Диффузные модели, часть 1
дискретное время

Григорий Бартош

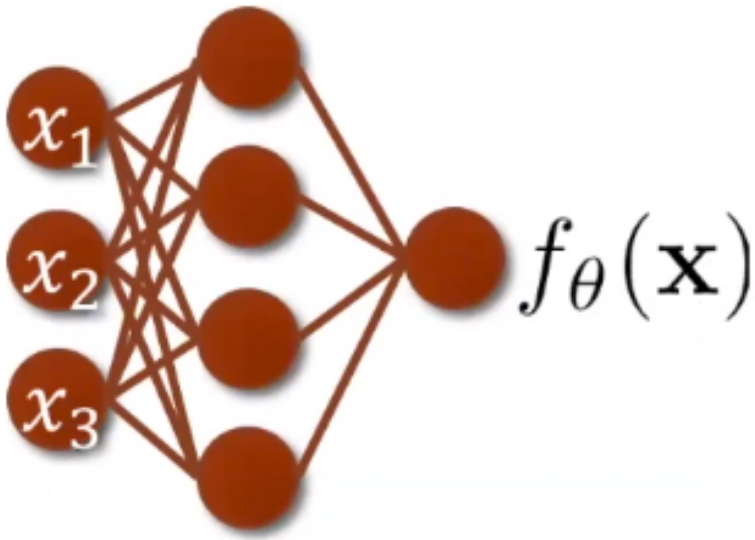
8 октября 2021

Energy-Based Models

- Energy-Based Models (EBMs)

$$p_\theta(\mathrm{x}) = \frac{e^{-f_\theta(\mathrm{x})}}{Z_\theta}, \quad f_\theta(\mathrm{x}) \in \mathbb{R}$$

$$Z_\theta = \int e^{-f_\theta(\mathrm{x})} \mathrm{dx}$$



- **Плюсы** Простая модель
- **Минусы** Сложно обучать θ через максимизацию правдоподобия

$$\mathbb{E}_{p_{\text{data}}(\mathrm{x})}[-\log p_\theta(\mathrm{x})] = \mathbb{E}_{p_{\text{data}}(\mathrm{x})}[f_\theta(\mathrm{x}) + \log Z_\theta]$$

Обучение EBMs через score matching

- Score function

$$s(\mathrm{x}) = \nabla_{\mathrm{x}} \log p(\mathrm{x})$$

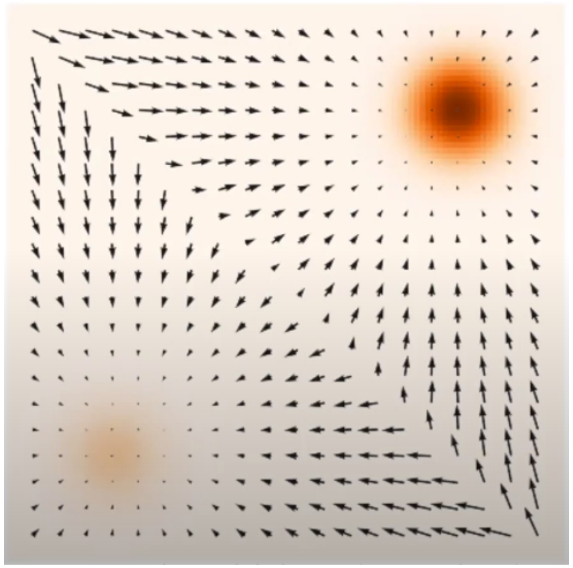
- Score function для EBM модели

$$\mathrm{s}_\theta(\mathrm{x}) = \nabla_{\mathrm{x}} \log p_\theta(\mathrm{x}) = -\nabla_{\mathrm{x}} f_\theta(\mathrm{x}) - \underbrace{\nabla_{\mathrm{x}} \log Z_\theta}_{=0} = -\nabla_{\mathrm{x}} f_\theta(\mathrm{x})$$

- Score matching (Fisher divergence)

$$\mathbb{E}_{p_{\text{data}}(\mathrm{x})}[\|\mathrm{s}_\theta(\mathrm{x}) - \nabla_{\mathrm{x}} \log p_{\text{data}}(\mathrm{x})\|_2^2]$$

Score matching иллюстрация



$$\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{\theta}(\mathbf{x})$$

Denoising score matching

- Зашумим данные

$$p_\sigma(\tilde{x}) = \int_x q_\sigma(\tilde{x}|x)p_{\text{data}}(x)dx, \quad q_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$$

- Score matching для зашумленных данных

$$\mathbb{E}_{p_\sigma(x)}[\|s_\theta(x) - \nabla_x \log p_\sigma(x)\|_2^2]$$

- Можно переписать как

$$\mathbb{E}_{q_\sigma(\tilde{x}|x)p_{\text{data}}(x)}[\|s_\theta(x) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|_2^2]$$

Sliced score matching – I

- Многомерный случай

$$\frac{1}{2}\mathbb{E}_{p_{\text{data}}(x)}[\|\nabla_x \log p_{\text{data}}(x) - s_\theta(x)\|_2^2]$$

