Estimation séquentielle de quantiles conditionnels dans les codes stochastiques

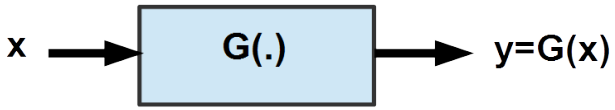
T. Labopin-Richard F. Gamboa A. Garivier

Institut de mathématiques de Toulouse

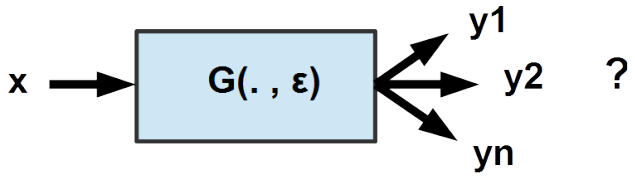
29 Septembre 2015

Plan

Qu'est-ce qu'un code stochastique ?

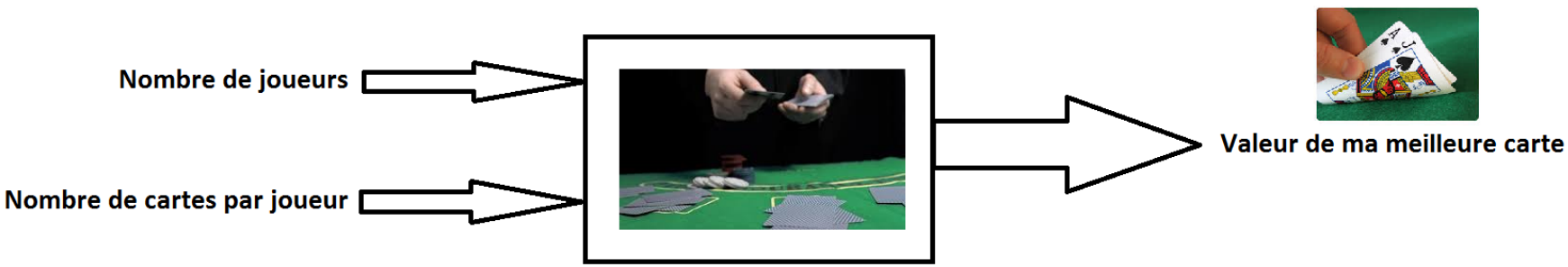


Numerical code : Y=G(X)
G(x) is a real number

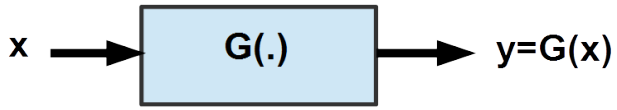


Stochastic code : Y=G(X, ε)
G(x, ε) is a random variable

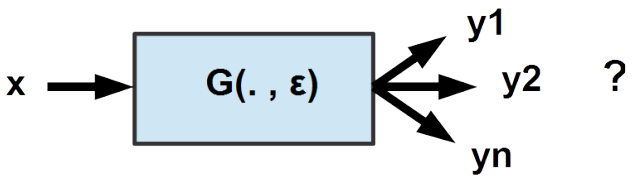
Un exemple simple



Qu'est-ce qu'un code stochastique ?



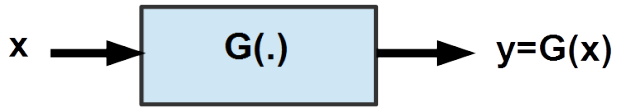
Numerical code : Y=G(X)
G(x) is a real number



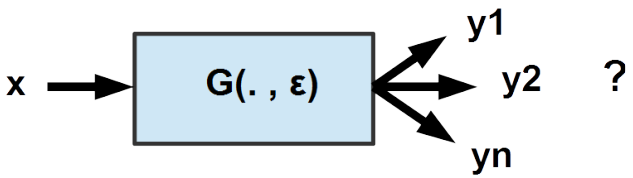
Stochastic code : Y=G(X, ε)
G(x, ε) is a random variable

Notations : $X \in \mathbb{R}^d$ vecteur des entrées. α fixé le niveau du quantile à estimer.

Qu'est-ce qu'un code stochastique ?



Numerical code : Y=G(X)
G(x) is a real number



Stochastic code : Y=G(X, ε)
G(x, ε) is a random variable

Notations : $X \in \mathbb{R}^d$ vecteur des entrées. α fixé le niveau du quantile à estimer.

