

# Generative Adversarial Networks (GAN) and related mathematics

A/Prof Richard Yi Da Xu  
richardxu.com

University of Technology Sydney (UTS)

July 23, 2019

1. Traditional GAN
2. Mathematics on W-GAN
3. Duality and KKT conditions
4. info-GAN
5. Bayesian GAN

This lecture is referenced heavily from:

- ▶ <https://vincentherrmann.github.io/blog/wasserstein/>
- ▶ <https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html>
- ▶ <https://towardsdatascience.com/infogan-generative-adversarial-networks-part-iii-380c0c6712cd>
- ▶ <http://www.math.ubc.ca/~israel/m340/kkt2.pdf>
- ▶ <https://spaces.ac.cn/archives/6280>
- ▶ <https://spaces.ac.cn/archives/6051>
- ▶ <https://arxiv.org/pdf/1705.09558.pdf>

## Traditional GAN

- ▶ look at GAN objective:

$$\begin{aligned}\min_G \max_D \left( \mathcal{L}(D, G) \equiv \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \right) \\ = \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \underbrace{\mathbb{E}_{x \sim p_g(x)} [\log(1 - D(\mathbf{x}))]}_{\text{alternative expression}}\end{aligned}$$

- ▶ note that only  $p_g(x)$  is parameterized, you can **not** learn  $p_r(\mathbf{x})$
- ▶ **traditional view of  $D$ :**  $D$  maximize the difference between  $p_r(\mathbf{x})$  and  $p_g(\mathbf{x})$ , and  $G$  minimize the difference between  $p_r(\mathbf{x})$  and  $p_g(\mathbf{x})$
- ▶ **critic view of  $D$ :**  $D$  gives a critic between  $p_r(\mathbf{x})$  and  $p_g(\mathbf{x})$  in terms the largest of their distance (i.e, the most strict critic/judge), by maximize the difference measure between  $p_r$  and  $p_g$   
 $G$  tries to make it better ( $p_g(\mathbf{x})$  to look like  $p_r(\mathbf{x})$ ) using the current measure  
**moral of story:**  $D$  presents a way to measure between  $p_r$  and  $p_g$ , i.e., some kind of divergence

$$\left( \max_D (\mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{x \sim p_g(x)} [\log(1 - D(\mathbf{x}))]) \right) \text{ gives the strictest critic!}$$

# GAN Objective - many representations

- ▶ be careful of the signs:
- ▶ using  $-\log(D)$  trick:  $\mathcal{L}(D, G) \approx \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_g} [\log(D(G(.)))]$ :
- ▶ let  $U(\mathbf{x}) \equiv -\log D(\mathbf{x})$  and to fix  $G$ : (comes later for Energy GAN representation)

$$\begin{aligned} D &= \arg \max_D \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [-U] - \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [-U] \\ &= \arg \max_D -\mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [U(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [U(\mathbf{x})] \\ &= \arg \min_D \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [U(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [U(\mathbf{x})] \\ &= \arg \min_U \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [U(\mathbf{x})] - \mathbb{E}_{z \sim q(z)} [U(G(z))] \end{aligned}$$

$$\text{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

► in cases where  $p(x) \rightarrow 0$ , but  $q(x) \gg 0$ , effect of  $q(x)$  is disregarded

1.  $p = 0.000001$ ;  $q = 0.999999$ ; print  $p^* \text{ np.log}(p/q)$ : -1.3815509557963774e-05
2.  $p = 0.000001$ ;  $q = 0.100000$ ; print  $p^* \text{ np.log}(p/q)$ : -1.1512925464970228e-05

- ▶ we try to find  $q$  as a proposal distribution for  $\pi$
- ▶ it may turn into a **PRO** when finding approximations for  $\pi(x)$  by proposal  $q(x)$  by minimizing their KL:

$$\text{KL}(q||\pi) = \int_x q(x) \log \frac{q(x)}{\pi(x)} dx \quad \text{note the order of } \pi \text{ and } q$$

- ▶ make sure any  $x$  very *improbable* to be drawn from  $\pi(x)$  would also be very *improbable* to be drawn from  $q(x)$ :
  1. when  $q(x) \gg 0$  AND  $\pi(x) \rightarrow 0 \implies \text{KL} \rightarrow \text{high}$ :  
**prevents** draw samples where  $\pi(x)$  is low **prohibitive**  
`pi = 0.000001; q = 0.999999; print q* np.log(q/pi): 13.81549574245421`
  2. when  $q(x) \rightarrow 0$  AND  $\pi(x) \gg 0 \implies \text{KL} \rightarrow 0$ :  
**prevents** draw samples where  $\pi(x)$  is high **more forgiven**  
`pi = 0.999999; q = 0.000001; print q* np.log(q/pi): -1.3815509557963774e-0`

# KL divergence for GAN setting

- **same** as previous page, we change  $q \rightarrow p_g$ , and  $\pi \rightarrow p_r$ :

$$\text{KL}(p_g \| p_r) = \int_{\mathbf{x}} p_g \log \frac{p_g(\mathbf{x})}{p_r(\mathbf{x})} d\mathbf{x}$$

1. when  $p_g(\mathbf{x}) \gg 0$  AND  $p_r(\mathbf{x}) \rightarrow 0 \implies \text{KL} \rightarrow \text{high}$ :  
**prohibitive** for Generator to generate “unreal” image ( $p_r$  is low)  
`pr = 0.000001; pg = 0.999999; print pg* np.log(pg/pr): 13.81549574245421`  
**consequence** Generator generate **less diverse** samples  
may lead towards mode collapse
2. when  $p_g(\mathbf{x}) \rightarrow 0$  AND  $p_r(\mathbf{x}) \gg 0 \implies \text{KL} \rightarrow 0$ :  
**more forgiven** if Generator unable to generate “real” samples ( $p_r$  is high)  
`pr = 0.999999; pg = 0.000001; print pg* np.log(pg/pr): -1.3815509557963774e-0`



► **JS divergence:**

$$\text{JS}(p\|q) = \frac{1}{2}\text{KL}\left(p\left\|\frac{p+q}{2}\right.\right) + \frac{1}{2}\text{KL}\left(q\left\|\frac{p+q}{2}\right.\right)$$

# Find optimal $D^*$ after fixed $G$ (part 1)

- fix  $G$  first:

$$\begin{aligned}\min_G \max_D \mathcal{L}(D, G) &= \underbrace{\mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})}[\log(1 - D(\mathbf{x}))]}_{\mathcal{L}(G, D)} \\ \implies \mathcal{L}(G, D) &= \int_{\mathbf{x}} \left( \underbrace{p_r(\mathbf{x}) \log(D(\mathbf{x})) + p_g(\mathbf{x}) \log(1 - D(\mathbf{x}))}_{F(\mathbf{x}, D(\mathbf{x}))} \right) d\mathbf{x}\end{aligned}$$

- look at functional  $J = \int_{\mathbf{x}} \left( \underbrace{p_r(\mathbf{x}) \log(D(\mathbf{x})) + p_g(\mathbf{x}) \log(1 - D(\mathbf{x}))}_{F(\mathbf{x}, D(\mathbf{x}))} \right) d\mathbf{x}$ :

- Euler Lagrange says: to find stationary function  $\mathbf{f}$  of functional  $F$ :

$$\int_a^b F(x, \mathbf{f}(x), \mathbf{f}'(x)) dx$$

- then  $\mathbf{f}$  of a real argument  $x$ , a stationary point of the functional  $F$  when:

$$\frac{\partial F}{\partial \mathbf{f}} - \frac{d}{dx} \frac{\partial F}{\partial \mathbf{f}'} = 0$$

- in our case, we have  $x$  and  $\mathbf{f} \equiv D(x)$  and **not** have  $D'(x)$ :

$$\frac{\partial F}{\partial D(x)} = 0$$

## Find optimal $D^*$ after fixed $G$ (part 2)

► let:  $J = \int_x \underbrace{\left( p_r(x) \log(D(x)) + p_g(x) \log(1 - D(x)) \right)}_{F(x, D(x))} dx$

$$\begin{aligned} F(x, D(x)) &= p_r(x) \log D(x) + p_g(x) \log(1 - D(x)) \\ \frac{\partial F(x, D(x))}{\partial D(x)} &= p_r(x) \frac{1}{D(x)} - p_g(x) \frac{1}{1 - D(x)} = \left( \frac{p_r(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)} \right) \\ &= \frac{p_r(x) - (p_r(x) + p_g(x))D(x)}{D(x)(1 - D(x))} \end{aligned}$$

► Let  $\frac{dF(x, D(x))}{dD(x)} = 0$ :

$$\begin{aligned} \frac{p_r(x) - (p_r(x) + p_g(x))D(x)}{D(x)(1 - D(x))} &= 0 \\ \implies p_r(x) - (p_r(x) + p_g(x))D(x) &= 0 \\ D^*(x) &= \frac{p_r(x)}{p_r(x) + p_g(x)} \end{aligned}$$

► can be thought of as  $p(z|x)$  in mixture density. visualize 1-d diagram

## substitute Optimal $D^*$ into $\mathcal{L}$ :

- substitute  $D^*(\mathbf{x}) = \frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})}$  into  $\mathcal{L}(G, D^*) \equiv \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})}[\log(1 - D(\mathbf{x}))]$ , we get:

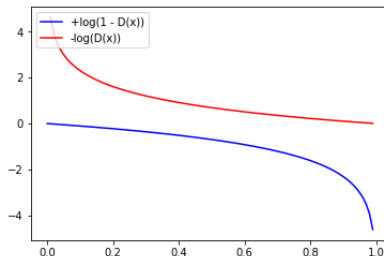
$$\mathcal{L}(G, D^*) = \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} \left[ \log \frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})} \right] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} \left[ \log \left( 1 - \frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})} \right) \right]$$

