

Stabilizing generative adversarial networks with adaptive noise injection

Yamin Zhang ¹, Fabing Duan ^{1,*}, François Chapeau-Blondeau ² and Derek Abbott ³

¹*Institute of Complexity Science, Qingdao University, Qingdao 266071, People's Republic of China*

²*Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS),
Université d'Angers, 62 Avenue Notre Dame du Lac, 49000 Angers, France*

³*Centre for Biomedical Engineering (CBME) and School of Electrical & Electronic Engineering,
The University of Adelaide, Adelaide, SA 5005, Australia*

 (Received 11 September 2025; revised 30 May 2026; accepted 27 August 2026; published 16 September 2026)

In conventional generative adversarial networks (GANs), the discriminator typically employs a sigmoid activation to map features to probabilities. However, this activation suffers from the vanishing-gradient problem, which can lead to training instability and mode collapse. This work reformulates the final-layer activation of the discriminator as cumulative distribution functions (CDFs) of random variables, thereby introducing a learnable noise-scale parameter. We show analytically that, for CDF families with bounded reversed hazard rates in the negative tail, a noise-scale value smaller than unity scales the gradient of both the saturating and nonsaturating generator losses by a factor proportional to its reciprocal. Rather than fully resolving the vanishing-gradient problem, this mechanism provides more informative gradient signals to the generator when the discriminator is close to its optimum. Experiments on a two-dimensional Gaussian mixture and on MNIST show that the proposed CDF-based activation effectively improves mode coverage and training stability. Meanwhile, experiments on CIFAR-10 using a deep convolutional GAN (DCGAN) indicate that, when gradient flow is already stabilized by architectural features, the choice of final-layer activation plays a more limited role. An ablation study on MNIST further distinguishes the benefits of the proposed method from that of conventional noise regularization techniques. These results demonstrate that the proposed CDF-based activation also contributes a meaningful scheme for stabilizing adversarial training.

DOI: [10.1103/ch8n-wrv6](https://doi.org/10.1103/ch8n-wrv6)

I. INTRODUCTION

Injecting noise into neural networks to avoid overfitting [1,2] and improve generalization [3,4] is a well-established method in the field of machine learning. This strategy highlights the beneficial role of noise, as well as the phenomenon of stochastic resonance [5,6] originally studied in statistical physics for explaining the periodic transitions between glacial and interglacial periods in the Earth's climate [5]. The underlying mechanism behind noise injection in neural networks [1–4,7] can be explained as a Tikhonov regularizer [1] on the loss function. This explanation is derived under the assumption that the noise variance is small relative to the scale of the input data, so that higher-order terms beyond the second-order Taylor expansion of the loss can be neglected [1]. However, as observed in previous studies of noise-boosted neural networks [4,8], the converged noise variance in optimized activation functions is frequently much larger than unity, indicating that the Tikhonov regularization interpretation does not always hold. This motivates a more general analytical framework for understanding the role of noise in neural network activations.

We here argue that the constructive role of noise is more naturally interpreted through the framework of stochastic resonance, in which an appropriate level of additive noise

amplifies the response of a nonlinear system to weak inputs and thereby yields cooperative interplay between signal and fluctuations [5]. Importantly, unlike the Tikhonov regularization framework, the stochastic resonance mechanism does not require the noise intensity to be infinitesimal, but instead implies an optimal, potentially large noise level at which the system response is maximally enhanced [5,6]. This provides a more general perspective for understanding the constructive role of noise in neural networks over a wide range of noise intensities.

Recently, Kosko *et al.* treated the backpropagation as a special case of the generalized Expectation-Maximization algorithm [9], and injected optimal noise into a neural network, particularly into hidden and visible neurons, to accelerate its convergence. An alternative perspective [10] interpreted the suprathreshold stochastic resonance model [6], which comprises an ensemble of parallel McCulloch-Pitts neurons [11], as a noise-enhanced activation function. Then the threshold neural networks based on noise enhanced activations manifest stochastic resonance effects through the interdependent relationship among hidden neurons, noise intensity, and the target function [4,10]. For an infinite-sized array of the rectified linear units (ReLUs), the average response as activation function and its smooth derivative functions are both explicitly dependent on injected noise [12]. These studies [4,6,7,9,10,12] associating noise injection with stochastic resonance represent two distinct modes: one explicitly injects noise directly into neural networks [6,9], while the latter [4,10,12] implicitly transforms

