CEDL

Fall 2017

HT

TOWARDS PRINCIPLED METHODS FOR TRAINING GENERATIVE ADVERSARIAL NETWORKS

Martin Arjovsky
Courant Institute of Mathematical Sciences
martinarjovsky@gmail.com

Léon Bottou
Facebook AI Research
leonb@fb.com

- Why are GANs hard to train?
  - Mode collapse
  - Unstable training behavior
  - Loss function lacking information about convergence
- Theory?

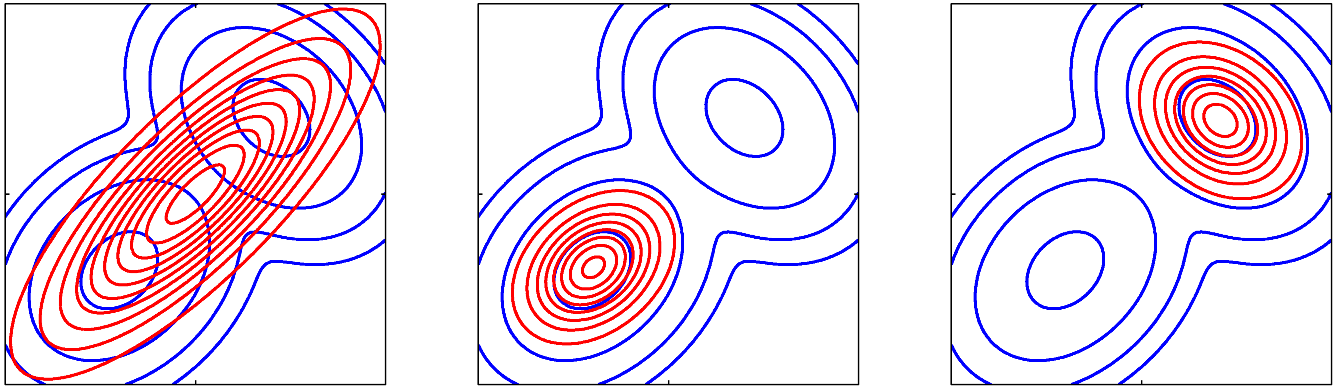
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int_{\mathcal{X}} P_r(x) \log \frac{P_r(x)}{P_g(x)} \mathrm{d}x$$

- Maximum likelihood
- Minimizing the Kullback-Leibler divergence

generator's distribution $\mathbb{P}_g$ (that depends of course on θ)

- KL = 0 if $\mathbb{P}_g = \mathbb{P}_r$
- Asymmetrical
  - $P_r(x) > P_g(x)$ mode dropping
  - $P_r(x) < P_g(x)$ low cost for fake samples

PRML Fig. 10.3 (Bishop)



p: blue

q: red

KL(p||q)

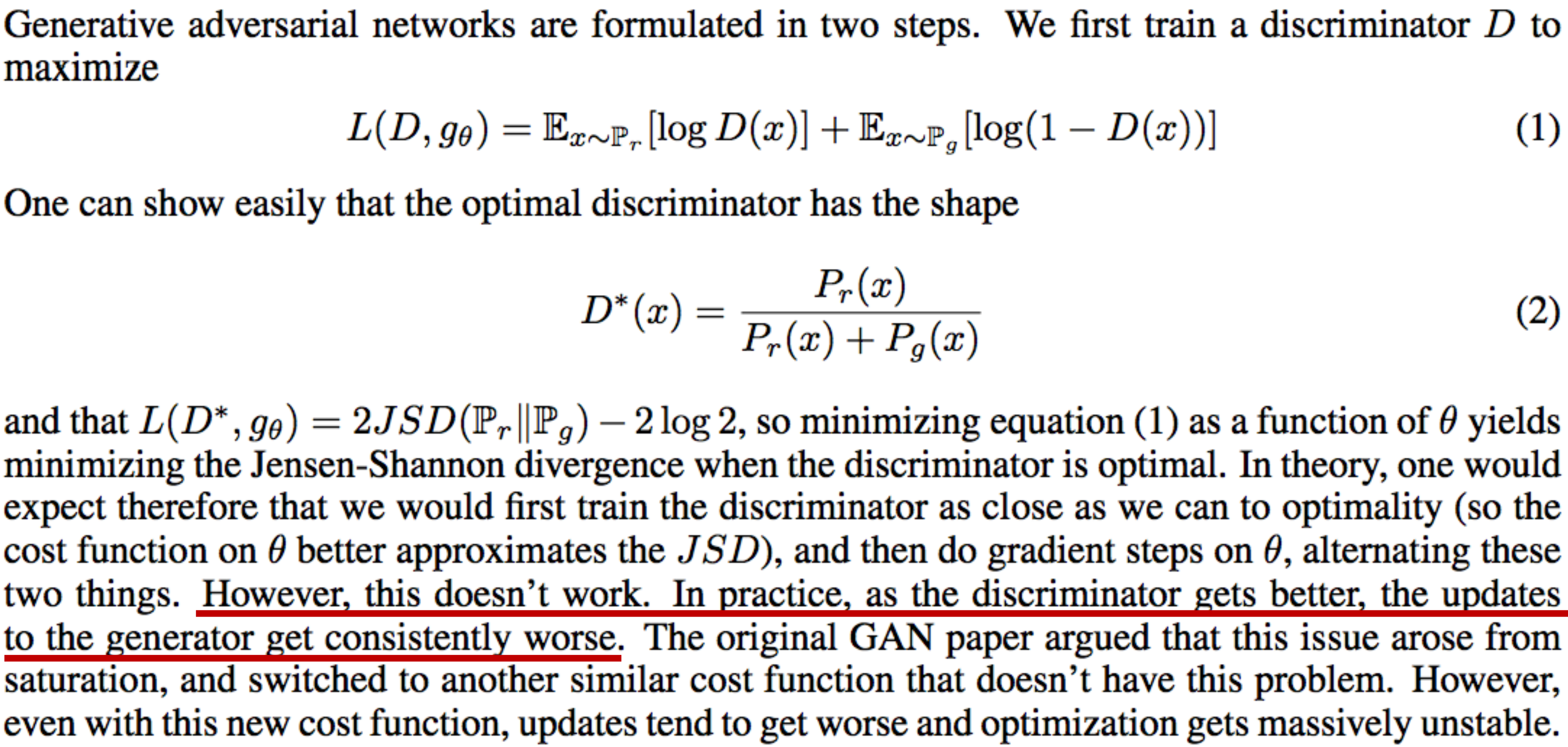
KL(q||p)

Jensen-Shannon Divergence

$$JSD(\mathbb{P}_r \| \mathbb{P}_g) = \frac{1}{2} KL(\mathbb{P}_r \| \mathbb{P}_A) + \frac{1}{2} KL(\mathbb{P}_g \| \underset{\underset{\frac{P_r+P_g}{2}}{\big\downarrow}}{\mathbb{P}_A})$$

upper bound: 2 log2

原文



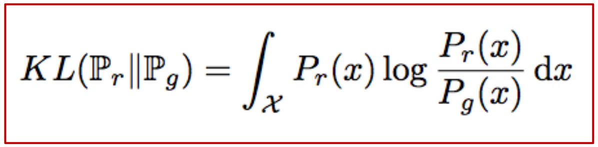
Generative adversarial networks are formulated in two steps. We first train a discriminator D to maximize

$$L(D, g_\theta) = \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{x\sim\mathbb{P}_g}[\log(1 - D(x))] \tag{1}$$

One can show easily that the optimal discriminator has the shape

$$D^*(x) = \frac{P_r(x)}{P_r(x) + P_g(x)} \tag{2}$$

and that $L(D^*, g_\theta) = 2JSD(\mathbb{P}_r\|\mathbb{P}_g) - 2\log 2$, so minimizing equation (1) as a function of θ yields minimizing the Jensen-Shannon divergence when the discriminator is optimal. In theory, one would expect therefore that we would first train the discriminator as close as we can to optimality (so the cost function on θ better approximates the JSD), and then do gradient steps on θ, alternating these two things. However, this doesn't work. In practice, as the discriminator gets better, the updates to the generator get consistently worse. The original GAN paper argued that this issue arose from saturation, and switched to another similar cost function that doesn't have this problem. However, even with this new cost function, updates tend to get worse and optimization gets massively unstable.

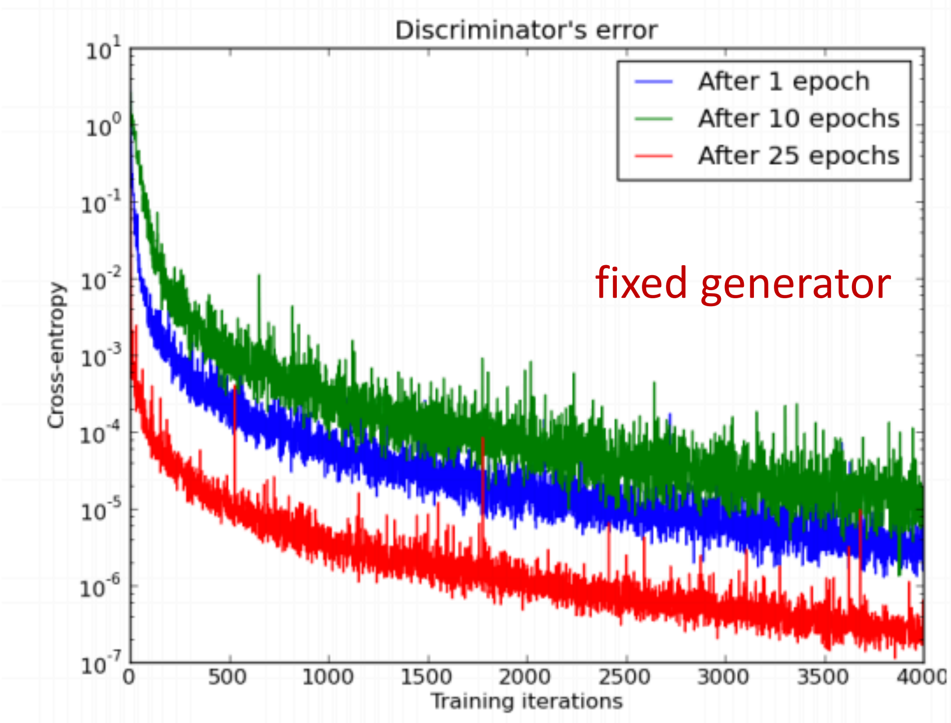


$$KL(\mathbb{P}_r\|\mathbb{P}_g) = \int_{\mathcal{X}} P_r(x)\log\frac{P_r(x)}{P_g(x)}\,\mathrm{d}x$$