- A better way to find its relationship with JS divergence:

$$\begin{aligned} \text{JS}(p_r \| p_g) &= \frac{1}{2} \text{KL} \left( p_r \parallel \frac{p_r + p_g}{2} \right) + \frac{1}{2} \text{KL} \left( p_g \parallel \frac{p_r + p_g}{2} \right) \\ &= \frac{1}{2} \left( \int_{\mathbf{x}} p_r(\mathbf{x}) \log \frac{p_r(\mathbf{x})}{\frac{p_r(\mathbf{x}) + p_g(\mathbf{x})}{2}} d\mathbf{x} \right) + \frac{1}{2} \left( \int_{\mathbf{x}} p_g(\mathbf{x}) \log \frac{p_g(\mathbf{x})}{\frac{p_r(\mathbf{x}) + p_g(\mathbf{x})}{2}} d\mathbf{x} \right) \\ &= \frac{1}{2} \left( \log 2 + \int_{\mathbf{x}} p_r(\mathbf{x}) \log \frac{p_r(\mathbf{x})}{p_r + p_g(\mathbf{x})} d\mathbf{x} \right) + \frac{1}{2} \left( \log 2 + \int_{\mathbf{x}} p_g(\mathbf{x}) \log \frac{p_g(\mathbf{x})}{p_r + p_g(\mathbf{x})} d\mathbf{x} \right) \\ &= \frac{1}{2} \left( \log 4 + \int_{\mathbf{x}} p_r(\mathbf{x}) \log \frac{p_r(\mathbf{x})}{p_r + p_g(\mathbf{x})} d\mathbf{x} + \int_{\mathbf{x}} p_g(\mathbf{x}) \log \frac{p_g(\mathbf{x})}{p_r + p_g(\mathbf{x})} d\mathbf{x} \right) \\ &= \frac{1}{2} \left( \log 4 + \mathcal{L}(G, D^*) \right) \\ \implies \mathcal{L}(G, D^*) &= 2\text{JS}(p_r \| p_g) - 2 \log 2 \end{aligned}$$

- $\mathcal{L}(D, G)$  can be approximated by:

$$\begin{aligned}\mathcal{L}(D, G) &= \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [\log(1 - D(\mathbf{x}))] \\ &\approx \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [-\log(D(\mathbf{x}))] \\ &\approx \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [\log(D(\mathbf{x}))]\end{aligned}$$



# substitute $D^*$ in $-\log(D)$ trick

$$\begin{aligned}\mathcal{L}(G, D^*) &\equiv \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D^*(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [\log(1 - D^*(\mathbf{x}))] = 2\text{JS}(p_r \| p_g) - 2 \log 2 \\ &\implies \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} [\log(1 - D^*(\mathbf{x}))] = -\mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D^*(\mathbf{x})] + 2\text{JS}(p_r \| p_g) - 2 \log 2\end{aligned}$$

► see how we can put KL into the picture:

$$\begin{aligned}\text{KL}(p_g \| p_r) &= \mathbb{E}_{\mathbf{x} \sim p_g} \left[ \log \frac{p_g(\mathbf{x})}{p_r(\mathbf{x})} \right] \\ &= \mathbb{E}_{\mathbf{x} \sim p_g} \left[ \log \frac{p_g(\mathbf{x})}{p_r(\mathbf{x})} \right] = \mathbb{E}_{\mathbf{x} \sim p_g} \left[ \log \frac{\frac{p_g(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})}}{\frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})}} \right] \\ &= \mathbb{E}_{\mathbf{x} \sim p_g} \left[ \log \frac{1 - D^*(\mathbf{x})}{D^*(\mathbf{x})} \right] \\ &= \mathbb{E}_{\mathbf{x} \sim p_g} [\log(1 - D^*(\mathbf{x}))] - \mathbb{E}_{\mathbf{x} \sim p_g} [D^*(\mathbf{x})] \\ \implies \mathbb{E}_{\mathbf{x} \sim p_g} [-D^*(\mathbf{x})] &= \text{KL}(p_g \| p_r) - \mathbb{E}_{\mathbf{x} \sim p_g} [\log(1 - D^*(\mathbf{x}))] \\ &= \text{KL}(p_g \| p_r) - \mathbb{E}_{\mathbf{x} \sim p_g} [\log(1 - D^*(\mathbf{x}))] \\ &= \text{KL}(p_g \| p_r) + \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [\log D^*(\mathbf{x})] - 2\text{JS}(p_r \| p_g) + 2 \log 2\end{aligned}$$

- ▶ see how it works with  $-\log(D)$  trick:

$$\begin{aligned}\mathbb{E}_{\mathbf{x} \sim p_g}[-D^*(\mathbf{x})] &= \underbrace{\text{KL}(p_g \| p_r) - 2\text{JS}(p_r \| p_g)}_{\text{depends on } p_g} + \underbrace{2 \log 2 + \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[\log D^*(\mathbf{x})]}_{\text{not depend on } p_g} \\ &\propto \text{KL}(p_g \| p_r) - 2\text{JS}(p_r \| p_g)\end{aligned}$$

# What is the Optimal $\mathcal{L}$ when have both $G^*$ and $D^*$

- ▶ knowing  $D^*(x) = \frac{p_r(x)}{p_r(x) + p_g(x)}$ , then optimal  $p_g^{\theta^*}(x)$  is when it becomes identical to  $p_r(x)$ :
- ▶ from previous page:

$$\begin{aligned}\mathcal{L}(G, D^*) &= 2\text{JS}(p_r \| p_g) - 2\log 2 \\ \implies \mathcal{L}(G^*, D^*) &= -2\log 2\end{aligned}$$

- ▶ we can check again by deriving it directly from  $\mathcal{L}$ :

$$p_r(x) = p_g^{\theta^*}(x) \implies D^*(x) = \frac{1}{2}$$

- ▶ to find max value of  $\mathcal{L}$ :

$$\begin{aligned}\mathcal{L}(G^*, D^*) &= \int_x \left( p_r(x) \log(D^*(x)) + p_g^{\theta^*}(x) \log(1 - D^*(x)) \right) dx \\ &= \int_x \left( p_r(x) \log\left(\frac{1}{2}\right) + p_g^{\theta^*}(x) \log\left(1 - \frac{1}{2}\right) \right) dx \\ &= \log \frac{1}{2} \int_x p_r dx + \log \frac{1}{2} \int_x p_g^{\theta^*}(x) dx \\ &= -2\log 2\end{aligned}$$



- ▶ Given distributions  $P$  and  $Q$  of two vertical bars:

$$P: \quad x = 0$$

$$y \sim U(0, 1)$$

$$Q: \quad x = \theta, 0 \leq \theta \leq 1$$

$$y \sim U(0, 1)$$

# Problems with traditional GAN

- it turns out the distances are:

$$KL(P\|Q) = \sum_{\underbrace{x=0, y \in (0,1)}_{\forall(x,y)P(x,y)>0}} \underbrace{1}_{P(x,y)} \cdot \log \frac{\overbrace{1}^{P(x,y)}}{\underbrace{0}_{Q(x,y)}} = +\infty$$

$$KL(Q\|P) = \sum_{\underbrace{x=\theta, y \in (0,1)}_{\forall(x,y)Q(x,y)>0}} \underbrace{1}_{Q(x,y)} \cdot \log \frac{\overbrace{1}^{Q(x,y)}}{\underbrace{0}_{P(x,y)}} = +\infty$$

$$\begin{aligned} D_{JS}(P, Q) &= \frac{1}{2} \left( \sum_{x=0, y \in U(0,1)} \underbrace{1}_{P(x,y)} \cdot \log \frac{\overbrace{1}^{P(x,y)}}{\underbrace{1/2}_{\frac{P(x,y)+Q(x,y)}{2}}} + \sum_{x=\theta, y \in U(0,1)} \underbrace{1}_{Q(x,y)} \cdot \log \frac{\overbrace{1}^{Q(x,y)}}{\underbrace{1/2}_{\frac{P(x,y)+Q(x,y)}{2}}} \right) \\ &= \log 2 \end{aligned}$$

## Wasserstein-GAN

$$\min_G \left[ \underbrace{\max_{f, \|f\|_L \leq 1} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}[f(G_\theta(\mathbf{z}))]}_{\text{critic}} \right]$$

- ▶ it's pretty intuitive to see what **critic** objective does
- ▶ so why the heck we call it “Wasserstein-GAN”?
- ▶ Because discriminator/critic can be proven to be dual of Wasserstein Distance! - we prove it the other way around from primal  $\rightarrow$  dual
- ▶ and it turns out that:

$$\mathcal{W}(P, Q) = |\theta|$$

it doesn't have the “zero-jump” effect like KL or JS distance

# Wasserstein GAN and Earth Mover Distance

- ▶ Wasserstein distances between  $p_r$  and  $p_g$  are:

$$\text{EMD}(p_r, p_g) = \inf_{\gamma \in \Pi} \sum_{x,y} \|x - y\| \gamma(x, y) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y) \sim \gamma} \|x - y\|$$

- ▶ try find a transport schedule  $\gamma(x, y)$ : to “move” amount of earth from one place  $x \sim p_g(x)$  (generated) distributed from over the domain of  $y \sim p_r(y)$  (real) or vice versa
- ▶ needs to ensure marginal distributions are still there:

$$\sum_x \gamma(x, y) = p_r(y) \qquad \sum_y \gamma(x, y) = p_g(x)$$

- ▶ this is our new **critic**

► **GAN and W-GAN:**

1. GAN:

$$\text{Discriminator: } \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D_{\theta_d}(\mathbf{x}_i) + \log (1 - D_{\theta_d}(G_{\theta_g}(\mathbf{z}_i)))]$$

$$\text{Generator: } \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log (D_{\theta_d}(G_{\theta_g}(\mathbf{z}_i)))$$

2. if we can change GAN into W-GAN:

$$\text{find a critic: } \gamma^* = \inf_{\gamma \in \Pi} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma, \mathbf{x} \sim p_g, \mathbf{y} \sim p_r} \|\mathbf{x} - \mathbf{y}\|$$

$$\text{Generator: } \nabla_{\theta} \frac{1}{m} \sum_{i=1}^m \log (D_{\gamma^*}(G_{\theta}(\mathbf{z}_i)))$$

that is all we need to do. However, it is impractical to compute:

$$\gamma^* = \inf_{\gamma \in \Pi} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma, \mathbf{x} \sim p_g, \mathbf{y} \sim p_r} \|\mathbf{x} - \mathbf{y}\|$$

we need a lot of tricks!

