Generative Modeling

Distances and metrics

Denis Derkach, Artem Ryzhikov, Maxim Artemev

Laboratory for methods of big data analysis

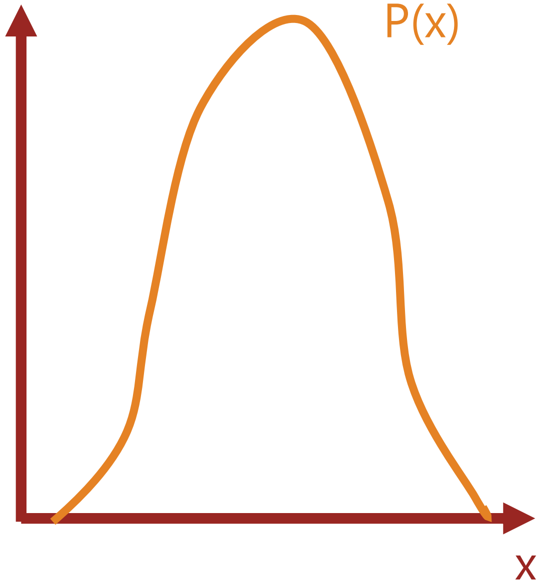
Spring 2022

In this Lecture

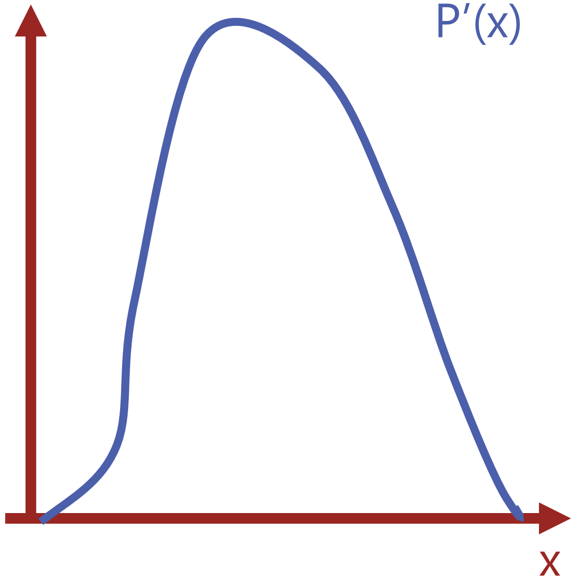
- f-divergences
  - total Variation Distance;
  - Kullback-Leibler Divergence;
  - Jensen-Shannon Divergence;
  - divergence inequalities;
  - variational lower bound.
- Metrics (seminar)

Total Variation Distance

What we measure



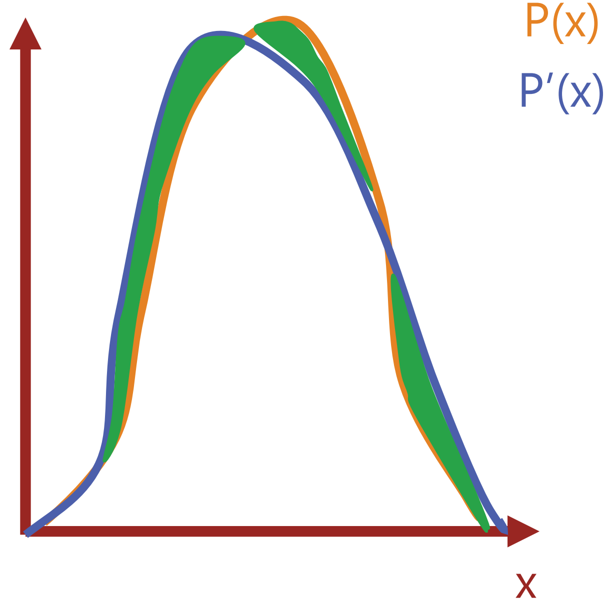
True Probability Density



Fitted Probability Density

P’(x) is similar to P(x)?

First idea: absolute difference



$$\int |P(x) - P'(x)|\, dx$$

Total Variation Distance

For $p(x)$ and $q_\theta(x)$ being PDFs:

$$D(p(x), q_\theta(x)) = \frac{1}{2}\int |p(x) - q_\theta(x)| dx$$

This can be rewritten using Scheffe's theorem

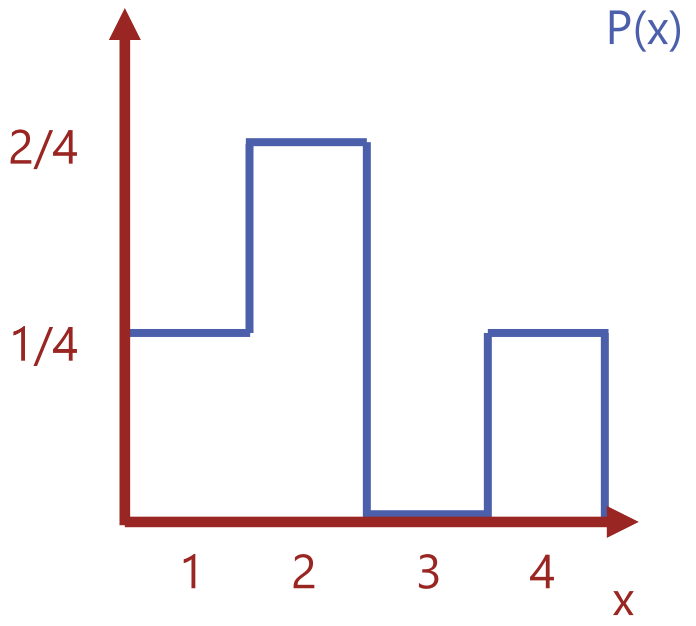
$$\boxed{D(p(x), q_\theta(x)) = \sup_{\mathrm{A}} \left| \int_{\mathrm{A}} \mathrm{p(x)dx} - \int_{\mathrm{A}} \mathrm{q}_\theta\mathrm{(x)dx} \right|}$$

Where A is any measurable set.

