Langevin Dynamics

Loucas Pillaud-Vivien

November 7, 2019

Introduction

- Sampling **distribution** over high-dimensional space is an important topic in computational statistics and machine learning
- Example of application : Bayesian inference for high-dimensional models
- **Problems**:
  1. Most of sampling techniques do not scale to high-dimension. **Big d**.
  2. And to large number of data (recall HMC, need the full gradient). **Big N**.

Example: Bayesian setting

- A Bayesian model is specified by:
  1. sampling distribution of observed data: **likelihood** $Y \sim L(\cdot|\theta)$
  2. a **prior** distribution p on the parameter space $\theta \in \mathbb{R}^d$
- The inference is based on the posterior distribution

$$\pi(d\theta) = \frac{p(d\theta)L(Y|\theta)}{\int L(Y|u)p(du)}$$

- The normalizing constant is often not tractable (too high dimensional), we can only compute:

$$\pi(d\theta) \propto p(d\theta)L(Y|\theta)$$

Outline

Framework

- We want to sample the following measure that has a density w.r.t Lebesgue *known up to a normalization* factor.

$$d\mu(x) = \frac{e^{-V(x)}dx}{\int_{\mathbb{R}^d} e^{-V(y)}dy}$$

- We assume that V is L-smooth : i.e. continuously differentiable and $\exists L > 0$ s.t.

$$\|\nabla V(x) - \nabla V(y)\| \leqslant L\|x - y\|$$

Convergence to equilibrium for Diffusions

Let us consider the overdamped Langevin diffusion in $\mathbb{R}^d$:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t,$$

- L-smoothness of V gives **existence and unicity** of a solution
- **Stationnary measure**: $d\mu(x) = \frac{e^{-V(x)}dx}{\int_{\mathbb{R}^d} e^{-V(y)}dy}$.
- **Semi-group**: $P_t(f)(x) = \mathbb{E}[f(X_t)|X_0 = x] \longrightarrow$ "law of X_t".
- **Infinitesimal generator**: $\mathcal{L}\phi = \Delta\phi - \nabla V \cdot \nabla\phi$.

We can verify that the semi-group follows the dynamics:

$$\frac{d}{dt}P_t(f) = \mathcal{L}P_t(f).$$

$\longrightarrow$ **Question** : what speed of convergence then??? ?

Convergence to equilibrium for Diffusions

Theorem (Poincaré implies convergence to equilibrium)

With the notations above, the following propositions are equivalent:

- *μ satisfies a Poincaré Inequality with constant $\mathcal{P}$*
- *For all f smooth, $\mathrm{Var}_\mu(P_t(f)) \leqslant e^{-2t/\mathcal{P}}\mathrm{Var}_\mu(f)$ for all $t \geqslant 0$.*

Proof: Integration by part formula (μ is reversible),

$$-\int f(\mathcal{L}g)\,d\mu = \int \nabla f \cdot \nabla g\,d\mu = -\int (\mathcal{L}f)g\,d\mu, \quad \text{hence,}$$

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}_\mu(P_t(f)) = \frac{d}{dt}\int (P_t(f))^2 d\mu &= 2\int P_t(f)(\mathcal{L}P_t(f))d\mu \\ &= -2\int \|\nabla P_t(f)\|^2 d\mu \\ &\leqslant -2/\mathcal{P}\ \mathrm{Var}_\mu(P_t(f))\end{aligned}$$

Poincaré inequalities: definition in modern language

Definition (Poincaré inequality)

$\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies a Poincaré Inequality with constant $\mathcal{P}$ if

$$\mathrm{Var}_\mu(f) \leqslant \mathcal{P}_\mu \int \|\nabla f\|^2 d\mu,$$

for all (bounded) $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ of class $\mathcal{C}^1$.

Recall that :

- $\mathrm{Var}_\mu(f) = \int f^2 d\mu - \left(\int f d\mu\right)^2 = \int \left(f - \int f d\mu\right)^2 d\mu$
- $\int \|\nabla f\|^2 d\mu = \mathcal{E}(f)$ is the *Dirichlet Energy*.

Spectral interpretation: $\mathcal{E}(f) = \int \nabla f \cdot \nabla f d\mu = \int f(-\mathcal{L}f) d\mu$

$\longrightarrow 1/\mathcal{P} = \lambda_2$, first non-trivial eigenvalue of $\mathcal{L}$.

