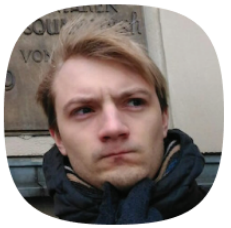
Vladislav Belavin, Maxim Borisyak, Denis Derkach



Introduction to distances: f-divergence

2021

MLHEP 2021
HIGHER SCHOOL OF ECONOMICS NATIONAL RESEARCH UNIVERSITY
LAMBDA • HSE
Yandex
SCHOOL OF DATA ANALYSIS
EPFL
SIT Schaffhausen Institute of Technology

Motivation

Quote

All models are wrong, but some useful.

-George Box

Quote

All models are wrong, but some useful.

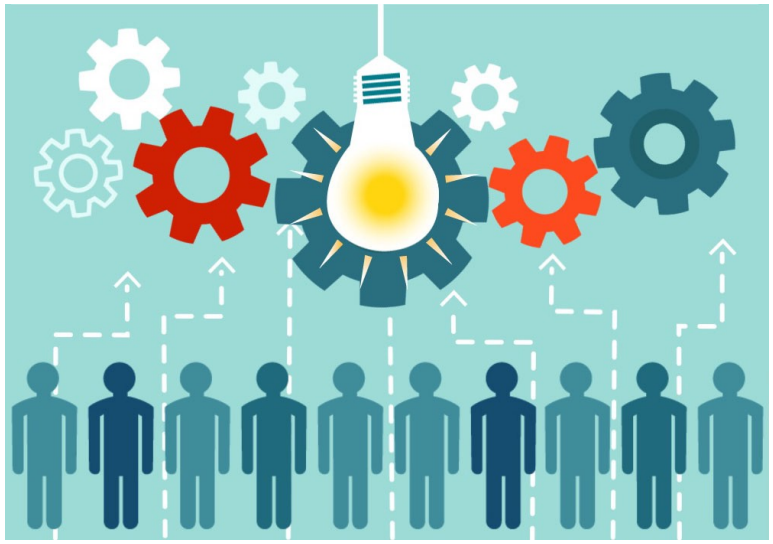
-George Box

What is a good generative model and what is a bad model?

Pic from: `https://bit.ly/2ZXrr2u`

Human assessment

> A natural choice is to ask another person to take a look at a result and decide if the generated image looks real.



From: `https://bit.ly/2Zm1QkG`

Human-driven metric

Pros:

- the most intuitive choice of metric.

Cons:

- biased towards models that overfit, i.e. favor mode collapsing of generative models;
- difficult to use for training:
  - expensive to collect;
  - noisy;
  - not differentiable.

Desired properties of metric

- favour models that generate high fidelity samples*;
- favour models that generate diverse samples*;
- be differentiable w.r.t. generative model parameters and easy to compute;
- agree (correlate) with human assessment.

> * The model that should produce pictures of cats and dogs only generates pictures of dogs. The photos generated are very close to the training kit. Thus, we have an undertrained and overtrained network at the same time.

Citation from: `http://www.deeplearningbook.org/contents/generative_models.html`

f-divergence

Notation

From now on, lets denote

- $p(x)$ as true probability density function;
- $q_\theta(x)$ as its estimate;
  - θ are parameters of the distribution q.

f-divergence

Definition

Let P and Q - be “good enough” (in terms of continuity and differentiability) probability distributions over some space Ω. p and q are theirs probability densities respectively, i.e. dP = p(x)dx and dQ $=$ q(x)dx. Than for a convex function f : $(0, \infty) \to \mathbb{R}$, such that f$(1) = 0$, f-divergence between P and Q is defined as:

$$D_f(P||Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)dx$$

We thus only need to choose a function f to define a divergence.

f-divergence: properties

- Non-negativity:

$$D_f(P||Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)dx \geq f\left(\int_\Omega \frac{p(x)}{q(x)} q(x)dx\right) = f(1) = 0$$

- For non-redundant $p(x)$ and $q(x)$, $D_f(P||Q)$ equal zero in two cases:
  - $f(t) = c(t-1), \;\; c \in R$
  - $p(x) = q(x), \;\; \forall x$

