Wasserstein GAN

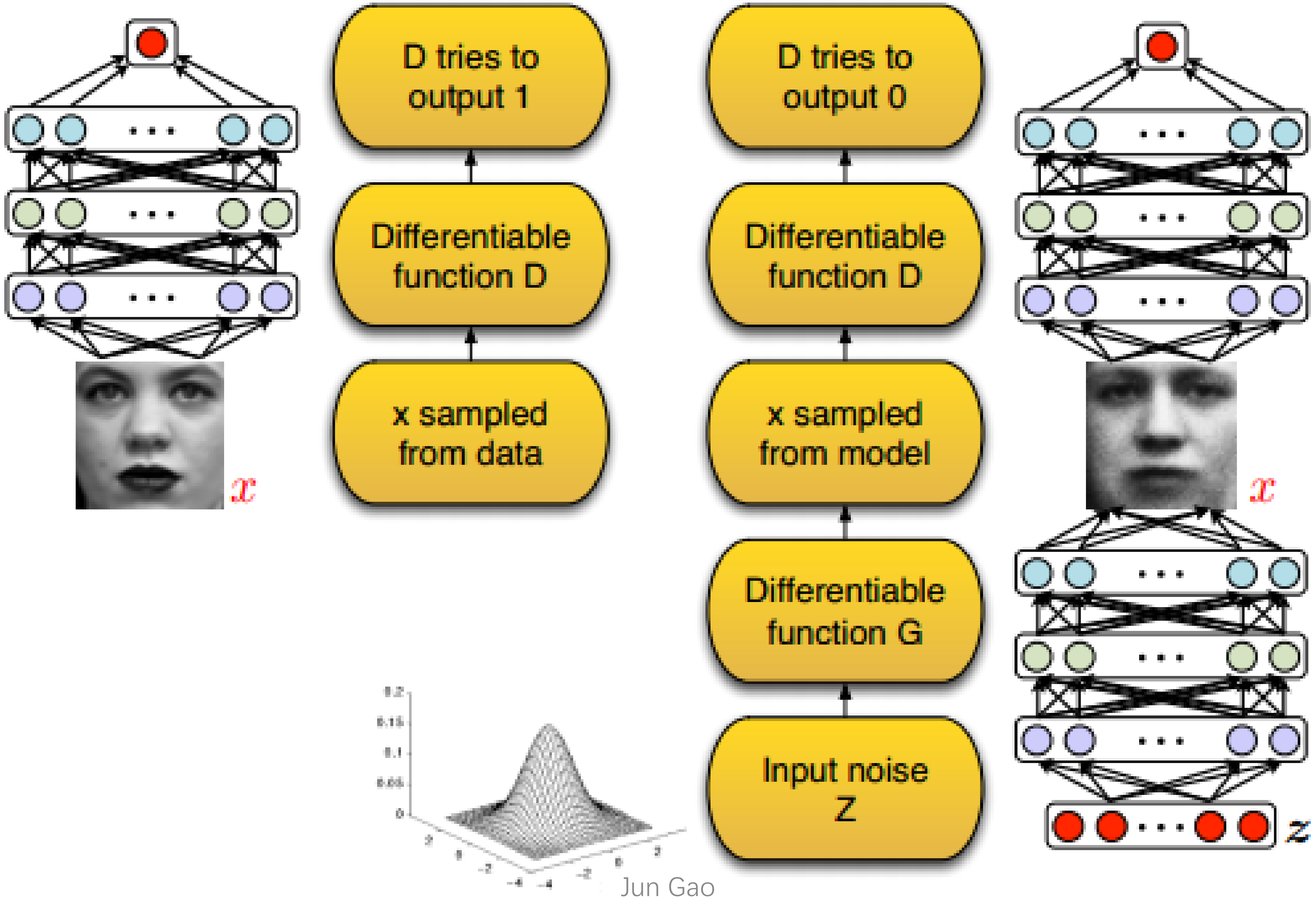
Jun Gao

Outline

- Introduction of GAN
- Training Phenomenon.
- Source of Instability.
- Compare Wasserstein Distance and KL divergence.
- Wasserstein GAN.

1. Towards Principled Methods for
Training Generative Adversarial Networks
-- Martin-ICLR2017 Oral
2. Wasserstein GAN
-- Martin-Arxiv (submitted to ICML)

Introduction-GAN



D tries to output 1

Differentiable function D

x sampled from data

D tries to output 0

Differentiable function D

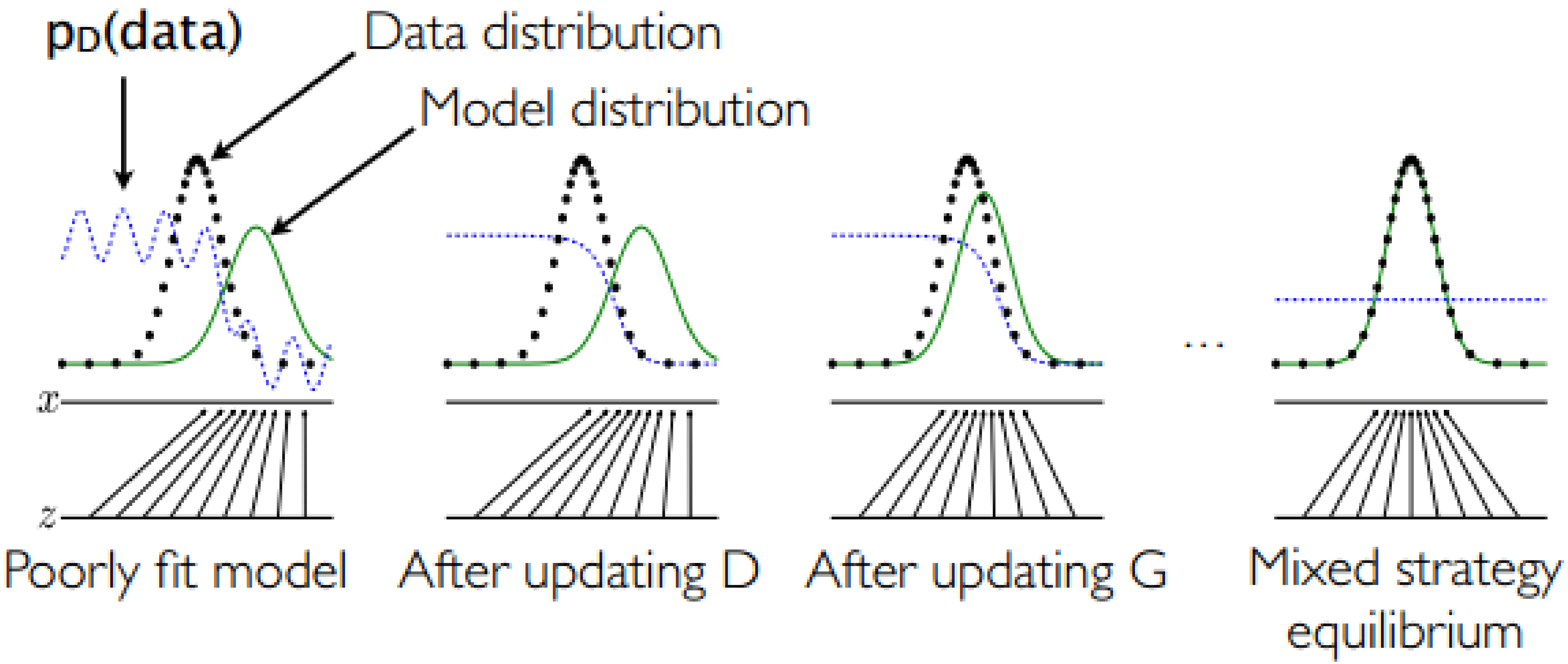
x sampled from model

Differentiable function G

Input noise Z

NIPS 2014 workshop
---Ian Goodfellow

Introduction-Overview



NIPS 2014 workshop
---Ian Goodfellow

Theoretical results

- Min-max Objective Function

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

- Unique global optimum:
  - For G fixed, the optimal discriminator D is :

$$D^*_G(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$$

  - Reformulate training criterion of G:

$$C(G) = -\log(4) + KL\left(p_{\text{data}} \left\| \frac{p_{\text{data}} + p_g}{2}\right.\right) + KL\left(p_g \left\| \frac{p_{\text{data}} + p_g}{2}\right.\right)$$

