

Deep Generative Models

Lecture 9

Roman Isachenko

Moscow Institute of Physics and Technology

Autumn, 2021

Recap of previous lecture

Autoregressive flow prior

$$\log p(\mathbf{z}|\boldsymbol{\lambda}) = \log p(\boldsymbol{\epsilon}) + \log \det \left| \frac{d\boldsymbol{\epsilon}}{d\mathbf{z}} \right|; \quad \mathbf{z} = g(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) = f^{-1}(\boldsymbol{\epsilon}, \boldsymbol{\lambda})$$

Theorem

VAE with the AF prior for latent code \mathbf{z} is equivalent to using the IAF posterior for latent code $\boldsymbol{\epsilon}$.

$$\begin{aligned} \mathcal{L}(q, \theta) &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \underbrace{\left(\log p(f(\mathbf{z}, \boldsymbol{\lambda})) + \log \left| \det \frac{\partial f(\mathbf{z}, \boldsymbol{\lambda})}{\partial \mathbf{z}} \right| \right)}_{\text{AF prior}} - \log q(\mathbf{z}|\mathbf{x}) \right] \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(f(\mathbf{z}, \boldsymbol{\lambda})) - \underbrace{\left(\log q(\mathbf{z}|\mathbf{x}) - \log \left| \det \frac{\partial f(\mathbf{z}, \boldsymbol{\lambda})}{\partial \mathbf{z}} \right| \right)}_{\text{IAF posterior}} \right] \end{aligned}$$

Recap of previous lecture

ELBO

$$p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \rightarrow \max_{\boldsymbol{\phi}, \boldsymbol{\theta}}.$$

- ▶ Normal variational distribution
 $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}_{\boldsymbol{\phi}}^2(\mathbf{x}))$ is poor (e.g. has only one mode).
- ▶ Flows models convert a simple base distribution to a complex one using an invertible transformation with simple Jacobian.

Flow model in latent space

$$\log q_K(\mathbf{z}_K|\mathbf{x}, \boldsymbol{\phi}_*) = \log q(\mathbf{z}_0|\mathbf{x}, \boldsymbol{\phi}) - \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \boldsymbol{\phi}_k)}{\partial \mathbf{z}_{k-1}} \right) \right|.$$

Let's use $q_K(\mathbf{z}_K|\mathbf{x}, \boldsymbol{\phi}_*)$, $\boldsymbol{\phi}_* = \{\boldsymbol{\phi}, \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K\}$ as a variational distribution. Here, $\boldsymbol{\phi}$ – encoder parameters, $\{\boldsymbol{\phi}_k\}_{k=1}^K$ – flow parameters.

Recap of previous lecture

Variational distribution

$$\log q_K(\mathbf{z}_K | \mathbf{x}, \phi_*) = \log q(\mathbf{z}_0 | \mathbf{x}, \phi) - \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \phi_k)}{\partial \mathbf{z}_{k-1}} \right) \right|.$$

ELBO objective

$$\begin{aligned} \mathcal{L}(\phi, \theta) = \mathbb{E}_{q(\mathbf{z}_0 | \mathbf{x}, \phi)} & \left[\log p(\mathbf{x}, \mathbf{z}_K | \theta) - \log q(\mathbf{z}_0 | \mathbf{x}, \phi) + \right. \\ & \left. + \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \phi_k)}{\partial \mathbf{z}_{k-1}} \right) \right| \right]. \end{aligned}$$

- ▶ Obtain samples \mathbf{z}_0 from the encoder.
- ▶ Apply flow model $\mathbf{z}_K = g(\mathbf{z}_0, \{\phi_k\}_{k=1}^K)$.
- ▶ Compute likelihood for \mathbf{z}_K using the decoder, base distribution for \mathbf{z}_0 and the Jacobian.
- ▶ We do not need an inverse flow function if we use flows in variational inference.

Recap of previous lecture

Images are discrete data flow is a continuous model. We need to convert a discrete data distribution to a continuous one.

