CS 8501 Advanced Topics in Machine Learning

Lecture 13: Diffusion Models

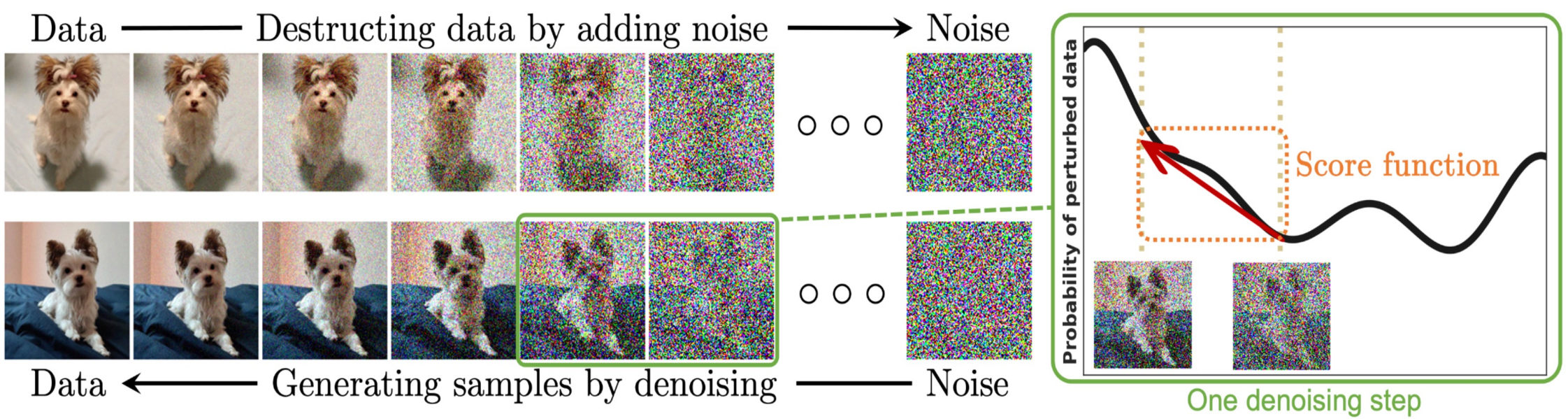
Yangfeng Ji
Information and Language Processing Lab
Department of Computer Science
University of Virginia
https://yangfengji.net/

Diffusion Processes

- Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. 2015
- Yang et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. 2022

Illustration

Diffusion models smoothly perturb data by adding noise, then reverse this process to generate new data from noise.



Each denoising step in the reverse process typically requires estimating the score function.

Forward Trajectory

Diffusion process:

$$\pi(y) = \int T_\pi(y|y';\beta)\pi(y')dy'$$

Diffusion kernel:

$$q(x^{(t)}|x^{(t-1)}) = T_\pi(x^{(t)}|x^{(t-1)};\beta_t)$$

where β_t is the diffusion rate

Binomial Diffusion Process

The forward trajectory is defined as

$$q(x^{0\dots T}) = q(x^{(0)}) \prod_{t=1}^{T} q(x^{(t)}|x^{(t-1)})$$

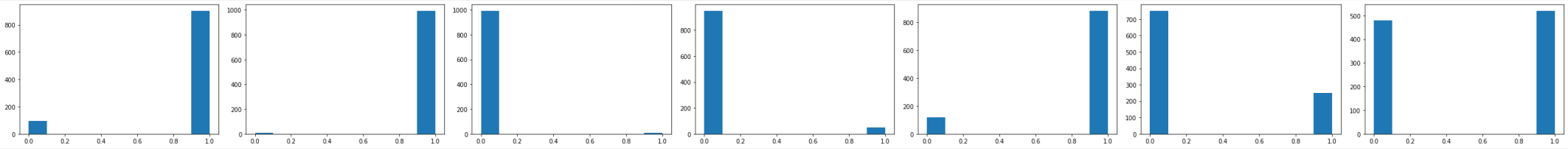
with

- $\pi(x^{(T)}) = \mathcal{B}(x^{(T)}; 0.5)$
- $q(x^{(t)}|x^{(t-1)}) = \mathcal{B}(x^{(t)}; (1-\beta_t)x^{(t-1)} + 0.5\beta_t)$ with $\beta = \frac{1}{T-t+1}$

