CPSC 486/586: Probabilistic Machine Learning *Out: March 1, 2023*

Problem Set 4

Instructor: Andre Wibisono *Due: March 29, 2023*

(P1) (Rejection sampling)

Consider the target probability distribution $\nu(x) \propto \exp(-f(x))$ on $\mathbb{R}^d$ where

$$f(x) = \frac{1}{2}\|x\|^2 + \|x\|_1 + \log\left(\sum_{i=1}^{d} e^{x_i}\right) \tag{1}$$

where $\|x\|^2 = x_1^2 + \cdots + x_d^2$ and $\|x\|_1 = |x_1| + \cdots + |x_d|$.

(a) Show that for all $x \in \mathbb{R}^d$,

$$f(x) \geq \frac{1}{2}\|x\|^2 + \frac{1}{d}x^\top \mathbf{1} + \log d$$

where $\mathbf{1} \in \mathbb{R}^d$ is the vector of all 1's.

(b) Implement rejection sampling to sample from ν using a Gaussian distribution μ as a proposal distribution (with mean and covariance that you can determine from part (b)). Draw $N = 1000$ samples from ν in dimension $d \in \{1, 2, \ldots, 9, 10\}$. Record and plot the number of samples drawn from μ as a function of d. How does it scale with d?

(P2) (Metropolis Random Walk)

For (P2), (P3), and (P4), we want to sample from the target distribution ν which is a mixture of Gaussians on $\mathbb{R}$:

$$\nu = \frac{1}{2}\mathcal{N}(-m, 1) + \frac{1}{2}\mathcal{N}(m, 1) \tag{2}$$

for some $m \in \mathbb{R}$. Below, we will use $m \in \{0, 1, 2\}$. (Note that when $m = 0$, ν is strongly log-concave; when $m = 1$, ν is log-concave; and when $m = 2$, ν is not log-concave.)

For this problem, implement the **Metropolis Random Walk (MRW)** algorithm to sample from ν, with step size $\eta > 0$ for $K = \lfloor \frac{T}{\eta} \rfloor$ iterations. Set $T = 100$ and $\eta = 0.1$. Here we start from $\rho_0 = \mathcal{N}(5, 10)$.

Metropolis Random Walk (MRW) to sample from ν:

(1) Start from $x_0 \sim \mathcal{N}(5, 10)$.

(2) For $k = 0, \ldots, K-1$:

i. From current $x_k \in \mathbb{R}$, draw from the proposal distribution $y_k \mid x_k \sim p(y_k \mid x_k)$. Recall for MRW, the proposal distribution is Gaussian with covariance $2\eta I$:

$$p(y \mid x) = \mathcal{N}(y \mid x, 2\eta) = \frac{1}{\sqrt{4\pi\eta}} e^{-\frac{(y-x)^2}{4\eta}}.$$

ii. Accept y_k with probability

$$\alpha(y_k) = \min\left\{1, \frac{\nu(y_k)}{\nu(x_k)}\right\}.$$

(If we accept y_k, then $x_{k+1} = y_k$; else, $x_{k+1} = x_k$.)

(3) Return x_K as an approximate sample from ν. Also return the acceptance rate A_K (the fraction of iterations k we accepted y_k).

Protocol $\mathsf{MRW}(m, \eta)$**:** Given $m \in \mathbb{R}$ and $\eta > 0$, run MRW above to draw $N = 500$ independent samples $x_K^{(n)}$, $1 \le n \le N$ (each with acceptance rate $A_K^{(n)}$). Let $\hat{\rho} = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_K^{(n)}}$ be the empirical distribution from the samples. Report:

(a) The mean and standard deviation of the acceptance rates $\{A_K^{(n)} : 1 \le n \le N\}$.

(b) Show a histogram of the empirical distribution vs. the target density (1) at initialization, and (2) at the end.

Task: For each $m \in \{0, 1, 2\}$ and $\eta = 0.1$, run the protocol $\mathsf{MRW}(m, \eta)$ above. Report the results and summarize what you can conclude about MRW.

(P3) (Metropolis-Adjusted Langevin Algorithm)

We want to sample from the same target distribution ν in (2).

In this problem, implement the following **Metropolis-Adjusted Langevin Algorithm (MALA)** with step size $\eta > 0$ for $K = \lfloor \frac{T}{\eta} \rfloor$ iterations. Set $T = 100$ and $\eta = 0.1$. Here we start from $\rho_0 = \mathcal{N}(5, 10)$.