*Contact author: fabingduan@qdu.edu.cn

noise intensity as an embedded parameter within the activation function. We further generalized the noise-enhanced activation function [6,10] into two categories [8]: double-directed saturation activations, which comprise the sigmoid activation, and single-saturating linear units, which include ReLUs [13] and Gaussian error linear units (GELUs) [14]. Furthermore, we demonstrated that the noise-scale parameter in these noise enhanced activations can be adaptively optimized through stochastic gradient descent [4,8,15] or can be efficiently searched using Bayesian optimization algorithm [15]. These investigations [4,8–10,12,15] established that the underlying mechanism of noise injection in neural networks is fundamentally consistent with the principles of stochastic resonance, whereby an optimal amount of noise elicits a more coherent network response to input data.

While noise injection is commonly understood as a regularizer [1] or an accelerator [9] for neural networks, it has also been demonstrated to play constructive roles in generative adversarial networks (GANs) [16]. Although GANs [16] have become a prominent generative modeling technique in computer vision and machine learning owing to their capacity to learn arbitrary data distributions without requiring explicit density estimation through adversarial training, their practical applicability is frequently constrained by training instabilities and mode collapse [16–21]. Training instabilities denote the oscillatory and nonconvergent dynamics commonly encountered in adversarial optimization [17,18]. Mode collapse refers to the phenomenon in which the generator learns to produce samples from only a limited subset of the modes present in the target data distribution, thereby failing to capture its full diversity [16,19]. These problems are generally attributed to multiple interacting factors, including the adversarial training dynamics, the choice of objective function, and insufficient regularization [17,22]. Among these factors, the vanishing-gradient effect induced by the final-layer activation of the discriminator is of particular relevance, as it reduces the informative gradient signals available to the generator and limits its ability to explore multiple modes.

To address this, injecting noise into both generated and real data acts as a regularizer that extends the support of their distributions and, when fed into the discriminator, leads to a smoother decision boundary and a more stable training process [16–18,20,21]. Furthermore, Style-GAN [23] incorporates single-channel Gaussian noise independently across multiple scales (resolutions) throughout its generator architecture, thereby yielding high-fidelity synthetic data with controllable stochastic variations [23–25]. Beyond noise injection, several gradient-based and Lipschitz-constraint methods have been developed to stabilize GAN training, including gradient penalties on the discriminator output [26–28] and spectral normalization of discriminator weights [29].

However, these investigations into noise injection within GANs [16–18,20,21,23–25] mainly focus on explicit noise sample injection. Conversely, research on implicit noise injection, particularly those on generalized noise-enhanced activation models [4,8,15], the tempered softmax activation [30,31] or sigmoid activation [31–34], has received little attention. Moreover, a unified formulation that generalizes the final activation as a noise-boosted transformation and investigates its role in gradient propagation and compensation

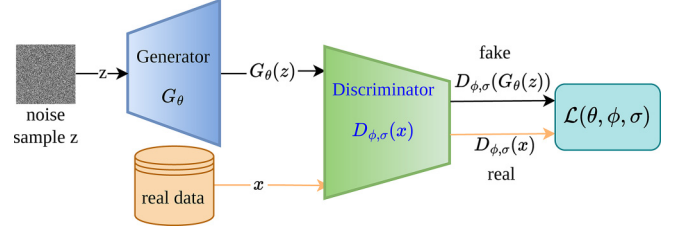


FIG. 1. Diagram of the GAN model. The generator G_θ , parameterized by θ , maps a latent noise vector $\mathbf{z}$ to a synthetic output $G_\theta(\mathbf{z})$. The discriminator $D_{\phi,\sigma}(x)$ in Eq. (2) takes as input both real data $x \sim P_r$ and generated data $G_\theta(\mathbf{z})$, and outputs the probability that the input is real. These outputs are then used to compute the objective function $\mathcal{L}(\theta, \phi, \sigma)$ defined in Eq. (8). Note that the shape of $\mathbf{z}$ shown here corresponds to the image-generation setting, but in general, $\mathbf{z} \in \mathbb{R}^{d_z}$ may have an arbitrary dimensionality.