Binomial Diffusion Process (II)

Consider a very simple example

- $q(x^{(0)}) = \text{Bern}(x^{(0)}; 0.9)$
- $q(x^{(t)}|x^{(t-1)}) = \mathcal{B}(x^{(t)}; (1-\beta_t)x^{(t-1)} + 0.5\beta_t)$
- $\beta = \frac{1}{T-t+1}$
- $T = 100$
- Sample size: 1000 at each time step



$t = 0, 10, 50, 90, 98, 100$

Binomial Diffusion Process (III)

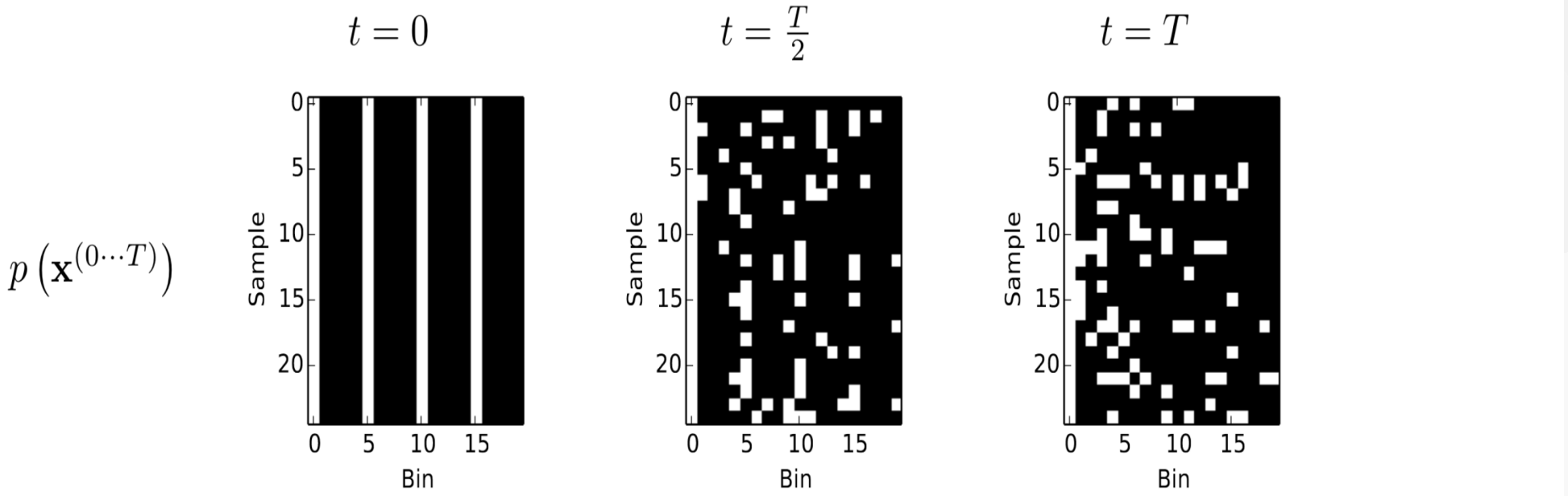


Figure 2. Binary sequence learning via binomial diffusion. A binomial diffusion model was trained on binary 'heartbeat' data, where a pulse occurs every 5th bin. Generated samples (left) are identical to the training data. The sampling procedure consists of initialization at independent binomial noise (right), which is then transformed into the data distribution by a binomial diffusion process, with trained bit flip probabilities. Each row contains an independent sample. For ease of visualization, all samples have been shifted so that a pulse occurs in the first column. In the raw sequence data, the first pulse is uniformly distributed over the first five bins.

