Paper Review:

"WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points [No et al., 2021]"

Speaker : Hanseul Cho

OptiML Lab, GSAI, KAIST

April 8, 2022

Why this paper?

WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points

Albert No [1] **TaeHo Yoon** [2] **Sehyun Kwon** [2] **Ernest K. Ryu** [2]

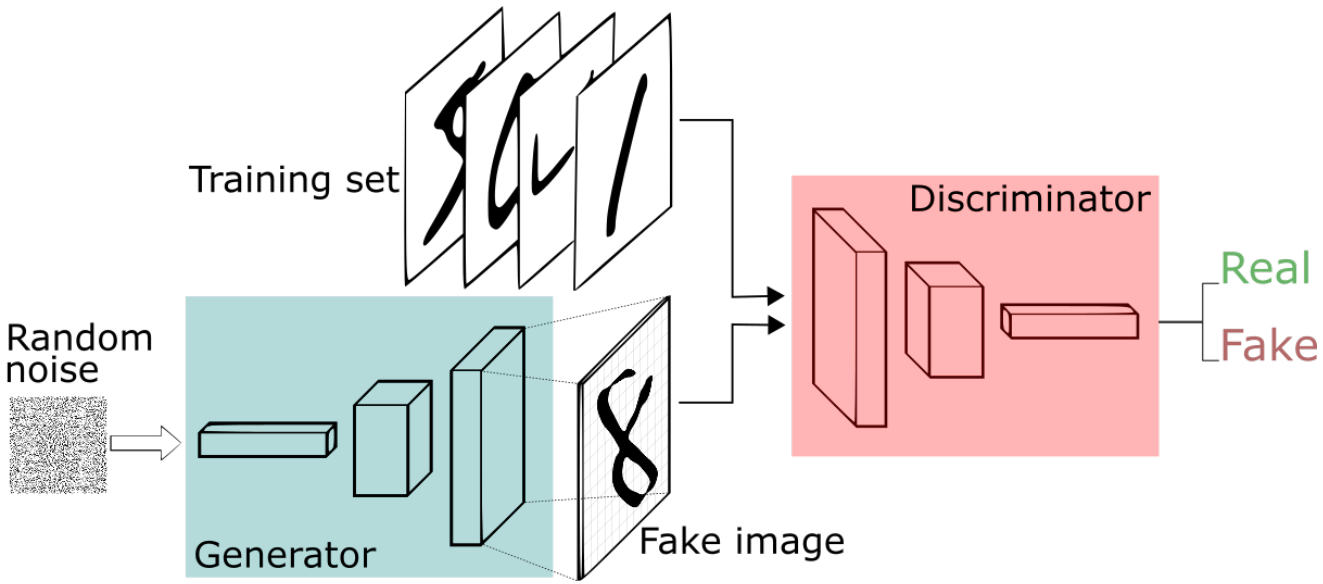
Abstract

Generative adversarial networks (GAN) are a widely used class of deep generative models, but their minimax training dynamics are not understood very well. In this work, we show that GANs with a 2-layer infinite-width generator and a 2-layer finite-width discriminator trained with stochastic gradient ascent-descent have no spurious stationary points. We then show that when the width of the generator is finite but wide, there are no spurious stationary points within a ball whose radius becomes arbitrarily large (to cover the entire parameter space) as the width goes to infinity.

- GAN: empirically well-known example of minimax problem.
- *Crucial skill* of the proofs: "Universal Approximation" results.

Outline

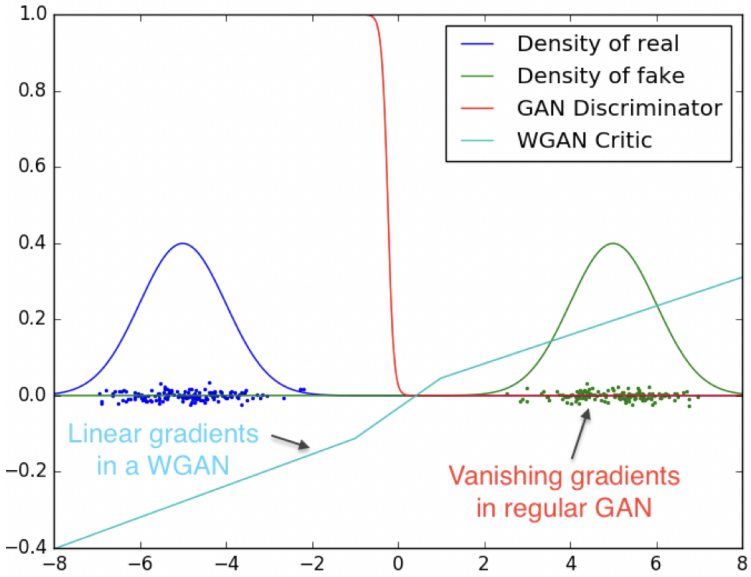
GAN(Generative Adversarial Network)



- **Generator** wants to maximize the loss (to fake Discriminator),
- while **Discriminator** wants to minimize the loss (to teach Generator).

