Implicit generative models

Dmitry Ulyanov[1,2]

Deep | Bayes
Moscow, 2017

1 Skoltech
Skolkovo Institute of Science and Technology
2 Yandex

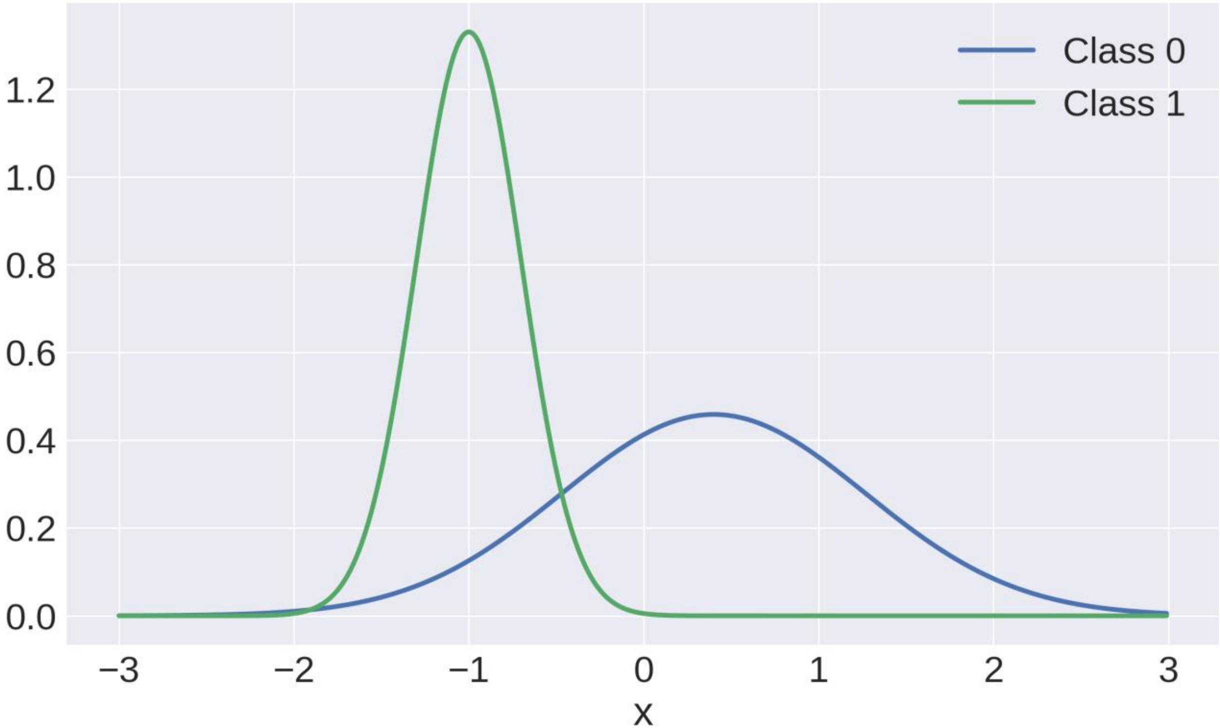
Outline

- Vanilla GAN intuition
- Distribution divergences
- Learning in implicit models
- Alpha GAN

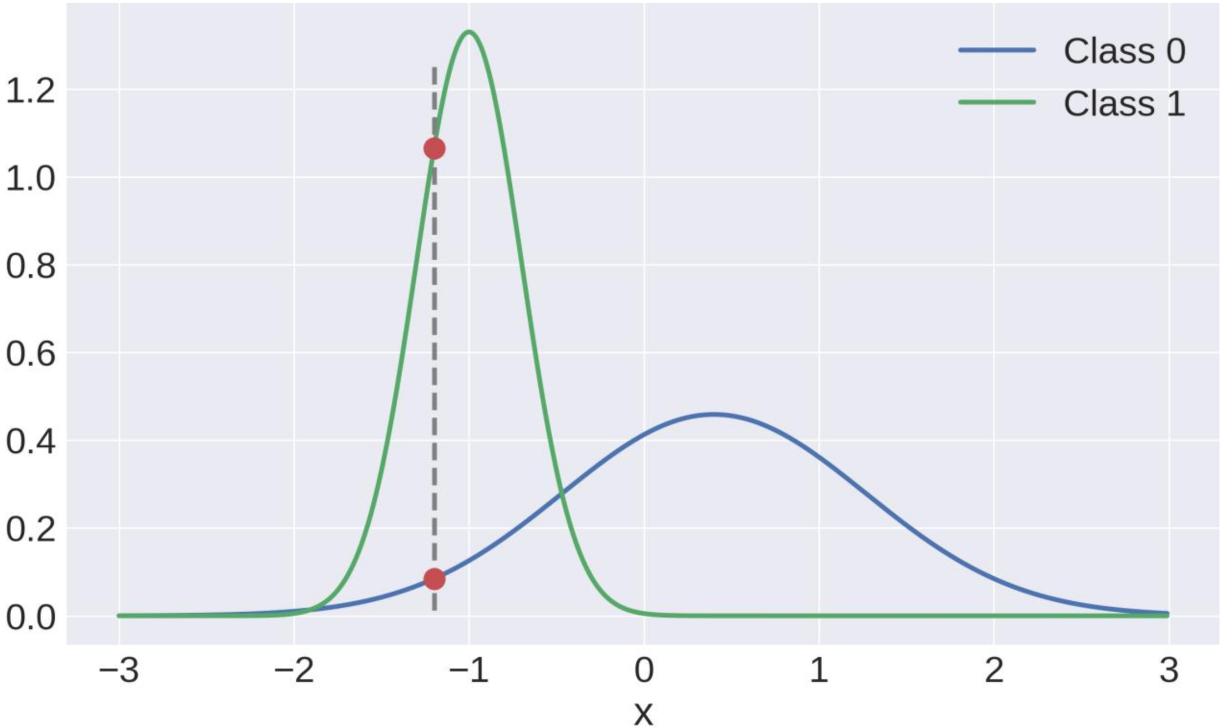
Outline

- **Vanilla GAN intuition**
- Distribution divergences
- Learning in implicit models
- Alpha GAN

