Langevin Monte Carlo for Log-Concave Densities

Sampling Algorithms and Geometries on Probability Distributions

Arnak Dalalyan
(joint with A. Karagulyan, L. Riou-Durand)

CREST, ENSAE Paris, IP Paris

The problem of sampling

- **The target density**

$$\pi(\boldsymbol{\theta}) \propto \exp\big(-f(\boldsymbol{\theta})\big), \qquad f:\mathbb{R}^p \to \mathbb{R},$$

$$\mu_q(\pi) = \bigg(\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^q\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\bigg)^{1/q} < +\infty, \quad q = 2.$$

- **Conditions on f**: gradient Lipschitz + convex

$$\big\|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\big\|_2 \leq M\|\boldsymbol{u}-\boldsymbol{v}\|_2 \qquad \& \qquad \nabla^2 f(\boldsymbol{u}) \succeq 0.$$

- **Example:** posterior of the multivariate logistic regression.
- **Goal**: find a distr. ν easy to sample and s.t. $W_q(\nu,\pi)$ is small.
- **Constant sampling:** if $\nu = \delta_0$, then $W_q(\nu,\pi) \leq \mu_q(\pi)$.
- **Equiv. of moments:** there is A_q s.t. for all log-concave π,

$$\mu_q(\pi) \leq A_q \mu_2(\pi)$$

(explicit expression leading to $A_3 \leq 3.5, A_4 \leq 4.6$).

First-order MCMC methods

Gradient oracle: assume that at any $\boldsymbol{\theta} \in \mathbb{R}^p$, we can evaluate $\nabla f(\boldsymbol{\theta})$.

- Langevin Monte Carlo (LMC)

$$\boldsymbol{\vartheta}_{k+1}^{\mathsf{LMC}} = \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}}) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \quad k = 0, 1, \ldots \quad \text{(LMC)}$$

where $\{\boldsymbol{\xi}_k\}$ is iid $\mathcal{N}_p(0, \mathbf{I}_p)$ indep. of $\boldsymbol{\vartheta}_0$. Set $\nu_k^{\mathsf{LMC}} = \mathcal{L}(\boldsymbol{\vartheta}_k^{\mathsf{LMC}})$.

- Langevin Monte Carlo with averaging (LMCa)

$$\boldsymbol{\vartheta}_k^{\mathsf{aLMC}} = \boldsymbol{\vartheta}_\tau^{\mathsf{LMC}}; \quad \tau \sim \mathtt{Unif}(1, \ldots, k), \quad k = 0, 1, \ldots \quad \text{(LMCa)}$$

and set $\nu_k^{\mathsf{aLMC}} = \mathcal{L}(\boldsymbol{\vartheta}_k^{\mathsf{aLMC}})$.

- Kinetic Langevin Monte Carlo (KLMC, aka underdamped LMC)

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} \\ \boldsymbol{\xi}_{k+1}^{(2)} \end{bmatrix} \quad \text{(KLMC)}$$

- Metropolis adjusted Langevin Algorithm (MALA)

First-order MCMC methods

Gradient oracle: assume that at any $\boldsymbol{\theta} \in \mathbb{R}^p$, we can evaluate $\nabla f(\boldsymbol{\theta})$.

- Langevin Monte Carlo (LMC)

$$\boldsymbol{\vartheta}_{k+1}^{\mathsf{LMC}} = \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}}) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \quad k = 0, 1, \ldots \quad \text{(LMC)}$$

  where $\{\boldsymbol{\xi}_k\}$ is iid $\mathcal{N}_p(0, \mathbf{I}_p)$ indep. of $\boldsymbol{\vartheta}_0$. Set $\nu_k^{\mathsf{LMC}} = \mathcal{L}(\boldsymbol{\vartheta}_k^{\mathsf{LMC}})$.

- Langevin Monte Carlo with averaging (LMCa)

$$\boldsymbol{\vartheta}_k^{\mathsf{aLMC}} = \boldsymbol{\vartheta}_\tau^{\mathsf{LMC}}; \quad \tau \sim \mathtt{Unif}(1, \ldots, k), \quad k = 0, 1, \ldots \quad \text{(LMCa)}$$

  and set $\nu_k^{\mathsf{aLMC}} = \mathcal{L}(\boldsymbol{\vartheta}_k^{\mathsf{aLMC}})$.

