Week 8

Problem 1: Distributions closed under convolution

1. Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$. Define the convolution as:

$$(p * q)(x) = \int_{-\infty}^{\infty} p(x-y)q(y)dy$$

   Show that $(p * q)(x) = \mathcal{N}(x; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2. Let $q(x_k \mid x_{k-1}) = \mathcal{N}\left(x_k; \alpha_k x_{k-1}, \sigma_k^2 I\right)$ for $k = 1, \ldots, n$. Starting from a fixed data point x_0, show that the conditional distribution $q(x_k \mid x_0)$ is a Gaussian of the form $\mathcal{N}(x_k; \tilde{\mu}_k, \tilde{\sigma}_k^2 I)$. Provide the explicit expressions for $\tilde{\mu}_k$ and $\tilde{\sigma}_k^2$ in terms of α_i, σ_i^2, and x_0.

3. Let $p = \text{Cauchy}(\mu_1, \gamma_1)$ and $q = \text{Cauchy}(\mu_2, \gamma_2)$, where the density is given by:

$$p(x; \mu, \gamma) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x-\mu}{\gamma}\right)^2\right)}$$

   Show that $p * q = \text{Cauchy}(\mu_1 + \mu_2, \gamma_1 + \gamma_2)$, and we can therefore use Cauchy distributions instead of Gaussians in the definition of the noising process.

   Note: for recent work on diffusion models with heavy-tailed distributions, refer to [1] and [2].

1. Naïvely, this could be done by direct computation:

$$\begin{aligned}
(p * q)(x) &= \int_{-\infty}^{\infty} p(x-y)q(y)dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-y-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} dy \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{(x-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)} dy \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\left(y - \frac{\sigma_1^2\mu_2 + \sigma_2^2(x-\mu_1)}{\sigma_1^2+\sigma_2^2}\right)^2 + \frac{(x-\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right)} dy \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\left(y - \frac{\sigma_1^2\mu_2 + \sigma_2^2(x-\mu_1)}{\sigma_1^2+\sigma_2^2}\right)^2\right)} dy \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} \sqrt{\frac{2\pi\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{(x-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} \\
&= \mathcal{N}(x; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).
\end{aligned}$$

However, a simpler proof could use that convolution is equivariant to **shift** of either distribution and to **scaling** of both distributions. This fact can be used to reduce the problem to the case where $\mu_1 = \mu_2 = 0$ (by shifting both distribution) and where $\sigma_1^2 = 1$ (by scaling both distributions by $\frac{1}{\sigma_1}$). In this case, we simply need to show that $\mathcal{N}(0,1) * \mathcal{N}(0, \sigma_2^2) = \mathcal{N}(0, 1 + \sigma_2^2)$, for which the above computation is far less cumbersome. See also discussion in the third part below.

2. We start by writing out the first few transitions:

$$
\begin{aligned}
q(x_1 \mid x_0) &= \mathcal{N}(x_1; \alpha_1 x_0, \sigma_1^2 I).\\
q(x_2 \mid x_0) &= \mathcal{N}(x_2; \alpha_2\alpha_1 x_0, (\alpha_2^2\sigma_1^2 + \sigma_2^2) I),\\
q(x_3 \mid x_0) &= \mathcal{N}(x_3; \alpha_3\alpha_2\alpha_1 x_0, (\alpha_3^2\alpha_2^2\sigma_1^2 + \alpha_3^2\sigma_2^2 + \sigma_3^2) I).
\end{aligned}
$$

We can see (and easily show by induction) that the mean and variance of $q(x_k \mid x_0)$ are given by

$$
\begin{aligned}
\tilde{\mu}_k &= \alpha_k\alpha_{k-1}\dots\alpha_1 x_0,\\
\tilde{\sigma}_k^2 &= \sigma_k^2 + \alpha_k^2\sigma_{k-1}^2 + \alpha_k^2\alpha_{k-1}^2\sigma_{k-2}^2 + \dots + \alpha_k^2\alpha_{k-1}^2\dots\alpha_2^2\sigma_1^2\\
&= \sum_{i=1}^{k} \sigma_i^2 \prod_{j=i+1}^{k} \alpha_j^2.
\end{aligned}
$$

3. Just as in the first part, we can reduce to the case where $\mu_1 = \mu_2 = 0$ and $\gamma_1 = 1$ by using the shift and scaling equivariance of convolution (note that μ and γ control the position and scale of any distribution whose distribution is expressed in terms of $\frac{x-\mu}{\gamma}$ – this is called a 'location-scale family' and Gaussians are similarly an example of such a family, as the density is expressed in terms of $\frac{x-\mu}{\sigma}$).

