On Riemannian Stochastic Approximation Schemes with Fixed Step-Size

Alain Durmus, **Pablo Jiménez**, Éric Moulines, Salem Said
alain.durmus@ens-paris-saclay.fr, pablo.jimenez-moreno@polytechnique.edu,
eric.moulines@polytechnique.edu, salem.said@u-bordeaux.fr

1. Motivation: Riemannian spaces

■ Solving the **root-finding problem**:

$$\text{find } \theta \in \Theta \text{ satisfying } h(\theta) = 0 \qquad (1)$$

where Θ is a complete and connected Riemannian manifold, $\mathrm{T}\Theta$ is the tangent bundle, the **mean-field**, $h : \Theta \to \mathrm{T}\Theta$, is a smooth vector field.

■ Our goals :

- approximate a solution iteratively ⇝ first-order method.
- use the **geometry** of Θ ⇝ curved space.
- find convergence bounds & asymptotic results.

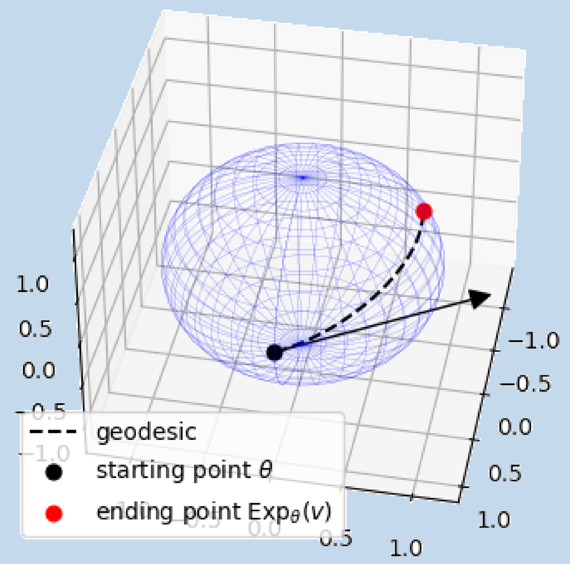
2. Presentation of the scheme

■ SA scheme to approximate (1) ⇝ extension of the **Robbins-Monro** algorithm for Riemannian manifolds [2, 1], for any $n \in \mathbb{N}$,

$$\theta_{n+1} = \mathrm{Exp}_{\theta_n}\{\eta H_{\theta_n}(X_{n+1})\} ,$$

where:

- $H_{\theta_n}(X_{n+1}) = h(\theta_n) + e_{\theta_n}(X_{n+1})$ is a noisy observation of h,
- $\mathbb{E}[e_\theta(X_1)] = 0$ with a bounded second moment,
- $\eta > 0$ is a step-size,
- $(X_n)_{n\in\mathbb{N}}$ random i.i.d. process on $(\mathsf{X}, \mathcal{X})$,
- $\theta^\star \in \Theta$ is a solution to (1),



- $\mathrm{Exp} : \mathrm{T}\Theta \to \Theta$ is the Riemannian exponential, roughly $\mathrm{Exp}_\theta(v) = \theta + v$.
- Extra assumptions ⇝ regularity conditions on e,

⇝ Lipschitz gradient Lyapunov function V s.t.

1. $\|h(\theta)\|^2 + C_2\langle \mathrm{grad}\, V(\theta), h(\theta)\rangle_\theta \leqslant C_1$, *i.e.* $-\mathrm{grad}\, V$ and h are “close”,
2. $\langle \mathrm{grad}\, V(\theta), h(\theta)\rangle_\theta \leqslant -\lambda V(\theta)\mathbb{1}_{\Theta\setminus \mathrm{B}(\theta^\star, r)}(\theta)$, *i.e.* h points towards $\theta^\star$ when far from it.

■ Special case $h = -\mathrm{grad}\, f$ corresponds to SGD optimization for a smooth $f : \Theta \to \mathbb{R}$.

3. The scheme is a Markov chain

■ For any $\eta > 0$, $(\theta_n)_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain.

■ Lyapunov conditions + Taylor expansion gives

Theorem 1 *(Ergodicity & stationary measures)*
There exists $\bar\eta > 0$ s.t. for any $\eta \in (0, \bar\eta]$, the Markov chain is geometrically **ergodic** and admits a unique **stationary measure** μ^η. In addition

$$\lim_{\eta\to 0} \mu^\eta \overset{d}{=} \delta_{\theta^\star} ,$$

where $\delta_{\theta^\star}$ is the Dirac mass on $\theta^\star$.

