General multilevel adaptations for stochastic approximation algorithms

Steffen Dereich

Westfälische Wilhelms-Universität Münster

http://wwwmath.uni-muenster.de/statistik/dereich/

joint work with Thomas Müller-Gronbach

Multilevel Monte Carlo @ Paris

7/7/2016

Agenda

I Introduction

II Preliminaries

III Local error analysis (main results)

IV Conclusion

I Introduction

Given:

- random variable U with values in a measurable space $\mathcal{U}$
- $F : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$ measurable such that $F(\theta, U)$ is integrable $\forall \theta \in \mathbb{R}^d$

Aim: Find zeroes of $f : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$f(\theta) = \mathbb{E}[F(\theta, U)].$$

I Introduction

Given:

- random variable U with values in a measurable space $\mathcal{U}$
- $F : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$ measurable such that $F(\theta, U)$ is integrable $\forall \theta \in \mathbb{R}^d$

Aim: Find zeroes of $f : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$f(\theta) = \mathbb{E}[F(\theta, U)].$$

E.g.: Computation of quantiles

- $F(\theta, U) = \alpha - \mathbf{1}_{\{U \leq \theta\}}$ for a $\alpha \in (0, 1)$ and a $\mathbb{R}$-valued random variable U
  $\Rightarrow$ zero of $f(\theta) = \mathbb{E}[F(\theta, U)] = \alpha - \mathbb{P}(U \leq \theta)$ is α-quantile

Computation of extremes

- F now is a mapping $F : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}$ and an extremal value of $f(\theta) = \mathbb{E}[F(\theta, U)]$ corresponds to a zero of

$$g(\theta) = \nabla f(\theta) = \mathbb{E}[\nabla_\theta F(\theta, U)].$$

I Examples

Focus: on the case where $F(\theta, U)$ is not simulatable.

Ex 1: SDE. $F(\theta, U) = f(\theta, X_T^{(\theta,U)})$, where $U = (U_t)$ is a Brownian motion and $(X_t^{(\theta,U)})_{t \geq 0}$ solves an integral equation

$$X_t^{(\theta,U)} = x_0^{(\theta)} + \int_0^t a(X_s^{(\theta,U)}, \theta) \, \mathrm{d}U_s + \int_0^t b(X_s^{(\theta,U)}, \theta) \, \mathrm{d}s.$$

I Examples

Focus: on the case where $F(\theta, U)$ is not simulatable.

Ex 1: SDE. $F(\theta, U) = f(\theta, X_T^{(\theta,U)})$, where $U = (U_t)$ is a Brownian motion and $(X_t^{(\theta,U)})_{t\geq 0}$ solves an integral equation

$$X_t^{(\theta,U)} = x_0^{(\theta)} + \int_0^t a(X_s^{(\theta,U)}, \theta)\,\mathrm{d}U_s + \int_0^t b(X_s^{(\theta,U)}, \theta)\,\mathrm{d}s.$$

Ex 2: PDE with random coefficients. $F(\theta, U)$ value of a PDE with random coefficients U at a certain point.

I Introduction

Central concepts: (to be introduced on the next slides)

- Robbins-Monro algorithm (Robbins, Monro '51)
- Polyak-Ruppert averaging (Ruppert '91, Polyak, Juditsky '92)
- Multilevel paradigm (Heinrich '98, Giles '08)

Aim: Multilevel stochastic approximation algorithms in the spirit of Giles '08

Rel. research: N. Frikha '13+, Multi-level stochastic approximation algorithms

II Robbins-Monro algorithms (L-attractors)

Def: We call a zero θ^* of f L-attractor for an $L > 0$ if

$$f(\theta) = H(\theta - \theta^*) + o(|\theta - \theta^*|), \quad \text{as } \theta \to \theta^*,$$

where H is in $\mathbb{R}^{d \times d}$ with

$$\max\{\mathrm{Re}(\lambda) : \lambda \text{ eigenvalue of } H\} \leq -L.$$

II Robbins-Monro algorithms (L-attractors)

