Generative Modeling

How to Use Deep Neural Networks to Produce a Cat

Denis Derkach

Laboratory for methods of big data analysis

HSE-Yandex Autumn School 2021, Moscow

LAMBDA • HSE

November 2021

In this Lecture

- What's a generative model.
- What it does.
- What are the main components.
- f-divergences
  - total Variation Distance;
  - Kullback-Leibler Divergence;
  - Jensen-Shannon Divergence.

Results of Generative Modeling

This X Does Not Exist!

This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.

Created by Phillip Wang.

This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.

Created by Ryan Hoover.

This Rental Does Not Exist

Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.

Created by Christopher Schmidt.

https://thisxdoesnotexist.com/

Video Modifications

We can **automatically** remove snow in video

https://incrussia.ru/news/ii-nauchilsya-poddelyvat-video/

More Tricks for Your Brain

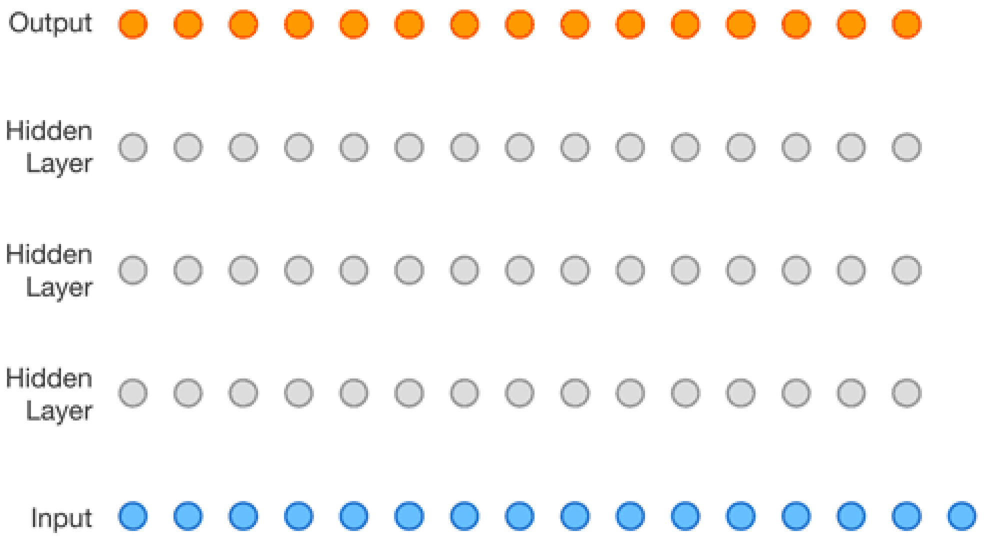
- Text generation.

Two men happily working on a plastic computer.
The toilet in the bathroom is filled with a bunch of ice.
A bottle of wine near stacks of dishes and food.
A large airplane is taking off from a runway.
Little girl wearing blue clothing carrying purple bag sit
SeqGAN (Baseline)
A baked mother cake sits on a street with a rear of it.
A tennis player who is in the ocean.
A highly many fried scissors sits next to the older.
A person that is sitting next to a desk.
Child jumped next to each other.
RankGAN (Ours)
Three people standing in front of some kind of boats.
A bedroom has silver photograph desk.
The bears standing in front of a palm state park.
This bathroom has brown bench.
Three bus in a road in front of a ramp.

More Tricks for Your Brain

- Text generation.

- Voice from text generation.



More Tricks for Your Brain

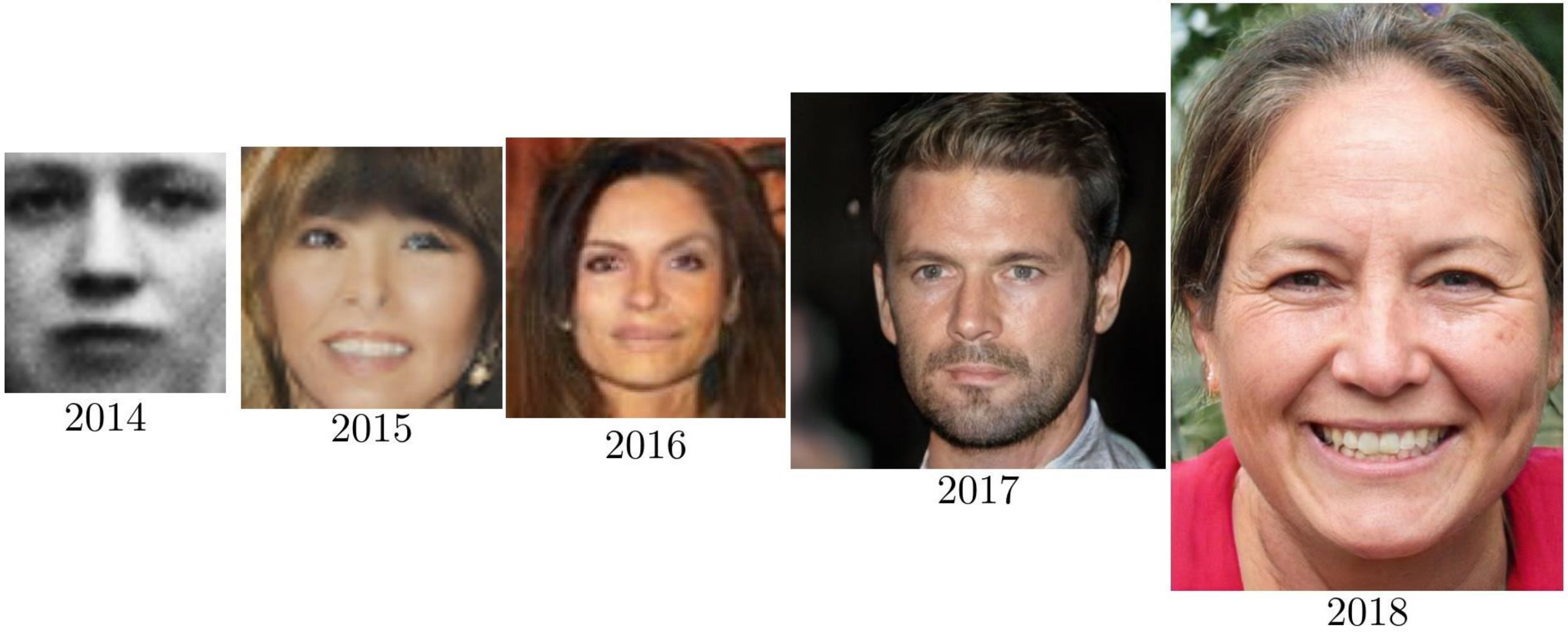
- Text generation.
- Voice from text generation.
- Style transfer.

More Tricks for Your Brain: Links

- Text generation.
  - https://www.tensorflow.org/tutorials/text/text_generation
- Voice from text generation.
  - https://deepmind.com/blog/article/wavenet-generative-model-raw-audio
- Style transfer.
  - https://towardsdatascience.com/style-transfer-with-gans-on-hd-images-88e8efcf3716

Generative Models Progress

The news are well motivated.



- Enormous progress in recent years.
- Technology is ready for new tasks.

https://twitter.com/goodfellow_ian/status/1084973596236144640

Generative Models Failures

- Image is created as **interpolation** between existing ones.


FAST COMPANY

02-08-19

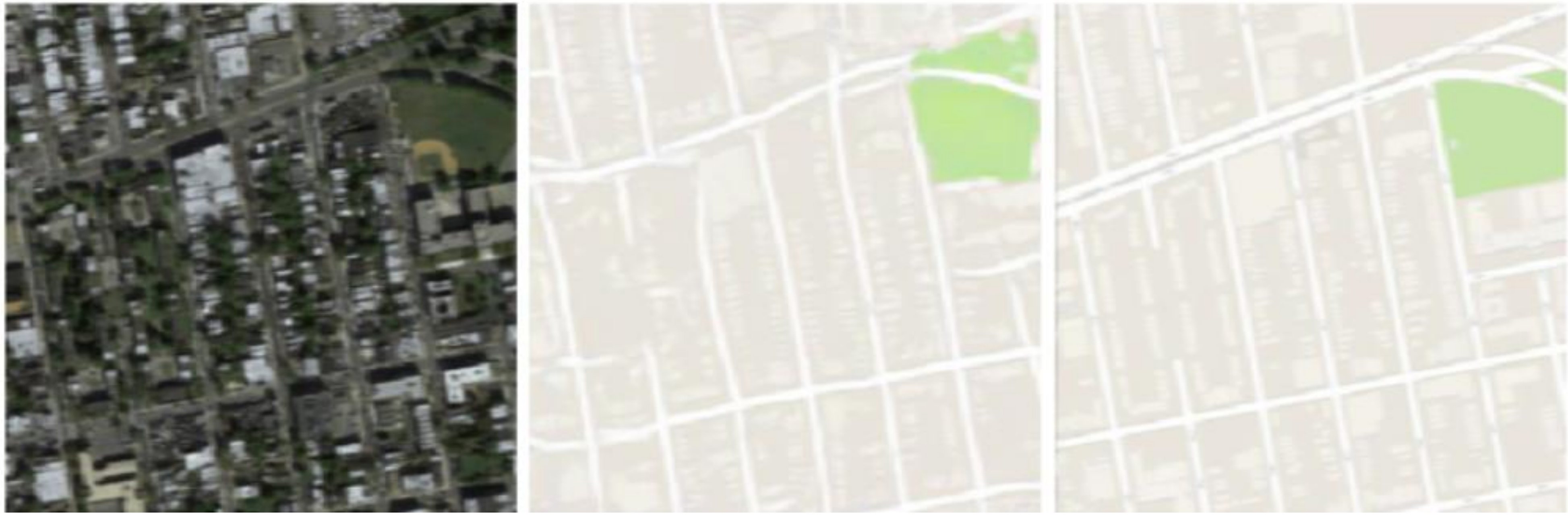
This AI dreams about cats–and they'll haunt your nightmares

Nvidia's new AI is capable of generating everything from human faces to kittens. But the development process left behind plenty of. . .errors.

https://www.fastcompany.com/90303908/this-ai-dreams-about-cats-and-theyll-haunt-your-nightmares

Dealing with Maps: generating map

- Image-to-image style transfer.
- Creates map on-the-fly from satellite image.



Input **Generated** **True**

https://github.com/ChengBinJin/pix2pix-tensorflow

Dealing with Maps: generating satellite image

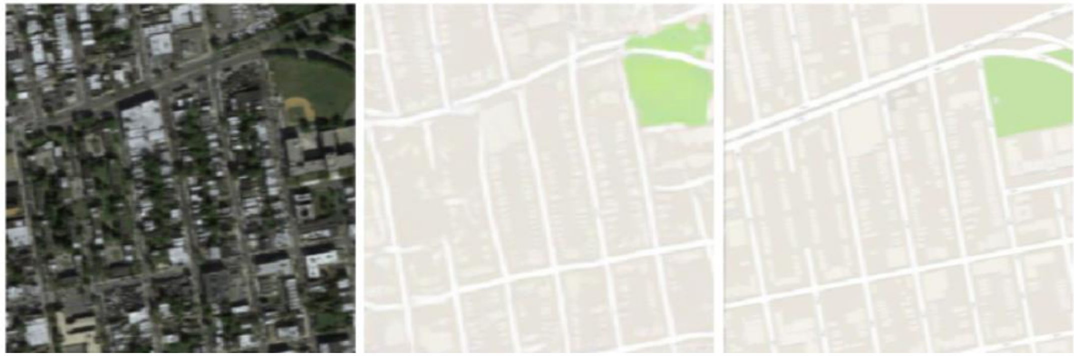
- Image-to-image style transfer
- Creates map on-the-fly from satellite image and vice versa.