- Kinetic Langevin Monte Carlo (KLMC, aka underdamped LMC)

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} \\ \boldsymbol{\xi}_{k+1}^{(2)} \end{bmatrix} \quad \text{(KLMC)}$$

- Metropolis adjusted Langevin Algorithm (MALA)

An illustration and the objective

Sampling guarantees

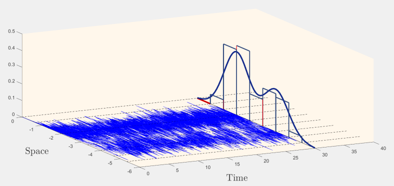

Figure: The blue lines represent different paths of a discretized Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density (the dark blue line).

Main goal: number of gradient evaluations that are sufficient to get ε-accuracy in W_q (especially for $q = 1$ and $q = 2$).

Mixing time of an approximate sampling algorithm Alg:

$$K_{\mathsf{Alg},W_q}(p,\varepsilon) = \min\left\{k \in \mathbb{N} : W_q(\nu_k^{\mathsf{Alg}}, \pi) \le \varepsilon\mu_q(\pi),\ \forall\, \pi \in \mathcal{P}\right\}.$$

Quick overview

Order of magnitude of the mixing time of various first-order samplers.

$\kappa = M\mu_2^2/p$	LMCa	MALA	α-LMC	α-KLMC
W_2	—	—	$\kappa p^2/\varepsilon^6$	$\kappa^{1.5} p^2/\varepsilon^5$
W_1	—	—	$\kappa p^2/\varepsilon^4$	$\kappa^{1.5} p^2/\varepsilon^3$
d_{TV}	$\kappa p^2/\varepsilon^4$ $\triangle$	$p^3(\kappa/\varepsilon)^{3/2}\diamond$	$\kappa^2 p^3/\varepsilon^4$ $\square$	—

- $\triangle$ behavior of the LMC with averaging [Durmus et al., 2019].
- $\diamond$ derived from [Dwivedi et al., 2018]
- $\square$ behavior of the LMC [Dalalyan, 2017].

First approach

Wasserstein from TV

- **Proposition** For any pair of probability measures (ν, ν'), and for any $q \geq 1$, we have:
$$W_q(\nu, \nu') \leq \inf_{r \geq q} \left\{ (\mu_r(\nu) + \mu_r(\nu'))\, d_{\mathrm{TV}}(\nu, \nu')^{\frac{1}{q} - \frac{1}{r}} \right\}.$$
- **Proof** optimal coupling $X \sim \nu$ and $Y \sim \nu'$ for the TV-distance: $d_{\mathrm{TV}}(\nu, \nu') = \mathrm{P}(X \neq Y)$ and $W_q(\nu, \nu') \leq \mathrm{E}[\|X - Y\|_2^q]^{1/q}$
- If π log-concave, $\mu_1(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{1+1/(r-1)}$ then
$$W_1(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$
- If π log-concave, $\mu_2(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{2+4/(r-2)}$ then
$$W_2(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$

First approach

Wasserstein from TV

- **Proposition** For any pair of probability measures (ν, ν'), and for any $q \geq 1$, we have:

$$W_q(\nu, \nu') \leq \inf_{r \geq q} \left\{ (\mu_r(\nu) + \mu_r(\nu'))\, d_{\mathrm{TV}}(\nu, \nu')^{\frac{1}{q} - \frac{1}{r}} \right\}.$$

- **Proof** optimal coupling $X \sim \nu$ and $Y \sim \nu'$ for the TV-distance: $d_{\mathrm{TV}}(\nu, \nu') = \mathrm{P}(X \neq Y)$ and $W_q(\nu, \nu') \leq \mathrm{E}[\|X - Y\|_2^q 1_{X \neq Y}]^{1/q}$. Use the Hölder inequality to conclude.

- If π log-concave, $\mu_1(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{1+1/(r-1)}$ then

$$W_1(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$

- If π log-concave, $\mu_2(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{2+4/(r-2)}$ then

$$W_2(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$

First approach

Wasserstein from TV

- **Proposition** For any pair of probability measures (ν, ν'), and for any $q \geq 1$, we have:

$$W_q(\nu, \nu') \leq \inf_{r \geq q} \left\{ (\mu_r(\nu) + \mu_r(\nu'))\, d_{\mathrm{TV}}(\nu, \nu')^{\frac{1}{q} - \frac{1}{r}} \right\}.$$

- **Proof** optimal coupling $X \sim \nu$ and $Y \sim \nu'$ for the TV-distance: $d_{\mathrm{TV}}(\nu, \nu') = \mathrm{P}(X \neq Y)$ and
- If π log-concave, $\mu_1(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{1+1/(r-1)}$ then

$$W_1(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$

- If π log-concave, $\mu_2(\nu) \lesssim \mu_2(\pi)$ and $d_{\mathrm{TV}}(\nu, \pi) \leq (\varepsilon / A_r)^{2+4/(r-2)}$ then

$$W_2(\nu, \pi) \leq \varepsilon \mu_2(\pi).$$

Quick overview

Order of magnitude of the mixing time of various first-order samplers.

