Deep Generative Models

Lecture 10

Roman Isachenko

Moscow Institute of Physics and Technology

2022 – 2023

Recap of previous lecture

ELBO objective

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

What is the problem to make the variational posterior model an **implicit** model?
We have to estimate density ratio (in KL term)

$$
r(\mathbf{x}, \mathbf{z}) = \frac{q_1(\mathbf{x}, \mathbf{z})}{q_2(\mathbf{x}, \mathbf{z})} = \frac{q(\mathbf{z}|\mathbf{x}, \phi)\pi(\mathbf{x})}{p(\mathbf{z})\pi(\mathbf{x})} = \frac{D(\mathbf{x}, \mathbf{z})}{1 - D(\mathbf{x}, \mathbf{z})}.
$$

Adversarial Variational Bayes

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

Mescheder L., Nowozin S., Geiger A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks, 2017

Recap of previous lecture

Main problems of standard GAN

- Vanishing gradients (solution: non-saturating GAN);
- Mode collapse (caused by Jensen-Shannon divergence).

Standard GAN

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

Informal theoretical results

The real images distribution $\pi(\mathbf{x})$ and the generated images distribution $p(\mathbf{x}|\boldsymbol{\theta})$ 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

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

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

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 ($f : \mathcal{X} \to \mathbb{R}$).

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

Recap of previous lecture

WGAN objective

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

- Function f in WGAN is usually called *critic*.
- If parameters ϕ lie in a compact set $\boldsymbol{\Phi} \in [-c, c]^d$ then $f(\mathbf{x}, \phi)$ will be K-Lipschitz continuous function.

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

"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"

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

Outline

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

2. f-divergence minimization

3. Evaluation of likelihood-free models

Outline

1. Lipschitzness of Wasserstein GAN critic
   - WGAN with Gradient Penalty
   - Spectral Normalization GAN
2. f-divergence minimization
3. Evaluation of likelihood-free models

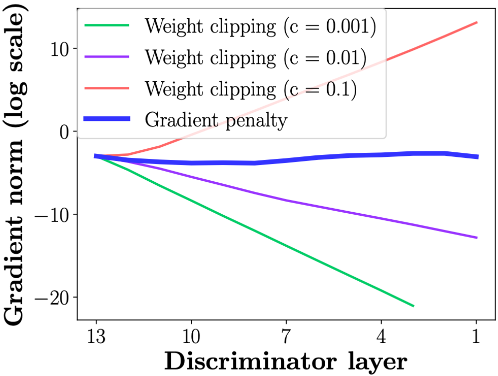
Outline

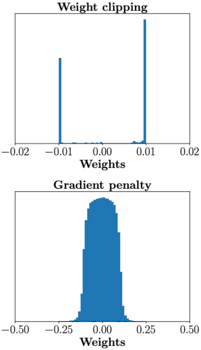
1. Lipschitzness of Wasserstein GAN critic
   WGAN with Gradient Penalty
   Spectral Normalization GAN

2. f-divergence minimization

3. Evaluation of likelihood-free models

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

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

- 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 $p(\mathbf{x}|\boldsymbol{\theta})$.
- 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

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

2. f-divergence minimization

3. Evaluation of likelihood-free models

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 $f(\mathbf{x}, \phi)$ 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.

- 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 $f(\mathbf{x}, \phi)$ 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

Outline

1. Lipschitzness of Wasserstein GAN critic
   - WGAN with Gradient Penalty
   - Spectral Normalization GAN
2. f-divergence minimization
3. Evaluation of likelihood-free models

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 $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ 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

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

Chalenge

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

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$

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

f-divergence family

Fenchel conjugate

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

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

f-divergence

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

$$= \int p(\mathbf{x}) \sup_{t \in \mathrm{dom}_{f*}} \left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})} t - f^*(t)\right) d\mathbf{x} =$$

$$= \int \sup_{t \in \mathrm{dom}_{f*}} \left(\pi(\mathbf{x})t - p(\mathbf{x})f^*(t)\right) d\mathbf{x}.$$

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

$$D_f(\pi||p) = \int \sup_{t \in \mathrm{dom}_{f*}} (\pi(\mathbf{x})t - p(\mathbf{x})f^*(t))\, d\mathbf{x} \geq$$

$$\geq \sup_{T \in \mathcal{T}} \int (\pi(\mathbf{x})T(\mathbf{x}) - p(\mathbf{x})f^*(T(\mathbf{x})))\, d\mathbf{x} =$$

$$= \sup_{T \in \mathcal{T}} \left[\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))\right]$$

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

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

f-divergence family

Variational divergence estimation

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

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

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

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

f-divergence family

Variational divergence estimation

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

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

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

Outline

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

2. f-divergence minimization

3. Evaluation of likelihood-free models

Evaluation of likelihood-free models

How to evaluate generative models?

Likelihood-based models

- Split data to train/val/test.
- Fit model on the train part.
- Tune hyperparameters on the validation part.
- Evaluate generalization by reporting likelihoods on the test set.

Not all models have tractable likelihoods

- VAE: compare ELBO values.
- GAN: ???

Evaluation of likelihood-free models

Let take some pretrained image classification model to get the conditional label distribution $p(y|\mathbf{x})$ (e.g. ImageNet classifier).

What do we want from samples?

- **Sharpness**

Low sharpness

High sharpness

The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).

- **Diversity**

Low diversity

High diversity

The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).

---

image credit: https://deepgenerativemodels.github.io

Evaluation of likelihood-free models

What do we want from samples?

- **Sharpness.** The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).
- **Diversity.** The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).

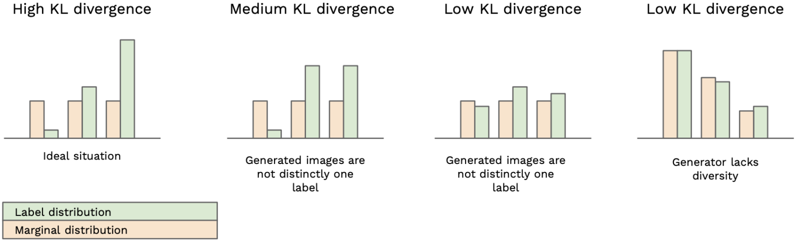


image credit: https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a

Summary

- 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.
- f-divergence family is a unified framework for divergence minimization, which uses variational approximation. Standard GAN is a special case of it.
- We need measure of quality for implicit models like GANs. One way is to analyze sharpness and diversity of samples.