- Одномерный случай

$$\begin{aligned}
\mathcal{L}(\theta) =& \frac{1}{2}\mathbb{E}_{p_{\text{data}}(x)}[(\nabla_x \log p_{\text{data}}(x) - s_\theta(x))^2] \\
=& \frac{1}{2}\int p_{\text{data}}(x)[(\nabla_x \log p_{\text{data}}(x) - s_\theta(x))^2]dx \\
=& \frac{1}{2}\int p_{\text{data}}(x)(\nabla_x \log p_{\text{data}}(x))^2 dx - \\
& \int p_{\text{data}}(x)\nabla_x \log p_{\text{data}}(x) s_\theta(x) dx + \\
& \frac{1}{2}\int p_{\text{data}}(x) s_\theta(x)^2 dx
\end{aligned}$$

Sliced score matching – II

$$\int p_{\text{data}}(x)\nabla_x \log p_{\text{data}}(x)s_\theta(x)dx =$$

$$= \int p_{\text{data}}(x)\frac{\nabla_x p_{\text{data}}(x)}{p_{\text{data}}(x)}s_\theta(x)dx$$

$$= \int s_\theta(x)\nabla_x p_{\text{data}}(x)dx$$

$$= s_\theta(x)p_{\text{data}}(x)\Big|_{-\infty}^{+\infty} - \int p_{\text{data}}(x)\nabla_x s_\theta(x)dx$$

$$= -\int p_{\text{data}}(x)\nabla_x s_\theta(x)dx$$

Sliced score matching – III

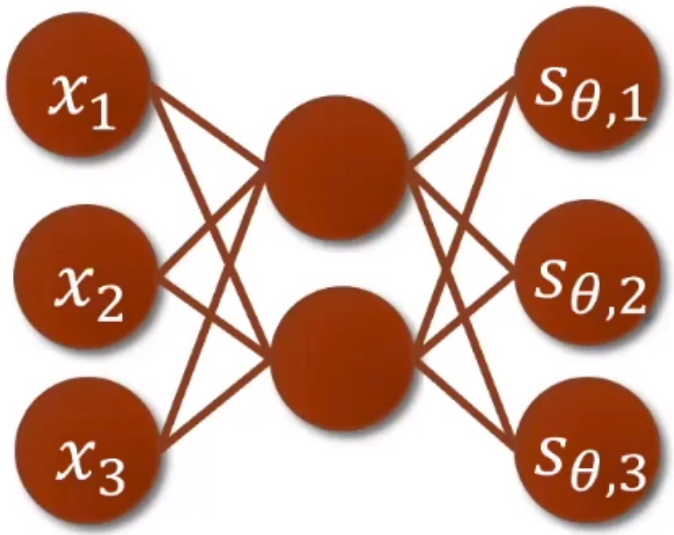
$$\mathcal{L}(\theta) = C + \int p_{\text{data}}(x)\nabla_x s_\theta(x)dx + \frac{1}{2}\int p_{\text{data}}(x)s_\theta(x)^2 dx$$
$$= \mathbb{E}_{p_{\text{data}}(x)}\left[\frac{1}{2}s_\theta(x)^2 + \nabla_x s_\theta(x)\right]$$

- Обратно к многомерному случаю

$$\mathbb{E}_{p_{\text{data}}(x)}\left[\frac{1}{2}\|s_\theta(x)\|_2^2 + tr(\nabla_x s_\theta(x))\right]$$

Sliced score matching – IV

- Score model



- Требует $\mathcal{O}(D)$ backpropagation

Sliced score matching – V

- Hutchinson trace estimator

$$\mathbb{E}_{p(v)}\left[\mathrm{v}^T \mathrm{A}\mathrm{v}\right] = tr(\mathrm{A}), \quad v \sim \{-1, 1\}^D$$

- Можно перейти к

$$\mathbb{E}_{p(v)}\mathbb{E}_{p_{\text{data}}(\mathrm{x})}\left[\frac{1}{2}\|\mathrm{s}_\theta(\mathrm{x})\|_2^2 + \mathrm{v}^T \nabla_{\mathrm{x}} \mathrm{s}_\theta(\mathrm{x})\mathrm{v}\right]$$

- Vector jacobian product (vjp) можно посчитать за $\mathcal{O}(1)$ backpropagation

$$\mathrm{v}^T \nabla_{\mathrm{x}} \mathrm{s}_\theta(\mathrm{x}) = \nabla_{\mathrm{x}} \mathrm{v}^T \mathrm{s}_\theta(\mathrm{x})$$

Langevin dynamics

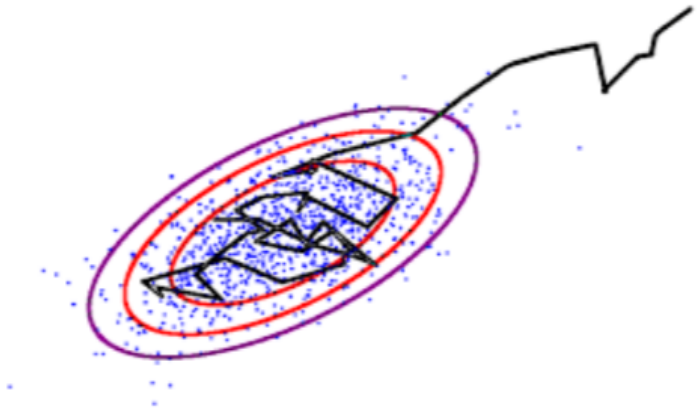
- Имеем доступ к

$$p(\mathrm{x}), \quad \nabla_{\mathrm{x}} \log p(\mathrm{x})$$

- Запустим динамику Ланжевена

$$\mathrm{x}_{i+1} \leftarrow \mathrm{x}_i + \epsilon \nabla_{\mathrm{x}} \log p(\mathrm{x}) + \sqrt{2\epsilon}\, \mathrm{z}_i, \quad i = 0, 1, \cdots, K,$$