WGAN(Wasserstein GAN [Arjovsky et al., 2017])

We want a **good generator** that estimates **true distribution** well.



That is, we want to reduce the "distance" between true / generated distributions.

WGAN(Wasserstein GAN [Arjovsky et al., 2017])

[Arjovsky et al., 2017] used Wasserstein-1 distance to measure the difference of distributions:

$$W(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma} \|x - y\|$$
$$= \sup_{f:\text{1-Lipschitz}} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x'\sim\mathbb{Q}}[f(x')]$$

... applying some duality argument (Kantorovich-Rubinstein). Refer to the Remark 6.5 of [Villani, 2008] for detailed descriptions.

- **Generator** $g_\theta(z)$ wants to minimize this distance;
  - it solves another 'minimax' problem!
- Lipschitz continuous $f \to$ **Discriminator** $f_\eta(\cdot)$!
  - it wants to maximize (a kind of) 'distance'.

Want to solve: where the discriminator f_η is Lipschitz continuous,

$$\inf_\theta \sup_\eta \Big(\mathbb{E}_X[f_\eta(X)] - \mathbb{E}_Z[f_\eta(g_\theta(Z))] \Big)$$

Regularized WGAN loss

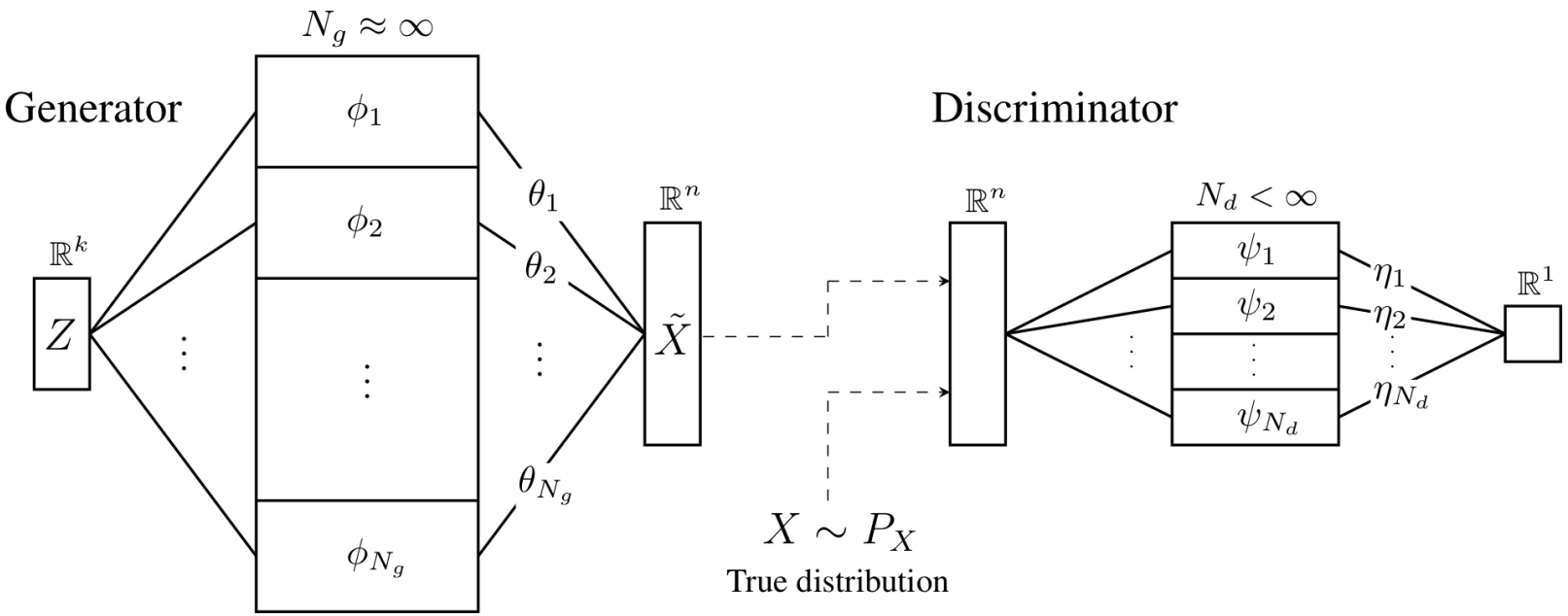
In [Arjovsky et al. 2017], they use *weight clipping* in the training process to maintain the discriminator to be Lipschitz.
Rather, we could explicitly **regularize the discriminator** [Lei et al. 2020]:

(A variant of) Loss Function

$$\inf_{\theta} \sup_{\eta} L(\theta, \eta) := \mathbb{E}_X[f_\eta(X)] - \mathbb{E}_Z[f_\eta(g_\theta(Z))] - \frac{1}{2}\|\eta\|^2 .$$

