Uncertainty Quantification for Stochastic Approximation Limits

Uladzislau Stazhynski[1]

joint work with S. Crépey[2], G. Fort [3] and E. Gobet[4]

1,4 - Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France,
2 - LaMME, Univ. Evry, CNRS, Université Paris-Saclay, 91037, Evry, France,
3 - CNRS, Institut de Mathmatiques de Toulouse (IMT), 31062 Toulouse Cedex 9, France.

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

Introduction

Stochastic Approximation (SA) method

- Want to find a solution $z^\star$ of the equation

$$\mathbb{E}[H(z,V)] = 0,$$

  where V is random noise, for which i.i.d. simulations are available, and H is known.
- **Stochastic Approximation** method – introduced by Robbins and Monro in [4]:

$$z^{k+1} = z^k - \gamma_{k+1} H(z^k, V^{k+1})$$

  where $(V^k)_k$ are i.i.d. $V^k \sim V$.
- Under suitable assumptions on $(\gamma_k)_{k\ge 0}$ and H we obtain $\lim_{k\to+\infty} z^k = z^\star$.
- Particular cases of SA: Monte Carlo and Stochastic Gradient Descent.
- **SA applications**: optimization, parameter estimation, signal processing, adaptive control, Monte Carlo optimization of stochastic systems, stochastic gradient descent methods in machine learning, adaptive Monte Carlo sampler, efficient tail computations, etc. (see e.g. [3, 1]).

Uncertainty Quantification problem for SA limits

- Assume that V follows a distribution $\mu(\theta, \mathrm{d}v)$ which depends on an **uncertain parameter** $\theta \in \Theta$, for which some prior distribution $\pi(\mathrm{d}\theta)$ on Θ is available.
- The limit $\phi^\star$ of the corresponding SA procedure

$$\phi^{k+1} = \phi^k - \gamma_{k+1} H(\phi^k, V^{k+1})$$

  depends on θ, i.e. $\phi^\star = \phi^\star(\theta)$.
- Our goal is to **reconstruct the function** $\phi^\star(\cdot)$ as an element of the Hilbert space corresponding to the scalar product induced by π, i.e. $\langle f; g\rangle_\pi := \int_\Theta f(\theta) g(\theta) \pi(\mathrm{d}\theta)$.

Methodology & Results

Formalization of the problem

- Assume that:
  - $\mathcal{V}$ is a metric space, $\Theta \subset \mathbb{R}^d$, and $H : \mathbb{R}^q \times \mathcal{V} \times \Theta \to \mathbb{R}^q$.
  - π is a probability distribution on Θ, μ is a transition kernel from Θ to $\mathcal{V}$.
  - $L^\pi_{2,q}$ is the Hilbert space of functions $f : \Theta \to \mathbb{R}^q$ with the norm $\|f\|_\pi := \sqrt{\sum_{i=1}^q \langle f_i; f_i\rangle_\pi}$. We fix an orthogonal basis $\{B_i(\cdot), i \in \mathbb{N}\}$ of $L^\pi_{2,q}$.
- The main problem writes as:

> Find $\phi^\star$ in $L^\pi_{2,q}$ such that
> $$\int_{\mathcal{V}} H(\phi^\star(\theta), v, \theta)\mu(\theta, \mathrm{d}v) = 0, \qquad \pi\text{-a.s.} \qquad (1)$$

- This is equivalent to finding $(u_i^\star)_{i\in\mathbb{N}}$ such that $\phi^\star = \sum_i u_i^\star B_i$.

