Deep Generative Models

Lecture 8

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

Recap of previous lecture

Gumbel-max trick

Let $g_k \sim \text{Gumbel}(0, 1)$ for $k = 1, \dots, K$. Then

$$c = \arg\max_k [\log \pi_k + g_k]$$

has a categorical distribution $c \sim \text{Categorical}(\boldsymbol{\pi})$.

Gumbel-softmax relaxation

Concrete distribution = **con**tinuous + dis**crete**

$$\hat{c}_k = \frac{\exp\left(\frac{\log q(k|\mathbf{x}, \phi) + g_k}{\tau}\right)}{\sum_{j=1}^K \exp\left(\frac{\log q(j|\mathbf{x}, \phi) + g_j}{\tau}\right)}, \quad k = 1, \dots, K.$$

Reparametrization trick

$$\nabla_\phi \mathbb{E}_{q(c|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{e}_c, \boldsymbol{\theta}) = \mathbb{E}_{\text{Gumbel}(0,1)} \nabla_\phi \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}),$$

where $\mathbf{z} = \sum_{k=1}^K \hat{c}_k \mathbf{e}_k$ (all operations are differentiable now).

Maddison C. J., Mnih A., Teh Y. W. The Concrete distribution: A continuous relaxation of discrete random variables, 2016

Jang E., Gu S., Poole B. Categorical reparameterization with Gumbel-Softmax, 2016

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):

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

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 $\mathbf{x} = G(\mathbf{z})$, 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

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

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

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

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.

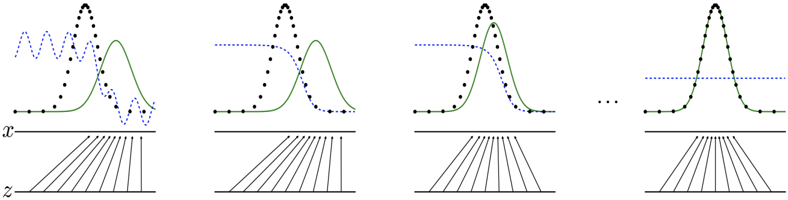
Goodfellow I. J. et al. Generative Adversarial Networks, 2014

Recap of previous lecture

- Generator updates are made in parameter space, discriminator is not optimal at every step.
- Generator and discriminator loss keeps oscillating during GAN training.

Objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} \log D_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D_{\boldsymbol{\phi}}(G_{\boldsymbol{\theta}}(\mathbf{z}))) \right]$$



(a) (b) (c) (d)

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

Outline

1. Wasserstein distance

2. Wasserstein GAN

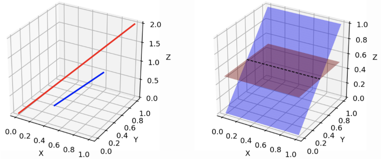
3. Lipschitzness of Wasserstein GAN critic
   - WGAN with Gradient Penalty
   - Spectral Normalization GAN

Outline

1. Wasserstein distance
2. Wasserstein GAN
3. Lipschitzness of Wasserstein GAN critic
   - WGAN with Gradient Penalty
   - Spectral Normalization GAN

Informal theoretical results

- Since $\mathbf{z}$ usually has lower dimensionality compared to $\mathbf{x}$, manifold $G_{\boldsymbol{\theta}}(\mathbf{z})$ has a measure 0 in $\mathbf{x}$ space. Hence, support of $p(\mathbf{x}|\boldsymbol{\theta})$ 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 $p(\mathbf{x}|\boldsymbol{\theta})$ 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