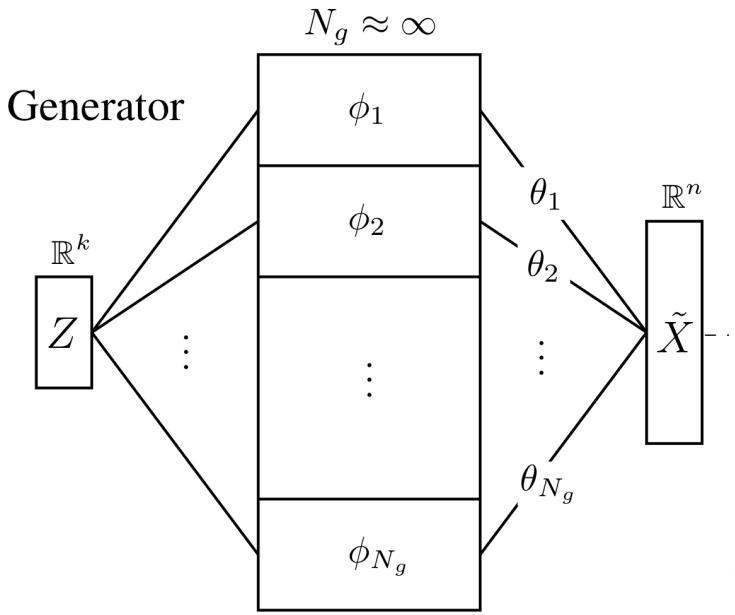
...we will come back to this formulation soon.

Architecture of WGAN (this paper)



- Generator and discriminator are both 2-layer networks.
  - Each 1st layer is nonlinear, randomly initialized, and not to be trained.
  - Each 2nd layer is linear, randomly initialized, and to be trained.
- Generator is sufficiently (and even infinitely-) wide.
- Discriminator is of finite-width.

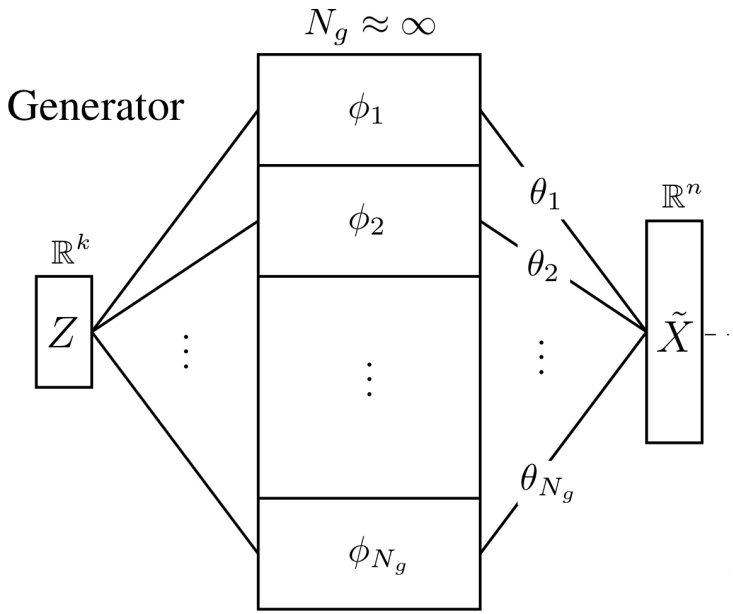
Generator of Finite-Width



- $Z \in \mathbb{R}^k$: Latent variable (noise) // $\tilde{X} \in \mathbb{R}^n$: Generated (fake) sample
- $\mathcal{G} = \{\phi(\cdot;\kappa) : \mathbb{R}^k \to \mathbb{R}^n | \kappa \in \mathbb{R}^p\}$: Generator feature functions.
  - nonlinear, randomly initialized, and not to be trained.
- Generator of width $N_g < \infty$:
  - $\theta \in \mathbb{R}^{N_g}$: randomly initialized and to be trained.

$$g_\theta(z) = \sum_{i=1}^{N_g} \theta_i \phi_i(z;\kappa_i), \ \phi_i \in \mathcal{G}$$

Generator of **Infinite-Width**



- $\mathcal{G} = \{\phi(\cdot;\kappa) : \mathbb{R}^k \to \mathbb{R}^n | \kappa \in \mathbb{R}^p\}$: Generator feature functions.
  - nonlinear, randomly initialized, and not to be trained.
- Generator of width $N_g = \infty$:
  - $\theta \in \mathcal{M}(\mathbb{R}^p)$: randomly initialized and to be trained.

$$g_\theta(z) = \int_{\mathbb{R}^p} \phi(z;\kappa) d\theta(\kappa) \left(\approx \int_{\mathbb{R}^p} \phi(z;\kappa)\theta'(\kappa) d\kappa \right)$$

Assumputions (1): Latent var. & Generator feature ftn.

(AL) The latent vector $Z \in \mathbb{R}^k$ has a Lipschitz continuous probability density function $q_Z(z)$ satisfying $q_Z(z) > 0$ for all $z \in \mathbb{R}^k$.

