Deep Generative Models
Lecture 9

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Wasserstein GAN

Wasserstein distance

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

The infimum across all possible joint distributions in $\prod(\pi, p)$ is intractable.

Kantorovich-Rubinstein duality

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

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

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

https://arxiv.org/abs/1701.07875

Wasserstein GAN

Kantorovich-Rubinstein duality

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

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

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

---

https://arxiv.org/abs/1701.07875

Wassestein GAN

Vanilla GAN objective

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

WGAN objective

$$\min_G W(\pi, p) = \min_G \frac{1}{K} \max_{\phi \in \mathbf{\Phi}} \left[\mathbb{E}_{\mathbf{x} \sim \pi} f(\mathbf{x}, \phi) - \mathbb{E}_{\mathbf{z} \sim p} f(G(\mathbf{z}), \phi)\right].$$

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

https://arxiv.org/abs/1701.07875

Wasserstein GAN

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

```
while θ has not converged do
    for t = 0, ..., n_critic do
        Sample {x^(i)}_{i=1}^m ~ P_r a batch from the real data.
        Sample {z^(i)}_{i=1}^m ~ p(z) a batch of prior samples.
        g_w ← ∇_w [ (1/m) Σ_{i=1}^m f_w(x^(i)) − (1/m) Σ_{i=1}^m f_w(g_θ(z^(i))) ]
        w ← w + α · RMSProp(w, g_w)
        w ← clip(w, −c, c)
    end for
    Sample {z^(i)}_{i=1}^m ~ p(z) a batch of prior samples.
    g_θ ← −∇_θ (1/m) Σ_{i=1}^m f_w(g_θ(z^(i)))
    θ ← θ − α · RMSProp(θ, g_θ)
end while
```

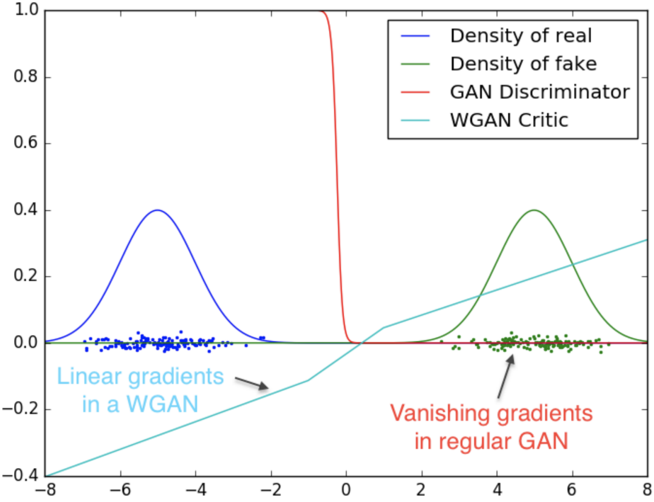
Where:

$$g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))\right]$$

$$g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$$

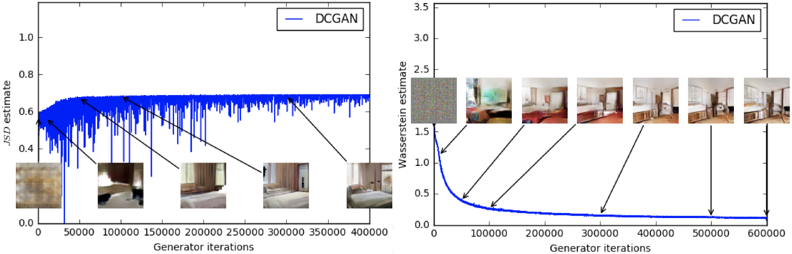
https://arxiv.org/abs/1701.07875

Wasserstein GAN



https://arxiv.org/abs/1701.07875

Wasserstein GAN

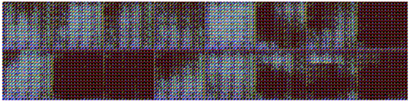


- *JSD* correlates poorly with the sample quality. Stays contast nearly maximum value $\log 2 \approx 0.69$.
- W is highly correlated with the sample quality.