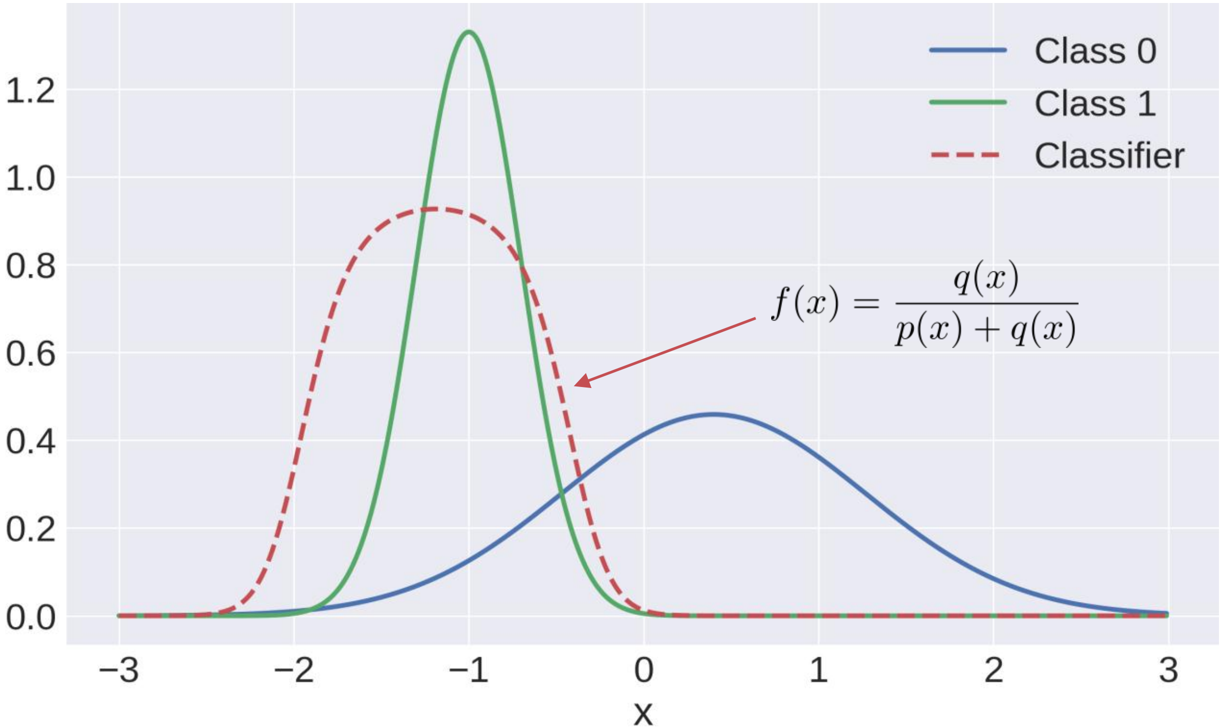
Some intuition behind implicit models



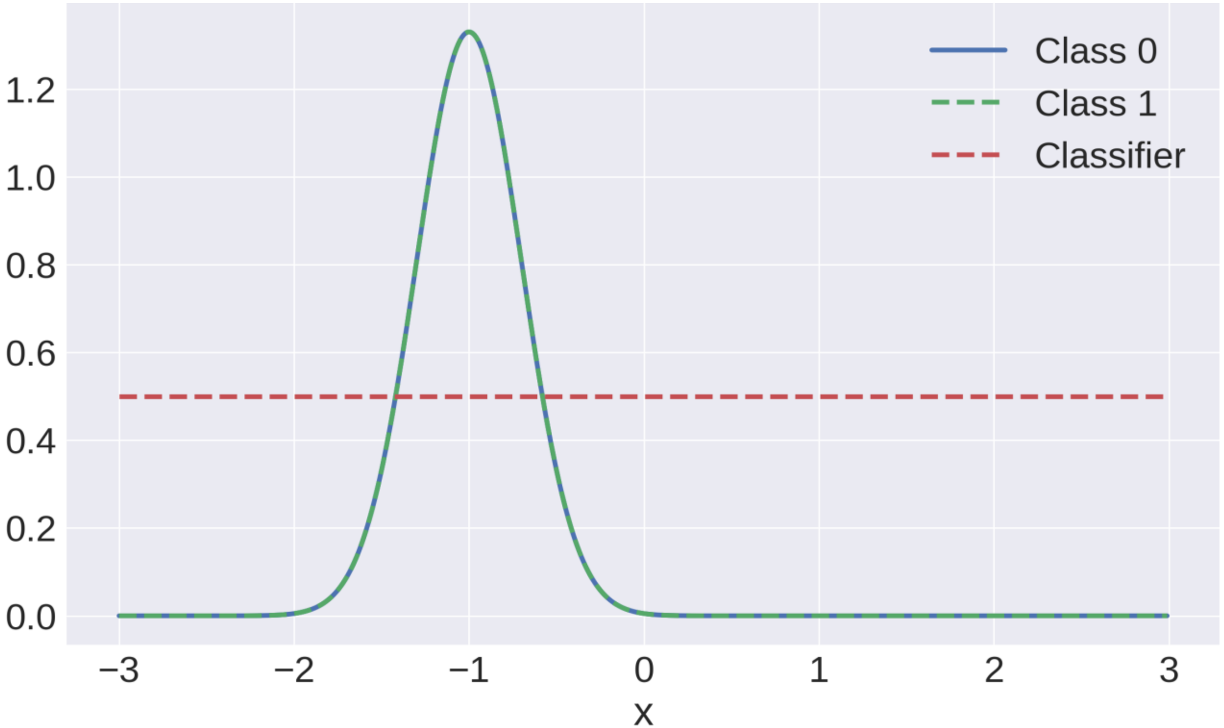
Some intuition behind implicit models