In this case, we need to show that $\text{Cauchy}(0,1) * \text{Cauchy}(0,\gamma_2) = \text{Cauchy}(0, 1+\gamma_2)$, which is a direct computation:

$$
\begin{aligned}
(p * q)(x) &= \int_{-\infty}^{\infty} p(x-y)q(y)dy\\
&= \int_{-\infty}^{\infty} \frac{1}{\pi\left(1+(x-y)^2\right)} \cdot \frac{1}{\pi\gamma_2\left(1+\left(\frac{y}{\gamma_2}\right)^2\right)} dy\\
&= \frac{1}{\pi^2\gamma_2} \int_{-\infty}^{\infty} \frac{1}{\left(1+(x-y)^2\right)\left(1+\left(\frac{y}{\gamma_2}\right)^2\right)} dy\\
&= \frac{\gamma_2}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{\left(1+y^2\right)\left(y-x)^2+\gamma^2\right)} dy\\
&= \frac{\gamma_2}{\pi^2} \frac{\pi(1+\gamma_2)}{\gamma_2(x^2+(1+\gamma_2)^2)}\\
&= \frac{1}{\pi(1+\gamma_2)\left(1+\left(\frac{x}{1+\gamma}\right)\right)}\\
&= \text{Cauchy}(x; 0, 1+\gamma_2).
\end{aligned}
$$

In the fifth line we have used an integral identity:

$$
\int_{-\infty}^{\infty} \frac{1}{(y^2+1)((y-x)^2+c^2)} dy = \frac{\pi(1+c)}{c(x^2+(1+c)^2)}.
$$

To prove this identity, we can use a partial fraction decomposition of the left side and do an (unpleasant) direct computation. Alternatively, one can use residue calculus: observe that the poles of the integrand are at i, $-i$, $x+ic$ and $x-ic$. The integral is $2\pi i$ times the sum of the residues in the upper half-plane, which gives a similar computation to the partial fraction method.

Yet another solution – to both this and the first part – uses the characteristic function $f_X(t) = \mathbb{E}[e^{itX}]$ (a restriction of the Fourier transform on the density). Convolution becomes multiplication in the transformed domain t, that is, if X and Y are independent and have densities p and q,

so $X + Y$ has density $p * q$, then $f_{X+Y}(t) = f_X(t) f_Y(t)$. It can be shown that the characteristic function of a $\mathcal{N}(\mu, \sigma)^2$ random variable is $e^{it\mu - \frac{1}{2}\sigma^2 t^2}$ and that of a Cauchy(μ, γ) is $e^{it\mu - \gamma|t|}$, showing clearly that multiplication of charactiristic function is the same as addition of parameters μ and γ (for Cauchy) or μ and σ^2 (for Gaussian).

Problem 2: Closed form for ELBO

As shown in the lecture, the ELBO for hierarchical VAE can be optimised using the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0 \sim p(z_0), z_{1,\dots,N} \sim q(z_{1,\dots,N} | z_0)} \left[\log p(z_N) \prod_{n=1}^{N} p_\theta(z_{n-1} \mid z_n) \right]$$

1. Assuming $p_\theta(z_{n-1} \mid z_n) = \mathcal{N}(z_n; f_\theta(z_n, n), \sigma_{n-1}^2 I)$ and $p(z_N)$ does not depend on θ, show that maximising $\mathcal{L}(\theta)$ is equivalent to minimising the following objective:

$$\widetilde{\mathcal{L}}(\theta) = \mathbb{E}_{\substack{n \sim \text{Unif}(\{1,\dots,N\}), \\ z_0 \sim p(z_0), \\ z_n, z_{n-1} \sim q(z_n, z_{n-1} \mid z_0)}} \left[\frac{1}{\sigma_{n-1}^2} \| z_{n-1} - f_\theta(z_n,\ n) \|^2 \right]$$

2. How can we sample from a model trained with the objective $\widetilde{\mathcal{L}}(\theta)$? Write the response as a probabilistic program.

3. What are the possible disadvantages of using the objective $\widetilde{\mathcal{L}}(\theta)$?

1. We rewrite the ELBO using the Gaussian form of the transitions $\pi_\theta(z_{n-1} \mid z_n)$:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\theta) &= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\dots,N} \sim q(z_{1,\dots,N} | z_0)} \left[\log p(z_N) + \sum_{n=1}^{N} \log p_\theta(z_{n-1} \mid z_n) \right] \\ &= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\dots,N} \sim q(z_{1,\dots,N} | z_0)} \left[\log p(z_N) - \sum_{n=1}^{N} \frac{1}{2\sigma_{n-1}^2} \| z_{n-1} - f_\theta(z_n, n) \|^2 \right] + \text{const.} \end{aligned}$$