Gaussian Diffusion Process

The joint distribution is defined as

$$q(x^{0\dots T}) = q(x^{(0)}) \prod_{t=1}^{T} q(x^{(t)}|x^{(t-1)})$$

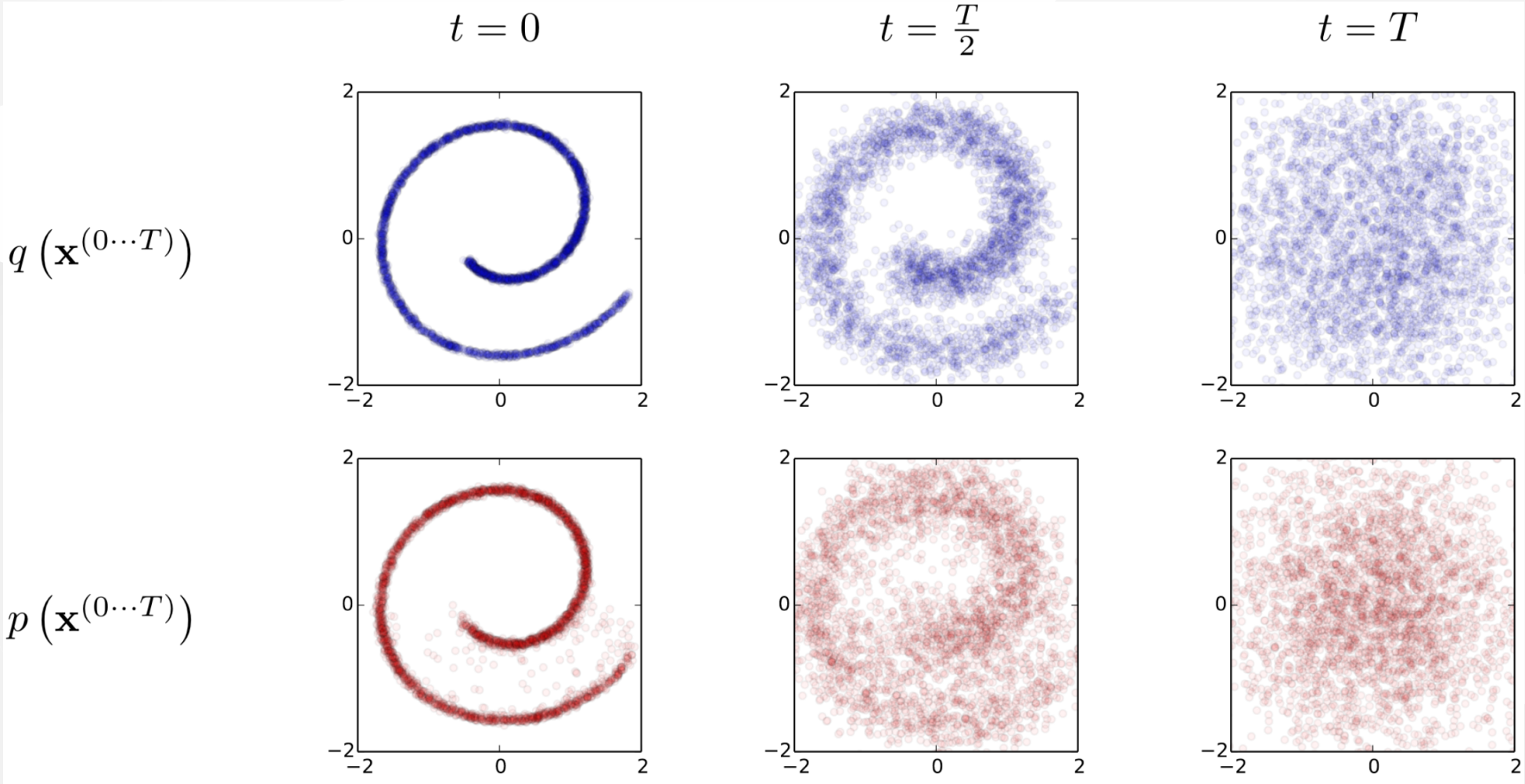
with

- $q^{(T)} = \mathcal{N}(x^{(T)}; 0, I)$
- $q(x^{(t)}|x^{(t-1)}) = \mathcal{N}(x^{(t)}; \sqrt{1-\beta_t}x^{(t-1)}, \beta_t I)$