But : Estimer le quantile de la loi $\mathcal{L}\left(G(X,\epsilon)|X=x\right)$ en utilisant le moins possible d'appels au code.

Pour estimer un quantile d'une loi Z, on peut...

1) Construire un échantillon $(Z_1, \ldots, Z_n)$ de la loi Z.
2) Utiliser un estimateur du quantile, par exemple :
   a) Le quantile empirique $Z_{(\lfloor n\alpha \rfloor + 1)}$.
   b) L'algorithme stochastique suivant :

$$\begin{cases} \theta_0 \in \mathbb{R} \\ \theta_{n+1} = \theta_n - \dfrac{1}{n^\gamma} \left(\mathbf{1}_{Z_{n+1} \leq \theta_n} - \alpha \right). \end{cases}$$

Application à notre problème

On veut estimer le quantile de la loi $\mathcal{L}\left(G(X,\epsilon)|X=x\right)$, on pourrait donc :

1) Fournir plusieurs fois l'entrée x au code stochastique, pour construire un échantillon de notre loi.
2) Utiliser un des deux estimateurs précédents.

Problème

On veut estimer le quantile conditionnel pour plusieurs entrées x à la fois.
Chaque appel au code coûte cher, on ne peut pas se permettre de construire un échantillon de la loi conditionnelle pour chacun de ces x.

Stratégie adoptée

1) On fixe un budget N.
2) On tire un échantillon d'entrées $(X_1, \ldots, X_N)$.
3) On observe les réponses correspondantes $(Y_1, \ldots, Y_N)$.
4) On applique un algorithme qui permettra, pour tout x et n'utilisant que le passé, de calculer un estimateur du quantile conditionnel.

Un algorithme solution

Notre algorithme

$$\begin{cases} \theta_0(x) \in \mathbb{R} \\ \theta_{n+1}(x) = \theta_n(x) - \dfrac{1}{n^\gamma}\left(\mathbf{1}_{Y_{n+1} \leq \theta_n(x)} - \alpha\right)\mathbf{1}_{X_{n+1} \in kNN_n(x)} \end{cases}$$

où $k_n = \lfloor n^\beta \rfloor$.

Un algorithme solution

Notre algorithme

$$\begin{cases} \theta_0(x) \in \mathbb{R} \\ \theta_{n+1}(x) = \theta_n(x) - \dfrac{1}{n^\gamma}\left(\mathbf{1}_{Y_{n+1} \leq \theta_n(x)} - \alpha\right)\mathbf{1}_{X_{n+1} \in kNN_n(x)} \end{cases}$$

où $k_n = \lfloor n^\beta \rfloor$.

- Quels sont les paramètres optimaux γ et β ?
- Sous quelles hypothèses l'algorithme est-il convergent ?
- Peut-on montrer des résultats non-asymptotiques ?

Hypothèse de continuité

On introduit les notations suivantes :

1) $B_n(x)$ est le plus petit segment contenant les k plus proches voisins de x.
2) $F_{Y^{B_n(x)}}$ est la fonction de répartition de la loi $\mathcal{L}(Y|X \in B_n(x))$.
3) F_{Y^x} est la fonction de répartition de la loi de $\mathcal{L}(Y|X = x)$.

Hypothèse de continuité

On introduit les notations suivantes :

1) $B_n(x)$ est le plus petit segment contenant les k plus proches voisins de x.
2) $F_{Y^{B_n(x)}}$ est la fonction de répartition de la loi $\mathcal{L}(Y|X \in B_n(x))$.
3) F_{Y^x} est la fonction de répartition de la loi de $\mathcal{L}(Y|X = x)$.

Hypothèse A1 Il existe une constante de Lipschitz M telle que $\forall n$, $\forall x \in \text{Supp}(X)$, $\forall t \in \mathbb{R}$:

$$|F_{Y^{B_n(x)}}(t) - F_{Y^x}(t)| \leq M \max_{a \in B_n(x)} ||x - a|| = M||X - x||_{(k_n,n)}$$

Hypothèses techniques

Hypothèse A2 La loi des entrées est à densité et cette fonction de densité est minorée sur son support par une constante $C_{inputs} > 0$.

