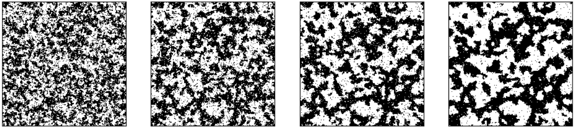
From Maximum Likelihood Principle, Variational Inference to Probabilistic Diffusion Models

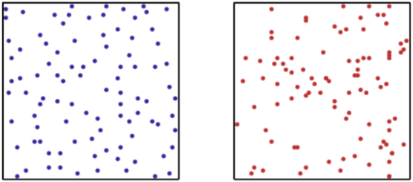
Jiaxi Zhao

23rd April, 2023

Sampling

Sampling is a fundamental computational task in lots of area, including applied mathematics, physics, and computer science.





DPP Independent

Maximum Likelihood Principle

In classical parametric statistics, suppose one has a statistical model $p(\mathbf{x};\theta)$ and a set of data samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, the maximum likelihood principle provide us with a natural estimator given by:

$$\widehat{\theta} = \arg\max_{\theta} \log \prod_{i=1}^{N} p(\mathbf{x}_i;\theta).$$

It is well-known that this estimator can also be viewed as the minimizer of the KL-divergence between the empirical measure and $p(\mathbf{x};\theta)$.

Real World Distribution

How about real world distribution? i.e. distribution of all the images, distribution of all the texts. How can we model these distributions and do sampling from them?

1. Classical non-parametric approach, e.g. mixture Gaussian model.
2. Bayesian inference, e.g. variational inference.
3. Generative adversarial network (GAN), energy-based method (Langevin dynamics), etc.

ELBO and Variational Inference

In order to estimate the unknown likelihood function $p(\mathbf{x})$, we introduce "Evidence lower bound" (ELBO)

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]$$

Here ϕ may be any statistical models, either parametric or non-parametric.

ELBO

To prove this, one has

$$
\begin{aligned}
\log p(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x})\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right] \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right],
\end{aligned}
$$

where the last inequality is based on the observation that $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$.

First generative model: Variational Autoencoder

Based on the simple idea of ELBO, one can introduce our first generative model in this series, i.e. variational autoencoder:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]$$
$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})).$$

Why such a structure is called autoencoder? Encoder: $q_\phi(\mathbf{z}|\mathbf{x})$, decoder: $p_\theta(\mathbf{x}|\mathbf{z})$.

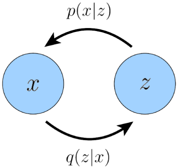


Figure: An illustration of VAE[1]

[1] Understanding Diffusion Models: A Unified Perspective, Calvin Luo

VAE

The remaining question is: how do we estimate the expectation of ELBO which is divided into two parts? We will pose following ansatz on the form of encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and prior $p(\mathbf{z})$:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})\mathbf{I}),$$
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}).$$

This will reduce D_{KL} term to an analytic form and we will calculate the first term via Monte Carlo sampling, i.e.

$$\arg\max_{\phi,\theta} \sum_{i=1}^{l} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),$$

where $\mathbf{z}^{(l)}$ is sampling from $\mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})\mathbf{I})$.

VAE

One more question appears: How do we do auto-differentiation in gradient-based optimization?

Here comes the reparameterization trick, by using following computational graph:

$$\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Parameters in ϕ is included in the calculation of loss function, therefore auto-differentiation becomes possible.

Second generative model: Hierarchical Variational Autoencoder

$$p(\mathbf{x}, \mathbf{z}_{1:T}) = p(\mathbf{z}_T)p(\mathbf{x}|\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_{t-1}|\mathbf{z}_t),$$

$$q_\phi(\mathbf{z}_{1:T}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x})\prod_{t=2}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}).$$

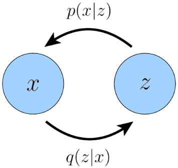


Figure: A Markovian Hierarchical Variational Autoencoder[3]with T hierarchical latents. The generative process is modeled as a Markov chain, where each latent $\mathbf{z}_t$ is generated only from the previous latent $\mathbf{z}_{t+1}$.