   Since $p(z_N)$ does not depend on θ, it does not affect the optimisation problem. Since n is uniformly distributed over $\{1, \dots, N\}$, we can rewrite the sum as an expectation over n and get

$$\mathcal{L}_{\text{ELBO}}(\theta) = -\mathbb{E}_{\substack{n \sim \text{Unif}(\{1,\dots,N\}), \\ z_0 \sim p(z_0), \\ z_{1,\dots,N} \sim q(\cdot \mid z_0)}} \left[\frac{1}{2\sigma_{n-1}^2} \| z_{n-1} - f_\theta(z_n, n) \|^2 \right] + \text{const.}$$

   It remains to see that the expression inside the expectation depends only on z_{n-1} and z_n.

   This shows that maximising $\mathcal{L}_{\text{ELBO}}(\theta)$ is equivalent to minimising $\widetilde{\mathcal{L}}(\theta)$.

2. To sample from the model, we sample z_N from $p(z_N)$ and then iteratively sample z_{n-1} from $p_\theta(z_{n-1} \mid z_n)$ for $n = N, \dots, 1$.

3. As shown in Problems 1 and 2 of the next homework, and in the lecture, the optimal predicted mean of the Gaussian $p_\theta(z_{n-1} \mid z_n)$ can be expressed in terms of $\mathbb{E}[z_0 \mid z_n]$, which is a consequence of the Gaussianity assumption on the noising process. The regression target for the one-step denoising mean has additional noise arising from the variance of $q(z_{n-1} \mid z_0, z_n)$, which results in a higher-variance optimisation objective.

   This is in addition to the advantage of predicting the score or noise relative to the (one-step or full) denoising mean, discussed in Problem 2 or the next homework.

Problem 3: Optimal denoiser

Consider the following training objective:

$$\mathcal{L}(f) = \mathbb{E}_{x\sim p(x), z\sim q(z|x)}\left[\|x - f(z)\|^2\right]$$

1. Show that the function f^* that minimizes $\mathcal{L}(f)$ is given by $f^*(z) = \mathbb{E}[x \mid z]$.
2. Assume that the dataset consists of two points, x_1 and x_2, and the empirical distribution over the dataset is given by $p(x) = \frac{1}{2}\delta_{x_1}(x) + \frac{1}{2}\delta_{x_2}(x)$. Let the noise model be Gaussian: $q(z \mid x) = \mathcal{N}(z; x, \sigma^2 I)$. Find a closed-form expression for the optimal denoiser $f^*(z)$.

1. (We assume all distributions have densities and full support. A rigourous proof in the general case requires some measure theory and calculus of variations.)

   Let $f^*(z) = \mathbb{E}[x \mid z]$. Then, for any f,

$$\begin{aligned}\mathcal{L}(f) &= \mathbb{E}_{x\sim p(x), z\sim q(z|x)}\left[\|x - f(z)\|^2\right] \\ &= \mathbb{E}_{x\sim p(x), z\sim q(z|x)}\left[\|x - f^*(z) + f^*(z) - f(z)\|^2\right] \\ &= \mathbb{E}_{x\sim p(x), z\sim q(z|x)}\left[\|x - f^*(z)\|^2 + \|f^*(z) - f(z)\|^2 + 2(x - f^*(z)) \cdot (f^*(z) - f(z))\right].\end{aligned}$$

   Considering just the last term, and factorising the joint density $p(x)q(z \mid x)$ in the other order as $q(z)q(x \mid z)$,

$$\begin{aligned}\mathbb{E}_{x\sim p(x), z\sim q(z|x)}\left[2(x - f^*(z)) \cdot (f^*(z) - f(z))\right] &= \mathbb{E}_{z\sim p(z)}\left[\underbrace{\mathbb{E}_{x\sim q(x|z)}\left[2(x - f^*(z))\right]}_{=0} \cdot (f^*(z) - f(z))\right] \\ &= 0.\end{aligned}$$

   Thus we have

$$\mathcal{L}(f) = \mathcal{L}(f^*) + \mathbb{E}([f^* - f]).$$

   This quantity is nonnegative and equals 0 if and only if $f^* = f$ for almost every z under the marginal distribution of z.