$\Rightarrow$ *Permet de gérer des quantités comme* $\mathbb{E}(||X - x||_{(k_n,n)})$ *ou* $\mathbb{P}(X \in kNN_n(x))$.

Hypothèses techniques

Hypothèse A2 La loi des entrées est à densité et cette fonction de densité est minorée sur son support par une constante $C_{inputs} > 0$.

$\Rightarrow$ *Permet de gérer des quantités comme* $\mathbb{E}(||X - x||_{(k_n,n)})$ *ou* $\mathbb{P}(X \in kNN_n(x))$.

Hypothèse A3 La fonction code g est à valeurs dans le compact $[a, b]$.

$\Rightarrow$ $\forall x$, $\theta_n(x)$ est bornée uniformément en ω. On appelle R la borne uniforme de $(\theta_n - \theta^*)^2$.

Hypothèses techniques

Hypothèse A2 La loi des entrées est à densité et cette fonction de densité est minorée sur son support par une constante $C_{inputs} > 0$.

$\Rightarrow$ *Permet de gérer des quantités comme* $\mathbb{E}(||X - x||_{(k_n,n)})$ *ou* $\mathbb{P}(X \in kNN_n(x))$.

Hypothèse A3 La fonction code g est à valeurs dans le compact $[a, b]$.

$\Rightarrow$ $\forall x$, $\theta_n(x)$ est bornée uniformément en ω. On appelle R la borne uniforme de $(\theta_n - \theta^*)^2$.

Hypothèse A4 Pour tout x, la loi $g(X, \epsilon)|X = x$ est à densité minorée par une constante $D(x) > 0$.

$\Rightarrow$ *Il existe une constante* $D_{code}(x)$ *telle que :*

$$\forall \theta_n(x),\ [F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))]\,[\theta_n(x) - \theta^*(x)] \geq D_{code}(x)\,[\theta_n(x) - \theta^*(x)]^2.$$

Résultats théoriques

Théorème : convergence presque-sûre

Soit x une entrée fixée. Sous les hypothèses **A1** and **A2**, l'algorithme en x est convergent presque sûrement si et seulement si $\frac{1}{2} < \gamma < \beta < 1$.

Esquisse de preuve

1) On décompose H en un terme de martingale et un reste, en posant :
$h_n(\theta_n) = \mathbb{E}(H(\theta_n, X_{n+1}, Y_{n+1})|\mathcal{F}_n)$ and $\xi_{n+1} = H(\theta_n, X_{n+1}, Y_{n+1}) - h_n(\theta_n)$.

   On a
$$T_n = \theta_n(x) + \sum_{j=1}^{n} \gamma_j h_{j-1}(\theta_{j-1}(x))$$
   martingale bornée dans L^2 donc qui converge presque-sûrement.
2) On montre la convergence presque-sûre de $(\theta_n(x))_n$.
   a) $(\theta_n(x))$ ne diverge pas vers $+\infty$ ou $-\infty$.
   b) $(\theta_n(x))$ converge p.s vers une limite finie.
3) La limite est $\theta^*(x)$ le quantile conditionnel que nous voulons estimer.

Théorème : vitesse de convergence

Soit x une entrée fixée. Sous les hypothèses **A1**, **A2**, **A3** et **A4**, pour tout $0 < \gamma < 1$, $0 < \beta < 1$ and $1 > \epsilon > 1 - \beta$, et pour $n \geq 2^{\frac{1}{\epsilon-(1-\beta)}} := N_0$,

$$\mathbb{E}\left[(\theta_n(x) - \theta^*(x))^2\right] \leq R \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + a_0(x) \exp\left(-2D_{code}(x) \sum_{k=1}^{n} \frac{1}{k^{\gamma+\epsilon}}\right) + \sum_{k=1}^{n} \exp\left(-2D_{code}(x) \sum_{i=k}^{n} \frac{1}{i^{\gamma+\epsilon}}\right) \beta_k$$

avec

$$\beta_n = R \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + 2\sqrt{R} M D(d) \gamma_{n+1} \left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1} + \gamma_{n+1}^2 \frac{k_n}{n+1},$$

$$D(d) = \sqrt[d]{2}\left(1 + \frac{8}{3d} + \frac{1}{\sqrt[d]{C_{input} H(d)}}\right) \text{ et } H(d) = \frac{\pi^{\frac{5}{2}}}{\Gamma\left(\frac{d}{2}\right)+1}.$$