4. Choosing the Lyapunov function

■ For Euclidean spaces, e.g. $\mathbb{R}^d$, typically $V_0 : \theta \mapsto \|\theta - \theta^\star\|^2$. It is smooth, its gradient points to the solution & is Lipschitz.

■ Riemannian schemes cannot use ρ^2_Θ as the Hessian is not bounded ⇝ not Lipschitz gradient. Instead, interpolate V_0 with another function.

■ Multiplying with a bump function χ on $\theta^\star$,

$$V_2 : \theta \mapsto \chi(\theta)\rho^2_\Theta(\theta^\star, \theta) + (1 - \chi(\theta))C .$$

■ Linearizing V_0, when far from $\theta^\star$ for some $\delta > 0$,

$$V_1 : \theta \mapsto \delta^2\{(\rho_\Theta(\theta^\star, \theta)/\delta)^2 + 1\}^{1/2} - \delta^2 .$$

5. Variance estimation at equilibrium

■ For **SGD**, we derive an expansion of the **mean error at stationarity**:

$$\int_\Theta \|\mathrm{grad}\, f(\theta)\|^2_\theta \mathrm{d}\mu^\eta(\theta) = (\eta/2)\,\mathrm{Tr}\,(\mathrm{Hess}_{\theta^\star} f\, \Sigma(\theta^\star)) + o(\eta) ,$$

where $\Sigma(\theta)$ is the covariance matrix of $e_\theta(X_1)$.
⇝ The square norm of grad f is **linear** w.r.t. the step-size η.

■ **Central limit theorem** to find the rate of convergence of $(\mu^\eta)_{\eta\in(0,\bar\eta]}$. Assume:

- Θ is a **Hadamard** manifold, *i.e.* complete and simply connected, with non-positive curvature,
- $e_\theta(X_1)$ has a finite moment of order $2 + \varepsilon$,
- a Taylor expansion of h at $\theta^\star$, roughly $h(\theta) = \mathbf{A}(\theta^\star - \theta) + o(\|\theta^\star - \theta\|)$.

Define a **renormalized** family of measures $(\bar\nu^\eta)_{\eta\in(0,\bar\eta]}$ by a factor $\eta^{1/2}$, *i.e.* for any $A \in \mathcal{B}(\mathrm{T}_{\theta^\star}\Theta)$: $\bar\nu^\eta(A) = \mu^\eta(\mathrm{Exp}_{\theta^\star}[\eta^{1/2}A])$.

Theorem 2 *(Central Limit Theorem)*
The family $(\bar\nu^\eta)_{\eta\in(0,\bar\eta]}$ converges weakly to $\mathrm{N}(0, \mathbf{V})$, where $\mathbf{V}$ is solution to the Lyapunov equation

$$\mathbf{AV} + \mathbf{VA}^\top = \Sigma(\theta^\star).$$

6. Applications

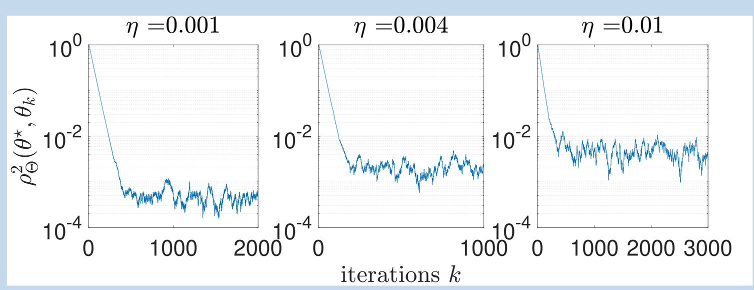
■ SGD **without boundedness** conditions. Assume

- f twice continuously differentiable & Lipschitz gradient,
- f is λ_f-strongly geodesically convex.

⇝ We obtain the **exponential forgetting** of the initial condition, with $O(\eta)$ oscillations:

$$\mathbb{E}\left[f(\theta_n) - f(\theta^\star)\right] \leqslant (1 - \eta\lambda_f/2)^n \left[f(\theta_0) - f(\theta^\star)\right] + \eta\, C . \qquad (2)$$

Fig. 1: Paths of the algorithm



■ With weaker assumptions, we study the **distance-like** function $D^2_\Theta(\theta_1, \theta_2) = \rho^2_\Theta(\theta_1, \theta_2)/[1 + \rho^2_\Theta(\theta_1, \theta_2)]$.

