On sampling from a log-concave density using kinetic Langevin diffusions

Lionel Riou-Durand
University of Warwick

Young Reasearchers' Meeting
14/01/2022

Joint work with

Arnak Dalalyan

ENSAE - Paris

Sampling

- Goal: to sample from a distribution Π on $\mathbb{R}^d$

Sampling

- Goal: to sample from a distribution Π on $\mathbb{R}^d$
- Sampling = transforming a uniform distribution into Π

$$(U_1, ..., U_k) \sim \mathcal{U}_{[0,1]^k}$$

Sampling

- Goal: to sample from a distribution Π on $\mathbb{R}^d$
- Sampling = transforming a uniform distribution into Π

$$(U_1, ..., U_k) \sim \mathcal{U}_{[0,1]^k}$$

- Our task: finding a map $T : [0, 1] \mapsto \mathbb{R}^d$ such that

$$\boldsymbol{\theta}_k \triangleq T(U_1, ..., U_k), \qquad \mathscr{D}(\boldsymbol{\theta}_k) = \Pi.$$

Sampling

- Goal: to sample from a distribution Π on $\mathbb{R}^d$
- Sampling = transforming a uniform distribution into Π

$$(U_1, ..., U_k) \sim \mathcal{U}_{[0,1]^k}$$

- Our task: finding a map $T : [0, 1] \mapsto \mathbb{R}^d$ such that

$$\boldsymbol{\theta}_k \triangleq T(U_1, ..., U_k), \qquad \mathscr{D}(\boldsymbol{\theta}_k) = \Pi$$

Approximate sampling

- A relaxed objective: to sample approximately from Π.

Approximate sampling

- A relaxed objective: to sample approximately from Π.
- A sampling algorithm = a sequence of tractable maps (T_k) such that for some $k \geq 1$ not too large

$$\boldsymbol{\theta}_k \triangleq T_k(U_1, ..., U_k), \qquad \mathscr{D}(\boldsymbol{\theta}_k) \approx \Pi.$$

Approximate sampling

- A relaxed objective: to sample approximately from Π.
- A sampling algorithm = a sequence of tractable maps (T_k) such that for some $k \geq 1$ not too large

$$\boldsymbol{\theta}_k \triangleq T_k(U_1, ..., U_k), \qquad \mathscr{D}(\boldsymbol{\theta}_k) \approx \Pi.$$

- How to measure the approximation error ?

Approximate sampling

- A relaxed objective: to sample approximately from Π.
- A sampling algorithm = a sequence of tractable maps (T_k) such that for some $k \geq 1$ not too large

$$\boldsymbol{\theta}_k \triangleq T_k(U_1, ..., U_k), \qquad \mathscr{D}(\boldsymbol{\theta}_k) \approx \Pi.$$

- How to measure the approximation error ?
- How many iterations shall I run to reach a given precision level ?

Mixing time

- Let $\mathrm{d}(\cdot, \cdot)$ be a metric between probability distributions, and $\varepsilon > 0$ be a precision level. Define the mixing time

$$\mathcal{K}_\varepsilon = \inf\{k \geq 0 : \mathrm{d}(\mathscr{D}(\boldsymbol{\theta}_k), \Pi) \leq \varepsilon\}$$

Mixing time

- Let $\mathrm{d}(\cdot, \cdot)$ be a metric between probability distributions, and $\varepsilon > 0$ be a precision level. Define the mixing time

$$\mathcal{K}_\varepsilon = \inf\{k \geq 0 : \mathrm{d}(\mathscr{D}(\boldsymbol{\theta}_k), \Pi) \leq \varepsilon\}$$

- Total Variation metric

$$\mathrm{TV}(\mu, \nu) = \sup_{\mathrm{A} \in \mathscr{B}(\mathbb{R}^{\mathrm{d}})} |\mu(\mathrm{A}) - \nu(\mathrm{A})|$$

Mixing time

- Let $\mathrm{d}(\cdot, \cdot)$ be a metric between probability distributions, and $\varepsilon > 0$ be a precision level. Define the mixing time

$$\mathcal{K}_\varepsilon = \inf\{k \geq 0 : \mathrm{d}(\mathscr{D}(\boldsymbol{\theta}_k), \Pi) \leq \varepsilon\}$$

- Total Variation metric

$$\mathrm{TV}(\mu, \nu) = \sup_{\mathrm{A} \in \mathscr{B}(\mathbb{R}^{\mathrm{d}})} |\mu(\mathrm{A}) - \nu(\mathrm{A})|$$