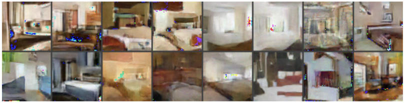
https://arxiv.org/abs/1701.07875

Wasserstein GAN

WGAN converged without batch norm and constant number of filters



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



---

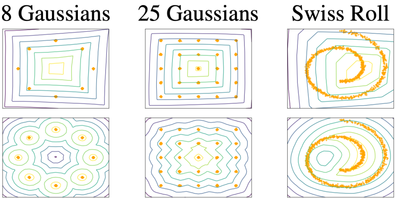
https://arxiv.org/abs/1701.07875

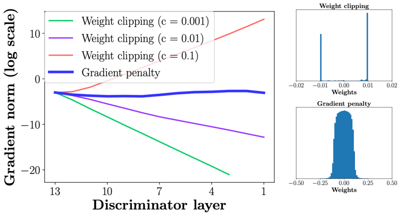
Wasserstein GAN with Gradient Penalty

The generator distribution is fixed and equals to the real distribution + Gaussian noise.
Problems with weight clipping:

- The critic ignores higher moments of the data distribution.
- The gradient either grow or decay exponentially.

Gradient penalty makes the gradients more stable.





https://arxiv.org/abs/1704.00028

Wasserstein GAN with Gradient Penalty

Theorem

Let $\pi(\mathbf{x})$ and $p(\mathbf{x})$ be two distribution in $\mathcal{X}$, a compact metric space. Then, there is 1-Lipschitz function f^* which is the optimal solution of

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

Let γ be the optimal transportation plan between $\pi(\mathbf{x})$ and $p(\mathbf{x})$. Then, if f^* is differentiable $\gamma(x = y) = 0$ and $\mathbf{x}_t = t\mathbf{x} + (1-t)\mathbf{y}$ with $t \in [0,1]$ it holds that

$$\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[\nabla f^*(\mathbf{x}_t) = \frac{\mathbf{y} - \mathbf{x}_t}{\|\mathbf{y} - \mathbf{x}_t\|} = 1\right].$$

Corollary

f^* has gradient norm 1 almost everywhere under $\pi(\mathbf{x})$ and $p(\mathbf{x})$.

---

https://arxiv.org/abs/1704.00028

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

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

- Samples $\hat{\mathbf{x}}$ are uniformly sampled along straight lines between pairs of points from the data distribution $\pi(\mathbf{x})$ and the generator distribuiton $p(\mathbf{x}|\boldsymbol{\theta})$.
- Enforcing the unit gradient norm constraint everywhere is intractable, it turns out sifficient to enforce it only along these straight lines.

---

https://arxiv.org/abs/1704.00028

Wasserstein GAN with Gradient Penalty

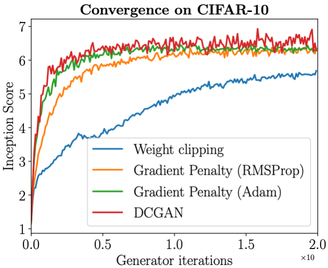
Algorithm 1 WGAN with gradient penalty. We use default values of $\lambda = 10$, $n_{\text{critic}} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$.

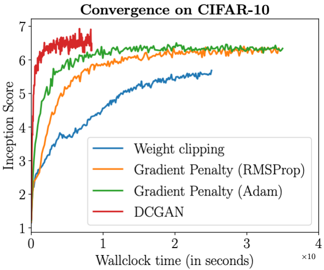
Require: The gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, Adam hyperparameters α, β_1, β_2.
Require: initial critic parameters w_0, initial generator parameters θ_0.

while θ has not converged **do**
  for $t = 1, \ldots, n_{\text{critic}}$ **do**
    for $i = 1, \ldots, m$ **do**
      Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a random number $\epsilon \sim U[0, 1]$.
      $\tilde{\boldsymbol{x}} \leftarrow G_\theta(\boldsymbol{z})$
      $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon)\tilde{\boldsymbol{x}}$
      $L^{(i)} \leftarrow D_w(\tilde{\boldsymbol{x}}) - D_w(\boldsymbol{x}) + \lambda(\|\nabla_{\hat{\boldsymbol{x}}} D_w(\hat{\boldsymbol{x}})\|_2 - 1)^2$
    end for
    $w \leftarrow \text{Adam}(\nabla_w \frac{1}{m} \sum_{i=1}^{m} L^{(i)}, w, \alpha, \beta_1, \beta_2)$
  end for
  Sample a batch of latent variables $\{\boldsymbol{z}^{(i)}\}_{i=1}^{m} \sim p(\boldsymbol{z})$.
  $\theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} -D_w(G_\theta(\boldsymbol{z})), \theta, \alpha, \beta_1, \beta_2)$
end while

https://arxiv.org/abs/1704.00028

Wasserstein GAN with Gradient Penalty





Nonlinearity (G)	[ReLU, LeakyReLU, $\frac{\text{softplus}(2x+2)}{2} - 1$, tanh]
Nonlinearity (D)	[ReLU, LeakyReLU, $\frac{\text{softplus}(2x+2)}{2} - 1$, tanh]
Depth (G)	[4, 8, 12, 20]
Depth (D)	[4, 8, 12, 20]
Batch norm (G)	[True, False]
Batch norm (D; layer norm for WGAN-GP)	[True, False]
Base filter count (G)	[32, 64, 128]
Base filter count (D)	[32, 64, 128]