  - The global minimum of C(G) is achieved if and only if

$$p_g = p_{data}.$$

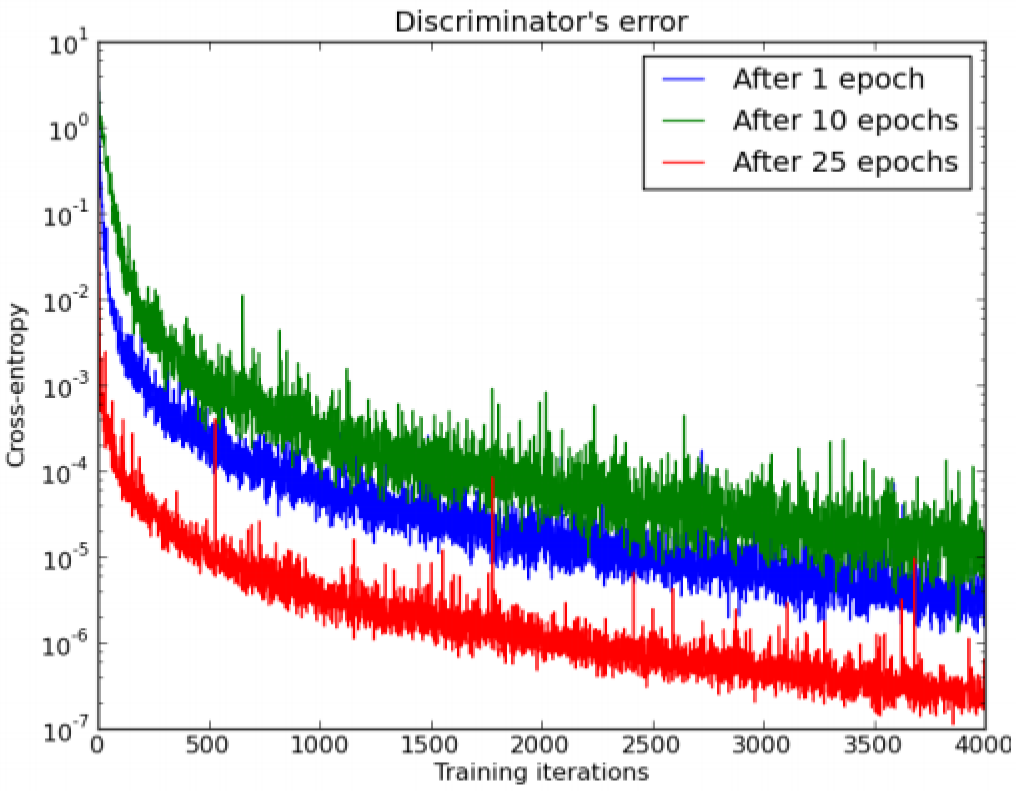
Training Phenomenon

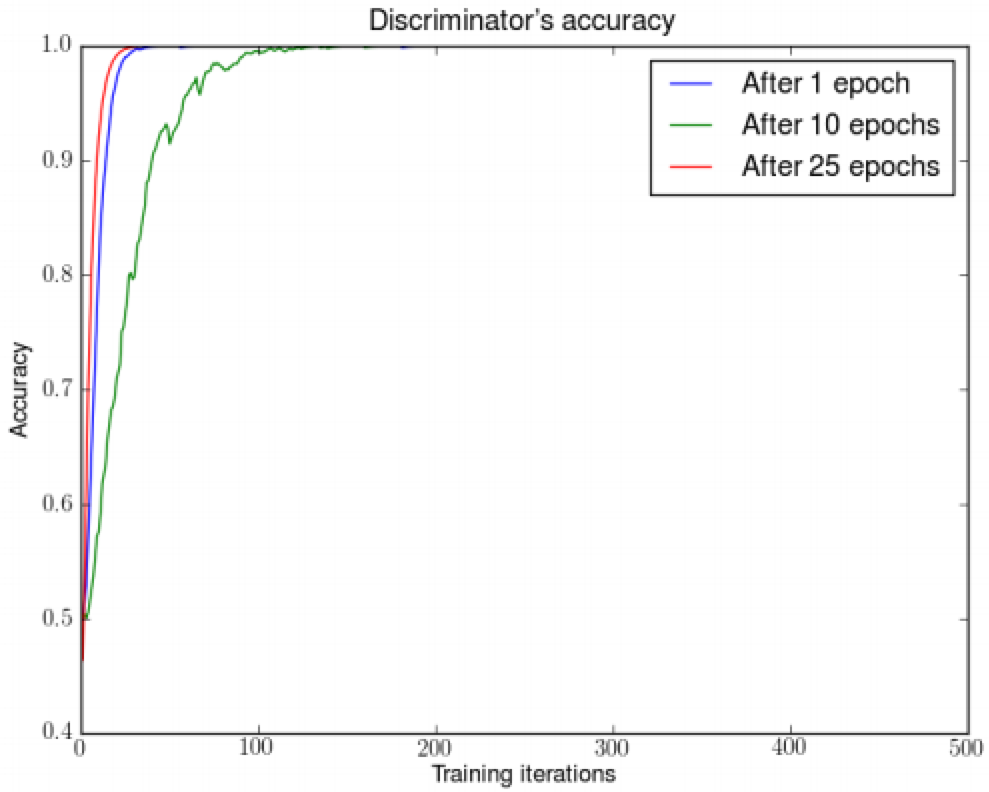
- Hard to train
  - Sensitive to hyperparameters and training paradigm， Even initialization.
- Balance between D and G
  - As D gets better, the updates to G get worse. （Paradox!）
  - Could not train D until optimal.
- Mode collapse

Source of Instability

$$L(D^*, g_\theta) = 2JSD(\mathbb{P}_r \| \mathbb{P}_g) - 2\log 2$$

- **1**. JSD is maxed out

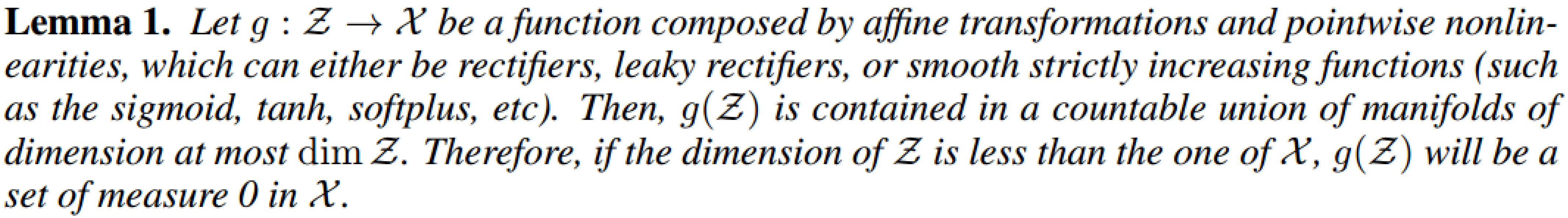




Train DCGAN after 1,10,25 epoch, then train D with G fixed

Source of Instability

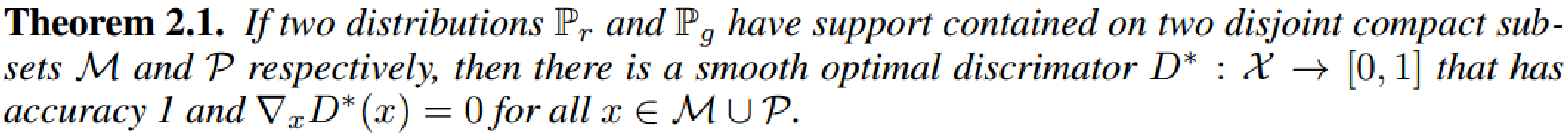
- **2.** Supports of P_r and P_g lies in low dimension manifolds.



