On the Generalization and Approximability in Generative Adversarial Networks (GANs)

Tengyu Ma
Stanford University

Sanjeev Arora
(Princeton)

Yu Bai
(Stanford)

Rong Ge
(Duke)

Yingyu Liang
(UW Madison)

Andrej Risteski
(MIT)

Yi Zhang
(Princeton/Google)
Standard parametric approach:
- a parametrized family of densities \( \{p_\theta : \theta \in \Omega\} \)
- maximize the log-likelihood

\[
\max_\theta \frac{1}{n} \sum \log p_\theta(x_i)
\]
Learning Latent Variable Models

Density $p_\theta(x)$ involves marginalization

$$p_\theta(x) = \int p_\theta(x, h) dh$$

- mixture models, noisy-or networks, Bayes nets
- energy-based models: deep Boltzmann machine, deep belief networks
- variational auto-encoders

Algorithms: MCMC, EM, contrastive divergence, variational inference

Challenge:

- Difficult to produce sharp edges
- What if the distribution doesn’t have a proper density?
  - E.g., the distribution of real images has low-dimensional support
- Optimization may be challenging
GANs: Learning Parameterized Generators/Samplers

- $Z \sim N(0, I_{k \times k})$
- Neural net $G_\theta(\cdot)$
- $X = G_\theta(Z) \in \mathbb{R}^d$

- Learn $\theta$, without attempts to compute the density of $X$
- $X$ may have low-dimensional support since $k < d$
Other Applications of Learning Parameterized Samplers

- **Style Transfer**
  - Monet ⇆ Photos
  - Zebras ⇆ Horses
  - Summer ⇆ Winter

- **Imitation learning** (imitating the expert policy)
  - Match the state distribution induced by the policy with that by the expert policy; the distribution only has a parametric sampler

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [Zhu et al.’16]
Generative Adversarial Imitation Learning [Ho & Ermon’16]
Algorithms for Learning Generators

1. Define a loss function $L(\theta)$
   - distance between distributions of $X = G_\theta(Z)$ and real images

2. Minimize $L(\theta)$ (by gradient descent and variants)

Notation:
- Given samples $x_1, \ldots, x_n \sim P$; $\hat{P}$: uniform dist. over \{ $x_1, \ldots, x_n$ \}
- $P_\theta$: distribution of $X = G_\theta(Z)$; $\hat{P}_\theta$ = uniform dist. over \{ $G_\theta(z_1), \ldots, G_\theta(z_n)$ \}, $z_i \in N(0,I)$

$$\text{min } L(\theta) := d(\hat{P}_\theta, \hat{P})$$

- TV, KL don’t work because $\hat{P}_\theta$ and $\hat{P}$ may not share support
- Loss function should be geometry-aware
Learning a Loss Function

\[ \min L(\theta) := d(\hat{P}_\theta, \hat{P}) \]

- A classifier/discriminator class \( \mathcal{F} \) (functions that map samples to or 0/1 (or \( \mathbb{R} \))

- \( d(\hat{P}_\theta, \hat{P}) = \) maximum accuracy to classify samples in \( \hat{P}_\theta \) and \( \hat{P} \)
  
  \[ = \) maximum discrepancy of the outputs of \( f \) on \( \hat{P}_\theta \) and \( \hat{P} \) (over \( f \in \mathcal{F} \))

- Optimal \( f \in \mathcal{F} \) is learned by optimization
Wasserstein Distance and W-GANs [AB'16]

- Wasserstein distance (earth mover distance, dual form)
  \[ W(P, Q) = \sup_{f : f \text{ is } 1-\text{Lipschitz}} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{X \sim Q} f(X)| \]

- Distance between empirical dist.
  \[ W(\hat{P}_\theta, \hat{P}) = \sup_{f : f \text{ is } 1-\text{Lipschitz}} \frac{1}{n} \sum_{i=1}^{n} f(G_\theta(z_i)) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \]

- Reality: Loss \( L(\theta) = \text{Integral Probability Metric (IPM)} \)
  \[ W_\mathcal{F}(\hat{P}_\theta, \hat{P}) = \sup_{f_\phi \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f_\phi(G_\theta(z_i)) - \frac{1}{n} \sum_{i=1}^{n} f_\phi(x_i) \]

- \( \mathcal{F} = \text{a parameterized family of functions, called discriminators (e.g., neural nets)} \)
  taking \( \mathcal{F} \subseteq \{1-\text{Lipschitz fun.}\} \quad \rightarrow \quad W_\mathcal{F} \leq W \)
Wasserstein GANs Training

- min max problem

\[
\min_{\theta} \max_{\phi} \mathcal{W}_f(\hat{P}_\theta, \hat{P}) = \min_{\theta} \max_{f\phi \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f_\phi(G_\theta(z_i)) - \frac{1}{n} \sum_{i=1}^{n} f_\phi(x_i) \right|
\]

- Differential function of \( \theta, \phi \)
- Alternating stochastic gradient descent
This Talk: Statistical Theory of GANs
Do W-GANs Learn the Distribution?