A. B. Tsybakov, Introduction to Nonparametric Estimation, sec 2.4

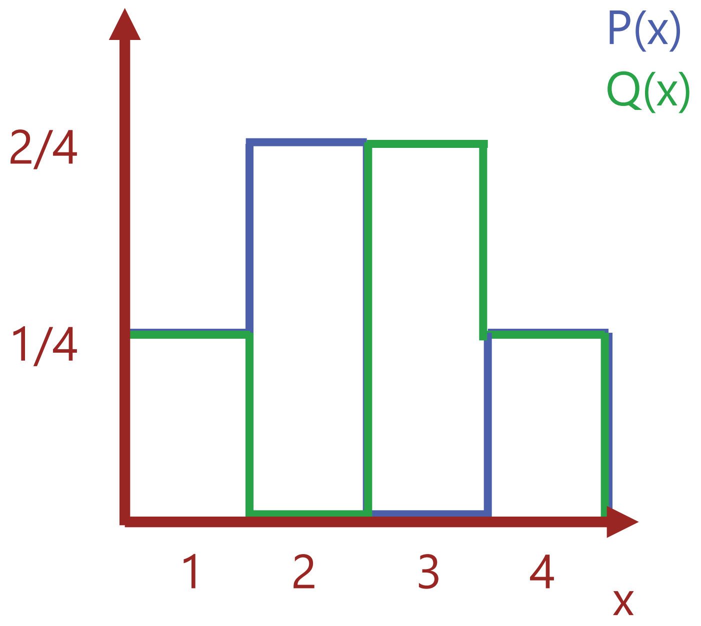
Total Variation Distance: example 1D

- discrete case for two PDFs



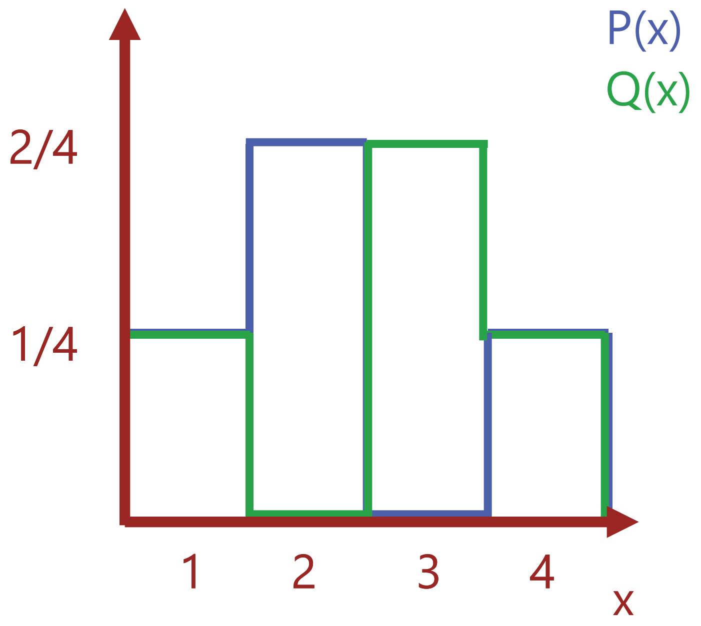
Total Variation Distance: example 1D

- discrete case for two PDFs



Total Variation Distance: example 1D

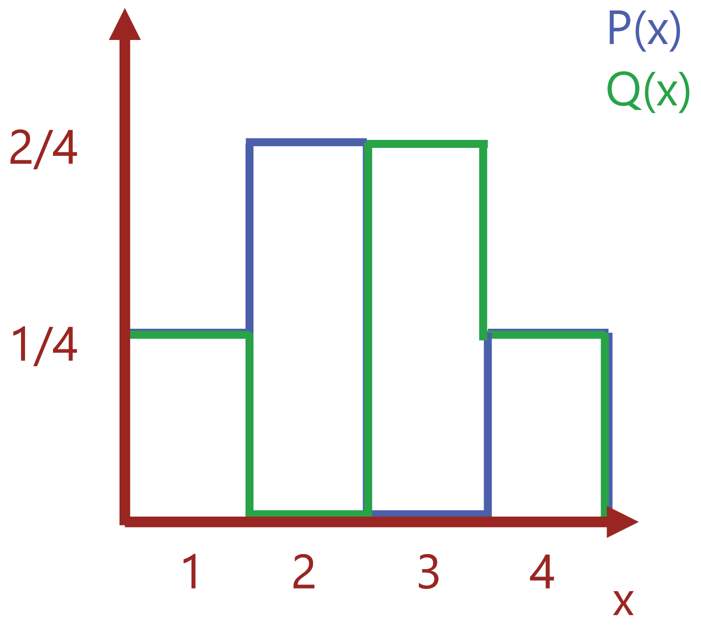
- discrete case for two PDFs
- calculate in two ways:



Total Variation Distance: example 1D

- discrete case for two PDFs
- calculate in two ways:
  - construct all possible subsets:

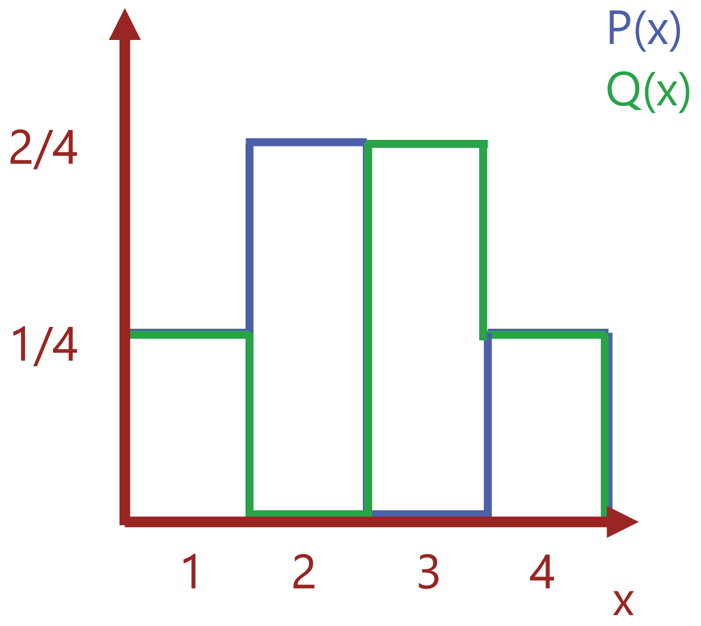
    {1}, {2}. {3}, {4}, {1;2}, {1;3}, {1;4}, {2;3}, {2;4}, {3;4}, {1;2;3}, {1;2;4}, {1;3;4}, {1,2,3,4}.



Total Variation Distance: example 1D

- discrete case for two PDFs
- calculate in two ways:
  - construct all possible subsets:

    D(p,q) = 0.5

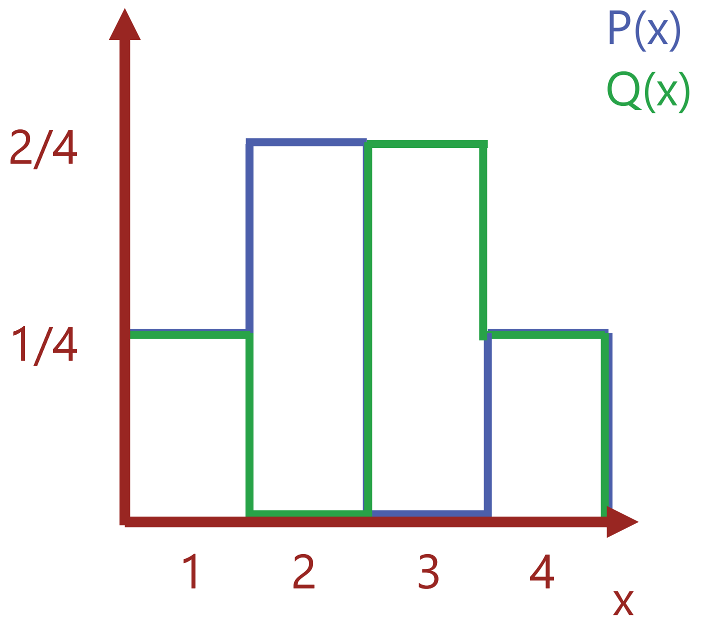


Total Variation Distance: example 1D

- discrete case for two PDFs
- calculate in two ways:
  - construct all possible subsets:

    D(p,q) = 0.5
  - integrate over full range:

    D(p,q) = 0.5



Total Variation Distance: observations

- Symmetric D(p, q) = D(q, p)
- Interpretable (using Scheffe lemma)
- Connected to hypothesis testing (D is the sum of errors)

Total Variation Distance: observations

- Symmetric D(p, q) = D(q, p)
- Interpretable (using Scheffe's theorem)
- Connected to hypothesis testing (D is the sum of errors)
- Too strong:

The distance might ignore the growing number of trials.