$\kappa = M\mu_2^2/p$	LMCa	MALA	α-LMC	α-KLMC
W_2	$A_r^8 p^2/\varepsilon^{8+8/r}$	$A_r^3 p^3/\varepsilon^{3+3/r}$	p^2/ε^6	p^2/ε^5
W_1	$A_r^4 p^2/\varepsilon^{4+4/r}$	$A_r^{3/2} p^3/\varepsilon^{3/2+1/r}$	p^2/ε^4	p^2/ε^3
d_{TV}	p^2/ε^4 $\triangle$	$p^3/\varepsilon^{3/2}$ $\diamond$	p^3/ε^4 $\square$	—

- dependence on p is not better than for the penalized LMC.
- The results for MALA involve very large constants.
- Good dependence on ε requires large r, but then the constants A_r blow up

Second approach: Poincaré inequality

- π satisfies the Poincaré inequality if

$$\mathsf{Var}_\pi[h] \le \mathcal{C}_\mathsf{P} \int \|\nabla h(\boldsymbol{\theta})\|_2^2\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

  We call $\mathcal{C}_\mathsf{P}$ the Poincaré constant.

- For any log-concave π, $\mathcal{C}_\mathsf{P} < \infty$.

- The Langevin diffusion satisfies ([Chewi et al., 2020, Lehec, 2021, Vempala and Wibisono, 2019]

$$W_2(\nu_t, \pi)^2 \le 2\mathcal{C}_\mathsf{P}\chi^2(\nu_0, \pi)e^{-t/\mathcal{C}_\mathsf{P}}, \qquad t \ge 0.$$

Theorem [Lehec, 2021, Thm 3] If π is log-concave and f is L-Lipschitz, then LMC with step-size $h \le p/L^2$ satisfies

$$W_2(\nu_k, \pi) \le 4\mathcal{C}_\mathsf{P}\chi^2(\nu_0, \pi)e^{-kh/\mathcal{C}_\mathsf{P}} + 3Lkh^{3/2}\sqrt{p}.$$

If $\boldsymbol{\vartheta}_0 \sim \mathcal{N}(x_*, (p/L^2)\mathsf{I})$ then $\log \chi^2(\nu_0, \pi) \le p + p\log(L^2\mathcal{C}_\mathsf{P}/p)$.

Quick overview

Order of magnitude of the mixing time of various first-order samplers.

$\kappa = M\mu_2^2/p$	LMCa	MALA	α-LMC	α-KLMC
W_2	$A_r^8 p^2/\varepsilon^{8+8/r}$	$A_r^3 p^3/\varepsilon^{3+3/r}$	p^2/ε^6	p^2/ε^5
W_1	$A_r^4 p^2/\varepsilon^{4+4/r}$	$A_r^{3/2} p^3/\varepsilon^{3/2+1/r}$	p^2/ε^4	p^2/ε^3
d_{TV}	p^2/ε^4 $\triangle$	$p^3/\varepsilon^{3/2}$ $\diamond$	p^3/ε^4 $\square$	—

- Lehec's result leads to

$$K_{LMC,W_2} = \Theta\left(\frac{\mathcal{C}_{\mathrm{P}}^3 L^2 p^2}{\varepsilon^4}\right)$$

- Mathematically elegant result, but dependence on $\mathcal{C}_{\mathrm{P}}$ is annoying.
- f global-Lipschitz assumption might be violated.

Third approach

Adding a quadratic penalty

- If f is α-strongly log-concave, then one has $\mathcal{C}_{\mathsf{P}} \leq 1/\alpha$ and $W_2(\nu_t, \pi) \leq e^{-\alpha t} W_2(\nu_0, \pi)$.
- Define the str. convex surrogate $f_\alpha(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \alpha \|\boldsymbol{\theta}\|_2^2/2$ and
$$\pi_\alpha(\boldsymbol{\theta}) := \frac{e^{-f_\alpha(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f_\alpha(\boldsymbol{v})} d\boldsymbol{v}}.$$
- **Proposition** We have
$$d_{\mathsf{TV}}(\pi, \pi_\alpha) \leq \alpha\, \mu_2^2(\pi) \qquad W_q^q(\pi, \pi_\alpha) \leq C_q \alpha \mu_2(\pi)^{q+2}.$$
In particular, $C_1 \leq 22$ and $C_2 \leq 111$.
- Define α-LMC as LMC for f_α.
- Use the triangle inequality
$$\mathrm{dist}(\nu_{k,\alpha}^{\mathsf{Alg}}, \pi) \leq \mathrm{dist}(\nu_{k,\alpha}^{\mathsf{Alg}}, \pi_\alpha) + \mathrm{dist}(\pi_\alpha, \pi). \tag{1}$$