2. By the previous part, we need to compute $\mathbb{E}[x \mid z]$:

$$\begin{aligned}\mathbb{E}[x \mid z] &= \mathbb{P}[x = x_1 \mid z]x_1 + \mathbb{P}[x = x_2 \mid z]x_2 \\ &= \frac{\mathcal{N}z; x_1, \sigma^2 I)}{\mathcal{N}z; x_1, \sigma^2 I) + \mathcal{N}z; x_2, \sigma^2 I)}x_1 + \frac{\mathcal{N}z; x_2, \sigma^2 I)}{\mathcal{N}z; x_1, \sigma^2 I) + \mathcal{N}z; x_2, \sigma^2 I)}x_2 \\ &= \frac{e^{-\frac{1}{2\sigma^2}\|z-x_1\|^2}}{e^{-\frac{1}{2\sigma^2}\|z-x_1\|^2} + e^{-\frac{1}{2\sigma^2}\|z-x_2\|^2}}x_1 + \frac{e^{-\frac{1}{2\sigma^2}\|z-x_2\|^2}}{e^{-\frac{1}{2\sigma^2}\|z-x_1\|^2} + e^{-\frac{1}{2\sigma^2}\|z-x_2\|^2}}x_2 \\ &= \text{softmax}\left(-\frac{1}{2\sigma^2}\|z - x_1\|^2, -\frac{1}{2\sigma^2}\|z - x_2\|^2\right) \cdot (x_1, x_2),\end{aligned}$$

   where the second line used that $\mathbb{P}[x = x_i \mid z] \propto \mathbb{P}[x = x_i]q(z \mid x_i)$.

Week 9

Assumptions for all problems

1. $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I)$, or $z_n = z_{n-1} + \sigma_n \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$.
2. $z_0, \ldots, z_{n-1}$ are conditionally independent of $z_{n+1}, \ldots, z_N$ given z_n.
3. $V_n = \sigma_1^2 + \ldots + \sigma_n^2$, so, by the previous two assumptions, $q(z_n \mid z_0) = \mathcal{N}(z_n; z_0, V_n I)$, or $z_n = z_0 + \sqrt{V_n}\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$.
4. Unless stated otherwise, $\mathbb{E}$ is taken over $p(z_0)$, $q(\cdot \mid z_0)$, and $n \sim \text{Unif}(\{1, \ldots, N\})$.

Problem 1: Closed form for ELBO continued

In Lecture 8 we showed that the diffusion model ELBO can be written as

$$\mathcal{L}_{\text{ELBO}}(\theta) = \mathbb{E}_{z_0 \sim p(z_0), z_{1,\ldots,N} \sim q(\cdot \mid z_0)} \left[\log p(z_N) + \sum_{n=1}^{N} \log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_n \mid z_{n-1})} \right]$$

Show the following Monte-Carlo estimator of $\mathcal{L}_{\text{ELBO}}(\theta)$:

$$\mathcal{L}_{\text{ELBO}}(\theta) = \underbrace{-\mathbb{E}\left[\frac{1}{2\gamma_{n-1}^2} \left\| \mu_n(z_n, z_0) - \mu_n(z_n; \theta) \right\|^2 \right]}_{L_n} + \underbrace{\mathbb{E}\left[\log p_\theta(z_0 \mid z_1) \right]}_{L_0} + \text{const.}$$

where $\mu_n(z_n, z_0)$ is given by a linear combination of z_n and z_0, $\mu_n(z_n; \theta)$ is the predicted mean of the Gaussian defining $p_\theta(z_{n-1} \mid x_n)$, and the γ_{n-1}^2 are some constants. Use the following steps:

1. Using the fact that $q(z_n \mid z_{n-1}) = \frac{q(z_{n-1} \mid z_n, z_0) q(z_n \mid z_0)}{q(z_{n-1} \mid z_0)}$ for $2 \leq n \leq N$, show that $q(z_{n-1} \mid z_n, z_0) = \mathcal{N}(z_{n-1}; \mu_n(z_n, z_0), \gamma_{n-1}^2 I)$. Provide the expressions for $\mu(z_n, z_0)$ and γ_{n-1}^2.
2. Compute $\log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_{n-1} \mid z_n, z_0)}$, assuming that $p_\theta(z_{n-1} \mid z_n) = \mathcal{N}(z_{n-1}; \mu_n(z_n; \theta), \gamma_{n-1}^2 I)$;

How is this version of the ELBO related to the one derived in Problem 2 of Week 8?