Def: We call a zero θ^* of f L-attractor for an $L > 0$ if

$$f(\theta) = H(\theta - \theta^*) + o(|\theta - \theta^*|), \quad \text{as } \theta \to \theta^*,$$

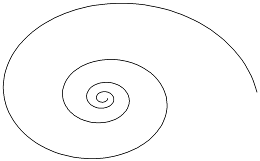
where H is in $\mathbb{R}^{d \times d}$ with

$$\max\{\mathrm{Re}(\lambda) : \lambda \text{ eigenvalue of } H\} \leq -L.$$

Motivation: For an L-attractor θ^* the solution $u : [0, \infty) \to \mathbb{R}^d$ of the differential equation

$$\dot{u}(t) = f(u(t))$$

looks in the vicinity of θ^* like



II Robbins-Monro algorithms

The evolution equation "finds" attracting zeroes. To get from the evolution equation

$$\dot{u}(t) = f(u(t))$$

to stochastic approximation algorithms one does

- Euler steps with step width $\gamma_1, \gamma_2, \ldots$
- with f replaced by a random variable having the "right" expectation.

II Robbins-Monro algorithms

The evolution equation “finds” attracting zeroes. To get from the evolution equation

$$\dot{u}(t) = f(u(t))$$

to stochastic approximation algorithms one does

- Euler steps with step width $\gamma_1, \gamma_2, \dots$
- with f replaced by a random variable having the “right” expectation.

Robbins-Monro system: Given a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of strictly positive reals and a starting value θ_0 we set

$$\theta_n = \theta_{n-1} + \gamma_n F(\theta_{n-1}, U_n)$$

for $n \in \mathbb{N}$, where $U_1, U_2, \dots$ are independent copies of U.

II Robbins-Monro algorithms

The evolution equation "finds" attracting zeroes. To get from the evolution equation

$$\dot{u}(t) = f(u(t))$$

to stochastic approximation algorithms one does

- Euler steps with step width $\gamma_1, \gamma_2, \ldots$
- with f replaced by a random variable having the "right" expectation.

Robbins-Monro system: Given a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of strictly positive reals and a starting value θ_0 we set

$$\theta_n = \theta_{n-1} + \gamma_n \, F(\theta_{n-1}, U_n)$$

for $n \in \mathbb{N}$, where $U_1, U_2, \ldots$ are independent copies of U.

Note: Natural assumptions on (γ_n) are

- $\gamma_n \to 0$: randomness of U_n should lose its impact
- $\sum_{n\in\mathbb{N}} \gamma_n = \infty$: the associated "Euler time" should tend to infinity.

Refs: Robbins, Monro '51, ..., Pelletier '98, ..., Duflo '96, Kushner, Yin '03

II Polyak-Ruppert averaging

Note: The fastest convergence of (θ_n) is obtained for (γ_n) of the form

$$\gamma_n = \frac{\gamma_0}{n} \quad \text{with} \quad \gamma_0 > (2L)^{-1}$$

Then

$$|\theta_n - \theta^*| \text{ ``}\approx\text{''} \ n^{-1/2}.$$

Problem: The strength of attraction L is typically not known!

II Polyak-Ruppert averaging

Note: The fastest convergence of (θ_n) is obtained for (γ_n) of the form

$$\gamma_n = \frac{\gamma_0}{n} \quad \text{with} \quad \gamma_0 > (2L)^{-1}$$

Then

$$|\theta_n - \theta^*| \text{ ``}\approx\text{''} \; n^{-1/2}.$$

Problem: The strength of attraction L is typically not known!

Remedy: Use $\gamma_n = n^{-\eta}$ with $\eta \in (1/2, 1)$ instead and consider as approximation

$$\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} \theta_k.$$

- requires stronger assumptions on f
- same order of convergence

Refs: Ruppert '91, Polyak, Juditsky '92

III Multilevel stochastic approximation

Aim: Compute zero of

$$f(\theta) = \mathbb{E}[F(\theta, U)]$$

for a nonsimulatable $F(\theta, U)$!

