f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Sebastian Nowozin, Botond Cseke, Ryota Tomioka

Machine Intelligence and Perception Group, Microsoft Research, Cambridge, UK

2016

О чем статья

- Обобщили objective function GAN для любых f-дивергенций
- Предложили более простой алгоритм с доказательством локальной сходимости
- Продемонстрировали работу с разными дивергенциями

f-divergence: определение

Для двух распределений P и Q с абсолютно непрерывными плотностями p и q определим f-дивергенцию, также известную как Ali-Silvey distance:

$$D_f(P||Q) = \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx,$$

где $f : \mathbb{R}_+ \to \mathbb{R}$ - выпуклая, полунепрерывная снизу функция с условием $f(1) = 0$, называемая generator function.

Сопряженная по Фенхелю функция: определение

Для каждой выпуклой, полунепрерывной снизу функции f существует сопряженная функция (convex conjugate), также называемая сопряженная по Фенхелю (Fenchel conjugate), определяемая как:

$$f^*(t) = \sup_{u \in \mathrm{dom}_f} \{ut - f(u)\} dx,$$

также выпуклая и полунепрерывная снизу.

$$f^{**} = f$$

Вывод задачи f-GAN

Представим f как сопряженную f^*:

$$D_f(P||Q) = \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx =$$
$$= \int_X q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx$$

Вывод задачи f-GAN

Проблема: как вытащить супремум.

$$
\begin{aligned}
D_f(P||Q) &= \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx = \\
&= \int_X q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx = \\
&\geq \sup_{T \in \tau} \left(\int_X p(x) T(x) dx - \int_X q(x) f^*(T(x)) dx \right)
\end{aligned}
$$

τ - произвольный класс функций $T : X \to \mathbb{R}$

Вывод задачи f-GAN

Перейдем к матожиданиям

$$
\begin{aligned}
D_f(P||Q) &= \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx = \\
&= \int_X q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx = \\
&\geq \sup_{T \in \mathcal{T}} \left(\int_X p(x) T(x) dx - \int_X q(x) f^*(T(x)) dx \right) = \\
&= \sup_{T \in \mathcal{T}} (\mathbb{E}_{x \sim P}[T(X)] - \mathbb{E}_{x \sim Q}[f^*(T(x))])
\end{aligned}
$$

Вывод задачи f-GAN

Name	$D_f(P\|Q)$	Generator $f(u)$	$T^*(x)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$	$u \log u$	$1 + \log \frac{p(x)}{q(x)}$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}x$	$(u-1)^2$	$2(\frac{p(x)}{q(x)} - 1)$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$\left(\sqrt{u}-1\right)^2$	$(\sqrt{\frac{p(x)}{q(x)}} - 1) \cdot \sqrt{\frac{q(x)}{p(x)}}$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$	$\log \frac{2p(x)}{p(x)+q(x)}$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$	$\log \frac{p(x)}{p(x)+q(x)}$

Рис.: Функции-генераторы и оптимальные вариационные функции для различных дивергенций

Постановка задачи f-GAN

Параметризуем Q вектором θ, T вектором ω
Генеративная модель Q_θ получается при нахождении седловой точки следующей функции:

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}[T_\omega(x)] - \mathbb{E}_{x \sim Q_\theta}[f^*(T_\omega(x))]$$

Минимизируем по θ, максимизируем по ω.

Сравнение objectives GAN и f-GAN

GAN:

$$\min_\theta \max_\omega (\mathbb{E}_{x\sim P}[\log(D_\omega(x))] - \mathbb{E}_{x\sim Q_\theta}[\log(1 - D_\omega(x))])$$

f-GAN:

$$\min_\theta \max_\omega (\mathbb{E}_{x\sim P}[T_\omega(x)] - \mathbb{E}_{x\sim Q_\theta}[f^*(T_\omega(x))])$$

- GAN - частный случай, соответствующий $\log(D_\omega(x)) = T_\omega(x)$
- GAN минимизируют дивергенцию Йенсена-Шеннона

