Strong Approximations for Robbins-Monro Procedures.

Valentin Konakov[1] and Enno Mammen[2]

[1]Higher School of Economics, Moscow

[2]Heidelberg University

Mathematical Statistics in the Information Age, Statistical Efficiency and Computational Ability

September 17–19, 2025
University of Vienna

Topic of talk I

Robbins-Monro algorithm (1951)

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, \eta_{n+1}), \ \theta_0 \in \mathbb{R}^d,$$

with

- $(\gamma_k)_{k\geq 0}$ decreasing step sequence, (today: $\gamma_k = \frac{A}{k+B}$ with $A > 0, B \geq 0$),
- $(\eta_k)_{k\geq 0}$ i.i.d. random variables,
- H function from $\mathbb{R}^d \times \mathcal{X}$ to $\mathbb{R}^d$,
- $\mathcal{X}$ support of η_i.

The Robbins-Monro procedure is used to approximate the zeros of the function: $h(\theta) = \mathbb{E}[H(\theta, \eta)]$.

Topic of talk II

Set-up: Have observed the algorithm for $n \leq N$ with $N \in \mathbb{N}$, i.e. $\theta_0, ..., \theta_N$ fixed, non-random.
Want to understand stochastics of **shifted Robbins-Monro algorithm:**

$$\left(\theta_n^N\right)_{n\geq 0} = (\theta_{N+n})_{n\geq 0}$$

These algorithm satisfy the following recurrence equation:

$$\theta_{n+1}^N = \theta_n^N - \gamma_{n+1}^N H(\theta_n^N, \eta_{n+1}^N)$$

with starting value $\theta_0^N = \theta_N$, where $\eta_{n+1}^N = \eta_{N+n+1}$, and $\gamma_{n+1}^N = \gamma_{N+n+1}$.

Topic of talk III

Time change: k number of iterations of Robbins-Monro algorithm after N,

New time: t

$$t_k^N = \gamma_1^N + \cdots + \gamma_k^N,$$

$$k_t^N = \inf\{k \in \mathbb{N} \; ; \; t_k^N \geq t\}.$$

For $\gamma_k = \frac{A}{k+B}$ with $A = 1, B = 0$ we have for $N \to \infty$

$$t_k^N \approx \ln\left(1 + \frac{k}{N}\right),$$

$$k_t^N \approx (\exp(t) - 1)N.$$

Robbins-Monro algorithm in new time

$$\check{\theta}_t^N = \theta_k^N \text{ for } t_k^N \leq t < t_{k+1}^N.$$

Topic of talk IV

Asymptotics of $\check{\theta}_t^N$ for $N \to \infty$

For t in a finite interval $[0, T]$ it holds that

$$\check{\theta}_t^N = \bar{\theta}_t^N + O_P(N^{-1/2})$$

with

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\theta}_t^N = -h(\bar{\theta}_t^N),$$
$$\bar{\theta}_0^N = \theta_0^N.$$

In old time:

$$\theta_k^N = \bar{\theta}_{t_k^N}^N + O_P(N^{-1/2})$$

for $k \in [0, k_T^N]$.

Topic of talk V

Weak convergence:

It holds for fixed T

$$\left(U_t^N : 0 \leq t \leq T\right) \rightarrow \left(X_t^N : 0 \leq t \leq T\right)$$

in distribution. Here:

$$U_t^N = \frac{\check{\theta}_t^N - \bar{\theta}_t^N}{\sqrt{\gamma_k^N}} \text{ for } t_k^N \leq t < t_{k+1}^N,$$

$$\mathrm{d}X_t^{N,i} = \bar{\alpha} X_t^{N,i} \mathrm{d}t - \sum_{j=1}^{d} \frac{\partial h_i}{\partial x_j}(\bar{\theta}_t^N) \cdot X_t^{N,j} \mathrm{d}t + \sum_{j=1}^{d} R_{ij}^{1/2}(\bar{\theta}_t^N) \mathrm{d}W_t^j,$$

for $i = 1, ..., d$, with $X_0 = U_0^N$ where $\bar{\alpha} = (2A(B+1))^{-1}$ and $R(\theta)$ covariance matrix of $H(\theta, \eta)$.

Topic of talk VI

For more details

- Nevel'son and Khas'minskiĭ (1973),
- Benveniste, Métivier, and Priouret (1990),
- Duflo (1997),
- Kushner and Yin (2003).

Topic of talk VII

Strong approximation

Put $\tau_N = \inf\{k : \|U^N_{t^N_k}\| \geq a_N\}$ with $a_N = C_a \ln(N)$.