Uniform dequantization bound

$$\mathbf{x} \sim \text{Categorical}(\boldsymbol{\pi}), \quad \mathbf{u} \sim U[0, 1], \quad \mathbf{y} = \mathbf{x} + \mathbf{u} \sim \text{Continuous}$$

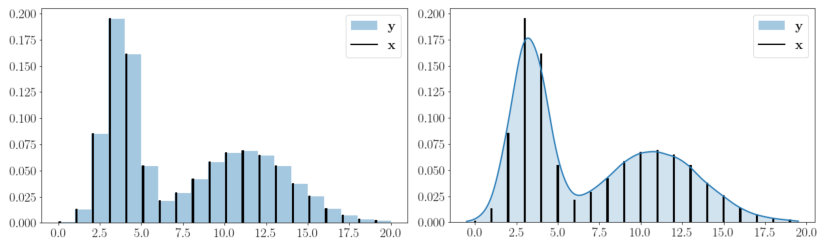
$$\log P(\mathbf{x}|\boldsymbol{\theta}) \geq \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\boldsymbol{\theta}) d\mathbf{u}.$$

Variational dequantization bound

Introduce variational dequantization noise distribution $q(\mathbf{u}|\mathbf{x})$ and treat it as an approximate posterior.

$$\log P(\mathbf{x}|\boldsymbol{\theta}) \geq \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\boldsymbol{\theta})}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} = \mathcal{L}(q, \boldsymbol{\theta}).$$

Recap of previous lecture



Flow model for dequantization

$$q(\mathbf{u}|\mathbf{x}) = p(h^{-1}(\mathbf{u}, \phi)) \cdot \left| \det \frac{\partial h^{-1}(\mathbf{u}, \phi)}{\partial \mathbf{u}} \right|.$$

Variational dequantization bound

$$\mathcal{L}(q, \theta) = \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u}.$$

Ho J. et al. Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design, 2019

Recap of previous lecture

Disentanglement learning

A disentangled representation is a one where single latent units are sensitive to changes in single generative factors, while being invariant to changes in other factors.

β -VAE

$$\mathcal{L}(q, \theta, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta) - \beta \cdot KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})).$$

Representations becomes disentangled by setting a stronger constraint with $\beta > 1$. However, it leads to poorer reconstructions and a loss of high frequency details.

ELBO surgery

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta, \beta) = \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta)}_{\text{Reconstruction loss}} - \underbrace{\beta \cdot \mathbb{I}_q[\mathbf{x}, \mathbf{z}]}_{\text{MI}} - \underbrace{\beta \cdot KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

DIP-VAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}) = \prod_{j=1}^d q(z_j)$$

DIP-VAE Objective

$$\begin{aligned}\mathcal{L}_{\text{DIP}}(q, \theta) &= \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z})) = \\ &= \frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z}))] - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z})) = \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta)]}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x}, \mathbf{z}]}_{\text{MI}} - \underbrace{(1 + \lambda) \cdot KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}\end{aligned}$$

DIP-VAE

$$\mathcal{L}_{\text{DIP}}(q, \theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) - \lambda \cdot \underbrace{KL(q(\mathbf{z}) || p(\mathbf{z}))}_{\text{intractable}}$$

Let match the moments of $q(\mathbf{z})$ and $p(\mathbf{z})$:

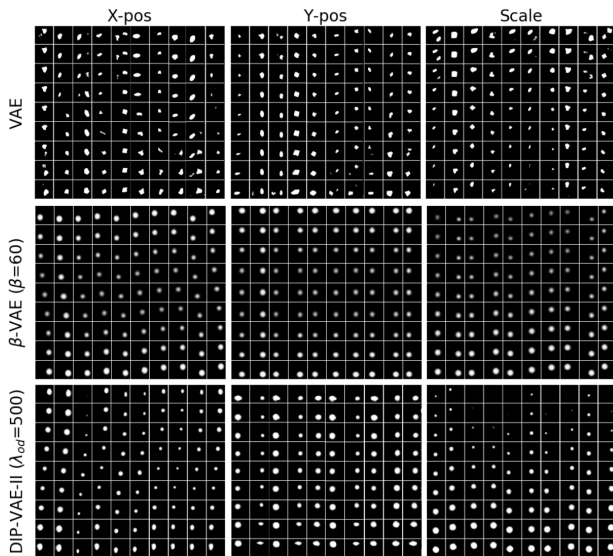
$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{q(\mathbf{z})} \left[(\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z}))(\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z}))^T \right]$$

DIP-VAE regularizes $\text{cov}_{q(\mathbf{z})}(\mathbf{z})$ to be close to the identity matrix.