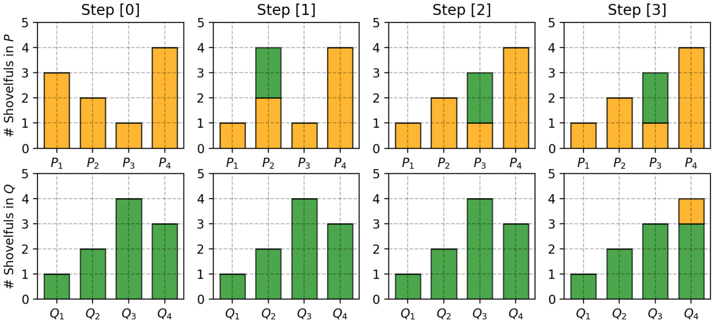
$$KL(\pi||p) = KL(p||\pi) = \infty, \quad JSD(\pi||p) = \log 2$$

- 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**. 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.



$$W(P,Q) = 2(\text{step } 1) + 2(\text{step } 2) + 1(\text{step } 3) = 5$$

Weng L. From GAN to WGAN, 2019

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 π and p.
- $\gamma(\mathbf{x}, \mathbf{y})$ – the amount, $\|\mathbf{x} - \mathbf{y}\|$– the distance.

For better understanding of transportation plan function γ, try to write down the plan for previous discrete case.

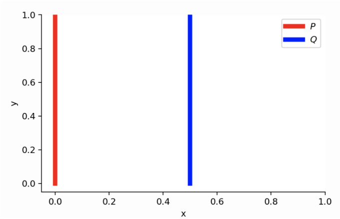
Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

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 $G_{\theta}(\mathbf{z})$ 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 π be a distribution on a compact space $\mathcal{X}$ and $\{p_t\}_{t=1}^{\infty}$ be a sequence of distributions on $\mathcal{X}$.

$$KL(\pi||p_t) \to 0 \,(\text{or } KL(p_t||\pi) \to 0) \tag{1}$$

$$JSD(\pi||p_t) \to 0 \tag{2}$$

$$W(\pi||p_t) \to 0 \tag{3}$$

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

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

Outline

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)

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

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

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

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

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

Wasserstein GAN

Theorem (Kantorovich-Rubinstein duality)

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

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

$$K \cdot W(\pi || p) = \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right] \geq$$

$$\geq \max_{\phi \in \mathbf{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f_\phi(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f_\phi(\mathbf{x}) \right]$$

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

Wasserstein GAN

Standard GAN objective

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

WGAN objective

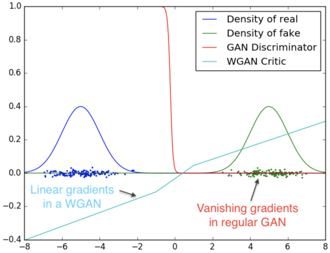
$$\min_{\boldsymbol{\theta}} W(\pi || p) \approx \min_{\boldsymbol{\theta}} \max_{\phi \in \boldsymbol{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f_{\phi}(\mathbf{x}) - \mathbb{E}_{p(\mathbf{z})} f_{\phi}(G_{\boldsymbol{\theta}}(\mathbf{z})) \right].$$

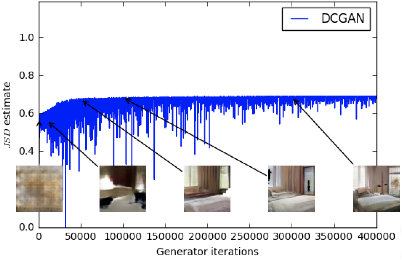
- Discriminator D is similar to the function f, but not the same (it is not a classifier anymore). In the WGAN model, function f is usually called **critic**.
- *"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"*. If the clipping parameter c is too large, it is hard to train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

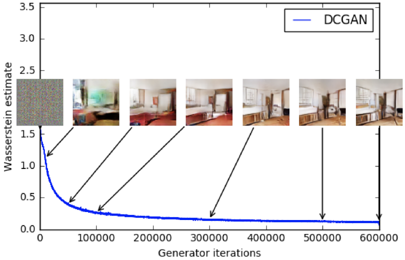
Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

