10707
Deep Learning: Spring 2020

Andrej Risteski

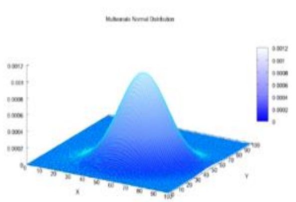
Machine Learning Department

Lecture 17:
Generative adversarial networks
Part II: Statistical Issues surrounding GANs

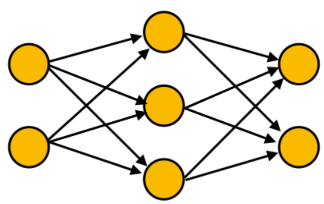
The GAN paradigm (Goodfellow et al. ‘14)

Goal: **Learn** a distribution close to some distribution we have few samples from. (Additionally, we will be able to sample efficiently from distribution.)

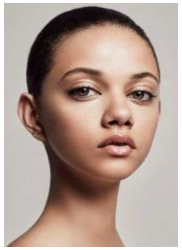
Approach: **Fit** distribution P_g parametrized by **neural network g**

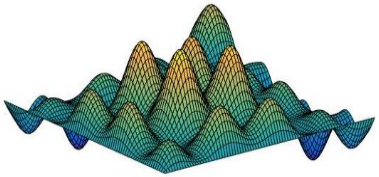


$Z \sim N(0, I_{k\times k})$



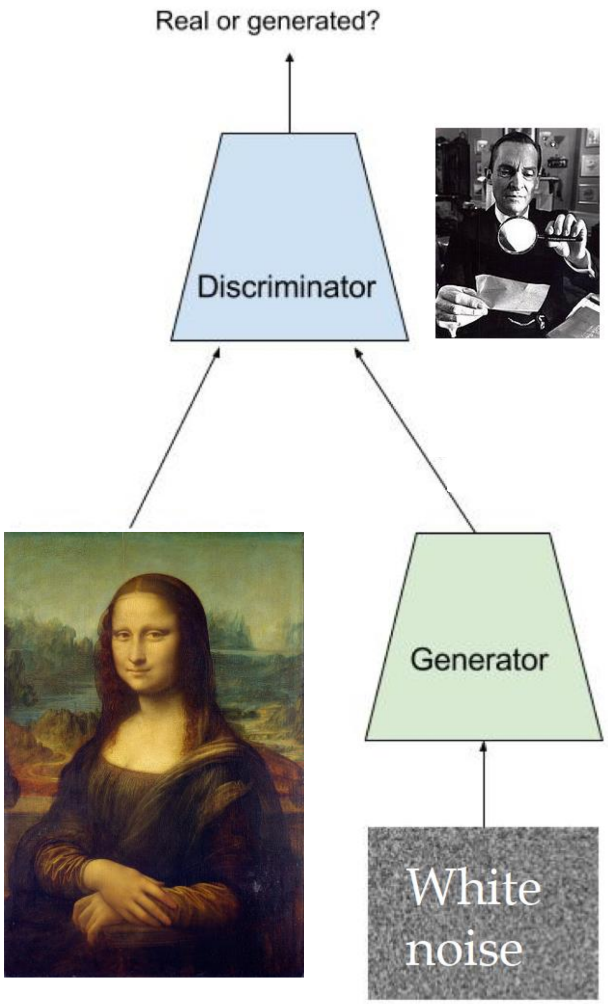
Neural network $g(\cdot)$





$X = g\,(Z)$

The GAN paradigm (Goodfellow et al. ‘14)



Game theoretic idea:

Generator trained to **fool** discriminator.

Discriminator trained to **beat** generator.