Представление вариационной функции

Проблема: f* для нек. дивергенций определена не на R, а на части
Поэтому на практике нужна вспомогательная функция.
Определим

$$T_\omega(x) = g_f(V_\omega(x)),$$

где $V_\omega : X \to \mathbb{R}$, $g_f : \mathbb{R} \to \mathrm{dom}_{f^*}$
Тогда целевая функция будет

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

Представление вариационной функции: примеры

Name	Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$	$f'(1)$
Total variation	$\frac{1}{2}\tanh(v)$	$-\frac{1}{2} \leq t \leq \frac{1}{2}$	t	0
Kullback-Leibler (KL)	v	$\mathbb{R}$	$\exp(t-1)$	1
Reverse KL	$-\exp(v)$	$\mathbb{R}_-$	$-1-\log(-t)$	-1
Pearson χ^2	v	$\mathbb{R}$	$\frac{1}{4}t^2+t$	0
Neyman χ^2	$1-\exp(v)$	$t<1$	$2-2\sqrt{1-t}$	0
Squared Hellinger	$1-\exp(v)$	$t<1$	$\frac{t}{1-t}$	0
Jeffrey	v	$\mathbb{R}$	$W(e^{1-t})+\frac{1}{W(e^{1-t})}+t-2$	0
Jensen-Shannon	$\log(2)-\log(1+\exp(-v))$	$t<\log(2)$	$-\log(2-\exp(t))$	0
Jensen-Shannon-weighted	$-\pi\log\pi-\log(1+\exp(-v))$	$t<-\pi\log\pi$	$(1-\pi)\log\frac{1-\pi}{1-\pi e^{t/\pi}}$	0
GAN	$-\log(1+\exp(-v))$	$\mathbb{R}_-$	$-\log(1-\exp(t))$	$-\log(2)$
α-div. ($\alpha<1, \alpha\neq 0$)	$\frac{1}{1-\alpha}-\log(1+\exp(-v))$	$t<\frac{1}{1-\alpha}$	$\frac{1}{\alpha}(t(\alpha-1)+1)^{\frac{\alpha}{\alpha-1}}-\frac{1}{\alpha}$	0
α-div. ($\alpha>1$)	v	$\mathbb{R}$	$\frac{1}{\alpha}(t(\alpha-1)+1)^{\frac{\alpha}{\alpha-1}}-\frac{1}{\alpha}$	0

Рис.: Функции активации и сопряженные функции, соответствующие разным дивергенциям

Алгоритм: Double-Loop vs Single-Loop

Double-loop алгоритм (Goodfellow et at., 2014):

- Внутренний цикл - k шагов по градиенту лосса дискриминатора
- Внешний цикл - отрицательный шаг по градиенту лосса генератора
- На практике во внутреннем цикле выполняется одна итерация (в итоге два бэкпропа)
- Недостаток теоретического анализа алгоритма

Single-loop алгоритм:

- Градиент по θ и ω считается за один бэкпроп, нет внутреннего цикла
- Доказали локальную геометрическую скорость сходимости

Algorithm 1 Single-Step Gradient Method

function SINGLESTEPGRADIENTITERATION($P, \theta^t, \omega^t, B, \eta$)
  Sample $X_P = \{x_1, \ldots, x_B\}$ and $X_Q = \{x'_1, \ldots, x'_B\}$, from P and Q_{θ^t}, respectively.
  Update: $\omega^{t+1} = \omega^t + \eta \nabla_\omega F(\theta^t, \omega^t)$.
  Update: $\theta^{t+1} = \theta^t - \eta \nabla_\theta F(\theta^t, \omega^t)$.
end function

Эксперимент: смесь гауссиан

- Модель Q_θ принимает на вход $z \sim N(0, 1)$ и возвращает $G_\theta(z) = \mu + \sigma z$
- Вариационная функция - двухслойная NN с *tanh* активациями
- В случае с гауссианами можно непосредственно минимизировать $D_f(P||Q_\theta)$ по θ