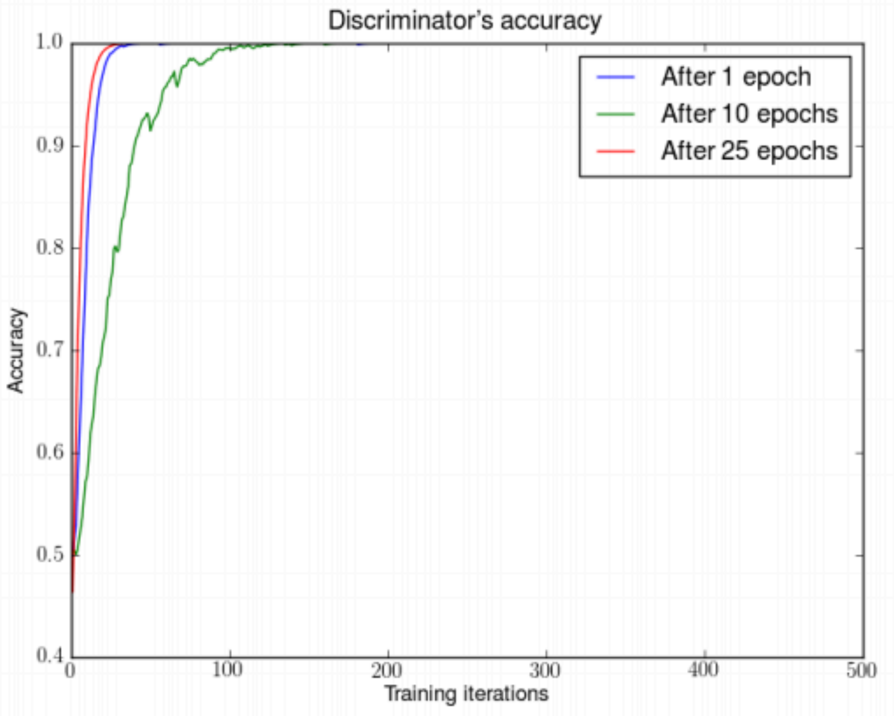
Sources of Instability

- If we just train D till convergence,

$2\log 2 - 2JSD(\mathbb{P}_r\|\mathbb{P}_g)$ goes to 0

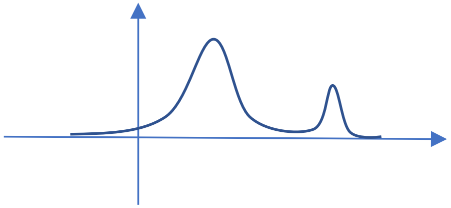


fixed generator

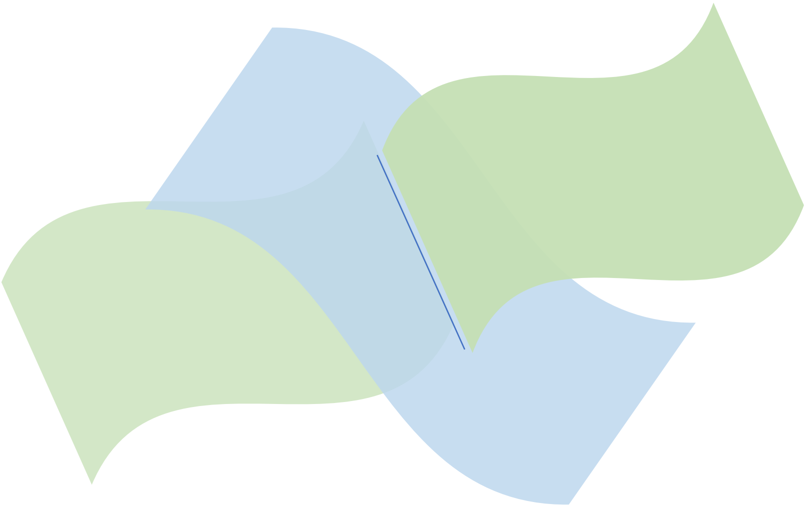


Why?

- The distributions not continuous
  - Their supports lie on low dimensional manifolds

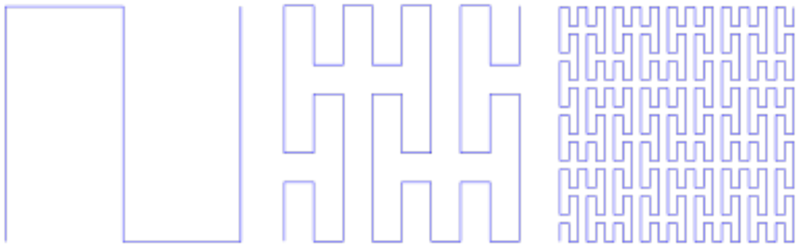


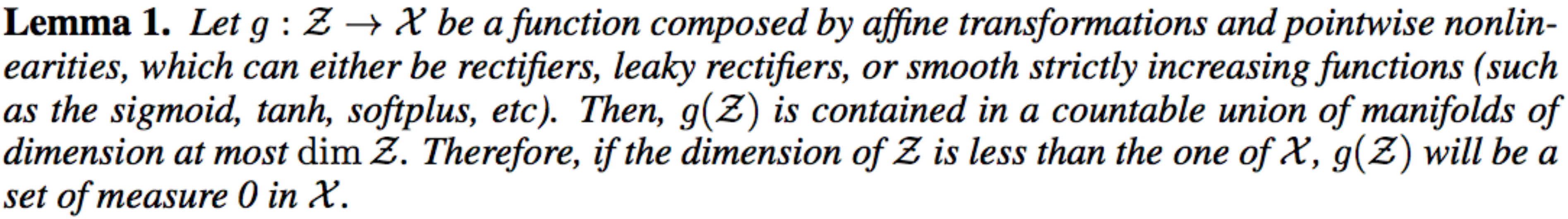
  - Support?
  - Manifolds? $\mathbb{P}_r$ $\mathbb{P}_g$



  - Measure 0 in $\mathcal{X}$

Nontrivial



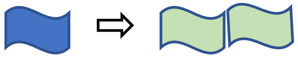


Lemma 1. *Let $g : \mathcal{Z} \to \mathcal{X}$ be a function composed by affine transformations and pointwise nonlinearities, which can either be rectifiers, leaky rectifiers, or smooth strictly increasing functions (such as the sigmoid, tanh, softplus, etc). Then, $g(\mathcal{Z})$ is contained in a 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 $\mathcal{X}$.*

g = neural network
Well behaved; not to worry

Proof of Lemma 1. We first consider the case where the nonlinearities are rectifiers or leaky rectifiers of the form $\sigma(x) = \mathbb{1}[x < 0]c_1 x + \mathbb{1}[x \geq 0]c_2 x$ for some $c_1, c_2 \in \mathbb{R}$. In this case, $g(z) = \mathbf{D}_n \mathbf{W}_n \dots \mathbf{D}_1 \mathbf{W}_1 z$, where $\mathbf{W}_i$ are affine transformations and $\mathbf{D}_i$ are some diagonal matrices dependent on z that have diagonal entries c_1 or c_2. If we consider $\mathcal{D}$ to be the (finite) set of all diagonal matrices with diagonal entries c_1 or c_2, then $g(\mathcal{Z}) \subseteq \bigcup_{D_i \in \mathcal{D}} \mathbf{D}_n \mathbf{W}_n \dots \mathbf{D}_1 \mathbf{W}_1 \mathcal{Z}$, which is a finite union of linear manifolds.

The proof for the second case is technical and slightly more involved. When σ is a pointwise smooth strictly increasing nonlinearity, then applying it vectorwise it's a diffeomorphism to its image. Therefore, it sends a countable union of manifolds of dimension d to a countable union of manifolds of dimension d. If we can prove the same thing for affine transformations we will be finished, since $g(\mathcal{Z})$ is just a composition of these applied to a $\dim \mathcal{Z}$ dimensional manifold. Of course, it suffices to prove that an affine transformation sends a manifold to a countable union of manifolds without increasing dimension, since a countable union of countable unions is still a countable union. Furthermore, we only need to show this for linear transformations, since applying a bias term is a diffeomorphism.

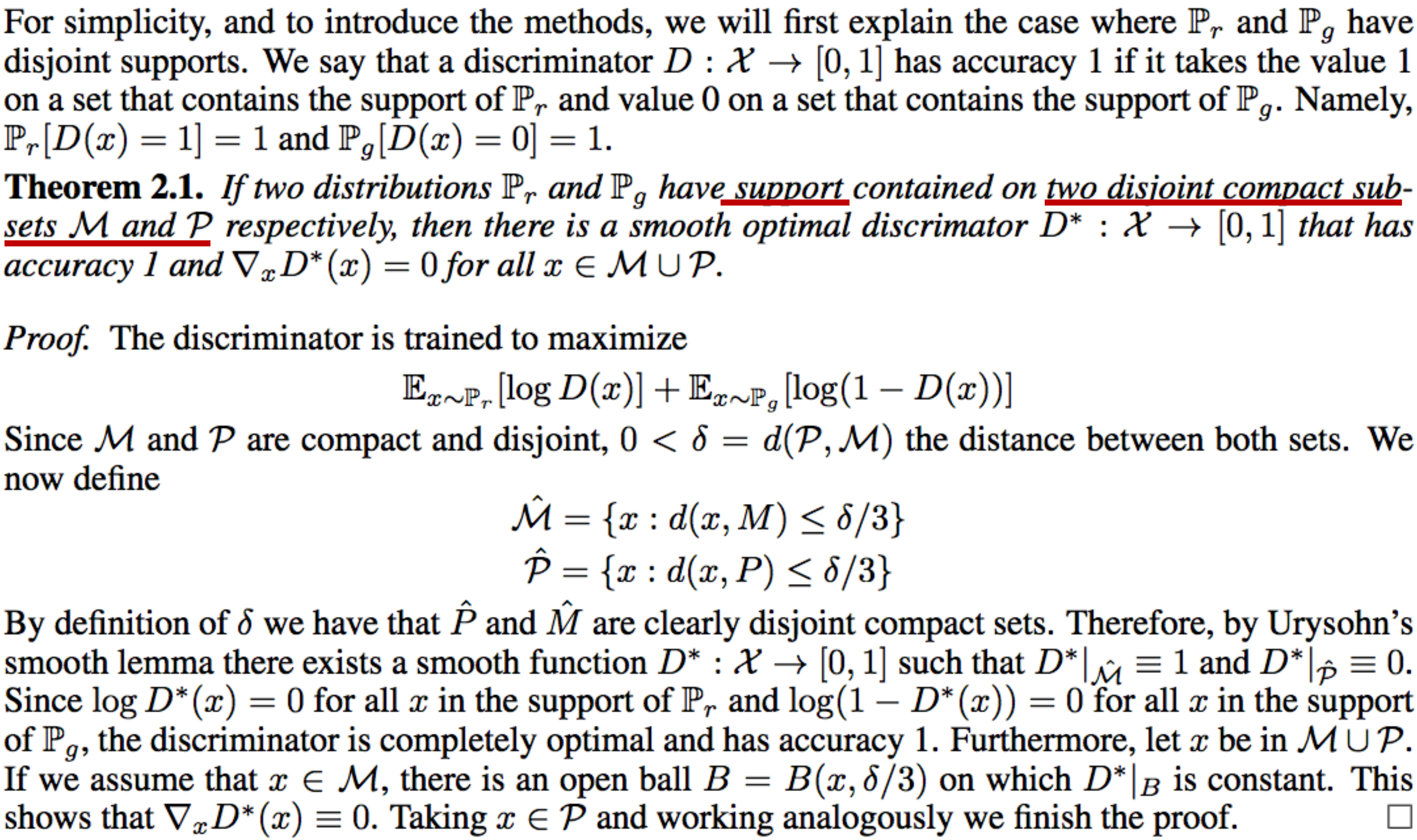


Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ be a matrix. Note that by the singular value decomposition, $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$, where $\boldsymbol{\Sigma}$ is a square diagonal matrix with diagonal positive entries and $\mathbf{U}, \mathbf{V}$ are compositions of changes of basis, inclusions (meaning adding 0s to new coordinates) and projections to a subset of the coordinates. Multiplying by Σ and applying a change of basis are diffeomorphisms, and adding 0s to new coordinates is a manifold embedding, so we only need to prove our statement for projections onto a subset of the coordinates. Let $\pi : \mathbb{R}^{n+k} \to \mathbb{R}^n$, where $\pi(x_1, \dots, x_{n+k}) = (x_1, \dots, x_n)$ be our projection and $\mathcal{M} \subseteq \mathbb{R}^{n+k}$ our d-dimensional manifold. If $n \leq d$, we are done since the image of π is contained in all $\mathbb{R}^n$, a manifold with at most dimension d. We now turn to the case where $n > d$. Let $\pi_i(x) = x_i$ be the projection onto the i-th coordinate. If x is a critical point of π, since the coordinates of π are independent, then x has to be a critical point of a π_i. By a consequence of the Morse Lemma, the critical points of π_i are isolated, and therefore so are the ones of π, meaning that there is at most a countable number of them. Since π maps the non-critical points onto a d dimensional manifold (because it acts as an embedding) and the countable number of critical points into a countable number of points (or 0 dimensional manifolds), the proof is finished. □

The Perfect Discrimination Theorems

- The supports of Pr and Pg
  - Disjoint compact subsets
  - Low dimensional manifolds
    - Intersect
    - Align

The Perfect Discrimination Theorems



For simplicity, and to introduce the methods, we will first explain the case where $\mathbb{P}_r$ and $\mathbb{P}_g$ have disjoint supports. We say that a discriminator $D : \mathcal{X} \to [0,1]$ has accuracy 1 if it takes the value 1 on a set that contains the support of $\mathbb{P}_r$ and value 0 on a set that contains the support of $\mathbb{P}_g$. Namely, $\mathbb{P}_r[D(x) = 1] = 1$ and $\mathbb{P}_g[D(x) = 0] = 1$.

