Challenges in Generative Models and Latent Variable Models

Xiao Wang

Department of Statistics
Purdue University

Recent Advances in Statistical Analysis of Imaging Data Workshop
Dec. 4—5, 2020

Introduction to Generative Models

Generative Models

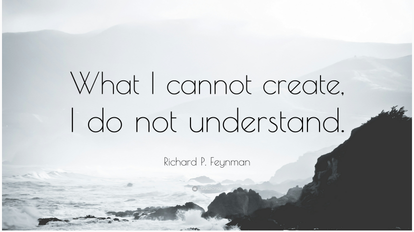
- Given training data, generate new samples from same distribution

(a) Training data $\sim p_{\text{data}}(x)$

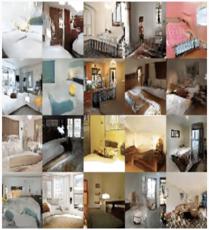
(b) Generated samples $\sim p_{\text{model}}(x)$

- Address density estimation: Explicit density estimation vs. implicit density estimation

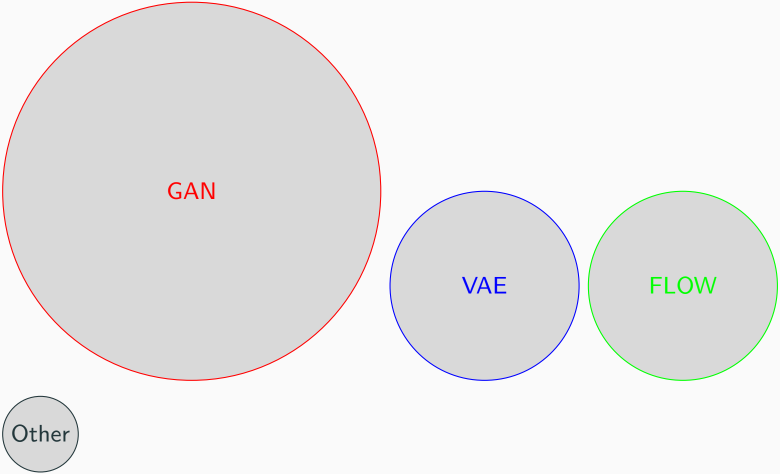
Why Generative Models are Important



- High dimensional data analysis, unsupervised learning, latent representation, dimension reduction, embedding, etc.
- Challenging tasks: artwork, super-resolution, NLP, cyber-security, etc.



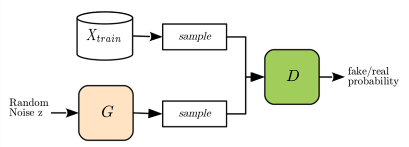
Generative Model Frameworks



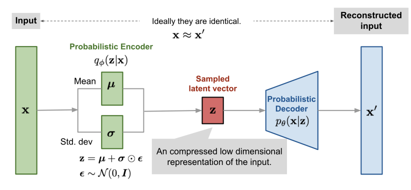
Most generative models belong to latent variable models!

Generative Model Frameworks

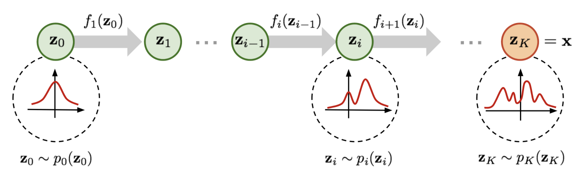
- GAN (Goodfellow et al. 2014):



- VAE (Kingma and Welling, 2013):



- FLOW (Rezende and Mohamed, 2015):



- GAN: $X \sim P_X$ and $G(Z) \sim P_G$ with $Z \sim N(0, I)$,

$$D(P_X, P_G) = \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}_{X \sim P_X} \phi_1(f(X)) - \mathbb{E}_{Y \sim P_G} \phi_2(f(Y)) \Big\},$$

