Problems vs (Vanilla) GANs

- Convergence: min-max unstable
  - Varying gradients
  - If $\theta$ too powerful, can overshoot:
    - 100% separable
    - No useful gradients for $G$
  - If $\theta$ too weak, random guessing
    - Also no gradients for $G$

  GAN never converges
  - Good performance quickly destabilized

- Model selection
  - $G$ & $\theta$ not correlated w/ quality
  - In industry, natural images
    - Often rely on "by eye" test!
    - Hard to know when to stop training

- Mode collapse
  - $G$ & $\theta$ stuck in vicious cycle
    - $G$ learns to produce 1s
      - $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$
        - $\theta$ fooled for a while
        - Then realizes 2 are all real, guess on 1s
        - $G$ will switch to making 1s
Beyond vanilla GANs

- Many issues of GANs stem from loss fn
  \[ \min_G \mathbb{E}_{x \sim p_{data}} [\mathcal{L}(G(z))] \]
  \[ \mathcal{L}(G(z)) = \mathbb{E}_{z \sim p_z} [\text{JS}_D(p_d, p_G)] \]
  \[ \text{JS}_D = 1 \quad \text{whenever } p_d \text{ and } p_G \text{ disjoint} \]
  \[ \text{saturates} \rightarrow \text{vanish gradient!} \]
  \[ \rightarrow \text{when } G \text{ is very bad, } D \text{ not helpful} \]
  \[ \text{also causes mode collapse, stems from } \mathbb{E}_{z \sim p_z} \]

- Want: better GAN loss, unbounded from above
  \[ \text{informative even for widely separated clusters} \]
  \[ \rightarrow \text{one SOTA approach: Wasserstein GANs (1701.07875)} \]

  Based on optimal transport theory
  \[ \text{"earth mover's distance" aka "Wasserstein distance"} \]
  \[ p \Rightarrow q \quad \text{"work it takes to transform } p \text{ to } q" \]
  \[ \rightarrow \text{right property: more separated } p \text{ and } q \text{, more work.} \]
Example:

\[ P: \begin{array}{c|c|c|c}
1 & 2 & 3 & 4 \\
\hline
3 & 2 & 1 & 4 \\
\end{array} \quad \text{and} \quad \begin{array}{c|c|c|c}
1 & 2 & 3 & 4 \\
\hline
4 & 3 & 2 & 1 \\
\end{array} \]

- transport 2 units 1 \rightarrow 2, so 1 matches
- transport 2 units 2 \rightarrow 3, so 2 matches
- transport 1 unit 4 \rightarrow 3, so 3 and 4 match

\[ \text{CMD: } 2 \cdot 1 + 2 \cdot 1 + 1 \cdot 1 = 5 \]

Lots of OT theory later...

\[ \text{CMD}(P,A) = W(P,A) = \min_{\pi \in \Pi(P,A)} \mathbb{E}_{(x,y) \sim \pi} \| x - y \|_1 \]

Like MSE in prob.

dist. space

\[ W = \text{min}_{\pi \in \Pi(P,A)} \mathbb{E}_{(x,y) \sim \pi} \| x - y \|_1 \]

\[ W, \text{ totally tractable...} \]

\[ \rightarrow \text{Brilliant idea: use Kantorovich-Rubinstein duality} \]

Amazing formula
\[ W_1(P; Q) = \max_{\|h\|_1 \leq 1} \left[ E[h(x)] - E[h(y)] \right] \]

space of 1-Lipschitz

actual value for K-Lipschitz
\[ |h(x) - h(y)| \leq k_{x-y} \quad \forall x, y \]
\[ |h'(x)| \leq k \quad \text{infinitesimal} \quad \text{by mean value theorem} \]

comes from rough idea of $p$: Lagrange multipliers
\[
E(h(x,y)) = \int_{x\in\mathcal{X}} \int_{y\in\mathcal{Y}} \delta(x,y) |x-y| \\
+ \int \left( P(x) - \int \delta(x,y) dy \right) f(x) dx \\
+ \int \left( Q(y) - \int \delta(x,y) dx \right) g(y) dy \\
= E[F] + E[G] + \int \delta(x,y) \left( f(x) + g(y) \right) (|x-y| - f(x) - g(y))
\]

w/ K-R duality, $W_1$ becomes tractable!

Approx $h$ w/ NN $\rightarrow$ "critic"

\[ \max_{\|h\|_1 \leq 1} \quad \text{train the critic}! \]