CPSC 661: Sampling Algorithms in Machine Learning	*Out: April 21, 2021*
Problem Set 3	
Lecturer: Andre Wibisono	*Due: May 12, 2021*

Instruction

Solve at least **2** of the following problems (feel free to solve as many as you'd like). Each problem has equal worth, so you can choose the ones that are most interesting to you. Collaboration is allowed and encouraged, but please write your own solution and acknowledge your collaborators. Submit the solution as a single PDF file via Canvas. If there are questions, please post a discussion on Canvas or email `andre.wibisono@yale.edu`.

1 Continuity equation

Recall the continuity equation for an ODE $\dot{X}_t = v(X_t)$ is the PDE $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$ satisfied by the density $\rho_t \colon \mathbb{R}^n \to \mathbb{R}$ of $X_t \in \mathbb{R}^n$. In this problem you will check that the continuity equation holds for linear vector field $v(x) = -\alpha x$ for some $\alpha > 0$.

1. Find the solution X_t (in terms of X_0) of the ODE:
$$\dot{X}_t = -\alpha X_t.$$
2. Suppose X_0 has density ρ_0. Find the density ρ_t of X_t above (in terms of ρ_0).
3. Check that the continuity equation holds: Compute both $\frac{\partial \rho_t}{\partial t}(x)$ and $-\nabla \cdot (\rho_t v)(x)$, and show they are equal.

2 Heat equation

Show that Gaussian convolution:
$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$$
satisfies the heat equation:
$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t.$$
That is, given $\rho_0 \colon \mathbb{R}^n \to \mathbb{R}$, define $\rho_t \colon \mathbb{R}^n \to \mathbb{R}$ for each $t > 0$ by
$$\rho_t(x) = \frac{1}{(4\pi t)^{n/2}} \int_{\mathbb{R}^n} \rho_0(y) \exp\left(-\frac{\|x-y\|^2}{4t}\right) dy.$$
Compute $\frac{\partial \rho_t(x)}{\partial t}$ and $\Delta \rho_t(x)$, and show they are equal.

3 Ornstein-Uhlenbeck

Recall the Ornstein-Uhlenbeck process $dX_t = -\alpha X_t\, dt + \sqrt{2}\, dW_t$ has solution

$$X_t \stackrel{d}{=} e^{-\alpha t} X_0 + \sqrt{\frac{1-e^{-2\alpha t}}{\alpha}} Z, \quad Z \sim \mathcal{N}(0, I).$$

Show that the density ρ_t of X_t:

$$\rho_t(x) = \frac{1}{(2\pi(\frac{1-e^{-2\alpha t}}{\alpha}))^{n/2}} \int_{\mathbb{R}^n} \rho_0(y) \exp\left(-\frac{\alpha(x - e^{-\alpha t} y)^2}{2(1-e^{-2\alpha t})}\right) dy$$

satisfies the Fokker-Planck equation:

$$\frac{\partial \rho_t(x)}{\partial t} = \alpha \nabla \cdot (\rho_t(x)\, x) + \Delta \rho_t(x).$$

4 Fisher information bound

Let ρ be a probability distribution with differentiable density function $\rho\colon \mathbb{R}^n \to \mathbb{R}$. Recall the Fisher information of ρ is

$$J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2] = \int_{\mathbb{R}^n} \rho(x) \|\nabla \log \rho(x)\|^2\, dx$$

and the second-order Fisher information of ρ is

$$K(\rho) = \mathbb{E}_\rho[\|\nabla^2 \log \rho\|_{\mathrm{HS}}^2] = \int_{\mathbb{R}^n} \rho(x) \|\nabla^2 \log \rho(x)\|_{\mathrm{HS}}^2\, dx$$

where $\|A\|_{\mathrm{HS}}^2 = \mathrm{Tr}(A^\top A)$ is the Hilbert-Schmidt norm of a matrix A.

1. By integration by parts, show that $J(\rho)$ can also be written as

   $$J(\rho) = -\mathbb{E}_\rho[\Delta \log \rho]$$

   where $\Delta = \mathrm{Tr}(\nabla^2)$ is the Laplacian.

