# Implicit generative models

Dmitry Ulyanov

*PhD student at Skoltech*

Deep|Bayes

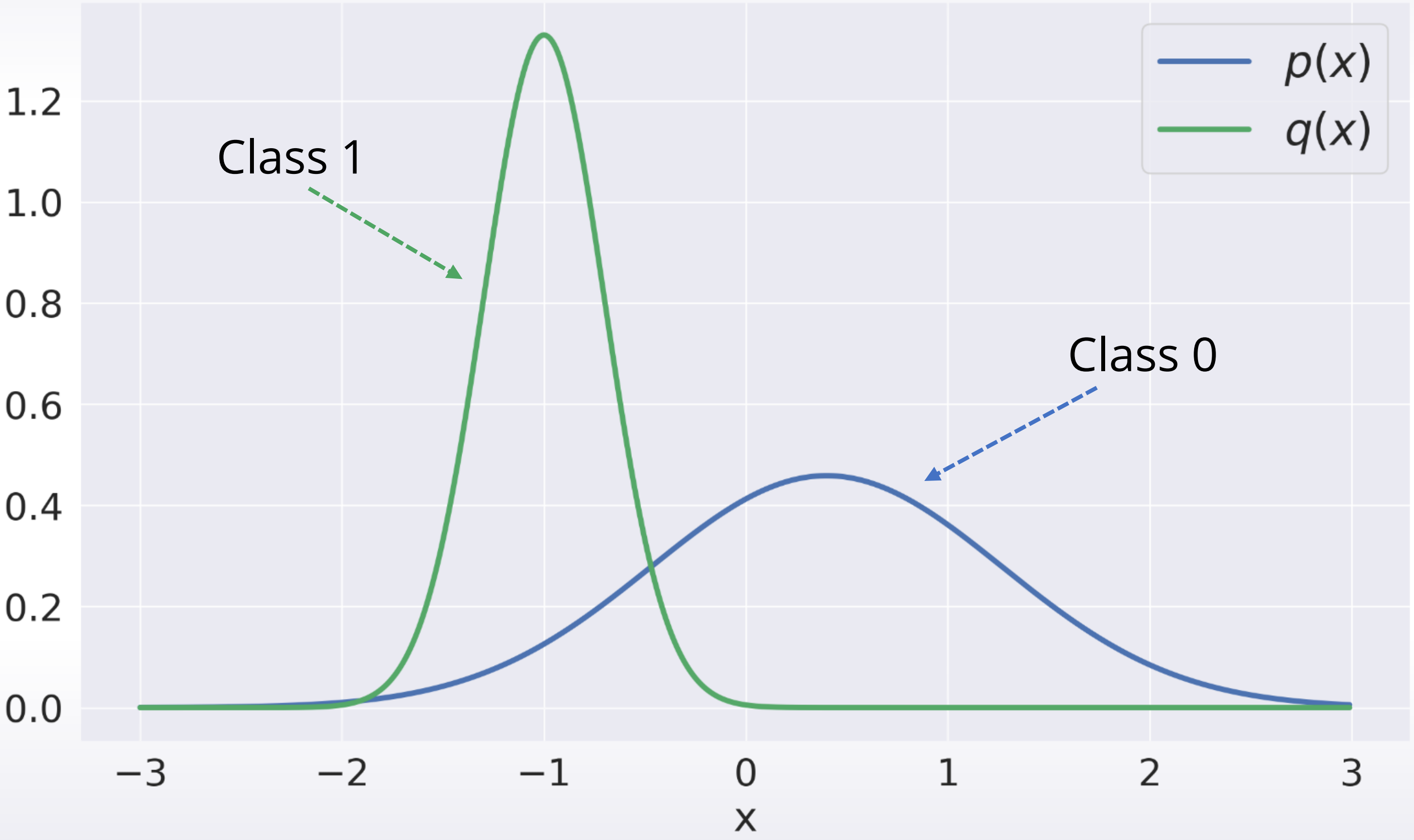
# Outline

- Implicit generative models
- Distribution divergences
- Learning in implicit models

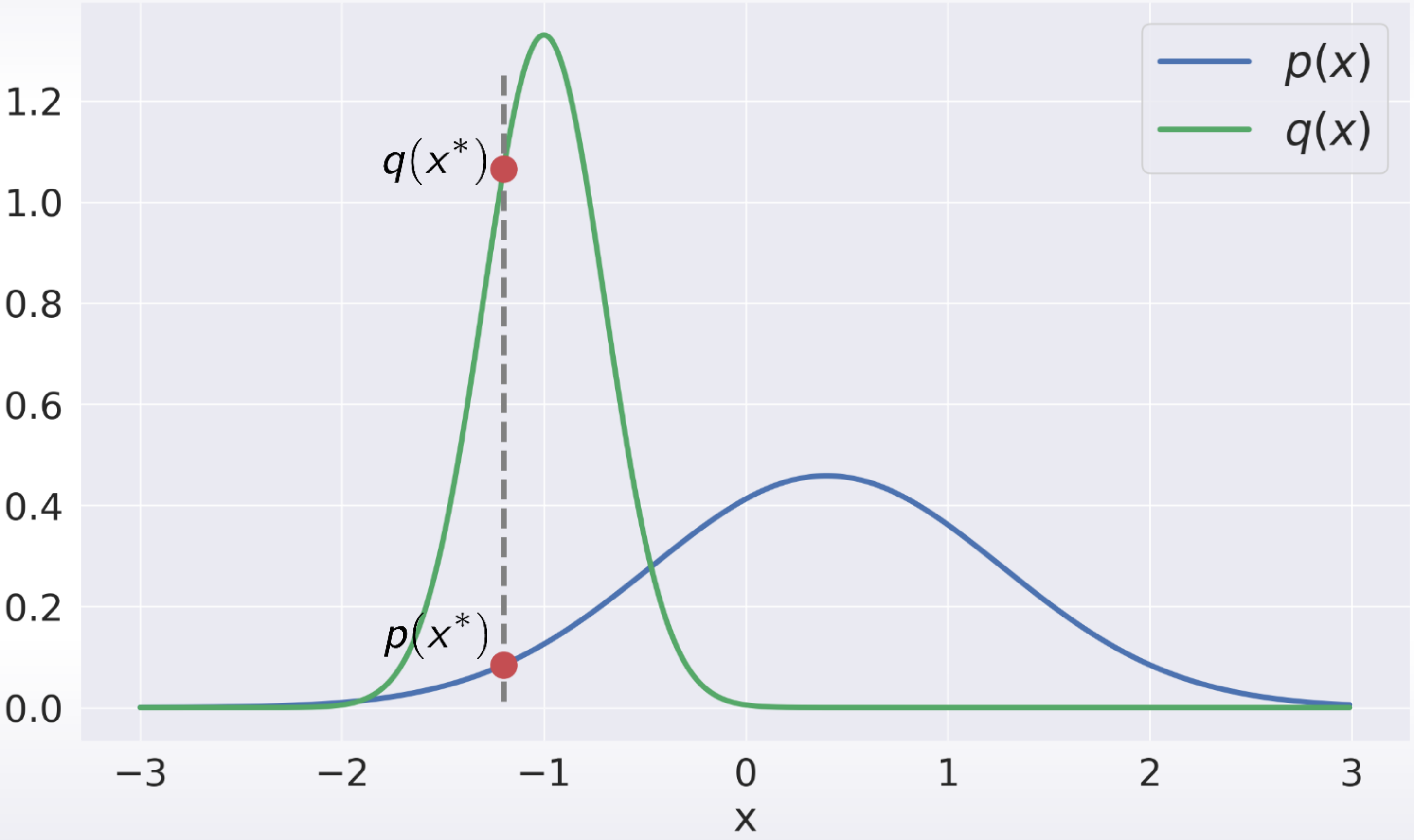
# Outline

- **Implicit generative models**
- Distribution divergences
- Learning in implicit models