W-GAN formalization (Arjovsky et al. ‘17)

Min-max problem:

- Min-player: generators $g \in G$; Max-player: discriminators f $\in F$.
- Samples from image distr. P_{real}. Unif. distribution over samples: $P_{samples}$
- P_g - generator distribution: $Z \sim N(0, I) \rightarrow g(Z)$

Training loss:

$$\min_{g \in G} \max_{f \in F} \left| \mathbb{E}_{P_g}[f] - \mathbb{E}_{P_{\text{samples}}}[f] \right|$$

Difference of expectation of f on **samples vs generated** images

Examples of distances d_F

$$\underbrace{\max_{f \in F} \left| \mathbb{E}_{P_g}[f] - \mathbb{E}_{P_{\text{samples}}}[f] \right|}_{d_F(P_{\text{samples}}, P_g)}$$

Absolute value can be removed (-f is Lip if f is Lip)

$F = \{f : |f|_\infty \leq 1\}$: **Total variation distance**

Measures differences of bounded functions

$F = \{f : \mathrm{Lip}(f) \leq 1\}$: **$W_1$ (Wasserstein, earthmover) distance**

Measures differences of 1-Lipschitz functions

What affects our choice of F?

Statistical considerations: very powerful discriminators (e.g. large neural networks) will require a lot of samples. Weak discriminators will specify a very weak metric: very "different" distributions will look very "similar" to metric.

Our understanding here is much better.

Algorithmic considerations: if discriminators are very powerful, gradient information for generator is too weak and can vanish. If they are too weak – metric is weak.

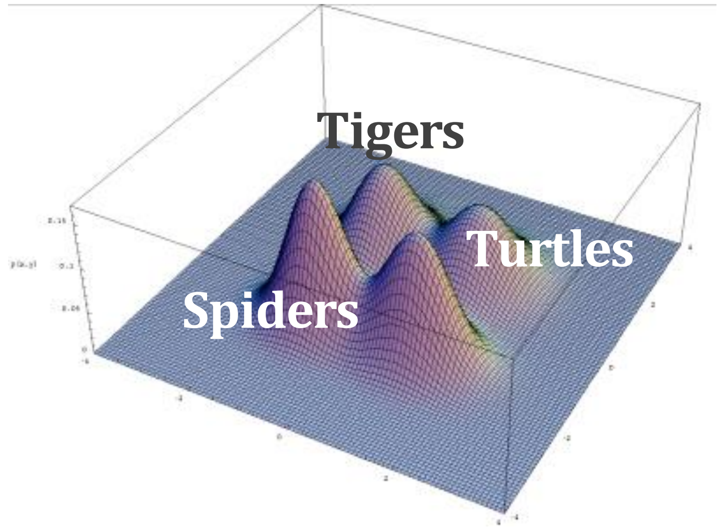
Our understanding of training dynamics is very poor.

Statistical questions

<u>Tension</u>: strength of discriminators?

Small (weak) discriminators ⇒ mode collapse:

Neural net discriminators with **≤ m parameters** **fooled** by generator w/ support size ≈ **m**.
[Arora et al'17, Arora-Risteski-Zhang **ICLR'18**]



Real-life distributions have large support!

Tension: strength of discriminators?

Small (weak) discriminators ⇒ mode collapse.

Happens for any P_{real}

Neural net discriminators with **≤ m parameters fooled** by generator w/ support size **≈ m**.
[Arora et al'17, Arora-Risteski-Zhang **ICLR'18**]

Not memorization!
More training samples
don't help.

Discriminators too weak: $\mathbf{d}_{\mathbf{F}}$ cannot distinguish between small-support distr. and P_{real}.

Real-life distributions have large support!

Weak discriminators ⇒ mode collapse

Small (weak) discriminators ⇒ mode collapse:

Neural net discriminators with **≤ m parameters fooled** by generator w/ support size **≈ m**.
[Arora et al'17, Arora-Risteski-Zhang **ICLR'18**]

Thm (*Arora et al '17*): Let F contain neural networks w/ some architecture w/ at most m trainable weights, are L-Lipschitz and outputs in [0,1]. Let the weights $\theta \in \Theta \subseteq \mathbb{B}^m$. Let $P_{generator}$ be the uniform distribution over $N \geq c\,\frac{m \log\left(\frac{Lm}{\epsilon}\right)}{\epsilon^2}$ iid samples from P_{real} for some absolute const. c. Let number of training samples be at least N. Then, whp over the choice of $P_{generator}$ and training data , we have:

$$\forall f \in F\colon\ |\mathbb{E}_{P_{generator}} f - \mathbb{E}_{P_{samples}} f| \leq \epsilon$$

