Mathematics Behind Generative Diffusion Models

Presenter: Ye Yuan

2023.09.26

Catalog

- Generative Model Formulation
- Maximum Likelihood Estimation
- Preliminaries of Diffusion Models
- Evidence Lower Bound (ELBO)
- Optimization Objective
- Summary and Discussion

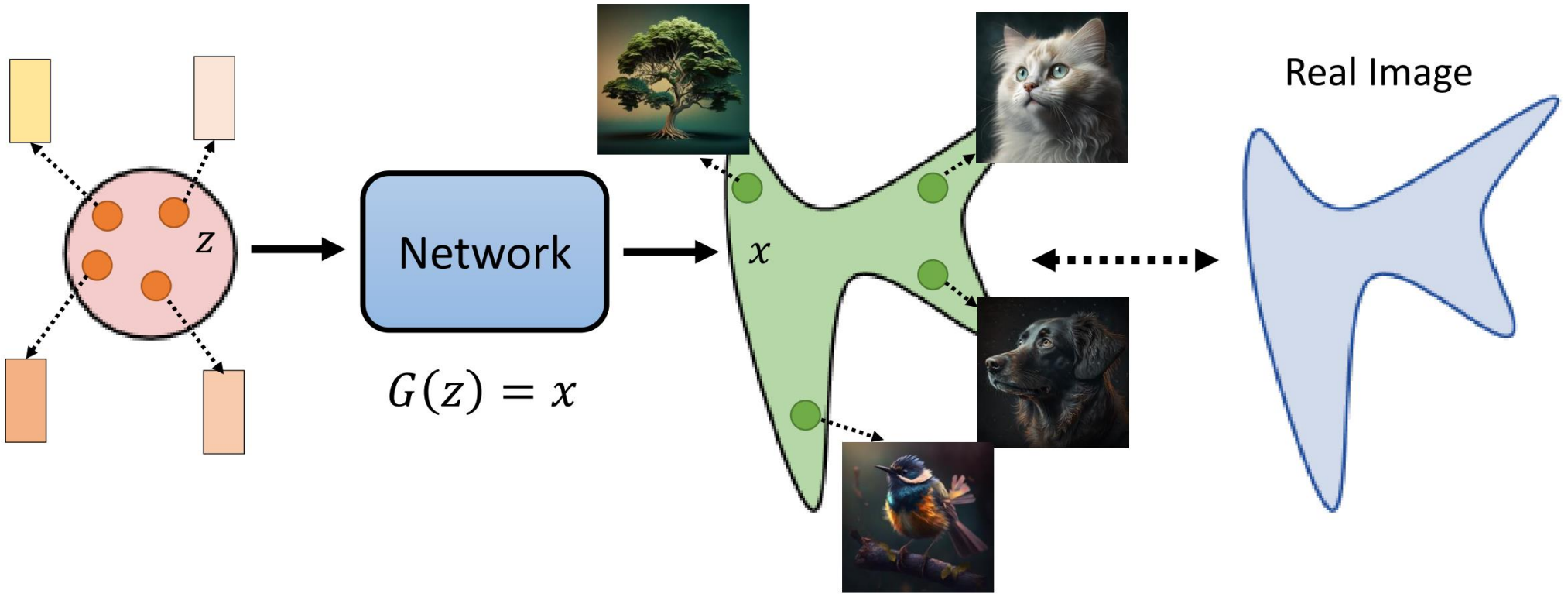
Generative Model Formulation

- Observed Data (x): the data we directly observed
- Latent Variable (z): a vector captures the features of the observed data.

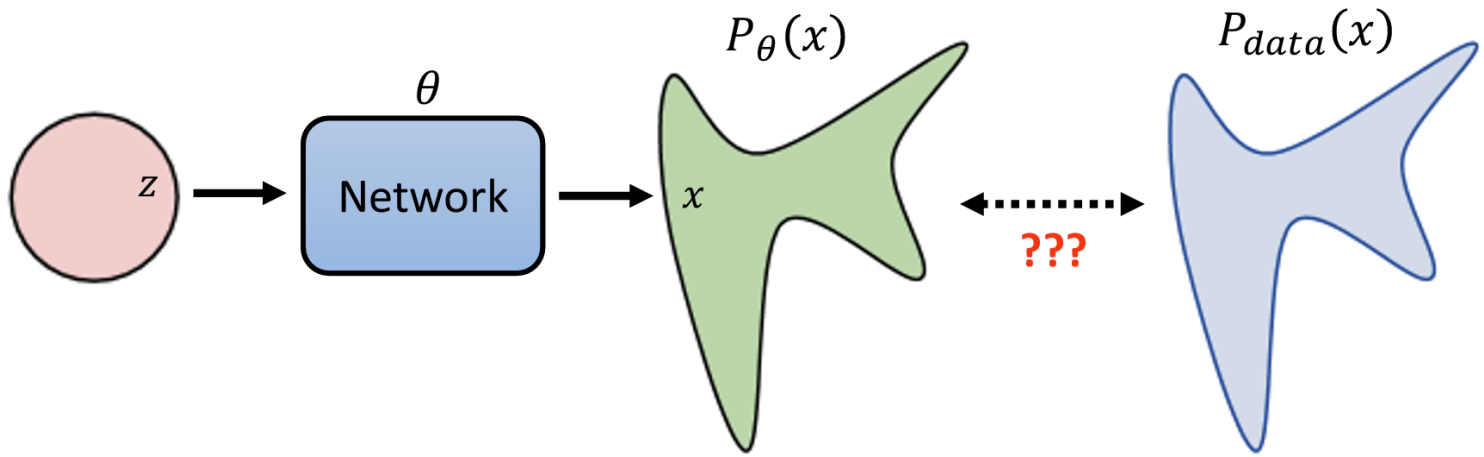
- Assumption: The observed data are generated by some latent variables.
- Mathematically, the observed data and the latent variables can be model as a joint distribution p(x, z).

