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
Introduction

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

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

○ A scalable framework: First-order sampling (assuming access to $\nabla V$).

  Step 1. **Langevin Dynamics**

  $$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \implies X_\infty \sim e^{-V}.$$
Introduction

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

○ A scalable framework: First-order sampling (assuming access to \( \nabla V \)).

  Step 1. **Langevin Dynamics**

  \[
  dX_t = -\nabla V(X_t) dt + \sqrt{2}dB_t \quad \Rightarrow \quad X_\infty \sim e^{-V}.
  \]

  Step 2. **Discretize**

  \[
  x^{k+1} = x^k - \beta_k \nabla V(x^k) + \sqrt{2\beta_k} \xi^k
  \]

  ▶ \( \beta_k \) step-size, \( \xi^k \) standard normal

  ▶ strong analogy to gradient descent method
Recent progress: Unconstrained distributions are easy

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

<table>
<thead>
<tr>
<th>Assumption</th>
<th>$W_2$</th>
<th>$d_{TV}$</th>
<th>KL</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LI \preceq \nabla^2 V \preceq mI$</td>
<td>$\tilde{O}(\epsilon^{-2}d)$</td>
<td>$\tilde{O}(\epsilon^{-2}d)$</td>
<td>$\tilde{O}(\epsilon^{-1}d)$</td>
<td>[Cheng and Bartlett, 2017] [Dalalyan and Karagulyan, 2017] [Durmus et al., 2018]</td>
</tr>
<tr>
<td>$LI \preceq \nabla^2 V \preceq 0$</td>
<td>$-$</td>
<td>$\tilde{O}(\epsilon^{-4}d)$</td>
<td>$\tilde{O}(\epsilon^{-2}d)$</td>
<td>[Durmus et al., 2018]</td>
</tr>
</tbody>
</table>

Note: $W_2(\mu_1, \mu_2) := \sqrt{\inf_{X \sim \mu_1, Y \sim \mu_2} \mathbb{E}\|X - Y\|^2}$, $d_{TV}(\mu_1, \mu_2) := \sup_{\mu_1(A), \mu_2(A)} \frac{|\mu_1(A) - \mu_2(A)|}{A}$ Borel
Recent progress: Unconstrained distributions are easy

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

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

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

- What about constrained distributions?
  - include many important applications, such as Latent Dirichlet Allocation (LDA).
A challenge: Constrained distributions are hard

- When \( \text{dom}(V) \) is compact, convergence rates deteriorate significantly.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>( \mathcal{W}_2 ) or KL</th>
<th>( d_{TV} )</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>( LI \gtrless \nabla^2 V \gtrless mI )</td>
<td>?</td>
<td>( \tilde{O}(\epsilon^{-6}d^5) )</td>
<td>[Brosse et al., 2017]</td>
</tr>
<tr>
<td>( LI \gtrless \nabla^2 V \gtrless 0 )</td>
<td>?</td>
<td>( \tilde{O}(\epsilon^{-6}d^5) )</td>
<td>[Brosse et al., 2017]</td>
</tr>
</tbody>
</table>

- 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.

  \[
  \min_{x \in \Delta_d} V(x)
  \]

  - Choose \( h \) to be the entropic mirror map, \( h^* \) its dual
Unconstrained optimization of constrained problems

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

\[
\min_{x \in \Delta_d} V(x)
\]

- Choose \( h \) to be the entropic mirror map, \( h^* \) its dual
- Mirror vs primal image: \( y = \nabla h(x) \iff x = \nabla h^*(y) \)

\[
y^{k+1} = y^k - \beta_k \nabla V(x^k) \Rightarrow \text{no projection since } \text{dom}(h^*) = \mathbb{R}^d.
\]
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, \( h^* \) its dual

▷ Mirror vs primal image: \( \mathbf{y} = \nabla h(\mathbf{x}) \iff \mathbf{x} = \nabla h^*(\mathbf{y}) \)

\[
\mathbf{y}^{k+1} = \mathbf{y}^k - \beta_k \nabla V(\mathbf{x}^k) \quad \Rightarrow \quad \text{no projection since } \text{dom}(h^*) = \mathbb{R}^d.
\]

- A “mirror descent theory” for Langevin Dynamics?
Mirrored Langevin Dynamics (MLD)

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

$$
MLD \equiv \begin{cases} 
\text{d}Y_t = -\nabla W \circ \nabla h(X_t) \text{d}t + \sqrt{2} \text{d}B_t \\
X_t = \nabla h^*(Y_t) 
\end{cases} \Rightarrow X_\infty \sim e^{-V}.
$$
Mirrored Langevin Dynamics (MLD)

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

$$\text{MLD} \equiv \begin{cases} \text{d} Y_t = -\nabla W \circ \nabla h(X_t) \text{d} t + \sqrt{2} \text{d} B_t \\ X_t = \nabla h^*(Y_t) \end{cases} \Rightarrow X_\infty \sim e^{-V}. \quad \text{(1)}$$

- Discretize: \[
\begin{align*}
y^{k+1} &= y^k - \beta_k \nabla W(y^k) + \sqrt{2} \xi^k \\
x^{k+1} &= \nabla h^*(y^{k+1})
\end{align*}
\]
Mirrored Langevin Dynamics (MLD)

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

$$\text{MLD} \equiv \begin{cases} \ \ \ \ \ \ \ \ dY_t = -\nabla W \circ \nabla h(X_t)dt + \sqrt{2}dB_t \\ X_t = \nabla h^*(Y_t) \end{cases} \Rightarrow X_\infty \sim e^{-V}.$$

- Discretize:

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

- 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!!
Sampling from a log-concave distribution with compact support with proximal langevin monte carlo.

Sampling from a log-concave distribution with projected langevin monte carlo.

Convergence of langevin mcmc in kl-divergence.

User-friendly guarantees for the langevin monte carlo with inaccurate gradient.

Analysis of langevin monte carlo via convex optimization.