Application to the Ornstein-Uhlenbeck process

The diffusion of the **Ornstein-Uhlenbeck process** follows the SDE in $\mathbb{R}^d$:

$$dX_t = -X_t dt + \sqrt{2} dB_t,$$

Denote $\mathcal{L}$ the operator $\mathcal{L}\phi = \Delta\phi - x \cdot \nabla\phi$, then

1. For $d\mu(x) = \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2} dx$, $\mathcal{L}$ is self adjoint in L^2_μ
2. μ stationnary measure of O-U process
3. μ verifies Poincaré inequality with constant 1.
4. for all f smooth, for all $t \geqslant 0$.

$$\mathrm{Var}_\mu(P_t(f)) \leqslant e^{-2t} \mathrm{Var}_\mu(f).$$

Poincaré inequalities

Long story short:

$$\text{Poincaré inequality} \iff \text{Spectral gap for } \mathcal{L}$$
$$\iff$$
$$\text{Exponential convergence for the diffusion}$$

Poincaré inequalities

For what distribution do they occur?

- When V is m-stongly convex: $\mathcal{P} = 1/m$ (linear convergence of gradient descent)

Poincaré inequalities

For what distribution do they occur?

- When V is m-stongly convex: $\mathcal{P} = 1/m$ (linear convergence of gradient descent)
- When V is **only** convex: yes but no bound...

Poincaré inequalities

For what distribution do they occur?

- When V is m-stongly convex: $\mathcal{P} = 1/m$ (linear convergence of gradient descent)
- When V is **only** convex: yes but no bound...
- A *generic condition* for non necessarily convex potential :

$$\frac{1}{2}|\nabla V|^2 - \Delta V \geqslant \alpha$$

Poincaré inequalities

For what distribution do they occur?

- When V is m-stongly convex: $\mathcal{P} = 1/m$ (linear convergence of gradient descent)
- When V is **only** convex: yes but no bound...
- A *generic condition* for non necessarily convex potential :

$$\frac{1}{2}|\nabla V|^2 - \Delta V \geqslant \alpha$$

- For mixture of Gaussian $\mathcal{P}$ explodes exponentially.

Ok, fine. But how do I get back to the real world and draw samples ?

Discretized Langevin Diffusion

- Idea: Sample the diffusion paths, using Euler-Maruyama scheme

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$$
$$X_{k+1} = X_k - \gamma_{k+1}\nabla V(X_k) + \sqrt{2\gamma_{k+1}}\xi_{k+1}$$

  where
  - $(\xi_k)_k$ is i.i.d $\mathcal{N}(0, I_d)$
  - $(\gamma_k)_k$ is a sequence of stepsizes, either constant or decreasing to 0
- Note the similarity with gradient descent or its stochastic counterpart.
- This algorithm is referred to *Unajusted Langevin Algorithm*, *Langevin Monte Carlo* or *Gradient Langevin Dynamics*.

Discretized Langevin Diffusion: constant stepsize

- When $\forall k,\ \gamma_k = \gamma$, then $(X_k)_k$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some mild assumptions R_γ is **irreducible**, **positive recurrent** and hence has an **invariant distribution** $d\mu_\gamma \neq d\mu$.
- Typical questions:
  - For a given precision how do we choose the stepsize γ and the number of iterations such that

$$dist(\delta_x R_\gamma^n, d\mu) \leqslant \epsilon$$

  - How do we choose x ?
  - How do we quantify $dist(d\mu_\gamma, d\mu)$?

Outline

Result for a strongly convex potential

Theorem (Durmus, Moulines 2016)

Assume that V is m-strongly convex and L smooth. Set $\gamma \in (0, 1/(m+L)])$ and $\kappa = mL/(m+L)$ then for all $x \in \mathbb{R}^d$,

$$W_2^2(\delta_x R_\gamma^n, \pi) \leqslant 2(1-\kappa\gamma)^n W_2^2(\delta_x, \pi) + Cd\gamma$$

Remarks :

- Decomposition **bias + variance** as for SGD.
- Geometric rate then distance from $d\mu$ to $d\mu_\gamma$
- One may choose γ s.t. for $n = \Theta\left(\frac{d}{\epsilon^2}\right)$ iterations $W_2^2(\delta_x R_\gamma^n, \pi) \leqslant \epsilon$
- Explicit way of choosing γ (it was a problem! –see MALA)

Result for a strongly convex potential : remarks

Remarks :

- Exactly same results for
  - Total variation (Dalalyan 2014)
  - KL divergence (Bartlett et al. 2017)
- Same result with decreasing step sizes but no parameter to tune!
- Quadratic improvement by Jordan et. al 2018 by considering underdamped Langevin (similar to HMC) for $n = \Theta\left(\frac{\sqrt{d}}{\epsilon}\right)$ iterations $W_2^2(\delta_x R_\gamma^n, \pi) \leqslant \epsilon$ (needed also only strong convexity **outside of a ball**).