Min. score	**Only GAN**	**Only WGAN-GP**	**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

https://arxiv.org/abs/1704.00028

Wasserstein GAN with Gradient Penalty

https://arxiv.org/abs/1704.00028

Spectral Normalization GAN

How else could we enforce Lipschitzness?

Fact 1

$$\|\mathbf{g}\|_L = \sup_{\mathbf{x}} \sigma(\nabla \mathbf{g}(\mathbf{x}))$$

Here $\sigma(\mathbf{A})$ – spectral norm of matrix $\mathbf{A}$.

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

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

Fact 2

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

---

https://arxiv.org/abs/1802.05957

Spectral Normalization GAN

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

$$f(\mathbf{x}, \phi) = \mathbf{W}_{K+1} a_K(\mathbf{W}_K a_{K-1}(\dots a_1(\mathbf{W}_1 \mathbf{x}) \dots)).$$

This feedforward network is a superposition of simple functions.

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

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

Critic spectral norm

$$\|f\|_L \leq \prod_{k=1}^{K+1} \sigma(\mathbf{W}_k).$$

If we replace the weights in the critic by $\mathbf{W}_k^{SN} = \mathbf{W}_k / \sigma(\mathbf{W}_k)$, we will get $\|f\|_L \leq 1$.

https://arxiv.org/abs/1802.05957

Spectral Normalization GAN

If we apply singular value decomposition to compute the $\sigma(\mathbf{W})$ at each round of the algorithm, the algorithm becomes computationally heavy.

Power iteration

- $\mathbf{u}$ – random vector.
- repeat

$$\mathbf{v} = \frac{\mathbf{W}^T\mathbf{u}}{\|\mathbf{W}^T\mathbf{u}\|}, \quad \mathbf{u} = \frac{\mathbf{W}\mathbf{v}}{\|\mathbf{W}\mathbf{v}\|}$$

- approximate the spectral norm

$$\sigma(\mathbf{W}) \approx \mathbf{u}^T\mathbf{W}\mathbf{v}$$

---

https://arxiv.org/abs/1802.05957

Spectral Normalization GAN

Algorithm 1 SGD with spectral normalization

- Initialize $\tilde{\boldsymbol{u}}_l \in \mathcal{R}^{d_l}$ for $l = 1, \ldots, L$ 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:

$$\tilde{\boldsymbol{v}}_l \leftarrow (W^l)^{\mathrm{T}} \tilde{\boldsymbol{u}}_l / \|(W^l)^{\mathrm{T}} \tilde{\boldsymbol{u}}_l\|_2 \tag{20}$$

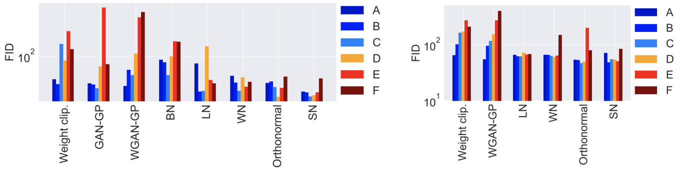
$$\tilde{\boldsymbol{u}}_l \leftarrow W^l \tilde{\boldsymbol{v}}_l / \|W^l \tilde{\boldsymbol{v}}_l\|_2 \tag{21}$$

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

$$\bar{W}^l_{\mathrm{SN}}(W^l) = W^l / \sigma(W^l), \text{ where } \sigma(W^l) = \tilde{\boldsymbol{u}}_l^{\mathrm{T}} W^l \tilde{\boldsymbol{v}}_l \tag{22}$$

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

$$W^l \leftarrow W^l - \alpha \nabla_{W^l} \ell(\bar{W}^l_{\mathrm{SN}}(W^l), \mathcal{D}_M) \tag{23}$$



(a) CIFAR-10 (b) STL-10

Summary

- Likelihood is not a perfect criteria to measure quality of generative model.
- Adversarial learning suggest to solve minimax problem with generator and discriminator.
- Vanila GAN tries to optimize (in some sense) Jensen-Shannon divergence.
- Mode collapse and vanishing gradients are the two main problems of vanila GAN.
- Lots of tips and tricks has to be used to make the GAN training is stable and scalable.
- Wasserstein distance is more appropriate objective function for distribution matching problem.
- Gradient Penalty and Spectral Normalization allows to make Wasserstein GAN even more stable.