$X_1, \ldots, X_n \sim \pm 1$, $S_n = \sum_n X_i$. Than

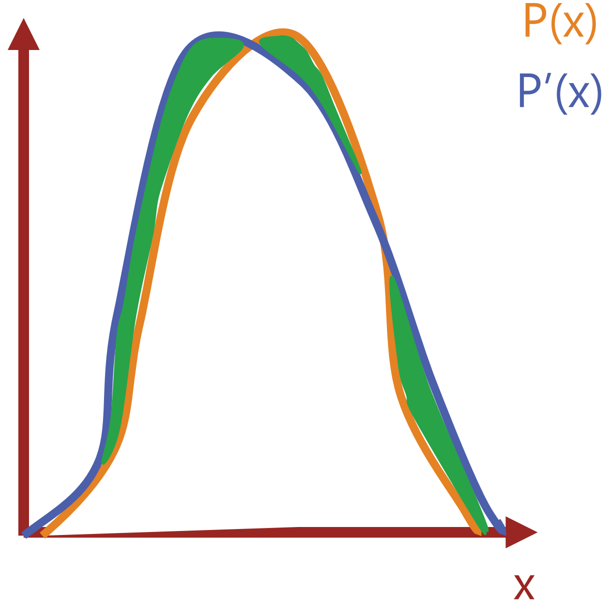
$$S_n/\sqrt{n} \to \mathcal{N}(0,1),$$

but $D(S_n, \mathcal{N}(0,1)) = 1$ for any n).

A. L. Gibbs, F. E. Su On Choosing and Bounding Probability Metrics
F Pollard, Total variation distance between measures

Kullback-Leibler Divergence

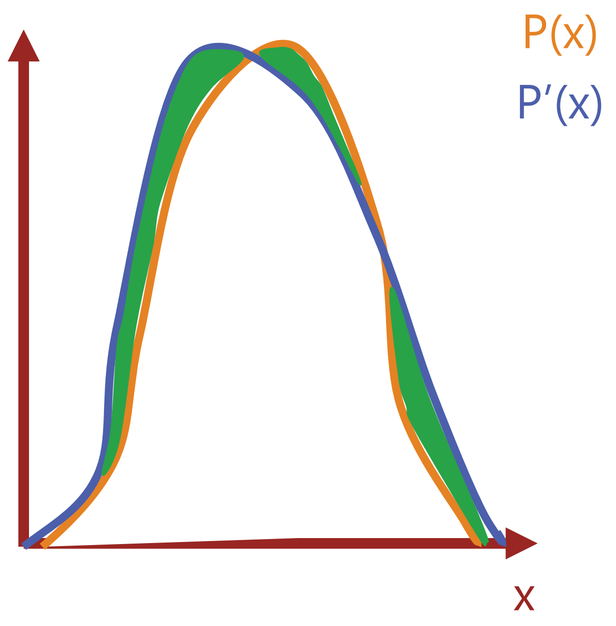
Kullback-Leibler divergence: ideas



Previously:

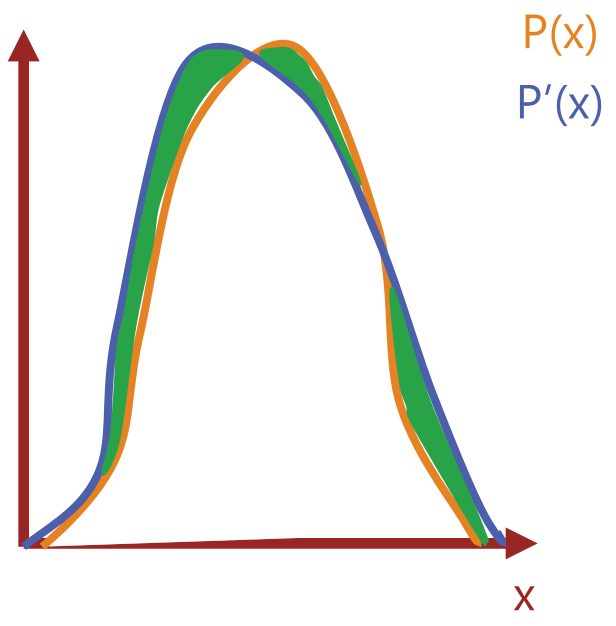
$$\int |P(x) - P'(x)|\, dx$$

Kullback-Leibler divergence: ideas



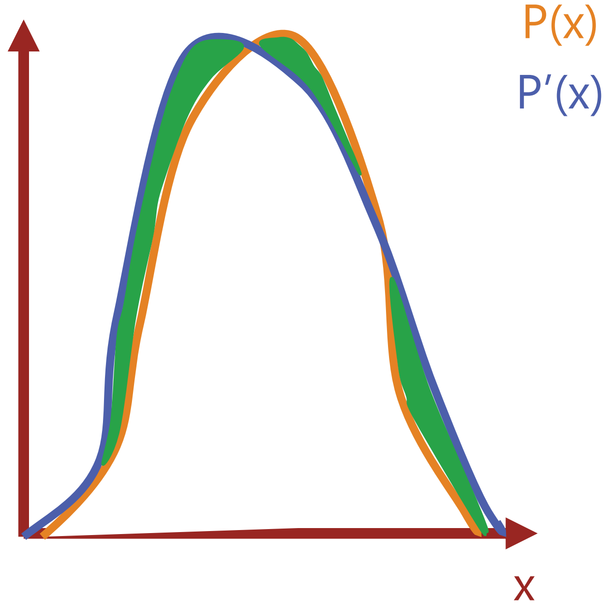
$$\frac{P(x)}{P'(x)}$$

Kullback-Leibler divergence: ideas



$$\ln \frac{P(x)}{P'(x)}$$

Kullback-Leibler divergence: ideas



$$\int P(x) \ln \frac{P(x)}{P'(x)} dx$$

Kullback-Leibler divergence: definition

For p(x) and q(x), two probability distributions,

$$KL(p||q_\theta) = \int p(x) \log \left(\frac{p(x)}{q_\theta(x)} \right) dx$$

Kullback-Leibler divergence: definition

For p(x) and q(x), two probability distributions,

$$KL(p||q_\theta) = \int p(x) \log \left(\frac{p(x)}{q_\theta(x)} \right) dx$$

- not symmetric $KL(P||Q) \neq KL(Q||P)$
- invariant under change of variables
- additive for independent variables
- nonnegative