- Assume f is geodesically quasi-convex: $-\langle \mathrm{Exp}^{-1}_\theta(\theta^\star), \mathrm{grad}\, f(\theta)\rangle_\theta \geqslant \tilde\lambda_f V_1(\theta)$,

⇝ Convergence as $O(1/n)$ until $O(\eta)$ oscillation:

$$n^{-1}\textstyle\sum_{k=0}^{n-1} \mathbb{E}\left[D^2_\Theta(\theta^\star, \theta_k)\right] \leqslant 4V_1(\theta_0)\Big/\left(n\eta\tilde\lambda_f\right) + \eta\, C . \qquad (3)$$

■ We study & implement the Riemannian **barycenter problem**: for a distribution π on Θ, minimize $f_\pi : \theta \mapsto (1/2)\int_\Theta \rho^2_\Theta(\theta, \nu)\pi(\mathrm{d}\nu)$,
⇝ $\mathrm{grad}\, f_\pi(\theta) = -\int_\Theta \mathrm{Exp}^{-1}_\theta(\nu)\pi(\mathrm{d}\nu)$.

Fig. 2: Paths of the algorithm

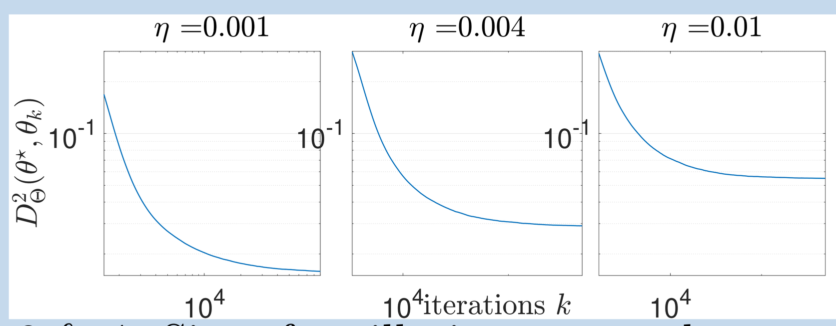
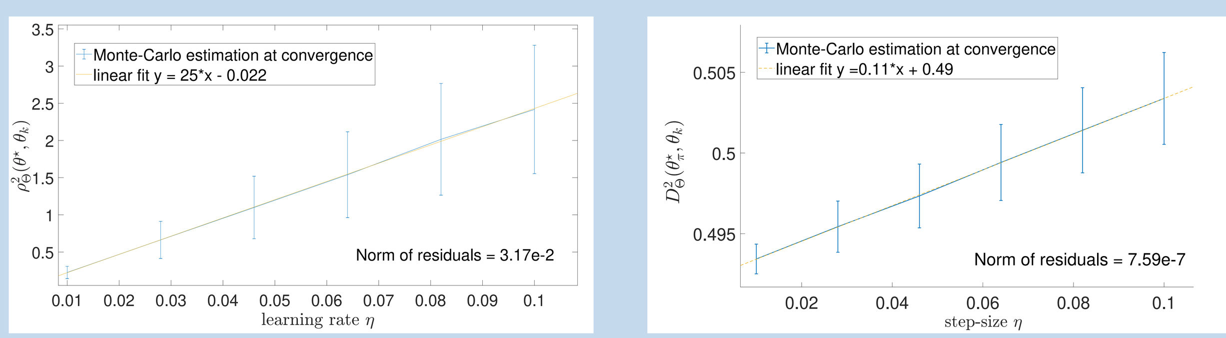


Fig. 3 & 4: Size of oscillations w.r.t. the step-size η



7. Experiments on the barycenter

On $\Theta = \mathrm{Sym}^+_{50}(\mathbb{R}) \subset \mathbb{R}^{50\times 50}$, the SPD manifold.

- Discrete case: $\pi = M_\pi^{-1}\sum_{i=1}^{M_\pi} \delta_{\bar\theta_i}$.

Apply (2), see Fig. 1 & 3.

- Continuous case: tame grad f_π by taking $H_\theta(X) = (1/2)\mathrm{Exp}^{-1}_\theta(X^{(1)})\{\rho^2_\Theta(\theta, X^{(2)})/2 + 1\}^{-1/2}$, where $X^{(1)}, X^{(2)} \sim \pi$ are i.i.d. copies.

Apply (3), see Fig. 2 & 4.

Bibliography

[1] S. Bonnabel. “Stochastic gradient descent on Riemannian manifolds”. In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2217–2229.

[2] H. Robbins and S. Monro. “A stochastic approximation method”. In: *The Annals of mathematical Statistics* 22.3 (1951), pp. 400–407.