Wasserstein Control of Mirror Langevin Monte Carlo

Kelvin Shuangjian Zhang†, Gabriel Peyré‡, Jalal Fadili§, Marcelo Pereyra¶

† ‡ CNRS and Département de Mathématiques et Applications, École Normale Supérieure / Université PSL, Paris, France   Contact: † szhang@ens.fr

§ Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, France   ¶ School of Mathematical and Computer Sciences, Heriot-Watt University, UK

GOAL

GOAL: Sample from a probability distribution π supported on $\mathcal{X} \subset \mathbb{R}^p$ in a high dimensional setting (i.e., for a large p).

APPLICATIONS: Bayesian inference, generative modeling, etc.

KNOWN: $f \stackrel{\text{def.}}{=} -\log(\frac{d\pi}{d\mathbf{x}})$. ($f \in C^2(\mathcal{X})$)

(Euclidean) Langevin Monte Carlo

- (Overdamped) Langevin dynamics:

$$d\mathbf{X}_t = -\nabla f(\mathbf{X}_t)dt + \sqrt{2}d\mathbf{B}_t, \tag{LD}$$

where $\{\mathbf{B}_t\}_{t\geq 0}$ is a standard p-dimensional Brownian motion.

- Euler-Maruyama discretization:

$$\mathbf{X}_{k+1} = \mathbf{X}_k - h_{k+1}\nabla f(\mathbf{X}_k) + \sqrt{2h_{k+1}}\boldsymbol{\xi}_{k+1}; \quad k = 0, 1, 2, \ldots \tag{LMC}$$

- See (Dalalyan and Karagulyan, 2019) for a review

Assumptions: $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \|\mathbf{x} - \mathbf{x}'\|_2^2$; (strong convexity)
$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\|_2 \leq M \|\mathbf{x} - \mathbf{x}'\|_2$. (Lipschitz smoothness)

Convergence: Let μ_k be the law of $\mathbf{X}_k$, $W_2(\cdot,\cdot)$ the Wasserstein 2-distance, and $h_k \equiv h \leq 2/(m+M)$, then

$$W_2(\mu_k, \pi) \leq (1 - mh)^k W_2(\mu_0, \pi) + 1.65\left(\frac{M}{m}\right) p^{\frac{1}{2}} h^{\frac{1}{2}}. \tag{1}$$

Iteration Complexity: needs $K_\varepsilon \approx \frac{M^2 p}{m^3 \varepsilon^2} \log\left(\frac{1}{\varepsilon}\right)$ steps to reach ε-precision.

Examples excluded by the above assumptions

Case 1: Gamma distribution. $f = \sum_{i=1}^p (1 - a_i)\log(x_i) + b_i x_i + C$;

Case 2: Dirichlet distribution. $f = (1 - a_1)\log(x) + (1 - a_2)\log(1 - x) + C$.

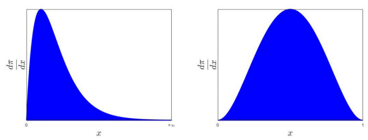

Figure 1: Probability density functions. LEFT: Gamma distribution. RIGHT: Dirichlet distribution.

Relaxation of strong convexity and Lipschitz-smoothness

- Equivalent assumptions to *strong convexity* and *Lipschitz-smoothness*: Let $\phi = \frac{\mathbf{x}^2}{2}$,

$$\begin{aligned} &\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}') \rangle \geq m \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2^2; \\ &\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\|_2 \leq M \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2. \end{aligned} \tag{2}$$

- Relative strong convexity and Lipschitz-smoothness:
$\exists$ some $C^2(\mathcal{X})$ *Legendre-type* convex entropy ϕ on $\mathcal{X}$, such that

$$\begin{aligned} &\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}') \rangle \geq m \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2^2; \\ &\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\|_2 \leq M \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2. \end{aligned} \tag{3}$$

Hessian Riemannian Langevin Monte Carlo algorithm

- Riemannian Langevin dynamics on Hessian Manifold $(\mathcal{X}, D^2\phi)$ (Roberts and Stramer (2002)):

$$d\mathbf{X}_t = \left(\theta(\mathbf{X}_t) - [D^2\phi(\mathbf{X}_t)]^{-1}\nabla f(\mathbf{X}_t)\right)dt + \sqrt{2[D^2\phi(\mathbf{X}_t)]^{-1}}d\mathbf{B}_t, \tag{4}$$

where $\theta(\mathbf{X}_t) \stackrel{\text{def.}}{=} -[D^2\phi(\mathbf{X}_t)]^{-1}\mathrm{Tr}\left(D^3\phi(\mathbf{X}_t)[D^2\phi(\mathbf{X}_t)]^{-1}\right)$.

