Mathematical Algorithms II, M1407.000500
E. Ryu
Fall 2022

Homework 1
Due 5pm, Monday, September 19, 2022

Problem 1: *Control variates.* Let X and Y be scalar-valued random variables such that

$$\mathbb{E}[X] = I, \qquad \mathbb{E}[Y] = 0$$

and

$$\mathbb{E}[(X-I)^2] = \Sigma_{XX}, \qquad \mathbb{E}[(Y)^2] = \Sigma_{YY}, \qquad \mathbb{E}[(X-I)Y] = \Sigma_{XY}.$$

Assume $0 < \Sigma_{XX} < \infty$ and $0 < \Sigma_{YY} < \infty$. Our goal is to estimate I with small variance. Clearly,

$$\mathbb{E}[X + \gamma Y] = I$$

for any $\gamma \in \mathbb{R}$. Find the solution to

$$\underset{\gamma \in \mathbb{R}}{\text{minimize}} \quad \mathrm{Var}(X + \gamma Y).$$

Remark. The point is that if X and Y are correlated, i.e., if $\Sigma_{XY} \neq 0$, then the optimal γ is non-zero. In such setups, Y is referred to as a *control variate*, as it is a random variable (variate) one can use to control (reduce) the variance. Of course, the variance is reduced only when γ is chosen well.

Problem 2: *Tweedie's formula.* Consider the vector-valued continuous random variables

$$Y = X + Z \in \mathbb{R}^n,$$

where $X \sim p_X$ and $Z \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$ are independent. (To clarify, p_X is a probability density function.) Write p_Y to denote the probability density function of Y. Show that

$$\mathbb{E}[X \,|\, Y] = Y + \Sigma \nabla \log p_Y(Y).$$

You may swap the order of derivatives and integrals without proof.

Hint. Start with the scalar case (so $n = 1$) with $\Sigma = 1$. Define

$$\ell(y) = \frac{p_Y(y)}{p_Z(y)} = \frac{\int_{\mathbb{R}} p_{Y|X}(y \,|\, x) p_X(x)\, dx}{p_Z(y)}$$

and show

$$\frac{d}{dy}\ell(y) = \mathbb{E}[X \,|\, Y]\ell(y).$$

Then, use the formula

$$\mathbb{E}[X \,|\, Y] = \frac{d}{dy}\log \ell(y).$$

Clarification. We do not assume X is a Gaussian.

Problem 3: Let $\mu_\theta(s) \in \mathbb{R}^n$ and $\Sigma_\theta(s) \in \mathbb{R}^{n\times n}$ be neural networks parameterized by $\theta \in \mathbb{R}^P$. Assume $\Sigma_\theta(s)$ is symmetric and strictly positive definite for any $s \in \mathcal{S}$ and $\theta \in \mathbb{R}^P$. Given $s \in \mathcal{S}$, let

$$a = \tanh(z), \qquad z \sim \mathcal{N}(\mu_\theta(s), \Sigma_\theta(s)).$$

Let $\pi_\theta(a \,|\, s)$ be the implicitly defined probability density function of the random variable $a \in \mathbb{R}^n$. Show that

$$\begin{aligned} z &= \tanh^{-1}(a) \\ \log \pi_\theta(a \,|\, s) &= -\frac{1}{2}\log\det\Sigma_\theta(s) - \frac{1}{2}(z-\mu_\theta(s))^\intercal \Sigma_\theta^{-1}(s)(z-\mu_\theta(s)) \\ &\quad - \frac{n}{2}\log(2\pi) - \sum_{i=1}^n \log(1-\tanh^2(z_i)). \end{aligned}$$

Problem 4: Let $X_1, \dots, X_T$ be a sequence with the hidden Markov property with respect to $h_1, \dots, h_T \in \mathcal{H}$, where $|\mathcal{H}| = m < \infty$. Define

$$\rho_T(h_T) = 1, \qquad \forall\, h_T \in \mathcal{H}$$

and

$$\rho_{t-1}(h_{t-1}) = \mathbb{P}(X_t, \dots, X_T \,|\, h_{t-1}), \qquad \forall h_{t-1} \in \mathcal{H}.$$

Show that

$$\rho_{t-1} = g(X_t, \rho_t)$$

for some function $g \colon \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}^m$.

Problem 5: Consider the setup of Problem 4 and let $s_1, \dots, s_T$ be as defined in the lecture. Let

$$\mu_t(X_t) = \sum_{h_t \in \mathcal{H}} s_t(h_t)\rho_t(h_t)\mathbb{P}(X_t \,|\, h_t).$$

Assume $X_t \in \mathcal{X}$ and $|\mathcal{X}| < \infty$, i.e., X_t is a discrete random variable with finite possible realizations, for $t = 1, \dots, T$. Show that

$$\mathbb{P}(X_t \,|\, X_1, \dots, X_{t-1}, X_{t+1}, \dots, X_T) = \mu_t^\flat(X_t),$$

where $\mu_t^\flat$ is the normalized probability mass function corresponding to μ_t.

