Optimizing Probability Distributions for Learning: Sampling meets Optimization

Peter Bartlett

Computer Science and Statistics UC Berkeley

Joint work with Yasin Abbasi-Yadkori, Niladri Chatterji, Xiang Cheng, Mike Jordan

February 22, 2019

Sampling Problems



Compute $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$. Write the density of $P(\theta|D)$ as $\frac{\exp(-U(\theta))}{\int \exp(-U(\theta)) d\theta}$.



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	Exponential weights strategy
• Player chooses action $a_t \in \mathcal{A}$,	
• Adversary chooses outcome y_t ,	$p_t(a) \propto \exp\left(-U(a) ight),$
• Player incurs loss $\ell(a_t, y_t)$.	t-1
Aim to minimize regret : $\sum_{t} \ell(a_t, y_t) - \min_a \sum_{t} \ell(a, y_t).$	with $U(a) := \eta \sum_{s=1}^{\infty} \ell(a, y_s).$

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

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

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

Discrete Time: Langevin MCMC Sampler (Euler-Maruyama)

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

- 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 (x₁, y₁), ..., (x_n, y_n) ∈ X × Y to choose parameters θ of a deep neural network f_θ : X → 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 , $\theta_{k+1} = \theta_k \eta_k \nabla \hat{U}_{\xi_k}(\theta_k)$
- ... with minibatch gradient estimates, $\hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i=1}^{k} \ell(y_i, f_{\theta}(x_i))$
- ... and decreasing stepsizes η_k .
 - What is the distribution of θ_k ?
 - View stochastic gradient methods as sampling algorithms.

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

• 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

Langevin diffusion

Stochastic differential equation:

$$dx_t = \underbrace{-\nabla U(x_t) dt}_{\text{drift}} + \sqrt{2} \, dB_t,$$

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

Paul Langevin wikipedia.org

Define p_t as the density of x_t .

Under mild regularity assumptions, $p_t \rightarrow p^*$; $p^*(x) \propto \exp(-U(x))$.

Discretization of the Langevin Diffusion

Langevin Markov Chain

Choose step-size η and simulate the Markov chain:

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

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 Markov Chain

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

How does the density p_k of x_k evolve?

Asymptotic results

Under regularity conditions, for shrinking step-size η_k , $\|p_k - p^*\|_{TV} \to 0$.

e.g., (Gelfand and Mitter, 1991), (Roberts and Tweedie, 1996)



Arnak Dalalyan mediamax.am

Quantitative results

For suitably small (fixed) η and $k = \tilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$, $\|p_k - p^*\|_{TV} \le \epsilon$. (Dalalyan, 2014) $W_2(p_k, p^*) \le \epsilon$. (Durmus and Moullines, 2016)

Langevin Diffusion as Gradient Flow



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



Richard Jordan



David Kinderlehrer

Felix Otto



Luigi Ambrosio



Nicola Gigli



Giuseppe Savaré



.

0

- 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

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

How does the density p_k of x_k evolve?

Theorem

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

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 $k = \tilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$,

 $\|p_k - p^*\|_{TV} \le \epsilon.$

 $W_2(p_k, p^*) < \epsilon$.

(Durmus and Moullines, 2016)

(Dalalyan, 2014)



Minimization over \mathbb{R}^d

Minimize $f : \mathbb{R}^d \to \mathbb{R}$ using gradient flow $y_t : \mathbb{R}^+ \to \mathbb{R}^d$ wrt Euclidean norm:

min
$$\left(\frac{d}{dt}f(y_t) + \frac{1}{2}\left\|\frac{d}{dt}y_t\right\|^2\right)$$

$$\frac{d}{dt}f(y_t) = \left\langle \nabla f(y_t), \frac{d}{dt}y_t \right\rangle$$
$$\frac{d}{dt}f(y_t^*) = -\|\nabla f(y_t^*)\|_2^2$$