Generative Model Formulation

- $P_{data}(x)$: probability distribution of the data
- $P_\theta(x)$: approximate probability distribution of the data
- P(z): probability distribution of the latent variable, usually a Gaussian distribution



Maximum Likelihood Estimation



Sample $\{x^1, x^2, \ldots, x^m\}$ from $P_{data}(x)$

$$\theta^* = arg\max_\theta \prod_{i=1}^{m} P_\theta(x^i)$$

Maximum Likelihood Estimation

Sample $\{x^1, x^2, \dots, x^m\}$ from $P_{data}(x)$

$$\theta^* = arg\max_\theta \prod_{i=1}^{m} P_\theta(x^i) = arg\max_\theta log \prod_{i=1}^{m} P_\theta(x^i)$$

$$= arg\max_\theta \sum_{i=1}^{m} logP_\theta(x^i) \approx arg\max_\theta E_{x\sim P_{data}}[logP_\theta(x)]$$

$$= arg\max_\theta \int_x P_{data}(x)logP_\theta(x)dx - \int_x P_{data}(x)logP_{data}(x)dx$$

(not related to θ)

$$= arg\max_\theta \int_x P_{data}(x)log\frac{P_\theta(x)}{P_{data}(x)}dx = arg\min_\theta KL(P_{data}||P_\theta)$$

Difference between P_{data} and P_θ

Maximum Likelihood = Minimize KL Divergence

Computational Intractability of Maximum Likelihood Estimation

- The marginal likelihood of the observed data x with respect to the model parameters θ is given by:

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

$$= \int p_\theta(x \mid z)\, p(z)\, dz$$

- Challenges:
  - **High Dimensionality:** the latent space z might be high-dimensional.
  - **Non-linearity & Model Parameters:** usually involves passing z through non-linear transformations defined by the parameters θ. This increases the complexity of the integral.
  - **Parameter Learning:** On top of computing the integral, we often want to learn the best set of parameters θ that maximizes the likelihood of the observed data.

Maximum Likelihood Estimation

- In practice, we usually maximize the log likelihood.
- Sometimes, the log likelihood is also called evidence.
- We can maximize the ELBO instead.

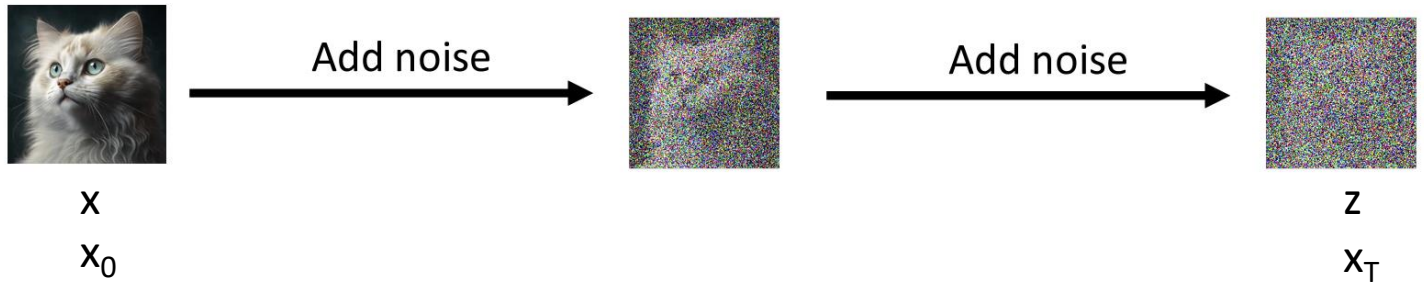
$$
\begin{aligned}
\log p(x) &= \log p(x) \int q_{\phi}(z|x) dz \\
&= \int q_{\phi}(z|x)(\log p(x)) dz \\
&= \mathop{\mathbb{E}}_{q_{\phi}(z|x)} [\log p(x)] \\
&= \mathop{\mathbb{E}}_{q_{\phi}(z|x)} \left[\log \frac{p(x,z)}{p(z|x)} \right] \\
&= \mathop{\mathbb{E}}_{q_{\phi}(z|x)} \left[\log \frac{p(x,z)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