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 $x \in \mathcal{M} \cup \mathcal{P}$.*

Proof. The discriminator is trained to maximize

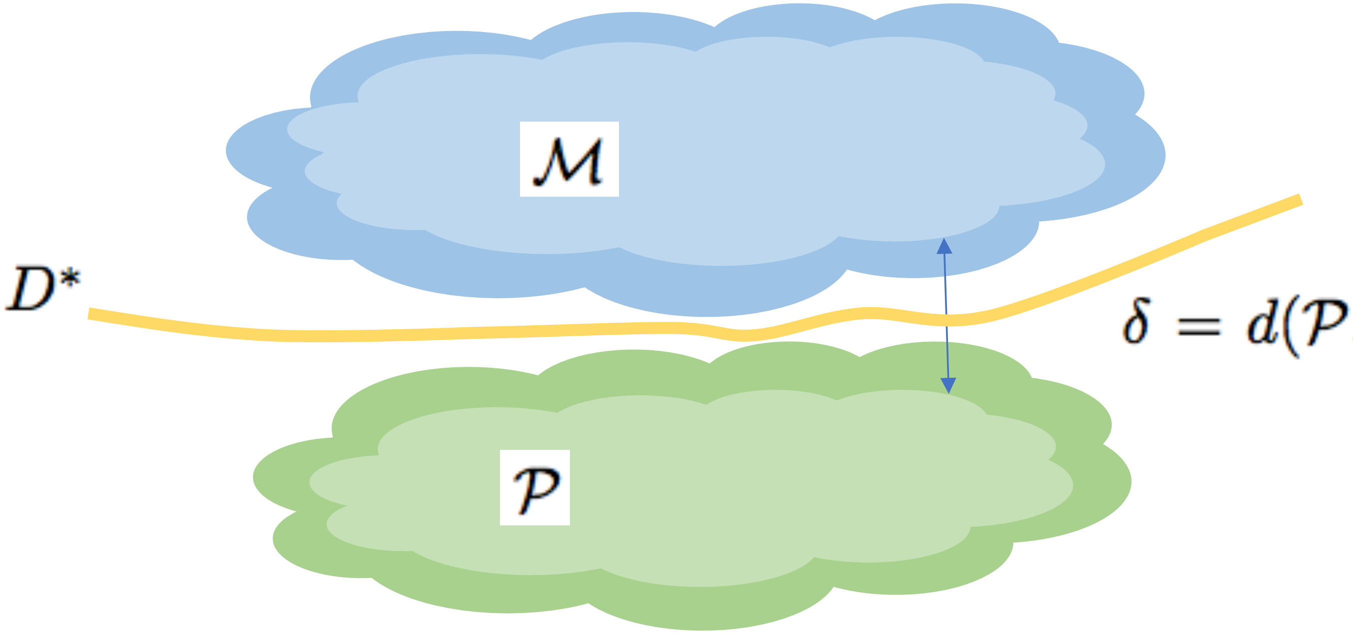
$$\mathbb{E}_{x \sim \mathbb{P}_r}[\log D(x)] + \mathbb{E}_{x \sim \mathbb{P}_g}[\log(1 - D(x))]$$

Since $\mathcal{M}$ and $\mathcal{P}$ are compact and disjoint, $0 < \delta = d(\mathcal{P}, \mathcal{M})$ the distance between both sets. We now define

$$\hat{\mathcal{M}} = \{x : d(x, M) \leq \delta/3\}$$
$$\hat{\mathcal{P}} = \{x : d(x, P) \leq \delta/3\}$$

By definition of δ we have that $\hat{P}$ and $\hat{M}$ are clearly disjoint compact sets. Therefore, by Urysohn's smooth lemma there exists a smooth function $D^* : \mathcal{X} \to [0,1]$ such that $D^*|_{\hat{\mathcal{M}}} \equiv 1$ and $D^*|_{\hat{\mathcal{P}}} \equiv 0$. Since $\log D^*(x) = 0$ for all x in the support of $\mathbb{P}_r$ and $\log(1 - D^*(x)) = 0$ for all x in the support of $\mathbb{P}_g$, the discriminator is completely optimal and has accuracy 1. Furthermore, let x be in $\mathcal{M} \cup \mathcal{P}$. If we assume that $x \in \mathcal{M}$, there is an open ball $B = B(x, \delta/3)$ on which $D^*|_B$ is constant. This shows that $\nabla_x D^*(x) \equiv 0$. Taking $x \in \mathcal{P}$ and working analogously we finish the proof. □

$$\hat{\mathcal{M}} = \{x : d(x, M) \leq \delta/3\}$$

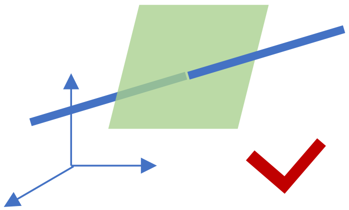


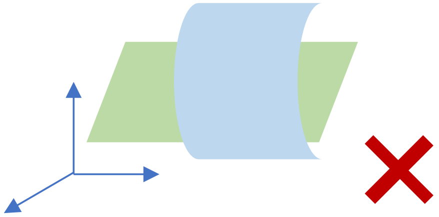
$$\delta = d(\mathcal{P}, \mathcal{M})$$

$$\hat{\mathcal{P}} = \{x : d(x, P) \leq \delta/3\}$$

The case of the supports being two different manifolds

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.

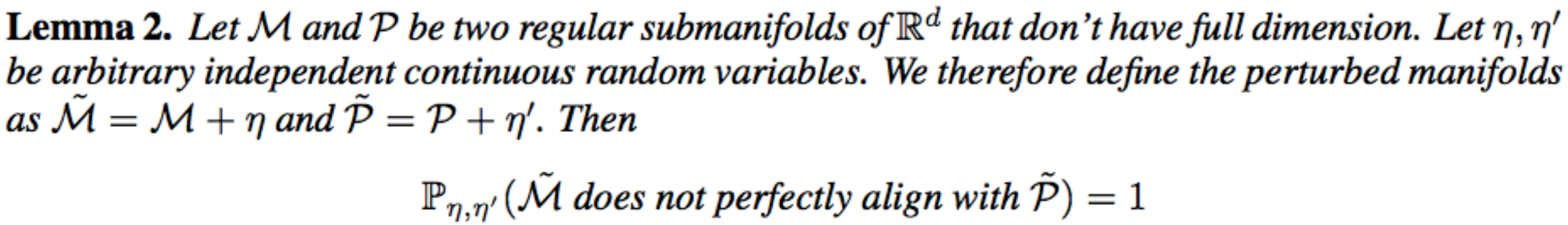




dim M + dim P – dim F

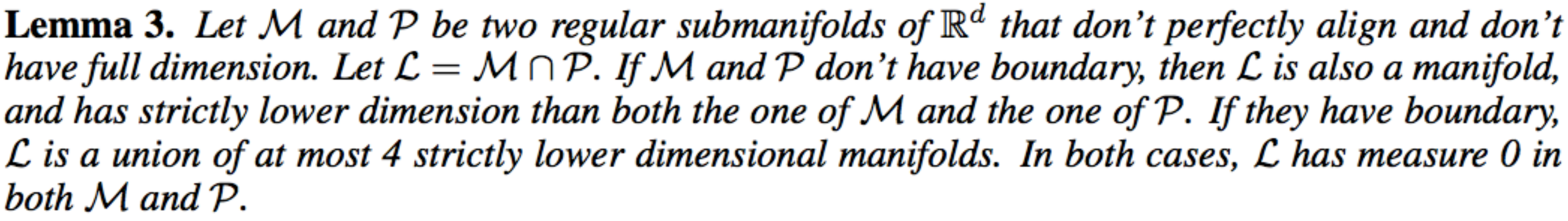
Definition 2.2. We say that two manifolds without boundary $\mathcal{M}$ and $\mathcal{P}$ **perfectly align** if there is an $x \in \mathcal{M} \cap \mathcal{P}$ such that $\mathcal{M}$ and $\mathcal{P}$ don't intersect transversally in x.
We shall note the boundary and interior of a manifold $\mathcal{M}$ by ∂M and Int M respectively. We say that two manifolds $\mathcal{M}$ and $\mathcal{P}$ (with or without boundary) perfectly align if any of the boundary free manifold pairs $(\text{Int } \mathcal{M}, \text{Int } \mathcal{P})$, $(\text{Int } \mathcal{M}, \partial\mathcal{P})$, $(\partial\mathcal{M}, \text{Int } \mathcal{P})$ or $(\partial\mathcal{M}, \partial\mathcal{P})$ perfectly align.

The interesting thing is that we can safely assume in practice that any two manifolds never perfectly align. This can be done since an arbitrarilly small random perturbation on two manifolds will lead them to intersect transversally or don't intersect at all. This is precisely stated and proven in Lemma 2.

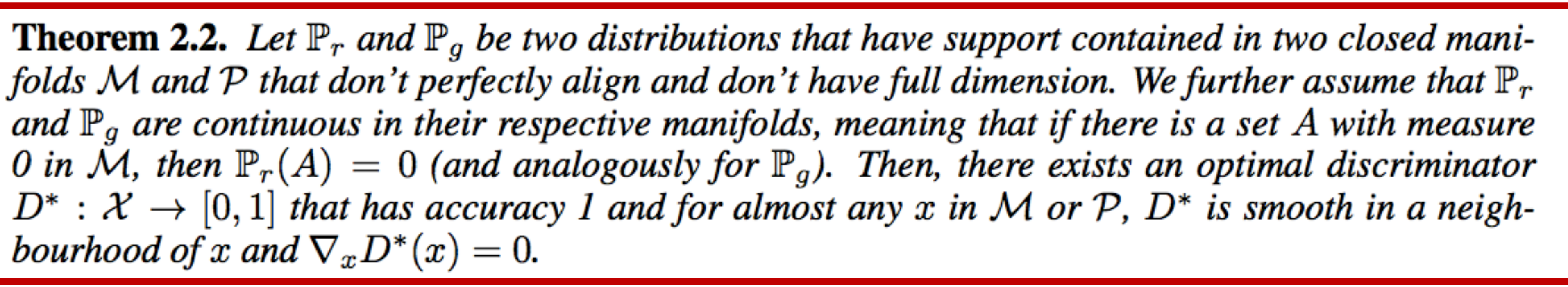


Lemma 2. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't have full dimension. Let η, η' be arbitrary independent continuous random variables. We therefore define the perturbed manifolds as $\tilde{\mathcal{M}} = \mathcal{M} + \eta$ and $\tilde{\mathcal{P}}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