Objective

$$\begin{aligned} \max_{q, \theta} & \left[\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) - \right. \\ & \left. - \lambda_1 \sum_{i \neq j} [\text{cov}_{q(\mathbf{z})}(\mathbf{z})]_{ij}^2 - \lambda_2 \sum_i \left([\text{cov}_{q(\mathbf{z})}(\mathbf{z})]_{ii} - 1 \right)^2 \right] \end{aligned}$$

DIP-VAE



Kumar A., Sattigeri P., Balakrishnan A. *Variational Inference of Disentangled Latent Concepts from Unlabeled Observations*, 2017

Challenging Disentanglement Assumptions

Whether unsupervised disentanglement learning is even possible for arbitrary generative models?

Theorem

For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^d p(z_i)$. Then, there exists an infinite family of bijective functions $f : \text{supp}(\mathbf{z}) \rightarrow \text{supp}(\mathbf{z})$ such that

- ▶ $\frac{\partial f_i(\mathbf{z})}{\partial z_j} \neq 0$ almost everywhere for all i and j (i.e., \mathbf{z} and $f(\mathbf{z})$ are completely entangled);
- ▶ and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \text{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).

Theorem claims that unsupervised disentanglement learning is impossible for arbitrary generative models with a factorized prior.

Challenging Disentanglement Assumptions

Assume we have $p(\mathbf{z})$ and some $p(\mathbf{x}|\mathbf{z})$ defining a generative model. Consider any unsupervised disentanglement method and assume that it finds a representation that is perfectly disentangled with respect to \mathbf{z} in the generative model.

- ▶ Theorem claims that $\exists \hat{\mathbf{z}} = f(\mathbf{z})$ where $\hat{\mathbf{z}}$ is completely entangled with respect to \mathbf{z} .
- ▶ Since the (unsupervised) disentanglement method only has access to observations \mathbf{x} , it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

Challenging Disentanglement Assumptions

Proof (1)

1. Consider the function $g : \text{supp}(\mathbf{z}) \rightarrow [0, 1]^d$:

$$g_i(\mathbf{u}) = P(z_i \leq u_i), \quad i = 1, \dots, d.$$

- ▶ g is bijective (since $p(\mathbf{z}) = \prod_{i=1}^d p(z_i)$).
- ▶ $\frac{\partial g_i(\mathbf{u})}{\partial u_i} \neq 0$, for all i and $\frac{\partial g_i(\mathbf{u})}{\partial u_j} = 0$ for all $i \neq j$.
- ▶ $g(\mathbf{z})$ is an independent d -dimensional uniform distribution.

2. Consider $h : (0, 1]^d \rightarrow \mathbb{R}^d$

$$h_i(\mathbf{u}) = \psi^{-1}(u_i), \quad i = 1, \dots, d.$$

Here ψ denotes the CDF of a standard normal distribution.

- ▶ h is bijective.
- ▶ $\frac{\partial h_i(\mathbf{u})}{\partial u_i} \neq 0$, for all i and $\frac{\partial h_i(\mathbf{u})}{\partial u_j} = 0$ for all $i \neq j$.
- ▶ $h(g(\mathbf{z}))$ is a d -dimensional standard normal distribution.

Challenging Disentanglement Assumptions

Proof (2)

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be an arbitrary orthogonal matrix with $A_{ij} \neq 0$ for all i, j . The family of such matrices is infinite.

- ▶ \mathbf{A} is orthogonal, it is invertible and thus defines a bijective linear operator.
- ▶ $\mathbf{A}h(g(\mathbf{z})) \in \mathbb{R}^d$ is hence an independent, multivariate standard normal distribution.
- ▶ $h^{-1}(\mathbf{A}h(g(\mathbf{z}))) \in \mathbb{R}^d$ is an independent d -dimensional uniform distribution.