# Primal function for EMD linear programming

$$\underbrace{\begin{pmatrix}
 \textcolor{red}{1} & \textcolor{red}{1} & \dots & \textcolor{red}{0} & \textcolor{red}{0} & \dots & \dots & \textcolor{red}{0} & \textcolor{red}{0} & \dots & \dots \\
 0 & 0 & \dots & 1 & 1 & \dots & \dots & 0 & 0 & \dots & \dots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots \\
 0 & 0 & \dots & 0 & 0 & \dots & \dots & 1 & 1 & \dots & \dots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots \\
 \hline
 \textcolor{blue}{1} & \textcolor{blue}{0} & \dots & \textcolor{blue}{1} & \textcolor{blue}{0} & \dots & \dots & \textcolor{blue}{1} & \textcolor{blue}{0} & \dots & \dots \\
 0 & 1 & \dots & 0 & 1 & \dots & \dots & 0 & 1 & \dots & \dots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots \\
 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & \dots & \dots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots
 \end{pmatrix}}_A
 \underbrace{\begin{pmatrix}
 \gamma(\mathbf{x}_1, \mathbf{y}_1) \\
 \gamma(\mathbf{x}_1, \mathbf{y}_2) \\
 \vdots \\
 \gamma(\mathbf{x}_2, \mathbf{y}_1) \\
 \gamma(\mathbf{x}_2, \mathbf{y}_2) \\
 \vdots \\
 \vdots \\
 \gamma(\mathbf{x}_n, \mathbf{y}_1) \\
 \gamma(\mathbf{x}_n, \mathbf{y}_2) \\
 \vdots \\
 \vdots
 \end{pmatrix}}_\Gamma
 =
 \underbrace{\begin{pmatrix}
 \textcolor{red}{p_g(\mathbf{x}_1)} \\
 p_g(\mathbf{x}_2) \\
 \vdots \\
 p_g(\mathbf{x}_n) \\
 \vdots \\
 \vdots \\
 \textcolor{blue}{p_r(\mathbf{y}_1)} \\
 p_r(\mathbf{y}_2) \\
 \vdots \\
 p_r(\mathbf{y}_n) \\
 \vdots \\
 \vdots
 \end{pmatrix}}_P$$

- ▶ look at the **RED** line:

$$\sum_{\mathbf{y}} \gamma(\mathbf{x}_1, \mathbf{y}) = \textcolor{red}{p_g(\mathbf{x}_1)}$$

- ▶ look at the **BLUE** line:

$$\sum_{\mathbf{x}} \gamma(\mathbf{x}, \mathbf{y}_1) = \textcolor{blue}{p_r(\mathbf{y}_1)}$$

- ▶  $\Gamma \equiv \gamma(x, y)$  acts like a vectorized joint distribution, each element  $\geq 0$
- ▶  $\mathbf{C} \equiv \text{vec}(\mathbf{D}(x, y))$  acts like a vectorized cost
- ▶  $\mathbf{P} = \begin{bmatrix} p_r(y) \\ p_g(x) \end{bmatrix}$

**primal form :**

$$\begin{aligned} \min (z = \mathbf{C}^\top \Gamma) \\ \text{s.t. } \mathbf{A}\Gamma = \mathbf{P} \\ \text{and } \Gamma \geq \mathbf{0} \end{aligned}$$

**dual form :**

$$\begin{aligned} \max (\tilde{z} = \mathbf{P}^\top \mathbf{F}) \\ \text{s.t. } \mathbf{A}^\top \mathbf{F} \leq \mathbf{C} \end{aligned}$$

$\mathbf{F}$  is variable in dual function

**Question** why dual in linear programming is in such form?

# Primal to Dual for Linear Programming (1)

- ▶ from <http://www.onmyphd.com/?p=duality.theory>
- ▶ let  $\mathbf{x} \equiv \mathbf{r}$  and  $\mathbf{b} \equiv \mathbf{P}$ :

$$\min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \underbrace{\mathbf{Ax} = \mathbf{b}}_{\mathbf{h}(\mathbf{x})}, \mathbf{x} \geq 0]$$

$$\mathcal{L}(\mathbf{x}, \mathbf{F}, \mu) = f(\mathbf{x}) + \mathbf{F}^T \mathbf{h}(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x}) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}, \mu \geq \mathbf{0}, \mathbf{F}$$

$$q(\mathbf{F}, \mu) = \inf_{\mathbf{x} \geq \mathbf{0}} [\mathcal{L}(\mathbf{x}, \mathbf{F}, \mu)]$$

$$= \inf_{\mathbf{x} \geq \mathbf{0}} [\mathbf{C}^T \mathbf{x} + \mathbf{F}^T (\mathbf{Ax} - \mathbf{b})]$$

$$= \inf_{\mathbf{x} \geq \mathbf{0}} [(\mathbf{C}^T + \mathbf{F}^T \mathbf{A})\mathbf{x} - \mathbf{F}^T \mathbf{b}]$$

- ▶ **task** only include  $(\mathbf{F}, \mu)$  space which **avoid** making  $q(\mathbf{F}, \mu) = -\infty$  (maximization) constraints should be put to avoid these regions.

$$(\mathbf{C}^T + \mathbf{F}^T \mathbf{A}) < \mathbf{0} \implies \mathbf{x} \text{ can be made arbitrarily large to make } q(\mathbf{F}, \mu) \rightarrow -\infty$$

$$\text{if } \mathbf{C}^T + \mathbf{F}^T \mathbf{A} \geq \mathbf{0} \implies \mathbf{x}^* = \mathbf{0} \implies q(\mathbf{F}, \mu) = -\mathbf{F}^T \mathbf{b}$$

- ▶ which means:

$$\max_{\mathbf{F}} [-\mathbf{F}^T \mathbf{b} \mid \mathbf{C}^T + \mathbf{F}^T \mathbf{A} \geq \mathbf{0}]$$

or let  $\mathbf{F}' = -\mathbf{F}$ :

$$\max_{\mathbf{F}'} [\mathbf{F}'^T \mathbf{b} \mid \mathbf{C}^T \geq \mathbf{F}'^T \mathbf{A}]$$



# Primal to Dual for Linear Programming (2)

let  $\mathbf{x} \equiv \mathbf{r}$  and  $\mathbf{b} \equiv \mathbf{P}$ :

assume the condition  $\mathbf{F}^T \mathbf{A} \leq \mathbf{C}^T \forall \mathbf{F}$ :

$$\mathbf{F}^T \mathbf{Ax}^* \leq \mathbf{C}^T \mathbf{x}^* \forall \mathbf{F} \text{ since } \mathbf{x}^* \geq 0, \text{ after multiplication, no change sign}$$

$$\Rightarrow \mathbf{F}^T \underbrace{\mathbf{b}} \leq \mathbf{C}^T \mathbf{x}^* \forall \mathbf{F} \text{ assume } \mathbf{Ax}^* = \mathbf{b}$$

$$= \min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$$

$$\Rightarrow \underbrace{\max_{\mathbf{F}} [\mathbf{F}^T \mathbf{b}]}_{\mathbf{F}^*} \leq \min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$$

$$\Rightarrow \max_{\mathbf{F}} [\mathbf{F}^T \mathbf{b} \mid \mathbf{F}^T \mathbf{A} \leq \mathbf{C}^T \forall \mathbf{F}] \leq \min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0] \quad \text{write the condition in}$$

## Lagrangian Duality and KKT condition

- ▶ a constrained optimization is in the following form (ignore the equality for now):

$$\begin{aligned} \min f(\mathbf{x}) \\ \text{s.t. } g_i(\mathbf{x}) \leq 0 \quad \forall i \in 1, \dots, m \end{aligned}$$

- ▶ after defined  $I(u) = \begin{cases} 0, & \text{if } u \leq 0 \\ \infty, & \text{otherwise} \end{cases}$
- ▶ we can specify a constrained equation using **unconstrained** equation:

$$J(x) = f(x) + \sum_i I[g_i(x)]$$

- ▶ it words, it makes infeasible region so large, i.e.,  $\infty$  making it impossible to find a **minimization** solution
- ▶ similarly, in **maximization**, infeasible region are assigned value of  $-\infty$  making it impossible to find a maximum solution

- ▶ replace  $\mathbb{I}[g_i(x)]$  by its lower bound  $\lambda_i g_i(\mathbf{x})$ , with  $\lambda_i \geq 0$ :

$$\begin{aligned} \left( \mathcal{L}(x, \lambda) \equiv f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \right) &\leq J(\mathbf{x}) \\ \text{i.e.,} \quad \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) &= J(\mathbf{x}) \end{aligned}$$

- ▶ if we were to minimize both side for  $\mathbf{x}$ :

$$\Rightarrow \left( p^* \equiv \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) \right) = \min_{\mathbf{x}} J(\mathbf{x})$$

# When constraints are **all satisfied**

When constraints are **all satisfied**:  $g_i(\mathbf{x}) \leq 0 \ \forall$

$$J(\mathbf{x}) = f(\mathbf{x}) + \sum_i \mathbf{I}[g_i(\mathbf{x})] \quad \mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x})$$

- ▶  $\mathcal{L}(\mathbf{x}, 0) = f(\mathbf{x})$
- ▶ best  $\lambda_i$  is when:

$$\lambda_i^* = \arg \max_{\lambda_i} \mathcal{L}(\mathbf{x}, \lambda_i) = 0$$

- ▶ this is because  $\lambda_i \geq 0$ , in case:

$$g_i(\mathbf{x}) \leq 0 \text{ and } \lambda_i > 0 \implies \lambda_i g_i(\mathbf{x}) \leq 0$$

- ▶ so **max** occur when  $\lambda_i = 0$

# When constraints are **not all satisfied**

When constraints are **not all satisfied**:  $\exists_i g_i(\mathbf{x}) > 0$

$$J(\mathbf{x}) = f(\mathbf{x}) + \sum_i \mathbf{I}[g_i(\mathbf{x})] \quad \mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x})$$

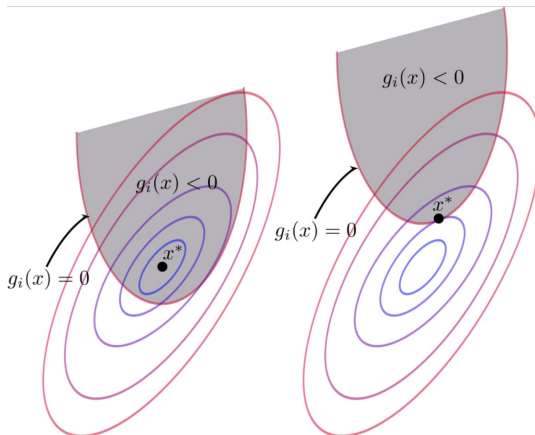
- ▶ we can **maximize**  $\mathcal{L}(\mathbf{x}, \lambda)$  by taking  $\lambda_i \rightarrow \infty$
- ▶ we can see that way to prevent going to infinity is to  $g_i(\mathbf{x}) = 0$
- ▶ so here is one way to look at **complimentary slackness**

combine with previous page: either  $\lambda_i$  or  $g_i(\mathbf{x})$  needs to be zero, i.e.,:

$$\lambda_i g_i(\mathbf{x}) = 0$$

# Diagrammatic illustration of **complimentary slackness**

from wikipedia:



- relationship between  $\lambda_i g_i(\mathbf{x})$  and  $\mathbb{I}[g_i(\mathbf{x})]$ :

$$\begin{aligned} \underbrace{f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x})}_{\mathcal{L}(\mathbf{x}, \lambda)} &\leq \underbrace{f(\mathbf{x}) + \sum_i \mathbb{I}[g_i(\mathbf{x})]}_{J(\mathbf{x})} \\ \Rightarrow \left( g(\lambda) \equiv \min_x \mathcal{L}(x, \lambda) \right) &\leq \underbrace{\left( p^* \equiv \min_x J(x) \right)}_{\text{no } \lambda} \\ \Rightarrow \left( d^* \equiv \max_{\lambda} g(\lambda) \right) &\leq p^* \end{aligned}$$

summarize:  $\left( d^* \equiv \max_{\lambda} \min_x \mathcal{L}(x, \lambda) \right) \leq \left( p^* \equiv \min_x J(x) = \min_x \max_{\lambda} \mathcal{L}(x, \lambda) \right)$

- if strong duality holds:

$$d^* = p^*$$



# Max-min inequality

- ▶ Max-min inequality

$$\sup_{z \in Z} \inf_{w \in W} f(z, w) \leq \inf_{w \in W} \sup_{z \in Z} f(z, w)$$

- ▶ “the greatest of all minima” is less or equal to “the least of all maxima”
- ▶ **proof**

$$\begin{aligned}\text{Let } g(z) &\triangleq \inf_{w \in W} f(z, w) \\ \implies g(z) &\leq f(z, w), \forall z \forall w \\ \implies \sup_z g(z) &\leq \sup_z f(z, w), \forall w \\ \implies \sup_z \inf_w f(z, w) &\leq \sup_z f(z, w), \forall w \\ \implies \sup_z \inf_w f(z, w) &\leq \inf_w \sup_z f(z, w)\end{aligned}$$

- ▶ this also applied to duality theorem:

$$(d^* \equiv \max_{\lambda} \min_x L(x, \lambda)) \leq (p^* \equiv \min_x \max_{\lambda} L(x, \lambda))$$

# Max-min inequality

- Max-min inequality

$$\sup_{\lambda \in \Lambda} \inf_{x \in \mathcal{X}} f(\lambda, x) \leq \inf_{x \in \mathcal{X}} \sup_{\lambda \in \Lambda} f(\lambda, x)$$

- “the greatest of all minima” is less or equal to “the least of all maxima”
- **proof**

$$\text{Let } g(\lambda) \triangleq \inf_{x \in \mathcal{X}} f(\lambda, x)$$

$$\implies g(\lambda) \leq f(\lambda, x), \forall \lambda \forall x$$

$$\implies \sup_{\lambda} g(\lambda) \leq \sup_{\lambda} f(\lambda, x), \forall x$$

$$\implies \sup_{\lambda} \inf_x f(\lambda, x) \leq \sup_{\lambda} f(\lambda, x), \forall x$$

$$\implies \sup_{\lambda} \inf_x f(\lambda, x) \leq \inf_x \sup_{\lambda} f(\lambda, x)$$

- this also applied to duality theorem:

$$(d^* \equiv \max_{\lambda} \min_x \mathcal{L}(x, \lambda)) \leq (p^* \equiv \min_x \max_{\lambda} \mathcal{L}(x, \lambda))$$

- ▶ Let  $X \subset \mathbb{R}^n$  and  $Y \subset \mathbb{R}^m$  be compact convex sets.
- ▶ If  $f : X \times Y \rightarrow \mathbb{R}$  is a continuous function that is convex-concave:

$f(\cdot, y) : X \rightarrow \mathbb{R}$  is convex for fixed  $y$

$f(x, \cdot) : Y \rightarrow \mathbb{R}$  is concave for fixed  $x$

- ▶ then:

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y)$$

summary of KKT condition:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x})$$

subject to  $h_i(\mathbf{x}) = 0, \forall i = 1, \dots, m$

subject to  $g_i(\mathbf{x}) \leq 0, \forall i = 1, \dots, n$

**minimization obj :** 
$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^m \mu_i \nabla_{\mathbf{x}} h_i(\mathbf{x}) + \sum_{i=1}^n \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) = 0$$

**maximization obj :** 
$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^m \mu_i \nabla_{\mathbf{x}} h_i(\mathbf{x}) - \sum_{i=1}^n \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) = 0$$

**equality constraints :** 
$$\nabla_{\mu} f(\mathbf{x}) + \sum_{i=1}^m \mu_i \nabla_{\mu} h_i(\mathbf{x}) + \sum_{i=1}^n \lambda_i \nabla_{\mu} g_i(\mathbf{x}) = 0$$

$$\implies \sum_{i=1}^m \mu_i \nabla_{\mu} h_i(\mathbf{x}) = 0$$

**Inequality constraints a.k.a. complementary slackness condition**

$$\lambda_i g_i(\mathbf{x}) = 0, \forall i = 1, \dots, n$$

$$\lambda_i \geq 0, \forall i = 1, \dots, n$$

$$g_i(\mathbf{x}) \leq 0, \forall i = 1, \dots, n$$

# KKT Example

- ▶ from <http://www.math.ubc.ca/~israel/m340/kkt2.pdf>

$$\text{maximize } f(x, y) = xy$$

$$\text{subject to } x + y^2 \leq 2$$

$$x, y \geq 0$$

- ▶ note the feasible region is bounded, so a global maximum must exist: a continuous function on a closed and bounded set has a maximum there. **why?**
- ▶ write constraints as:

$$g_1(x, y) \equiv x + y^2 \leq 2$$

$$g_2(x, y) \equiv -x \leq 0$$

$$g_3(x, y) \equiv -y \leq 0$$

- ▶ **maximization obj**

$$\nabla_x(xy) - [\nabla_x \lambda_1(x + y^2) + \nabla_x \lambda_2(-x) + \nabla_x \lambda_3(-y)] = 0 \implies y - \lambda_1 + \lambda_2 = 0$$

$$\nabla_y(xy) - [\nabla_y \lambda_1(x + y^2) + \nabla_y \lambda_2(-x) + \nabla_y \lambda_3(-y)] \implies x - 2y\lambda_1 + \lambda_3 = 0$$

- ▶ **inequality constraints, complementary slackness:**

$$\lambda_1(2 - x - y^2) = 0 \quad \text{or, } \lambda_1(x + y^2 - 2) = 0$$

$$\lambda_2 x = 0 \quad \text{or, } \lambda_2(-x) = 0$$

$$\lambda_3 y = 0 \quad \text{or, } \lambda_3(-y) = 0$$

$$\lambda_1, \lambda_2, \lambda_3 \geq 0$$

- ▶ **carried from primal constraints**

$$x + y^2 \leq 2$$

$$x, y \geq 0$$

# Karush-Kuhn-Tucker Example

- ▶ in each of “complementary slackness” equations,  $\lambda_i g_i(x_1, \dots, x_n) = 0$ , at least one of the two factors must be 0. With  $n$  such conditions, there would potentially be  $2^n$  possible cases to consider
- ▶ However, with some thought we might be able to reduce that considerably:

$$\left\{ \begin{array}{l} \text{case 1: suppose } \lambda_1 = 0 \\ \quad y - \lambda_1 + \lambda_2 = 0 \implies y + \lambda_2 = 0 \qquad x - 2y\lambda_1 + \lambda_3 = 0 \implies x + \lambda_3 = 0 \\ \quad \text{since each term is nonnegative, only way to happen } x = y = \lambda_2 = \lambda_3 = 0 \\ \quad \text{although KKT conditions satisfied when } x = y = \lambda_1 = \lambda_2 = \lambda_3 = 0 \\ \quad \text{but it is not a local maximum since:} \\ \qquad f(0, 0) = 0 \text{ while } f(x, y) > 0 \text{ at points in the interior of the feasible region} \\ \text{case 2: suppose } x + y^2 - 2 = 0 \qquad x = 2 - y^2 \\ \left\{ \begin{array}{l} \text{case 2a } x > 0 : \\ \quad (\because \lambda_2 x = 0) \implies \lambda_2 = 0 \\ \quad (\because y - \lambda_1 + \lambda_2 = 0) \implies \lambda_1 = y \\ \quad (\because x - 2y\lambda_1 + \lambda_3 = 0) \implies 2 - y^2 - 2y\lambda_1 + \lambda_3 = 0 \implies 2 - 3y^2 + \lambda_3 = 0 \\ \quad \text{consequently: } 3y^2 = 2 + \lambda_3, (\because \lambda_3 \geq 0) \implies y > 0, (\because \lambda_3 y = 0) \implies \lambda_3 = 0 \\ \quad \text{all KKT conditions are satisfied} \\ \text{case 2b } x = 0 : \\ \quad x + y^2 - 2 = 0 \implies y = \sqrt{2} \qquad \because y > 0 \implies \lambda_3 = 0 \\ \quad x - 2y\lambda_1 + \lambda_3 = 0 \implies \lambda_1 = 0, \text{ takes us back to case 1} \\ \quad \text{only two candidates for a local max: } (0, 0) \text{ and } \left(\frac{4}{3}, \sqrt{\frac{2}{3}}\right), \text{ global maximum at } \left(\frac{4}{3}, \sqrt{\frac{2}{3}}\right) \end{array} \right. \end{array} \right.$$

# Strong duality this time!

- ▶ we have proved that:

$$\max_F [F^\top \mathbf{b} \mid F^\top \mathbf{A} \leq \mathbf{C}^\top \ \forall F] \leq \min_{\mathbf{x}} [\mathbf{C}^\top \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$$

- ▶ but we are greedy, we want to prove in w-GAN setting, it has strong duality:

$$\max_F [F^\top \mathbf{b} \mid F^\top \mathbf{A} \leq \mathbf{C}^\top \ \forall F] = \min_{\mathbf{x}} [\mathbf{C}^\top \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$$

- ▶ we can use **Farkas Lemma** to prove this

# Farkas Lemma Proof Sketch

$$\text{prove} \quad \max_F [F^T \mathbf{b} \mid F^T \mathbf{A} \leq \mathbf{C}^T \ \forall F] \underbrace{=}_{\mathbf{x}} \min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$$

where  $z^* = \min_{\mathbf{x}} [\mathbf{C}^T \mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0]$  is min in primal

1. extend cleverly everything by a single dimension (1):

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ -\mathbf{C}^T \end{bmatrix}, \quad \hat{\mathbf{b}}_\epsilon = \begin{bmatrix} \mathbf{b} \\ -z^* + \epsilon \end{bmatrix}, \quad \hat{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \alpha \end{bmatrix} \text{ where } \epsilon, \alpha \in \mathbb{R}$$

