Lecture 10: Sampling Methods (cont.d)

0. (Last week): MCMC & Gibbs Sampling

Goal: obtain samples $x^{(1)}, \dots, x^{(n)}$ s.t.

$$\frac{1}{n} \sum_{i=1}^{n} f(x^{(i)}) \approx \mathbb{E}_{x \sim p}[f(x)]$$

$p(x)$: target density

MCMC: Construct a Markov chain, ie. transition operator $T(x'|x)$, with stationary distribution $p(x)$

Then, sample from the chain : $x^{(t+1)} \sim T(\cdot | x^{(t)})$

If the chain converges, we have

$x^{(n)} \sim p$ for n large.

→ Sufficient condition for having p = stat. distr. :

$p(x)\, T(x'|x) = p(x')\, T(x|x')$ (detailed balance / reversible M.C.)

Gibbs Sampling: At time t :

- select variable j
- $x_j^{(t+1)} \sim p(x_j | x_{-j}^{(t)})$

→ Satisfies detailed balance !

1. Metropolis-Hastings

Q: Can we do MCMC with more general proposals?
$x^{(t+1)} \sim q(\cdot | x^{(t)})$ for some q?

→ In general, does not converge to $p(x)$
→ But, we can "correct" this to ensure detailed-balance using "Metropolis adjustments"

Metropolis-Hastings scheme:

At step t:

- Sample $y \sim q(\cdot | x^{(t)})$
- With probability

$$\alpha(x^{(t)}, y) := \min\left\{1, \frac{p(y)\, q(x^{(t)} | y)}{p(x^{(t)})\, q(y | x^{(t)})}\right\},$$

set $x^{(t+1)} = y$. Otherwise, set $x^{(t+1)} = x^{(t)}$.

Remarks:

- The proposal q can be arbitrary! But good proposals lead to faster convergence.
- $p(x)$ can be un-normalized (Z cancels out)
- If q is symmetric: $q(x'|x) = q(x|x')$ then $\alpha(x^{(t)}, y) = \min\left(1, \frac{p(y)}{p(x^{(t)})}\right)$

→ encourages moves towards higher $p(x)$.

- If q is reversible : $p(x)\,q(x'|x) = p(x')\,q(x|x')$

  then $\alpha \equiv 1$: always accept ⇒ faster convergence

- In the discrete case, the transition Kernel is :

$$T(x'=j|x=i) = T_{ij} = \begin{cases} q(j|i)\,\alpha(i,j) & \text{if } i \neq j \\ q(i|i) + \sum_{j' \neq i} q(j'|i)\,(1-\alpha(i,j')) & \text{o/w} \end{cases}$$

Fact: M-H satisfies detailed balance

proof: We want $p(i)\,T_{ij} = p(j)\,T_{ji}$ for any i,j

- If $i=j$, ok ✓
- If $i \neq j$, we have

$$p(i)\,q(j|i)\,\alpha(i,j) = p(i)\,q(j|i)\,\min\left(1, \frac{p(j)\,q(i|j)}{p(i)\,q(j|i)}\right)$$

$$= \min\{p(i)\,q(j|i),\ p(j)\,q(i|j)\}$$

$$= p(j)\,q(i|j)\,\alpha(j,i)$$

that is, $p(i)\,T_{ij} = p(j)\,T_{ji}$

□

Remark: In general, sampling is difficult in high dimension

$$x = (x_1, \dots, x_d) \in \mathbb{R}^d$$

Assume: $p(x) = p(x_1) \cdots p(x_d)$

$q(y|x) = q(y_1|x_1) \cdots q(y_d|x_d)$

It can be shown that there exists $c > 0$ such that:

$$\lim_{d \to \infty} \frac{1}{d} \log\left(1 - p_{\text{reject}}\right) \leq -c$$

($p_{\text{reject}} = \iint (1 - \alpha(x,y)) \, q(y|x) \, p(x) \, dx \, dy$

average reject prob.)

$\rightarrow$ $p_{\text{reject}} \geq 1 - e^{-cd}$

rejection w.p. exponentially close to 1 !

