سوال ۱

گام 1 :

الف) $$P_\theta(x) = \int_z P_\theta(z)\, P_\theta(x|z)$$

ب) چالش اصلی در محاسبه این انتگرال intractible بودن آن به دلیل ابعاد زیاد z هست

ج) $$P_\theta(x) = \int_z P_\theta(x,z)\, dz = \int_z P_\theta(z) P_\theta(x|z)\, dz = E_{z\sim P_\theta(z)}[P_\theta(x|z)]$$

د) چون حساب کردن این انتگرال ممکن نیست سراغ تخمین زدن آن می رویم. برای اینکار از روش monte carlo استفاده می کنیم

$$P_\theta(x) \simeq \frac{1}{n}\sum_{i=1}^{n} P_\theta(x|z_i), \quad z_i \sim P_\theta(z)$$

گام 2 :

الف) $$P_\theta(x) = \int_z P_\theta(z) P_\theta(x|z) = \int_z P_\theta(x,z)\, dz = \int_z q(z) \frac{P_\theta(x,z)}{q(z)}\, dz$$

$$= E_{z\sim q(z)}\left[\frac{P_\theta(x,z)}{q(z)}\right]$$ ← همونطور که مشاهده میشه توزیعی که داریم ازش sample می گیریم متفاوته

$$\longrightarrow P_\theta(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{P_\theta(x,z_i)}{q(z_i)} \qquad z_i \sim q(z)$$

ب) یک انتخاب شهودی برای $q(z)$ میتونه $P_\theta(z|x)$ باشه

چرا که $P_\theta(z|x)$ توزیع latent variable ، z روشن میکنه که بیشترین شباهت رو به توزیع x داره و بهترین حالت برای کاهش واریانس است

گام 3 :

الف) مشتق گیری از یک جمع بسیار ساده تر از مشتق گیری از یک حاصل ضرب است. از آنجایی که الگوریتم های بهینه سازی مبتنی بر گرادیان (مشتق گیری) هستند، این ویژگی محاسبات را به شدت کاهش می دهد
به علاوه تابع لگاریتم یک تابع صعودی است این یعنی هر مقداری که $p_\theta(x)$ را maximize می کند $\log p_\theta(x)$ را نیز maximize می کند

ب) همانطور که در گام ۱ مشاهده شد، محاسبه مستقیم $P_\theta(x)$ به دلیل انتگرال‌گیری روی latent space z، intractable است. به طور مشابه وقتی محاسبه خود یک عبارت غیرعملی است، محاسبه لگاریتم آن نیز غیرعملی خواهد بود. حال سراغ یه lowerbound که همون ELBO باشه میریم

$$\log P_\theta(x) = \log E_{z\sim q(z)}\left[\frac{P_\theta(x,z)}{q(z)}\right]$$

از نامساوی jensen استفاده می‌کنیم، $\log(E[Y]) \geq E[\log(Y)]$

$$\log P_\theta(x) \geq E_{z\sim q(z)}\left[\log\frac{P_\theta(x,z)}{q(z)}\right]$$

$$= E_{z\sim q(z)}[\log P_\theta(x,z)] - E_{z\sim q(z)}[\log q(z)]$$

$$\mathcal{L}(x;\theta,q) = E_{q(z)}[\log P_\theta(x,z)] - E_{q(z)}[\log q(z)]$$

ج) maximize کردن ELBO هدف اصلی آموزش VAE است اما چگونگی اجرای آن با چالش همراه است و مسئله اصلی در محاسبه گرادیان نسبت به پارامترهای توزیع پسین ~~[illegible]~~ (encoder) است

د)

$$\mathcal{L}(x;\theta,\phi) = E_{z\sim q(z)}\left[\log\left(\frac{P_\theta(x,z)}{q(z)}\right)\right]$$

$$= E_{z\sim q_\phi(z)}\left[\log\left(\frac{P_\theta(z|x)\,P_\theta(x)}{q(z)}\right)\right]$$

$$= E_{q(z)}\left[\log P_\theta(z|x)\right] - \log q(z) + \log P_\theta(x)$$

$$\mathcal{L}(x;\theta,q) = \log P_\theta(x) - D_{KL}(q(z)\,\|\,P_\theta(z|x))$$

$$\log_\theta(x) - \mathcal{L}(x;\theta,q) = D_{KL}(q(z)\,\|\,P_\theta(z|x))$$

تفاضل بین ELBO و $\log P_\theta(x)$ با همین چیزی برابر است

سوال 4:

الف)

با توجه به فرض گفته شده داریم:

$$\log p_\theta(x) = \mathcal{L}(x;\theta,\lambda) + D_{KL}(q_\lambda(z|x) \| p_\theta(z|x))$$

برای کمینه کردن ترم KL به صورت غیر مستقیم، می توانیم ELBO را maximize کنیم دلیل این امر نیز دو چیز است

1- عبارت $\log p_\theta(X)$ به پارامترهای توزیع تقریبی (q_λ) یعنی λ بستگی ندارد

2- بنابراین وقتی ELBO را نسبت به λ بیشینه می کنیم، چون مجموع آن با D_{KL} ثابت است باعث کمینه شدن واگرایی KL یعنی $D_{KL}(q_\lambda(z|x) \| p_\theta(z|x))$ می شود

$$\nabla_\lambda \log p_\theta(x) = \nabla_\lambda \mathcal{L}(\theta,\lambda;x) + \nabla_\lambda D_{KL}(q_\lambda(z|x) \| p_\theta(z|X))$$

$$0 = \nabla_\lambda \mathcal{L}(\theta,\lambda;x) + \nabla_\lambda D_{KL}(q_\lambda(z|x) \| p_\theta(z|X))$$

$$\nabla_\lambda \mathcal{L}(\theta,\lambda;x) = -\nabla_\lambda D_{KL}(q_\lambda(z|x) \| p_\theta(z|x))$$

ب)

$$\cancel{\nabla_\theta} \mathcal{L}(x,\theta,\lambda) = E_{z\sim q_\lambda(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\lambda(z|x) \| p_\theta(z))$$

$$\nabla_\theta \mathcal{L}(x;\theta,\lambda) = E_{z\sim q_\lambda(z|x)}[\nabla_\theta \log p_\theta(x|z)] - \nabla_\theta D_{KL}(q_\lambda(z|x) \| p_\theta(z))$$