Lemma 1. *Let $g : \mathcal{Z} \rightarrow \mathcal{X}$ be a function composed by affine transformations and pointwise nonlinearities, which can either be rectifiers, leaky rectifiers, or smooth strictly increasing functions (such as the sigmoid, tanh, softplus, etc). Then, $g(\mathcal{Z})$ is contained in a countable union of manifolds of dimension at most $\dim \mathcal{Z}$. Therefore, if the dimension of $\mathcal{Z}$ is less than the one of $\mathcal{X}$, $g(\mathcal{Z})$ will be a set of measure 0 in $\mathcal{X}$.*

Source of Instability

- **3.1** The perfect discriminator

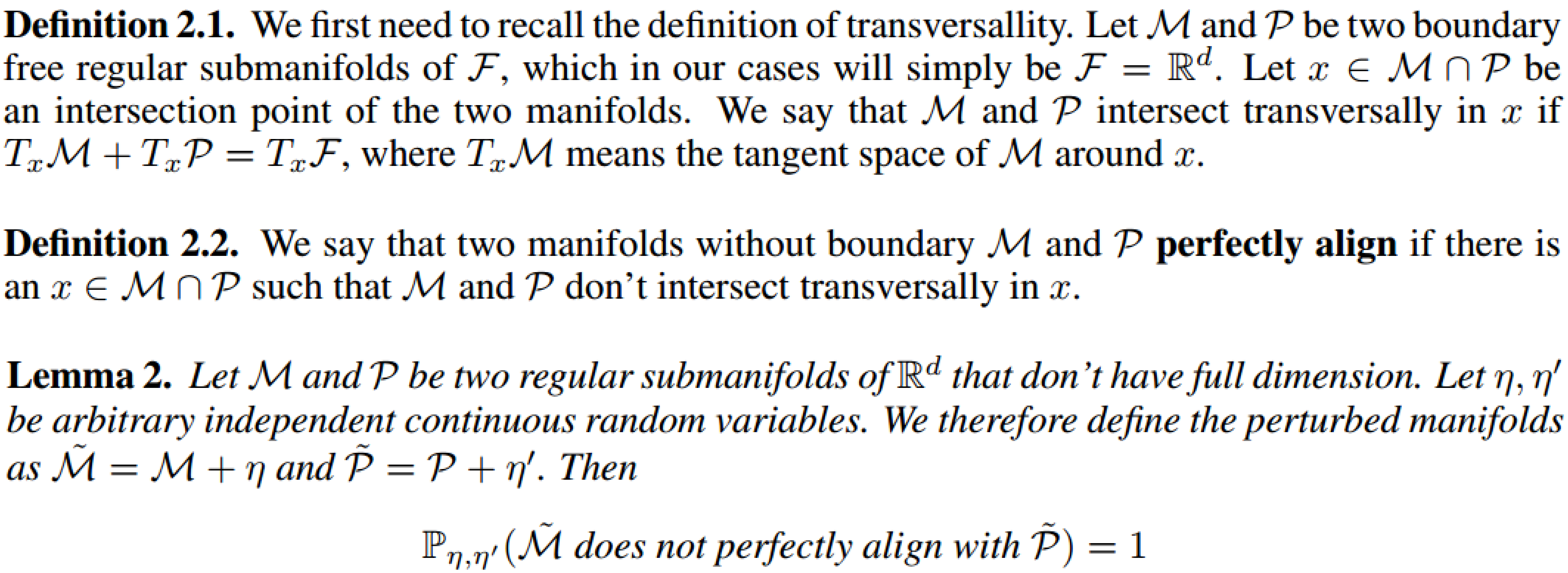


Theorem 2.1. *If two distributions $\mathbb{P}_r$ and $\mathbb{P}_g$ have support contained on two disjoint compact subsets $\mathcal{M}$ and $\mathcal{P}$ respectively, then there is a smooth optimal discriminator $D^* : \mathcal{X} \to [0,1]$ that has accuracy 1 and $\nabla_x D^*(x) = 0$ for all $x \in \mathcal{M} \cup \mathcal{P}$.*

Next? Take away disjoint assumption.

Source of Instability

- **3.2** Supports of $P_r \ and \ P_g$ never perfectly align.



Definition 2.1. We first need to recall the definition of transversallity. Let $\mathcal{M}$ and $\mathcal{P}$ be two boundary free regular submanifolds of $\mathcal{F}$, which in our cases will simply be $\mathcal{F} = \mathbb{R}^d$. Let $x \in \mathcal{M} \cap \mathcal{P}$ be an intersection point of the two manifolds. We say that $\mathcal{M}$ and $\mathcal{P}$ intersect transversally in x if $T_x\mathcal{M} + T_x\mathcal{P} = T_x\mathcal{F}$, where $T_x\mathcal{M}$ means the tangent space of $\mathcal{M}$ around x.

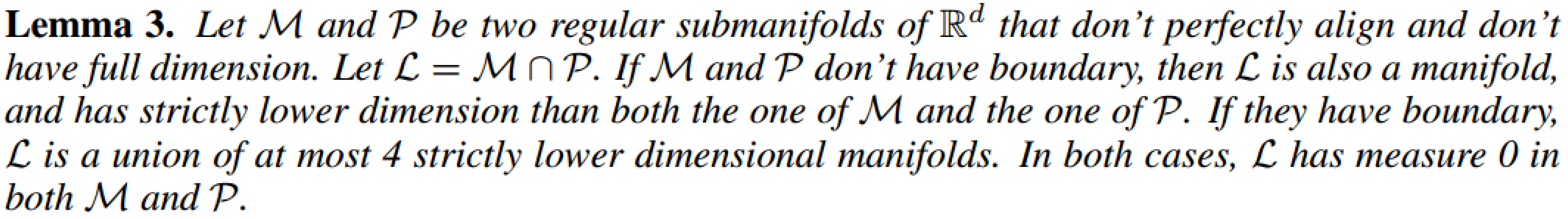
Definition 2.2. We say that two manifolds without boundary $\mathcal{M}$ and $\mathcal{P}$ **perfectly align** if there is an $x \in \mathcal{M} \cap \mathcal{P}$ such that $\mathcal{M}$ and $\mathcal{P}$ don't intersect transversally in x.

Lemma 2. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't have full dimension. Let η, η' be arbitrary independent continuous random variables. We therefore define the perturbed manifolds as $\tilde{\mathcal{M}} = \mathcal{M} + \eta$ and $\tilde{\mathcal{P}} = \mathcal{P} + \eta'$. Then*

$$\mathbb{P}_{\eta,\eta'}(\tilde{\mathcal{M}} \textit{ does not perfectly align with } \tilde{\mathcal{P}}) = 1$$

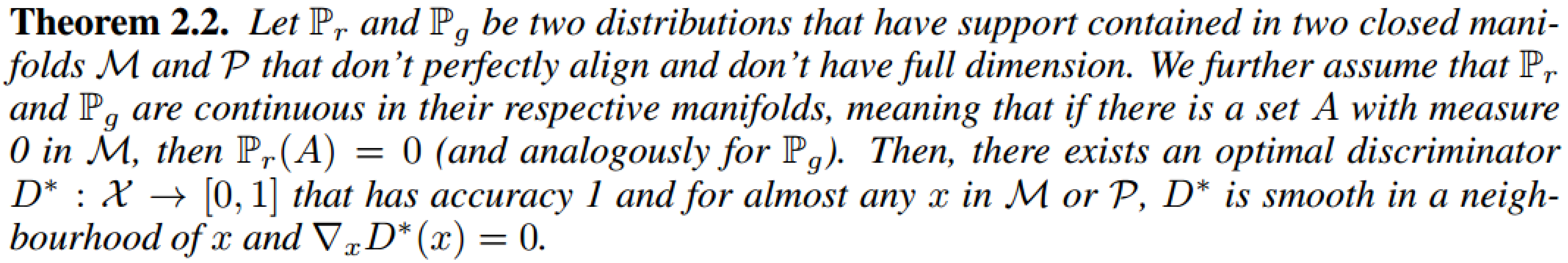
Source of Instability

- **3.3** Union of $P_r \ and \ P_g$ has strictly lower dimension



Lemma 3. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't perfectly align and don't have full dimension. Let $\mathcal{L} = \mathcal{M} \cap \mathcal{P}$. If $\mathcal{M}$ and $\mathcal{P}$ don't have boundary, then $\mathcal{L}$ is also a manifold, and has strictly lower dimension than both the one of $\mathcal{M}$ and the one of $\mathcal{P}$. If they have boundary, $\mathcal{L}$ is a union of at most 4 strictly lower dimensional manifolds. In both cases, $\mathcal{L}$ has measure 0 in both $\mathcal{M}$ and $\mathcal{P}$.*

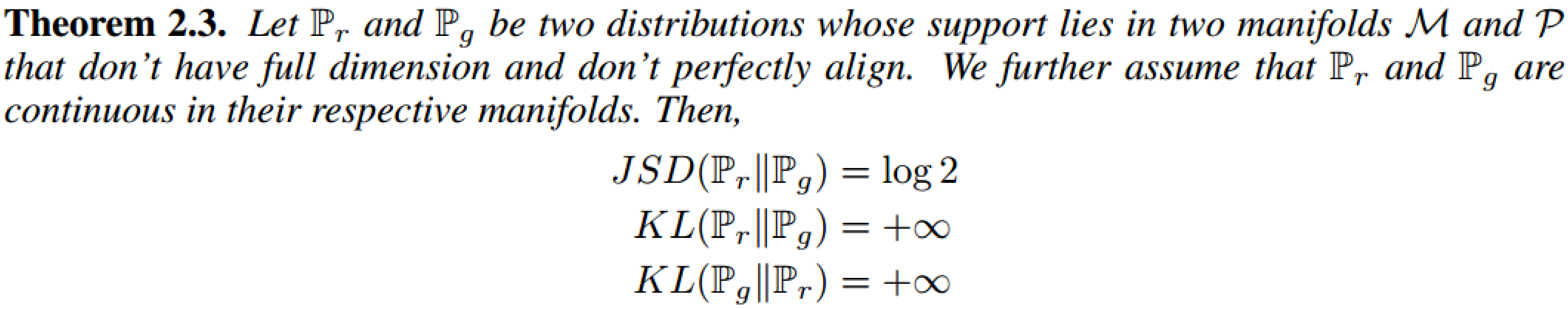
- **3.4.** The perfect discriminator



Theorem 2.2. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions that have support contained in two closed manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't perfectly align and don't have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds, meaning that if there is a set A with measure 0 in $\mathcal{M}$, then $\mathbb{P}_r(A) = 0$ (and analogously for $\mathbb{P}_g$). Then, there exists an optimal discriminator $D^* : \mathcal{X} \to [0, 1]$ that has accuracy 1 and for almost any x in $\mathcal{M}$ or $\mathcal{P}$, D^* is smooth in a neighbourhood of x and $\nabla_x D^*(x) = 0$.*

Source of Instability

- **3.5** Terrible distance measurement.