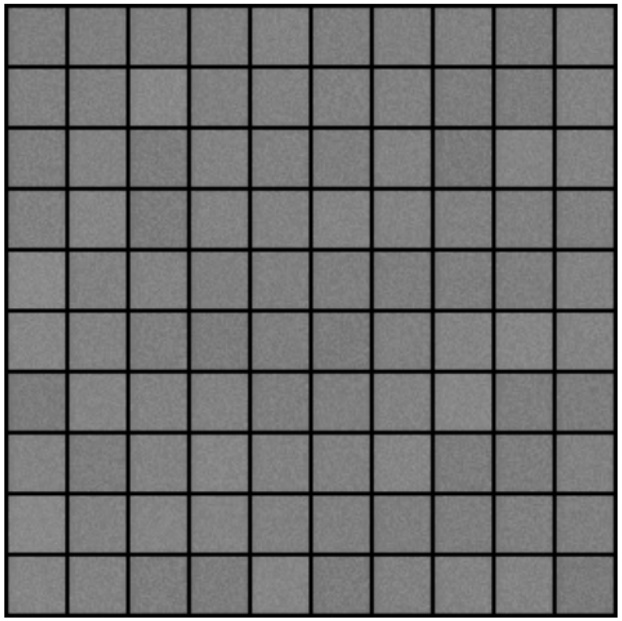
Где $\mathrm{z}_i \sim \mathcal{N}(0, I)$ и $K \to \infty$.



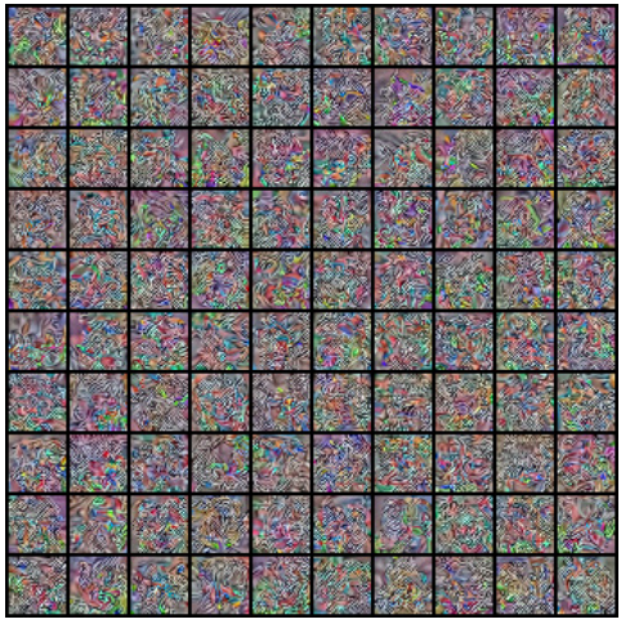
Score mutching промежуточные итоги

- Задаем модель
  - Как EBM модель и переходим к score function
  - Задаем score function
- Обучаем score mutching
  - Denoising score matching, зашумляя данные
  - Sliced score matching
- Сэмплируем
  - Динамикой Ланжевена