چون $q_\lambda(z|x)$ به θ وابسته نیست می توانیم گرادیان رو ببریم داخل

به فرم دیگه‌ای که بازنوشته شده:

$$\nabla_\theta \mathcal{L}(x,\theta,\lambda) = \nabla_\theta E_{z\sim q_\lambda(z|x)}[\log p_\theta(x,z)] - \cancelto{0}{\nabla_\theta E_{z\sim q_\lambda(z|x)}[\log q_\lambda(z)]}$$

$$= E_{z\sim q_\lambda(z|x)}[\nabla_\theta \log p_\theta(x,z)]$$

$$\simeq \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log p_\theta(x, z_i)$$

ج)

$$\nabla_\lambda \mathcal{L}(x;\theta,\lambda) = \nabla_\lambda E_{z\sim q_\lambda(z|x)}\left[\log P_\theta(x|z) - \log q_\lambda(z|x)\right]$$

چالش اصلی این است که توزیعی که از آن نمونه‌گیری می‌کنیم (q_λ) خود به پارامترهای (λ) وابسته است که می‌خواهیم نسبت به آن‌ها گرادیان بگیریم. این وابستگی مانع از آن می‌شود که عملگر گرادیان را به سادگی به داخل امید ریاضی منتقل کنیم زیرا تغییر λ خود فضای نمونه‌گیری را تغییر می‌دهد. این مشکل باعث می‌شود تخمین گرادیان با روش مونت کارلوی ساده، بایاس دار یا دارای واریانس بسیار بالا باشد

: Reparameterization Trick

راه حل این است که فرایند sample گرفتن را طوری بازنویسی کنیم که بخش تصادفی از پارامترهای λ جدا شود

$$z = \mu + \sigma \epsilon \qquad \epsilon \sim N(0, I)$$

$$\longrightarrow \mathcal{L}(x,\lambda;\theta) = E_{\epsilon\sim N(0,I)}\left[\log P(x, z)\right] - E_{\epsilon\sim N(0,I)}\left[\log q_\lambda(z)\right]$$

$$= E_{\epsilon\sim N(0,I)}\left[\nabla_\lambda\left[\log P_\theta(x,z) - \log q_\lambda(z)\right]\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \nabla_\lambda\left[\log P_\theta(x,z) - \log q_\lambda(z)\right]$$

5/b :

$$\log P_\theta(x) \geq \mathcal{L}(x;\theta,\lambda)$$

اگر فرض کنیم دیتاست ما $D=\{x_i\}_{i=1}^{N}$

$$\theta^* = \underbrace{\arg\max_\theta \frac{1}{N}\sum_{i=1}^{N}\log P_\theta(x_i)}$$

$$\log P_\theta(x_i) \geq \arg\max \mathcal{L}(x_i;\theta,\lambda_i)$$

$$\Rightarrow \sum_i \log P_\theta(x_i) \geq \sum_i \arg\max \mathcal{L}(x_i;\theta,\lambda)$$

$$\Rightarrow \frac{1}{N}\sum_i \log P_\theta(x_i) \geq \arg\max_{\lambda_1,\dots,\lambda_N} \frac{1}{N}\sum_i \mathcal{L}(x_i,\theta,\lambda_i)$$

(ب، 6b)

Initialize ϕ, θ repeat

sample randomly mini-batch $(x^{(1)} \dots x^{(M)}) \sim \mathcal{D}$ } I

sample noise m times $\varepsilon^{(i)} \sim p(\varepsilon)$

compute latent variable $z^{(i)} = \mu + \varepsilon^{(i)}\sigma$ } II

compute gradient of the min-batch estimator

$$g = \nabla_{\phi,\theta} \sum_{i=1}^{m} \tilde{L}_{ELBO}(\theta, \phi, x^{(i)})$$ } III

update parameters (with SGD or Adam)

$$\begin{bmatrix}\phi\\ \theta\end{bmatrix} \leftarrow \begin{bmatrix}\phi\\ \theta\end{bmatrix} + \alpha g$$

until convergence of ϕ, θ

I ابتدا پارامترهای encoder (ϕ) و decoder (θ) به صورت random، initialize می‌شود و از encoder برای رسیدن به فضای latent استفاده می‌کنیم
از decoder برای برگشتن از فضای latent به فضای ورودی استفاده می‌کنیم به این صورت که یه sample از latent گرفته میشه و میدیم به عنوان ورودی decode

II فرایند sample گیری به کمک reparameterization trick انجام می‌شود به این صورت که یه sample از $\varepsilon \sim N(0, I)$ می‌گیریم و $z = \mu + \varepsilon^{(i)}\sigma$ و این رو به عنوان ورودی به decoder میدیم

III تابع هدف هم با گرفتن گرادیان نسبت به θ و ϕ از elbo انجام میشه ابتدا جمع elbo ها رو تو مسیر حالی می‌زنیم و با استفاده از روشی مثل SGD یا Adam پارامترها را آپدیت می‌کنیم

بصورت کلی objective ما میشه

$$\arg\max_{\theta} \arg\max_{\lambda_1,\dots,\lambda_N} \sum_i \mathcal{L}(x_i;\theta,\lambda_i)$$

* گرادیان نسبت به θ برابر با مجموع گرادیان های هر ELBO نسبت به θ است یعنی وقتی به روزرسانی θ باید تاثیر آن بر بازسازی تمام داده ها در نظر گرفته شود

* گرادیان objective کلی نسبت به پارامترهای محلی λ_i فقط به همان نمونه داده x_i وابسته است این یعنی بهینه سازی λ_i تنها با هدف پیدا کردن λ برای x_i انجام می شود

این روش مقیاس پذیر نیست، زیرا به ازای هر داده ورودی، یک مجموعه پارامتر جدید (λ_i) ولی بهینه سازی انجام می شود

ب) فرض کنید پارامترهای encoder، Φ باشند و شبکه عصبی در encoder را به صورت $NN(x_i;\Phi)$ نمایش میدیم

$$\mu_\Phi, \sigma^2_\Phi = NN(x_i;\Phi)$$

از آنجایی که پارامترهای NN بین تمامی نمونه های داده مشترک هستند پس به جای استفاده از $\mathcal{L}(x_i,\theta,\lambda_i)$ داریم:

$$\mathcal{L}(x_i;\theta,\Phi) = E_{z\sim q_\Phi(z|x)}\left[\log P_\theta(x,z) - \log q_\Phi(z|x_i)\right]$$