Эксперимент: смесь гауссиан

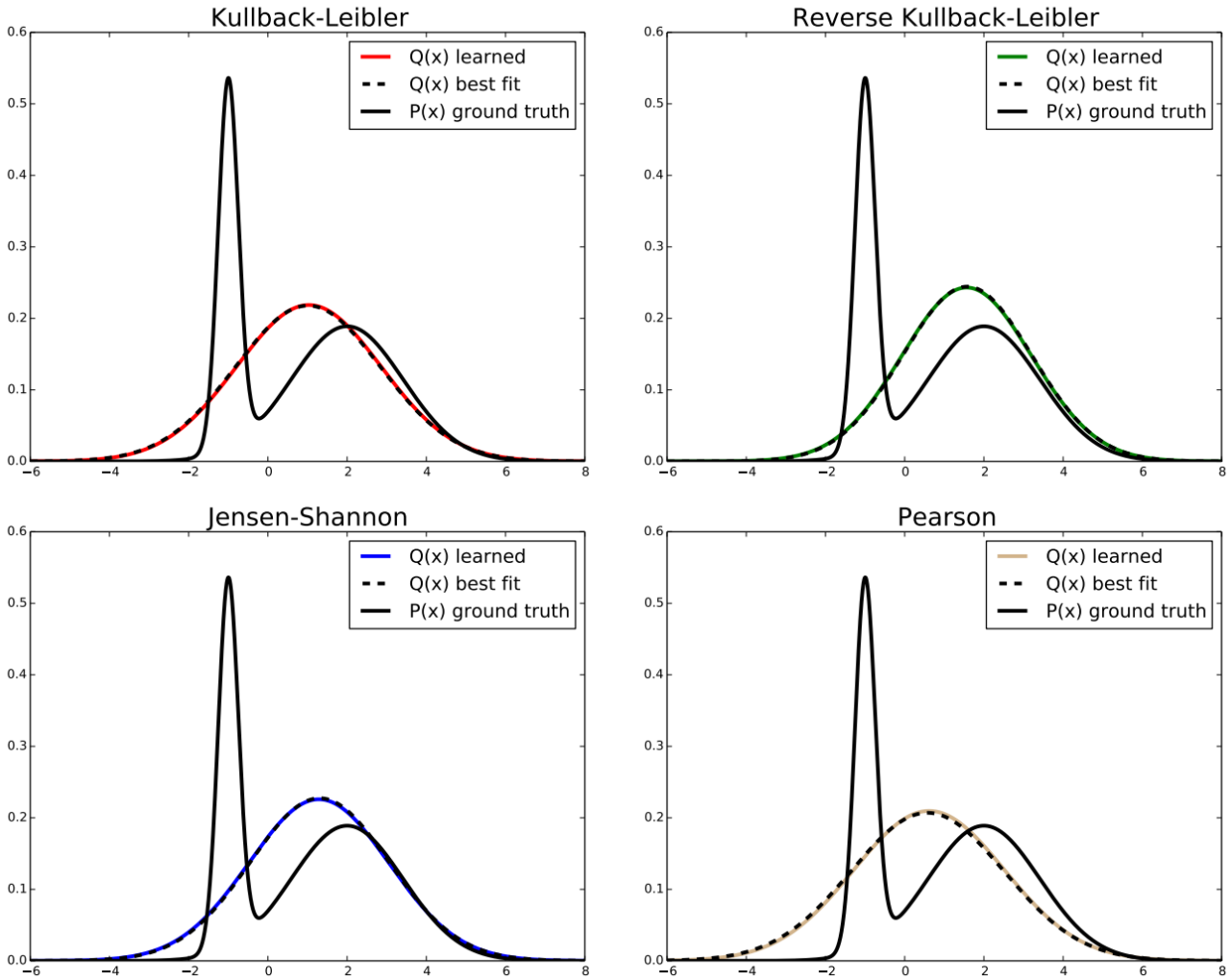


Рис.: Приближение смеси гауссиан путем прямой оптимизации дивергенции $D_f(p||q_{\theta*})$ (пунктир) и оптимизации функции $F(\omega, \theta)$