Théorème : paramètres optimaux

Sous les mêmes hypothèses, le risque quadratique décroit plus rapidement lorsque les paramètres sont $\gamma = \frac{1}{1+d}$ and $\beta = \gamma + \eta$ où $\eta > 0$ est le plus petit possible. Avec ces paramètres, nous obtenons, pour $n \geq \max(N_0, N_1)$

$$\mathbb{E}\left[(\theta_n(x) - \theta^*(x))^2\right] \leq \frac{C_1}{n^{\frac{1}{1+d}-\eta'}}$$

où les constantes sont connues explicitement.

Esquisse de preuve

- L'idée de la preuve est d'établir une inégalité du genre

$$a_{n+1}(x) \leq a_n(x)(1-\alpha_n) + \beta_n$$

- On commence donc pas développer le carré de $a_{n+1}(x)$:

$$\begin{aligned}(\theta_{n+1}(x) - \theta^*(x))^2 = (\theta_n(x) - \theta^*(x))^2 + \gamma_{n+1}^2 \left[(1-2\alpha)\mathbf{1}_{Y_{n+1}\leq\theta_n(x)} + \alpha^2\right] \mathbf{1}_{X_{n+1}\in kNN_n(x} \\ - 2\gamma_{n+1}(\theta_n(x) - \theta^*(x)) \left(\mathbf{1}_{Y_{n+1}\leq\theta_n(x)} - \alpha\right) \mathbf{1}_{X_{n+1}\in kNN_n(x)}\end{aligned}$$

- On prend l'espérance conditionnelle et on utilise la formule de Bayes

$$\begin{aligned}\mathbb{E}_n\left(\theta_{n+1}(x) - \theta^*(x))^2\right) \leq \mathbb{E}_n\left((\theta_n(x) - \theta^*(x))^2\right) + \gamma_{n+1}^2 P_n \\ - 2\gamma_{n+1}\left(\theta_n(x) - \theta^*(x)\right) P_n \left[F_{Y^{B_n(x)}}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right]\end{aligned}$$

Esquisse de preuve

- $F_{Y^{B_n(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))$, erreur **type variance**.

  Par **A1**,

$$|F_{Y^{B_n(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))| \leq M \sup\{||y-x||, y \in B_n(x)\} = M||X-x||_{(k_n,n)}$$

  et par **A3**, $|\theta_n(x) - \theta^*(x)| \leq \sqrt{R}$ donc

$$\begin{aligned} &- 2\gamma_{n+1}(\theta_n(x) - \theta^*(x))P_n\left[F_{Y^{B_n^{(x)}}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))\right] \\ &\leq 2\gamma_{n+1}\sqrt{R}MP_n||X-x||_{(k_n,n)} \end{aligned}$$

- $F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*)$, erreur **type biais**.

  Par **A4** nous avons

$$(\theta_n - \theta^*)\left[F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] \geq D_{code}(x)\left[\theta_n(x) - \theta^*(x)\right]^2.$$

  donc

$$\begin{aligned} &- 2\gamma_{n+1}(\theta_n(x) - \theta^*(x))P_n\left[F_{Y^x}(\theta^*(x)) - F_{Y^x}(\theta^*(x))\right] \\ &\leq -2\gamma_{n+1}D_{code}(x)(\theta_n(x) - \theta^*(x))^2 P_n \end{aligned}$$

Esquisse de preuve

- Finalement,

$$\mathbb{E}_n\left(\theta_{n+1}(x)-\theta^*(x))^2\right) \leq (\theta_n(x)-\theta^*(x))^2 - 2\gamma_{n+1}D_{code}(x)(\theta_n(x)-\theta^*(x))^2 P_n + \gamma_{n+1}^2 P_n + 2\gamma_{n+1}M\sqrt{R}||X-x||_{(k_n,n)}P_n.$$

- On prend l'espérance

$$a_{n+1}(x) \leq a_n(x) - 2\gamma_{n+1}D_{code}(x)\mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2 P_n\right] + \gamma_{n+1}^2\mathbb{E}(P_n) + 2\gamma_{n+1}M\sqrt{R}\mathbb{E}(||X-x||_{(k_n,n)}P_n).$$

