STAT 542: Homework 7

Ahmadreza Eslaminia NetID: ae15

Please make sure that your solutions are readable and the file size is reasonable. Typing the answers is highly encouraged.

Problem 1.

The purpose of this exercise is to understand why it helps to introduce ELBO in VAE. See p6 in s17_VAE.pdf for the definition of ELBO. Consider the following example: $p_\theta(x, z) = p(z)p_\theta(x \mid z)$, where $p(z = \cdot)$ is $\mathcal{N}(0, 1)$, and $p_\theta(x = \cdot \mid z)$ is $\mathcal{N}(z + \theta, 1)$. In other words, in the notation of p5 in s17_VAE.pdf, we have $\mu_\theta(z) = z+\theta$ and $\Sigma_\theta(z) = 1$. We chose this simplified example to streamline the calculations, but we can see that the intuition here (that taking the log reduces the variation of a function) extends to more general settings.

- Can you find an explicit expression for $p_\theta(z \mid x)$?
- Consider the variational approach, whereby we consider a variational family $q_\lambda = \mathcal{N}(\lambda_1, \lambda_2)$ of distributions, with $\lambda = (\lambda_1, \lambda_2) \in (0, \infty)^2$. In order to calculate $p_\theta(x, z)$, a naive approach is to consider

$$p_\theta(x, z) = \mathbb{E}_{q_\lambda}\left[p_\theta(x, z)/q_\lambda(z)\right] \tag{1}$$

$$\approx \frac{1}{n}\sum_{i=1}^{k} p_\theta\left(x, z^{(i)}\right)/q_\lambda\left(z^{(i)}\right) \tag{2}$$

where $z^{(i)}, i = 1, \ldots, k$ are samples drawn from q_λ. Here (2) is an unbiased estimator of (1) due to linearity of expectation. To make it a good estimator we also need to ensure its variance is small. Question: show that there exists a λ^* such that the variance of (2) is infinite whenever $\lambda_2 < \lambda^*$. What is the largest choice of λ^* ?

- The ELBO method addresses the above issue: consider the method on p6 and p7 in s17_VAE.pdf. What is the range of λ for which the variance of $\text{ELBO}(x, \theta, \lambda)$ is finite?

Solution 1.

Part 1.

Given the model $p_0(x, z) = p(z)p_0(x|z)$, where $p(z) = \mathcal{N}(0, 1)$ and $p_0(x|z) = \mathcal{N}(z + \theta, 1)$, we need to find the explicit expressions for various distributions.

The conditional distribution of x given z is defined as:

$$p_0(x|z) = \mathcal{N}(x|z + \theta, 1)$$

The joint distribution is the product of the marginal and conditional distributions:

$$p_0(x, z) = \mathcal{N}(z|0, 1)\mathcal{N}(x|z + \theta, 1)$$

To find the marginal distribution $p_0(x)$, integrate out z from the joint distribution:

$$p_0(x) = \int \mathcal{N}(z|0, 1)\mathcal{N}(x|z + \theta, 1)\, dz$$

Using the convolution of two normal distributions, the resulting marginal distribution of x is:

$$p_0(x) = \mathcal{N}(x|\theta, 2)$$

Applying Bayes' theorem:

$$p_0(z|x) = \frac{p_0(x|z)p(z)}{p_0(x)}$$

Substituting in the distributions:

$$p_0(z|x) = \frac{\mathcal{N}(x|z + \theta, 1)\mathcal{N}(z|0, 1)}{\mathcal{N}(x|\theta, 2)}$$

Part 2.

We consider the implementation of a Gaussian variational family $q_\lambda = \mathcal{N}(\lambda_1, \lambda_2)$ to model the latent variable z. This model is employed to estimate $p_0(x, z)$ using a naive estimator approach.

The estimator is defined as:

$$p_0(x, z) \approx \frac{1}{k} \sum_{i=1}^{k} \frac{p_0(x, z^{(i)})}{q_\lambda(z^{(i)})} \tag{1}$$

where $z^{(i)}$, $i = 1, \ldots, k$, are sampled from q_λ. The effectiveness of this estimator is heavily reliant on its variance remaining small.

