

# Deep Generative Models

## Lecture 8

Roman Isachenko



AI Masters

2024, Spring

## Recap of previous lecture

- ▶ Standard Gaussian  $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}) \Rightarrow$  over-regularization;
- ▶  $p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i) \Rightarrow$  overfitting and highly expensive.

### ELBO revisiting

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) = \text{RL} - \text{MI} - \text{KL}(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z}|\lambda))$$

It is Forward KL with respect to  $p(\mathbf{z}|\lambda)$ .

### ELBO with flow-based VAE prior

$$\begin{aligned} \mathcal{L}(\phi, \theta) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} [\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(\mathbf{z}|\lambda) - \log q(\mathbf{z}|\mathbf{x}, \phi)] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \left[ \log p(\mathbf{x}|\mathbf{z}, \theta) + \underbrace{\left( \log p(\mathbf{f}_{\lambda}(\mathbf{z})) + \log |\det(\mathbf{J}_{\mathbf{f}})| \right)}_{\text{flow-based prior}} - \log q(\mathbf{z}|\mathbf{x}, \phi) \right] \\ \mathbf{z} &= \mathbf{f}_{\lambda}^{-1}(\mathbf{z}^*) = \mathbf{g}_{\lambda}(\mathbf{z}^*), \quad \mathbf{z}^* \sim p(\mathbf{z}^*) = \mathcal{N}(0, \mathbf{I}) \end{aligned}$$

# Recap of previous lecture

## Likelihood-free learning

- ▶ Likelihood is not a perfect quality measure for generative model.
- ▶ Likelihood could be intractable.

Imagine we have two sets of samples

- ▶  $\mathcal{S}_1 = \{\mathbf{x}_i\}_{i=1}^{n_1} \sim \pi(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_2 = \{\mathbf{x}_i\}_{i=1}^{n_2} \sim p(\mathbf{x}|\boldsymbol{\theta})$  – generated (or fake) samples.

Let define discriminative model (classifier):

$$p(y = 1|\mathbf{x}) = P(\{\mathbf{x} \sim \pi(\mathbf{x})\}); \quad p(y = 0|\mathbf{x}) = P(\{\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})\})$$

## Assumption

Generative distribution  $p(\mathbf{x}|\boldsymbol{\theta})$  equals to the true distribution  $\pi(\mathbf{x})$  if we can not distinguish them using discriminative model  $p(y|\mathbf{x})$ . It means that  $p(y = 1|\mathbf{x}) = 0.5$  for each sample  $\mathbf{x}$ .

## Recap of previous lecture

- ▶ **Generator:** generative model  $\mathbf{x} = \mathbf{G}(\mathbf{z})$ , which makes generated sample more realistic.
- ▶ **Discriminator:** a classifier  $D(\mathbf{x}) \in [0, 1]$ , which distinguishes real samples from generated samples.

### GAN optimality theorem

The minimax game

$$\min_G \max_D \underbrace{\left[ \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(\mathbf{G}(\mathbf{z}))) \right]}_{V(G,D)}$$

has the global optimum  $\pi(\mathbf{x}) = p(\mathbf{x}|\theta)$ , in this case  $D^*(\mathbf{x}) = 0.5$ .

$$\min_G V(G, D^*) = \min_G [2JSD(\pi||p) - \log 4] = -\log 4, \quad \pi(\mathbf{x}) = p(\mathbf{x}|\theta).$$

If the generator could be **any** function and the discriminator is **optimal** at every step, then the generator is **guaranteed to converge** to the data distribution.



# Outline

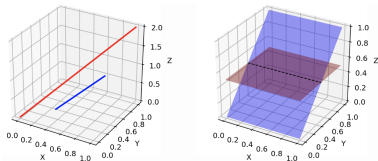
1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic
  - Wasserstein GAN
  - WGAN with Gradient Penalty
3.  $f$ -divergence minimization

# Outline

1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic
  - Wasserstein GAN
  - WGAN with Gradient Penalty
3. f-divergence minimization

# Informal theoretical results

- ▶ Since  $\mathbf{z}$  usually has lower dimensionality compared to  $\mathbf{x}$ , manifold  $\mathbf{G}_\theta(\mathbf{z})$  has a measure 0 in  $\mathbf{x}$  space. Hence, support of  $p(\mathbf{x}|\theta)$  lies on low-dimensional manifold.
- ▶ Distribution of real images  $\pi(\mathbf{x})$  is also concentrated on a low dimensional manifold.