Esquisse de preuve

- $\mathbb{E}(P_n) = \frac{k_n}{n+1}$ car

$$P_n = \mathbb{P}(||X-x|| \leq ||X-x||_{(k_n,n)}) = F_{||X-x||}(||X-x||_{(k_n,n)}) \sim \beta(k_n, n-k_n+1)$$

- $\mathbb{E}(||X-x||_{(k_n,n)} P_n) \leq D(d) \left(\frac{k_n}{n+1}\right)^{1+\frac{1}{d}}$ car

1) $\mathbb{E}(||X-x||_{(k_n,n)} P_n) = \frac{k_n}{n+1} \mathbb{E}(||X-x||_{(k_n+1,n+1)})$.
2) Inégalités type Bernstein dans

$$\mathbb{E}(||X-x||_{(k_n+1,n+1)}) \leq \int_0^A \mathbb{P}(\mathcal{B}(n+1, q_u) < k_n + 1) du$$

Esquisse de preuve

- $\mathbb{E}(P_n(\theta_n(x) - \theta^*(x))^2) = ?$

Solution : On va utiliser une borne sur cette espérance. Pour cela, on doit considérer $b_n(x) = \mathbb{E}((\theta_n(x) - \theta^*(x))^2 \mathbf{1}_{P_n > \epsilon_n})$ avec $\epsilon_n = \frac{1}{n^\epsilon}$.

On peut faire les mêmes raisonnements sur cette quantité pour trouver :

$$b_{n+1}(x) \leq b_n(x)(1 - \alpha_n) + \beta_n$$

Pour se ramener au même type d'inégalité sur $a_n(x)$, on gère le terme $\overline{b_n(x)} := \mathbb{E}((\theta_n(x) - \theta^*(x))^2 \mathbf{1}_{P_n \leq \epsilon_n})$ avec des inégalités de concentration.

Modèle en dimension 1-Convergence presque-sûre

Nous avons testé deux modèles pour $X \sim \mathcal{U}([-1,1])$, $\epsilon \sim \mathcal{U}([-0.5,0.5])$ et $x = 0$:

$$g(X,\epsilon) = X^2 + \epsilon \text{ et } g(X,\epsilon) = |X| + \epsilon$$

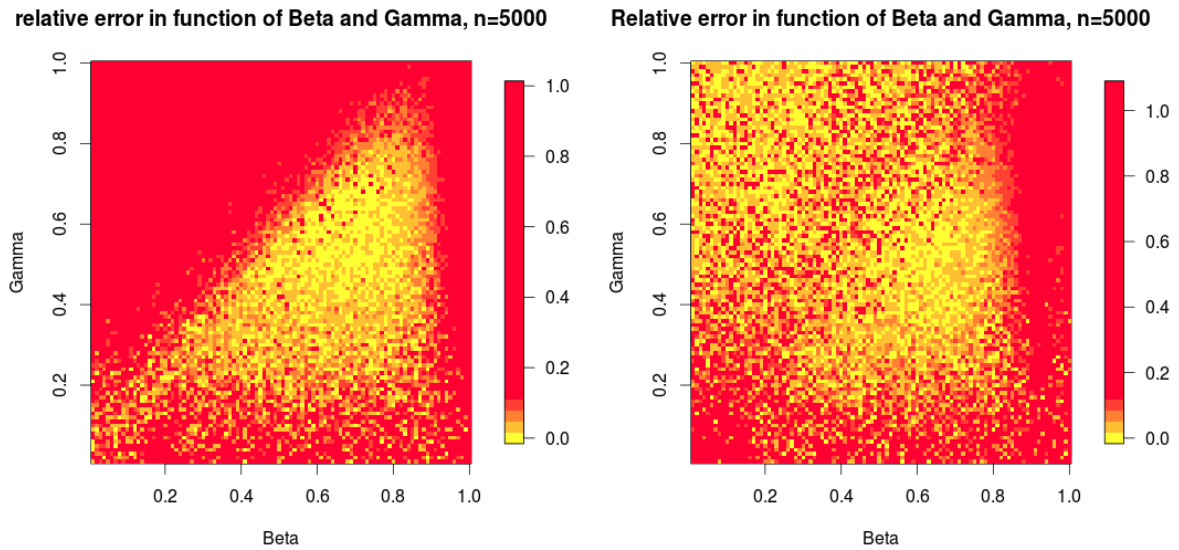


FIGURE – Convergence presque-sûre en fonction de β et γ.