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}}(\theta) &= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\ldots,N} \sim q(\cdot \mid z_0)} \left[\log p(z_N) + \sum_{n=1}^{N} \log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_n \mid z_{n-1})} \right] \\
&= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\ldots,N} \sim q(\cdot \mid z_0)} \left[\log p(z_N) + \sum_{n=2}^{N} \log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_{n-1} \mid z_n, z_0)} \cdot \frac{q(z_{n-1} \mid z_0)}{q(z_n \mid z_0)} + \frac{p_\theta(z_0 \mid z_1)}{q(z_1 \mid z_0)} \right] \\
&= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\ldots,N} \sim q(\cdot \mid z_0)} \left[\log \frac{p(z_N)}{q(z_N \mid z_0)} + \sum_{n=2}^{N} \log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_{n-1} \mid z_n, z_0)} + \log p_\theta(z_0 \mid z_1) \right] \\
&= \mathbb{E}_{z_0 \sim p(z_0), z_{1,\ldots,N} \sim q(\cdot \mid z_0)} \left[\sum_{n=2}^{N} \log \frac{p_\theta(z_{n-1} \mid z_n)}{q(z_{n-1} \mid z_n, z_0)} \right] + \mathbb{E}_{z_0 \sim p(z_0), z_1 \sim q(z_1 \mid z_0)} \left[\log p_\theta(z_0 \mid z_1) \right] + \text{const.} \\
&= -\mathbb{E}\left[\sum_{n=2}^{N} \text{KL}\left(q(z_{n-1} \mid z_n, z_0) \| p_\theta(z_{n-1} \mid z_n) \right) \right] + \mathbb{E}_{z_0 \sim p(z_0), z_1 \sim q(z_1 \mid z_0)} \left[\log p_\theta(z_0 \mid z_1) \right] + \text{const.}
\end{aligned}$$

Using the expressions for $q(z_n \mid z_0)$ and $q(z_n \mid z_{n-1})$ and the result shown in lecture about

conditioning the sum of two Gaussians, we get

$$q(z_{n-1} \mid z_n,\, z_0) = \mathcal{N}\left(z_{n-1}; \underbrace{\frac{V_{n-1}z_n + \sigma_n^2 z_0}{V_n}}_{\mu(z_n, z_0)},\, \underbrace{\frac{\sigma_n^2 V_{n-1}}{V_n} I}_{\gamma_{n-1}^2 I}\right).$$

Finally, using the fact that KL between two Gaussians can be expressed (up to an additive constant) as a squared norm of the difference of their means divided by the variance of the first one, we obtain the final expression.

Problem 2: Equivalent losses

Using the fact that $\mu_n(z_n, z_0)$ is a linear combination of z_n and z_0, show that L_n can be equivalently rewritten as

1. **(denoiser)** $\mathbb{E}\left[\omega(n)\|z_0 - f_n(z_n;\theta)\|^2\right]$ using that $\mu_n(z_n, z_0)$ is a linear combination of z_n and z_0;
2. **(noise predictor)** $\mathbb{E}\left[\omega(n)\|\varepsilon - f_n(z_n;\theta)\|^2\right]$ using $z_n = z_0 + \sqrt{V_n}\varepsilon$; $\varepsilon \sim \mathcal{N}(0, I)$;
3. **(score predictor)** $\mathbb{E}\left[\omega(n)\|s_n - f_n(z_n;\theta)\|^2\right]$ using, given fixed z_0, $s_n = \frac{1}{V_n}(z_0 - z_n)$;

where in each case $f_n(z_n;\theta)$ is a neural network. In each case, express $\mu_n(z_n;\theta)$ via $f_n(z_n;\theta)$ and compute the weight $\omega(n)$ that gives the equivalent loss.
Note: for the continuous versions of the provided losses, the weight ω can be given a particular meaning. To learn more about it, see the following papers: [1], [2].

From the previous part, we can express the optimal one-step denoising mean $\mu_n^*(z_n)$ in terms of z_n and $\mathbb{E}[z_0 \mid z_n]$ as

$$\frac{V_{n-1}}{V_n} z_n + \frac{\sigma_n^2}{V_n}\mathbb{E}[z_0 \mid z_n],$$

so

$$\mathbb{E}[z_0 \mid z_n] = \frac{V_n}{\sigma_n^2}\mu_n^*(z_n) - \frac{V_{n-1}}{\sigma_n^2} z_n.$$

Using this expression, we can express $\mu_n^*(z_n)$ in terms of the optimal $f_n^*(z_n)$ for each of the three cases:

1. Denoiser:
$$f_n^*(z_n) = \mathbb{E}[z_0 \mid z_n] = \frac{V_n}{\sigma_n^2}\mu_n^*(z_n) - \text{const}(z_n).$$
2. Noise predictor:
$$f_n^*(z_n) = \frac{1}{\sqrt{V_n}}(z_n - \mathbb{E}[z_0 \mid z_n)) = -\frac{\sqrt{V_n}}{\sigma_n^2}\mu_n^*(z_n) + \text{const}(z_n).$$
3. Score predictor:
$$f_n^*(z_n) = \frac{1}{V_n}(\mathbb{E}[z_0 \mid z_n] - z_n) = \frac{1}{\sigma_n^2}\mu_n^*(z_n) + \text{const}(z_n).$$