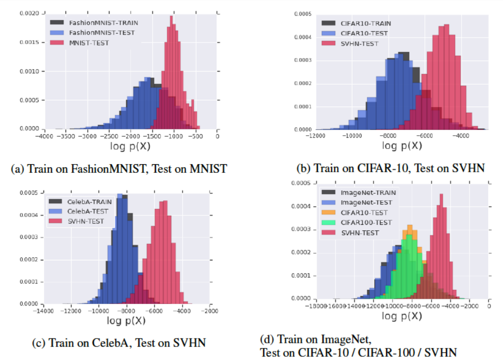
- VAE: Using $q_\phi(z|x) = N[\mu(x), \sigma(x)^2]$ to approximate the true posterior $p_\theta(z|x)$,

$$\log p_\theta(x) = \mathbb{E}_z\Big[\log p_\theta(x|z)\Big] - D_{KL}(q_\phi(z|x) \| p_\theta(z)) + D_{KL}(q_\phi(z|x) \| p_\theta(z|x))$$

- FLOW: $X = G(Z)$ with G being invertible and $Z \sim N(0, I)$,

$$\log p(x) = log \ \ p_Z(G^{-1}(x)) + \log \left| \det \frac{dG^{-1}}{dx} \right|$$

- GAN: Unstable training; Mode collapsing
- VAE: Maximizing lower bound of likelihood; Low quality blurrier sample
- FLOW: Too high latent dimension; Invertible neural networks



(a) Train on FashionMNIST, Test on MNIST

(b) Train on CIFAR-10, Test on SVHN

(c) Train on CelebA, Test on SVHN

(d) Train on ImageNet, Test on CIFAR-10 / CIFAR-100 / SVHN

Figure 2: image credits: Nalisnick et al., ICLR 2019

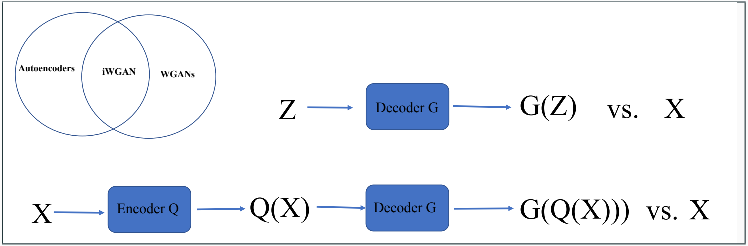
How to check the performance? How to learn the intrinsic dimension of the data? How to perform out-of-distribution detection?

Inferential Wasserstein GANs (iWGAN)

What's New

- Can we propose a model which provides a unifying framework combining the best of VAEs and GANs in a principal way?
- Do there even exist these two mappings, the encoder Q and the decoder G, for any high-dimensional random variable X such that $Q(X) \sim Z$ and $G(Z) \sim X$?
- Is there any probabilistic interpretation such as the maximum likelihood principle on encoder-decoder GANs?
- Developments in this direction:
  — VAE-GAN (Larsen, 2016)
  — Adversarially Learned Inference (ALI) (Dumoulin, 2016)
  — Auto-encoding GANs (α-GAN) (Rosca, 2017)
  — Adversarial Generator Encoders (AGE) (Ulyanov, 2018)

Inferential Wasserstein GANs (Chen, Gao, Wang, 2020)



- Objectives:
  - Match the distribution of the latent space with the prior distribution
  - Match the decoded distribution with the data distribution
  - Match the reconstructed distribution with the data distribution

Encoder and Decoder

- Meaningful encoding and feasible decoder
- Nash's embedding theorem (Nash, 1956)

Theorem

Consider a continuous random variable $X \in \mathcal{X}$, where $\mathcal{X}$ is a d-dimensional smooth Riemannian manifold. Then, there exist two mappings $Q^* : \mathcal{X} \to \mathbb{R}^p$ and $G^* : \mathbb{R}^p \to \mathcal{X}$, with $p = \max\{d(d+5)/2, d(d+3)/2+5\}$, such that $Q^*(X)$ follows a multivariate normal distribution with zero mean and identity covariance matrix and $G^* \circ Q^*$ is an identity mapping, i.e., $X = G^*(Q^*(X))$.

