CS 329M: Topics in Artificial Intelligence **Spring 2017**

Lecture 9: June 5

Lecturer: James Zou *Scriber: Qijia Jiang*

In this note, we discuss generative models with adversaries.

9.1 Different Measures

Say we have true distribution P_r (real distribution for data) and we want to somehow approximate it with P_θ, i.e. distribution parameterized by θ. A natural question to ask is what metric to use for measuring closeness between two distributions. Below we present several potential choices.

- KL divergence, defined as

$$\mathrm{KL}(P_r||P_\theta) = \int P_r(x)\log P_r(x)dx - \int P_r(x)\log P_\theta(x)dx$$

  To find an appropriate θ, we simply pick θ that minimizes the KL divergence, which essentially reduces to the MLE estimator in this case

$$X_i \sim P_r \Rightarrow C - \frac{1}{m}\sum_i \log P_\theta(x_i)$$

- Total variation distance

$$\delta(P_r, P_\theta) = \frac{1}{2}\int |P_r(x) - P_\theta(x)|\, dx$$

- Earthmover distance / Wasserstein distance

$$W(P_r, P_\theta) = \inf_\Gamma \int P_r(x)dx \int \Gamma(y|x)\|y-x\|_2 dy$$

To compare these 3 metrics, below we illustrate with an example.

Example: Let P_r, P_1 and P_2 be 3 different distributions with $P_r \sim U[0,1], P_1 \sim U[1,2], P_2 \sim U[2,3]$.

In this case, we have

- $\mathrm{KL}(P_r||P_1) = \infty = \mathrm{KL}(P_r||P_2)$ due to the disjoint support
- $\delta(P_r, P_1) = 1 = \delta(P_r, P_2)$ follows from a simple calculation of ℓ_1 distance
- $W(P_r, P_1) = 1 \neq 2 = W(P_r, P_2)$

More generally, if $P_\epsilon \sim U[\epsilon, \epsilon+1]$, we have $\forall \epsilon \geq 0$,

$$\mathrm{KL}(P_r||P_\epsilon) = \infty, \quad \delta(P_r, P_\epsilon) = 1, \quad W(P_r, P_\epsilon) = \epsilon$$

Now there's another definition of the Earthmover distance, which can be viewed as a dual version of the definition presented above

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

where $\|\cdot\|_L$ denotes the Lipschitz constant of function. Here we have $\forall x, y$,

$$|f(x) - f(y)| \le \|x - y\|_2$$

Theorem 1 *Let P_r be a distribution on compact set χ, let $\{P_h\}_{h\in\mathbb{H}}$ be a set of distributions (think sequence of approximations), then as $h \to \infty$,*

$$KL(P_h||P_r) \to 0 \overset{(1)}{\Longrightarrow} \delta(P_h, P_r) \to 0 \overset{(2)}{\Longrightarrow} W(P_h, P_r) \to 0$$

Remark. Converse doesn't hold. Example above provides an counter-example.

Proof Sketch: (1) First claim essentially follows from Pinsker's inequality,

$$\delta(P_h, P_r) \le \sqrt{2\mathrm{KL}(P_h||P_r)}$$

(2) Second claim can be reasoned through plot. (Look at the pdf's of the 2 distributions, l_1 distance goes to 0 implies EMD goes to 0 as well.)

9.2 Distance Computation

In the previous section, we showed that EMD in some sense is a better measure due to its "sensitivity". In this section, we show how to make the definition "operational", i.e. how to compute and optimize it efficiently.

Let $\Omega = \{f : \|f\|_L \le 1\}$ and f' be the sup function, now

$$X \sim P_\theta \Longleftrightarrow \text{sample } z \sim P(Z),\ X = g_\theta(Z)$$

where $P(Z)$ can simply be taken as $N(0, I)$ and $g_\theta(\cdot)$ is a nonlinear function parameterized by θ.

Going back to the definition of EMD, we can rewrite it as

$$W(P_r, P_\theta) = \mathbb{E}_{X\sim P_r}[f'(X)] - \mathbb{E}_{Z\sim P}[f'(g_\theta(Z)]$$

However, we don't have access to f'. But we can approximate Ω by $\{f_w : \|f_w\| \le 1\}$, where neural network weights are used to parameterize the functions, and the condition imposed on the norm of the weights ensures that we get a smooth function.

This way, the objective becomes

$$\min_\theta \max_w \quad \mathbb{E}_{X\sim P_r}[f_w(X)] - \mathbb{E}_{Z\sim P}[f_w(g_\theta(Z)]$$

9.3 Towards an Actual Algorithm (WGAN)

Consider the following algorithm.

Input: $\{X^{(i)}\} \sim P_r$
Output: θ

Outerloop: optimize θ
  Innerloop: optimize w
    Sample batch $\{X^{(i)}\}$
    Sample $\{Z^{(i)}\} \sim P(Z) = N(0, I)$
    Compute $\nabla_w[\frac{1}{m}\sum f_w(x^{(i)}) - \frac{1}{m}\sum f_w(g_\theta(z^{(i)}))]$
    Update w: $w \leftarrow$ threshold w (to ± 0.01 say)
  Sample $\{Z^{(i)}\} \sim N(0, I)$
  Compute $-\nabla_\theta \frac{1}{m}\sum f_w(g_\theta(z^{(i)}))$
  Update θ

Note the inner-loop here is essentially computing an approximation to EMD. The other advantage of using EMD as metric is its differentiability.

9.4 Some Examples & Extensions

There are variations of this algorithm that uses other notions of distance like cross-entropy. And the objective becomes

$$\min_\theta \max_w \quad \mathbb{E}_{X \sim P_r}[\log f_w(X)] - \mathbb{E}_{Z \sim P}[\log(1 - f_w(g_\theta(Z))]$$

There exists natural interpretation of this as a 2-player minimax game between a generator and a discriminator. There are also connections to reinforcement learning and robust algorithm in general.

Here are some pictures showing the examples generated by WGAN [Arjovsky, Chintala, Bottou '17]. Most of them look reasonable if zoomed out.

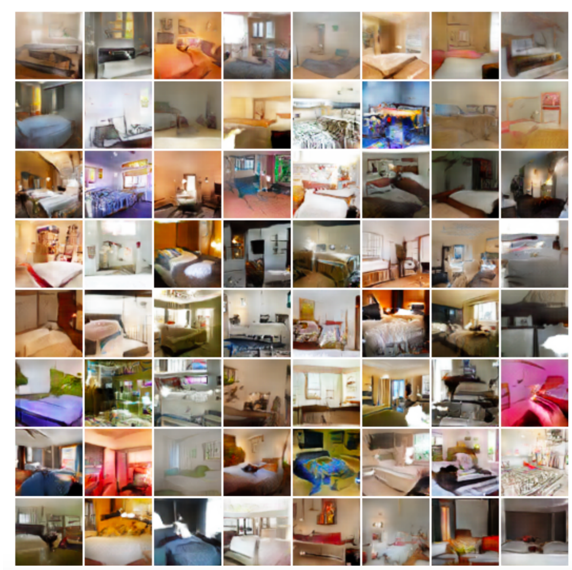

Figure 9.1: WGAN examples