سوال 6

الف)

$$\mathcal{L}(x_i;\theta,\Phi) = E_{z\sim q_\Phi(z|x)}\left[\log \frac{P_\theta(x, z)}{q_\Phi(z|x_i)}\right]$$

$$\mathcal{L}(x_i;\theta,\Phi) = E_{z\sim q_\Phi(z|x)}\left[\log \frac{P_\theta(x|z)\,P_\theta(z)}{q_\Phi(z|x_i)}\right]$$

$$\mathcal{L}(x_i;\theta,\Phi) = E_{z\sim q_\Phi(z|x)}\left[\log P_\theta(x|z)\right] - E_{z\sim q_\Phi(z|x)}\left[\frac{q_\Phi(z|x_i)}{P_\theta(z)}\right]$$

$$\mathcal{L}(x;\theta,\Phi) = E_{z\sim q_\Phi(z|x)}\left[\log P_\theta(x|z)\right] - KL\left(q_\Phi(z|x)\,\|\,P_\theta(z)\right)$$

سوال 2:

الف)

برای اینکار ابتدا log liklihood ای $P(x)$ را می نویسیم

$$E_{z\sim q(z_1,z_2|x)}[\log P(x)] = E_{z\sim q(z_1,z_2|x)}\left[\log\frac{P(x,z_1,z_2)}{P(z_1,z_2|x)}\right] = E_{z\sim q(z_1,z_2|x)}\left[\log\frac{P(x|z_1)P(z_1|z_2)P(z_2)}{P(z_1,z_2|x)}\right]$$

$$= E_{z\sim q(z_1,z_2|x)}\left[\log\frac{P(x|z_1)P(z_1|z_2)P(z_2)}{P(z_1|x)P(z_2|x)}\right] = E_{z\sim q(z_1,z_2|x)}\left[\log\frac{P(x|z_1)P(z_1|z_2)P(z_2)}{P(z_1|x)P(z_2|x)}\times\frac{q(z_1,z_2|x)}{q(z_1,z_2|x)}\right]$$

$$E_{z\sim q(z_1,z_2|x)}\left[\log\left(P(x|z_1)\times\left(\frac{q(z_1|x)\,q(z_2|z_1)}{P(z_1|z_2)P(z_2)}\right)^{-1}\times\frac{q(z_1|x)\,q(z_2|z_1)}{P(z_1|x)\,P(z_2|x)}\right)\right]$$

$$= E_{z\sim q(z_1,z_2|x)}[\log(P(x|z_1))] \underbrace{- E_{z\sim q(z_1,z_2|x)}\left[\log\frac{q(z_1|x)\,q(z_2|z_1)}{P(z_1|z_2)\,P(z_2)}\right]}_{e} + KL[q(z_1,z_2|x)\,||\,P(z_1,z_2|x)]$$

$$\to -E_{z\sim q(z_1|x)q(z_2|z_1)}\left[\log\frac{q(z_1|x)\,q(z_2|z_1)}{P(z_1|z_2)\,P(z_2)}\right]$$

$$= -\left[E_{z\sim q(z_1|x)q(z_2|z_1)}\left[\log\frac{q(z_2|z_1)}{P(z_2)}\right] + E_{z\sim q(z_1|x)q(z_2|z_1)}\left[\log\frac{q(z_1|x)}{P(z_1|z_2)}\right]\right]$$

$$= -\left[E_{z\sim q(z_1|x)}\left[E_{z\sim q(z_2|z_1)}\left[\log\frac{q(z_2|z_1)}{P(z_2)}\right]\right] + E_{z\sim q(z_1|x)q(z_2|z_1)}\left[\log\frac{q(z_1|x)}{P(z_1|z_2)}\right]\right]$$

$$= -E_{z\sim q(z_1|x)}[KL(q(z_2|z_1)\,||\,P(z_2))] \underbrace{- E_{z\sim q(z_1|x)q(z_2|z_1)}\left[\log\frac{q(z_1|x)}{P(z_1|z_2)}\right]}_{e}$$

$$\to -E_{q(z_1|x)}[\log(z_1|x) - E_{q}(z_2|z_1)[\log(P(z_1|z_2))]]$$

$$\longrightarrow \log\tilde{P}(z_1) = E_{q}(z_2|z_1)[\log P(z_1|z_2)]$$

$$* \; -E_{q(z_1|x)}\left[\log\frac{q(z_1|x)}{\tilde{P}(z_1)}\right] = -KL(q(z_1|x)\,||\,\tilde{P}(z_1))$$

$$\to E_{z\sim q(z_2,z_1|x)}[\log P(x)] = E_{z\sim q(z_1,z_2|x)}[\log P(x|z_1)] - E_{z\sim q(z_1|x)}[KL(q(z_2|z_1)\,||\,P(z_2))]$$

$$- E_{z\sim q(z_2|z_1)}[KL(q(z_1|x))\,||\,\tilde{P}(z_1)] + \underbrace{KL[q(z_1,z_2|x)\,||\,P(z_1,z_2|x)]}_{\text{intractible}}$$

ما فقط ترم ها elbo را داریم

subject : date :

ب)

ترم اول:

از آنجایی که $p(x|z_1)$ تنها به z_1 وابسته است، این عبارت $E_{z\sim q(z|x)}[\log p(x|z_1)]$ ساده می‌شود. این بخش نشان می‌دهد که مدل چقدر خوب می‌تواند داده ورودی x را از روی نمایش لایه latent z_1 بازسازی کند. که به این جمله reconstruction نامیده می‌شود و به عملکرد decoder مرتبط است.
بیشینه کردن این ترم باعث می‌شود که مدل یاد بگیرد تا از latent z_1، عکس‌های شبیه به ورودی اصلی تولید کند.

