

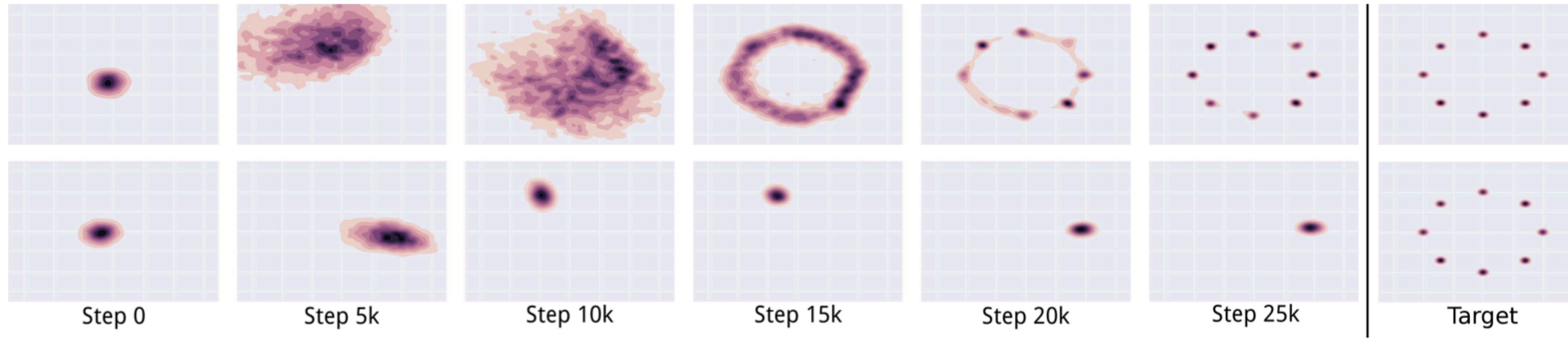
Проблемы GAN и их решения в более сложных архитектурах

Ким Михаил, 182

Основные проблемы

- Генерация однообразных примеров (модальный коллапс)
- Неустойчивость обучения
- Затухание градиента

Модальный коллапс



Источник: [Unrolled Generative Adversarial Networks](#)

- Генератор учит наиболее вероятный пример с точки зрения дискrimинатора
- Дискриминатор научится отклонять этот пример и, вероятно, застрянет в локальном оптимуме
- Генератор обучается под конкретный дискриминатор и наоборот

Неустойчивость обучения

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_r(x)} \log D(x) + \mathbb{E}_{x \sim p_g(x)} \log(1 - D(x))$$

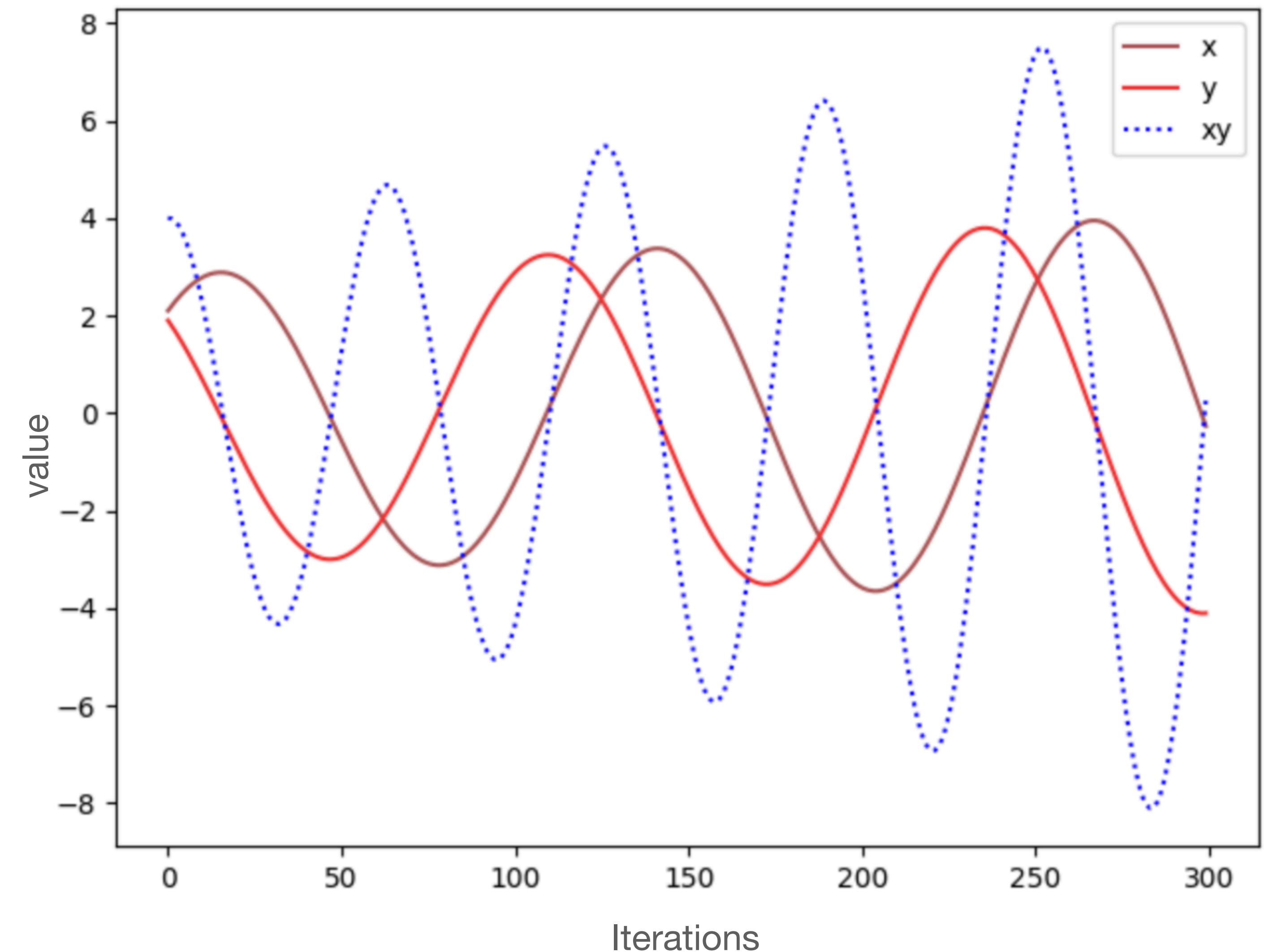
Равновесие Нэша: набор стратегий, при которых ни один игрок не может увеличить выигрыш путём изменения своей стратегии, если другие игроки свои стратегии не меняют

Неустойчивость обучения

$$\min_y \max_x V(x, y) = xy$$

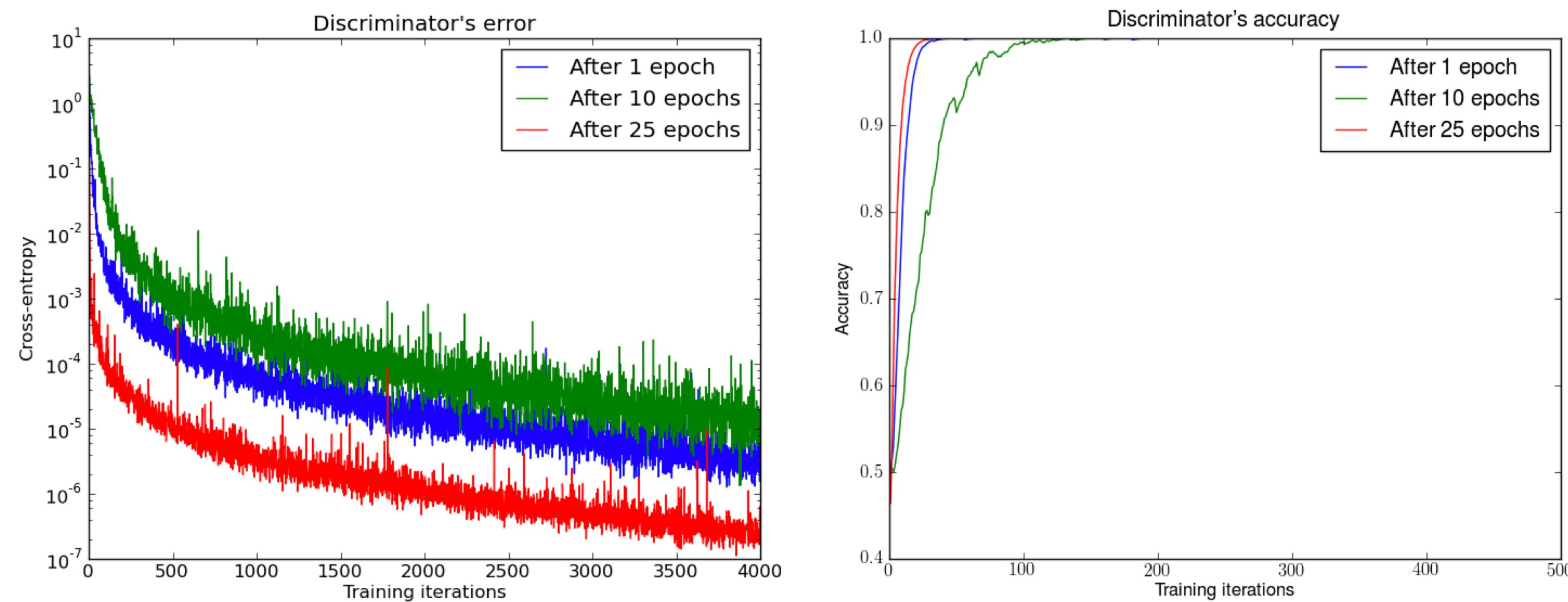
$$y = y - \eta x$$

$$x = x + \eta y$$



Затухание градиента

Можно показать, что многообразия, содержащие support распределений ($\{x : p(x) \neq 0\}$), не совпадают, из этого следует, что существует оптимальный дискриминатор



Источник: [Towards Principled Methods for Training Generative Adversarial Networks](#)

Затухание градиента

Что в этом плохого?

Затухание градиента

Что в этом плохого?

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_r(x)} \log D(x) + \mathbb{E}_{x \sim p_g(x)} \log(1 - D(x))$$

- При оптимальном дискриминаторе значение функции потерь будет близко к 0
- При обучении генератор не будет получать полезную информацию от дискриминатора

Затухание градиента

Что в этом плохого?

