Wasserstein Control of Mirror Langevin Monte Carlo

Kelvin Shuangjian Zhang[*], Gabriel Peyré[†], Jalal Fadili[‡], Marcelo Pereyra[§]

[*] [†] CNRS and DMA, École Normale Supérieure, Université PSL, Paris, France
[‡] Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, France
[§] School of Mathematical and Computer Sciences, Heriot-Watt University, UK

[*] [†] This work is supported by the ERC project NORIA.

Optimal Transport and Applications, CMS 2020 Winter Meeting
December 6, 2020

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).

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

Applications: Bayesian inference, generative modeling, etc.

(Euclidean) Langevin Monte Carlo

(Overdamped) Langevin dynamics

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

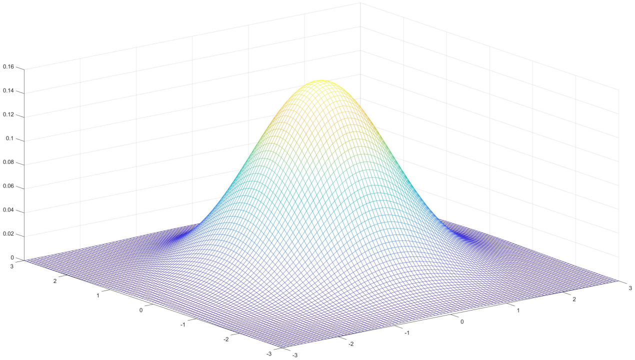
where $\{\mathbf{B}_t\}_{t\geq 0}$ is a standard p-dimensional Brownian motion.

(Euclidean) Langevin Monte Carlo

(Overdamped) Langevin dynamics

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

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

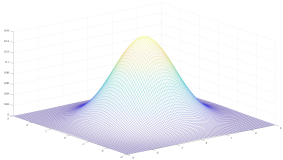


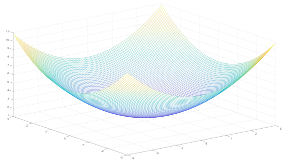
(Euclidean) Langevin Monte Carlo

(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.





$$\frac{d\pi}{dx} = C \cdot e^{-\frac{1}{2}(x-x^*)^T\Sigma(x-x^*)}$$

$$f(x) = -\log\left(\frac{d\pi}{dx}\right) = -\log C + \frac{1}{2}(x-x^*)^T\Sigma(x-x^*)$$

(Euclidean) Langevin Monte Carlo

(Overdamped) Langevin dynamics

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

Euler-Maruyama discretization

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

Convergence scheme

- The continuous dynamics $\mathbf{X}_t$ has π as its unique invariant measure.
- A discretization algorithm ensure the convergence of $\mathbf{X}_k$.

(Euclidean) Langevin Monte Carlo

Theorem (Dalalyan and Karagulyan, 2019)

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

"User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." Dalalyan and Karagulyan, *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

(Euclidean) Langevin Monte Carlo

Theorem (Dalalyan and Karagulyan, 2019)

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

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ; \qquad \text{(strong convexity)}$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 . \qquad \text{(Lipschitz smoothness)}$$

Then

"User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." Dalalyan and Karagulyan, *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

(Euclidean) Langevin Monte Carlo

Theorem (Dalalyan and Karagulyan, 2019)

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

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2; \qquad \textit{(strong convexity)}$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2. \qquad \textit{(Lipschitz smoothness)}$$

Then

Convergence

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

"User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." Dalalyan and Karagulyan, *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

(Euclidean) Langevin Monte Carlo

Theorem (Dalalyan and Karagulyan, 2019)

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

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ; \qquad \text{(strong convexity)}$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 . \qquad \text{(Lipschitz smoothness)}$$

