HW3

Faezeh Pouya Mehr

May 9, 2023

1 Question1

1.1

adding and subtracting evidence lower bound to the data log-likelihood, we have:

$$\ln p(x) = \ln p(x) + \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz - \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz$$

$$= \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz - \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz + \int q(z|x) \ln p(x)\, dz$$

$$= \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz - \int q(z|x)[\ln \frac{p(x,z)}{q(z|x)} - \ln p(x)]\, dz$$

$$= \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz - \int q(z|x)[\ln \frac{p(x,z)}{q(z|x) \ln p(x)}]\, dz$$

$$= \int q(z|x) \ln \frac{p(x,z)}{q(z|x)}\, dz - \int q(z|x) \ln \frac{p(z|x)}{q(z|x)}\, dz$$

$$\ln p(x) = \mathcal{L}(\theta, \phi; x) + KL(q(z|x)||p(z|x))$$

1.2

Variational inference converts the posterior inference problem into the optimization problem of finding an approximate probability distribution $q(z|x)$ that is as close as possible to $p(z|x)$. This can be formalized as solving the following optimization problem:

$$min_\phi KL(q_\phi(z|x)||p(z|x))$$

where ϕ parameterizes the approximation q and $KL(q|p)$denotes the Kullback-Leibler divergence between q and p and is given by:

$$KL(q|p) = \int_x q(x) log \frac{q(x)}{p(x)}$$

However, this optimization problem is no easier than our original problem ($max_\theta p_\theta(x)$)because it still requires us to evaluate $p(z|x)$. Plugging in the definition of KL, we can write:

$$KL(q_\phi(z|x)||p(z|x)) = \int_x q_\phi(z|x) log \frac{q_\phi(z|x)}{p(z|x)}$$

$$= \int_x q_\phi(z|x) log \frac{q_\phi(z|x)p(x)}{p(x,z)}$$

$$= \int_x q_\phi(z|x) log \frac{q_\phi(z|x)}{p(x,z)} + \int_x q_\phi(z|x) log p(x)$$

$$= -\mathcal{L}(\phi) + log p(x)$$

where $\mathcal{L}(\phi)$ is the ELBO. As $KL(q_\phi(z|x)||p(z|x)) \geq 0$ an important property of non-negativity of KL divergence: $\mathcal{L}(\phi) \leq \log p(x)$.

which means that minimizing $KL(q_\phi(z|x)||p(z|x))$ is equivalent to maximizing $\mathcal{L}(\phi)$.

1.3

$$\mathcal{L}(\theta, q_\phi^*; x_i) = \log(p_\theta(x_i)) - D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i)$$

$$\mathcal{L}(\theta, q_i^*; x_i) = \log(p_\theta(x_i)) - D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i)$$

$$\log(p_\theta(x_i)) = \log(p_\theta(x_i))$$
$$\mathcal{L}(\theta, q_\phi^*; x_i) + D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i) = \mathcal{L}(\theta, q_i^*; x_i) + D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i)$$

$$\mathcal{L}(\theta, q_\phi^*; x_i) - \mathcal{L}(\theta, q_i^*; x_i) = D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i) - D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i)$$

for comparing ELBO, note that necessarily $\mathcal{L}(\theta, q_\phi^*; x_i) \leq \mathcal{L}(\theta, q_i^*; x_i)$, because otherwise it should be the case that $\mathcal{L}(\theta, q_i^*; x_i) < \mathcal{L}(\theta, q_\phi^*; x_i)$, and this results in contradiction, because q_i^* supposed to be $argmax_q \mathcal{L}(\theta, \phi; x_i)$, but now we have found another solution q_ϕ^* that makes instance-dependent ELBO even larger which contradicts the fact that q_i^* resulted in the maximum ELBO. so $\mathcal{L}(\theta, q_\phi^*; x_i) \leq \mathcal{L}(\theta, q_i^*; x_i)$ and since we had:

$$\mathcal{L}(\theta, q_\phi^*; x_i) - \mathcal{L}(\theta, q_i^*; x_i) = D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i) - D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i)$$

then $\mathcal{L}(\theta, q_\phi^*; x_i) \leq \mathcal{L}(\theta, q_i^*; x_i)$ is equivalent to the following:

$$D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i) - D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i) \leq 0$$

$$D_{KL}(q_i^*(z|x_i)||p_\theta(z|x_i) \leq D_{KL}(q_\phi^*(z|x_i)||p_\theta(z|x_i)$$

so KL divergence for q_ϕ^* is bigger than q_i^*.

