Duality for f-Divergences

Primer: Convex Conjugates

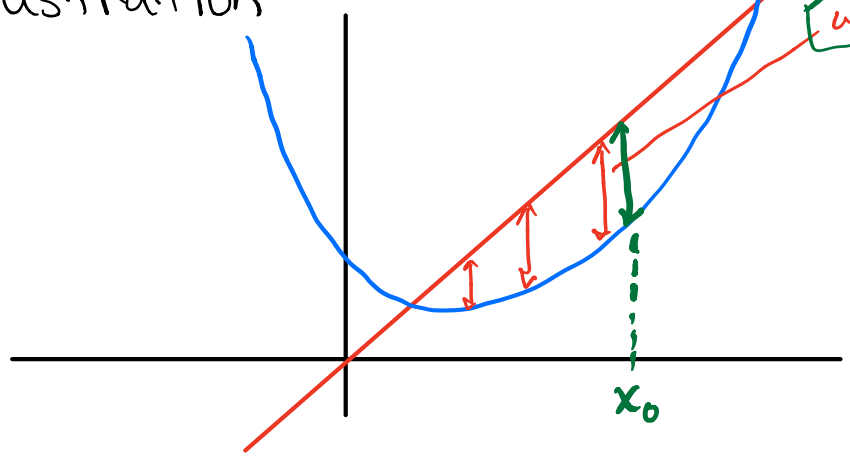
Definition (Convex Conjugate): Let $f: I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an interval, be a convex function. The convex conjugate of f is another function $f^*: I^* \to \mathbb{R}$ given by

$$f^*(y) = \sup_{x \in I} (y \cdot x - f(x))$$

where

$$I^* := \left\{ y \in \mathbb{R} \mid \sup_{x \in I} (yx - f(x)) < \infty \right\}$$

Illustration



Proposition (properties of convex conjugate)
The convex conjugate f^* of f satisfies

(i) f^* is continuous in the domain
(ii) f^* is convex
(iii) biconjugation $(f^*)^* = f$

Duality:

Recall

$$D_f(P\|Q) = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}(x)\right) dQ(x)$$

and consider f^* the convex conjugate of f above.

By property (iii),

$$f(x) = \sup_{y \in I^*} (xy - f^*(y))$$

$$\Rightarrow \quad D_f(P\|Q) = \int_{\mathcal{X}} f\left(\underbrace{\frac{dP}{dQ}(x)}_{\text{"x in above expression"}}\right) dQ(x)$$

$$= \int_{\mathcal{X}} \sup_{y \in I^*} \left(\frac{dP}{dQ}(x) \cdot y - f^*(y)\right) dQ(x)$$

pick any $g: \mathcal{X} \to I^*$ and set $y = g(x)$ to get a lower bound

$$\geq \int_{\mathcal{X}} \left(\frac{dP}{dQ}(x) \cdot g(x) - f^*(g(x))\right) dQ(x)$$

$$= \int_{\mathcal{X}} g(x) \frac{dP}{dQ}(x)\, dQ(x) - \int_{\mathcal{X}} f^*(g(x))\, dQ(x) = \mathbb{E}_P[g(x)] - \mathbb{E}_Q[f^*(g(x))]$$

$\Rightarrow$ (supremizing over all measurable $g: \mathcal{X} \to \mathcal{I}^*$)

$$D_f(P\|Q) \geq \sup_{g:\mathcal{X}\to\mathcal{I}^*} \mathbb{E}_P[g(x)] - \mathbb{E}_Q[f^*(g(x))]$$

It can be shown that the above lower bound is tight and achieved by

$$g(x) = f'\left(\frac{dP}{dQ}(x)\right)$$

where f' is the 1^{st} order derivative of f.

Theorem (f-divergence duality)
For any f-divergence, we have

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(x)] - \mathbb{E}_Q[f^*(g(x))]$$

where the sup is over all g for which both expectations are finite.