- ▶ If  $\pi(\mathbf{x})$  and  $p(\mathbf{x}|\theta)$  have disjoint supports, then there is a smooth optimal discriminator. We are not able to learn anything by backproping through it.
- ▶ For such low-dimensional disjoint manifolds
$$KL(\pi||p) = KL(p||\pi) = \infty, \quad JSD(\pi||p) = \log 2$$
- ▶ Adding continuous noise to the inputs of the discriminator smoothes the distributions of the probability mass.

Weng L. *From GAN to WGAN*, 2019

Arjovsky M., Bottou L. *Towards Principled Methods for Training Generative Adversarial Networks*, 2017

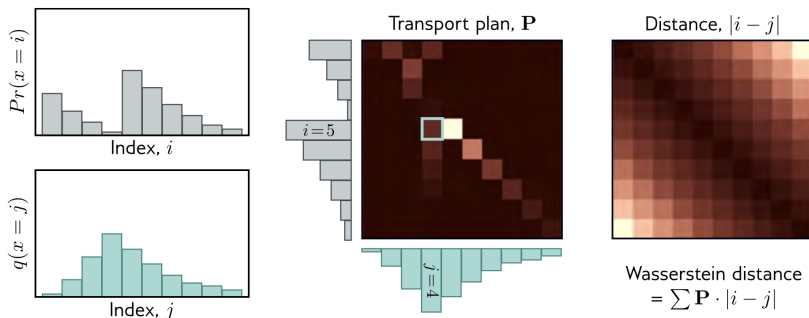


# Wasserstein distance (discrete)

A.k.a. **Earth Mover's distance**.

## Optimal transport formulation

The minimum cost of moving and transforming a pile of dirt in the shape of one probability distribution to the shape of the other distribution.



## Wasserstein distance (continuous)

$$W(\pi, p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x} - \mathbf{y}\| \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$

- ▶  $\gamma(\mathbf{x}, \mathbf{y})$  – transportation plan (the amount of "dirt" that should be transported from point  $\mathbf{x}$  to point  $\mathbf{y}$ )

$$\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} = p(\mathbf{y}); \quad \int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \pi(\mathbf{x}).$$

- ▶  $\Gamma(\pi, p)$  – the set of all joint distributions  $\gamma(\mathbf{x}, \mathbf{y})$  with marginals  $\pi$  and  $p$ .
- ▶  $\gamma(\mathbf{x}, \mathbf{y})$  – the amount,  $\|\mathbf{x} - \mathbf{y}\|$  – the distance.

## Wasserstein metric

$$W_s(\pi, p) = \inf_{\gamma \in \Gamma(\pi, p)} \left( \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|^s \right)^{1/s}$$

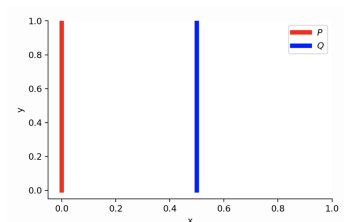
Here we will use  $W(\pi, p) = W_1(\pi, p)$  that corresponds to the optimal transport formulation.

# Wasserstein distance vs KL vs JSD

Consider 2d distributions

$$\pi(x, y) = (0, U[0, 1])$$

$$p(x, y|\theta) = (\theta, U[0, 1])$$



- $\theta = 0$ . Distributions are the same

$$KL(\pi||p) = KL(p||\pi) = JSD(p||\pi) = W(\pi, p) = 0$$

- $\theta \neq 0$

$$KL(\pi||p) = \int_{U[0,1]} 1 \log \frac{1}{0} dy = \infty = KL(p||\pi)$$

$$JSD(\pi||p) = \frac{1}{2} \left( \int_{U[0,1]} 1 \log \frac{1}{1/2} dy + \int_{U[0,1]} 1 \log \frac{1}{1/2} dy \right) = \log 2$$

$$W(\pi, p) = |\theta|$$

---

Weng L. From GAN to WGAN, 2019

Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

# Wasserstein distance vs KL vs JSD

## Theorem 1

Let  $\mathbf{G}_\theta(\mathbf{z})$  be (almost) any feedforward neural network, and  $p(\mathbf{z})$  a prior over  $\mathbf{z}$  such that  $\mathbb{E}_{p(\mathbf{z})} \|\mathbf{z}\| < \infty$ . Then therefore  $W(\pi, p)$  is continuous everywhere and differentiable almost everywhere.