1.4

1.4.1

for estimating marginal likelihood we can use importance sampling:

$$p(x) \sim \frac{1}{K} \sum_{k=1}^{K} \frac{p(x, z_k)}{q(z_k, x)}$$

where $z_k \sim q(z|x)$

therefore, the log marginal likelihood can be estimated:

$$logp(x) \sim \sum_{k=1}^{K} [\log p(x, z_k)q - \log(z_k, x)] - \log K$$

Based on this, both approaches will perform the same as the same set of equations is applied to both.

1.4.2

if we were to estimate the marginal log-likelihood for all n samples in the training data set, then using q_ϕ^*, we only do $argmax_{q_\phi \in \mathcal{Q}} \sum_{i=1}^{n} \mathcal{L}(\theta, \phi; x_i)$ once for a summation, and then use it n times to obtain the log-likelihood of each sample. But using q_i^*, we need to do $q_i^* = argmax_{q_\phi \in \mathcal{Q}} \mathcal{L}(\theta, \phi; x_i)$ for every sample, and then use the results to get the log-likelihood estimate of each sample by ELBO which makes this approach computationally inefficient.

1.4.3

As was mentioned above, for q_ϕ^* we only need one neural net that gives $argmax_{q_\phi \in \mathcal{Q}} \sum_{i=1}^{n} \mathcal{L}(\theta, \phi; x_i)$ so we need to store only parameters for this network. Whereas in the other approach we need to store the parameter of n networks, one per instance variational approximator, so it needs more storage.

2

2.1

we can trace reverse conditional probability we have by using a condition on x_0 as well as x_t as below:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t, I)$$

using Bayes rule, we have:

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\alpha\, exp(-\frac{1}{2}(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}))$$

$$= exp(-\frac{1}{2}(\frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha_t x_{t-1}^2}{\beta_t} + \frac{x_{t-1}^2 - 2\sqrt{\bar{\alpha_{t-1}}}x_0 x_{t-1} + \bar{\alpha_{t-1}}x_0^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}))$$

$$= exp(-\frac{1}{2}((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})x_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}})x_{t-1} + C(x_t, x_0)))$$

where $C(x_t, x_0)$ is some function not involving x_{t-1} and detail are omitted. Following the standard Gaussian density function, the mean and variance can be parameterized as follows (recall that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$):

$$\tilde{\beta}_t = \frac{1}{(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})}$$

$$= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}.\beta_t$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}}}{\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}}$$

$$= (\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}})\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0$$

2.2

We computed an analytical solution for $q(x_{t-1}|x_t, x_0)$ in the previous section. But in fact, we want to reverse the forward process and sample from $q(x_{t-1}|x_t)$ to be able to recreate the true samples from a gaussian noise input, $x_T \sim \mathcal{N}(0, I)$. Since we don't have the x_0 in this reverse process, we must compute $q(x_{t-1}|x_t)$. Note that if β_T is small enough $q(x_{t-1}|x_t)$ will also be gaussian. Unfortunately, we cannot easily estimate $q(x_{t-1}|x_t)$ because it needs to use the entire dataset, which requires complex computations that grow exponentially with the number of steps. As a result, It is computationally intractable and impractical. Therefore we need to learn a model $p_\theta(x_{t-1}|x_t)$ to approximate these conditional probabilities in order to run the reverse diffusion process.