mechanisms has yet to be sufficiently explored. In this context it is noted that temperature scaling [30–34] is a widely used technique in which the logits of a neural network are divided by a deterministic parameter $T > 0$ before being passed through a sigmoid or softmax activation, thereby controlling the sharpness of the output distribution. This operation is purely deterministic and does not involve explicit stochastic noise injection. However, within the theoretical framework of our generalized noise-enhanced activation models [4,8,15], the sigmoid function $1/(1 + e^{-x/T})$ with temperature T is mathematically identical to the logistic CDF with scale parameter $\sigma = T$. Based on this correspondence, we will hereafter refer to this parameter T as the noise scale σ throughout this work, while emphasizing that the resulting activation remains a deterministic function of the input. Building upon this perspective, we aim to investigate implicit noise injection in GANs by formulating the final-layer activations of the discriminator as CDF-based transformations and analyzing their theoretical and practical effects on adversarial training stability.

II. GAN MODEL

We consider a general formulation of the GAN, as depicted in Fig. 1 (see, e.g., Ref. [35] for a measure-theoretic description of the GAN framework). Let $\mathcal{Z} = \mathbb{R}^{d_z}$ denote the latent space and $\mathcal{X} = \mathbb{R}^{d_x}$ the data space, and two dimensions d_z and d_x are generally distinct. The generator is defined as a mapping $G_\theta : \mathcal{Z} \rightarrow \mathcal{X}$, parameterized by θ , which transforms a latent noise vector $\mathbf{z} \sim P_z$ into a synthetic sample $G_\theta(\mathbf{z}) \sim P_g$. In practice, P_z is typically chosen as a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$ or a uniform distribution over $[-1, 1]^{d_z}$. The discriminator is defined as a mapping $D : \mathcal{X} \rightarrow [0, 1]$, which takes as input either a real sample $\mathbf{x} \sim P_r$ or a generated sample $G_\theta(\mathbf{z})$ and outputs the probability that the input originates from the real data distribution P_r . Here P_z , P_r , and P_g represent the distributions of latent noise vectors $\mathbf{z}$, real data $\mathbf{x}$, and generated data $G_\theta(\mathbf{z})$, respectively.

In general, a feedforward neural network with L layers can be expressed iteratively as

$$\mathbf{x}^{(l)} = \psi_l(h_l(\mathbf{x}^{(l-1)})), \quad l = 1, \dots, L, \quad (1)$$

where $h_l(\mathbf{x}) = W_l \mathbf{x} + \mathbf{b}_l$ denotes the affine transformation at the l th layer with weight matrix W_l and bias vector $\mathbf{b}_l$, ψ_l is the activation function applied elementwise, and $\mathbf{x}^{(0)} = \mathbf{x} \in \mathbb{R}^{d_x}$ is the input. Here we decompose the discriminator into two components: a feature-mapping network $V_\phi : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$, parameterized by $\phi = \{W_l, \mathbf{b}_l\}_{l=1}^L$ and defined as $V_\phi(\mathbf{x}) = h_L(\mathbf{x}^{(L-1)})$, and a scalar activation F_η that replaces ψ_L in Eq. (1), resulting in the discriminator output

$$D(\mathbf{x}) = F_\eta[V_\phi(\mathbf{x})]. \quad (2)$$

Specifically, we take F_η to be the cumulative distribution function (CDF) of a zero-mean random variable η , which generalizes the conventional sigmoid activation and introduces a tunable noise scale through the distribution of η , such as the logistic CDF

$$F_\eta(x) = [1 + e^{-\frac{x}{\sigma}}]^{-1}, \quad (3)$$

the Gaussian CDF

$$F_\eta(x) = \frac{1}{2} [1 + \operatorname{erf}(\frac{x}{\sqrt{2}\sigma})], \quad (4)$$

or the Laplace CDF

$$F_\eta(x) = \begin{cases} \frac{1}{2} e^{\frac{x}{\sigma}}, & x < 0, \\ 1 - \frac{1}{2} e^{-\frac{x}{\sigma}}, & x \geq 0, \end{cases} \quad (5)$$

mapping $V_\phi(x)$ into a probability. Here all CDFs are defined with zero mean and $\sigma > 0$ acts as the scale parameter. For a centered unit-scale random variable η_0 with CDF F_{η_0} , the scale-family random variable $\eta = \sigma \eta_0$ has CDF $F_\eta(x) = F_{\eta_0}(x/\sigma)$. Accordingly, the discriminator in Eq. (2) can be rewritten as

$$D_{\phi,\sigma}(x) = F_{\eta_0}\left(\frac{V_\phi(x)}{\sigma}\right), \quad (6)$$

where σ is a learnable noise-scale parameter that is adaptively updated during training alongside the network parameters ϕ and θ .

