Mirrored Langevin Dynamics

Ya-Ping Hsieh
https://lions.epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

NeurIPS Spotlight

[Dec 6th, 2018]

Joint work with

Ali Kavis, Paul Rolland, Volkan Cevher @ LIONS

lions@epfl

Introduction

- Task: given a target distribution $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$, generate samples from μ.
  - ▷ Fundamental in machine learning/statistics/computer science/etc.

Introduction

- Task: given a target distribution $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$, generate samples from μ.
  - Fundamental in machine learning/statistics/computer science/etc.

- A scalable framework: First-order sampling (assuming access to ∇V).

  Step 1. **Langevin Dynamics**

$$\mathrm{d}\mathbf{X}_t = -\nabla V(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \quad \Rightarrow \quad X_\infty \sim e^{-V}.$$

Introduction

- Task: given a target distribution $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$, generate samples from μ.
  - ▷ Fundamental in machine learning/statistics/computer science/etc.

- A scalable framework: First-order sampling (assuming access to ∇V).

  Step 1. **Langevin Dynamics**

$$\mathrm{d}\mathbf{X}_t = -\nabla V(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \quad \Rightarrow \quad X_\infty \sim e^{-V}.$$

  Step 2. **Discretize**

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \beta_k \nabla V(\mathbf{x}^k) + \sqrt{2\beta_k}\boldsymbol{\xi}^k$$

  - ▷ β_k step-size, $\boldsymbol{\xi}^k$ standard normal
  - ▷ strong analogy to gradient descent method

Recent progress: Unconstrained distributions are easy

- State-of-the-art: When $\mathrm{dom}(V) = \mathbb{R}^d$,

Assumption	$\mathcal{W}_2$	d_{TV}	KL	Literature
$LI \succeq \nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-1}d\right)$	[Cheng and Bartlett, 2017] [Dalalyan and Karagulyan, 2017] [Durmus et al., 2018]
$LI \succeq \nabla^2 V \succeq 0$	-	$\tilde{O}\left(\epsilon^{-4}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	[Durmus et al., 2018]

Note: $\mathcal{W}_2(\mu_1,\mu_2) \coloneqq \sqrt{\inf_{X\sim\mu_1, Y\sim\mu_2} \mathbb{E}\|X-Y\|^2}$, $\quad d_{\mathsf{TV}}(\mu_1,\mu_2) \coloneqq \sup_{A \text{ Borel}} |\mu_1(A) - \mu_2(A)|$

Recent progress: Unconstrained distributions are easy

∘ State-of-the-art: When $\text{dom}(V) = \mathbb{R}^d$,

Assumption	$\mathcal{W}_2$	d_{TV}	KL	Literature
$LI \succeq \nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-1}d\right)$	[Cheng and Bartlett, 2017] [Dalalyan and Karagulyan, 2017] [Durmus et al., 2018]
$LI \succeq \nabla^2 V \succeq 0$	-	$\tilde{O}\left(\epsilon^{-4}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	[Durmus et al., 2018]

Note: $\mathcal{W}_2(\mu_1,\mu_2) \coloneqq \sqrt{\inf_{X\sim\mu_1, Y\sim\mu_2} \mathbb{E}\|X-Y\|^2}, \quad d_{\mathsf{TV}}(\mu_1,\mu_2) \coloneqq \sup_{A \text{ Borel}} |\mu_1(A)-\mu_2(A)|$

∘ What about **constrained** distributions?

▷ include many important applications, such as Latent Dirichlet Allocation (LDA).

A challenge: Constrained distributions are hard

◦ When $\mathrm{dom}(V)$ is compact, convergence rates deteriorate significantly.

Assumption	$\mathcal{W}_2$ or KL	d_{TV}	Literature
$LI \succeq \nabla^2 V \succeq mI$	?	$\tilde{O}\left(\epsilon^{-6} d^5\right)$	[Brosse et al., 2017]
$LI \succeq \nabla^2 V \succeq 0$	?	$\tilde{O}\left(\epsilon^{-6} d^5\right)$	[Brosse et al., 2017]

▷ *cf.*, when V is unconstrained, $\tilde{O}(\epsilon^{-4} d)$ convergence in d_{TV}.

▷ Projection is **not** a solution: slow rates [Bubeck et al., 2015], boundary issues.