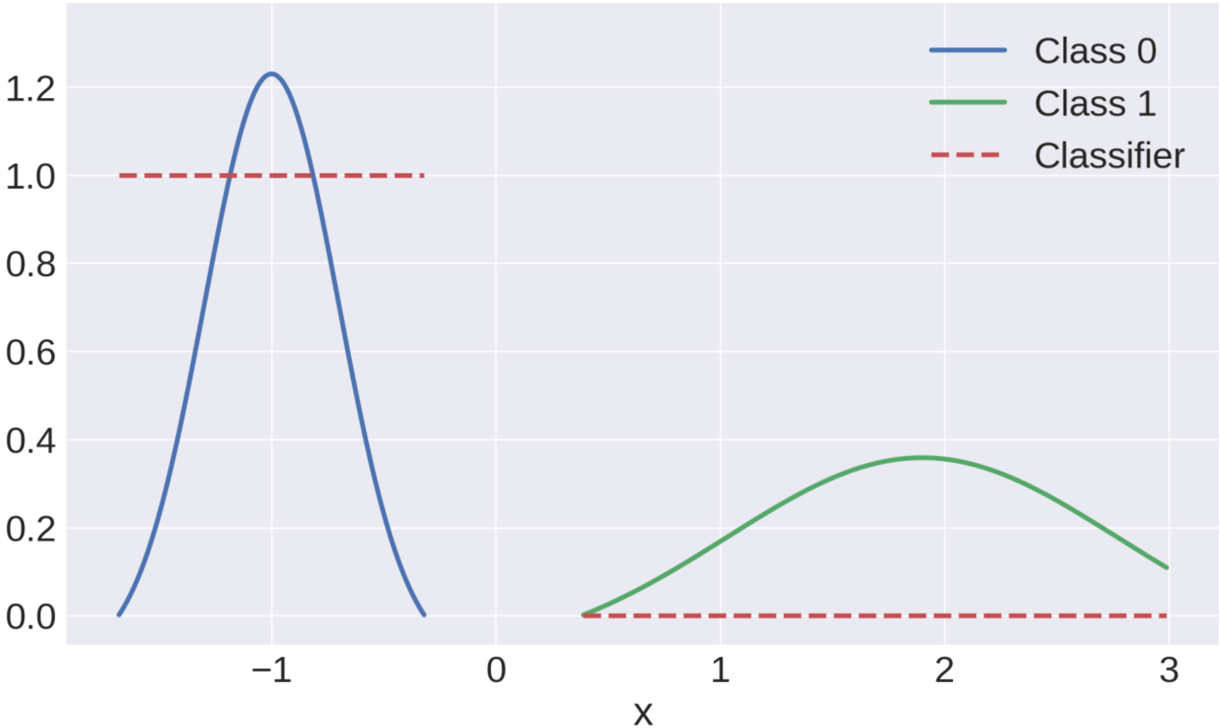
Some intuition



Some intuition



Some intuition

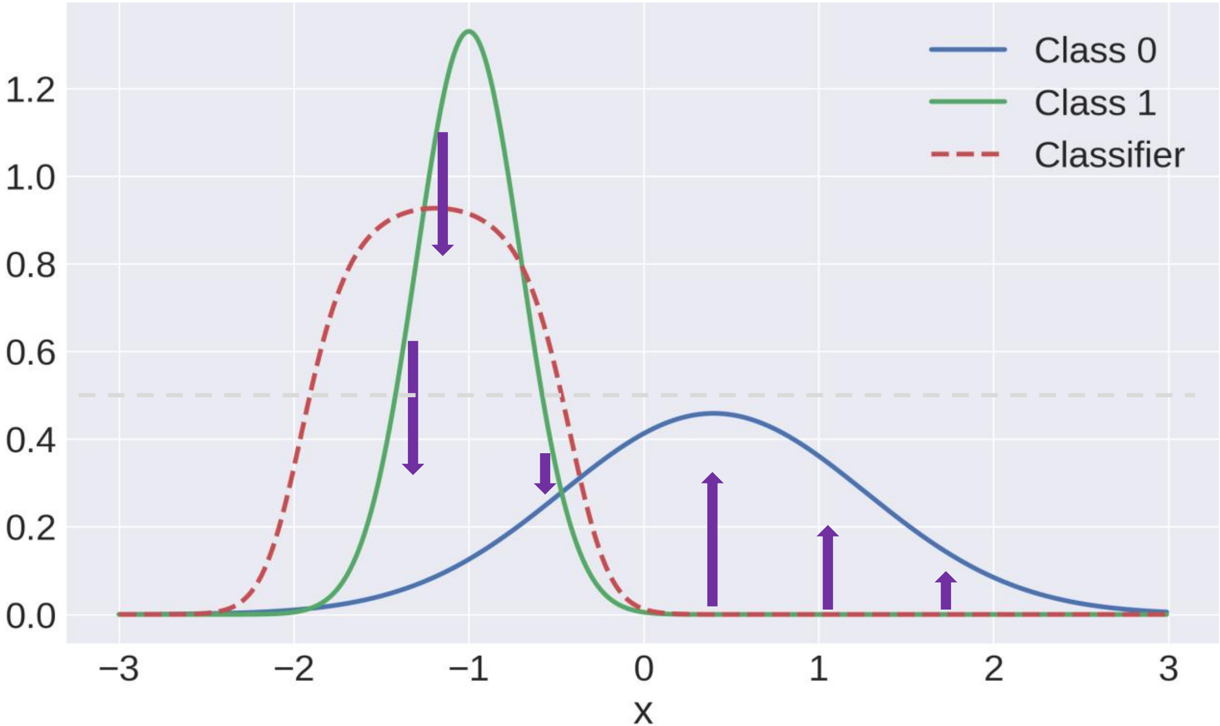


Some intuition