Input **Generated** **True**

https://github.com/ChengBinJin/pix2pix-tensorflow

Dealing with Maps: generating satellite image

- Image-to-image style transfer
- Creates map on-the-fly from satellite image and vice versa.
- The technology is the same as for “Monet” painting. Just need good representation.



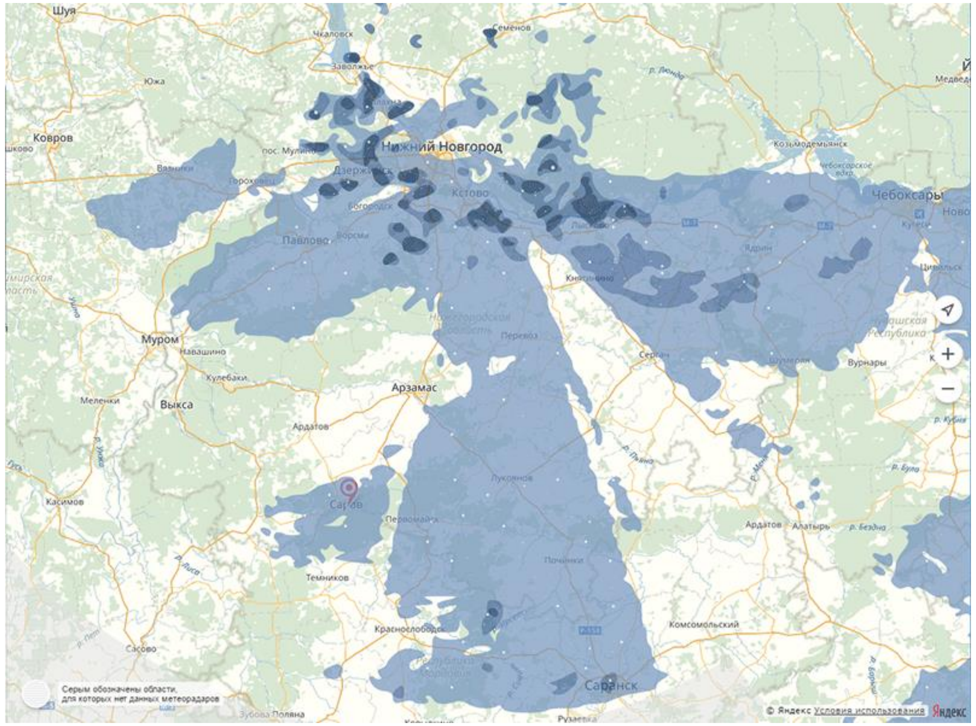
=

Dealing with Satellite Images: Super-resolution

- We can “create” a more appropriate map quality.
- This later can be used in segmentation task.

https://omdena.com/blog/super-resolution/

Weather prediction: nowcast



- Video prediction for precipitation.
- Generation of future state, based on the previous one.

https://www.kdd.org/kdd2019/accepted-papers/view/precipitation-nowcasting-with-satellite-imagery

Dirty Road Signs Generation

- Road signs from the book are too clean.
- Need to put mud and shadows on the signs.

https://arxiv.org/abs/1907.12902
https://www.hse.ru/sci/diss/426009543

What Generative Models **Do not** Produce

- No new information is created.
- All interpolations are done in some representation space.

Chapter outcome

- Generative models in machine learning were developing quickly in the last 6 years.
- Current state-of-the-art allows to implement generative models in more serious tasks than deceiving non-expert human.

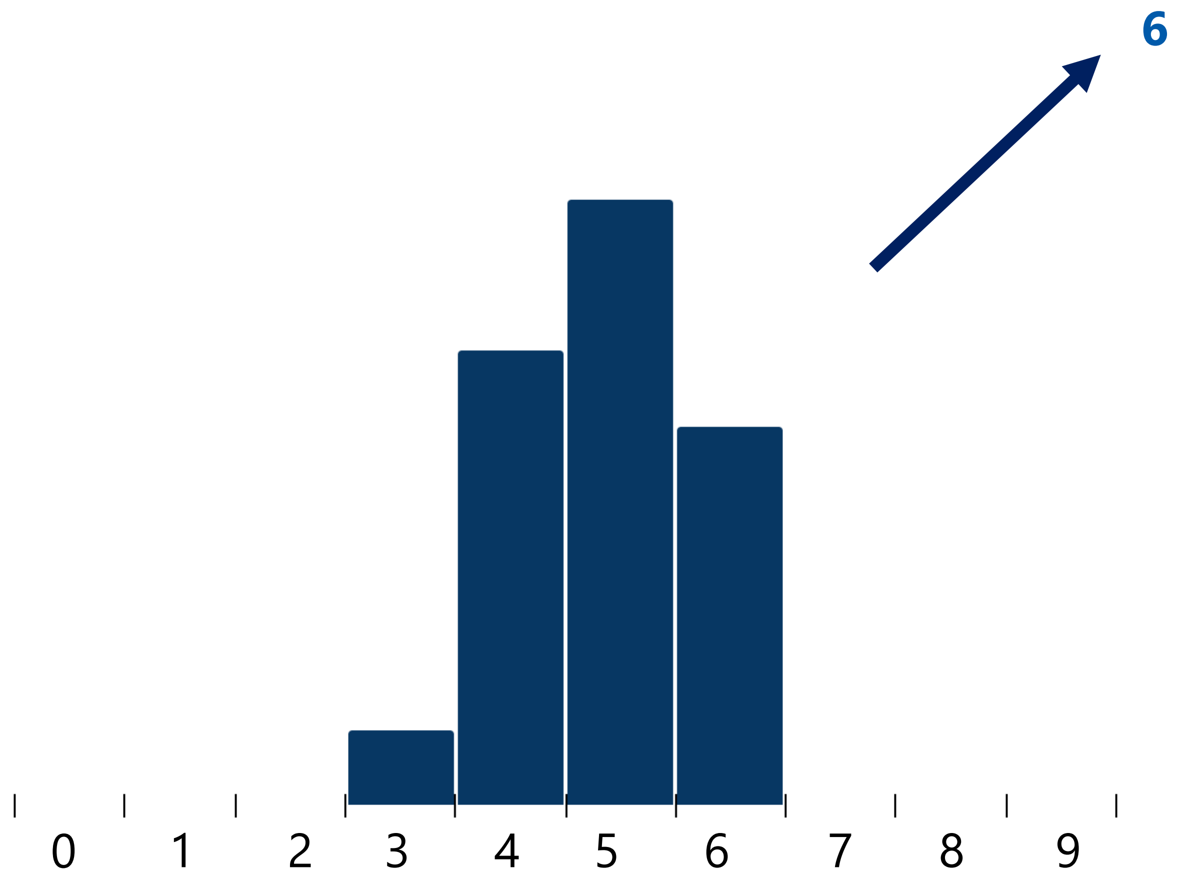
What is Generative Modeling

Random Number Generation

- We have sample with numbers:

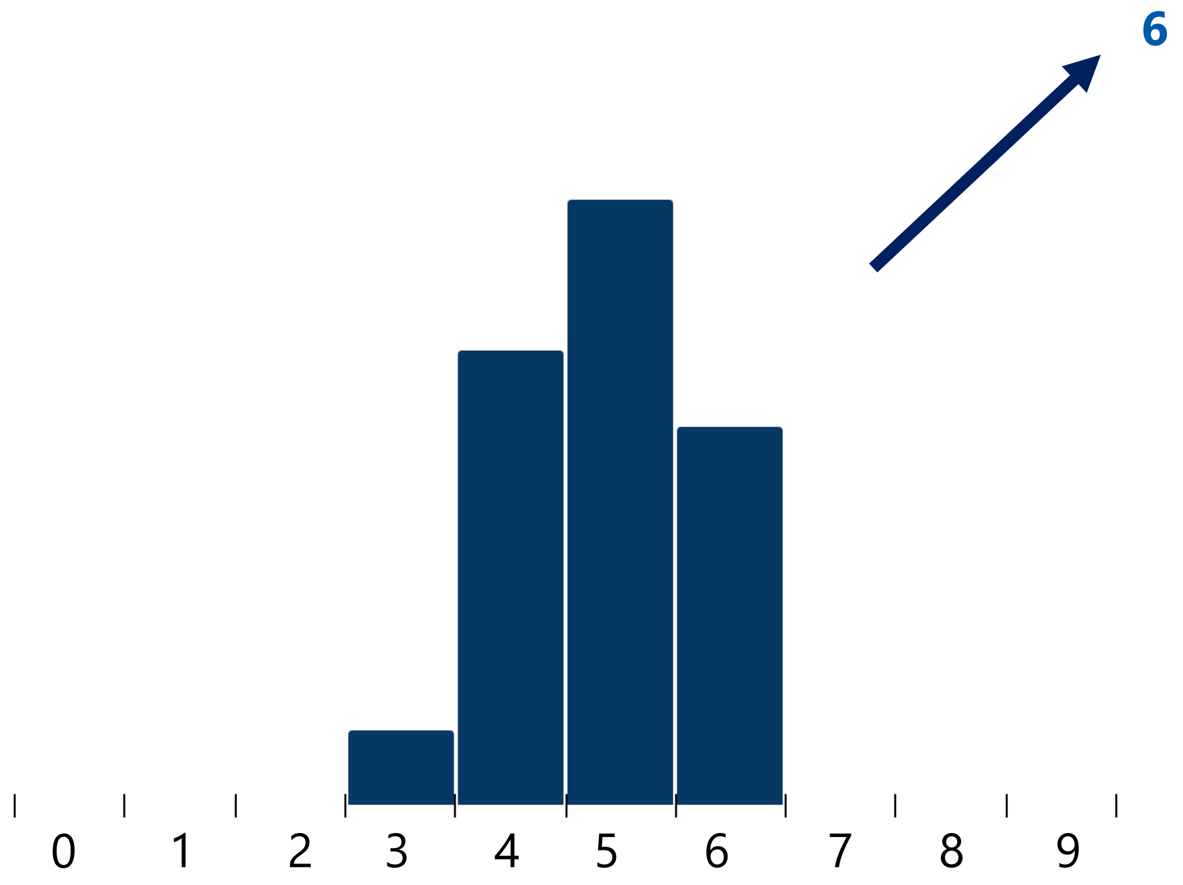
  3; 5; 4; 4; 4; 4; 5 ; 6 ; 5 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 6

- Want to create a new number alike.



How we did it?

- Assume there is a probability density $p_{true}(x)$.
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.