Divergences

What do we have?

- Forward KL divergence in maximum likelihood estimation
- Reverse KL in variational inference
- JS divergence in vanilla gan
- Wasserstein distance in WGAN

Divergence minimization

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

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 divertgence if

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

f-divergence family

f-divergence

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

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

Fenchel conjugate

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

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

https://arxiv.org/abs/1606.00709

f-divergence family

Variational divergence estimation

$$
\begin{aligned}
D_f(\pi||p) &= \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \\
&= \int p(\mathbf{x}) \sup_{t \in \mathrm{dom}_{f*}} \left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})} t - f^*(t)\right) d\mathbf{x} \\
&= \int \sup_{t \in \mathrm{dom}_{f*}} \left(\pi(\mathbf{x}) t - p(\mathbf{x}) f^*(t)\right) d\mathbf{x} \\
&\geq \sup_{T \in \mathcal{T}} \int \left(\pi(\mathbf{x}) T(\mathbf{x}) - p(\mathbf{x}) f^*(T(\mathbf{x}))\right) d\mathbf{x} \\
&= \sup_{T \in \mathcal{T}} \left[\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))\right]
\end{aligned}
$$

Here $\mathcal{T} : \mathcal{X} \to \mathbb{R}$ is an arbitrary class of functions.
The lower bound is tight for $T^*(\mathbf{x}) = f'\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right)$.

https://arxiv.org/abs/1606.00709

f-divergence family

Variational divergence estimation

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


(a) GAN (b) KL (c) Squared Hellinger

https://arxiv.org/abs/1606.00709

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

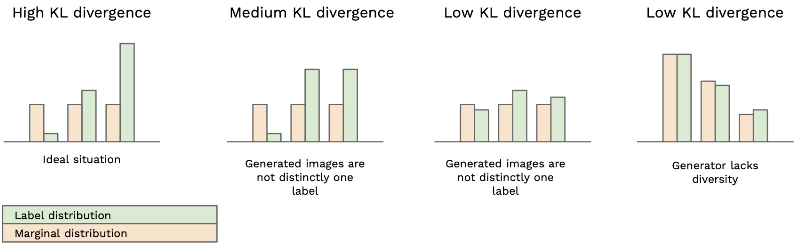
High diversity

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).

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).



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

Evaluation of likelihood-free models

What do we want from samples?

- Sharpness $\Rightarrow$ low $H(y|\mathbf{x}) = -\sum_y \int_{\mathbf{x}} p(y,\mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x}$.
- Diversity $\Rightarrow$ high $H(y) = -\sum_y p(y) \log p(y)$.

Inception Score

$$
\begin{aligned}
IS &= \exp(H(y) - H(y|\mathbf{x})) \\
&= \exp\left(-\sum_y p(y)\log p(y) + \sum_y \int_{\mathbf{x}} p(y,\mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x}\right) \\
&= \exp\left(\sum_y \int_{\mathbf{x}} p(y,\mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)} d\mathbf{x}\right) \\
&= \exp\left(\mathbb{E}_{\mathbf{x}} \sum_y p(y|\mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)}\right) = \exp\left(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x})||p(y))\right)
\end{aligned}
$$

---

https://arxiv.org/abs/1606.03498

Evaluation of likelihood-free models

Inception Score

$$IS = \exp\left(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x}) || p(y))\right)$$

IS limitations

- Inception score depends on the quality of the pretrained classifier $p(y|\mathbf{x})$.
- If generator produces images with a different set of labels from the classifier training set, IS will be low.
- If the generator produces one image per class, the IS will be perfect (there is no measure of intra-class diversity).
- IS only require samples from the generator and do not take into account the desired data distribution $\pi(\mathbf{x})$ directly (only implicitly via a classifier).

https://arxiv.org/abs/1801.01973

Evaluation of likelihood-free models

Theorem

If $\pi(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta})$ has moment generation functions then

$$\pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}) \Leftrightarrow \mathbb{E}_\pi \mathbf{x}^k = \mathbb{E}_p \mathbf{x}^k, \quad \forall k \geq 1.$$

This is intractable to calculate all moments.