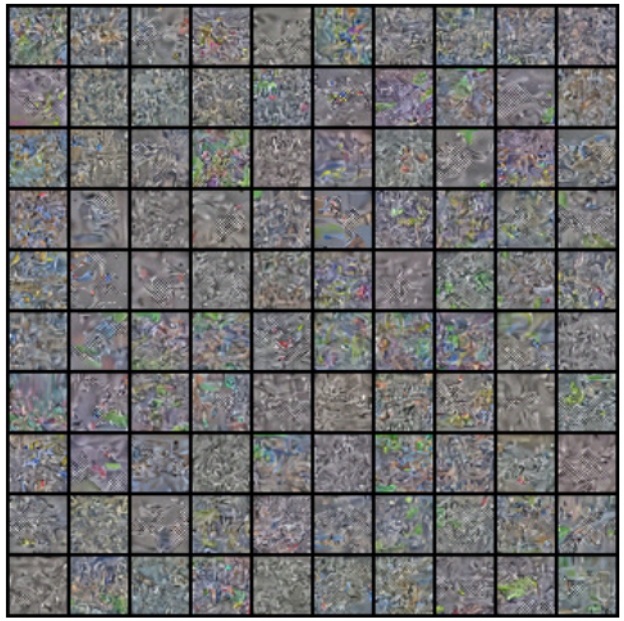
Наивный score mutching



(a) MNIST



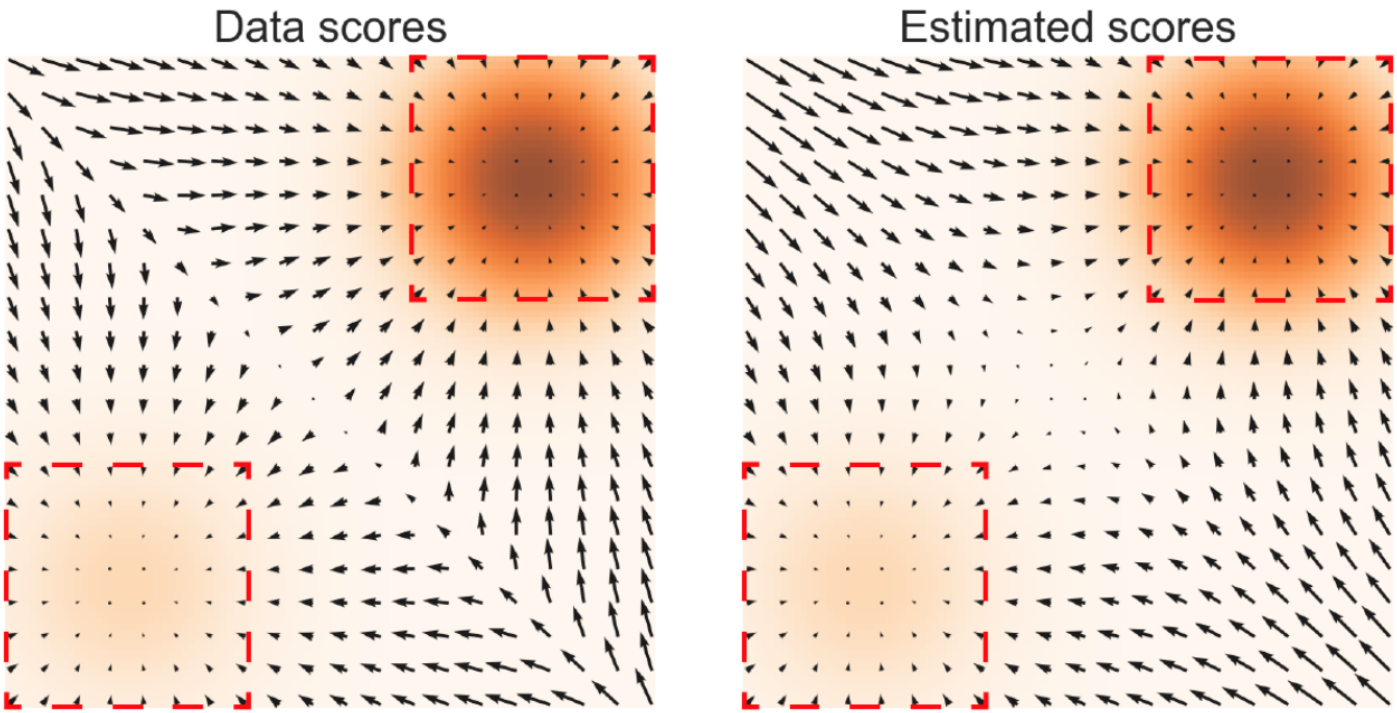
(b) CelebA



(c) CIFAR-10

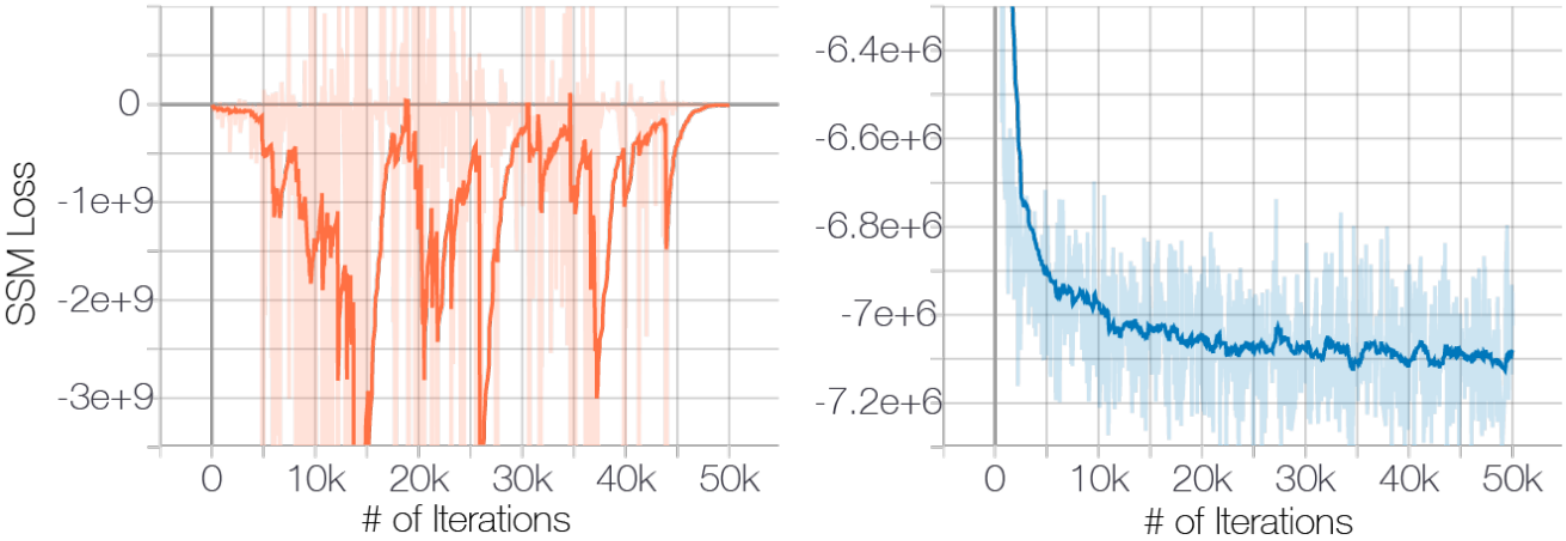
