Local Stability of Wasserstein GANs With Abstract Gradient Penalty

Cheolhyeong Kim, Seungtae Park, and Hyung Ju Hwang

***Abstract*— The convergence of generative adversarial networks (GANs) has been studied substantially in various aspects to achieve successful generative tasks. Ever since it is first proposed, the idea has achieved many theoretical improvements by injecting an instance noise, choosing different divergences, penalizing the discriminator, and so on. In essence, these efforts are to approximate a real-world measure with an idle measure through a learning procedure. In this article, we provide an analysis of GANs in the most general setting to reveal what, in essence, should be satisfied to achieve successful convergence. This work is not trivial since handling a converging sequence of an abstract measure requires a lot more sophisticated concepts. In doing so, we find an interesting fact that the discriminator can be penalized in a more general setting than what has been implemented. Furthermore, our experiment results substantiate our theoretical argument on various generative tasks.**

***Index Terms*— Abstract measure, gradient penalty, local stability, measure-valued differentiation (MVD), Wasserstein generative adversarial network (WGAN).**

I. Introduction

Generative adversarial networks (GANs) have achieved remarkable improvements in both practical and theoretical fields ever since it is first proposed. It has been able to sample from not only real-like images [1] but also from meaningful joint distributions, such as text-to-image generation, image-to-text generation, and low-quality-to-high-quality image generation [2]–[5].

However, although GANs can generate real-like data, it is not sufficient to argue that GANs can generate any samples we can expect from a real-world distribution. Therefore, many theoretical studies have attempted to fix such anomalies by injecting an instance noise [6] and selecting different divergences [7], [8]. In addition, an equivalence between the two aforementioned approaches is revealed [9], [10].

The Wasserstein GAN (WGAN) is well-known to resolve the problems of generic GANs by selecting the Wasserstein distance as the divergence [7]. However, WGAN often fails with simple examples because the Lipschitz constraint on discriminator is rarely achieved during the optimization process and weight clipping. Thus, mimicking the Lipschitz constraint on the discriminator by using a gradient penalty was proposed by Gulrajani *et al.* [11]. Also, a noise injection and regularizing with a gradient penalty appear to be equivalent. The addition of instance noise in f-GAN can be approximated to adding a zero centered gradient penalty [10]. Thus, regularizing GAN with a simple gradient penalty term was suggested by Mescheder *et al.* [9] who provided proof of its stability.

Based on a theoretical analysis of the dynamic system, Nagarajan and Kolter [12] first proved local exponential stability of the gradient-based optimization dynamics in GANs by treating the simultaneous gradient descent algorithm with a dynamic system approach. These previous studies were useful because they showed that the local behavior of GANs can be explained using dynamic system tools and the related Jacobian's eigenvalues.

From the gradient penalty terms [9], [11] and the scope of dynamic system viewpoint [12], various methods of regularizing WGAN have been proposed. These studies lead to a simple but essential question: What sort of abstract properties of penalizing methods should be required to ensure the local stability of dynamics of WGAN with a simple gradient penalty term? This is certainly not a trivial question since analyzing a converging sequence of an abstract measure requires a lot more sophisticated notions and methods. In this article, we provide an analysis of GANs in the most general setting to disclose what, in essence, should be satisfied to achieve successful convergence. Our contributions are the following.

1) We propose an abstract property of the gradient penalty measure to ensure a convergence of the model near an equilibrium. We generalize the common property of gradient penalty measures as an abstract form and give this as an additional assumption. We provide rigorous proof for the local stability of the dynamic system with general penalty measures under suitable assumptions.
2) We exploit the measure-valued differentiation (MVD), which makes it possible to deal with abstract terms, which cannot be written in an integral form with

Manuscript received 16 October 2019; revised 21 May 2020 and 16 September 2020; accepted 2 February 2021. Date of publication 19 February 2021; date of current version 2 September 2022. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant NRF-2017R1E1A1A03070105 and Grant NRF-2019R1A5A1028324 and in part by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korean Government (MSIP) through the Artificial Intelligence Graduate School Program, Pohang University of Science and Technology (POSTECH), under Grant 2019-0-01906. *(Corresponding author: Hyung Ju Hwang.)*

Cheolhyeong Kim is with the Department of Mathematics, Pohang University of Science and Technology (POSTECH), Pohang 37673, South Korea (e-mail: tyty4@postech.ac.kr).

Hyung Ju Hwang is with the Department of Mathematics and Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), Pohang 37673, South Korea (e-mail: hjhwang@postech.ac.kr).

Seungtae Park is with the Department of Mechanical Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea (e-mail: swash21@postech.ac.kr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3057885.

Digital Object Identifier 10.1109/TNNLS.2021.3057885

probability density functions. The concept of MVD gives a theoretical and technical foundation for dealing with an integral over abstract measure in the stability analysis. This also makes it possible to deal with an abstract measure's derivative with respect to finite-dimensional parameters while proving the local stability of the system.

3) We explain a reason for the success of previous penalty measures based on the proof of the local stability. We claim that the support of a penalty measure will be strongly related.
4) We experimentally examine general convergence results by applying three test penalty measures to several examples. The proposed test measures are unintuitive, but two of them still satisfy the assumptions, which also achieve the successful convergence results.

II. Preliminaries

We interpret the updating procedure of GANs as a continuous dynamic system. The continuous dynamic system approach, which is so-called the ODE method, analyzes the GAN optimization problem with a simultaneous gradient descent algorithm, as described by Nagarajan and Kolter [12].

Furthermore, the analysis of GANs requires a concept of a converging sequence of a probability measure. Rigorously speaking, this requires a firm definition of a converging sequence of an abstract measure. Second, we need a concept of a derivative of an expectation with respect to a related probability measure. This concept is required since we will investigate a smooth behavior of an expectation of a penalty term in a continuous dynamic system.

In Section II-A, we will introduce the aforementioned measure-theoretic concepts. In Section II-B, we provide our formulation of GANs with a gradient penalty as a continuous dynamic system.

A. Notations and Preliminaries Regarding Measure Theory

$D(x;\psi) : \mathcal{X} \to \mathbb{R}$ is a discriminator function with its parameter ψ, and $G(z;\theta) : \mathcal{Z} \to \mathcal{X}$ is a generator function with its parameter θ. p_d is the distribution of real data, and $p_g = p_\theta$ is the distribution of the generated samples in $\mathcal{X}$, which is induced from the generator function $G(z;\theta)$ and a known initial distribution $p_{\text{latent}}(z)$ in the latent space $\mathcal{Z}$. $\|\cdot\|$ denotes the L^2 Euclidean norm if no special subscript is present.

In this section, we define: 1) measures that we are interested in; 2) convergence of such measures; and 3) a derivative of an expectation with respect to the measure. Throughout this study, we assume that the measures over the sample space are all finite and bounded.

Definition 1: For a set of finite measures $\{\mu_i\}_{i\in\mathcal{I}}$ in $(\mathbb{R}^n, d)$ with the Euclidean distance d, $\{\mu_i\}_{i\in\mathcal{I}}$ is referred to as bounded if there exists some $M > 0$ such that for all $i \in \mathcal{I}$

$$\mu_i\left(\mathbb{R}^n\right) \leq M. \tag{1}$$

Now, we introduce the convergence of measures that satisfy Definition 1. Roughly speaking, we say a sequence of measures in Definition 1 weakly converges when its expectation over every continuous bounded function converges accordingly.

Definition 2 (Weak Convergence of a Finite Measure): For a bounded sequence of finite measures $\{\mu_n\}_{n\in\mathbb{N}}$ on the Euclidean space $\mathbb{R}^n$ with a σ-field of Borel subsets $\mathcal{B}(\mathbb{R}^n)$, μ_n converges weakly to μ if and only if, for every continuous bounded function ϕ on $\mathbb{R}^n$, its integrals with respect to μ_n converge to $\int \phi d\mu$, that is

$$\mu_n \to \mu \Longleftrightarrow \int \phi d\mu_n \to \int \phi d\mu. \tag{2}$$

Taking the derivative of an expectation with respect to its measure is challenging. In the most general setting, measures are not necessarily absolutely continuous. That is, we cannot always differentiate an expectation with respect to its parametric probability measure in a closed form as usual. We claim that such generalization is not only theoretical but also realistic since it is widely observed that real-world data are distributed over lower dimensional supports. Hence, we introduce the weak derivatives of a probability measure [13] as the following.