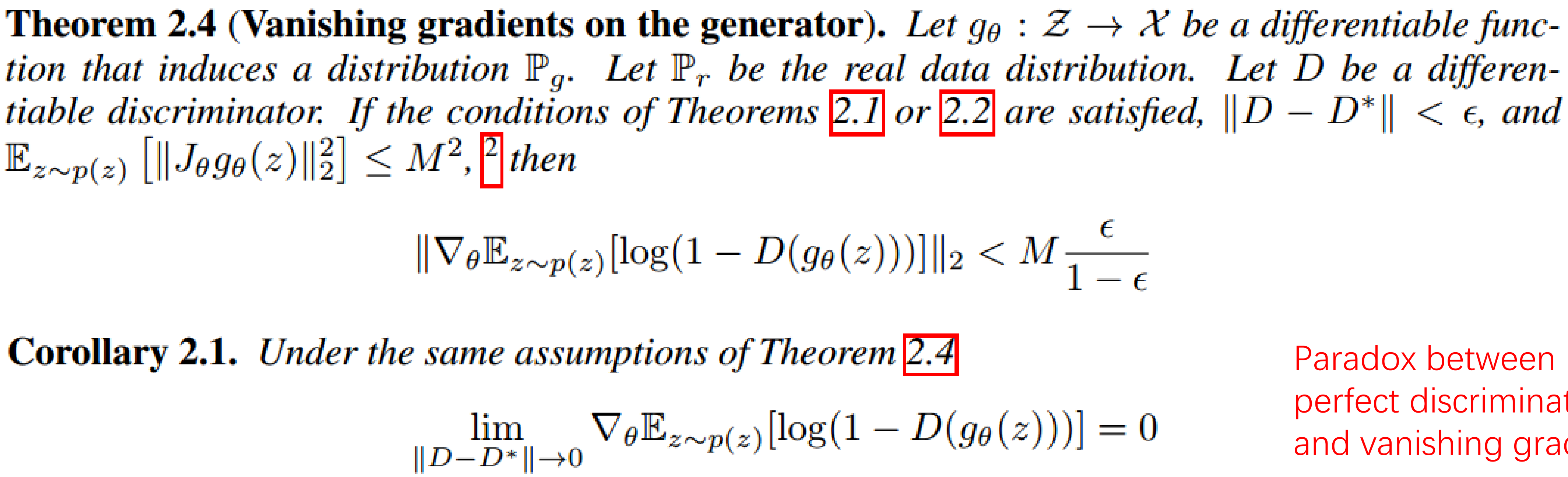


Theorem 2.3. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions whose support lies in two manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't have full dimension and don't perfectly align. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds. Then,*

$$\begin{aligned} JSD(\mathbb{P}_r \| \mathbb{P}_g) &= \log 2 \\ KL(\mathbb{P}_r \| \mathbb{P}_g) &= +\infty \\ KL(\mathbb{P}_g \| \mathbb{P}_r) &= +\infty \end{aligned}$$

Source of Instability

- Theorem 2.1 and 2.2 shows the perfect discriminator whose gradient will be zero almost everywhere.
- **4.** Vanishing gradient



Theorem 2.4 (Vanishing gradients on the generator). *Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a differentiable function that induces a distribution $\mathbb{P}_g$. Let $\mathbb{P}_r$ be the real data distribution. Let D be a differentiable discriminator. If the conditions of Theorems 2.1 or 2.2 are satisfied, $\|D - D^*\| < \epsilon$, and $\mathbb{E}_{z\sim p(z)}\left[\|J_\theta g_\theta(z)\|_2^2\right] \leq M^2$,* [2] *then*

$$\|\nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))]\|_2 < M\frac{\epsilon}{1-\epsilon}$$

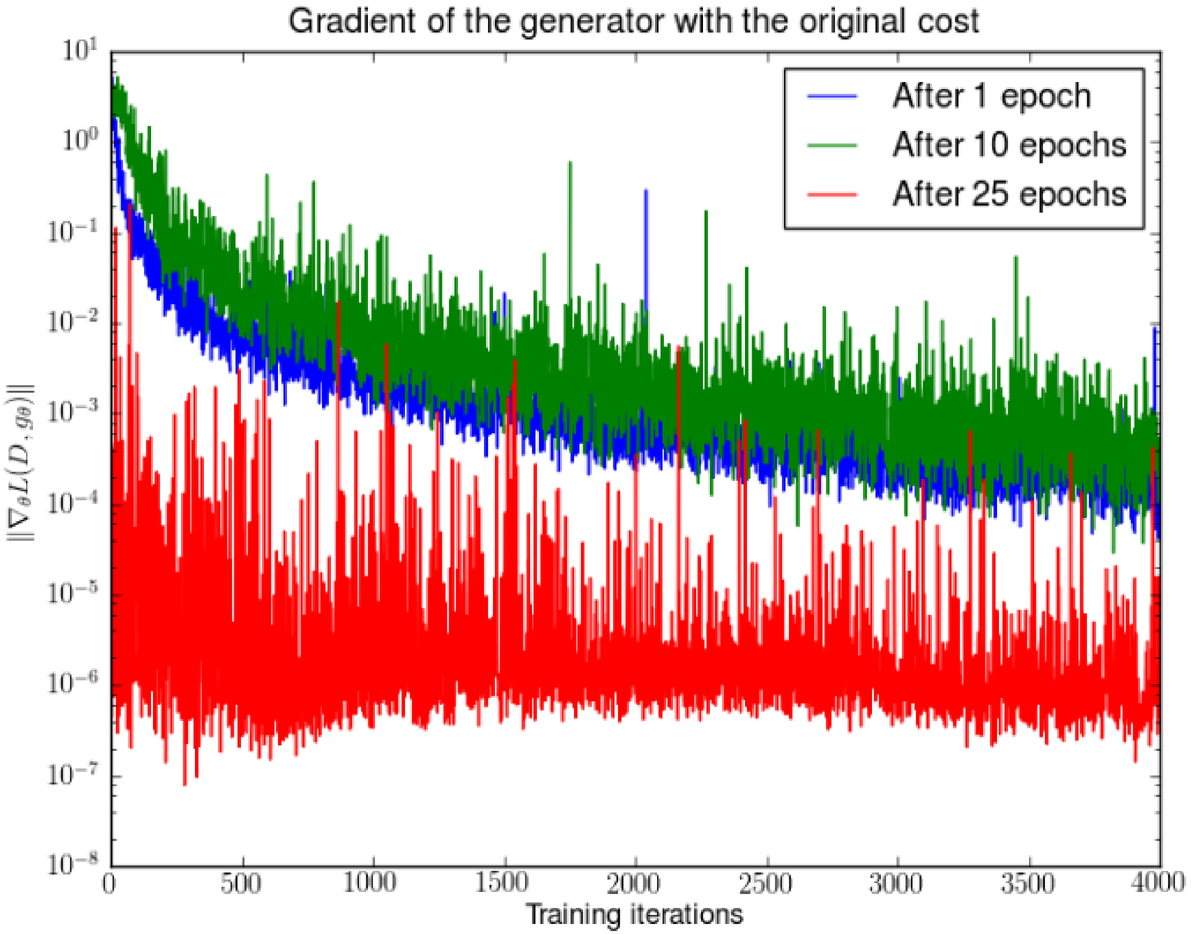
Corollary 2.1. *Under the same assumptions of Theorem 2.4*

$$\lim_{\|D-D^*\|\to 0} \nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))] = 0$$

Paradox between perfect discriminator and vanishing gradient!