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

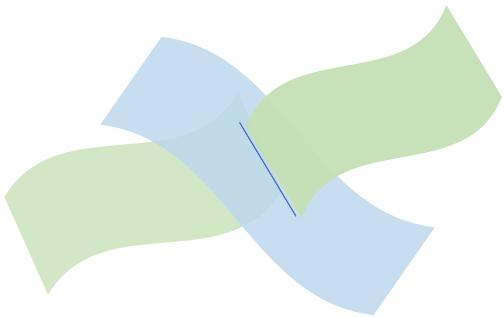


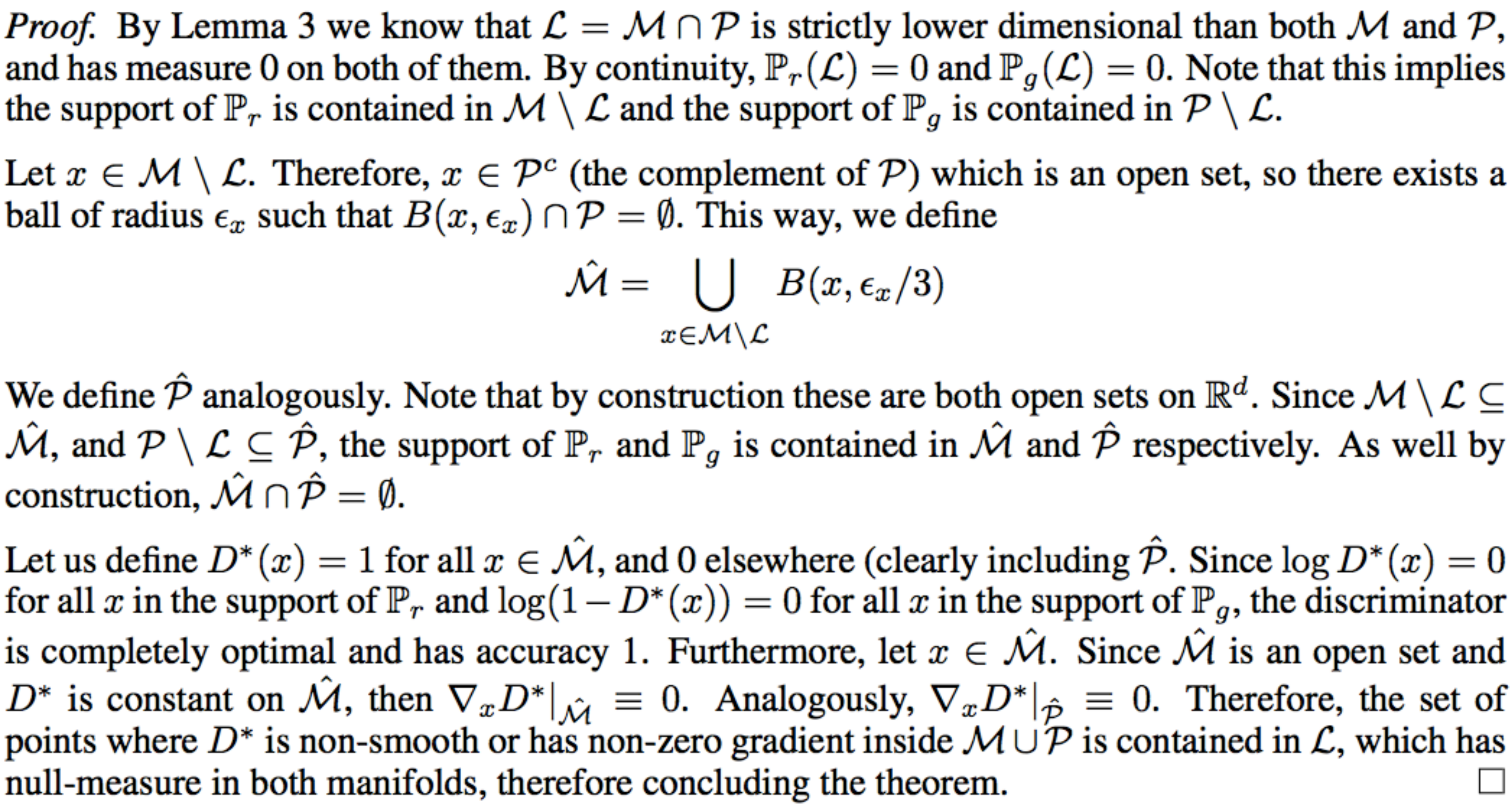
Lemma 3. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't perfectly align and don't have full dimension. Let $\mathcal{L} = \mathcal{M} \cap \mathcal{P}$. If $\mathcal{M}$ and $\mathcal{P}$ don't have boundary, then $\mathcal{L}$ is also a manifold, and has 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 $\mathcal{P}$.*



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 $\nabla_x D^*(x) = 0$.*

Proof of Theorem 2.2 Skip?





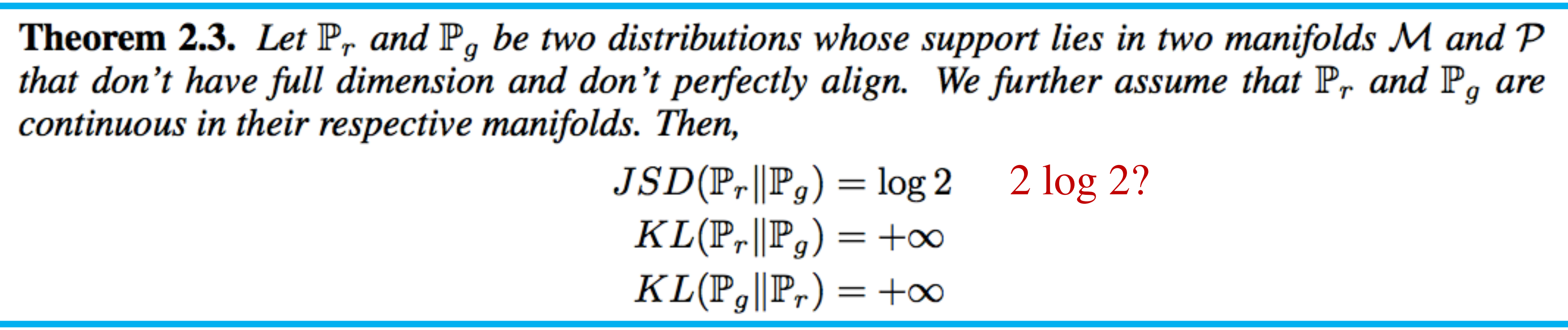
Proof. By Lemma 3 we know that $\mathcal{L} = \mathcal{M} \cap \mathcal{P}$ is strictly lower dimensional than both $\mathcal{M}$ and $\mathcal{P}$, and has measure 0 on both of them. By continuity, $\mathbb{P}_r(\mathcal{L}) = 0$ and $\mathbb{P}_g(\mathcal{L}) = 0$. Note that this implies the support of $\mathbb{P}_r$ is contained in $\mathcal{M} \setminus \mathcal{L}$ and the support of $\mathbb{P}_g$ is contained in $\mathcal{P} \setminus \mathcal{L}$.

Let $x \in \mathcal{M} \setminus \mathcal{L}$. Therefore, $x \in \mathcal{P}^c$ (the complement of $\mathcal{P}$) which is an open set, so there exists a ball of radius ϵ_x such that $B(x, \epsilon_x) \cap \mathcal{P} = \emptyset$. This way, we define

$$\hat{\mathcal{M}} = \bigcup_{x \in \mathcal{M} \setminus \mathcal{L}} B(x, \epsilon_x / 3)$$

We define $\hat{\mathcal{P}}$ analogously. Note that by construction these are both open sets on $\mathbb{R}^d$. Since $\mathcal{M} \setminus \mathcal{L} \subseteq \hat{\mathcal{M}}$, and $\mathcal{P} \setminus \mathcal{L} \subseteq \hat{\mathcal{P}}$, the support of $\mathbb{P}_r$ and $\mathbb{P}_g$ is contained in $\hat{\mathcal{M}}$ and $\hat{\mathcal{P}}$ respectively. As well by construction, $\hat{\mathcal{M}} \cap \hat{\mathcal{P}} = \emptyset$.

Let us define $D^*(x) = 1$ for all $x \in \hat{\mathcal{M}}$, and 0 elsewhere (clearly including $\hat{\mathcal{P}}$. Since $\log D^*(x) = 0$ for all x in the support of $\mathbb{P}_r$ and $\log(1 - D^*(x)) = 0$ for all x in the support of $\mathbb{P}_g$, the discriminator is completely optimal and has accuracy 1. Furthermore, let $x \in \hat{\mathcal{M}}$. Since $\hat{\mathcal{M}}$ is an open set and D^* is constant on $\hat{\mathcal{M}}$, then $\nabla_x D^*|_{\hat{\mathcal{M}}} \equiv 0$. Analogously, $\nabla_x D^*|_{\hat{\mathcal{P}}} \equiv 0$. Therefore, the set of points where D^* is non-smooth or has non-zero gradient inside $\mathcal{M} \cup \mathcal{P}$ is contained in $\mathcal{L}$, which has null-measure in both manifolds, therefore concluding the theorem. □



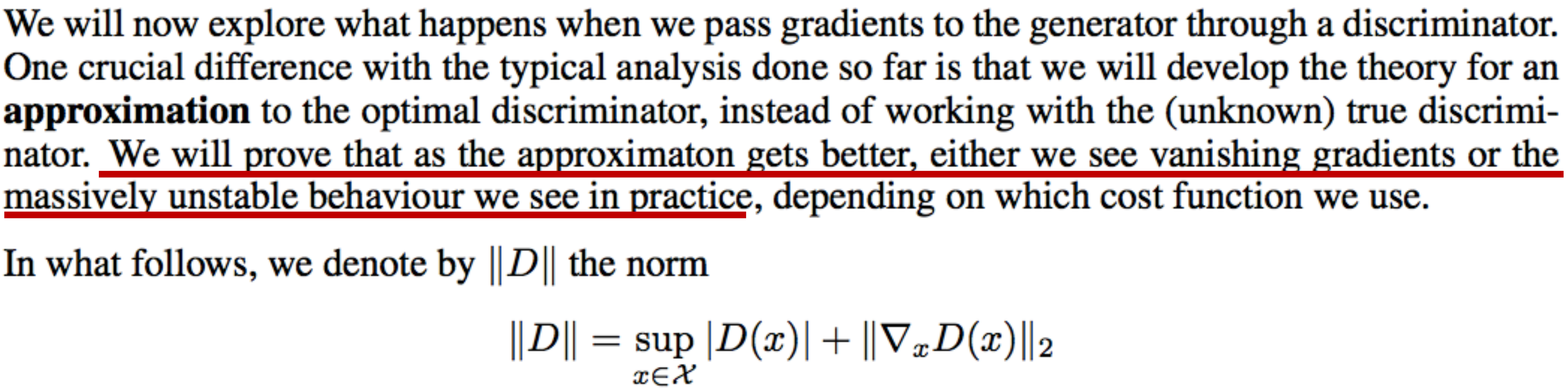
Theorem 2.3. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions whose support lies in two manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't have full dimension and don't perfectly align. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds.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

$$JSD(\mathbb{P}_r \| \mathbb{P}_g) = \log 2 \qquad 2 \log 2?$$
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = +\infty$$
$$KL(\mathbb{P}_g \| \mathbb{P}_r) = +\infty$$

To test similarities between the distributions?

The key idea for the next paper

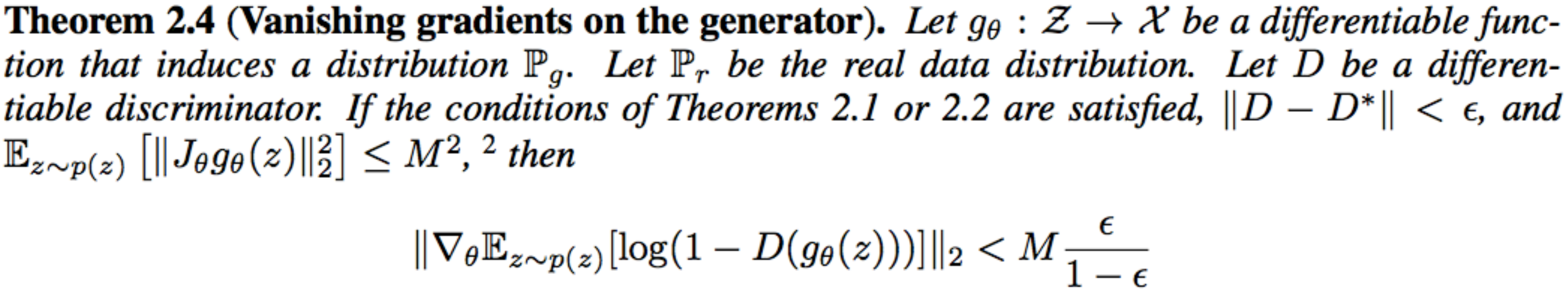
The Problems of the Cost Functions



We will now explore what happens when we pass gradients to the generator through a discriminator. One crucial difference with the typical analysis done so far is that we will develop the theory for an **approximation** to the optimal discriminator, instead of working with the (unknown) true discriminator. We will prove that as the approximaton gets better, either we see vanishing gradients or the massively unstable behaviour we see in practice, depending on which cost function we use.