Score mutching, регионы без данных

- Ничего не знаем про регионы, где нет данных



Score mutching, нестабильность обучения

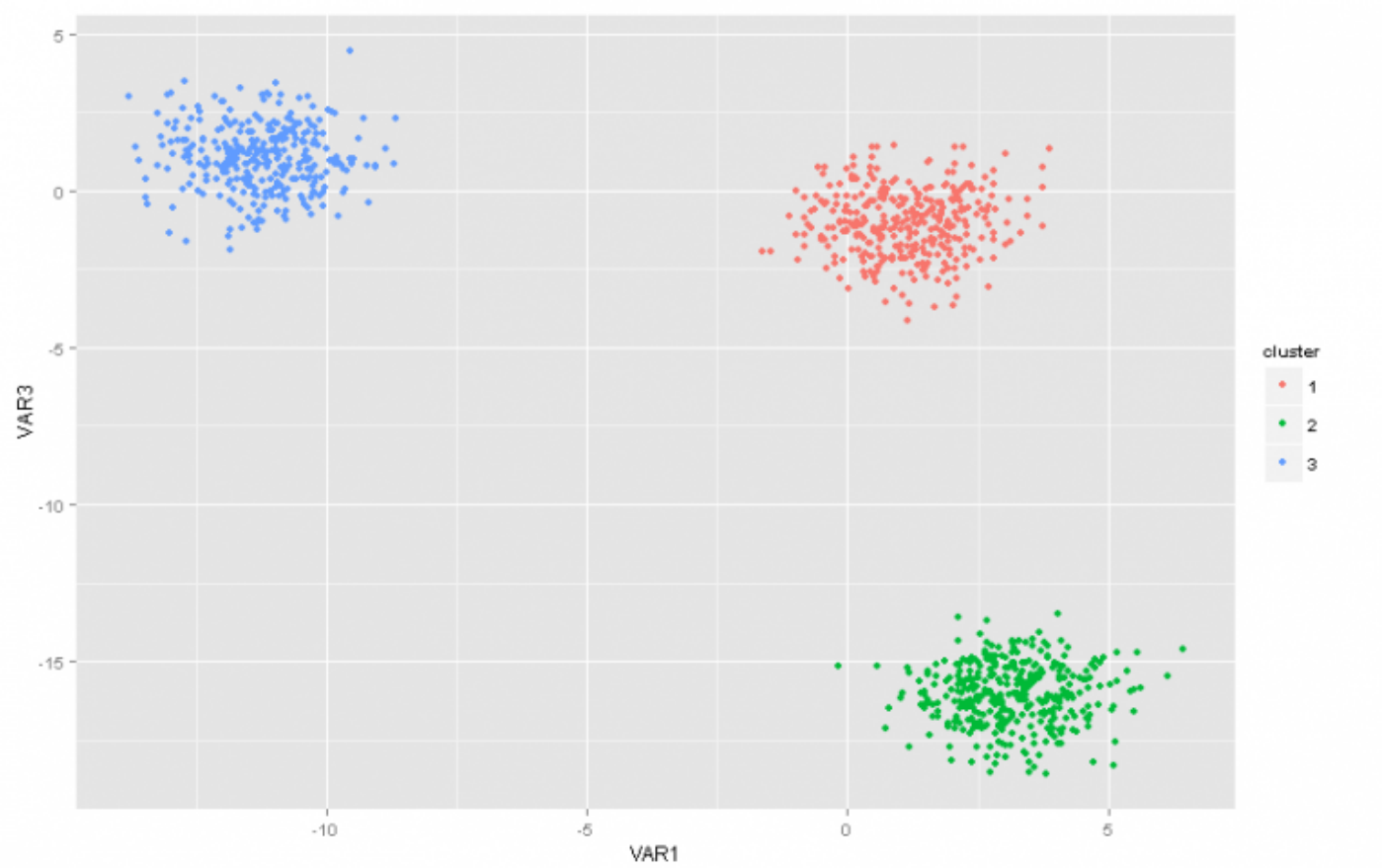
- Sliced score matching дает нестабильное обучение