ترم دوم:

میزان شباهت $q(z_2|z_1)$ (سطح بالاتر latent variable که از z_1 استنتاج می‌شود) به توزیع $p(z_2)$ (معمولاً توزیع ساده مثل $\mathcal{N}(0, I)$) را ارزیابی می‌کند.
این جمله به عنوان یک regularizer برای latent space در z_2 عمل می‌کند. با کمینه کردن این ترم مدل به سمتی می‌رود که z_2 را به گونه‌ای استنتاج کند که با prior dist آن سازگار باشد. این کار امکان نمونه‌برداری معنادار mean از z_2 برای تولید داده‌های جدید را فراهم می‌آورد

ترم سوم:

این بخش میزان شباهت توزیع $q(z_1|x)$ (یعنی توزیع latent variable در z_1 که از ورودی x به دست می‌آید) به $p(z_1|z_2)$ با Expectation روی $q(z_2|z_1)$ که خود از توزیع $p(z_2)$ به دست می‌آید
این ترم به عنوان regularizer برای latent space در z_1 عمل می‌کند. این بخش اطمینان می‌دهد که توزیع z_1 که از داده‌ها یاد گرفته می‌شود با ساختار hierarchical مدل و نحوه ارتباط آن با z_2 سازگار باشد.

ج)

۱) دلایل وقوع (posterior collapse)

۱. اگر به ویژه decoder ($p(x|z_1)$ یا $p(z_i|z_{i+1})$) بسیار قدرتمند باشد می‌تواند داده‌ها یا لایه‌های latent پایینی را با دقت خوبی بازسازی کند، حتی اگر اطلاعات کمی از latent variable های بالایی دریافت کند

۲. تابع ELBO شامل دو بخش اصلی است: reconstruction و KL-Divergence. KL ها توزیع posterior $q(z|\cdot)$ را مجبور می‌کند تا به توزیع prior $p(z)$ نزدیک شود. اگر decoder بتواند بدون نیاز به اطلاعات کمی از latent var کار بازسازی خوبی انجام دهد مدل برای بیشینه کردن ELBO راه آسانی را انتخاب می‌کند؛ یعنی KL ها را مینیمم کند و متغیر را با نزدیک کردن $q(z|\cdot)$ به $p(z)$ به سمت صفر میل می‌دهد. این امر باعث می‌شود latent variable از ورودی غیر مستقل شود و اطلاعاتی از داده را کد نکند. این پدیده برای لایه‌های بالاتر ... و z_2, z_3 محتمل‌تر است

سوال 3 :

ابتدا سراغ عبارت زیر می رویم :

$$\inf_{D \in D} \left\{ \int \left[(sth)\, p_d(x) + (sth)\, p_g(x) \right] dx \right\}$$

در واقع این میانگین جمع loss ها از داده های واقعی و فیک است :

① اگر sample از توزیع واقعی یعنی p_d آمده باشد که احتمال آن 0.5 برچسب واقعی برای این داده ها $y=1$ است بنابراین loss برابر است با $l(1.D(x))$.

② برای داده های فیک بصورت مشابه اگر sample از توزیع داده های جنریت شده یعنی p_g آمده باشد که احتمال آن 0.5 ، loss برابر است با $l(-1.D(x))$

بنابراین $R_I(D)$ را بصورت زیر بازنویسی می کنیم

$$R_I(D) = \int_x 0.5\, p_d(x).\, l(D(x)) + 0.5\, p_g(x).\, l(-D(x))\, dx$$

$$\longrightarrow \inf_{D \in D} \int \left[0.5.\, l(D(x)).\, p_d(x) + 0.5.\, l(-D(x))\, p_g(x) \right] dx$$

حالا باید بریم سراغ حساب کردن اینفیمم.

برای سادگی از نوتیشن $d = D(x)$ استفاده می کنیم پس داریم :

$$L(d,x) = p_d(x)\, l(d) + p_g\, l(-d)$$

$$R_I(D) = \int \frac{1}{2} \left[\inf L(d,x) \right] dx$$ پس چون اینفیمم به d بستگی داره نه x پس میبریمش داخل انتگرال

می توانیم یک loss function ساده که شرایط مسئله رو ارضا می کنه تعریف کنیم $l(z) = e^{-z}$

$$L(d,x) = p_d(x)\, e^{-d} + p_g(x)\, e^{d}$$

برای پیدا کردن d ای که optimal است از L نسبت به d مشتق می گیریم و برابر با صفر می گذاریم

$$\frac{\partial L}{\partial d} = -p_d(x)\, e^{-d} + p_g(x)\, e^{d} = 0$$

$$\Rightarrow p_g(x)\, e^{d} = p_d(x)\, e^{-d} \Rightarrow e^{2d} = \frac{p_d(x)}{p_g(x)}$$

$$\Longrightarrow \boxed{d^* = \frac{1}{2} \log\left(\frac{p_d(x)}{p_g(x)}\right)}$$

2) اثر بر جریان اطلاعات:

1) این پدیده نشان می‌دهد که اطلاعات از داده ورودی به طور موثر به لایه‌های latent بالاتر منتقل نشود یا برعکس، اطلاعات از لایه‌های latent بالای به طور موثری تولید لایه‌های پایین‌تر یا خود داده استفاده نمی‌شوند

2) latent varible‌هایی که دچار این پدیده شده‌اند، در یادگیری و مدل‌سازی ویژگی‌های مهم داده و تغییرات موجود در داده‌ها به کار گرفته نمی‌شوند. مدل عملاً آنها را نادیده می‌گیرد

3) هدف از ساختار Hierarchial، یادگیری ویژگی‌ها در سطوح مختلف انتزاع است. این پدیده به این معنی است که ساختار Hierarchial به درستی شکل نگرفته و برخی لایه‌ها «مرده» یا غیرکاربردی شده‌اند

3) دو تکنیک برای جلوگیری از این پدیده

1) KL Annealing:

در این روش وزن جمله‌ی KL در تابع هزینه ELBO در طول فرآیند آموزش به تدریج افزایش داده می‌شود. آموزش با وزن صفر یا بسیار کم برای جمله KL آغاز شده و این وزن به آرامی تا مقدار نهایی خود یعنی 1 افزایش می‌یابد