WGAN

- Wasserstein distance: The natural geometry for probability measures (Kantorovich, Koopmans, Nobel'75; Villani, Fields'10)
- WGAN:

$$W_1(P_X, P_G) = \inf_{\pi \in \Pi(P_X, P_Z)} \mathbb{E}_{(X,Z)\sim\pi} \|X - G(Z)\|$$

- By the Kantorovich-Rubinstein duality,

$$W_1(P_X, P_G) = \sup_{f \in \mathcal{F}} \mathbb{E}_{X\sim P_X}\left[f(X)\right] - \mathbb{E}_{Z\sim P_Z}\left[f(G(Z))\right]$$

- Both are difficult constrained optimization problems.

Primal and Dual Optimal Values

- The primal variable π for the primal problem is also a dual variable for the dual problem, and the primal variable f for the dual problem is also a dual variable for the primal problem.
- Introduce the encoder Q to approximate the posterior distribution $p(z|x)$.
- The optimal value of the primal problem satisfies

$$\begin{aligned}&\inf_{\pi}\sup_{f}\mathbb{E}_{\pi}\left\|X-G(Z)\right\|+\int_{x}f(x)\Big(P_X(x)-\int_{z}\pi(x,z)dz\Big)dx-\int_{z}f(G(z))\Big(P_Z(z)-\int_{x}\pi(x,z)dx\Big)dz\\&=\inf_{Q}\sup_{f}\mathbb{E}_X\|X-G(Q(X))\|+\mathbb{E}_X\big[f(G(Q(X)))\big]-\mathbb{E}_Z\big[f(G(Z))\big],\end{aligned}$$

- The optimal value of the dual problem satisfies

$$\begin{aligned}&\sup_{f}\inf_{\pi}\mathbb{E}_X\big[f(X)\big]-\mathbb{E}_Z\big[f(G(Z))\big]-\int_{\mathcal{X}\times\mathcal{Z}}\pi(x,z)\Big(f(x)-f(G(z))-\|x-G(z)\|\Big)dxdz\\&=\sup_{f}\inf_{Q}\mathbb{E}_X\|X-G(Q(X))\|+E_X\Big[f(G(Q(X)))\Big]-\mathbb{E}_Z\big[f(G(Z))\big].\end{aligned}$$

iWGAN Objective

- iWGAN:

$$\overline{W}_1(P_X, P_G) = \inf_{Q \in \mathcal{Q}} \sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim P_X} \|X - G(Q(X))\| + \mathbb{E}_{X \sim P_X}\big[f(G(Q(X)))\big] - \mathbb{E}_{Z \sim P_Z}\big[f(G(Z))\big].$$

  - The iWGAN objective is equivalent to

$$\overline{W}_1(P_X, P_G) = \inf_{Q \in \mathcal{Q}} W_1(P_X, P_{G(Q(X))}) + W_1(P_{G(Q(X))}, P_G),$$

    and $W_1(P_X, P_G) \le \overline{W}_1(P_X, P_G)$.
  - This upper bound is tight. If there exists a $Q^* \in \mathcal{Q}$ such that $Q^*(X)$ has the same distribution with P_Z, then $W_1(P_X, P_G) = \overline{W}_1(P_X, P_G)$. We have $\overline{W}_1(P_X, P_G) = 0 \Longleftrightarrow P_X = P_{G(Q^*(X))} = P_G$.

Generalization Bound for iWGANs

- For supervised learning, the generalization error is the difference between the expected loss (test error) and the empirical loss (training error).
- In practice, we minimize the empirical version, $\widehat{\overline{W}}_1(P_X, P_G)$, of $\overline{W}_1(P_X, P_G)$ to learn both the encoder and the decoder.

Theorem

Given a generator $G \in \mathcal{G}$, and given n samples $(x_1, \ldots, x_n)$ from $\mathcal{X} = \{x : \|x\| \leq B\}$, with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, we have

$$W_1(P_X, P_G) \leq \widehat{\overline{W}}_1(P_X, P_G) + 2\widehat{\mathfrak{R}}_n(\mathcal{F}) + 3B\sqrt{\frac{2}{n}\log\left(\frac{2}{\delta}\right)},$$

where $\widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^n \epsilon_i f(x_i)\right]$ is the empirical Rademacher complexity of the 1-Lipschitz function set $\mathcal{F}$, in which ϵ_i is the Rademacher variable.