Third generative model: Variational Diffusion Model

We can finally go to the description of the variational diffusion model, one can think of VDM as a MHVAE with following requirement:

1. The latent dimension is exactly equal to the data dimension, we use $\mathbf{x}_{1:T}$ to denote latent variables.
2. The structure of the latent encoder at each time step is pre-defined as a linear Gaussian model, i.e. $q_\phi(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$.
3. $p(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Figure: Sampling along the trajectories of a diffusion model[4].

[4] Denoising Diffusion Probabilistic Models, Jonathan Ho et. al.

VDM

Based on this framework, one can calculate the ELBO as follows:

$$
\begin{aligned}
\log p(\mathbf{x}) &\geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \\
&= \mathbb{E}\left[\log \frac{p(\mathbf{x}_T)p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})}\right] + \mathbb{E}\left[\log \prod_{t=1}^{T-1} \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] + \mathbb{E}_{q(\mathbf{x}_T,\mathbf{x}_{T-1}|\mathbf{x}_0)}\left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})}\right] \\
&\quad + \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t+1},\mathbf{x}_{t-1},\mathbf{x}_t|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] - \mathbb{E}_{q(\mathbf{x}_{T-1}|\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_{T-1})||p(\mathbf{x}_T)) \\
&\quad - \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t+1},\mathbf{x}_{t-1}|\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1})||p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})).
\end{aligned}
$$

VDM

Let us have a brief interpretation of the terms in this expansion:

1. $\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right]$: reconstruction term, same as VAE.
2. $\mathbb{E}_{q(\mathbf{x}_{T-1}|\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_{T-1})||p(\mathbf{x}_T))$: prior matching term, no trainable parameters.
3. $\sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t+1},\mathbf{x}_{t-1}|\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1})||p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}))$: consistency term, related to reversibility.

VDM

Using Bayesian rule, one can rearrange the term to simplify the expectation we are dealt with:

$$\log p(\mathbf{x}) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] - D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) \\ - \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t+1},\mathbf{x}_0)} D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t+1},\mathbf{x}_0)||p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})).$$

The last term is now called denoising matching term.

VDM

Again the practical question is: How to optimize such a complicated objective function? We will significant reduce the calculation by leveraging the Gaussian transition assumption, i.e.

$$q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}.$$

Hence, it suffices to derive $q(\mathbf{x}_t|\mathbf{x}_0)$:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}\right) + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon.
\end{aligned}$$

VDM

Based on the previous calculation, one obtains

$$q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{\sqrt{\bar{\alpha}_t}(1-\alpha_t)\mathbf{x}_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_t)\mathbf{x}_t}{1-\bar{\alpha}_t}, \Sigma_q(t)\right),$$

where $\bar{\alpha} = \prod_{t=1}^{T} \alpha_t, \Sigma_q(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_t)}{1-\bar{\alpha}_t}\mathbf{I}$. Therefore, it makes sense to set

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_q(t)), \\ \mu_\theta(\mathbf{x}_t, t) &= \frac{\sqrt{\bar{\alpha}_t}(1-\alpha_t)\mathbf{x}_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_t)\mathbf{x}_\theta(\mathbf{x}_t, t)}{1-\bar{\alpha}_t}. \end{aligned}$$

Advantages of Diffusion models

Diffusion models have following advantages:

1. Comparing to GAN, it is more robust to mode collapse, successfully apply to multimodal distribution. Some combination with sampling technique in determinantal point process is also transplanted to diffusion models.

2. It processes solid theoretic foundation and shares lots of connection with energy-based sampler, score-matching sampler.

Disadvantages of Diffusion models

On the other hand, several disadvantages remain

1. Since one requires the prior for $\mathbf{x}_T$ to be standard Gaussian, the time horizon is usually taken to be a large number, which makes the training and sampling much more expansive than other methods such as GAN.

2. The accuracy of the sampling is not comparable to SOTA, i.e. various modification of GAN.

3. Currently, there is no satisfying theory on choosing the diffusion norse parameters α_t.

Reference

1. Luo, Calvin. "Understanding diffusion models: A unified perspective." arXiv preprint arXiv:2208.11970 (2022).

2. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

3. Tutorial on Denoising Diffusion-based Generative Modeling: Foundations and Applications: https://www.youtube.com/watch?v=cS6JQpEY9cs&t=1948s