Define $f : \text{supp}(\mathbf{z}) \rightarrow \text{supp}(\mathbf{z})$:

$$f(\mathbf{u}) = g^{-1}(h^{-1}(\mathbf{A}h(g(\mathbf{z}))))).$$

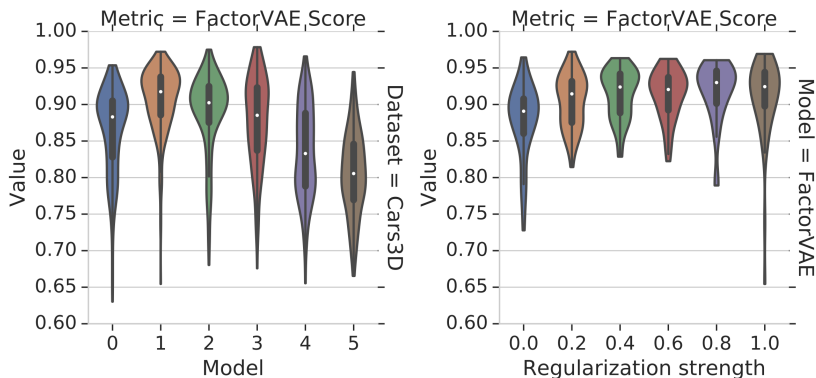
By definition $f(\mathbf{z})$ has the same marginal distribution as \mathbf{z} :

$$P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u}) \text{ and } \frac{\partial f_i(\mathbf{z})}{\partial z_j} \neq 0.$$

Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

Importance of different models and hyperparameters for disentanglement



Locatello F. et al. *Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations*, 2018

Challenging Disentanglement Assumptions

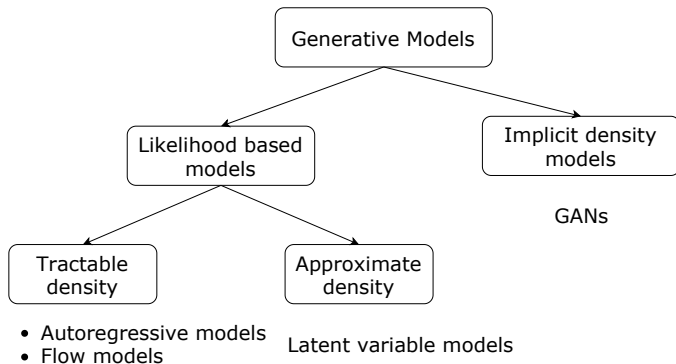
Agreement of different disentanglement metrics

Dataset = Noisy-dSprites

BetaVAE Score (A)	100	80	44	41	46	37
FactorVAE Score (B)	80	100	49	52	25	38
MIG (C)	44	49	100	76	6	42
DCI Disentanglement (D)	41	52	76	100	-8	38
Modularity (E)	46	25	6	-8	100	13
SAP (F)	37	38	42	38	13	100
	(A)	(B)	(C)	(D)	(E)	(F)

Locatello F. et al. *Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations*, 2018

Generative models zoo



Likelihood based models

Is likelihood a good measure of model quality?

Poor likelihood
Great samples

$$p_1(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \epsilon \mathbf{I})$$

For small ϵ this model will generate samples with great quality, but likelihood will be very poor.

Great likelihood
Poor samples

$$p_2(\mathbf{x}) = 0.01p(\mathbf{x}) + 0.99p_{\text{noise}}(\mathbf{x})$$

$$\begin{aligned} \log [0.01p(\mathbf{x}) + 0.99p_{\text{noise}}(\mathbf{x})] &\geq \\ &\geq \log [0.01p(\mathbf{x})] = \log p(\mathbf{x}) - \log 100 \end{aligned}$$

Noisy irrelevant samples, but for high dimensions $\log p(\mathbf{x})$ becomes proportional to m .

Likelihood-free learning

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

Where did we start

We would like to approximate true data distribution $\pi(\mathbf{x})$. Instead of searching true $\pi(\mathbf{x})$ over all probability distributions, learn function approximation $p(\mathbf{x}|\theta) \approx \pi(\mathbf{x})$.