The USA algorithm:

- In [2] we propose the following algorithm to solve (1):
  - **Inputs:** sequences $\{\gamma_k, k \ge 1\}$ (step-size), $\{m_k, k \ge 1\}$ (growing dimension), $\{M_k, k \ge 1\}$(number of simulations at each iteration); initial point $\{u_i^0, i = 0, \dots, m_0\}$, total number of iterations $K \in \mathbb{N}$; $(\theta^s_{k+1}, V^s_{k+1})$, $s = 1\dots, M_{k+1}$, – i.i.d. simulations w.r.t. $\pi(\mathrm{d}\theta)\mu(\theta, \mathrm{d}v)$.
  - **Repeat for** $k = 1, \dots, K$**:** for $i = 0, \dots, m_{k+1}$

$$u_i^{k+1} = u_i^k - \gamma_{k+1} M_{k+1}^{-1} \sum_{s=1}^{M_{k+1}} H\left(\sum_{j=0}^{m_k} u_j^k B_j(\theta^s_{k+1}), V^s_{k+1}, \theta^s_{k+1}\right) B_i(\theta^s_{k+1})$$

  and $u_i^k = 0$ for $i > m_{k+1}$.
  - **Output:** the vector $\{u_i^K, i = 0, \dots, m_K\}$.
- This gives the following approximation of $\phi^\star$: $\phi^K := \sum_{i=0}^{m_K} u_i^K B_i$.
- Main features of the USA algorithm:
  - USA is a **single iterative procedure without nested calculations** (as opposed to naive Monte Carlo UQ) – this yields much lower computational cost.
  - It is an iterative procedure **in growing dimension** (dim. $\to \infty$), yet it is fully implementable.

Convergence analysis

- USA has a specific form, since it is a stochastic approximation procedure in growing dimension.
- As argued in [2], existing works on SA in finite dimension or in Hilbert spaces cannot be applied to show the convergence.
- The main contribution of [2] is the original convergence proof of the USA algorithm:

> **Thm 1.** Under suitable assumptions (see [2]) there exists a random variable ϕ^∞ taking values in the solution set of (1) such that
> $$\lim_{k\to\infty} \left\|\phi^k - \phi^\infty\right\|_\pi = 0 \text{ a.s.}, \qquad \lim_{k\to\infty} \mathbb{E}\left[\left\|\phi^k - \phi^\infty\right\|_\pi^p\right] = 0 \quad \text{for any } p \in (0,2)\ .$$

Numerical Tests (details of the numerical examples are not given here, see [2])

- Illustration of the convergence $\phi^K \to \phi^\star$:

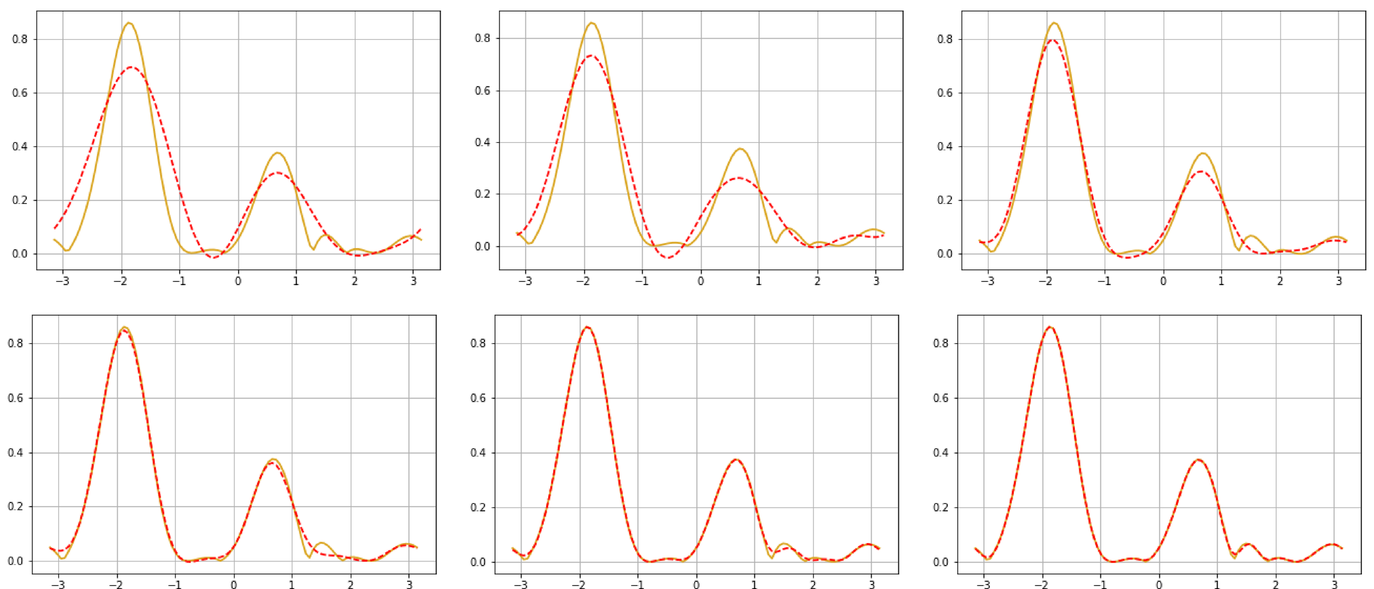

Figure 1: The functions $\phi^\star$ and ϕ^K are displayed in respective solid line and dashed lines, as a function of $\theta \in [-\pi, \pi]$, $K \in \{64, 128, 256, 512, 1024, 2048\}$.

- Dimension growth feature of the USA algorithm is important for:
  - Asymptotic convergence to the solution $\phi^\star$;
  - Bias-free estimation of low-order coefficients;
  - Optimization of the convergence trajectory;

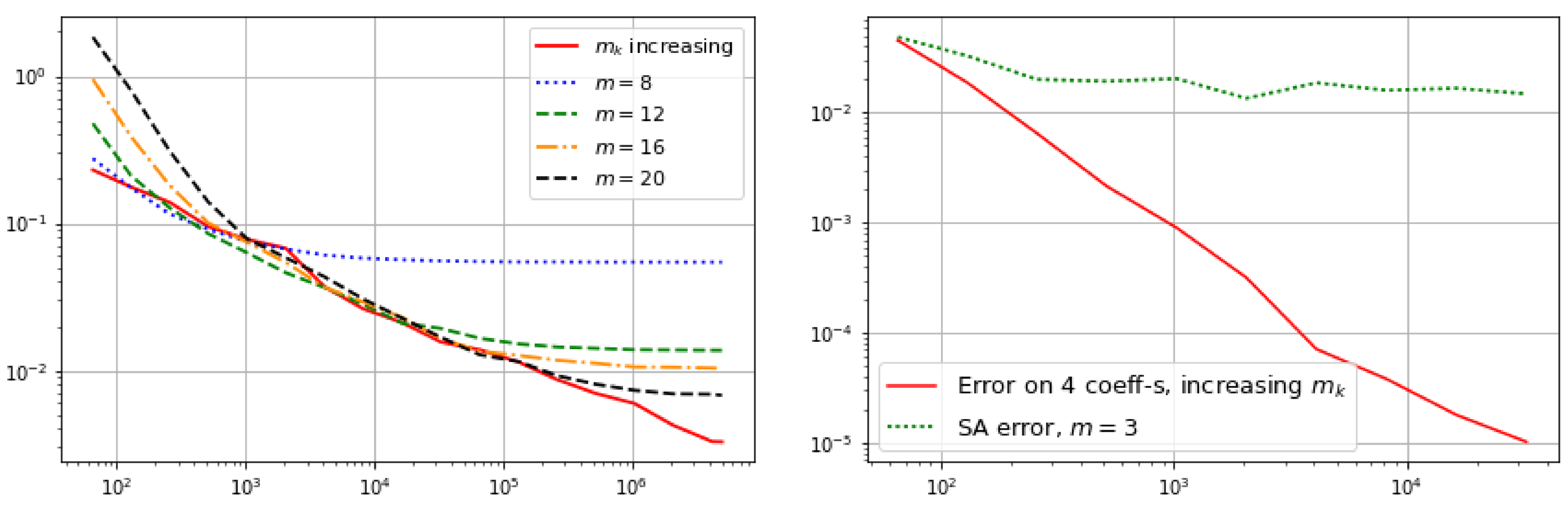

Figure 2: Left: $\mathbb{E}\left[\left\|\phi^K - \phi^\star\right\|_\pi^2\right]^{1/2}$ as a function of the number of iterations, for different choices of the sequence $\{m_k, k \in \mathbb{N}\}$: m_k increasing (solid line) and $m_k = m$ fixed (other lines). Right: In the case $m_k \to \infty$ (solid line) and $m_k = m = 3$ (dotted line), the error on the first 4 coefficients $\mathbb{E}\left[\sum_{i=0}^3 (u_i^K - u_i^\star)^2\right]^{1/2}$ as a function of the number of iterations K.