در مراحل اولیه training، با کم بودن وزن جمله KL، مدل بیشتر روی جمله reconstruction تمرکز می‌کند. این امر مدل را مجبور می‌کند تا برای reconstruct دقیق داده‌ها از latent var‌ها استفاده کرده و اطلاعات مفیدی را در آن کد کند. قبل از اینکه فشار جمله KL برای نزدیک کردن توزیع Postrior به prior خیلی قوی شود، encoder فرصت پیدا می‌کند تا نمایش‌های معنی‌داری یاد بگیرد

2) استفاده از decoder ضعیف‌تر:

ظرفیت یا قدرت بخش decoder مدل ($P(x|z_1)$ یا $P(z_i|z_{i+1})$) کاهش داده می‌شود. مثلاً از مدل‌های محدود شده با پیچیدگی کمتر استفاده می‌شود یا تعداد پارامترهای آن کم می‌شود

decoder با قدرت کمتر برای بازسازی دقیق داده‌ها، نیازمند اطلاعات بیشتری از latent variable‌ها خواهد بود. در نتیجه، نمی‌تواند به راحتی latent variable‌ها را نادیده بگیرد و این امر از غیرفعال شدن آن‌ها جلوگیری می‌کند

$$R(D) = \frac{1}{2}\int[(\rho_d+\rho_g)(\log(\rho_g+\rho_d)-\log_{\rho_g}) - \rho_d(\log\rho_d - \log\rho_g)]dx$$

$$R(D) = \frac{1}{2}\int[\rho_d(\log(\rho_d+\rho_g) - \log\rho_d) + \rho_g(\log(\rho_d+\rho_g) - \log\rho_g)]dx$$

$$R(D) = \frac{1}{2}\int[\rho_d\log\left(\frac{\rho_d+\rho_g}{\rho_d}\right) + \rho_g\log\log\left(\frac{\rho_d+\rho_g}{\rho_g}\right)]dx$$

$$\xrightarrow[\rho_d+\rho_g = 2m]{m(x) = \frac{\rho_d(x)+\rho_g(x)}{2}} R(D) = \frac{1}{2}\int[\rho_d\log\left(\frac{2m}{\rho_d}\right) + \rho_g\log\left(\frac{2m}{\rho_g}\right)]dx$$

$$= \frac{1}{2}\int[\rho_d(\log 2 + \log\frac{m}{\rho_d}) + \rho_g(\log 2 + \log\frac{m}{\rho_g})]dx$$

$$R(D) = \underbrace{\frac{1}{2}\int(\rho_d+\rho_g)\log 2\, dx}_{I} + \underbrace{\frac{1}{2}\int\left(\rho_d\log\frac{m}{\rho_d} + \rho_g\log\frac{m}{\rho_g}\right)dx}_{II}$$

I: $\int(\rho_d+\rho_g)dx = \int\rho_d dx + \int\rho_g dx = 1+1 \rightarrow \frac{1}{2}\times 2\times\log 2 = \log 2$

II: $I_{JS} = \frac{1}{2}D_{KL}(\rho_d \| m) + \frac{1}{2}D_{KL}(\rho_g \| m) = \frac{1}{2}\int\rho_d\log\frac{\rho_d}{m}dx + \frac{1}{2}\int\rho_g\log\frac{\rho_g}{m}dx$

$$\rightarrow \frac{1}{2}\int\left(\rho_d\log\frac{m}{\rho_d} + \rho_g\log\frac{m}{\rho_g}\right)dx = -I_{JS}$$

$$\boxed{R(D) = \log 2 - I_{JS}(P_d, P_g)}$$

ب)

ما می‌خواهیم $g(d,u) = u(1-d)^2 + (1+d)^2$ را minimize کنیم

$$\frac{\partial g}{\partial d} = -2u(1-d) + 2(1+d) = 0 \Rightarrow d^* = \frac{u-1}{u+1}$$

حالا این d^* را در رابطه می‌گذاریم

$$-f(u) = u\left(1-\frac{u-1}{u+1}\right)^2 + \left(1+\frac{u-1}{u+1}\right)^2$$

$$-f(u) \overset{?}{=} \frac{4u}{u+1} \longrightarrow f(u) = -\frac{4u}{u+1}$$

→ نزولی ✓

→ محدب ✓

$$D^*(x) = \frac{\frac{P_d(x)}{P_g(x)} - 1}{\frac{P_d(x)}{P_g(x)} + 1} = \frac{P_d(x) - P_g(x)}{P_d(x) + P_g(x)} \rightarrow \text{optimal Discriminator}$$

$$R(D) = \frac{1}{2}\int P_g(x)\left(\frac{4u}{u+1}\right)dx = 2\int \frac{P_d(x)P_g(x)}{P_d(x)+P_g(x)}\,dx$$

$$\longrightarrow R(D) = \frac{1}{2} I_g(P_d, P_g) \quad , \quad g(t) = \frac{4t}{t+1}$$

ج)

ما می‌خواهیم $g(d,u) = u\log(1-e^{-d}) + \log(1+e^{d})$

$$\frac{\partial g}{\partial d} = u\left(\frac{-e^{-d}}{1+e^{-d}}\right) + \frac{e^{d}}{1+e^{d}} = 0 \longrightarrow u = e^{d} \Rightarrow d^* = \log u$$

$$-f(u) = u\log(1+e^{-\log u}) + \log(1+e^{\log u}) = (u+1)\log(u+1) - u\log u$$

نزولی

محدب

$$D^*(x) = \log\left(\frac{P_d(x)}{P_g(x)}\right)$$

$$R(D) = \frac{1}{2}\int P_g(x)\left[(u+1)\log(u+1) - u\log u\right]dx$$

$$R(D) = \frac{1}{2}\int P_g(x)\left[\left(\frac{P_d}{P_g}+1\right)\log\left(\frac{P_d}{P_g}+1\right) - \frac{P_d}{P_g}\log\left(\frac{P_d}{P_g}\right)\right]dx$$

$$= \frac{1}{2}\int\left[(P_d+P_g)\log\left(\frac{P_d+P_g}{P_g}\right) - P_d\log\left(\frac{P_d}{P_g}\right)\right]dx$$

بخش دوم

برای اینکه کار ساده‌تر شود تابع $f(u)$ را به صورت زیر بازنویسی می‌کنیم

$$-f(u) = \min\{u \cdot I(d) + I(-d)\}$$

$$\rightarrow u = \frac{p_d(x)}{p_g(x)}, \quad d = D(x), \quad R_I(D) = \frac{1}{2}\int p_g(x)[-f(x)]\,dx$$

