

# Deep Generative Models

## Lecture 10

Roman Isachenko



Autumn, 2022

## Recap of previous lecture

### ELBO objective

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(z|x, \phi)} [\log p(x|z, \theta) - KL(\log q(z|x, \phi) || p(z))] \rightarrow \max_{\phi, \theta} .$$

What is the problem to make the variational posterior model an **implicit** model?

We have to estimate density ratio

$$r(x, z) = \frac{q_1(x, z)}{q_2(x, z)} = \frac{q(z|x, \phi)\pi(x)}{p(z)\pi(x)}.$$

## Adversarial Variational Bayes

$$\max_D \left[ \mathbb{E}_{\pi(x)} \mathbb{E}_{q(z|x, \phi)} \log D(x, z) + \mathbb{E}_{\pi(x)} \mathbb{E}_{p(z)} \log(1 - D(x, z)) \right]$$

## Recap of previous lecture

### Standard GAN

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

### Main problems

- ▶ Vanishing gradients (non-saturating GAN does not suffer of it);
- ▶ Mode collapse (caused by behaviour of Jensen-Shannon divergence).

### Informal theoretical results

The real images distribution  $\pi(\mathbf{x})$  and the generated images distribution  $p(\mathbf{x}|\theta)$  are low-dimensional and have disjoint supports.

In this case

$$KL(\pi||p) = KL(p||\pi) = \infty, \quad JSD(\pi||p) = \log 2.$$

---

Goodfellow I. J. et al. Generative Adversarial Networks, 2014

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

## Recap of previous lecture

### Wasserstein distance

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

- ▶  $\Gamma(\pi, p)$  – the set of all joint distributions  $\gamma(\mathbf{x}, \mathbf{y})$  with marginals  $\pi$  and  $p$  ( $\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} = p(\mathbf{y})$ ,  $\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \pi(\mathbf{x})$ )
- ▶  $\gamma(\mathbf{x}, \mathbf{y})$  – transportation plan (the amount of "dirt" that should be transported from point  $\mathbf{x}$  to point  $\mathbf{y}$ ).
- ▶  $\|\mathbf{x} - \mathbf{y}\|$  – the amount,  $\|\mathbf{x} - \mathbf{y}\|$  – the distance.

### 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  $\|f\|_L \leq K$  are  $K$ -Lipschitz continuous functions  $(f : \mathcal{X} \rightarrow \mathbb{R})$ .