Theorem

Choose $r > 6$ and $C_a, C_b > 0$. Assume $\|\theta^N_0\| \leq a_N N^{-1/2}$ and $C > 0$ large enough. Make higher order smoothness assumptions on H and h and moment conditions on $H(\theta_n, \eta)$.

Then for $N > 1$ there exist Robbins-Monro algorithms θ^N_k on $k \in \{1, 2, 3, ...\}$ and diffusion processes X^N_t on $t \in [0, \infty)$ with

$$\mathbb{P}\Big(\|(\theta^N_k - \bar{\theta}^N_{t^N_k}) - \gamma_k^{1/2} X_{t^N_k}\| \leq C(\ln(k+N))^r/(k+N)$$

$$\text{for } 0 \leq k \leq \tau_N \wedge N^{C_b} \Big) \to 1 \text{ for } N \to \infty.$$

Topic of talk VIII

Equivalent formulation in new time scale.

Theorem

Choose $r > 6$ and $C_a, C_b > 0$. Assume $\|\theta_0^N\| \leq a_N N^{-1/2}$ and $C > 0$ large enough. Make higher order smoothness assumptions on H and h and moment conditions on $H(\theta_n, \eta)$.

Then for $N > 1$ there exist scaled Robbins-Monro algorithms U_t^N on $t \in [0, \infty)$ and diffusion processes X_t^N on $t \in [0, \infty)$ with

$$\mathbb{P}\Big(\|U_t^N - X_t^N\| \leq C(\ln(k_t^N + N))^r/(k_t^N + N) \\ \text{for } 0 \leq t \leq t^N_{\tau_N \wedge N^{C_b}} \Big) \to 1 \text{ for } N \to \infty.$$

Ingredients of the proof

(1) Truncation: truncated processes V_t^N and Y_t^N.

(2) Discretisation: consider V_t^N and Y_t^N on a grid of points $t \in J_n$

(3) bounds for differences of transition densities of V_t^N and Y_t^N for neighbored points of J_n

(4) L_1 bounds for differences of the joint densities $(V_t^N : t \in J_n)$ and $(Y_t^N : t \in J_n)$

(5) Conclude that there exist processes V_t^N and Y_t^N with

$$\mathbb{P}\Big(V_t^N = Y_t^N \text{ for all } t \in J_n\Big) \to 1 \text{ for } N \to \infty.$$

(6) Conclude that the theorem holds with U_t^N and X_t^N replaced by V_t^N and Y_t^N, respectively.

(7) Conclude that the theorem holds (with U_t^N and X_t^N)

Remark on (4) $\implies$ (5)

(4) L_1 bounds for differences of the joint densities $(V_t^N : t \in J_n)$ and $(Y_t^N : t \in J_n)$

(5) Conclude that there exist processes V_t^N and Y_t^N with

$$\mathbb{P}\Big(V_t^N = Y_t^N \text{ for all } t \in J_n\Big) \to 1 \text{ for } N \to \infty.$$

Standard argument: For densities f and g with

$$\xi = \int \min\{f(x), g(x)\}\mathrm{d}x = 1 - \frac{1}{2}\int |f(x) - g(x)|\mathrm{d}x$$

there exist random variables X and Y with densities f and g, respectivlely, on the same probability space with

$$\mathbb{P}(X = Y) = \xi.$$

Remark on (5) $\implies$ (6)$\implies$ (7)

(5) Conclude that there exist processes V_t^N and Y_t^N with

$$\mathbb{P}\Big(V_t^N = Y_t^N \text{ for all } t \in J_n \Big) \to 1 \text{ for } N \to \infty.$$

(6) Conclude that the theorem holds with U_t^N and X_t^N replaced by V_t^N and Y_t^N, respectively.

(7) Conclude that the theorem holds (with U_t^N and X_t^N)

(5) $\implies$ (6): The gird J_N is chosen such that

$$|k_t^N - k_{t'}^N| \leq C(\ln k_t^N)^r$$

for two neighbored elements t and t' of J_N

(6) $\implies$ (7): direct arguments

Remaining steps

(1) Truncation: truncated processes V_t^N and Y_t^N.

(2) Discretisation: consider V_t^N and Y_t^N on a grid of points $t \in J_n$

(3) bounds for differences of transition densities of V_t^N and Y_t^N for neighbored points of J_n

(2) $\surd$

Remains (1) and (3).

Remark on (1)

(1) Truncation: truncated processes V_t^N and Y_t^N.