Эксперимент: смесь гауссиан

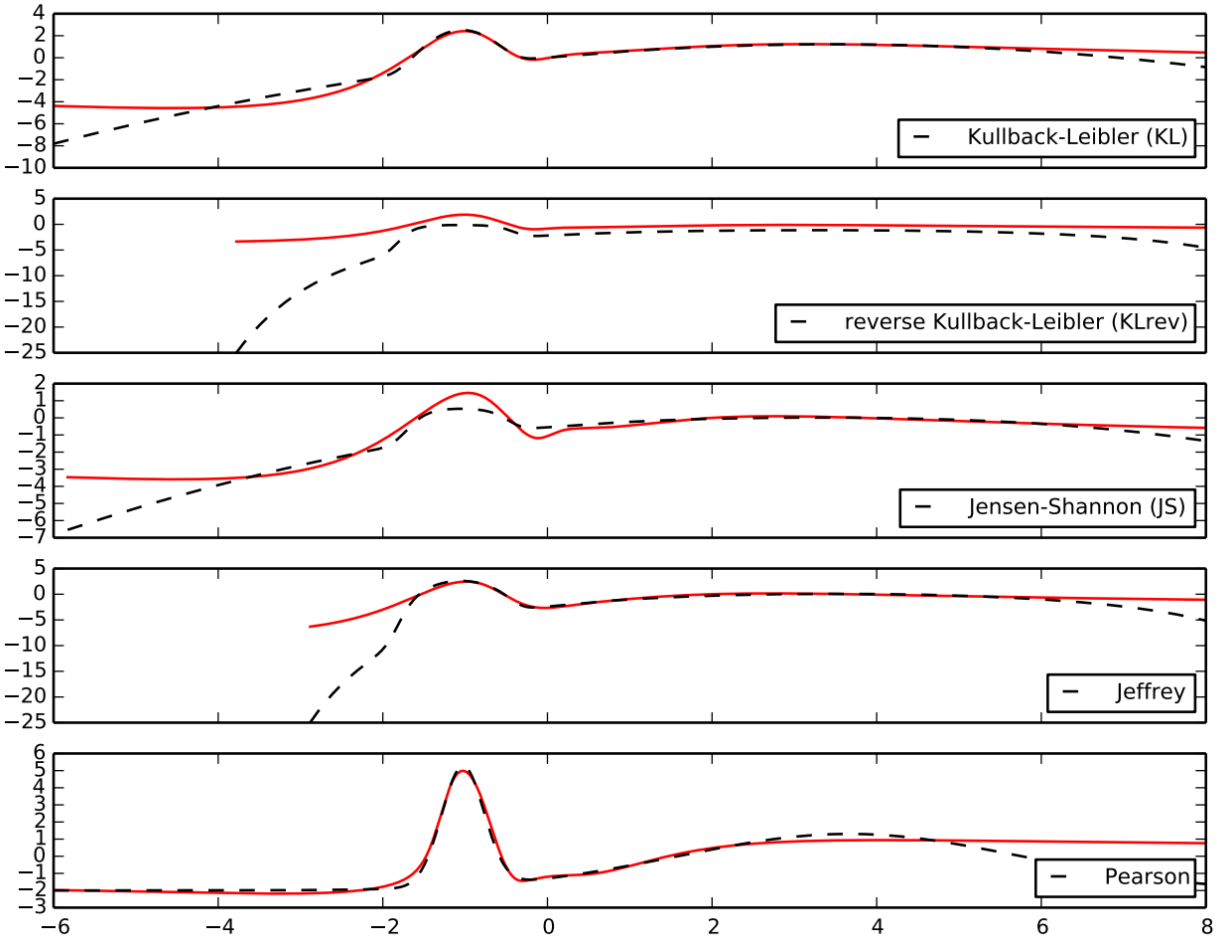


Рис.: Оптимальная вариационная функция T^* (пунктир) и T_ω (красная линия)

Эксперимент: смесь гауссиан

Эксперимент:

1. Обучили T_ω и Q_θ для определенной дивергенции
2. Взяли T_ω от другой дивергенции и обучили заново при фиксированной Q_θ

Результаты:

Наименьшие значения целевой функции достигаются при той дивергенции, на которой Q_θ была обучена.

train \ test	KL	KL-rev	JS	Jeffrey	Pearson
KL	**0.2808**	0.3423	0.1314	0.5447	0.7345
KL-rev	0.3518	**0.2414**	0.1228	0.5794	1.3974
JS	0.2871	0.2760	**0.1210**	0.5260	0.92160
Jeffrey	0.2869	0.2975	0.1247	**0.5236**	0.8849
Pearson	0.2970	0.5466	0.1665	0.7085	**0.648**

Рис.: Значения целевой функции для различных пар дивергенций

Эксперимент: генерация классных комнат

1. Использовали 168 тыс. фотографий классных комнат из базы LSUN
2. Оптимизационный алгоритм - ADAM (adaptive moment estimation)
3. Gradient Clipping - для предотвращения exploding gradients problem
4. Batch Normalization

Algorithm 1 Pseudo-code for norm clipping

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

Рис.: Пример gradient clipping

Эксперимент: генерация классных комнат

Взяли за основу DCGAN (Radford et. al), поменяли целевую функцию.

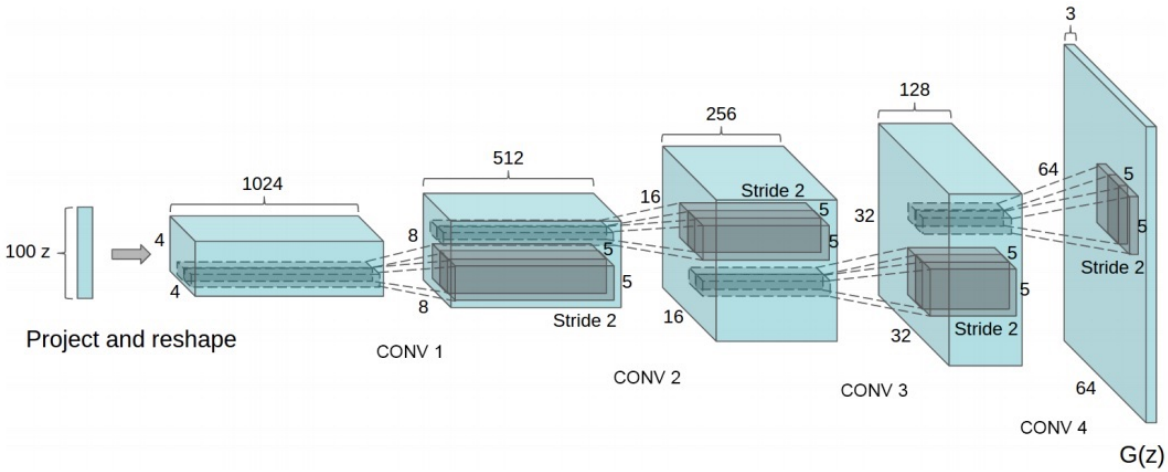


Рис.: Архитектура DCGAN

- Генератор: deconvolutional network, 3M параметров
- Вариационная функция: convnet, 3M параметров