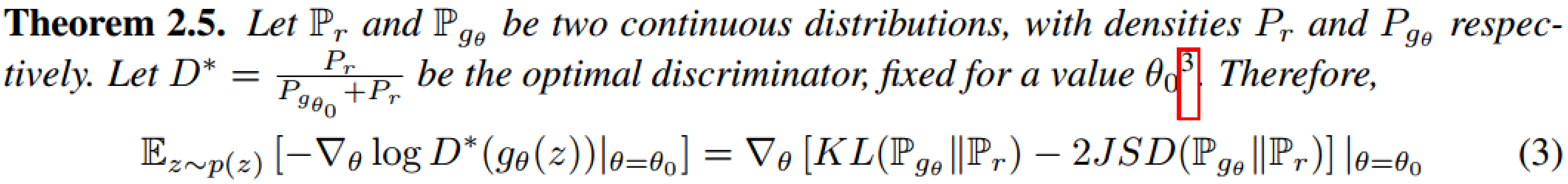
Source of Instability

- **4**. Vanishing Gradient



Source of Instability

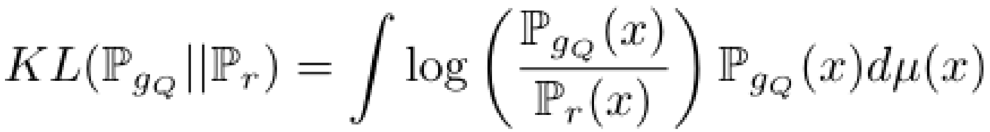
- 5. The $\log D$ alternative



Theorem 2.5. *Let $\mathbb{P}_r$ and $\mathbb{P}_{g_\theta}$ be two continuous distributions, with densities P_r and P_{g_θ} respectively. Let $D^* = \frac{P_r}{P_{g_{\theta_0}} + P_r}$ be the optimal discriminator, fixed for a value θ_0*[3]*. Therefore,*

$$\mathbb{E}_{z\sim p(z)}\left[-\nabla_\theta \log D^*(g_\theta(z))|_{\theta=\theta_0}\right] = \nabla_\theta \left[KL(\mathbb{P}_{g_\theta}\|\mathbb{P}_r) - 2JSD(\mathbb{P}_{g_\theta}\|\mathbb{P}_r)\right]|_{\theta=\theta_0} \qquad (3)$$

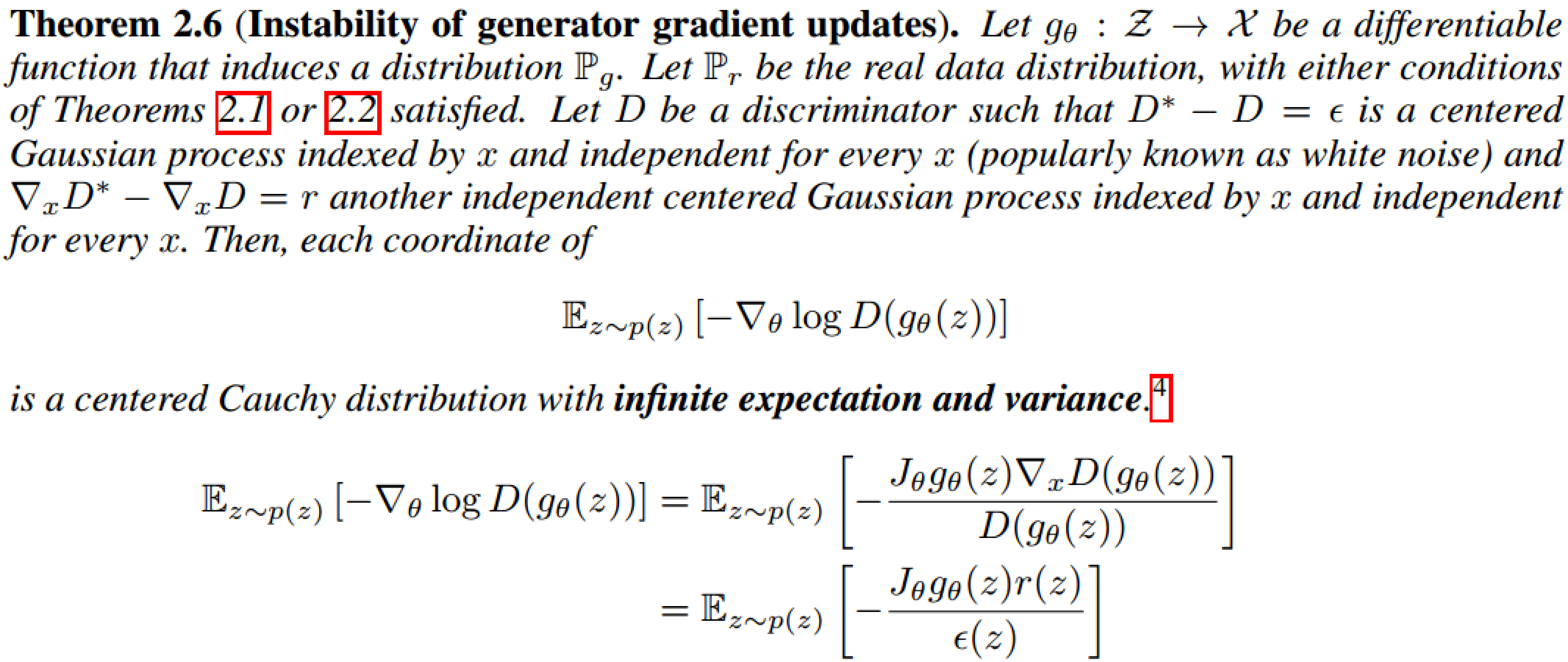
- Assign extremely high cost to generating fake looking examples, while an extremely low cost to mode dropping.



$$KL(\mathbb{P}_{g_Q}||\mathbb{P}_r) = \int \log\left(\frac{\mathbb{P}_{g_Q}(x)}{\mathbb{P}_r(x)}\right)\mathbb{P}_{g_Q}(x)d\mu(x)$$

Source of Instability

- 6. Instability of generator



Theorem 2.6 (Instability of generator gradient updates). *Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a differentiable function that induces a distribution $\mathbb{P}_g$. Let $\mathbb{P}_r$ be the real data distribution, with either conditions of Theorems 2.1 or 2.2 satisfied. Let D be a discriminator such that $D^* - D = \epsilon$ is a centered Gaussian process indexed by x and independent for every x (popularly known as white noise) and $\nabla_x D^* - \nabla_x D = r$ another independent centered Gaussian process indexed by x and independent for every x. Then, each coordinate of*