If the training succeeds

\[ W_F(\hat{P}_\theta, \hat{P}) \leq \epsilon \]

“training loss”
empirical distance
weak discriminators

\[ P_\theta \text{ is close to } P \]

\[ W(P_\theta, P) \leq g(\epsilon) \]

“test loss”
population distance
strong discriminators

\[ W_F(P_\theta, P) \Rightarrow W(P_\theta, P) \]

Generalization
(of discriminators)

Approximability/
Distinguishability
If $F$ has high-complexity, e.g., $F = \{1$-Lipschitz fun.\}

$$W(F(\hat{P}_\theta), \hat{P}) \neq W(\hat{P}_\theta, P) \quad \text{and} \quad W(F(P_\theta, P) = W(P_\theta, P)$$

- E.g., even if $P = Q = \text{Gaussian}; \hat{P}^n, \hat{Q}^n$ uniform dist. over $n = \text{poly samples from } P, Q$
- Discriminator class is too complex $\Rightarrow$ lack generalization (with poly samples)

$$W(\hat{P}_\theta, P) \gtrsim 1$$

$$W(Q, P) = 0$$

[Arora-Ge-Liang-M.-Zhang, ICML17]
If $\mathcal{F}$ has low-complexity $C$ (in e.g. VC-dim.):

W.h.p.,

$$W_\mathcal{F}(\hat{P}_\theta, \hat{P}) = W_\mathcal{F}(P_\theta, P) \pm O(\sqrt{C/n})$$

Proof follows standard concentration inequalities

true for any $P$ by concentration

E.g., $P = \text{Gaussian}; Q = \hat{P}^m$ (uniform dist. over $m$ samples from $P$ with $m = \text{poly}/\epsilon^2$)

$$W_\mathcal{F}(Q, P) \lesssim \epsilon$$

$$W(Q, P) \gtrsim 1$$

can not approximate $P$ by discrete distributions

For typical $P$,

$$\exists Q, W_\mathcal{F}(Q, P) \not\sim W(Q, P)$$

[Arora-Ge-Liang-M.-Zhang, ICML17]
Dilemma

Generalization (of discriminators) vs Approximability/Distinguishability

\[ W_\mathcal{F}(\hat{P}_\theta, \hat{P}) \quad \Rightarrow \quad W_\mathcal{F}(P_\theta, P) \quad \Rightarrow \quad W(P_\theta, P) \]

If \( \mathcal{F} \) has low-complexity \( C \) (in e.g. VC-dim.):

W.h.p.,
\[ W_\mathcal{F}(\hat{P}_\theta, \hat{P}) = W_\mathcal{F}(P_\theta, P) \pm O(\sqrt{C/n}) \]

- Low-complexity \( \mathcal{F} \) lacks distinguishing power: small discriminators cannot detect mode collapse
- Good generalization but poor approximability \( \rightarrow \) poor diversity

For typical \( P \),
\[ \exists Q, W_\mathcal{F}(Q, P) \not\approx W(Q, P) \]

- Potential Ex. 2., \( P = \) Gaussian; \( Q = \) “real-valued pseudo-random generators"
- \( W_\mathcal{F}(Q, P) \lesssim \epsilon \)
- \( W(Q, P) \gtrsim 1 \)

[Arora-Ge-Liang-M.-Zhang, ICML17]
Empirical test of diversity

- Birthday paradox test for the number of modes [Arora-Risteski-Zhang’18]
- Support of dist. = $N \Rightarrow$ duplicate found in $\sqrt{N}$ samples

Near-duplicates found among 500 samples
(Implying support size $500^2 \approx 250k$)
(Training set has size 200k)
Beyond the Dilemma: Restricted Approximability

Generalization of discriminators

\[ W_{\mathcal{F}}(\hat{P}_\theta, \hat{P}) \rightarrow W_{\mathcal{F}}(P_\theta, P) \rightarrow W(P_\theta, P) \]