2.3

we can use the variational lower bound to optimize the negative log-likelihood.

$$-\log p_\theta(x_0) \leq -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$= -\log p_\theta(x_0) + E_{x_{1:T}\sim q(x_{1:T}|x_0)}[\log \frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_{0:T})}{p_\theta(x_0)}}]$$

$$= -\log p_\theta(x_0) + E_q[\log \frac{q(x_{1:T}|x_0}{p_\theta(x_{0:T})} + \log p_\theta(x_0)]$$

$$= E_q[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}]$$

so far we have shown that :

$$E_{q(x_0)} \log p_\theta(x_0) \geq E_q[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}]$$

this objective can be further rewritten to be a combination of several KL divergence and entropy terms:

$$E_q[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] = E_q[\log \frac{p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}]$$

$$= E_q[\log p_\theta(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}]$$

we need to bring $q(x_{t-1}|x_t, x_0)$ in the picture. we can see that:

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

so:

$$q(x_t|x_{t-1}) = \frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

plugging this into the equation derived from the previous section, we get:

$$E_q[\log p_\theta(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}]$$

$$= E_q[\log p_\theta(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=1}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}]$$

$$= E_q[\log p_\theta(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log q(x_T|x_0)]$$

this is because :

$$\sum_{t=1}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \sum_{t=1}^{T} \log q(x_{t-1}|x_0) - \sum_{t=1}^{T} \log q(x_t|x_0) = \log q(x_T|x_0)$$

so we get :

$$E_q[\log p_\theta(x_T) + \log p_\theta(x_0|x_1) \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log q(x_T|x_0)]$$

$$= E_q[\log p_\theta(x_0|x_1) - \log \frac{q(x_T|x_0)}{p_\theta(x_T)} \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}]$$

$$= E_q[\log p_\theta(x_0|x_1) - KL[q(x_T|x_0)||p_\theta(x_T)] - \sum_{t=2}^{T} KL[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]]$$

so in conclusion: we have shown that:

$$E_{q(x_0)} \log p_\theta(x_0) \geq E_q[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}]$$

and further we have shown that:

$$E_q[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] = E_q[\log p_\theta(x_0|x_1) - KL[q(x_T|x_0)||p_\theta(x_T)] - \sum_{t=2}^{T} KL[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]]$$

so putting all these together:

$$E_{q(x_0)} \log p_\theta(x_0) \geq E_q[\log p_\theta(x_0|x_1) - KL[q(x_T|x_0)||p_\theta(x_T)] - \sum_{t=2}^{T} KL[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]]$$

the first term is $L_0(x_0) = -E_{q(x_1|x_0)} \log p_\theta(x_0|x_1)$
the second term is: $L_T(x_0) = KL(q(x_T|x_0)||p(x_T))$
the last term is: $L_{t-1}(x_0) = E_{q(x_t|x_0)} KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$

2.4

every KL term in the above objective function (except for L_0) compares two Gaussian distributions and therefore they can be computed in closed form. L_T which is the full forward process and reverse process which is "fixed" so we can just replace it with a constant and can be ignored during training because q has no learnable parameters and x_T is a Gaussian noise.

2.5

The $\log p_\theta(x \mid z)$ term in ELBO for vanilla VAE is the reconstruction term, and It measures how well the model can reconstruct the input data from the latent variables. The $L_0(x_0)$ term in ELBO for DDPM is the reconstruction term, and It estimates how well the model can generate the input data from the initial latent variables. The KL divergence term in VAE objective function measures the difference between the variational distribution and the prior distribution over the latent variables, and It promotes the variational distribution to be close to the prior distribution. The denoising matching terms $L_{t-1}(x_0)$ in DDPM objective function is the key difference between the ELBO for a VAE and a DDPM. In DDPM, the transition distributions are modeled by a neural network, which has a set of trainable parameters. This makes the model more expressive and capable of capturing complex dependencies between the latent variables. In contrast, vanilla VAE only has trainable parameters for the mean and variance of the variational distribution. As a result, DDPM usually has more trainable parameters than the vanilla VAE, which makes it more expressive but also more computationally expensive to train.