Examples

① KL: $f(x) = x \log x \implies f^*(y) = e^{y-1}$

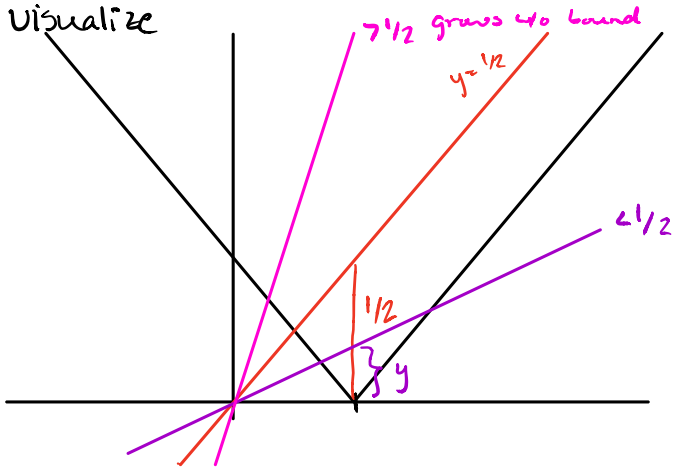
$$D_{KL}(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} 1 + \mathbb{E}_P[g(x)] - \mathbb{E}_Q[e^{g(x)}]$$

② TV

$$f(x) = \frac{1}{2}|x-1|$$

$$\Rightarrow f^*(y) = \begin{cases} y & |y| \le 1/2 \\ \infty & \text{o/w} \end{cases}$$

Visualize



$$\Rightarrow \delta_{TV}(P,Q) = \sup_{g: X \to \mathbb{R}} \mathbb{E}_P[g(x)] - \mathbb{E}_Q[g(x)]$$

$$\|g\|_\infty \le \frac{1}{2}$$

Generative Modeling

Generative modeling is an unsupervised learning task, where we are given unlabeled data

$$\{x_i\}_{i=1}^n \overset{iid}{\sim} P \in \mathcal{P}(\mathbb{R}^d)$$

and we are trying to learn some underlying structure (cluster / dim. reduction / estimate for P).

For us, we will aim to use our samples to "learn" a model Q_θ, from a class $\{Q_\theta\}_{\theta \in \Theta}$, where $\Theta \subseteq \mathbb{R}^{d'}$, such that $Q_\theta \approx P$.

Note: By "learning" Q_θ we do NOT require to explicitly know it, but we do want at the very least to be able to sample from it.

The state of the art systems for learning such "sampleable" generative models are Generative Adverserial Networks (GANs).

- GANS:

- Resources:
  (i) Real samples $\{X_i\}_{i=1}^n \overset{iid}{\sim} P \in \mathcal{P}(\mathbb{R}^d)$
  (ii) Noise: $Z \sim N(0, \sigma^2 I_{d_0})$, $d_0 \ll d$

- Generator: Z is reshaped through a parametrized (by DNN) function $g_\theta : \mathbb{R}^{d_0} \to \mathbb{R}^d$.
  - We will think about the generator's output as our generators output as our generated/synthesized/"fake" example
  - We denote the law of $g_\theta(Z)$ by Q_θ

