Estimation itérative en propagation d'incertitudes : réglage robuste de l'algorithme de Robbins-Monro

Bertrand Iooss

Estimation itérative en propagation d'incertitudes : réglage robuste de l'algorithme de Robbins-Monro

Bertrand Iooss [1]

[1] *EDF R&D, 6 Quai Watier, 78401 Chatou - bertrand.iooss@edf.fr*

Résumé.

En quantification d'incertitudes de modèles numériques, l'estimation de quantiles des sorties du modèle est réalisée usuellement par l'analyse statistique de l'échantillon complet de la variable étudiée. Cette approche n'est pas applicable lorsque des quantités prohibitives de données sont générées à chaque simulation. Ce problème peut être résolu grâce à une technique d'estimation à la volée (itérative) basée sur l'algorithme de Robbins-Monro. Nous étudions numériquement cet algorithme afin d'estimer une fonction quantile discrétisée à partir d'échantillons de taille limitée (quelques centaines d'observations). En pratique, la distribution de la variable sous-jacente étant inconnue, il est essentiel de définir des valeurs "robustes" des paramètres de l'algorithme, afin que les estimations des quantiles soient raisonnablement bonnes dans la plupart des situations.

Mots-clés. Quantification d'incertitudes, Quantile, Estimation itérative, Robbins-Monro, Moyennisation

Abstract.

In uncertainty quantification of numerical simulation models, the classical approach for quantile estimation requires availability of the full sample of the studied variable. This approach is not suitable at exascale as large ensembles of simulation runs would need to gather a prohibitively large amount of data. This problem can be solved thanks to an on-the-fly (iterative) approach based on the Robbins-Monro algorithm. We numerically study this algorithm for estimating a discretized quantile function from samples of limited size (a few hundreds observations). As in practice, the distribution of the underlying variable is unknown, the goal is to define "robust" values of the algorithm parameters, which means that quantile estimates have to be reasonably good in most situations.

Keywords. Uncertainty Quantification, Quantile, Iterative estimation, Robbins-Monro, Averaging

1 Introduction

Lors du développement et de l'utilisation des modèles de simulation numérique, les analyses d'incertitudes et de sensibilité sont des outils précieux (Smith, 2014). Elles nécessitent d'exécuter plusieurs (voire de nombreuses) fois le modèle de simulation avec différentes

valeurs des entrées du modèle (suivant des lois de probabilité prédéfinies) afin de calculer des quantités statistiques d'intérêt (notées QoI) sur les sorties du modèle (i.e. leur moyenne, variance, quantiles, indices de sensibilité, ...). La pratique usuelle consiste à stocker tous les résultats de simulation avant de calculer les QoI. Dans certains cas où des variables d'état dépendant du temps et de l'espace sont simulées, la masse de données produites rend prohibitifs leur stockage et leurs temps de lecture (nécessaires à l'estimation des QoI). Une solution proposée récemment dans Terraz et al. (2017) consiste à ne pas stocker les sorties des simulations en calculant les QoI à la volée. Cela amène à considérer des problèmes d'estimation statistique itérative, sujet relativement classique dans le traitement des gros volumes de données mais peu explorés dans les études d'incertitudes de modèles numériques.

Dans ce travail, nous nous intéressons à l'élaboration d'un algorithme d'estimation itératif en propagation d'incertitudes (dans la suite de Ribés et al., 2019), alors que l'analyse de sensibilité itérative a été étudiée dans Terraz et al. (2017). Nous nous focalisons sur l'estimation de quantiles, éléments essentiels pour le calcul d'intervalles de prédiction ou de tolérance, et pour la détection d'outliers, en particulier dans les études de sûreté (voir un exemple dans le domaine de l'ingénierie nucléaire dans Iooss and Marrel, 2019). En se restreignant (par souci de concision) à une sortie scalaire, nous cherchons un estimateur $\hat{q}_\alpha$ des α-quantiles q_α (de la variable aléatoire $Y \in \mathbb{R}$) définis par:

$$q_\alpha = \inf\{y \in \mathbb{R} \,|\, \mathbb{P}(Y \leq y) \geq \alpha\} \,, \tag{1}$$