Definition 3 (Weak Derivatives of a Probability Measure): Consider the Euclidean space and its σ-field of Borel subsets $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Let P_θ be a probability measure on $\mathbb{R}^d$, which depends on its 1-D parameter θ. The probability measure P_θ is called weakly differentiable at θ if a signed finite measure P'_θ exists where

$$\begin{aligned}\frac{d}{d\theta}\int \phi(x) dP_\theta &= \lim_{\Delta\to 0}\frac{1}{\Delta}\left\{\int \phi(x) dP_{\theta+\Delta} - \int \phi(x) dP_\theta\right\} \\ &= \int \phi(x) dP'_\theta\end{aligned} \tag{3}$$

is satisfied for every continuous bounded function ϕ on $\mathbb{R}^n$. For the multidimensional parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, this can be defined similarly as $(\partial/(\partial\theta_1)P_\theta, \partial/(\partial\theta_2)P_\theta, \ldots, \partial/(\partial\theta_n)P_\theta)$

It is also possible to extend the concept of weak derivatives of a probability measure to a general finite measure Q_θ. Note that the product rule for differentiating can also be applied in a similar manner to calculus

$$\frac{d}{d\theta}\int \phi(x;\theta) dP_\theta = \int \nabla_\theta \phi(x;\theta) dP_\theta + \int \phi(x;\theta) dP'_\theta. \tag{4}$$

Therefore, for the general finite measure $Q_\theta = M(\theta)P_\theta$, its derivative Q'_θ can be represented as below by introducing a normalizing coefficient $M(\theta) = \int 1 dQ_\theta < \infty$

$$Q'_\theta = M'(\theta)P_\theta + M(\theta)P'_\theta. \tag{5}$$

B. Problem Setting as a Dynamic System

The optimization process of the discriminator and the generator can be expressed as a system of ODEs [12]. Basically, the generator (or discriminator) minimizes (or maximizes) the Wasserstein distance through the Kantorovich–Rubinstein duality. Differentiation of each objective by respective parameters results in a system of ODEs.

In addition, the discriminator needs to be regularized to satisfy the Lipschitz condition of the Kantorovich–Rubinstein duality. The proper choice of the regularization term has been discussed. As discussed in [9], regularizing the discriminator through the Euclidean norm $\|\cdot\|$ is problematic since its derivative $\frac{\cdot}{\|\cdot\|}$ is not defined at the origin. Likewise, the penalty term $\mathbb{E}_{\mu_{\text{GP}}}[(\|\nabla_x D\|-1)^2]$ of WGAN-GP can trigger a discontinuity in its dynamic system. Therefore, we choose a squared, simple gradient penalty (SGP) term $\mathbb{E}_{\mu}[\|\nabla_x D\|^2]$ as our regularization term. As discussed in [9], this regularization term makes the resulting dynamic system differentiable. Note that this can be viewed as a soft regularization based on the size of the discriminator's gradient [10].

As a result, let a quadruple $(D(x;\psi), p_d, p_\theta, \mu)$ denote our resulting formulation defined as follows.

Definition 4: The SGP μ-WGAN optimization problem with a simple gradient penalty term $\|\nabla_x D\|^2$, penalty measure $\mu = \mu_{\psi,\theta}$ which that on the discriminator's parameter ψ and the generator's parameter θ, and penalty weight hyperparameter $\rho > 0$ is given as follows, where the penalty term is only introduced to update the discriminator:

$$\begin{aligned}&\max_{\psi} : \mathbb{E}_{p_d}[D(x;\psi)] - \mathbb{E}_{p_\theta}[D(x;\psi)] - \frac{\rho}{2}\mathbb{E}_{\mu}\left[\|\nabla_x D(x;\psi)\|^2\right]\\ &\min_{\theta} : \mathbb{E}_{p_d}[D(x;\psi)] - \mathbb{E}_{p_\theta}[D(x;\psi)].\end{aligned} \tag{6}$$

According to [12] and many other optimization problem studies, the simultaneous gradient descent algorithm for GAN updating can be viewed as an autonomous dynamic system of discriminator parameters and generator parameters, which we denote as ψ and θ. As a result, the corresponding dynamic system is given as follows:

$$\begin{aligned}\dot{\psi} &= \mathbb{E}_{p_d}\left[\nabla_\psi D\right] - \mathbb{E}_{p_\theta}\left[\nabla_\psi D\right] - \frac{\rho}{2}\nabla_\psi \mathbb{E}_{\mu}\left[\nabla_x^T D \nabla_x D\right]\\ \dot{\theta} &= \nabla_\theta \mathbb{E}_{p_\theta}[D].\end{aligned} \tag{7}$$

Note that the penalty measure μ determines the information provided to the discriminator during the optimization process. This raises an interesting question: what sort of information should be provided to the discriminator to ensure convergence to the equilibrium point? As we will discuss in Section IV-A, our work is, therefore, the equilibrium point is still achievable with a general condition (see Assumption 5). This also provides a theoretical ground for existing penalty measures.

III. Toy Examples

We investigate two examples considered in previous studies by Mescheder *et al.* [9] and Nagarajan and Kolter [12]. We then generalize the results to a finite measure case. The first example is the univariate Dirac GAN, which was introduced by Mescheder *et al.* [9].

Definition 5 (Dirac GAN): The Dirac GAN comprises a linear discriminator $D(x;\psi) = \psi x$, data distribution $p_d = \delta_0$, and sample distribution $p_\theta = \delta_\theta$.

The Dirac GAN with a gradient penalty with an arbitrary probability measure is known to be globally convergent [9]. We argue that this result can be generalized to a finite penalty measure case.

Lemma 1: Consider the Dirac GAN problem with the quadruple form $(D(x;\psi) = \psi x, \delta_0, \delta_\theta, \mu_{\psi,\theta})$. Suppose that some small $\eta > 0$ exists such that its finite penalty measure $\mu_{\psi,\theta}$ with mass $M(\psi,\theta) = \int 1 d\mu_{\psi,\theta} \geq 0$ satisfies either of the following.

1) $M(\psi,\theta) > 0$ for $(\psi,\theta) \in B_\eta((0,0))$.
2) $M(0,0) = 0$ and $\psi\nabla_\psi M(\psi,\theta) \geq 0$ for $(\psi,\theta) \in B_\eta((0,0))$.

Then, the SGP μ-WGAN optimization dynamics with $(D(x;\psi) = \psi x, \delta_0, \delta_\theta, \mu_{\psi,\theta})$ are locally stable at the origin, and the basin of attraction $B = B_R((0,0))$ is an open ball with radius R. Its radius is given as follows:

$$\begin{aligned}R = \max\big\{\eta \geq 0 \,|\, &2M(\psi,\theta) + \psi\nabla_\psi M(\psi,\theta) \geq 0\\ &\text{for all } (\psi,\theta) \text{ such that } \psi^2+\theta^2 \leq \eta^2\big\}.\end{aligned} \tag{8}$$

Proof: The related dynamic system of the SGP μ-WGAN optimization problem for Dirac GAN can be written as follows:

$$\begin{aligned}\dot{\psi} &= -\theta - \frac{\rho}{2}\nabla_\psi \mathbb{E}_{\mu_{\psi,\theta}}\left[\psi^2\right]\\ \dot{\theta} &= \psi.s.\end{aligned} \tag{9}$$

First, the only equilibrium point is given by $(\psi^*,\theta^*) = (0,0)$, from

$$\begin{aligned}0 &= -\theta - 2\psi M(\psi,\theta) - \psi^2\nabla_\psi M(\psi,\theta)\\ 0 &= \psi.\end{aligned} \tag{10}$$

The corresponding Jacobian matrix of the dynamic system at the equilibrium point (0, 0) is written as follows:

$$J = \begin{bmatrix} Z & -1 \\ 1 & 0 \end{bmatrix} \tag{11}$$

where

$$Z = -\frac{\rho}{2}\nabla_{\psi\psi}\mathbb{E}_{\mu_{\psi,\theta}}\left[\psi^2\right]\Big|_{\psi=0,\theta=0}. \tag{12}$$

Since $\nabla_\psi D(x;\psi) = \psi$ does not depend on x, this can be rewritten as follows:

$$\begin{aligned}Z &= -\frac{\rho}{2}\nabla_{\psi\psi}\left(\psi^2\mathbb{E}_{\mu_{\psi,\theta}}[1]\right)\Big|_{\psi=0,\theta=0}\\ &= -\frac{\rho}{2}\nabla_{\psi\psi}\left(\psi^2 M(\psi,\theta)\right)\Big|_{\psi=0,\theta=0}\\ &= -\frac{\rho}{2}\left(2M(\psi,\theta)+4\psi M_\psi(\psi,\theta) + \psi^2 M_{\psi\psi}(\psi,\theta)\right)\Big|_{\psi=0,\theta=0}\\ &= -\rho M(0,0).\end{aligned} \tag{13}$$

Therefore, if $M(0,0) > 0$, then the given system is locally stable since its linearlized system's eigenvalues have negative real parts. If $M(0,0) = 0$, then the stability of the system cannot be proved by the linearlization theorem. In this case, consider the Lyapunov function

$$L(\psi(t),\theta(t)) = \psi(t)^2 + \theta(t)^2. \tag{14}$$

Differentiating with t, we get

$$\begin{aligned}\dot{L} &= 2\left(\psi\psi' + \theta\theta'\right)\\ &= -\rho\psi\nabla_\psi\left(\psi^2 M(\psi,\theta)\right)\\ &= -\rho\psi\left(2\psi M(\psi,\theta) + \psi^2\nabla_\psi M(\psi,\theta)\right)\\ &= -\rho\psi^2\left(2M(\psi,\theta) + \psi\nabla_\psi M(\psi,\theta)\right) \leq 0.\end{aligned} \tag{15}$$

It is clear that $L(\psi,\theta)\geq 0$ and equality holds iff $\psi=\theta=0$. Also, $\dot{L}\leq 0$ since $M(\psi,\theta)\geq 0$ and $\psi\nabla_\psi M(\psi,\theta)\geq 0$ from the assumption. Furthermore, it is clear that if $(\psi(0),\theta(0))\in B_\eta((0,0))$, then $(\psi(\tau),\theta(\tau))\in B_\eta((0,0))$ for all $\tau\geq 0$ since the Lyapunov function (square of the distance between the origin and $(\psi(\tau),\theta(\tau))$) always decreases as $\tau\to\infty$. Therefore, the given system is stable from the Lyapunov stability theorem.