# Outline

1. Wasserstein GAN
2. WGAN with Gradient Penalty
3. Spectral Normalization GAN
4. f-divergence minimization

# Outline

1. Wasserstein GAN
2. WGAN with Gradient Penalty
3. Spectral Normalization GAN
4. f-divergence minimization

# Wasserstein GAN

## Theorem (Kantorovich-Rubinstein duality)

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

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

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

# Wasserstein GAN

## Standard GAN objective

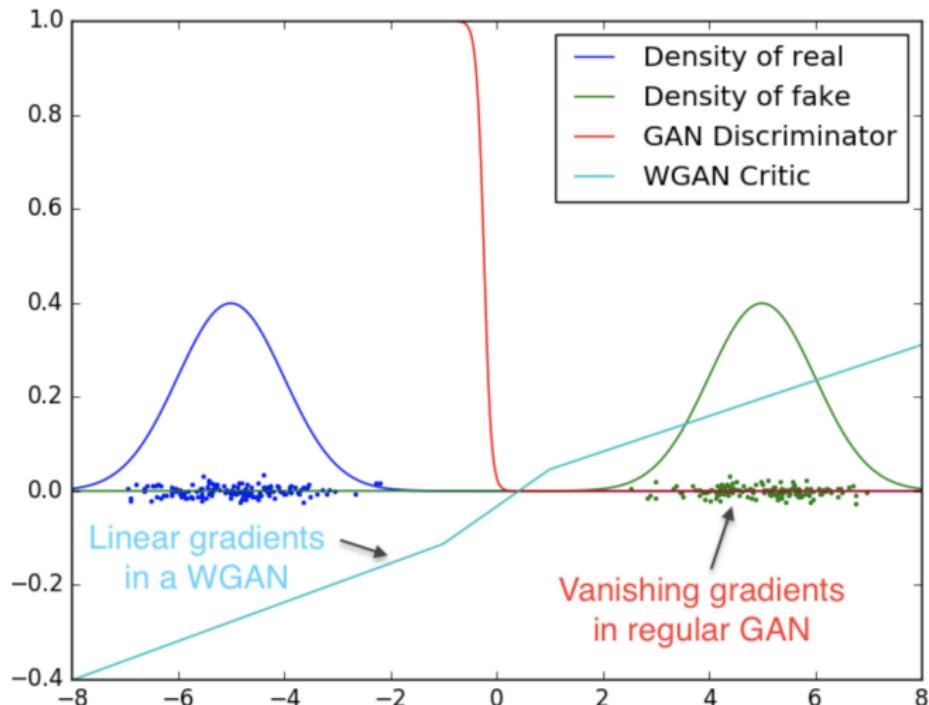
$$\min_{\theta} \max_{\phi} \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}, \phi) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z}, \theta), \phi))$$

## WGAN objective

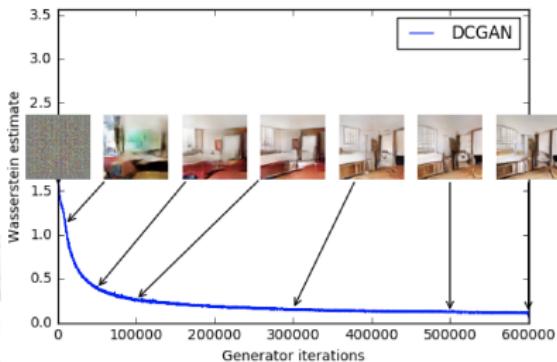
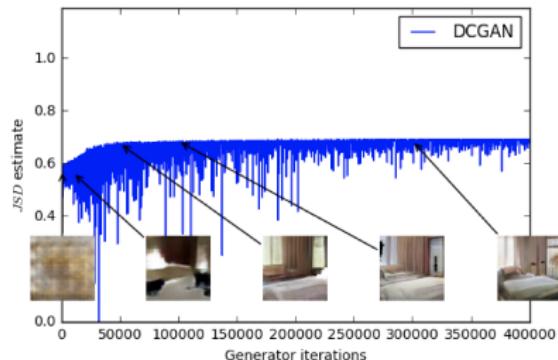
$$\min_{\theta} W(\pi || p) \approx \min_{\theta} \max_{\phi \in \Phi} [\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}, \phi) - \mathbb{E}_{p(\mathbf{z})} f(G(\mathbf{z}, \theta), \phi)] .$$

- ▶ 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 is large, it is hard to train the critic till optimality. If the clipping parameter is too small, it could lead to vanishing gradients.

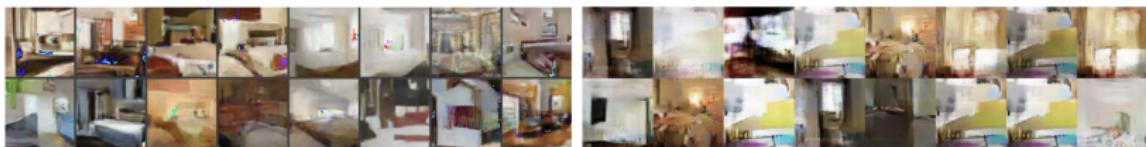
# Wasserstein GAN



# Wasserstein GAN



- ▶  $JSD$  correlates poorly with the sample quality. Stays constant nearly maximum value  $\log 2 \approx 0.69$ .
- ▶  $W$  is highly correlated with the sample quality.