Imagine we have two sets of samples

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

Two sample test

$$H_0 : \pi(\mathbf{x}) = p(\mathbf{x}|\theta), \quad H_1 : \pi(\mathbf{x}) \neq p(\mathbf{x}|\theta)$$

Define test statistic $T(\mathcal{S}_1, \mathcal{S}_2)$. The test statistic is likelihood free. If $T(\mathcal{S}_1, \mathcal{S}_2) < \alpha$, then accept H_0 , else reject it.

Likelihood-free learning

Two sample test

$$H_0 : \pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}), \quad H_1 : \pi(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$$

Desired behaviour

- ▶ $p(\mathbf{x}|\boldsymbol{\theta})$ minimizes the value of test statistic $T(\mathcal{S}_1, \mathcal{S}_2)$.
- ▶ It is hard to find an appropriate test statistic in high dimensions. $T(\mathcal{S}_1, \mathcal{S}_2)$ could be learnable.

GAN objective

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

- ▶ **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.

Vanilla GAN optimality

Theorem

The minimax game

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

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

Proof (fixed G)

$$\begin{aligned} V(G, D) &= \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{x}|\theta)} \log(1 - D(\mathbf{x})) \\ &= \int \underbrace{[\pi(\mathbf{x}) \log D(\mathbf{x}) + p(\mathbf{x}|\theta) \log(1 - D(\mathbf{x}))]}_{y(D)} d\mathbf{x} \end{aligned}$$

$$\frac{dy(D)}{dD} = \frac{\pi(\mathbf{x})}{D(\mathbf{x})} - \frac{p(\mathbf{x}|\theta)}{1 - D(\mathbf{x})} = 0 \quad \Rightarrow \quad D^*(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}$$

Vanilla GAN optimality

Proof continued (fixed $D = D^*$)

$$\begin{aligned} V(G, D^*) &= \mathbb{E}_{\pi(\mathbf{x})} \log \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)} + \mathbb{E}_{p(\mathbf{x}|\theta)} \log \frac{p(\mathbf{x}|\theta)}{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)} \\ &= KL\left(\pi(\mathbf{x}) \parallel \frac{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}{2}\right) + KL\left(p(\mathbf{x}|\theta) \parallel \frac{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}{2}\right) - 2 \log 2 \\ &= 2JSD(\pi(\mathbf{x}) \parallel p(\mathbf{x}|\theta)) - 2 \log 2. \end{aligned}$$

Jensen-Shannon divergence (symmetric KL divergence)

$$JSD(\pi(\mathbf{x}) \parallel p(\mathbf{x}|\theta)) = \frac{1}{2} \left[KL\left(\pi(\mathbf{x}) \parallel \frac{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}{2}\right) + KL\left(p(\mathbf{x}|\theta) \parallel \frac{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}{2}\right) \right]$$

Could be used as a distance measure!

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

Vanilla GAN optimality

Theorem

The minimax game

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

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

Proof

for fixed G :

$$D^*(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}$$

for fixed $D = D^*$:

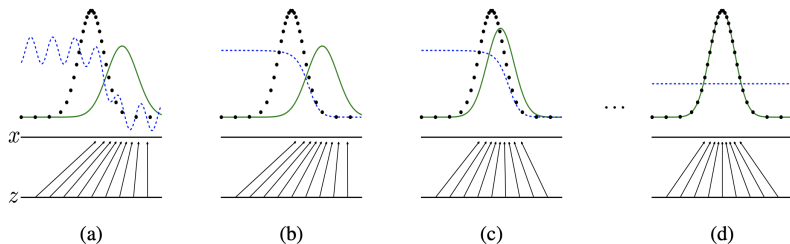
$$\min_G V(G, D^*) = \min_G [2JSD(\pi||p) - \log 4] = -\log 4, \quad \pi(\mathbf{x}) = p(\mathbf{x}|\theta).$$

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.

Vanilla GAN

Objective

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



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

Summary

- ▶ Majority of disentanglement learning models use heuristic objective or regularizers to achieve the goal, but the task itself could not be solved without good inductive bias.
- ▶ Likelihood is not a perfect criteria to measure quality of generative model.
- ▶ Adversarial learning suggest to solve minimax problem to match the distributions.
- ▶ Vanilla GAN tries to optimize Jensen-Shannon divergence (in theory).