Wasserstein GAN

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







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

Outline

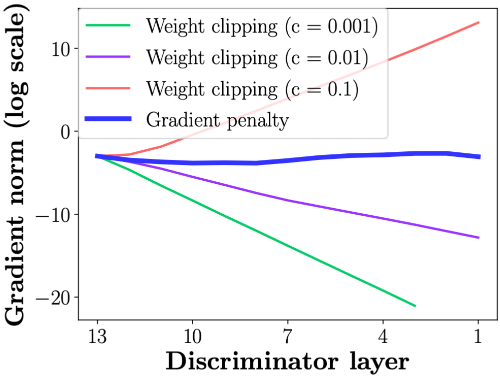
1. Wasserstein distance

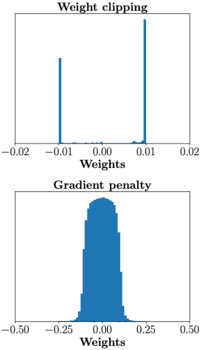
2. Wasserstein GAN

3. Lipschitzness of Wasserstein GAN critic
   - WGAN with Gradient Penalty
   - Spectral Normalization GAN

Outline

Wasserstein GAN with Gradient Penalty





Weight clipping analysis

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

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

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 γ 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 $\mathbf{y} \sim \pi(\mathbf{x})$, $\mathbf{z} \sim p(\mathbf{x}|\boldsymbol{\theta})$, $t \in [0, 1]$ 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 $p(\mathbf{x})$.

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

Wasserstein GAN with Gradient Penalty

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

Gradient penalty

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

- 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 $p(\mathbf{x}|\boldsymbol{\theta})$.
- Enforcing the unit gradient norm constraint everywhere is intractable, it turns out to be sifficient to enforce it only along these straight lines.

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

Outline

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 $\mathbf{g}$ is a K-Lipschitz vector function then

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

Statement 2

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

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

Miyato T. et al. Spectral Normalization for Generative Adversarial Networks, 2018

Spectral Normalization GAN

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

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

This feedforward network is a superposition of simple functions.

- σ_k is a pointwise nonlinearities. We assume that $\|\sigma_k\|_L = 1$ (it holds for ReLU).
- $\mathbf{g}(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$ 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 $f_\phi(\mathbf{x})$ by $\mathbf{W}_k^{SN} = \mathbf{W}_k / \|\mathbf{W}_k\|_2$, we will get $\|f\|_L \leq 1$.

Miyato T. et al. Spectral Normalization for Generative Adversarial Networks, 2018

Spectral Normalization GAN

How to compute $\|\mathbf{W}\|_2 = \sqrt{\lambda_{\max}(\mathbf{W}^T\mathbf{W})}$?
We are not able to apply SVD at each iteration.

Power iteration (PI) method

- $\mathbf{u}_0$ – random vector.
- for $m = 0, \dots, M-1$: (M is a fixed number of steps)
$$\mathbf{v}_{m+1} = \frac{\mathbf{W}^T\mathbf{u}_m}{\|\mathbf{W}^T\mathbf{u}_m\|}, \quad \mathbf{u}_{m+1} = \frac{\mathbf{W}\mathbf{v}_{m+1}}{\|\mathbf{W}\mathbf{v}_{m+1}\|}.$$
- approximate the spectral norm
$$\|\mathbf{W}\|_2 = \sqrt{\lambda_{\max}(\mathbf{W}^T\mathbf{W})} \approx \mathbf{u}_M^T\mathbf{W}\mathbf{v}_M.$$

SNGAN gradient update

- Apply PI method to get approximation of spectral norm.
- Normalize weights $\mathbf{W}_k^{SN} = \mathbf{W}_k / \|\mathbf{W}_k\|_2$.
- Apply gradient rule to $\mathbf{W}$.

Miyato T. et al. Spectral Normalization for Generative Adversarial Networks, 2018

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 neccessary conditions for optimal critic.
- Spectral normalization is a weight normalization technique to enforce Lipshitzness, which is helpful for generator and critic.