Kullback-Leibler divergence: observations

- **KL divergence is connected to cross-entropy:**

$$KL(p||q) = H(p) + H(p, q),$$

where $H(p, q) = \mathbb{E}_p(\log q).$

KL and Maximum Likelihood

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(p(x)||q_\theta(x))$$

KL and Maximum Likelihood

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(p(x)||q_\theta(x))$$

$$= \underset{\theta}{\operatorname{argmin}} (\mathbb{E}_{x\sim p}[\log p(x)] - \mathbb{E}_{x\sim p}[\log q_\theta(x)])$$

KL and Maximum Likelihood

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(p(x)||q_\theta(x))$$

$$= \underset{\theta}{\operatorname{argmin}} (\mathbb{E}_{x\sim p}[\log p(x)] - \mathbb{E}_{x\sim p}[\log q_\theta(x)])$$

$$= -\underset{\theta}{\operatorname{argmin}} \mathbb{E}_{x\sim p}[\log q_\theta(x)]$$

KL divergence: observations

- **KL divergence is connected to cross-entropy:**

$$KL(p||q) = H(p) + H(p, q),$$

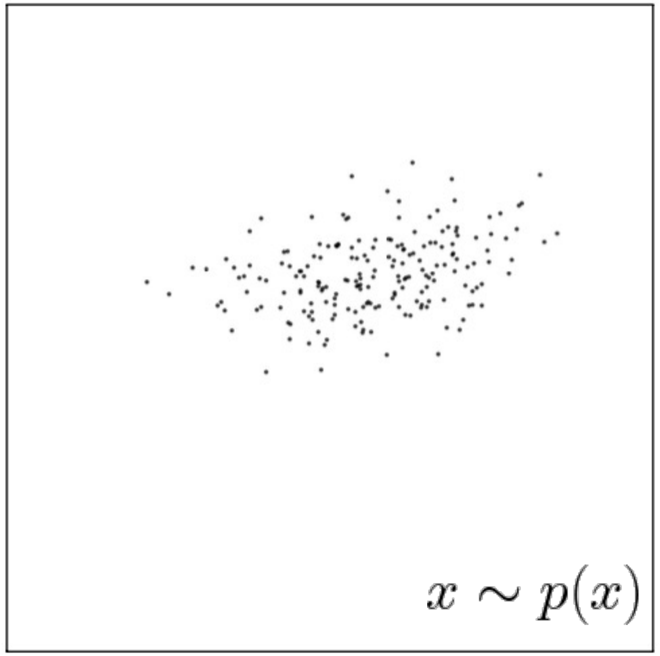
where $H(p, q) = \mathbb{E}_p(\log q)$.

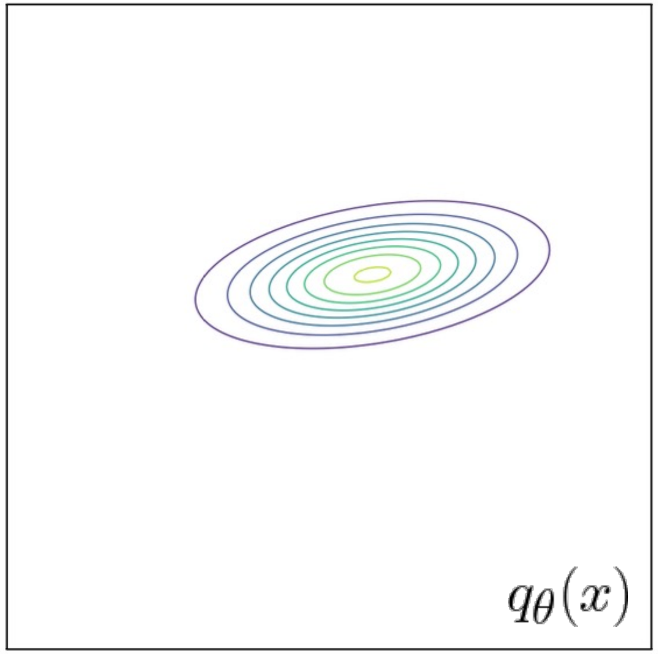
- **Minimizing KL divergence is equivalent to maximizing the likelihood.**

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(p(x)||q_\theta(x)) = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(q_\theta(x); x)$$