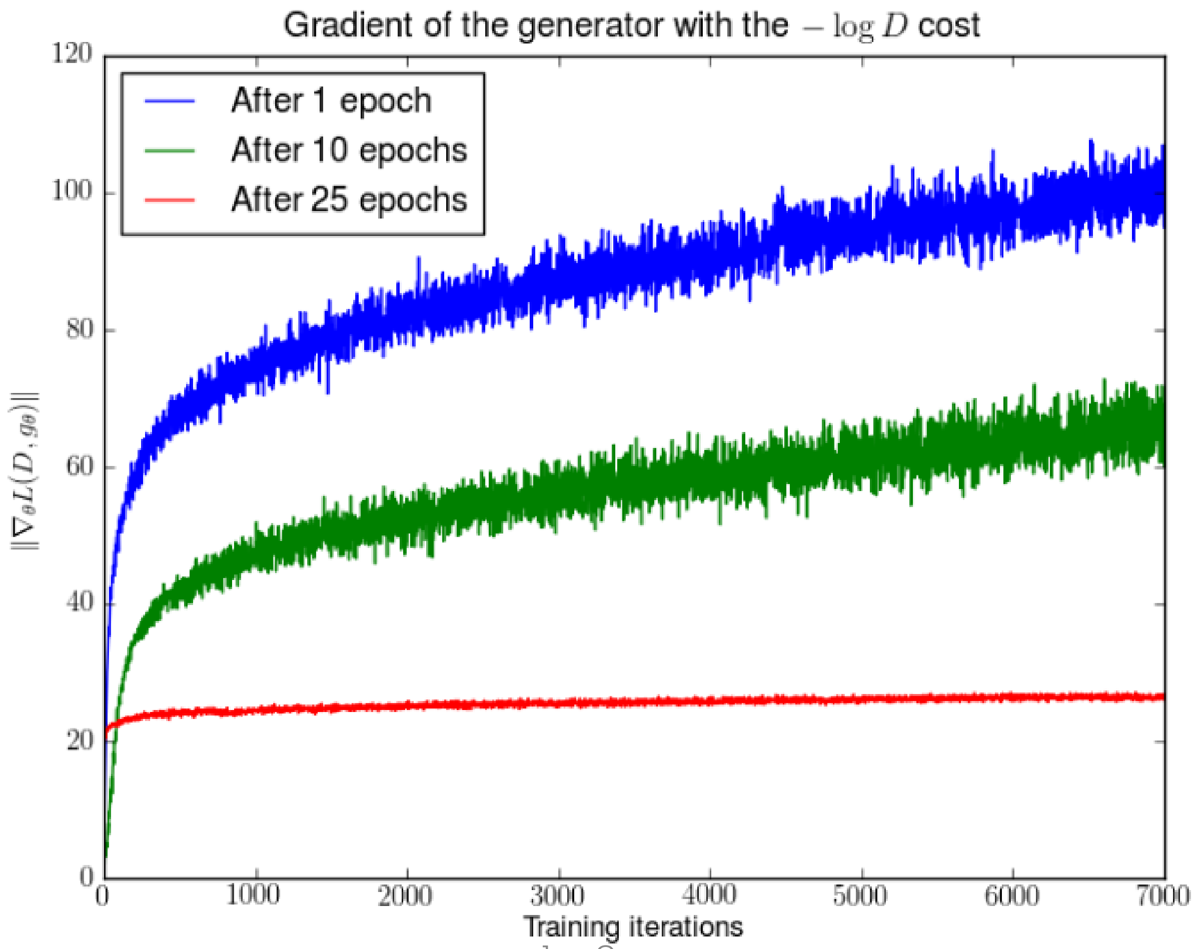
$$\mathbb{E}_{z\sim p(z)}\left[-\nabla_\theta \log D(g_\theta(z))\right]$$

is a centered Cauchy distribution with ***infinite expectation and variance.***[4]

$$\begin{aligned}\mathbb{E}_{z\sim p(z)}\left[-\nabla_\theta \log D(g_\theta(z))\right] &= \mathbb{E}_{z\sim p(z)}\left[-\frac{J_\theta g_\theta(z)\nabla_x D(g_\theta(z))}{D(g_\theta(z))}\right]\\ &= \mathbb{E}_{z\sim p(z)}\left[-\frac{J_\theta g_\theta(z) r(z)}{\epsilon(z)}\right]\end{aligned}$$

Source of Instability

- 6. Instability of generator



Summary

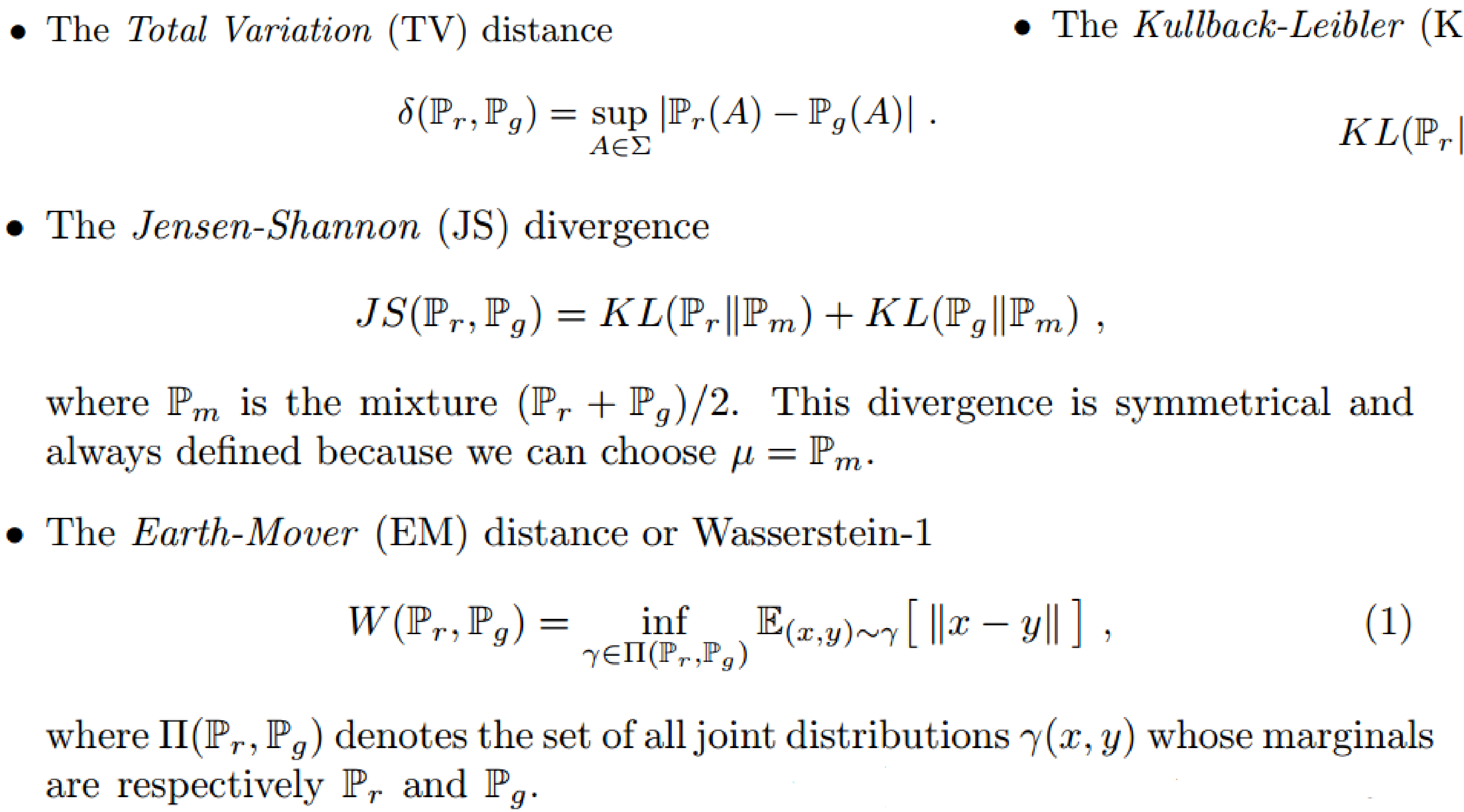
- Vanish gradient
  - Not perfectly align
  - Discriminator with zero gradient.
- Mode dropping
- Infinite variance

Outline

- Introduction of GAN
- Training Phenomenon.
- Source of Instability.
- **Compare Wasserstein Distance and KL divergence.**
- **Wasserstein GAN**.

1. Towards Principled Methods for Training Generative Adversarial Networks
   -- Martin-ICLR2017 Oral
2. Wasserstein GAN
   -- Martin-Arxiv (submitted to ICML)

Compare different distances



- The *Total Variation* (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ .$$

- The *Jensen-Shannon* (JS) divergence

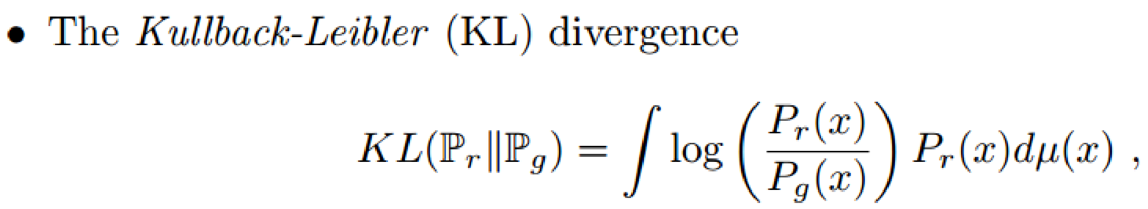
$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ ,$$

where $\mathbb{P}_m$ is the mixture $(\mathbb{P}_r + \mathbb{P}_g)/2$. This divergence is symmetrical and always defined because we can choose $\mu = \mathbb{P}_m$.

- The *Earth-Mover* (EM) distance or Wasserstein-1

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big] \ , \qquad (1)$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$.