It can be checked again that if $\mu_{\psi,\theta}$ is a probability measure, then the system is globally stable as pointed by Mescheder *et al.* [9]. Basin of attraction is given by whole $\mathbb{R}^2$ plane since $M(\psi,\theta)=1$, so

$$\dot{L}=-\rho\psi^2\big(2M+\psi\nabla_\psi M\big)=-2\rho\psi^2\leq 0 \tag{16}$$

for every $(\psi,\theta)\in\mathbb{R}^2$. □

Motivated by this example, we can extend this idea to another toy example given by Nagarajan and Kolter [12], where WGAN fails to converge to the equilibrium points $(\psi^*,\theta^*)=(0,\pm 1)$.

Lemma 2: Consider the toy example $(D(x;\psi)=\psi x^2, U(-1,1), U(-|\theta|,|\theta|), \mu_{\psi,\theta})$ where $U(0,0)=\delta_0$ and the ideal equilibrium points are given by $(\psi^*,\theta^*)=(0,\pm 1)$. For a finite measure $\mu_{\psi,\theta}=\mu_\theta$ on $\mathbb{R}$, which does not depend on ψ, suppose that $\mu_\theta\to\mu^*$ as $\theta\to\theta^*$ with $\mu^*\neq C\delta_0$ for $C\geq 0$. The dynamic system is locally stable near the desired equilibrium $(0,\pm 1)$, where the spectrum of the Jacobian at $(0,\pm 1)$ is given by $\lambda=-2\rho\mathbb{E}_{\mu^*}[x^2]\pm(4\rho^2\mathbb{E}_{\mu^*}[x^2]^2-(4/9))^{1/2}$.

Proof: From the general setup of the SGP μ-WGAN optimization problem

$$\begin{aligned}\dot{\psi}&=\mathbb{E}_{p_D}\big[D_\psi\big]-\mathbb{E}_{p_\theta}\big[D_\psi\big]-\frac{\rho}{2}\nabla_\psi\mathbb{E}_{\mu_{\psi,\theta}}\big[D_x^2\big]\\ \dot{\theta}&=\nabla_\theta\mathbb{E}_{p_\theta}[D]\end{aligned} \tag{17}$$

the corresponding dynamic system can be written as follows:

$$\begin{aligned}\dot{\psi}&=\frac{1}{3}-\frac{\theta^2}{3}-4\rho\psi\mathbb{E}_{\mu_{\psi,\theta}}\big[x^2\big]\\ \dot{\theta}&=\frac{2\psi\theta}{3}.\end{aligned} \tag{18}$$

Let $\mathbb{E}_{\mu^*}[x^2]=A^2$, and then, the Jacobian matrix at the equilibrium $(0,\pm 1)$ is given by

$$J=\begin{bmatrix}-4\rho A^2 & \mp\frac{2}{3}\\ \pm\frac{2}{3} & 0\end{bmatrix}. \tag{19}$$

Therefore, the given system is locally stable unless $A=0$. □

IV. Main Convergence Theorem

In this section, we propose assumptions to guarantee the local stability around the equilibrium point of our system of ODEs. We assume the existence of an equilibrium point $\theta=\theta^*$ since a large capacity of the generator will be able to achieve or almost achieve $p_d=p_{\theta^*}$. In Section IV-A, we provide the necessary assumptions for the local stability. In Section IV-B, we propose our main convergence theorem with a sketch of the proof. More detailed proofs are provided in the Appendix.

A. Assumptions

Our main goal of this section is to introduce an ideal behavior of gradient penalty near an ideal equilibrium. Assumptions 1–4 state the conditions of an ideal equilibrium, which were previously studied in [9] and [12], whereas Assumption 5 states the behavior of gradient penalty near the equilibrium, which are first discussed in our work.

The first assumption is made regarding a realizable case of equilibrium conditions for GANs, where we state ideal conditions for the discriminator parameter and generator parameter. As the parameters converge to the ideal equilibrium, the sample distribution (p_θ) converges to the real data distribution (p_d) and the discriminator cannot distinguish the generated sample and the real data.

Assumption 1: $p_\theta\to p_d$ weakly as $\theta\to\theta^*$ and $D(x;\psi^*)=0$ on $\text{supp}(p_d)$ and its small open neighborhood, i.e., there exists some $\epsilon=\epsilon(x')>0$, which depends on the data point so that $x\in\cup_{x'\in\text{supp}(p_d)}B_{\epsilon_{x'}}(x')$ implies $D(x;\psi^*)=0$. For simplicity, we denote $\cup_{x'\in\text{supp}(p_d)}B_{\epsilon_{x'}}(x')$ as $B(\text{supp}(p_d))$.

The second assumption ensures that the higher order terms cannot affect the stability of the SGP μ-WGAN. Compared with the previous study by Nagarajan and Kolter [12], conditions for the discriminator parameter are generalized to deal with the abstract penalty measure.

Assumption 2:

$$\begin{aligned}g(\theta)&=\big\|\mathbb{E}_{p_d}\big[\nabla_\psi D\big(x;\psi^*\big)\big]-\mathbb{E}_{p_\theta}\big[\nabla_\psi D\big(x;\psi^*\big)\big]\big\|^2\\ h(\psi)&=\mathbb{E}_{\mu_{\psi,\theta^*}}\big[\|\nabla_x D(x;\psi)\|^2\big]\end{aligned} \tag{20}$$

are locally constant along the nullspace of the Hessian matrix of $g(\theta)$ at $\theta=\theta^*$ and $h(\psi)$ at $\psi=\psi^*$, respectively. That is, there exists some small $r_g,r_d>0$ so that, for any vector u in the nullspace of the Hessian matrix of g with $\|u\|<r_g$, $g(\theta^*)=g(\theta^*+u)$. Respectively, for any v in the nullspace of the Hessian matrix of h with $\|v\|<r_d$, $h(\psi^*)=h(\psi^*+v)$.

The third assumption allows us to extend our results to discrete probability distribution cases, as described by Mescheder *et al.* [9]. Ideal discriminators are robust and flat under a small enough perturbation on the generator parameter.

Assumption 3: There exists $\epsilon_g>0$ such that $D(x;\psi^*)=0$ on $\cup_{|\theta-\theta^*|<\epsilon_g}\text{supp}(p_\theta)$.

The fourth assumption indicates that there are no other equilibriums that do not satisfy the given assumptions near (ψ^*,θ^*), which justifies the projection along the axis perpendicular to the null space.

Assumption 4: Either (ψ^*,θ^*) is an isolated equilibrium, or there exist $\delta_d,\delta_g>0$ such that all equilibrium points in $B_{\delta_d}(\psi^*)\times B_{\delta_g}(\theta^*)$ satisfy the other assumptions.

The proposed assumption (Assumption 5) is related to sufficient conditions for the penalty measure. A calculation of the gradient penalty based on samples from the data manifold and generator manifold or the interpolation of both was introduced in recent studies [9]–[11]. Therefore, it is plausible to assume that the penalty measure depends on discriminator's parameter ψ and generator's parameter θ.

Assumption 5: The finite penalty measure $\mu=\mu_{\psi,\theta}$ satisfies the following.

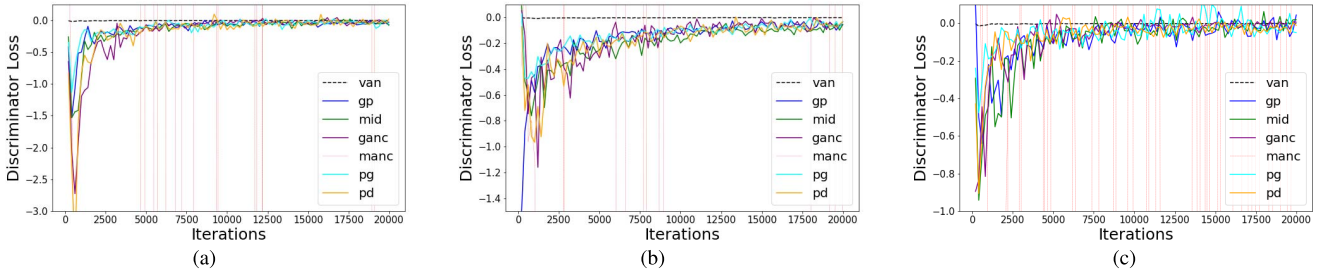


Fig. 1. Discriminator loss plots of 2-D examples. Except for $\mu_{m,\text{anc}}$ with the red dotted line, which fluctuates wildly outside of the given discriminator loss range, the others converge and generate the target distributions. (a) Eight Gaussians. (b) 25 Gaussians. (c) Swissroll.