الف)

ما باید تابع $g(d,u) = u\,I(d \leq 0) + I(-d < 0)$ را minimize کنیم

$d>0 \rightarrow g(d,u) = u \cdot 0 + 1 = 1$

$d<0 \rightarrow g(d,u) = u \cdot 1 + 0 = u$

$d=0 \rightarrow g(d,u) = u \cdot 1 + 1 = u+1$

بنابراین

$$-f(u) = \min(u,1) \rightarrow f(u) = -\min(u,1)$$

غیر نزولی ✓

محدب ✓

متمایزکننده بهینه $D^*(x)$:

اگر $u>1$ باشد (یعنی $p_d(x)>p_g(x)$) ، باید مقدار خروجی متمایزکننده بزرگتر از صفر ($d>0$) انتخاب شود

و اگر $u<1$ باشد (یعنی $p_d(x)<p_g(x)$) باید مقدار کوچکتر از صفر ($d<0$) انتخاب شود

بنابراین متمایزکننده بهینه یک تابع قاطع است یعنی به طور قطعی تصمیم می‌گیرد که داده واقعی یا جعلی است

$$\rightarrow R_I(D) = \frac{1}{2}\int p_g(x)\min\left(\frac{p_d(x)}{p_g(x)}, 1\right)dx = \frac{1}{2}\int \min(p_d(x), p_g(x))\,dx$$

$$\int \min(p,q)\,dx = 1 - \frac{1}{2}\int |p-q|\,dx \quad *$$

$$I_{TV}(P,Q) = \frac{1}{2}\int |P-Q|\,dx \quad *$$

$$\rightarrow R(D) = \frac{1}{2}\left(1 - I_{TV}(P_d, P_g)\right)$$

از اونجایی که مشتق دوم L یعنی $p_d(x)e^{-d}+p_g(x)e^{d}$ همیشه بزرگتر از صفر است پس این d^* minimum است

$$L(d^*,x)=p_d(x)e^{-d^*}+p_g(x)e^{d^*} \qquad \begin{cases} e^{d^*}=\sqrt{\frac{p_d(x)}{p_g(x)}} \\ e^{-d^*}=\sqrt{\frac{p_g(x)}{p_d(x)}} \end{cases}$$

$$L(d^*,x)=p_d(x)\sqrt{\frac{p_g(x)}{p_d(x)}}+p_g(x)\sqrt{\frac{p_d(x)}{p_g(x)}}$$

$$L(d^*,x)=2\sqrt{p_d(x)p_g(x)}$$

$$\rightarrow \inf R_I(D)=\int\frac{1}{2}\left(2\sqrt{p_d(x)p_g(x)}\right)dx=\int\sqrt{p_d(x)p_g(x)}\,dx$$

در صورت تمرین گفته شده که نشان دهیم نتیجه برابر است با

$$-\frac{1}{2}\int f\left(\frac{p_d(x)}{p_g(x)}\right)p_g(x)\,dx$$

اول باید فرم $\sqrt{p_d(x)p_g(x)}$ رو یکم عوض کنیم تا شبیه فرم بالا بشه

$$\sqrt{p_d(x)p_g(x)}=\sqrt{\frac{p_d(x)}{p_g(x)}p_g^2(x)}=\sqrt{\frac{p_d(x)}{p_g(x)}}\;p_g(x)$$

$$\longrightarrow \inf R_I(D)=\int\sqrt{\frac{p_d(x)}{p_g(x)}}\;p_g(x)\,dx$$

خب اگه $u=\frac{p_d(x)}{p_g(x)}$ باشد داریم

$$-\frac{1}{2}f(u)\,p_g(x)=\sqrt{u}\,p_g(x)$$

$$\rightarrow \boxed{f(u)=-2\sqrt{u}}$$

چک کردن ویژگی ها:

نزولی بودن:

$$f'(u)=-\frac{1}{\sqrt{u}}$$

از اونجایی که u تقسیم 2 احتمال است همیشه مثبته یعنی $(u>0)$ پس $f'(u)$ همیشه منفی است و منفی بودن مشتق اول یعنی $f(u)$ نزولی است

محدب بودن:

$$f''(u)=\frac{1}{2\sqrt{u^3}}$$

از اونجایی که محدبه پس مشتق دوم همیشه مثبته و این به معنی concave بودن است

سوال 4 بخش الف)

ابتدا عبارت داخل Expectation را باز می‌کنیم

$$\|s_\theta(x) - \nabla_x \log q(x)\|^2 = \|s_\theta(x)\|^2 - 2 s_\theta(x) \cdot \nabla_x \log q(x) + \|\nabla_x \log q(x)\|^2$$

$$l_1(\theta) = \underbrace{E_{q(x)}\left[\frac{1}{2}\|s_\theta(x)\|^2\right]}_{I} - \underbrace{E_{q(x)}\left[s_\theta(x) \cdot \nabla_x \log q(x)\right]}_{II} + \underbrace{E_{q(x)}\left[\frac{1}{2}\|\nabla_x \log q(x)\|^2\right]}_{III}$$

I این ترم در حکم ما قرار داره پس نیازی نیست تغییرش بدیم

III : این ترم تنها به توزیع داده $q(x)$ بستگی دارد و به پارامتر مدل θ بستگی ندارد پس می‌توان آن را به عنوان ترم ثابت C_1 در نظر گرفت

II : این ترم نیاز به تغییر داره تا به حکم سوال برسیم

$$E_{q(x)}\left[s_\theta(x) \cdot \nabla_x \log q(x)\right] = \int_x q(x) \left(s_\theta(x) \cdot \nabla_x \log (q(x)\right) dx$$

$$\xrightarrow{\nabla_x \log q(x) = \frac{\nabla_x q(x)}{q(x)}} \int_x q(x) \left(s_\theta(x) \cdot \frac{\nabla_x q(x)}{q(x)}\right) dx = \int_x s_\theta(x) \cdot \nabla_x q(x) dx$$