Using in fits

Fit data points from 2D Gaussian function

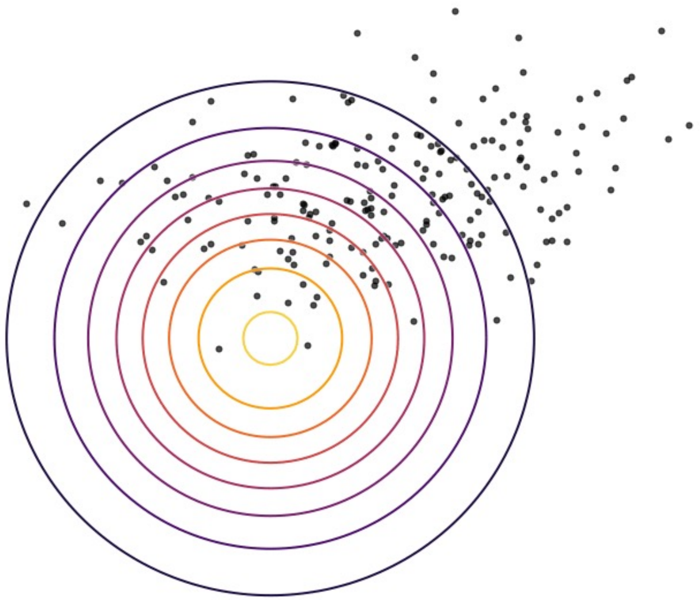




…with 2D Gaussian function

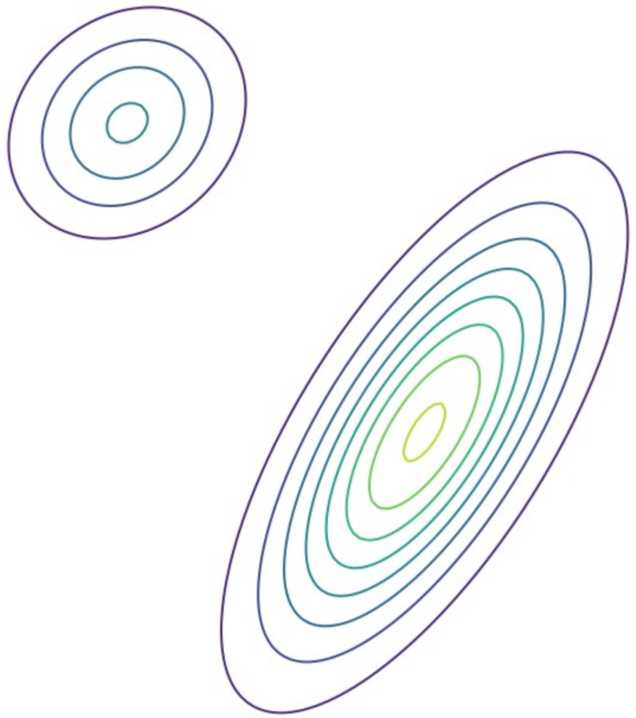
Here and Later: Colin Raffel's blog

Using in fits



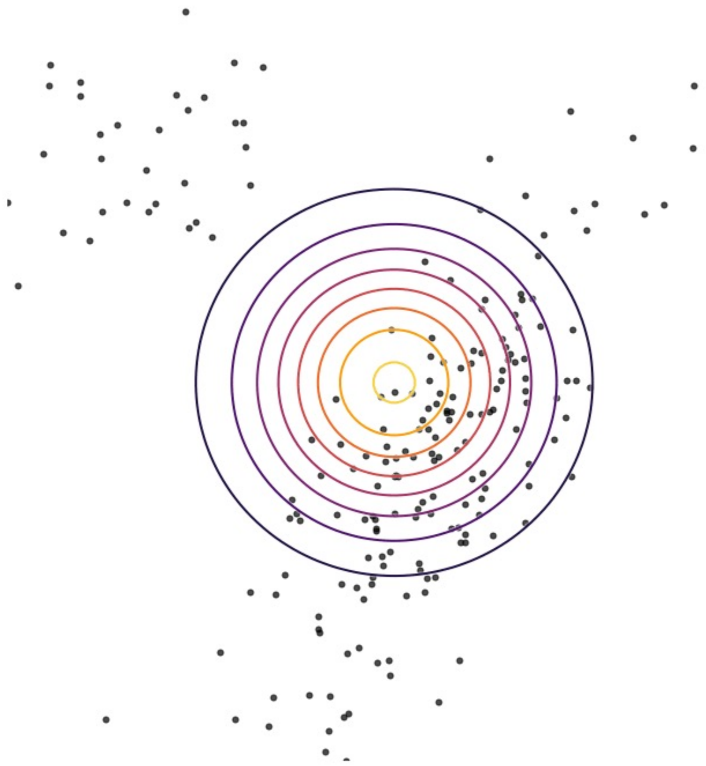
- Runs smoothly for simple data

Using in fits: Multimodal data



- Runs smoothly for simple data

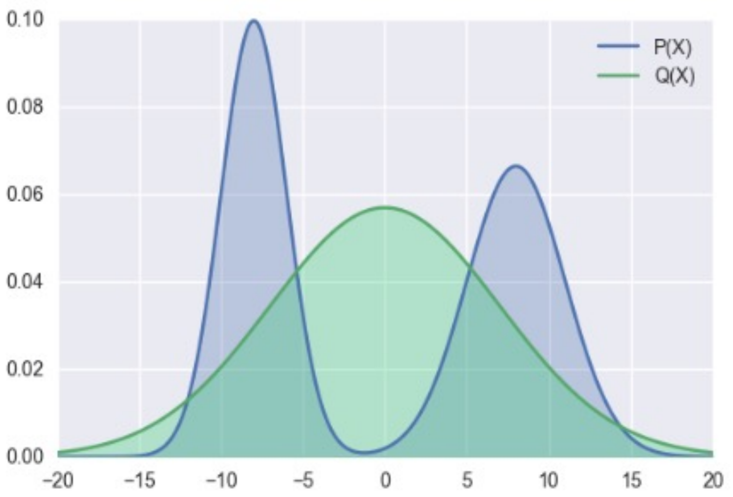
Using in fits: Multimodal data



- Runs smoothly for simple data
- Problems for multimodal data
- Covers significant amount of empty spaces

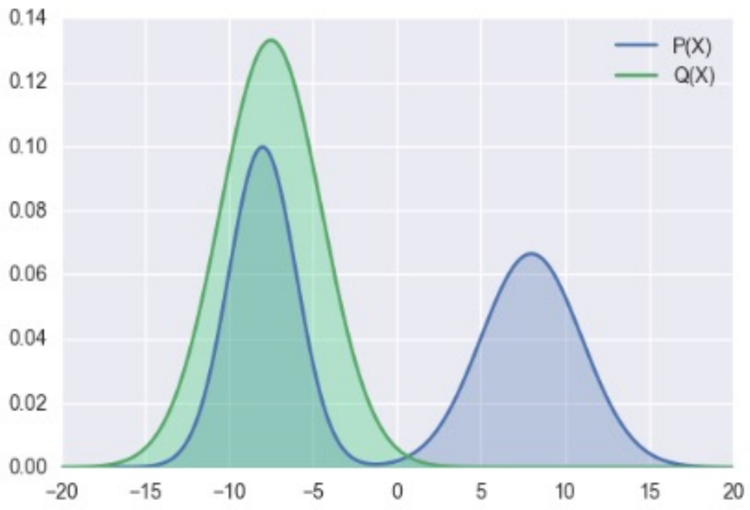
KL divergence: study

$$KL(p||q_\theta) = \int p(x) \log \left(\frac{p(x)}{q_\theta(x)} \right) dx$$



KL is zero avoiding, as it is avoiding q(x) = 0 whenever p(x) > 0

$$KL(q_\theta||p) = \int q_\theta(x) \log \left(\frac{q_\theta(x)}{p(x)} \right) dx$$



Reverse KL is zero forcing, as it forces q(X) to be 0 on some areas, even if p(X) > 0

Agustinus Kristiadi's Blog

Reverse KL divergence: fits

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(q_\theta(x)||p(x))$$

Reverse KL divergence: fits

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(q_\theta(x)||p(x))$$

$$= \underset{\theta}{\operatorname{argmin}} (\mathbb{E}_{\tilde{x}\sim q_\theta}[\log q_\theta(x)] - \mathbb{E}_{\tilde{x}\sim q_\theta}[\log p(x)])$$

Reverse KL divergence: fits

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(q_\theta(x)||p(x))$$

$$= \underset{\theta}{\operatorname{argmin}}(\mathbb{E}_{\tilde{x}\sim q_\theta}[\log q_\theta(x)] - \mathbb{E}_{\tilde{x}\sim q_\theta}[\log p(x)])$$

$$= \underset{\theta}{\operatorname{argmax}}(-\mathbb{E}_{\tilde{x}\sim q_\theta}[\log q_\theta(x)] + \mathbb{E}_{\tilde{x}\sim q_\theta}[\log p(x)])$$