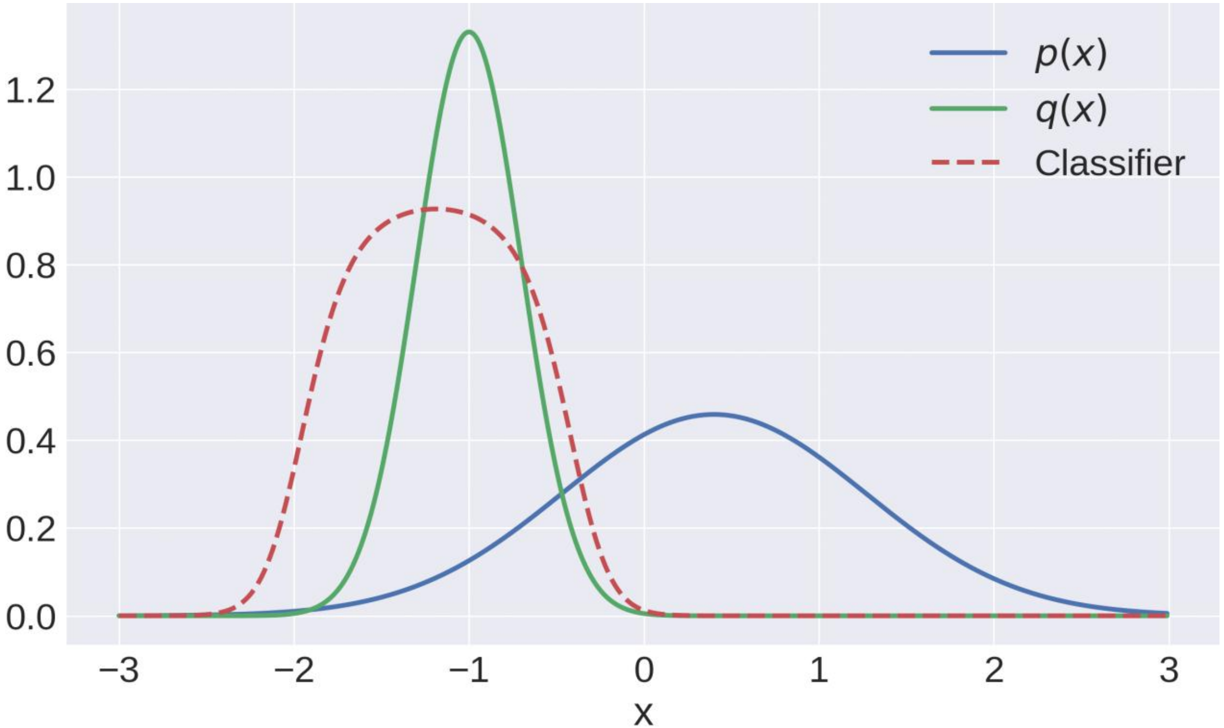
- So far we had fixed P, Q and only trained classifier.

- How do we use classifier's output to move Q towards P?