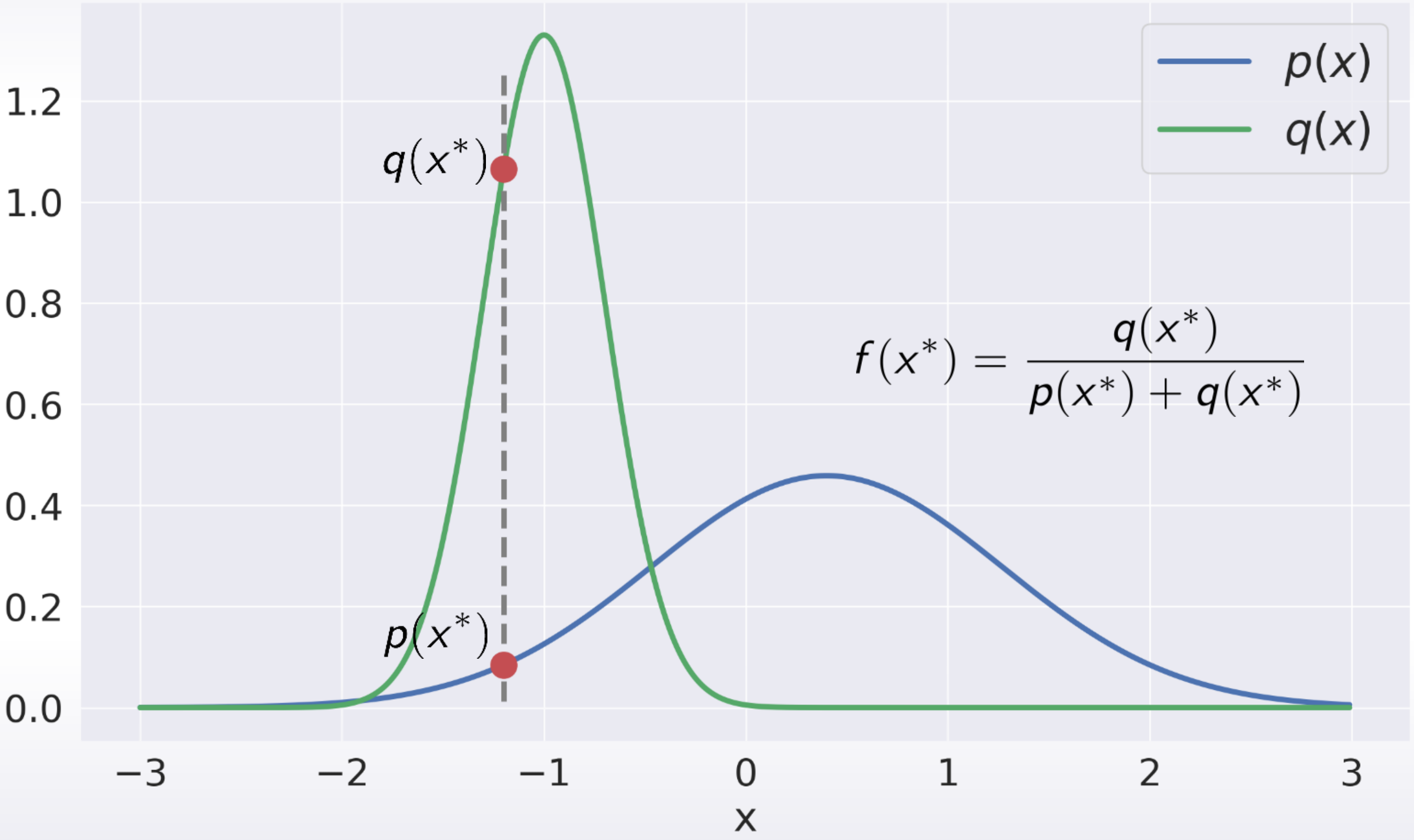
# Classification



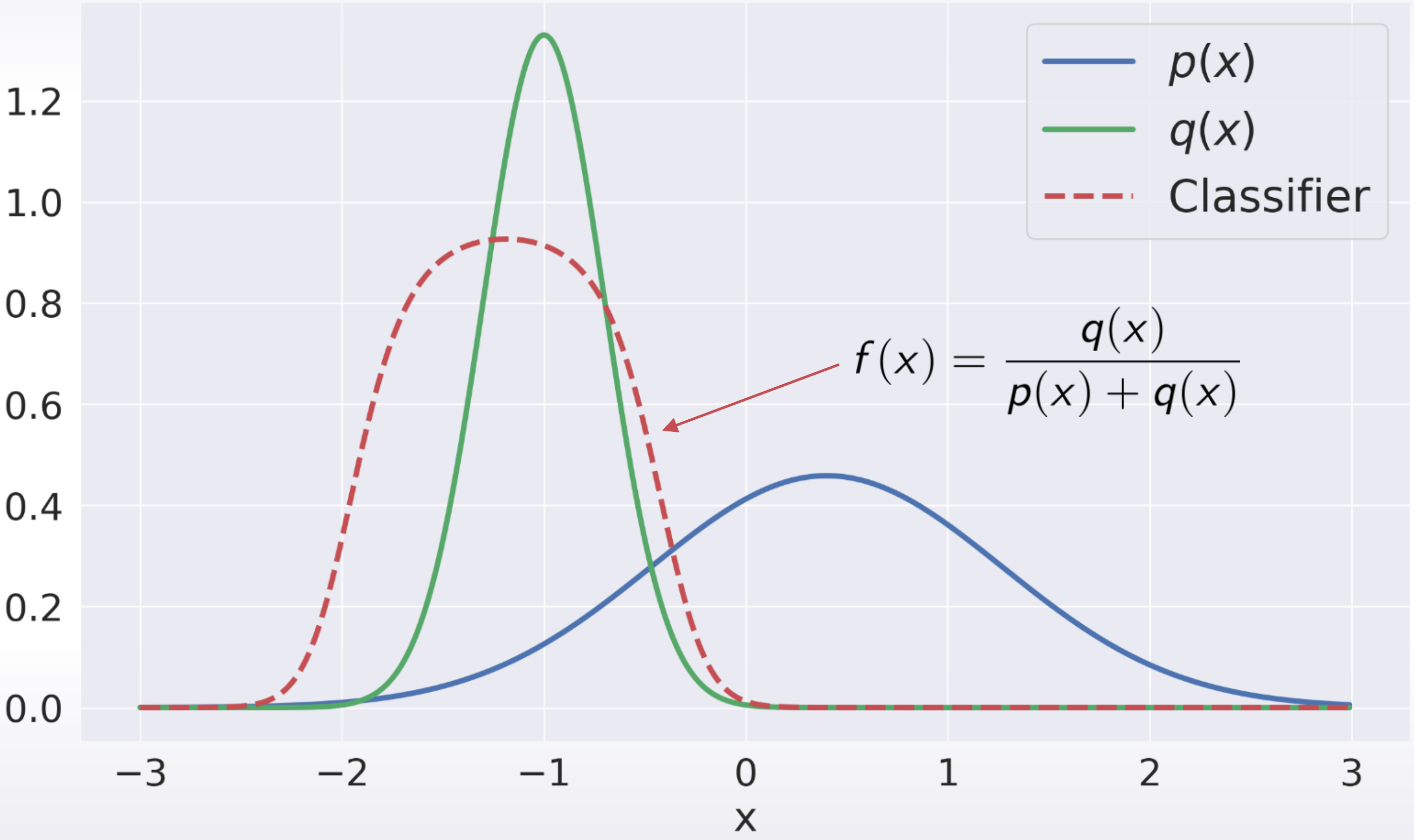
# Classification



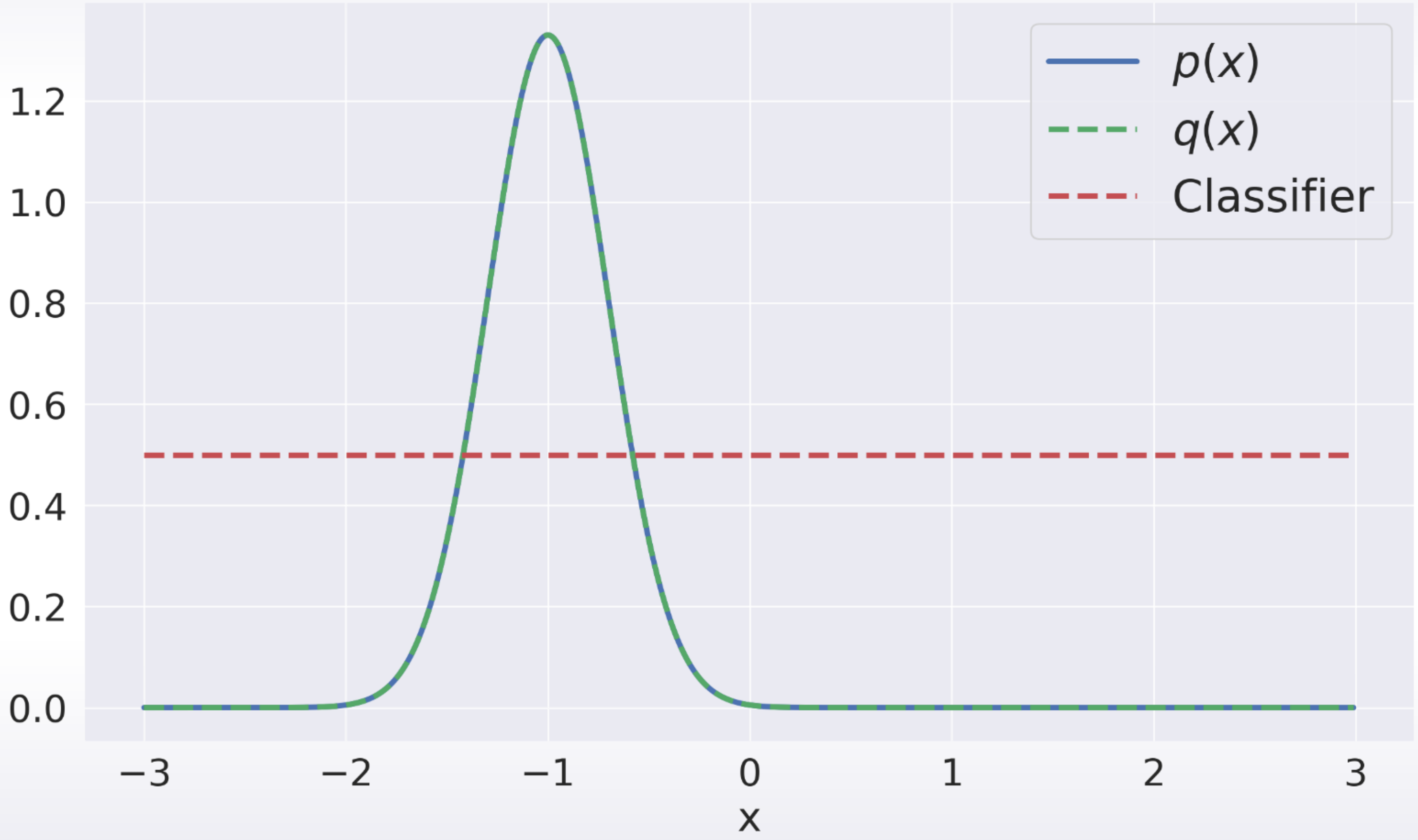