Then

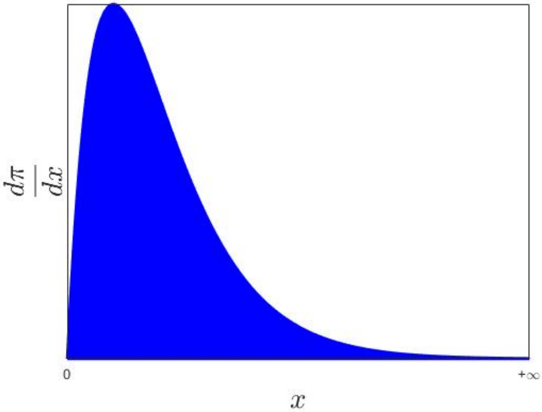
Iteration Complexity

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

"User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." Dalalyan and Karagulyan, *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

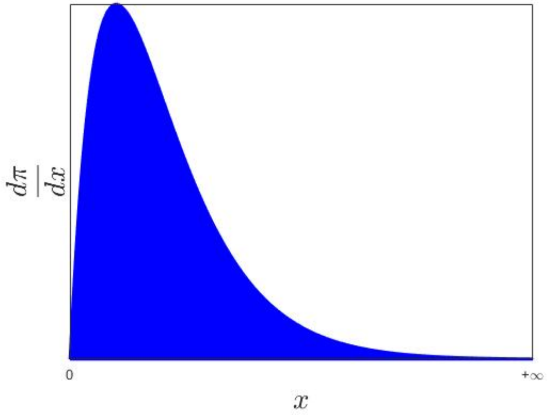
Gamma Distribution

- 1D-plot on density



Gamma Distribution

- 1D-plot on density



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

Gamma Distribution

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

Strong convexity and Lipschitz smoothness

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ;$$
$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 .$$

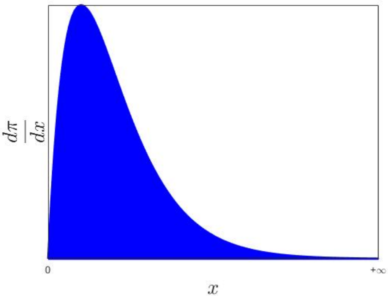
Gamma Distribution

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

Strong convexity and Lipschitz smoothness

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ;$$
$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 .$$

-

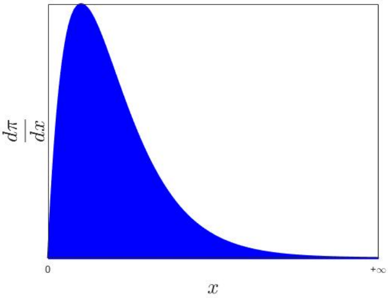


Gamma Distribution

Langevin Monte Carlo

$$\mathbf{X}_{k+1} = \mathbf{X}_k - h_{k+1}\nabla f(\mathbf{X}_k) + \sqrt{2h_{k+1}}\xi_{k+1}; \qquad k = 0, 1, 2, \ldots$$

-



Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ; \qquad \text{(strong-convexity)}$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 . \qquad \text{(Lipschitz smoothness)}$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \mathbf{x}' \rangle \geq m \left\| \mathbf{x} - \mathbf{x}' \right\|_2^2 ;$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 .$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\|\mathbf{x} - \mathbf{x}'\right\|_2^2;$

$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M \left\|\mathbf{x} - \mathbf{x}'\right\|_2.$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\| \nabla(\mathbf{x}^2/2) - \mathbf{x}' \right\|_2^2 ;$$

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \mathbf{x} - \mathbf{x}' \right\|_2 .$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2^2 ;$$

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M \left\|\mathbf{x} - \mathbf{x}'\right\|_2 .$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2^2 ;$$

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M \left\|\nabla(\mathbf{x}^2/2) - \mathbf{x}'\right\|_2 .$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2^2;$$
$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2. \tag{2}$$

Relaxation of Strong convexity and Lipschitz-smoothness

Previous assumptions

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\rangle \geq m \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2^2;$$
$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M \left\|\nabla(\mathbf{x}^2/2) - \nabla(\mathbf{x}'^2/2)\right\|_2. \quad (2)$$