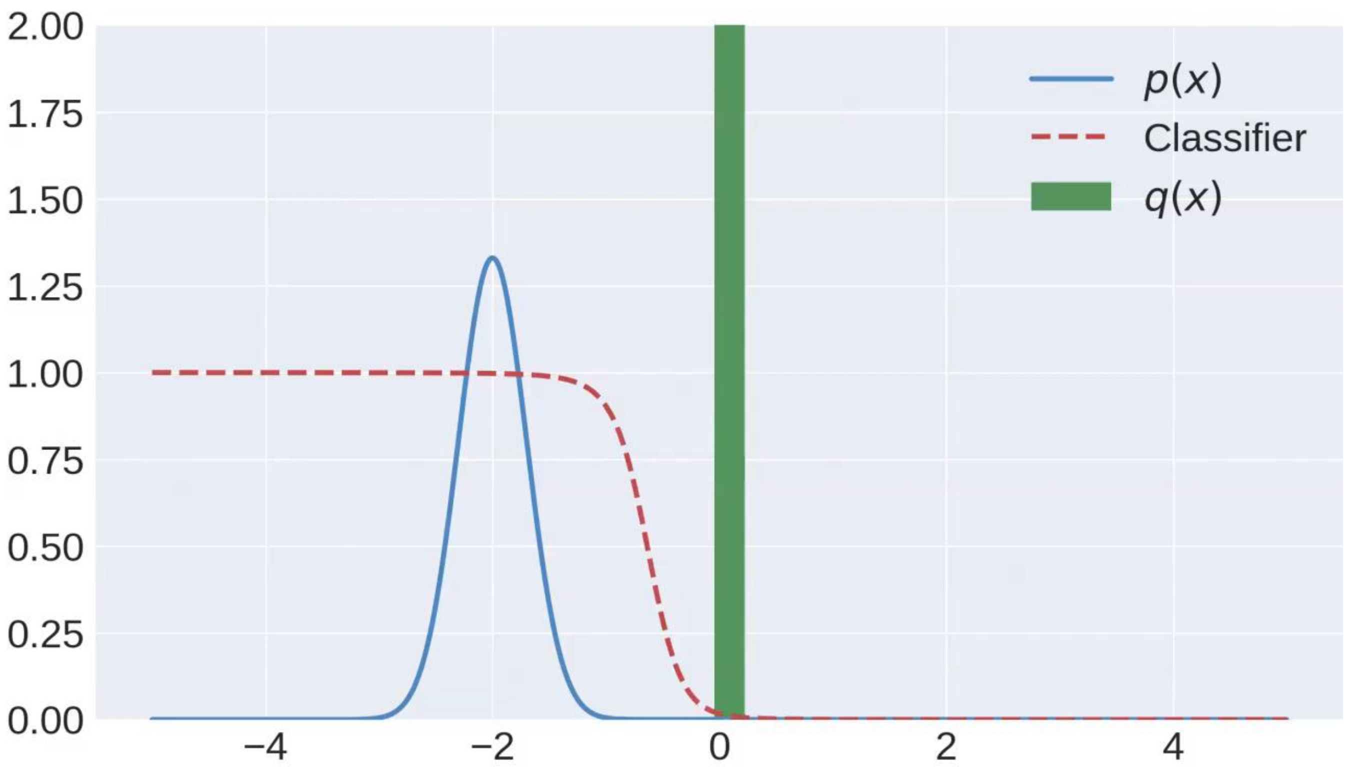
Some intuition



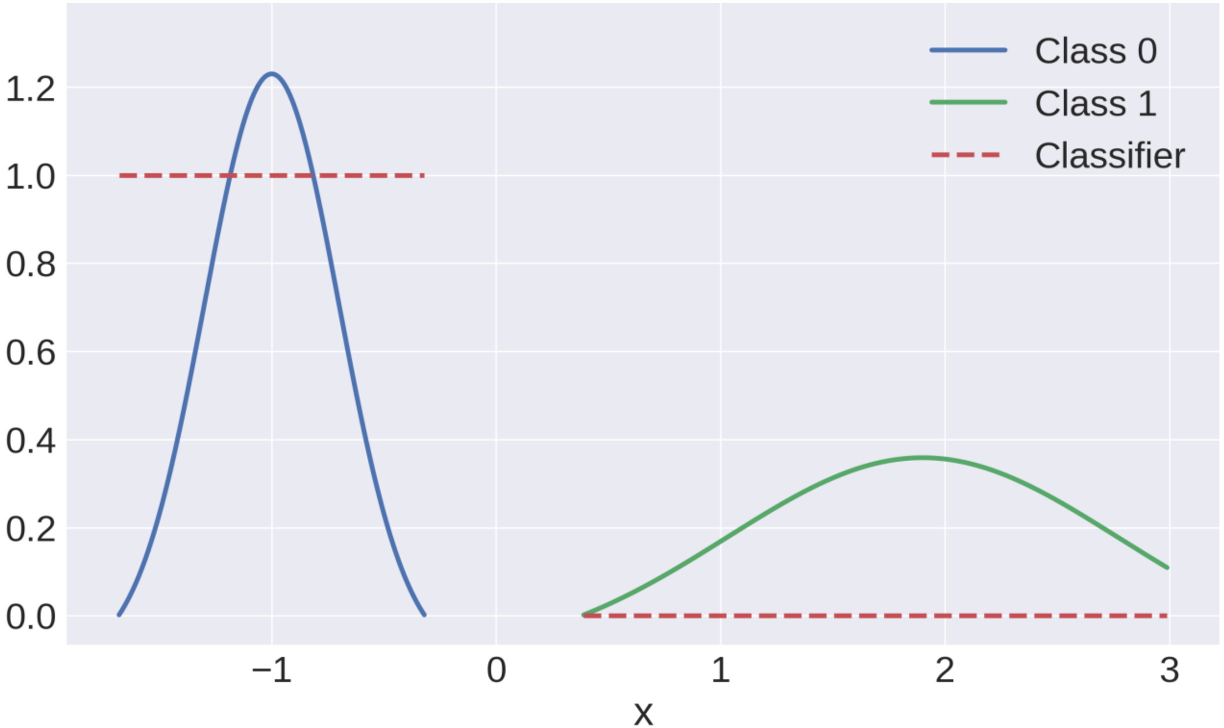
Some intuition



Simulation



A problem



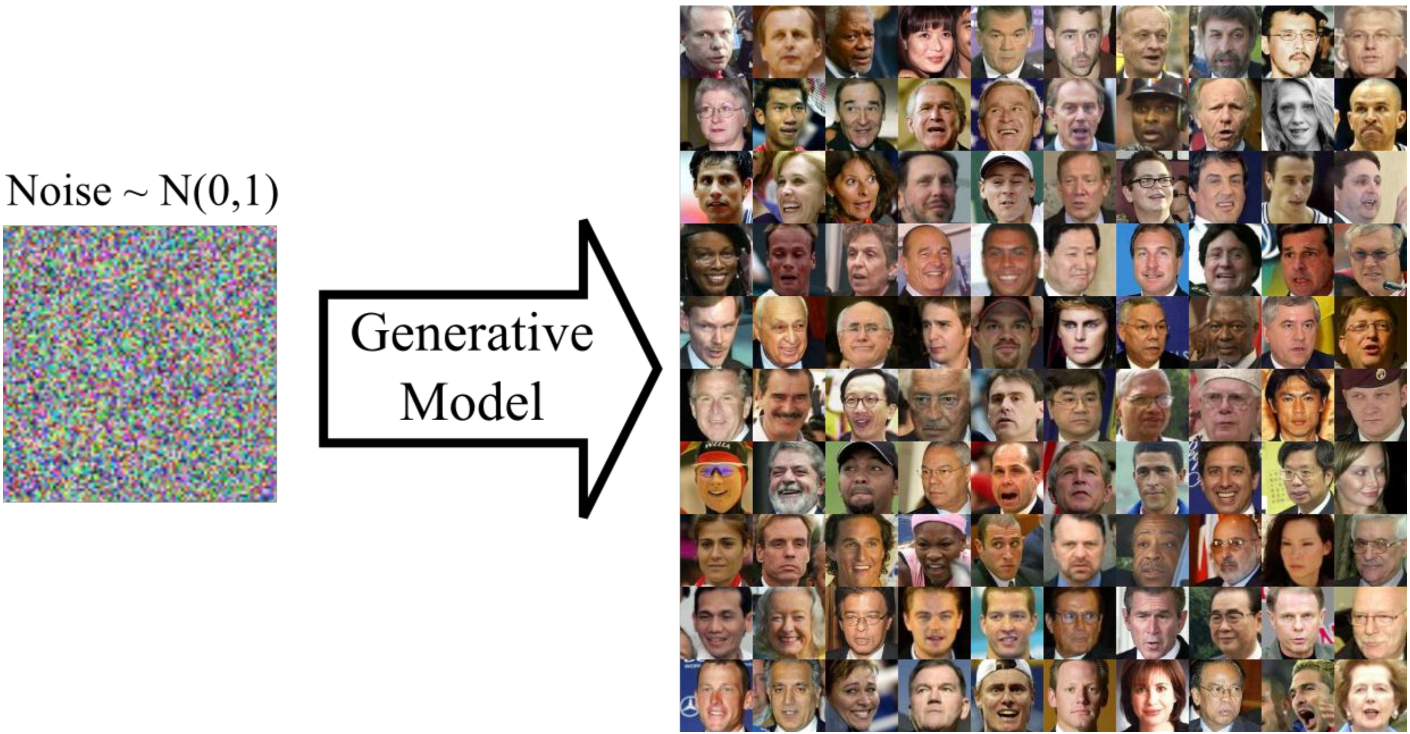
Summary

- What do we need to learn a classifier?
  - Only samples!

1. We are usually given samples from $p(\mathbf{x})$
2. How do we sample from $q(\mathbf{x})$?
   - $\mathbf{z} \sim N(0,1)$
   - $G(\mathbf{z}) \sim q(\mathbf{x})$

   - That is $G(\mathbf{z})$ implicitly defines $q(\mathbf{x})$

Some intuition



Implicit models

- Implicit models
  - Density function is intractable
  - There is a way to sample from them
    - Thus, we can compute expectations
  - Calculate gradients w.r.t. parameters
- GAN – is a particualar case of implicit generative models