Then the GAN training procedure is formulated as a min-max optimization over the noise-scale parameter $\sigma \in \mathbb{R}^+$, the hypothesis spaces $\mathcal{H}_G = \{G_\theta : \mathcal{Z} \rightarrow \mathcal{X} \mid \theta \in \mathbb{R}^{N_G}\}$ for the generator, and $\mathcal{H}_D = \{D_{\phi,\sigma} : \mathcal{X} \rightarrow (0, 1) \mid \phi \in \mathbb{R}^{N_D}, \sigma \in \mathbb{R}^+\}$ for the discriminator, where N_G and N_D denote the number of parameters of the generator and discriminator networks, respectively. Equivalently, the minimax optimization problem can be expressed as

$$\min_{\theta \in \mathbb{R}^{N_G}} \max_{\phi \in \mathbb{R}^{N_D}, \sigma \in \mathbb{R}^+} \mathcal{L}(\theta, \phi, \sigma), \quad (7)$$

where the objective function is defined as

$$\mathcal{L}(\theta, \phi, \sigma) = \mathbb{E}_{P_r}[\ln D_{\phi,\sigma}(\mathbf{x})] + \mathbb{E}_{P_z}[\ln[1 - D_{\phi,\sigma}(G_\theta(\mathbf{z}))]]. \quad (8)$$

The explicit update rule for the learnable noise-scale parameter σ is presented in Algorithm 1 of Sec. III D.

Since CDFs F_η are monotonically increasing functions that range from 0 to 1, probability outputs $D_{\phi,\sigma}(x)$ and $D_{\phi,\sigma}[G_\theta(\mathbf{z})]$ are in the interval $[0, 1]$. The formulation of Eq. (6) then establishes a generalized framework for final-layer activations of discriminators that map inputs to

ALGORITHM 1. GAN training with adaptive noise scale.

Require: Learning rate γ , batch size m , discriminator steps n_d

- 1: Initialize $\theta^0, \phi^0, \sigma^0$
- 2: **for** $t = 1, 2, \dots, T$ **do**
- 3: **for** $k = 1, \dots, n_d$ **do**
- 4: Sample $\{\mathbf{x}^{(i)}\}_{i=1}^m \sim P_r, \{\mathbf{z}^{(i)}\}_{i=1}^m \sim P_z$
- 5: $\phi \leftarrow \text{Adam}(\phi, \nabla_\phi \mathcal{L}_B)$
- 6: $\sigma \leftarrow \text{Adam}(\sigma, \nabla_\sigma \mathcal{L}_B)$
- 7: **end for**
- 8: Sample $\{\mathbf{z}^{(i)}\}_{i=1}^m \sim P_z$
- 9: $\theta \leftarrow \text{Adam}(\theta, \nabla_\theta \mathcal{L}_G^{\text{ns}})$
- 10: **end for**

probabilities. This framework does not directly apply to critic-based architectures such as Wasserstein GAN [17,26], where the discriminator outputs an unbounded scalar score without a probabilistic activation. In addition, the conventional GAN model [16,22,36,37], which employs a sigmoid activation function, is a special case of Eq. (3) with the noise-scale parameter fixed at $\sigma = 1$, i.e., the noise scale is not learnable in that setting. In contrast, the proposed framework introduces σ as a learnable parameter in Eqs. (3)–(5), enabling the discriminator to adaptively modulate the sharpness of its output activation during training. As will be shown in Sec. III, smaller values of σ amplify the gradient signal, while larger values yield a smoother discriminator response.

For example, Figs. 2(a) and 2(b) illustrate the log probability $\ln[1 - F_\eta(x)]$ in Eq. (7) using logistic, Gaussian, and Laplace CDFs, and their corresponding gradients under different noise scales, respectively. As illustrated in Fig. 2(b), compared to the gradient characteristics of GANs employing sigmoid activation (i.e., the logistic CDF with $\sigma = 1$) [16], logistic, Gaussian, and Laplace CDFs with $\sigma < 1$ exhibit enhanced gradient signals. This enhancement will be demonstrated to effectively address the vanishing-gradient problem in conventional GANs [16,17].

