Wasserstein GAN and Its Improved Training

May 4, 2019

This document is written as I read the “Wasserstein GAN” paper [Arjovsky et al., 2017] and the “Improved Training of Wasserstein GANs” paer [Gulrajani et al., 2017].

1 Introduction

- We are concerned with learning a probability density.
- We have real data examples $\{x^{(i)}\}_{i=1}^m$ generated from a probability distribution P_r (whose density is denoted by P_r) that we do not know.
- We define a parameteric familiy of probability density function $(P_\theta)_{\theta \in \mathbb{R}^d}$. We then would like to find the θ that yields the density matches that of P_r.
- In the GAN approach, we define a random variable Z with a fixed distribution $p(z)$. We pass it through a parameteric function $g_\theta : \mathcal{Z} \to \mathcal{X}$. We then attempt to find θ so that the distribution of $g_\theta(Z)$ matches that of P_r.
- We typically find θ by starting with a random initial parameter and evolve it until the parameter “converges.”
- In the above formulation, there is a hidden assumption that the mapping from θ to P_θ is *continuous*. This means that when a sequence of parameter $(\theta_t)_{t \in \mathbb{N}}$ converges to θ (say, in the Euclidean space), then P_{θ_t} should converge to P_θ. After all, there is no use if there parameter converges, but the distribution does not.
- Convergence of probability distributions depends on how we define the distance between them.
- Let $\rho(P_\theta, P_r)$ denote a distance function between P_θ and P_r.
- A sequence of distribution $(P_t)_{t \in \mathbb{N}}$ *converges* if and only if there is a distribution P_∞ such that $\rho(P_t, P_\infty)$ tends to zero.
- We say a distance ρ *induces a weaker topology* if it makes it easier for a sequence of distribution to converge. More precisely, the topology induces by ρ is weaker than that induced by ρ' if the set of convergence sequences of ρ is a superset of that under ρ'.
- From the discussion on convergence of P_θ, it is desirable that the distance function ρ we use induces a weaker topology because it would allow the optimization process to converge on more paths.
- Moreover, we can use $\rho(P_\theta, P_r)$ as the loss function for the optimization. Minimizing this function is indeed trying to make the model distribution match that of the real distribution.

2 Different Distances

- Let:
  - $\mathcal{X}$ be a compact metric set;
  - Σ denote the set of all Borel subsets of $\mathcal{X}$;
  - $\text{Prob}(\mathcal{X})$ denote the space of probability measure defined on $\mathcal{X}$.
- There are several well-known distance between two probability distribution $P_r, P_g \in \text{Prob}(\mathcal{X})$.
  - The *Total Variation* (TV) distance

    $$\delta(P_r, P_g) = \sum_{A \in \Sigma} \|P_r(A) - P_g(A)\|.$$

  - The *Kullback–Leibler* (KL) divergence

    $$KL(P_r \| P_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) \ \mathrm{d}\mu(x),$$

    where μ is a measure defined on $\mathcal{X}$. The KL divergence is asymmetric. It can also be infinite when there points such that $P_g(x) = 0$ and $P_r(x) > 0$.
  - The *Jensen–Shannon* (JS) divergence

    $$JS(P_r, P_g) = KL(P_r \| P_m) + KL(P_g \| P_m)$$

    where $P_m = (P_r + P_g)/2$. This divergence is symmetrical and always defined because μ can be chosen to make the KL divergence well-behaved.
  - The *Earth-Mover* (EM) distance or Wasserstain-1

    $$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} E_{(x,y)\sim\gamma}[\|x - y|]$$

    where $\Pi(P_r, P_g)$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are P_r and P_g, respectively.
    - $\gamma(x, y)$ indicates how much "mass" must be transported from x to y in order to transform P_r into P_g.
    - The EM distance is the "cost" of the optimal transport plan.
- The paper gives an example of where a simple sequence of probability distributions converges under EM but does not converge under other distances. The target distribution is defined on a low-dimensional manifold which intersects with the model distribution on a set of measure zero. All the distance functions other than EM are not continuous, so we cannot even optimize. The example shows that it is easier to learn a probability distribution under EM in this case.
- It can be shown that EM is much weaker than JS.
- Moreover, $W(P_r, P_g)$ is continuous under mild assumptions.
- Before we can discuss this assumption, though, let us remind ourselves of Lipschitz continuity.
  - Given two metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$, a function f is *Lipschitz* if there exists a real constant $K \geq 0$ such that

    $$d_\mathcal{Y}(f(x_1), f(x_2)) \leq K d_\mathcal{X}(x_1, x_2)$$

    for all $x_1, x_2 \in \mathcal{X}$.