Grrrrr...But you know... I do not like to compute all the gradients...

Stochastic Gradient Langevin Dynamics (SGLD)

- Recall: the **ULA** algorithm is a discretization of the overdamped Langevin diffusion, which leaves invariant the target distribution $d\mu$.
- To further reduce the computational cost, SGLD uses unbiased estimates of the gradient!
- Initially proposed by Welling, M. and Teh, Y.W. 2011.

SGLD algorithm

- Interested in situations where the distribution $d\mu$ arises as the **posterior distribution** in a Bayesian inference problem with prior $d\mu_0$ and a large number $N >> 1$ of i.i.f observations z_i with likelihoods $p(z_i|X)$:

$$d\mu(X|z_i) \propto d\mu_0(X) \prod_{i=1}^{N} p(z_i|X).$$

- Denote for $i \in \{1, \dots, N\}$,
  - $V_i(X) = -\log(p(z_i|X))$
  - $V_0(X) = -\log(d\mu_0(X))$
  - $V = \sum_{i=0}^{N} V_i$
- ULA cost of one iteration is Nd which is **prohibitively large**

SGLD algorithm

- Welling, M and Teh, Y.W. suggested to replace ∇V with an unbiased estimate

$$\nabla V_0 + (N/p) \sum_{i \in S} \nabla V_i,$$

where S is a minibatch of size p.

- A **single update** of SGLD is thus (cost pd):

$$X_{k+1} = X_k - \gamma \left(\nabla V_0(X_k) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla V_i(X_k) \right) + \sqrt{2\gamma} Z_{k+1}$$

- Same idea as SGD.
- Two *sources of* ***randomness***: estimates of the gradient and Gaussian added noise to sample.

SGLD algorithm: need for variance reduction

$$X_{k+1} = X_k - \gamma \left(\nabla V_0(X_k) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla V_i(X_k) \right) + \sqrt{2\gamma} Z_{k+1}$$

Two sources of noise. For $\gamma = \gamma_0 / N$:

1. Noise from gradient estimates too big $\Rightarrow$ no sampling.
2. Need to decrease the variance: assume x_* unique minimizer of V,

$$X_{k+1} = X_k - \gamma \left(\nabla V_0(X_k) - \nabla V_0(x_*) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla V_i(X_k) - \nabla V_i(x_*) \right) + \sqrt{2\gamma} Z_{k+1}$$

If $\gamma = \gamma_0 / N$, **SGLD** $\sim$ **SGD**. Use variance control to sample.
Precise analysis from Moulines et al. (2018).

Non-convex Learning via SGLD

Classical learning problem:

- Find the minimum of $F(w) := \mathbb{E}_P[f(w, Z)]$ where f is not necessarily convex.
- Call $F_Z(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$

Consider the Langevin diffusion and its associated discretization :

$$dX_t = -\nabla F_Z(X_t) + \sqrt{2\beta^{-1}}dB_t$$

$$X_{k+1} = X_k - \eta\nabla f(w, z_k) + \sqrt{2\eta\beta^{-1}}\xi_k$$

Converges to $d\mu_z(dw) \propto \exp(-\beta F_z(w))$, when $\beta \sim 1/T$ is big, it concentrates around minimizers of F_z and hence F.

Non-convex Learning via SGLD

$$X_{k+1} = X_k - \eta \nabla f(w, z_k) + \sqrt{2\eta\beta^{-1}}\xi_k$$

(X_k) converges to $d\mu_z(w) \propto \exp(-\beta F_z(w))$, $\beta \sim 1/T$.

Theorem (Raginsky, Rakhlin, Telgarsky (2018))

For $k \geqslant \epsilon^{-4}$, $\eta \leqslant \epsilon^4$,

$$\mathbb{E}F(X_k) - F^* \leqslant c\epsilon + \frac{(\beta + d)^2}{n} + \frac{d \log(\beta + 1)}{\beta}$$

Sketch of proof: control of three terms

- How far from the true diffusion + invariant measure $\exp(-\beta F_z(w))$
- How far F_Z is from F
- How far a sample from $\exp(-\beta F_z(w))$ is near a minimizer of F_Z in terms of β

Conclusion

We have seen how Langevin Dynamics can be used to derive new algorithm for:

- Sampling
- Bayesian Learning
- Non-convex optimization

Problem with non-convexity: **metastability** of the markov process
$\longrightarrow$ old problem in computational chemistry.
"Particle remain trap in wells for a long time before going out."
There has been a huge effort in this community to tackle this problem

Inspiration for Machine Learning ?