(AG) All generator feature functions $\phi \in \mathcal{G}$ are of form $\phi(z;\kappa) = \sigma_g(\kappa_w z + \kappa_b)$, where $\kappa = (\kappa_w, \kappa_b) \in \mathbb{R}^{n\times k} \times \mathbb{R}^n$, and $\sigma_g : \mathbb{R} \to \mathbb{R}$ is a bounded continuous activation function satisfying $\lim_{r\to-\infty} \sigma_g(r) < \lim_{r\to\infty} \sigma_g(r)$. (So $p = nk + n$.)

Easier version:

- (**AL**) $Z \in \mathbb{R}^k$ can be Gaussian.
- (**AG**) $\phi(z; A, b) = \sigma_g(Az + b)$ where σ_g is sigmoidal.

Universal Approximation Property (UAP)

(Universal approximation property) For any function $f : \mathbb{R}^k \to \mathbb{R}^n$ such that $\mathbb{E}_Z \left[\|f(Z)\|_2 \right] < \infty$ and $\varepsilon > 0$, there exists $\theta_\varepsilon \in \mathcal{M}(\mathbb{R}^p)$ such that

$$\mathbb{E}_Z \left[\|g_{\theta_\varepsilon}(Z) - f(Z)\|_2 \right] < \varepsilon.$$

Lemma 1 (Denseness of $\mathcal{G}$.)

With infinite-width generator, **(AG)** $\Rightarrow$ **(UAP)**. (Due to [Hornik, 1991].)

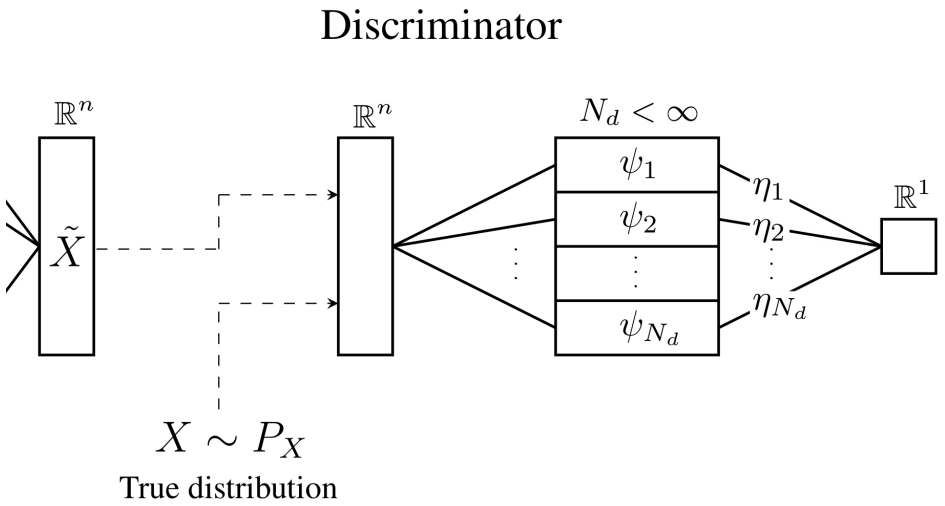
Lemma 2 (Dual of "Denseness of $\mathcal{G}$.")

Assume **(AL)** and **(UAP)**. If a bounded continuous function $h : \mathbb{R}^k \to \mathbb{R}^n$ holds

$$\mathbb{E}_Z[\phi(Z)^\top h(Z)] = 0 \quad \text{for every } \phi \in \mathcal{G},$$

then $h \equiv 0$.

Discriminator



- Input $\in \mathbb{R}^n$: either X (true data) or $\tilde{X} = g_\theta(Z)$. // Output $\in \mathbb{R}$.
- $\mathcal{D} = \{\psi_j : \mathbb{R}^n \to \mathbb{R} | j = 1, \ldots, N_d\}$: Discriminator feature functions.
  - nonlinear, randomly initialized, and not to be trained.
- Discriminator of width $N_d < \infty$:
  - $\eta \in \mathbb{R}^{N_d}$: randomly initialized and to be trained.

$$f_\eta(x) = \sum_{j=1}^{N_d} \eta_j \psi_j(x) = \eta^\top \Psi(x).$$

Assumputions (2): Discriminator feature ftn.

(AD) For all $1 \leq j \leq N_d$, the discriminator feature functions are of form $\psi_j(x) = \sigma(a_j^\intercal x + b_j)$ for some $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$. The twice differentiable activation function σ satisfies $\sigma'(x) > 0$ for all $x \in \mathbb{R}$ and $\sup_{x\in\mathbb{R}} |\sigma(x)| + |\sigma'(x)| + |\sigma''(x)| < \infty$. The weights $a_1, \ldots, a_{N_d}$ and biases $b_1, \ldots, b_{N_d}$ are sampled (IID) from a distribution with a probability density function.

Easier version:

- **(AD)** $\psi_j(x) = \sigma_d(a_j^\top x + b_j)$, where σ can be sigmoid or $\tanh$ and weights (a_j, b_j) are sampled i.i.d. (e.g. from Gaussian).