The variance of the estimator can be expressed as:

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^{k} \frac{p_0(x, z^{(i)})}{q_\lambda(z^{(i)})}\right] &= \frac{1}{k}\mathrm{Var}\left[\frac{p_0(x, z)}{q_\lambda(z)}\right] \\
&= \frac{1}{k}\left(\mathbb{E}_{q_\lambda}\left[\left(\frac{p_0(x, z)}{q_\lambda(z)}\right)^2\right] - \left(\mathbb{E}_{q_\lambda}\left[\frac{p_0(x, z)}{q_\lambda(z)}\right]\right)^2\right)
\end{aligned}$$

Given the model specifications $p_0(x,z) = p(z)p_0(x|z)$, $p(z) = \mathcal{N}(0,1)$, and $p_0(x|z) = \mathcal{N}(z+\theta,1)$, and using the variational distribution $q_\lambda(z)$, we substitute and further elaborate:

$$\begin{aligned}\operatorname{Var}\left[\frac{p_0(x,z)}{q_\lambda(z)}\right] = \mathbb{E}_{q_\lambda}\left[\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(z^2+(x-z-\theta)^2)}{2}}\right)^2 \cdot \left(\frac{1}{\sqrt{2\pi\lambda_2}}e^{\frac{(z-\lambda_1)^2}{2\lambda_2}}\right)^{-2}\right] \\ - \left(\mathbb{E}_{q_\lambda}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{(z^2+(x-z-\theta)^2)}{2}} \cdot \frac{1}{\sqrt{2\pi\lambda_2}}e^{-\frac{(z-\lambda_1)^2}{2\lambda_2}}\right]\right)^2\end{aligned}$$

To determine the largest λ^* such that the variance is infinite whenever $\lambda_2 < \lambda^*$, we analyze the behavior of the variance expression. Focusing on the first term:

$$\frac{1}{2\pi\lambda_2}e^{\frac{\lambda_1^2}{\lambda_2}} \cdot e^{\frac{(x-\theta-\lambda_1)^2}{\lambda_2}}$$

As $\lambda_2 \to 0$, the exponential term $e^{\frac{\lambda_1^2}{\lambda_2}}$ increases without bound for any non-zero λ_1. Consequently, the entire first term tends towards infinity as $\lambda_2 \to 0$.

The second term involves an expectation over the Gaussian distribution $q_\lambda(z)$. As λ_2 reduces, the distribution becomes sharply peaked around λ_1, and the expectation becomes increasingly significant but finite. Therefore, for sufficiently small λ_2, the first term dominates, making the variance infinite.

Consequently, we find that $\lambda^* = \infty$, indicating that the variance of the estimator is infinite whenever $\lambda_2 < \infty$. In conclusion, $\lambda^* = \infty$ is the threshold below which the variance of the estimator in equation (1) becomes unbounded.

Part 3.

We examine the conditions under which the variance of the Evidence Lower Bound (ELBO) remains finite when using a Gaussian variational family $q_\lambda = \mathcal{N}(\lambda_1, \lambda_2)$ to approximate the posterior in variational inference. The focus is on identifying the range of λ_2 that guarantees this property.

The ELBO is given by:

$$\operatorname{ELBO}(x,\theta,\lambda) = \mathbb{E}_{q_\lambda(z)}\left[\log\frac{p_\theta(x,z)}{q_\lambda(z)}\right]$$

which can be approximated using a Monte Carlo estimator:

$$\operatorname{ELBO}(x,\theta,\lambda) \approx \frac{1}{k}\sum_{i=1}^{k}\log\frac{p_\theta(x,z^{(i)})}{q_\lambda(z^{(i)})}$$

where $z^{(i)} \sim q_\lambda(z)$.

The variance of the ELBO estimator is crucial for ensuring the stability and accuracy of variational inference:

$$\operatorname{Var}(\operatorname{ELBO}) = \frac{1}{k}\operatorname{Var}\left(\log\frac{p_\theta(x,z)}{q_\lambda(z)}\right)$$

Focusing on a single term, we need to understand the variance of:

$$\log \frac{p_\theta(x,z)}{q_\lambda(z)} = \log p_\theta(x,z) - \log q_\lambda(z)$$

Expanding $\log q_\lambda(z)$ considering $q_\lambda(z) = \mathcal{N}(\lambda_1, \lambda_2)$, we have:

$$\log q_\lambda(z) = -\frac{1}{2}\log(2\pi\lambda_2) - \frac{(z-\lambda_1)^2}{2\lambda_2}$$