Equivalent assumptions

Let $\phi = \frac{\mathbf{x}^2}{2}$,

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

Relaxation of Strong convexity and Lipschitz-smoothness

Equivalent assumptions

Let $\phi = \frac{\mathbf{x}^2}{2}$,

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

Current assumptions (weaker)

$\exists$ some $C^2(\mathcal{X})$ *Legendre-type* convex entropy ϕ on $\mathcal{X}$, such that

$$\begin{aligned} \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \rangle &\geq m \left\| \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \right\|_2^2 ; \\ \left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 &\leq M \left\| \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \right\|_2 . \end{aligned} \tag{4}$$

Relaxation of Strong convexity and Lipschitz-smoothness

Equivalent assumptions

Let $\phi = \frac{\mathbf{x}^2}{2}$,

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

Relative strong convexity and Lipschitz-smoothness

$\exists$ some $C^2(\mathcal{X})$ *Legendre-type* convex entropy ϕ on $\mathcal{X}$, such that

$$\begin{aligned} &\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \rangle \geq m \left\| \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \right\|_2^2 ; \\ &\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}') \right\|_2 \leq M \left\| \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{x}') \right\|_2 . \end{aligned} \tag{4}$$

Hessian Riemannian Langevin Monte Carlo (**HRLMC**)

- Riemannian Langevin dynamics on Hessian Manifold $(\mathcal{X}, D^2\phi)$[1]:

$$d\mathbf{X}_t = \left(\theta(\mathbf{X}_t) - [D^2\phi(\mathbf{X}_t)]^{-1}\nabla f(\mathbf{X}_t)\right)dt + \sqrt{2[D^2\phi(\mathbf{X}_t)]^{-1}}d\mathbf{B}_t, \quad (5)$$

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

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Hessian Riemannian Langevin Monte Carlo (**HRLMC**)

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

$$d\mathbf{X}_t = \big(\theta(\mathbf{X}_t) - [D^2\phi(\mathbf{X}_t)]^{-1}\nabla f(\mathbf{X}_t)\big)\,dt + \sqrt{2[D^2\phi(\mathbf{X}_t)]^{-1}}d\mathbf{B}_t, \quad (5)$$

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

- Denoting $\mathbf{Y}_t \overset{\text{def.}}{=} \nabla\phi(\mathbf{X}_t)$, SDE (5) reads

$$d\mathbf{Y}_t = -\nabla f \circ \nabla\phi^*(\mathbf{Y}_t)dt + \sqrt{2[D^2\phi^*(\mathbf{Y}_t)]^{-1}}d\mathbf{B}_t, \quad (6)$$

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

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Hessian Riemannian Langevin Monte Carlo (**HRLMC**)

- Denoting $\mathbf{Y}_t \stackrel{\text{def.}}{=} \nabla\phi(\mathbf{X}_t)$, SDE (5) reads

$$d\mathbf{Y}_t = -\nabla f \circ \nabla\phi^*(\mathbf{Y}_t)dt + \sqrt{2[D^2\phi^*(\mathbf{Y}_t)]^{-1}}d\mathbf{B}_t, \tag{6}$$

  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 (6) :

$$\mathbf{Y}_{k+1} \stackrel{\text{def.}}{=} \mathbf{Y}_k - h_{k+1}\nabla f(\nabla\phi^*(\mathbf{Y}_k)) + \sqrt{2h_{k+1}[D^2\phi^*(\mathbf{Y}_k)]^{-1}}\xi_{k+1}. \tag{7}$$

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

- The Euler-Maruyama discretization of SDE (6) :

$$\mathbf{Y}_{k+1} \stackrel{\text{def.}}{=} \mathbf{Y}_k - h_{k+1}\nabla f(\nabla\phi^*(\mathbf{Y}_k)) + \sqrt{2h_{k+1}[D^2\phi^*(\mathbf{Y}_k)]^{-1}}\xi_{k+1}. \quad (7)$$