Each case has the form $f_n^*(z_n) = c(n)\mu_n^*(z_n) + \text{const}(z_n)$. The objective is a least-squares regression with respect to a linear transformation of the regression variable; regressing on μ_n with coefficient $\frac{1}{2\gamma_{n-1}^2} = \frac{V_n}{2\sigma_n^2 V_{n-1}}$ is equivalent to regressing on f_n with coefficient $\omega(n) = \frac{V_n}{2\sigma_n^2 V_{n-1} c(n)^2}$. In our three cases:

1. Denoiser: $c(n) = \frac{V_n}{\sigma_n^2}$, $\omega(n) = \frac{\sigma_n^2}{2V_{n-1}V_n}$.
2. Noise predictor: $c(n) = \frac{-\sqrt{V_n}}{\sigma_n^2}$, $\omega(n) = \frac{\sigma_n^2}{2V_{n-1}}$.
3. Score predictor: $c(n) = \frac{1}{\sigma_n^2}$, $\omega(n) = \frac{V_n\sigma_n^2}{2V_{n-1}}$.

Problem 3: Discrete sampling using score / noise / denoiser predictions

For each of the equivalent definitions of L_n from the previous problem and corresponding trained models f_θ, write the probabilistic program for sampling z_0.

Procedures for sampling z_{n-1} from z_n:

1. Denoiser: $z_{n-1} = f_n^*(z_n) + \sqrt{V_{n-1}}\varepsilon,\ \varepsilon \sim \mathcal{N}(0,1)$
2. Noise predictor: $z_{n-1} = \left(z_n - \sqrt{V_n} f_n^*(z_n)\right) + \sqrt{V_{n-1}}\varepsilon,\ \varepsilon \sim \mathcal{N}(0,1)$
3. Score predictor: $z_{n-1} = z_n + V_n f_n^*(z_n) + \sqrt{V_{n-1}}\varepsilon,\ \varepsilon \sim \mathcal{N}(0,1)$

Week 10

Problem 1: From VP SDE to VE SDE through a variable change

Consider the following two SDE types, frequently used in the context of diffusion models:

$$\text{VP SDE}: \; dz_t = -\alpha_t z_t \, dt + \gamma_t \, d\omega_t, \qquad \text{VE SDE}: \; dy_t = \sigma_t \, d\omega_t.$$

In this problem, we will show that one can obtain VE SDE from VP SDE through a change of variables. To do this, we define a new process $y_t = z_t e^{\int_0^t \alpha_s \, ds}$ and compute dy_t.

1. The Euler-Maruyama approximation to $z_{t+\Delta t}$ is

   $$z_{t+\Delta t} = z_t - \alpha_t z_t \Delta t + \gamma_t \sqrt{\Delta t}\varepsilon + o(\Delta t), \quad \varepsilon \sim \mathcal{N}(0, I).$$

   Use the definition of y_t to show that

   $$y_{t+\Delta t} = (y_t - \alpha_t y_t \Delta t) e^{\int_t^{t+\Delta t} \alpha_s \, ds} + \gamma_t e^{\int_0^{t+\Delta t} \alpha_s \, ds} \sqrt{\Delta t}\varepsilon + o(\Delta t).$$

2. Use that $e^{\int_t^{t+\Delta t} \alpha_s \, ds} = 1 + \alpha_t \Delta t + O(\Delta t^2)$ to show that

   $$y_{t+\Delta t} = y_t + \gamma_t e^{\int_0^{t+\Delta t} \alpha_s \, ds} \sqrt{\Delta t}\varepsilon + o(\Delta t).$$

3. The above equation is, up to $o(\Delta t)$, an Euler-Maruyama integration scheme for a VE SDE in y_t. What is this SDE's diffusion coefficient σ_t in terms of α_t and γ_t?

1. Substituting the expression of z_t in terms of y_t on both sides of the Euler-Maruyama approximation, we get

   $$y_{t+\Delta t} e^{-\int_0^{t+\Delta t} \alpha_s \, ds} = (1 - \alpha_t \Delta t) y_t e^{-\int_0^t \alpha_s \, ds} + \gamma_t \sqrt{\Delta t}\varepsilon + o(\Delta t).$$

   Multiplying through by $e^{\int_0^{t+\Delta t} \alpha_s \, ds}$ and using that $\int_0^{t+\Delta t} \alpha_s \, ds = \int_0^t \alpha_s \, ds + \int_t^{t+\Delta t} \alpha_s \, ds$, we get the desired equation.

2. Using the given expansion of the exponential, we get that the first term on the right side of the equation in the previous part is

   $$\begin{aligned}(y_t - \alpha_t y_t \Delta t) e^{\int_t^{t+\Delta t} \alpha_s \, ds} &= (y_t - \alpha_t y_t \Delta t)(1 + \alpha_t \Delta t + O(\Delta t^2)) \\ &= y_t (1 - \alpha_t \Delta t)(1 + \alpha_t \Delta t + O(\Delta t^2)) \\ &= y_t (1 + O(\Delta t^2)) \qquad\qquad = y_t + o(\Delta t).\end{aligned}$$

   Substituting this back into the equation, we get the desired result.