avec $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ où $\alpha_{\min}$ $(\in]0,1[)$ et $\alpha_{\max}$ $(\in]0,1[)$ sont les valeurs minimale et maximale des ordres des quantiles estimés. Dans notre étude, $\alpha_{\min}$ (resp. $\alpha_{\max}$) sera égal à 5% (resp. 95%). L'estimateur empirique de q_α, associant à l'échantillon i.i.d. $(Y_1, \ldots, Y_N)$ l'échantillon ordonné $(Y_{(1)}, \ldots, Y_{(N)})$, s'écrit $\hat{q}_\alpha^N = Y_{(\lfloor \alpha N \rfloor + 1)}$.

A la place de cet estimateur, nous étudions l'algorithme de Robbins-Monro (RM) (Robbins and Monro, 1951) bien connu pour l'estimation itérative de quantile. L'une des spécificités de notre étude, comme dans Tierney (1983), réside dans la faiblesse de la taille de l'échantillon disponible (quelques centaines d'observations). Ainsi, les propriétés asymptotiques de l'estimateur considéré, quoique apportant des garanties de convergence essentielles, seront peu exploitables pour le réglage des algorithmes.

L'algorithme RM consiste à mettre à jour l'estimateur courant du quantile (noté $q_\alpha(n)$) à chaque nouvelle observation Y_{n+1} avec $n \geq 1$ par la formule de récurrence

$$q_\alpha(n+1) = q_\alpha(n) - \frac{C}{n^\gamma} \left(\mathbb{1}_{Y_{n+1} \leq q_\alpha(n)} - \alpha \right) \,, \tag{2}$$

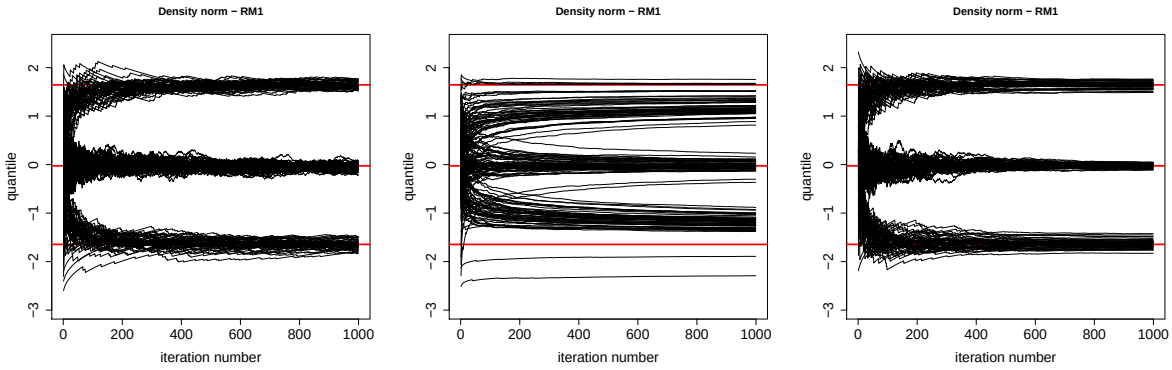
avec $q_\alpha(1) = Y_1$ (étape d'initialisation issue de la première donnée), $C > 0$ une constante et $\gamma \in]0,1]$ régissant la vitesse de descente de l'algorithme stochastique. A taille d'échantillon N finie, l'estimateur RM du α-quantile de Y est donc $\hat{q}_\alpha = q_\alpha(N)$. Cet estimateur est consistant et asymptotiquement gaussien pour $\gamma \in]0.5,1]$ (Duflo, 1997). La valeur de γ ne semble donc pas d'une importance cruciale mais, pour des N peu élevés, nous allons

voir dans la section 2 que son réglage est important. La section 3 discute du réglage de la constante C qui est primordial pour que le pas de descente soit d'amplitude convenable. La section 4 présente finalement une version moyennée de l'algorithme RM.

