

Ускорение семплирования из диффузионных моделей с использованием состязательных сетей

Охотников Никита Владимирович

МФТИ

2023

Цели исследования

Цель

Модифицировать классическую диффузионную модель для существенного ускорения процесса семплирования

Задача

Проанализировать способы моделирования мультимодального распределения в обратном диффузионном процессе

Предлагается

Использовать неявную генеративную модель – состязательную сеть – на каждом шаге диффузионного процесса

Необходимо

Рассмотреть различные постановки минимизационной задачи для используемой модели

- ▶ Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020
- ▶ Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans, 2021.
- ▶ Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.
- ▶ Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization, 2016.
- ▶ Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Неявное моделирование обратного диффузионного процесса

Диффузионный процесс

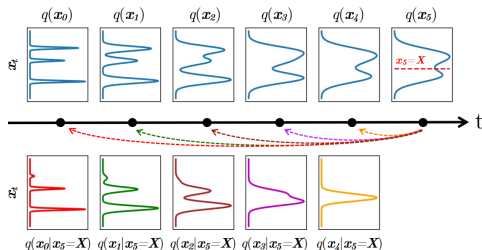
▶ Прямой:

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

где $t = \overline{0}, \overline{T}$, \mathbf{x}_0 – семпл из исходного распределения, \mathbf{x}_t – семпл на шаге t , $\beta_t \in (0, 1)$

▶ Обратный:

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) \underset{T \gg 1}{\approx} \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\theta}(\mathbf{x}_t, t), \Sigma_{\theta}(\mathbf{x}_t, t))$$



Основные предположения

- ▶ Марковость обратного процесса
- ▶ Нормальность и следовательно унимодальность $p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)$

Предложение

- ▶ Использовать неявную модель для восстановления распределения

Мотивация

- ▶ Моделирование мультимодального распределения для существенного уменьшения T

Диффузионная модель

Описание

В основе модели лежит постепенное добавление случайного нормального шума с коэффициентом $\beta_t \in (0, 1)$ в семпл \mathbf{x}_0 из исходного распределения в прямом процессе и постепенное восстановление распределения в обратном.

Прямой процесс

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

где $t = \overline{0, T}$, \mathbf{x}_t – семпл на шаге t . В таком случае, принимая $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{i=1}^t \alpha_i$ можно записать:

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

Таким образом, при достаточно больших T со сколь угодно большой точностью $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, а значит обратный процесс начинается с нормального шума.

После некоторых математических преобразований получаем минимизационную задачу:

$$\sum_{t=1}^n \mathbb{E}_{\mathbf{x}_1 \dots \mathbf{x}_T} KL(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)) \rightarrow \min$$

Обратный процесс

В приближении $T \gg 1$ распределение каждого следующего семпла в обратном процессе обусловлено только на предыдущий, а также нормально.

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) \underset{T \gg 1}{\approx} \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\theta}(\mathbf{x}_t, t), \Sigma_{\theta}(\mathbf{x}_t, t))$$

Если $\mathbf{X} = (\mathbf{x}_0^1 \dots \mathbf{x}_0^n) \sim p_0(\mathbf{x})$, то из метода максимального правдоподобия:

$$\theta = \underset{\theta}{\operatorname{argmax}} p(\mathbf{X} | \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \log p(\mathbf{x}_0^i | \theta)$$

Постановка задачи

Проблема

При существенном уменьшении числа шагов обратного диффузионного процесса ($T \gtrsim 1$) предположения марковости и тем более нормальности $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ очевидно не верны. Кроме того, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ мультимодальное.

Задача

Предложить неявную модель для аппроксимации мультимодального распределения в обратном процессе.

Метод

По аналогии с классической диффузионной моделью будем минимизировать некоторую меру близости между распределениями D_{adv} , но при помощи неявной модели

$$\sum_{t=1}^n \mathbb{E}_{\mathbf{x}_1 \dots \mathbf{x}_T} D_{adv} (q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \xrightarrow{\theta} \min$$

где D_{adv} , в случае состязательных сетей, есть некоторая f-дивергенция или метрика Вассерштайна.

