Bayesian inference

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

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

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

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

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

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

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

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)

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

- 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.
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 \(\theta\) of a deep neural network \(f_\theta : \mathcal{X} \rightarrow \mathcal{Y}\).

- 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 \(\theta_0\),
  \[
  \theta_{k+1} = \theta_k - \eta_k \nabla \hat{U}_{\xi_k}(\theta_k)
  \]

- ... with minibatch gradient estimates,
  \[
  \hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i \in \xi_k} \ell(y_i, f_\theta(x_i))
  \]

- ... and decreasing stepsizes \(\eta_k\).

- What is the distribution of \(\theta_k\)?

- View stochastic gradient methods as sampling algorithms.
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
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
Langevin diffusion

Stochastic differential equation:

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

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

Define \( p_t \) as the density of \( x_t \).

Under mild regularity assumptions, \( p_t \to p^* \); \( p^*(x) \propto \exp(-U(x)) \).
Discretization of the Langevin Diffusion

Langevin Markov Chain

Choose step-size $\eta$ and simulate the Markov chain:

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

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 Markov Chain

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

How does the density \( p_k \) of \( x_k \) evolve?

**Asymptotic results**

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

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

**Quantitative results**

For suitably small (fixed) \( \eta \) 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)
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 \( \frac{d}{dt} \mathcal{H}(p_t) + \frac{1}{2} |p'_t|^2. \)

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?

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

Convergence of Langevin MCMC in KL-divergence.
Xiang Cheng and PB.
Langevin MCMC

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

How does the density \( p_k \) of \( x_k \) evolve?

**Theorem**

For smooth, strongly convex \( U \), that is, \( \forall x, ml \leq \nabla^2 U(x) \leq LI \), suitably small \( \eta \) and \( k = \tilde{\Omega} \left( \frac{d}{\epsilon} \right) \) ensure that \( \mathcal{KL}(p_k \| p^*) \leq \epsilon. \)

**Implies older bounds for TV and \( W_2 \):**

For suitably small \( \eta \) 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)
Analog with gradient flow over $\mathbb{R}^d$

**Minimization over $\mathbb{R}^d$**

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

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

$$\frac{d}{dt} f(y_t) = \left\langle \nabla f(y_t), \frac{d}{dt} y_t \right\rangle$$

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

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

Minimize $\mathcal{H}(p) = \mathcal{K} \mathcal{L}(p \parallel p^*)$ using gradient flow $p_t : \mathbb{R}^+ \rightarrow \mathcal{P}(\mathbb{R}^d)$ wrt $W_2$:

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

$$\frac{d}{dt} \mathcal{H}(p_t) = \mathbb{E}_{x \sim p_t} \left[ \left\langle \nabla_x \frac{\partial \mathcal{H}}{\partial p}(p_t)(x), v_t(x) \right\rangle \right]$$

$$\frac{d}{dt} \mathcal{H}(p_t^*) = -\mathbb{E}_{x \sim p_t^*} \left[ \left\| \nabla_x \frac{\partial \mathcal{H}}{\partial p}(p_t^*)(x) \right\|^2_2 \right]$$
\( \mathcal{P}(\mathbb{R}^d) \): set of densities over \( \mathbb{R}^d \).

\[ H(p) = KL(p \parallel p^*) = \int \log \frac{p(x)}{p^*(x)} p(x) \, dx. \]

\[ W_2^2(p, q) = \inf_{\gamma \in \Gamma(p, q)} \mathbb{E}(x, y) \sim_{\gamma} \| x - y \|_2^2, \] with \( \Gamma(p, q) \): all joint distributions on \( \mathbb{R}^d \times \mathbb{R}^d \) with marginals \( p \) and \( q \).

For a curve \( p_t : \mathbb{R}^+ \rightarrow \mathcal{P}(\mathbb{R}^d) \), the metric derivative is

\[ |p'_t| = \lim_{h \to 0} \frac{W_2(p_t, p_{t+h})}{h}. \]

If \( v_t \) is tangent to \( p_t \), then \( |p'_t|^2 = \mathbb{E}_{x \sim p_t} \| v_t(x) \|_2^2 \).

Fréchet derivative:

\[ \frac{\partial H}{\partial p_t}(p_t) = 1 + \log \left( \frac{p_t}{p^*} \right). \]

\[ \frac{d}{dt} H(p_t) = \mathbb{E}_{x \sim p_t} \left[ \langle \nabla_x \frac{\partial H}{\partial p_t}(p_t)(x), v_t(x) \rangle \right] \]
Analog with gradient flow over $\mathbb{R}^d$

**Minimization over $\mathbb{R}^d$**

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

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 $\mathcal{H}(p)$ in $W_2$, which implies $\mathcal{H}(p) - \mathcal{H}(p^*) \leq \frac{1}{m} \mathbb{E}_{x \sim p} \left[ \|\nabla \frac{\partial \mathcal{H}}{\partial p}(p)(x)\|_2^2 \right]$.

Hence $\frac{d}{dt}(\mathcal{H}(p_t) - \mathcal{H}(p^*)) \leq -m(\mathcal{H}(p_t) - \mathcal{H}(p^*))$. 
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 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?

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

Underdamped Langevin MCMC: A non-asymptotic analysis.
Xiang Cheng, Niladri Chatterji, PB and Mike Jordan.
Kramers’ Equation (1940)

Stochastic differential equation:

\[ dx_t = v_t \, dt, \]
\[ dv_t = -v_t \, dt - \nabla U(x_t) \, dt + \sqrt{2} \, dB_t, \]

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

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

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

Underdamped Langevin Markov Chain

Choose step-size \( \eta \) and simulate the SDE:

\[
d\tilde{x}_t = \tilde{v}_t \, dt \\
\frac{d\tilde{v}_t}{dt} = -\tilde{v}_t \, dt - \nabla U(\tilde{x}_{k\eta}) \, dt + \sqrt{2} \, dB_t
\]

for \( k\eta \leq t < (k+1)\eta \).

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

(Durmus and Moullines, 2016)

Related work

HMC

(Lee and Vempala, 2017)

With separability assumption

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

- Multi-modal $p$ (nonconvex $U$)?

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 $\eta$ 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.
Nonconvex potentials

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)*
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
Quantitative central limit theorems for discrete stochastic processes.
Xiang Cheng, PB and Mike Jordan.
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 \(\theta\) of a deep neural network \(f_\theta : \mathcal{X} \rightarrow \mathcal{Y}\).

- Aim to minimize loss \(U(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))\).

- Stochastic gradient: Random \(\theta_0\), \(\theta_{k+1} = \theta_k - \eta \nabla \hat{U}_{\xi_k}(\theta_k)\),

- ... with minibatch gradient estimates, \(\hat{U}_{\xi_k}(\theta) = \frac{1}{\xi_k} \sum_{i \in \xi_k} \ell(y_i, f_\theta(x_i))\)
This has the form:

\[ x_{k+1} = x_k - \eta \nabla \hat{U}_{\xi_k}(x_k) \]

\[ = x_k - \eta \nabla U(x_k) + \sqrt{\eta} T_{\xi_k}(x_k), \]

... 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

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

Theorem

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

- If \( \eta \) is sufficiently small, \( W_2(\hat{p}, p^*) \leq \epsilon \), \( (x_\infty \sim \hat{p}) \)
- and for \( k = \tilde{\Omega} \left( \frac{d^7}{\epsilon^2} \right) \), \( W_2(p_k, p^*) \leq \epsilon \), \( (x_k \sim p_k) \)

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

- Optimization theory for sampling methods
  - 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?
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