3. The equation in the previous part is the Euler-Maruyama integration scheme for a VE SDE with $\sigma_t = \gamma_t e^{\int_0^t \alpha_s \, ds}$. Notice that the integral in the exponent goes to t and not $t + \Delta t$ because the difference between the two is $O(\Delta t)$, which is absorbed in the $o(\Delta t)$ term when multiplied by $\sqrt{\Delta t}$.

Problem 2: From Hierarchical VAE to SDE

Consider a noising scheme in which $z_n \sim \mathcal{N}(z_0, V_n)$ and each z_{n+1} is obtained from z_n by independent Gaussian increments. In this problem, we will show that this noising scheme defines the same distribution as the following SDE:

$$dz_t = \sqrt{\frac{\partial}{\partial t} V(t)}\, d\omega_t,$$

where $V : [0, N] \to \mathbb{R}$ is a monotonic differentiable function such that $V(n) = V_n$ for integers n.

1. Start by finding the closed form for the transition from z_n to z_{n+1}.
2. Assume that we integrate the SDE from time n to $n+1$ using the Euler-Maruyama scheme with step $\Delta t = \frac{1}{K}$. Show that the marginal distribution of z_{n+1} given z_n under this scheme is

$$z_{n+1} \sim \mathcal{N}\left(z_n, \frac{1}{K}\left(\frac{\partial V}{\partial t}(n) + \frac{\partial V}{\partial t}\left(n + \frac{1}{K}\right) + \cdots + \frac{\partial V}{\partial t}\left(n + \frac{K-1}{K}\right)\right)\right).$$

3. Take the limit $K \to \infty$ and show the variance in the Part 2 approaches $V(n+1) - V(n)$, recovering the expression in Part 1.

1. The transition from z_n to z_{n+1} is given by $z_{n+1} = z_n + \mathcal{N}(0, V_{n+1} - V_n)$, so $z_{n+1} \sim \mathcal{N}(z_n, V_{n+1} - V_n)$.

2. The Euler-Maruyama scheme for integrating the SDE from time n to $n+1$ with step $\Delta t = \frac{1}{K}$ is given by

$$z_{t+1/K} = z_t + \sqrt{\frac{\partial V}{\partial t}(t)}\sqrt{\Delta t}\varepsilon_t, \quad \varepsilon \sim \mathcal{N}(0, I).$$

   Iterating this scheme gives

$$\begin{aligned} z_{n+k/K} &= z_n + \sum_{i=0}^{k-1} \sqrt{\frac{\partial V}{\partial t}\left(n + \frac{i}{K}\right)}\sqrt{\Delta t}\varepsilon_{n+i/K}, \qquad k = 1, \ldots, K, \varepsilon_{n+\frac{i}{K}} \sim \mathcal{N}(0, I) \\ &= z_n + \mathcal{N}\left(0, \frac{1}{K}\sum_{i=0}^{k-1} \frac{\partial V}{\partial t}\left(n + \frac{i}{K}\right)\right). \end{aligned}$$

   In particular, for $k = K$, we get the expression in the problem statement.

3. The variance in the previous problem is the left-endpoint Riemann sum for the integral $\int_n^{n+1} \frac{\partial V}{\partial t}(t)dt$ for a K-part uniform partition of the interval $[n, n+1]$. As $K \to \infty$, this Riemann sum approaches the integral, which is equal to $V(n+1) - V(n) = V_{n+1} - V_n$ by the fundamental theorem of calculus.

Problem 3: From SDE to ODE flow

Consider the following forward SDE: $dz_t = \sigma_t\, d\omega_t$. Denote its marginal densities at time t by p_t.

1. Use the identities from the lecture to write the corresponding reverse SDE and probability flow ODE.
2. Show that the Fokker-Planck-Kolmogorov (FPK) equation for the forward SDE, the FPK equation for the reverse SDE, and the continuity equation for the probability flow ODE all coincide.