2.6

- $$\tilde{\mu}_t(x_t, x_0) = \frac{\frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\tilde{\alpha}_{t-1}} x_0}{1-\tilde{\alpha}_{t-1}}}{\frac{\alpha_t}{\beta_t} + \frac{1}{1-\tilde{\alpha}_{t-1}}}$$
$$= (\frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\tilde{\alpha}_{t-1}} x_0}{1-\tilde{\alpha}_{t-1}}) \frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t} \beta_t$$
$$\frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t} x_t + \frac{\sqrt{\tilde{\alpha}_{t-1}} \beta_t}{1-\tilde{\alpha}_t} x_0$$

  we can represent $x_0 = \frac{1}{\sqrt{\tilde{\alpha}_t}}(x_t - \sqrt{1-\tilde{\alpha}_t}\epsilon_t)$ and plug it into the above equation and obtain:

$$\tilde{\mu}_t = \frac{\sqrt{\alpha_t}(1-\tilde{\alpha}_{t-1})}{1-\tilde{\alpha}_t} x_t + \frac{\sqrt{\tilde{\alpha}_{t-1}} \beta_t}{1-\tilde{\alpha}_t} \frac{1}{\sqrt{\tilde{\alpha}_t}}(x_t - \sqrt{1-\tilde{\alpha}_t}\epsilon_t)$$
$$= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_t)$$

- recall that we need to learn a neural network to approximate the conditioned probability distributions in the reverse diffusion process, $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$. we would like to train μ_θ to predict $\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_t)$. because x_t is available as input at training time, we can reparameterize the Gaussian noise term instead to make it predict ϵ_t from the input x_t at time step t:

$$\mu_\theta(x_t,t)=\frac{1}{\sqrt{\alpha_t}}\big(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(x_t,t)\big)$$

thus

$$x_{t-1}=\mathcal{N}(x_{t-1};\frac{1}{\sqrt{\alpha_t}}\big(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(x_t,t)\big),\sigma_t^2 I)$$

we've shown that: $L_{t-1}(x_0)=E_{q(x_t|x_0)}KL(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))$ we are given $p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\sigma_t^2 I)$ and $q(x_{t-1}|x_t,x_0)=\mathcal{N}(x_{t-1}|\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I)$ where $\tilde{\beta}_t=\sigma_t^2$, hence $q(x_{t-1}|x_t,x_0)=\mathcal{N}(x_{t-1}|\tilde{\mu}_t(x_t,x_0),\sigma_t^2 I)$

the loss term L_{t-1} is parameterized to minimize the difference from $\tilde{\mu}$:

$$L_{t-1}=E_{x_0,\epsilon}[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)||^2]+const$$

$$=E_{x_0,\epsilon}[\frac{1}{2\sigma_t^2}||\frac{1}{\sqrt{\alpha_t}}\big(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\big)-\frac{1}{\sqrt{\alpha_t}}\big(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(x_t,t)\big)||^2]+const$$

$$=E_{x_0,\epsilon}[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\sigma_t^2}||\epsilon_t-\epsilon_\theta(x_t,t)||^2]+const$$

$$=E_{x_0,\epsilon}[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\sigma_t^2}||\epsilon_t-\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon_t,t)||^2]+const$$

$$=E_{x_0,\epsilon}[\frac{\beta_t^2}{2\alpha_t(1-\bar{\alpha}_t)\sigma_t^2}||\epsilon_t-\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon_t,t)||^2]+const$$

$$=E_{x_0,\epsilon}[\lambda_t||\epsilon_t-\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon_t,t)||^2]+const$$

- $$p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\sigma_t^2 I)$$

$$L_0(x_0)=-E_{q(x_1|x_0)}\log p_\theta(x_0|x_1)=-E_{q(x_1|x_0)}\log\mathcal{N}(x_0;\mu_\theta(x_1,1),\sigma_1^2 I)$$

$$=-\frac{1}{2\sigma_1^2}||x_0-\mu_\theta(x_1,1)||^2+const$$

$$\mu_\theta(x_1,1)=\frac{1}{\sqrt{\alpha_1}}\big(x_1-\frac{\beta_1}{\sqrt{1-\bar{\alpha}_1}}\boldsymbol{\epsilon}_\theta(x_1,1)\big)$$

