Learning Generative Models

- We assume data comes from some unknown distribution $P_r$.
- We want to learn a distribution $P_\theta$, where $\theta$ are the parameters of the distribution.
- One approach is to learn a function $g_\theta$ that transforms an existing distribution $Z$ into $P_\theta$. Here, $g_\theta$ is some differentiable function, $Z$ is usually a Gaussian distribution, and $P_\theta = g_\theta(Z)$.
- To train $g_\theta$ (and by extension $P_\theta$), we need a measure of distance between distributions.
Different distances

Consider a probability distributions defined over $R^2$ where the true data
distribution is $(0, y)$, with $y$ sampled uniformly from $U[0,1]$ (right figure).
We try to model it with the family of distributions $P_\theta$, where
$P_\theta = (\theta, y)$, with $y$ also sampled from $U[0,1]$.

Below are some distance measures:

- The Total Variation (TV) distance is
  \[ \delta(P_r, P_g) = \sup_A |P_r(A) - P_g(A)| \]

- The Kullback-Leibler (KL) divergence is
  \[ KL(P_r || P_g) = \int_x \log \left( \frac{P_r(x)}{P_g(x)} \right) P_r(x) \, dx \]
  This isn't symmetric. The reverse KL divergence is defined as $KL(P_g || P_r)$.

- The Jenson-Shannon (JS) divergence: Let $M$ be the mixture distribution $M = P_r/2 + P_g/2$.
  Then
  \[ JS(P_r, P_g) = \frac{1}{2} KL(P_r || P_m) + \frac{1}{2} KL(P_g || P_m) \]

Wasserstein Distance

The Wasserstein distance is the *minimum cost* of transporting mass in converting the data
distribution $P_g$ to the data distribution $P_r$.

\[ W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}(x,y) \sim \gamma \left[ \|x - y\| \right] \]

$\Pi(P_r, P_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $P_r$
and $P_g$.

---

**Fig. 1**: Probability distribution $P_r$ and $P_\theta$, each with ten states.
Earth Mover’s Distance

Move boxes from left configuration to the configuration in the right.

- Moving cost = weight*distance.
- Box 1 is moved from location 1 to 7, so distance is 6.
- Move 2 boxes from location 3 to 10 and the entry γ(3, 10) is therefore set to 2. Cost is 2x7 = 14.

Examples of Transport Plans

- Wasserstein Distance is the cost for the cheapest plan.
- Π contains all the possible transport plan γ.
- The amount of mass that leaves x is ∫γ(x, y) dy and it must equal amount of mass originally at x i.e. Pr(x).
- The amount of mass that enters y is ∫x γ(x, y) dx and it must equal amount the amount of mass that ends up at y i.e. Pg(y).
What is the Earth Mover’s Distance in our example?

Because the two distributions are just translations of one another, the best way transport plan moves mass in a straight line from $(0, y)$ to $(\theta, y)$. This gives: $W(P_0, P_\theta) = |\theta|$

- There exist sequences of distributions that don’t converge under the JS, KL, reverse KL, or TV divergence, but which do converge under the EM distance.
- For the JS, KL, reverse KL, and TV divergence, there are cases where the gradient is always 0. EM plot is continuous and provides usable gradient everywhere.

Figure 1: These plots show $\rho(P_0, P_\theta)$ as a function of $\theta$ when $\rho$ is the EM distance (left plot) or the JS divergence (right plot). The EM plot is continuous and provides a usable gradient everywhere. The JS plot is not continuous and does not provide a usable gradient.

EM distance vs others in our example

Real and fake distribution when $\theta = 1$
**Wasserstein GAN**

Computing Wasserstein distance exactly is intractable:

\[
W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} \left[ \|x - y\| \right]
\]

How do we approximate this?

1. A result from Kantorovich-Rubinstein duality shows \( W \) is equivalent to:

\[
W(P_r, P_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_\theta}[f(x)]
\]

(where the supremum is taken over all 1-Lipschitz functions.)