Unit ball

In the **model parameters**:

$$\forall x\colon\ \left|\mathrm{f}_\theta(\mathrm{x}) - \mathrm{f}_{\hat\theta}(\mathrm{x})\right| \leq L\,\left\|\theta - \hat\theta\right\|_2$$

Can grow to infinity

Weak discriminators $\Rightarrow$ mode collapse

Thm (*Arora et al '17*): Let F contain neural networks w/ some architecture w/ at most m trainable weights, are L-Lipschitz and outputs in [0,1]. Let the weights $\theta \in \Theta \subseteq \mathbb{B}^m$. Let $P_{generator}$ be the uniform distribution over $N \geq c\,\frac{m \log\left(\frac{Lm}{\epsilon}\right)}{\epsilon^2}$ iid samples from P_{real} for some absolute const. c. Let number of training samples be at least N. Then, whp over the choice of $P_{generator}$ and training data , we have:

$$\forall f \in F\colon\ |\mathbb{E}_{P_{generator}} f - \mathbb{E}_{P_{samples}} f| \leq \epsilon$$

Proof: Let $P_{generator}$ be a distribution over N random samples from P_{real}.

Consider a **fixed** $f \in F$. By **Chernoff's inequality**, with we have:

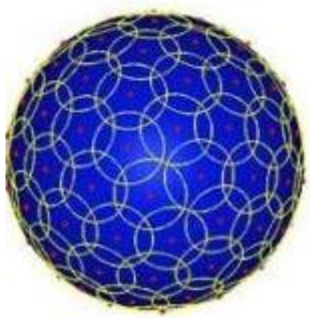
$$\Pr\left[\left|\mathbb{E}_{P_{real}} f - \mathbb{E}_{P_{generator}} f\right| \geq \frac{\epsilon}{4}\right] \leq 2\exp\left(-\frac{\epsilon^2 N}{2}\right)$$

We will perform a **union bound**, along with an **epsilon net argument**.

Why can't we immediately do a union bound? F is not discrete!

Epsilon nets

How many "mostly different" neural nets are there?



Def: An ϵ −net for Θ is a set Θ_ϵ s.t.
for every $\theta \in F, \exists \hat{\theta} \in \Theta_\epsilon$: $\left|\left|\theta - \hat{\theta}\right|\right|_2 \leq \epsilon$

Easy construction: there exists an ϵ −net of the m-dim unit ball w/ size $O\left(\left(\frac{1}{\epsilon}\right)^m\right)$
(Intuitive: the volume of a ϵ-radius ball is $\sim \epsilon^m$)

Why is this useful? By *Lipschitzness*, if we have two discriminators $f_\theta, f_{\hat{\theta}}$

$$\forall x: \left|f_\theta(x) - f_{\hat{\theta}}(x)\right| \leq L\,\epsilon$$

Weak discriminators ⇒ mode collapse

Let $P_{generator}$ be a distribution over N random samples from P_{real}.

Consider a fixed $f \in F$. By Chernoff's inequality, with we have:

$$\Pr\left[\left|\mathbb{E}_{P_{real}} f - \mathbb{E}_{P_{generator}} f\right| \geq \frac{\epsilon}{4}\right] \leq 2\exp\left(-\frac{\epsilon^2 N}{2}\right)$$

Consider an $\frac{\epsilon}{4L}-$ net of F, which has size $\exp\left(O\left(m\log\left(\frac{L}{\epsilon}\right)\right)\right)$.