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_r(x)} \log D(x) + \mathbb{E}_{x \sim p_g(x)} \log(1 - D(x))$$

- При оптимальном дискриминаторе значение функции потерь будет близко к 0
- При обучении генератор не будет получать полезную информацию от дискриминатора

Особенно плохо влияет в начале обучения

NSGAN

Non-saturating GAN:

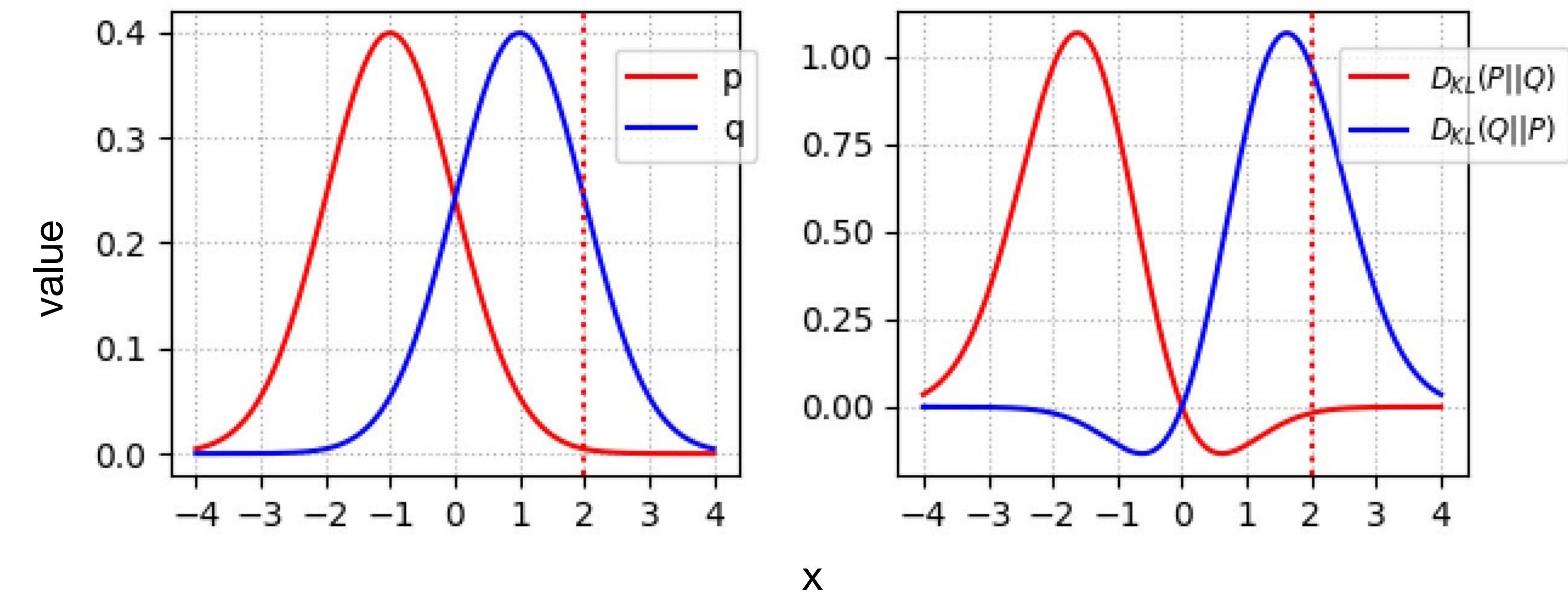
$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_r(x)} \log D(x) - \mathbb{E}_{x \sim p_g(x)} \log(D(x))$$

Минимизация $\mathbb{E}_{z \sim p(z)} \log(1 - D(G(z)))$ по $G(z)$ эквивалентна максимизации $\mathbb{E}_{z \sim p(z)} \log(D(G(z)))$ или минимизации $-\mathbb{E}_{z \sim p(z)} \log(D(G(z)))$.

WGAN: Сравнение разных дивергенций

$$KL(\mathbb{P}_r \parallel \mathbb{P}_g) = \int p_r(x) \log \frac{p_r(x)}{p_g(x)} dx$$

- $p_r > 0, p_g \rightarrow 0$ очень большой штраф за модальный коллапс
- $p_g > 0, p_r \rightarrow 0$ небольшой штраф за нереалистичные изображения



Источник: [Why it is so hard to train Generative Adversarial Networks!](#)

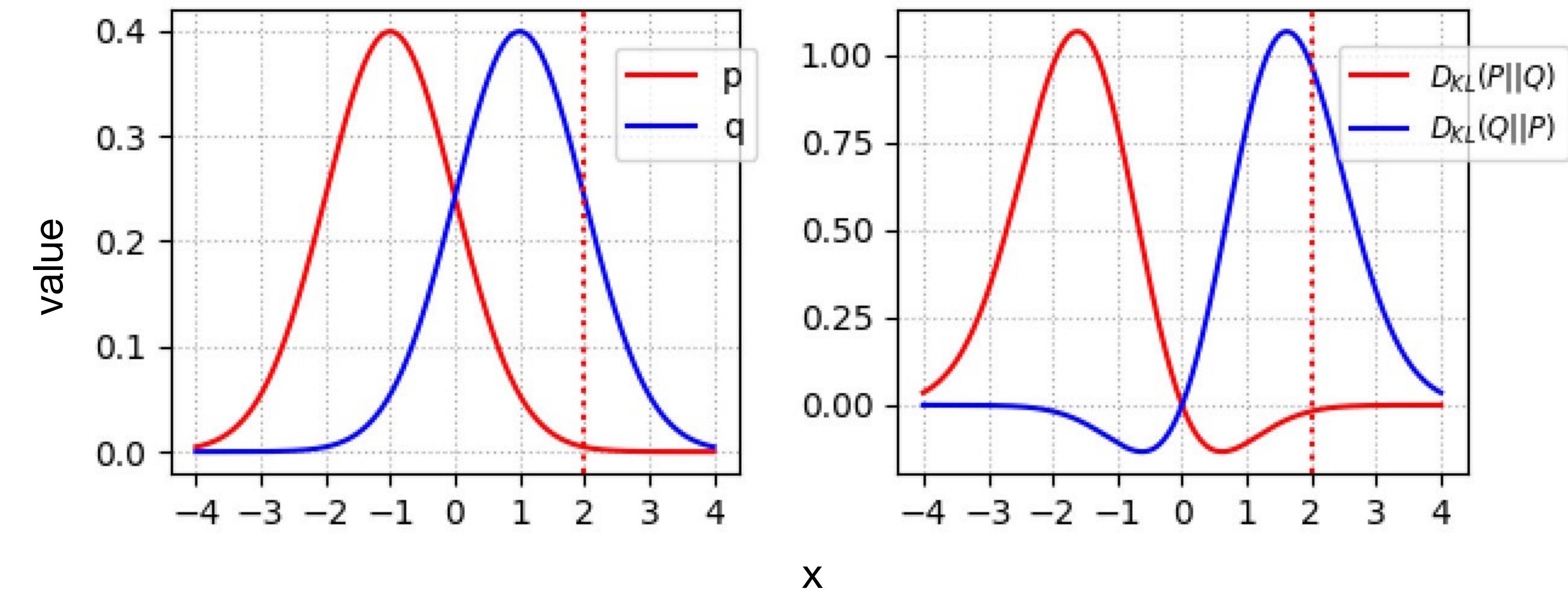
WGAN: Сравнение разных дивергенций

$$KL(\mathbb{P}_r \parallel \mathbb{P}_g) = \int p_r(x) \log \frac{p_r(x)}{p_g(x)} dx$$

- $p_r > 0, p_g \rightarrow 0$ очень большой штраф за модальный коллапс
- $p_g > 0, p_r \rightarrow 0$ небольшой штраф за нереалистичные изображения

$$KL(\mathbb{P}_g \parallel \mathbb{P}_r) = \int p_g(x) \log \frac{p_g(x)}{p_r(x)} dx$$

- $p_r > 0, p_g \rightarrow 0$ небольшой штраф за модальный коллапс
- $p_g > 0, p_r \rightarrow 0$ очень большой штраф за нереалистичные изображения



Источник: [Why it is so hard to train Generative Adversarial Networks!](#)

WGAN: Сравнение разных дивергенций

Jensen-Shannon Divergence

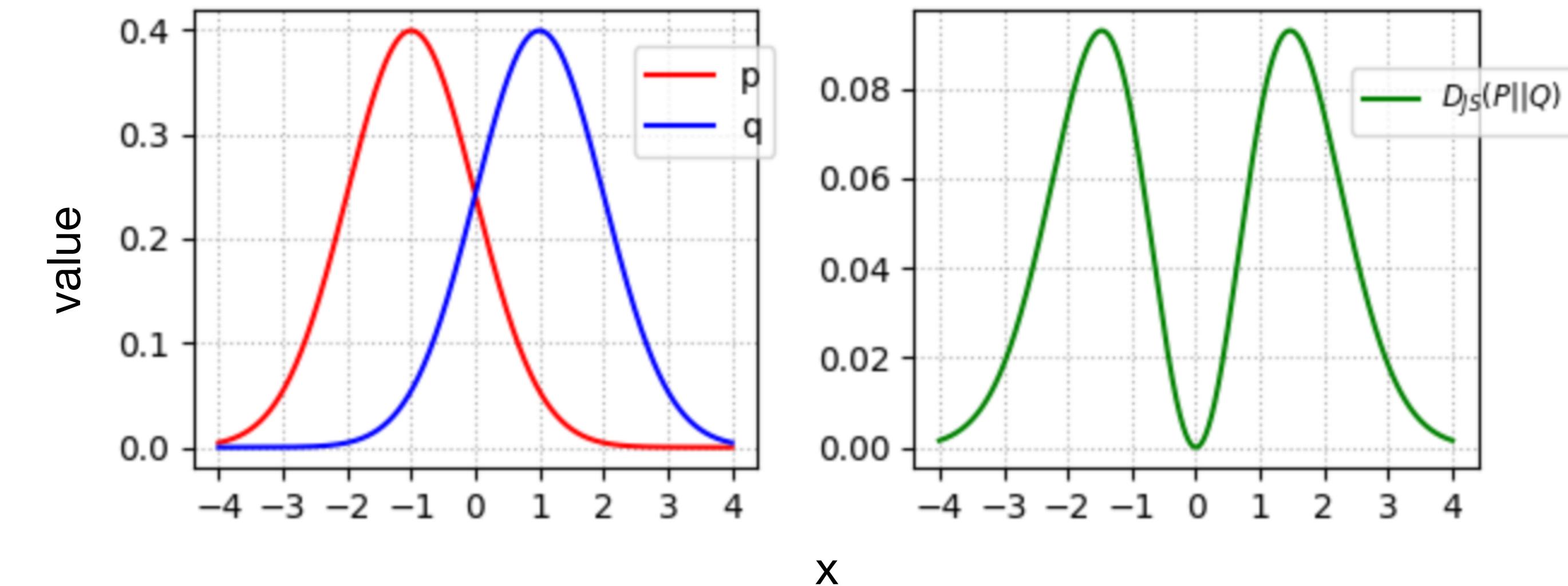
$$JS(p_r, p_g) = KL(p_r \| p_m) + KL(p_g \| p_m)$$

$$\text{где } p_m = \frac{p_r + p_g}{2}$$

- Симметрична
- Тесно связана с обучением GAN:

Пусть $D_G^*(x)$ – оптимальный дискриминатор среди всех возможных дискриминаторов при фиксированном генераторе $G(x)$, тогда:

$$V(G, D^*) = 2JS(p_r, p_g) - \log 4$$



Источник: [Why it is so hard to train Generative Adversarial Networks!](#)

WGAN: Сравнение разных дивергенций

Расстояние Wasserstein-1 (Earth Mover distance)

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} \|x - y\|$$

где $\Pi(x, y)$ – множество всех таких совместных распределений $\gamma(x, y)$, что маргинальные распределения, получаемые из него, равны $\mathbb{P}_r, \mathbb{P}_g$ соответственно