2. If we replace the supremum over 1-Lipschitz functions with the supremum over \( K \)-Lipschitz functions, then the supremum is \( K \cdot W(P_r, P_\theta) \) instead. The supremum over \( K \)-Lipschitz functions \( \{f : \|f\|_L \leq K\} \) is still intractable.

3. Suppose we have a parametrized function family \( \{f_w\}_{w \in \mathcal{W}} \), where \( w \) are the weights and \( \mathcal{W} \) is the set of all possible weights and suppose these functions are all \( K \)-Lipschitz functions for some \( K \). Then we have:

\[
\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim P_r}[f_w(x)] - \mathbb{E}_{x \sim P_\theta}[f_w(x)] \leq \sup_{\|f\|_L \leq K} \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_\theta}[f(x)]
\]

\[
= K \cdot W(P_r, P_\theta)
\]

**Training in WGAN**

- Given a fixed \( g_\theta \), we can compute the optimal \( f_w \) for the Wasserstein distance. We can then backprop through \( W(P_r, g_\theta(Z)) \) to get the gradient for \( \theta \).

\[
\nabla_\theta W(P_r, P_\theta) = \nabla_\theta \left( \mathbb{E}_{x \sim P_r}[f_w(x)] - \mathbb{E}_{z \sim Z}[f_w(g_\theta(z))] \right)
\]

\[
= -\mathbb{E}_{z \sim Z}[\nabla_\theta f_w(g_\theta(z))]
\]

- Training Process has three steps:
  - For a fixed \( \theta \), compute an approximation of \( W(P_r, P_\theta) \) by training \( f_w \) (Discriminator) to convergence.
  - Once we find the optimal \( f_w \), compute the \( \theta \) gradient \( -\mathbb{E}_{z \sim Z}[\nabla_\theta f_w(g_\theta(z))] \) by sampling several \( z \sim Z \).
  - Update \( \theta \), and repeat the process.
- This entire derivation only works when the function family \( \{f_w\}_{w \in \mathcal{W}} \) is \( K \)-Lipschitz. The weights \( w \) are constrained to lie within \([-c, c]\), by clipping \( w \), after every update for \( w \).
WGAN Algorithm

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: $\alpha$, the learning rate. $c$, the clipping parameter. $m$, the batch size.
          $n_{\text{critic}}$, the number of iterations of the critic per generator iteration.

Require: $w_0$, initial critic parameters. $\theta_0$, initial generator’s parameters.

1: while $\theta$ has not converged do
   2: for $t = 0,...,n_{\text{critic}}$ do
      3: Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.
      4: Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
      5: $g_w \leftarrow \nabla_w \left[ \frac{1}{m} \sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_w(g_{\theta}(z^{(i)})) \right]$
      6: $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
      7: $w \leftarrow \text{clip}(w, -c, c)$
   8: end for
   9: Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
  10: $g_{\theta} \leftarrow -\nabla_{\theta} \frac{1}{m} \sum_{i=1}^m f_w(g_{\theta}(z^{(i)}))$
  11: $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_{\theta})$
12: end while

Experiment to showcase difference between GAN and WGAN

- There are two 1D Gaussian distributions (real and fake).
- Train a GAN discriminator and WGAN critic to optimality and plot their values.

WGAN gives a reasonably nice gradient over everything!
Wasserstein loss seems to correlate well with image quality.

JSD loss does not correlate with image quality
Improved Stability

- If we remove batch norm from the generator, WGAN still generates okay samples, but DCGAN fails completely.
- No evidence of **Mode Collapse** in Author’s experiment

Problems with WGAN

- The difficulty in WGAN is to enforce the Lipschitz constraint.
- WGAN may still result in bad quality images and doesn’t converge, especially when the hyperparameter ‘c’ is not tuned correctly.
- It results in either vanishing or exploding gradients.
- Weight clipping biases the critic towards much simpler Functions.
- Given a fixed critic architecture and fixed `c` for clamping, can we quantitatively compare different generators by computing the Wasserstein estimate of both?
- Because of the wide range of performance under different hyperparameters, hyperparameter tuning is particularly important for any cost functions, and therefore it will have a better return of investments.
References


https://medium.com/@sunnerli/the-story-about-wgan-784be5acd84c