2 Réglage robuste de γ

Nous cherchons une valeur de γ qui donne des résultats "acceptables" quelle que soit la distribution de Y (inconnue en pratique). Notre test numérique considère les cas $Y \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{U}[0,1]$, avec $N = 1000$, $C = 1$ et trois ordres de quantile α (0.05, 0.5 et 0.95). Pour chacun de ces cas, la Figure 1 montre 50 trajectoires indépendantes de l'estimateur RM $q_\alpha(n)$ pour $n = 1, \ldots, N$ en considérant trois étalonnages différents de γ : 0.6, 1 et une variation linéaire en fonction de n qui s'écrit

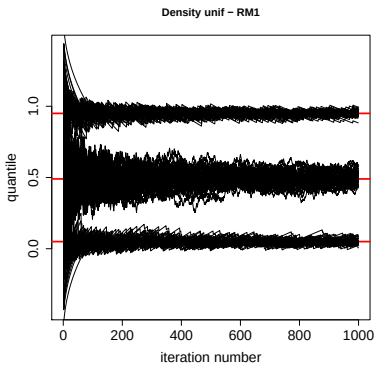
$$\gamma(n) = 0.5 + 0.5\frac{n-1}{N-1} \, . \tag{3}$$



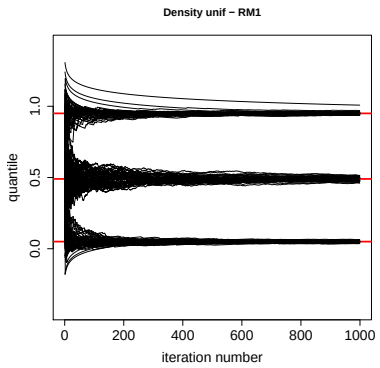
(a) $\gamma = 0.6$.

(b) $\gamma = 1$.

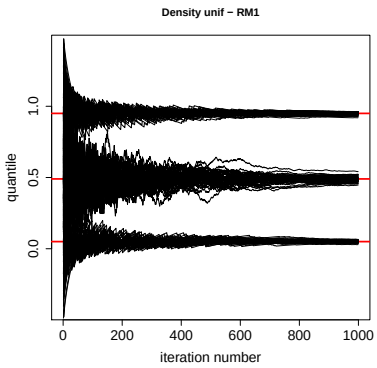
(c) Linear γ evolution.



(d) $\gamma = 0.6$.



(e) $\gamma = 1$.



(f) Linear γ evolution.

Figure 1: Simulations de trajectoires de l'algorithme RM ($N = 1000$, haut : $Y \sim \mathcal{N}(0,1)$, bas : $Y \sim \mathcal{U}[0,1]$). Les lignes rouges donnent les quantiles exacts d'ordre 0.05, 0.5 et 0.95.

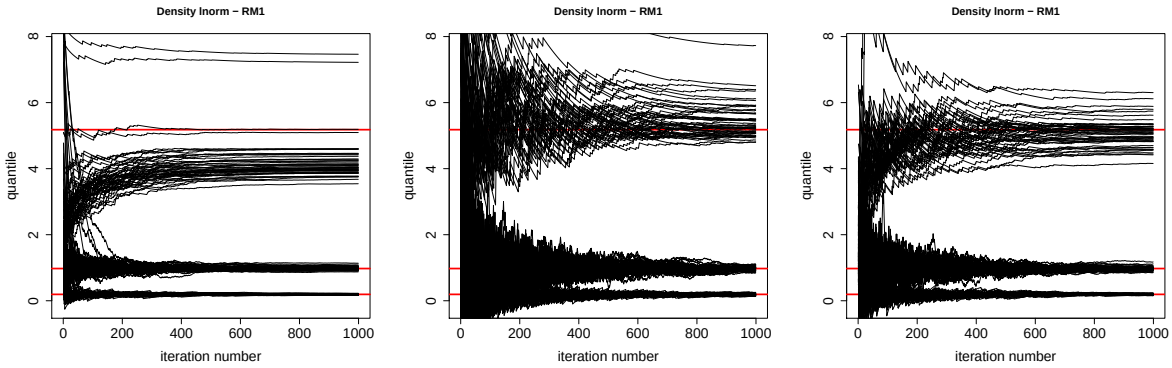
L'idée du profil de $\gamma(n)$ donné par l'Eq. (3) est d'avoir des fluctuations fortes de l'estimateur au début de l'algorithme (pour effacer sa dépendance aux valeurs de Y tirées en premier) puis des fluctuations faibles en fin d'algorithme (pour stabiliser l'estimateur aux dernières itérations). En effet, nous pouvons constater que les fluctuations avec $\gamma = 1$ sont trop faibles dans le cas gaussien ($\gamma = 0.6$ est satisfaisant dans ce cas-là) et les fluctuations avec $\gamma = 0.6$ sont trop fortes dans le cas uniforme ($\gamma = 1$ est satisfaisant dans ce cas-là). Le profil d'une variation linéaire de γ réalise un compromis entre ces deux cas extrêmes (et dans les nombreux autres tests réalisés). Par ailleurs, les propriétés théoriques asymptotiques de l'algorithme RM sont conservées avec le réglage de l'Eq. (3).