When $\beta_t \to 1$, we have

$$q(x^{(t)}) \to \mathcal{N}(0, I)$$

Gaussian Diffusion Process (II)



Backward Trajectory

The central problem targetted by the backward trajectory is estimate the distribution $x^{(t-1)}$ given the whole (forward) diffusion process

$$q(x^{(0,\ldots,T)}) : x^{(0)} \rightarrow x^{(1)} \rightarrow \cdots \rightarrow x^{(T-1)} \rightarrow x^{(T)}$$

Conceptually, we have

$$q(x^{(t-1)}|x^{(0,\ldots,t-2)}, x^{(t,\ldots,T)}) =$$

Backward Trajectory (II)

We use another distribution p to approximate the backward process

$$p(x^{(T)}) = \pi(x^{(T)})$$

$$p(x^{(0\dots T)}) = p(x^{(T)}) \prod_{t=1}^{T} p(x^{(t-1)}|x^{(t)})$$

where

$$p(x^{(t-1)}|x^{(t)}) = \mathcal{N}(x^{(t-1)}; f_\mu(x^{(t)}, t), f_\Sigma(x^{(t)}, t))$$

Training

Original objective function

$$
\begin{aligned}
L &= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \\
&= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \\
&\qquad \log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} | \mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}\right)} \end{array} \right]
\end{aligned}
$$

The same trick has been used in variational inference

Training (II)

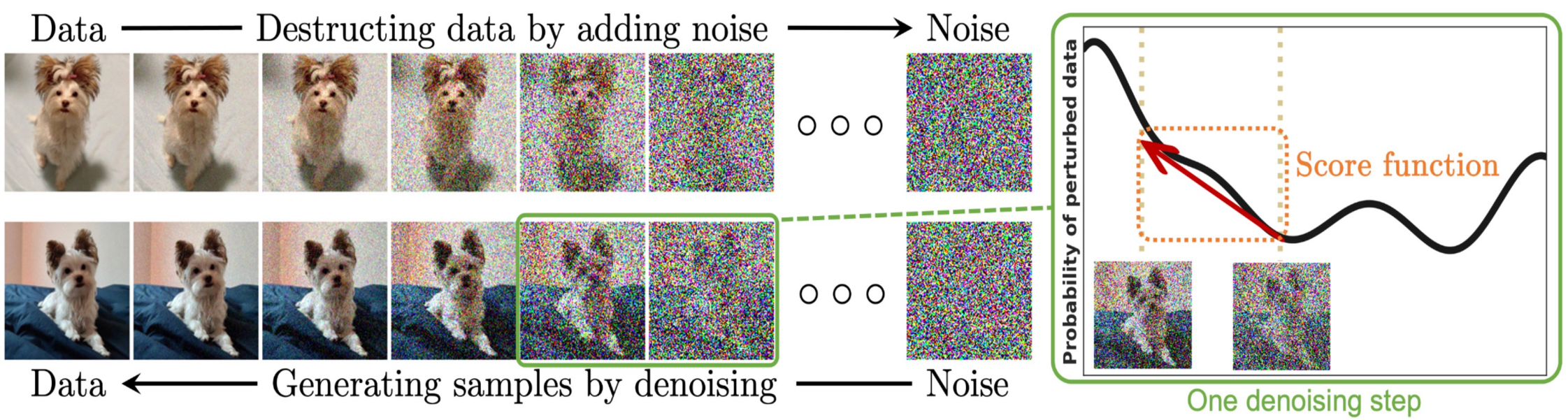
Its variational lower bound K

$$K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}\right) \cdot \\ D_{KL}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right) \middle|\middle| p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right) \\ + H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right).$$