Reformulation to Minimization (loss function revisit)

$$L(\theta, \eta) = \mathbb{E}_X[f_\eta(X)] - \mathbb{E}_Z[f_\eta(g_\theta(Z))] - \frac{1}{2}\|\eta\|^2$$
$$= \mathbb{E}_X[\eta^\top \Psi(X)] - \mathbb{E}_Z[\eta^\top \Psi(g_\theta(Z))] - \frac{1}{2}\|\eta\|^2 .$$

Note: due to regularization term, $L(\theta, \eta)$ is concave w.r.t. η. Moreover, our minimax problem is equivalent to a **minimization** problem as follows.

$$\inf_\theta \sup_\eta L(\theta, \eta) = \inf_\theta J(\theta),$$

where

$$J(\theta) \triangleq \sup_\eta L(\theta, \eta) = \frac{1}{2}\|\mathbb{E}_X[\Psi(X)] - \mathbb{E}_Z[\Psi(g_\theta(Z))]\|^2 .$$

So far we've seen... (+Q&A?)

WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points

Albert No [1] **TaeHo Yoon** [2] **Sehyun Kwon** [2] **Ernest K. Ryu** [2]

Spurious/Non-Spurious Stationary Point

Definition. (Stationary point)

$\theta_s \in \mathbb{R}^N$ is a *stationary point* of a differentiable function $J : \mathbb{R}^N \to \mathbb{R}$ if

$$\nabla J(\theta_s) = 0.$$

A stationary point is **spurious** if it is NOT a global minimum.

However, What if $\theta \in \mathcal{M}(\mathbb{R}^p)$ (a Lebesgue measure)?

Stationary Point of $J(\theta)$

Recall:

$$J(\theta) \triangleq \sup_{\eta} L(\theta, \eta) = \frac{1}{2} \left\| \mathbb{E}_X[\Psi(X)] - \mathbb{E}_Z[\Psi(g_\theta(Z))] \right\|^2 .$$

Definition.

$\theta_s \in \mathcal{M}(\mathbb{R}^p)$ is a *stationary point* of a $J(\theta)$ if

$$J(\theta_s + \lambda\mu), \text{ as a function of } \lambda \in \mathbb{R},$$

is differentiable and has zero gradient at $\lambda = 0$ for any $\mu \in \mathcal{M}(\mathbb{R}^p)$,

$$\text{i.e., } \frac{\partial}{\partial \lambda} J(\theta_s + \lambda\mu)\Big|_{\lambda=0} = 0.^{(1)}$$

[1] The L.H.S. of the last equation is actually to the definition of '*Gâteaux derivative*', a generalization of directional derivative for the function of Lebesgue measure. Refer to [this link].

Main results of [No et al., 2021]

Note: $\begin{cases} N_g \text{ is the width of Generator,} \\ N_d \text{ is the width of Discriminator, and} \\ n \text{ is the size of sample } X \text{ \& } \tilde{X}. \end{cases}$

1. Infinite-width Generator & Small Discriminator ($N_d \leq n$)
   - Any stationary point is non-spurious, with probability 1.

2. Infinite-width Generator & Large Discriminator ($n < N_d < \infty$)
   - If the range of generator function contains an open ball, then any stationary point is non-spurious with probability 1.

3. Finite-width Generator ($n < N_g < \infty$) & Small Discriminator
   - If N_g is sufficiently large, then any stationary point in an (arbitrarily large) ball is non-spurious with high probability.

Infinite-width Generator & Small Discriminator ($N_d \leq n$)

Theorem 4 ($N_g = \infty$, $N_d \leq n$)

Assume (AL), (AG), and (AD). Then, with probability 1, any stationary point θ_s of J satisfies $J(\theta_s) = 0$.

Remark:

- $J(\theta_s) = 0$ implies θ_s is a global minimum of J: non-spurious.
- "$N_d \leq n$" is important: (AD) & $N_d \leq n$ implies that $\nabla\psi_1(x), \cdots, \nabla\psi_{N_d}(x)$ are linearly independent (with probability 1).

Proof of Theorem 4

Let $r(\theta) \triangleq \mathbb{E}_X[\Psi(X)] - \mathbb{E}_Z[\Psi(g_\theta(Z))]$; then, $J(\theta) = \frac{1}{2}\|r(\theta)\|^2$. Let θ_s be a stationary point of J. Then, for any $\mu \in \mathcal{M}(\mathbb{R}^p)$,

$$0 = \frac{\partial}{\partial \lambda} J(\theta_s + \lambda\mu)\Big|_{\lambda=0} = -\mathbb{E}_Z\left[\sum_{j=1}^{N_d} r_j(\theta_s)\nabla\psi_j(g_{\theta_s}(Z))^\top g_\mu(Z)\right].$$

By Lemma 2 (Dual of Denseness of $\mathcal{G}$, which is implied by (AL), (AG)),

$$\sum_{j=1}^{N_d} r_j(\theta_s)\nabla\psi_j(g_{\theta_s}(Z)) = 0.$$

Note that (AD) & $N_d \leq n$ implies that, with probability 1, $\nabla\psi_1(x)$,...,$\nabla\psi_{N_d}(x)$ are linearly independent. Therefore, $r(\theta_s) = 0$, and $J(\theta_s) = 0$, with probability 1.