3 Réglage robuste de C

Dans la section précédente, la constante C a été fixée à 1. Ce choix s'avère catastrophique lorsque la variable considérée a une dispersion qui n'est pas de cet ordre de grandeur. Il faut rappeler qu'en pratique cette dispersion est inconnue. La Figure 2 montre 50 trajectoires indépendantes de l'estimateur RM $q_\alpha(n)$ pour $n = 1, \ldots, 1000$ et Y de loi lognormale ($\log(Y) \sim \mathcal{N}(0,1)$). γ est de profil linéaire et trois étalonnages différents de C sont testés : 1, 10 et un réglage adaptatif qui s'écrit

$$C(n) = |q_{\alpha_{\max}}(n-1) - q_{\alpha_{\min}}(n-1)| \,, \tag{4}$$

où $n \geq 2$ et $C(1) = |Y_2 - Y_1|$. Sur la Figure 2, il est clair que, pour le quantile d'ordre 0.95, C doit être suffisamment grand pour que les fluctuations soient importantes dès le début de l'algorithme RM. Le réglage adaptatif de C via l'Eq. (4) permet de réguler automatiquement ces fluctuations. De nombreux autres tests numériques sur des distributions de différents types ont permis de confirmer la justesse de ce choix.



(a) $C = 1$. (b) $C = 10$. (c) C adaptatif.

Figure 2: Simulations de trajectoires de l'algorithme RM ($N = 1000$, $Y \sim \mathcal{LN}(0,1)$). Les lignes rouges donnent les quantiles exacts d'ordre 0.05, 0.5 et 0.95.

4 Version moyennée de Robbins-Monro

Il est connu que la version moyennée de RM (notée ici RMM) converge plus rapidement que celle de l'Eq. (2). Nous avons cependant constaté que, si le quantile moyenné est introduit dans (2), les fluctuations de l'estimateur le long des itérations ne sont pas d'ampleur suffisante pour converger vers la valeur exacte. Il faut donc conserver la formulation (2) pour $q_\alpha(n)$ et stocker en plus, à chaque itération, l'estimateur moyenné (noté $\bar{q}_\alpha(n)$) :

$$\bar{q}_\alpha(n+1) = \bar{q}_\alpha(n) + \frac{q_\alpha(n+1) - \bar{q}_\alpha(n)}{n+1} \ , \tag{5}$$

avec $n \geq 1$ et $\bar{q}_\alpha(1) = Y_1$.

La Figure 3 compare les algorithmes RM et RMM pour $Y \sim \mathcal{N}(0,1)$, $N = 1000$ et le réglage adaptatif de C (cf. Eq. (4)) . Les quantiles sont estimés pour des ordres α discrétisés dans l'intervalle $[0.05, 0.95]$ par pas de 0.01. La métrique utilisée (en ordonnée) est l'erreur quadratique moyenne entre les quantiles exacts et les quantiles estimés. Les estimations sont répétées 100 fois de manière indépendante afin de capturer la variabilité des erreurs due à l'échantillonnage. L'estimateur de référence est l'estimateur empirique (qui n'est pas itératif). Sur cet exemple, les performances de RM et RMM avec un γ de profil linéaire sont similaires et proches de celles de l'estimateur empirique. Un γ constant et faible (égal à 0.6) donne des résultats encore meilleurs avec RMM (mais pas avec RM). En fait, la moyennisation dans RMM (qui fait converger plus rapidement l'estimateur du quantile) rend inutile l'augmentation du γ vers 1 que l'on a avec le profil linéaire. D'autres tests avec différentes distributions, non montrés ici, présentent des conclusions similaires.