When $q(x^{(t-1)}|x^{(t)}, x^{(0)})$ is given (aka, the diffusion process is fixed), each reverse diffusion step can be estimated independently

[Sohl-Dickstein et al., 2015]

Illustration



Linear Gaussian Systems

- Bishop. Pattern Recognition and Machine Learning. 2006
  - Chapter 02

Multivariate Gaussian

For T-dimensional random vector x, the multivariate Gaussian distribution is defined as

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{T/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

where μ is the mean vector and Σ is the covariance matrix.

Sometimes, we use $\Lambda = \Sigma^{-1}$ instead of Σ for convenience.

Conditional Distribution

Consider the following

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}; \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}; \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

where $\Sigma_{ab} = \Sigma_{ba}^{\top}$

The conditional distribution

$$p(x_a|x_b) = \mathcal{N}(x_a; \mu_{a|b}, \Sigma_{a|b})$$

- $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$
- $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

Marginal Distribution

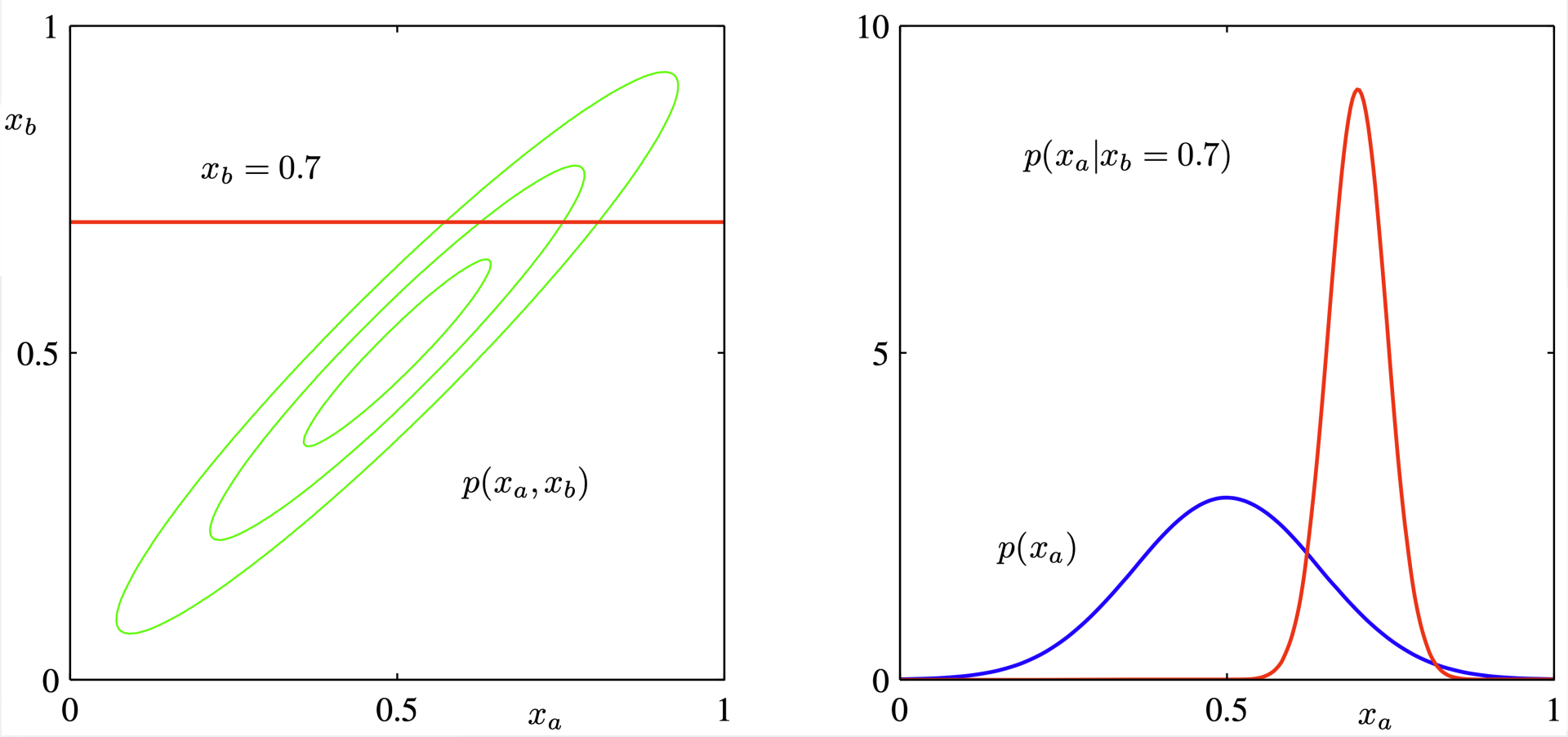
The marginal distribution x_a is given by