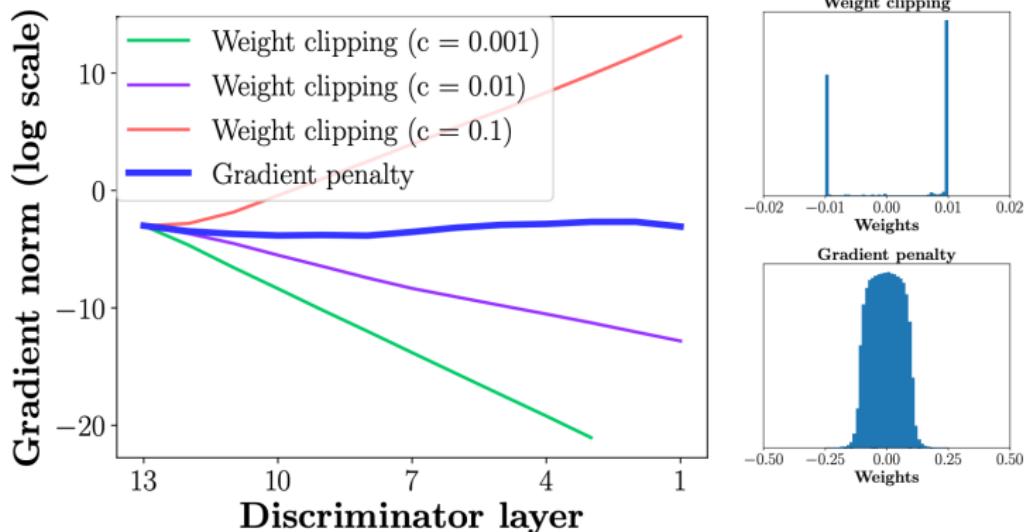


"In no experiment did we see evidence of mode collapse for the WGAN algorithm."

# Outline

1. Wasserstein GAN
2. WGAN with Gradient Penalty
3. Spectral Normalization GAN
4. 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 distribution 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

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

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

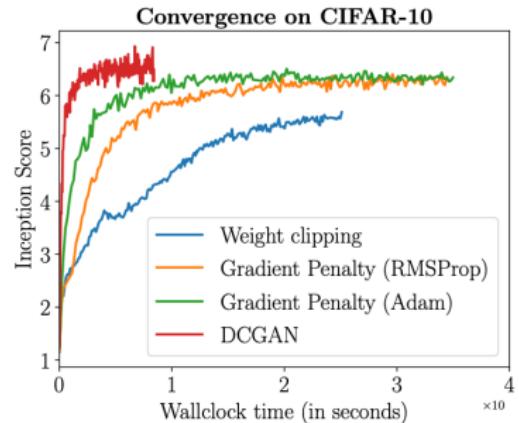
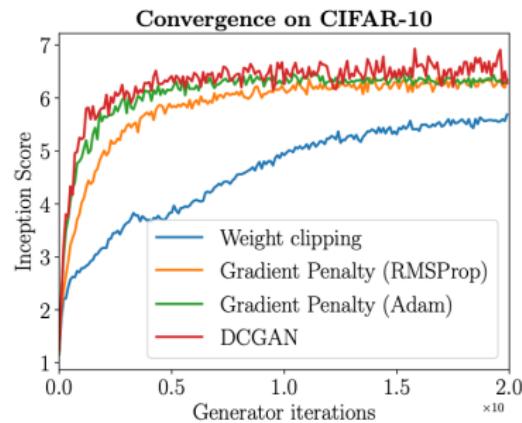
A differentiable function is 1-Lipschitz if and only if it has gradients with norm at most 1 everywhere.

### Gradient penalty

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

- ▶ Samples  $\hat{\mathbf{x}}_t = t\mathbf{y} + (1 - t)\mathbf{z}$  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.