$$L_0(x_0)=-\frac{1}{2\sigma_1^2}||x_0-\frac{1}{\sqrt{\alpha_1}}\big(x_1-\frac{\beta_1}{\sqrt{1-\bar{\alpha}_1}}\boldsymbol{\epsilon}_\theta(x_1,1)\big)||^2+const$$

$$x_1=\sqrt{\alpha_1}x_0+\sqrt{1-\alpha_1}\epsilon$$

$$L_0(x_0)=-\frac{1}{2\sigma_1^2}||x_0-\frac{1}{\sqrt{\alpha_1}}\big(\sqrt{\alpha_t}x_0+\sqrt{1-\alpha_1}\epsilon-\frac{\beta_1}{\sqrt{1-\bar{\alpha}_1}}\boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0+\sqrt{1-\alpha_1}\epsilon,1)\big)||^2+const$$

$$=-\frac{1}{2\sigma_1^2}||-\frac{\sqrt{1-\alpha_1}}{\sqrt{\alpha_1}}\epsilon+\frac{\beta_1}{\sqrt{1-\bar{\alpha}_1}*\sqrt{\alpha_1}}\boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0+\sqrt{1-\alpha_1}\epsilon,1)\big)||^2$$

$$\bar{\alpha_1}=\alpha_1$$

$$=-\frac{1}{2\sigma_1^2}||-\frac{\sqrt{1-\alpha_1}}{\sqrt{\alpha_1}}\epsilon+\frac{\beta_1}{\sqrt{1-\alpha_1}*\sqrt{\alpha_1}}\boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0+\sqrt{1-\alpha_1}\epsilon,1)\big)||^2$$

$$= -\frac{1}{2\sigma_1^2}|| - \frac{\sqrt{1-\alpha_1}}{\sqrt{\alpha_1}}\epsilon + \frac{\sqrt{1-\alpha_1}}{\sqrt{\alpha_1}}\boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon, 1))||^2$$

$$= \frac{1}{2\sigma_1^2} * (\frac{\sqrt{1-\alpha_1}}{\sqrt{\alpha_1}})^2||\epsilon - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon, 1))||^2$$

$$= \frac{1}{2\sigma_1^2} * \frac{1-\alpha_1}{\alpha_1}||\epsilon - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon, 1))||^2$$

$$= \frac{1}{2\sigma_1^2} * \frac{\beta_1^2}{2\alpha_1(1-\bar{\alpha}_1)}||\epsilon - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_1}x_0 + \sqrt{1-\bar{\alpha}_1}\epsilon, 1))||^2$$

where in the previous part we have shown that

$$E_{q_(x_0)}L_{t-1}(x_0) = E_{x_0,\epsilon}[\frac{\beta_1^2}{2\alpha_1(1-\bar{\alpha}_1)\sigma_1^2}||\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_1}x_0 + \sqrt{1-\bar{\alpha}_1}\epsilon_t, 1)||^2] + const$$

these two equations are equal when $\epsilon_1 = \epsilon$

$$E_{q_(x_0)}L_0(x_0) = E_{x_0,\epsilon}[\frac{\beta_1^2}{2\alpha_1(1-\bar{\alpha}_1)\sigma_1^2}||\epsilon_1 - \epsilon_\theta(\sqrt{\bar{\alpha}_1}x_0 + \sqrt{1-\bar{\alpha}_1}\epsilon_1, 1)||^2] + const$$

2.7

As we've seen in section 2.3:

$$\mathcal{L}_{DDPM}(\theta; x_0) = -L_0(x_0) - \sum_{t=2}^{T} L_{t-1}(x_0) - L_T(x_0)$$

We've discussed that $L_T(x_0)$ has no learnable parameters so we can ignore it in the DDPM objective function. As a result, we can write the following equation as DDPM loss function:

$$\mathcal{L}_{DDPM}(\theta; x_0) = \sum_{t=1}^{T} E_{x_0,\epsilon}[\lambda_t||\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)||^2] + const =$$

$$E_{t,x_0,\epsilon}[\lambda_t||\epsilon_t - \epsilon_\theta(x_t, t)||^2] + const$$