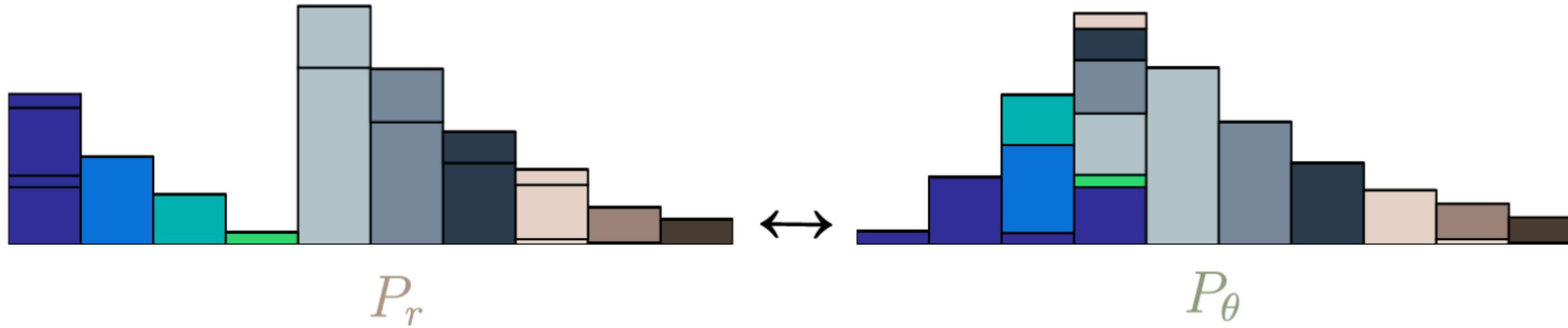
WGAN: Сравнение разных дивергенций

Задача: перераспределить «массу» так, чтобы распределения совпадали



Источник: [Wasserstein GAN and the Kantorovich-Rubinstein Duality](#)

WGAN: Сравнение разных дивергенций



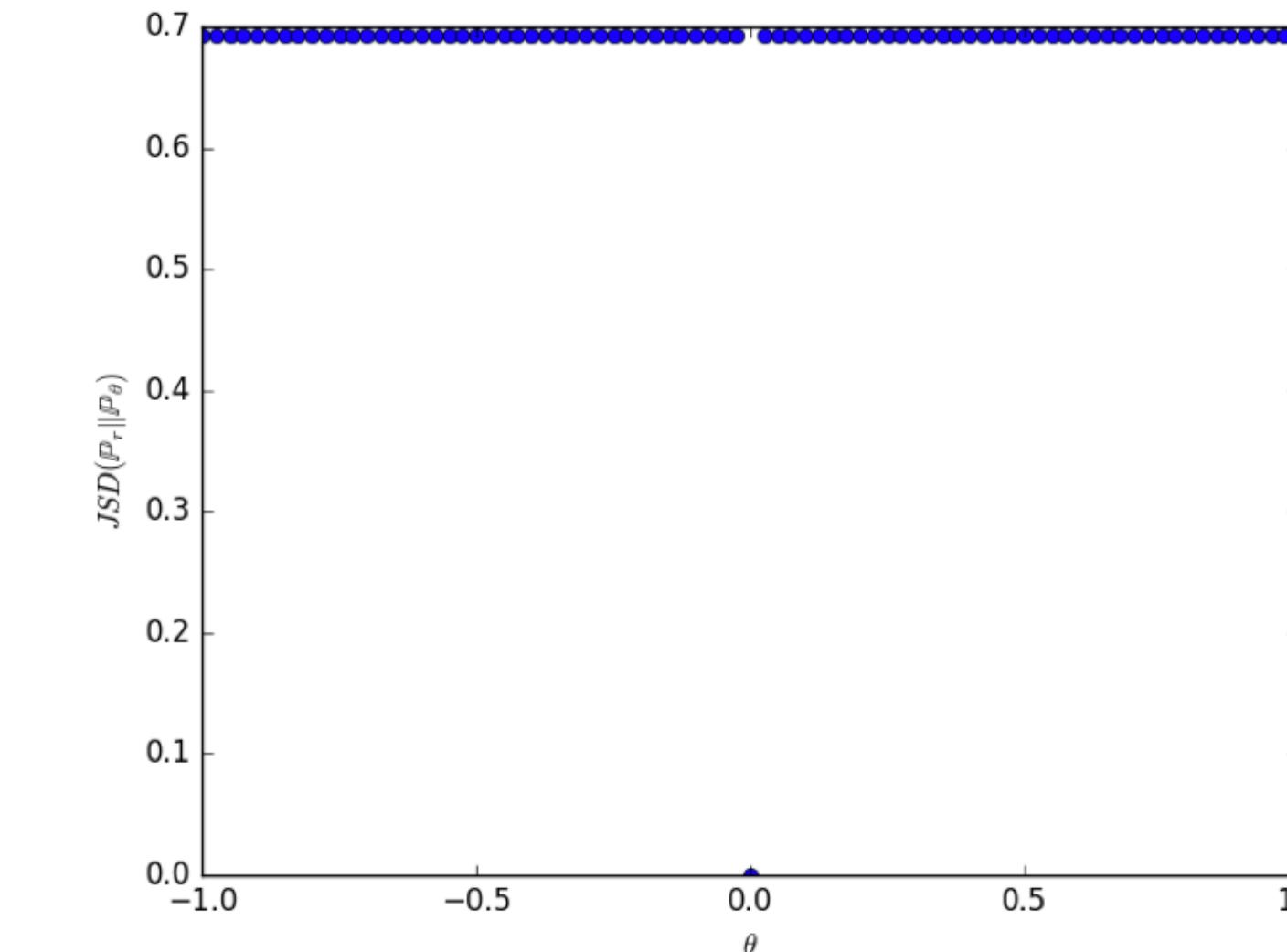
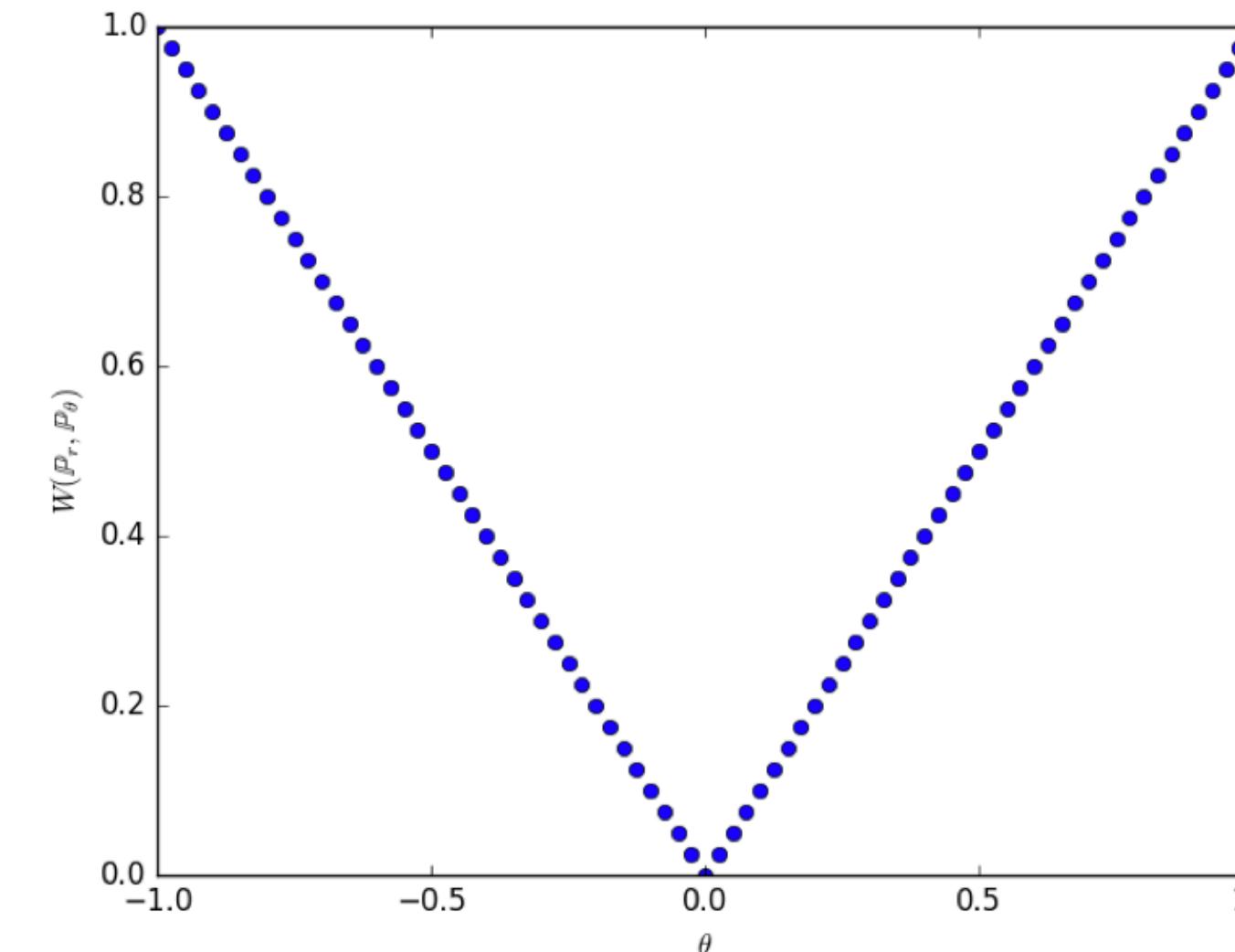
Источник: [Wasserstein GAN and the Kantorovich-Rubinstein Duality](#)

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} \|x - y\|; \|x - y\| - \text{«длина» перемещения, } \gamma - \text{«вес»}$$

WGAN: Сравнение разных дивергенций

Пусть $Z \sim U[0,1]$, \mathbb{P}_0 – распределение $(0, Z)$ и $g_\theta(z) = (\theta, z)$, а \mathbb{P}_θ – распределение $g_\theta(Z)$, тогда:

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \log 2$, если $\theta \neq 0$ и 0 иначе
- $KL(\mathbb{P}_0 \parallel \mathbb{P}_\theta) = KL(\mathbb{P}_\theta \parallel \mathbb{P}_0) = \infty$, если $\theta \neq 0$ и 0 иначе



WGAN: Свойства

При определённых несильных ограничениях на $g_\theta(z)$ и $p(z)$ $W(\mathbb{P}_0, \mathbb{P}_\theta)$ непрерывная и почти всюду дифференцируемая функция.

Проблема: сложно считать по определению

WGAN: Свойства

При определённых несильных ограничениях на $g_\theta(z)$ и $p(z)$ $W(\mathbb{P}_0, \mathbb{P}_\theta)$ непрерывная и почти всюду дифференцируемая функция.

Проблема: сложно считать по определению, но можно показать, что:

$$W(\mathbb{P}_0, \mathbb{P}_\theta) = \sup_{\|f\| \leq 1} \mathbb{E}_{x \sim \mathbb{P}_0}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

где $\|f\| \leq 1$ – множество 1-липшицевых отображений

WGAN: ограничение на критика

Функция $f(x)$ называется K -липшицевой, если:

$$\forall x, y \in Dom(f) \mid f(x) - f(y) \mid \leq K \mid x - y \mid$$

При обучении приближения $f(x)$ функцией $f_w(x)$ будет ограничивать w так, чтобы он лежал в компактном множестве, например, в $[-c, c]^l$

WGAN: ограничение на критика

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α , the learning rate. c , the clipping parameter. m , the batch size.
 n_{critic} , the number of iterations of the critic per generator iteration.