Generalization Bound for iWGANs

- The 1-Wasserstein distance between P_X and P_G can be dominantly upper bounded by the empirical $\widehat{W}_1(P_X, P_G)$ and Rademacher complexity of $\mathcal{F}$.
- The capacity of $\mathcal{Q}$ determines the value of $\widehat{W}_1(P_X, P_G)$.
- When $\mathcal{F}$ is a set of 1-Lipschitz neural network, Bartlett et al. (2017) established $\widehat{\mathfrak{R}}_n(\mathcal{F})$ of order $\mathcal{O}(B\sqrt{L^3/n})$, where L denotes the depth of network $f \in \mathcal{F}$, and Li et al. (2018) showed a similar upper bound with an order of $\mathcal{O}(B\sqrt{Ld^2/n})$ can be obtained by utilizing the results from , where d is the width of the network.

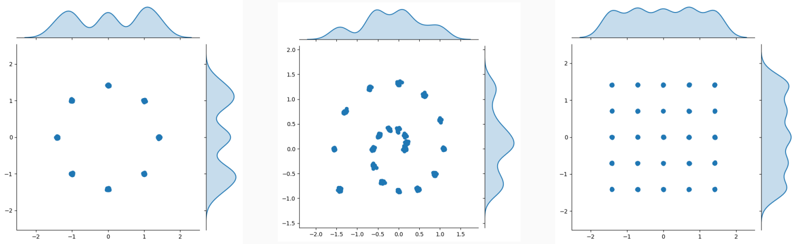
The Algorithm

- When to stop training:
  - The duality gap can be defined as

$$\text{DualGap}(\widetilde{G}, \widetilde{Q}, \widetilde{f}) = \sup_{f \in \mathcal{F}} L(\widetilde{G}, \widetilde{Q}, f) - \inf_{G \in \mathcal{G}, Q \in \mathcal{Q}} L(G, Q, \widetilde{f}),$$

    where $L(G, Q, f) = \mathbb{E}_X \|X - G(Q(X))\| + \mathbb{E}_X[f(G(Q(X)))] - \mathbb{E}_Z[f(G(Z))]$.
  - If $\widetilde{G}$ outputs the same distribution as X and $\widetilde{Q}$ outputs the same distribution as Z, the duality gap is zero and $X = \widetilde{G}(\widetilde{Q}(X))$ for $X \sim P_X$.

Mixture of Gaussian



(a) RING **(b)** Swiss Roll **(c)** GRID

Figure 3: Three toy datasets with an increasing difficulty.

Mixture of Gaussian: Generated Samples

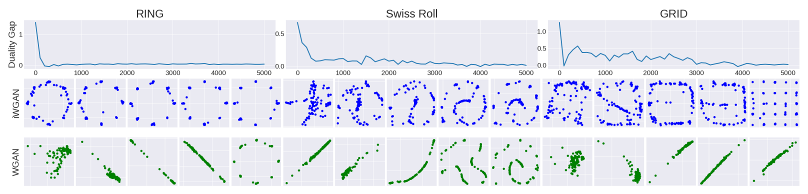
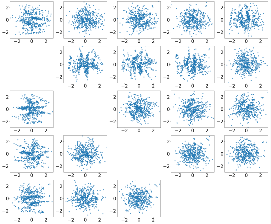


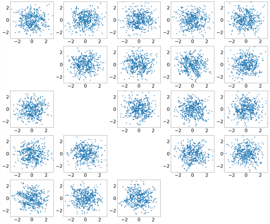
Figure 4: Duality gap and generated samples from iWGANs on mixture of Gaussians

- The duality gap converges to 0
- Our model converges to the true distribution very fast without the mode collapse.