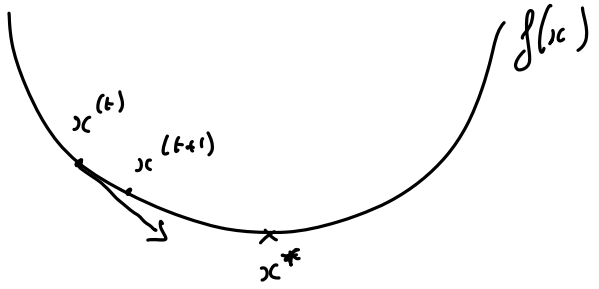
$\rightarrow$ Usually, $p(x)$ has some low-dimensional structure that should be exploited through a good proposal q.

$\rightarrow$ Common rule of thumb: adjust proposal size (e.g. # of variables resampled, width of a Gaussian) s.t. $p_{\text{rej}} \approx 25\%$.

[2.] Langevin Dynamics

Q: How can we obtain efficient sampling methods for high-dimensional, continuous distributions?

Motivation: Continuous optimization: using gradients can lead to faster, dimension-free convergence



$$x^{(t+1)} = x^{(t)} - h \nabla f(x^{(t)})$$

$$\|x^{(t)} - x^*\|^2 \leq (1-c)^t \|x^{(0)} - x^*\|^2$$

with c indep. of d!

→ How can we leverage gradient information in sampling?

e.g. $p(x) \propto e^{-f(x)}$: we want to sample more near minimizers of $f(x)$ (larger $p(x)$)

Langevin dynamics:

$$x^{(t+1)} = x^{(t)} + h \nabla (\log p(x^{(t)})) + \sqrt{2h}\, \xi^{(t)} \quad (LD)$$

with $\xi^{(t)} \sim \mathcal{N}(0, I)$

Note: if $p(x) \propto e^{-f(x)}$, we have $\nabla(\log p) = -\nabla f$

→ similar to gradient descent + Gaussian noise.

- **Metropolis adjusted Langevin** (MALA)

Idea: Use (LD) as a proposal in Metropolis-Hastings

$$\text{Accept } x^{(t+1)} \text{ w.p. } \min\left\{1, \frac{p(x^{(t+1)})\, q(x^{(t)} \mid x^{(t+1)})}{p(x^{(t)})\, q(x^{(t+1)} \mid x^{(t)})}\right\}$$

where $q(y \mid x) = \mathcal{N}(y \mid x + h \nabla(\log p)(x), 2h)$

Notes: ☺ · Unbiased thanks to M-H → converge to correct $p(x)$
· Leverage gradient for faster convergence

☹ · Could still lead to many rejections in high dim.

- **Unadjusted / Overdamped Langevin dynamics**

Idea: We can use (LD) without Metropolis adjustment
→ This could be biased (i.e. does not converge to p)
→ But for small step size h, the bias is small, and avoids large rejection rate in high dimension

Key insight: View (LD) as the discretization of an underlying continuous-time process.

Optimization view:

Gradient descent: $x^{(t+1)} = x^{(t)} - h \nabla f(x^{(t)})$
(discrete dynamics)

$\downarrow h \to 0$

Gradient flow: $\dot{x} = -\nabla f(x)$
(continuous dynamics, ODE)

($x(t)$: trajectory, $\dot{x} = \frac{dx}{dt}$)

Convergence: $x(t) \to x^* \in \operatorname{argmin} f$

Sampling with Langevin diffusions:

Discrete: $x^{(t+1)} = x^{(t)} - h \nabla f(x^{(t)}) + \sqrt{2h}\, \xi^{(t)}$

$\xi^{(t)} \sim \mathcal{N}(0, I)$

$\downarrow h \to 0$

Continuous: $dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t$

Diffusion Process
SDE (stochastic diff. eq.)

"drift" term

"Wiener process"
"Brownian motion"

X_t, W_t are random (processes)

Equivalent view: consider the time-evolution of the density of X_t, denoted $\rho(x, t)$

It can be shown that ρ obeys the PDE:

(FP) $$\partial_t \rho = \nabla \cdot (\rho \nabla f) + \Delta \rho$$ (Fokker-Planck equation)

Here, $\partial_t \rho := \frac{\partial}{\partial t} \rho(x,t)$

$$\nabla \cdot (\rho \nabla f) = \operatorname{div}(\rho \nabla f) = \sum_i \frac{\partial}{\partial x_i} \left(\rho(x,t) \frac{\partial f}{\partial x_i}(x) \right)$$

$$\Delta \rho = \sum_i \frac{\partial^2}{\partial x_i^2} \rho(x,t) = \nabla \cdot \nabla \rho$$

(Δ: Laplacian)

Fact: $p(x) \propto e^{-f(x)}$ is a stationary solution of (FP)

proof:

stationary: time independent: $\rho(x,t) \equiv \bar{\rho}(x) = p(x)$ so that $\partial_t \bar{\rho} = 0$.

