Optimizing Probability Distributions for Learning: Sampling meets Optimization

Peter Bartlett

Computer Science and Statistics
UC Berkeley

March 25, 2019

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

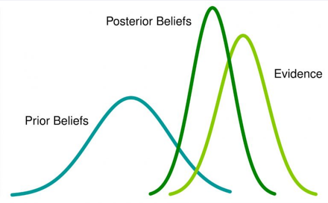
Sampling Problems

Bayesian inference

Compute $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$.

Write the density of $P(\theta|D)$ as

$$\frac{\exp(-U(\theta))}{\int \exp(-U(\theta))\, d\theta}.$$



www.analyticsvidhya.com

Langevin diffusion

Simulate a stochastic differential equation:

$$dx_t = -\nabla U(x_t)\, dt + \sqrt{2}\, dB_t.$$

Stationary distribution has density $p^*(\theta) \propto \exp(-U(\theta))$.

Sampling Problems

Prediction as a repeated game

- Player chooses action $a_t \in \mathcal{A}$,
- Adversary chooses outcome y_t,
- Player incurs loss $\ell(a_t, y_t)$.

Aim to minimize **regret**:
$\sum_t \ell(a_t, y_t) - \min_a \sum_t \ell(a, y_t)$.

Exponential weights strategy

$$p_t(a) \propto \exp\left(-U(a)\right),$$

$$\text{with} \quad U(a) := \eta \sum_{s=1}^{t-1} \ell(a, y_s).$$

Langevin diffusion

Simulate a stochastic differential equation:

$$dx_t = -\nabla U(x_t)\, dt + \sqrt{2}\, dB_t.$$

Stationary distribution has density $p^*(a) \propto \exp(-U(a))$.

Sampling Algorithms

Langevin diffusion

$$\text{SDE:} \qquad dx_t = -\nabla U(x_t)\,dt + \sqrt{2}\,dB_t.$$

Stationary distribution has density $p^*(\cdot) \propto \exp(-U(\cdot))$.

Discrete Time: Langevin MCMC Sampler (Euler-Maruyama)

$$x_{k+1} = x_k - \eta \nabla U(x_k) + \sqrt{2\eta}\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I).$$

- How close to the desired p^* is p_k (the density of x_k)?
- How rapidly does it converge?

Viewpoint

Sampling as optimization over the space of probability distributions.

Sampling Algorithms for Optimization

Parameter optimization in deep neural networks

- Use training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ to choose parameters θ of a deep neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$.
- Aim to minimize loss $U(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$.
- Gradient: $\theta_{k+1} = \theta_k - \eta_k \nabla U(\theta_k)$
- Stochastic gradient: Random θ_0, $\qquad \theta_{k+1} = \theta_k - \eta_k \nabla \hat{U}_{\xi_k}(\theta_k)$
- ... with minibatch gradient estimates, $\hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i \in \xi_k} \ell(y_i, f_\theta(x_i))$

- What is the distribution of θ_k?
- View stochastic gradient methods as sampling algorithms.
- How can we improve their performance?

Outline

- The Langevin diffusion
- Optimization theory for sampling methods
  - Convergence of Langevin MCMC in KL-divergence
  - Nesterov acceleration in sampling
  - The nonconvex case
- Sampling methods for optimization
  - Stochastic gradient methods as SDEs

Discretization of the Langevin Diffusion

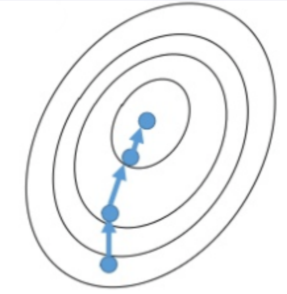
Langevin Markov Chain

Choose step-size η and simulate the Markov chain:

$$x_{k+1} = x_k - \eta \nabla U(x_k) + \sqrt{2\eta}\xi_k, \qquad \xi_k \overset{iid}{\sim} \mathcal{N}(0, I_d).$$

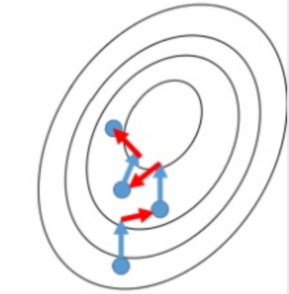
Gradient descent

$$x_{k+1} = x_k - \eta \nabla U(x_k)$$



Langevin Markov Chain

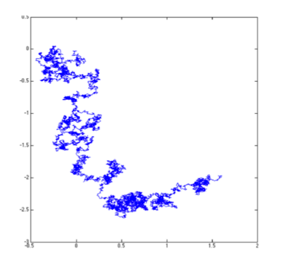
$$x_{k+1} = x_k - \eta \nabla U(x_k) + \sqrt{2\eta}\xi_k$$



Langevin Diffusion as Gradient Flow

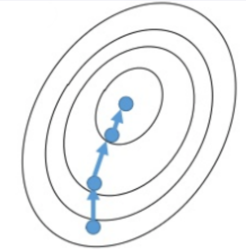
Langevin Diffusion in $\mathbb{R}^d$

$$dx_t = -\nabla U(x_t)\,dt + \sqrt{2}\,dB_t.$$



Gradient flow in $\mathcal{P}(\mathbb{R}^d)$

p_t minimizes $\frac{d}{dt}\mathcal{H}(\mathbf{p}_t) + \frac{1}{2}|\mathbf{p}_t'|^2$.