2. when  $\epsilon > 0$ : after **proved**  $\alpha > 0$  (2.1) using Farkas Lemma, we then prove:

$$\tilde{z} = \max_F [\mathbf{b}^T \mathbf{F} \mid \mathbf{A}^T \mathbf{F} \leq \mathbf{C}] > z^* - \epsilon \quad (\text{using Farkas Lemma again!}) \quad (2.2)$$

3. then it is obvious  $\tilde{z} \in ((z^* - \epsilon), z^*)$   
making  $\epsilon$  infinitely small, we get

$$\tilde{z} = z^*$$



# Convex and Conic combination

- ▶ matrix  $\mathbf{A} \in \mathbb{R}^{d \times n} \triangleq (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$
- ▶ **def** Convex combination:

$$C = \{\mathbf{a} \mid \mathbf{a} = \alpha_1 \mathbf{a}_1 + \dots + \alpha_k \mathbf{a}_k, \alpha_1 + \dots + \alpha_k = 1, \alpha_i \geq 0\}$$

for example  $\mathbf{A} \in \mathbb{R}^{2 \times 3}$ , then it looks like a painted triangle

- ▶ **def** Conic combination is:

$$C = \{\mathbf{a} \mid \mathbf{a} = \alpha_1 \mathbf{a}_1 + \dots + \alpha_k \mathbf{a}_k, \alpha_i \geq 0\}$$

for example  $\mathbf{A} \in \mathbb{R}^{2 \times 3}$ , it looks painted cone from the origin

► **Farkas Lemma** say, for a vector **b**, there are exactly two **mutually exclusive** possibilities:

1. **b inside** the cone:

$$\exists \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0 \text{ (in every dimension) s.t. } \mathbf{Ax} = \mathbf{b}$$

2. **b outside** the cone:

$$\nexists \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0 \text{ (in every dimension) s.t. } \mathbf{Ax} = \mathbf{b}$$
$$\forall \mathbf{x} \geq 0, \text{ (in every dimension) s.t. } \mathbf{Ax} \neq \mathbf{b}$$

these are not the most useful definitions, we use instead:

$$\exists \mathbf{F} \in \mathbb{R}^m, \text{ s.t. } \mathbf{A}^\top \mathbf{F} \leq 0 \text{ and } \mathbf{b}^\top \mathbf{F} > 0$$

note that  $\mathbf{y} \in \mathbb{R}^m$ , and  $\mathbf{x} \in \mathbb{R}^n$ , they are not the same dimension

$$\exists F \in \mathbb{R}^m, \text{ s.t. } \mathbf{A}^\top F \leq 0 \text{ and } \mathbf{b}^\top F > 0$$

where  $F \in \mathbb{R}^m$ , and  $\mathbf{x} \in \mathbb{R}^n$   
the geometry can be thought as:

- ▶  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  forms a cone, each  $\mathbf{x}_i$  to be either an internal or external wall.
- ▶  $F$  is the outer door, swing about the origin, that is more than  $\frac{\pi}{2}$  away from each and every wall ( $\mathbf{A}$ ), as  $\mathbf{A}^\top F \leq 0$
- ▶  $\mathbf{b}$  is an inner door, swing about the origin that is less than  $\frac{\pi}{2}$  from outer door ( $F$ ), as  $\mathbf{b}^\top F \geq 0$
- ▶ can made much clearer by include  $h$  (orthogonal to  $b$ ):  
there is always a  $F$  and together with its orthogonal pair  $h$  to contain  $\mathbf{b}$

# 1 Clever Extension

- ▶ here comes a tricky bit: extend  $\mathbf{a}_i$  by one dimension, i.e.,  $m \rightarrow m + 1$ , so the rest variables  $(\mathbf{A}, \mathbf{b}, F)$  has an additional dimension:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ -\mathbf{C}^\top \end{bmatrix}, \quad \hat{\mathbf{b}}_\epsilon = \begin{bmatrix} \mathbf{b} \\ -z^* + \epsilon \end{bmatrix} \quad \hat{F} = \begin{bmatrix} F \\ \alpha \end{bmatrix} \text{ where } \epsilon, \alpha \in \mathbb{R}$$

note that  $\mathbf{x}$  does **not** extend, so it can be applied in both systems

- ▶ also note that:

$$\hat{\mathbf{b}}_0 = \begin{bmatrix} \mathbf{b} \\ -z^* + 0 \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -z^* \end{bmatrix}$$

## 2.1 Proving $\alpha > 0$ using Farkas Lemma (1)

- for  $\epsilon = 0$ , can prove  $\exists \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0$  s.t.  $\hat{\mathbf{A}}\mathbf{x} = \mathbf{b}_0 \implies \mathbf{b}_0$  inside cone, i.e., Farkas case (1): obviously  $\mathbf{x} = \mathbf{x}^*$  works!

$$\hat{\mathbf{A}}\mathbf{x}^* = \begin{bmatrix} \mathbf{A} \\ -\mathbf{c}^\top \end{bmatrix} \mathbf{x}^* = \begin{bmatrix} \mathbf{A}\mathbf{x}^* \\ -\mathbf{c}^\top \mathbf{x}^* \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -z^* + 0 \end{bmatrix} = \hat{\mathbf{b}}_0$$

1. **b inside** the cone:  $\exists \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0$  (in every dimension) s.t.  $\mathbf{Ax} = \mathbf{b}$
2. **b outside** the cone:  $\exists \mathbf{F} \in \mathbb{R}^m$ , s.t.  $\mathbf{A}^\top \mathbf{F} \leq 0$  and  $\mathbf{b}^\top \mathbf{F} > 0$

since it's Farkas (1), then Farkas (2) can **not** exist, i.e.,:

$$\forall \hat{\mathbf{A}}^\top \hat{\mathbf{F}} \leq 0 \implies \underline{\hat{\mathbf{b}}_0^\top \hat{\mathbf{F}} \leq 0}$$

- $\alpha$ -condition 1:  $\epsilon = 0 : \forall \hat{\mathbf{A}}^\top \hat{\mathbf{F}} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\mathbf{F}} \leq 0$



## 2.1 Proving $\alpha > 0$ using Farkas Lemma (3)

- ▶  **$\alpha$ -condition 1:**  $\epsilon = 0 : \forall \hat{\mathbf{A}}^\top \hat{\mathbf{F}} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\mathbf{F}} \leq 0$
- ▶  **$\alpha$ -condition 2:**  $\epsilon > 0 : \forall \hat{\mathbf{A}}^\top \hat{\mathbf{F}} \leq 0 \quad \exists \hat{\mathbf{b}}_0^\top \hat{\mathbf{F}} + \alpha \epsilon > 0$
- ▶ since  $\exists \hat{\mathbf{F}}$  satisfy both  $\alpha$ -**conclusions**, it only works when  $\alpha > 0$
- ▶ note that not every  $\alpha > 0$  works, but it's a necessary conditions!

## 2.2 Prove $\tilde{z} > z^* - \epsilon$ using Farkas Lemma (1)

- ▶ we just proved that  $\alpha > 0$ , which implies by it won't change sign
- ▶ we saw when  $\epsilon > 0$ , there exists no non-negative solution, the **implication** is Farkas case (2):  
meaning when  $\epsilon > 0$ , there exist  $\hat{F} \equiv \begin{bmatrix} \mathbf{F} \\ \alpha \end{bmatrix}$  solution such that:

$$\underbrace{\begin{bmatrix} \mathbf{A} \\ -\mathbf{C}^\top \end{bmatrix}^\top \begin{bmatrix} \mathbf{F} \\ \alpha \end{bmatrix} \leq 0}_{\Rightarrow \mathbf{A}^\top \mathbf{F} \leq \alpha \mathbf{C}} \quad \underbrace{\begin{bmatrix} \mathbf{b} \\ -z^* + \epsilon \end{bmatrix} \begin{bmatrix} \mathbf{F} \\ \alpha \end{bmatrix} > 0}_{\Rightarrow \mathbf{b}^\top \mathbf{F} > \alpha(z^* - \epsilon)}$$

$$\mathbf{A}^\top \mathbf{F} \leq \alpha \mathbf{C} \Rightarrow \mathbf{A}^\top \frac{\mathbf{F}}{\alpha} \leq \mathbf{C}$$

$$\mathbf{b}^\top \mathbf{F} > \alpha(z^* - \epsilon) \Rightarrow \mathbf{b}^\top \frac{\mathbf{F}}{\alpha} > (z^* - \epsilon)$$



## 2.2 Prove $\tilde{z} > z^* - \epsilon$ using Farkas Lemma (2)

- ▶ now we have:  $\mathbf{A}^\top \frac{\mathbf{F}}{\alpha} \leq \mathbf{C}$  and  $\mathbf{b}^\top \frac{\mathbf{F}}{\alpha} > (z^* - \epsilon)$

$$\underbrace{\mathbf{A}^\top \mathbf{F} \leq \mathbf{C}}_{\text{constraint}} \quad \text{and} \quad \underbrace{\mathbf{b}^\top \mathbf{F} > (z^* - \epsilon)}_{\text{obj}}$$

- ▶ combine the two above, we have:

$$\tilde{z} = \max_{\mathbf{F}} [\mathbf{b}^\top \mathbf{F} \mid \mathbf{A}^\top \mathbf{F} \leq \mathbf{C}] > z^* - \epsilon$$

- ▶ we can make  $\epsilon$  arbitrarily small, to make  $\tilde{z} = z^*$ , so we have **strong** duality!