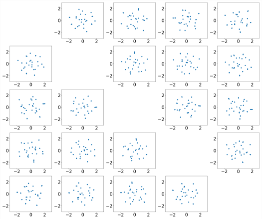
Mixture of Gaussian: Latent Space



(a) RING



(b) Swiss Roll



(c) GRID

Figure 5: Latent Space of Mixture of Gaussians

CelebA: Generated Samples

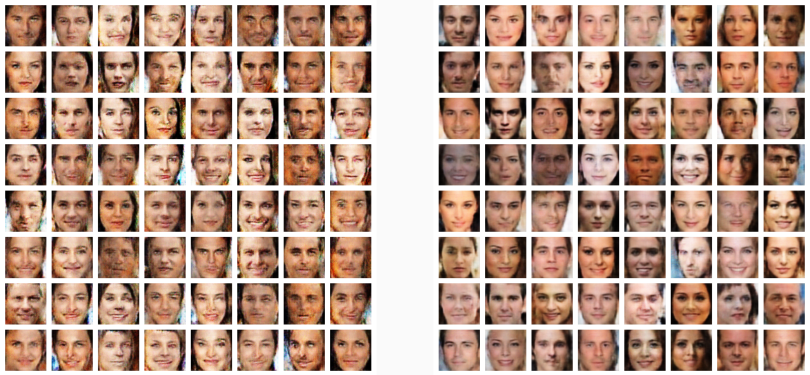


Figure 6: Left:WGAN-GP; Right:iWGAN

CelebA: Latent Space

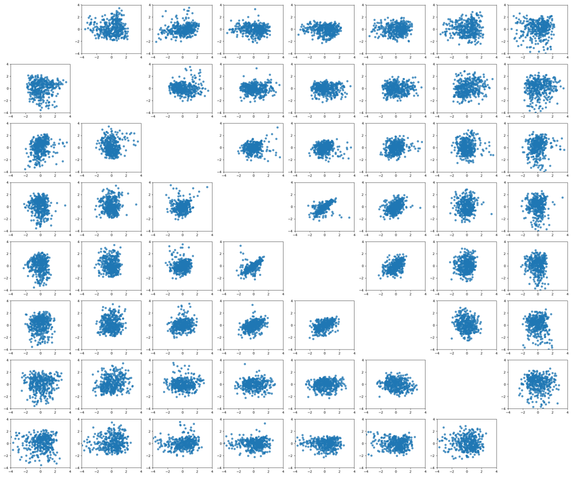


Figure 7: Latent Space of CelebA dataset: the first 8 dimensions of the latent space calculated by $Q(x)$.

CelebA: Interpolations

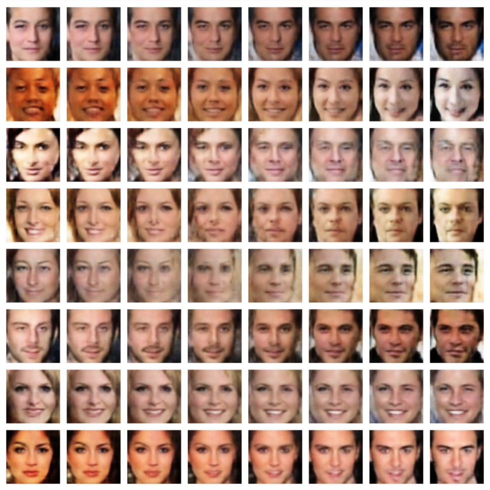


Figure 8: Interpolations between two images

CelebA: Quality Check

(a) Samples with high quality scores

(b) Samples with lower quality scores

Figure 9: Sample quality check by iWGAN on CelebA

Conclusions

- We have compared iWGAN with WGAN-GP, WAE, ALI both visually and numerically, in terms of reconstruction, generative sample quality, latent distribution.
- iWGAN is a unified framework to fuse the best of VAEs and WGANs.
- Similar to rejection sampling, latent distribution can be refined to produce the generative distribution which is the same as data distribution (Che et al. 2020).
- Adaptively learn the intrinsic dimension of data manifold.