Unconstrained optimization of constrained problems

- **Entropic Mirror Descent**: Unconstrained optimization within the simplex.

$$\min_{\mathbf{x} \in \Delta_d} V(\mathbf{x})$$

  ▷ Choose h to be the entropic mirror map, $h^\star$ its dual

Unconstrained optimization of constrained problems

- **Entropic Mirror Descent**: Unconstrained optimization within the simplex.

$$\min_{\mathbf{x}\in\Delta_d} V(\mathbf{x})$$

  - ▷ Choose h to be the entropic mirror map, $h^\star$ its dual
  - ▷ Mirror vs primal image: $\mathbf{y} = \nabla h(\mathbf{x}) \Leftrightarrow \mathbf{x} = \nabla h^\star(\mathbf{y})$

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \beta_k \nabla V(\mathbf{x}^k) \;\Rightarrow\; \text{no projection since } \mathrm{dom}(h^\star) = \mathbb{R}^d.$$

Unconstrained optimization of constrained problems

◦ **Entropic Mirror Descent**: Unconstrained optimization within the simplex.

$$\min_{\mathbf{x}\in\Delta_d} V(\mathbf{x})$$

▷ Choose h to be the entropic mirror map, $h^\star$ its dual

▷ Mirror vs primal image: $\mathbf{y} = \nabla h(\mathbf{x}) \Leftrightarrow \mathbf{x} = \nabla h^\star(\mathbf{y})$

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \beta_k \nabla V(\mathbf{x}^k) \;\Rightarrow\; \text{no projection since } \mathrm{dom}(h^\star) = \mathbb{R}^d.$$

◦ A "mirror descent theory" for Langevin Dynamics?

Mirrored Langevin Dynamics (MLD)

- Given e^{-V} and h, compute $e^{-W} := \nabla h \# e^{-V}$

$$\text{MLD} \equiv \begin{cases} \mathrm{d}\mathbf{Y}_t = -\nabla W \circ \nabla h(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \\ \mathbf{X}_t = \nabla h^\star(\mathbf{Y}_t) \end{cases} \quad \Rightarrow \quad X_\infty \sim e^{-V}.$$

Mirrored Langevin Dynamics (MLD)

∘ Given e^{-V} and h, compute $e^{-W} := \nabla h \# e^{-V}$

$$\text{MLD} \equiv \begin{cases} \mathrm{d}\mathbf{Y}_t = -\nabla W \circ \nabla h(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \\ \mathbf{X}_t = \nabla h^\star(\mathbf{Y}_t) \end{cases} \quad \Rightarrow \quad X_\infty \sim e^{-V}.$$

∘ Discretize: $\begin{cases} \mathbf{y}^{k+1} = \mathbf{y}^k - \beta_k \nabla W(\mathbf{y}^k) + \sqrt{2}\boldsymbol{\xi}^k \\ \mathbf{x}^{k+1} = \nabla h^\star(\mathbf{y}^{k+1}) \end{cases}$.

Mirrored Langevin Dynamics (MLD)

- Given e^{-V} and h, compute $e^{-W} := \nabla h \# e^{-V}$

$$\text{MLD} \equiv \begin{cases} \mathrm{d}\mathbf{Y}_t = -\nabla W \circ \nabla h(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \\ \mathbf{X}_t = \nabla h^\star(\mathbf{Y}_t) \end{cases} \quad \Rightarrow \quad X_\infty \sim e^{-V}.$$

- Discretize: $\begin{cases} \mathbf{y}^{k+1} = \mathbf{y}^k - \beta_k \nabla W(\mathbf{y}^k) + \sqrt{2}\boldsymbol{\xi}^k \\ \mathbf{x}^{k+1} = \nabla h^\star(\mathbf{y}^{k+1}) \end{cases}$.

- The dual distribution e^{-W} can be unconstrained even if e^{-V} is constrained.
  - ▷ Convergence rates for e^{-W} are easy.

Benefits of MLD

- Improved rates for constrained sampling.

- Can turn non-convex problems into convex ones!!
  - We provide the first $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ rate for Latent Dirichlet Allocation.

- Works well in practice.

For more details...

Welcome to our poster #43!!

[0] Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. (2017).
Sampling from a log-concave distribution with compact support with proximal langevin monte carlo.
arXiv preprint arXiv:1705.08964.

[0] Bubeck, S., Eldan, R., and Lehec, J. (2015).
Sampling from a log-concave distribution with projected langevin monte carlo.
arXiv preprint arXiv:1507.02564.

[0] Cheng, X. and Bartlett, P. (2017).
Convergence of langevin mcmc in kl-divergence.
arXiv preprint arXiv:1705.09048.

[0] Dalalyan, A. S. and Karagulyan, A. G. (2017).
User-friendly guarantees for the langevin monte carlo with inaccurate gradient.
arXiv preprint arXiv:1710.00095.

[0] Durmus, A., Majewski, S., and Miasojedow, B. (2018).
Analysis of langevin monte carlo via convex optimization.
arXiv preprint arXiv:1802.09188.