# Classification



# Classification



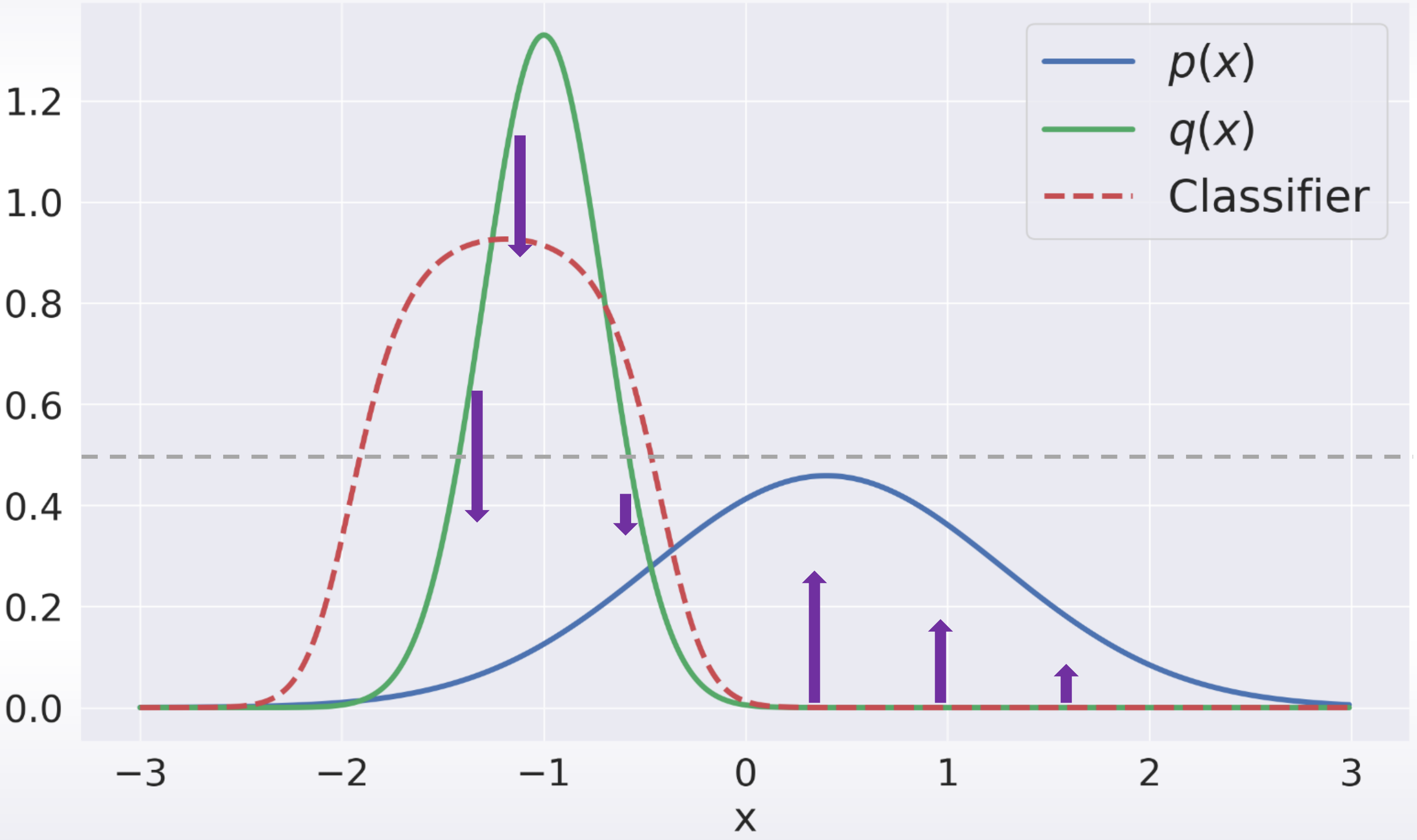
# Classification



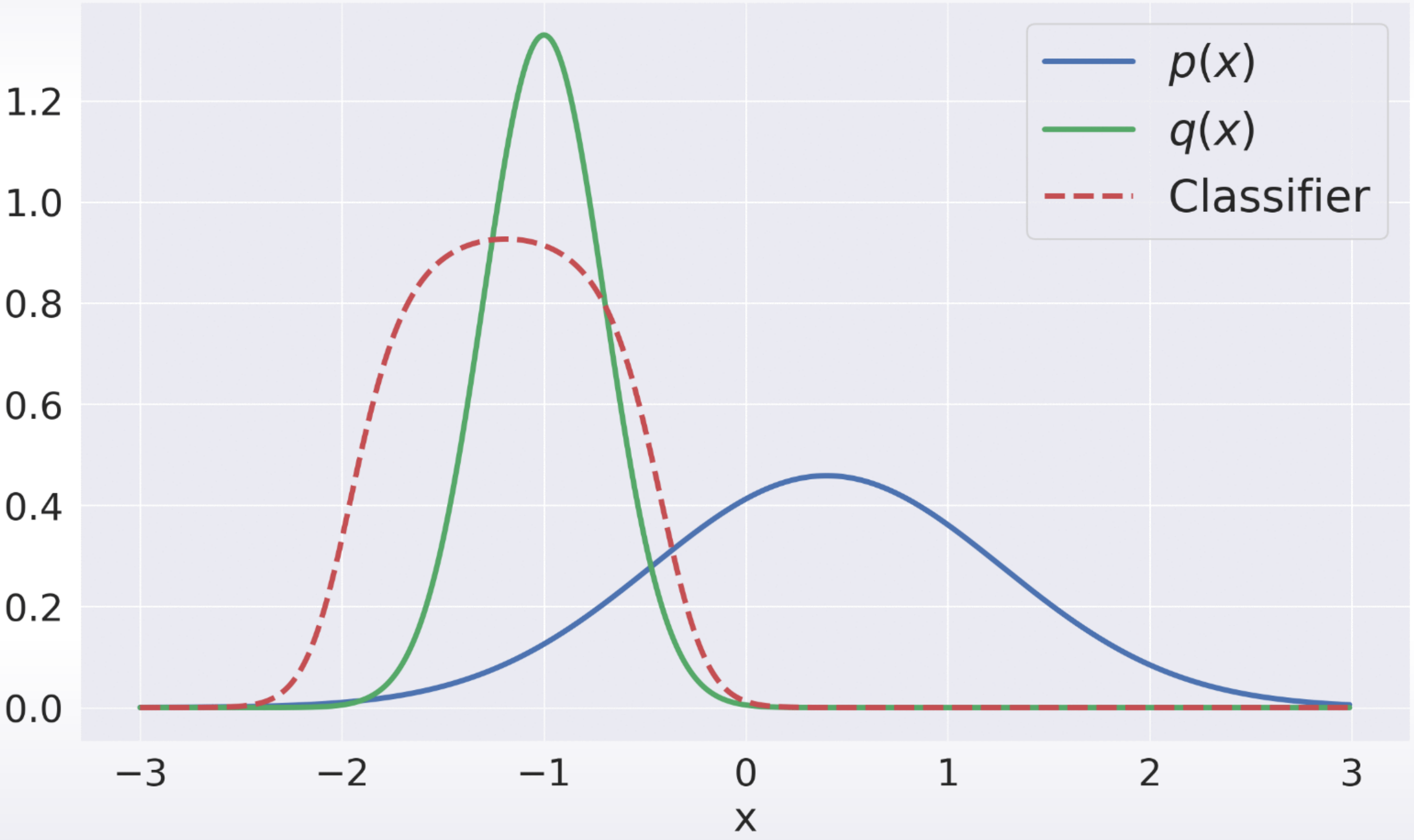
# Some intuition

- So far we had fixed $p(x)$, $q(x)$ and only trained classifier
- How do we use classifier's output to move $q(x)$ towards $p(x)$?