Use: Hierarchical scheme of approximations:

$$F_1, F_2, \dots : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$$

measurable functions. Further $F_0 \equiv 0$.

- Establish stochastic approximation schemes that show the same rates of convergence as one has in the computation of a single expectation.
- Employ similar assumptions as in Giles '08.

III Assumptions

Assu A=A($\alpha, \beta, \theta^*, L, p$): θ^* is a L-attractor of f and there exist $\delta > 0$ and $c \in (0, \infty)$ such that for $\theta \in B(\theta^*, \delta)$

- $|\mathbb{E}[F(\theta, U)] - \mathbb{E}[F_k(\theta, U)]| \leq c\,(M^k)^{-\alpha}$
- $\mathbb{E}[|F(\theta, U) - F_k(\theta, U)|^p]^{2/p} \leq c\,(M^k)^{-\beta}$
- one simulation of $F_k(\theta, U) - F_{k-1}(\theta, U)$ is assigned the cost $C_k = M^k$

(with $\alpha, \beta, L \in (0, \infty)$, $\theta^* \in \mathbb{R}^d$ and $p \in [2, \infty)$).

Note:

- The assumptions only require an error control for fixed θ uniformly in a neighbourhood of θ^*.

III Assumptions

Assu A=A($\alpha, \beta, \theta^*, L, p$): θ^* is a L-attractor of f and there exist $\delta > 0$ and $c \in (0, \infty)$ such that for $\theta \in B(\theta^*, \delta)$

- $|\mathbb{E}[F(\theta, U)] - \mathbb{E}[F_k(\theta, U)]| \leq c\,(M^k)^{-\alpha}$
- $\mathbb{E}[|F(\theta, U) - F_k(\theta, U)|^p]^{2/p} \leq c\,(M^k)^{-\beta}$
- one simulation of $F_k(\theta, U) - F_{k-1}(\theta, U)$ is assigned the cost $C_k = M^k$

(with $\alpha, \beta, L \in (0, \infty)$, $\theta^* \in \mathbb{R}^d$ and $p \in [2, \infty)$).

Note:

- The assumptions only require an error control for fixed θ uniformly in a neighbourhood of θ^*.
- In the SDE example one may for instance choose Euler approximations with M^k steps.
- The error analysis done for classical multilevel algorithms can directly be transferred.
- More elaborate multilevel algorithms such as antithetic Milstein can also be combined with our approach.

III Scheme with deterministic choice of levels (A)

Algorithms are specified by an initial vector $\theta_0 \in \mathbb{R}^d$,

(i) $(\gamma_n)_{n\in\mathbb{N}} \subset (0,\infty)$: decreasing sequence determining step sizes
(ii) $(m_n)_{n\in\mathbb{N}} \subset \mathbb{N}$: increasing sequence determining maximal levels, and
(iii) $(a_k)_{k\in\mathbb{N}} \subset (0,\infty)$: decreasing sequence determining iteration numbers

$$N_{n,k} = \lceil a_k / a_{m_n} \rceil, \quad \text{for } k = 1, \dots, m_n \text{ and } n \in \mathbb{N}.$$

Innovation: Using iid copies $(U_{n,k,\ell})$ of U we set

$$Z_n(\theta) = \sum_{k=1}^{m_n} \frac{1}{N_{n,k}} \sum_{\ell=1}^{N_{n,k}} \big(F_k(\theta, U_{n,k,\ell}) - F_{k-1}(\theta, U_{n,k,\ell})\big)$$

Robbins-Monro step: adapted dynamical system $(\theta_n)_{n\in\mathbb{N}}$ such that

$$\theta_n = \theta_{n-1} + \gamma_n Z_n(\theta_{n-1}).$$

Cost:

$$\mathrm{cost}_n = \sum_{j=1}^{n} \sum_{k=1}^{m_j} N_{j,k}\, C_k$$

III Local error analysis (Robbins-Monro)

