Class 10

Shikhar Saxena

February 20, 2023

Contents

Reference: [WGAN-Paper](#)

Brief Review of Duality

- Lagrangian $L(x, \lambda, \nu)$
- Dual Function $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$
- $\sup_{\lambda, \nu} \ g(\lambda, \nu) = d^*$
- $\min_x \ f_0(x) = p^*$

---

1. Weak Duality: $d^* \leq p^*$
2. Strong Duality: $d^* = p^*$

Proof.

$$\sup_{\lambda \geq 0} L(x, \lambda) = \sup_{\lambda \geq 0} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \right)$$

We know $\lambda_i \geq 0$, $f_i(x) \leq 0 \ \forall i$

$$= \begin{cases} f_0(x), & f_i(x) \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

Otherwise case arises when x is not feasible and some $f_i(x) > 0$.

Thus, we see that

$$p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda)$$

And we know that

$$d^* = \sup_\lambda \inf_x L(x, \lambda)$$

Thus weak duality holds. $\qquad \square$

Issues with KL (or f) Divergence

The problem with KL and f-divergences is that they are dependent ont the supports of the densities.

$supp(f) = \{x \in dom(f), \ f(x) \neq 0\}$

If the densities don't have the same support, then KL (or f) divergence will go to ∞ (arises from the non-overlapping supports).

Thus, we can't really compare/measure distributions that have non-overlapping supports using KL or f-divergence.

Remedy: Smoothing

We add a noise ϵ to the input vector. This noise is supposed to have high bandwidth and is added to both distributions so as they have overlapping supports. (It should be minimum density so as to not change the original densities a lot).

This is disadvantageous as we ourselves are introducing error and hence degrading the quality of the samples.

Some More Measures

- Total Variation (TV) Measure

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

- Wasserstein Distance (W-1) or Earth Mover (EM) Distance

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

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

Intuitively, $\gamma(x, y)$ indicates how much "mass" must be transported from x to y in order to transform the distributions P_r into the distribution P_g. The EM distance then is the <u>cost</u> of the optimal transport.

Variational Approach on EM-Distance

Estimation of $\inf_{\gamma \in \Pi(P_r, P_g)} E_{(x,y)\sim\gamma} \left[\, \|x - y\| \,\right]$ is **intractable**.

Here, we use duality to get a variational form. We reformulate EM-distance as a solution of max over 1-Lipschitz functions.

Definition 1 (L-Lipschitz). *Function $f : A \to B$ is L-Lip if*

$$|f(x) - f(y)| \le L\|x - y\| \quad \forall x, y \in A$$

We use Kantorovich-Rubinstein Duality here.

Essentially, EM-distance is transformed to the following variational form:

$$W(P_r, P_g) = \sup_{\|h\|_L \le 1} \left[E_{x\sim P_r}[h(x)] - E_{x\sim P_g}[h(x)] \right]$$

Refer the proof of KR Duality here.

WGAN

Consider a parameterized family of functions $\{h_w\}$, $w \in W$ are all K-Lipschitz (we generalize the 1-Lipschitz condition to K-Lipschitz condition here).

The Problem being solved here thus becomes:

$$\inf_{P_g} W(P_r, P_g) = \inf_{\theta} \max_{w \in W} E_{x\sim P_r}[h_w(x)] - E_{z\in p(z)}[h_w(g_\theta(z))]$$

The first expectation is called the **critic** while the second expectation is called the **generator**.