Generative AI and Foundation Models, M3309.001800
E. Ryu
Spring 2024

Homework 1
Due 5pm, Monday, April 15, 2024

Problem 1: *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

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

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

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

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

and show

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

Then, use the formula

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

Clarification. We do not assume X is a Gaussian.

Problem 2: *D_{KL} of Gaussian random variables.* Show that

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu_0, \sigma_0^2 I) \| \mathcal{N}(\mu_1, \sigma_1^2 I)\right) = \frac{1}{2\sigma_1^2}\|\mu_1 - \mu_0\|^2 + \frac{(\sigma_0^2/\sigma_1^2 - 1)d}{2} + d \log\left(\frac{\sigma_1}{\sigma_0}\right),$$

where d is the underlying dimension of the random variables, $\mu_0, \mu_1 \in \mathbb{R}^d$, $\sigma_0 > 0$, and $\sigma_1 > 0$.

Remark. In the context of deep learning, if σ_0 and σ_1 are not trainable parameters, then we can write

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu_0, \sigma_0^2 I) \| \mathcal{N}(\mu_1, \sigma_1^2 I)\right) = \frac{1}{2\sigma_1^2}\|\mu_1 - \mu_0\|^2 + C.$$

Problem 3: *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

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

where p_Z is a suitable latent distribution. We first sample a random $\nu \in \mathbb{R}^D$ such that $\mathbb{E}[\nu \nu^\intercal] = I$ and solve

$$\begin{aligned} \log \widehat{p_\theta^{(\text{gen})}}(X) &= \log p_Z(z(0)) - \hat{\ell}(0) \\ \frac{d}{dt} \begin{bmatrix} z(s) \\ \hat{\ell}(s) \end{bmatrix} &= \begin{bmatrix} f\left(z(s), \theta, s\right) \\ -\nu^\intercal \frac{\partial f}{\partial z} \nu(z(s), \theta, s) \end{bmatrix} \qquad \text{for } s \in [0, 1] \\ \begin{bmatrix} z(1) \\ \hat{\ell}(1) \end{bmatrix} &= \begin{bmatrix} X \\ 0 \end{bmatrix} \end{aligned}$$

in reverse pseudo-time. In class, we have established that

$$\mathop{\mathbb{E}}_\nu [\log \widehat{p_\theta^{(\text{gen})}}(X)] = \log p_\theta^{(\text{gen})}(X).$$

Show that solving

$$\begin{aligned} \frac{\partial \log \widehat{p_\theta^{(\text{gen})}}(X)}{\partial \theta} &= b(1) \\ \dot{a}(s) &= -a \frac{\partial f}{\partial z}(z(s), \theta, s) - \frac{\partial}{\partial z} \nu^\intercal \frac{\partial f}{\partial z}(z(s), \theta, s) \nu, \qquad s \in [0, 1] \\ \dot{b}(s) &= -a \frac{\partial f}{\partial \theta}(z(s), \theta, s) - \frac{\partial}{\partial \theta} \nu^\intercal \frac{\partial f}{\partial z}(z(s), \theta, s) \nu, \qquad s \in [0, 1] \\ a(0) &= \frac{\partial \log p_0(z)}{\partial z} \Big|_{z = z(0)} \in \mathbb{R}^{1 \times D}, \qquad b(0) = 0 \in \mathbb{R}^{1 \times P} \end{aligned}$$

in forward pseudo-time yields a stochastic gradient of the log-likelihood, i.e., show that

$$\mathop{\mathbb{E}}_\nu \left[\frac{\partial \log \widehat{p_\theta^{(\text{gen})}}(X)}{\partial \theta} \right] = \frac{\partial}{\partial \theta} \log p_\theta^{(\text{gen})}(X).$$

Hint. Apply the adjoint state method with

$$\tilde{z} = \begin{bmatrix} z \\ \lambda \end{bmatrix}, \quad \tilde{f}(z(s), \theta, s) = \begin{bmatrix} f \\ -\nu^\intercal \frac{\partial f}{\partial z} \nu \end{bmatrix} (z(s), \theta, s), \quad \mathcal{L}(\tilde{z}(0)) = \log p_Z(z(0)) - \lambda(0)$$

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

Problem 4: *Equivalence of graph-form backward passes.* Let $G = (V, E)$ be a DAG representing a computation graph as discussed in the backdrop lecture. Show that the graph-form backdrop code version 1

```
# Forward pass given u.value for source nodes
for v in V :   # In linear topological order
  v.value = v.fn( [u.value for u->v] )

for v in V :   # .zero_grad()
  v.grad = 0

# Backward pass
v_out.grad = 1
for v in V :   # In reversed linear topological order
  for w such that v->w :
    v.grad += w.grad @ w.fn.grad(v)
```

and version 2

```
# Forward pass given u.value for source nodes
for v in V :
  v.value = v.fn( [u.value for u->v] )

for v in V :   # .zero_grad()
  v.grad = 0

v_out.grad = 1
for v in V :   # In reversed linear topological order
  for u such that u->v :
    u.grad += v.grad @ v.fn.grad(u)
```

are equivalent.

Hint. First, transform the loop

```
for v in V :   # In reversed linear topological order
  for w such that v->w :
    v.grad += w.grad @ w.fn.grad(v)
```

into

```
for v in V :   # In reversed linear topological order
  for w in V :   # In any order
    if v->w :
      v.grad += w.grad @ w.fn.grad(v)
```

Problem 5: 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 6: Let $\sigma_t \geq 0$ be a smooth non-decreasing function for $0 \leq t \leq T$. 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 7: *Sampling SDE family.* Consider the forward-time SDE

$$dX_t = f(X_t, t)dt + g(t)dW_t$$

with $X_0 \sim p_0$. Write $\{p_t\}_{t\geq 0}$ to denote the marginal densities of $\{X_t\}_{t\geq 0}$. Show that the reverse-time SDEs defined by

$$d\overline{X}_t = \left(f(\overline{X}_t, t) - \left(1 - \frac{\lambda}{2}\right) g^2(t) \nabla_{\overline{X}_t} \log p_t(\overline{X}_t)\right) dt + \sqrt{1-\lambda}\, g(t) d\overline{W}_t$$

for $t \in [0,T]$ with $\overline{X}_T \sim p_T$ have the same marginals $\{p_t\}_{t\in[0,T]}$ for all $\lambda \leq 1$. For simplicity, assume $\overline{X}_t \in \mathbb{R}$ and $\overline{W}_t$ is the 1-dimensional reverse-time Brownian motion.

Remark. Note that $\lambda = 0$ corresponds to the standard SDE sampling while $\lambda = 1$ corresponds to the standard ODE sampling of diffusion models.

Remark. This result holds more generally for $s_\theta\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ and $X_t \in \mathbb{R}^d$, but we assume $d = 1$ for simplicity.