- We say that f is *locally Lipschitz* if, for every $x \in \mathcal{X}$, there exists a neighborhood U of x such that f, restricted to U is Lipschitz continuous.

- Recall the GAN generator function $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$. We will denote the latent vector by z and the generator's parameter by θ. The function is denoted by $g_\theta(z)$.

- The assumption for EM's distance continuity is as follows:

  Assumption 1. *Let the GAN generator function be locally Lipschitz. We say that g satisfies this assumption for a certain probability p over $\mathcal{Z}$ if there are local Lipschitz constants $L(\theta, z)$ such that*
  $$E_{z\sim p}[L(\theta, z)] < \infty.$$
  That is, the expected value of the local Lipschitz constant is well defined.

- We are now ready for the main result.

  Theorem 2. *Let P_r be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g. Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathcal{R}^d \to \mathcal{X}$ be a function denoted by $g_\theta(z)$ where $z \in \mathcal{Z}$ and $\theta \in \mathbb{R}^d$. Let P_θ denote the distribution of $g_\theta(Z)$. Then,*

  1. *If g is continuous in θ, then so is $W(P_r, P_\theta)$.*
  2. *If g is locally Lipschitz and satisfies Assumption 1, then $W(P_r, P_g)$ is continuous everywhere, and differentiable almost everywhere.*
  3. *The previous two statements are false for JS and KL.*

  Corollary 3. *Let g_θ be any feedforward neural network parameterized by θ, and $p(z)$ a prior over z such that $E_{z\sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.). Then Assumption 1 is satisfied and therefore $W(P_r, P_\theta)$ is continuous everywhere and differentiable almost everywhere.*

- The following theorem states that EM is weaker than all the other distances.

  Theorem 4. *Let P be a distribution on a compact space $\mathcal{X}$ and $(P_n)_{n\in\mathbb{N}}$ a sequence of distribution on $\mathcal{X}$. Then, considering all limits as $n \to \infty$:*

  1. *The following statements are equivalent:*
     - $\delta(P_n, P) \to 0$.
     - $JS(P_n, P) \to 0$.
  2. *The following statements are equivalent:*
     - $W(P_n, P) \to 0$.
     - $P_n \to P$ *in the sence of convergence in distribution for random variables.*
  3. $KL(P_n \| P) \to$ *or* $KL(P \| P_n) \to 0$ *implies the statement in 1.*
  4. *The statements in 1 imply the statements in 2.*

3 Wassersteing GAN

- We now consider the problem of incorporating the Wasserstein distance into GAN training. The first problem is how we might evaluate it.

- The Kantorovich–Rubinstein duality theorem gives another formula for the Wasserstein distance:
  $$W(P_r, P_\theta) = \sup_{\|f\|_L \le 1} E_{x\sim P_r}[f(x)] - Ex \sim P_\theta[f(x)] \tag{1}$$
  where the supermum is over all the 1-Lipshitz functions $f : \mathcal{X} \to \mathcal{R}$.

- Recall that a real-valued function f is K-Lipshitz if
$$|f(x_1) - f(x_1)| \leq K d_{\mathcal{X}}(x_1, x_2)$$
for all $x_1, x_2 \in \mathcal{X}$.
- In Equation (1), if we replace $\|f\|_L \leq 1$ with $\|f\|_L \leq K$, then the LHS becomes $K \cdot W(P_r, P_\theta)$.
- To approximate the supremum in Equation (1), we can prepare a family of K-Lipschitz functions $\{f_w\}_{w \in \mathcal{W}}$. Then, we can compute:
$$\max_{w \in \mathcal{W}} E_{w \sim P_r}[f_w(x)] - E_{z \in p(z)}[f_w(g_\theta(z))].$$
- To optimize θ, we can also find the gradient of the expression with respect to θ by estimating:
$$E_{z \sim p(z)}[\nabla_\theta f_w(g_\theta(z))].$$
- **Theorem 5.** *Let P_r be any distribution. Let P_θ be the distribution of $g_\theta(Z)$ with Z a random variable with density p and g_θ a function satisfying Assumption 1. Then, there is a solution $f : \mathcal{X} \to \mathcal{R}$ to the problem*
$$\max_{\|f\|_L \leq 1} E_{x \sim P_r}[f(x)] - E_{x \sim P_\theta}[f(x)]$$
and we have
$$\nabla_\theta W(P_r, P_\theta) = -E_{z \quad p(z)}[\nabla_\theta f(g_\theta(z))]$$
when both terms are well-defined.
- So, how do we find such an f in the theorem? We can have f_w be a neural network parametermized with weight w lying in a compact space $\mathcal{W}$ to maximize $E_{x \sim P_r}[f_w(x)] - E_{x \sim P_\theta}[f_w(x)]$.
- Note that, since $\mathcal{W}$ is a compact space, all function f_w will be K-Lipshitz for some constant K. So, maximizing the function would approximate the supermum upto a scaling factor and the capacity of the function f_w. (The paper calls f_w the "critic," not the discriminator.)
- To make sure that w lies in a compact space, we can clamp the weights to a fixed box, say $\mathcal{W} = [-0.01, 0.01]^l$, after each gradient update.
- The WGAN algorithm builds on the above idea. It requires the following parameter:
    - α is the learning rate. The paper uses $\alpha = 5 \times 10^{-5}$.
    - c is the weight clamping constant. The paper uses $c = 0.01$.
    - m is the minibatch size. The paper uses $m = 64$.
    - n_{critic} is the number of times to update the critic's weight for each generator weight update. The paper uses $n_{critic} = 5$.