In what follows, we denote by $\|D\|$ the norm

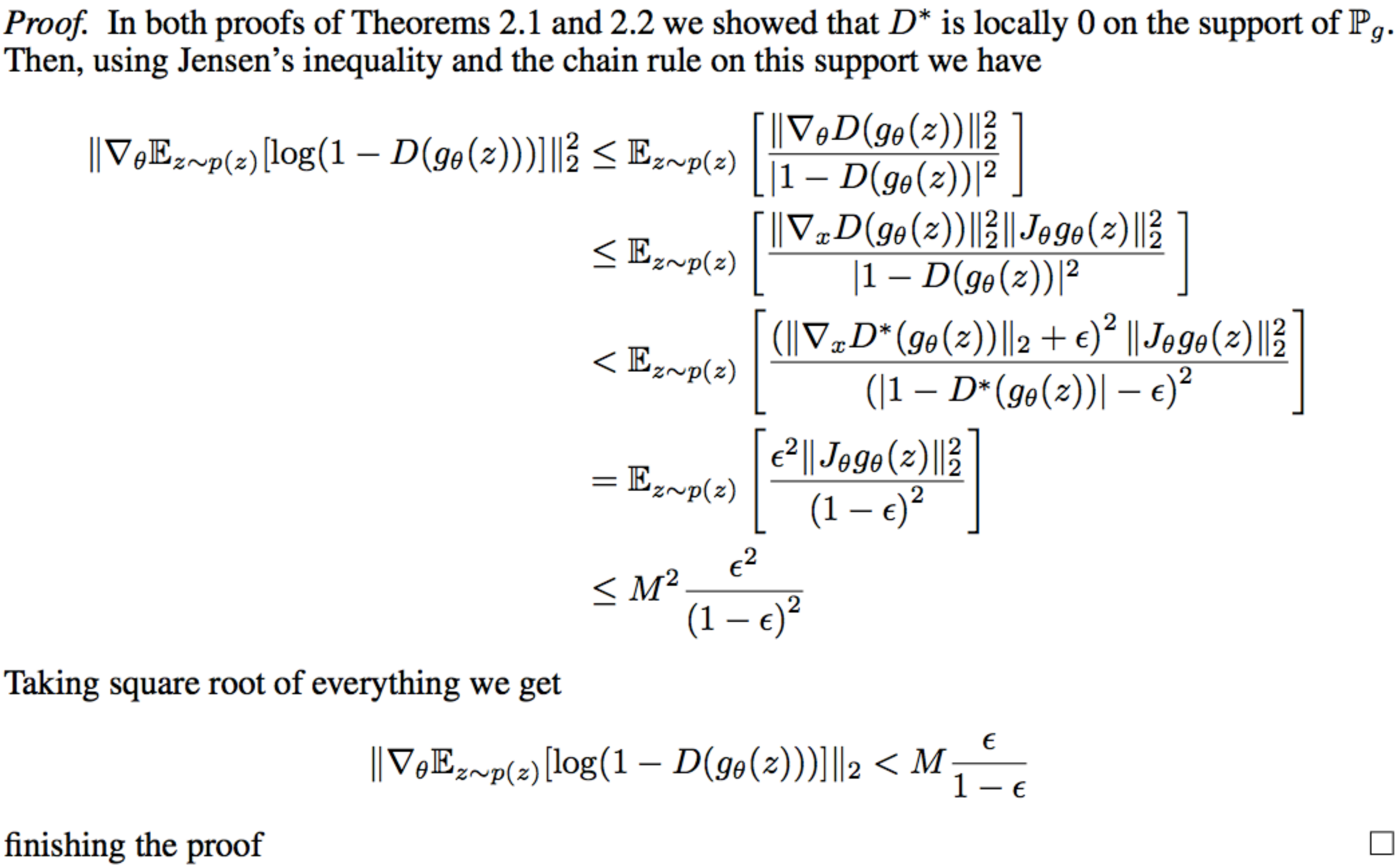
$$\|D\| = \sup_{x \in \mathcal{X}} |D(x)| + \|\nabla_x D(x)\|_2$$



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*

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

證明不難



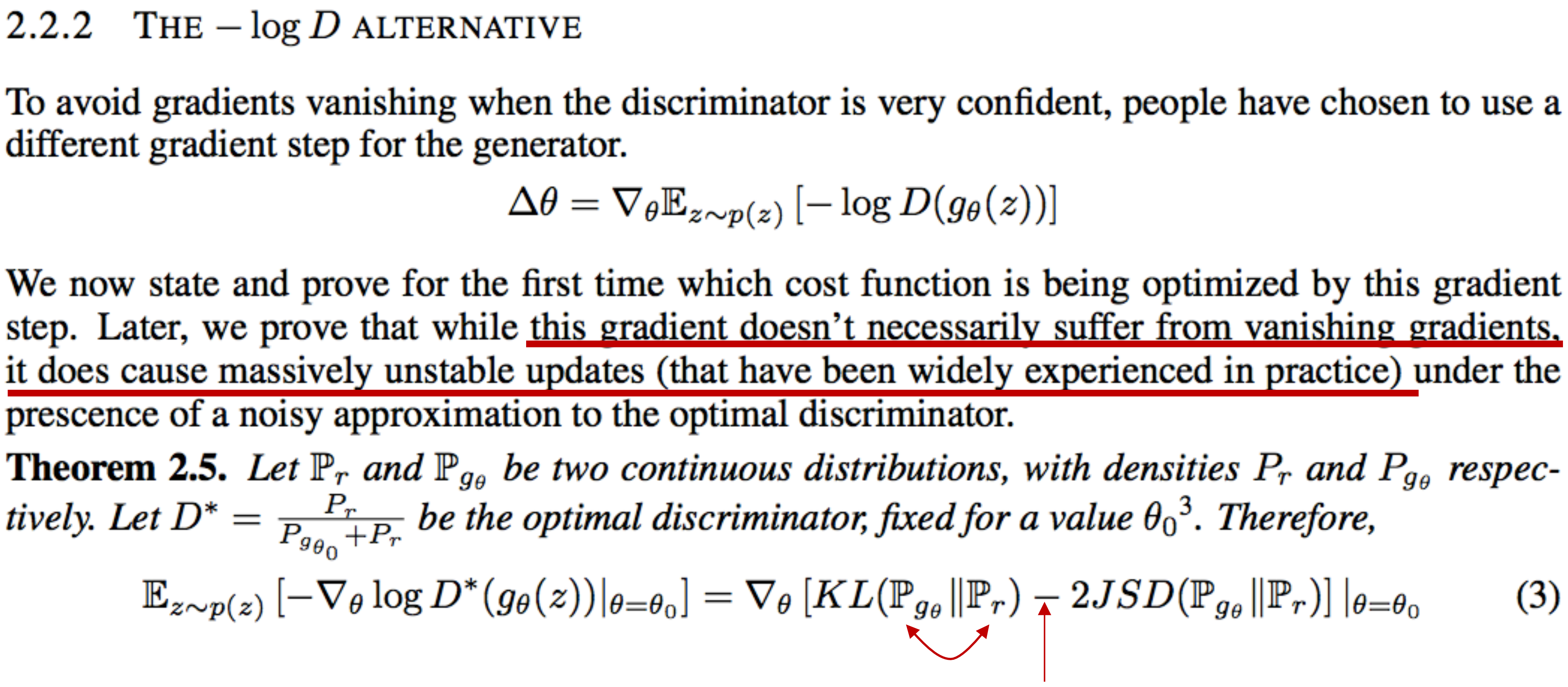
Proof. In both proofs of Theorems 2.1 and 2.2 we showed that D^* is locally 0 on the support of $\mathbb{P}_g$. Then, using Jensen's inequality and the chain rule on this support we have

$$\begin{aligned}
\|\nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1-D(g_\theta(z)))]\|_2^2 &\leq \mathbb{E}_{z\sim p(z)}\left[\frac{\|\nabla_\theta D(g_\theta(z))\|_2^2}{|1-D(g_\theta(z))|^2}\right] \\
&\leq \mathbb{E}_{z\sim p(z)}\left[\frac{\|\nabla_x D(g_\theta(z))\|_2^2 \|J_\theta g_\theta(z)\|_2^2}{|1-D(g_\theta(z))|^2}\right] \\
&< \mathbb{E}_{z\sim p(z)}\left[\frac{\left(\|\nabla_x D^*(g_\theta(z))\|_2+\epsilon\right)^2 \|J_\theta g_\theta(z)\|_2^2}{\left(|1-D^*(g_\theta(z))|-\epsilon\right)^2}\right] \\
&= \mathbb{E}_{z\sim p(z)}\left[\frac{\epsilon^2 \|J_\theta g_\theta(z)\|_2^2}{(1-\epsilon)^2}\right] \\
&\leq M^2 \frac{\epsilon^2}{(1-\epsilon)^2}
\end{aligned}$$

Taking square root of everything we get

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

finishing the proof □



2.2.2 The $-\log D$ alternative

To avoid gradients vanishing when the discriminator is very confident, people have chosen to use a different gradient step for the generator.

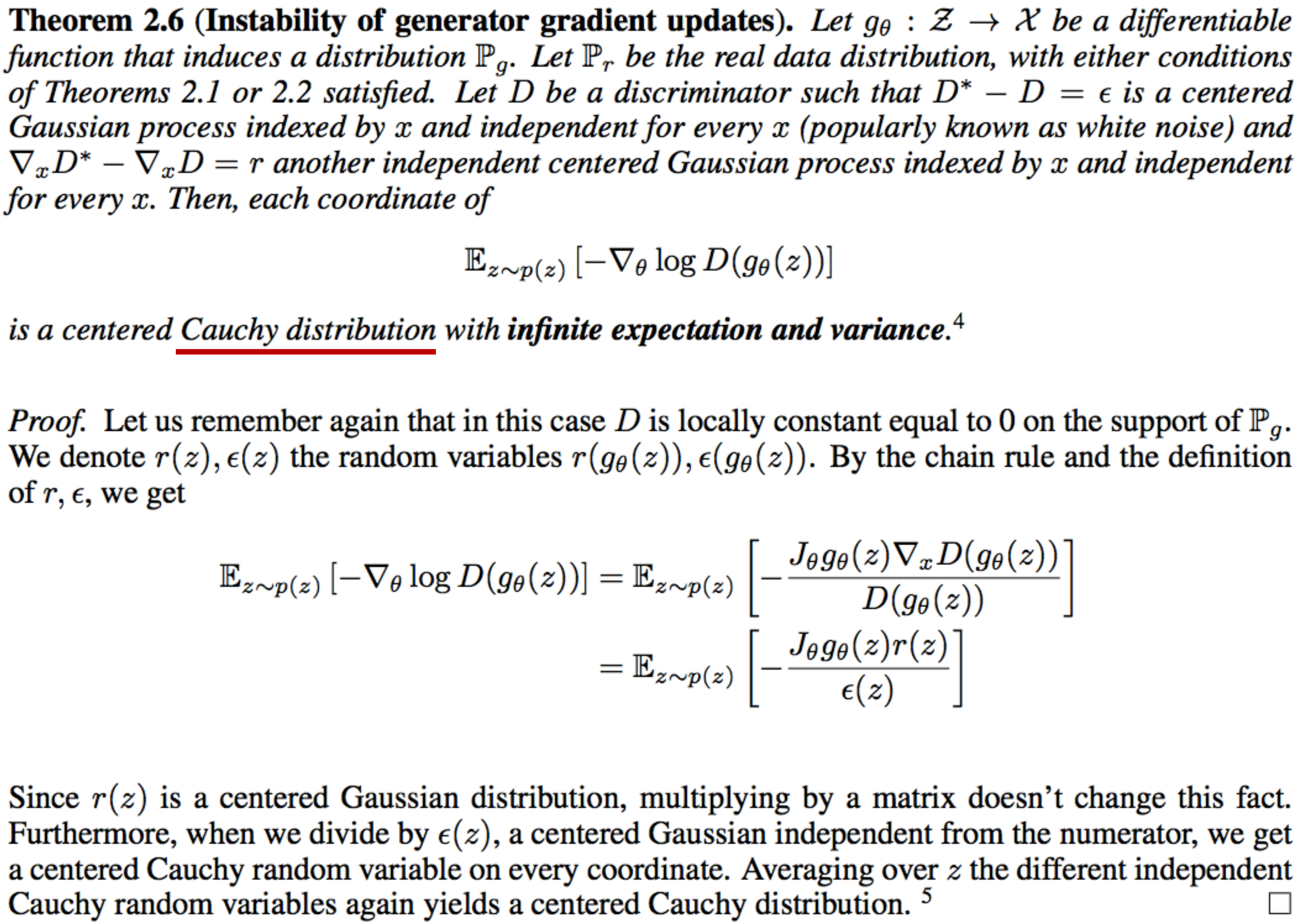
$$\Delta\theta = \nabla_\theta \mathbb{E}_{z\sim p(z)}\left[-\log D(g_\theta(z))\right]$$

We now state and prove for the first time which cost function is being optimized by this gradient step. Later, we prove that while this gradient doesn't necessarily suffer from vanishing gradients, it does cause massively unstable updates (that have been widely experienced in practice) under the prescence of a noisy approximation to the optimal discriminator.