Etude du risque quadratique

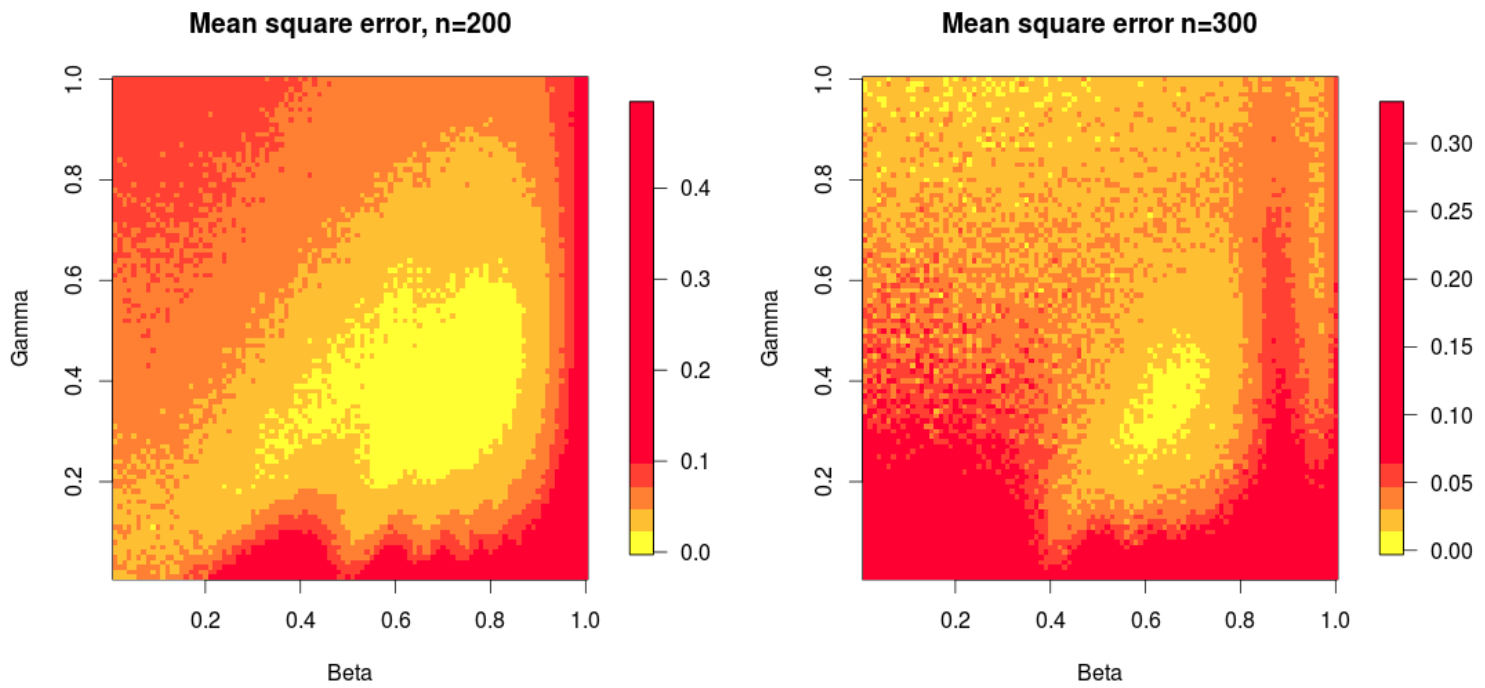


FIGURE – Convergence du risque quadratique en fonction de β et γ.

Dimensions 2 et 3

Nous avons testé les modèles $g(X, \epsilon) = ||X||^2 + \epsilon$ for $X \sim \mathcal{U}([-1, 1]^d)$, $\epsilon \sim \mathcal{U}([-0.5, 0.5])$ et $x = 0_{\mathbb{R}^d}$.