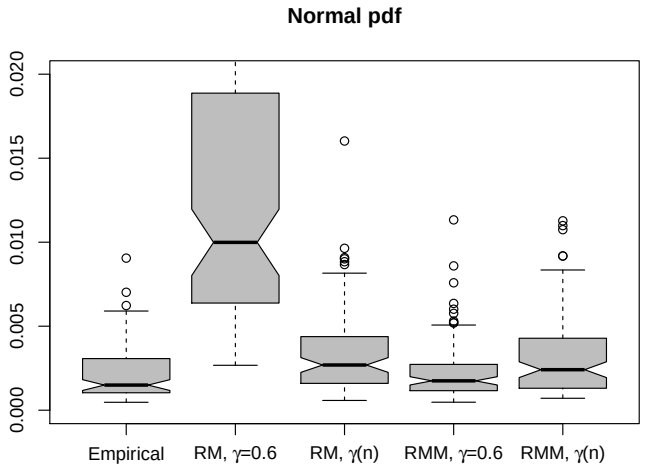


Figure 3: Erreurs quadratiques moyennes des fonctions quantiles discrétisées pour les estimateurs empirique, RM et RMM. $\gamma(n)$ est le réglage de γ en profil linéaire.

5 Conclusions

Ce travail a permis de dégager quelques heuristiques pour l'estimation itérative de quantile par l'algorithme RM avec un échantillon de faible taille. Le choix d'un C adaptatif est bénéfique dans tous les cas et le choix d'un γ de profil linéaire est robuste et doit être privilégié pour RM. Par contre, pour l'algorithme RMM, γ faible donne de meilleurs résultats. Enfin, l'utilisation de séquences de points bien réparties au lieu d'échantillons i.i.d (tests non montrés ici) permettent d'améliorer la précision des estimateurs de manière drastique, avec γ de profil linéaire et C adaptatif. Cette idée semble judicieuse et sera étudiée en profondeur dans le cas où la variable Y provient d'un modèle dont les entrées sont de grande dimension et où le choix d'un bon plan d'expériences (de type "space filling design") est important. Dans le même ordre d'idée, il sera fructueux de combiner l'algorithme RM et les techniques de simulation d'événements rares (cf. e.g. Kohler et al., 2014). Une autre perspective majeure de ce travail sera d'avoir accès à un intervalle de confiance sur le quantile estimé, quantité indispensable dans les applications pratiques.

6 Remerciements

Ce travail émane en partie du projet ANR international INDEX (ANR-18-CE91-0007) dédiés aux plans d'expériences incrémentaux. L'auteur remercie Luc Pronzato, Bernard Bercu, Alejandro Ribés et Clément Gauchy pour de fructueuses discussions.

Bibliographie

Duflo, M. (1997). *Random iterative models*. Springer, Berlin.

Iooss, B. and Marrel, A. (2019). Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *Nuclear Technology*, 205:1588–1606.

Kohler, M., Krzyżak, A., and Walk, H. (2014). Nonparametric recursive quantile estimation. *Statistics and Probability Letters*, 93:102–107.

Ribés, A., Terraz, T., Fournier, Y., Iooss, B., and Raffin, B. (2019). Large scale in transit computation of quantiles for ensemble runs. *Preprint, https://hal.inria.fr/hal-02016828*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407.

Smith, R. (2014). *Uncertainty quantification*. SIAM.

Terraz, T., Ribes, A., Fournier, Y., Iooss, B., and Raffin, B. (2017). Large scale in transit global sensitivity analysis avoiding intermediate files. In *Proceedings the International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing)*, Denver, USA.

Tierney, L. (1983). A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4:706–711.