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

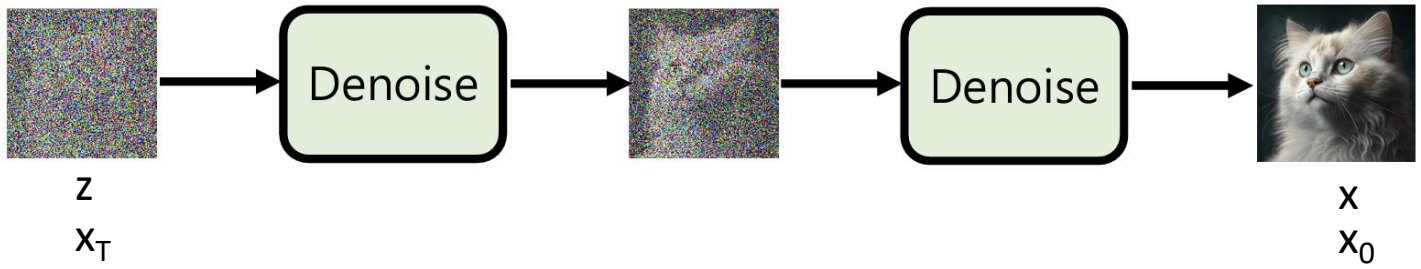
Evidence Lower Bound (ELBO)

Denoising Diffusion Models

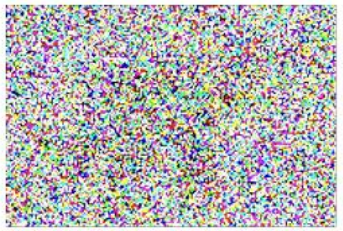
- Diffusion Process (Forward Process)



- Denoising Process (Backward Process)



Diffusion Process

$\sqrt{a_t}$  $+ \sqrt{1 - a_t}$  $=$ 

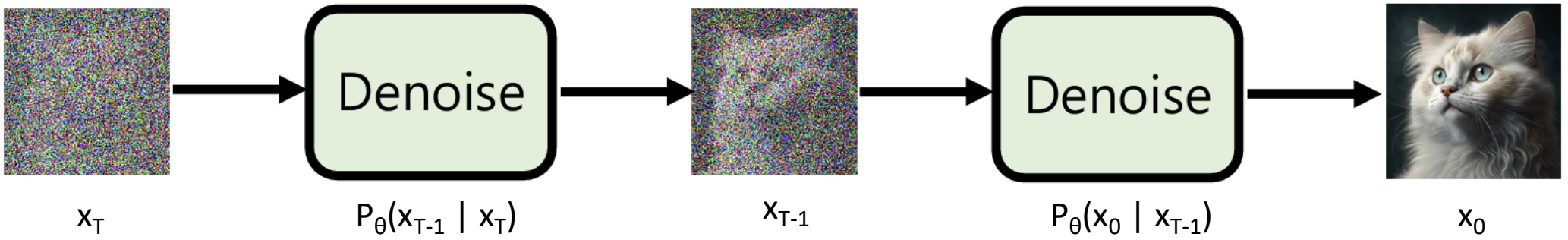
x_{t-1} ε x_t

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$$

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

Denoising Process



$$p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

$$p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$$

ELBO in Diffusion Model

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$$

$$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$$

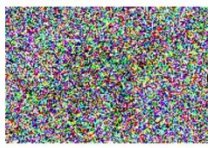
$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] - D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right].$$

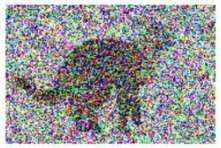
$\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]$: a reconstruction term, predicting the log probability of the original data sample given the first-step latent.

$D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$: represents how close the distribution of the final noisified input is to the standard Gaussian prior.