III. GRADIENT SIGNALS OF THE NOISE-SCALED GAN

In this section we investigate the influence of the noise-scale parameter σ on the training dynamics of the proposed GAN. We begin by analyzing the gradient signals received by the generator under both the saturating and nonsaturating loss functions, and subsequently analyze the local convergence behavior in the vicinity of the Nash equilibrium. Finally, we present the adaptive update rule for σ , which is detailed in Algorithm 1.

Let P_r and P_g denote the continuous distributions of real and generated data, respectively. We assume their supports are closed manifolds $\mathcal{M}$ and $\mathcal{P}$ embedded as lower-dimensional subsets of the data space $\mathcal{X} = \mathbb{R}^{d_x}$. For brevity, the discriminator $D_{\phi,\sigma}$ in Eq. (6) is denoted as D in the following derivations. When $\mathcal{M}$ and $\mathcal{P}$ are either disjoint or intersect only on a set of measure zero in $\mathcal{X}$ [17], there exists an optimal discriminator $D^*(x) : \mathcal{X} \rightarrow [0, 1]$ that perfectly separates them. Specifically, $D^*(x) = 1$ on $\mathcal{M}$ and $D^*(x) = 0$ on $\mathcal{P}$ [16,17]. The optimal discriminator D^* is differentiable almost

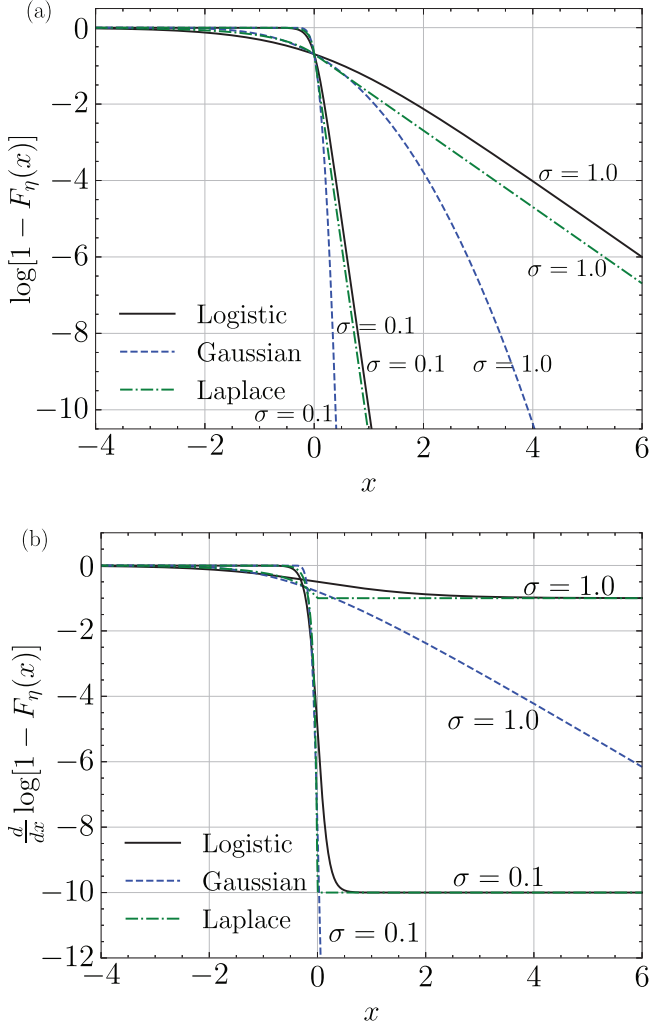


FIG. 2. (a) Log-probability $\ln[1 - F_\eta(x)]$ and (b) its corresponding gradients under different noise scales of logistic, Gaussian, and Laplace noise.

everywhere in $\mathcal{M} \cup \mathcal{P}$, and since it is locally constant on these supports, its gradient satisfies $\nabla_x D^*(x) = 0$ [16,17].