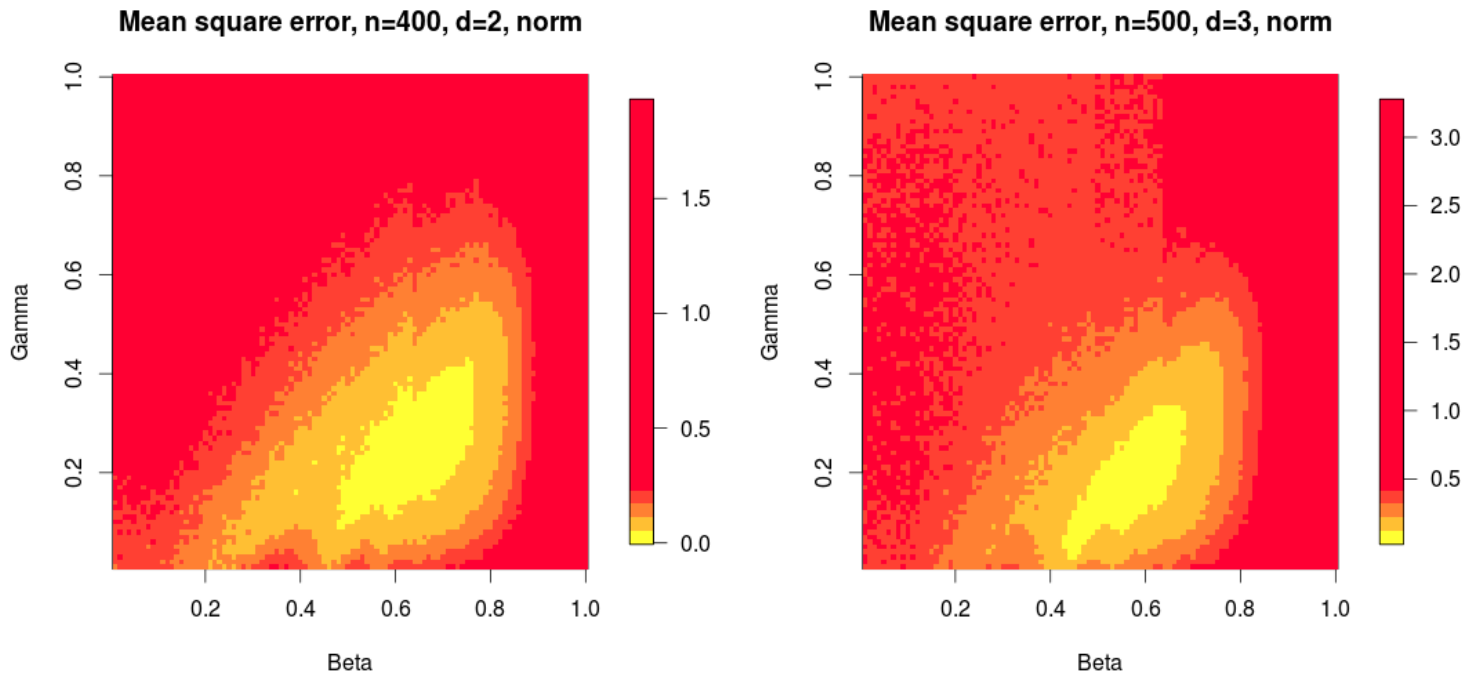


FIGURE – Convergence du risque quadratique en fonction de β and γ.

Lorsque les supports ne sont pas compacts

Pour le modèle $g(X, \epsilon) = X^2 + \epsilon$ avec $X \sim \mathcal{E}(1)$ et $\epsilon \sim \mathcal{N}(0, 1)$, nous avons :