# Rewrite the Dual function $\mathbf{P}^\top \mathbf{F}$

- ▶ switching generic symbols back:  $\boldsymbol{\Gamma} \equiv \mathbf{x}$  and  $\mathbf{P} \equiv \mathbf{b}$
- ▶ we know primal and dual are equal then:

$$\min_{\boldsymbol{\Gamma}} [\boldsymbol{\Gamma}^\top \mathbf{C} \mid \mathbf{A}\boldsymbol{\Gamma} = \mathbf{P}, \boldsymbol{\Gamma} \geq 0] = \max_{\mathbf{F}} [\mathbf{P}^\top \mathbf{F} \mid \mathbf{A}^\top \mathbf{F} \leq \mathbf{C}]$$

- ▶ by breaking up  $\mathbf{F}$  into  $\mathbf{F}_w = \begin{bmatrix} f_g^w \\ f_r^w \\ f_g^w \end{bmatrix}$  to match with  $\mathbf{P}$ :

$$\mathbf{F} = \begin{bmatrix} f_g^w(\mathbf{x}_1) \\ f_g^w(\mathbf{x}_2) \\ \vdots \\ f_g^w(\mathbf{x}_n) \\ \vdots \\ f_r^w(\mathbf{y}_1) \\ f_r^w(\mathbf{y}_2) \\ \vdots \\ f_r^w(\mathbf{y}_n) \\ \vdots \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} p_g(\mathbf{x}_1) \\ p_g(\mathbf{x}_2) \\ \vdots \\ p_g(\mathbf{x}_n) \\ \vdots \\ p_r(\mathbf{y}_1) \\ p_r(\mathbf{y}_2) \\ \vdots \\ p_r(\mathbf{y}_n) \\ \vdots \end{bmatrix}$$

- ▶ we have:

$$\mathbf{P}^\top \mathbf{F} = \sum_n p_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + \sum_n p_r(\mathbf{y}_n) f_r^w(\mathbf{y}_n)$$

# Rewrite the Dual function $\mathbf{P}^\top \mathbf{F}$ (2)

► further refining:

$$\begin{aligned}\mathbf{P}^\top \mathbf{F} &= \sum_n p_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + \sum_n p_r(\mathbf{y}_n) f_r^w(\mathbf{y}_n) \\ &= \sum_n p_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + \sum_n p_r(\mathbf{x}_n) f_r^w(\mathbf{x}_n) \\ &= \sum_n [p_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + p_r(\mathbf{x}_n) f_r^w(\mathbf{x}_n)] \\ &= \sum_n [p_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + p_r(\mathbf{x}_n) f_r^w(\mathbf{x}_n)]\end{aligned}$$

# Finishing Farkas theorem (2)

$$\min_{\Gamma} [\Gamma^T \mathbf{C} \mid \mathbf{A}\Gamma = \mathbf{P}, \Gamma \geq 0] = \max_{\mathbf{F}} [\mathbf{P}^T \mathbf{F} \mid \mathbf{A}^T \mathbf{F} \leq \mathbf{C}]$$

$$\underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 & \cdots \\ 1 & 0 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ 0 & 1 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 & \cdots \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ 0 & 0 & \cdots & 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ 0 & 0 & \cdots & 1 & \cdots & 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \end{bmatrix}}_{\mathbf{A}^T} \underbrace{\begin{bmatrix} f_g^w(\mathbf{x}_1) \\ f_g^w(\mathbf{x}_2) \\ \vdots \\ f_g^w(\mathbf{x}_n) \\ \vdots \\ f_r^w(\mathbf{y}_1) \\ f_r^w(\mathbf{y}_2) \\ \vdots \\ f_r^w(\mathbf{y}_n) \\ \vdots \end{bmatrix}}_{\mathbf{F}} \leq \underbrace{\begin{bmatrix} d(\mathbf{x}_1, \mathbf{y}_1) \\ d(\mathbf{x}_1, \mathbf{y}_2) \\ \vdots \\ d(\mathbf{x}_2, \mathbf{y}_1) \\ d(\mathbf{x}_2, \mathbf{y}_2) \\ \vdots \\ \vdots \\ d(\mathbf{x}_n, \mathbf{y}_1) \\ d(\mathbf{x}_n, \mathbf{y}_2) \\ \vdots \\ \vdots \end{bmatrix}}_{\mathbf{C}}$$

pick any row of  $\mathbf{A}^T$ , gives you:

$$f_g^w(\mathbf{x}_i) + f_r^w(\mathbf{y}_j) \leq d(\mathbf{x}_i, \mathbf{y}_j)$$

write it more generically,  $\forall \mathbf{x}, \mathbf{y}, f_g^w(\mathbf{x}) + f_r^w(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})$

# reduce argument to only $f$ instead of $f_g^w$ and $f_r^w$

- ▶ dual function:

$$\begin{aligned}\mathcal{W}(p_g, p_r) &= \max_{\mathbf{F}} [\mathbf{P}^\top \mathbf{F} \mid \mathbf{A}^\top \mathbf{F} \leq \mathbf{C}] \\ &= \max_{f_r^w, f_g^w} \left\{ \underbrace{\sum_n [\rho_g(\mathbf{x}_n) f_g^w(\mathbf{x}_n) + \rho_r(\mathbf{x}_n) f_r^w(\mathbf{x}_n)]}_{\mathbf{P}^\top \mathbf{F}} \mid \underbrace{f_g^w(\mathbf{x}) + f_r^w(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})}_{\mathbf{A}^\top \mathbf{F} \leq \mathbf{D}} \right\}\end{aligned}$$

- ▶ objective function is

$$\mathbf{P}^\top \mathbf{F} = \rho_r(\mathbf{x}) f_r^w(\mathbf{x}) + \rho_g(\mathbf{x}) f_g^w(\mathbf{x})$$

- ▶ constraint is  $f_g^w(\mathbf{x}) + f_r^w(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})$ :
- ▶ a special case is when:

$$f_g^w(\mathbf{x}) + f_r^w(\mathbf{x}) \leq d(\mathbf{x}, \mathbf{x}) = 0$$

# reduce argument to only $f$ instead of $f_g^w$ and $f_r^w$

- ▶  $\mathbf{P}^\top \mathbf{F} = p_r(\mathbf{x})f_r^w(\mathbf{x}) + p_g(\mathbf{x})[f_r^w(\mathbf{x})]$  since  $p_r, p_g \geq 0$ , for each  $\mathbf{x}_i$ , the term can be equivalently expressed as:

$$\max_{x_1, x_2} [p_1 x_1 + p_2 x_2 \mid x_1 + x_2 \leq 0] \quad p_1, p_2 \geq 0$$

- ▶ wlog,  $x_1 \geq x_2$ , i.e.  $x_1 \geq 0$  and  $x_2 \leq 0$ :
- ▶ so max occur at:

$$\max(p_1, p_2)x_1 + \min(p_1, p_2)x_2 \quad \forall x_1, x_2 \quad \text{note } x_1, x_2 \text{ can has very large absolute value}$$

- ▶  $x_1 + x_2 \leq 0 \implies |x_1| \leq |x_2|$  makes the above less than  $x_1 + x_2 = 0$ , i.e.,  $|x_1| = |x_2|$ , therefore:

$$\max_{f_r^w, f_g^w} [\mathbf{P}^\top \mathbf{F}] = \max_{f_r^w} \int_{\mathbf{x}} [p_r(\mathbf{x})f_r^w(\mathbf{x}) + p_g(\mathbf{x}) - f_r^w(\mathbf{x})]$$

- ▶ substitute  $f \equiv f_r^w(x) = -f_g^w(\mathbf{x})$ :

$$\max_{f_r^w, f_g^w} [\mathbf{P}^\top \mathbf{F}] = \max_f \int_{\mathbf{x}} [p_r(\mathbf{x})f(\mathbf{x}) - p_g(\mathbf{x})f(\mathbf{x})] \quad \forall f(\mathbf{x})$$

$$\begin{aligned}\Rightarrow \mathcal{W}(p_g, p_r) &= \max_{\substack{f \\ \|f\|_L \leq 1}} \left\{ \int [p_r(\mathbf{x})f(\mathbf{x}) - p_g(\mathbf{x})f(\mathbf{x})] d\mathbf{x} \mid f(\mathbf{x}) - f(\mathbf{y}) \leq d(\mathbf{x}, \mathbf{y}) \right\} \\ &= \max_{\substack{f \\ \|f\|_L \leq 1}} \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})}[f(\mathbf{x})]\end{aligned}$$

- put all together:

$$\begin{aligned}\mathcal{W}(p_g, p_r) &= \max_{\substack{f \\ \|f\|_L \leq 1}} \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})}[f(\mathbf{x})] \\ \Rightarrow \mathcal{L}(G, f) &= \min_G \left[ \max_{\substack{f \\ \|f\|_L \leq 1}} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}[f(G_\theta(\mathbf{z}))] \right]\end{aligned}$$

- in words, the discriminator/critic is try to find a 1-Lipschitz function  $f$  that best aligns with real data from  $p_r$  and aligns poorly with generated data  $p_g$

# Apply Minimax theorem to WGAN formulation

$$\begin{aligned} W(p_r, p_\theta) &= \inf_{\gamma \in \pi} \mathbb{E}_{x, y \sim \gamma} [\|x - y\|] \\ &= \inf_{\gamma} \mathbb{E}_{x, y \sim \gamma} [\|x - y\| + \underbrace{\sup_f \mathbb{E}_{s \sim p_r} [f(s)] - \mathbb{E}_{t \sim p_\theta} [f(t)] - (f(x) - f(y))}_{\begin{cases} 0, & \text{if } \gamma \in \pi \\ +\infty & \text{else} \end{cases}}] \end{aligned}$$

as  $x, y \not\sim \gamma \implies \mathbb{E}_{s \sim p_r} [f(s)] \neq f(x)$  and  $\mathbb{E}_{t \sim p_\theta} [f(t)] \neq f(y)$ , can apply some extreme  $f$  to make it  $\infty$

$$\begin{aligned} &= \inf_{\gamma} \sup_f \mathbb{E}_{x, y \sim \gamma} [\|x - y\| + \mathbb{E}_{s \sim p_r} [f(s)] - \mathbb{E}_{t \sim p_\theta} [f(t)] - (f(x) - f(y))] \\ &= \sup_f \inf_{\gamma} \mathbb{E}_{x, y \sim \gamma} [\|x - y\| + \mathbb{E}_{s \sim p_r} [f(s)] - \mathbb{E}_{t \sim p_\theta} [f(t)] - (f(x) - f(y))] \end{aligned}$$

can swap inf, sup due to convex-concave

$$\begin{aligned} &= \sup_f \mathbb{E}_{s \sim p_r} [f(s)] - \mathbb{E}_{t \sim p_\theta} [f(t)] + \underbrace{\inf_{\gamma} \mathbb{E}_{x, y \sim \gamma} [\|x - y\| - (f(x) - f(y))]}_{\begin{cases} 0, & \text{if } \|f\|_L \leq 1 \\ -\infty & \text{else} \end{cases}} \end{aligned}$$

in the case of  $\|f\|_L \leq 1$ :

$$\begin{aligned} \|f(x_1) - f(x_2)\| &\leq \underbrace{K}_{=1} \|x_1 - x_2\| \\ \implies \|x_1 - x_2\| &\geq (f(x_1) - f(x_2)) \\ \implies \|x_1 - x_2\| - (f(x_1) - f(x_2)) &\geq 0 \end{aligned}$$

think  $4 - 3 > 0$  and  $4 - (-3) > 0$



- ▶ remaining question is about  $L$ -Lipschitz function:

$$\max_{f, \|f\|_L \leq 1} \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[f(\mathbf{x})]$$

- ▶ the key is to know why:

$$L = \max_x |f'(x)|$$

- ▶ i.e., a differentiable function  $f$  is  $L$ -Lipschitz if and only if it has gradients with norm at most  $L$  everywhere.
- ▶ we can then do both Gradient-Clipping and Gradient-Penalty!