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,*

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

High cost on generating fake looking samples
Low cost on mode dropping



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]

Proof. Let us remember again that in this case D is locally constant equal to 0 on the support of $\mathbb{P}_g$. We denote $r(z), \epsilon(z)$ the random variables $r(g_\theta(z)), \epsilon(g_\theta(z))$. By the chain rule and the definition of r, ϵ, we get

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

Since $r(z)$ is a centered Gaussian distribution, multiplying by a matrix doesn't change this fact. Furthermore, when we divide by $\epsilon(z)$, a centered Gaussian independent from the numerator, we get a centered Cauchy random variable on every coordinate. Averaging over z the different independent Cauchy random variables again yields a centered Cauchy distribution. [5] □

Solutions?

- Softer metrics
    - Add noise
    - Wasserstein metric
        - Earth Mover’s Distance (EMD)

Corollary 3.2. *Let $\epsilon, \epsilon' \sim \mathcal{N}(0, \sigma^2 I)$ and $\tilde{g}_\theta(z) = g_\theta(z) + \epsilon'$, then*

$$\begin{aligned}
\mathbb{E}_{z\sim p(z),\epsilon'} & \left[\nabla_\theta \log(1 - D^*(\tilde{g}_\theta(z)))\right] \\
&= \mathbb{E}_{z\sim p(z),\epsilon'} \left[a(z) \int_{\mathcal{M}} P_\epsilon(\tilde{g}_\theta(z) - y) \nabla_\theta \|\tilde{g}_\theta(z) - y\|^2 \, \mathrm{d}\mathbb{P}_r(y) \right. \\
&\left. - b(z) \int_{\mathcal{P}} P_\epsilon(\tilde{g}_\theta(z) - y) \nabla_\theta \|\tilde{g}_\theta(z) - y\|^2 \, \mathrm{d}\mathbb{P}_g(y) \right] \\
&= 2 \nabla_\theta JSD(\mathbb{P}_{r+\epsilon} \| \mathbb{P}_{g+\epsilon})
\end{aligned}$$

$$\begin{aligned}
P_{X+\epsilon}(x) &= \mathbb{E}_{y\sim \mathbb{P}_X} \left[P_\epsilon(x - y)\right] \\
&= \int_{\mathcal{M}} P_\epsilon(x - y) \, \mathrm{d}\mathbb{P}_X(y)
\end{aligned}$$

This theorem therefore tells us that the density $P_{X+\epsilon}(x)$ **is inversely proportional to the average distance to points in the support of $\mathbb{P}_X$, weighted by the probability of these points**. In the case of the support of $\mathbb{P}_X$ being a manifold, we will have the weighted average of the distance to the points along the manifold. How we choose the distribution of the noise ϵ will impact the notion

AMORTISED MAP INFERENCE FOR IMAGE SUPER-RESOLUTION

Casper Kaae Sønderby[1 2]*, Jose Caballero[1], Lucas Theis[1], Wenzhe Shi[1] & Ferenc Huszár[1]
casperkaae@gmail.com, {jcaballero,ltheis,wshi,fhuszar}@twitter.com
[1]Twitter Cortex, London, UK
[2]University of Copenhagen, Denmark

3.2.1 INSTANCE NOISE

The theory suggests that GANs should be a convergent algorithm. If a unique optimal discriminator exists and it is reached by optimising D to perfection at each step, technically the whole algorithm corresponds to gradient descent on an estimate of $\mathrm{KL}[q_\theta \| p_Y]$ with respect to θ. In practice, however, GANs tend to be highly unstable. So where does the theory go wrong? We think the main reason for the instability of GANs stems from q_θ and p_Y being concentrated distributions whose support does not overlap. The distribution of natural images p_Y is often assumed to concentrate on or around a low-dimensional manifold. In most cases, q_θ is degenerate and manifold-like by construction, such as in AffGAN. Therefore, odds are that especially before convergence is reached, q_θ and p_Y can be perfectly separated by several Ds violating a condition for the convergence proof. We try to remedy this problem by adding *instance noise* to both SR and true image samples. This amounts to minimising the divergence $d_\sigma(q_\theta, p_Y) = \mathrm{KL}\left[p_\sigma * q_\theta \| p_\sigma * p_Y\right]$, where $p_\sigma * q_\theta$ denotes convolution of q_θ with the noise distribution p_σ. The noise level σ can be annealed during training, and the noise allows us to safely optimise D until convergence in each iteration. The trick is related to one-sided label noise introduced by Salimans et al. (2016), however without introducing a bias in the optimal discriminator, and we believe it is a promising technique for stabilising GAN training in general. For more details please see Appendix C

Wasserstein GAN

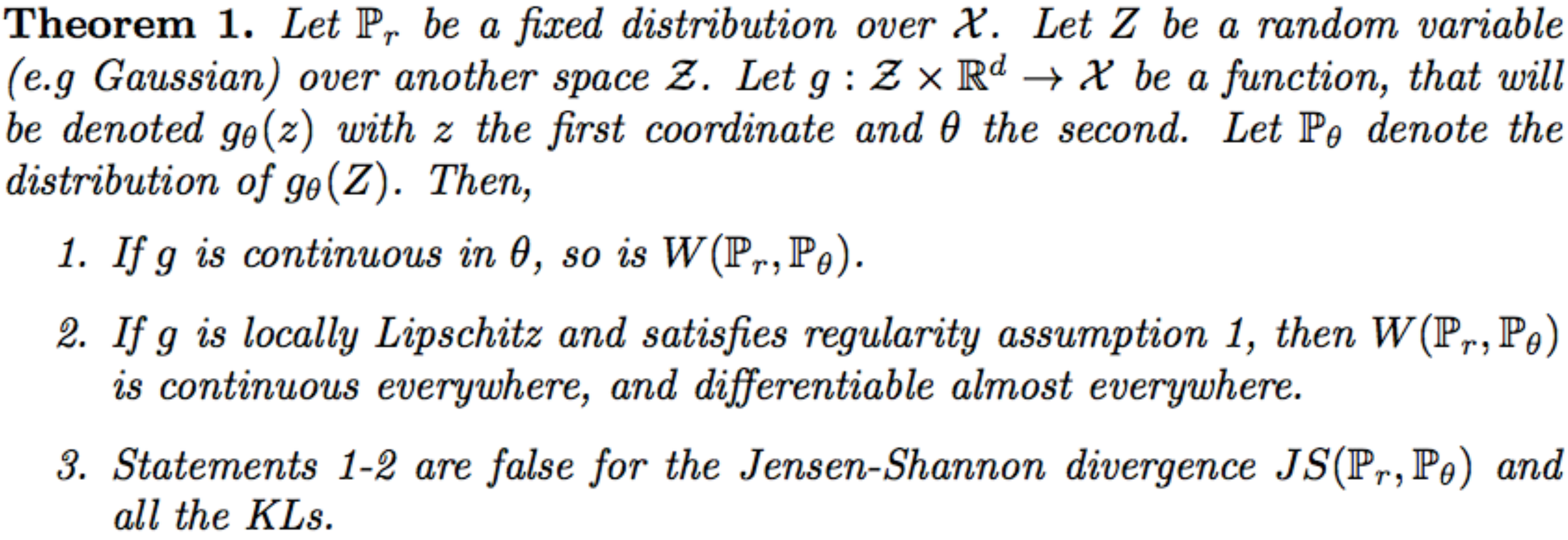
Martin Arjovsky[1], Soumith Chintala[2], and Léon Bottou[1,2]

[1]Courant Institute of Mathematical Sciences
[2]Facebook AI Research

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

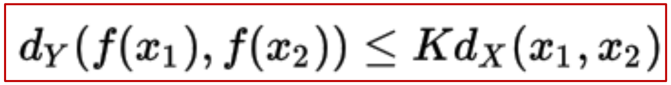
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big] , \tag{1}$$

  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 $\mathbb{P}_g$. Intuitively, $\gamma(x, y)$ indicates how much "mass" must be transported from x to y in order to transform the distributions $\mathbb{P}_r$ into the distribution $\mathbb{P}_g$. The EM distance then is the "cost" of the optimal transport plan.

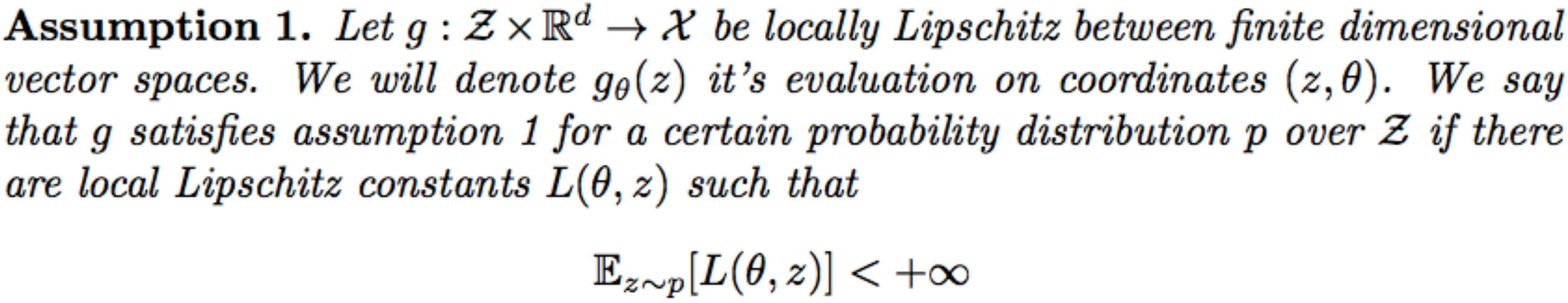


Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second.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



$$\boxed{d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2)}$$

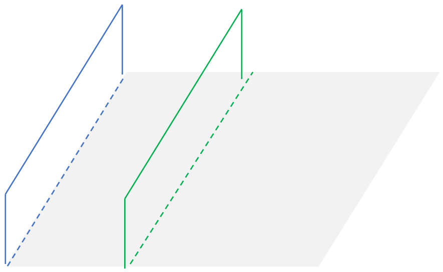


Assumption 1. *Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be locally Lipschitz between finite dimensional vector spaces. We will denote $g_\theta(z)$ it's evaluation on coordinates (z, θ). We say that g satisfies assumption 1 for a certain probability distribution p over $\mathcal{Z}$ if there are local Lipschitz constants $L(\theta, z)$ such that*

$$\mathbb{E}_{z \sim p}[L(\theta, z)] < +\infty$$

Example 1 (Learning parallel lines)**.** Let $Z \sim U[0,1]$ the uniform distribution on the unit interval. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$ (a 0 on the x-axis and the random variable Z on the y-axis), uniform on a straight vertical line passing through the origin. Now let $g_\theta(z) = (\theta, z)$ with θ a single real parameter. It is easy to see that in this case,