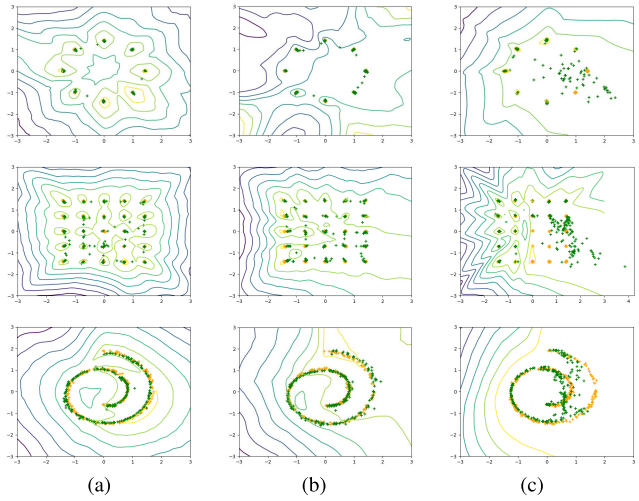


Fig. 2. 2-D examples generated with μ_{mid}, $\mu_{g,\text{anc}}$, and $\mu_{m,\text{anc}}$. μ_{mid} and $\mu_{g,\text{anc}}$ succeeded to generate the target distributions, whereas $\mu_{m,\text{anc}}$ failed to generate the samples near $(2,-1)$. (a) μ_{mid}. (b) $\mu_{g,\text{anc}}$. (c) $\mu_{m,\text{anc}}$.

1) $\mu_{\psi,\theta} \to \mu_{\psi^*,\theta^*} = \mu^*$, where $\text{supp}(\mu_{\psi,\theta})$ only depends on θ. (We will denote $\text{supp}(\mu_{\psi,\theta}) = \text{supp}(\mu_\theta)$ since its support only depends on θ.) Near the equilibrium, $\mu_{\psi,\theta}$ can be weakly differentiated twice with respect to ψ. In addition, its mass $M(\psi,\theta) = \int 1 d\mu_{\psi,\theta}$ is a twice-differentiable function of ψ and bounded near the equilibrium.
2) $\text{supp}(p_d) \subset \text{supp}(\mu^*)$.
3) There exists $\epsilon_\mu > 0$ such that $\text{supp}(\mu_\theta) \subset V$ for $|\theta - \theta^*| < \epsilon_\mu$, where $V = \{x | \nabla_x D(x;\psi^*) = 0\}$.

Assumption 5(a)[1] is technically required to take the derivative of the integral $\mathbb{E}_{\mu_{\psi,\theta}}[\|\nabla_x D(x;\psi)\|^2]$ with respect to ψ.

The Assumption 5 can be described in detail as follows: (5a) the penalty measure's support approaches to a data manifold and its weight changes smoothly with respect to ψ and θ; (5b) at the equilibrium, the penalty measure's support contains the data manifold; and (5c) the ideal discriminator will remain flat near $\text{supp}(\mu^*)$ and its small open neighborhood. This is an extension of Assumption 3 and a quite plausible situation that we can expect from the gradient penalty of the ideal discriminator on $\text{supp}(\mu^*)$ and the flatness of the ideal discriminator on the data manifold.

In summary, the gradient penalty regularization term with any penalty measure where the support approaches $B(\text{supp}(p_d))$ in a smooth manner works well, and this main result can explain the regularization effect of previously proposed penalty measures, such as μ_{GP}, p_d, p_θ, and their mixtures.

B. Main Convergence Theorem

According to the modified assumptions given above, we prove that the related dynamic system is locally stable near the equilibrium. The tools used for analyzing the stability are mainly based on those described by Nagarajan and Kolter [12]. Our main contributions comprise proposing the ideal behavior of the penalty measure and proving the local stability for all penalty measures that satisfy Assumption 5.

Theorem 1: Suppose that our SGP μ-WGAN optimization problem $(D, p_d, p_\theta, \mu_{\psi,\theta})$ with equilibrium point (ψ^*, θ^*) satisfies the Assumptions above. Then, the related dynamic system is locally stable at the equilibrium.

A detailed proof of the main convergence theorem is given in the Appendix. A sketch of the proof is given in three steps. First, it is enough to check that all nonzero eigenvalues of the Jacobian of the dynamic system have negative real parts. For the zero-eigenvalues and corresponding eigenvectors, it is enough to show that the system is still locally stable along these eigenvectors. Therefore, it is enough to observe the Jacobian of the dynamic system at the equilibrium point. Next, after removing some zero terms, the Jacobian matrix at the equilibrium is given by

$$\begin{bmatrix} -\rho Q & -R \\ R^T & 0 \end{bmatrix} \tag{21}$$

where $Q = \mathbb{E}_{\mu^*}[\nabla_{\psi x} D \nabla_{\psi x}^T D]$ and $R = \nabla_\theta \mathbb{E}_{p_\theta}[\nabla_\psi D]|_{\theta=\theta^*}$. The system is locally stable when both Q and $R^T R$ are positive definite. We can complete the proof by dealing with zero eigenvalues by showing that $N(Q^T) \subset N(R^T)$, and the projected system's stability implies the original system's stability. Our analysis mainly focuses on WGAN, which is the simplest case of the following general GAN minimax

[1]This condition is technically required to handle the derivative of the measure in a convenient manner using the general formulation. Even if the measure is not differentiable, it may be possible to differentiate the integral. For instance, δ_ψ is continuous, but it does not have its weak derivative. However, it is still possible to differentiate $\mathbb{E}_{\delta_\psi}[\omega(x)] = \omega(\psi)$ if the function ω is differentiable at ψ.

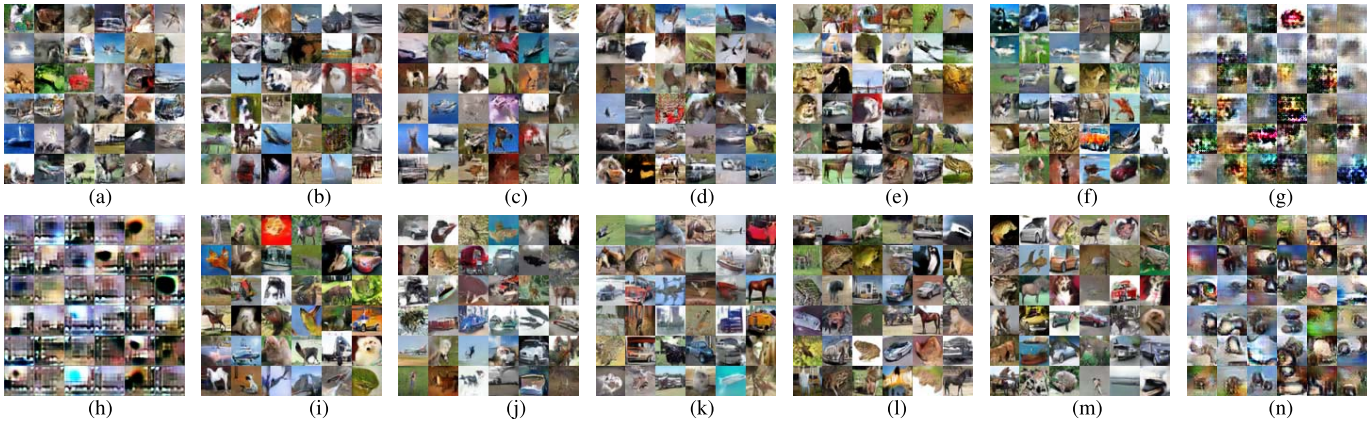


Fig. 3. Generated CIFAR-10 examples with DCGAN (first row) and ResNet (second row) architectures. Note that the penalty measure $\mu_{m,\text{anc}}$ and WGAN with ResNet failed to generate images, which can be found in (g), (h), and (n). (a) WGAN. (b) p_g. (c) p_d. (d) μ_{GP}. (e) μ_{mid}. (f) $\mu_{g,\text{anc}}$. (g) $\mu_{m,\text{anc}}$. (h) WGAN. (i) p_g. (j) p_d. (k) μ_{GP}. (l) μ_{mid}. (m) $\mu_{g,\text{anc}}$. (n) $\mu_{m,\text{anc}}$.

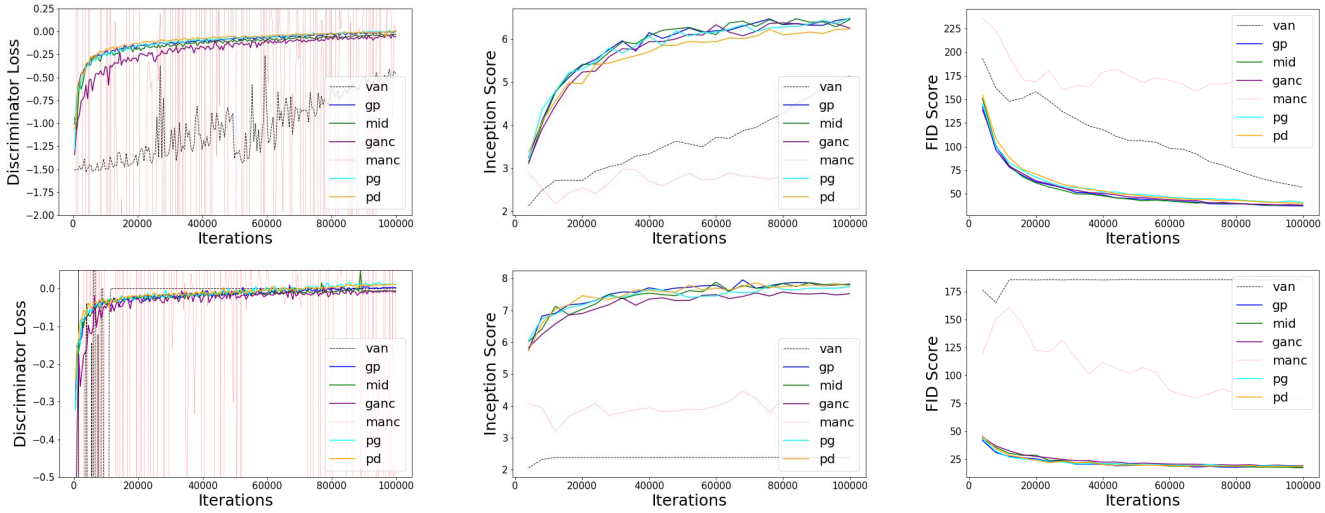


Fig. 4. Plots for the discriminator loss, inception score, and FID score of the generated CIFAR-10 images with DCGAN (first row) and ResNet (second row) architectures. Note that WGAN with ResNet (denoted as van, with black dashed line) failed to generate target images and the discriminator loss plot of the penalty measure $\mu_{m,\text{anc}}$ (with red dotted line) fluctuates wildly outside of the given discriminator loss range, whereas the others perform well.