3

we will first derive the closed form solution for D^* using calculus of variations. we know that for differentiable functions f(x) and differentiable functions g(y, x) with continuous derivative we have:

$$\frac{\partial}{\partial f(x)} \int g(f(x), x)\, dx = \frac{\partial}{\partial y} g(f(x), x)$$

to find D^*, we need to compute the following:

$$argmax_D E_{x\sim p_1}[\log D(x)] + E_{x\sim p_0}[\log(1 - D(x))].$$

$$= max_D \int f_1(x) log D(x)\, dx + \int f_0(x) log(1 - D(x)) dx$$

taking the functional derivative w.r.t D and setting to zero and find D^*, we get (using the corresponding densities f_0, f_1 for distributions p_0, p_1)

$$\frac{\partial}{\partial D} \int f_1(x) log D(x)\, dx + \frac{\partial}{\partial D} \int f_0(x) log(1 - D(x))\, dx = 0$$

now using the functional derivative property from the deep learning book, by defining y= D(x), g_1(y, x) = $f_1(x) \log y$, $g_0(y, x) = f_0(x) \log y$ we have:

$$\frac{\partial}{\partial D} \int f_1(x) \log D(x)\, dx + \frac{\partial}{\partial D} \int f_0(x) \log(1 - D(x))\, dx$$

$$= \frac{\partial}{\partial y}(f_1(x) \log y) + \frac{\partial}{\partial y}(f_0(x) \log(1 - y))$$

$$= \frac{f_1(x)}{y} - \frac{f_0(x)}{1 - y}$$

replacing y=D(x)

$$= \frac{f_1(x)}{D(X)} - \frac{f_0(x)}{1 - D(x)}$$

$$= 0$$

$$\frac{f_1(x)}{D(x)} = \frac{f_0(x)}{1 - D(x)}$$

$$\frac{f_0(x)}{f_1(x)} = \frac{1 - D(x)}{D(x)}$$

$$\frac{f_0(x)}{f_1(x)} = \frac{1}{D(x)} - 1$$

$$D^*(x) = \frac{f_1(x)}{f_0(x) + f_1(x)}$$

Now for estimating the JSD, let's write and manipulate it:

$$JSD(p_0, p_1) = \frac{1}{2}(KL(p_0||\mu) + KL(p_1||\mu)) = \frac{1}{2} \int f_0(x) \log \frac{f_0(x)}{\frac{f_0(x)+f_1(x)}{2}}\, dx + \frac{1}{2} \int f_1(x) \log \frac{f_1(x)}{\frac{f_0(x)+f_1(x)}{2}}\, dx$$

but note that from the equation for $D^*(x)$ we can easily get:

$$\frac{f_1(x)}{\frac{f_0(x)+f_1(x)}{2}} = 2D^*(x)$$

$$\frac{f_0(x)}{\frac{f_0(x)+f_1(x)}{2}} = \frac{f_0(x)}{f_1(x)} \frac{f_1(x)}{\frac{f_0(x)+f_1(x)}{2}} = (\frac{1}{D^*(x)} - 1)2D^*(x) = 2 - 2D^*(x)$$

plugging these results into the equation for JSD(p_0, p_1):

$$JSD(p_0, p_1) = \frac{1}{2} \int f_0(x) \log \frac{f_0(x)}{\frac{f_0(x)+f_1(x)}{2}}\, dx + \frac{1}{2} \int f_1(x) \log \frac{f_1(x)}{\frac{f_0(x)+f_1(x)}{2}}\, dx$$

$$= \frac{1}{2} \int f_0(x) \log 2(1 - D^*(x))\, dx + \frac{1}{2} \int f_1(x) \log 2D^*(x)\, dx$$

$$= \frac{1}{2} \int f_0(x) \log 2\, dx + \frac{1}{2} \int f_0(x) \log(1 - D^*(x))\, dx + \frac{1}{2} \int f_1(x) \log 2\, dx + \frac{1}{2} \int f_1(x) \log D^*(x)\, dx$$