If \( \mathcal{F} \) has low-complexity \( C \)

\[ W_{\mathcal{F}}(\hat{P}_\theta, \hat{P}) = W_{\mathcal{F}}(P_\theta, P) \pm O(\sqrt{C/n}) \]

\[ \exists Q, W_{\mathcal{F}}(Q, P) \not\approx W(Q, P) \]

➢ But we only need:

\[ \forall Q = P_\theta, W_{\mathcal{F}}(Q, P) \approx W(Q, P) \]

➢ For a particular generator class, it’s possible to design corresponding parameterized discriminator class \( \mathcal{F} \) with restricted approximability:

\[ \forall Q = P_\theta, W(Q, P)^c \lesssim W_{\mathcal{F}}(Q, P) \lesssim W(Q, P) \]

Generator classes with restricted approximability:

➢ Gaussian, mixture of Gaussian, and exponential family

➢ Injective neural networks generators (next slide) [Bai-M.-Risteski’18]
Assume $G_\theta$ is an injective function. E.g., an $\ell$-layer neural nets with leaky relu activation and full-rank weight matrices.

Define $\mathcal{F} = \{\ell + 2$-layer neural nets with a special structure$\}$. Then,

$$\forall \theta, W(P_\theta, P)^c \lesssim W_\mathcal{F}(P_\theta, P) \lesssim W_\mathcal{F}(\hat{P}_\theta, \hat{P})$$

First polynomial sample complexity result for GANs.

Prior result [Liang’18] works in non-parametric setting with strong smoothness assumptions (requires exponential samples) [Bai-M.-Risteski’18]
Proof Sketch of Restricted Approximability

Goal: $\forall \theta, W(P_\theta, P)^c \preceq W_F(P_\theta, P) \preceq W(P_\theta, P)$

Lemma [Zhang-Liu-Zhou-Xu-He’17]

Suppose

- $P, P_\theta$ both have proper density $p$ and $p_\theta$
- $\mathcal{F}$ can approximate $\log p - \log p_\theta$

Then,

$$KL(P_\theta || P) \leq W_F(P_\theta, P)$$

Proof: let $f = \log p - \log p_\theta$ be the discriminator

$$KL(P_\theta || P) + KL(P || P_\theta) = E_P[f] - E_{P_\theta}[f] \leq W_F(P_\theta, P)$$
Proof Sketch of Restricted Approximability

Goal: \[ \forall \theta, W(P_\theta, P)^c \lesssim W_\mathcal{F}(P_\theta, P) \lesssim W(P_\theta, P) \]

If generator \( G_\theta \) is an invertible neural net (\( k = d \))

- Design a neural net that contains linear combination of \( \log p, \log p_\theta \)

Then,

\[ KL(P_\theta \| P) \leq W_\mathcal{F}(P_\theta, P) \]

Transportation inequality: Bobkov-Gotze
Proof Sketch of Restricted Approximability

- **Goal:** \( \forall \theta, W(P_\theta, P)^c \lesssim W_\mathcal{F}(P_\theta, P) \lesssim W(P_\theta, P) \)

If generator \( G_\theta \) is an injective neural net \((k < d)\)

- \( \log p, \log p_\theta \) don’t exist

- \( P^\delta, P^\delta_\theta \): convolution of \( P \) and \( P_\theta \) with a small Gaussian of variance \( \delta \)

Approximating \( \log p^\delta, \log p^\delta_\theta \) by \( \mathcal{F} \) (non-trivial) gives

\[
W(P_\theta, P)^2 \approx W(P^\delta_\theta, P^\delta)^2 \lesssim KL(P^\delta_\theta \| P^\delta) \lesssim W_\mathcal{F}(P^\delta_\theta, P^\delta) \\
\lesssim 1/\delta^4 \cdot W(P^\delta, P^\delta_\theta) \\
\leq 1/\delta^4 \cdot W(P, P_\theta)
\]

Quantifying the error gives:

\[
W(P_\theta, P)^2 \lesssim \inf_\delta (W_\mathcal{F}(P^\delta, P^\delta_\theta) + \delta \log(1/\delta)) \lesssim \text{poly}(d) \cdot W(p, q)^{1/3}
\]

IPM between smoothed distributions
Conclusions

This talk: statistical theory for GANs
- Generalization vs Approximability Tradeoff
- Restricted Approximability

Looking ahead:
- Stronger restricted approximability result?
  - beyond the curse of dimensionality: a natural target function class = \{near-optimal discriminators\}
  - connection to metric embeddings?
- Other formulation of GANs (e.g., cycleGANs)
- Understanding optimization of GANs (and in general non-convex optimization)

Thank you!