Эксперимент: генерация классных комнат

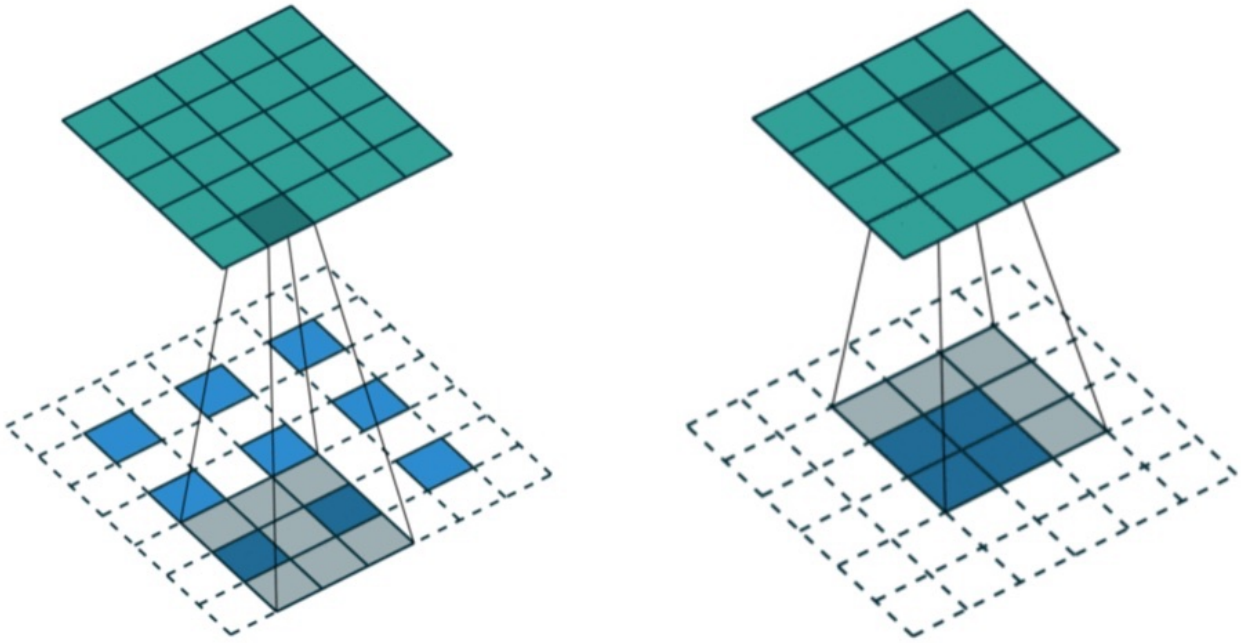


Рис.: Fractionally-strided convolutions or transposed convolutional layers (GOOD NAME) or deconvolutions (BAD NAME)

Эксперимент: генерация классных комнат

(a) GAN (b) KL (c) Squared Hellinger

Рис.: Результаты экперимента с генерацией классных комнат

Влияет ли дивергенция на результат?

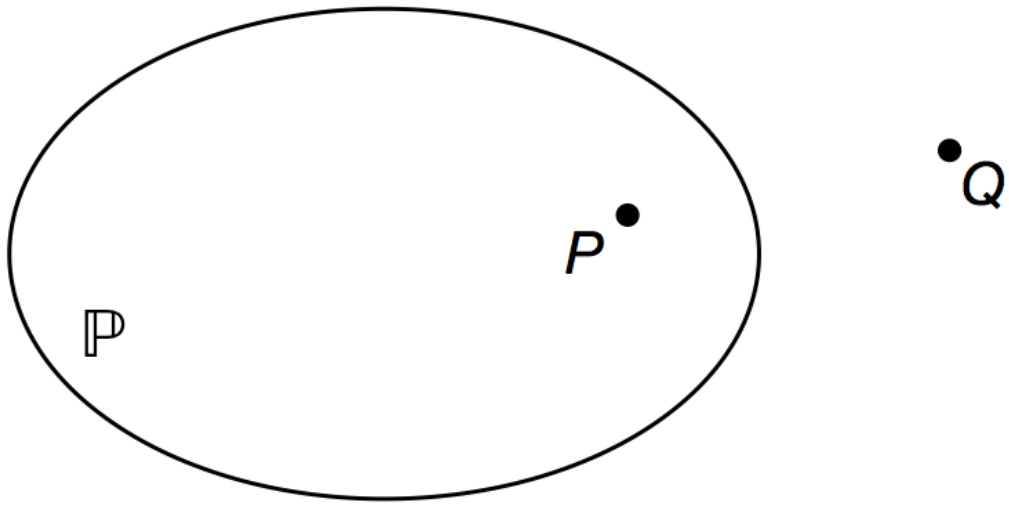
Играет ли выбор дивергенции серьезную роль?