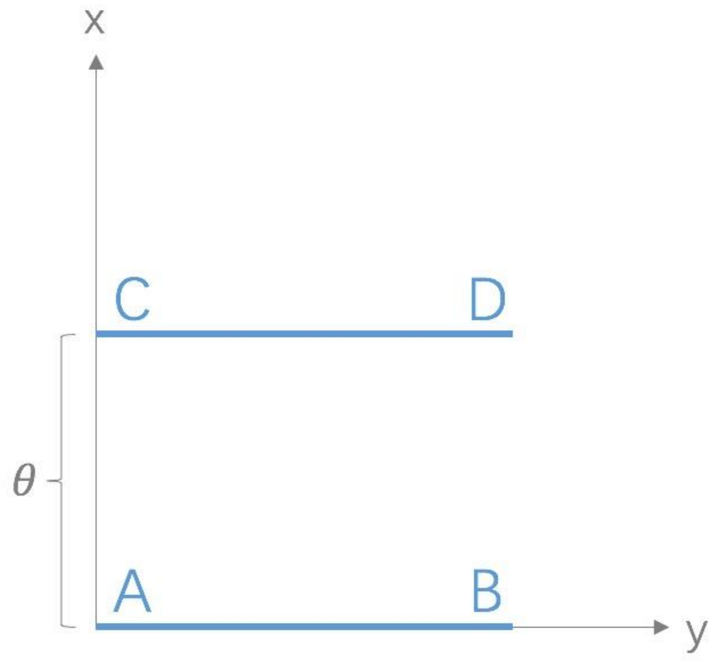


- The *Kullback-Leibler* (KL) divergence

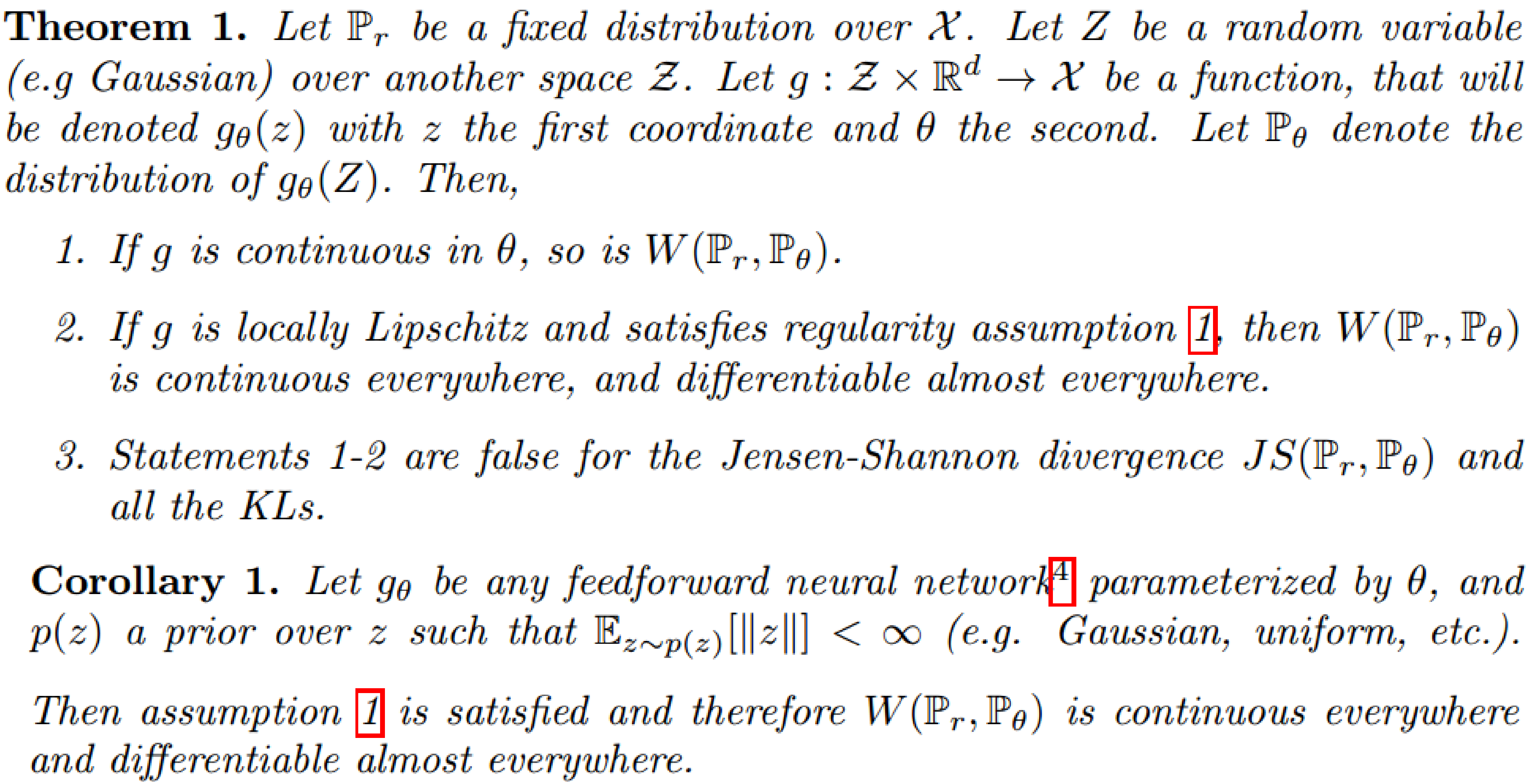
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \ ,$$

Continuity property

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|,$
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$



Continuity property



Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.*
2. *If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

Corollary 1. *Let g_θ be any feedforward neural network[4] parameterized by θ, and $p(z)$ a prior over z such that $\mathbb{E}_{z\sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.).*

Then assumption 1 is satisfied and therefore $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.

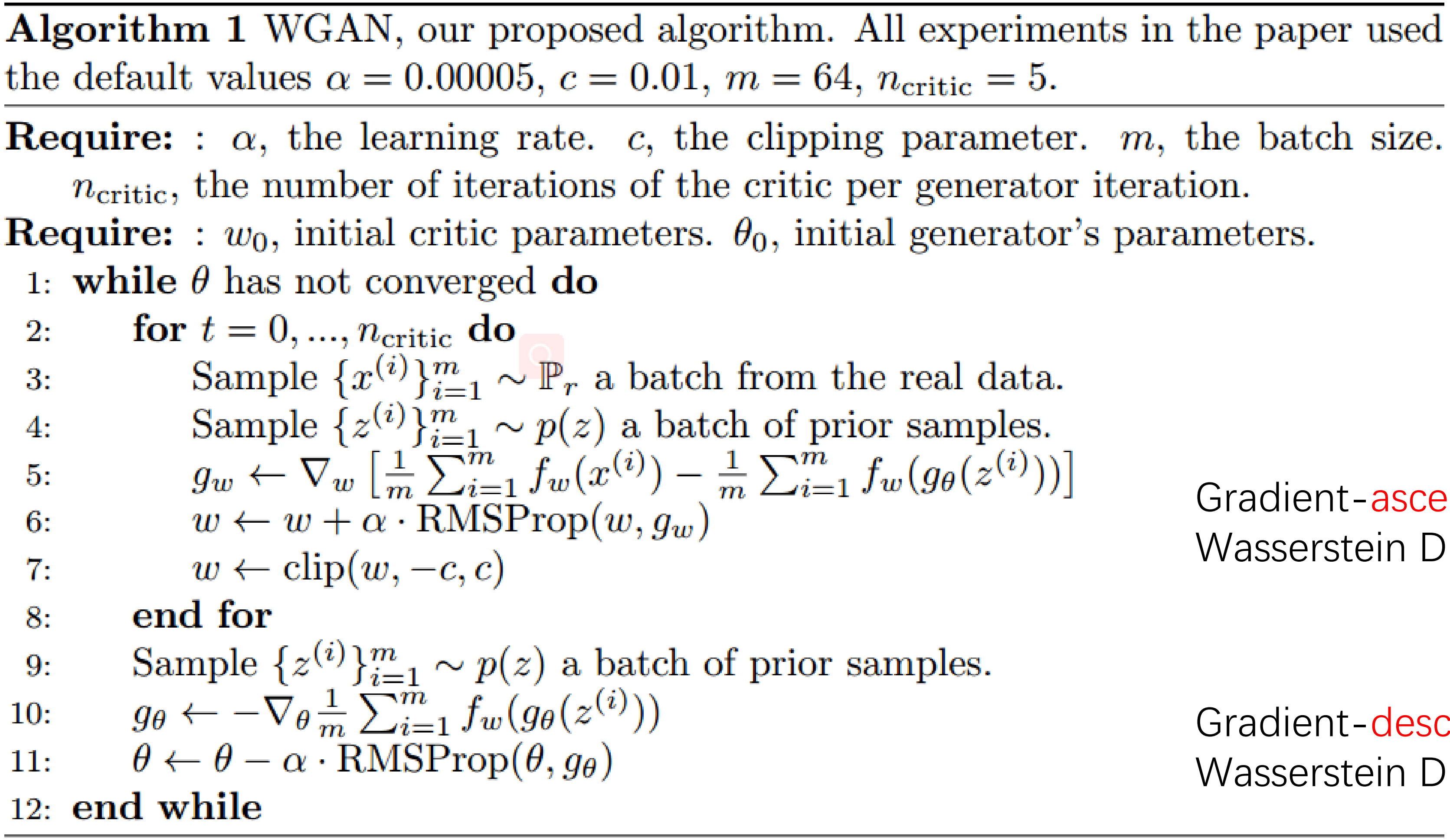
Wasserstein GAN

- Kantorovich-Rubinstein duality

$$W(P_r, P_g) = \frac{1}{K} \sup_{||f||_L \le K} \mathbb{E}_{x\sim P_r}[f(x)] - \mathbb{E}_{x\sim P_g}[f(x)]$$

- Using a neural network to approximate f(x).
- Clamping weights to a fixed box such that K-Lipschitz.