We need to check: $\Delta p + \nabla \cdot (p \nabla f) = 0$

We have $\Delta p = \nabla \cdot \nabla p = \nabla \cdot (-p \nabla f)$ □

$\Rightarrow$ For $h \to 0$, we can hope to converge to p, even without Metropolis adjustments!

Convergence of (FP) to $p \propto e^{-f}$

Lemma: Let $(\rho_t)_t$ be a solution of (FP).

We have $\frac{d}{dt} H(\rho_t | p) \leq -I(\rho_t | p)$

where $H(q|p) = \int_{\mathcal{X}} \log\left(\frac{q}{p}\right) q$ (KL divergence)

$I(q|p) = \int_{\mathcal{X}} \|\nabla \log\left(\frac{q}{p}\right)\|^2 q$ (Fisher divergence)

proof:

Assume $p(x) = e^{-f(x)}$ (i.e. $Z=1$, w.l.o.g.)

$$\frac{d}{dt} H(\rho_t|p) = \frac{d}{dt} \int (\log \rho_t + f)\rho_t$$

$$= \int \frac{\partial_t \rho_t}{\rho_t} \rho_t + \int (\log \rho_t + f)\partial_t \rho_t$$

$$= \underbrace{\partial_t \int \rho_t(x)dx}_{=0} + \int (\log \rho_t + f)\nabla \cdot (\rho_t \nabla f + \nabla \rho_t)$$

$$= -\int \nabla(\log \rho_t + f) \cdot (\rho_t \nabla f + \nabla \rho_t) \quad \text{(Integration by parts)}$$

$$= -\int (\nabla \log \rho_t + \nabla f) \cdot (\rho_t \nabla f + \rho_t \nabla \log \rho_t)$$

$$= -\int \|\nabla \log \rho_t + \nabla f\|^2 \rho_t = -I(\rho_t | p)$$

□

Corollary. If we have $H(\rho|e^{-f}) \le \frac{1}{2\lambda} I(\rho|e^{-f})$ (*)

Then we obtain a convergence rate for (FP)

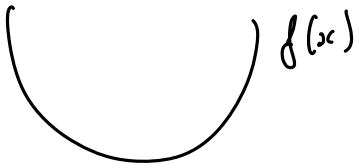
$$H(\rho_t|p) \le H(\rho_0|p)\, e^{-2\lambda t}$$

proof: use $\varphi'(t) \le -2\lambda\varphi(t) \Rightarrow \varphi(t) \le \varphi(0)\, e^{-2\lambda t}$ □

Remark: (*) is known as a **Log-Sobolev Inequality**

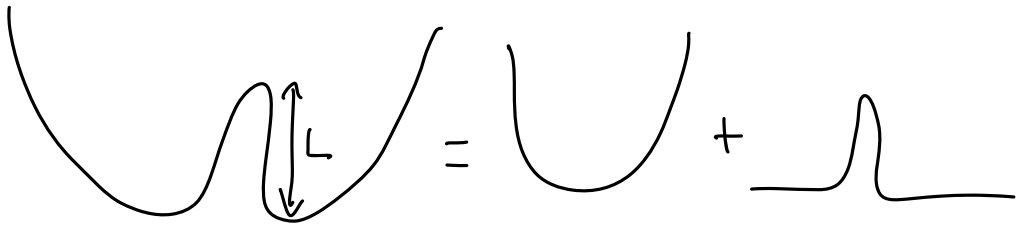
- if f is **λ-strongly convex**, then we have (*)

  $(\nabla^2 f \ge \lambda I)$

  $\Rightarrow$ sampling is easy!

- if $f = g + r$ with g strongly convex and $r(x) \in [0, L]$
  then λ in (*) is exponentially bad in L!!

  $\Rightarrow$ t needs to be exponential in L for convergence!

- Motivates other methods, e.g. **parallel tempering**
  (sample at different temperatures in parallel)