$\int f_0(x) \log 2\, dx = \log 2 \int f_0(x)\, dx = \log 2$
$\int f_1(x) \log 2\, dx = \log 2 \int f_1(x)\, dx = \log 2$
therefore:

$$JSD(p_0, p_1) = \frac{1}{2} \int f_0(x) \log 2\, dx + \frac{1}{2} \int f_0(x) \log(1 - D^*(x))\, dx + \frac{1}{2} \int f_1(x) \log 2\, dx + \frac{1}{2} \int f_1(x) \log D^*(x)\, dx$$

$$= \log 2 + \frac{1}{2}(E_{x \sim p_1}[\log D(x)] + E_{x \sim p_0}[\log(1 - D(x))])$$

we can use the above equation to estimate the JSD.

3.1

from the equation of the optimal discriminator, we have that:

$$D^*(x) = \frac{f_1(x)}{f_0(x) + f_1(x)}$$

with a bit of manipulation, we get:

$$\frac{1}{D^*(x)} = \frac{f_0(x) + f_1(x)}{f_1(x)} = 1 + \frac{f_0(x)}{f_1(x)}$$

$$\frac{f_0(x)}{f_1(x)} = \frac{1}{D^*(x)} - 1 = \frac{1 - D^*(x)}{D^*(x)}$$

$$\frac{f_1(x)}{f_0(x)} = \frac{D^*(x)}{1 - D^*(x)}$$

$$f_1(x) = f_0(x)\frac{D^*(x)}{1 - D^*(x)}$$

this concludes the derivation.

4

4.1

$$J(W, W_p) = \frac{1}{2}E_{x_1,x_2}\left[\|W_p f_1 - \text{Stop-Grad}(f_2)\|_2^2\right]$$

$$\frac{1}{2}E_{x_1,x_2}[(W_p f_1 - f_2)^T(W_p f_1 - f_2)]$$

$$\frac{1}{2}E_{x_1,x_2}[((W_p f_1)^T - f_2^T)(W_p f_1 - f_2)]$$

$$\frac{1}{2}E_{x_1,x_2}[f_1^T W_p^T W_p f_1 - f_1^T W_p^T f_2 - f_2^T W_p f_1 + f_2^T f_2]$$

$$\frac{1}{2}E_{x_1,x_2}[tr(f_1^T W_p^T W_p f_1) - tr(f_1^T W_p^T f_2) - tr(f_2^T W_p f_1) + tr(f_2^T f_2)]$$

$$\frac{1}{2}E_{x_1,x_2}[tr(W_p^T W_p f_1^T f_1) - tr(W_p^T f_2 f_1^T) - tr(f_1 f_2^T W_p) + tr(f_2 f_2^T)]$$

$$\frac{1}{2}[E[tr(W_p^T W_p f_1^T f_1)] - E[tr(W_p^T f_2 f_1^T)] - E[tr(f_1 f_2^T W_p)] + E[tr(f_2 f_2^T)]]$$

$$\frac{1}{2}[tr(E[(W_p^T W_p f_1^T f_1)]) - tr(E[(W_p^T f_2 f_1^T)]) - tr(E[(f_1 f_2^T W_p)]) + tr(E[(f_2 f_2^T)])]$$

$$\frac{1}{2}[tr(W_p^T W_p E[f_1^T f_1]) - tr(W_p^T E[f_2 f_1^T]) - tr(E[f_1 f_2^T]W_p) + tr(E[f_2 f_2^T])]$$

$$\frac{1}{2}[tr(W_p^T W_p F_1) - tr(W_p^T F_{21}) - tr(F_{21} W_p) + tr(F_2)]$$

$$\frac{1}{2}[tr(W_p^T W_p F_1) - tr(W_p F_{12}) - tr(F_{21} W_p) + tr(F_2)]$$

4.2

$$W_p = -\frac{\partial J}{\partial W_p}$$

$$= -\frac{\partial}{\partial W_p}\frac{1}{2}[tr(W_p^T W_p F_1) - tr(F_{12}W_p) - tr(W_p F_{12}) - tr(F_2]$$