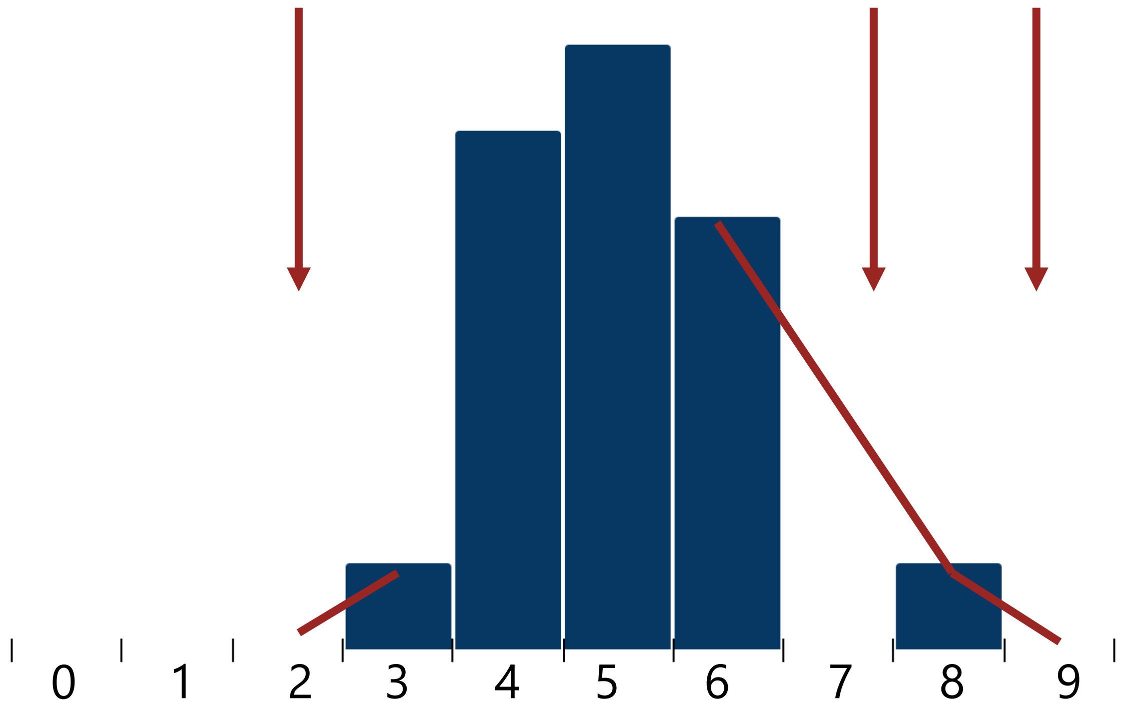


Random Number Generation

- We have **different** sample with numbers:

  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.

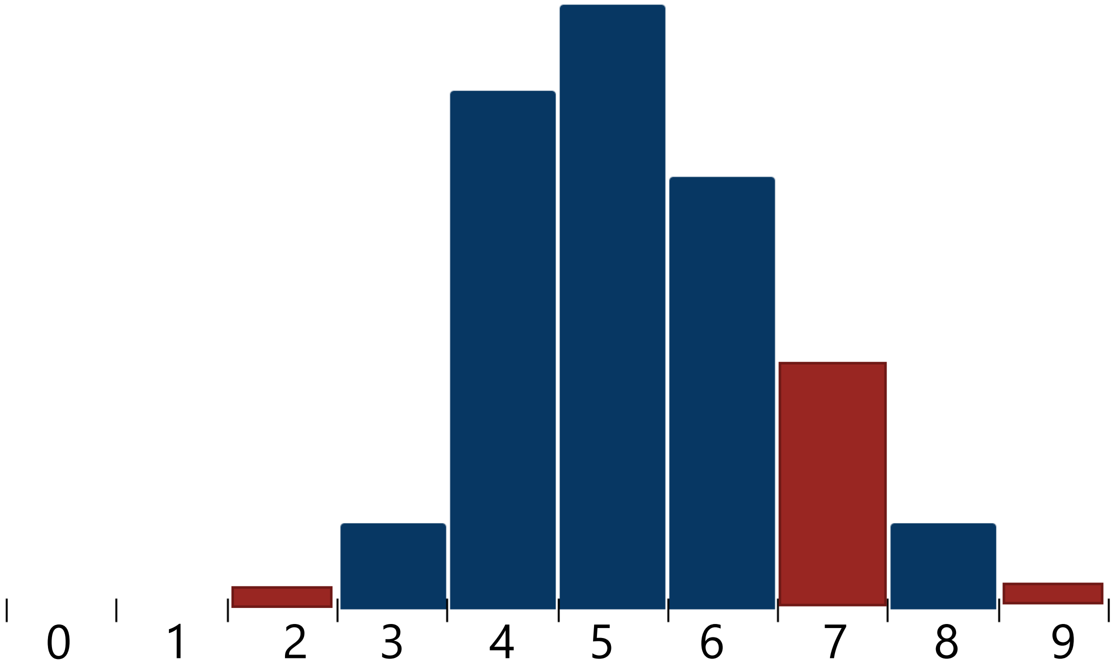


Random Number Generation

- We have **different** sample with numbers:

  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.

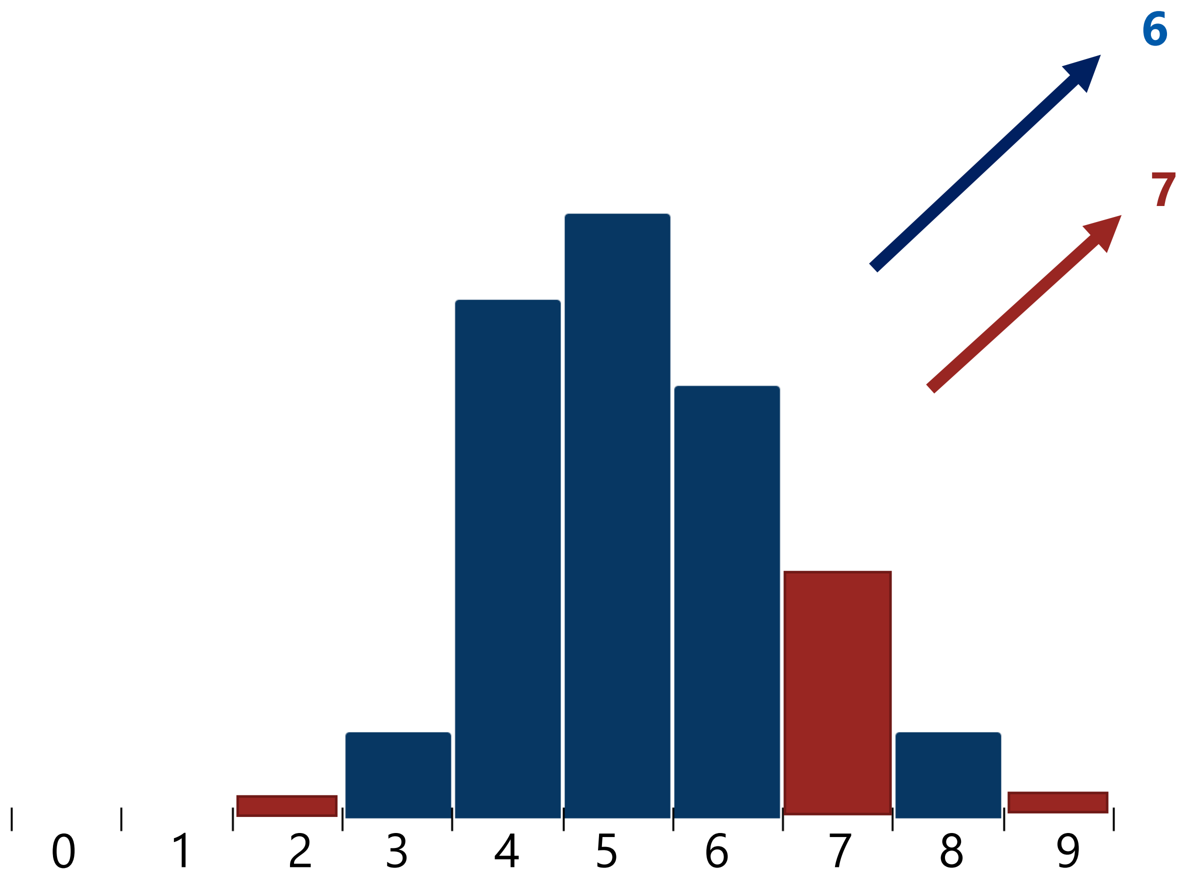


Random Number Generation

- We have **different** sample with numbers:

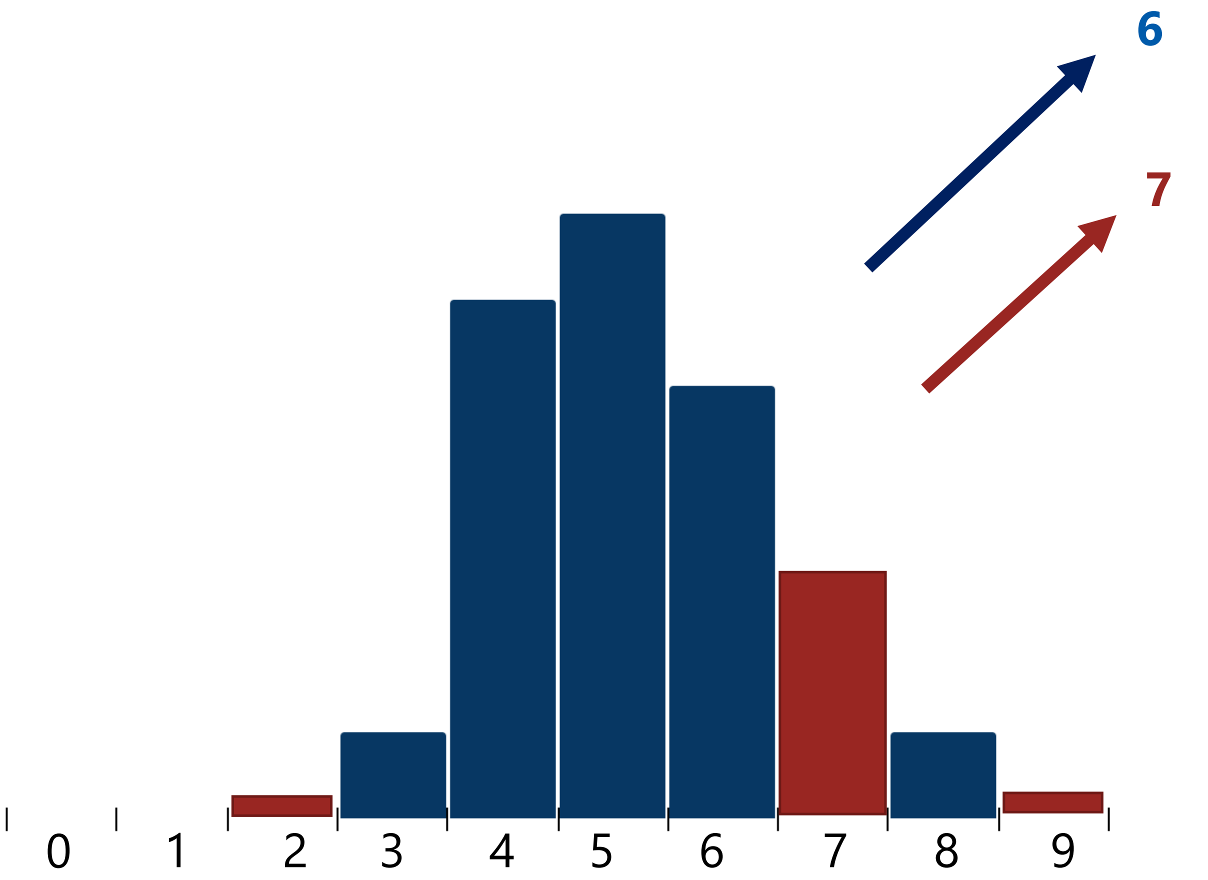
  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.



Random Number Generation

- Assume there is a probability density $p_{true}(x)$.
- **Choose interpolation model.**
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.

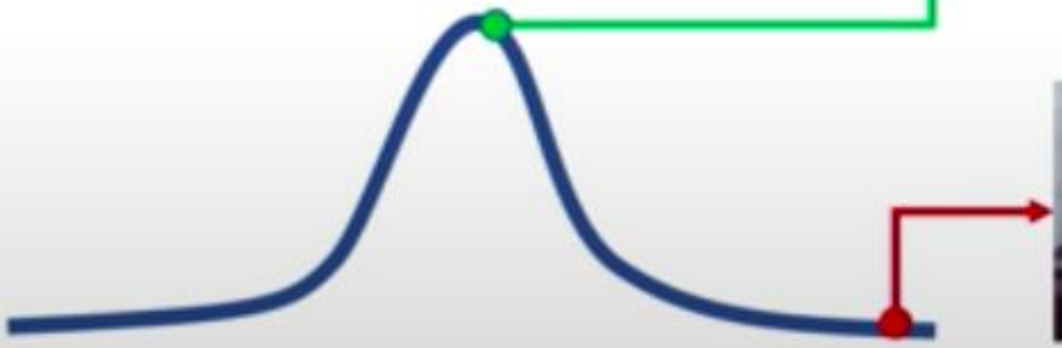


Case Study: Anomaly Detection

- **Problem:** How can we detect when we encounter something new or rare?
- **Strategy:** Leverage generative models, detect outliers in the distribution
- Use outliers during training to improve even more!