# Using classifier's feedback



# Using classifier's feedback



# Parametrization

1. What do we need to learn a classifier?

**Only samples from p(x) and q(x)!**

2. How do we parametrize model distribution q(x) ?

## Parametrize density function

e. g. $q_\theta(x) = \mathcal{N}(x; \theta, I)$

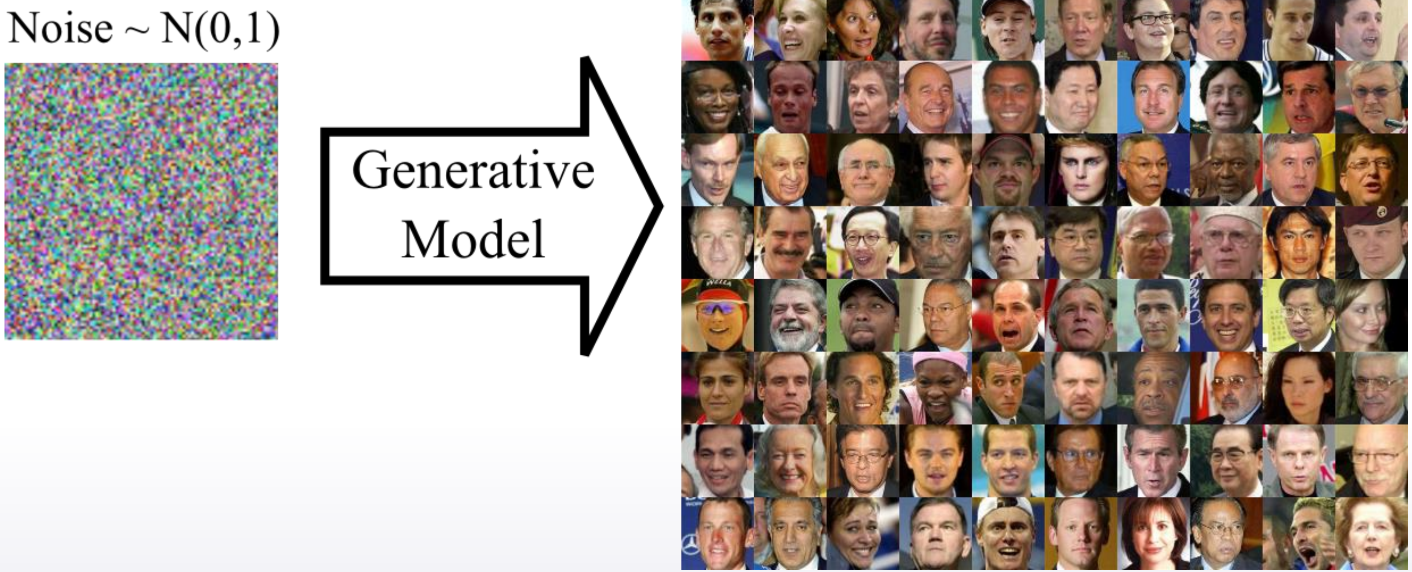
- We should be able to sample from $q_\theta$
- Have access to density at any point.

## Define implicitly

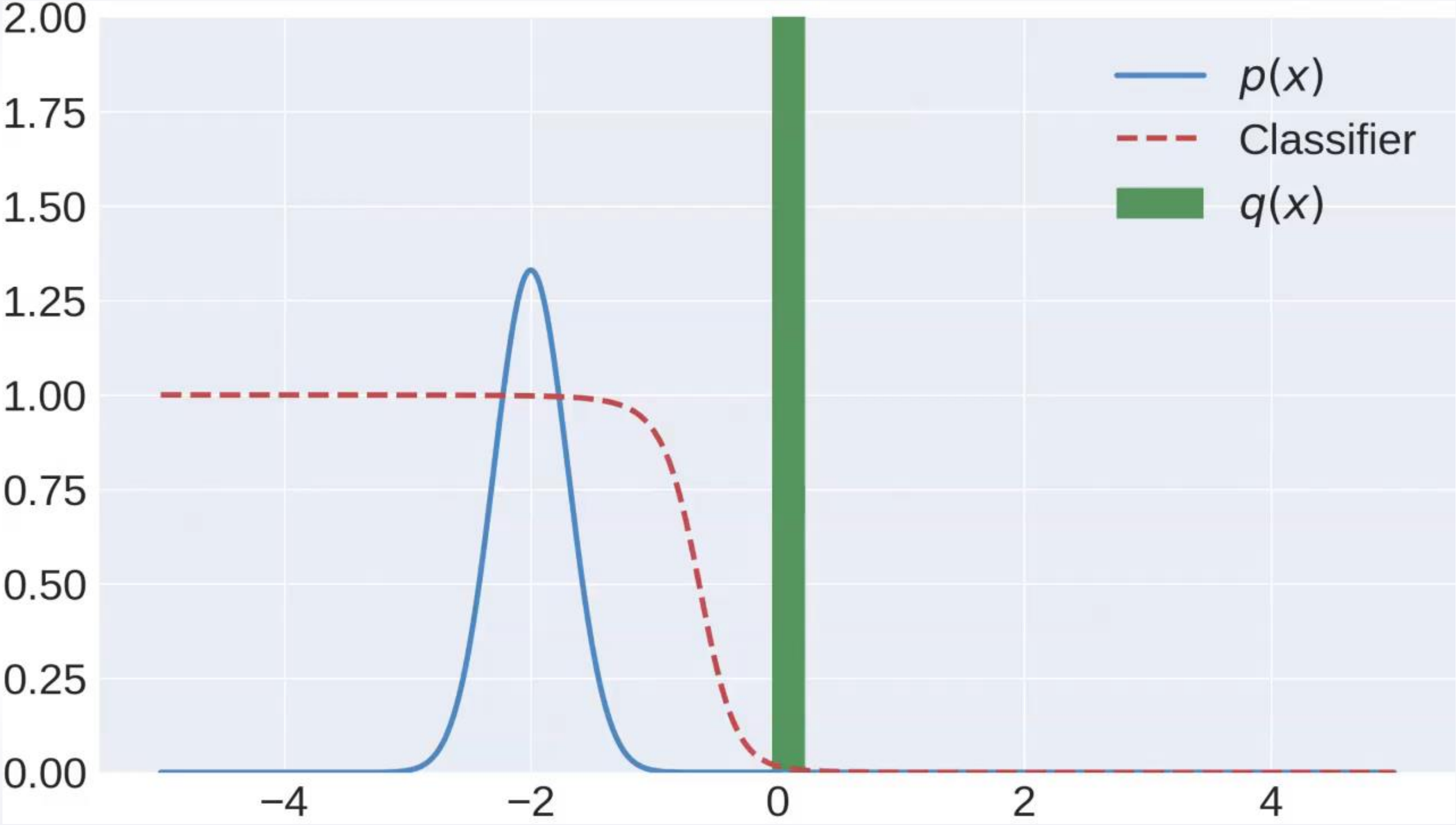
$z \sim \mathcal{N}(0, I), \quad G_\theta(z) \sim q_\theta(x)$

- Sampling is always easy
- Hard to evaluate point density $q_\theta(x)$

# In case of images



# Simulation



# GAN Game

- Classifier

$$f_\phi(\mathbf{x}) = p_\phi(y = 1|\mathbf{x})$$

- Classification loss

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p(\mathbf{x})}[-\log f_\phi(\mathbf{x})] + \mathbb{E}_{p(\mathbf{z})}[-\log(1 - f_\phi(G_\theta(\mathbf{z})))]$$

## Algorithm

1. **Update classifier**

$$\phi^* = \arg\min_\phi \mathcal{L}(\phi, \theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# In general

## General learning scheme:

1. **Update guide**
    - $\frac{p(\mathbf{x})}{p(\mathbf{x})+q(\mathbf{x})}$
    - $\frac{p(\mathbf{x})}{q(\mathbf{x})}$
    - $p(\mathbf{x}) - q(\mathbf{x})$
    - $\mathcal{D}(p \| q)$
2. Use guide to **update generator**
    - Move $q(\mathbf{x})$ closer to $p(\mathbf{x})$
3. **Repeat**