Введение GAN моделей

Дискриминатор

Зададим дискриминатор как $D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t)$, где φ – обучаемые параметры. Для начала используем схему тренировки стандартного GAN¹, как ранее было предложено², тогда задача минимизации для дискриминатора:

$$\min_{\varphi} \sum_{t \geq 1}^n \mathbb{E}_{q(\mathbf{x}_t)} [\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\log(D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\log(1 - D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))]]$$

Генератор

Введем генератор $G_\theta(\mathbf{x}_{t-1}^{fake}, \mathbf{z}, t)$ с параметрами θ , латентной переменной $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, обусловленный на \mathbf{x}_{t-1}^{fake} и порождающий семплы из исходного распределения. В таком случае целевое распределение в обратном диффузионном процессе:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \int p_\theta(\mathbf{x}_0|\mathbf{x}_t) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_0 = \int p(\mathbf{z}) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = G_\theta(\mathbf{x}_t, \mathbf{z}, t)) d\mathbf{z}$$

При известном дискриминаторе тренируем генератор на максимизацию

$$\max_{\theta} \sum_{t \geq 1}^n \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\log(D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))]$$

¹<https://doi.org/10.48550/arxiv.1406.2661>

²<https://doi.org/10.48550/arxiv.2112.07804>

Альтернативные схемы тренировки

Семейство F-дивергенций

F-дивергенции – семейство функций, определяемых как

$$D_f(Q||P) = \int_{\mathcal{X}} p(\mathbf{x}) f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x},$$

где f – выпуклая непрерывная слева функция, такая что $f(1) = 0$ называемая порождающей.

Произвольную f-дивергенцию можно оценить снизу, используя сопряженную к f функцию f^* :

$$D_f(Q||P) \geq \sup_{T \in \mathcal{T}} (\mathbb{E}_{\mathbf{x} \sim Q}[T(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P}[f^*(T(\mathbf{x}))]),$$

где \mathcal{T} – произвольный класс функций $T : \mathcal{X} \rightarrow \mathbb{R}$.

Минимизируемый функционал

Генератор распределения $P: G = G_\theta$,

модель V_ω приближающая функцию $T: T = T_\omega = g_f(V_\omega)$, где g_f – функция активации.

В таком случае целевой функционал:

$$F(\theta, \omega) = \mathbb{E}_{\mathbf{x} \sim Q}[g_f(V_\omega(\mathbf{x}))] - \mathbb{E}_{\mathbf{x} \sim P_\theta}[f^*(g_f(V_\omega(\mathbf{x})))].$$

Будем подставлять частные случаи этого функционала в схему тренировки диффузионной модели аналогично стандартному GAN.

Альтернативные схемы тренировки

Рассматриваемые частные случаи

- ▶ Квадрат расстояния Хелингера

$$H^2(Q||P) - \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{q(\mathbf{x})} - \sqrt{p(\mathbf{x})} \right)^2 d(\mathbf{x}), \quad f_{H^2}^*(t) = \frac{t}{1-t}, \quad g_{f_{H^2}}(V) = 1 - \exp(V).$$

$$\min_{\theta} \max_{\varphi} \sum_{t \geq 1}^n \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\exp(V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] - \mathbb{E}_{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\exp(-V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] \right]$$

- ▶ Обратная KL-дивергенция

$$\text{Reverse-KL}(Q||P) = \text{KL}(P||Q) = \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}, \quad f_{R-KL}^*(t) = -1 - \log(-t), \quad g_{f_{R-KL}}(V) =$$

$$\min_{\theta} \max_{\varphi} \sum_{t \geq 1}^n \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\exp(V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] + \mathbb{E}_{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)} [1 + V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t)] \right].$$

- ▶ Total variation distance

$$\delta(Q||P) = \sup_{\mathbf{x}} |Q(\mathbf{x}) - P(\mathbf{x})|, \quad f_{\delta}^*(t) = t, \quad g_{f_{\delta}}(V) = \frac{1}{2} \tanh(V).$$

$$\min_{\theta} \max_{\varphi} \sum_{t \geq 1}^n \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\tanh(V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] - \mathbb{E}_{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\tanh(V_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] \right].$$

Альтернативные схемы тренировки

Wasserstein distance

Отдельно рассматриваем wasserstein distance – другую меру близости распределений, не входящую в семейство f-дивергенций

$$W(Q||P) = \inf_{\gamma \in \Gamma(Q,P)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|,$$

Где Γ – множество всех совместных распределений $\gamma(\mathbf{x}, \mathbf{y})$ таких, что $\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} = P(\mathbf{y})$, $\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y} = Q(\mathbf{x})$.