# why $L$ -Lipschitz $f$ has gradients with norm at most $L$ everywhere

- ▶ for  $L$ -Lipschitz function in general, i.e., include non-convex  $f$ :
- ▶ Given  $x < y$  in interval  $(a, b)$ , (prove the case of  $y < x$  is equally easy):

$$\begin{aligned} |f(x) - f(y)| &= \underbrace{\left| \int_x^y f'(t) dt \right|}_{|a+b| \leq |a| + |b|} \leq \int_x^y |f'(t)| dt \\ &\leq \max_{t \in [x, y]} |f'(t)| \int_x^y 1 dt = \underbrace{\max_{t \in [x, y]} |f'(t)|}_L |x - y| \end{aligned}$$

- ▶ we conclude that:

$$|f(x) - f(y)| \leq L|x - y| \implies L = \max_{t \in [x, y]} |f'(t)|$$

# Ensure function $f$ is 1-Lipschitz: Weight Clipping

- ▶ since the the weights  $w$  are written as  $w^\top \mathbf{x}$  in neural network, derivative w.r.t input  $\mathbf{x}$   $\frac{\partial \mathcal{W}}{\partial \mathbf{x}}$  will be in terms of  $w$ , so:
- ▶ need to limit all weights  $w_i \in [-c, c]$

# Ensure function $f$ is 1-Lipschitz: Weight Clipping Gradient Penalty

- ▶ since largest of gradient of a 1-Lipschitz function  $\nabla$ ,

$$\mathcal{W}_{\text{GP}} = \underbrace{\mathbb{E}_{\tilde{x} \sim p_g}[f(\tilde{x})] - \mathbb{E}_{x \sim p_r}[f(x)]}_{\text{critic loss}} + \underbrace{\lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}} \left[ (\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1)^2 \right]}_{\text{Gradient Penalty}}$$

- ▶ the above **critic loss** is a minimization instead of maximization, so we switched the term around, i.e., instead of:

$$\mathbb{E}_{x \sim p_r}[f(x)] - \mathbb{E}_{\tilde{x} \sim p_g}[f(\tilde{x})]$$

where

$$\hat{x} = t\tilde{x} + (1 - t)x \quad 0 \leq t \leq 1$$

# Lipschitz property and norms of matrix parameters neural networks

- ▶ what if we add some norm based regularizer to the matrix parameter  $\|W\|$ ?
- ▶ when kind of  $L$ -Lipschitz does it correspond to?

# Lipschitz property for Neural Networks

- ▶ given  $f = \sigma(W^\top \mathbf{x} + b)$ , we may want to have a look at what  $L$ -Lipschitz is this?

$$\begin{aligned}\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| &\leq L\|\mathbf{x}_1 - \mathbf{x}_2\| \\ \implies \|\sigma(W^\top \mathbf{x}_1 + b) - \sigma(W^\top \mathbf{x}_2 + b)\| &\leq L\|\mathbf{x}_1 - \mathbf{x}_2\|\end{aligned}$$

Let

$$\begin{aligned}f(\mathbf{x}) &\approx \nabla_{\mathbf{x}} f(\mathbf{x})(\mathbf{x}_1 - \mathbf{x}_2) && \text{where } \mathbf{x}_1 \leq \mathbf{x} \leq \mathbf{x}_2 \\ &= \nabla_{\mathbf{x}} \sigma(W^\top \mathbf{x} + b)(\mathbf{x}_1 - \mathbf{x}_2) && \text{and } \nabla_{\mathbf{x}} \sigma(W^\top \mathbf{x} + b) = \sigma'(\underbrace{W^\top \mathbf{x} + b}_z) \times \underbrace{W}_{\frac{dz}{d\mathbf{x}}} \\ &= \sigma'(W^\top \mathbf{x} + b)W(\mathbf{x}_1 - \mathbf{x}_2)\end{aligned}$$

$\sigma'(W^\top \mathbf{x} + b)$  can be chosen to be bounded!

- ▶ so we need to look at:

$$\begin{aligned}\|W^\top(\mathbf{x}_1 - \mathbf{x}_2)\| &\leq L\|\mathbf{x}_1 - \mathbf{x}_2\| \\ \text{wlof : } \|W(\mathbf{x}_1 - \mathbf{x}_2)\| &\leq L\|\mathbf{x}_1 - \mathbf{x}_2\|\end{aligned}$$

► definition:

$$\begin{aligned}\|W\|_F &= \sqrt{\left(\sum_{i,j=1}^n |w_{ij}|^2\right)} \\ &= \sqrt{\text{tr}(WW^\top)} = \sqrt{\text{tr}(W^\top W)} \\ &= \text{is the L2 regularizer!}\end{aligned}$$

► it's a matrix norm, therefore:

$$\|WB\|_F \leq \|W\|_F \|B\|_F$$

► unitary invariant, for all unitary vector,  $U$  and  $V$ , where  $U^\top = U^{-1}$

$$\|W\|_F = \|UW\|_F = \|WV\|_F = \|UWV\|_F$$

► can prove the following:

$$\|W\|_2 = \sqrt{\sigma_{\max}(W^\top W)} \leq \|W\|_F = \sqrt{n} \sqrt{\sigma_{\max}(W^\top W)}$$

► Frobenius norm is an upper-bound of spectral norm!

- ▶ using cauchy schwarz:

$$\begin{aligned}\|W\mathbf{x}\|^2 &= \sum_{i=1}^m \left| \sum_{j=1}^n w_{ij} x_j \right|^2 \leq \sum_{i=1}^m \left\{ \left( \sum_{j=1}^n |w_{ij}|^2 \right) \left( \sum_{j=1}^n |x_j|^2 \right) \right\} \\ &= \left( \sum_{i=1}^m \sum_{j=1}^n |w_{ij}|^2 \right) \|\mathbf{x}\|^2 \\ &= \|W\|_F^2 \|\mathbf{x}\|^2 \\ \implies \|W\mathbf{x}\| &\leq \|W\|_F \|\mathbf{x}\| \quad \forall \mathbf{x} \\ \implies \|W(\mathbf{x}_1 - \mathbf{x}_2)\| &\leq \underbrace{\|W\|_F}_L \|\mathbf{x}_1 - \mathbf{x}_2\|\end{aligned}$$

- ▶ adding  $\|W\|_F^2$ , a.k.a, L2 regularizer helps with neural network with a  $(L = \|W\|_F)$ -Lipschitz, but it may not be tight enough
- ▶ since  $\|W\|_2 = \sqrt{\sigma_{\max}(W^T W)} \leq \|W\|_F$ , let's see if we can use  $L = \|W\|_2$ , aka, spectral norm



- ▶ Given a linear function  $f_z(\cdot)$ , how "big" is its output, i.e., how big is the number  $f_z(x) = z^T x$  relative to the size (norm) of  $x$ ? This is exactly the number:

$$\frac{z^T x}{\|x\|}$$

we need to normalize by  $\|x\|$  to remove the effects of input  $x$

- ▶ We say that norm of  $z$  is the largest this quantity can possibly be:

$$\|z\|_* = \sup_{x \neq 0} \frac{z^T x}{\|x\|}$$

- ▶ or more generically:

$$\underbrace{\|z\|_*}_{\text{dual norm}} = \sup \left\{ x^T z \mid \underbrace{\|x\|}_{\text{"ordinary" norm}} \leq 1 \right\}$$

- ▶ Dual norm of  $L_2$  norm is the  $L_2$  norm. Dual norm of  $L_1$  norm is  $L_\infty$  norm

- Dual norm of  $L_2$  norm is the  $L_2$  norm:

$$\sup\{z^\top x \mid \|x\|_{L_2} \leq 1\} = \|z\|_{L_2}$$

max occurs when  $x$  is a unit vector pointing in the same direction as  $z$

- Dual norm of  $L_1$  norm is  $L_\infty$  norm and vice versa:

$$\sup\{z^\top x \mid \underbrace{\|x\|_{L_\infty}}_{\max(|x_1|, \dots, |x_n|)} \leq 1\} = \|z\|_{L_2}$$

max occurs when  $x$  is in corner of a square where signs of each dimension matches between  $z$  and  $x$

for example,  $z = (-5, 5)^\top \implies x = (-1, 1)$

# Matrix norm: p-norm vector

1. in general:

$$\begin{aligned}\|A\|_p &= \sup_{\|x\| \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \\ &= \sup\{\|Ax\|_p \mid \|x\|_p = 1\}\end{aligned}$$

2.  $p = 1$ :

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

the “chosen”  $x$  will be a one hot vector: like a column selector to find a column with max sum of absolute value

3.  $p = \infty$ :

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

the “chosen”  $x$  will be a vector of  $\{+1, -1\}$  to suit the row with max sum of absolute values

4.  $p = 2$ : **spectral norm**

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\lambda_{\max}(AA^T)}$$

## 1. $p = 2$ : spectral norm

$$\begin{aligned}\|A\|_2^2 &= \sup_{\|x\|_2=1} \|Ax\|_2^2 \\&= \sup_{\|x\|_2=1} (x^T A^T A x) \\&= \max_{\|x\|_2=1} x^T U \operatorname{diag}(\lambda_1, \dots, \lambda_n) V^T x \\&= \max_{\|x\|_2=1} x^T \operatorname{diag}(\lambda_1, \dots, \lambda_n) x \\&= \max_{\|x\|_2=1} \lambda_1 x_1^2 + \dots + \lambda_n x_n^2 \\&= \max\{\lambda_1, \dots, \lambda_n\} \text{ the chosen } x \text{ is when } (x_1^2, \dots, x_n^2) \text{ is a one hot corresponding to largest } \lambda \\&= \lambda_{\max}(A^T A)\end{aligned}$$

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\lambda_{\max}(AA^T)}$$

# Compute Spectral Norm

- ▶ compute  $\sigma_{\max}(\mathbf{A}^\top \mathbf{A})$  is hard!
- ▶ however, we can approximate it by:

repeat :

$$u \leftarrow \frac{(\mathbf{A}^\top \mathbf{A})u}{\|(\mathbf{A}^\top \mathbf{A})u\|}$$

$$\|\mathbf{A}\|_2^2 \approx u^\top \mathbf{A}^\top \mathbf{A} u$$

- ▶ or

repeat :

$$v \leftarrow \frac{\mathbf{A}^\top u}{\|\mathbf{A}^\top u\|}, \quad u \leftarrow \frac{\mathbf{A} v}{\|\mathbf{A} v\|}$$

$$\|\mathbf{A}\|_2^2 \approx u^\top \mathbf{A}^\top \mathbf{A} v$$

- ▶ why it works?