- Denoting $\mathbf{Y}_t \stackrel{\text{def.}}{=} \nabla\phi(\mathbf{X}_t)$, SDE (4) reads

$$d\mathbf{Y}_t = -\nabla f \circ \nabla\phi^*(\mathbf{Y}_t)dt + \sqrt{2[D^2\phi^*(\mathbf{Y}_t)]^{-1}}d\mathbf{B}_t, \tag{5}$$

here $\phi^*(\mathbf{y}) \stackrel{\text{def.}}{=} \sup_{\mathbf{x}\in\mathcal{X}}\langle \mathbf{x}, \mathbf{y}\rangle - \phi(\mathbf{x})$ is the Legendre-Fenchel conjugate of ϕ.

- The Euler-Maruyama discretization of SDE (5):

$$\mathbf{Y}_{k+1} \stackrel{\text{def.}}{=} \mathbf{Y}_k - h_{k+1}\nabla f(\nabla\phi^*(\mathbf{Y}_k)) + \sqrt{2h_{k+1}[D^2\phi^*(\mathbf{Y}_k)]^{-1}}\boldsymbol{\xi}_{k+1}. \tag{6}$$

- Using $\mathbf{X}_k = \nabla\phi^*(\mathbf{Y}_k)$, we derive the Hessian Riemannian Langevin Monte Carlo algorithm

$$\mathbf{X}_{k+1} \stackrel{\text{def.}}{=} \nabla\phi^*\left(\nabla\phi(\mathbf{X}_k) - h_{k+1}\nabla f(\mathbf{X}_k) + \sqrt{2h_{k+1}[D^2\phi(\mathbf{X}_k)]}\boldsymbol{\xi}_{k+1}\right). \tag{HRLMC}$$

- Ignoring the randomness term, **HRLMC** algorithm reduces to the *Mirror Descent* algorithm:

$$\mathbf{X}_{k+1} \stackrel{\text{def.}}{=} \nabla\phi^*\left(\nabla\phi(\mathbf{X}_k) - h_{k+1}\nabla f(\mathbf{X}_k)\right). \tag{Mirror Descent}$$

Other assumptions on ϕ and f

- Self-concordance-like condition on ϕ:

$$\sqrt{2}\left\|D^2\phi(\mathbf{x})^{\frac{1}{2}} - D^2\phi(\mathbf{x}')^{\frac{1}{2}}\right\|_F \leq \kappa \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2. \tag{7}$$

- Bound on the commutator of $D^2\phi$ and $D^2 f$:

$$\left\|\left[(D^2\phi(\mathbf{x}))^{-1}, D^2 f(\mathbf{x})\right]\right\|_2 \leq \delta. \tag{8}$$

- Moment condition on the Hessian of ϕ:

$$R \stackrel{\text{def.}}{=} \mathbf{E}_{\mathbf{X}\sim\pi}\left[\left\|D^2\phi(\mathbf{X})\right\|_2\right] = \int_{\mathcal{X}} \left\|D^2\phi(\mathbf{x})\right\|_2 e^{-f(\mathbf{x})}d\mathbf{x} < +\infty. \tag{9}$$

- Interaction of key parameters:

$$\tilde{\kappa} \stackrel{\text{def.}}{=} \sqrt{\kappa^2 + \frac{\delta(4M+\delta)}{2(m+M)}} < \sqrt{2m}. \tag{10}$$

Note: Our theory covers not only all the strongly convex and Lipschitz-smooth cases, but also new cases not known in the literature.

