{R}^d \rightarrow \mathbf{R}$, on définit son support comme l'ensemble fermé

$$S_g = \overline{\{x \in \mathbf{R}^d \mid g(x) \neq 0\}}$$

L'échantillonnage d'importance se base sur la formule suivante : pour toute densité f telle que $S_f \supset S_\varphi$,

$$\int \varphi(x) dx = \int_{S_\varphi} \varphi(x) dx = \int_{S_f} \varphi(x) dx = \int_{S_f} \frac{\varphi(x)}{f(x)} f(x) dx = \mathbf{E}_{X \sim f} \left[\frac{\varphi(X)}{f(X)} \right]$$

L'échantillonnage d'importance "naïf" consiste à :

- générer $X_1, \dots, X_n \sim f$ i.i.d
- faire du MC : $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \frac{\varphi(X_i)}{f(X_i)}$

Prop. Si $\int |\varphi| < \infty$, $\hat{\mu}_n$ est sans biais, $\hat{\mu}_n \xrightarrow{\text{p.s.}} \int \varphi$. Si $\int \frac{\varphi^2}{f} < \infty$, $\sqrt{n}(\hat{\mu}_n - \int \varphi) \rightarrow \mathcal{N}(0, \sigma^2)$, $\sigma^2 = \text{Var}\left(\frac{\varphi}{f}\right)$. On a également $\frac{\sqrt{n}}{\hat{\sigma}_n}(\hat{\mu}_n - \int \varphi) \rightarrow \mathcal{N}(0, 1)$, avec $\hat{\sigma}_n = \frac{1}{n} \sum \left(\frac{\varphi(X_i)}{f(X_i)} - \hat{\mu}_n \right)^2$.

Cette méthode est naïve car f n'est pas choisie par rapport à φ .

Ex. Soit φ la densité d'une gaussienne centrée réduite et $f \sim \mathcal{N}(\theta, 1)$. Ici $I(\varphi) = 1$,

$$\sigma^2 + 1 = \int \frac{\varphi^2}{f^2} f d\lambda = \int \frac{\varphi^2}{f} d\lambda = \frac{1}{\sqrt{2\pi}} \int e^{-x^2 + (x-\theta)^2/2} dx = e^{\theta^2} \underbrace{\frac{1}{\sqrt{2\pi}} \int e^{-(x+\theta)^2} dx}_{=1}$$

Deux cas de figure sont possibles :

- $\theta = 0$, d'où $\sigma^2 = 0$. Plus généralement, si φ n'est pas nécessairement une densité, alors le choix $f \propto \varphi$ (mais positif, e.g $f = \frac{\varphi}{\int \varphi}$) est optimal. Ce choix dépend de la solution à notre problème de départ, donc il est impossible à réaliser.
- $\theta \gg 1$, d'où $\sigma^2 \gg 1$.

La question est : comment choisir f en pratique ?

Rem. f est appelée la distribution d'échantillonnage, ou bien l'échantillonneur.

Rem. Si le but est d'estimer une espérance par rapport à g , alors prendre $\varphi \cdot g$ à la place de φ .

Par ailleurs il existe deux méthodes de réduction de la variance :

- variable de contrôle : approcher φ dans une certaine base \rightarrow pas de choix d'échantillonneur.
- changer la mesure d'échantillonnage.

Dans les deux cas on s'adapte à φ .

Réduction de la variance

On caractérise ici l'échantillonneur optimal.

On fait la remarque suivante :

$$\text{Var}(\hat{\mu}_n) = 0 \iff \int \left(\frac{\varphi}{f} - I(\varphi) \right)^2 f d\lambda = 0 \iff \frac{\varphi}{f} = I(\varphi) \text{ p.p.}$$

Si φ change de signe sur des ensembles de mesures non-nulles, alors prendre $\frac{\varphi}{f} = I(\varphi)$ p.p est impossible car est une densité, donc positive. Obtenir $\sigma^2 = 0$ est alors impossible.

Si φ est de signe constant alors $f = \frac{|\varphi|}{\int |\varphi|}$ donne une variance nulle. En fait, ce $f = \frac{|\varphi|}{\int |\varphi|}$ est optimal dans tous les cas.

Th. Parmi les densités f tq $\int \frac{\varphi^2}{f} d\lambda < \infty$, le minimiseur de $\text{Var}\left(\frac{\varphi}{f}\right) = \int \left(\frac{\varphi}{f} - I(\varphi) \right)^2 f d\lambda$ est $f^* = \frac{|\varphi|}{\int |\varphi|}$ et

$$\sigma^{*2} = \text{Var}_{X \sim f^*} \left(\frac{\varphi(X)}{f^*(X)} \right) = \int \left(\frac{\varphi}{f^*} - I(\varphi) \right)^2 d\lambda$$

Échantillonnage d'importance paramétrique

En pratique et en dimension 1 ou 2, on peut représenter $|\varphi|$ et en déduire un f "proche" (visuellement) de $|\varphi|$. On se donne une famille de lois par rapport auxquelles on sait générer des v.a :

$$\mathcal{P} = \{f_\theta, \theta \in \Theta\}$$

où $\Theta \subset \mathbf{R}^q, q \geq 1$, ce qui fait de \mathcal{P} une famille paramétrique.

On aimerait calculer $\theta^* \in \arg \min_{\theta \in \Theta} \text{Var}_{X \sim f_\theta} \left(\frac{\varphi(X)}{f_\theta(X)} \right)$ mais $\theta \mapsto \text{Var}_{X \sim f_\theta} \left(\frac{\varphi}{f_\theta} \right)$ est inconnue.

Donc on cherche à estimer par simulation cette variance.

Soit f_0 l'échantillon initial.

a) Générer $Z_1, \dots, Z_{n_1} \sim f_0$ i.i.d

$$\hat{\mu}_{f_0} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi(Z_i)}{f_0(Z_i)}, \quad \hat{\sigma}_\theta = \frac{1}{n} \sum_{i=1}^n \frac{\varphi(Z_i)^2}{f_\theta(Z_i)f_0(Z_i)}, \quad \mathbb{E} \left[\frac{\varphi^2}{f_\theta f_0} \right] = \int \frac{\varphi^2}{f_\theta} d\lambda$$

b) $\hat{\theta}_n = \arg \min_{\theta \in \Theta} \hat{\sigma}_\theta^2$

c) $X_1, \dots, X_n \sim f_{\hat{\theta}_n}$ i.i.d, $\hat{\mu}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n \frac{\varphi(X_i)}{f_{\hat{\theta}_n}(X_i)}$