- LSUN эксперимент: нет
- Theis et al., (2015), Huszar (2015): да

Почему разница не наблюдается в эксперименте с аудиториями?

Влияет ли дивергенция на результат?

Задача: выбрать наилучшую модель из параметрического семейства согласно выбранной метрике



Влияет ли дивергенция на результат?

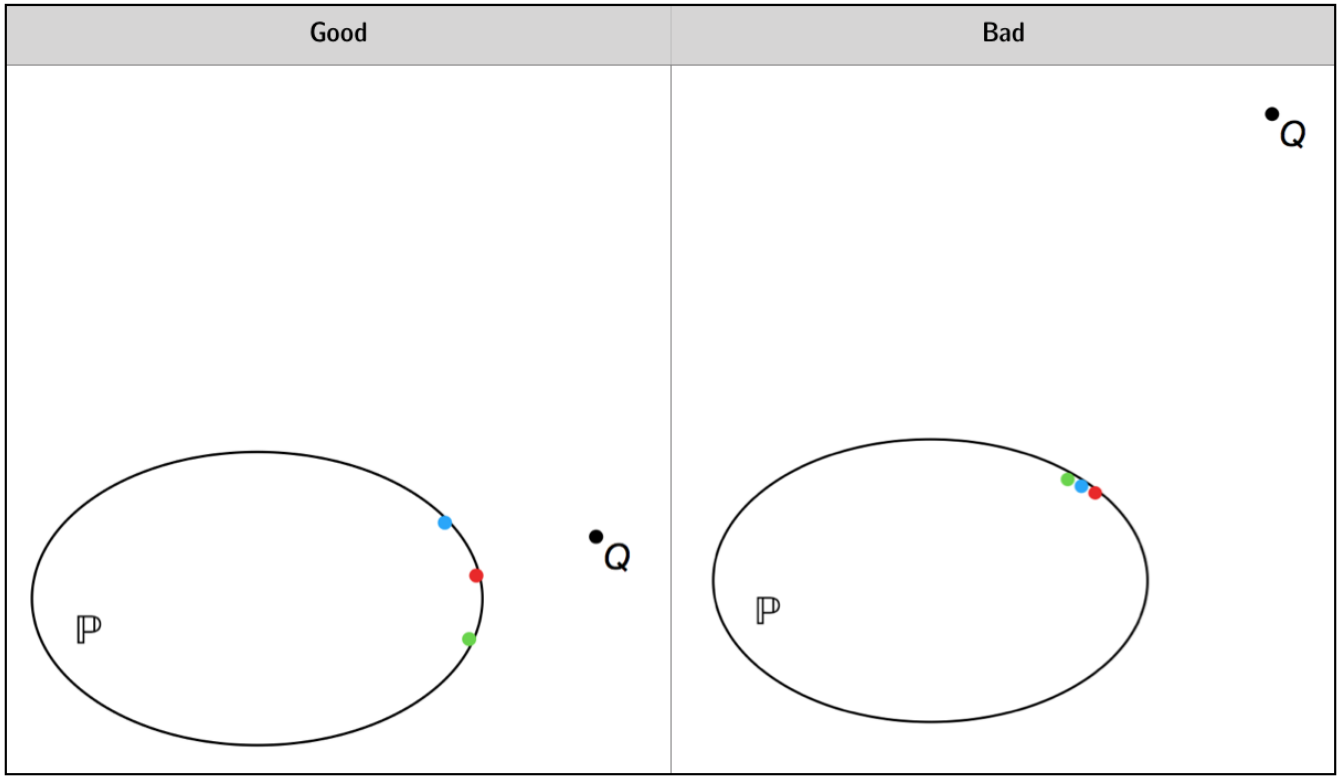


Рис.: Выбор модели: разные точки соответствуют разным дивергенциям

Предполагаемая причина отстутствия визуальных различий в эксперименте LSUN: bias семейства генераторов.

Ссылка на статью

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
https://arxiv.org/abs/1606.00709