

# Deep Generative Models

## Lecture 12

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

## Recap of previous lecture

### Forward gaussian diffusion process

Let  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ,  $\beta_t \in (0, 1)$ ,  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ .

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

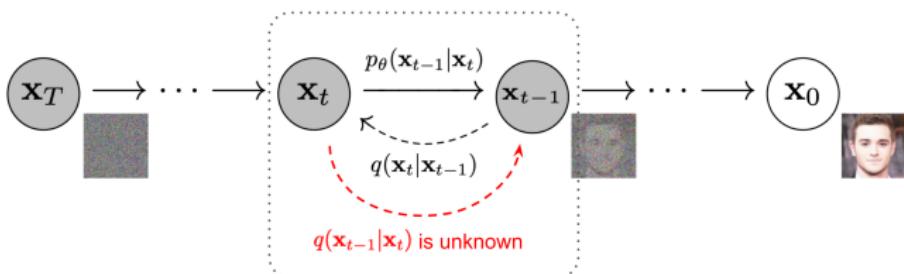
$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t | \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \cdot \mathbf{I});$$

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0, (1 - \bar{\alpha}_t) \cdot \mathbf{I}).$$

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;
2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon}$ , where  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ ,  $t \geq 1$ ;
3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, 1)$ , where  $T \gg 1$ .

If we are able to invert this process, we will get the way to sample  $\mathbf{x} \sim \pi(\mathbf{x})$  using noise samples  $p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

## Recap of previous lecture



### Reverse gaussian diffusion process

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

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} = \mathcal{N}(\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

- ▶  $q(\mathbf{x}_{t-1})$ ,  $q(\mathbf{x}_t)$  are intractable.
- ▶ If  $\beta_t$  is small enough,  $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$  will be Gaussian (Feller, 1949).
- ▶  $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$  defines how to denoise a noisy image  $\mathbf{x}_t$  with access to the completely denoised image  $\mathbf{x}_0$ .

# Recap of previous lecture

Let's define the reverse process

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t) \approx p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t))$$

Forward process

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon}$ ,  
where  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ ,  $t \geq 1$ ;

3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

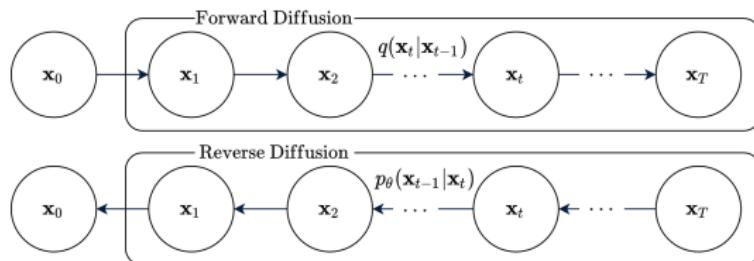
Reverse process

1.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ ;

2.  $\mathbf{x}_{t-1} = \sigma_\theta(\mathbf{x}_t, t) \cdot \boldsymbol{\epsilon} + \mu_\theta(\mathbf{x}_t, t)$ ;

3.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

Gaussian diffusion model as VAE



## Recap of previous lecture

- ▶  $\mathbf{z} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$  is a latent variable.
- ▶ Variational posterior distribution

$$q(\mathbf{z}|\mathbf{x}) = q(\mathbf{x}_1, \dots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}).$$

- ▶ Generative distribution and prior

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}); \quad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{t=2}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}) \cdot p(\mathbf{x}_T)$$

## ELBO

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})} = \mathcal{L}(q, \boldsymbol{\theta}) \rightarrow \max_{q, \boldsymbol{\theta}}$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}) - \textcolor{violet}{KL}(q(\mathbf{x}_T|\mathbf{x}_0) || p(\mathbf{x}_T)) - \\ &\quad - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \textcolor{violet}{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}))}_{\mathcal{L}_t} \end{aligned}$$

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. Langevin dynamic and SDE basics

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. Langevin dynamic and SDE basics

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. Langevin dynamic and SDE basics

# Reparametrization of DDPM

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))$$

$\mathcal{L}_t$  is the mean of KL between two normal distributions:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1} | \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$$

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t) \mathbf{I})$$

Here

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

$$\tilde{\beta}_t = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} = \text{const.}$$

Let assume

$$\sigma_\theta^2(\mathbf{x}_t, t) = \tilde{\beta}_t \mathbf{I} \quad \Rightarrow \quad p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I}).$$

# Reparametrization of DDPM

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1} | \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I});$$
$$p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I}).$$

Use the formula for KL between two normal distributions:

$$\begin{aligned}\mathcal{L}_t &= \mathbb{E}_{q(\mathbf{x}_t | \mathbf{x}_0)} KL\left(\mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) || \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I})\right) \\ &= \mathbb{E}_{q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \frac{1}{2\tilde{\beta}_t} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\|^2 \right]\end{aligned}$$

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

$$\begin{aligned}\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \mathbf{x}_0 \\ &= \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \boldsymbol{\epsilon}\end{aligned}$$

# Reparametrization of DDPM

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \left[ \frac{1}{2\tilde{\beta}_t} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\|^2 \right]$$

## Reparametrization

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon$$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon_\theta(\mathbf{x}_t, t)$$

$$\mathcal{L}_t = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right]$$

$$= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right]$$

At each step of reverse diffusion process we try to predict the noise  $\epsilon$  that we used in the forward diffusion process!