Frechet Inception Distance

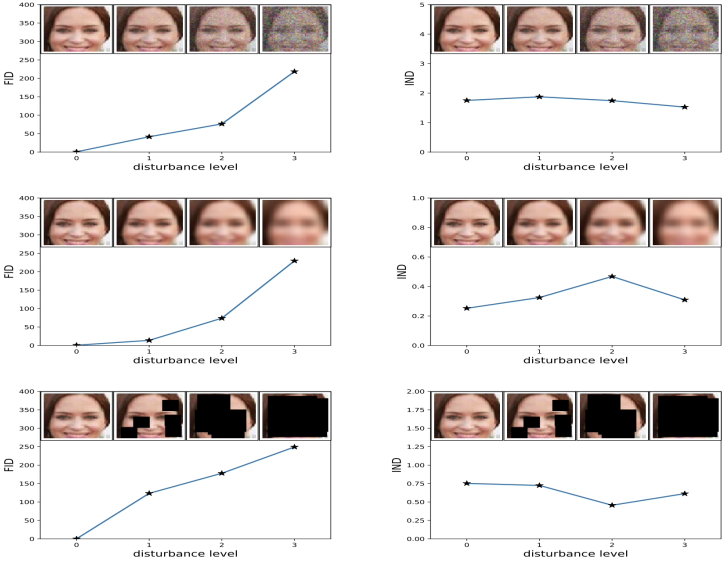
$$D^2(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \mathrm{Tr}\left(\mathbf{C}_\pi + \mathbf{C}_p - 2\sqrt{\mathbf{C}_\pi \mathbf{C}_p}\right)$$

- $\mathbf{m}_\pi$, $\mathbf{C}_\pi$ are mean vector and covariance matrix of feature representations for real samples from $\pi(\mathbf{x})$
- $\mathbf{m}_p$, $\mathbf{C}_p$ are mean vector and covariance matrix of feature representations for generated samples from $p(\mathbf{x}|\boldsymbol{\theta})$.
- Representations are output of intermediate layer from pretrained classification model.

---

https://arxiv.org/abs/1706.08500

Evaluation of likelihood-free models



https://arxiv.org/abs/1706.08500

Evaluation of likelihood-free models

Frechet Inception Distance

$$D^2(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \mathrm{Tr}\left(\mathbf{C}_\pi + \mathbf{C}_p - 2\sqrt{\mathbf{C}_\pi \mathbf{C}_p}\right)$$

FID limitations

- FID depends on the pretrained classification model.
- FID needs a large samples size for evaluation.
- Calculation of FID is slow.
- FID extimates only two sample moments.

---

https://arxiv.org/abs/1706.08500

Summary

References

- **WGAN**: Wasserstein GAN
  https://arxiv.org/abs/1701.07875
  Summary: Proposed Earth-Mover distance as divergence measure. It shows theoretically that EM distance could be better than KL, JS and TV measures. Kantorovich-Rubinstein duality gives the new objective. All we need to enforce is Lipschitz continuity (e.x. by gradient clipping).
- **WGAN-GP**: Improved Training of Wasserstein GANs
  https://arxiv.org/abs/1704.00028
  Summary: Proves that optimal critic has unit gradient almost everywhere. Gradient penalty proposed to enforce Lipchitzness and constraint the critic norm. It greatly improves stability.
- **SN-GAN**: Spectral Normalization for Generative Adversarial Networks
  https://arxiv.org/abs/1802.05957
  Summary: Constrain Lipschitz norm by spectral norm pf weights. Spectral norm of superpositionis less or equal to spectral norms product. Usually works better than WGAN-GP.
- **f-GAN**: Training generative neural samplers using variational divergence minimization
  https://arxiv.org/abs/1606.00709
  Summary: Extend adversarial learning pipeline to any f-divergence. Variational divergence minimization frameworkis derived.
- Improved Techniques for Training GANs
  https://arxiv.org/abs/1606.03498
  Summary: Inception Score for GAN evaluation was proposed.
- A Note on the Inception Score
  https://arxiv.org/abs/1801.01973
  Summary: Inception Score is not an ideal metric.
- GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium
  https://arxiv.org/abs/1706.08500
  Summary: Frechet inception distance was proposed for GAN evaluation.