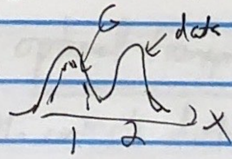
Problems w/ (Vanilla) GANs

- Convergence : min-max unstable
  vanishing gradients
  - if D to powerful, can overwhelm gen (100% seperable, no useful gradients for G)

    (can't train to completion)
  - if D too weak, random guessing
    also no gradients for G

    GAN never convergence
    good performance quickly destabilized
- Model selection
  - D & G losses not correlated w/ quality
  - in industry / natural images
    often rely on "by eye" test!
  - hard to know when to stop training
- Mode collapse

  G & D stuck in vicious cycle

  ex 

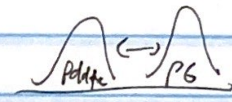
  G learns to produce 1's
  D fooled for a while
  then realizes 2 are all real, guess on 1's
  → G will switch to making 2's

(ganlab demo)

Beyond vanilla GANs

- Many issues w/ GANs stem from loss fn

if D trained to optimality → $\min_G \text{JSD}(p_{data}, p_G)$



JSD = 1

JSD = 1 whenever p_{data} p_G disjoint
bounded from above
saturates
→ vanishing gradients!

~~also mode collapse due to vanishing gradients~~
~~JSD = const~~

→ when G is very bad → no ability to improve!

↓

So can't train D to optimality for vanilla GAN

also causes mode collapse, stems from suboptimal D.

- Want: better GAN loss, unbounded from above
informative even for widely separated distr'ns

→ one SOTA approach: Wasserstein GANS (1701.07875
Arjovsky, Chintala, Bottou)

Based on optimal transport theory

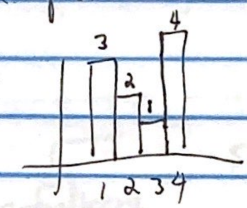
dist b/w distr'ns P, Q

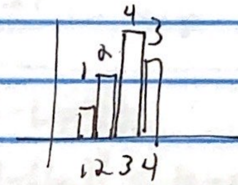
"earth mover's distance" aka "Wasserstein distance"

"work it takes to transform P into Q"

↳ right property: more separated P & Q, more work

Example:

P: 

Q 

· transport 2 units 1 → 2 so 1 matches
· transport 2 units 2 → 3 so 2 matches
· transport 1 unit 4 → 3 so 3 & 4 match ✓

EMD: $2\cdot 1 + 2\cdot 1 + 1\cdot 1 = 5$.

Lots of OT theory later...

expectation value over all dist'ns γ

Euclidean dist'ce

$$\boxed{EMD(P,Q) = W_1(P,Q) = \min_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\left[\,\|x-y\|\,\right]}$$

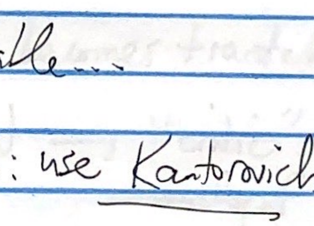
like MSE in prob dist'n space!

arb dim $P(x), Q(x)$, $x \in \mathbb{R}^d$

"Wasserstein-1"

Wasserstein-k is w/ $\|x-y\|^k$

space of all prob dist'ns $\gamma(x,y)$ on $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$



marginal $\gamma(x) = P(x)$
$\gamma(y) = Q(y)$

$W_1 = 0 \iff P = Q$
unbounded for $P \neq Q$
$\Lambda \iff \Lambda$ JSD → 1
$W_1 \to \infty$

W_1 totally intractable...

→ Brilliant idea: use Kantorovich–Rubinstein duality

Amazing formula

K-R duality

$$\boxed{W_1(P,Q) = \max_{\|h\|_L \le 1} \left[\mathbb{E}_{x\sim P}[h(x)] - \mathbb{E}_{y\sim Q}[h(y)] \right]}$$

space of 1-Lipschitz

actually true for K-Lipschitz $\|h\|_L \le K$.

$$|h(x_1)-h(x_2)| \le |x_1-x_2| \quad \forall x_1,x_2$$

$$\Updownarrow$$

$$|h'(x)| \le 1 \text{ infinitesimal}$$ by mean value thm

rough idea of pf: comes from Lagrange multipliers

$$\mathbb{E}_{x,y\sim\gamma}(\|x-y\|) = \int dx\,dy\, \gamma(x,y)\,\|x-y\|$$
$$+ \int \left(P(x) - \int \gamma(x,y)\,dy\right) f(x)\,dx$$
$$+ \int \left(Q(y) - \int \gamma(x,y)\,dx\right) g(y)\,dy$$
$$= \mathbb{E}_{x\sim P}[f] + \mathbb{E}_{y\sim Q}[g] + \int dx\,dy\,\gamma(x,y)\left(\|x-y\| - f(x) - g(y)\right)$$

$\vdots$

---

w/ K-R duality, W_1 becomes tractable!

Approx h w/ NN $\longrightarrow$ "critic"
analog of discriminator

$\max_{\|h\|_L \le 1} (\cdots) \longrightarrow$ training the critic!