# Wasserstein GAN with Gradient Penalty



## WGANGP convergence

Min. score	Only GAN	Only WGANGP	Both succeeded	Both failed
1.0	0	8	192	0
3.0	1	88	110	1
5.0	0	147	42	11
7.0	1	104	5	90
9.0	0	0	0	200

# Outline

1. Wasserstein GAN
2. WGAN with Gradient Penalty
3. Spectral Normalization GAN
4. f-divergence minimization

# Spectral Normalization GAN

## Definition

$\|\mathbf{A}\|_2$  is a *spectral norm* of matrix  $\mathbf{A}$ :

$$\|\mathbf{A}\|_2 = \max_{\mathbf{h} \neq 0} \frac{\|\mathbf{A}\mathbf{h}\|_2}{\|\mathbf{h}\|_2} = \max_{\|\mathbf{h}\|_2 \leq 1} \|\mathbf{A}\mathbf{h}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T \mathbf{A})},$$

where  $\lambda_{\max}(\mathbf{A}^T \mathbf{A})$  is the largest eigenvalue value of  $\mathbf{A}^T \mathbf{A}$ .

## Statement 1

if  $g$  is a K-Lipschitz function then

$$\|\mathbf{g}\|_L \leq K = \sup_{\mathbf{x}} \|\nabla g(\mathbf{x})\|_2.$$

## Statement 2

Lipschitz norm of superposition is bounded above by product of Lipschitz norms

$$\|\mathbf{g}_1 \circ \mathbf{g}_2\|_L \leq \|\mathbf{g}_1\|_L \cdot \|\mathbf{g}_2\|_L$$

## Spectral Normalization GAN

Let consider the critic  $f(\mathbf{x}, \phi)$  of the following form:

$$f(\mathbf{x}, \phi) = \mathbf{W}_{K+1} \sigma_K (\mathbf{W}_K \sigma_{K-1} (\dots \sigma_1 (\mathbf{W}_1 \mathbf{x}) \dots)).$$

This feedforward network is a superposition of simple functions.

- ▶  $\sigma_k$  is a pointwise nonlinearities. We assume that  $\|\sigma_k\|_L = 1$  (it holds for ReLU).
- ▶  $\mathbf{g}(\mathbf{x}) = \mathbf{W}\mathbf{x}$  is a linear transformation ( $\nabla \mathbf{g}(\mathbf{x}) = \mathbf{W}$ ).

$$\|\mathbf{g}\|_L \leq \sup_{\mathbf{x}} \|\nabla \mathbf{g}(\mathbf{x})\|_2 = \|\mathbf{W}\|_2.$$

### Critic spectral norm

$$\|f\|_L \leq \|\mathbf{W}_{K+1}\|_2 \cdot \prod_{k=1}^K \|\sigma_k\|_L \cdot \|\mathbf{W}_k\|_2 = \prod_{k=1}^{K+1} \|\mathbf{W}_k\|_2.$$

If we replace the weights in the critic  $f(\mathbf{x}, \phi)$  by

$\mathbf{W}_k^{SN} = \mathbf{W}_k / \|\mathbf{W}_k\|_2$ , we will get  $\|f\|_L \leq 1$ .

# Spectral Normalization GAN

How to compute  $\|\mathbf{W}\|_2 = \sqrt{\lambda_{\max}(\mathbf{W}^T \mathbf{W})}$ ?

If we apply SVD to compute the  $\|\mathbf{W}\|_2$  at each iteration, the algorithm becomes intractable.