Reverse KL divergence: fits

Find the optimal parameter, θ^*:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} KL(q_\theta(x)||p(x))$$

entropy for the
fitted model

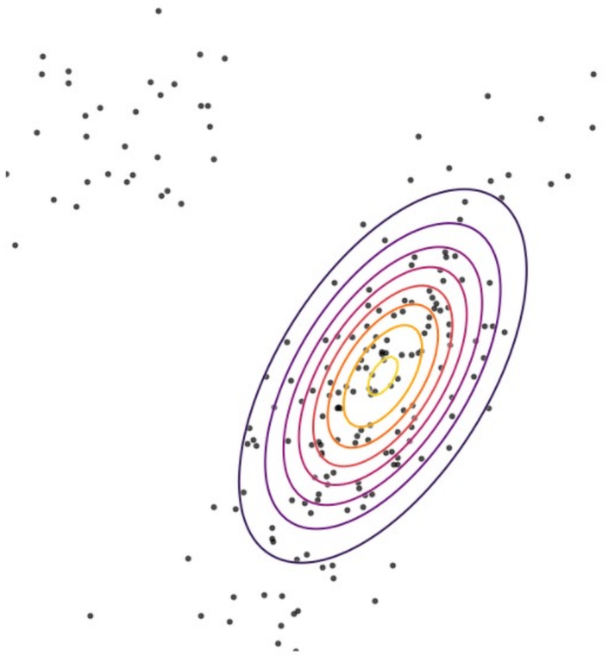
$$= \underset{\theta}{\operatorname{argmax}}(-\mathbb{E}_{\tilde{x}\sim q_\theta}[\log q_\theta(x)] + \mathbb{E}_{\tilde{x}\sim q_\theta}[\log p(x)])$$

relation between
fitted and generated

Reverse KL divergence: fits

- $q_\theta(x)$ covers only regions with data

- reasonable in multi-modal data for one solution



Critical: we do not have direct access to p(x).

Jensen-Shannon Divergence

Jensen-Shannon Divergence: idea

- KL divergence is asymmetric

Jensen-Shannon Divergence: idea

- KL divergence is asymmetric

$$KL(p||q) + KL(q||p)$$

Jensen-Shannon Divergence: idea

- KL divergence is asymmetric
- KL can become infinite

$$KL(p||q) + KL(q||p)$$

Jensen-Shannon Divergence: idea

- KL divergence is asymmetric
- KL can become infinite

$$KL(p(x)||\frac{p(x)+q_\theta(x)}{2})+KL(q_\theta(x)||\frac{p(x)+q_\theta(x)}{2})$$

Jensen-Shannon Divergence: Definition

For p(x) and q(x), two probability distributions,

$$JS(p,q) = \frac{1}{2}\left(KL(p(x)||\frac{p(x)+q_\theta(x)}{2}) + KL(q_\theta(x)||\frac{p(x)+q_\theta(x)}{2})\right)$$

- symmetric
- nonnegative $0 \leq JS(P,Q) \leq \ln(2)$
- can be transformed to a true distance $\sqrt{JS(p,q)}$

J. Lin Divergence measures based on the Shannon entropy

f-divergences

Definition

- Let $f: (0; \infty) \to \mathbb{R}$ be a convex function with f(1) = 0.
- P and Q - two probability distributions on a measurable space $(\mathcal{X}, \mathcal{F})$.
- p and q - absolutely continuous with respect to a base measure dx defined on $\mathcal{X}$.
- f-divergence is defined:

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

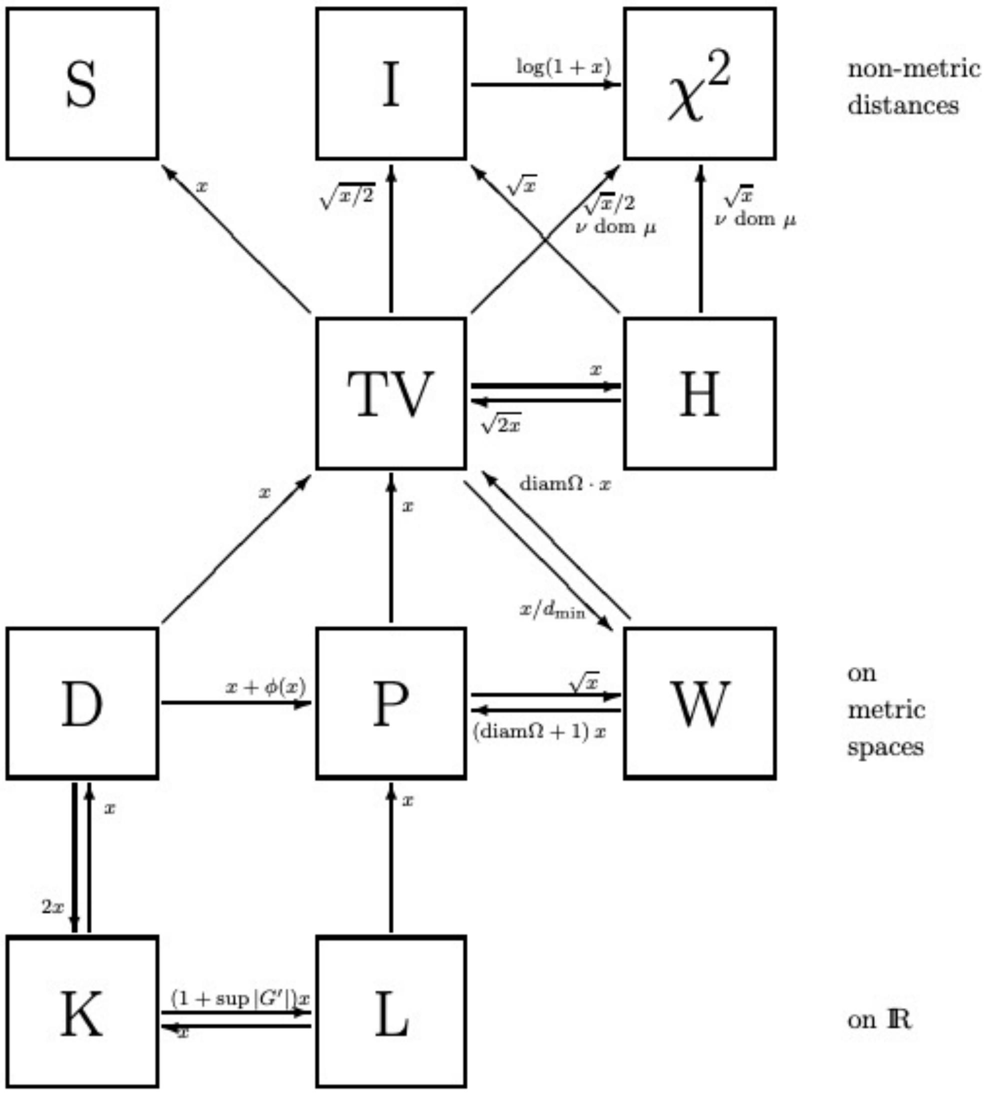
- f is called generator.