Infinite-width Generator & Large Discriminator ($n < N_d < \infty$)

If $n < N_d$, then the n dimensional vectors $\nabla\psi_1(x)$, ..., $\nabla\psi_{N_d}(x)$ CANNOT be linearly independent. However, a similar result still holds.

Theorem 6 ($N_g = \infty$, $n < N_d < \infty$)

Assume (AL), (AG), and (AD). Assume σ_d is either sigmoid or the $\tanh$ function. Then the following statement holds with probability 1: for any stationary point θ_s of J, if the range of g_{θ_s} contains an open-ball in $\mathbb{R}^n$, then $J(\theta_s) = 0$.

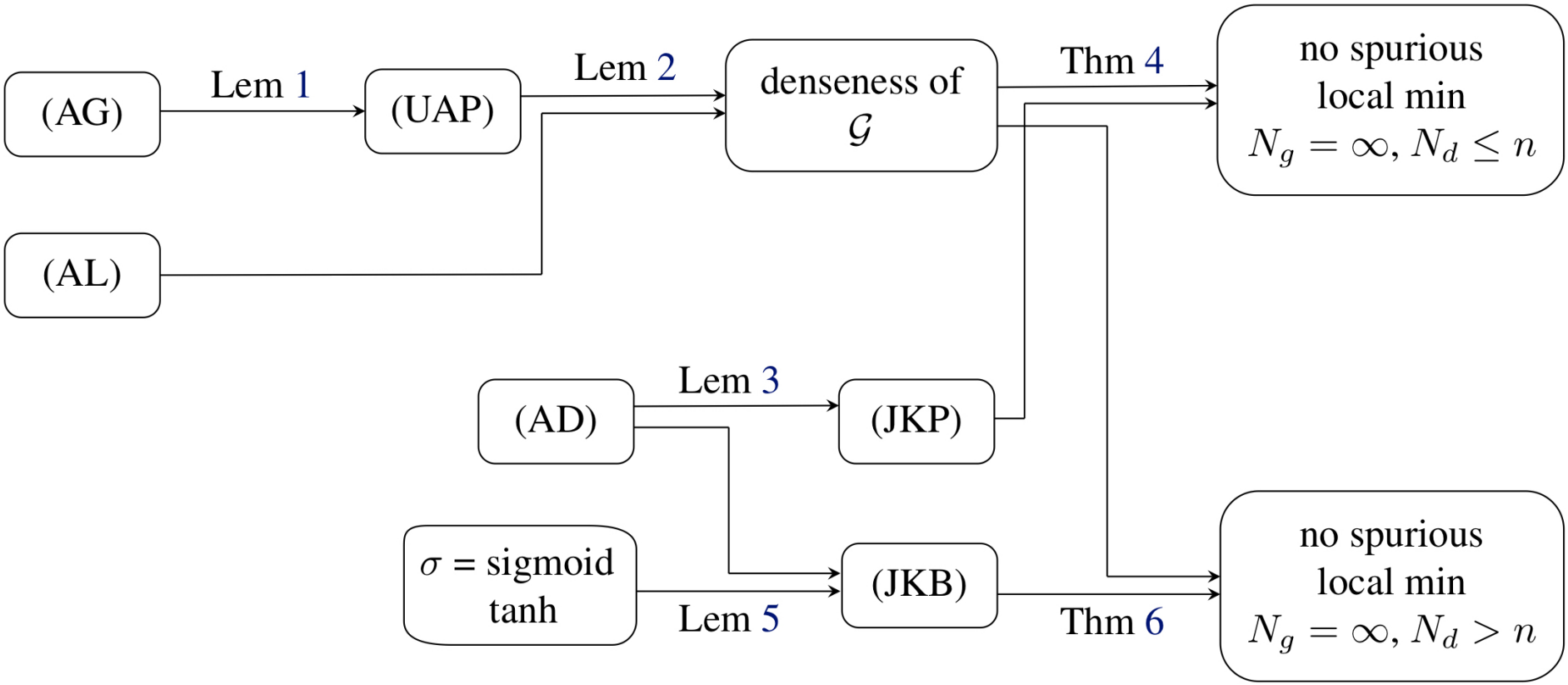
Remark:

- The proof is almost the same as that of Theorem 4.
- "Open-ball" condition is sufficient:
  - If $n < N_d$ and σ_d is either sigmoid or $\tanh$, (AD) implies a weaker condition (Lemma 5): for any open ball $B \subset \mathbb{R}^n$,

$$\bigcap_{x \in B} \ker(D\Psi(x)^\top) = \{0\}.$$

Infinite-width Generator

Diagrammatic summary of proofs



Finite-width Generator & Small Discriminator ($N_d \leq n$)

Theorem 8 ($N_g < \infty$, $N_d \leq n$) [2]

For any $C > 0$ and $\zeta > 0$, there exists a large enough $N_g \in \mathbb{N}$ such that the following statement holds with probability at least $1 - \zeta$: any stationary point $\theta_s \in \mathbb{R}^{N_g}$ satisfying $\|\theta_s\|_1 \leq C$ is a global minimum.