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

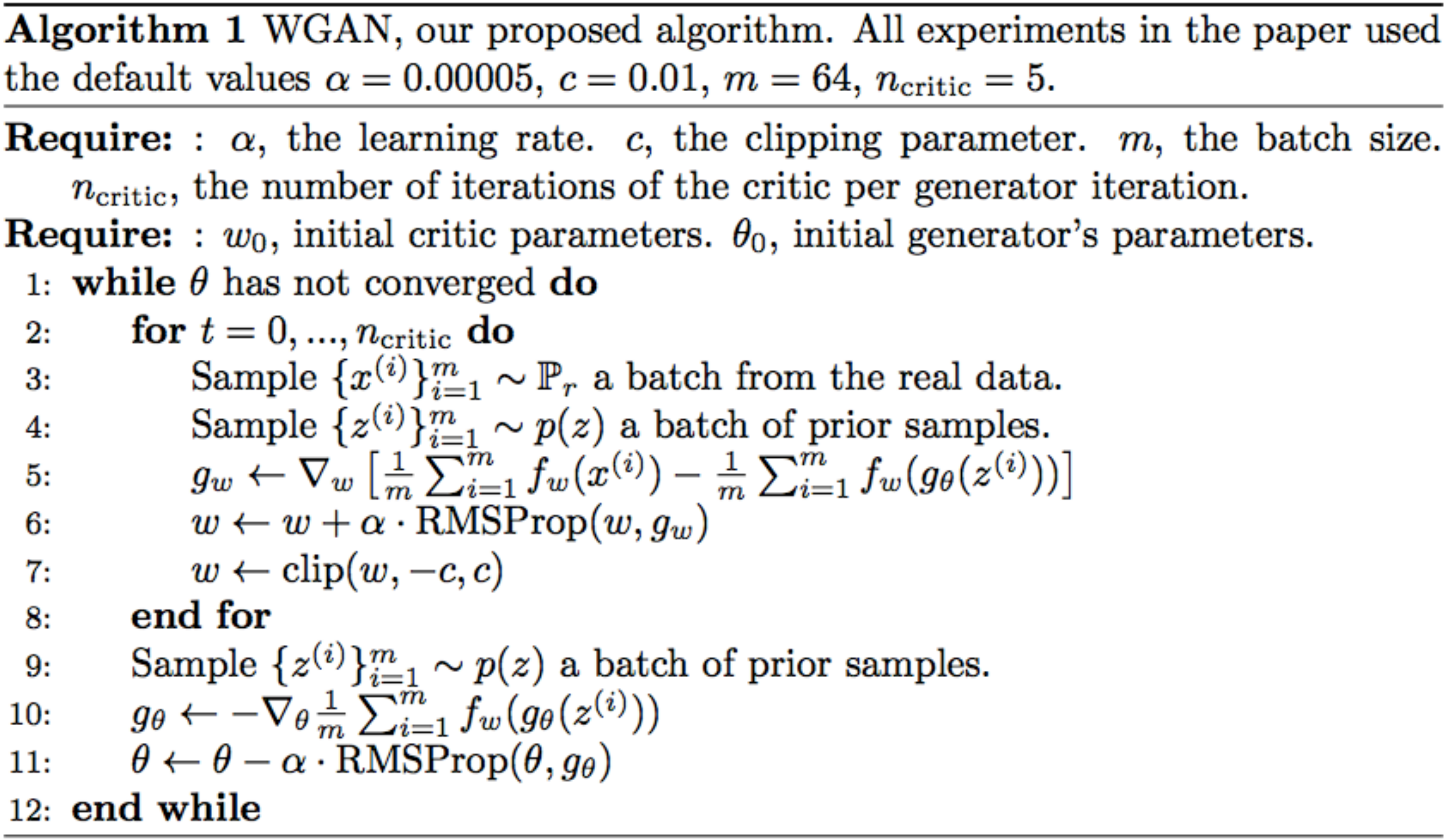


When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL, or TV divergences. Figure 1 illustrates this for the case of the EM and JS distances.

EMD intractable

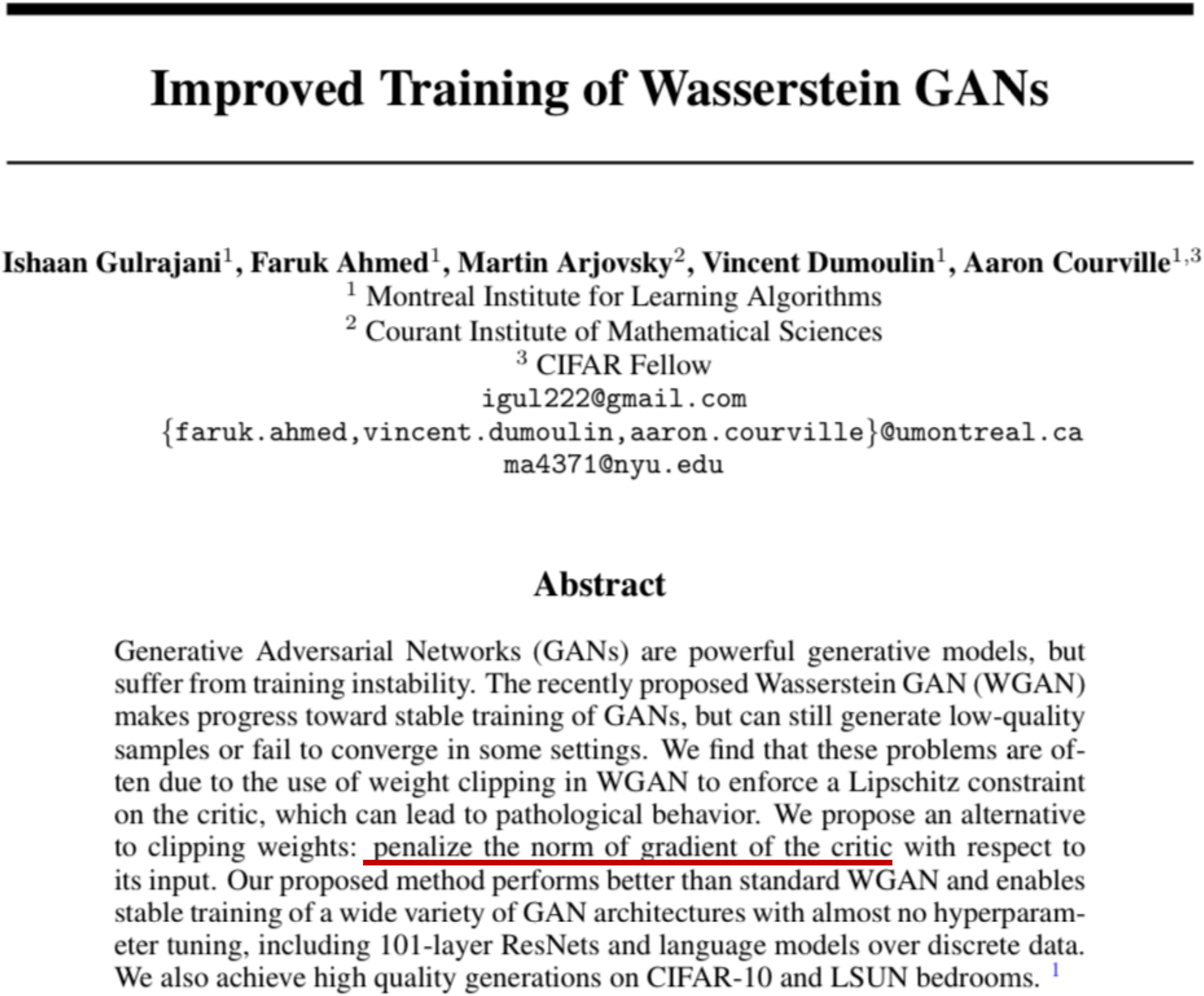
$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))]$$