(Jordan, Kinderlehrer and Otto, 1998), (Ambrosio, Gigli and Savaré, 2005)

Richard Jordan

David Kinderlehrer

Felix Otto

Luigi Ambrosio

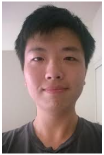
Nicola Gigli

Giuseppe Savaré

$W_2(\mathbf{p}_t, \mathbf{p}_{t+h})$

Sampling as Optimization

- Sampling algorithms can be viewed as deterministic optimization procedures over a space of probability distributions.
- Can we apply tools and techniques from optimization to sampling?



Xiang Cheng

An Optimization Analysis in $\mathcal{P}(\mathbb{R}^d)$

Convergence of Langevin MCMC in KL-divergence.
Xiang Cheng and PB.
arXiv:1705.09048[stat.ML]; ALT 2018.

Langevin MCMC

$$x_{k+1} = x_k - \eta \nabla U(x_k) + \sqrt{2\eta}\xi_k, \qquad \xi_k \overset{iid}{\sim} \mathcal{N}(0, I_d).$$

How does the density p_k of x_k evolve?

Theorem

For smooth, strongly convex U, that is, $\forall x, mI \preceq \nabla^2 U(x) \preceq LI$,

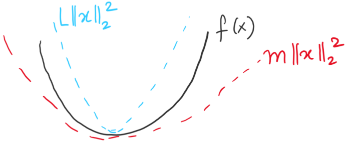
suitably small η and $k = \tilde{\Omega}\left(\frac{d}{\epsilon}\right)$ ensure that $\mathcal{KL}\left(\mathbf{p}^k \| \mathbf{p}^*\right) \leq \epsilon$.

Implies older bounds for TV and W_2:

For suitably small η and $k = \tilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$,

$\|p_k - p^*\|_{TV} \leq \epsilon$. (Dalalyan, 2014)

$W_2(p_k, p^*) \leq \epsilon$. (Durmus and Moullines, 2016)



Sampling as Optimization

- Sampling algorithms can be viewed as deterministic optimization procedures over the probability space.
- Can we apply tools and techniques from optimization to sampling?

Xiang Cheng

Niladri Chatterji

Mike Jordan

Nesterov acceleration in $\mathcal{P}(\mathbb{R}^d)$

Underdamped Langevin MCMC: A non-asymptotic analysis.
Xiang Cheng, Niladri Chatterji, PB and Mike Jordan.
arXiv:1707.03663 [stat.ML]; COLT18.

Nesterov acceleration in sampling

Kramers' Equation (1940)

Stochastic differential equation:

$$\begin{aligned} dx_t &= v_t\, dt, \\ dv_t &= \underbrace{-v_t\, dt}_{\text{friction}} - \underbrace{\nabla U(x_t)\, dt}_{\text{acceleration}} + \sqrt{2}\, dB_t, \end{aligned}$$

where $x_t, v_t \in \mathbb{R}^d$, $U : \mathbb{R}^d \to \mathbb{R}$, dB_t is standard Brownian motion on $\mathbb{R}^d$.

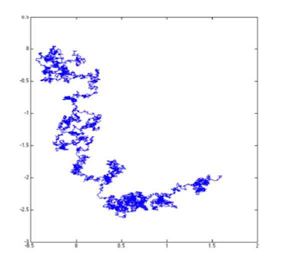
Define p_t as the density of (x_t, v_t).
Under mild regularity assumptions, $p_t \to p^*$:

$$p^*(x) \propto \exp\left(-U(x) - \frac{1}{2}\|v\|_2^2\right).$$

Nesterov acceleration in sampling

(Overdamped) Langevin Diffusion

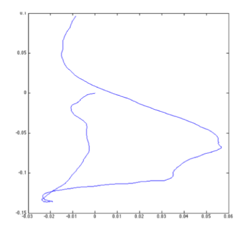
$$dx_t = -\nabla U(x_t)\, dt + \sqrt{2}\, dB_t.$$



Underdamped Langevin Diffusion

$$dx_t = v_t\, dt,$$

$$dv_t = -v_t\, dt - \nabla U(x_t)\, dt + \sqrt{2}\, dB_t.$$



Nesterov acceleration in sampling

Theorem

For smooth, strongly convex U, suitably small η and $k = \tilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\right)$, underdamped Langevin MCMC gives $W_2(p_k, p^*) \leq \epsilon$.