Minimization over $\mathcal{P}(\mathbb{R}^d)$

Minimize $\mathcal{H}(\mathbf{p}) = \mathcal{KL}(\mathbf{p} \| \mathbf{p}^*)$ using gradient flow $\mathbf{p}_t : \mathbb{R}^+ \to \mathcal{P}(\mathbb{R}^d)$ wrt W_2 :

min
$$\left(\frac{d}{dt}\mathcal{H}(\mathbf{p}_t) + \frac{1}{2}\left|\mathbf{p}_t'\right|^2\right)$$

$$\frac{d}{dt}\mathcal{H}(\mathbf{p}_t) = \mathbb{E}_{\mathbf{x}\sim\mathbf{p}_t} \left[\left\langle \nabla_{\mathbf{x}} \frac{\partial \mathcal{H}}{\partial \mathbf{p}}(\mathbf{p}_t)(\mathbf{x}), \mathbf{v}_t(\mathbf{x}) \right\rangle \right]$$
$$\frac{d}{dt}\mathcal{H}(\mathbf{p}_t^*) = -\mathbb{E}_{\mathbf{x}\sim\mathbf{p}_t^*} \left[\left\| \nabla \frac{\partial \mathcal{H}}{\partial \mathbf{p}}(\mathbf{p}_t^*)(\mathbf{x}) \right\|_2^2 \right]$$

Notation

- $\mathcal{P}(\mathbb{R}^d)$: set of densities over \mathbb{R}^d .
- $\mathcal{H}(\mathbf{p}) = \mathcal{KL}(\mathbf{p} || \mathbf{p}^*) = \int \log \frac{p(x)}{p^*(x)} p(x) dx.$
- $W_2^2(p,q) = \inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{(x,y) \sim \gamma} ||x y||_2^2$, with $\Gamma(p,q)$: all joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals p and q.
- For a curve $\mathbf{p}_t: \mathbb{R}^+ o \mathcal{P}(\mathbb{R}^d)$, the metric derivative is

$$|\mathbf{p}_t'| = \lim_{h \to 0} \frac{W_2(\mathbf{p}_t, \mathbf{p}_{t+h})}{h}$$

- If v_t is tangent to \mathbf{p}_t , then $|\mathbf{p}'_t|^2 = \mathbb{E}_{x \sim \mathbf{p}_t} \left[\|v_t(x)\|_2^2 \right]$.
- Fréchet derivative: $\frac{\partial \mathcal{H}}{\partial \mathbf{p}_t}(\mathbf{p}_t) = 1 + \log\left(\frac{\mathbf{p}_t}{\mathbf{p}^*}\right)$.
- $\frac{d}{dt}\mathcal{H}(\mathbf{p}_t) = \mathbb{E}_{x \sim \mathbf{p}_t}\left[\left\langle \nabla_x \frac{\partial \mathcal{H}}{\partial \mathbf{p}_t}(\mathbf{p}_t)(x), v_t(x)\right\rangle\right]$

Minimization over \mathbb{R}^d

m-strong convexity of f implies $f(y) - f(y^*) \le \frac{1}{m} \|\nabla f(y)\|_2^2$.

Hence
$$\frac{d}{dt}(f(y_t) - f(y^*)) \leq -m(f(y_t) - f(y^*)).$$

Minimization over $\mathcal{P}(\mathbb{R}^d)$

m-strong convexity of *U* implies *m*-geodesic-convexity of $\mathcal{H}(\mathbf{p})$ in W_2 , which implies $\mathcal{H}(\mathbf{p}) - \mathcal{H}(\mathbf{p}^*) \leq \frac{1}{m} \mathbb{E}_{x \sim \mathbf{p}} \left[\left\| \nabla \frac{\partial \mathcal{H}}{\partial \mathbf{p}}(\mathbf{p})(x) \right\|_2^2 \right].$ Hence $\frac{d}{dt} \left(\mathcal{H}(\mathbf{p}_t) - \mathcal{H}(\mathbf{p}^*) \right) \leq -m(\mathcal{H}(\mathbf{p}_t) - \mathcal{H}(\mathbf{p}^*)).$

- 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 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:

$$dx_t = v_t dt,$$

$$dv_t = \underbrace{-v_t dt}_{\text{friction}} - \underbrace{\nabla U(x_t) dt}_{\text{acceleration}} + \sqrt{2} dB_t$$



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 .

Hendrick A. Kramers wikipedia.org

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

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

(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

Underdamped Langevin Markov Chain Choose step-size η and simulate the SDE: $d\tilde{x}_t = \tilde{v}_t dt$ $d\tilde{v}_t = -\tilde{v}_t dt - \nabla U(\tilde{x}_{k\eta}) dt + \sqrt{2} dB_t$ for $k\eta < t < (k+1)\eta$.