انتگرال یه دات پروداکت است که مجموع روی dim های i است، $\sum_i \int_x s_{\theta,i}(x) \frac{\partial q(x)}{\partial x_i} dx$

$$\xrightarrow{\text{انتگرال جزء به جزء}} \int_x s_{\theta,i}(x) \frac{\partial q(x)}{\partial x_i} dx = \cancel{\left[s_{\theta,i}(x)\, q(x)\right]_{-\infty}^{+\infty}}^{\,0} - \int_x q(x) \frac{\partial s_{\theta,i}(x)}{\partial x_i} dx$$

طبق فرض سوال

$$\sum_i \int_x s_{\theta,i}(x) \frac{\partial q(x)}{\partial x_i} dx = \sum_i \left(-\int_x q(x) \frac{\partial s_{\theta,i}(x)}{\partial x_i} dx\right) = -\int_x q(x) \left(\sum_i \frac{\partial s_{\theta,i}(x)}{\partial x_i}\right) dx$$

$$= -E_{q(x)}\left[Tr\left(\nabla_x s_\theta(x)\right)\right]$$

$$\rightarrow \boxed{l_1(\theta) = E_{q(x)}\left[\frac{1}{2}\|s_\theta(x)\|^2 + Tr\left(\nabla_x s_\theta(x)\right)\right] + C_1}$$

این تابع هزینه برای آموزش عملی مناسب نیست چون محاسبه‌ی $Tr(\nabla_x s_\theta(x))$ بسیار پرهزینه است که این یک ماتریس ژاکوبین است. اگر ورودی ما یک عکس با ابعاد بالا باشد ژاکوبین یک ماتریس غول پیکر خواهد بود محاسبه تمام درایه‌های این ماتریس بزرگ برای هر نمونه در هر مرحله از آموزش، به توان محاسباتی فوق‌العاده بالایی نیاز دارد که در عمل غیرممکن است.

بخش ب)

$$l_3(\theta) = E_{x\sim q(x),\, \epsilon\sim N(0,I)}\left[\frac{1}{2}\left\| s_\theta(x+\sigma\epsilon) + \frac{\epsilon}{\sigma}\right\|^2\right]$$

$$\xrightarrow[\epsilon = \frac{\tilde{x}-x}{\sigma}]{\tilde{x} = x+\sigma\epsilon} l_3(\theta) = E_{q(x,\tilde{x})}\left[\frac{1}{2}\left\| s_\theta(\tilde{x}) + \frac{\tilde{x}-x}{\sigma^2}\right\|^2\right]$$

$$\xrightarrow[= -\frac{\tilde{x}-x}{\sigma^2}]{\nabla_{\tilde{x}} \log q(\tilde{x}|x) = \frac{1}{\sigma^2}(x-\tilde{x})} l_3(\theta) = E_{q(x,\tilde{x})}\left[\frac{1}{2}\left\| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q(\tilde{x}|x)\right\|^2\right]$$

حال با گسترش DSM objective میپردازیم به رابطه آن با score matching برسیم

$$L_{sm}(\theta) = E_{q(\tilde{x})}\left[\frac{1}{2}\left\| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q(\tilde{x})\right\|^2\right]$$

$$\xrightarrow{DSM\ obj} l_3(\theta) = E_{q(x,\tilde{x})}\left[\underbrace{\frac{1}{2}\| s_\theta(\tilde{x})\|^2}_{I} - \underbrace{s_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q(\tilde{x}|x)}_{II} + \underbrace{\frac{1}{2}\|\nabla_{\tilde{x}} \log q(\tilde{x}|x)\|^2}_{III}\right]$$

I این ترم که اول sm هست پس کاریش نداریم

II این فقط به نویز بستگی داره و به پارامترهای مدل وابسته نیست پس میتونه ثابت باشه

III اینم از بخش اول نشون دادیم که برابر $E_{q(\tilde{x})}[Tr(\nabla_{\tilde{x}} s_\theta(\tilde{x}))]$ است

بنابراین DSM objective بصورت زیر میشود

$$l_3(\theta) = E_{q(\tilde{x})}\left[\frac{1}{2}\| s_\theta(\tilde{x})\|^2 + Tr(\nabla_{\tilde{x}} s_\theta(\tilde{x}))\right] + c$$

$$L_{sm}(\theta) = E_{q(\tilde{x})}\left[\frac{1}{2}\| s_\theta(\tilde{x})\|^2 + Tr(\nabla_{\tilde{x}} s_\theta(\tilde{x}))\right] + c'$$

بنابراین هر دوی $l_3(\theta)$ (DSM) و $L_{sm}(\theta)$ دقیقاً به عبارت یکسانی شامل θ ساده می‌شوند پس مینیمم کردن یکی از آنها برابر با مینیمایز کردن دیگری است مقدار آنها فقط به اندازه ی مقداری ثابتی فرق دارد که تاثیری روی مینیمم نخواهد داشت

سوال ۵ :

الف) ~~توجه~~ به کمک استقرا باید ثابت کنیم ~~$q_\sigma(x_t|x_0)=N(\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)I)$~~

پایه استقرا:

برای آخرین گام زمانی یعنی $t=T$ توزیع حاشیه‌ای مستقیماً در تعریف q_σ به ما داده شده است

$$q_\sigma(x_T|x_0)=N(x_T;\sqrt{\bar\alpha_T}\,x_0,(1-\bar\alpha_T)I)$$

این رابطه با فرمول مطابقت دارد، پس پایه استقرا برقرار است

فرض استقرا

فرض می‌کنیم که برای یک گام زمانی دلخواه t (که در آن $1<t\le T$)، توزیع حاشیه‌ای، شکل مورد نظر ما را دارد

$$q_\sigma(x_t|x_0)=N(x_t;\sqrt{\bar\alpha_t}\,x_0,(1-\bar\alpha_t)I)$$

گام استقرا:

اکنون باید نشان بدیم که توزیع حاشیه‌ای برای گام زمانی قبلی، یعنی $t-1$ نیز از همین فرمول تبعیت می‌کند. برای یافتن $q_\sigma(x_{t-1}|x_0)$ از توزیع $q_\sigma(x_{t-1},x_t|x_0)$ روی متغیر x_t انتگرال می‌گیریم

$$q_\sigma(x_{t-1}|x_t)=\int q_\sigma(x_{t-1}|x_t,x_0)\,q_\sigma(x_t|x_0)\,dx_t$$