Examples

Name	$D_f(P\|Q)$	Generator $f(u)$
Total variation	$\frac{1}{2}\int \lvert p(x) - q(x)\rvert \, \mathrm{d}x$	$\frac{1}{2}\lvert u-1\rvert$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$(\sqrt{u}-1)^2$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x$	$(u-1)\log u$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2} + u\log u$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)} + (1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u - (1-\pi+\pi u)\log(1-\pi+\pi u)$

f-divergence inequalities

Abbreviation	Metric
D	Discrepancy
H	Hellinger distance
I	Relative entropy (or Kullback-Leibler divergence)
K	Kolmogorov (or Uniform) metric
L	Lévy metric
P	Prokhorov metric
S	Separation distance
TV	Total variation distance
W	Wasserstein (or Kantorovich) metric
χ^2	χ^2 distance



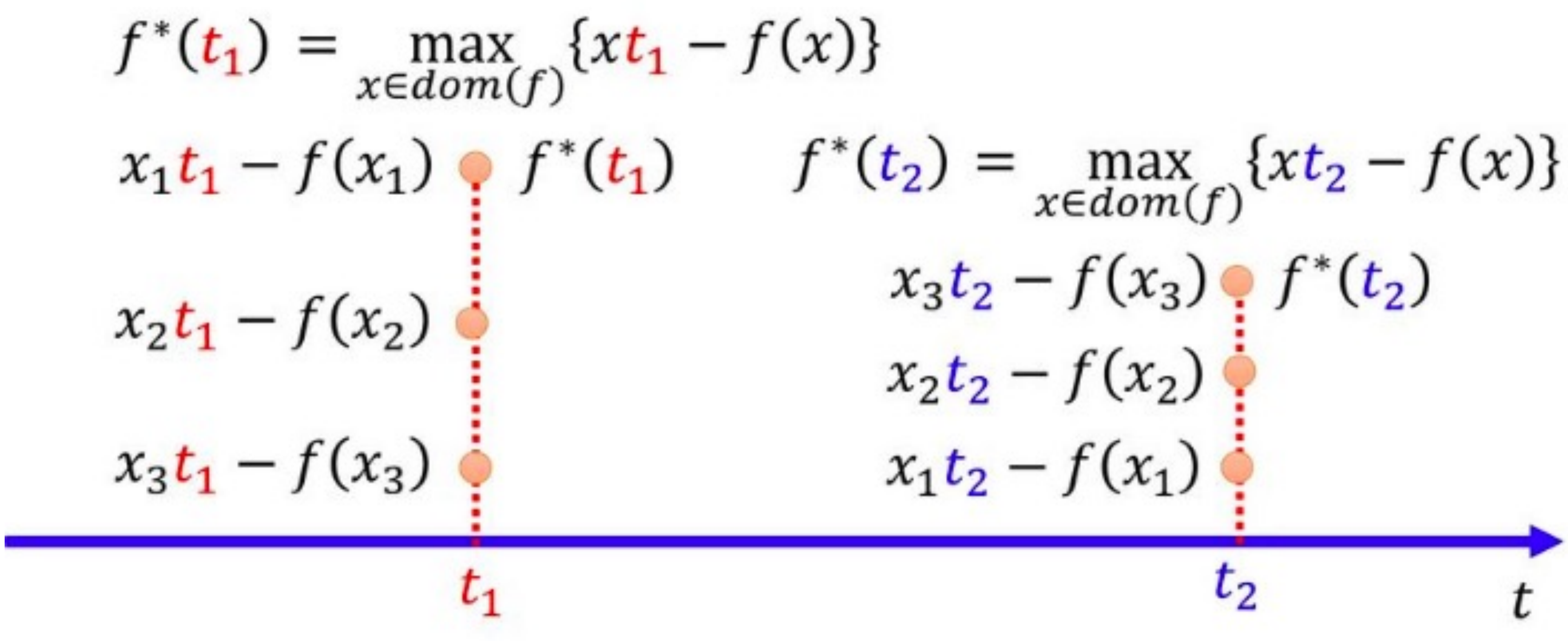
A. L. Gibbs, F. E. Su On Choosing and Bounding Probability Metrics

Fenchel conjugate

- Each generator has a **Fenchel conjugate** function:

$$f^*(t) = \sup_x (xt - f(x))$$

if f is convex, then $(f^*)^* = f$.



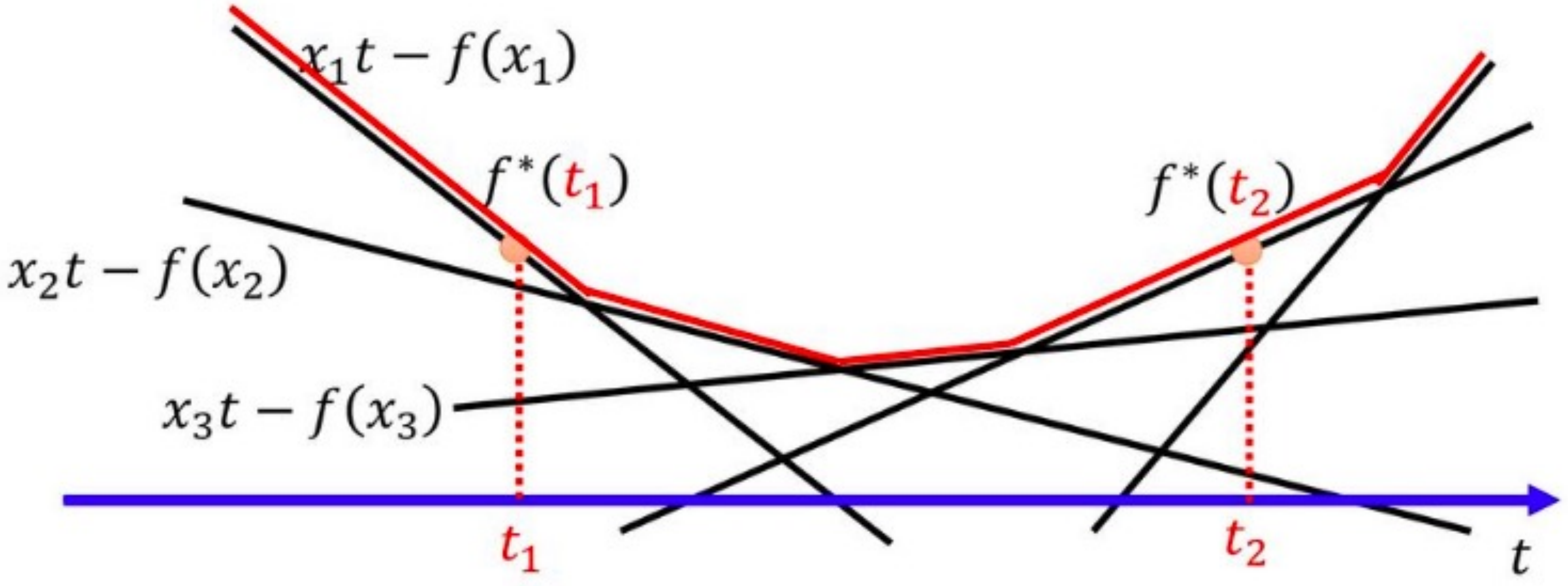
Hung-yi Lee's lecture course

Fenchel conjugate

- Each generator has a **Fenchel conjugate** function:

$$f^*(t) = \sup_x (xt - f(x))$$

if f is convex, then $(f^*)^* = f$.