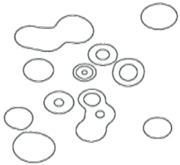
Significantly faster than overdamped Langevin:

For suitably small η and $k = \tilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$, $W_2(p_k, p^*) \leq \epsilon$.

(Durmus and Moullines, 2016)

Nonconvex potentials

- Multi-modal p (nonconvex U)?



Xiang Cheng

Niladri Chatterji

Yasin Abbasi-Yadkori

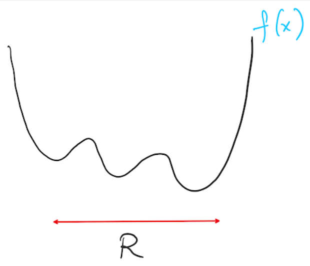
Mike Jordan

Sharp convergence rates for Langevin dynamics in the nonconvex setting.
Xiang Cheng, Niladri Chatterji, Yasin Abbasi-Yadkori, PB and Mike Jordan.
arXiv:1805.01648 [stat.ML].

Nonconvex potentials

Assumptions

- Smooth everywhere: $\nabla^2 U \preceq LI$.
- Strongly convex outside a ball:
  $\forall x, y,\ \|x - y\|_2 \geq R \Rightarrow U(x) \geq U(y) + \langle U(y), x - y \rangle + \frac{m}{2}\|x - y\|_2^2$.



Nonconvex potentials

Theorem

Suppose U is L-smooth and strongly convex outside a ball of radius R and η is suitably small.

1. If $k = \tilde{\Omega}\left(\frac{d}{\epsilon^2}\exp(LR^2)\right)$, then overdamped Langevin MCMC has $W_1(p_k, p^*) \leq \epsilon$.
2. If $k = \tilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\exp(LR^2)\right)$, then underdamped Langevin MCMC has $W_1(p_k, p^*) \leq \epsilon$.

- We can think of LR^2 is a measure of non-convexity of U.
- The improvement from overdamped to underdamped is the same as in the convex case.

Outline

- The Langevin diffusion
- Optimization theory for sampling methods
  - Convergence of Langevin MCMC in KL-divergence
  - Nesterov acceleration in sampling
  - The nonconvex case
- Sampling methods for optimization
  - Stochastic gradient methods as SDEs

Sampling Algorithms for Optimization

Xiang Cheng

Mike Jordan

Quantitative central limit theorems for discrete stochastic processes.
Xiang Cheng, PB and Mike Jordan.
arXiv:1902.00832 [math.ST].

Sampling Algorithms for Optimization

Parameter optimization in deep neural networks

- Use training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ to choose parameters θ of a deep neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$.
- Aim to minimize loss $U(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$.
- Stochastic gradient: Random θ_0, $\qquad \theta_{k+1} = \theta_k - \eta \nabla \hat{U}_{\xi_k}(\theta_k)$,
- ... with minibatch gradient estimates, $\hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i \in \xi_k} \ell(y_i, f_\theta(x_i))$

Sampling Algorithms for Optimization

Definitions

- $x_{k+1} = x_k - \eta \nabla U(x_k) + \sqrt{\eta} T_{\xi_k}(x_k)$, with $\xi_k \overset{iid}{\sim} q$.
- Define the covariance of the noise: $\sigma_x^2 := \mathbb{E}_\xi \left[T_\xi(x) T_\xi(x)^\top \right]$.
- Consider the SDE: $dx_t = -\nabla U(x_t)\, dt + \sqrt{2} \sigma_{x_t}\, dB_t$.
- Let p^* denote its stationary distribution.

Theorem

For U smooth, strongly convex, bounded third derivative, σ_x^2 uniformly bounded, $T_\xi(\cdot)$ smooth, bounded third derivatives, $\log p^*$ with bounded third derivatives,

If η is sufficiently small, $W_2(\hat{p}, p^*) \le \epsilon$, $(x_\infty \sim \hat{p})$

and for $k = \tilde{\Omega}\left(\frac{d^7}{\epsilon^2}\right)$, $W_2(p_k, p^*) \le \epsilon$. $(x_k \sim p_k)$

The classical CLT (with U quadratic) shows that the $1/\sqrt{k}$ rate is optimal.

Outline

- The Langevin diffusion
- Optimization theory for sampling methods
  - Convergence of Langevin MCMC in KL-divergence
  - Nesterov acceleration in sampling
  - The nonconvex case
- Sampling methods for optimization
  - Stochastic gradient methods as SDEs