Since $N \geq c\,\frac{m\log\left(\frac{Lm}{\epsilon}\right)}{\epsilon^2}$, the probability on the RHS is bounded by $2\exp\left(-\frac{cm\log\left(\frac{Lm}{\epsilon}\right)}{2}\right)$

Thus, **union bounding** over the $\frac{\epsilon}{4L}-$ net, we have, for a sufficiently large c, that

$$\forall \theta \in \Theta_{\frac{\epsilon}{4L}}:\ \Pr\left[\left|\mathbb{E}_{P_{real}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right| \geq \frac{\epsilon}{4}\right] \leq \exp(-m)$$

Weak discriminators $\Rightarrow$ mode collapse

$$\forall \theta \in \Theta_{\frac{\epsilon}{4L}}:\ \Pr\left[\left|\mathbb{E}_{P_{real}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right| \geq \frac{\epsilon}{4}\right] \leq \exp(-m)$$

By **exactly** the same argument, we have

$$\forall \theta \in \Theta_{\frac{\epsilon}{4L}}:\ \Pr\left[\left|\mathbb{E}_{P_{real}} f_\theta - \mathbb{E}_{P_{samples}} f_\theta\right| \geq \frac{\epsilon}{4}\right] \leq \exp(-m)$$

Since $\left|\mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right| =$

$$\left|\mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{real}} f_\theta + \mathbb{E}_{P_{real}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right| \leq$$

$$\left|\mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{real}} f_\theta\right| + \left|\mathbb{E}_{P_{real}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right|$$

Hence, with probability at least $1 - 2\exp(-m)$

$$\forall \theta \in \Theta_{\frac{\epsilon}{4L}}:\ \left|\mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta\right| \leq \frac{\epsilon}{2}$$

Weak discriminators ⇒ mode collapse

Hence, with probability at least $1 - 2\exp(-m)$

$$\forall \theta \in \Theta_{\frac{\epsilon}{4L}}: \quad \left| \mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta \right| \leq \frac{\epsilon}{2}$$

Consider any $\theta \in \Theta$. By the definition of an $\frac{\epsilon}{4L}$-net, there exists a $\hat{\theta} \in \Theta_{\frac{\epsilon}{4L}}$, s.t. $\forall x: |f_\theta(x) - f_{\hat{\theta}}(x)| \leq \epsilon/4$. Hence,

$$\left| \mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{generator}} f_\theta \right|$$

$$= \left| \mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{samples}} f_{\hat{\theta}} + \mathbb{E}_{P_{samples}} f_{\hat{\theta}} - \mathbb{E}_{P_{generator}} f_{\hat{\theta}} + \mathbb{E}_{P_{generator}} f_{\hat{\theta}} - \mathbb{E}_{P_{generator}} f_\theta \right|$$

$$\leq \left| \mathbb{E}_{P_{samples}} f_\theta - \mathbb{E}_{P_{samples}} f_{\hat{\theta}} \right| + \left| \mathbb{E}_{P_{samples}} f_{\hat{\theta}} - \mathbb{E}_{P_{generator}} f_{\hat{\theta}} \right| + \left| \mathbb{E}_{P_{generator}} f_{\hat{\theta}} - \mathbb{E}_{P_{generator}} f_\theta \right|$$

$$\leq \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon$$

Which is indeed what we want.

Tension: strength of discriminators

Large discriminators $\Rightarrow$ poor generalization:

Loss with small # samples differs a lot from loss with infinite # samples.

$$d_F(P_{samples}, P_g) \not\approx d_F\left(P_{real}, P_g\right)$$

This is a problem even for distributions as simple as a standard Gaussian!