- Добавление шума стабилизирует обучение

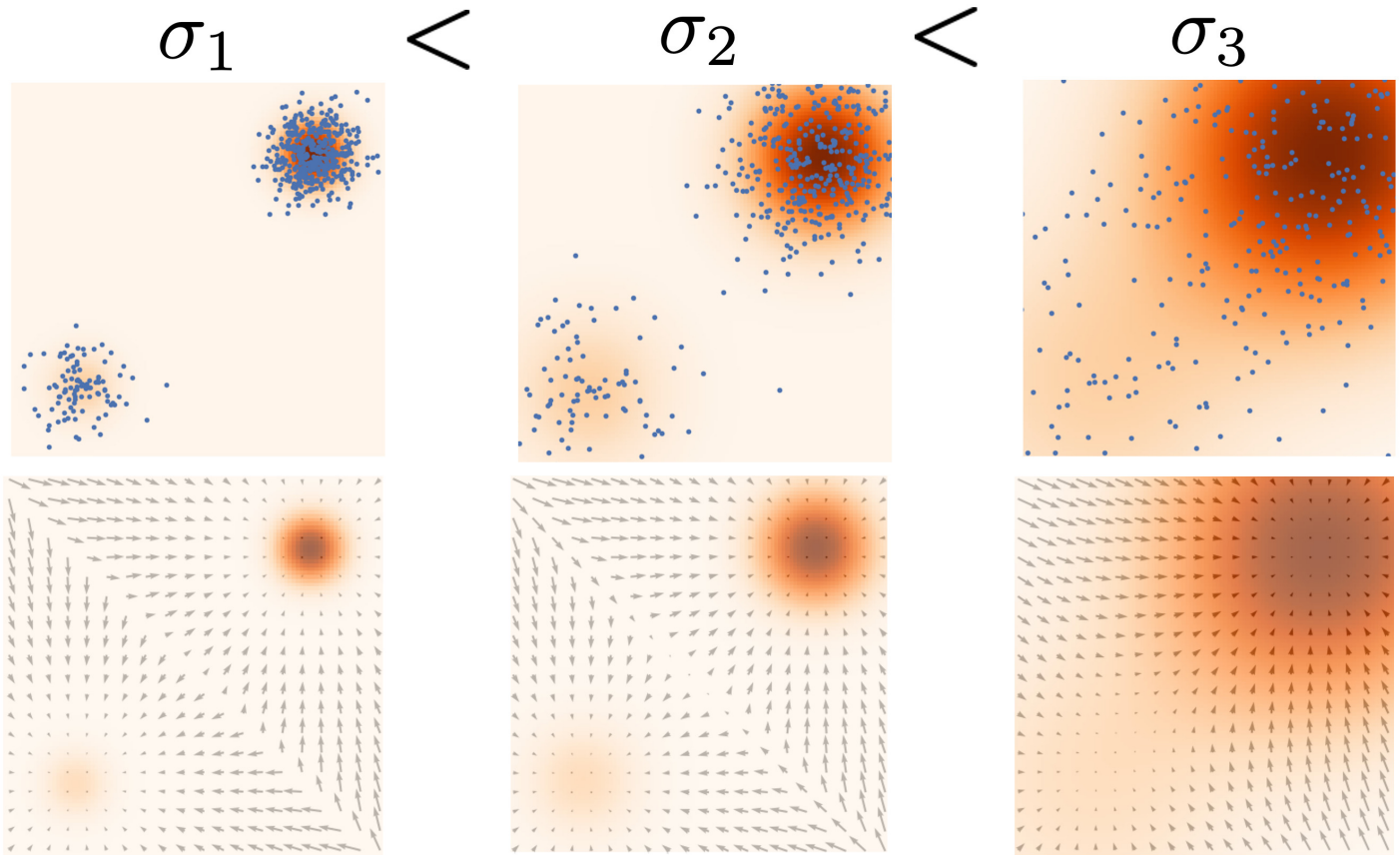
Score mutching, ограничения динамики Ланжевена

- Динамика Ланжевена с трудом переходит между модами



Несколько уровней зашумления

- Введем несколько уровней зашумления



Noise Conditiona Score Network NCSN

- Обусловим score function $\mathrm{s}_\theta(\mathrm{x})$ на уровень i.
- На каждом уровне м хотим

$$\mathrm{s}_\theta(\mathrm{x}, i) \approx \nabla_{\mathrm{x}} \log p_{\sigma_i}(\mathrm{x})$$

- Будем обучать

$$\sum_{i=1}^{L} \lambda(i) \mathbb{E}_{p_{\sigma_i}(\mathrm{x})}[\|\nabla_{\mathrm{x}} \log p_{\sigma_i}(\mathrm{x}) - \mathrm{s}_\theta(\mathrm{x}, i)\|_2^2], \quad \lambda(i) = \sigma_i^2$$

NCSN сэмплирование

- Авторы брали $L = 10$ и $T = 100$.

Algorithm 1 Annealed Langevin dynamics.

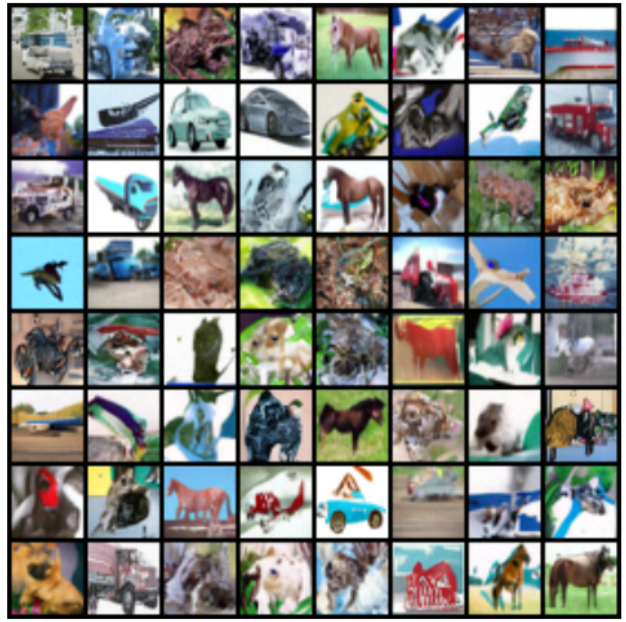
Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2 / \sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2} \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, \mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