Analyzing the variance of $\log q_\lambda(z)$:

$$\text{Var}\left(\log q_\lambda(z)\right) = \text{Var}\left(\frac{(z-\lambda_1)^2}{2\lambda_2}\right)$$

Given $z - \lambda_1 \sim \mathcal{N}(0, \lambda_2)$, we calculate:

$$\mathbb{E}\left[\left(\frac{(z-\lambda_1)^2}{2\lambda_2}\right)^2\right] = \frac{1}{4\lambda_2^2}\mathbb{E}[(z-\lambda_1)^4]$$

Using the fourth moment of a normal distribution:

$$\mathbb{E}[(z-\lambda_1)^4] = 3\lambda_2^2$$

Thus:

$$\frac{1}{4\lambda_2^2} \cdot 3\lambda_2^2 = \frac{3}{4}$$

This computation shows that the variance is indeed finite.

The variance of the ELBO is finite as long as $\lambda_2 > 0$. There is no upper limit on λ_2 from this analysis, indicating that the ELBO's variance remains bounded for all positive values of λ_2.

Problem 2.

Suppose that ϕ is a smooth nonnegative function supported on $[-1, 1]$ satisfying $\int \phi = 1$, and $\phi(x) > 0$ for all $x \in [-1, 1]$. Let $p_1(x) = 0.5\phi(x+2) + 0.5\phi(x-2)$ and $p_2(x) = 0.1\phi(x+2) + 0.9\phi(x-2)$ be two density functions. An illustration can be found on p12 of s18_diffusion.pdf.

- Suppose that we use a "cold start" of the Langevin dynamics, i.e. set $X_0 = -2$ as initialization, and then follow the dynamics on p7 of s18_diffusion.pdf with $\nabla \log p \leftarrow \nabla \log p_2$ for a very large time t (i.e. achieving the stationary distribution). What is the distribution of X_t ? Why?

1 bonus point Suppose that we use a "warm start" of the Langevin dynamics, i.e. set $X_0 \sim p_1$ to be random. Then follow the Langevin dynamics with $\nabla \log p \leftarrow \nabla \log p_2$ for a very large time t (i.e. achieving the stationary distribution). What is the distribution of X_t ? Why?

Solution 2.

Part 1.

Given a stochastic differential equation (SDE) governing the Langevin dynamics with an initial condition and specific drift and diffusion terms, we analyze the behavior and the eventual stationary distribution of the process X_t.

-We consider the Langevin dynamics defined by:

$$dX_t = \nabla \log p_2(X_t)\, dt + \sqrt{2}\, dW_t$$

with the initialization $X_0 = -2$. The distribution $p_2(x)$ is given as:

$$p_2(x) = 0.1\phi(x+2) + 0.9\phi(x-2)$$

where ϕ is a smooth nonnegative function supported on $[-1, 1]$ and normalized to integrate to 1 over its support.

-The density p_2 is a mixture of two translations of ϕ, heavily biased towards $x = -2$. This suggests a higher probability density around this point due to the mixture weights.

Gradient of Logarithm of p_2 The dynamics are driven by the gradient of the logarithm of p_2, which can be computed as:

$$\nabla \log p_2(x) = \frac{\nabla p_2(x)}{p_2(x)} = \frac{0.1\phi'(x+2) - 0.9\phi'(x-2)}{0.1\phi(x+2) + 0.9\phi(x-2)}$$

This gradient term encourages movement towards the regions where $p_2(x)$ has higher values, predominantly around $x = -2$.

The Langevin dynamics aim to converge to the distribution from which the drift term $\nabla \log p_2(x)$ is derived. The stochastic component $\sqrt{2}dW_t$ introduces variability, allowing exploration but still biased towards areas of higher $p_2(x)$. Given the initialization at $X_0 = -2$, the process starts near the major mode of p_2 and is likely to remain in this region.

The stationary distribution of the Langevin dynamics, when it exists and is unique, corresponds to the distribution p_2 itself. The Fokker-Planck equation for the Langevin dynamics at stationarity becomes:

$$0 = -\nabla \cdot (\nabla \log p_2 \cdot p_2) + \Delta p_2$$

ensuring that p_2 satisfies this stationary condition due to the balance of the drift and the diffusion terms. As $t \to \infty$, X_t is expected to be distributed according to p_2, particularly concentrating around $x = -2$ due to the strong localization effects of the dynamics and the initial state. The process robustly demonstrates the principles of stochastic processes converging to a stationary distribution, governed by the specified SDE dynamics.