(Not the standard Euler-Maruyama discretization.)

• A version of Hamiltonian Monte Carlo

(Duane, Kennedy, Pendleton and Roweth, 1987), (Neal, 2011)

• How does the density p_k of $(\tilde{x}_{k\eta}, \tilde{v}_{k\eta})$ evolve?

Theorem

For smooth, strongly convex U, suitably small η and $k = \tilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\right)$,

underdamped Langevin MCMC gives $W_2(p_k, p^*) \leq \epsilon$.

Idea of proof:

uses tools from (Eberle, Guillin and Zimmer, 2017)

Synchronous coupling (shared Brownian motion); strong convexity.

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)

Related work

HMC With separability assumption

(Lee and Vempala, 2017)

(Mangoubi and Smith, 2017), (Mangoubi and Vishnoi, 2018).

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

• 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:

 $\forall x, y, \|x-y\|_2 \geq R \Rightarrow U(x) \geq U(y) + \langle U(y), x-y \rangle + \frac{m}{2} \|x-y\|_2^2.$



Theorem

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

- 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.

Idea of proof

- Synchronous coupling when far away; exploits strong convexity.
- Eberle's (2016) reflection coupling (1-D Brownian motion along the line between) when close: this 1-D random walk couples.
- Since it is in 1-D, the rate is not exponential in dimension.

Related work

- Weaker assumptions; exponential in dimension. (Raginsky, Rakhlin and Telgarsky, 2017)
- Stronger assumptions: mixtures of Gaussians. (Ge, Lee and Risteski, 2017)
- Metropolis-Hastings version.

(Bou-Rabee, Eberle and Zimmer, 2018)

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



Xiang Cheng



Mike Jordan

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

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 , $\theta_{k+1} = \theta_k \eta \nabla \hat{U}_{\varepsilon_k}(\theta_k)$,
- ... with minibatch gradient estimates, $\hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i \in I} \ell(y_i, f_{\theta}(x_i))$

This has the form:

$$egin{aligned} & \mathbf{x}_{k+1} = \mathbf{x}_k - \eta
abla \hat{U}_{\xi_k}(\mathbf{x}_k) \ & = \mathbf{x}_k - \eta
abla U(\mathbf{x}_k) + \sqrt{\eta} T_{\xi_k}(\mathbf{x}_k). \end{aligned}$$

- ... which is suggestive of a Langevin diffusion but ...
- The noise $T_{\xi_k}(x) = \sqrt{\eta} \left(\nabla U(x) \nabla \hat{U}_{\xi_k}(x) \right)$ is not Gaussian, and depends on x.
- What is the stationary distribution of x_k?
- How rapidly is it approached?

Sampling Algorithms for Optimization

Definitions

- $x_{k+1} = x_k \eta \nabla U(x_k) + \sqrt{\eta} T_{\xi_k}(x_k)$,
- Define the covariance of the noise:
- 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,

 $\begin{array}{l} T_{\xi}(\cdot) \text{ smooth, bounded third derivatives, } \log p^{*} \text{ with bounded third derivatives,} \\ \text{If } \eta \text{ is sufficiently small, } W_{2}(\hat{p},p^{*}) \leq \epsilon, \qquad \qquad (x_{\infty} \sim \hat{p}) \\ \text{and for } k = \tilde{\Omega}\left(\frac{d^{7}}{\epsilon^{2}}\right), \ W_{2}(p_{k},p^{*}) \leq \epsilon. \qquad \qquad (x_{k} \sim p_{k}) \\ \end{array}$

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

with $\xi_{k} \overset{iid}{\sim} a$.

 $\sigma_{\mathbf{x}}^2 := \mathbb{E}_{\mathcal{E}} \left[T_{\mathcal{E}}(\mathbf{x}) T_{\mathcal{E}}(\mathbf{x})^\top \right].$

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

- Large scale problems: stochastic gradient estimates
- Variance reduction with stochastic gradient estimates
- Convergence in KL for underdamped Langevin, nonconvex
- With constraints
- Lower bounds
- Sampling methods for optimization
 - Stochastic gradient with momentum?
 - Nonconvex loss U?
 - Role of noise covariance in behavior of stochastic gradient method?

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