NCSN сэмплирование, примеры

(a) MNIST

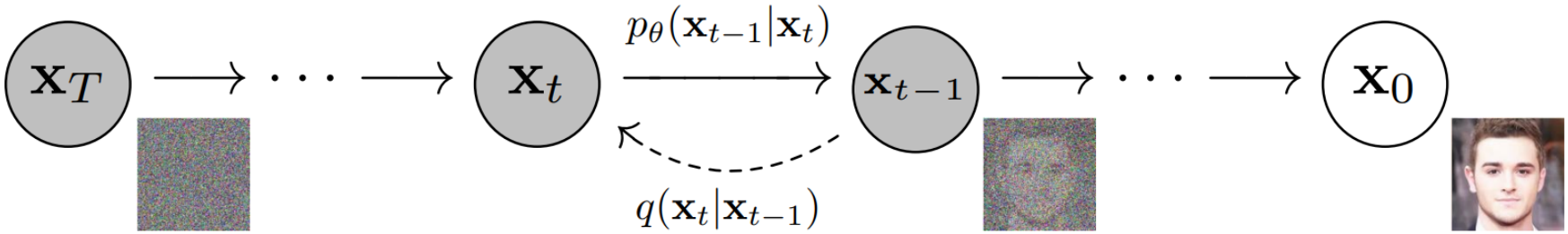
(b) CelebA



(c) CIFAR-10

DDPM, зашумление

- Общая схема



- Зашумление данных

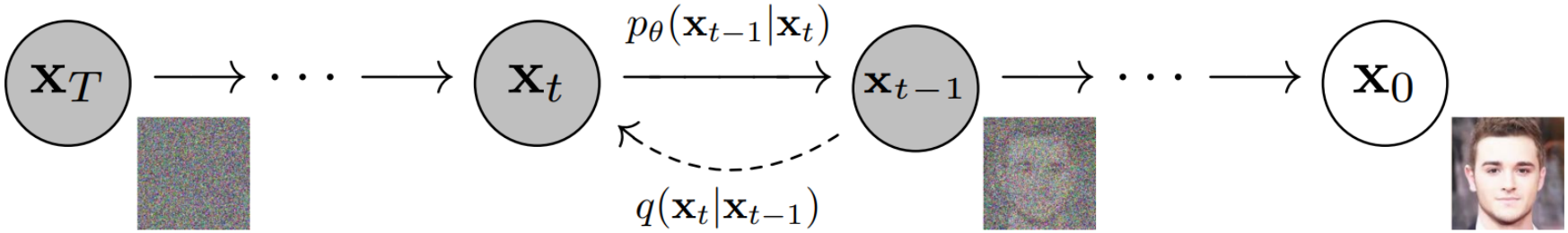
$$q(\mathsf{x}_{1:T}|\mathsf{x}_0) = \prod_{t=1}^{T} q(\mathsf{x}_t|\mathsf{x}_{t-1}), \quad q(\mathsf{x}_t|\mathsf{x}_{t-1}) = \mathcal{N}(\mathsf{x}_t; \sqrt{1-\beta_t}\mathsf{x}_{t-1}, \beta_t I)$$

- Зашумление на несколько шагов

$$q(\mathsf{x}_t|\mathsf{x}_0) = \mathcal{N}(\mathsf{x}_t; \sqrt{\bar{\alpha}_t}\mathsf{x}_0, (1-\bar{\alpha}_t)I), \quad \alpha_t = 1-\beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

DDPM, удаление шума

- Общая схема



- Удаление шума

$$p_\theta(\mathrm{x}_{0:T}|\mathrm{x}_0) = p(\mathrm{x}_T)\prod_{t=1}^{T} p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t), \quad p_\theta(\mathrm{x}_T) = \mathcal{N}(\mathrm{x}_T; 0, I)$$

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

DDPM, вероятность в точки

$$
\begin{aligned}
p_\theta(\mathrm{x}_0) &= \int p_\theta(\mathrm{x}_{0:T})\mathrm{d}\mathrm{x}_{1:T} \\
&= \int p_\theta(\mathrm{x}_{0:T})\frac{q(\mathrm{x}_{1:T}|\mathrm{x}_0)}{q(\mathrm{x}_{1:T}|\mathrm{x}_0)}\mathrm{d}\mathrm{x}_{1:T} \\
&= \int q(\mathrm{x}_{1:T}|\mathrm{x}_0)\frac{p_\theta(\mathrm{x}_{0:T})}{q(\mathrm{x}_{1:T}|\mathrm{x}_0)}\mathrm{d}\mathrm{x}_{1:T} \\
&= \int q(\mathrm{x}_{1:T}|\mathrm{x}_0)p_\theta(\mathrm{x}_T)\prod_{t=1}^{T}\frac{p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)}{q(\mathrm{x}_t|\mathrm{x}_{t-1},\mathrm{x}_0)}\mathrm{d}\mathrm{x}_{1:T}
\end{aligned}
$$