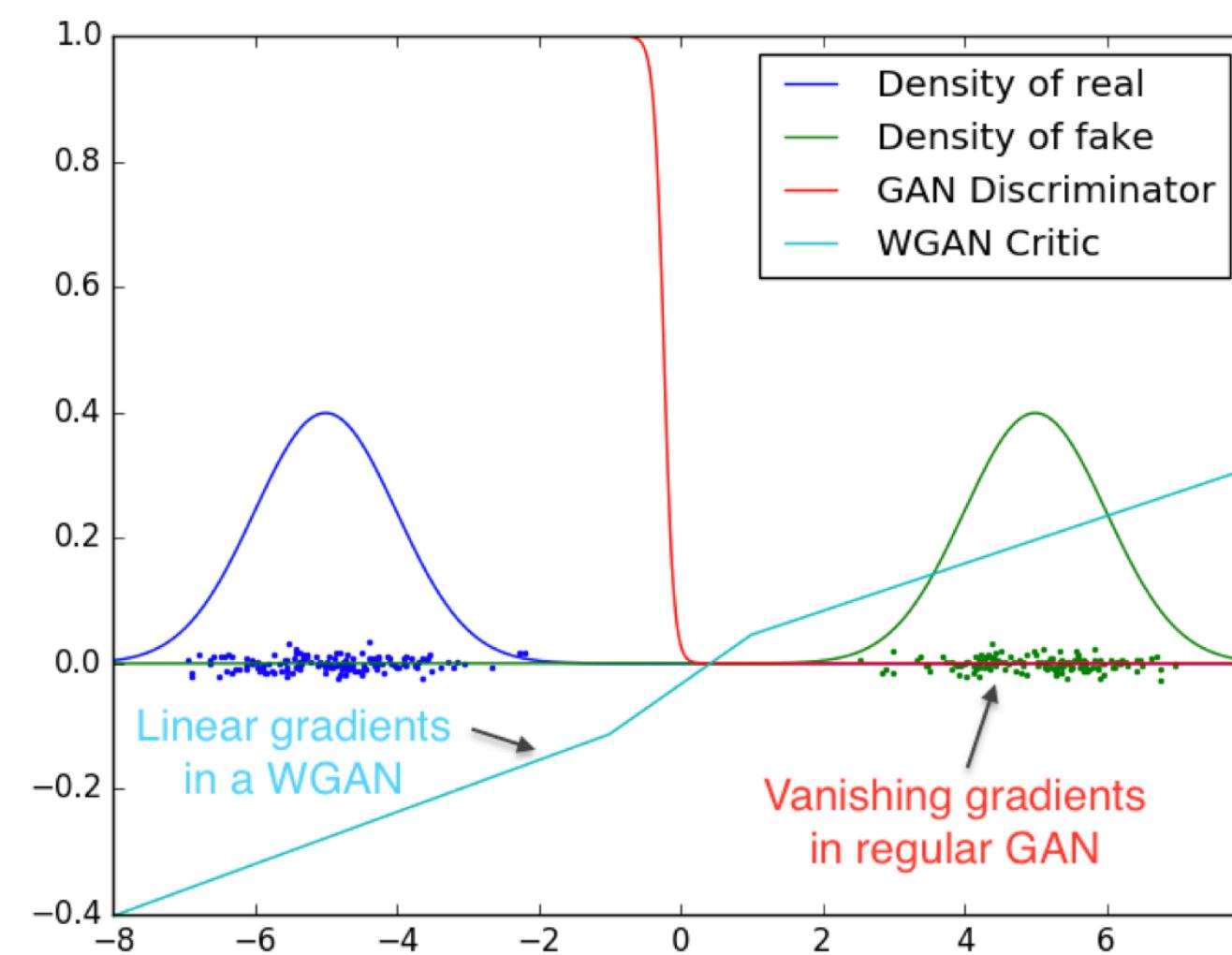
Require: : w_0 , initial critic parameters. θ_0 , initial generator's parameters.

```
1: while  $\theta$  has not converged do
2:   for  $t = 0, \dots, n_{\text{critic}}$  do
3:     Sample  $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$  a batch from the real data.
4:     Sample  $\{z^{(i)}\}_{i=1}^m \sim p(z)$  a batch of prior samples.
5:      $g_w \leftarrow \nabla_w \left[ \frac{1}{m} \sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)})) \right]$ 
6:      $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$ 
7:      $w \leftarrow \text{clip}(w, -c, c)$ 
8:   end for
9:   Sample  $\{z^{(i)}\}_{i=1}^m \sim p(z)$  a batch of prior samples.
10:   $g_\theta \leftarrow -\nabla_\theta \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)}))$ 
11:   $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$ 
12: end while
```

WGAN: достоинства

Можно обучать критика, пока он не станет оптимальным:

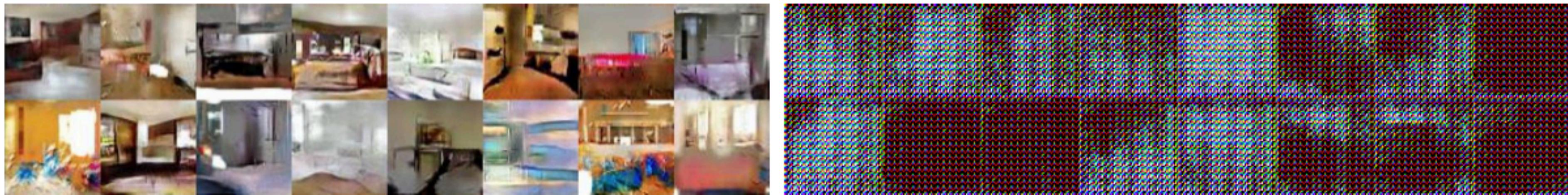
- Нет затухания градиента
- Нет модального коллапса
- Значение функции потерь коррелирует с качеством генерируемых примеров



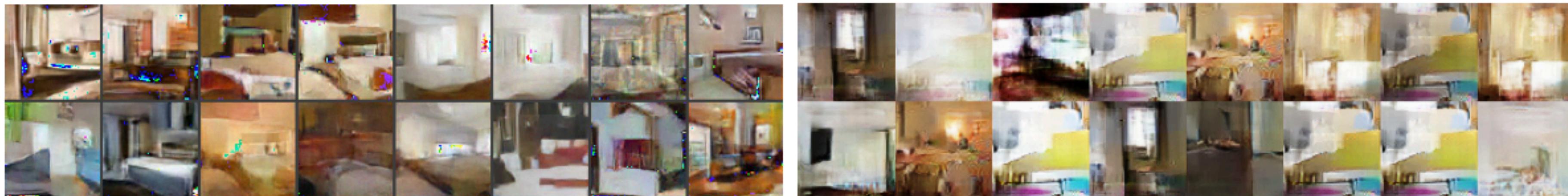
Сравнение дискриминатора и критика

Источник: [Wasserstein GAN](#)

WGAN: достоинства



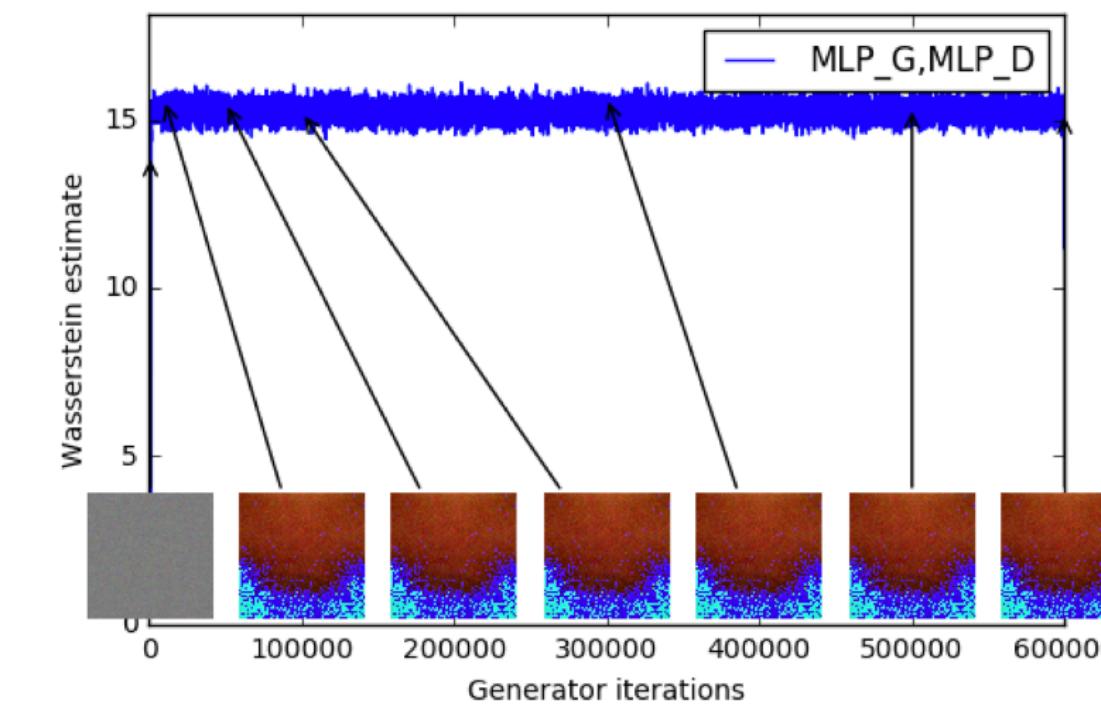
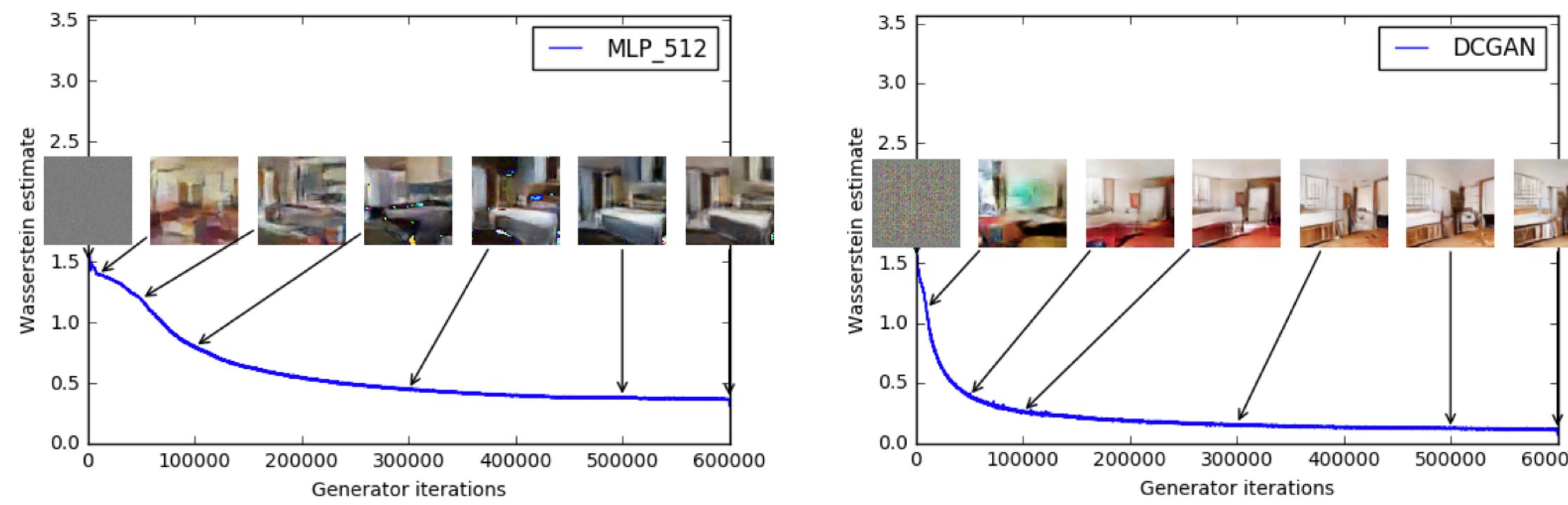
DCGAN без batchnorm, с фиксированным количеством фильтров на каждом слое.
Слева: WGAN, Справа: классический GAN