# Implicit models

## In an implicit model:

- Density function $q_\theta(\mathbf{x})$ is intractable
- But there is a way to sample from $q_\theta(\mathbf{x})$
  - Thus, we can compute expectations over $q_\theta(\mathbf{x})$
- Should be able to calculate gradients w.r.t. parameters $\theta$

**GAN – is a particular case of implicit generative models**

# Prescribed vs implicit models

## Prescribed (think of VAE)

- $p(z)$
- $q(x)$
- $p(x|z)$
- $q(z|x)$
- $q(x,z)$

**Evaluate and sample**

## Implicit (think of GAN)

- Evaluate and sample from $p(z)$
- Sample from $p(x), q(x)$
- Approximate $q(x)$ using samples
- Approximate $q(z|x)$

# GAN Game

- Classification loss

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p(\mathbf{x})}[-\log f_\phi(\mathbf{x})] + \mathbb{E}_{p(\mathbf{z})}[-\log(1 - f_\phi(G_\theta(\mathbf{z})))]$$

## Algorithm

1. **Update classifier**

$$\min_\phi \mathcal{L}(\phi, \theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# GAN Game

- Classification loss

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p(\mathbf{x})}[-\log f_\phi(\mathbf{x})] + \mathbb{E}_{p(\mathbf{z})}[-\log(1 - f_\phi(G_\theta(\mathbf{z})))]$$

## Algorithm

1. **Update classifier**

$$\max_{\phi} -\mathcal{L}(\phi, \theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# GAN Game

- Classification loss

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{p(\mathbf{x})}[-\log f_\phi(\mathbf{x})] + \mathbb{E}_{p(\mathbf{z})}[-\log(1 - f_\phi(G_\theta(\mathbf{z})))]$$

## Algorithm

1. **Update classifier**

$$\max_\phi -\mathcal{L}(\phi, \theta) = -\log(4) + 2\mathcal{D}_{JS}(p \| q_\theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# GAN Game

- Classification loss

$$\mathcal{L}(\phi,\theta) = \mathbb{E}_{p(\mathbf{x})}[-\log f_\phi(\mathbf{x})] + \mathbb{E}_{p(\mathbf{z})}[-\log(1 - f_\phi(G_\theta(\mathbf{z})))]$$

## Algorithm

1. **Update classifier**

Estimate (kind of) distance between $p(\mathbf{x})$ and $q(\mathbf{x})$

$$\max_\phi -\mathcal{L}(\phi,\theta) = -\log(4) + 2\mathcal{D}_{JS}(p\|q_\theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

Minimize the distance

# (GAN) Game

- (Classification) loss

$$\mathcal{L}(\phi, \theta) = ?$$

## Algorithm

Estimate (kind of) distance between $p(\mathbf{x})$ and $q(\mathbf{x})$

1. **Update classifier**

$$\boxed{\max_{\phi} -\mathcal{L}(\phi, \theta) = \mathcal{D}(p \| q_\theta)}$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

Minimize the distance

# Let’s discuss some divergences that allow dual formulation

# Outline

- Implicit generative models
- Distribution divergences
- Learning in implicit models

# Outline

- Implicit generative models
- **Distribution divergences**
- Learning in implicit models

# Divergences : plan

- f-Divergence
- Integral Probability Metrics
- Optimal transport

# Divergences : plan

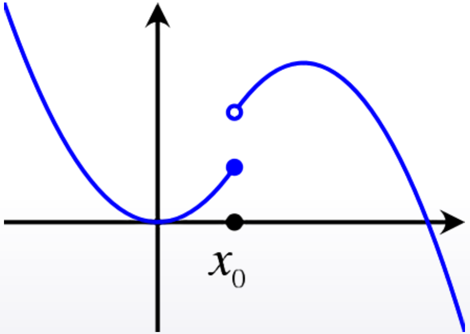
- **f-Divergence**
- Integral Probability Metrics
- Optimal transport

# f-Divergence

- For distributions $P$ and $Q$ **f-divergence** is defined as:

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

where the **generator function** $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex lower semicontinuous function satisfying $f(1) = 0$.




*Lower semi-continuous function (but non-convex)*

# f-Divergence

$$D_f\left(P \parallel Q\right) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\, \mathrm{d}x$$

- **KL-divergence:** $f(t) = t\, log(t)$

$$D_f\left(P \parallel Q\right) = KL\left(P \parallel Q\right)$$

- **Reversed KL-divergence:** $f(t) = -log(t)$

$$D_f\left(P \parallel Q\right) = KL\left(Q \parallel P\right)$$

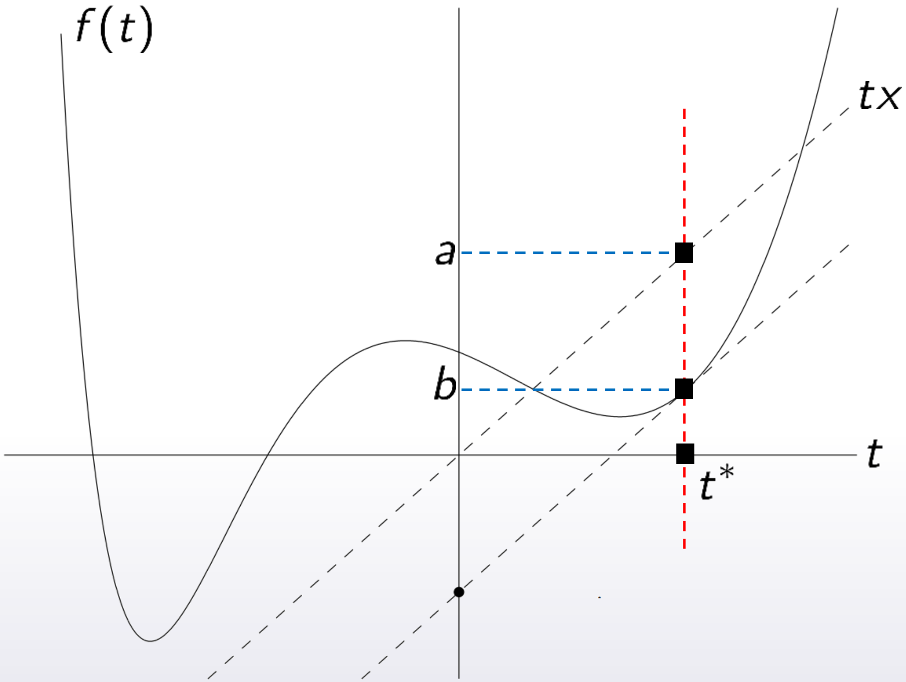
- **Total variation:** $f(t) = \frac{1}{2}|t - 1|$

$$D_f\left(P \parallel Q\right) = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)|\, \mathrm{d}x$$