Part 2.

This analysis explores the behavior of a stochastic process initialized according to one distribution p_1 and evolving under dynamics driven by the gradient of the logarithm of another distribution p_2.

Consider a stochastic process X_t defined by the Langevin dynamics:

$$dX_t = \nabla \log p_2(X_t)dt + \sqrt{2}dW_t$$

with an initial condition where X_0 is distributed according to p_1. We aim to determine the stationary distribution of X_t as $t \to \infty$.

Langevin dynamics typically lead to a distribution converging to the one associated with its drift term. Here, the dynamics are governed by:

$$dX_t = \nabla \log p_2(X_t)dt + \sqrt{2}dW_t$$

implying that the process is driven to explore the state space according to p_2.

The Fokker-Planck equation associated with the given Langevin dynamics is:

$$\frac{\partial p}{\partial t} = -\nabla \cdot (p \nabla \log p_2) + \Delta p$$

At stationarity, this simplifies to:

$$0 = -\nabla \cdot (p \nabla \log p_2) + \Delta p$$

indicating that p must be p_2 for this differential equation to hold true, as this satisfies the condition of zero net flux across every point in the state space.

Though the process starts with $X_0 \sim p_1$, the dynamics specified strongly guide X_t towards areas where p_2 is maximized, due to the deterministic drift term $\nabla \log p_2(X_t)$. The random fluctuations introduced by $\sqrt{2}dW_t$ enable the process to explore the state space but do not affect the ultimate convergence to p_2.

Regardless of the initial distribution p_1, the Langevin dynamics described here will result in X_t converging to the stationary distribution p_2 as $t \to \infty$. This outcome is driven

Problem 3.

Suppose that W_t is a standard Brownian motion process p6 of s18_diffusion.pdf, and $X \sim p$ is independent of the Brownian motion, where p is a distribution on $\mathbb{R}^d$. Let $p_t \sim p * \mathcal{N}(0, 2tI)$ be the distribution of $X + \sqrt{2}W_t$.

- If $p = \mathcal{N}(0, I)$, which of the following is/are correct? Why?

1. For the stochastic differential equation (SDE) $dX_t = \sqrt{2}dW_t$ with initialization $X_0 \sim p$, we have that $X_t \sim p_t$.

2. For the differential equation $dX_t = -\nabla \log p_t(X_t)\, dt$ with initialization $X_0 \sim p$, we have that $X_t \sim p_t$.

3. For the SDE $dX_t = 9\nabla \log p_t\left(X_t\right) dt + 10\sqrt{2} dW_t$ with initialization $X_0 \sim p$, we have that $X_t \sim p_t$.
4. For the SDE $dX_t = 99\nabla \log p_t\left(X_t\right) dt + 10\sqrt{2} dW_t$ with initialization $X_0 \sim p$, we have that $X_t \sim p_t$.
5. The distribution of the whole sample paths $\left(X_t\right)_{t \geq 0}$ (not just the marginal at time t) in part 1) and 2) are the same.

- [1 bonus point for correct choice and 2 bonus point for correct derivations] If p is a general distribution, not necessarily Gaussian. Make the selection again, and please explain.

Hint: we can use a similar integration by parts method on the Langevin process slide p7 of s18_diffusion.pdf. Taylor expand to the second order terms and use $E\left[dW^2\right] = dt$ and $\mathbb{E}\left[dW_t \mid X_t\right] = 0$.

Solution 3.

Part 1.

1.

True.
The SDE given is $dX_t = \sqrt{2} dW_t$. Since $X_0 \sim \mathcal{N}(0, I)$, the solution to this SDE is $X_t = X_0 + \sqrt{2} W_t$. Given that $W_t \sim \mathcal{N}(0, tI)$, the distribution of X_t is:

$$X_t \sim \mathcal{N}(0, I) + \mathcal{N}(0, 2tI) = \mathcal{N}(0, (1 + 2t)I),$$

showing that X_t is normally distributed with mean 0 and a variance that grows linearly with time.

2.