Wasserstein GAN



Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

```
Require: : $\alpha$, the learning rate. $c$, the clipping parameter. $m$, the batch size.
    $n_{\text{critic}}$, the number of iterations of the critic per generator iteration.
Require: : $w_0$, initial critic parameters. $\theta_0$, initial generator's parameters.
while $\theta$ has not converged do
    for $t = 0, ..., n_{\text{critic}}$ do
        Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.
        Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
        $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))\right]$
        $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
        $w \leftarrow \text{clip}(w, -c, c)$
    end for
    Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
    $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$
    $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while
```

Gradient-ascend to approximate Wasserstein Distance

Gradient-descend to minimize Wasserstein Distance.

Changes relative to GAN

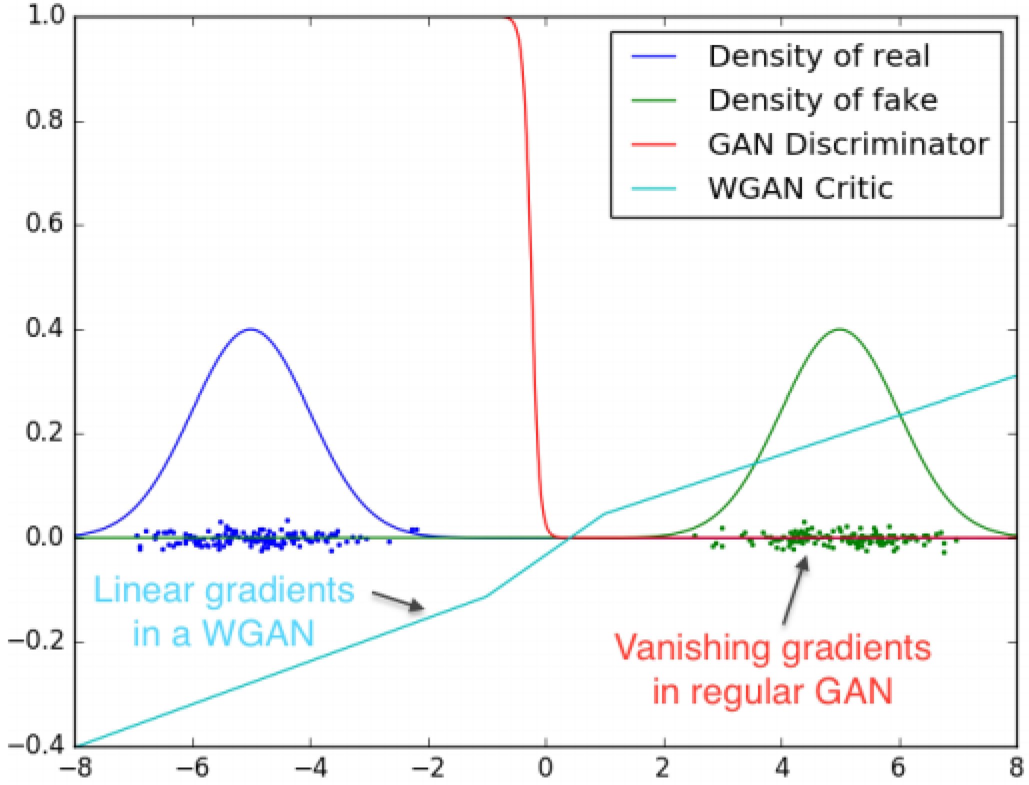
- Remove sigmoid and log loss.
- Clamp weights.
- Replace momentum-based optimizer with RMSProp or SGD.

Comments

- Wasserstein distance is continuous and differentiable a.e.
- We can (and should) train the critic till optimality.
- Meaningful loss metric.
- Not adversarial but a measure of distance. Something like Actor-Critic Policy Gradient Method in RL… …

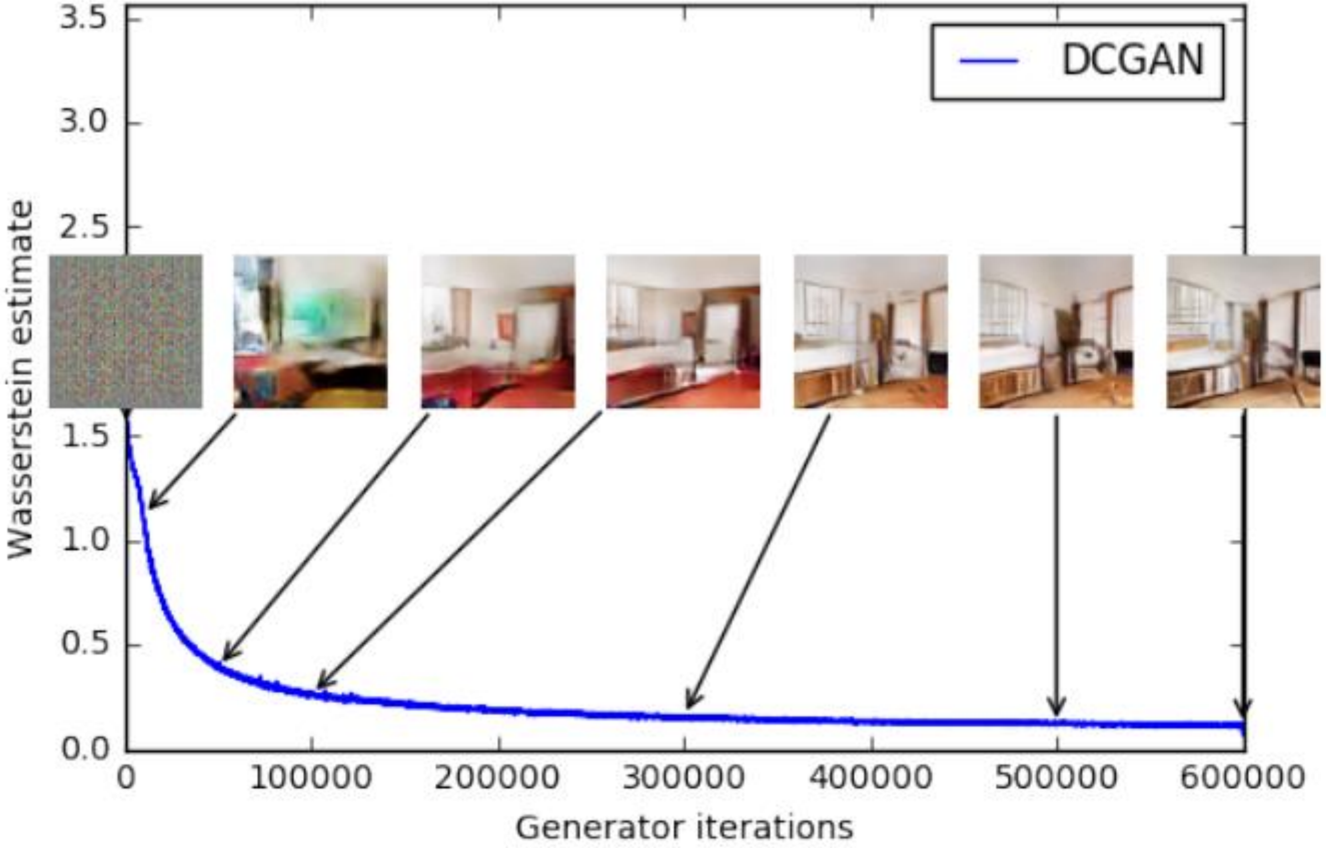
Experiment results

- Not saturate gradients



Experiment results

- Meaningful loss metric



Summary

- First work from a theoretical view of GAN’s problems.
  - Training instability.
  - Mode dropping.
  - G and D paradox.
- Propose Wasserstein GAN to come over shortcomings.
  - Improved stability.
  - No mode dropping.
  - Train D until optimal.

Thank You