Пользуясь, двойственностью Канторовича-Рубинштейна получаем:

$$\begin{cases} \max_{\varphi} (\mathbb{E}_{\mathbf{x} \sim Q} [f_{\varphi}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P_{\theta}} [f_{\varphi}(\mathbf{x})]) , \\ \|f\|_L \leq K, \end{cases}$$

где $\|f_{\varphi}\|_L \leq K$ – K -липшицевы функции, приближаемые нейросетью с параметрами φ , θ – параметры генератора для распределения P .

Итоговая задача оптимизации:

$$\begin{cases} \min_{\theta} \max_{\varphi} \sum_{t=1}^n \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [f_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t)] - \mathbb{E}_{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)} [f_{\varphi}(\mathbf{x}_{t-1}, \mathbf{x}_t, t)] \right] , \\ \|f_{\varphi}\|_L \leq K. \end{cases}$$

Вычислительный эксперимент

Цели

- ▶ Анализ качества семплов в зависимости от количества шагов обратного процесса T для DDPM
- ▶ Достижение сравнимого качества для малых $T \leq 10$ с использованием GAN
- ▶ Анализ применимости альтернативных состязательных сетей

Метрика качества

Fréchet inception distance или FID-score, в предположение нормальности вычисляется как:

$$\text{FID}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma'))^2 = \|\mu - \mu'\|_2^2 + \text{tr} \left(\Sigma + \Sigma' - 2 \left(\Sigma^{\frac{1}{2}} \cdot \Sigma' \cdot \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right).$$

В данное выражение подставляем выход модели InceptionV3³ для реальных и сгенерированных данных.

Данные

Fashion-MNIST⁴ – 60000 черно-белых картинок 28×28

Архитектура

Генеративная сеть – собственная реализация U-Net⁵ модели.

Дискриминативная – сжимающая половина генеративной и выходной линейный слой.

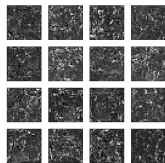
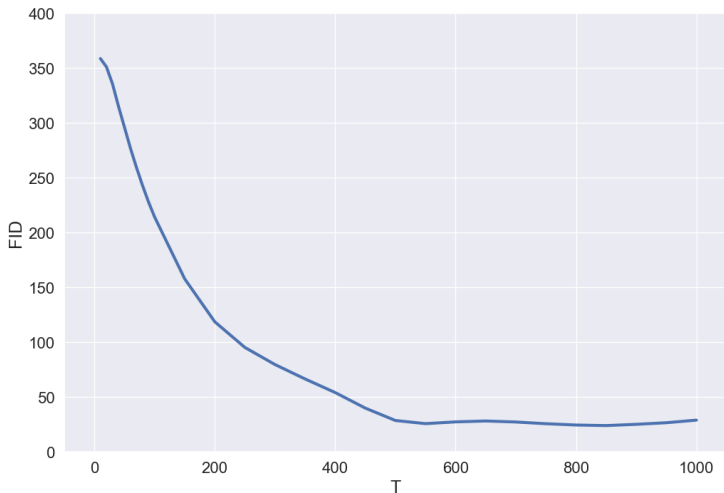
³<https://doi.org/10.48550/arXiv.1512.00567>