$$p(x_a) = \int p(x_a, x_b) dx_b$$

More explicitly, we have

$$p(x_a) = \mathcal{N}(x_a; \mu_a, \Sigma_{aa})$$

Illustration



High-level Message

The marginal/conditional distribution of a multivariate Gaussian is still a Gaussian

- This property will be heavily exploited in the following section.

Denoising Diffusion Model

- Ho et al. Denoising Diffusion Probabilistic Models. 2020

Diffusion Process, Revisited

Starting from $q(x_0)$, the diffusion process can be described as the following joint distribution

$$q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

with each component defined as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Multivariate Gaussian

Given

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

we have

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I)$$

with

- $\alpha_t = 1 - \beta_t$
- $\overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$
- When T is sufficiently large, $q(x_t|x_0) \approx \mathcal{N}(x_t; 0, I)$

Objective Function

Reconsider the following component in the objective function

$$D_{KL}\left(q(x^{(t-1)}|x^{(t)}, x^{(0)})||p_\theta(x^{(t-1)}|x^{(t)})\right)$$

- $q(x^{(t-1)}|x^{(t)}, x^{(0)})$: the posterior distribution $x^{(t-1)}$ given the whole diffusion process
- $p_\theta(x^{(t-1)}|x^{(t)})$: the approximation distribution to reverse the diffusion process

Posterior Distribution

$q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)$ again is a Gaussian distribution

$$q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) = \mathcal{N}\left(x^{(t-1)}; \tilde{\mu}_t\left(x^{(t)}, x^{(0)}\right), \tilde{\beta} I\right)$$

where

$$\tilde{\mu}_t\left(x^{(t)}, x^{(0)}\right) = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} x^{(0)} + \frac{\sqrt{\alpha_t}\left(1-\overline{\alpha}_{t-1}\right)}{1-\overline{\alpha}_t} x^{(t)}$$

$$\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$$

[Ho et al., 2020]

Approximation Distribution

In [Ho et al., 2020], $p_\theta(x^{(t-1)}|x^{(t)})$ is defined as

$$p_\theta(x^{(t-1)}|x^{(t)}) = \mathcal{N}(x^{(t-1)}; \mu_\theta(x^{(t)}, t), \Sigma_\theta(x^{(t)}, t))$$

with $\Sigma_\theta(x^{(t)}, t) = \beta_t I$

Minimize the KL divergence is reduced as

$$E_q\left[\frac{1}{2\beta_t}\|\tilde{\mu}_t(x^{(t)}, x^{(0)}) - \mu_\theta(x^{(t)}, t)\|\right]$$

which is a score function

Further Discussion

- Kreis et al. Denoising Diffusion-based Generative Modeling: Foundations and Applications. 2022
- Yang et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. 2022

Thank You!