For instance, if P_{real} is a standard d-dimensional Gaussian, with any poly(d) number of samples, with high probability $W_1\left(P_{samples}, P_{real}\right) \geq 1.1$

(Like sampling random pts on the unit sphere: in high dimensions they will be far away with high probability)

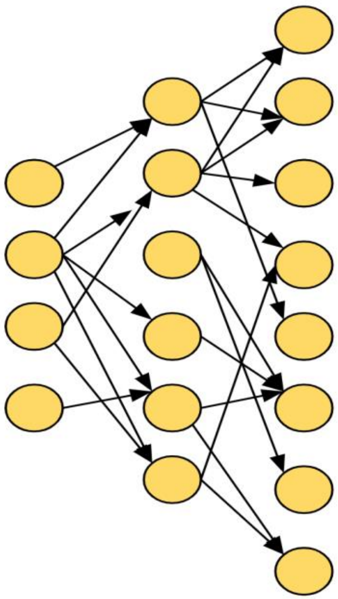
In other words, the class of all Lipschitz function is too large!!!

Sweet spot for natural distributions

Let P_{real} **itself** be generated by neural net. ($P_{real} = P_g$, $g \in G$)

Let G = { **1-to-1 neural networks** of bounded size }

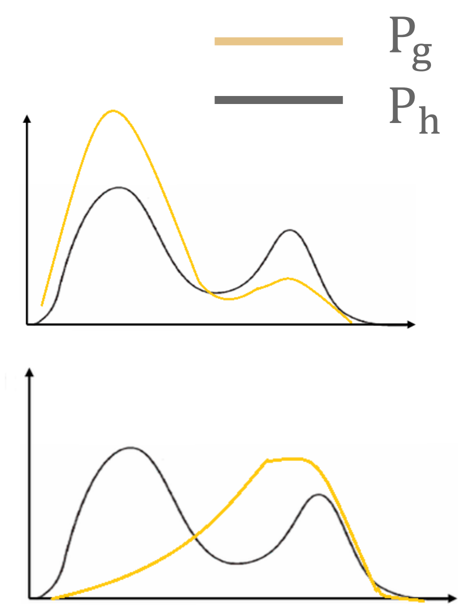
Design **small** discriminators F w/ good distinguishing power.



- Less general than arbitrary neural-net generators ✗
- Allows data to lie on **low-dim. manifold**. ✓

Distinguishing power

Discriminators F have **distinguishing power** against generators G, if: $\forall \mathrm{g}, \mathrm{h} \in \mathrm{G} : d_F\left(\mathrm{P_g}, \mathrm{P_h}\right) \gtrsim JS(P_g, P_h)$



Distance d_F is not too much weaker than JS, **only** for distributions in our class.

$JS(\mathrm{p}, \mathrm{q}) = \mathrm{KL}(\mathrm{q}||\mathrm{p}) + \mathrm{KL}(\mathrm{p}||\mathrm{q})$

Mai

Neural nets of **slightly larger** depth/size than generators.
(Suggestion for practice!)

Thm (Bai-Ma-Risteski ICLR'19**) :** **Small** discriminators F with **distinguishing power** for G = {**1-to-1 neural nets** of bdd size} exist.

So, if P_{real} generated by 1-to-1 neural net with **d** params,

w/ **poly(d)** samples, $d_F\left(P_{samples}, P_g\right) \leq \epsilon \Rightarrow JS\left(P_{real}, P_g\right) \leq O(\epsilon)$

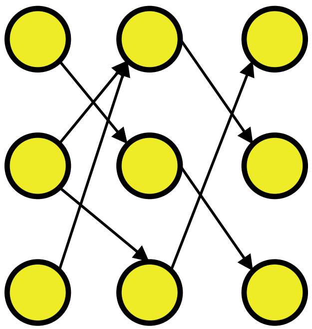
Training was successful

True distribution learned.

Natural distributions: more formally

Let P_{real} **itself** be generated by neural net. ($P_{real} = P_g$, $g \in G$)

Let G be the set of neural networks $\mathbb{R}^d \to \mathbb{R}^d$ that are:



Parametrized by weight matrices $W_i \in \mathbb{R}^{d \times d}$, biases $b_i \in \mathbb{R}^d$

Invertible: all W_i are *full-rank*, non-linearity σ is *invertible* and differentiable.

Number of layers is l. (Size is clearly bdd by $l\ d^2$)

Main result

Thm (Bai-Ma-Risteski ICLR'19**) :** Let P_{real} generated by 1-to-1 neural net with depth bounded by l and invertible, differentiable activation.