optimization:

$$\begin{aligned}\max_{\psi} : &\mathbb{E}_{p_d}[f(D(x;\psi))] + \mathbb{E}_{p_\theta}[f(-D(x;\psi))] \\ &- \frac{\rho}{2}\mathbb{E}_{\mu}\left[\|\nabla_x D(x;\psi)\|^2\right] \\ \min_{\theta} : &\mathbb{E}_{p_d}[f(D(x;\psi))] + \mathbb{E}_{p_\theta}[f(-D(x;\psi))] \end{aligned} \tag{22}$$

with $f(x) = x$. A similar approach is still valid for general GANs with a function f with $f''(x) < 0$ and $f'(0) \neq 0$.

V. Experimental Results

We claim that every penalty measure that satisfies the assumptions can regularize the WGAN and generate similar results to recently proposed gradient penalty methods with a simple gradient penalty term. Six penalty measures were tested on 2-D problems [11] (mixture of eight Gaussians, mixture of 25 Gaussians, and swissroll) and image generation tasks (CIFAR-10 and CelebA-HQ data sets with resolution 128 $\times$ 128) using a simple gradient penalty term. The penalty measures and its detailed sampling methods are listed in Table I, where $x_d \sim p_d, x_g \sim p_\theta$, and $\alpha \sim U(0,1)$. $\mathcal{A}$ indicates a fixed anchor point in $\mathcal{X}$. Throughout this section, we will only discuss on WGAN with a simple gradient penalty term since WGAN-GP is already known to perform well on 2-D examples and image generation tasks [11].

SGP μ-WGAN was examined with various penalty measures comprising three recently proposed measures and three artificially generated measures. p_θ and p_d were proposed by Mescheder *et al.* [9], and μ_{GP} was introduced from the WGAN-GP. We proposed and analyzed the artificial penalty measures μ_{mid}, $\mu_{g,\text{anc}}$, and $\mu_{m,\text{anc}}$ as test penalty measures. Note that five penalty measures p_d, p_g, μ_{GP}, μ_{mid}, and $\mu_{g,\text{anc}}$ satisfy the assumptions, whereas $\mu_{m,\text{anc}}$ does not.

The experiments were conducted based on the implementation of [11]. The loss function was modified from a nonzero centered gradient penalty to a simple gradient penalty. Throughout this section, the number of discriminator updates

Fig. 5. Generated CelebA-HQ 128 × 128 examples with the DCGAN architecture. Note that the penalty measures $\mu_{m,\text{anc}}$ and WGAN failed to generate images. (a) WGAN. (b) p_g. (c) p_d. (d) μ_{GP}. (e) μ_{mid}. (f) $\mu_{g,\text{anc}}$. (g) $\mu_{m,\text{anc}}$.

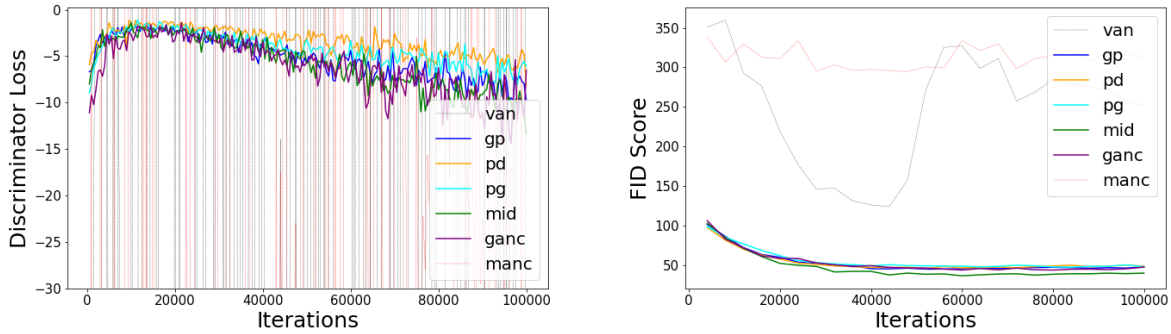


Fig. 6. Plots for discriminator loss and FID score of generated CelebA-HQ 128 × 128 images with DCGAN architecture. Note that WGAN (denoted as van, with black dashed line) and $\mu_{m,\text{anc}}$ (with red dotted line) do not converge, whereas the others perform well. Discriminator loss is reported for every 500 iterations, and FID score is reported for every 4000 iterations.

per generator update was chosen as five [7], and the Adam optimizer [14] with its learning rate 10^{-4} was used as a discriminator/generator's optimizer.

A. 2-D Examples

We checked the convergence of p_θ on the 2-D examples (eight Gaussians, swissroll data, and 25 Gaussians) for the SGP-WGANs with six penalty measures. Each data set was trained over 20 000 iterations. The anchor $\mathcal{A}$ for $\mu_{g,\text{anc}}$ was set as $(2, -1)$ for the 2-D examples. Overall, five penalty measures (μ_{GP}, μ_{mid}, p_d, p_g, and $\mu_{g,\text{anc}}$) succeeded to generate the target distributions, whereas $\mu_{m,\text{anc}}$ failed. Plots of the discriminator loss for 2-D examples were reported for every 200 iterations, which can be found in Fig. 1. We present generated results with μ_{mid}, $\mu_{g,\text{anc}}$, and $\mu_{m,\text{anc}}$ in Fig. 2.

B. CIFAR-10

We trained WGAN and the SGP-WGANs with six penalty measures for the CIFAR-10 data set. DCGAN [1] and ResNet [15] were used to construct discriminators and generators in this section, which are previously constructed in [11]. The models were trained for 100 000 iterations. The anchor $\mathcal{A}$ for $\mu_{g,\text{anc}}$ and $\mu_{m,\text{anc}}$ during CIFAR-10 generation was set as a black monochrome image. The images generated with WGAN and six penalty measures with DCGAN [1] and ResNet [15] architectures are shown in Fig. 3. We can observe that WGAN

TABLE I
LIST OF BENCHMARK WGANS (WGAN AND SIX PENALTY MEASURES WITH A SIMPLE GRADIENT PENALTY TERM)

Penalty	Penalty term	Sampling method
WGAN	None (Weight Clipping)	None
p_g	$\|\nabla_x D\|^2$	$\hat{x} = x_g$
p_d	$\|\nabla_x D\|^2$	$\hat{x} = x_d$
μ_{GP}	$\|\nabla_x D\|^2$	$\hat{x} = \alpha x_d + (1-\alpha)x_g$
μ_{mid}	$\|\nabla_x D\|^2$	$\hat{x} = 0.5x_d + 0.5x_g$
$\mu_{g,anc}$	$\|\nabla_x D\|^2$	$\hat{x} = \alpha\mathcal{A} + (1-\alpha)x_g$
$\mu_{m,anc}$	$\|\nabla_x D\|^2$	$\hat{x} = 0.5\mathcal{A} + 0.5x_g$

TABLE II
BENCHMARK SCORE RESULTS ON THE GENERATED SAMPLES UNDER DCGAN AND RESNET ARCHITECTURES

Penalty	CIFAR-10 DCGAN		CIFAR-10 ResNet	
	Inception	FID	Inception	FID
WGAN	5.15 ± 0.08	56.9	2.38 ± 0.01	185.4
p_g	6.49 ± 0.07	40.7	7.74 ± 0.07	18.9
p_d	6.23 ± 0.05	39.8	7.78 ± 0.09	19.0
μ_{GP}	6.47 ± 0.04	37.3	7.83 ± 0.06	17.6
μ_{mid}	6.45 ± 0.05	37.9	7.80 ± 0.10	17.7
$\mu_{g,anc}$	6.25 ± 0.07	37.8	7.52 ± 0.05	19.3
$\mu_{m,anc}$	2.99 ± 0.03	159.6	4.04 ± 0.05	78.1

failed to generate images under the ResNet architecture and $\mu_{m,\text{anc}}$ failed to converge.

Results from WGAN and six penalty measures were evaluated based on the inception score [16] and the FID score [17], as shown in Table II, which are useful tools for scoring the quality of generated images. For the CIFAR-10 image generation task, the inception score [16], [18] and FID score [17] were used as benchmark scores to evaluate the generated images. The higher inception score and lower FID score indicate the good quality of the generated images. We generated 50 000 samples in total. The number of samples for evaluating an inception score is 100. Compared with WGAN, generated images and benchmark scores of five penalty measures with a simple gradient penalty show similar regularization performances from the results in Table II, whereas $\mu_{m,\text{anc}}$ failed to generate the original images. Plots for the discriminator loss, inception score, and FID score can be found in Fig. 4. Discriminator loss is reported for every 500 iterations. Inception score and FID score are reported for every 4000 iterations.

