DIVERGENCES

KL, JS, AND WASSERSTEIN 1

FROM GAN TO WGAN

- Ref. Weng, Lilian. "From GAN to WGAN." arXiv preprint arXiv:1904.08994 (2019).

FROM GAN TO WGAN

- 두 분포 간의 유사성 혹은 거리를 측정하는 방법
  - Kullback-Leibler (KL) Divergence
  - Jensen-Shannon (JS) Divergence
  - Wasserstein 1

KL-DIVERGENCE

KULLBACK-LEIBLER (KL) DIVERGENCE

- 어느 분포 p가 다른 분포 q로부터 얼마나 떨어져 있는가
- 어느 분포 p가 다른 분포 q의 정보량을 얼마나 잘 보존하는가
  - 정보량을 잘 보존할 수록 서로 비슷한 분포

KULLBACK-LEIBLER (KL) DIVERGENCE

- 정보이론에서 정보량
  - 놀람의 정도(degree of surprise)
  - $h(x) = -\log p(x)$
  - P(x)는 확률분포에서의 값: 0~1 사이의 실수

KULLBACK-LEIBLER (KL) DIVERGENCE

- 정보이론에서 엔트로피(entropy)
  - 놀람의 정도의 평균(기대값), 불확실성의 정도
  - $Entropy = E[-\log p(x)]$

KULLBACK-LEIBLER (KL) DIVERGENCE

- KL-divergence
  - 상대적인 엔트로피(relative entropy)
  - $D_{KL} = E[-\log q(x)] - E[-\log p(x)]$
  - $$D_{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)}\, dx$$

KULLBACK-LEIBLER (KL) DIVERGENCE

- $D_{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)} \, dx$
- $p(x)$와 $q(x)$가 같을 때 0으로 최솟값

KULLBACK-LEIBLER (KL) DIVERGENCE

- $D_{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)}\, dx$
- 비대칭적: $D_{KL}(p\,||\,q) \neq D_{KL}(q\,||\,p)$

KULLBACK-LEIBLER (KL) DIVERGENCE

- $D_{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)}\, dx$
- $p(x)$가 0에 가까워지면 $q(x)$의 효과가 무시됨
  - 분자가 0이면 분모에 상관없이 0
  - 이 경우 분포사이의 유사성 측정이 힘들어짐

JS-DIVERGENCE

JENSEN-SHANNON (JS) DIVERGENCE

- JS-divergence
  - [0, 1] 범위로 한정됨
  - 대칭적
  - 부드러움
  - $D_{JS}(p\|q) = \frac{1}{2} D_{KL}(p\| \frac{p+q}{2}) + \frac{1}{2} D_{KL}(q\| \frac{p+q}{2})$

JENSEN-SHANNON (JS) DIVERGENCE

- GAN의 손실 함수를 최적화하는 것은 D_{JS}를 최적화하는 것과 같음
  - GAN이 성공할 수 있었던 것은 D_{KL} 대신 D_{JS} 을 사용했기 때문

JENSEN-SHANNON (JS) DIVERGENCE

- Training GANs:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

  - Min-max game

JENSEN-SHANNON (JS) DIVERGENCE

- Training GANs:

$$\min_{\theta_g} \max_{\theta_d} \boxed{\left[\mathbb{E}_{x\sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))\right]}$$

$$= \mathbb{E}_{x\sim p_r(x)} [\log D(x)] + \mathbb{E}_{x\sim p_g(x)} [1 - \log D(x)]$$

  - 노이즈 분포에서의 샘플링이 아닌
  - 생성기 분포에서의 샘플링으로 표현

JENSEN-SHANNON (JS) DIVERGENCE

- $L^{(D)} = -\mathbb{E}_{x \sim p_{data}} \log D(x) - \mathbb{E}_{x \sim p_g} \log(1 - D(x))$

  $L^{(D)} = -\int_x p_{data}(x) \log D(x)\, dx - \int_x p_g(x) \log(1 - D(x))\, dx$

  $L^{(D)} = -\int_x (p_{data}(x) \log D(x) + p_g(x) \log(1 - D(x)))\, dx$

JENSEN-SHANNON (JS) DIVERGENCE

$$L^{(D)} = -\int_x \boxed{\left(p_{data}(x)\log D(x) + p_g(x)\log(1 - D(x))\right)} dx$$

- $y = a\log y + b\log(1-y)$ 형태
- $a, b \in R^2$ 와 $y \in [0,1]$ 에 대해 최댓값은 $\dfrac{a}{a+b}$
- 따라서 최적의 판별기 $D^*(x) = \dfrac{p_{data}}{p_{data} + p_g}$