این عبارت، انتگرال روی حاصل ضرب دو توزیع گوسی است و نتیجه آن نیز یک توزیع گوسی خواهد بود. کافی است میانگین و کوواریانس این توزیع جدید را محاسبه کنیم

بر اساس فرض $x_t\sim N(\mu_t,\Sigma_t)$ ⟵

$$\mu_t=\sqrt{\bar\alpha_t}\,x_0$$
$$\Sigma_t=(1-\bar\alpha_t)I$$

بر اساس تعریف مسئله $q_\sigma(x_{t-1}|x_t,x_0)$ نیز گوسی است $x_{t-1}\sim N(Ax_t+b,\Sigma_{t-1|t})$

$$A=\frac{\sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\bar\alpha_t}}\,I$$

$$b=\sqrt{\bar\alpha_{t-1}}\,x_0-A\sqrt{\bar\alpha_t}\,x_0$$

$$\Sigma_{t-1|t}=\sigma_t^2 I$$

subject : date :

با استفاده از خواص تبدیل خطی متغیرهای گوسی، توزیع حاشیه‌ای حاصل برای x_{t-1} یک توزیع گوسی با Σ_{t-1} و μ_{t-1} خواهد بود

μ_{t-1}:

$$\mu_{t-1} = E[x_{t-1}] = A\mu_t + b$$

$$\mu_{t-1} = \left(\frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} I\right)\left(\sqrt{\bar{\alpha}_t}\, x_0\right) + \left(\sqrt{\bar{\alpha}_{t-1}}\, x_0 - \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\bar{\alpha}_t}\, x_0\right)$$

$$\mu_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x_0$$

Σ_{t-1}:

$$\Sigma_{t-1} = \mathrm{cov}(x_{t-1}) = A\Sigma_t A^T + \Sigma_{t-1|t}$$

$$\Sigma_{t-1} = \left(\frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} I\right)\left((1-\bar{\alpha}_t) I\right)\left(\frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} I\right)^T + \sigma_t^2 I$$

$$\Sigma_{t-1} = \frac{1-\bar{\alpha}_{t-1}-\sigma_t^2}{1-\bar{\alpha}_t}(1-\bar{\alpha}_t) I + \sigma_t^2 I$$

$$\Sigma_{t-1} = (1-\bar{\alpha}_{t-1}-\sigma_t^2) I + \sigma_t^2 I$$

$$\Sigma_{t-1} = (1-\bar{\alpha}_{t-1}) I$$

میانگین و کوواریانس محاسبه شده برای $q_\sigma(x_{t-1}|x_0)$ دقیقاً با فرمول مطلوب ما مطابقت دارد بنابراین استقراء ثابت شد

ب)

ابتدا elbo را می‌نویسیم

$$L_{ELBO} = E_{q_\sigma(x_{0:T})}\left[\log \frac{p_\theta(x_{0:T})}{q_\sigma(x_{1:T}|x_0)}\right]$$

$$L_{ELBO} = \underbrace{E_{q_\sigma}[\log p_\theta(x_0|x_1)]}_{L_0} + \underbrace{\sum_{t=2}^{T} E_{q_\sigma}\left[\log \frac{p_\theta(x_{t-1}|x_t)}{q_\sigma(x_{t-1}|x_t, x_0)}\right]}_{L_{t-1}} - \underbrace{E_{q_\sigma}[D_{KL}(q_\sigma(x_T|x_0)\|p_\theta(x_T))]}_{L_T}$$

ترم L_T به پارامترهای مدل (θ) وابسته نیست و در بهینه سازی نادیده گرفته می‌شود

حال می‌رویم سراغ L_{t-1}

$q_\sigma(x_{t-1}|x_t, x_0)$ که توزیع مطلوب است

$p_\theta(x_{t-1}|x_t)$ که همان $q_\sigma(x_{t-1}|x_t, f_\theta^{(t)}(x_t))$ است و توسط مدل پیش‌بینی می‌شود

واریانس این 2 توزیع گوسی یکسان و برابر با $\sigma_t^2 I$ است D_{KL} بین دو توزیع گوسی با کوواریانس یکسان متناسب با توان دوی فاصله میانگین‌هایشان است. میانگین این دو توزیع تنها در بخش مربوط به x_0 با هم تفاوت دارند (در یکی x_0 واقعی و در دیگری پیش‌بینی مدل یعنی $f_\theta^{(t)}(x_t)$ قرار دارد)

$$L_{t-1} \propto -E_{q_\sigma}\left[\|f_\theta^{(t)}(x_t) - x_0\|^2\right]$$

$$f_\theta^{(t)}(x_t) - x_0 = \frac{x_t - \sqrt{1-\bar\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\bar\alpha_t}} - \frac{x_t - \sqrt{1-\bar\alpha_t}\,\epsilon}{\sqrt{\bar\alpha_t}} = -\frac{\sqrt{1-\bar\alpha_t}}{\sqrt{\bar\alpha_t}}\left(\epsilon_\theta^{(t)} - \epsilon\right)$$

ما را این هر ترم همونه L_{t-1} مستقیما به حداقل سازی فاصله بین نویز واقعی و نویز پیش‌بینی شده توسط مدل تبدیل می‌شود

$$L_{total} = \min_\theta \sum_{t=1}^{T} \gamma_t E_{x_0, \epsilon}\left[\|\epsilon_\theta(\sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,\epsilon, t) - \epsilon\|^2\right]$$

این همان objective ساده شده در مدل‌های دیفیوژن استاندارد است

جواب پرسش،

این دیدگاه با صفر در نظر گرفتن واریانس ($\sigma_t \to 0$)، فرایند تولید نمونه را از حالت تصادفی به یک فرایند کاملاً قطعی تبدیل می‌کند. این فرایند قطعی همان معروف DDIm است.

مزیت اصلی این کار، افزایش سرعت چشمگیر است. از آنجایی که مسیر تولید قطعی است، دیگر نیازی به پیمودن تمام گام‌های زمانی (مثلاً 1000 گام) نیست. در عوض می‌توان با برداشتن گام‌های بزرگتر و پیمودن یک زیردنباله از مراحل، نمونه‌های را با سرعت بسیار بیشتری تولید کرد و هزینه زمانی را به شدت کاهش داد.