Let F be the set of neural networks of depth l+1, size O$(l\ d^2)$ activations $\sigma^{-1}, (\cdot)^2, \log\ (\sigma^{-1})'$.

Then, if we have $N \geq poly(d, l, 1/\epsilon)$ training samples,

$d_F\left(P_{samples}, P_g\right) \leq \epsilon \Rightarrow JS\left(P_{real}, P_g\right) \leq O(\epsilon)$

Distinguishing power: main idea

Discriminators F have **distinguishing power** against generators G,

if: $\forall \mathrm{g}, \mathrm{h} \in \mathrm{G} : d_F\left(\mathrm{P_g}, \mathrm{P_h}\right) \gtrsim JS(P_g, P_h)$

Claim: if F is chosen as the set of neural networks of depth l+1, size $\mathrm{O}(l\ d^2)$ activations $\sigma^{-1}, (\cdot)^2, \log\ (\sigma^{-1})'$, then F has distinguishing power against G.

Distinguishing power

What does this buy us?

Remember, small training error means $d_F\left(\mathrm{P}_{\mathrm{samples}}, \mathrm{P}_{\mathrm{g}}\right)$ is small.

Since the neural networks in F are bounded in size (i.e. the capacity of the class is bounded): one can use similar techniques as the ones we saw in the section on generalization to show that if we have N training samples

$$d_F\left(\mathrm{P}_{\mathrm{samples}}, \mathrm{P}_{\mathrm{g}}\right) = d_F\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right) \pm \frac{poly(d)}{N}$$

Taking $N \geq poly\left(d, \frac{1}{\epsilon}\right)$, $\left|d_F\left(\mathrm{P}_{\mathrm{samples}}, \mathrm{P}_{\mathrm{g}}\right) - d_F\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right)\right| \leq \epsilon$

But, by what we showed, we also have $d_F\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right) \geq JS\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right)$. Hence:

$$JS\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right) \leq d_F\left(\mathrm{P}_{\mathrm{real}}, \mathrm{P}_{\mathrm{g}}\right) + \epsilon$$

Distinguishing power: main idea

Discriminators F have **distinguishing power** against generators G,

if: $\forall \mathrm{g}, \mathrm{h} \in \mathrm{G} : d_F\left(\mathrm{P_g}, \mathrm{P_h}\right) \gtrsim JS(P_g, P_h)$

Claim: if F is chosen as the set of neural networks of depth l+1, size $\mathrm{O}(l\ d^2)$ activations $\sigma^{-1}, (\cdot)^2, \log\ (\sigma^{-1})'$, then F has distinguishing power against G.

Proof: Remember that $d_F\left(\mathrm{P_g}, \mathrm{P_h}\right) = \max_{f \in F} |\mathbb{E}_{P_g} f - \mathbb{E}_{P_h} f|$

On the other hand, we also have $JS\left(P_g, P_h\right) = KL(P_g || P_h) + KL(P_h \| P_g)$

$$= \mathbb{E}_{P_g} (\log P_g - \log P_h) - \mathbb{E}_{P_h} (\log P_g - \log P_h)$$

Suppose it were the case that $\log P_g - \log P_h \in F$: then, we'd have

$$\max_{f \in F} |\mathbb{E}_{P_g} f - \mathbb{E}_{P_h} f| \geq |\mathbb{E}_{P_g}(\log P_g - \log P_h) - \mathbb{E}_{P_h}\left(\log P_g - \log P_h\right)| \geq JS\left(P_g, P_h\right)$$

Distinguishing power: the density

Discriminators F have **distinguishing power** against generators G,

if: $\forall \mathrm{g}, \mathrm{h} \in \mathrm{G} : d_F\left(\mathrm{P_g}, \mathrm{P_h}\right) \gtrsim JS(P_g, P_h)$