- Choice of the dimension (m_k) growth speed: here $m_k = k^b$ for different values of b

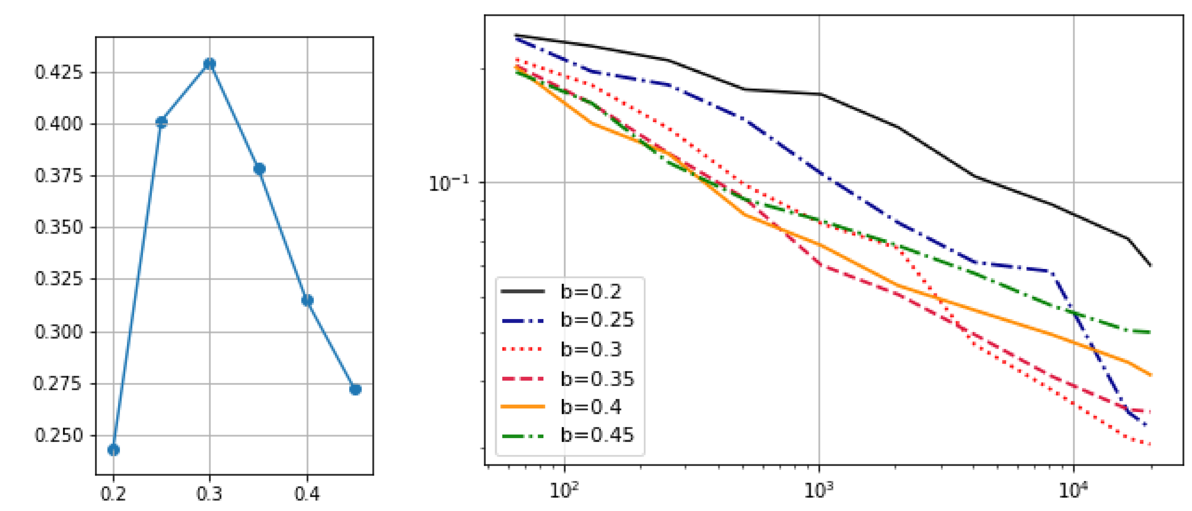

Figure 3: Left: empirical L^2 convergence rate for $b \in \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$. Right: total error $\mathbb{E}\left[\left\|\phi^K - \phi^\star\right\|_\pi^2\right]^{1/2}$ as a function of the number of iterations K, for different values of b.

- Impact of the choice of b (dimension growth speed):
  - Bias-variance trade-off as b increases.
  - For $b \le 0.3$ the total error is dominated by the truncation error $\sum_{i>m_K} (u_i^\star)^2$ (i.e. bias).
  - Optimal choice of b: exact result in the upcoming paper on the L^2-convergence rate of the USA algorithm.

Conclusions

- The USA algorithm is efficient, fully constructive and easy to implement.
- It is given by a single procedure without nested calculations, which leads to much higher efficiency with respect to naive methods.
- The convergence assumptions are given in terms of finite dimensional problems for fixed values of θ, as opposed to abstract assumptions involving Hilbert space notions, often hard to check in practice.

Future prospects

- Upcoming paper on the L^2-convergence rate of the USA algorithm.
- Applications to the calculation of model sensitivities w.r.t. θ.
- Applications to parametric risk measure calculation, risk measure sensitivities and XVAs calculation in finance.

References

[1] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

[2] S. Crépey, G. Fort, E. Gobet, and U. Stazhynski. Uncertainty quantification for stochastic approximation limits using chaos expansion. *Preprint, available at* `https://hal.archives-ouvertes.fr/hal-01629952`, 2017.

[3] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Application of Mathematics*. Springer, 1997.

[4] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.