1. The reverse SDE is given by

$$dz_t = \left(-\sigma_t^2 \nabla \log p_t(z_t)\right)\, dt + \sigma_t\, d\bar{\omega}_t$$

and the probability flow ODE is given by

$$dz_t = -\frac{1}{2}\sigma_t^2 \nabla \log p_t(z_t)\, dt.$$

2. The FPK equation for the forward SDE is given by

$$\frac{\partial}{\partial t} p_t(z) = \frac{1}{2}\sigma_t^2 \Delta p_t(z)$$

and for the reverse SDE by

$$\begin{aligned}
\frac{\partial}{\partial t} p_t(z) &= -\nabla \cdot \left(-\sigma_t^2 \nabla \log p_t(z) p_t(z)\right) - \frac{1}{2}\sigma_t^2 \Delta p_t(z) \\
&= -\nabla \cdot \left(-\sigma_t^2 \nabla p_t(z)\right) - \frac{1}{2}\sigma_t^2 \Delta p_t(z) \\
&= \sigma_t^2 \Delta p_t(z) - \frac{1}{2}\sigma_t^2 \Delta p_t(z) \\
&= \frac{1}{2}\sigma_t^2 \Delta p_t(z).
\end{aligned}$$

The continuity equation for the probability flow ODE is

$$\begin{aligned}
\frac{\partial}{\partial t} p_t(z) &= -\nabla \cdot \left(-\frac{1}{2}\sigma_t^2 \nabla \log p_t(z) p_t(z)\right) \\
&= -\nabla \cdot \left(-\frac{1}{2}\sigma_t^2 \nabla p_t(z)\right) \\
&= \frac{1}{2}\sigma_t^2 \Delta p_t(z).
\end{aligned}$$

Problem 4: From ODE flow to Normalising Flows

In exercise 2 of Week 4 we showed that for a transition of the form $z_{t+\Delta t} = z_t + u_t(z_t)\Delta t$ – the form of an Euler integration step for an ODE – the change of density can be written as

$$\log p_{t+\Delta t}(z_{t+\Delta t}) - \log p_t(z_t) = -\mathrm{Tr}\left[\mathrm{J}_{u_t}\right] \Delta t + O(\Delta t^2),$$

where J_{u_t} is the Jacobian of u_t at z_t.
In this problem, we use this result to derive the continuity equation.

1. First, check using Taylor expansions that

$$\log p_{t+\Delta t}(z_{t+\Delta t}) = \log p_t(z_t) + \frac{\partial}{\partial t} \log p_t(z_t)\Delta t + \nabla \log p_t(z_t) \cdot (z_{t+\Delta t} - z_t) + O(\Delta t^2).$$

2. Using Part 1 and the transition formula, convert the change of density formula to

$$\frac{\partial}{\partial t} \log p_t(z_t) = -\nabla \cdot u_t - \nabla \log p_t(z_t) \cdot u_t(z_t).$$

   You will need to use that $\mathrm{Tr}\left[\mathrm{J}_{u_t}\right] = \nabla \cdot u_t$.

3. Multiply both sides by p_t and derive the usual form of the continuity equation.
   (You will need the product rule for divergences: for a scalar function f and vector field v, $\nabla \cdot (fv) = f(\nabla \cdot v) + (\nabla f) \cdot v$.)

1. The first-order Taylor expansion of $\log p_t(z_t)$ at (t, z_t) is given by

$$\log p_{t+\Delta t}(z_t + \Delta z_t) = \log p_t(z_t) + \frac{\partial}{\partial t} \log p_t(z_t)\Delta t + \nabla \log p_t(z_t) \cdot \Delta z_t + O(\Delta t^2, \|\Delta z_t\|^2).$$

Writing $z_{t+\Delta t} = z_t + \Delta z_t$, so $\Delta z_t = u_t(z_t)\Delta t$, we have $\|\Delta z_t\|^2 = O(\Delta t^2)$, so the error term is $O(\Delta t^2)$, recovering the desired expression.

2. Substituting the previous part into the change of density formula and rearranging gives

$$\frac{\partial}{\partial t} \log p_t(z_t)\Delta t = -\mathrm{Tr}\left[\mathrm{J}_{u_t}\right] \Delta t - \nabla \log p_t(z_t) \cdot (z_{t+\Delta t} - z_t) + O(\Delta t^2).$$

Substituting $(z_{t+\Delta t} - z_t) = u_t(z_t)\Delta t$ and $\mathrm{Tr}\left[J_{u_t}\right] = \nabla \cdot u_t$, then dividing by Δt and taking $\Delta t \to 0$ gives the result.

3. Multiplying both sides by p_t and rearranging gives

$$\begin{aligned} p_t \frac{\partial}{\partial t} \log p_t(z_t) &= -p_t(\nabla \cdot u_t - \nabla \log p_t(z_t) \cdot u_t(z_t)) \\ \frac{\partial}{\partial t} p_t(z_t) &= -p_t \nabla \cdot u_t - \nabla p_t(z_t) \cdot u_t(z_t) \\ \frac{\partial}{\partial t} p_t(z_t) &= -\nabla \cdot (p_t u_t). \end{aligned}$$