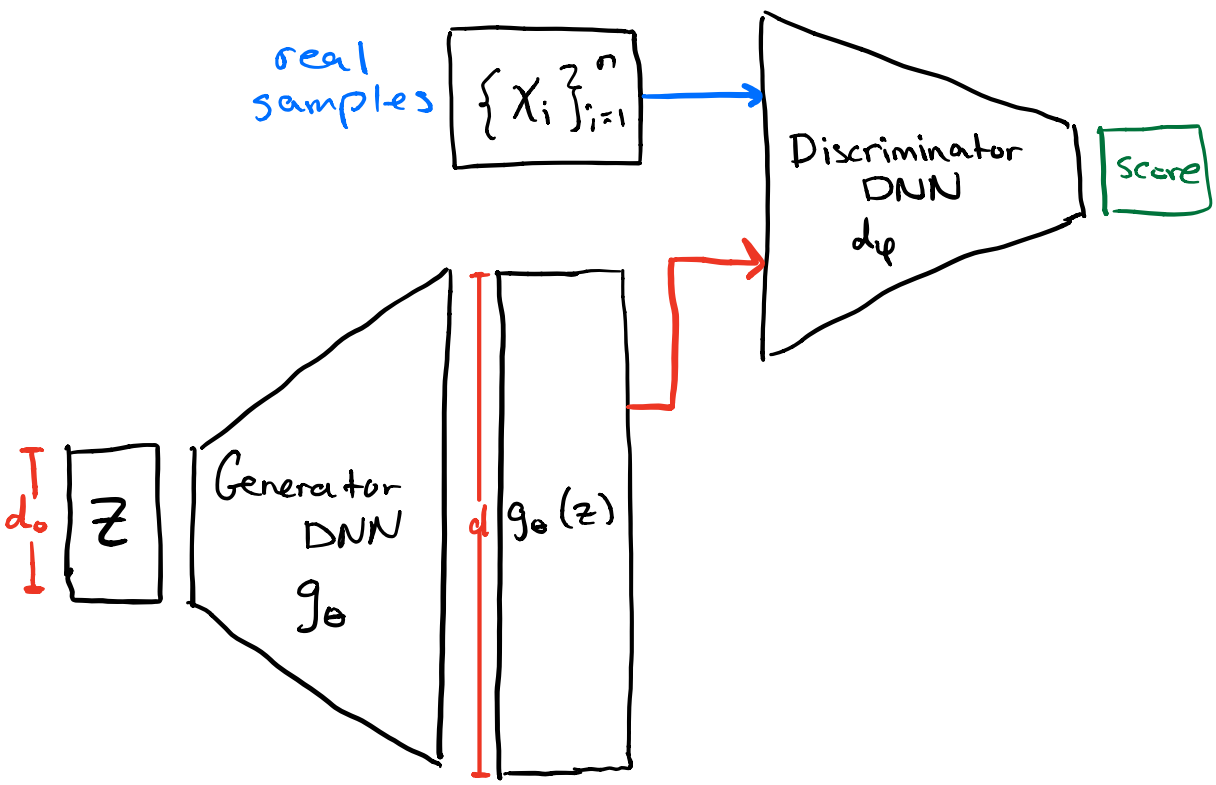
- **Discrimination** 2^{nd} DNN $d_\varphi : \mathbb{R}^d \to \mathbb{R}$ that takes in both "real" and "fake" samples and tries to tell them apart.

- **Optimization**: By iteratively optimizing g_θ and d_φ via an alternating optimization procedure, once converged we obtain a generator that "is able to fool even the best discriminator possible".

- Objective Example

$$\inf_{\theta \in \Theta} \sup_{\varphi \in \Phi} \mathbb{E}\left[d_\varphi(x)\right] - \mathbb{E}\left[d_\varphi(g_\theta(z))\right]$$

Block Diagram



- Principle Generative Modeling Objective:

Pick an f divergence D_f and solve $\inf_{\theta \in \Theta} D_f(P \| Q_\theta)$

Plugging in the dual form of D_f into the above recovers the min max game formulation of GANs.

Example

$$\inf_{\theta \in \Theta} \delta_{TV}(P, Q_\theta) = \inf_{\theta \in \Theta} \sup_{\|d\|_\infty \leq \frac{1}{2}} \mathbb{E}[d(X)] - \mathbb{E}[d(g_\theta(Z))]$$

$$\simeq \inf_{\theta \in \Theta} \sup_{\substack{\varphi \in \Phi \\ \|d_\varphi\|_\infty \leq \frac{1}{2}}} \mathbb{E}[d(X)] - \mathbb{E}[d(g_\theta(Z))]$$

Concluding Remarks

(i) GANs are very useful in practice but hard to study theoretically

(ii) $\inf_\theta D_f(P \| Q_\theta)$ is a well-posed math question that is useful for studying sample complexity, feasability, etc.

(iii) Wasserstein GAN (2017): $\delta = W_1$

the 1-Wasserstein distance, start from $\inf_{\theta \in \Theta} W_1(P, Q_\theta)$ and

end up with an implementable min-max game, and get the "best" GAN constructions known to date.