95% of Driving Data:
(1) sunny, (2) highway, (3) straight road

Detect outliers to avoid unpredictable behavior when training



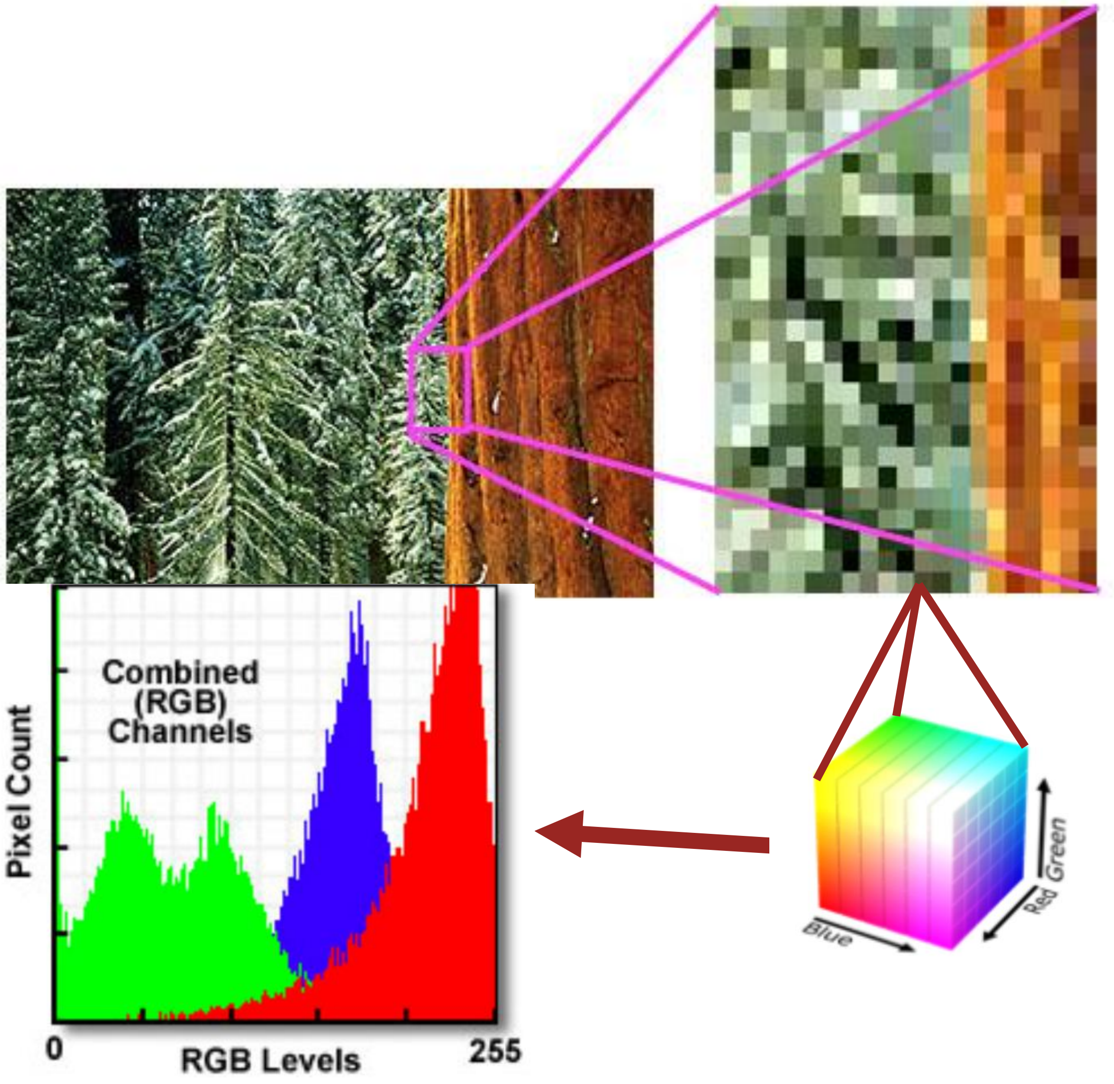
Edge Cases

Harsh Weather

Pedestrians

http://introtodeeplearning.com/

More Complicated Case: Figures



- Figure consists of pixels.
- One can use this representation.
- Each pixel is encoded by 3 colours.
- **Multi-modal distribution.**
- **Multidimensional problem.**

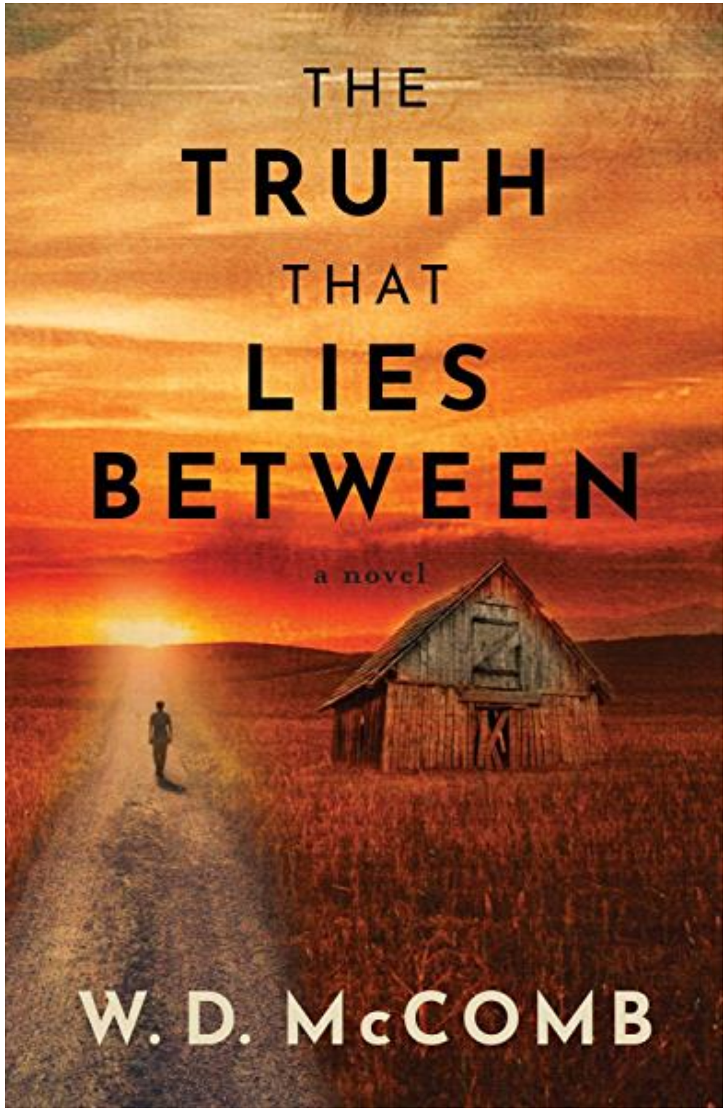
Number of Parameters

- Handwritten digits dataset.
- Only black and white pixels.
- Number of pixels 28X28.
- Number of possible states:

  $2x2x2x....x2 = 2^n.$
- **Number of parameters:**

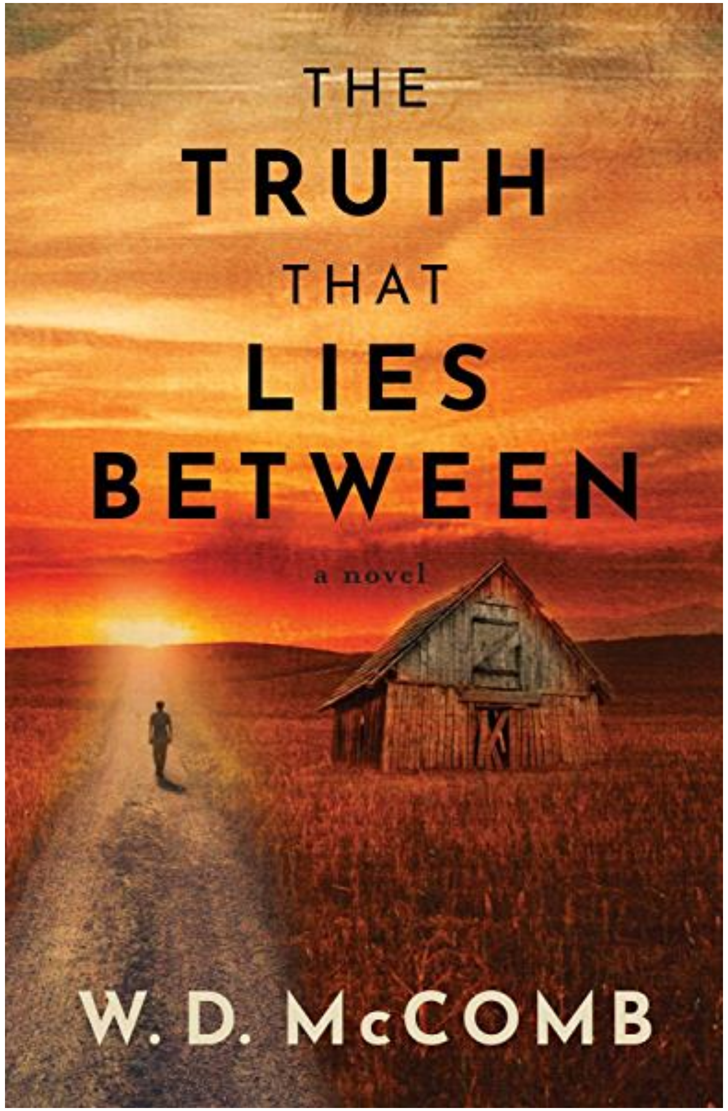
  $2^n - 1.$
- **For Independent pixels:**

  n.



Generative model: Final Touch

- Assume there is a probability density $p_{true}(x)$.
- Choose interpolation model.
- **Reduce number of dimensions.**
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.



Amazon

Generative model: Problem Statement

Three major tasks, given a generative model f from a class of models $\mathcal{F}$:

- **Estimation**: find the f in $\mathcal{F}$ that best matches observed data.
- **Evaluate Likelihood**: compute $f(z)$ for a given z.
- **Sampling**: drawing from f.

S. Nowozin et al. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Generative model vs Discriminative model

Discriminative models

- learn $\mathbb{P}(y|x)$
- Directly characterizes the decision boundary between classes only
- Examples: Logistic Regression, SVM, etc

Generative models

- learn $\mathbb{P}(x|y)$ (and eventually $\mathbb{P}(y, x)$)
- Characterize how data is generated (distribution of individual class)
- Examples: Naive Bayes, HMM, etc.

https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

Chapter outcome

- Generative modeling is a distinct task in machine learning.
- Mathematically, it aims to reconstruct the probability density, from which the given dataset was sampled.

Early Generative Models

First ideas

For parametric model.