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
```



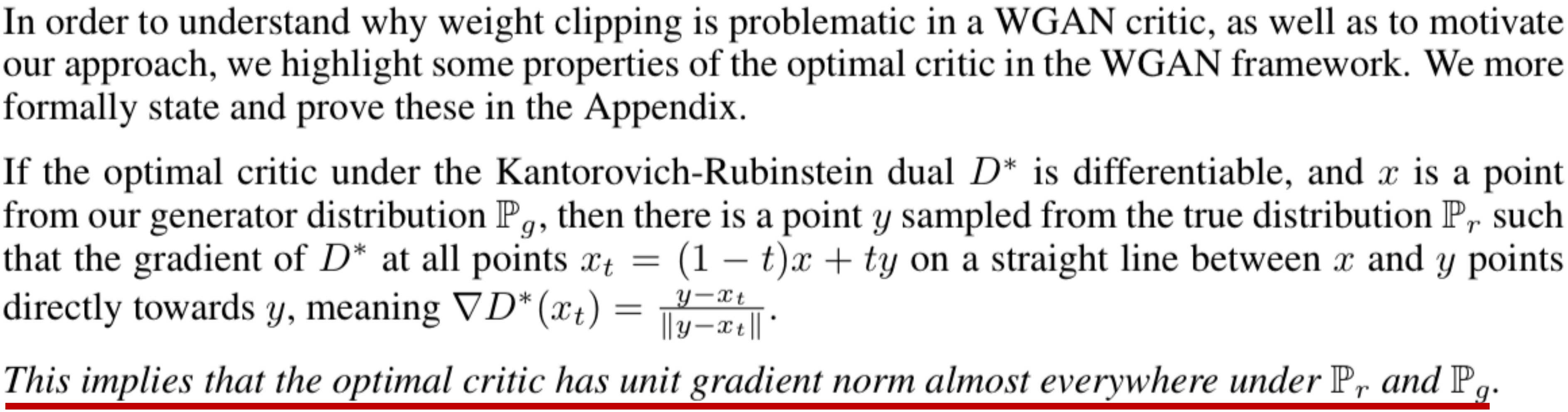
Improved Training of Wasserstein GANs

Ishaan Gulrajani[1], Faruk Ahmed[1], Martin Arjovsky[2], Vincent Dumoulin[1], Aaron Courville[1,3]
[1] Montreal Institute for Learning Algorithms
[2] Courant Institute of Mathematical Sciences
[3] CIFAR Fellow
igul222@gmail.com
{faruk.ahmed,vincent.dumoulin,aaron.courville}@umontreal.ca
ma4371@nyu.edu

Abstract

Generative Adversarial Networks (GANs) are powerful generative models, but suffer from training instability. The recently proposed Wasserstein GAN (WGAN) makes progress toward stable training of GANs, but can still generate low-quality samples or fail to converge in some settings. We find that these problems are often due to the use of weight clipping in WGAN to enforce a Lipschitz constraint on the critic, which can lead to pathological behavior. We propose an alternative to clipping weights: penalize the norm of gradient of the critic with respect to its input. Our proposed method performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning, including 101-layer ResNets and language models over discrete data. We also achieve high quality generations on CIFAR-10 and LSUN bedrooms. [1]

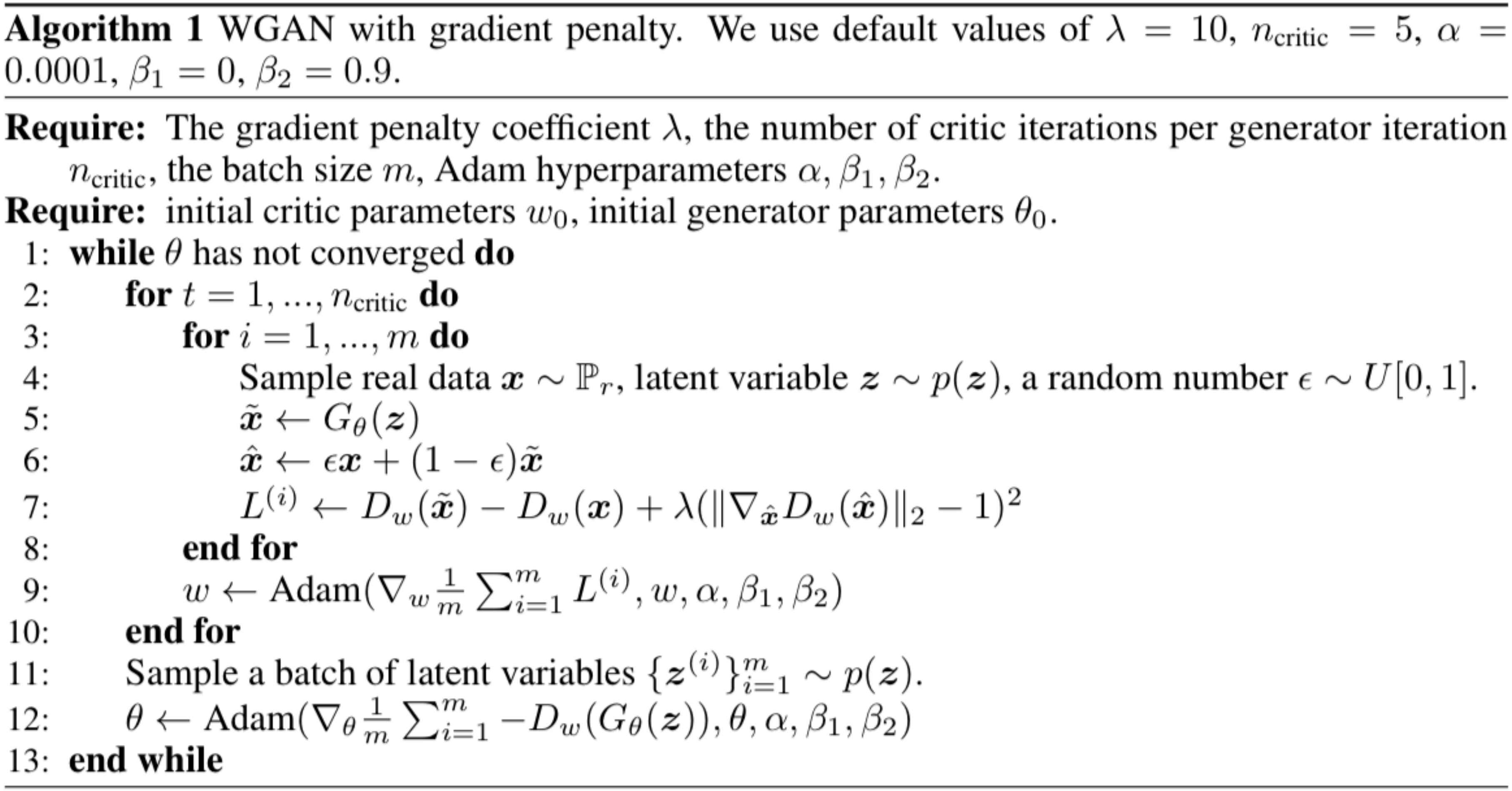
Properties of the Optimal WGAN Critic



In order to understand why weight clipping is problematic in a WGAN critic, as well as to motivate our approach, we highlight some properties of the optimal critic in the WGAN framework. We more formally state and prove these in the Appendix.

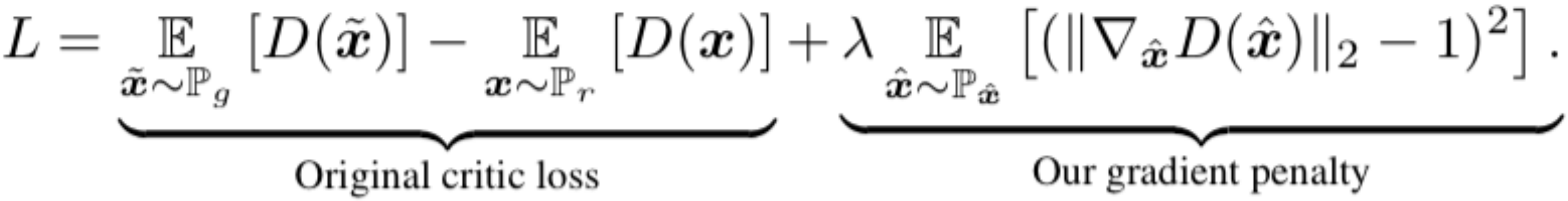
If the optimal critic under the Kantorovich-Rubinstein dual D^* is differentiable, and x is a point from our generator distribution $\mathbb{P}_g$, then there is a point y sampled from the true distribution $\mathbb{P}_r$ such that the gradient of D^* at all points $x_t = (1-t)x + ty$ on a straight line between x and y points directly towards y, meaning $\nabla D^*(x_t) = \frac{y - x_t}{||y - x_t||}$.

This implies that the optimal critic has unit gradient norm almost everywhere under $\mathbb{P}_r$ and $\mathbb{P}_g$.

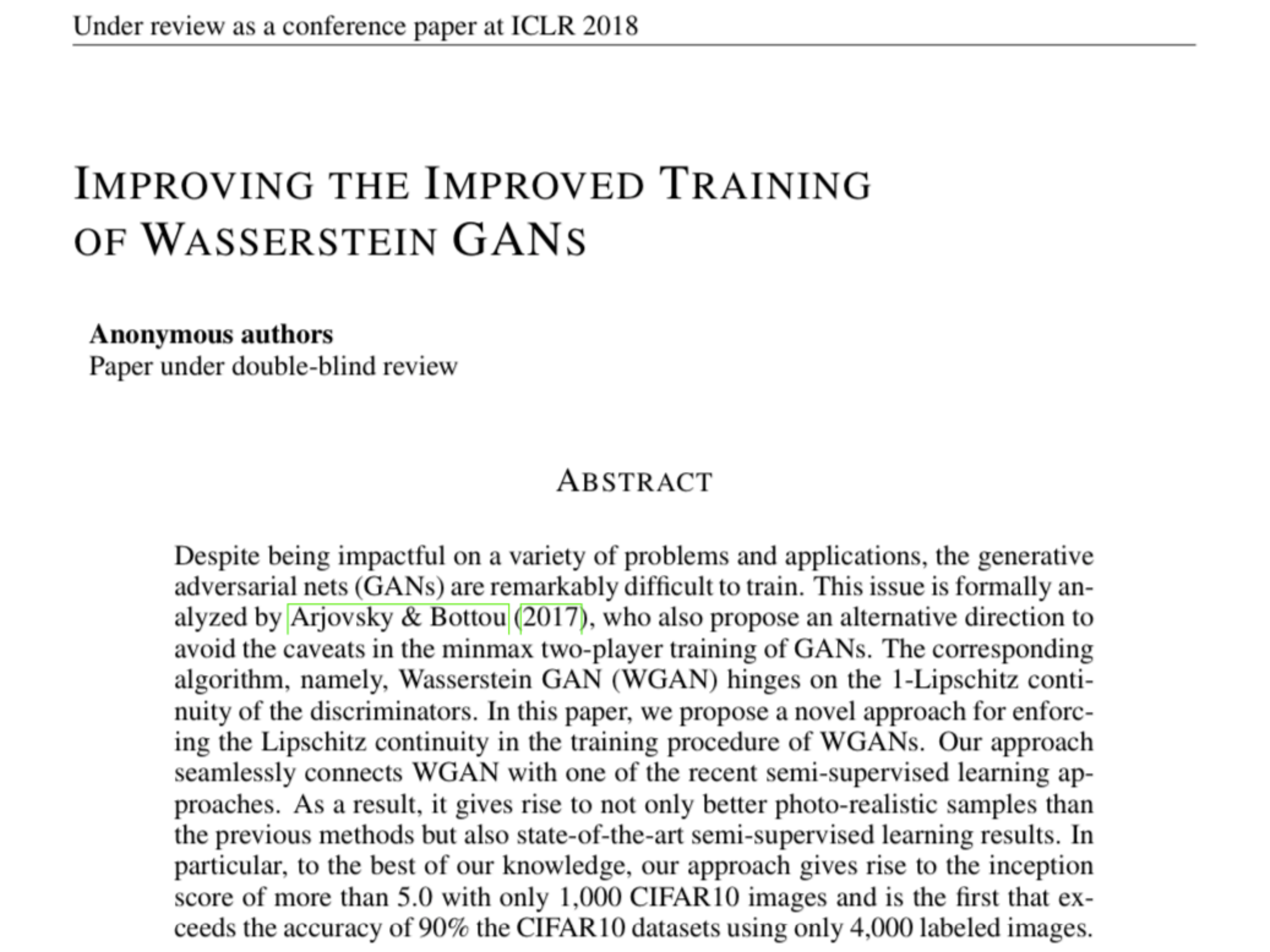


Algorithm 1 WGAN with gradient penalty. We use default values of $\lambda = 10$, $n_{\text{critic}} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$.

```
Require: The gradient penalty coefficient $\lambda$, the number of critic iterations per generator iteration
    $n_{\text{critic}}$, the batch size $m$, Adam hyperparameters $\alpha, \beta_1, \beta_2$.
Require: initial critic parameters $w_0$, initial generator parameters $\theta_0$.
while $\theta$ has not converged do
    for $t = 1, ..., n_{\text{critic}}$ do
        for $i = 1, ..., m$ do
            Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a random number $\epsilon \sim U[0, 1]$.
            $\tilde{\boldsymbol{x}} \leftarrow G_\theta(\boldsymbol{z})$
            $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon)\tilde{\boldsymbol{x}}$
            $L^{(i)} \leftarrow D_w(\tilde{\boldsymbol{x}}) - D_w(\boldsymbol{x}) + \lambda(\|\nabla_{\hat{\boldsymbol{x}}} D_w(\hat{\boldsymbol{x}})\|_2 - 1)^2$
        end for
        $w \leftarrow \text{Adam}(\nabla_w \frac{1}{m}\sum_{i=1}^m L^{(i)}, w, \alpha, \beta_1, \beta_2)$
    end for
    Sample a batch of latent variables $\{\boldsymbol{z}^{(i)}\}_{i=1}^m \sim p(\boldsymbol{z})$.
    $\theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m}\sum_{i=1}^m -D_w(G_\theta(\boldsymbol{z})), \theta, \alpha, \beta_1, \beta_2)$
end while
```



$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} \left[D(\tilde{\boldsymbol{x}})\right] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} \left[D(\boldsymbol{x})\right]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$



IMPROVING THE IMPROVED TRAINING OF WASSERSTEIN GANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite being impactful on a variety of problems and applications, the generative adversarial nets (GANs) are remarkably difficult to train. This issue is formally analyzed by Arjovsky & Bottou (2017), who also propose an alternative direction to avoid the caveats in the minmax two-player training of GANs. The corresponding algorithm, namely, Wasserstein GAN (WGAN) hinges on the 1-Lipschitz continuity of the discriminators. In this paper, we propose a novel approach for enforcing the Lipschitz continuity in the training procedure of WGANs. Our approach seamlessly connects WGAN with one of the recent semi-supervised learning approaches. As a result, it gives rise to not only better photo-realistic samples than the previous methods but also state-of-the-art semi-supervised learning results. In particular, to the best of our knowledge, our approach gives rise to the inception score of more than 5.0 with only 1,000 CIFAR10 images and is the first that exceeds the accuracy of 90% the CIFAR10 datasets using only 4,000 labeled images.

2.1 IMPROVING THE IMPROVED TRAINING OF WGAN

Let d denote the ℓ_2 metric on an input space used in this paper. A discriminator $D : \mathcal{X} \mapsto \mathcal{Y}$ is Lipschitz continuous if there exists a real constant $M \geqslant 0$ such that, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$,

$$d(D(\boldsymbol{x}_1), D(\boldsymbol{x}_2)) \leqslant M \cdot d(\boldsymbol{x}_1, \boldsymbol{x}_2). \tag{3}$$

Immediately, we can add the following soft consistency term (CT) to the value function of WGAN in order to penalize the violations to the inequality in eq. (3),

$$CT|_{\boldsymbol{x}_1,\boldsymbol{x}_2} = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\max\left(0, \frac{d(D(\boldsymbol{x}_1), D(\boldsymbol{x}_2))}{d(\boldsymbol{x}_1, \boldsymbol{x}_2)} - M'\right)\right] \tag{4}$$

Remarks. Here we face the same snag as in (Gulrajani et al., 2017), *i.e.*, it is impractical to substitute all the possibilities of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ pairs into eq. (4). What pairs and which regions of the input set $\mathcal{X}$ should we check for eq. (4)? Arguably, it is fairly safe to limit our scope to the manifold that supports the real data distribution $\mathbb{P}_r$ and its surrounding regions. After all, the distribution of the generative model $\mathbb{P}_G$ is desired to be as close as possible to $\mathbb{P}_r$. We use the notation M in eq. (3) and a different M' in eq. (4) to reflect the fact that the continuity will be checked only sparsely at some data points in practice.

What Have We Learned?

- Why are GANs hard to train?
    - Mode collapse
    - Unstable training behavior
    - Loss function lacking information about convergence
- Theory?