JENSEN-SHANNON (JS) DIVERGENCE

- 최적의 판별기 $D^*(x) = \dfrac{p_{data}}{p_{data} + p_g}$

$$L^{(D^*)} = -\mathbb{E}_{x\sim p_{data}} \log \frac{p_{data}}{p_{data} + p_g} - \mathbb{E}_{x\sim p_g} \log\left[1 - \frac{p_{data}}{p_{data} + p_g}\right]$$

$$L^{(D^*)} = 2\log 2 - D_{KL}\left[p_{data} \,\|\, \frac{p_{data} + p_g}{2}\right] - D_{KL}\left[p_g \,\|\, \frac{p_{data} + p_g}{2}\right]$$

$$L^{(D^*)} = 2\log 2 - 2D_{JS}(p_{data} \,\|\, p_g)$$

JENSEN-SHANNON (JS) DIVERGENCE

$$L^{(D^*)} = 2\log 2 - 2D_{JS}(p_{data} \parallel p_g)$$

- L을 최소화하는 것은 D_{JS}를 최대화하는 것
  - 실제 분포와 가짜 분포의 거리를 최대화

JENSEN-SHANNON (JS) DIVERGENCE

- 최적의 생성기 $G^*(x)$는 $p_g = p_{data}$를 만듦
  - D_{JS}를 최소화
- 최적의 생성기에서 $p_g = p_{data}$이므로 판별기는
  - $D^*(x) = \dfrac{p_{data}}{p_{data} + p_g} = \dfrac{1}{2}$
- 최적의 값 $L^{(D^*)} = 2\log 2 - 2D_{JS}(p_{data} \parallel p_g) = -2\log 2$

WASSERSTEIN 1

WASSERSTEIN 1

- μ와 ν 사이의 p-Wasserstein distance
- $W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\chi \times \chi} d(x, y)^p d\pi(x, y) \right)^{1/p}$

WASSERSTEIN 1

- 1-Wasserstein 혹은 Wasserstein 1 (EMD)
- $W(p_r, p_g) = \inf_{\gamma \sim \Pi(p_r, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$

WASSERSTEIN 1

- 1-Wasserstein 혹은 Wasserstein 1
- $$W(p_r, p_g) = \inf_{\gamma \sim \Pi(p_r, p_g)} \boxed{\mathbb{E}_{(x,y)\sim\gamma} [\|x - y\|]}$$

$$\sum_{x,y} \gamma(x, y)\|x - y\| = \mathbb{E}_{x,y\sim\gamma} \|x - y\|$$

  - $\gamma(x, y)$: 질량
  - $||x - y||$: 거리

WASSERSTEIN 1

- 1-Wasserstein 혹은 Wasserstein 1
- $W(p_r, p_g) = \inf_{\boxed{\gamma \sim \Pi(p_r, p_g)}} \mathbb{E}_{(x,y)\sim\gamma} [\|x - y\|]$
    - $\Pi(p_r, p_g)$: p_r 과 p_g 사이에서 가능한 모든 결합 분포
    - $\gamma(x, y)$은 $\Pi(p_r, p_g)$ 공간에 존재하는 결합 분포

WASSERSTEIN 1

- 1-Wasserstein 혹은 Wasserstein 1
- $W(p_r, p_g) = \boxed{\inf_{\gamma \sim \Pi(p_r, p_g)}} \mathbb{E}_{(x,y)\sim\gamma} [\|x - y\|]$
  - inf는 greatest lower bound
  - EMD 이동 계획 중 가장 작은 비용