- Using $\mathbf{X}_k = \nabla\phi^*(\mathbf{Y}_k)$, we derive the **HRLMC** algorithm

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

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Hessian Riemannian Langevin Monte Carlo (**HRLMC**)

- Using $\mathbf{X}_k = \nabla\phi^*(\mathbf{Y}_k)$, we derive the **HRLMC** algorithm

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

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Hessian Riemannian Langevin Monte Carlo (**HRLMC**)

$$\mathbf{X}_{k+1} \stackrel{\text{def.}}{=} \nabla\phi^*\Big(\nabla\phi(\mathbf{X}_k) - h_{k+1}\nabla f(\mathbf{X}_k) \qquad\qquad\qquad \Big).$$

(Mirror Descent)

[1] "Langevin diffusions and Metropolis-Hastings algorithms." Roberts and Stramer, *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Other assumptions on ϕ and f

- Self-concordance-like condition on ϕ:

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

Other assumptions on ϕ and f

- Self-concordance-like condition on ϕ:

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

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

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

Other assumptions on ϕ and f

- Self-concordance-like condition on ϕ:

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

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

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

- Moment condition on the Hessian of ϕ:

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

Other assumptions on ϕ and f

- Self-concordance-like condition on ϕ:

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

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

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

- Moment condition on the Hessian of ϕ:

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

- Interaction of key parameters:

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

Examples

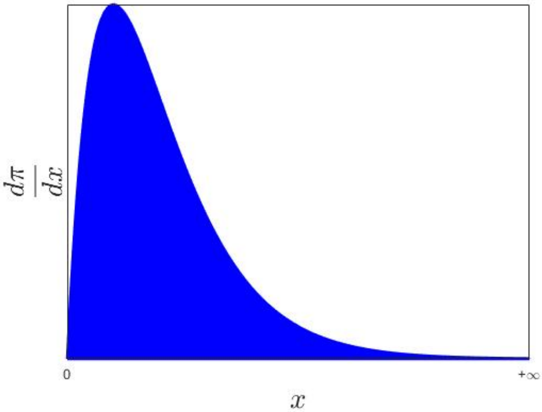
Case 1

Gamma distribution. $f = \sum_{i=1}^{p}(1 - a_i)\log(x_i) + b_i x_i + C$; take $\phi = -\sum_{i=1}^{p}\log(x_i)$. (Burg's entropy)

Examples

Case 1

Gamma distribution. $f = \sum_{i=1}^{p}(1 - a_i)\log(x_i) + b_i x_i + C$; take $\phi = -\sum_{i=1}^{p}\log(x_i)$. (Burg's entropy)



Examples

Case 1

Gamma distribution. $f = \sum_{i=1}^{p}(1 - a_i)\log(x_i) + b_i x_i + C$; take $\phi = -\sum_{i=1}^{p}\log(x_i)$. (Burg's entropy)

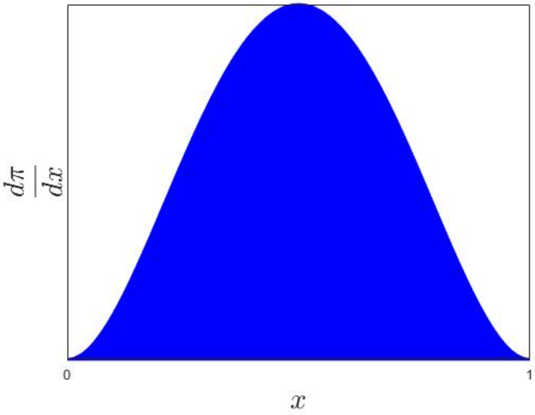
Case 2

Dirichlet distribution. $f = (1 - a_1)\log(x) + (1 - a_2)\log(1 - x) + C$; take $\phi = -\log(x) - \log(1 - x)$. (Burg's entropy on 1D Simplex)