# Reparametrization of DDPM

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - KL(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) - \\ - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))}_{\mathcal{L}_t}$$

$$\mathcal{L}_t = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right]$$

Simplified objective

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t \sim U[2, T]} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2$$

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. Langevin dynamic and SDE basics

# Denoising diffusion probabilistic model (DDPM)

DDPM is a VAE model

- ▶ Encoder is a fixed Gaussian Markov chain  $q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)$ .
- ▶ Latent variable is a hierarchical (in each step the dim. of the latent equals to the dim of the input).
- ▶ Decoder is a simple Gaussian model  $p(\mathbf{x}_0 | \mathbf{x}_1, \theta)$ .
- ▶ Prior distribution is given by parametric Gaussian Makov chain  $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta)$ .

Forward process

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;
2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon}$ ,  
where  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ ,  $t \geq 1$ ;
3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

Reverse process

1.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ ;
2.  $\mathbf{x}_{t-1} = \sigma_\theta(\mathbf{x}_t, t) \cdot \boldsymbol{\epsilon} + \mu_\theta(\mathbf{x}_t, t)$ ;
3.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

# Denoising diffusion probabilistic model (DDPM)

## Training

1. Get the sample  $\mathbf{x}_0 \sim \pi(\mathbf{x})$ .
2. Sample timestamp  $t \sim U[1, T]$  and the noise  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .
3. Get noisy image  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ .
4. Compute loss  $\mathcal{L}_{\text{simple}} = \|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t)\|^2$ .

## Sampling

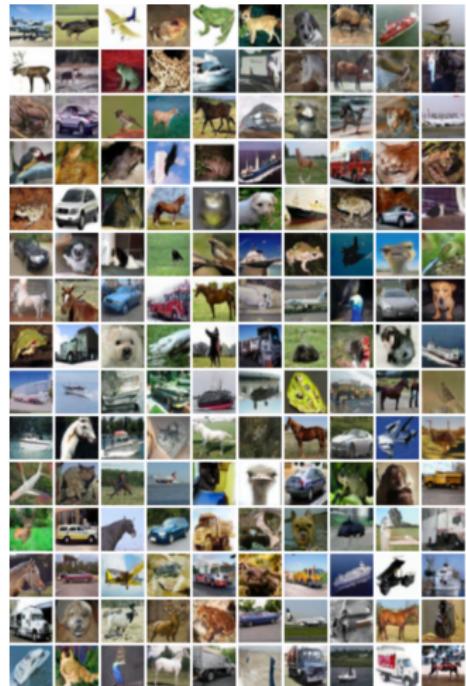
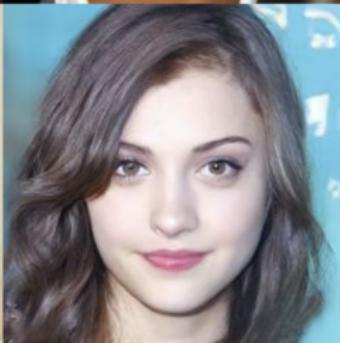
1. Sample  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ .
2. Compute mean of  $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mu_{\theta}(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I})$ :

$$\mu_{\theta}(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon_{\theta}(\mathbf{x}_t, t)$$

3. Get denoised image  $\mathbf{x}_{t-1} = \mu_{\theta}(\mathbf{x}_t, t) + \sqrt{\tilde{\beta}_t} \cdot \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .

# Denoising diffusion probabilistic model (DDPM)

## Samples



# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. Langevin dynamic and SDE basics

## Langevin dynamic

Imagine that we have some generative model  $p(\mathbf{x}|\theta)$ .

### Statement

Let  $\mathbf{x}_0$  be a random vector. Then under mild regularity conditions for small enough  $\eta$  samples from the following dynamics

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{1}{2} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \theta) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

will comes from  $p(\mathbf{x}|\theta)$ .

What do we get if  $\boldsymbol{\epsilon} = \mathbf{0}$ ?

### Energy-based model

$$p(\mathbf{x}|\theta) = \frac{\hat{p}(\mathbf{x}|\theta)}{Z_\theta}, \quad \text{where } Z_\theta = \int \hat{p}(\mathbf{x}|\theta) d\mathbf{x}$$

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\theta) - \nabla_{\mathbf{x}} \log Z_\theta = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\theta)$$

Gradient of normalized density equals to gradient of unnormalized density.