## Theorem 2

Let  $\pi$  be a distribution on a compact space  $\mathcal{X}$  and  $\{p_t\}_{t=1}^\infty$  be a sequence of distributions on  $\mathcal{X}$ .

$$KL(\pi \| p_t) \rightarrow 0 \text{ (or } KL(p_t \| \pi) \rightarrow 0) \quad (1)$$

$$JSD(\pi \| p_t) \rightarrow 0 \quad (2)$$

$$W(\pi \| p_t) \rightarrow 0 \quad (3)$$

Then, considering limits as  $t \rightarrow \infty$ , (1) implies (2), (2) implies (3).

# Wasserstein GAN

## Wasserstein distance

$$W(\pi||p) = \inf_{\gamma \in \Gamma(\pi,p)} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\| = \inf_{\gamma \in \Gamma(\pi,p)} \int \|\mathbf{x} - \mathbf{y}\| \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$

The infimum across all possible joint distributions in  $\Gamma(\pi, p)$  is intractable.

## Theorem (Kantorovich-Rubinstein duality)

$$W(\pi||p) = \frac{1}{K} \max_{\|f\|_L \leq K} [\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})],$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\|f\|_L \leq K$  are  $K$ -Lipschitz continuous functions ( $f : \mathcal{X} \rightarrow \mathbb{R}$ )

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq K \|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}.$$

Now we need only samples to get Monte Carlo estimate for  $W(\pi||p)$ .

# Outline

1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic
  - Wasserstein GAN
  - WGAN with Gradient Penalty
3. f-divergence minimization

# Outline

1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic  
Wasserstein GAN  
WGAN with Gradient Penalty
3. f-divergence minimization

# Wasserstein GAN

## Theorem (Kantorovich-Rubinstein duality)

$$W(\pi||p) = \frac{1}{K} \max_{\|f\|_L \leq K} [\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})],$$

- ▶ Now we have to ensure that  $f$  is  $K$ -Lipschitz continuous.
- ▶ Let  $f_\phi(\mathbf{x})$  be a feedforward neural network parametrized by  $\phi$ .
- ▶ If parameters  $\phi$  lie in a compact set  $\Phi$  then  $f_\phi(\mathbf{x})$  will be  $K$ -Lipschitz continuous function.
- ▶ Let the parameters be clamped to a fixed box  $\Phi \in [-c, c]^d$  (e.x.  $c = 0.01$ ) after each gradient update.

$$\begin{aligned} K \cdot W(\pi||p) &= \max_{\|f\|_L \leq K} [\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})] \geq \\ &\geq \max_{\phi \in \Phi} [\mathbb{E}_{\pi(\mathbf{x})} f_\phi(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f_\phi(\mathbf{x})] \end{aligned}$$



# Wasserstein GAN

## Standard GAN objective

$$\min_{\theta} \max_{\phi} \mathbb{E}_{\pi(\mathbf{x})} \log D_{\phi}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D_{\phi}(\mathbf{G}_{\theta}(\mathbf{z})))$$

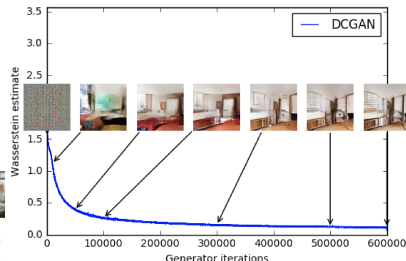
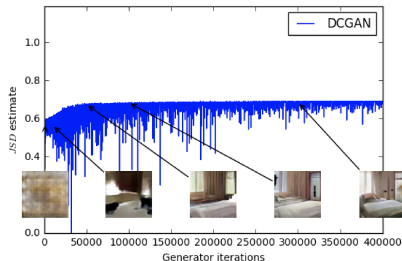
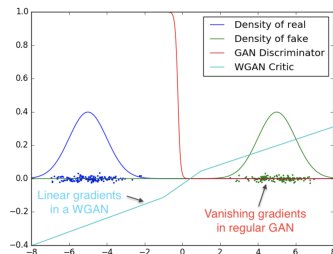
## WGAN objective

$$\min_{\theta} W(\pi || p) \approx \min_{\theta} \max_{\phi \in \Phi} [\mathbb{E}_{\pi(\mathbf{x})} f_{\phi}(\mathbf{x}) - \mathbb{E}_{p(\mathbf{z})} f_{\phi}(\mathbf{G}_{\theta}(\mathbf{z}))].$$

- ▶ Discriminator  $D$  is similar to the function  $f$ , but not the same (it is not a classifier anymore). In the WGAN model, function  $f$  is usually called **critic**.
- ▶ "*Weight clipping is a clearly terrible way to enforce a Lipschitz constraint*". If the clipping parameter  $c$  is too large, it is hard to train the critic till optimality. If the clipping parameter  $c$  is too small, it could lead to vanishing gradients.