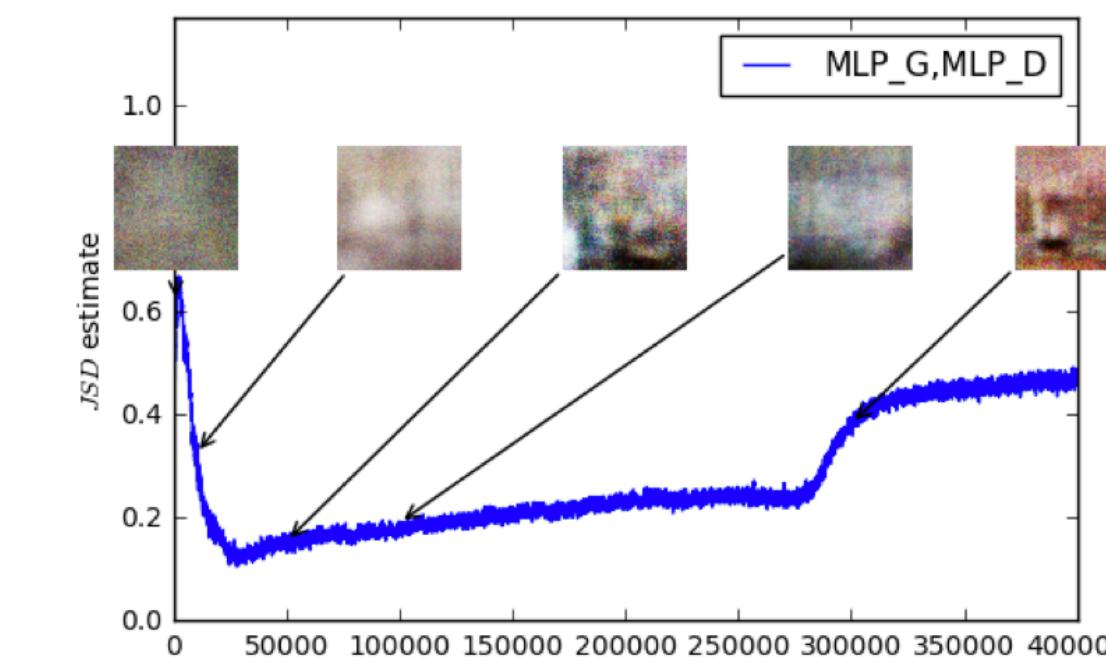
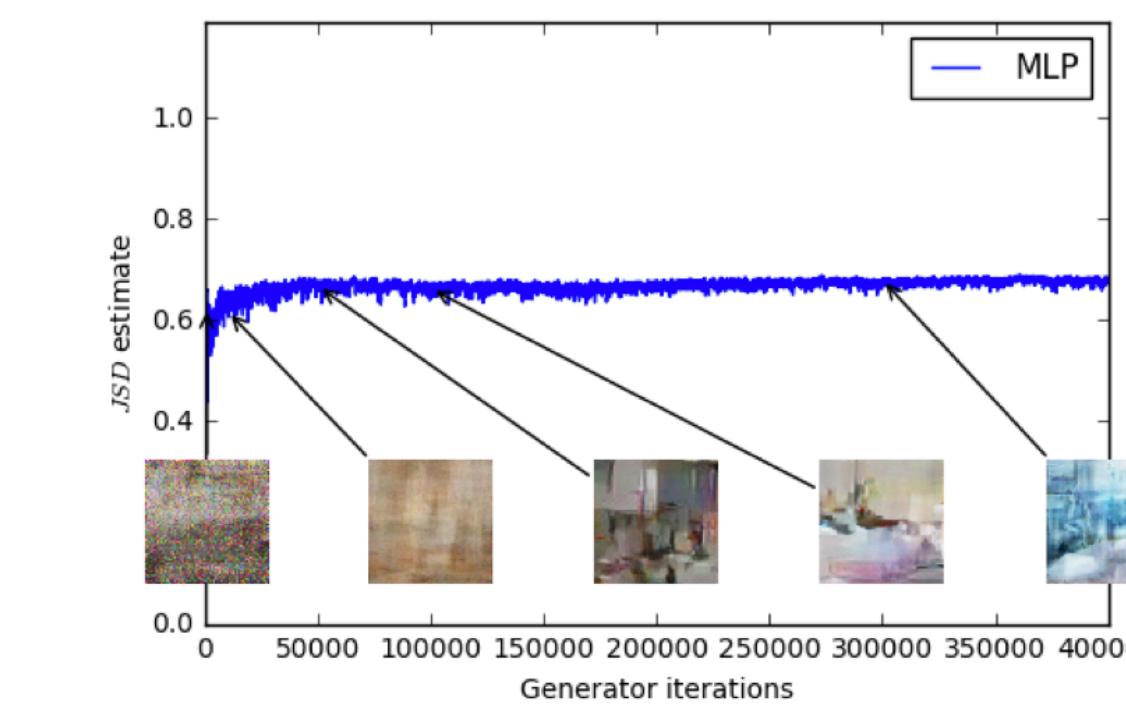
MLP. Слева: WGAN, Справа: классический GAN

Источник: [Wasserstein GAN](#)

WGAN: достоинства



Кривые обучения при WGAN

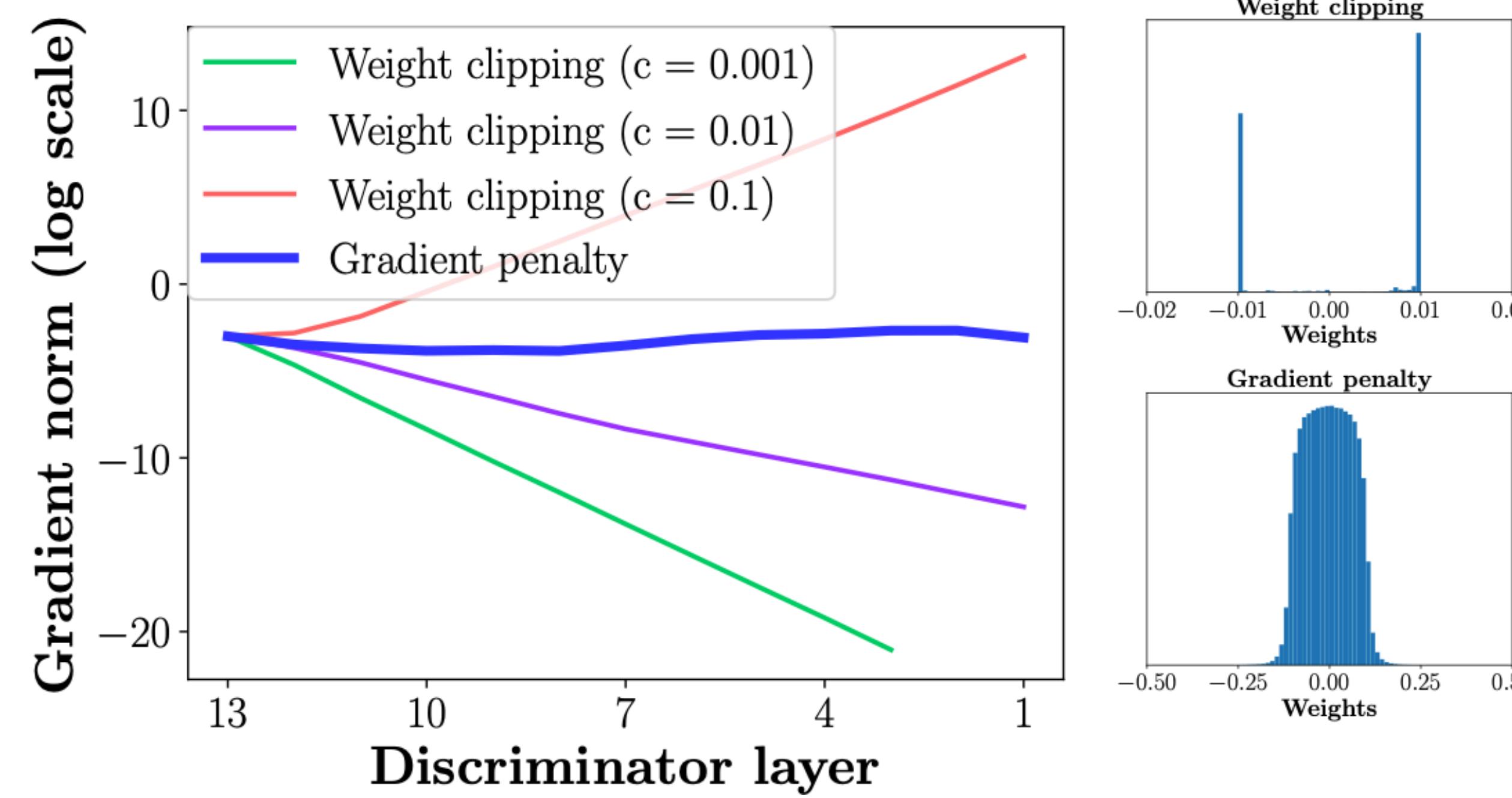


Кривые обучения при JS

Источник: [Wasserstein GAN](#)

WGAN: недостатки

- Ограничение на веса критика – очень грубый способ сделать функцию K -липшицевой
- Нестабильность обучения при использовании оптимизаторов с инерцией (momentum)



Источник: [Improved Training of Wasserstein GANs](#)

WGAN-GP

$$\|f(x)\| \leq 1 \Leftrightarrow \|\nabla_x f(x)\| \leq 1$$

WGAN-GP

$$\|f(x)\| \leq 1 \Leftrightarrow \|\nabla_x f(x)\| \leq 1$$

$$L = \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_t}[(\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1)^2]$$

-WGAN loss

Gradient Penalty

где \mathbb{P}_t – распределение точек на отрезке между $x \sim \mathbb{P}_r$ и $g_\theta(z)$

WGAN-GP

$$\|f(x)\| \leq 1 \Leftrightarrow \|\nabla_x f(x)\| \leq 1$$

$$L = \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_t}[(\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1)^2]$$