# Fenchel Conjugate

- For every function $f$ we can define its **Fenchel conjugate** function $f^*$:

$$f^*(x) = \sup_{t \in \mathrm{dom} f} \{tx - f(t)\}$$




1. For a fixed $x$

   We look for the largest difference
   Between linear function and $f(t)$

2. Optimality condition for sup:

$$\left.\frac{\mathrm{d}(tx - f(t))}{\mathrm{d}t}\right|_{t=t^*} = 0$$

$$x = \frac{\mathrm{d}f(t^*)}{\mathrm{d}t}$$

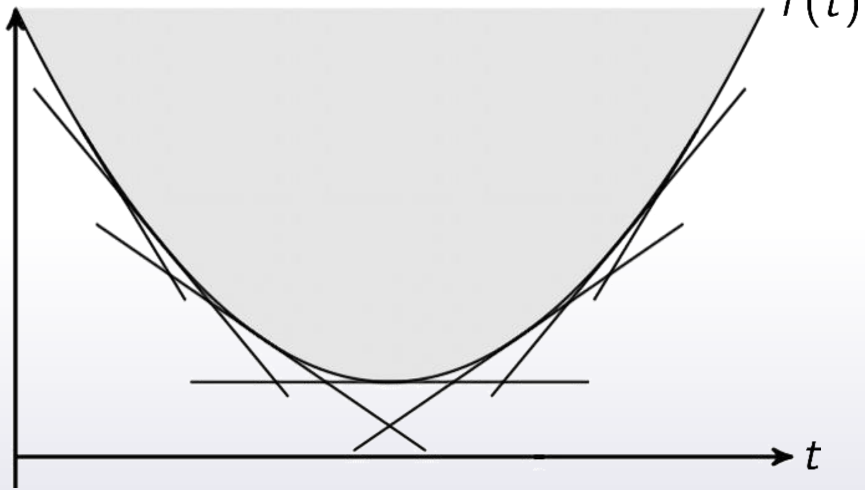
*Image from http://www.machinelearning.ru/wiki/images/2/2d/Figurnov-fenchel-slides.pdf*

# Fenchel Conjugate

- For every function $f$ we can define its **Fenchel conjugate** function $f^*$:

$$f^*(x) = \sup_{t \in \mathrm{dom} f} \{tx - f(t)\}$$

$t^* = f^*(x)$ **Tells us that a line with a slope** $x$ **supports** $f(t)$ **at** $t^*$



# Fenchel Conjugate

- For every function $f$ we can define its **Fenchel conjugate** function $f^*$:

$$f^*(x) = \sup_{t \in \mathrm{dom} f} \{tx - f(t)\}$$

- and **biconjugate**

$$f^{**}(x) = \sup_{t \in \mathrm{dom} f^*} \{tx - f^*(t)\}$$

- For convex, lower-semicontinuous functions $f$ biconjugate is equal to $f$:

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

# f-Divergence dual form

- For our $f$:

$$f(x) = \sup_{t \in \mathrm{dom} f^*} \{tx - f^*(t)\}$$

- **Derivation**:

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

- **Fact:** the bound is tight for

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

# Divergences : plan

- f-Divergence
- **Integral Probability Metrics**
- Optimal transport

# Integral Probability Metrics (IPM)

- Let $\mathcal{F}$ be any class of bounded real-valued functions.

$$IPM(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim Q} f(x) \right|$$

Basically: largest difference between statistics.

- Different choices of $\mathcal{F}$ lead to different measures:
  - **Kantorovich metric** (Wasserstein distance)

$$\mathcal{F} = \{f : \|f\|_L \leq 1\}$$

  - **Total variation distance**

$$\mathcal{F} = \{f : \|f\|_{\text{inf}} \leq 1\}$$

# f-Divergence vs IPM

- $f$**-Divergence**

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

- **IPM**

$$IPM(P, Q) = \sup_{f \in \mathcal{F}} \left|\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right|$$

# Divergences : plan

- f-Divergence
- Integral Probability Metrics
- **Optimal transport**

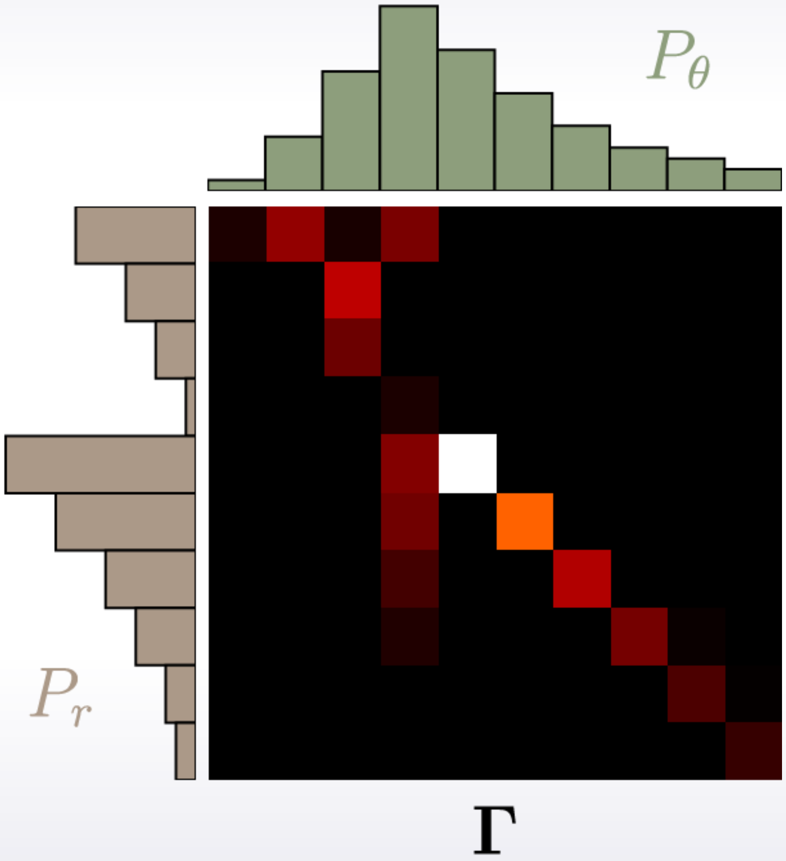
# Optimal transport


- Define a **cost** of transporting from $x$ to $y$ as $c(x, y)$
  - e.g. $c(x, y) = ||x - y||$
- **Optimal transport** cost is then defined as:

$$T(P, Q) = \inf_{\Gamma \in \mathcal{P}(x \sim P, y \sim Q)} \mathbb{E}_{(x,y) \sim \Gamma} [c(x, y)]$$

  - where $\mathcal{P}(x \sim P, y \sim Q)$ is a set of all joint distributions of $(x, y)$ with marginals $P$ and $Q$ respectively.