DDPM, вариационная нижняя оценка – I

$$
\begin{aligned}
\mathcal{L}_\theta &= \mathbb{E}_{q(\mathrm{x}_0)}\left[-\log p_\theta(\mathrm{x}_0)\right] \\
&= \int q(\mathrm{x}_0)\left[-\log\left[\int q(\mathrm{x}_{1:T}|\mathrm{x}_0)p_\theta(\mathrm{x}_T)\prod_{t=1}^{T}\frac{p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)}{q(\mathrm{x}_t|\mathrm{x}_{t-1},\mathrm{x}_0)}\mathrm{d}\mathrm{x}_{1:T}\right]\right]\mathrm{d}\mathrm{x}_0 \\
&\leq -\int q(\mathrm{x}_0)\left[\int q(\mathrm{x}_{1:T}|\mathrm{x}_0)\log\left[p_\theta(\mathrm{x}_T)\prod_{t=1}^{T}\frac{p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)}{q(\mathrm{x}_t|\mathrm{x}_{t-1},\mathrm{x}_0)}\right]\mathrm{d}\mathrm{x}_{1:T}\right]\mathrm{d}\mathrm{x}_0 \\
&= -\int q(\mathrm{x}_{0:T})\log\left[p_\theta(\mathrm{x}_T)\prod_{t=1}^{T}\frac{p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)}{q(\mathrm{x}_t|\mathrm{x}_{t-1},\mathrm{x}_0)}\right]\mathrm{d}\mathrm{x}_{0:T} \\
&= \mathbb{E}_{q(\mathrm{x}_{0:T})}\left[-\log p_\theta(\mathrm{x}_T) - \sum_{t=1}^{T}\log\frac{p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)}{q(\mathrm{x}_t|\mathrm{x}_{t-1},\mathrm{x}_0)}\right]
\end{aligned}
$$

DDPM, вариационная нижняя оценка – II

$$
\begin{aligned}
&= \mathbb{E}_{q(x_{0:T})} \left[-\log p_\theta(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}, x_0)} \right] \\
&= \mathbb{E}_{q(x_{0:T})} \left[-\log p_\theta(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}, x_0)} - \log \frac{p_\theta(x_0|x_1)}{q(x_t|x_1, x_0)} \right] \\
&= \mathbb{E}_{q(x_{0:T})} \left[-\log p_\theta(x_T) - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \frac{p_\theta(x_{t-1}|x_0)}{q(x_t|x_0)} - \log \frac{p_\theta(x_0|x_1)}{q(x_t|x_1, x_0)} \right] \\
&= \mathbb{E}_{q(x_{0:T})} \left[-\log \frac{p_\theta(x_T)}{q(x_T|x_0)} - \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log p_\theta(x_0|x_1) \right] \\
&= \mathbb{E}_{q(x_{0:T})} \left[D_{KL}(q(x_T|x_0)||p_\theta(x_T)) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \right]
\end{aligned}
$$

DDPM, вариационная нижняя оценка – III

- ELBO

$$\mathbb{E}_{q(\mathrm{x}_{0:T})}\left[D_{KL}(q(\mathrm{x}_T|\mathrm{x}_0)||p_\theta(\mathrm{x}_T)) + \sum_{t>1} D_{KL}(q(\mathrm{x}_{t-1}|\mathrm{x}_t,\mathrm{x}_0)||p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t)) - \log p_\theta(\mathrm{x}_0|\mathrm{x}_1)\right]$$

- Распишем $q(\mathrm{x}_{t-1}|\mathrm{x}_t,\mathrm{x}_0)$

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

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

$$\tilde{\beta}_t(\mathrm{x}_t,\mathrm{x}_0) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

DDPM, репараметризация – I

- Распишем $\mathcal{L}_{t-1}$

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

- Заметим

$$\mathrm{x}_t(\mathrm{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\mathrm{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$\mathrm{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathrm{x}_t(\mathrm{x}_0, \epsilon) - \sqrt{1 - \bar{\alpha}_t}\epsilon \right)$$

- Перепишем

$$\begin{aligned} \mathcal{L}_{t-1} &= \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t \left(\mathrm{x}_t(\mathrm{x}_0, \epsilon), \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathrm{x}_t(\mathrm{x}_0, \epsilon) - \sqrt{1 - \bar{\alpha}_t}\epsilon \right) \right) - \mu_\theta(\mathrm{x}_t, t)\|^2 \right] \\ &= \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \| \frac{1}{\sqrt{\alpha_t}} \left(\mathrm{x}_t(\mathrm{x}_0, \epsilon) - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon \right) - \mu_\theta(\mathrm{x}_t, t)\|^2 \right] \end{aligned}$$

DDPM, репараметризация – II

- Перепишем μ_θ

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(\mathrm{x}_t(\mathrm{x}_0, \epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathrm{x}_t, t)\right)$$

- Снова распишем $\mathcal{L}_{t-1}$

$$\mathcal{L}_{t-1} = \mathbb{E}_{q,\epsilon}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\|\epsilon - \epsilon_\theta(\mathrm{x}_t, t)\|^2\right]$$

DDPM сэмплирование, примеры