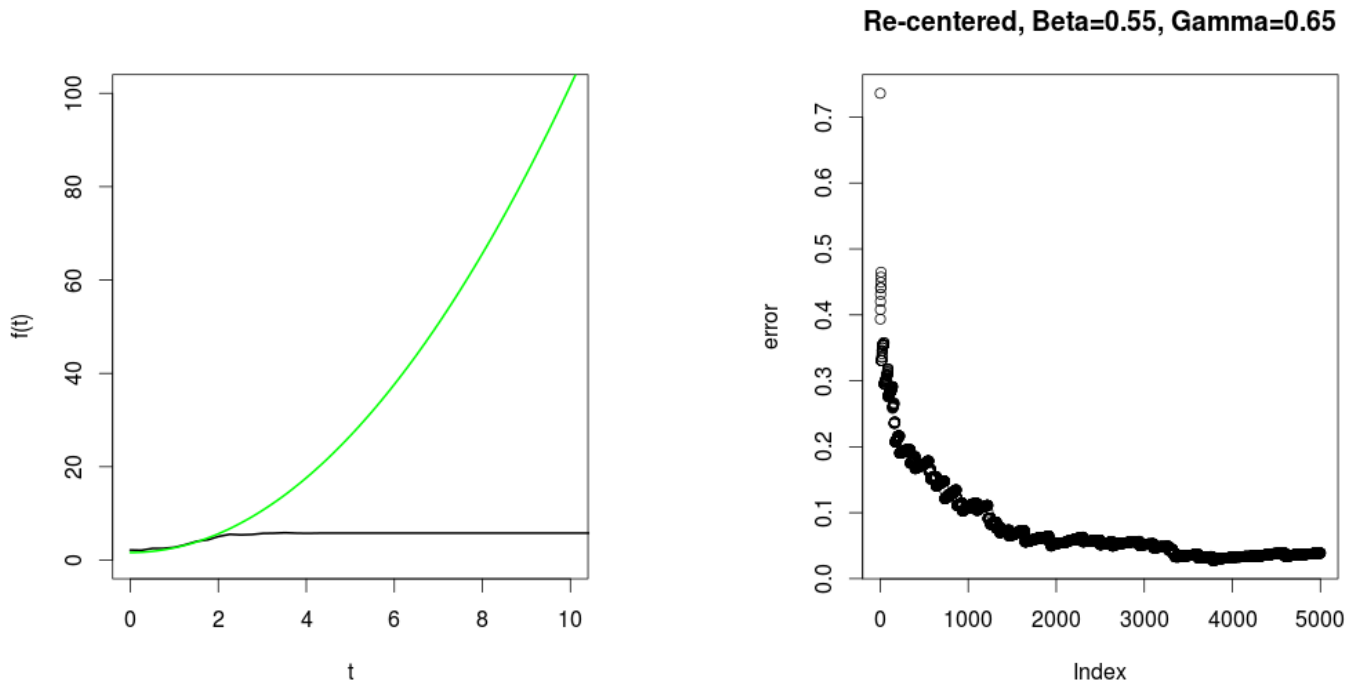


FIGURE – Vers des supports non-compacts

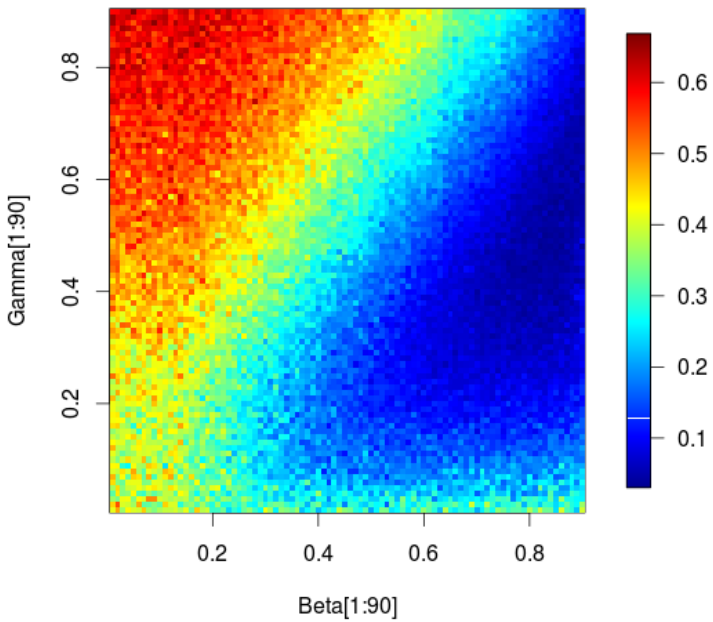


FIGURE – Vers des supports non-compacts

Conclusion et perspectives

Conclusion :

- Nous avons présenté un agorithme pour estimer le quantile conditionnel de la sortie d'un code stochastique.
- Nous avons mis en avant les paramètres optimaux de cet algorithme, pour qu'il ait la meilleure vitesse de convergence.
- Les simulations numériques montrent que notre algorithme est un outil puissant pour répondre au problème.

Perspectives :

- Que se passe-t-il en sortant des hypothèses de support compact ?
- Trouver une manière intelligente de choisir l'échantillon de départ.
- Appliquer cet algorithme à des vrais jeux de données.

Merci pour votre attention.