Can write:

$$\begin{aligned} U^N_{t^N_{k+1}} &= U^N_{t^N_k} + G_N(t^N_k, U^N_{t^N_k})\gamma^N_{k+1} U^N_{t^N_k} \\ &\quad -\sqrt{\gamma^N_{k+1}}\xi\left(\bar\theta^N_{t^N_k} + \sqrt{\gamma^N_k}U^N_{t^N_k}, \eta^N_{k+1}\right) + \beta^N_{k+1}, \end{aligned}$$

where

$$G_N(t^N_k, x) = \alpha^N_{t^N_k} I - \sqrt{\frac{\gamma^N_k}{\gamma^N_{k+1}}}\int_0^1 \mathcal{D}h\left(\bar\theta^N_{t^N_k} + \delta x\sqrt{\gamma^N_k}\right)\mathrm{d}\delta.$$

and $\beta^N_k \to 0$, $\alpha^N_{t^N_k} \to (2A(B+1))^{-1}$ and $\mathcal{D}h$ derivative of h ($d \times d$ matrix).

Remark on (1)

This representation motivates the following truncated process $V^N_{t_k}$:

$$\begin{aligned} V^N_{t^N_{k+1}} &= V^N_{t^N_k} + F_N(t^N_k, V^N_{t^N_k})\gamma^N_{k+1} V^N_{t^N_k} \\ &\quad - \sqrt{\gamma^N_{k+1}}\,\xi\left(\bar{\theta}^N_{t^N_k} + \sqrt{\gamma^N_k}\chi_N(V^N_{t^N_k}), \eta^N_{k+1}\right), \end{aligned}$$

where

$$F_N(t,x) = (2A(B+1))^{-1}\, I - \int_0^1 \mathcal{D}h(\bar{\theta}_t + \delta\chi_N(x)\sqrt{\gamma^N_1}\,)\mathrm{d}\delta,$$

χ_N smooth with $\chi_N(x) = x$ for $\|x\| \leq a_N$ and $\chi_N(x) = 0$ for $\|x\| \geq 2a_N$.

Major change: Replace x by $\chi_N(x)$ at some places.

Truncation: Make the structure of V^N_t simple if $\|V^N_t\|$ is large.

Remark on (1)

Smoothly "truncated" diffusion Y_t^N:

$$\mathrm{d}Y_t^N = F_N(t, Y_t^N)Y_t^N\mathrm{d}t + R^{1/2}(\bar{\theta}_t^N)\mathrm{d}W_t$$

with the same function F_N as defined in the last slide.

Remark on (3)

Itt remains to comment on

(3) bounds for differences of transition densities of V_t^N and Y_t^N for neighbored points of J_n

For this step we will apply the following result:

Remark on (3)

Theorem

For $s < t$ and $x, z \in \mathbb{R}^d$ it holds that

$$\int_{\mathbb{R}^d} |p_N - q_N|(s,t,x,z)\mathrm{d}z \\ \leq C(\ln N)^2 N^{-1/2}\sqrt{t-s},$$

where

$$q_N(s,t,x,z) \text{ conditional density of } Y_t^N \text{ at } z \text{ given } Y_s^N = x,$$
$$p_N(s,t,x,z) \text{ conditional density of } V_t^N \text{ at } z \text{ given } V_s^N = x.$$

The theorem is an extension of a result in Konakov, M. and Huang (2025) with stricter bounds and modified processes Y_t^N and V_t^N.

Remark on (3)

For the proof we make use of the **parametrix method**.
Short introduction to parametrix approach (Levi (1907), McKean and Singer (1967)): Consider SDE in $\mathbb{R}^d$ of the form

$$Z_t = z + \int_0^t b(s, Z_s)\mathrm{d}s + \int_0^t \sigma(s, Z_s)\mathrm{d}W_s.$$

Additionally, consider the equation with coefficients "frozen" at the point y and put $\tilde{p}(s,t,x,y) = p^y(s,t,x,y)$ where $p^z(s,t,x,y)$ is the Gaussian transition density of

$$\tilde{Z}_v = \tilde{Z}_0 + \int_0^v b(u, z)\mathrm{d}u + \int_0^v \sigma(u, z)\mathrm{d}W_u.$$

Remark on (3)

We now use the backward and forward Kolmogorov equations:

$$\frac{\partial \tilde{p}}{\partial s} + \tilde{L}\tilde{p} = 0, \quad \frac{\partial p}{\partial s} + Lp = 0, \quad -\frac{\partial \tilde{p}}{\partial t} + \tilde{L}^* \tilde{p} = 0, \quad -\frac{\partial p}{\partial t} + L^* p = 0.$$

Together with the initial conditions

$$\tilde{p}(t, t, x, y) = \delta(x - y) \text{ and } p(t, t, x, y) = \delta(x - y).$$

we can write the basic equality for the parametrix method:

Remark on (3)

$$
\begin{aligned}
& p(s,t,x,y) - \tilde{p}(s,t,x,y) \\
& \quad = \int_s^t \mathrm{d}u \frac{\partial}{\partial u}\left[\int_{\mathbb{R}^k} p(s,u,x,z)\tilde{p}(u,t,z,y)\mathrm{d}z\right] \\
& \quad = \int_s^t \mathrm{d}u \int_{\mathbb{R}^k} \Big[\tilde{p}(u,t,z,y)L^* p(s,u,x,z) \\
& \qquad\qquad - p(s,u,x,z)\tilde{L}\tilde{p}(u,t,z,y)\Big]\mathrm{d}z \\
& \quad = \int_s^t \mathrm{d}u \int_{\mathbb{R}^k} \left[p(s,u,x,z)(L-\tilde{L})\tilde{p}(u,t,z,y)\right]\mathrm{d}z \\
& \quad = p \otimes H(s,t,x,y)
\end{aligned}
$$

where $H = [L - \tilde{L}]\tilde{p}$ and the convolution type binary operation $\otimes$

$$
(f \otimes g)(s,t,x,y) = \int_s^t \mathrm{d}u \int_{\mathbb{R}^d} f(s,u,x,z)g(u,t,z,y)\mathrm{d}z.
$$

Remark on (3)

Iterative application of $p - \tilde{p} = p \otimes H$ gives an infinite series

$$p = \sum_{r=0}^{\infty} \tilde{p} \otimes H^{(r)},$$

where $\tilde{p} \otimes H^{(0)} = \tilde{p}$ and $\tilde{p} \otimes H^{(r+1)} = (\tilde{p} \otimes H^{(r)}) \otimes H$ for $r = 0, 1, 2, \dots$. An important property of this representation is that it allows us to express the **non-Gaussian density p in terms of Gaussian densities** $\tilde{p}$. For our diffusion Y_t^N we get:

$$q_N(t, s, x, y) \;=\; \sum_{r=0}^{\infty} \tilde{q}_N \otimes H_N^{(r)}(t, s, x, y)$$

with an appropriate choice of a Gaussian density $\tilde{q}_N$ and operator H_N.

Remark on (3)

In Konakov, M. (2000) a similar series representation has been proposed for discrete time Markov processes. For the transition densities of V_t^N it is given by

$$p_N(t_l^N, t_k^N, x, y) \;\; = \;\; \sum_{r=0}^{N} \tilde{p}_N \otimes_N \mathcal{K}_N^{(r)}(t_l^N, t_k^N, x, y),$$

with $\tilde{p}_N$ density of sum of independent variables and discretized time convolution

$$(f \otimes_N g)(t_i^N, t_j^N, x, y) = \sum_{k=i}^{j-1} \gamma_{k+1}^N \int_{\mathbb{R}^d} f(t_i^N, t_k^N, x, z) g(t_k^N, t_j^N, z, y) \mathrm{d}z$$

and $g \otimes_N \mathcal{K}_N^{(r)} = (g \otimes_N \mathcal{K}_N^{(r-1)}) \otimes_N \mathcal{K}_N$ with $g \otimes_N \mathcal{K}_N^{(0)} = g$.

Remark on (3)

Back to our theorem that states a bound on $|p_N - q_N|(s,t,x,y)$.

Idea: compare:

$$p_N(t_l^N, t_k^N, x, y) \;=\; \sum_{r=0}^{N} \tilde{p}_N \otimes_N K_N^{(r)}(t_l^N, t_k^N, x, y),$$

and

$$q_N(t, s, x, y) \;=\; \sum_{r=0}^{\infty} \tilde{q}_N \otimes H_N^{(r)}(t, s, x, y)$$

Remark on (3)

We have to compare in a series expansion:

(a) operator $H_N^{(r)}(t, s, x, y)$ with operator $K_N^{(r)}$ for all number r of concolutions,

(b) discretized time convolution $\otimes_N$ with continuous version $\otimes$,

(c) $\tilde{p}_N$ density of sum of independent variables with Gaussian densities $\tilde{q}_N$

Con (a) and (b) need many technical considerations and many smoothness assumptions on H, h and density of H.

Pro Need no conderations on the transition densities of Markov processes and diffusions.

Thank you!

To Do Strong approximations for Robbins-Monro algorithm

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, \eta_{n+1}),\ \theta_0 \in \mathbb{R}^d,$$

with other choices of γ_k, e.g. $\gamma_k = \frac{A}{k^\beta + B}$ with $A > 0, B \geq 0$ and $\frac{1}{2} < \beta < 1$.