- Wasserstein distance (w.r.t the Euclidean norm $\|.\|$)

$$\mathscr{C}(\mu, \nu) = \{\text{distributions on } \mathbb{R}^d \times \mathbb{R}^d \text{ with marginals } \mu \text{ and } \nu\}$$

$$\mathrm{W}_{\mathrm{q}}(\mu, \nu) = \inf_{\gamma \in \mathscr{C}(\mu, \nu)} \mathbb{E}_{(\mathrm{X}, \mathrm{Y}) \sim \gamma} \left[\|\mathrm{X} - \mathrm{Y}\|^{\mathrm{q}} \right]$$

Geometric ergodicity

- A sampling sequence $(\boldsymbol{\theta}_k)$ is called geometrically ergodic if there exists $C > 0$ and $\rho \in (0, 1)$ such that

$$\mathrm{d}(\mathscr{D}(\boldsymbol{\theta}_k), \Pi) \leq C\rho^k.$$

Geometric ergodicity

- A sampling sequence $(\boldsymbol{\theta}_k)$ is called geometrically ergodic if there exists $C > 0$ and $\rho \in (0, 1)$ such that

$$\mathrm{d}(\mathscr{D}(\boldsymbol{\theta}_k), \Pi) \leq C\rho^k.$$

- A qualitative guarantee: not explicit if C and ρ are unknown.

Mixing time

- How to compare two sampling algorithms on the computational basis ?

Mixing time

- How to compare two sampling algorithms on the computational basis ?
- How does the accuracy of a sampler scale with the dimension ?

Mixing time

- How to compare two sampling algorithms on the computational basis ?
- How does the accuracy of a sampler scale with the dimension ?
- Can be answered by a quantitative study of the mixing time.

Mixing time

- How to compare two sampling algorithms on the computational basis ?
- How does the accuracy of a sampler scale with the dimension ?
- Can be answered by a quantitative study of the mixing time.

Focus on log-concave distributions

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map such that

$$\int_{\mathbb{R}^d} \exp\{-f(\boldsymbol{\theta})\} d\boldsymbol{\theta} < \infty.$$

Focus on log-concave distributions

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map such that

$$\int_{\mathbb{R}^d} \exp\{-f(\boldsymbol{\theta})\} d\boldsymbol{\theta} < \infty.$$

- Goal: to sample approximately from Π with density

$$\pi(\boldsymbol{\theta}) \propto \exp\{-f(\boldsymbol{\theta})\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^d.$$

Focus on log-concave distributions

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map such that

$$\int_{\mathbb{R}^d} \exp\{-f(\boldsymbol{\theta})\} d\boldsymbol{\theta} < \infty.$$

- Goal: to sample approximately from Π with density

$$\pi(\boldsymbol{\theta}) \propto \exp\{-f(\boldsymbol{\theta})\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^d.$$

- Log-concave distributions: a convenient framework for quantitative results.

A parallel with optimization

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map, which has a unique minimizer $\boldsymbol{\theta}^*$.

A parallel with optimization

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map, which has a unique minimizer $\boldsymbol{\theta}^*$.
- Goal: to approximate $\boldsymbol{\theta}^*$ with an optimization algorithm (e.g. gradient descent).

A parallel with optimization

- Assume that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex map, which has a unique minimizer $\boldsymbol{\theta}^*$.
- Goal: to approximate $\boldsymbol{\theta}^*$ with an optimization algorithm (e.g. gradient descent).
- Most quantitative results (e.g. Boyd and Vandenberghe 2004) established for (strongly-)convex functions.

Quantitative results for sampling

- Let $(\boldsymbol{W}_t)$ be the standard Brownian motion on $\mathbb{R}^d$.

Quantitative results for sampling

- Let $(\boldsymbol{W}_t)$ be the standard Brownian motion on $\mathbb{R}^d$.
- Focus on sampling algorithms based on time discretizations of diffusion processes.

Quantitative results for sampling

- Let $(\boldsymbol{W}_t)$ be the standard Brownian motion on $\mathbb{R}^d$.
- Focus on sampling algorithms based on time discretizations of diffusion processes.
- Overdamped Langevin diffusion

$$\mathrm{d}\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)dt + \sqrt{2}\mathrm{d}\boldsymbol{W}_t.$$

Kinetic Langevin diffusion

- Damping parameter $\gamma \geq 0$, a.k.a friction.
- Langevin SDE for $t \geq 0$:

$$\mathrm{d}\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{V}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_t \\ -\nabla f(\boldsymbol{X}_t) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \mathbf{0}_d \\ -\gamma \boldsymbol{V}_t \, \mathrm{d}t + \sqrt{2\gamma} \, \mathrm{d}\boldsymbol{W}_t \end{bmatrix}.$$

- Langevin dynamics = Hamiltonian dynamics with momentum refreshment continuously induced by a Brownian Motion $(\boldsymbol{W}_t)_{t \geq 0}$.
- Invariant measure on $\mathbb{R}^{2d}$

$$\pi_* = \pi \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$$

.

