APPLICATION OF O.T.

- DEMYSTIFYING GANS;
- APPROXIMATING $KL(p_\theta \| q)$
- APPROXIMATING $W_p^p(\mathbb{P}_\theta \| \mathbb{Q})$

NOTATION
p_θ DENSITY OF $\mathbb{P}$, "ν"
q DENSITY OF $\mathbb{Q}$, "μ"

EXPERIMENT SETUP

- WE OBSERVE DATA $x \sim q(x)$
- p_θ APPROXIMATES q

$$\boxed{\begin{array}{ccc} KL(p_\theta \| q) & \text{v.s.} & KL(q \| p_\theta) \\ & & \parallel\!| \\ \mathbb{E}_{p_\theta}\left[\log \frac{p_\theta}{q}\right] & \text{v.s.} & \mathbb{E}_q\left[\log \frac{q}{p_\theta}\right] \end{array}}$$

THE CLASSIC EXAMPLE (USUALLY CONDITIONALS)

- OBSERVE $\{(x_i, y_i)\}_{i=1}^N$
- FIND $p_\theta(y|x)$ TO APPROXIMATE $q(y|x)$

USUALLY WE PICK $KL(q \| p_\theta)$

THUS $\mathbb{E}_q\left[\log \frac{q}{p_\theta}\right] = \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p_\theta]$

$$\theta^* = \arg\min_\theta \mathbb{E}_q\left[\log \frac{q}{p_\theta}\right] = -\arg\min_\theta \mathbb{E}_q[\log p_\theta]$$

$$= \arg\max_\theta \mathbb{E}_q[\log p_\theta] \approx \sum_{i=1}^N \log p_\theta(x_i) \equiv \sum_{i=1}^N \log \mu(\delta x_i)$$

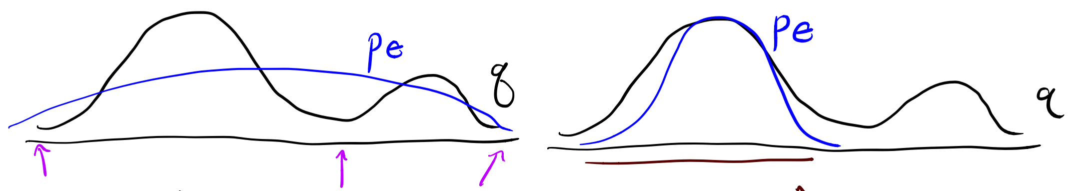
$p_\theta(x_i)$ IS EASY. USUALLY $x_i = (\tilde{x}_i, \tilde{y}_i)$;

$$p_\theta(\tilde{X}, \tilde{Y}) \underset{\text{INDEPENDENT}}{=} \prod_{i=1}^N p_\theta(\tilde{x}_i, \tilde{y}_i) \underset{\text{"PREDICTIVE TASK"}}{=} \prod_{i=1}^N p_\theta(\tilde{y}_i \mid \tilde{x}_i) \simeq$$

$$\underset{\text{LOSS FUNCTION CHOICE}}{=} \prod_{i=1}^N \left[f_\theta(x_i)\right]^{y_i}\left[1 - f_\theta(x_i)\right]^{1-y_i} = \underset{\text{PYTORCH CODE}}{\text{nn.BinaryCrossEntropy}}$$

TLDR; THIS IS STANDARD AND common.

IMPACT OF $KL(q \| p_\theta)$ V.S. $KL(p_\theta \| q)$



$KL(q_\theta \| p)$ COVERS UNLIKELY REGIONS ($\Rightarrow$ BAD SAMPLES)

$KL(p \| q_\theta)$ STAYS CONTAINED TO SUPPORT OF q.

RE-WRITING $KL(q_\theta \| p)$

$$\theta^* = \arg\min_\theta \mathbb{E}_{x \sim p_\theta}\left[\log \frac{p_\theta(x)}{q(x)}\right] = \arg\max_\theta \mathbb{E}_{x \sim p_\theta}\left[\log \frac{q(x)}{p_\theta(x)}\right]$$

$$= \arg\max_\theta \mathbb{E}_{x \sim p_\theta}\left[\log \frac{q(x)}{p_{\theta^*}(x)}\right] = \arg\max_\theta \mathbb{E}_{x \sim p_\theta}\left[g(x)\right]$$

ASSUME OKAY.

WITH $g(x) = \log \frac{q(x)}{p_{\theta^*}(x)}$ & θ^* AS THE "TRUE" PARAMETER.

OBVIOUSLY, WE DON'T HAVE θ^* SO NO, WE CANNOT COMPUTE $g(x)$.

APPROXIMATING $KL(q_\theta \| p)$ [STANDARD GAN]

$$\operatorname*{argmin}_\theta KL(p_\theta \| q) = \operatorname*{argmax}_\theta \mathbb{E}_{x \sim p_\theta}\left[g(x) \right]$$

ISSUES

1.) HOW DO WE SAMPLE $x \sim p_\theta$?

2.) HOW DO WE COMPUTE $g(x)$?

SOLUTIONS

(REPARMETERIZATION TRICK)

1.) SIMULATE $x = f_\theta(z)$ WITH $z \sim p(z)$. THEN $x \sim p_\theta(x)$.

$$\theta^* = \operatorname*{argmax} \mathbb{E}_{z \sim p(z)}\left[g(f_\theta(z)) \right]$$

2.) APPROXIMATE g WITH g_ϕ

HOW TO TRAIN g_ϕ?

SINCE g IS THE THEORTICALLY OPTIMAL CLASSIFIER, TRAIN g_ϕ TO CLASSIFY BETWEEN $x \sim p_\theta(x)$ & $x \sim q(x)$.

ISSUES WITH $KL(p_\theta || q)$

- REQUIRES DENSITIES OF $\mathbb{Q}$ & $\mathbb{P}_\theta$ TO EXIST
- REQUIRES OVERLAP OF SUPPORT BETWEEN q & p_θ
- BOTH OF THESE ARE UNLIKELY FOR IMAGES.
- ASSUME WE THINK "$W_p^p(\mathbb{P}_\theta, \mathbb{Q})$" IS BETTER

APPROXIMATING $W_p^p(\mathbb{P}_\theta, \mathbb{P}_r)$

[$\mathbb{Q}$ UPDATED TO $\mathbb{P}_r$; ITS NO BIG DEAL]

$$W_p^p(\mathbb{P}_\theta, \mathbb{P}_r) = \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}_{(x,y)\sim\gamma}\left[c(x,y)\right]$$ IS INTRACTIBLE

$$= \sup_{\substack{\varphi \in L^1(\mathbb{P}_\theta),\ \psi \in L^1(\mathbb{P}_r) \\ \varphi(x)+\psi(y) \leq c(x,y)}} \mathbb{E}_{x\sim\mathbb{P}_\theta}\left[\varphi(x)\right] + \mathbb{E}_{x\sim\mathbb{P}_r}\left[\psi(x)\right]$$

... WITH SOME MATH ...

$$= \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_\theta}\left[f(x)\right] - \mathbb{E}_{x\sim\mathbb{P}_r}\left[f(x)\right]$$

THIS WE CAN COMPUTE DIRECTLY WITH FAKE ($\mathbb{P}_\theta$) AND REAL ($\mathbb{P}_r$) SAMPLES BUT WE NEED TO LEARN f. THUS f IS APPROXIMATE BY f_ϕ, SOME NETWORK WITH WEIGHT CLIPPING. [WASSERSTEIN GAN]