Wasserstein Distance

- Let d be the Riemannian distance associated with the squared Hessian metric $[D^2\phi(\mathbf{x})]^2$. Then the natural associated geometric distance on the space of probability distributions on $\mathcal{X}$ is the Wasserstein distance

$$W_{2,\phi}^2(\mu,\nu) \stackrel{\text{def.}}{=} \inf_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\nu} \mathbf{E}\left[d^2(\mathbf{x},\mathbf{x}')\right] = \inf_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\nu} \mathbf{E}\left[\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2^2\right].$$

Note: When $\phi(\mathbf{x}) = \|\mathbf{x}\|^2/2$, one recovers the standard W_2 distance used in the Euclidean Langevin Monte Carlo (1).

Main Result

Theorem (Constant step-size)**.** Under the assumptions (3),(7)–(10), assume $h_k \equiv h$ is sufficiently small. Then

$$W_{2,\phi}(\mu_k,\pi) \leq \rho^k W_{2,\phi}(\mu_0,\pi) + h^{\frac{3}{2}}(1-\rho)^{-1}p^{\frac{1}{2}}M^{\frac{1}{2}}R^{\frac{1}{2}}\left(1.65\sqrt{M} + \kappa/\sqrt{3}\right) + h(1-\rho)^{-1}p^{\frac{1}{2}}\kappa R^{\frac{1}{2}},$$

where the contraction ratio $\rho \stackrel{\text{def.}}{=} \sqrt{(1-mh)^2 + h\tilde{\kappa}^2} < 1$.

- **Contraction:** Under vanishing step-sizes, the **HRLMC** algorithm contracts toward a Wasserstein ball centered at the target distribution π with radius

$$r_0 \stackrel{\text{def.}}{=} \frac{2\kappa p^{\frac{1}{2}}R^{\frac{1}{2}}}{2m - \tilde{\kappa}^2}.$$

Note: when $\phi = \frac{\mathbf{x}^2}{2}$, $r_0 = 0$ implies convergence.

- **Iteration Complexity:** $K_\varepsilon \approx \frac{pMR\left(\sqrt{M}+\kappa\right)^2}{(2m-\tilde{\kappa}^2)^3}\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)$ steps to reach $(r_0+\varepsilon)$-precision.

Numerics

- Dirichlet distribution $d\pi \propto x^2(1-x)^2dx$ on 1D Simplex:

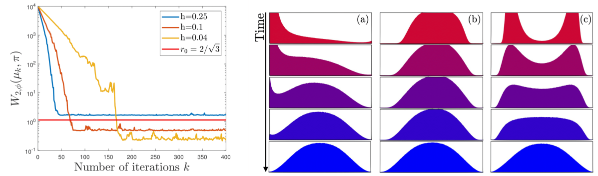

Figure 2: Left: Evolution in time of the sampling error for various constant step-sizes. A horizontal line at $r_0 = \frac{2}{\sqrt{3}}$ materializes the size of the bias term. Right: Visual display of the evolution of the empirical distribution of $\mathbf{X}_k$ for three different initializations.

- Dirichlet distribution $d\pi \propto x_1^2 x_2^2(1 - x_1 - x_2)^2 dx_1 dx_2$ on 2D Simplex:

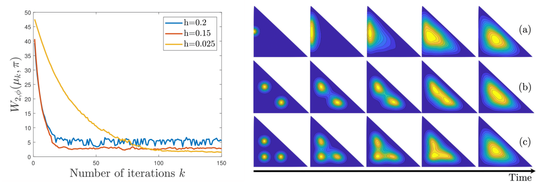

Figure 3: Left: evolution in time of the sampling error for various constant step-sizes. Right: visual display of the evolution of the empirical distribution of $\mathbf{X}_k$ shown as contour plots at different times, for three different initializations.

Conclusion:

- Contributions:
  - (a) First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.
  - (b) Our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance for the quadratic case.
  - (c) Numerics also support our theory.
- Future work:
  - (a) We conjecture that the bias term is inevitable. How to prove it?
  - (b) What is a provably good discretization of the Riemannian Langevin dynamics for general manifolds?