C. CelebA-HQ 128

We ran the CelebA-HQ image generation task for WGAN and the SGP-WGANs with six penalty measures. A resolution of CelebA-HQ images was resized to 128×128. DCGAN [1] was used to build a generator and a discriminator. Their detailed architectures can be found in the Appendix. The models were trained for 100 000 iterations. The anchor $\mathcal{A}$ for $\mu_{g,\text{anc}}$ and $\mu_{m,\text{anc}}$ was set as a black monochrome image as on the CIFAR-10 tasks. The batch size was set to 64. Verified our main convergence theorem. Observing the results of WGAN and $\mu_{m,\text{anc}}$, their discriminator losses fluctuate, and their FID scores do not decrease. Generated images in Fig. 5 also show that the assumptions discussed in Section IV-A are indeed necessary. For the CelebA-HQ image generation task, the FID score [17] was used as a benchmark score to evaluate the trained WGANs. As on the CIFAR-10 tasks, we generated 50 000 samples in total. Table III shows FID scores after 100 000 iterations. Plots of the discriminator loss and FID score in Fig. 6 empirically.

TABLE III
FID SCORE OF 50 000 SAMPLES GENERATED FROM WGAN AND SIX PENALTY MEASURES

Penalty	WGAN	p_g	p_d	μ_{GP}	μ_{mid}	$\mu_{g,anc}$	$\mu_{m,anc}$
FID	252.3	48.2	49.0	48.1	40.2	47.7	311.6

VI. CONCLUSION

In this study, we proposed an additional assumption on an abstract property of the gradient penalty measure to ensure the local stability, and then, we proved the local stability of a simple gradient penalty μ-WGAN optimization problem with the MVD tool. This proof provides insights into the good behavior of gradient penalty and the success of regularization with previously proposed penalty measures. Furthermore, our theoretical approach was supported by relevant experiments with the previously proposed penalty measure and our unintuitive penalty measures. In future research, our works can be extended to an alternative gradient descent algorithm and its related optimal hyperparameters. Stability at nonrealizable equilibrium points is one of the important topics on the stability of GANs. Optimal penalty measures for achieving the best convergence speed can be also investigated using a spectral theory, which provides a mathematical analysis on the stability of GAN with precise information on the convergence.

APPENDIX A
PROOF OF THE MAIN CONVERGENCE THEOREM

Proof: Let us consider the Jacobian matrix

$$J = \begin{bmatrix} K_{\text{DD}} & K_{\text{DG}} \\ K_{\text{GD}} & K_{\text{GG}} \end{bmatrix} \tag{23}$$

at the equilibrium (ψ^*, θ^*),[2] where the each block matrix can be represented as follows:

$$\begin{aligned} K_{\text{DD}} &= \mathbb{E}_{p_d}\left[\nabla_{\psi\psi} D\right] - \mathbb{E}_{p_\theta}\left[\nabla_{\psi\psi} D\right] - \frac{\rho}{2}\nabla_{\psi\psi}\mathbb{E}_{\mu_{\psi,\theta}}\left[\|\nabla_x D\|^2\right] \\ K_{\text{DG}} &= -\nabla_{\theta\psi}\mathbb{E}_{p_\theta}[D] - \frac{\rho}{2}\nabla_{\theta\psi}\mathbb{E}_{\mu_{\psi,\theta}}\left[\|\nabla_x D\|^2\right] \\ K_{\text{GD}} &= \nabla_{\psi\theta}\mathbb{E}_{p_\theta}[D] \\ K_{\text{GG}} &= \nabla_{\theta\theta}\mathbb{E}_{p_\theta}[D]. \end{aligned} \tag{24}$$

[2]In standard notation, $\nabla_\psi g$ is the dim(range of g) $\times$ dim(ψ) matrix. For a real-valued function f, we consider the first derivative as the column vector instead of the row vector. $\nabla_\psi f$ is considered to be the dim(ψ) $\times$ 1 matrix (column vector) of the total derivative. For the second derivative, $\nabla_{\psi\theta} f = (\nabla_\psi)(\nabla_\theta f)$ is the dim(θ) $\times$ dim(ψ) matrix. The transpose notation is used in a similar manner to the matrix.

First, Assumption 1 implies

$$\mathbb{E}_{p_d}\left[\nabla_{\psi\psi} D\right] - \mathbb{E}_{p_{\theta^*}}\left[\nabla_{\psi\psi} D\right] = 0 \tag{25}$$

since $p_\theta \to p_d$ as $\theta \to \theta^*$. From Assumption 3, $D(x;\psi^*)$ is zero on the $\text{supp}(p_\theta)$ with $|\theta - \theta^*| < \epsilon_g$, which implies that

$$K_{GG} = \nabla_{\theta\theta}\mathbb{E}_{p_\theta}\left[D(x;\psi^*)\right]\Big|_{\theta=\theta^*} = 0. \tag{26}$$

We still need to evaluate $\nabla_{\psi\psi}\mathbb{E}_{\mu_{\psi,\theta}}[\|\nabla_x D\|^2]$ and $\nabla_{\theta\psi}\mathbb{E}_{\mu_{\psi,\theta}}[\|\nabla_x D\|^2]$ at the equilibrium. According to Assumption 5a, finite signed measures $\mu'_{\psi,\theta}$ and $\mu''_{\psi,\theta}$ exist,[3] so they are the first and second weak derivatives of $\mu_{\psi,\theta}$ with respect to the parameter ψ at (ψ^*,θ^*). Therefore, the expectations given above can be rewritten as follows:

$$\begin{aligned} I &= \nabla_{\psi\psi}\int_{\text{supp}(\mu_{\psi,\theta})} \|\nabla_x D\|^2 d\mu_{\psi,\theta}\Big|_{\psi=\psi^*,\theta=\theta^*} \\ &= \int_{\text{supp}(\mu_{\psi,\theta})} \left(2\nabla^T_{\psi x} D \nabla_{\psi x} D + 2K_0\right) d\mu_{\psi,\theta} \\ &\quad + \int_{\text{supp}(\mu_{\psi,\theta})} 2\left(\nabla^T_{\psi x} D \nabla_x D\right) d\mu'_{\psi,\theta} \\ &\quad + \int_{\text{supp}(\mu_{\psi,\theta})} \|\nabla_x D\|^2 d\mu''_{\psi,\theta}\Big|_{\psi=\psi^*,\theta=\theta^*} \end{aligned} \tag{27}$$

$$\begin{aligned} II &= \nabla_{\theta\psi}\int_{\text{supp}(\mu_{\psi,\theta})} \|\nabla_x D\|^2 d\mu_{\psi,\theta}\Big|_{\psi=\psi^*,\theta=\theta^*} \\ &= \nabla_\theta\Bigg(\int_{\text{supp}(\mu_{\psi,\theta})} 2\left(\nabla^T_{\psi x} D \nabla_x D\right) d\mu_{\psi,\theta} \\ &\qquad + \int_{\text{supp}(\mu_{\psi,\theta})} \|\nabla_x D\|^2 d\mu'_{\psi,\theta}\Bigg)\Big|_{\psi=\psi^*,\theta=\theta^*} \end{aligned} \tag{28}$$

where

$$K_0(x;\psi) = \left[\sum_k \frac{\partial^3}{\partial\psi_i \partial\psi_j \partial x_k} D(x;\psi) \frac{\partial}{\partial x_k} D(x;\psi)\right]_{ij}. \tag{29}$$

From Assumption 5c, the fact that the weak derivative of $\mu_{\psi,\theta}$ vanishes outside of $\text{supp}(\mu_{\psi,\theta})$, $\nabla_x D(x;\psi^*) = 0$ on V that includes $\text{supp}(\mu_{\psi,\theta})$ for all θ with $|\theta - \theta^*| < \epsilon_\mu$, and $\mu'_{\psi,\theta} = \mu''_{\psi,\theta} = 0$ on the outside of $\text{supp}(\mu_{\psi,\theta})$, which leads to the desired results

$$\begin{aligned} I &= \int_{\text{supp}(\mu^*)} 2\left(\nabla^T_{\psi x} D(x;\psi^*) \nabla_{\psi x} D(x;\psi^*)\right) d\mu^* \\ II &= 0. \end{aligned} \tag{30}$$

After canceling the undesired terms, the Jacobian matrix at the equilibrium (ψ^*,θ^*) is given as

$$J = \begin{bmatrix} -\rho Q & -R \\ R^T & 0 \end{bmatrix} \tag{31}$$

where

$$\begin{aligned} Q &= \mathbb{E}_{\mu^*}\left[\nabla^T_{\psi x} D \nabla_{\psi x} D\right] \\ R &= \nabla_\theta \mathbb{E}_{p_\theta}\left[\nabla_\psi D\right]\Big|_{\theta=\theta^*}. \end{aligned} \tag{32}$$

[3] $\mu'_{\psi,\theta}$ and $\mu''_{\psi,\theta}$ will be considered as row vector($1 \times \dim(\psi)$ matrix) and $\dim(\psi) \times \dim(\psi)$ matrix of finite signed measures, respectively. $\mu'_{\psi,\theta} = \left[\frac{\partial}{\partial\psi_1}\mu_{\psi,\theta} \quad \cdots \quad \frac{\partial}{\partial\psi_{\dim(\psi)}}\mu_{\psi,\theta}\right]$ and $\mu''_{\psi,\theta} = \left[\frac{\partial^2}{\partial\psi_i\partial\psi_j}\mu_{\psi,\theta}\right]_{ij}$.