2. Prove that

   $$K(\rho) \geq \frac{J(\rho)^2}{n}.$$

Hint: If A has eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\|A\|_{\mathrm{HS}}^2 = \sum_{i=1}^n \lambda_i^2$ and $\mathrm{Tr}(A) = \sum_{i=1}^n \lambda_i$.

5 Convergence rate of entropy along heat equation

Consider the heat equation

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$

from an initial $\rho_0 \colon \mathbb{R}^n \to \mathbb{R}$. Recall that we have $\frac{d}{dt} H(\rho_t) = J(\rho_t)$ and $\frac{d}{dt} J(\rho_t) = -2K(\rho_t)$.

1. Using the inequality from Problem 4, prove that along the heat equation,

$$J(\rho_t) \le \left(\frac{1}{J(\rho_0)} + \frac{2t}{n} \right)^{-1}.$$

2. Integrate the inequality above to show that along the heat equation,

$$H(\rho_t) \ge H(\rho_0) - \frac{n}{2} \log \left(1 + \frac{2t}{n} J(\rho_0) \right).$$

6 LSI implies Poincaré via linearization

Suppose ρ is close to ν, in the sense that $\rho(x) = (1 + \epsilon g(x))\nu(x)$ for some $\epsilon > 0$ and $g : \mathbb{R}^n \to \mathbb{R}$. Recall from Problem Set 1 that as $\epsilon \to 0$, relative entropy recovers the χ^2-divergence:

$$H_\nu(\rho) = \frac{1}{2} \mathrm{Var}_\nu(g) + O(\epsilon^3).$$

1. Show that as $\epsilon \to 0$, relative Fisher information recovers the Dirichlet energy:

$$J_\nu(\rho) = \mathbb{E}_\nu \left[\|\nabla g\|^2 \right] + O(\epsilon^3).$$

2. Conclude that log-Sobolev inequality:

$$J_\nu(\rho) \ge 2\alpha H_\nu(\rho) \quad \text{for all } \rho \in \mathcal{P}(\mathbb{R}^n)$$

   implies the Poincaré inequality:

$$\mathbb{E}_\nu \left[\|\nabla g\|^2 \right] \ge \alpha \mathrm{Var}_\nu(g) \quad \text{for all } g \colon \mathbb{R}^n \to \mathbb{R}.$$

7 Bias of Forward-Flow algorithm

Consider minimizing a composite objective function:

$$\min_{x \in \mathbb{R}} f(x) + g(x)$$

where f and g are both quadratic functions on $\mathbb{R}$:

$$f(x) = \frac{1}{2}(x-1)^2, \qquad g(x) = \frac{1}{2}(x+1)^2.$$

Note the minimum of $f + g$ is at $x^* = 0$.

1. Write down the iteration of the Forward-Flow algorithm for minimizing $f + g$ (obtained by applying gradient descent for f with step size η, followed by the exact gradient flow for g for time η).

2. Compute the biased limit of the Forward-Flow algorithm above.

3. Write down the iteration of the Forward-Backward algorithm for minimizing $f + g$ (obtained by applying gradient descent for f with step size η, followed by the proximal method for g with step size η).

4. Check that the Forward-Backward algorithm above is unbiased (the iterates converge to 0).

8 Convergence rate of heavy ball dynamics

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be an α-strongly convex function for some $\alpha > 0$. Consider the heavy ball dynamics:

$$\ddot{X}_t + 2\sqrt{\alpha}\,\dot{X}_t + \nabla f(X_t) = 0$$

starting from arbitrary $X_0 \in \mathbb{R}^n$, $\dot{X}_0 \in \mathbb{R}^n$. Show that f is minimized exponentially fast:

$$f(X_t) - f(x^*) \leq O(e^{-\sqrt{\alpha}t})$$

where $x^* = \arg\min_{x\in\mathbb{R}^n} f(x)$.

Hint: Show the following is a Lyapunov function (decreasing with time):

$$\mathcal{E}_t = e^{\sqrt{\alpha}t}\Big(f(X_t) - f(x^*) + \frac{\alpha}{2}\left\|X_t - x^* + \frac{1}{\sqrt{\alpha}}\dot{X}_t\right\|^2\Big).$$