- **Inversion sampling**. For x with CDF $F_X(x)$:

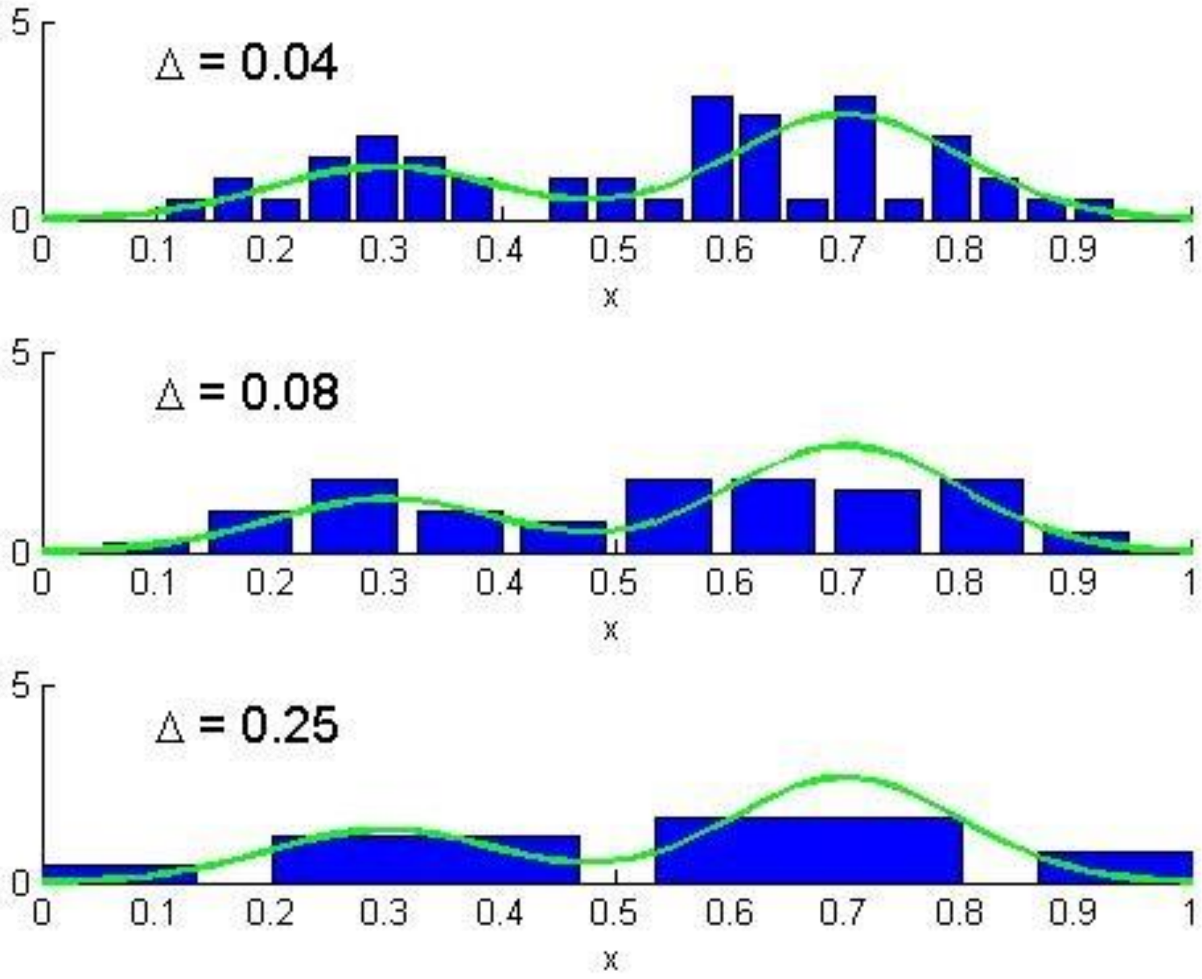
$$z \sim \mathrm{Unif}(0;1);\; x = F_X^{-1}(z).$$

- **Works in multidimensions**. Sample successively.
  - Generate X from the marginal $p_X(x) = \int p_{X,Y}(x,y)\mathrm{d}y$.
  - Generate Y given $X = x$ from the conditional $p_{Y|X}(y|x) = \frac{p_{x,y}(x,y)}{p_X(x)}$.

For 1D Gaussian model, the convergence is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

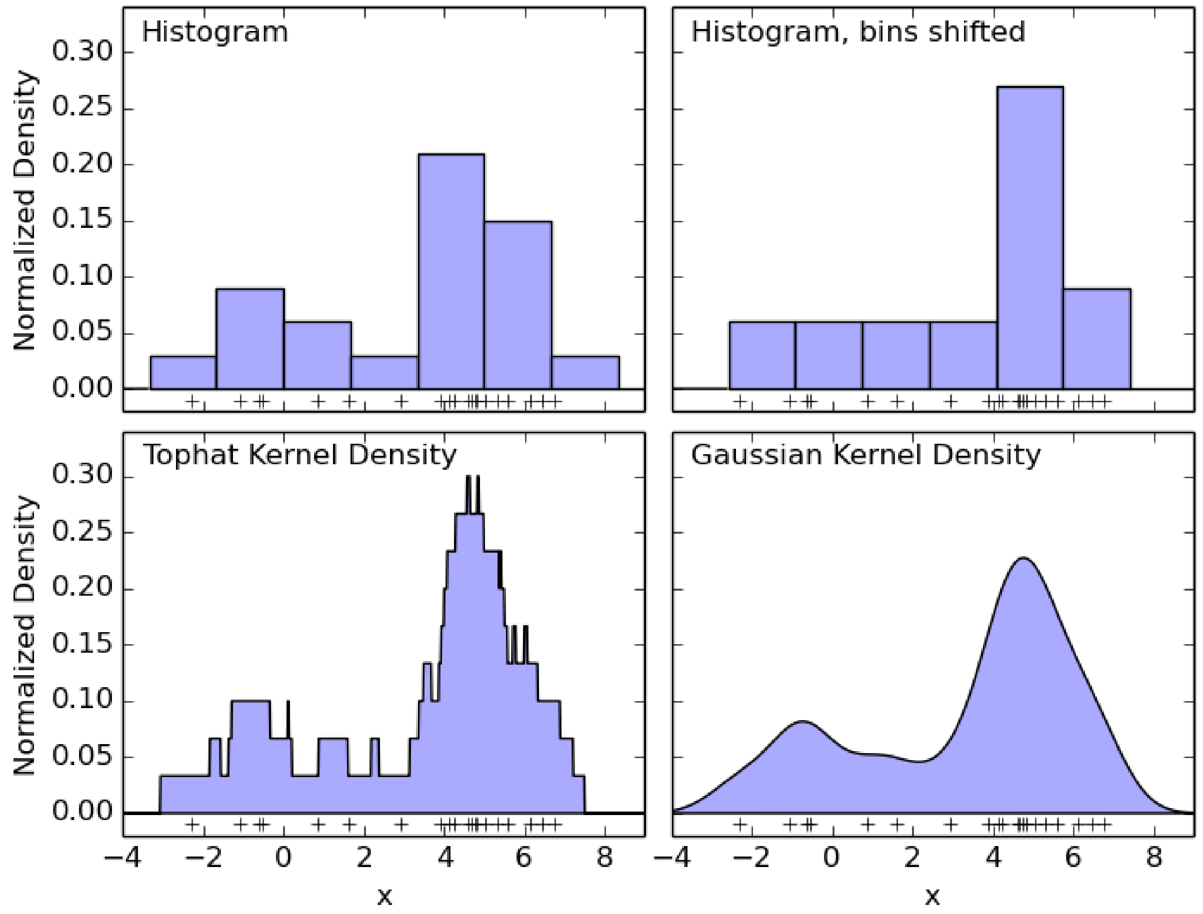
“Non-parametric” Approaches

- Histograms can be used.
- Need to choose optimal bin size.
- Smaller bins for approximate constant estimate.
- Larger bins for less fluctuations.
- Can be chosen using empirical risk.



SunLei's Linear Space

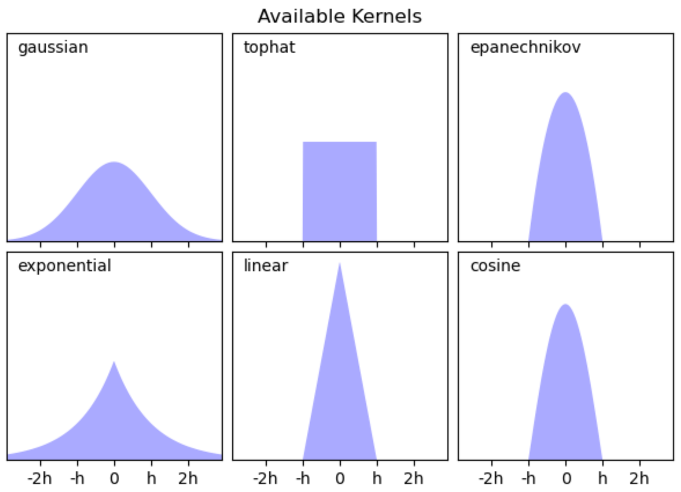
Kernel-density estimation



- Assign **every** event a weight.
- Smooth between events.
- Kernel Density Estimation:

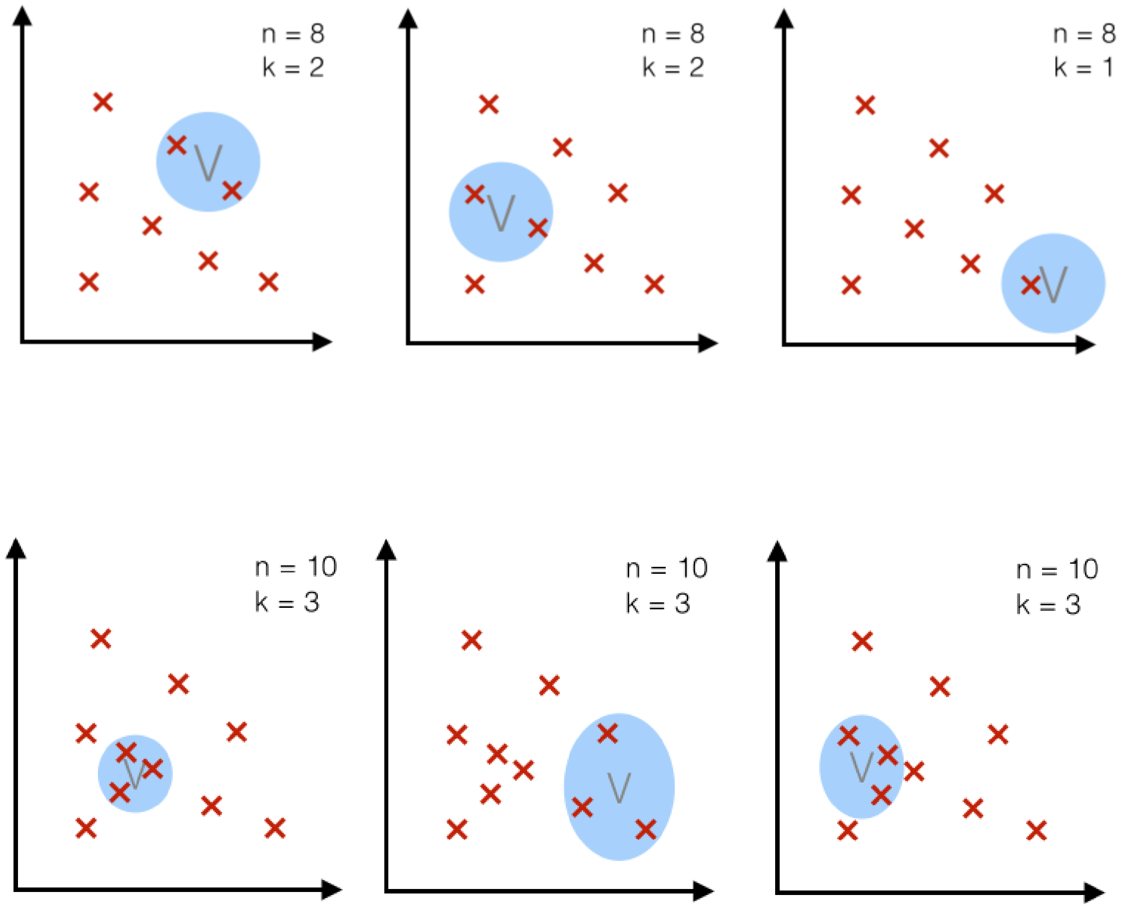
$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right),$$

K – some kernel, h – bandwidth.