From the definition of Q, it is easy to check that Q is at least positive semidefinite. It is known that, for a negative definite matrix A and full column rank matrix B, the block matrix

$$\begin{bmatrix} A & B \\ -B^T & 0 \end{bmatrix}$$

is Hurwitz, i.e., all eigenvalues of the matrix have a negative real part. Therefore, if Q is positive definite and R is full column rank, the proof is complete. We consider the complementary case.

Suppose that Q or $R^T R$ has some zero eigenvalues. Let $Q = U_D \Lambda_D U_D^T$ and $R^T R = U_G \Lambda_G U_G^T$ with $U_D = \begin{bmatrix} T_D & S_D \end{bmatrix}$ and $U_G = \begin{bmatrix} T_G & S_G \end{bmatrix}$, where T_D and T_G are the eigenvectors of Q and $R^T R$ that correspond to nonzero eigenvalues. First, we assume that T_D and T_G are not empty. We can show that $(\psi^* + \xi v, \theta^* + \nu w)$ is also an equilibrium point for a sufficiently small ξ, ν and $v \in N(Q)$, $w \in N(R^T R)$ by using the techniques given by [12]. If the system does not update at the equilibrium point (ψ^*,θ^*) and its small neighborhood $(\psi^* + \xi v, \theta^* + \nu w)$ is perturbed along $N(Q)$ and $N(R^T R)$, then it is reasonable to project the system orthogonal to $N(Q)$ and $N(R^T R)$.

First, we assume that $v \in N(Q)$ for a unit vector v. By Assumption 2, $h(\psi^* + \xi v) = h(\psi^*) = 0$ for $|\xi| < \xi_d$, which implies that $\nabla_x D(x;\psi^* + \xi v) = 0$ for $x \in \text{supp}(\mu_{\psi^*+\xi v,\theta^*}) = \text{supp}(\mu^*)$ and $|\xi| < \xi_d$. Thus, we obtain

$$\mathbb{E}_{\mu_{\psi^*+\xi v,\theta^*}}\left[\nabla^T_{\psi x} D(x;\psi^* + \xi v) \nabla_x D(x;\psi^* + \xi v)\right] = 0 \tag{33}$$

and

$$\int_{\text{supp}(\mu^*)} \left\|\nabla_x D(x;\psi^* + \xi v)\right\|^2 d\mu'_{\psi^*+\xi v,\theta^*} = 0. \tag{34}$$

From Assumption 4

$$\mathbb{E}_{p_d}\left[\nabla_\psi D(x;\psi^* + \xi v)\right] - \mathbb{E}_{p_{\theta^*}}\left[\nabla_\psi D(x;\psi^* + \xi v)\right] = 0. \tag{35}$$

By adding (33), (34), and (35), we obtain $\dot{\psi} = 0$. In addition

$$\begin{aligned} \dot{\theta} &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} D(x;\psi^* + \xi v) dp_\theta\Big|_{\theta=\theta^*} \\ &= \int_{\mathcal{Z}} \nabla^T_\theta G(z;\theta^*) \nabla_x D(G(z;\theta^*);\psi^* + \xi v)\, p_{\text{latent}}(z) dz \\ &= 0. \end{aligned} \tag{36}$$

Therefore, the point $(\psi^* + \xi v, \theta^*)$ with $|\xi| < \xi_d$ is an equilibrium point. According to Assumption 4, $D(x;\psi^* + \xi v)$ is an equilibrium discriminator for $|\xi| < \delta_d$, and thus, $D(x;\psi^* + \xi v)$ is already an optimal discriminator for $|\xi| < \min(\xi_d, \delta_d)$.

Suppose that $w \in N(R^T R)$ for a unit vector w. By Assumption 2, $g(\theta^*) = g(\theta^* + \nu w) = 0$ for $|\nu| < \nu_g$, and thus

$$\mathbb{E}_{p_d}\left[\nabla_\psi D(x;\psi^*)\right] - \mathbb{E}_{p_{\theta^*+\nu w}}\left[\nabla_\psi D(x;\psi^*)\right] = 0 \text{ for } |\nu| < \nu_g. \tag{37}$$

Furthermore, Assumption 3 gives $\mathbb{E}_{p_\theta}[D(x;\psi^*)] = 0$ for a sufficiently close θ with $|\theta - \theta^*| < \epsilon_g$, which implies that

$$\dot{\theta} = \nabla_\theta \mathbb{E}_{p_\theta}\left[D(x;\psi^*)\right]\Big|_{\theta=\theta^*+\nu w} = 0 \tag{38}$$

TABLE IV
DETAILED NETWORK ARCHITECTURES OF THE GENERATOR AND THE DISCRIMINATOR ON CELEBA-HQ GENERATION TASK. LRELU DENOTES LEAKYRELU WITH $\alpha = 0.2$

Layer	output size	filter	activations
Fully Connected	$1024 \cdot 4 \cdot 4$	$256 \to 1024 \cdot 4 \cdot 4$	-
ConvTranspose2D	$512 \times 8 \times 8$	$1024 \to 512$	ReLU
ConvTranspose2D	$256 \times 16 \times 16$	$512 \to 256$	ReLU
ConvTranspose2D	$128 \times 32 \times 32$	$256 \to 128$	ReLU
ConvTranspose2D	$64 \times 64 \times 64$	$128 \to 64$	ReLU
ConvTranspose2D	$3 \times 128 \times 128$	$64 \to 3$	Tanh
Conv2D	$64 \times 64 \times 64$	$3 \to 64$	LReLU
Conv2D	$128 \times 32 \times 32$	$64 \to 128$	LReLU
Conv2D	$256 \times 16 \times 16$	$128 \to 256$	LReLU
Conv2D	$512 \times 8 \times 8$	$256 \to 512$	LReLU
Conv2D	$1024 \times 4 \times 4$	$512 \to 1024$	LReLU
Conv2D	$1 \times 1 \times 1$	$1024 \to 1$	-

for $|\nu| < \epsilon_g$. Finally

$$\int_{\text{supp}\left(\mu_{\psi^*,\theta^*+\nu w}\right)} 2\nabla_{\psi x}^T D\left(x;\psi^*\right)\nabla_x D\left(x;\psi^*\right) d\mu_{\psi^*,\theta^*+\nu w} + \int_{\text{supp}\left(\mu_{\psi^*,\theta^*+\nu w}\right)} \left\|\nabla_x D\left(x;\psi^*\right)\right\|^2 d\mu'_{\psi^*,\theta^*+\nu w} = 0 \quad (39)$$

since $\text{supp}(\mu_{\psi^*,\theta^*+\nu w}) \subset V$ and $\nabla_x D(x;\psi^*) = 0$ on V for a sufficiently small $|\nu| < \epsilon_\mu$ (Assumption 5c). By adding (37) and (39), we obtain

$$\begin{aligned}\dot{\psi} &= \mathbb{E}_{p_d}\left[\nabla_\psi D\left(x;\psi^*\right)\right] - \mathbb{E}_{p_{\theta^*+\nu w}}\left[\nabla_\psi D\left(x;\psi^*\right)\right] \\ &\quad - \frac{\rho}{2}\int_{\text{supp}\left(\mu_{\psi^*,\theta^*+\nu w}\right)} 2\nabla_{\psi x}^T D\left(x;\psi^*\right)\nabla_x D\left(x;\psi^*\right) d\mu_{\psi^*,\theta^*+\nu w} \\ &\quad - \frac{\rho}{2}\int_{\text{supp}\left(\mu_{\psi^*,\theta^*+\nu w}\right)} \left\|\nabla_x D\left(x;\psi^*\right)\right\|^2 d\mu'_{\psi^*,\theta^*+\nu w} \\ &= 0. \end{aligned} \quad (40)$$

Therefore, the point $(\psi^*, \theta^* + \nu w)$ with $|\nu| < \min(\epsilon_\mu, \epsilon_g, \nu_g, \delta_g)$ is an equilibrium point, which implies that $p_{\theta^*+\nu w} = p_d$ according to Assumption 4.