Metropolis-Adjusted Langevin Algorithm (MALA) to sample from $\nu \propto e^{-f}$:

(1) Start from $x_0 \sim \mathcal{N}(0, 10)$.

(2) For $k = 0, \ldots, K-1$:

i. From current $x_k \in \mathbb{R}$, let

$$y_k = x_k - \eta \nabla f(x_k) + \sqrt{2\eta} z_k$$

where $z_k \sim \mathcal{N}(0, 1)$ is an independent Gaussian random variable in $\mathbb{R}$.

This is equivalent to using the following proposal distribution:

$$p(y \mid x) = \mathcal{N}(x - \eta \nabla f(x), \eta)(x) = \frac{1}{\sqrt{4\pi\eta}} e^{-\frac{(y-x+\eta\nabla f(x))^2}{4\eta}}.$$

ii. Accept y_k with probability

$$\alpha(y_k) = \min\left\{1, \frac{\nu(y_k)p(x_k \mid y_k)}{\nu(x_k)p(y_k \mid x_k)}\right\}.$$

(If we accept y_k, then $x_{k+1} = y_k$; else, $x_{k+1} = x_k$.)

(3) Return x_K as an approximate sample from ν. Also return the acceptance rate A_K (the fraction of iterations k we accepted y_k).

Repeat the task and protocol in (P2) with MALA in place of MRW. Report the results, and compare with MRW.

(P4) (Unadjusted Langevin Algorithm)

We want to sample from the same target distribution ν in (2).

In this problem, implement the following **Udjusted Langevin Algorithm (ULA)** with step size $\eta > 0$ for $K = \lfloor \frac{T}{\eta} \rfloor$ iterations. Set $T = 100$ and $\eta = 0.1$. Here we start from $\rho_0 = \mathcal{N}(5, 10)$.

Unadjusted Langevin Algorithm (ULA) to sample from $\nu \propto e^{-f}$:

(1) Start from $x_0 \sim \mathcal{N}(0, 10)$.

(2) For $k = 0, \ldots, K-1$:

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

where $z_k \sim \mathcal{N}(0, 1)$ is an independent Gaussian random variable in $\mathbb{R}$.

(3) Return x_K as an approximate sample from ν.

Repeat the task and protocol in (P2) with ULA in place of MRW. Report the results, and compare with MRW and MALA.

(P5) (Project)

(a) Remind us what is the main problem that you want to study in your project. Formulate the question as concretely as possible.

(b) Report any progress you have made in answering your question (e.g. what you have tried, what the results are, and what you will try next).

Additional questions for 586

(Q1) (Gaussian convolution decreases Wasserstein distance)

We will show that convolution with Gaussian density (which corresponds to adding Gaussian noise) decreases Wasserstein distance.[1] Concretely, given $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, for all $t \geq 0$ we define:

$$\rho(t) = \rho * \mathcal{N}(0, tI)$$

so that $\rho(0) = \rho$. Here $\mathcal{N}(0, tI)$ is the Gaussian density on $\mathbb{R}^d$ with mean 0 and covariance matrix tI, and $*$ is the convolution operator.

(a) Show that for all $\rho, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $t \geq 0$:

$$W_2(\rho(t), \nu(t)) \leq W_2(\rho, \nu).$$

(*Hint*: Use synchronous coupling.)

(b) Prove or disprove (i.e. provide a proof or a counterexample): For any $\rho, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\lim_{t\to\infty} W_2(\rho(t), \nu(t)) = 0.$$

(Q2) (Ornstein-Uhlenbeck process)

Recall the **Ornstein-Uhlenbeck (OU) process** is a stochastic process $(X_t)_{t\geq 0}$ in $\mathbb{R}^d$ that follows the stochastic differential equation (SDE):

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

for some $\alpha > 0$, where $(W_t)_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. Here the target distribution is $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$.

Recall that the solution $X_t \sim \rho_t$ of OU has the same distribution at each time $t \geq 0$ as:

$$X_t \overset{d}{=} e^{-\alpha t} X_0 + \sqrt{\frac{1 - e^{-2\alpha t}}{\alpha}} Z \tag{3}$$

where $Z \sim \mathcal{N}(0, I)$. Let $Q_t = \{Q_{t,x} \colon x \in \mathbb{R}^d\}$ denote the Markov chain that sends $X_0 \sim \rho_0$ to $X_t \sim \rho_t$ following (3), so $Q_{t,x} = \mathcal{N}(e^{-\alpha t} x, \frac{1-e^{-2\alpha t}}{\alpha} I)$.