Sci-kit

KDE2KNN



- With fixed volume kernel outliers can lead to fluctuations in $\hat{p}(x)$.
- Vary kernel volume to cover k nearest neighbors.
- Better coverage of tails.

S. Raschka's blog

KDE and kNN Optimal Parameter Choice

Minimize integral MSE (or L2 risk function) to determine optimal parameter and convergence:

$$MSE\big(\hat{p}_n(x_0)\big) = \mathrm{bias}^2(\hat{p}_n(x_0)) + Var(\hat{p}_n(x_0)).$$

$$MISE(\hat{p}_n) = \int MSE\big(\hat{p}_n(x)\big)\, dx$$

This is not a straightforward task (need cross-validation selector) but can be solved under some conditions.

$$MISE_{opt}(\hat{p}_n) = \mathcal{O}(\mathrm{n}^{-\frac{4}{4+d}}).$$

when *d* is large, the optimal convergence rate is very very slow.

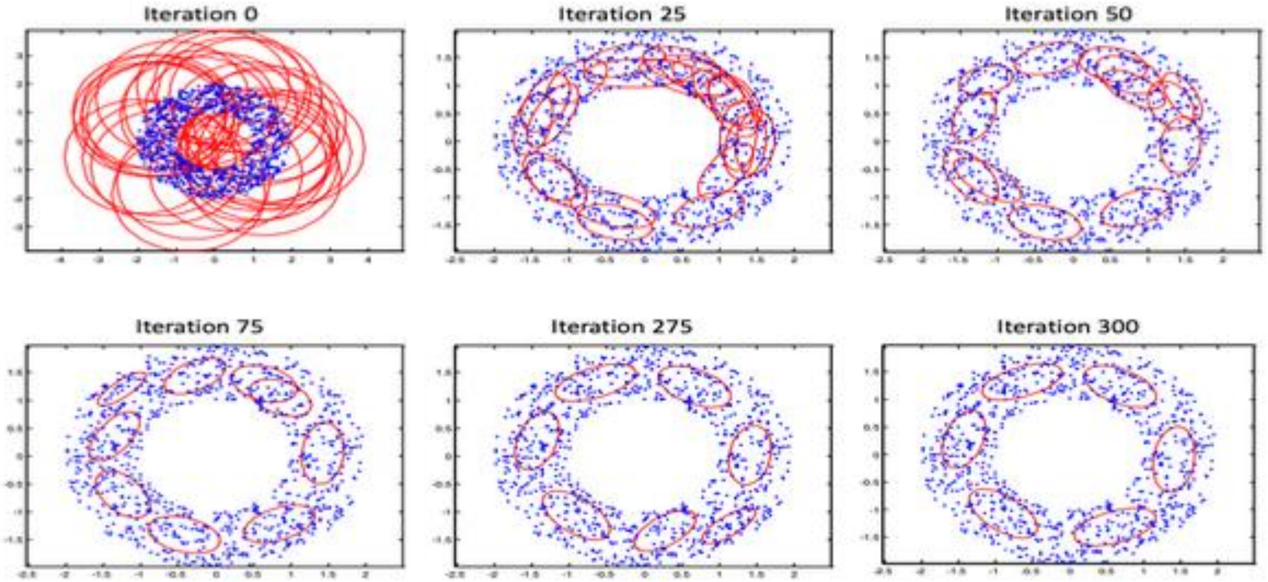
For example, L. Wasserman Lecture Notes

KDE and kNN Summary

- Efficient in low dimensional estimation.
- Controllable convergence rate for bias or variance but the overall rate is similar.
- Mixture of KDE and kNN are available.
- To speed up the convergence, once can attempt to find manifolds in the d-dimension.
- Fairly hard to sample and keep the model in memory.

Gaussian Mixture Model

Training set: $n = 900$ examples from a uniform pdf inside an annulus,
model: GMM with $K = 30$ Gaussian components



- Reduce number of Gaussians.
- Infer Gaussian parameters from data.
- Estimate density:

$$\hat{p}_n(x) = \sum_{l=1}^{K} \pi_l \phi(x; \mu_l, \sigma_l)$$

K Gaussian distribution ϕ is used.

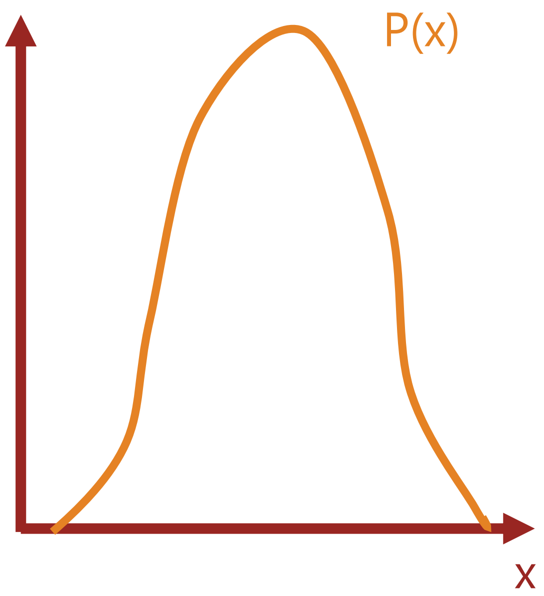
- **Need EM-algorithm**

GMM Summary

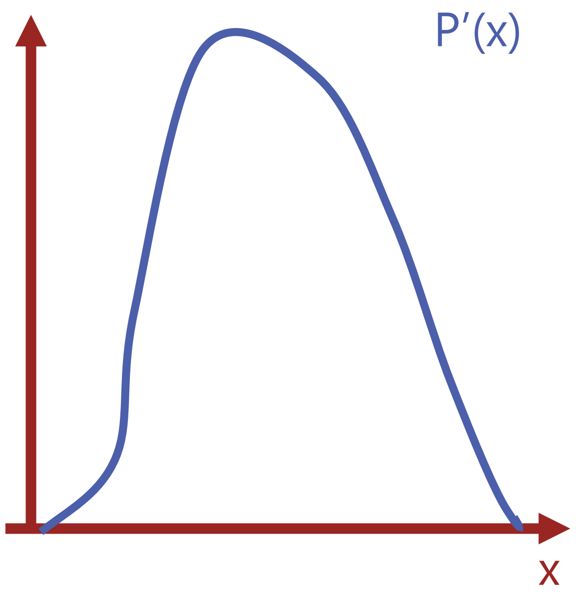
- Better convergence rate than non parametric.
- Identifiability problem. We cannot distinguish between two exchanged solutions. .
- Computation problem. We need to use EM algorithm to find solution.
- Choice of K. A very difficult task, one may use a model selection technique to choose it, however, no simple rule exists.
- Does not really converge to a true PDF.

Total Variation Distance

What we measure



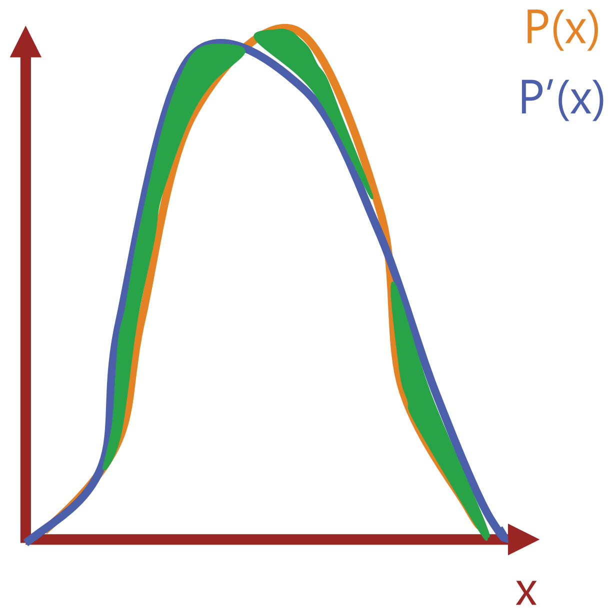
True Probability Density



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

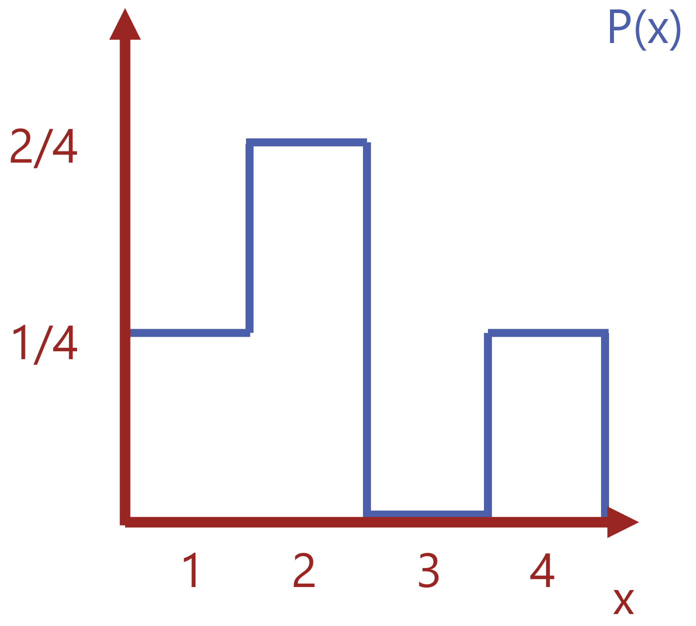
$$\boxed{D(p(x), q_\theta(x)) = \sup_{\mathrm{A}} \left| \int_{\mathrm{A}} \mathrm{p(x)dx} - \int_{\mathrm{A}} \mathrm{q_\theta(x)dx} \right|}$$

Where A is any measurable set.