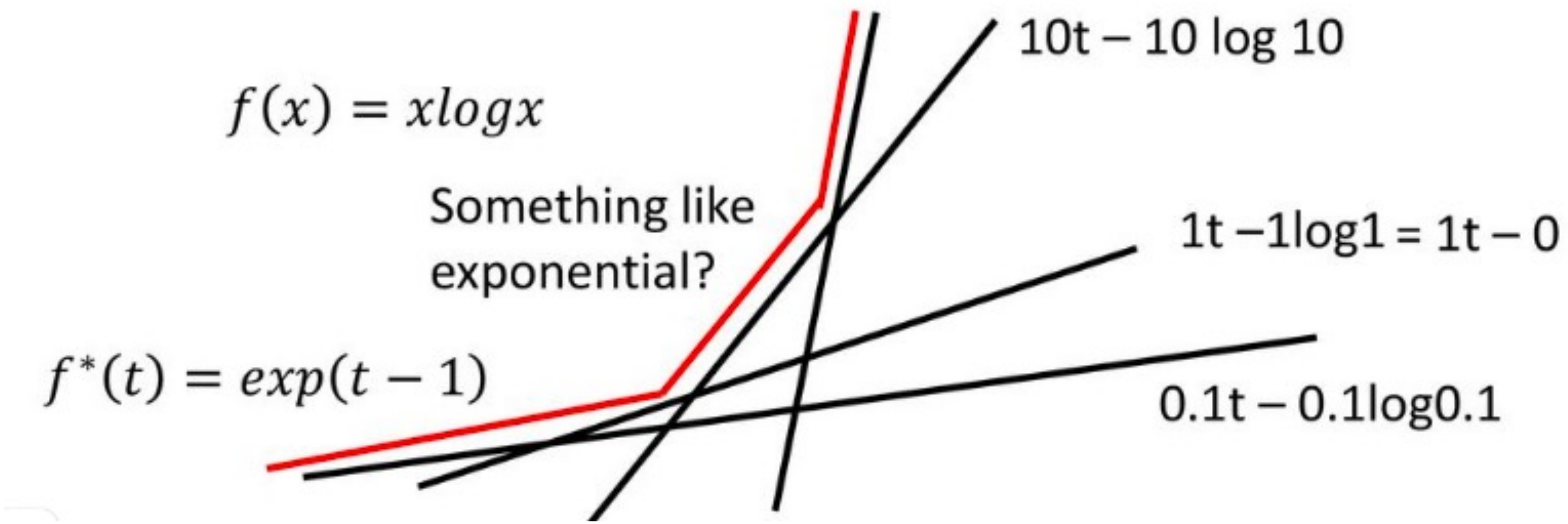


Fenchel conjugate

- Each generator has a **Fenchel conjugate** function:

$$f^*(t) = \sup_x (xt - f(x))$$

if f is convex, then $(f^*)^* = f$.



Fenchel conjugate

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

- Each generator has a **Fenchel conjugate** function:

$$f^*(t) = \sup_x (xt - f(x))$$

if f is convex, then $(f^*)^* = f$.

$$f(x) = (f^*)^*(x) = \sup_t (xt - f^*(t))$$

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

$$\geq \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{X}} p(x)\, T(x)\, \mathrm{d}x - \int_{\mathcal{X}} q(x)\, f^*(T(x))\, \mathrm{d}x \right)$$

$$= \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim P}\left[T(x)\right] - \mathbb{E}_{x \sim Q}\left[f^*(T(x))\right] \right),$$

Optimal Variational Function $T(x)$

- We can choose $T^*(x)$, an optimal variational function, that makes inequality tightest:

$$T^*(x) = f'\left(\frac{p(x)}{q(x)}\right).$$

X. Nguyen et al. Estimating divergence functionals and the likelihood ratio by convex risk minimization

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

Conclusions

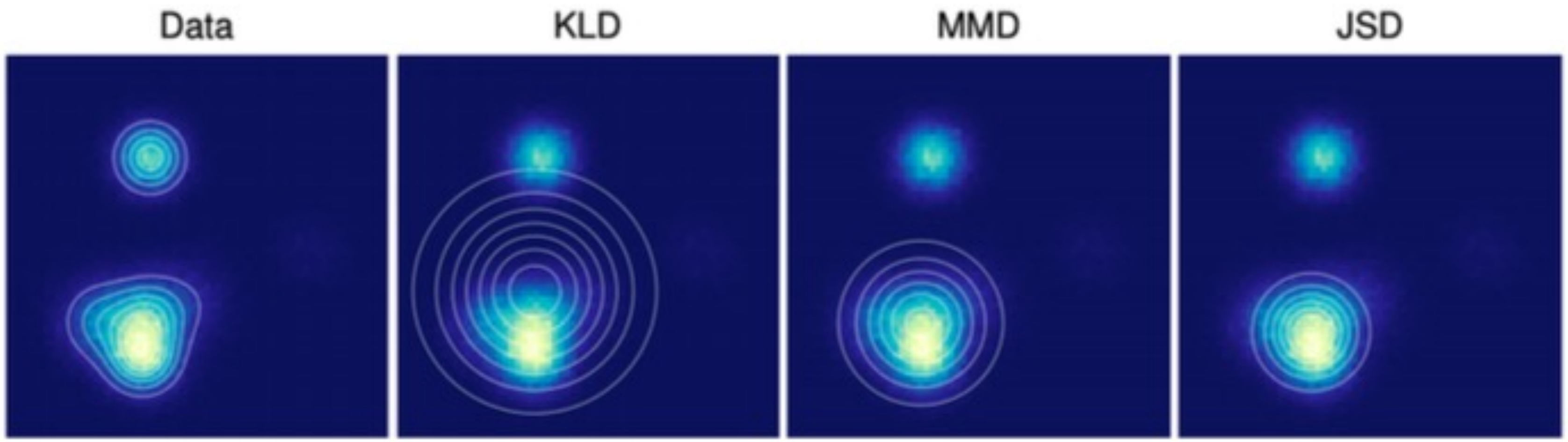


Figure 1: An isotropic Gaussian distribution was fit to data drawn from a mixture of Gaussians by either minimizing Kullback-Leibler divergence (KLD), maximum mean discrepancy (MMD), or Jensen-Shannon divergence (JSD). The different fits demonstrate different tradeoffs made by the three measures of distance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

Problems:

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

https://arxiv.org/abs/1506.05751

Conclusions

- f-divergences quantify dissimilarities between probability densities.
- Direct optimization of popular divergence requires the knowledge of true function.
- Special metrics should be developed to quantify quality of generative models (see seminar).