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)
- Assume that f is m-strongly convex, and

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)
- Assume that f is m-strongly convex, and
  - (C_1) smoothness: ∇f is M-Lipschitz

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)
- Assume that f is m-strongly convex, and
  - (C_1) smoothness: ∇f is M-Lipschitz
  - (C_2) further smoothness: $\nabla^2 f$ is Lipschitz for spectral norm

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)
- Assume that f is m-strongly convex, and
  - (C_1) smoothness: ∇f is M-Lipschitz
  - (C_2) further smoothness: $\nabla^2 f$ is Lipschitz for spectral norm
- Explicit bounds on the contraction and discretization errors, for total variation, Kullback Leibler, Wasserstein distances.

A new interest for Langevin type algorithms

- Recent contributions (Dalalyan 2017, Durmus and Moulines 2017, Eberle et al. 2017, Cheng et al. 2017)
- Assume that f is m-strongly convex, and
  - (C_1) smoothness: ∇f is M-Lipschitz
  - (C_2) further smoothness: $\nabla^2 f$ is Lipschitz for spectral norm
- Explicit bounds on the contraction and discretization errors, for total variation, Kullback Leibler, Wasserstein distances.
- How does the mixing time scale with the dimension p, the precision level ε, and the condition number $\kappa = M/m$?

Results for the Wasserstein distance

- Mixing times for W_2 (Durmus and Moulines 2017, Cheng et al. 2017, Durmus et al. 2018, Dalalyan and Karagulyan 2019)

$\tilde{\mathcal{O}}$	C_1	C_2
LMC	$\kappa(p/\varepsilon^2)$	$\kappa^{3/2}(p/\varepsilon)$
KLMC	$\kappa^2(p^{1/2}/\varepsilon)$	-

- Recall: $\kappa = M/m$ is the condition number of the Hessian
  (C_1): $mI_p \preccurlyeq \nabla^2 f \preccurlyeq MI_p$
  (C_2): $\nabla^2 f$ is Lipschitz for the spectral norm

Our work

- Several objectives, in the continuation of Cheng et al. 2017
- A mixing rate for the Kinetic Langevin diffusion for any $\gamma > 0$.
- Sharper mixing times for KLMC by optimizing with respect to $\gamma > 0$.
- To reduce the discretization error under (C_2): Lipschitz smoothness of the Hessian matrix.

Results for the Wasserstein distance

- Mixing times for W_2 (previous works & Dalalyan and R-D 2020)

$\tilde{\mathcal{O}}$	C_1	C_2
LMC	$\kappa(p/\varepsilon^2)$	$\kappa^{3/2}(p/\varepsilon)$
KLMC	$\kappa^{3/2}(p^{1/2}/\varepsilon)$	$\kappa^{5/4}(p/\varepsilon)^{1/2}$

- Recall: $\kappa = M/m$ is the condition number of the Hessian
  (C_1): $mI_p \preccurlyeq \nabla^2 f \preccurlyeq MI_p$
  (C_2): $\nabla^2 f$ is Lipschitz for the spectral norm

Typical result

- Let ν_k the distribution of KLMC after k iterates.
- **Theorem:** *Assume that* $m \cdot \mathbf{I}_d \preccurlyeq \nabla^2 f \preccurlyeq M \cdot \mathbf{I}_d$, *then* $\forall\, \gamma \geq \sqrt{m + L}$, *and* $h \leq m/(4\gamma L)$, *we have*

$$W_2(\nu_k, \pi) \leq \underbrace{\sqrt{2}\Big(1 - \frac{3mh}{4\gamma}\Big)^k W_2(\nu_0, \pi)}_{\text{contraction error}} + \underbrace{\frac{Mh\sqrt{2p}}{m}}_{\text{discretization error}}$$

References I

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Che+17] Xiang Cheng et al. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *arXiv preprint arXiv:1707.03663* (2017).

[Dal17] Arnak S Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

References II

[DK19] Arnak S Dalalyan and Avetik Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stochastic Processes and their Applications* (2019).

[DM17] Alain Durmus and Eric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.

[DR20] Arnak S Dalalyan and Lionel R-D. "On sampling from a log-concave density using kinetic Langevin diffusions". In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

References III

[Dur+18] Alain Durmus et al. "Analysis of Langevin Monte Carlo via convex optimization". In: *arXiv preprint arXiv:1802.09188* (2018).

[Ebe+17] Andreas Eberle et al. "Couplings and quantitative contraction rates for Langevin dynamics". In: *arXiv preprint arXiv:1703.01617* (2017).

Thank you !