Main result for α-LMC

Theorem

Suppose that the potential f is convex and M-Lipschitz. Let $q \in [1,2]$. Then, for every $\alpha \leq M/20$ and $h \leq 1/(M+\alpha)$, we have

$$W_q(\nu_K^{\alpha\text{-LMC}}, \pi) \leq \underbrace{\sqrt{\mu_2}(1-\alpha h)^{K/2}}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{(2.1hMp/\alpha)^{1/2}}_{\text{discretization error}} + \underbrace{\Big(C_q \alpha \mu_2^{q+2}\Big)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}}.$$

- **FAQ**: why α is in the discretization error as well ?
- Optimizing wrt to α and h, we get

$$K_{\alpha\text{-LMC},W_1} \leq 5 \times 10^4 M \frac{\mu_2^2 p}{\varepsilon^4} \log(100/\varepsilon)$$

$$K_{\alpha\text{-LMC},W_2} \leq 4 \times 10^6 M \frac{\mu_2^2 p}{\varepsilon^6} \log(100/\varepsilon).$$

Main result for α-KLMC

Theorem

Suppose f is convex and M-Lipschitz. Let $q \in [1,2]$. Then for every $\alpha \le M/20$, $\gamma \ge \sqrt{M+2\alpha}$ and $h \le \alpha/(4\gamma(M+\alpha))$, we have

$$W_q(\nu_K^{\alpha\text{-KLMC}}, \pi) \le \underbrace{\sqrt{2\mu_2}\left(1 - \frac{3\alpha h}{4\gamma}\right)^K}_{\text{error finite time}} + \underbrace{1.5Mp^{1/2}(h/\alpha)}_{\text{discretization error}} + \underbrace{\left(C_q \alpha \mu_2^{q+2}\right)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}} .$$

Optimizing wrt to α and h, we get

$$K_{\alpha\text{-KLMC}, W_1}(p, \varepsilon) \le 9.2 \times 10^3 (M\mu_2^2)^{3/2}(p^{1/2}/\varepsilon^3)\log(150/\varepsilon)$$
$$K_{\alpha\text{-KLMC}, W_2}(p, \varepsilon) \le 4.4 \times 10^5 (M\mu_2^2)^{3/2}(p^{1/2}/\varepsilon^5)\log(150/\varepsilon).$$

Conclusions and outlook

- Non-asymptotic sampling guarantees for convex (but not strongly convex) and gradient-Lipschitz potentials.
- The simple convexification trick is still "competitive".
- Faster rates are obtained under additional smoothness (Hessian Lipschitz) assumptions.
- Current work: variable step-size h_k + variable penalty α_k + randomized mid-point discretization [Shen and Lee, 2019].
- Time-continuous bound in [Karagulyan and Dalalyan, 2020]: if $\alpha(t) = 1/(t + \mu_2^2(\pi))$ then

$$W_2(\nu_t, \pi) \leq \frac{10\mu_2^2\big(1 + \log(1 + t/\mu_2^2)\big)}{\sqrt{t + \mu_2^2}}.$$

Conclusions and outlook

- Non-asymptotic sampling guarantees for convex (but not strongly convex) and gradient-Lipschitz potentials.
- The simple convexification trick is still "competitive".
- Faster rates are obtained under additional smoothness (Hessian Lipschitz) assumptions.
- Current work: variable step-size h_k + variable penalty α_k + randomized mid-point discretization [Shen and Lee, 2019].
- Time-continuous bound in [Karagulyan and Dalalyan, 2020]: if $\alpha(t) = 1/(t + \mu_2^2(\pi))$ then

$$W_2(\nu_t, \pi) \leq \frac{10\mu_2^2\big(1 + \log(1 + t/\mu_2^2)\big)}{\sqrt{t + \mu_2^2}}.$$

Interested in a postdoc in Paris: send me an email.

References I

S. Chewi, T. L. Gouic, C. Lu, T. Maunu, P. Rigollet, and A. Stromme. Exponential ergodicity of mirror-Langevin diffusions, 2020.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *J. R. Stat. Soc. B*, 79:651 – 676, 2017.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte-Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73–1, 2019.

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR, 2018.

References II

A. Karagulyan and A. Dalalyan. Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets. *Advances in Neural Information Processing Systems*, 33, 2020.

J. Lehec. The langevin monte carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pages 8094–8106. 2019.