So, it suffices to show that $\forall g, h$: $\log P_g - \log P_h \in F$

First, notice that if $x = g(z)$, then (inverting one layer at a time):

$$z = W_1^{-1}(\sigma^{-1}(W_2^{-1}(\sigma^{-1}(\ldots \underbrace{\sigma^{-1}(W_l^{-1}(x - b_l)}_{\text{Invert one layer}} - \cdots) - b_2) - b_1)$$

Let us denote the map above by g^{-1}. Let us denote by $\phi(z)$ the density of z under the standard Gaussian. Then, by the change of variables formula:

$$P_g(x) = \phi(g^{-1}(x))|\det(J_x(g^{-1}(x))|$$

Jacobian wrt x

Distinguishing power: the density

$$\text{So,}\ \log P_g(x) = \log \phi\left(g^{-1}(x)\right) + \log|\det\left(J_x\left(g^{-1}(x)\right)\right)|$$

Consider the first term: $g^{-1}(x)$ is a neural network of depth l, size O(l d^2) and activations σ^{-1}.

As $\phi\left(g^{-1}(x)\right) = Z + \exp\left(-\left|\left|g^{-1}(x)\right|\right|^2\right)$, we have $\log \phi(g^{-1}(x)) = -\left|\left|g^{-1}(x)\right|\right|^2$

$$\left|\left|g^{-1}(x)\right|\right|^2 = \sum_i \left(g_i^{-1}(x)\right)^2$$

Hence, $\phi(g^{-1}(x))$ can be represented by an extra layer on top of $g^{-1}(x)$ with activation $(\cdot)^2$.

Distinguishing power: the Jacobian

So, $\log P_g(x) = \log \phi\left(g^{-1}(z)\right) + \log\left|\det\left(J_x\left(g^{-1}(x)\right)\right|\right.$

$$g^{-1}(x) = W_1^{-1}(\sigma^{-1}(W_2^{-1}(\sigma^{-1}\left(\ldots\sigma^{-1}\left(W_l^{-1}(x - b_l) - \cdots\right) - b_2\right) - b_1)$$

Let us denote: $h_l = W_l^{-1}(x - b_l)$, $h_{l-1} = W_{l-1}^{-1}(\sigma^{-1}(h_l) - b_l)$, etc.

Claim: $J_x\left(g^{-1}(x)\right) = W_1^{-1} diag\left((\sigma^{-1})'(h_2)\right)W_2^{-1} \ldots W_{l-1}^{-1} diag\left((\sigma^{-1})'(h_l)\right)W_l^{-1}$

Pf: As a simple special case: $\frac{\partial}{\partial x_j}\sigma^{-1}\left(W_l^{-1}(x - b_l)\right)_i = (\sigma^{-1})'(h_l)\left(W_l^{-1}\right)_{ij}$

Writing it as a matrix: $\frac{\partial}{\partial x}\sigma^{-1}\left(W_l^{-1}(x - b_l)\right) = W_l^{-1}\, diag((\sigma^{-1})'(h_l))$

The claim follows by a similar calculation and the chain rule.

Distinguishing power: the Jacobian

Claim: $J_x\left(g^{-1}(x)\right) = W_1^{-1} diag\left((\sigma^{-1})'(h_2)\right) W_2^{-1} \dots W_{l-1}^{-1} diag\left((\sigma^{-1})'(h_l)\right) W_l^{-1}$

Since det(AB) = det(A) det(B), we have

$$\log\det\left(J_x\left(g^{-1}(x)\right)\right) = C + \sum_{k=1}^{l} \sum_{i=1}^{d} \log\ (\sigma^{-1})'(h_k)_i$$

Which clearly is expressible as a l-layer neural net with size $O(ld^2)$ and activations $\log\ (\sigma^{-1})'$.

Altogether, we get that $\forall g \in G, \log P_g \in F$, from which we get $\forall g, h\colon \log P_g - \log P_h \in F$

Hence, $d_F(\mathrm{P_g}, \mathrm{P_h}) \geq JS(P_g, P_h)$, i. e. F has distinguishing power wrt to F.