  The algorithm is as follows:

```
while $\theta$ has not converged do
    for $t = 0$ to $n_{critic}$ do
        Sample a batch $\{x^(i)\}_{i=1}^{m} \sim P_r$ of real data.
        Sample a batch $\{z^(i)\}_{i=1}^{m} \sim p(z)$ of latent vectors.
```

```
      Compute the gradient for the critic weight:
          $$g_w \leftarrow \nabla_w \left[ \frac{1}{m} \sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)})) \right]$$
      $w \leftarrow$ Weight-Update$(w, g_w, \alpha)$
    end for
    Sample a batch $\{z^(i)\}_{i=1}^{m} \sim p(z)$ of latent vectors.
    Compute the gradient of the generator parameter:
        $$g_\theta \leftarrow -\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$$
    $\theta \leftarrow$ Weight-Update$(\theta, \theta_w, \alpha)$
end while
```

4 Problems with Weight Clipping

- It is later found that weight clipping in WGAN leads to optimization problems [Gulrajani et al., 2017]. Other weight contraints such as L2 norm clipping, weight normalization, and soft contraints such as L1 nad L2 weight decay, also exhibit similar problems. The problems are
  - Capacity underuse.
  - Exploding and vanishing gradients.

4.1 Capacity Underuse

- Weight clipping biases the critic towards mcuh simpler functions.

  Proposition 6. *Let P_r and P_g be two distributions in a compact metric space $\mathcal{X}$. Let f^* be the 1-Lipschitz function which is the optimal solution of* $\max_{\|f\|_L \leq 1} E_{y \sim P_r}[f(y)] - E_{x \sim P_g}[f(x)]$. *Let π be the optimal joint distribution between P_r and P_g. If f^* is differentiable, $\pi(x = y) = 0$, and $x_t = tx + (1-t)y$ with $0 \leq t \leq 1$, it holds that*

  $$P_{(x,y) \sim \pi} \left[\nabla f^*(x_t) = \frac{y - x_t}{\|y - x_t\|} \right] = 1.$$

  Corollary 7. *f^* has gradient norm 1 almost everywhere under P_r and P_g.*

  In other words, the optimal WGAN critic has unit gradient norm almost everywhere.

- Experimentally, it is observed that the critic will end up being very simple functions.

4.2 Explding and Vanishing Gradients

- The authors of [Gulrajani et al., 2017] trained WGAN on Swiss Roll toy dataset, varying the clipping threshold c to values in $[10^{-1}, 10^{-2}, 10^{-3}]$, and plot the norm fo the gradient of the critic loss with respect to the layers in the network. They found that the gradient either grows or decays exponentially as they move back through the network.

5 Gradient Penalty

- A differentiable function is 1-Lipschitz if and only if it has gradients with norm at most 1 everywhere.
- A way to enforce this is to constrain the gradient norm of the critic with respect to the inputs (not the weights). This is done through having a panelty term on the magnitude of the gradient:

  $$L = E_{x \sim P_g}[f(x)] - E_{x \sim P_r}[f(x)] + \lambda E_{x \sim P_x}[(\|\nabla_x f(x)\|_2 - 1)^2].$$

  Enforcing the gradient to be 1 everywhere is nigh impossible. So, the probability distribution P_x samples uniformly along straight lines between pairs of points sampled from the data distribution P_r and the generator distribution P_g. This is motivated by Proposition 6.
- All experiments in the paper uses $\lambda = 10$.
- The paper does not use batch normalization in the critic because it changes the form of the critic function from one input to one output to one input batch to one output batch. The method works with any normalization techniques that introduces correlation between the samples.
- The lost encourages the norm to go towards 1 instead of just staying below 1. Empirially, this does not contrain the function too much because the optiomal WGAN critic has gradient norm of 1 almost everywhere.

References

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *CoRR*, abs/1704.00028.