-WGAN loss

Gradient Penalty

где \mathbb{P}_t – распределение точек на отрезке между $x \sim \mathbb{P}_r$ и $g_\theta(z)$

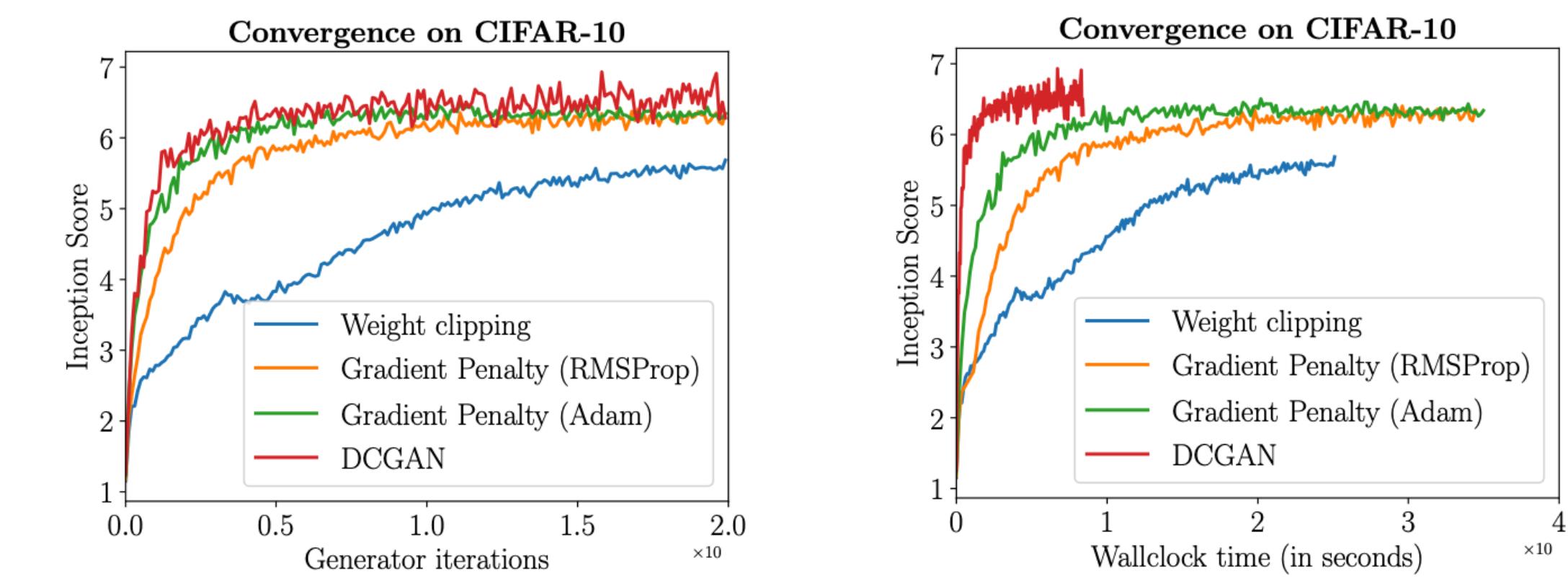
- Сэмплируем $x \sim \mathbb{P}_r, z \sim \mathbb{P}_z, \epsilon \sim U[0,1]$, тогда $\hat{x} = \epsilon x + (1 - \epsilon)g_\theta(z)$
- Считаем $L^{(i)}$
- Повторяем m раз

Обновляем $f_w(x)$ с помощью $\nabla_w \sum_i^m L^{(i)}$; Не используем BN в архитектуре критика

WGAN-GP



Сравнение изображений, сгенерированных разными моделями с разными функциями потерь



Сравнения кривых обучения

Источник: [Improved Training of Wasserstein GANs](#)

Progressive growing of GANs

Хотим генерировать изображения высокого разрешения:



Изображения размера 1024 × 1024, сгенерированные GAN

Источник: [*Progressive Growing of GANs for Improved Quality, Stability, and Variation*](#)

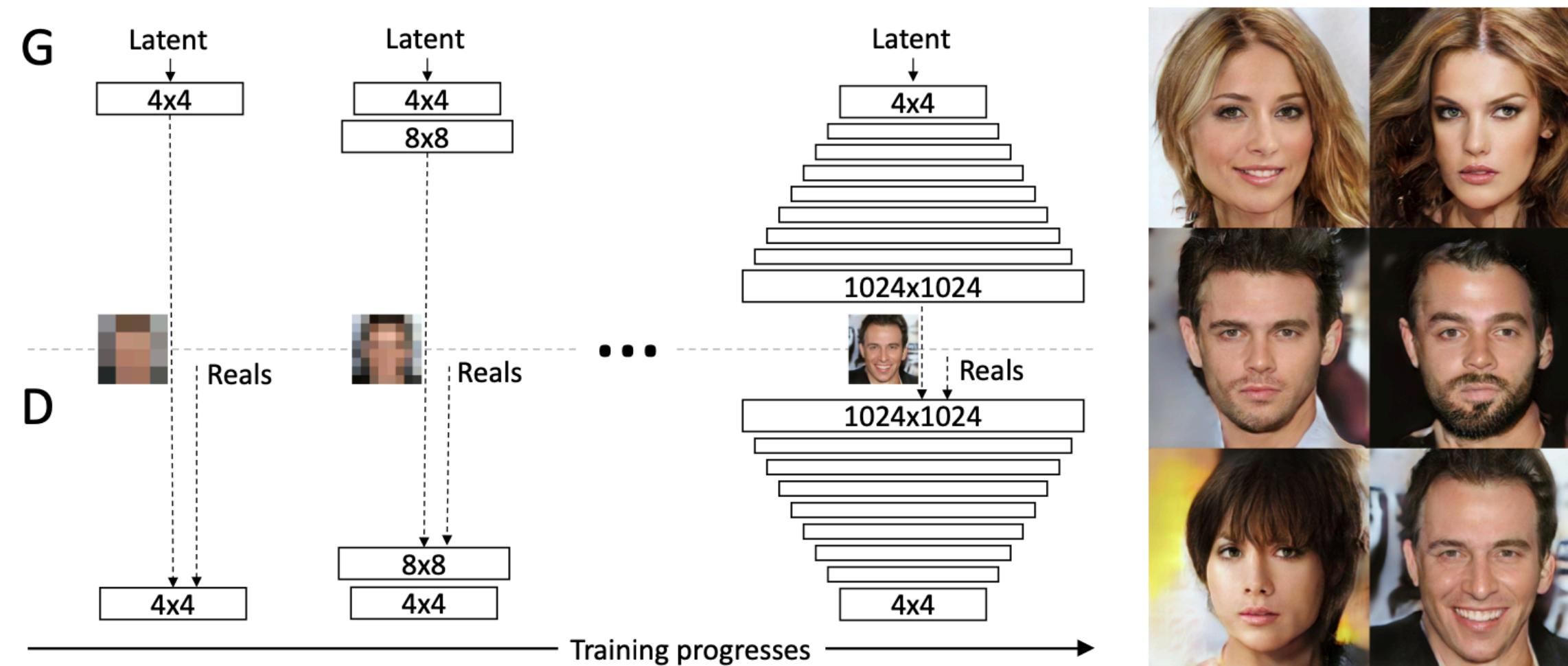
Progressive growing of GANs

Обучать генерировать сразу детализированные изображения сложно и долго.

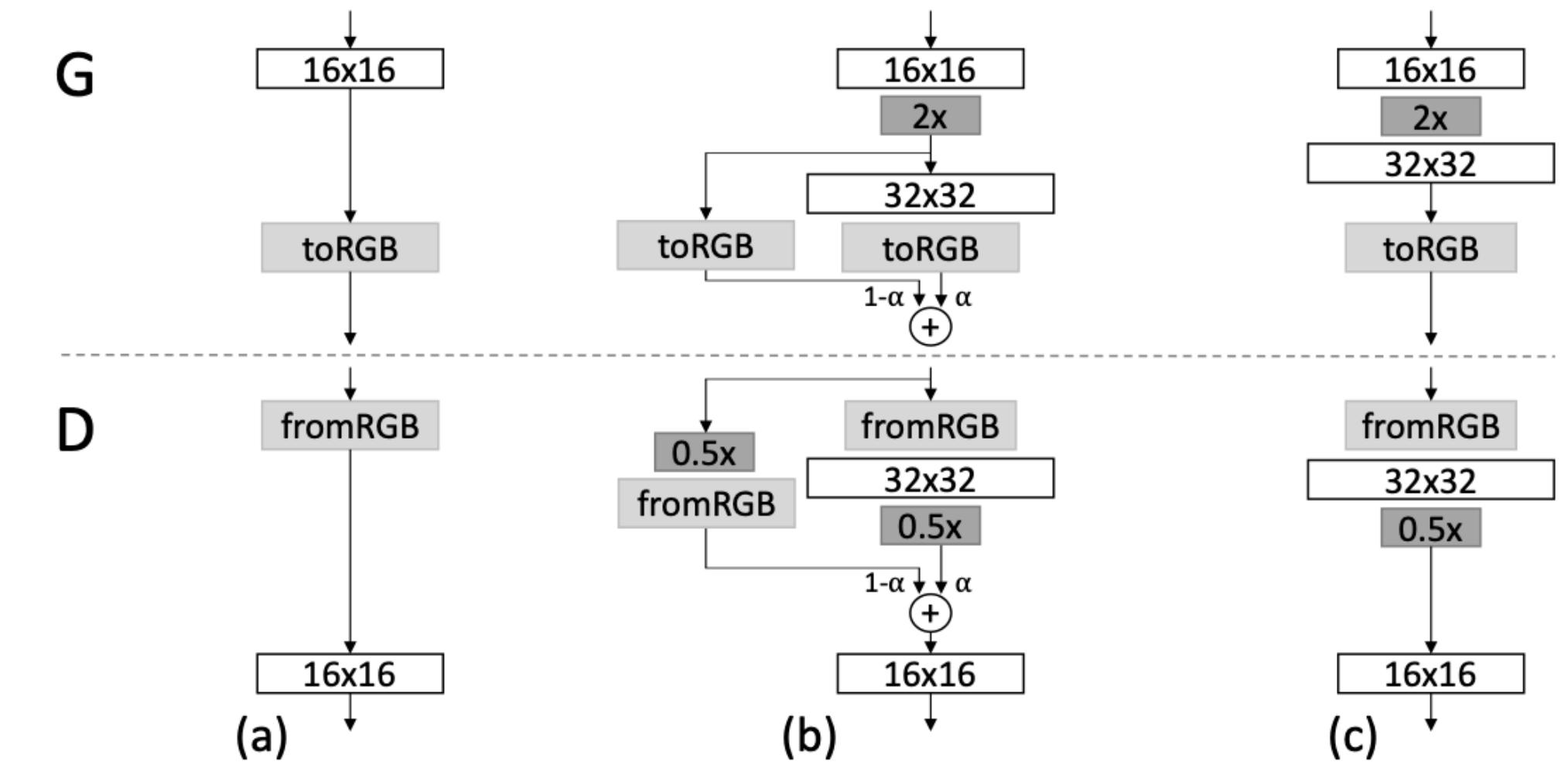
Вместо этого будем постепенно добавлять слои и увеличивать разрешение, почему это помогает:

- Усложняем задачу дискриминатору
- Модель сначала учится признакам большого масштаба, а затем уточняет изображение, добавляя детали
- По мере роста модели даём ей простые задачи, а не сразу одну сложную

Progressive growing of GANs



Начинаем с изображения 4x4. Когда GAN обучается, добавляем слои в G и D, повышая тем самым разрешение. Все предыдущие слои остаются обучаемыми



При добавлении новых слоев комбинируем изображения с разных этапом с учётом затемнения. Для того, чтобы размеры совпадали, используем NN filtering для upsampling, а для downsampling average pooling

Progressive growing of GANs: детали

Увеличение разнообразия с помощью `minibatch standard deviation`:

- В последнем блоке дискриминатора (перед финальными свёртками и линейным слоем) считаем в батче стандартное отклонение значений карт признаков
- Усредняем `std` по всем пикселям и всем измерениям
- Создаём карту признаков, заполненную полученным значением
- Добавляем её к исходному набору карт признаков

Progressive growing of GANs: детали

Инициализация весов

- Инициализируем веса с помощью стандартного нормального распределения
- Во время обучения на каждом слое умножаем веса слоя на $\sqrt{2/f_{in}}$
- «The benefit of doing this dynamically instead of during initialization is somewhat subtle»

Progressive growing of GANs: детали

Нормализация карт признаков:

$$b_{x,y} := \frac{a_{x,y}}{\sqrt{\frac{1}{N} \sum_{j=0}^{N-1} (a_{x,y}^j)^2 + \epsilon}}$$

где $a_{x,y}$ – вектор значений карт признаков в позиции (x, y) , N – количество карт признаков

Метрики качества

Fréchet Inception Distance

- Получаем векторные представления (embedding) изображений
- Рассматриваем их как многомерные нормальные векторы

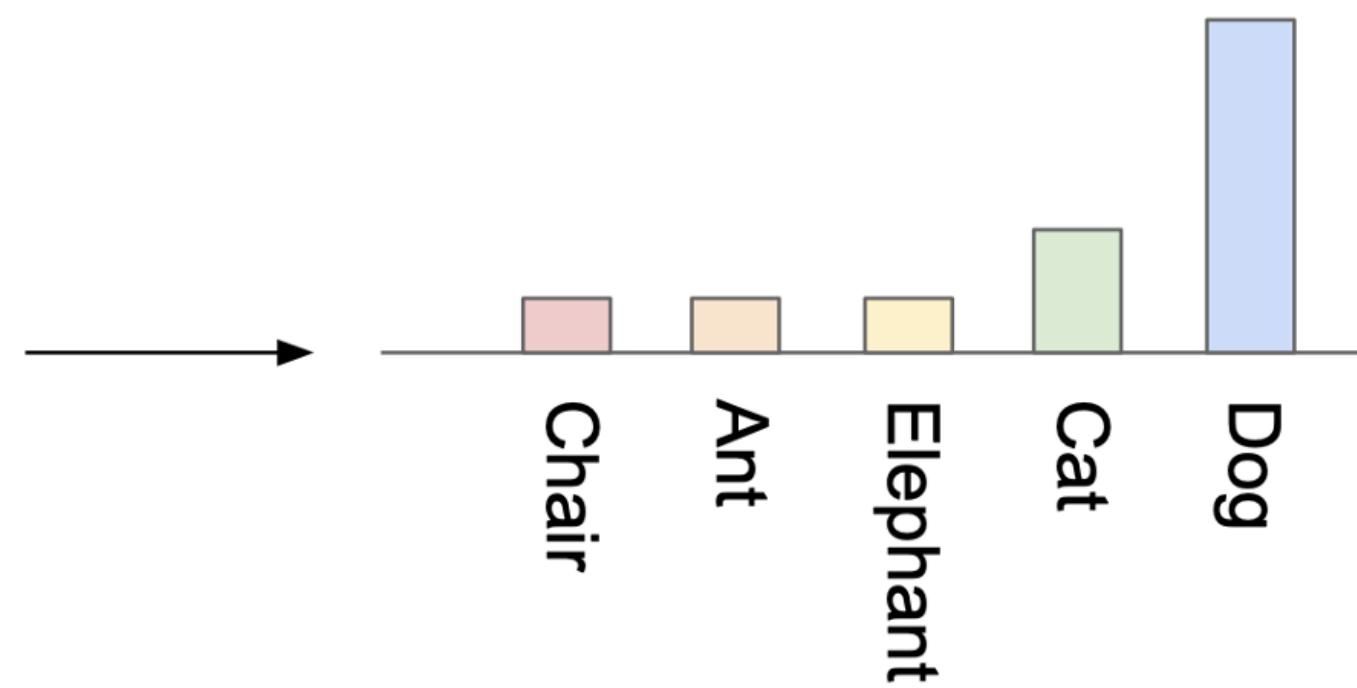
$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + \text{tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}),$$