If we consider the projected system $(\alpha, \beta) = (T_D^T\psi, T_G^T\theta)$, then the projected dynamic system's Jacobian at $(T_D^T\psi^*, T_G^T\theta^*)$ is given as follows:

$$\begin{aligned} J' &= \begin{bmatrix} -\rho T_D^T Q T_D & -T_D^T R T_G \\ T_G^T R^T T_D & 0 \end{bmatrix} \\ &= \begin{bmatrix} -\rho \Lambda_D^{(+)} & -T_D^T R T_G \\ T_G^T R^T T_D & 0 \end{bmatrix}. \end{aligned} \quad (41)$$

Therefore, we only need to prove that $T_D^T R T_G$ is of full column rank. Suppose that $u \in N(Q^T) = N(Q)$ for a unit vector u. According to Assumption 2, $h(\psi)$ is locally constant at ψ^* along the direction u. Therefore, for a sufficiently small scalar ξ with $|\xi| < \xi_u$

$$h\left(\psi^* + \xi u\right) = h\left(\psi^*\right) = 0 \quad (42)$$

where the second equality comes from Assumption 5. This implies that $\nabla_x D(x;\psi^* + \xi u) = 0$ on $x \in \text{supp}(\mu_{\psi^*+\xi u,\theta^*}) = \text{supp}(\mu^*)$ for a small value of $|\xi| < \epsilon_u$. By taking directional derivative with respect to ψ along the direction u, we obtain

$$u^T\nabla_{\psi x}^T D\left(x;\psi^*\right) = 0, x \in \text{supp}\left(\mu_{\psi^*+\xi u,\theta^*}\right) = \text{supp}\left(\mu^*\right) \quad (43)$$

and thus

$$u^T\nabla_{\psi x}^T D\left(x;\psi^*\right) = u^T\nabla_{x\psi} D\left(x;\psi^*\right) = 0 \quad (44)$$

for all $x \in \text{supp}(p_{\theta^*}) = \text{supp}(p_d)$ by Assumption 5b. By calculating $u^T R$ directly, we obtain

$$\begin{aligned} u^T R &= u^T \frac{\partial}{\partial\theta}\int_{\mathcal{X}} \nabla_\psi D\left(x;\psi^*\right) dp_\theta\Big|_{\theta=\theta^*} \\ &= u^T \frac{\partial}{\partial\theta}\int_{\mathcal{X}} \nabla_\psi D\left(G(z;\theta);\psi^*\right) p_{\text{latent}}(z) dz\Big|_{\theta=\theta^*} \\ &= \int_{\mathcal{X}} u^T\nabla_{x\psi} D\left(G\left(z;\theta^*\right);\psi^*\right)\nabla_\theta G\left(z;\theta^*\right) p_{\text{latent}}(z) dz \\ &= 0. \end{aligned} \quad (45)$$

Thus, we obtain $u \in N(R^T)$, which implies that $N(Q^T) \subset N(R^T)$ and $C(R) \subset C(Q)$. Now, we can check that RT_G is of full column rank since $T_G^T R^T R T_G = \Lambda_G^{(+)}$ is positive definite. Therefore

$$RT_G w = 0 \Rightarrow w = 0. \quad (46)$$

We note that the projection matrix on $C(Q)$ is given by $T_D(T_D^T T_D)^{-1}T_D^T = T_D T_D^T$. In addition, we know that $C(RT_G) \subset C(R) \subset C(Q)$. Therefore

$$\begin{aligned} T_D^T R T_G w &= 0 \\ &\Rightarrow T_D T_D^T R T_G w = 0 \\ &\Rightarrow \text{Projection of } w' \\ &= RT_G w \in C(RT_G) \text{ onto } C(Q) \text{ is zero} \\ &\Rightarrow w' = RT_G w = 0 \\ &\Rightarrow w = 0 \end{aligned} \quad (47)$$

which completes the proof that $T_D^T R T_G$ is a full column rank matrix.

Now, we only need to obtain proofs for the trivial cases where either one of T_D or T_G is empty. First, suppose that T_G is empty. Similar to the analysis given above, we can find that the point (ψ^*, θ) with $|\theta - \theta^*| < \min(\epsilon_\mu, \epsilon_g, \delta_g, \nu)$ is an equilibrium point, where $g(\theta^*) = g(\theta)$ for a sufficiently small $|\theta - \theta^*| < \nu$. We conclude that $p_\theta = p_d$ for $|\theta - \theta^*| < \min(\epsilon_\mu, \epsilon_g, \delta_g, \nu)$. Under the generator initialization that is sufficiently close according to θ^*, we can only observe the discriminator update

$$\dot{\psi} = -\frac{\rho}{2}\nabla_\psi \mathbb{E}_{\mu_{\psi,\theta}}\left[\|\nabla_x D(x;\psi)\|^2\right] \quad (48)$$

since $\mathbb{E}_{p_d}[D(x;\psi)] - \mathbb{E}_{p_\theta}[D(x;\psi)] = 0$ for any ψ and $|\theta - \theta^*| < \min(\epsilon_\mu, \epsilon_g, \delta_g, \nu)$. The discriminator update described above is a stable system near the equilibrium $\psi = \psi^*$ since the Jacobian of the update on ψ is given as $-\rho Q$ and the zero eigenvalues can be ignored in a similar manner to the previous step. Therefore, the given system is stable near the equilibrium.

Suppose that T_D is empty. Given that $N(Q^T) \subset N(R^T)$, $R = 0$, and then, the results are similar to those presented above, but our goal is to show that (ψ, θ) is an equilibrium point, where (ψ, θ) is sufficiently close to the original equilibrium point. We note that (ψ^*, θ) is also an equilibrium point that satisfies the assumptions.

By Assumption 2, $h(\psi) = h(\psi^*) = 0$ for $|\psi - \psi^*| < \xi$, which implies that $\nabla_x D(x; \psi) = 0$ for $x \in \text{supp}(\mu_{\psi,\theta^*}) = \text{supp}(\mu^*)$ and $|\psi - \psi^*| < \xi$. Thus, we obtain

$$\mathbb{E}_{\mu_{\psi,\theta^*}}\left[\nabla_{\psi x}^T D(x; \psi)\nabla_x D(x; \psi)\right] = 0 \tag{49}$$

$$\frac{\rho}{2}\int_{\text{supp}(\mu^*)} \|\nabla_x D\|^2 d\mu'_{\psi,\theta^*} dx = 0. \tag{50}$$

By Assumption 4, $\mathbb{E}_{p_d}[\nabla_\psi D(x; \psi)] - \mathbb{E}_{p_{\theta^*}}[\nabla_\psi D(x; \psi)] = 0$ since $p_d = p_{\theta^*}$. In addition

$$\begin{aligned}\dot{\theta} &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} D(x; \psi) dp_\theta\Big|_{\theta=\theta^*} \\ &= \int_{\mathcal{Z}} \nabla_\theta^T G\big(z; \theta^*\big)\nabla_x D\big(G\big(z; \theta^*\big); \psi\big) p_{\text{latent}}(z) dz \\ &= 0.\end{aligned} \tag{51}$$

Therefore, the point (ψ, θ^*) with $|\psi - \psi^*| < \min(\xi, \delta_d)$ is an equilibrium point. From Assumption 4, $D(x; \psi)$ is an equilibrium discriminator, and thus, $D(x; \psi)$ is already an optimal discriminator for $|\psi - \psi^*| < \min(\xi, \delta_d)$ and p_θ coincides with the data distribution p_d for $|\theta - \theta^*| < \min(\epsilon_\mu, \epsilon_g, \delta_g)$, which indicates that every discriminator and generator near (ψ^*, θ^*) is an equilibrium point, and this completes the proof of the main theorem. □

APPENDIX B
MODEL ARCHITECTURE FOR CELEBA-HQ GENERATION

See Table IV.

REFERENCES

[1] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[2] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976.

[4] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5908–5916.

[5] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, New York City, NY, USA, Jun. 2016, pp. 1060–1069. [Online]. Available: http://jmlr.org/proceedings/papers/v48/reed16.html

[6] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszar, "Amortised map inference for image super-resolution," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=S1RP6GLle

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[8] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[9] L. M. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3478–3487.

[10] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2015–2025.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.

[12] V. Nagarajan and J. Z. Kolter, "Gradient descent GAN optimization is locally stable," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5591–5600.

[13] B. Heidergott and F. J. Vázquez-Abad, "Measure-valued differentiation for Markov chains," *J. Optim. Theory Appl.*, vol. 136, no. 2, pp. 187–209, Feb. 2008.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

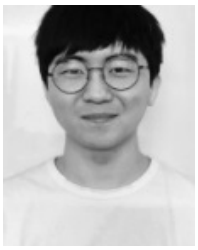
[16] S. Barratt and R. Sharma, "A note on the inception score," 2018, *arXiv:1801.01973*. [Online]. Available: http://arxiv.org/abs/1801.01973

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[18] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2226–2234.

Cheolhyeong Kim received the bachelor's degree in mathematics from the Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea, in 2016, where he is currently pursuing the Ph.D. degree focusing on mathematical analysis on deep learning algorithms.

His current research interests include neural networks, deep learning, active inference, and relevant mathematical analysis.



Seungtae Park received the bachelor's degree in mathematics from the Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, in 2017. He is currently pursuing the Ph.D. degree with the Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea.

His current research interests include representation learning, bioinformatics, and deep generative models.

Hyung Ju Hwang received the Ph.D. degree in mathematics from Brown University, Providence, RI, USA, in 2002.

She was a Post-Doctoral Researcher with the Max-Planck Institute in Leipzig, Leipzig, Germany, from 2002 to 2003, and a Research Assistant Professor with Duke University, Durham, NC, USA, from 2003 to 2005. She is currently the Chair Professor of the Department of Mathematics, Pohang University of Science and Technology (POSTECH), Pohang, South Korea. She has published more than 66 scientific articles in the fields of applied mathematics, artificial intelligence, machine learning, and interdisciplinary research. Her research interests include optimization, deep learning, applied mathematics, partial differential equations, and data analysis in applied fields.

Dr. Hwang is currently working as the Vice-President of the Korean Society for Industrial and Applied Mathematics (KSIAM) and the Director of the POSTECH Center for Applications of Mathematics (PCAM), Pohang.