Prescribed vs implicit models

Prescribed (think of VAE)

- p(**z**)
- p(**x**)
- p(**x** | **z**)
- p(**z** | **x**)
- p(**x**, **z**)

- In practice:
  - Sampling is not quite fair?

Implicit (think of GAN)

- p(**z**)
- Sample from p(**x**)

- p(**x**) ?
- p(**z** | **x**) ?

- In practice:
  - (More) fair sampling?

Outline

- Vanilla GAN intuition
- **Distribution divergences**
- Learning in implicit models
- Alpha GAN

Metrics : plan

- f-Divergence
- Integral Probability Metrics
- Optimal transport

Metrics : plan

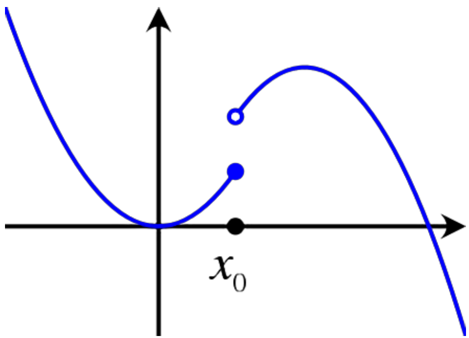
- **f-Divergence**
- Integral Probability Metrics
- Optimal transport

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 semi-continuous function satisfying $f(1) = 0$.