$$= -\frac{1}{2}[\frac{\partial}{\partial W_p}tr(W_p^T W_p F_1) - \frac{\partial}{\partial W_p}tr(F_{12}W_p) - \frac{\partial}{\partial W_p}tr(W_p F_{12}) - \frac{\partial}{\partial W_p}tr(F_2)]$$

$$\frac{\partial}{\partial W_p}tr(W_p^T W_p F_1) = W_p(F_1^T + F_1)$$

$$\frac{\partial}{\partial W_p}tr(F_{12}W_p) = F_{12}^T$$

$$\frac{\partial}{\partial W_p}tr(W_p F_{12}) = F_{12}^T$$

$$-\frac{\partial J}{\partial W_p} = -\frac{1}{2}[2W_p F_1^T - 2F_{12}^T]$$

since F_1 is symmetric:

$$-W_p F_1^T + F_{12}^T = -W_p F_1 + F_{12}^T$$

4.3

we have the following terms:

$$F_1 = W(X + X')W^T F_2 = W(X + X')W^T and F_{12} = WXW^T$$

so the gradient is given as

$$W(t) = -\frac{\partial J}{\partial W(t)}$$

$$= -W_p(t)^T W_p(t)W(t)(X + X') + (W_p(t)^T + W_p(t))W(t)X - W(t)(X + X')$$

$$= -(W_p(t)^T W_p(t) + I_{n2})W(t)X' - (W_p(t)W_p(t) - W_p(t)^T - W_p(t) + I_{n2})W(t)X$$

$$-(W_p(t)^T W_p(t) + I_{n2})W(t)X' - (W_p(t) - I_{n2}^T)(W_p(t) - I_{n2})W(t)X$$

$$= -(W_p(t)^T W_p(t) + I_{n2})W(t)X' - \tilde{W}_p(t)^T \tilde{W}_p(t)W(t)X$$

then using the definition of the Kronecker Product:

$$\frac{d}{dt}vec(W(t)) = -[X' \bigotimes (W_p(t)^T W_p(t) + I_{n2}) + X \bigotimes (\tilde{W}_p(t)^T \tilde{W}_p(t))]vec(W(t))$$

hence:

$$\frac{d}{dt}vec(W(t)) = -H(t)vec(W(t)$$

then as stated in the Note: For a time-varying positive definite matrix $H(t)$ whose minimal eigenvalues are bounded away from 0, the dynamics are shown below:

$$\frac{d}{dt}(t) = -H(t)(t),$$

satisfies the constraint $\|w(t)\|_2 = e^{-\lambda_0 t}\|w(0)\|_2$, implying that $w(t) \to 0$.

then using the property given that the minimal eigenvalue $\lambda_{min}(H(t))$ is bounded away from zero, i.e. $\inf_{t\geq 0} \lambda_{min}(H(t)) \geq \lambda_0 > 0$, we can prove that w(t) →0 and therefore collapse.

4.4

now that we don't have stop-gradient nor predictor we can replace $W_p = I_{n2}$, therefore:

$$W(t) = -\frac{\partial J}{\partial W(t)}$$

$$= -H(t)vec(W(t))$$

$$-[X'(t) \bigotimes (I_{n2} + I_{n2}) + X \bigotimes ((I_{n2} - I_{n2})^T (I_{n2} - I_{n2}))]vec(W(t))$$

$$= -[X' \bigotimes (I_{n2} + I_{n2})]vec(W(t))$$

$$-[X'(t) \bigotimes 2I_{n2}]vec(W(t))$$

so based on the proof in 4.3 and the fact that X' is a positive definite matrix, we can show that

$$\|vec(W(t))\|_2^2 \leq \exp(-\lambda_0 t)\|vec(W(0)\|_2 \to 0$$

hence: W(t) $\to$ 0.

4.5

The stop-gradient mechanism in SimSiam prevents the gradient from back-propagating through the predictor network, forcing the network to learn representations that are invariant to the specific output of the predictor. This encourages the network to capture higher-level features of the data, rather than memorizing specific details. By doing so, the stop-gradient and the predictor mechanism in SimSiam prevent representational collapse and help the network to learn more robust representations.

References