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

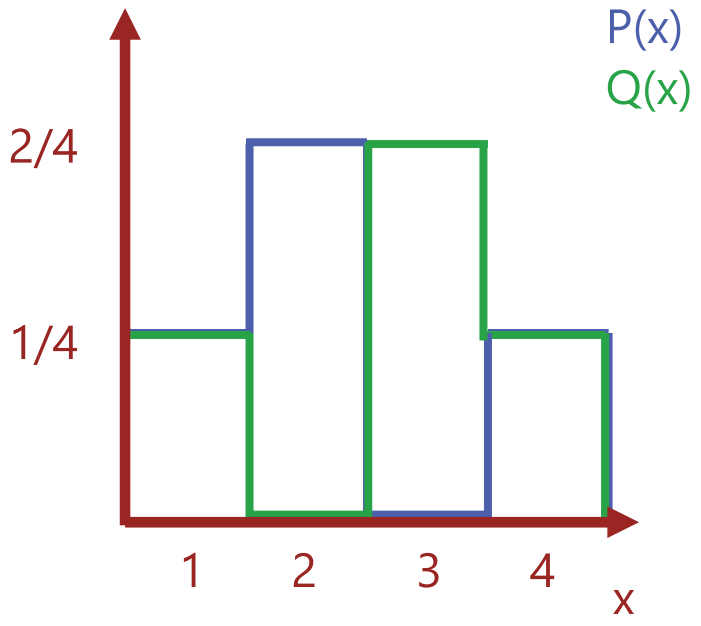
Total Variation Distance: example 1D

- discrete case for two PDFs



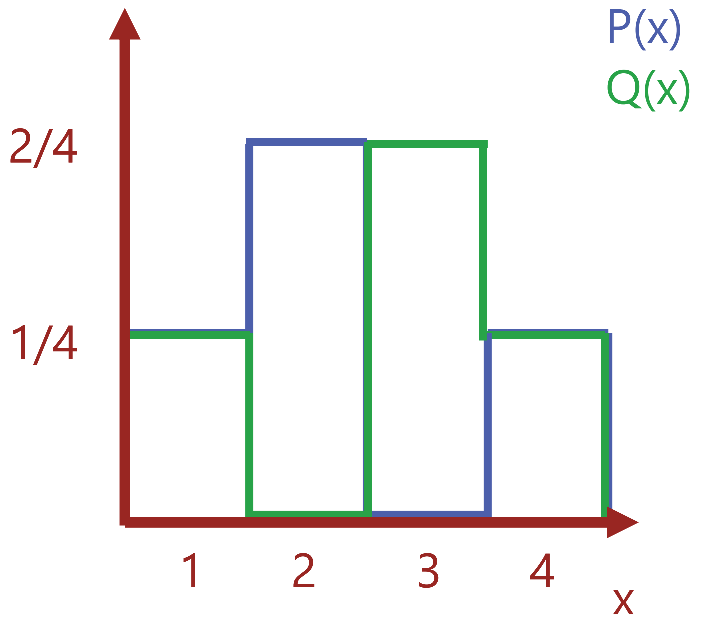
Total Variation Distance: example 1D

- discrete case for two PDFs



Total Variation Distance: example 1D

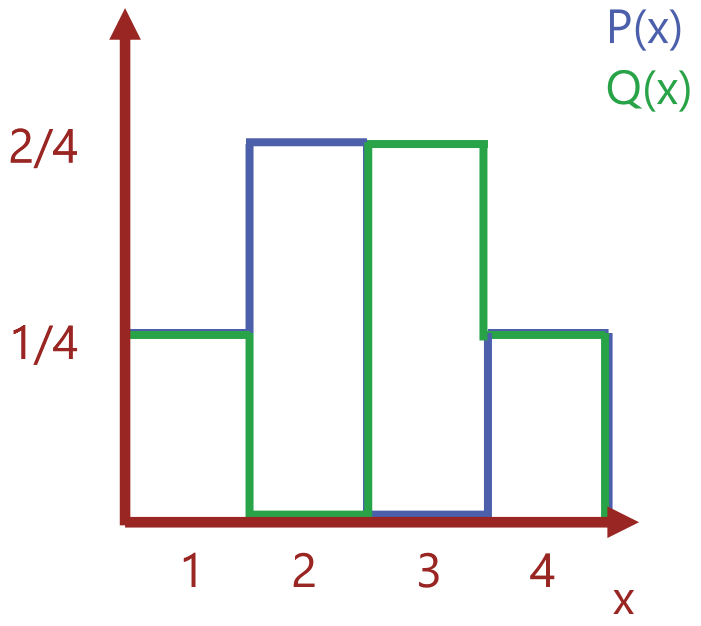
- discrete case for two PDFs
- calculate in two ways:



Total Variation Distance: example 1D

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

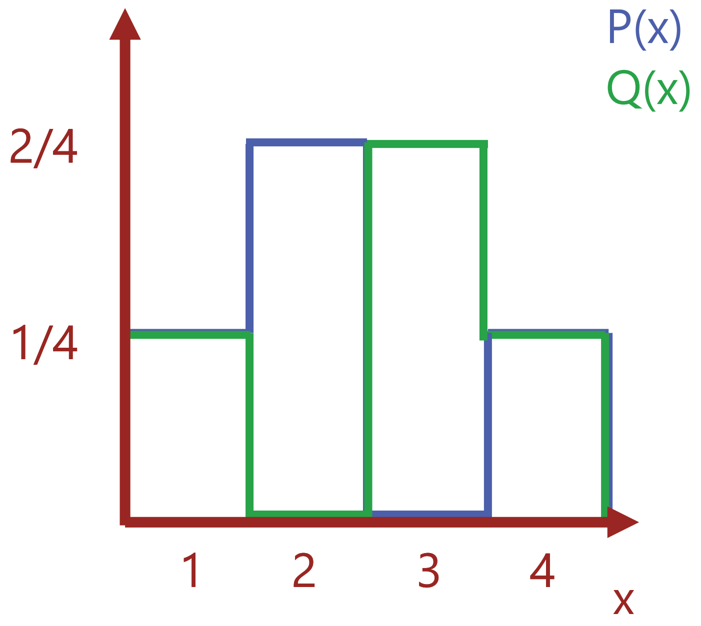
    {1}, {2}. {3}, {4}, {1;2}, {1;3}, {1;4}, {2;3}, {2;4}, {3;4}, {1;2;3}, {1;2;4}, {1;3;4}, {1,2,3,4}.



Total Variation Distance: example 1D

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

    $D(p,q) = 0.5$



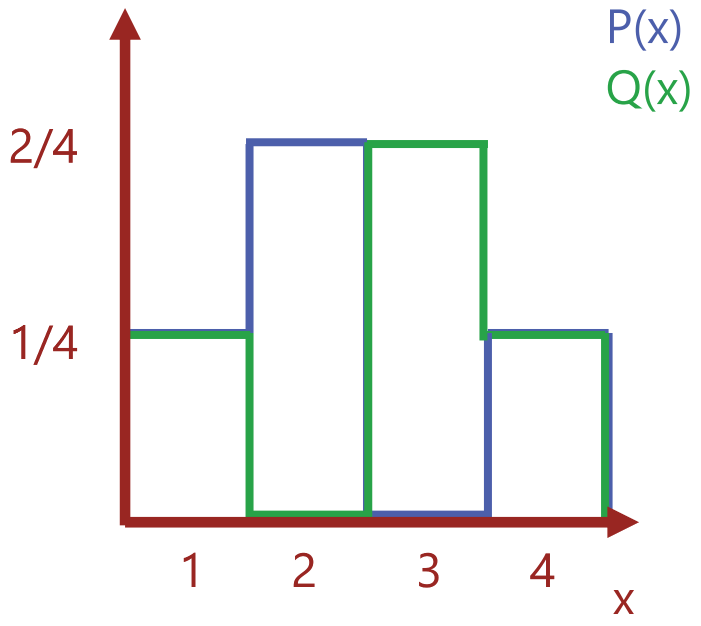
Total Variation Distance: example 1D

- 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

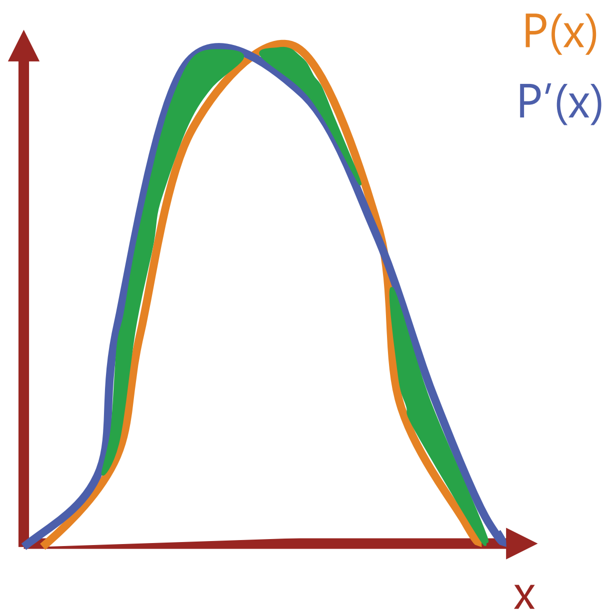
$$S_n/\sqrt{n} \to \mathcal{N}(0,1),$$

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

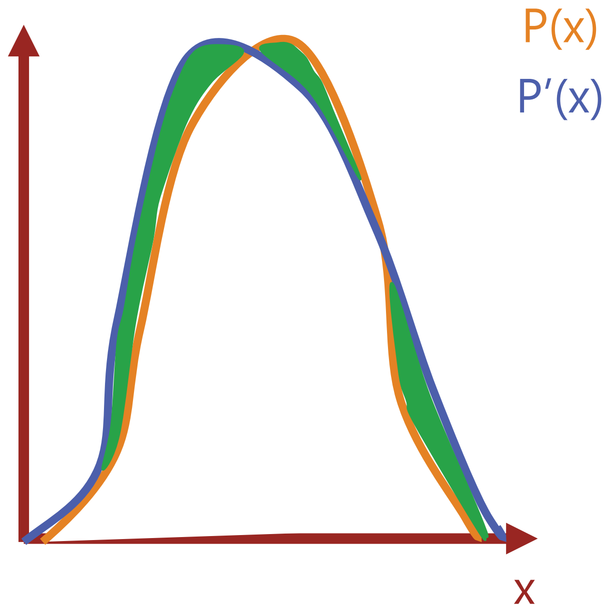
Kullback-Leibler divergence: ideas



Previously:

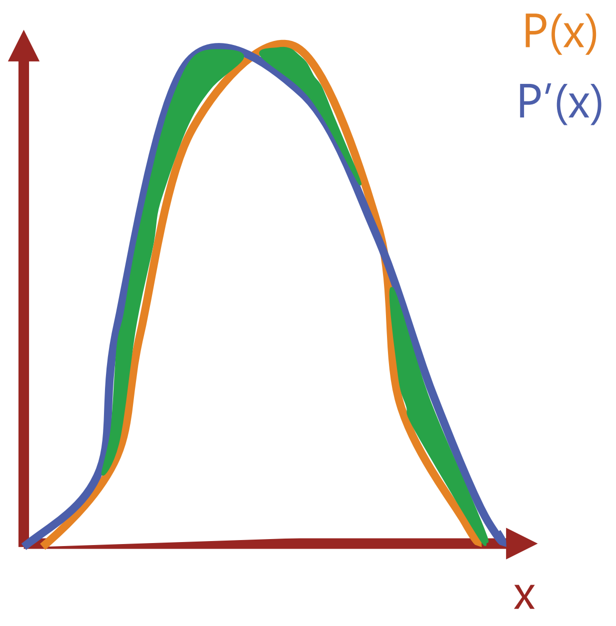
$$\int |P(x) - P'(x)|\, dx$$

Kullback-Leibler divergence: ideas



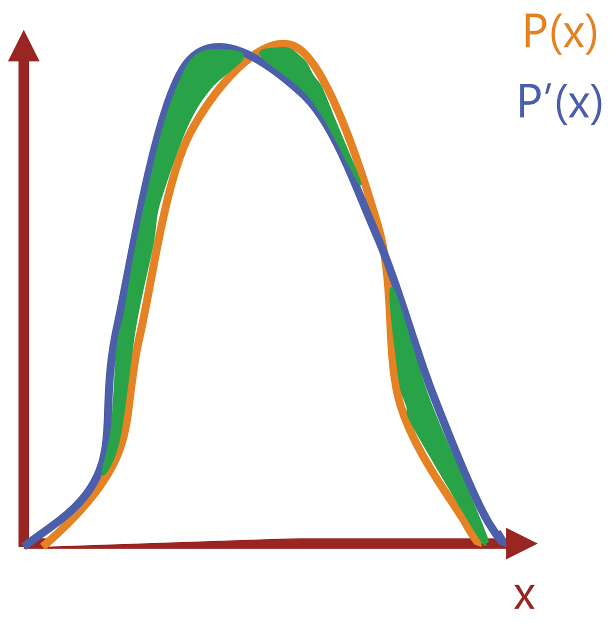
$$\frac{P(x)}{P'(x)}$$

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:**

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

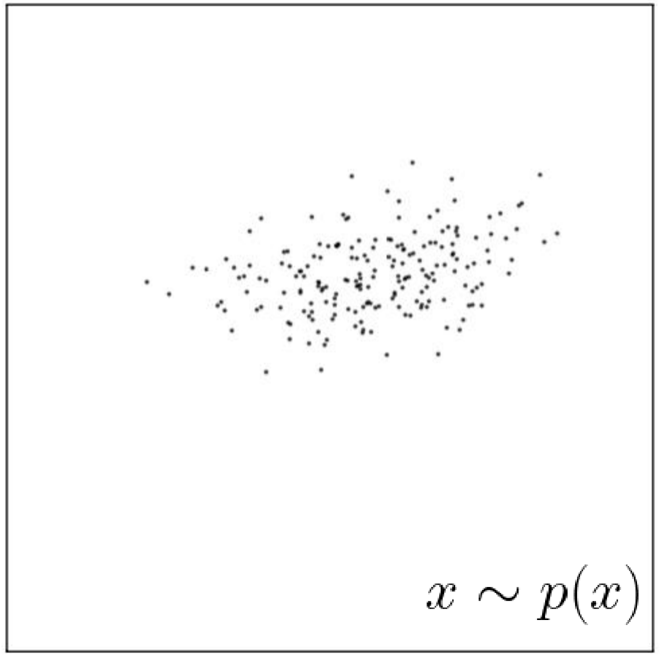
where $H(p,q) = \mathbb{E}_p(\log q)$.