# Wasserstein GAN

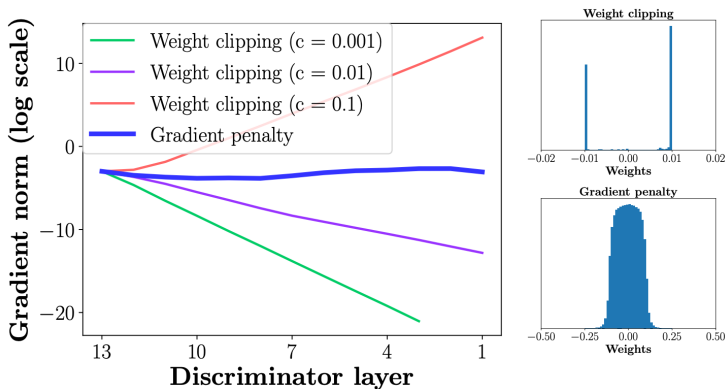
- ▶ WGAN has non-zero gradients for disjoint supports.
- ▶  $JSD(\pi||p)$  correlates poorly with the sample quality. Stays constant nearly maximum value  $\log 2 \approx 0.69$ .
- ▶  $W(\pi||p)$  is highly correlated with the sample quality.



# Outline

1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic  
Wasserstein GAN  
WGAN with Gradient Penalty
3. f-divergence minimization

# Wasserstein GAN with Gradient Penalty



## Weight clipping analysis

- ▶ The gradients either grow or decay exponentially.
- ▶ Gradient penalty makes the gradients more stable.

# Wasserstein GAN with Gradient Penalty

## Theorem

Let  $\pi(\mathbf{x})$  and  $p(\mathbf{x})$  be two distributions in  $\mathcal{X}$ , a compact metric space. Let  $\gamma$  be the optimal transportation plan between  $\pi(\mathbf{x})$  and  $p(\mathbf{x})$ . Then

1. there is 1-Lipschitz function  $f^*$  which is the optimal solution of

$$\max_{\|f\|_L \leq 1} \left[ \mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right].$$

2. if  $f^*$  is differentiable,  $\gamma(\mathbf{y} = \mathbf{z}) = 0$  and  $\hat{\mathbf{x}}_t = t\mathbf{y} + (1-t)\mathbf{z}$  with  $\mathbf{y} \sim \pi(\mathbf{x})$ ,  $\mathbf{z} \sim p(\mathbf{x}|\theta)$ ,  $t \in [0, 1]$  it holds that

$$\mathbb{P}_{(\mathbf{y}, \mathbf{z}) \sim \gamma} \left[ \nabla f^*(\hat{\mathbf{x}}_t) = \frac{\mathbf{z} - \hat{\mathbf{x}}_t}{\|\mathbf{z} - \hat{\mathbf{x}}_t\|} \right] = 1.$$

## Corollary

$f^*$  has gradient norm 1 almost everywhere under  $\pi(\mathbf{x})$  and  $p(\mathbf{x})$ .

# Wasserstein GAN with Gradient Penalty

A differentiable function is 1-Lipschitz if and only if it has gradients with norm at most 1 everywhere.

## Gradient penalty

$$W(\pi||p) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})}f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})}f(\mathbf{x})}_{\text{original critic loss}} + \lambda \underbrace{\mathbb{E}_{U[0,1]} \left[ (\|\nabla f(\hat{\mathbf{x}})\|_2 - 1)^2 \right]}_{\text{gradient penalty}},$$

- ▶ Samples  $\hat{\mathbf{x}}_t = t \cdot \mathbf{y} + (1 - t) \cdot \mathbf{z}$  with  $t \in [0, 1]$  are uniformly sampled along straight lines between pairs of points:  $\mathbf{y}$  from the data distribution  $\pi(\mathbf{x})$  and  $\mathbf{z}$  from the generator distribution  $p(\mathbf{x}|\theta)$ .
- ▶ Enforcing the unit gradient norm constraint everywhere is intractable, it turns out to be sufficient to enforce it only along these straight lines.

# Outline

1. Wasserstein distance
2. Lipschitzness of Wasserstein GAN critic
  - Wasserstein GAN
  - WGAN with Gradient Penalty
3. f-divergence minimization

# Divergences

- ▶ Forward KL divergence in maximum likelihood estimation.
- ▶ Reverse KL in variational inference.
- ▶ JS divergence in standard GAN.
- ▶ Wasserstein distance in WGAN.