## Power iteration method

- ▶  $\mathbf{u}_0$  – random vector.
- ▶ for  $k = 0, \dots, K - 1$ : ( $K$  is a fixed number of steps)

$$\mathbf{v}_{k+1} = \frac{\mathbf{W}^T \mathbf{u}_k}{\|\mathbf{W}^T \mathbf{u}_k\|}, \quad \mathbf{u}_{k+1} = \frac{\mathbf{W} \mathbf{v}_{k+1}}{\|\mathbf{W} \mathbf{v}_{k+1}\|}.$$

- ▶ approximate the spectral norm

$$\|\mathbf{W}\|_2 = \sqrt{\lambda_{\max}(\mathbf{W}^T \mathbf{W})} \approx \mathbf{u}_K^T \mathbf{W} \mathbf{v}_K.$$

# Spectral Normalization GAN

---

**Algorithm 1** SGD with spectral normalization

---

- Initialize  $\tilde{\mathbf{u}}_l \in \mathcal{R}^{d_l}$  for  $l = 1, \dots, L$  with a random vector (sampled from isotropic distribution).
- For each update and each layer  $l$ :
  1. Apply power iteration method to a unnormalized weight  $W^l$ :

$$\tilde{\mathbf{v}}_l \leftarrow (W^l)^T \tilde{\mathbf{u}}_l / \| (W^l)^T \tilde{\mathbf{u}}_l \|_2 \quad (20)$$

$$\tilde{\mathbf{u}}_l \leftarrow W^l \tilde{\mathbf{v}}_l / \| W^l \tilde{\mathbf{v}}_l \|_2 \quad (21)$$

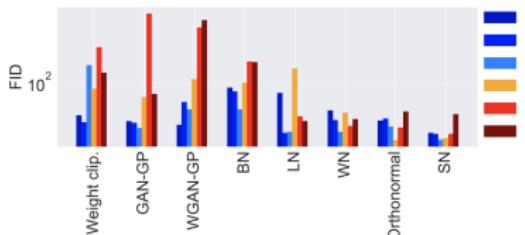
2. Calculate  $\bar{W}_{\text{SN}}$  with the spectral norm:

$$\bar{W}_{\text{SN}}^l(W^l) = W^l / \sigma(W^l), \text{ where } \sigma(W^l) = \tilde{\mathbf{u}}_l^T W^l \tilde{\mathbf{v}}_l \quad (22)$$

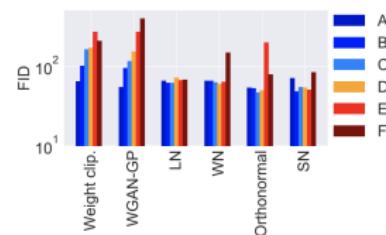
3. Update  $W^l$  with SGD on mini-batch dataset  $\mathcal{D}_M$  with a learning rate  $\alpha$ :

$$W^l \leftarrow W^l - \alpha \nabla_{W^l} \ell(\bar{W}_{\text{SN}}^l(W^l), \mathcal{D}_M) \quad (23)$$

---



(a) CIFAR-10



(b) STL-10

# Outline

1. Wasserstein GAN
2. WGAN with Gradient Penalty
3. Spectral Normalization GAN
4. 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  $\mathcal{S}$  be the set of all possible probability distributions. Then  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  is a divergence if

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

### General divergence minimization task

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

### Challenge

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

# 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	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx - \log(4)$	$u \log u - (u+1) \log(u+1)$

## 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  $\mathbf{x}$ .

## 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  $T(\mathbf{x}) = \log D(\mathbf{x})$ .

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

# f-divergence family

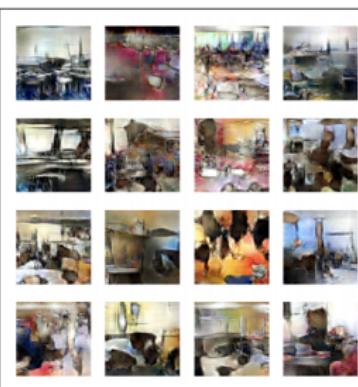
## Variational divergence estimation

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

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



(a) GAN



(b) KL



(c) Squared Hellinger

## Summary

- ▶ 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 works better.
- ▶ Spectral normalization is a weight normalization technique to enforce Lipshitzness, which is helpful for generator and critic.
- ▶ f-divergence family is a unified framework for divergence minimization, which uses variational approximation. Standard GAN is a special case of it.