Lower semi-continuous function (but non-convex)

f-Divergence

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

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

$$D_f(P \| Q) = KL(P \| Q)$$

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

$$D_f(P \| Q) = KL(Q \| P)$$

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

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

Fenchel Conjugate

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

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

- and biconjugate

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

- 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:

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

- The bound is tight for

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

Metrics: plan

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

Integral Probability Metrics (IPM)

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

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

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

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

  - **Total variation distance**

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

  - **Maximum mean discrepancy (MMD)**

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

MMD

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

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathop{\mathbb{E}}_{x \sim P} f(x) - \mathop{\mathbb{E}}_{x \sim Q} f(x)|$$

- A map $\phi : \mathcal{X} \to \mathcal{H}$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. $\mathcal{H} \iff k \iff \phi$.

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

- Has closed form solution! For a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\text{MMD}_k(P, Q) = \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim P}} [k(x, y)] - 2 \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim Q}} [k(x, y)] + \mathop{\mathbb{E}}_{\substack{x \sim Q \\ y \sim Q}} [k(x, y)]$$

- If we have a map $\phi(x)$:

$$\text{MMD}_k(P, Q) = \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim P}} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} - 2 \mathop{\mathbb{E}}_{\substack{x \sim P \\ y \sim Q}} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} + \mathop{\mathbb{E}}_{\substack{x \sim Q \\ y \sim Q}} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

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

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

Metrics : plan

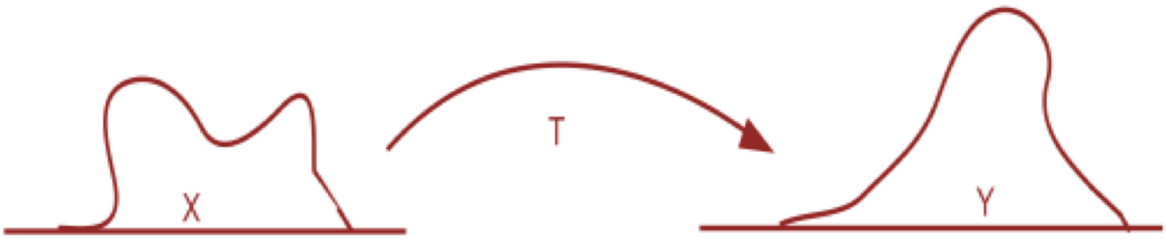
- f-Divergence
- Integral Probability Metrics
- **Optimal transport**

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

- 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:

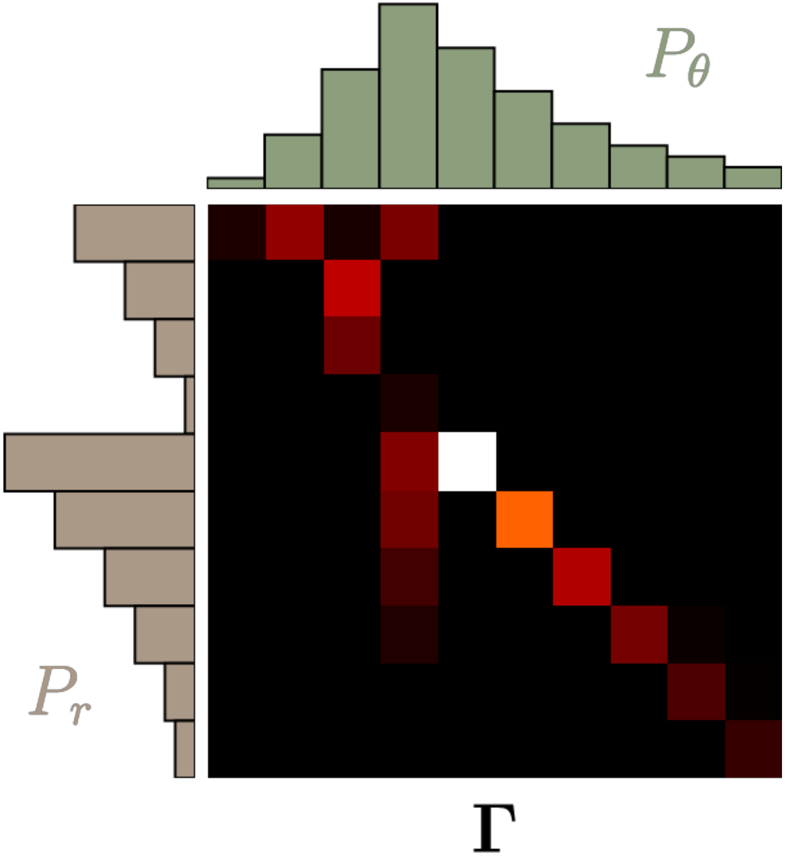
$$T(P, Q) = \inf_{\Gamma \in \mathcal{P}(x \sim P, y \sim Q)} \mathbb{E}_{(x,y) \sim \Gamma} [c(x, y)]$$

- 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.

Now a question:

- P = $\mathcal{N}(0, 1)$
- Q = $\mathcal{N}(0, 1)$
- What Γ minimizes $T(P, Q)$?