Problem 6: *Backprop for FFJORD.* Consider the neural ODE

$$\frac{d}{ds}z(s) = f(z(s), \theta, s), \qquad s \in [0,1].$$

Let $\mathcal{F}_\theta^{1,0} \colon \mathbb{R}^D \to \mathbb{R}^D$ be the flow operator from pseudo-time $s = 1$ to $s = 0$. Let $x \in \mathbb{R}^D$ be a given datapoint, and consider the problem of evaluating a stochastic gradient of

$$\log p(x) = \log p_0\left(\mathcal{F}_\theta^{1,0}(x)\right) - \int_0^1 \operatorname{Tr}\left(\frac{\partial f}{\partial z}(z(s), \theta, s)\right)\, ds,$$

where p_0 is a suitable latent distribution. To this end, sample a random $\nu \in \mathbb{R}^D$ such that $\mathbb{E}[\nu\nu^\intercal] = I$ and solve

$$\frac{d}{ds}\begin{bmatrix} z \\ \lambda \end{bmatrix}(s) = \begin{bmatrix} f \\ -\nu^\intercal \frac{\partial f}{\partial z}\nu \end{bmatrix}(z(s), \theta, s)$$

with terminal values $z(1) = x$ and $\lambda(1) = 0$ to obtain $z(0)$ and $\hat{\ell} = \log p_0(z_0) - \lambda(0)$. The argument with the Hutchinson estimator shows that

$$\hat{\ell} = \log p_0\left(\mathcal{F}_\theta^{1,0}(x)\right) - \int_0^1 \nu^\intercal \frac{\partial f}{\partial z}(z(s), \theta, s)\nu\, ds,$$

is an unbiased estimator of $\log p(x)$. Show that solving

$$\frac{da}{ds}(s) = -a\frac{\partial f}{\partial z}(z(s), \theta, s) - \frac{\partial}{\partial z}\nu^\intercal \frac{\partial f}{\partial z}(z(s), \theta, s)\nu, \qquad s \in [0,1]$$

and

$$\frac{db}{ds}(s) = -a\frac{\partial f}{\partial \theta}(z(s), \theta, s) - \frac{\partial}{\partial \theta}\nu^\intercal \frac{\partial f}{\partial z}(z(s), \theta, s)\nu, \qquad s \in [0,1]$$

with initial conditions $a(0) = \nabla \log p_0(z(0))$ and $b(0) = 0$ yields

$$b(1) = \frac{\partial \hat{\ell}}{\partial \theta}.$$

Hint. Apply the adjoint method theorem with reversed pseudo-time and

$$\tilde{z} = \begin{bmatrix} z \\ \lambda \end{bmatrix}, \qquad \tilde{f}(z(s), \theta, s) = \begin{bmatrix} f \\ -\nu^\intercal \frac{\partial f}{\partial z}\nu \end{bmatrix}(z(s), \theta, s), \qquad \mathcal{L}(\tilde{z}(0)) = \hat{\ell} = \log p_0(z(0)) - \lambda(0).$$

Then, simplify the dynamics using the fact that $\frac{\partial \tilde{f}(z(s), \theta, s)}{\partial \lambda} = 0$.

Problem 7: Let $\rho\colon [0,T] \to \mathbb{R}$. Consider the d-dimensional SDE

$$dX_t = f(X_t, t)dt + \rho(t)dW_t, \qquad t \in [0,T]$$

with initial condition $X_0 \sim p_0$. Let $\{p_t\}_{t=0}^T$ be the marginal marginal density functions. Show that $\{p_t\}_{t=0}^T$ satisfies the Fokker–Planck equation

$$\partial_t p_t = -\nabla_x \cdot (f p_t) + \frac{\rho^2}{2}\Delta p_t,$$

where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator.

Problem 8: Let $\sigma_t > 0$ be a smooth non-decreasing function for $0 \le t \le T$. Define

$$\rho(t) = \sqrt{\frac{d}{dt}\sigma_t^2}, \qquad t \in [0,T].$$

For simplicity, assume $d = 1$. Consider the SDE

$$dX_t = \rho(t)dW_t, \qquad t \in [0,T]$$

with initial condition $X_0 \sim p_0$. Show $X_t \,|\, X_0 \sim \mathcal{N}(X_0, \sigma_t^2)$ by verifying that

$$p_t(x) = \int_{\mathbb{R}^d} p_{t|0}(x \,|\, y) p_0(y)\, dy = \int_{\mathbb{R}^d} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(x-y)^2}{2\sigma_t^2}\right] p_0(y)\, dy$$

satisfies the Fokker–Planck equation.

Remark. It is actually sufficient to assume that σ_t is absolutely continuous, rather than smooth.

Problem 9: Consider the ODE

$$dX_t = \left(f(X_t, t) - \frac{g^2(t)}{2} \nabla_{X_t} \log p_t(X_t) \right) dt, \qquad t \in [0,T]$$

with terminal condition $X_T \sim p_T$. Let $\{p_t\}_{t=0}^T$ be the marginal marginal density functions. For simplicity, assume $d = 1$. Show that $\{p_t\}_{t=0}^T$ satisfies the Fokker–Planck equation

$$\partial_t p_t = -\partial_x (f p_t) + \frac{g^2}{2}\partial_x^2 p_t.$$

Hint. As with the derivation of the Fokker–Planck equation, start with

$$\partial_t \mathbb{E}_{X\sim p_t}[\varphi(X)] \approx \frac{1}{\varepsilon}\mathbb{E}_{X\sim p_t}\left[\varphi\left(X + \varepsilon\left(f(X,t) - \frac{g^2(t)}{2}\nabla_X \log p_t(X)\right)\right) - \varphi(X)\right].$$