## Stochastic differential equation (SDE)

Let define stochastic process  $\mathbf{x}(t)$  with initial condition  $\mathbf{x}(0) \sim p_0(\mathbf{x})$ :

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$$

- ▶  $\mathbf{f}(\mathbf{x}, t)$  is the **drift** function of  $\mathbf{x}(t)$ .
- ▶  $g(t)$  is the **diffusion** coefficient of  $\mathbf{x}(t)$ .
- ▶ If  $g(t) = 0$  we get standard ODE.
- ▶  $\mathbf{w}(t)$  is the standard Wiener process (Brownian motion):
  1.  $\mathbf{w}(0) = 0$  (almost surely);
  2.  $\mathbf{w}(t)$  has independent increments;
  3.  $\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, (t-s)\mathbf{I})$ .
- ▶  $d\mathbf{w} = \mathbf{w}(t+dt) - \mathbf{w}(t) = \mathcal{N}(0, \mathbf{I} \cdot dt) = \epsilon \cdot \sqrt{dt}$ , where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .

**Note:** In contrast to ODE, initial condition  $\mathbf{x}(0)$  does not uniquely determine the process trajectory.

# Stochastic differential equation (SDE)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad d\mathbf{w} = \boldsymbol{\epsilon} \cdot \sqrt{dt}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

- ▶ At each moment  $t$  we have the density  $p(\mathbf{x}(t), t)$ .
- ▶ How to get distribution  $p(\mathbf{x}, t)$  for  $\mathbf{x}(t)$ ?

## Theorem (Kolmogorov-Fokker-Planck)

Evolution of the distribution  $p(\mathbf{x}, t)$  is given by the following ODE:

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = \text{tr} \left( -\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t)p(\mathbf{x}, t)] + \frac{1}{2}g^2(t)\frac{\partial^2 p(\mathbf{x}, t)}{\partial \mathbf{x}^2} \right)$$

**Note:** This is the generalization of KFP theorem that we used in continuous-in-time NF.

## Langevin SDE (special case)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$$

$$d\mathbf{x} = \frac{1}{2}\frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}, t)dt + \mathbf{1} \cdot d\mathbf{w}$$

## Langevin SDE (special case)

$$d\mathbf{x} = \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}, t) dt + \mathbf{1} \cdot d\mathbf{w}$$

Let apply KFP theorem.

$$\begin{aligned}\frac{\partial p(\mathbf{x}, t)}{\partial t} &= \text{tr} \left( -\frac{\partial}{\partial \mathbf{x}} \left[ p(\mathbf{x}, t) \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}, t) \right] + \frac{1}{2} \frac{\partial^2 p(\mathbf{x}, t)}{\partial \mathbf{x}^2} \right) = \\ &= \text{tr} \left( -\frac{\partial}{\partial \mathbf{x}} \left[ \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} p(\mathbf{x}, t) \right] + \frac{1}{2} \frac{\partial^2 p(\mathbf{x}, t)}{\partial \mathbf{x}^2} \right) = 0\end{aligned}$$

The density  $p(\mathbf{x}, t) = \text{const}(t)!$

## Discretized Langevin SDE

$$\mathbf{x}_{t+1} - \mathbf{x}_t = \eta \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}, t) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \eta \approx dt.$$

## Langevin dynamic

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{1}{2} \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \eta \approx dt.$$

# Stochastic differential equation (SDE)

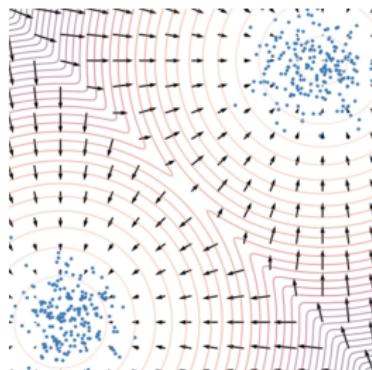
## Statement

Let  $\mathbf{x}_0$  be a random vector. Then samples from the following dynamics

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{1}{2} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \theta) + \sqrt{\eta} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

will come from  $p(\mathbf{x} | \theta)$  under mild regularity conditions for small enough  $\eta$  and large enough  $t$ .

The density  $p(\mathbf{x} | \theta)$  is a **stationary** distribution for this SDE.



## Summary

- ▶ DDPM is a VAE model with hierarchical latent variables.
- ▶ At each step DDPM predicts the noise that was used in the forward diffusion process.
- ▶ DDPM is really slow, because we have to apply the model  $T$  times.
- ▶ Langevin dynamics allows to sample from the energy-based model using the score function (due to the existence of stationary distribution of SDE).