## What is a divergence?

Let  $\mathcal{P}$  be the set of all possible probability distributions. Then  $D : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  is a divergence if

- ▶  $D(\pi||p) \geq 0$  for all  $\pi, p \in \mathcal{P}$ ;
- ▶  $D(\pi||p) = 0$  if and only if  $\pi \equiv p$ .

## General divergence minimization task

$$\min_p D(\pi||p)$$

## Challenge

We do not know the real distribution  $\pi(\mathbf{x})$ !



# f-divergence family

## f-divergence

$$D_f(\pi||p) = \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}.$$

Here  $f : \mathbb{R}_+ \rightarrow \mathbb{R}$  is a convex, lower semicontinuous function satisfying  $f(1) = 0$ .

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 $\chi^2$	$\int \frac{(q(x)-p(x))^2}{p(x)} dx$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$	$(\sqrt{u}-1)^2$
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	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx - \log(4)$	$u \log u - (u+1) \log(u+1)$

Nowozin S., Cseke B., Tomioka R. *f*-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

# f-divergence family

## Fenchel conjugate

$$f^*(t) = \sup_{u \in \text{dom}_f} (ut - f(u)), \quad f(u) = \sup_{t \in \text{dom}_{f^*}} (ut - f^*(t))$$

**Important property:**  $f^{**} = f$  for convex  $f$ .

## f-divergence

$$\begin{aligned} D_f(\pi || p) &= \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} = \\ &= \int p(\mathbf{x}) \sup_{t \in \text{dom}_{f^*}} \left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})} t - f^*(t)\right) d\mathbf{x} = \\ &= \int \sup_{t \in \text{dom}_{f^*}} (\pi(\mathbf{x}) t - p(\mathbf{x}) f^*(t)) d\mathbf{x}. \end{aligned}$$

Here we seek value of  $t$ , which gives us maximum value of  $\pi(\mathbf{x})t - p(\mathbf{x})f^*(t)$ , for each data point  $\mathbf{x}$ .

---

Nowozin S., Cseke B., Tomioka R. *f*-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

# f-divergence family

## f-divergence

$$D_f(\pi||p) = \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}.$$

## Variational f-divergence estimation

$$\begin{aligned} D_f(\pi||p) &= \int \sup_{t \in \text{dom}_{f^*}} (\pi(\mathbf{x})t - p(\mathbf{x})f^*(t)) d\mathbf{x} \geq \\ &\geq \sup_{T \in \mathcal{T}} \int (\pi(\mathbf{x})T(\mathbf{x}) - p(\mathbf{x})f^*(T(\mathbf{x}))) d\mathbf{x} = \\ &= \sup_{T \in \mathcal{T}} [\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))] \end{aligned}$$

This is a lower bound because of Jensen inequality and restricted class of functions  $\mathcal{T} : \mathcal{X} \rightarrow \mathbb{R}$ .

# f-divergence family

## Variational divergence estimation

$$D_f(\pi || p) \geq \sup_{T \in \mathcal{T}} [\mathbb{E}_{\pi} T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))]$$

The lower bound is tight for  $T^*(\mathbf{x}) = f' \left( \frac{\pi(\mathbf{x})}{p(\mathbf{x})} \right)$ .

### Example (JSD)

- ▶ Let define function  $f$  and its conjugate  $f^*$

$$f(u) = u \log u - (u + 1) \log(u + 1), \quad f^*(t) = -\log(1 - e^t).$$

- ▶ Let reparametrize  $T(\mathbf{x}) = \log D(\mathbf{x})$ .

$$\min_G \max_D [\mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(\mathbf{G}(\mathbf{z})))]$$

# f-divergence family

## Variational divergence estimation

$$D_f(\pi||p) \geq \sup_{T \in \mathcal{T}} [\mathbb{E}_{\pi} T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))]$$

**Note:** To evaluate lower bound we only need samples from  $\pi(\mathbf{x})$  and  $p(\mathbf{x})$ . Hence, we could fit implicit generative model.



Nowozin S., Cseke B., Tomioka R. *f*-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

## Summary

- ▶ KL and JS divergences work poorly as model objective in the case of disjoint supports.
- ▶ Earth-Mover distance is a more appropriate objective function for distribution matching problem.
- ▶ Kantorovich-Rubinstein duality gives the way to calculate the EM distance using only samples.
- ▶ Wasserstein GAN uses Kantorovich-Rubinstein duality for getting Earth Mover distance as model objective.
- ▶ Weight clipping is a terrible way to enforce Lipschitzness. Gradient Penalty adds regularizer to loss that uses necessary conditions for optimal critic.
- ▶ f-divergence family is a unified framework for divergence minimization, which uses variational approximation. Standard GAN is a special case of it.