Kullback-Leibler Divergence Definition

KL divergence is an f-divergence with $f(t) = t\log(t)$:

$$KL(P||Q) = \int_{\mathbb{R}^n} p(x) \log\left(\frac{p(x)}{q_\theta(x)}\right) dx.$$

Properties of KL divergence

- not symmetric $\mathrm{KL}(P||Q) \neq \mathrm{KL}(Q||P)$;
  - not a proper distance metric;
- invariant under change of variables;
- chain rule (derived from Bayes' theorem):
  $\mathrm{KL}(P(x,y)||Q(x,y)) = \mathrm{KL}(P(x|y)||Q(x|y)) + \mathrm{KL}(P(y)||Q(y))$;
  - useful for deriving upper/lower bounds*.

*Example: `https://arxiv.org/pdf/1907.11891.pdf`

KL and MLE

Let θ_* be the true value of θ. Denote

$$\theta_{KL} = \arg\min_{\theta} KL(p||q_\theta) = \arg\min_{\theta} \mathbb{E}_{x\sim p(x)}(\log p(x)) - \mathbb{E}_{x\sim p(x)}(\log q_\theta(x)) =$$

$$\arg\max_{\theta} \lim_{N\to\infty} \sum_{i=1}^{N} \log(q_\theta(x_i)) = \theta_{MLE}, \quad x_i \sim p(x)$$

I.e. θ_{MLE} and θ_{KL} are precisely the same. Thus KL divergence is closely connected with MLE and inherit nice properties from MLE like consistency and efficiency.

KL and CE

Given two distributions $p(x)$ and $q(x)$, cross-entropy(CE) is defined as

$$H(p, q) = -\mathbb{E}_{x \sim p(x)}(\log q(x))$$

We will denote $H(p, p)$ as $H(p)$, which usually called Shannon entropy in literature. KL divergence is connected to it:

$$KL(p, q) = \mathbb{E}_{x \sim p(x)}(\log p(x) - \log q(x)) = H(p, q) - H(p).$$

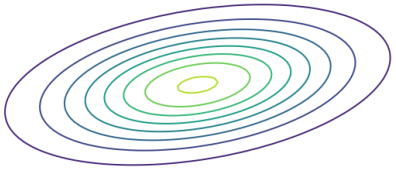
Since we normally optimise $L(\theta) = H(p_{data}, q_\theta)$, than optimisation of $KL(p_{data}, q_\theta)$ is equivalent to optimization of $H(p_{data}, q_\theta)$ w.r.t. to θ, because $H(p_{data})$ do not depend on θ.

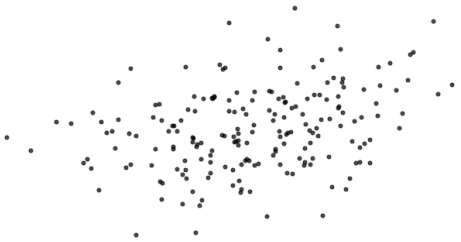
Convergence properties & issues

Trying to converge

Let's check the convergence properties.

Unfortunately, we do not have access to the true $p(x)$, so we must sample from it:

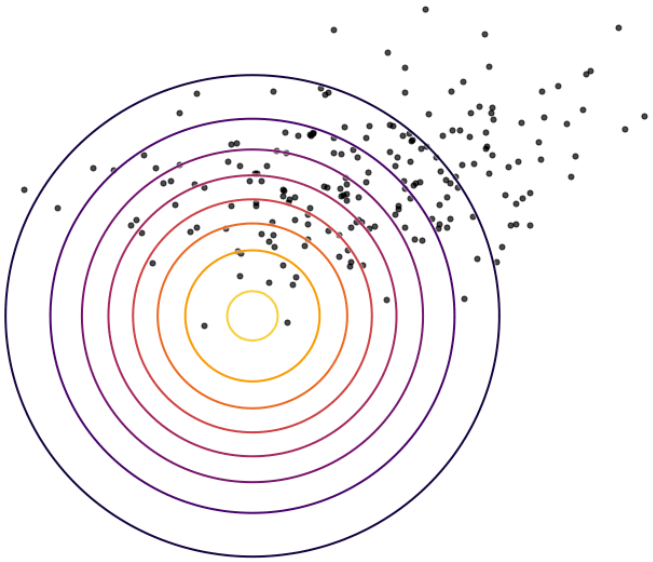




Here and later some examples are motivated by:

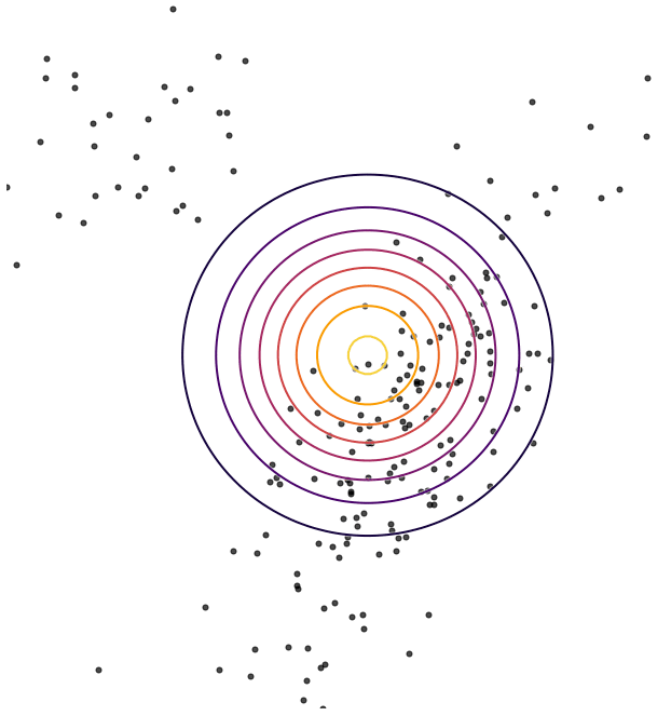
`https://colinraffel.com/blog/gans-and-divergence-minimization.html`

Converging with KL: unimodal case



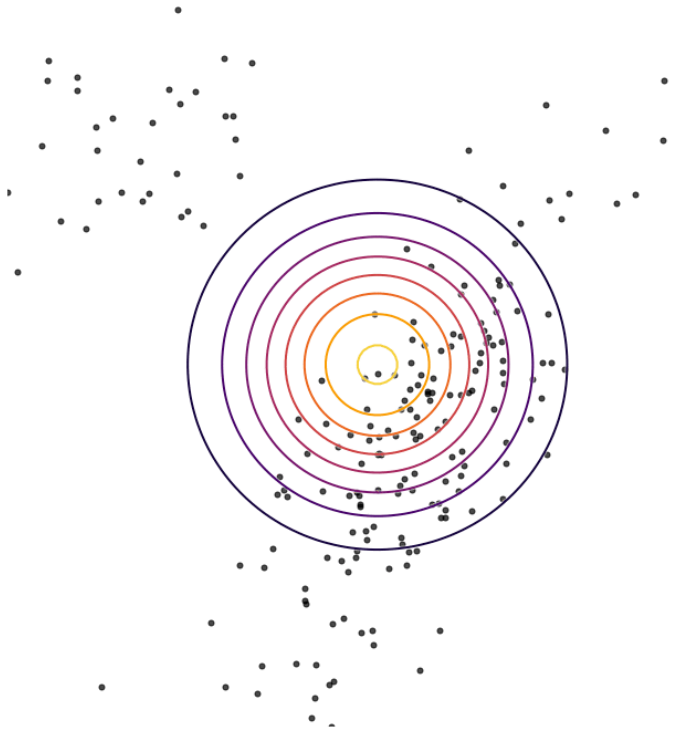
- The procedure works!
- Does it mean that the problem of building a model is solved?

Converging with KL: multimodal case



- The procedure works!
- Does it mean that the problem of building a model is solved?
- Not really, for the multimodal case, we will have problems.