Examples

Case 2

Dirichlet distribution. $f = (1 - a_1)\log(x) + (1 - a_2)\log(1 - x) + C$; take $\phi = -\log(x) - \log(1 - x)$. (Burg's entropy on 1D Simplex)



Examples

Case 1

Gamma distribution. $f = \sum_{i=1}^{p}(1-a_i)\log(x_i) + b_i x_i + C$; take $\phi = -\sum_{i=1}^{p}\log(x_i)$. (Burg's entropy)

Case 2

Dirichlet distribution. $f = (1-a_1)\log(x) + (1-a_2)\log(1-x) + C$; take $\phi = -\log(x) - \log(1-x)$. (Burg's entropy on 1D Simplex)

	Case 1	Case 2
m	$\min_i\{a_i - 1\}$	$\min\{a_1 - 1, a_2 - 1\}$
M	$\max_i\{a_i - 1\}$	$\max\{a_1 - 1, a_2 - 1\}$
κ	$\sqrt{2}$	$\sqrt{2}$
δ	0	0
R	$\sum_i (a_i-3)!/b_i^{a_i-2}$	$\frac{(a_1-3)!(a_2-1)!+(a_1-1)!(a_2-3)!}{(a_1+a_2-3)!}$
$\tilde{\kappa} < \sqrt{2m}$	$a_i > 2, \forall i$	$a_1, a_2 > 2$

Main result:

- Let d be the Riemannian distance associated with the squared Hessian metric $[D^2\phi(\mathbf{x})]^2$. Define

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

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:

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

Theorem (Z.-Peyré-Fadili-Pereyra, COLT2020)

Under the above assumptions, 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 $\rho \stackrel{\text{def.}}{=} \sqrt{(1-mh)^2 + h\tilde{\kappa}^2} < 1$.

Main result:

Theorem (Z.-Peyré-Fadili-Pereyra, COLT2020)

Under the above assumptions, 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 $\rho \overset{\text{def.}}{=} \sqrt{(1-mh)^2 + h\tilde{\kappa}^2} < 1.$

Contraction

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

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

Main result:

Theorem (Z.-Peyré-Fadili-Pereyra, COLT2020)

Under the above assumptions, 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 $\rho \overset{\text{def.}}{=} \sqrt{(1-mh)^2 + h\tilde{\kappa}^2} < 1$.

Contraction

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

$$r_0 \overset{\text{def.}}{=} \frac{2\kappa p^{\frac{1}{2}} R^{\frac{1}{2}}}{2m - \tilde{\kappa}^2} = 0, \text{ when } \phi = \frac{\mathbf{x}^2}{2}.$$

Main result:

Theorem (Z.-Peyré-Fadili-Pereyra, COLT2020)

Under the above assumptions, 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 $\rho \overset{\text{def.}}{=} \sqrt{(1-mh)^2 + h\tilde{\kappa}^2} < 1$.