EMD IS BETTER THAN KL OR JS

WHY GAN IS HARD TO TRAIN

- GAN의 훈련은 매우 어려움
  - Loss 함수와 분포를 보면 알 수 있음

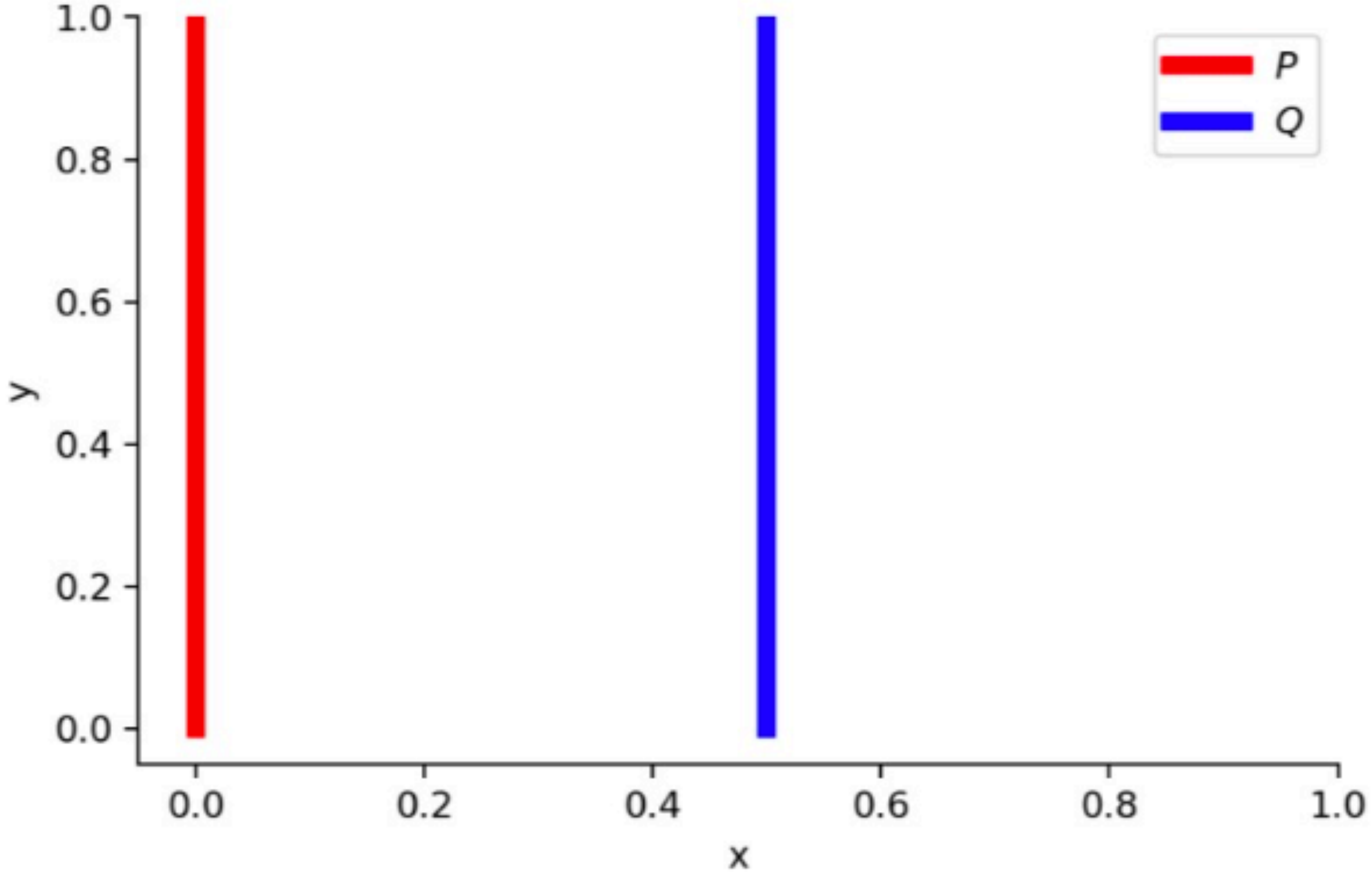
WHY GAN IS HARD TO TRAIN

- 판별기는 D_{JS}를 최대화하고자 하며
- 생성기는 D_{JS}를 최소화(0)하고자 함
  - 수렴하기 어려움

WHY GAN IS HARD TO TRAIN

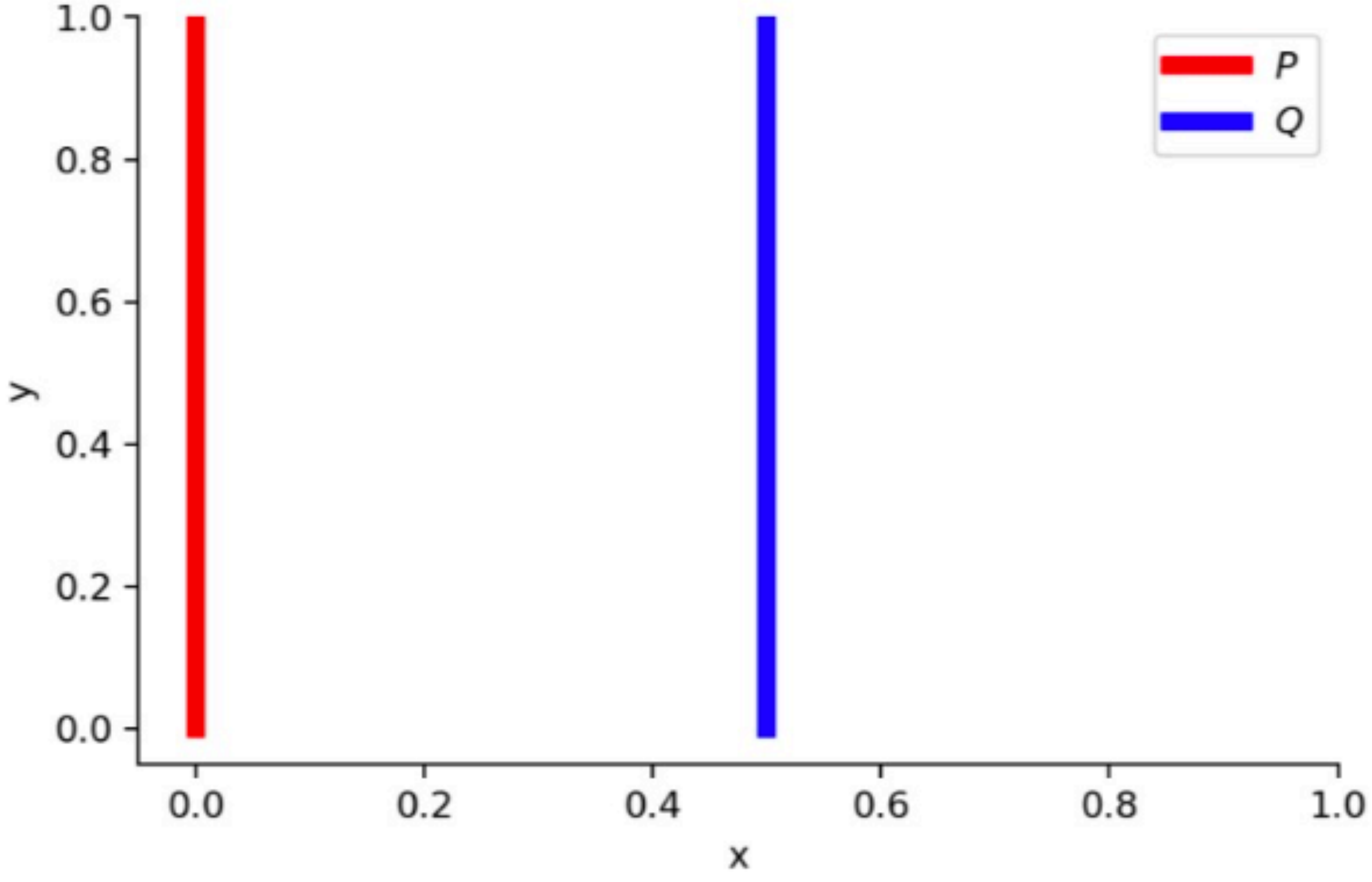
- 식별자의 성능이 나쁘면
  - 식별자가 정확한 피드백을 하지 못함
  - 손실 함수 L이 현실을 반영하지 못함
- 식별자의 성능이 너무 좋으면
  - 손실 함수가 0에 가까워짐
  - Vanishing gradient
  - 학습이 매우 느려지거나 불가

WHY GAN IS HARD TO TRAIN



- 두 분포가 겹치지 않으므로 Divergence 계산이 어려움

WHY GAN IS HARD TO TRAIN



- P는 x=0, $U(0,1)$ (균등 분포)
- Q는 x=θ, $U(0,1)$. 그림에서 $\theta = 0.5$ $(0 \leq \theta \leq 1)$

EMD IS BETTER THAN KL OR JS

- $\theta \neq 0$ 이라 하면 두 분포가 겹치지 않음
- KL-divergence:

$$D_{KL}(P\|Q) = \sum_{x=0, y\sim U(0,1)} 1 \cdot \log \frac{1}{0} = +\infty$$

$$D_{KL}(Q\|P) = \sum_{x=\theta, y\sim U(0,1)} 1 \cdot \log \frac{1}{0} = +\infty$$

EMD IS BETTER THAN KL OR JS

- JS-divergence:

$$D_{JS}(P,Q) = \frac{1}{2}(\sum_{x=0,y\sim U(0,1)} 1 \cdot \log \frac{1}{1/2} + \sum_{x=0,y\sim U(0,1)} 1 \cdot \log \frac{1}{1/2}) = \log 2$$

  - 상수이므로 경사 하강에 도움이 되지 않음

EMD IS BETTER THAN KL OR JS

- EMD:

$$W(P,Q) = |\theta|$$

  - Smooth function

EMD IS BETTER THAN KL OR JS

- $\theta = 0$ 이라 하면 두 분포가 완전히 겹침
- $D_{KL}(P\|Q) = D_{KL}(Q\|P) = D_{JS}(P, Q) = 0$
- $W(P, Q) = 0 = |\theta|$
  - 연속적

EMD IS BETTER THAN KL OR JS

- EMD는 smooth하고 연속적인 함수
  - 안정적 경사 하강법에 도움

DIVERGENCES

KL, JS, AND WASSERSTEIN 1