Theorem: (D, Müller-Gronbach '16+) Suppose that Assumption A is satisfied and that $2\alpha > \beta$ or $\beta > 1$. Let

$$\sigma = \frac{2\alpha}{4\alpha - \beta - \min(1,\beta)}, \quad \gamma_0 \in (\rho/(2L), \infty)$$

and

$$\gamma_n = \gamma_0\, n^{-1}, \quad m_n = \left\lceil \frac{\sigma}{\alpha \ln M} \ln(n+1) \right\rceil, \quad a_n = M^{-n^{\frac{(\beta+1)}{2}}}.$$

Then there exist $\delta, \kappa \in (0, \infty)$ such that

$$\limsup_{n\to\infty} \frac{1}{\varepsilon_n}\, \mathbb{E}[\mathbf{1}_{\{(\theta_n)_{n\geq k_0} \subset B(\theta^*,\delta)\}} |\theta_n - \theta^*|^p]^{1/p} \leq \kappa$$

for every $k_0 \in \mathbb{N}$ with

$$\varepsilon_n = \begin{cases} n^{-\sigma}, & \text{if } \beta \neq 1, \\ n^{-\sigma}\sqrt{\ln(n+1)}, & \text{if } \beta = 1, \end{cases} \quad \text{and} \quad \mathrm{cost}_n \leq \begin{cases} \kappa\, n^{2\sigma}, & \text{if } \beta > 1, \\ \kappa\, n^{2\sigma} \ln(n+1), & \text{if } \beta = 1, \\ \kappa\, n^{\sigma\left(1+\frac{1-\beta}{\alpha}\right)}, & \text{if } \beta < 1. \end{cases}$$

III Local error analysis (Polyak-Ruppert averaging)

Polyak-Ruppert averaging: Let $(b_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be an increasing sequence and consider

$$\bar{\theta}_n = \frac{1}{\sum_{k=1}^n b_k} \sum_{k=1}^n \mathbb{1}_{\{|\theta_k - \theta_n| \le C\}} \, b_k \, \theta_k .$$

III Local error analysis (Polyak-Ruppert averaging)

Polyak-Ruppert averaging: Let $(b_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be an increasing sequence and consider

$$\bar\theta_n = \frac{1}{\sum_{k=1}^n b_k} \sum_{k=1}^n \mathbb{1}_{\{|\theta_k-\theta_n|\le C\}}\, b_k\, \theta_k .$$

Assu B=B($\alpha,\beta,\theta^*,L,p,\lambda$): Assu A is satisfied and one has

$$f(\theta) = Df(\theta^*)(\theta-\theta^*) + o(|\theta-\theta^*|^{1+\lambda}) \text{ as } \theta \to \theta^*.$$

III Local error analysis (Polyak-Ruppert averaging)

Polyak-Ruppert averaging: Let $(b_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be an increasing sequence and consider

$$\bar{\theta}_n = \frac{1}{\sum_{k=1}^n b_k} \sum_{k=1}^n \mathbb{1}_{\{|\theta_k - \theta_n| \le C\}}\, b_k\, \theta_k .$$

Assu B=B($\alpha, \beta, \theta^*, L, p, \lambda$): Assu A is satisfied and one has

$$f(\theta) = Df(\theta^*)(\theta - \theta^*) + o(|\theta - \theta^*|^{1+\lambda}) \text{ as } \theta \to \theta^*.$$

Theorem: (D, Müller-Gronbach '16+) Suppose Assu B is satisfied with general $p \ge 2$ and suppose that $2\alpha > \beta$ or $\beta > 1$. Let σ, (m_n) and (a_n) be as in the previous theorem.

Take $q \in [p/(1+\lambda), p)$ and $\eta \in ((1 - 2\sigma(p-q)/p)_+, 1)$ and set $\gamma_n = \mathrm{const}\, n^{-\eta}$. Further take $\xi \in [\sigma - 1/2, \infty)$ and set $b_n = n^{\xi}$.