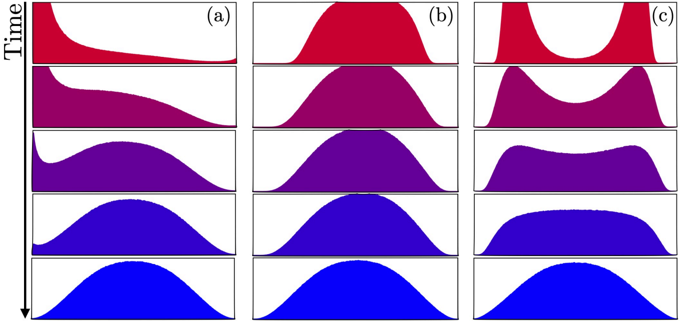
Iteration Complexity

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

Numerics

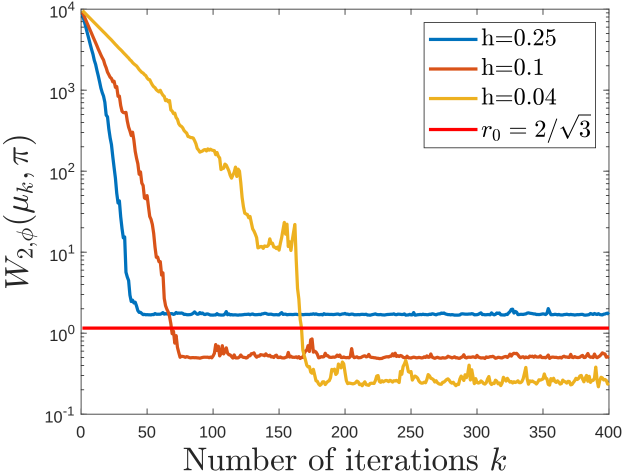
- Dirichlet distribution $d\pi \propto x^2(1-x)^2 dx$ on 1D Simplex:

  Visual display of the evolution of the empirical distribution of $\mathbf{X}_k$ for three different initializations: (a) Dirac measure at 10^{-4}; (b) uniform measure on $[0.3, 0.8]$; (c) two Dirac measures at 0.2 and 0.8.



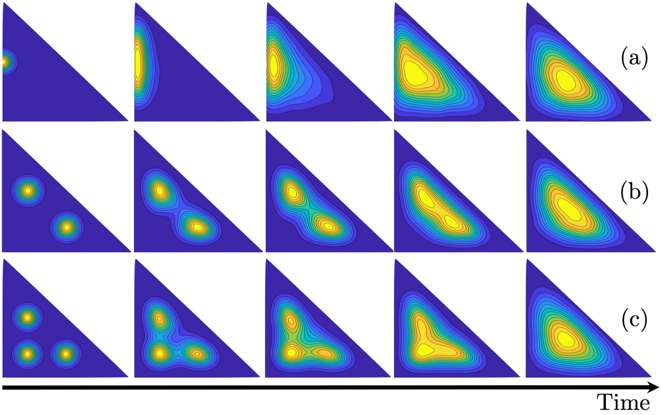
Numerics

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



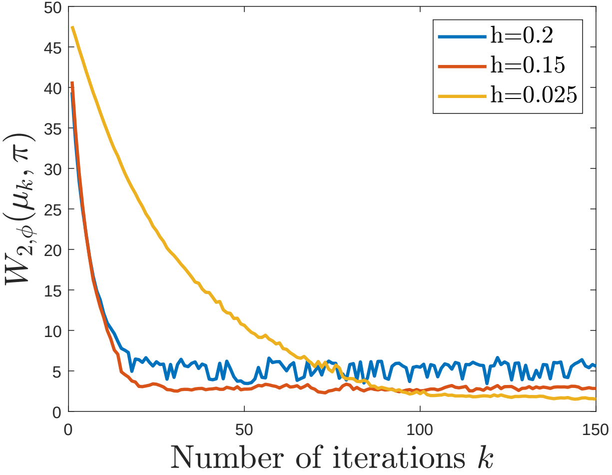
Numerics

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



Numerics

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



Conclusion:

Contributions

- First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.

Conclusion:

Contributions

- First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.
- Our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance for the quadratic case.

Conclusion:

Contributions

- First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.
- Our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance for the quadratic case.
- Numerics also support our theory.

Conclusion:

Contributions

- First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.
- Our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance for the quadratic case.
- Numerics also support our theory.

Future work

- We conjecture that the bias term is inevitable. How to prove it?

Conclusion:

Contributions

- First guarantees of HRLMC which show that it contracts into a Wasserstein ball centered at the desired invariant distribution.
- Our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance for the quadratic case.
- Numerics also support our theory.

Future work

- We conjecture that the bias term is inevitable. How to prove it?
- What is a provably good discretization of the Riemannian Langevin dynamics for general manifolds?

Thank you very much!