Converging with KL: multimodal case intuition

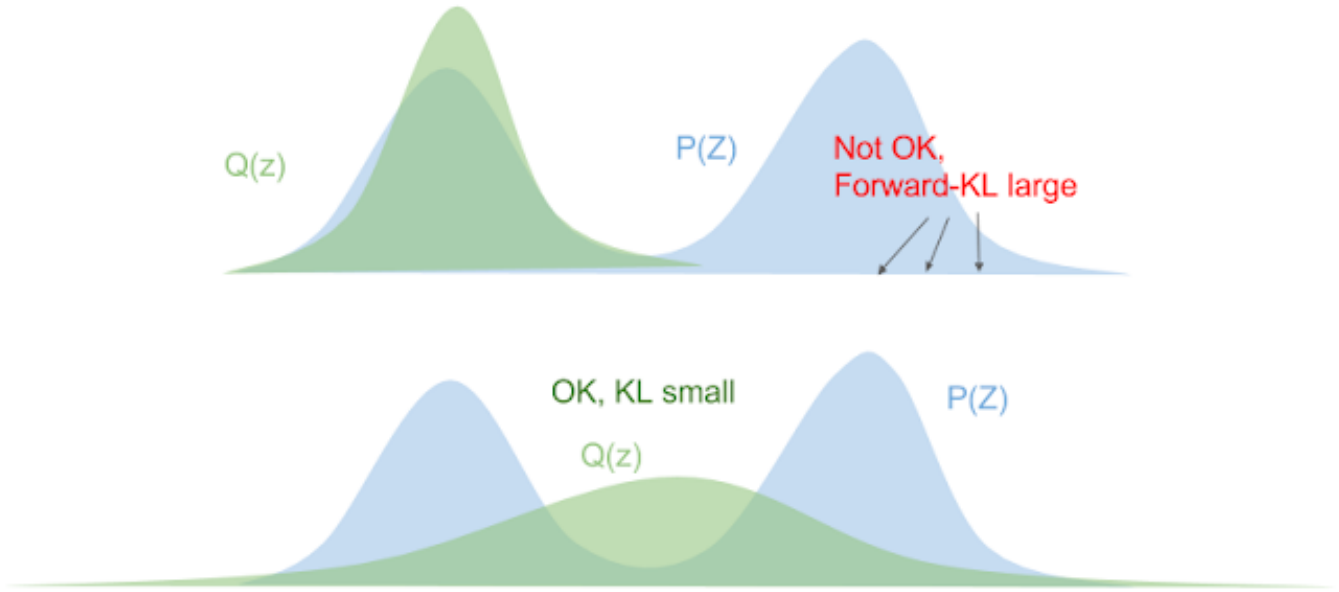


- It's natural to look at the quantity we optimize:

$$\arg\max_{\theta} \mathbb{E}_{x\sim p}[\log q_\theta(x)]$$

- If there is no $q_\theta(x)$ support in the place, where we have $x \sim p(x)$ than the loss function goes to ∞.
- We also have $q_\theta(x)$ support in places with no $x \sim p(x)$.

More intuition on KL



Forward KL(KL $= \int p(x) \log \frac{p(x)}{q_\theta(x)} dx$) is known as zero avoiding, as it is avoiding $q(x) = 0$ whenever $P(x) > 0$.