Then there exist $\delta, \kappa \in (0,\infty)$ such that

$$\limsup_{n\to\infty} \frac{1}{\varepsilon_n} \mathbb{E}[\mathbb{1}_{\{(\theta_n)_{n\ge k_0} \subset B(\theta^*,\delta)\}} |\bar{\theta}_n - \theta^*|^q]^{1/q} \le \kappa$$

for every $k_0 \in \mathbb{N}$ with (ε_n) as before.

IV Comparison with related work

Frikha '14+: CLT for multilevel stochastic approximation for SDEs

IV Comparison with related work

Frikha '14+: CLT for multilevel stochastic approximation for SDEs

Approach: Denote

$$f_n(\theta) = \mathbb{E}[F_n(\theta, U)]$$

and let θ_n^* denote the unique zero of f_n (assumption). For $m \in \mathbb{N}$ one has

$$\theta_m^* = (\theta_m^* - \theta_{m-1}^*) + \ldots + \theta_1^*$$

and to estimate $\theta_n^* - \theta_{n-1}^*$ one performs coupled stochastic approximation algorithms with F_n and F_{n-1} using the same U's.

IV Comparison with related work

Frikha '14+: CLT for multilevel stochastic approximation for SDEs

Approach: Denote

$$f_n(\theta) = \mathbb{E}[F_n(\theta, U)]$$

and let θ_n^* denote the unique zero of f_n (assumption). For $m \in \mathbb{N}$ one has

$$\theta_m^* = (\theta_m^* - \theta_{m-1}^*) + \ldots + \theta_1^*$$

and to estimate $\theta_n^* - \theta_{n-1}^*$ one performs coupled stochastic approximation algorithms with F_n and F_{n-1} using the same U's.

Comments:

- algorithms utilise Polyak-Ruppert averaging
- analysis cumbersome since one needs to analyse coupled stochastic approximation algorithms
- optimal results are obtained for (γ_n) of the form $\gamma_n = \gamma_0/n$
  $\Rightarrow$ estimate for L needed

IV Pros and cons

Robbins-Monro	**Polyak-Ruppert averaging**
estimates for L needed	value of L irrelevant
except differentiablility in θ^* no regularity assumptions on f	slightly stronger regularity assumptions on f in θ^*
original moment in error estimate	reduced moment in error estimate

IV Pros and cons

Robbins-Monro	**Polyak-Ruppert averaging**
estimates for *L* needed	value of *L* irrelevant
except differentiablility in θ^* no regularity assumptions on *f*	slightly stronger regularity assumptions on *f* in θ^*
original moment in error estimate	reduced moment in error estimate

Fixed choice of levels	**Random choice of levels**
general moments accessible	loss of efficiency for moments larger than 2
$2\alpha > \beta$ needed in slow regime	$2\alpha = \beta$ generally allowed

IV Concluding remarks

- With multilevel stochastic approximation the computation of L-attractors is as costly as the computation of a single expectation with multilevel.
- As for classical multilevel Monte Carlo one can replace the random variables $F_k(\theta, U) - F_{k-1}(\theta, U)$ by other random variables $P_k(\theta, U)$ having the same expectation. In particular, the antithetic Milshtein idea is applicable.
- The approach can easily be adapted to the computation of maxima.
- A combination with extrapolation methods is straight-forward.
- Central limit theorems can also be deduced by using standard theory
- In the fast regime ($\beta > 1$) one can replace in classical stochastic approximation algorithms $F(\theta, U)$ by

$$\frac{F_J(\theta, U) - F_{J-1}(\theta, U)}{a_J}.$$

with J being independent of U with appropriate prob. weights (a_k) (in the spirit of McLeish '11, Glynn, Rhee '12).

Main reference:

S. Dereich, T. Müller-Gronbach "General multilevel adaptations for stochastic approximation algorithms", arXiv:1506.05482

This and further related articles can be found on my homepage:

`http://wwwmath.uni-muenster.de/statistik/dereich/`

Thank you for your attention!