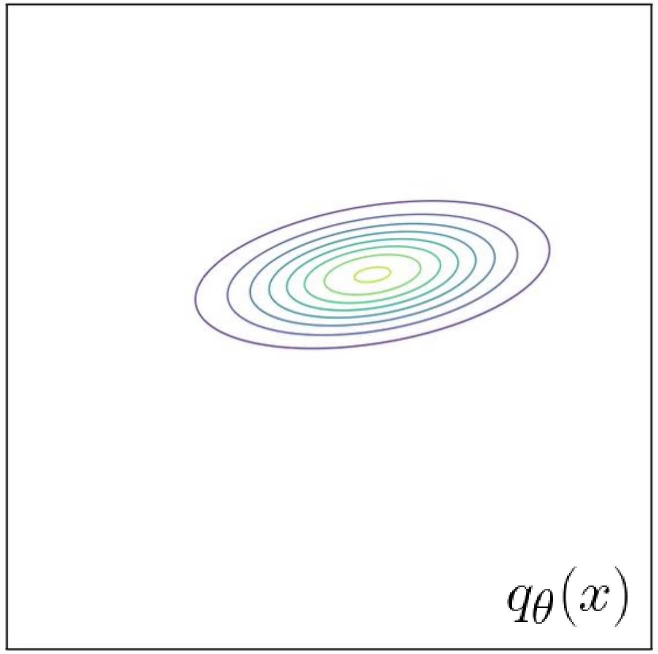
- **Minimizing KL divergence is equivalent to maximizing the likelihood.**

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

Using in fits

Fit data points from 2D Gaussian function

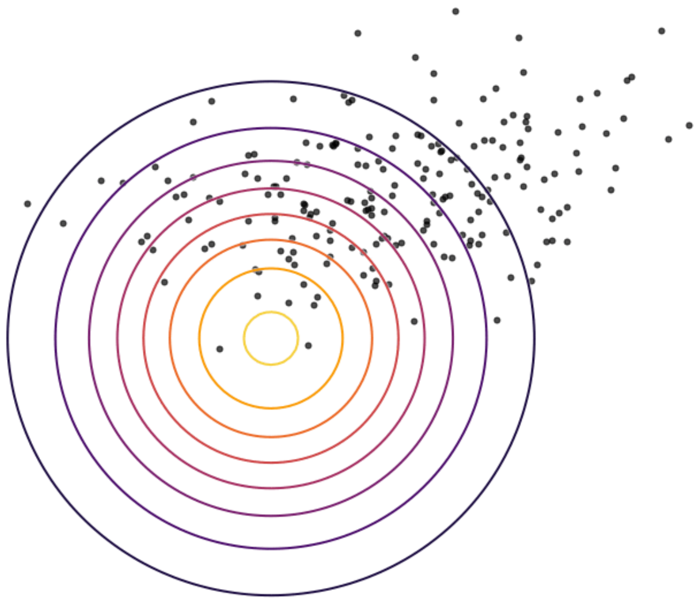




...with 2D Gaussian function

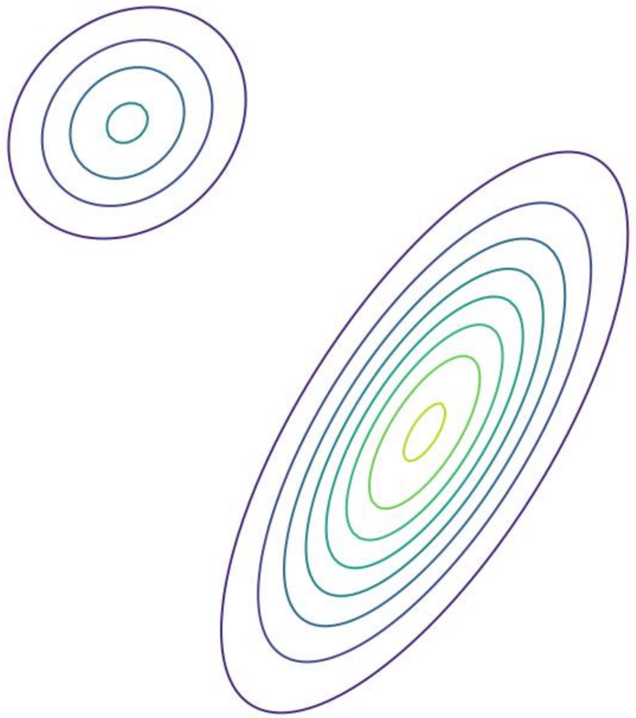
Here and Later: Colin Raffel's blog

Using in fits



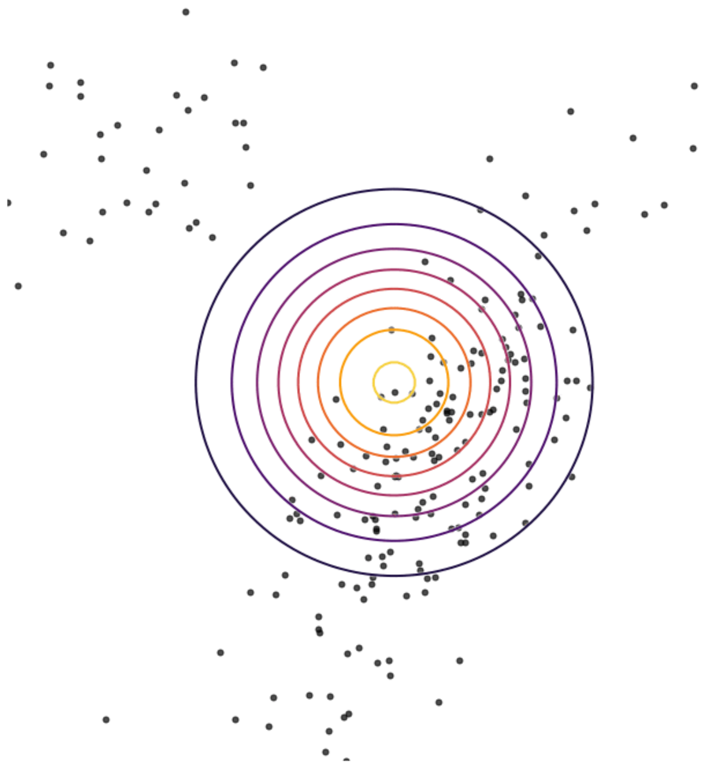
- Runs smoothly for simple data

Using in fits: Multimodal data



- Runs smoothly for simple data

Using in fits: Multimodal data

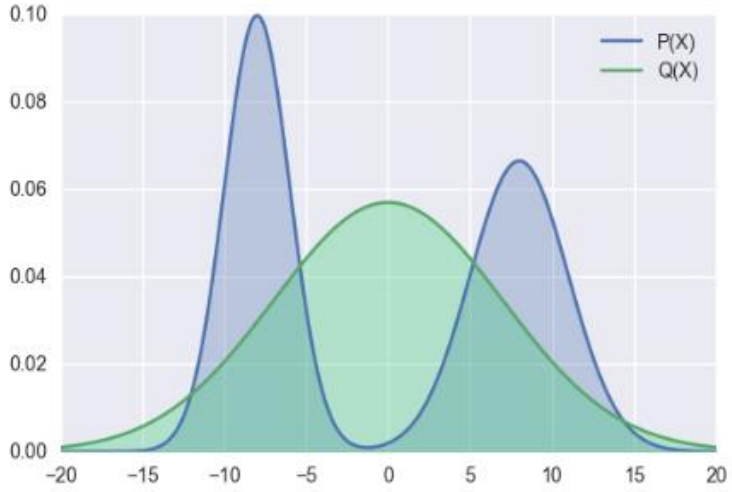


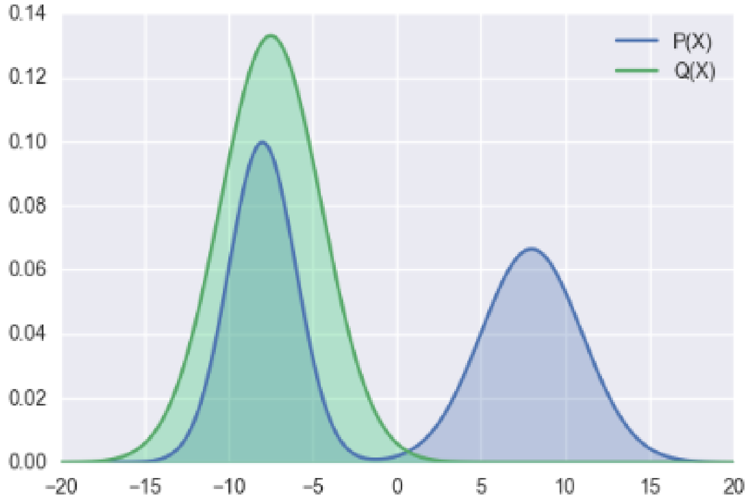
- Runs smoothly for simple data
- Problems for multimodal data
- Covers significant amount of empty spaces

KL divergence: study

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

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





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

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^*:

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

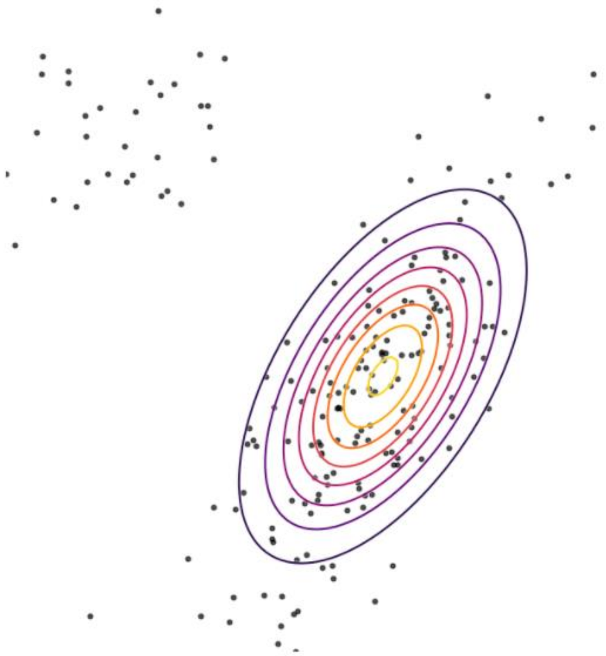
entropy for the
fitted model

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

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

Final Summary

- Generative modeling is a distinct task of machine learning.
- Several pre-deep learning algorithms can produce reasonable results in the low dimensional data.
- Denoising Autoencoder is one of the first pseudo-generative models.