⁴<http://arxiv.org/abs/1708.07747>

⁵<http://arxiv.org/abs/1505.04597>

Вычислительный эксперимент

Диффузионная модель



(a) $T = 50$



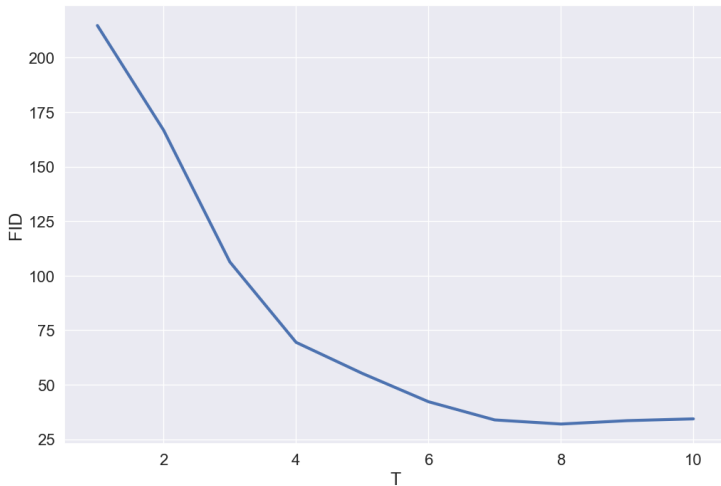
(b) $T = 300$



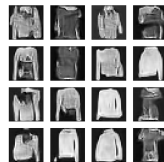
(c) $T = 1000$

Вычислительный эксперимент

DDGAN с JS-дивергенцией



(a) $T = 3$



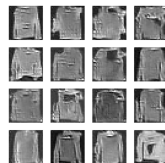
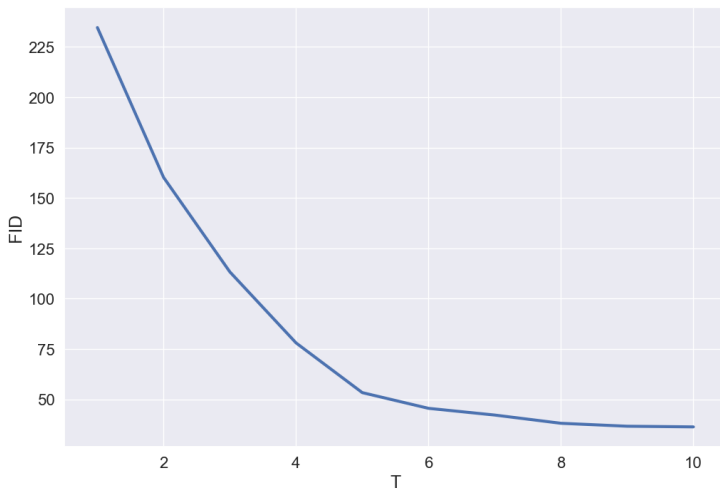
(b) $T = 5$



(c) $T = 10$

Вычислительный эксперимент

DDGAN с H^2



(a) $T = 3$



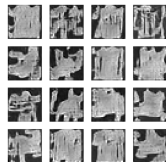
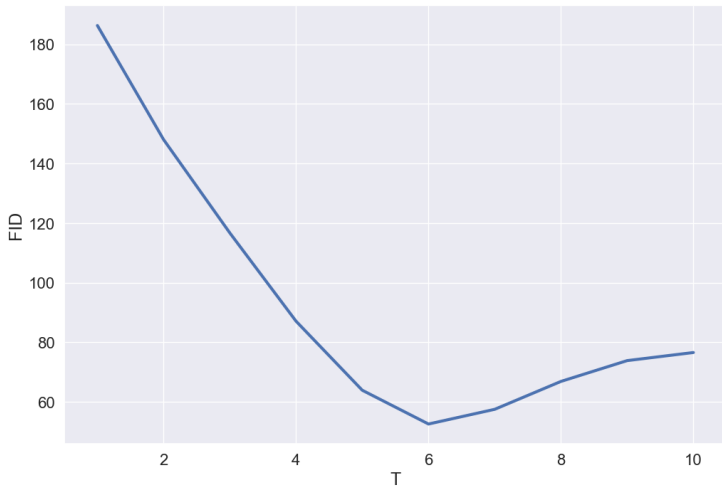
(b) $T = 5$



(c) $T = 10$

Вычислительный эксперимент

DDGAN с обратной KL дивергенцией



(a) $T = 3$



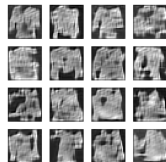
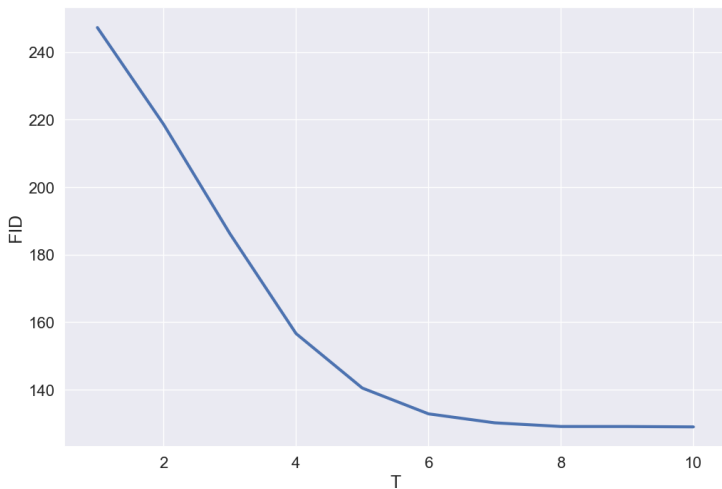
(b) $T = 5$