Implication:

- With 'high' probability, $J(\theta)$ has no spurious stationary points within a ℓ_1 ball whose radius C becomes arbitrarily large as $N_g \to \infty$.

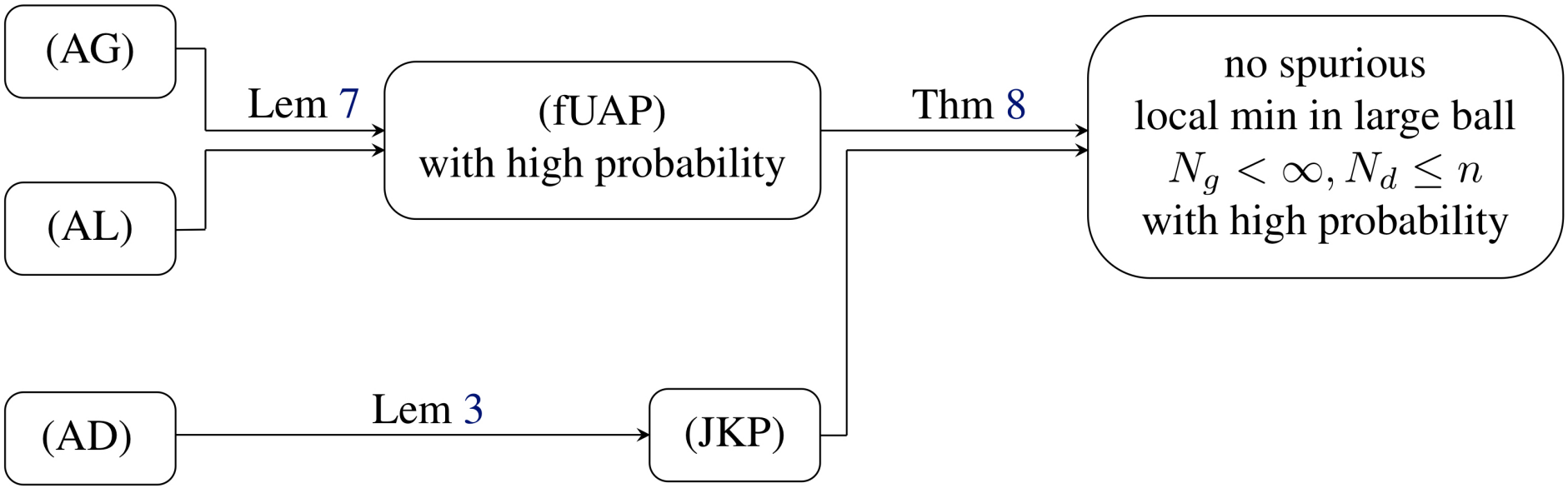
Proof Outline:

1. From (AG) and (AL), establish a finite version of UAP.
   - Approximate Dirac delta function.
2. Perform a finite version of the previous proofs.
   - Every arguments of form " $???= 0$ " is replaced to " $|???| < \epsilon$ ".

[2] Detailed assumption on parameters is omitted. See [No et al., 2021], Lemma 7.

Finite-width Generator & Small Discriminator ($N_d \leq n$)

Diagrammatic summary of proofs



Finite-width Generator

A finite version of UAP: approximation of Dirac delta functions

Let δ be Dirac delta function and $\delta^{(l)} : \mathbb{R}^k \to \mathbb{R}^n$ be

$$\left[\delta^{(l)}(z)\right]_i \begin{cases} 0 & \text{if } i \neq l \\ \delta(z) & \text{if } i = l \end{cases}$$

Finite Universal Approximation Property (fUAP)

For a given $\epsilon > 0$, there exists a large enough $N_g \in \mathbb{N}$ and $\phi_1, \ldots, \phi_{N_g} \in \mathcal{G}$ such that there exists $\{\theta_i^{(\epsilon,l)} \in \mathbb{R} | 1 \leq i \leq N_g, 1 \leq l \leq n\}$ satisfying

$$\left| \mathbb{E}_Z \left[\left(\sum_{i=1}^{N_g} \theta_i^{(\epsilon,l)} \phi_i(Z) - \delta^{(l)}(Z) \right)^\top f(Z) \right] \right| < \epsilon \sup_{z \in \mathbb{R}^k} \{ \|f(z)\|_2 + \|Df(z)\| \}$$

for all coordinates $l = 1, \cdots, n$ and for any continuously differentiable $f : \mathbb{R}^k \to \mathbb{R}$ such that $\sup_{z \in \mathbb{R}^k} \{ \|f(z)\|_2 + \|Df(z)\| \} < 0$.

