Contents

Daniel Roy, Shai Ben-David

1 Intro

We'll discuss:

1. Informal problem
2. Basic algorithm
3. Some properties
4. Problems
   (a) Vanishing gradients
   (b) Mode collapse
5. State-of-the-art: Unrolled GANs

1.1 Informal problem

We have a true distribution $P_{\text{true}} = P_{\text{data}}$ over $\mathbb{R}^D$ (e.g. $D = 1000$). We want to find a (neural network) function $G_\theta : \mathbb{R}^d \to \mathbb{R}^D$, called the **generator**, such that for Z some randomness, $Z \sim$(some given distribution/noise), $G_\theta(Z) \sim P_{\text{model}}(Z|\theta)$.

1.2 Basic algorithm

We introduce another neural network $D_\phi : \mathbb{R}^D \to \mathbb{R}$ which tries to determine whether the data point (image) is from the real model or from the generator.

The objective function is

$$\min_\theta \max_\phi \underbrace{\left[-\frac{1}{2} \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} \ln D_\phi(x) - \frac{1}{2} \mathop{\mathbb{E}}_{z} \ln(1 - D_\phi(G_\theta(z)))\right]}_{V(\phi,\theta)}.$$

The optimal D is

$$D_{\text{opt}}(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x|\theta)}.$$

In practice, we have a finite number of samples, so the objective is

$$\min_\theta \max_\phi \left[-\frac{1}{2} \sum_i \ln D_\phi(x_i) - \frac{1}{2} \sum_j \ln(1 - D_\phi(G_\theta(z_j))) \right].$$

This isn't a convex problem, so you can't switch min and max.

In practice, do simultaneous gradient descent. Take a step in ϕ, then θ, etc.[1] But the generator actually wants to perform well against the max. In unrolled GANs, do n steps of the discriminator update.

What is the goal? We want G to generate a distribution minimizing the JS divergence

$$\min_\theta JS(p_{\text{model}}(\cdot|\theta)||p_{\text{data}}) \tag{1}$$

$$JS(P||Q) = \frac{1}{2} KL(P||M) + \frac{1}{2} KL(Q||M) \tag{2}$$

$$M = \frac{1}{2}(P + Q). \tag{3}$$

You can also minimize f-divergences,

$$D_f(Q||P) = \int f\left(\frac{dQ}{dP}(x)\right) dP(x).$$

fffff You can get a lot of different divergences/metrics from different f. Ex. One direction of KL is maximum likelihood.

Solving for θ in the minimax against a powerful discriminator would be the same as minimizing the divergence.

Two versions of KL:

- $KL(p_{\text{data}}||p_{\text{model}})$ prefers to overgeneralize.
- $KL(p_{\text{model}}||p_{\text{data}})$ focuses on one mode.

You can interpolate these two cases.

The problem of mode collapse is some problem with the training procedure, not the

The theorist way is to look at TV (total variation, L^1) distance, corresponding to the ideal discriminator (binary).

[1] Contrast with alternating minimization: given the current generator, find the best discriminator. This is unstable.

1.3 Problems

1.3.1 Vanishing gradients

If you try to max in ϕ and min in θ, if the discriminator gets ahead of the generator, the gradients of generator are close to 0, and discriminator gets generator in situation where it doesn't know how to improve itself. So actually use a different objective function.

For other f divergences, the generator is also minimizing a different objective function.

2nd term: on fake data, minimize probability of mistake. Instead, minimize $J^{(G)} = -\frac{1}{2}\mathbb{E}_z \ln D(G(z))$. You tend not to have the vanishing gradient problem.

1.3.2 Mode collapse

This is serious. Fitting on mixtures of Gaussians, over time it fits a mode, jumps to another, and meanders around. The generator does well on one mode. The discriminator chases it around; the generator does not converge.

Unrolled GANs: see that it focused too much somewhere, and penalize.

Rather than discriminator take a single observation, get a bag of samples from the true and fake data; do a 3-sample test—which is more like the real distribution?

Mode problem is generally a problem with log likelihood: it rewards you for matching a mode.

Sometimes caps the discriminator to not make strong predictions.