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

Let

- $Z \sim U[0, 1]$
- $P = (0, Z)$
- $Q = (\theta, Z)$

Then

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

Divergences : summary

- **f-Divergences**
  - Primal

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

  - Dual

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

- **IPMs**

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

- **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

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

Outline

- Vanilla GAN intuition
- Distribution divergences
- **Learning in Implicit models**
- Alpha GAN

Learning in implicit models in general

- We are interested in:
  - Density ratio $r(x) = \frac{p(x)}{q(x)}$
  - Density difference $r(x) = p(x) - q(x)$
- Ratio loss
  - To find $r(x)$
- Generative loss
  - Move $q(x)$ closer to $p(x)$

Learning in implicit models: plan

- Class Probability Estimation
- Divergence minimization
- Ratio matching
- Moment matching

Learning in implicit models: plan

- **Class Probability Estimation**
- Divergence minimization
- Ratio matching
- Moment matching

Class-probability matching

$$r(x) = \frac{p^*(\mathbf{x})}{q_\theta(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Big/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} = \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})}$$

- Classifier

$$\mathcal{D}(\mathbf{x};\phi) = p(y=1|\mathbf{x})$$

- Proper scoring rule:

$$\mathcal{L}(\phi, \boldsymbol{\theta}) = \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}(\mathbf{x};\phi)] + \mathbb{E}_{q_\theta(\mathbf{x})}[-\log(1-\mathcal{D}(\mathbf{x};\phi))]$$

- Ratio loss

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

- Generative loss

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

Learning in implicit models: plan

- Class Probability Estimation
- **Divergence minimization**
- Ratio matching
- Moment matching

Divergence minimization I

- **1.** Variational estimate:

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

- Let’s parametrize $T(x)$ (with a neural net) directly.

Divergence minimization I

$$D_f(P\|Q) = \sup_{T(\mathbf{x})\in\mathcal{T}} \left(\mathbb{E}_{\mathbf{x}\sim p^*(\mathbf{x})}\left[T(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim q_{\boldsymbol{\theta}}}\left[f^*(T(\mathbf{x}))\right]\right)$$

- Now we will learn a neural net to output one number, that we interpret as ratio.

$$\mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}) = T(\mathbf{x})$$

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

- Ratio loss

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

- Generative loss

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

Divergence minimization II

- **1.** Variational estimate:

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

- **2.** Bound is tight for

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

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

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

Divergence minimization II

$$D_f(P\|Q) = \sup_{r(\mathbf{x})\in\mathcal{R}} \left(\mathbb{E}_{\mathbf{x}\sim p^*(\mathbf{x})}\left[f'(r(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{x}\sim q_{\boldsymbol{\theta}}}\left[f^*(f'(r(\mathbf{x})))\right]\right)$$

- Now we will learn a neural net to output one number, that we interpret as ratio.

$$\mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}) = r_{\boldsymbol{\phi}}(\mathbf{x})$$

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

- Ratio loss

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

- Generative loss

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

Learning in implicit models: plan

- Class Probability Estimation
- Divergence minimization
- **Ratio matching**
- Moment matching

Ratio matching

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

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

- Ratio loss

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

- Generative loss

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

Learning in implicit models: plan

- Class Probability Estimation
- Divergence minimization
- Ratio matching
- **Moment matching**

Moment matching

- Can be kenelized!

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \left(\mathbb{E}_{p^*(\mathbf{x})}[\phi(\mathbf{x})] - \mathbb{E}_{q_\theta(\mathbf{x})}[\phi(\mathbf{x})]\right)^2 \\
&\approx \left(\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{M}\phi(G(\boldsymbol{z}_i))\right)^2 \\
&= \frac{1}{N^2}\sum_{i,j=1}^{N} k(\mathbf{x}_i, \mathbf{x}_j) \\
&\quad - \frac{2}{NM}\sum_{i,j=1}^{N,M} k(\mathbf{x}_i, G(\mathbf{z}_j)) \\
&\quad + \frac{1}{M^2}\sum_{i,j=1}^{M} k(G_{\boldsymbol{\theta}}(\mathbf{z}_i), G(\mathbf{z}_j))
\end{aligned}
$$

- Note, that estimator above is biased (can be easily corrected)

Outline

- Vanilla GAN intuition
- Distribution divergences
- Learning in implicit models
- **Alpha GAN**

Alpha-GAN

- Density ratio trick

$$r(x) = \frac{p^*(\mathbf{x})}{q_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \frac{\mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x})}{1-\mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x})}$$

- ELBO:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathsf{KL}[q_\eta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$$

- Synthetic likelihood

$$\mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] = \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p^*(\mathbf{x})}\right] + \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p^*(\mathbf{x})]$$

$$\approx \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}\left[\log \frac{\mathcal{D}_{\boldsymbol{\phi}}(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}))}{1-\mathcal{D}_{\boldsymbol{\phi}}(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}))}\right] + \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p^*(\mathbf{x})]$$

- Implicit Variational Distributions

$$-\mathsf{KL}[q_\eta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] = \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{z})}{q_\eta(\mathbf{z}|\mathbf{x})}\right] \approx \mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}\left[\log \frac{\mathcal{C}_{\boldsymbol{\omega}}(\mathbf{z})}{1-\mathcal{C}_{\boldsymbol{\omega}}(\mathbf{z})}\right]$$

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/*