[1] Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability distributions ρ on $\mathbb{R}^d$ with finite second moment ($\mathbb{E}_\rho[\|X\|^2] < \infty$). Recall the *Wasserstein distance* between $\rho, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is

$$W_2(\rho, \nu)^2 = \inf_{(X,Y)\sim\pi^{XY}} \mathbb{E}[\|X - Y\|^2]$$

where the minimization is over all coupling π^{XY} of ρ and ν, i.e. a joint probability distribution of $(X, Y) \sim \pi^{XY}$ with the correct marginals $X \sim \rho$ and $Y \sim \nu$ (which means $\rho(x) = \int_{\mathbb{R}^d} \pi^{XY}(x, y)\, dy$ and $\nu(y) = \int_{\mathbb{R}^d} \pi^{XY}(x, y)\, dx$).

(a) Show that Q_t is reversible with respect to $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$, i.e. for all $x, y \in \mathbb{R}^d$:

$$\nu(x) Q_{t,x}(y) = \nu(y) Q_{t,y}(x).$$

(b) Show that OU is *variance-preserving*: If $\mathsf{Cov}(X_0) = \frac{1}{\alpha} I$, then $\mathsf{Cov}(X_t) = \frac{1}{\alpha} I$ for all $t \geq 0$.

(c) Let ρ_0 be arbitrary. Use synchronous coupling to show that for all $t \geq 0$:

$$W_2(\rho_t, \nu)^2 \leq e^{-2\alpha t} W_2(\rho_0, \nu)^2.$$

(Q3) (Discrete-time Ornstein-Uhlenbeck algorithm)

Suppose $x_k \in \mathbb{R}^d$ follows the discrete-time **Ornstein-Uhlenbeck (OU) algorithm**:

$$x_{k+1} = x_k - \eta \alpha x_k + \sqrt{2\eta} z_k$$

where $\eta > 0$ is step size, and where $z_k \sim \mathcal{N}(0, I)$ is independent. Starting from $x_0 \sim \rho_0$, we get $x_k \sim \rho_k$ along the OU algorithm.

Let $\nu = \mathcal{N}(0, \frac{1}{\alpha} I)$ which is the target distribution of the continuous-time OU process.

(a) Show that $x_k \sim \rho_k$ can be written as (where $\overset{d}{=}$ is equality in distribution):

$$x_k \overset{d}{=} (1 - \alpha\eta)^k x_0 + \sqrt{2A_k} \hat{z}_k$$

for some $A_k > 0$, where $\hat{z}_k \sim \mathcal{N}(0, I)$ is independent. Determine A_k explicitly.

(b) Assume $0 \leq \eta < \frac{2}{\alpha}$. Show that the stationary distribution of the OU algorithm is:

$$\nu_\eta = \mathcal{N}\left(0, \frac{1}{\alpha(1 - \frac{1}{2}\alpha\eta)} I\right).$$

That is, show that if $x_k \sim \nu_\eta$, then $x_{k+1} \sim \nu_\eta$. Is the OU algorithm reversible with respect to ν_η?

(c) Assume $0 \leq \eta < \frac{2}{\alpha}$. Show that for any ρ_0, along the OU algorithm, for any $k \geq 0$:

$$W_2(\rho_k, \nu_\eta) \leq (1 - \alpha\eta)^k W_2(\rho_0, \nu_\eta).$$

(*Hint*: Use synchronous coupling.)

(d) Note $\nu_\eta \neq \nu$, hence the OU algorithm is biased (does not converge to the target ν). Show that for $0 \leq \eta \leq \frac{1}{\alpha}$:

$$W_2(\nu, \nu_\eta) \leq \frac{\eta\sqrt{\alpha d}}{2}.$$

(*Hint:* Recall the Bures-Wasserstein distance between Gaussians.)

(e) Assume $0 \le \eta \le \frac{1}{\alpha}$. Show that for any ρ_0, along the OU algorithm, for any $k \ge 0$:

$$W_2(\rho_k, \nu) \le (1 - \alpha\eta)^k W_2(\rho_0, \nu_\eta) + \frac{\eta\sqrt{\alpha d}}{2}.$$

(f) Suppose we want to generate $x_K \sim \rho_K$ with error $W_2(\rho_K, \nu) \le \varepsilon$ for some $0 < \varepsilon \ll 1$. Show that it suffices to run the OU algorithm with step size $\eta = \frac{\varepsilon}{\sqrt{\alpha d}}$ for

$$T \ge \frac{\sqrt{d}}{\sqrt{\alpha}\,\varepsilon} \log \frac{2W_2(\rho_0, \nu_\eta)}{\varepsilon}$$

iterations.