Finite-width Generator

Lemma 7. *Assume (AL) and (AG). Assume the first n parameters $\{\kappa_i\}_{i=1}^{n}$ are chosen so that $\{\phi_i\}_{i=1}^{n}$ are constant functions spanning the sample space $\mathbb{R}^n$. Assume the remaining parameters $\{\kappa_i\}_{i=n+1}^{N_g}$ are sampled (IID) from a probability distribution that has a continuous and strictly positive density function. Then for any $\varepsilon > 0$ and $\zeta > 0$, there exists large enough [6] N_g such that (Finite universal approximation property) with ε holds with probability at least $1 - \zeta$.*

"We CAN approximate delta functions with (sufficiently but finitely many) randomly initialized feature functions (for generator)."

Finite-width Generator

Proof Outline of Lemma 7 ($n = 1$ case only)

"We can approximate delta functions with (sufficiently but finitely many) randomly initialized feature functions."

$$\begin{aligned}
\delta(z) &\approx \tilde{\delta}^{\epsilon}(z) = (C/\epsilon^k)\exp(-\|z/\epsilon\|^2) \\
&\approx \theta_1^{\epsilon}\phi_1(z;\kappa_1) + \int_{\mathbb{R}^{k+1}} \phi(z;\kappa)m(\kappa)d\kappa \quad (3) \\
&\approx \theta_1^{\epsilon}\phi_1(z;\kappa_1) + \int_{\|\kappa\|\leq K} \phi(z;\kappa)m(\kappa)d\kappa \quad (\text{for large } K < 0) \\
&\approx \theta_1^{\epsilon}\phi_1(z;\kappa_1) + \sum_{i=2}^{N_g} \theta_i^{\epsilon}\phi(z;\kappa_i) \quad (\text{with probability } 1-\zeta) \quad (4)
\end{aligned}$$

(3) [Barron, 1993], [Telgarsky, 2021]: Infinite-width Neural Nets

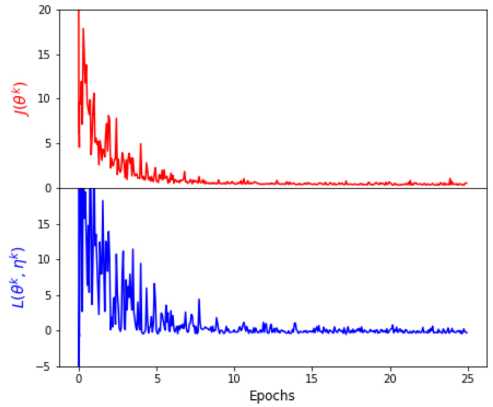
(4) [Rahimi and Recht, 2008]: Random Feature Learning

Experiment & Limitation

- Gaussian Mixture with $N_g = 5000$.

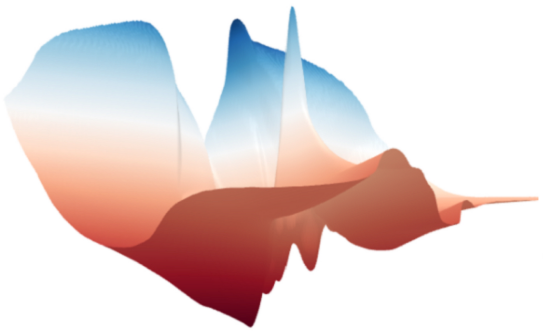
(a) Samples from true distribution P_X

(b) Samples from generator $g_\theta(Z)$

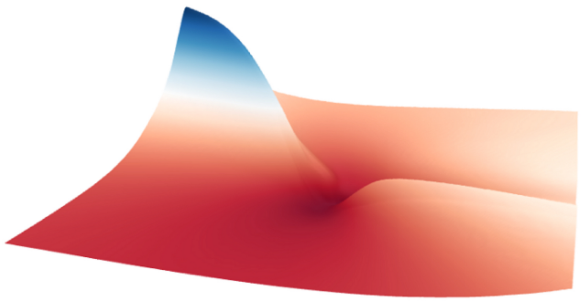


(c) Convergence of the loss functions J and L

- Loss landscape visualization, with $N_g = 2$ v.s. $N_g = 10$.



(a) Loss landscape with $N_g = 2$



(b) Loss landscape with $N_g = 10$

- Nearly Vanishing Gradient Problem can occur.
  - despite the absence of spurious stationary points.

References I

Arjovsky, M., Chintala, S., and Bottou, L. (2017).

Wasserstein gan.

Barron, A. (1993).

Universal approximation bounds for superpositions of a sigmoidal function.
IEEE Transactions on Information Theory, 39(3):930–945.

Hornik, K. (1991).

Approximation capabilities of multilayer feedforward networks.
Neural Networks, 4(2):251–257.

No, A., Yoon, T., Sehyun, K., and Ryu, E. K. (2021).

Wgan with an infinitely wide generator has no spurious stationary points.
In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8205–8215. PMLR.

Rahimi, A. and Recht, B. (2008).

Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning.
In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Telgarsky, M. (2021).

Deep learning theory lecture notes.
`https://mjt.cs.illinois.edu/dlt/`.
Version: 2021-10-27 v0.0-e7150f2d (alpha).

Villani, C. (2008).

Optimal Transport: Old and New.
Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.