# Why it works?

- ▶ this is very similar to Power Method:  
[https://github.com/roboticcam/machine-learning-notes/blob/master/stochastic\\_matrices.pdf](https://github.com/roboticcam/machine-learning-notes/blob/master/stochastic_matrices.pdf)
- ▶ however, this time,  $\lambda_{\max}(K \equiv \mathbf{A}^\top \mathbf{A}) \neq 1$ !
- ▶ but the same can still apply:

$$\begin{aligned} \mathbf{u}^{(0)} &= c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n \\ \implies K^t \mathbf{u}^{(0)} &= c_1 K^t \mathbf{v}_1 + \cdots + c_n K^t \mathbf{v}_n \\ &= c_1 \lambda_1^t \mathbf{v}_1 + \cdots + c_n \lambda_n^t \mathbf{v}_n \\ &\approx c_1 \lambda_1^t \mathbf{v}_1 \end{aligned}$$

means  $K^t \mathbf{u}^{(0)}$  gives a good approximation to un-normalized  $\mathbf{v}_1$

- ▶ which we can see the first term dominates! However, it may grow significantly big! We therefore, need a normalization term:

$$\tilde{\mathbf{v}}_1 \leftarrow \frac{K \mathbf{u}}{\|K \mathbf{u}\|}$$

- ▶ finally

$$\begin{aligned} A \tilde{\mathbf{v}}_1 &= \lambda_1 \tilde{\mathbf{v}}_1 \\ \implies \tilde{\mathbf{v}}_1^\top A \tilde{\mathbf{v}}_1 &= \lambda_1 \tilde{\mathbf{v}}_1^\top \tilde{\mathbf{v}}_1 = \lambda_1 = \|\mathbf{A}\|_2^2 \end{aligned}$$

# Spectral Norm for $L$ -Lipschitz

$$\begin{aligned}\|W\mathbf{x}\|_2 &= \max_{\mathbf{x} \neq 0} \frac{\|W\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\ \Rightarrow \|W\mathbf{x}\|_2 &\leq \|W\|_2 \|\mathbf{x}\|_2 \\ \|W\mathbf{x}\| &\leq \|W\|_2 \|\mathbf{x}\| \quad \forall \mathbf{x} \text{ drop the } L2 \text{ norm index for vector} \\ \Rightarrow \|W(\mathbf{x}_1 - \mathbf{x}_2)\| &\leq \underbrace{\|W\|_2}_L \|\mathbf{x}_1 - \mathbf{x}_2\|\end{aligned}$$

enforcing  $W$  to keep its norm value closer to  $\|W\|_2$ , makes the function more robust than Frobenius norm!

# Maximum Entropy Generators for Energy-Based Models

- ▶ Enough of W-GAN, talk something new!
- ▶ **Discriminator**

$$\begin{aligned} U &= \arg \min_U \mathbb{E}_{x \sim p(x)} [U(x)] - \mathbb{E}_{x \sim \hat{q}(x)} [U(x)] + \lambda \mathbb{E}_{x \sim p(x)} [\|\nabla_x U(x)\|^2] \\ &= \arg \min_U \mathbb{E}_{x \sim p(x)} [U(x)] - \mathbb{E}_{z \sim q(z)} [U(G(z))] + \lambda \mathbb{E}_{x \sim p(x)} [\|\nabla_x U(x)\|^2] \end{aligned}$$

- ▶ **Generator**

$$G = \arg \min_G \mathbb{E}_{z \sim q(z)} [U(G(z))]$$



- ▶ Original GAN:

$$\min_G \max_D \left( L(D, G) \equiv \mathbb{E}_{x \sim p_r(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \right)$$

$$\implies L(D, G) = \underbrace{\mathbb{E}_{x \sim p_r(x)} [\log D(x)] + \mathbb{E}_{x \sim p_g(x)} [\log(1 - D(x))]}_{V(D, G)}$$

problem is that the  $z$  sampled is not controllable. We need to append it with a code  $c$

- ▶ infoGAN:

$$\min_G \max_D L(D, G) = V(D, G) - \lambda I(c; G(z, c))$$

- ▶  $I(c, \mathbf{x})$  is mutual information, how much we know about  $c$  when we know  $\mathbf{x}$  and vice versa
- ▶ if  $\mathbf{x}$  and  $c$  are completely uncorrelated:  $\implies I(c, \mathbf{x})$  is low
- ▶ if  $\mathbf{x}$  and  $c$  are correlated:  $\implies I(c, \mathbf{x})$  is high

- ▶ Conditional entropy:

$$H(Y|X) = - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)} = - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(y|x)$$

- ▶ note that conditional entropy  $H(Y|X)$  and cross entropy  $H(P||Q)$  are not the same thing!

►  $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ :  $c \equiv Y$  and  $G(z, c) \equiv X$

$$I(c; G(z, c)) = H(c) - H(c|G(z, c))$$

$$= \mathbb{E}_{x \sim G(z, c)} \left[ \underbrace{\mathbb{E}_{c' \sim p(c|x)} [\log P(c'|x)]}_{-H(P)} \right] + H(c)$$

► use variational stuff, and remember, from <https://github.com/roboticcam/machine-learning-notes/blob/master/regression.pdf>:

$$H(P||Q) = H(P) + \text{KL}(P||Q) \implies H(P||Q) \geq H(P) \implies -H(P) \geq -H(P||Q)$$

$$\implies I(c; G(z, c)) \geq \mathbb{E}_{x \sim G(z, c)} \left[ \underbrace{\mathbb{E}_{c' \sim p(c|x)} [\log Q(c'|x)]}_{-H(P||Q)} \right] + H(c)$$

$$\mathcal{L}_I(G, Q) = \mathbb{E}_{c \sim P(c), x \sim G(z, c)} [\log Q(c|x)] + H(c)$$

- ▶ from previous page:

$$\begin{aligned} \mathcal{I}(c; G(z, c)) &\geq \mathbb{E}_{x \sim G(z, c)} \left[ \mathbb{E}_{c' \sim \underbrace{p(c|x)}_{\text{too hard!}}} [\log Q(c'|x)] \right] + H(c) \quad (1) \\ &= \mathcal{L}_I(G, Q) \end{aligned}$$

- ▶ so instead of sample  $p(x, c') = p(x)p(c'|x)$ , we make it  $p(x, c) = p(c)p(x|c)$ :

$$\mathcal{L}_I(G, Q) = \mathbb{E}_{\underbrace{c \sim P(c)}_{\text{easy to sample!}}, x \sim G(z, c)} [\log Q(c|x)] + H(c) \quad (2)$$

# Why sample ① and ② are the same?

- sample ① :  $x \sim p(x)$ , then sample  $y|x$ , then sample back  $x'|y$ . Finally, back and to compute  $f(x', y)$ :

$$\underbrace{E_{x \sim X, y \sim Y|x, x' \sim X|y} [f(x', y)]}_{\text{①}} = \int_x p(x) \int_y p(y|x) \int_{x'} p(x'|y) f(x', y) dx' dy dx$$

$$= \int_y p(y) \int_x p(x|y) \int_{x'} p(x'|y) f(x', y) dx' dx dy$$

$$= \int_y p(y) \int_{x'} p(x'|y) f(x', y) \underbrace{\int_x p(x|y) dx}_{=1} dx' dy$$

$$= \int_y p(y) \int_x p(x|y) f(x, y) dx dy$$

$$= \int_x p(x) \int_y p(y|x) f(x, y) dy dx$$

$$= \underbrace{E_{x \sim X, y \sim Y|x} [f(x, y)]}_{\text{②}}$$

- ② : it has the same effect of sample  $(x, y)$  directly from  $f(x, y)$ , and then to compute  $f(x, y)$

1. sample a noise  $z \sim p(z)$  and  $c \sim p(c)$
2. generate  $\mathbf{x} = G(c, z)$
3.  $D$  differentiates real and fake as usual
4. calculate  $Q(c|\mathbf{x})$

## ► Generator

$$p(\theta_g | \mathbf{z}, \theta_d) \propto \left( \prod_{i=1}^{n_g} D_{\theta_d} \left( G_{\theta_g}(\mathbf{z}^{(i)}) \right) \right) p(\theta_g | \alpha)$$

## ► Discriminator

$$p(\theta_d | \mathbf{z}, \mathbf{X}, \theta_g) \propto \prod_{i=1}^{n_d} D_{\theta_d}(\mathbf{x}^{(i)}) \times \prod_{i=1}^{n_g} \left( 1 - D_{\theta_d}(G_{\theta_g}(\mathbf{z}^{(i)})) \right) \times p(\theta_d | \alpha)$$

►  $p(\theta_g|\theta_d)$

$$\begin{aligned} p(\theta_g|\theta_d) &= \int p(\theta_g, \mathbf{z}|\theta_d) d\mathbf{z} = \int \underbrace{p(\theta_g|\mathbf{z}, \theta_d)}_{\text{independent of } \theta_d} \underbrace{p(\mathbf{z}|\theta_d)}_{\text{independent of } \theta_d} d\mathbf{z} \\ &= \int p(\theta_g|\mathbf{z}, \theta_d) p(\mathbf{z}) d\mathbf{z} \\ &\approx \frac{1}{N} \sum_{i=1}^N p(\theta_g|\mathbf{z}^{(i)}, \theta_d) \quad \mathbf{z}^{(i)} \sim p(\mathbf{z}) \end{aligned}$$

►  $p(\theta_d|\theta_g)$

$$\begin{aligned} p(\theta_d|\theta_g) &\equiv p(\theta_d|\mathbf{X}, \theta_g) = \int_{\mathbf{z}} p(\theta_d, \mathbf{z}|\mathbf{X}, \theta_g) d\mathbf{z} = \int \underbrace{p(\theta_d|\mathbf{z}, \mathbf{X}, \theta_g)}_{\text{independent of } \theta_g} \underbrace{p(\mathbf{z}|\mathbf{X}, \theta_g)}_{\text{independent of } \theta_g} d\mathbf{z} \\ &= \int_{\mathbf{z}} p(\theta_d|\mathbf{z}, \mathbf{X}, \theta_g) p(\mathbf{z}) d\mathbf{z} \\ &\approx \frac{1}{N} \sum_{i=1}^N p(\theta_d|\mathbf{z}^{(i)}, \mathbf{X}, \theta_g) \quad \mathbf{z}^{(i)} \sim p(\mathbf{z}) \end{aligned}$$