где μ_x, μ_g – выборочные средние, Σ_x, Σ_g – выборочные матрицы ковариаций

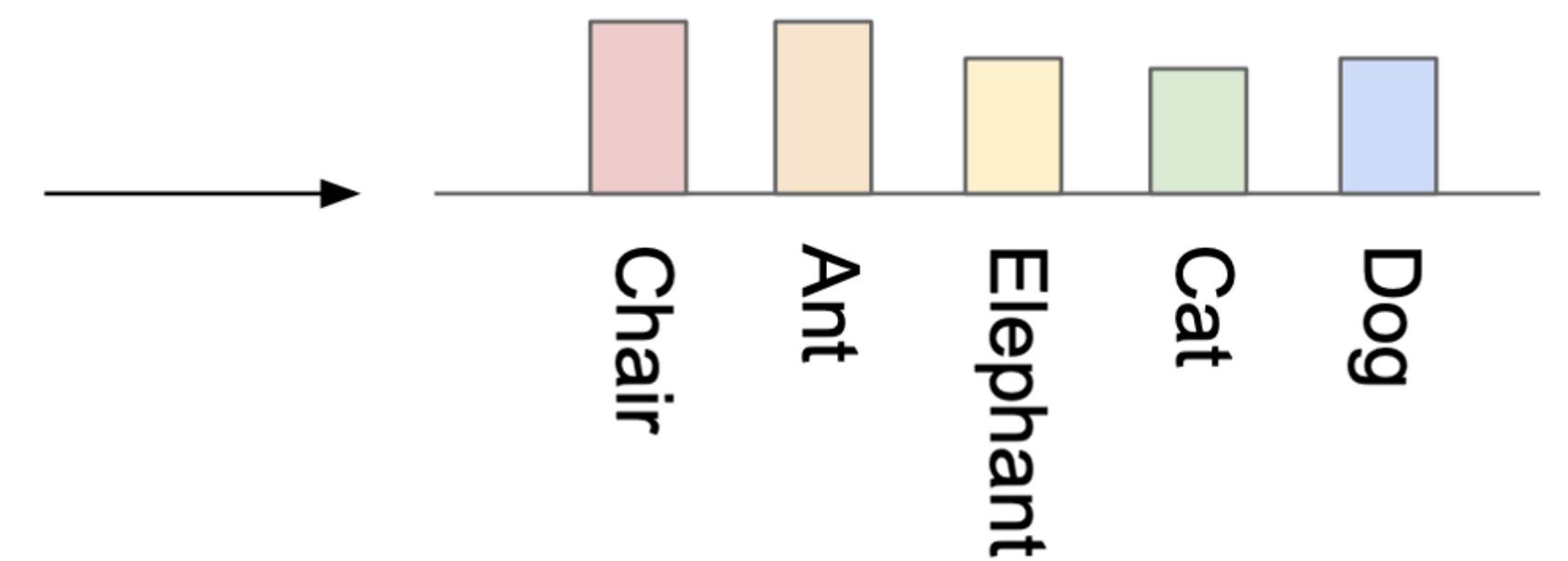
Метрики качества

Inception Score

Хотим измерять, насколько реалистичные и разнообразные картинки генерируются



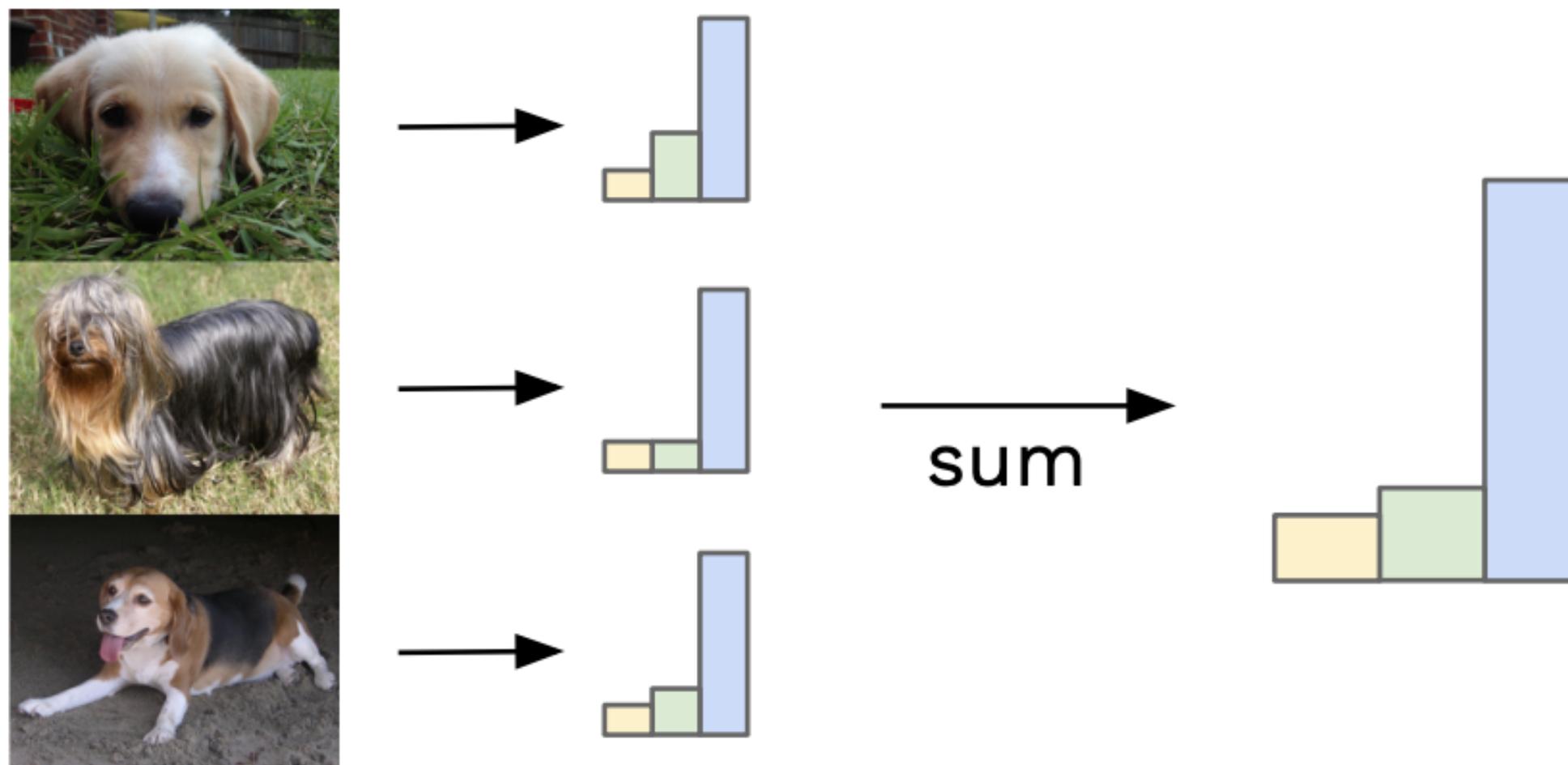
Отчётливо классифицируемый объект



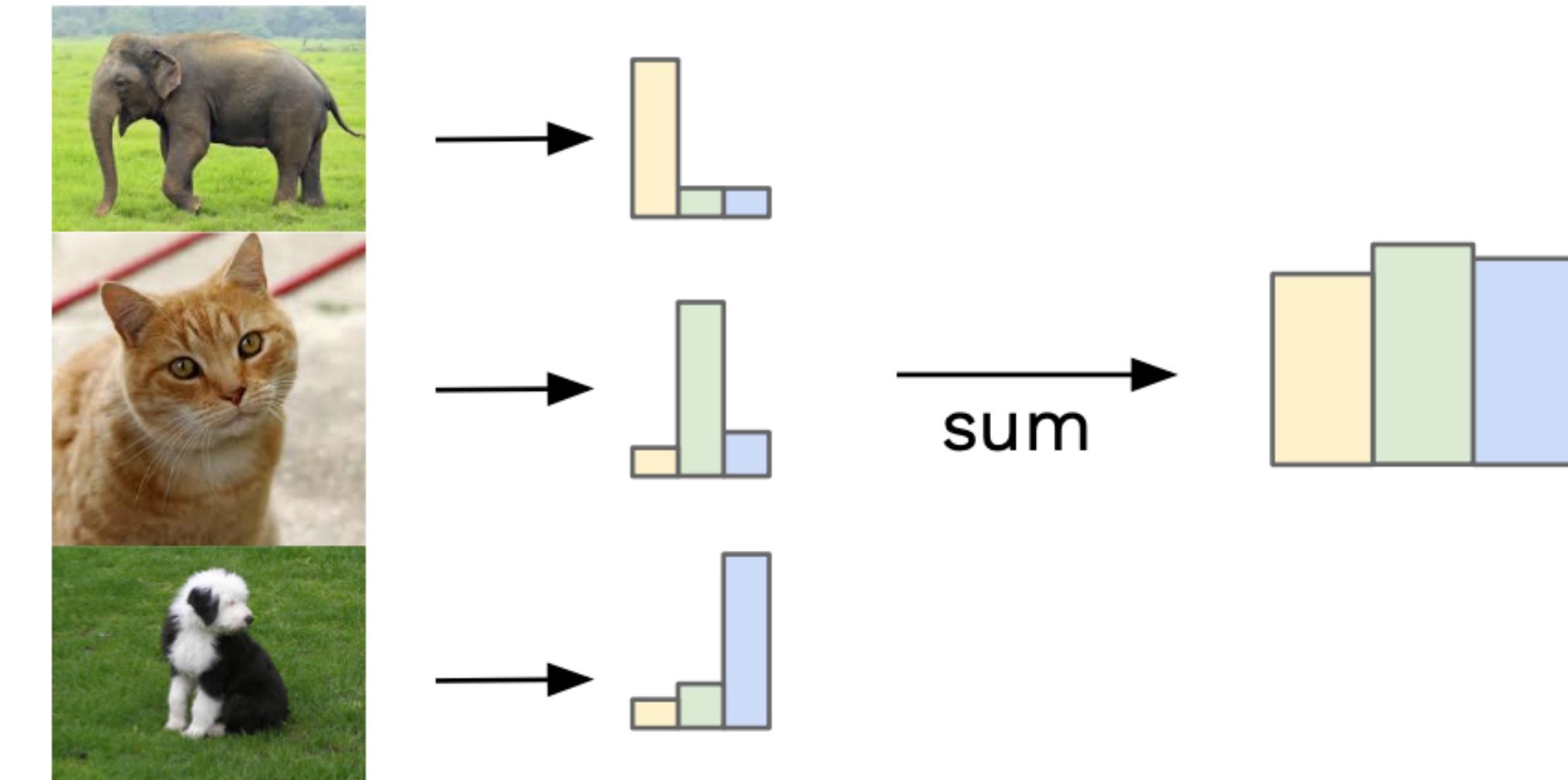
Нет конкретного объекта

Метрики качества

Similar labels sum to give focussed distribution



Different labels sum to give uniform distribution



Маргинальное распределение, когда изображения однообразные

Маргинальное распределение, когда изображения разнообразные

Источник: [A simple explanation of the Inception Score](#)

Метрики качества

Алгоритм подсчёта Inception Score:

- С помощью обученного классификатора получаем распределения $p(y | x = G(z))$
- Считаем маргинальное распределение усреднением $p(y | x = G(z))$
- $IS = \exp[KL(p(y | x) \| p(y))]$

Итог: имеет ли всё это смысл?

Использую NSGAN, думая, что тогда градиент не будет затухать



Использую WGAN, который устойчив к модальному коллапсу и затуханию градиента

Использую WGAN-GP, заменил weight clipping на осмысленную регуляризацию на градиент

Перебираю гиперпараметры всего* месяц и средней моделью бью по показателям state-of-the-art подходы

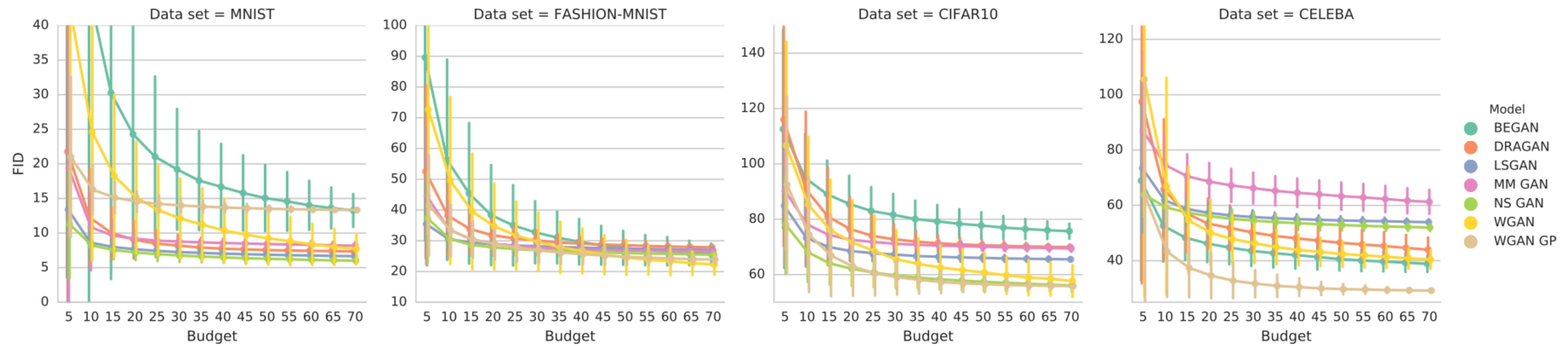
- β_1 : the parameter of the Adam optimization algorithm.
- Learning rate: generator/discriminator learning rate.
- λ : Multiplier of the gradient penalty for DRAGAN and WGAN GP. Learning rate for k_t in BEGAN.
- Disc iters: Number of discriminator updates per one generator update.
- batchnorm: If True, the batch normalization will be used in the discriminator.
- γ : Parameter of BEGAN.
- clipping: Parameter of WGAN, weights will be clipped to this value.

Перебираемые параметры в экспериментах Google

Источник: [Are All GANs Created Equal](#)

*по мнению исследователей из Google

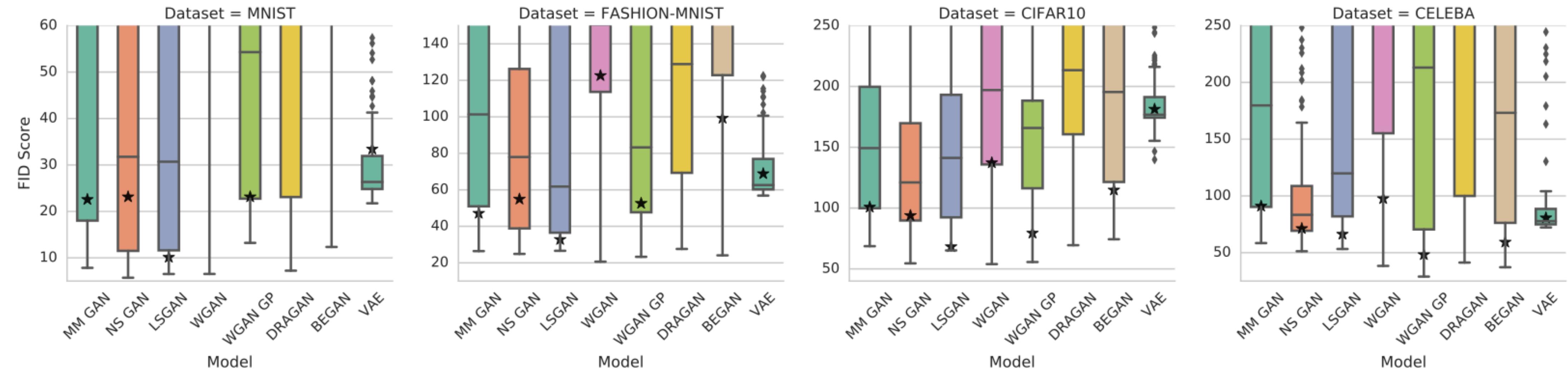
Итог: имеет ли всё это смысл?



Зависимость распределения минимального FID от бюджета. Для каждого «бюджета» были подсчитаны матожидания и стандартные отклонения, используя 5000 бутстррап выборок, полученных из 100 запусков

Источник: [Are All GANs Created Equal](#)

Итог: имеет ли всё это смысл?



Зависимость FID от гиперпараметров. Звёздочка – FID при параметрах, предложенных авторами

Источник: [Are All GANs Created Equal](#)

Что узнали?

- Основные проблемы GAN и их причины
- NSGAN, WGAN, WGAN-GP
- Progressive GAN
- Fréchet Inception Distance, Inception score
- Подбор гиперпараметров важен

Источники

- *Generative Adversarial Networks.* Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio
- *Wasserstein GAN.* Martin Arjovsky, Soumith Chintala, Léon Bottou
- *Improved Training of Wasserstein GANs.* Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville
- *Progressive Growing of GANs for Improved Quality, Stability, and Variation.* Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen
- *Are GANs Created Equal? A Large-Scale Study.* Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, Olivier Bousquet

Вопросы

- Напишите WGAN Loss, объясните, почему оптимизация производится по 1-липшицевым функциям и что нужно делать при обучении, чтобы это ограничение выполнялось?
- Напишите WGAN и WGAN-GP Loss и объясните зачем нужна регуляризация нормы градиента. Какая проблема WGAN решена в WGAN-GP?
- Опишите метод progressive training, а также метод увеличения разнообразия генерируемых изображений?
- Опишите способ подсчёта Inception Score. Что означает высокий/ средний/низкий показатель IS? (Если успеем)