True.
The equation $dX_t = -\nabla \log p(X_t) dt$ with $p(X_t) = \mathcal{N}(0, I)$ implies $\nabla \log p(X_t) = -X_t$. Hence, $dX_t = X_t dt$, an Ornstein-Uhlenbeck process. The stationary solution for this SDE, where $dp/dt = 0$, satisfies:

$$0 = -X_t + \text{noise term} = 0 \implies X_t \sim \mathcal{N}(0, I),$$

which maintains the initial Gaussian distribution due to the negative feedback towards the mean.

3.

False.
For the SDE $dX_t = 9\nabla \log p(X_t) dt + 10\sqrt{2} dW_t$, using $\nabla \log p(X_t) = -X_t$ (since $p = \mathcal{N}(0, I)$):

$$dX_t = -9X_t dt + 10\sqrt{2} dW_t.$$

The variance at stationarity can be derived from:

$$d(\text{Var}(X_t)) = 0 = -18\text{Var}(X_t)dt + 200dt \implies \text{Var}(X_t) = \frac{200}{18} \neq 1.$$

Thus, the variance does not match that of p, and X_t does not converge to $\mathcal{N}(0, I)$.

4.

False.
Similar to the previous analysis, the variance stabilization condition for $dX_t = 99\nabla \log p(X_t)dt + 10\sqrt{2}dW_t$ yields:

$$d(\text{Var}(X_t)) = 0 = -198\text{Var}(X_t)dt + 200dt \implies \text{Var}(X_t) = \frac{200}{198} \approx 1.01,$$

indicating the variance slightly exceeds that required for p.

5.

False.
The distributions of entire sample paths differ significantly between the scenarios. In part 1, X_t represents unbounded Brownian motion with growing variance, while in part 2, it describes an Ornstein-Uhlenbeck process that reverts to its mean and maintains a constant variance, leading to fundamentally different dynamic behaviors and path properties.

Part 2.

1. $dX_t = \sqrt{2}dW_t$ **with** $X_0 \sim p$

True.

- This SDE represents unforced Brownian motion, scaled by $\sqrt{2}$. The solution for X_t is $X_t = X_0 + \sqrt{2}W_t$. - Using the fact that W_t is standard Brownian motion:

$$X_t = X_0 + \sqrt{2}W_t$$

- Since $W_t \sim \mathcal{N}(0, t)$, and $X_0 \sim p$, X_t distributes as:

$$X_t \sim p * \mathcal{N}(0, 2t)$$

- X_t maintains a Gaussian spread centered around the initial distribution p, widening over time due to the diffusion term.

2. $dX_t = -\nabla \log p(X_t)dt$ **with** $X_0 \sim p$

False, under general conditions.

- The SDE can be rewritten using Itô's lemma, considering $f(X_t) = \log p(X_t)$:

$$df = \nabla \log p(X_t) \cdot dX_t + \frac{1}{2}\nabla^2 \log p(X_t) \cdot (dX_t)^2$$

- Plugging in the SDE dynamics:

$$dX_t = -\nabla \log p(X_t)dt \Rightarrow df = -(\nabla \log p(X_t))^2 dt + 0$$

- The lack of noise prevents X_t from exploring the state space effectively, potentially causing it to settle in local minima rather than conforming to p.

3. $dX_t = 9\nabla \log p(X_t)dt + 10\sqrt{2}dW_t$ **with** $X_0 \sim p$

False.

- Similar setup as before but adding noise:

$$df = 9(\nabla \log p(X_t))^2 dt + 10\sqrt{2}\nabla \log p(X_t)dW_t$$

- The presence of strong noise significantly affects the distribution of X_t, potentially causing it to diverge from p, especially in variance.

4. $dX_t = 99\nabla \log p(X_t)dt + 10\sqrt{2}dW_t$ **with** $X_0 \sim p$

False.

- Applying similar reasoning as in part 3, but with even stronger drift, the noise term still disrupts the distribution maintenance:

$$df = 99(\nabla \log p(X_t))^2 dt + 10\sqrt{2}\nabla \log p(X_t)dW_t$$

- Even with aggressive reversion to the mean, the high noise level overwhelms the system's ability to stabilize at p.

5. Comparison of sample paths in part 1) and 2)

False.

- Sample paths from part 1 evolve with increasing variance, exploring more of the state space, whereas in part 2, paths are driven strictly by gradient descent, potentially converging to a steady state or trapping in local structures of p.