# Optimal transport: example



# Optimal transport dual

- **Primal**:

$$T(P,Q) = \inf_{\Gamma \in \mathcal{P}(x \sim P, y \sim Q)} \mathbb{E}_{(x,y)\sim\Gamma}\left[c(x,y)\right]$$

- **Dual (Wasserstein-1 metric)**:

$$T(P,Q) = W_1(P,Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)$$

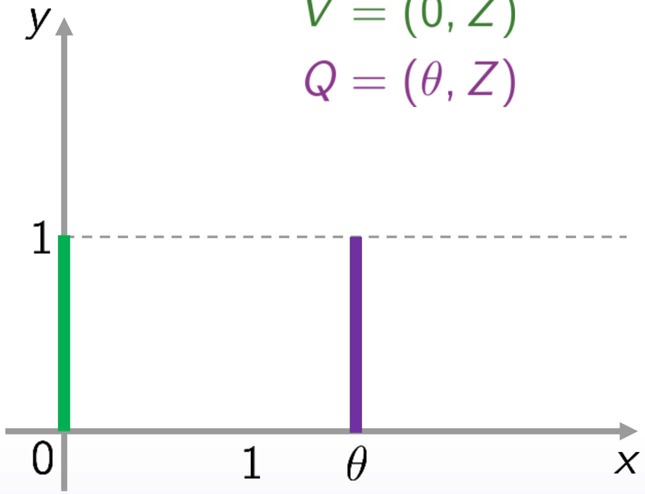
It is actually an IPM

# Optimal transport vs f-Divergence

$Z \sim U[0,1]$

$V = (0, Z)$

$Q = (\theta, Z)$




- $W_1(P, Q) = \theta$

- $JS(P \| Q) = \begin{cases} log(2), & \theta \neq 0 \\ 0, & \theta = 0 \end{cases}$

- $KL(P \| Q) = \begin{cases} \infty, & \theta \neq 0 \\ 0, & \theta = 0 \end{cases}$

# Divergences: summary

## f-Divergence

- **Primal**

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

- **Dual**

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

## Integral Probability Metric (IPM)

$$IPM(P,Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right|$$

## Optimal transport

- **Primal**

$$T(P,Q) = \inf_{\Gamma\in\mathcal{P}(x\sim P, y\sim Q)} \mathbb{E}_{(x,y)\sim\Gamma}\left[c(x,y)\right]$$

- **Dual**

$$T(P,Q) = W_1(P,Q) = \sup_{\|f\|_L\leq 1} \mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)$$

# Outline

- Implicit generative models
- Distribution divergences
- **Learning in Implicit models**

# Learning

- Loss

$$-\mathcal{L}(\phi, \theta) = \text{Dual for a divergence } \mathcal{D}(p \| q_\theta)$$

- Game

$$\min_\theta \max_\phi -\mathcal{L}(\phi, \theta) = \min_\theta \mathcal{D}(p \| q_\theta)$$

## Algorithm

1. **Update classifier**

$$\max_\phi -\mathcal{L}(\phi, \theta) = \mathcal{D}(p \| q_\theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# Example: f-divergence

- **Variational estimate**

$$D_f(P\|Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\,\mathrm{d}x$$
$$\geq \sup_{T\in\mathcal{T}} \left(\mathbb{E}_{x\sim P}\left[T(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*(T(x))\right]\right)$$

- Parametrize $T(x)$ directly

# Example: f-divergence

- Loss

$$-\mathcal{L}(\phi, \theta) = \mathbb{E}_{x \sim p(x)} \left[ T_\phi(x) \right] - \mathbb{E}_{x \sim q_\theta} \left[ f^*(T_\phi(x)) \right]$$

- Game

$$\min_\theta \max_\phi -\mathcal{L}(\phi, \theta) = \min_\theta \mathcal{D}_f(p \| q_\theta)$$

## Algorithm

1. **Update classifier**

$$\max_\phi -\mathcal{L}(\phi, \theta) = \mathcal{D}_f(p \| q_\theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*, \theta^{old})}{\partial \theta}$$

3. **Repeat**

# Another parametrization of f-divergence

- **Variational estimate**

$$
\begin{aligned}
\mathcal{D}_f(P\|Q) &= \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \\
&\geq \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x\sim P}\left[T(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*(T(x))\right]\right) \qquad (1)
\end{aligned}
$$

- **Fact:** Bound is tight for

$$
T^*(x) = f'(\frac{p(x)}{q(x)}) = f'(r^*(x)) \qquad (2)
$$

- Let's put **(2)** in **(1)**.

$$
\mathcal{D}_f(P\|Q) = \sup_{r(x)\in\mathcal{R}} \left(\mathbb{E}_{x\sim p(x)}\left[f'(r(x))\right] - \mathbb{E}_{x\sim q(x)}\left[f^*(f'(r(x)))\right]\right)
$$

# Another parametrization of f-divergence

- Loss

$$-\mathcal{L}(\phi,\theta) = \mathbb{E}_{x\sim p(x)}\left[f'(r_\phi(x))\right] - \mathbb{E}_{x\sim q_\theta(x)}\left[f^*(f'(r_\phi(x)))\right]$$

- Game

$$\min_\theta \max_\phi -\mathcal{L}(\phi,\theta) = \min_\theta \mathcal{D}_f(p\|q_\theta)$$

## Algorithm

1. **Update classifier**

$$\max_\phi -\mathcal{L}(\phi,\theta) = \mathcal{D}_f(p\|q_\theta)$$

2. **Update generator**

$$\theta^{new} = \theta^{old} + \frac{\partial \mathcal{L}(\phi^*,\theta^{old})}{\partial \theta}$$

3. **Repeat**

# Ratio matching

- Directly match $r_\phi(x)$ and $r^*(x) = \frac{p(x)}{q_\theta(x)}$

$$
\begin{aligned}
\mathcal{L}(\phi, \theta) &= \frac{1}{2}\mathbb{E}_{q_\theta(x)}(r_\phi(x) - r^*(x))^2 dx \\
&= \tfrac{1}{2}\mathbb{E}_{q_\theta(x)}[r_\phi(x)^2] - \int_x q_\theta(x)\frac{p(x)}{q_\theta(x)} r_\phi(x) dx + \underbrace{\tfrac{1}{2}\mathbb{E}_{q_\theta(x)}[r^{*2}(x)]}_{const(r_\phi)} \\
&= \tfrac{1}{2}\mathbb{E}_{q_\theta(x)}[r_\phi(x)^2] - \mathbb{E}_{p(x)}[r_\phi(x)]
\end{aligned}
$$

- **Ratio loss**

$$\min_\phi \; \mathcal{L}(\phi, \theta)$$

- **Generative loss**

$$\min_\theta -\mathcal{L}(\phi, \theta)$$

# That’s it

Thank you!

# References


- Nowozin, Cseke, Tomioka. f-GAN: *Training generative neural samplers using variational divergence minimization*, 2016
- Goodfellow et al. *Generative adversarial nets*, 2014.
- Arjovsky, Chintala, Bottou. *Wasserstein GAN*, 2017.
- Arjovsky, Bottou. *Towards Principled Methods for Training Generative Adversarial Networks*, 2017.
- Ilya Tolstikhin, *Implicit generative models: dual vs. primal approaches*" slides , 2017
- https://vincentherrmann.github.io/blog/wasserstein/
- http://www.machinelearning.ru/wiki/images/2/2d/Figurnov-fenchel-slides.pdf