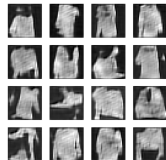
(c) $T = 10$

Вычислительный эксперимент

DDGAN с total variation



(a) $T = 3$



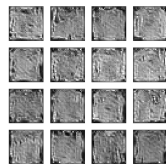
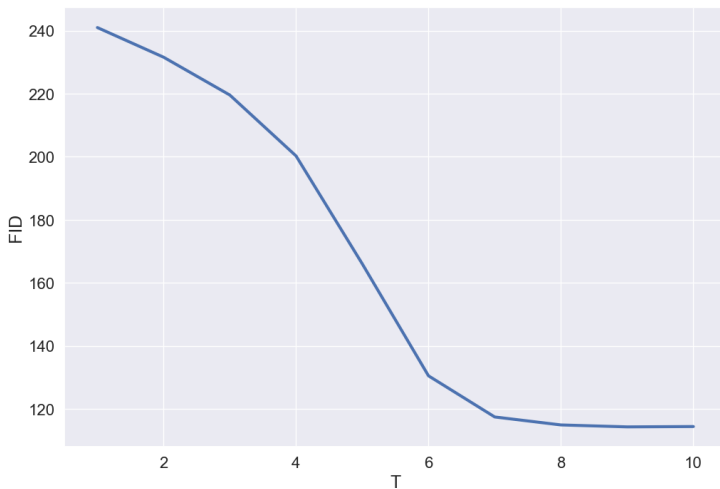
(b) $T = 5$



(c) $T = 10$

Вычислительный эксперимент

DDGAN с wasserstein distance



(a) $T = 3$



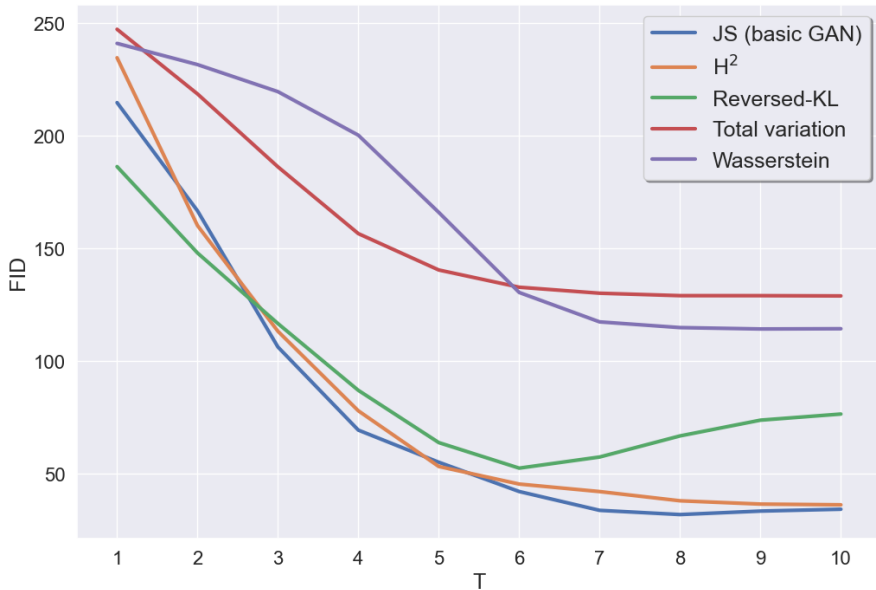
(b) $T = 5$



(c) $T = 10$

Вычислительный эксперимент

Сравнение различных схем тренировки



Вычислительный эксперимент

Сравнение различных схем тренировки

Модель	FID
DDPM	28.9
JS DDGAN	34.3
H ² DDGAN	36.3
Reversed-KL DDGAN	76.6
Total variation DDGAN	129.0
Wasserstein DDGAN	114.4

FID-score для максимального T

Итоги эксперимента

- ▶ Количество шагов в диффузионной модели успешно снижено на 2 порядка при незначительном падении качества семплов
- ▶ Стандартный (JS) DDGAN показывает себя лучше остальных
- ▶ DDGAN с H² достигает сравнимого со стандартным качества
- ▶ Обучение всех DDGAN кроме стандартного склонно к расбалансировке и требует тщательного подбора гиперпараметров

- ▶ Предложены различные способы моделирования мультимодального распределения в обратном диффузионном процессе
- ▶ Экспериментально подтверждена возможность использования различных подходов к обучению GAN моделей для использования их в диффузионной сети
- ▶ Проведено сравнение различных алгоритмов обучения на синтетических данных.
- ▶ В дальнейшем планируется исследовать способы выбора схемы тренировки в зависимости от входных данных