The point of generator model is that you hope that in learning a good generator, you somehow unlock the structure in the data. There are tricks to making the coordinates more interpretable.

1.4 Theoretical problem

We have a distribution p_{true}. We have a way of representing a distribution, a neural net, and want to get as close as possible.

Suppose we have a family of representations of distributions.

$$F = \{G(z) : z \sim \text{Uniform}, G \in (\text{class of NNs})\}.$$

- What is the input? $x_1, \ldots, x_m \sim p_{\text{true}}$.
- What do we mean by "minimize"? We define a distance between distributions D.

I will show this is too hard for any distance, and we need to restrict it to make it feasible. We want to find a mathematical valid formulation that reflects natural assumptions. (Find formalization that explains reality.)

What is the sample, computational complexity of this problem, ignoring algorithmic issues?

Task: Find $G \in F$ that minimizes $D(p_{\text{true}}, G(z))$.

We introduce the two sample problem. There are two unknown distributions P_1, P_2.

Problem 1.1 (Two sample problem)**:** Given samples $S_1 \sim P_1^m$, $S_2 \sim P_2^m$, determine: is $P_1 = P_2$ or $D(P_1, P_2) > \frac{1}{3} = \varepsilon$?

Consider a simple version.

1. Toy problem 1: $P_1 = \text{Uniform}[1, \ldots, N]$, $P_2 = \text{Uniform}[1, \ldots, \frac{N}{2}]$. Take a constant number $O(1)$ of samples, see if you have any points in $[\frac{N}{2} + 1, \ldots, N]$.

2. Toy problem 2: Now consider P_1 known, P_2 unknown.

   $P_1 = \text{Uniform}[1, \ldots, N]$, $P_2 = \text{Uniform}(A)$ where $A \subseteq [1, \ldots, N]$, $|A| = \frac{N}{2}$.

   Now $n = \Omega(\sqrt{N})$. As long as we don't see a point repeating, we cannot tell if it's uniform from the whole subset or a subset. We need $\Omega(\sqrt{N})$ to see repetitions.

The key difference between these 2 cases: is the problem symmetric?

To make the problem feasible (solve the blowup of sample complexity), I'm going to restrict the discriminator, not the distributions. I want the sample complexity $m(\varepsilon, \delta, d)$ to depend on just the distance ε, and probability of failure δ, and the power of discriminator d (ex. they are in a family of bounded VC dimension).

We want to estimate $D(p_{\text{true}}, G_1)$, $D(p_{\text{true}}, G_2)$.

We pick a notion of distance D based on our class of discriminators such that we have finite sample size approximation guarantees.

Let A be a family of subsets of X (domain of P),

$$D_A(P_1, P_2) = \sup_{a \in A} |P_1(a) - P_2(a)|.$$

This looks like total variation distance, except we restrict to $a \in A$.

This is a metric; it can be justified by taking a set of subsets we're interested in seeing differences in. It follows from VC theory that if A has some finite VC-dimension then the probability

$$\mathbb{P}_{S_1 \sim P_1^m, S_2 \sim P_2^m}[|D_A(S_1, S_2) - D_A(P_1, P_2)| > \varepsilon] < O\left(\frac{md}{l^{m\varepsilon^2}}\right).$$

You can generalize from sets to functions. Replace the sup by

$$\sup_{f \in A} |\mathbb{E}_{P_1}(f) - \mathbb{E}_{P_2}(f)|.$$

We get pseudodimension, or dimension of a family of functions.

This is actually the type of distance we are trying to minimize!

Ex. subsets are all intervals. P_1 is uniform over odd points, P_2 is uniform over even points, with respect to intervals they are very similar; no interval has many more/less odd than even points.

We assume the difference between distributions is detectable by the family of sets/functions.

We were looking at sensor networks; detect whether there was a real change/event.

Can we come up with a family that captures what humans can distinguish?

A discriminator is only a NN of certain size. We can talk about sample sizes. This doesn't say anything about the generator.

How general is this formulation of distance? Converse: If there is distance with finite sample size guarantee, does it have this form (or something similar)?