$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right]$: a denoising matching term. The $q(x_{t-1}|x_t, x_0)$ defines how to denoise a noisy image x_t with access to what the final, completely denoised image x_0 should be.



x_t



x_{t-1}

x_0

ELBO in Diffusion Model

- ELBO = $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] - D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right].$$

- $\theta^* = argmax$ $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] - D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right].$$

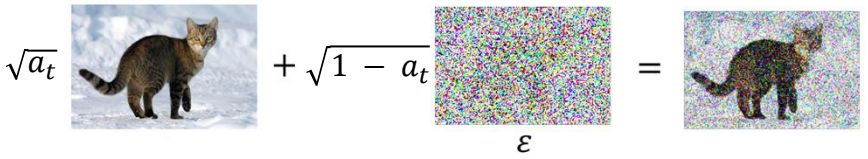
$q(x_t \mid x_0)$

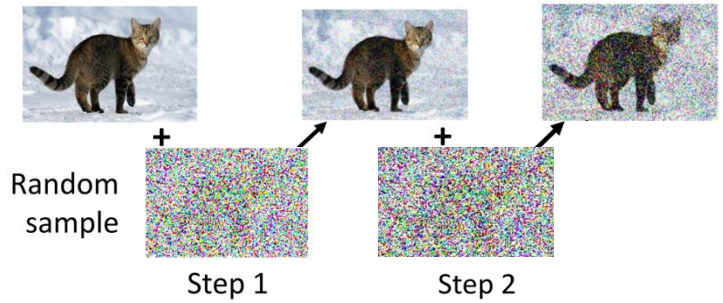
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$$

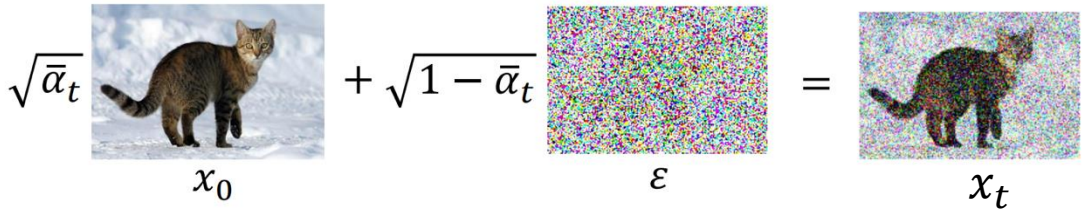
$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \dots \\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \\
&\sim \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\,\mathbf{I})
\end{aligned}$$

We assume:





Actually:



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

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) \\
\sigma_q^2(t) &= \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}
\end{aligned}
$$

Optimization Objective

- $\theta^* = argmax\ \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] - D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right].$$

- $\theta^* = argmin\ \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]$
- KL Divergence between two Gaussian Distribution:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \parallel \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x) + (\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)\right]$$

Optimization Objective

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\|_2^2\right]$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left[\|\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2\right]$$

We finally derive the loss function in DDPM [1].

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \quad \Rightarrow \quad \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0$$

Questions for Audience

Question1

1. We first claim that the generative diffusion model is likelihood-based, which can be optimized by maximum likelihood estimation.
2. We prove that we can maximize the ELBO as a proxy goal.
3. We derive the ELBO for diffusion models.
4. We conclude the same loss function in DDPM by some trivial substitutions and rearrangements.

Question2

1. We first claim that the generative diffusion model is likelihood-based, which can be optimized by maximum likelihood estimation.
2. We prove that we can maximize the ELBO as a proxy goal.
3. We derive the ELBO for diffusion models.
4. We conclude the same loss function in DDPM by some trivial substitutions and rearrangements.

Question3

1. We first claim that the generative diffusion model is likelihood-based, which can be optimized by maximum likelihood estimation.
2. We prove that we can maximize the ELBO as a proxy goal.
3. We derive the ELBO for diffusion models.
4. We conclude the same loss function in DDPM by some trivial substitutions and rearrangements.

Summary

1. We first claim that the generative diffusion model is likelihood-based, which can be optimized by maximum likelihood estimation.
2. We prove that we can maximize the ELBO as a proxy goal.
3. We derive the ELBO for diffusion models.
4. We conclude the same loss function in DDPM by some trivial substitutions and rearrangements.

Discussion

- Main take away:
  - $q(x_t|x_0)$ and $q(x_{t-1}|x_t, x_0)$ are important. See[5].
  - To derive different versions of $\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0)$, we have different formats but mathematically equivalent loss functions.

- Some arguments regarding diffusion models:
  - Humans don't generate images in this way.
  - The latents are restricted to the same dimensionality as the original input. (somehow mitigated by the latent diffusion model [6])
  - Sampling is an expensive procedure, as multiple denoising steps must be run. (can be mitigated by faster sampling[7] or distillation[8].)

Reference

1. Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (2020).
2. Luo, C. Understanding diffusion models: A unified perspective, 2022.
3. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning,ICML 2015 (2015).
4. Song, Y., and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems (2019).
5. Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces, 2023.
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022.
7. Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models, 2023.
8. Meng, C., Rombach, R., Gao, R., Kingma, D. P., Ermon, S., Ho, J., and Salimans, T. On distillation of guided diffusion models, 2023.

Thanks

- Questions and comments?