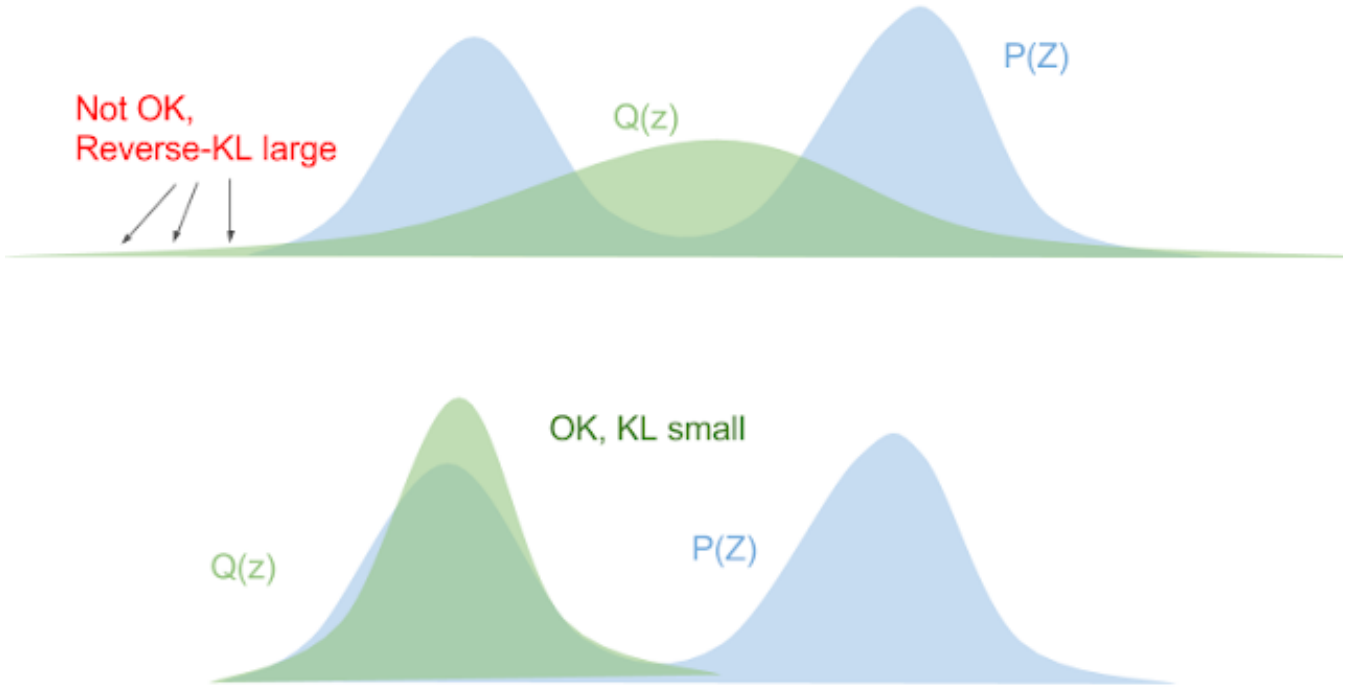
Picture credit: `https://bit.ly/329zHz9`

Reverse KL divergence

In order to overcome the problems, we can define a reverse divergence:

$$\mathrm{rKL}(q_\theta||p) = \mathrm{KL}(p||q_\theta) = \int_{\mathbb{R}^n} q_\theta(x) \log\left(\frac{q_\theta(x)}{p(x)}\right) dx.$$

Intuition on rKL



Reverse KL Divergence(rKL $= \int q_\theta(x) \log \frac{q_\theta(x)}{p(x)} dx$) is known as zero forcing, as it forces $Q(X)$ to be 0 on some areas, even if $P(X) > 0$.

rKL: optimisation

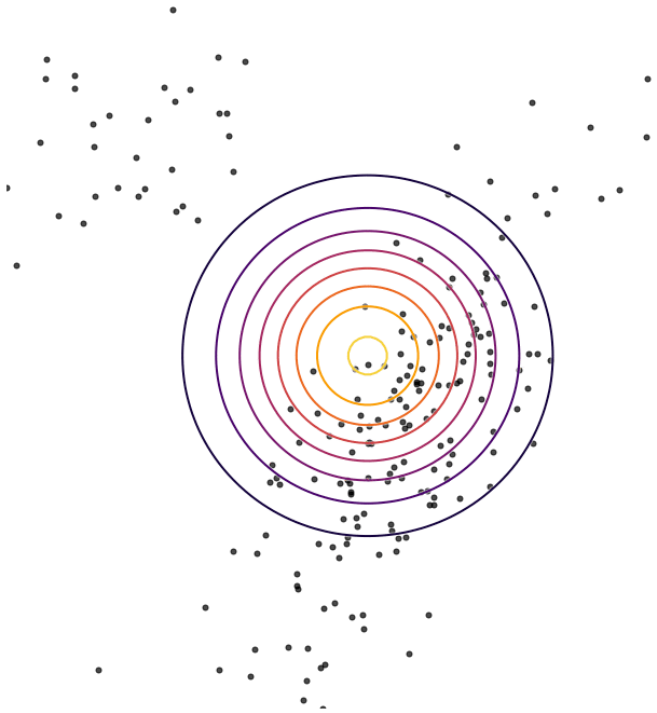
In fact, we are optimizing a very similar thing:

$$\begin{aligned}
\theta^* &= \arg\min_\theta \mathsf{KL}(\mathsf{q}_\theta || \mathsf{p}) = \\
&= \arg\min_\theta (\mathbb{E}_{\tilde{\mathsf{x}} \sim \mathsf{q}_\theta}[\log \mathsf{q}_\theta(\mathsf{x})] - \mathbb{E}_{\tilde{\mathsf{x}} \sim \mathsf{q}_\theta}[\log \mathsf{p}(\mathsf{x})]) = \\
&= \arg\max_\theta (-\mathbb{E}_{\mathsf{x} \sim \mathsf{q}_\theta}[\log \mathsf{q}_\theta(\mathsf{x})] + \mathbb{E}_{\mathsf{x} \sim \mathsf{q}_\theta}[\log \mathsf{p}(\mathsf{x})])
\end{aligned}$$

But we do not have the previous problem of likelihood going to infinity in unreasonable places.

The first term is related to entropy of the generating model, the second penalises generated samples that are not similar to real distribution.

Converging with rKL



- We no longer have $q_\theta(x)$ support in the regions with no $x \sim p(x)$ population.
- The converged distribution looks reasonable but only for one solution.

Jensen-Shannon Divergence

We can try to optimize different divergences however, the problems normally stay. A distinguishable attempt is to construct the mixture of KL and rKL:

$$\begin{aligned} \mathrm{JS}(\mathrm{p}(\mathrm{x})||\mathrm{q}_\theta(\mathrm{x})) = & \quad \tfrac{1}{2} \quad \mathrm{KL}\left(\mathrm{p}(\mathrm{x})||\frac{\mathrm{p}(\mathrm{x})+\mathrm{q}_\theta(\mathrm{x})}{2}\right) + \\ & + \quad \tfrac{1}{2} \quad \mathrm{KL}\left(\mathrm{q}_\theta(\mathrm{x})||\frac{\mathrm{p}(\mathrm{x})+\mathrm{q}_\theta(\mathrm{x})}{2}\right), \end{aligned}$$

It is symmetric and does not ignore zeroes like KL and does not ignore x like rKL.

Metrics from divergences

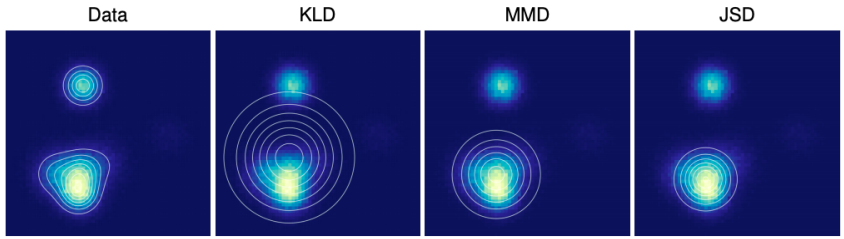


Figure 1: An isotropic Gaussian distribution was fit to data drawn from a mixture of Gaussians by either minimizing Kullback-Leibler divergence (KLD), maximum mean discrepancy (MMD), or Jensen-Shannon divergence (JSD). The different fits demonstrate different tradeoffs made by the three measures of distance between distributions.

Problems:

- need to use metrics different from optimised;
- difficulties in case of high dimension problem;
- not evident choice of a good metric.

From:https://arxiv.org/abs/1506.05751

Constructing the divergence

- We use f-divergence, for convex function $f(x) : \mathbb{R}^+ \mapsto \mathbb{R}$ and $f(1) = 0$:

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

$$\int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)dx \geq f\left(\int_{\mathcal{X}} \frac{p(x)}{q(x)} q(x)dx\right) = 0.$$

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

Conclusion

Key point:

- KL have nice properties from maximum likelihood perspective;
  - ...but tends to “smooth” distribution;
- rKL allows to generate high-fidelity samples;
  - ...but might lead to mode collapsing;
- JS somewhere in the middle between KL and rKL;
  - you also can play with mixing coefficients, i.e. choose α different from $\frac{1}{2}$ in $\alpha\mathrm{KL}(\mathrm{p}|\mathrm{q}) + (1-\alpha)\mathrm{KL}(\mathrm{q}|\mathrm{p})$;
- choice of $\mathrm{f}(\cdot)$ should depend on the problem and your prior knowledge.

Thank you for your attention!

Vladislav Belavin
vbelavin@hse.ru
SchattenGenie
hse_lambda