To train the generator, two loss formulations are commonly employed [16]. The first is the saturating generator loss

$$\mathcal{L}_G^{\text{sat}} = \mathbb{E}_{P_z} [\ln \{1 - D(G_\theta(z))\}], \quad (9)$$

which follows directly from the original minimax formulation of Eq. (8). To solve the problem of vanishing gradients when the generator performs poorly, Goodfellow *et al.* [16] proposed the alternative nonsaturating generator loss

$$\mathcal{L}_G^{\text{ns}} = \mathbb{E}_{P_z} [-\ln D(G_\theta(z))]. \quad (10)$$

Nevertheless, both loss formulations exhibit critical theoretical limitations when the discriminator reaches near-optimality. To formalize this, we define the Sobolev-type norm of the discriminator D as [17]

$$\|D\| = \sup_{x \in \mathcal{X}} [|D(x)| + \|\nabla_x D(x)\|_2], \quad (11)$$

where $\|\nabla_x D(x)\|_2$ denotes the ℓ_2 norm of the gradient vector. Consider a differentiable discriminator D that approximates

the optimal discriminator such that $\|D - D^*\| < \epsilon$ for an arbitrarily small tolerance $\epsilon > 0$. Arjovsky and Bottou [17] proved that under this near-optimal condition, the saturating loss $\mathcal{L}_G^{\text{sat}}$ fails to provide an effective gradient signal for the generator. Furthermore, they also demonstrated that minimizing the nonsaturating loss $\mathcal{L}_G^{\text{ns}}$ leads to training instability [16,17].

A. Gradient amplification of saturating generator loss

However, this analysis of the vanishing gradient for the generator G_θ [17] does not consider the scale parameter σ in the discriminator defined in Eq. (6). The saturating generator loss $\mathcal{L}_G^{\text{sat}}$ in Eq. (9) has a bound on its gradient norm

$$\begin{aligned} \|\nabla_\theta \mathcal{L}_G^{\text{sat}}\|_2^2 &\leq \mathbb{E}_{P_z} [\|\nabla_\theta \ln(1 - D[G_\theta(z)])\|_2^2] \\ &\leq \mathbb{E}_{P_z} \left[\frac{\|\nabla_x D[G_\theta(z)]\|_2^2 \|\nabla_\theta G_\theta(z)\|_2^2}{|1 - D[G_\theta(z)]|^2} \right]. \end{aligned} \quad (12)$$

The first inequality follows from the Jensen inequality applied to the convex function $\|\cdot\|_2^2$, namely, $\|\mathbb{E}[X]\|_2^2 \leq \mathbb{E}[\|X\|_2^2]$, and the second inequality holds for $\nabla_\theta D[G_\theta(z)] = [\nabla_\theta G_\theta(z)]^\top \nabla_x D[G_\theta(z)]$ and the matrix norm inequality $\|A^\top b\|_2 \leq \|A\|_2 \|b\|_2$. The inequality in Eq. (12) provides a preliminary bound on the gradient norm of the saturating generator loss. To obtain an explicit dependence on the noise-scale parameter σ , we further assume that the Jacobian of the generator is uniformly bounded, i.e., $\sup_z \|\nabla_\theta G_\theta(z)\|_2 \leq M$ for some constant $M > 0$. This assumption holds in practice for typical multilayer perceptron (MLP), convolutional, and ResNet-type generators with bounded activations and batch normalization. When the Jacobian is not naturally bounded, standard regularization techniques such as spectral normalization [29], weight decay, or gradient clipping can be employed. Weaker conditions, such as local Lipschitz continuity of G_θ on the support of P_z , will yield similar gradient bounds. For the sake of analytical tractability, we here assume the uniform bound for the Jacobian of the generator, as it explicitly reveals the role of the discriminator in controlling the gradient signal received by the generator.

Using the scaled discriminator defined in Eq. (6), its gradient with respect to the input can be expressed as

$$\nabla_x D(x) = \frac{1}{\sigma} f_{\eta_0} \left(\frac{V_\phi(x)}{\sigma} \right) \nabla_x V_\phi(x), \quad (13)$$

where the probability density function (PDF) $f_{\eta_0}(u) = dF_{\eta_0}(u)/du = F'_{\eta_0}(u)$. Since $f_{\eta_0}(u)$ is a PDF, then we assume it is bounded, i.e., $0 \leq f_{\eta_0}(u) \leq B$ for all $u \in \mathbb{R}$. Moreover, for the logistic, Gaussian, and Laplace distributions, $f_{\eta_0}(u) > 0$ for all $u \in \mathbb{R}$. Since $F_{\eta_0}(u) \in (0, 1)$ for all finite u , the CDF-parameterized discriminator cannot attain the values 0 or 1 exactly. We therefore interpret near-optimality within the CDF family as a finite regime in which $V_{\phi^*}(x)$ takes a large positive (negative) but finite value on the support of P_r (P_g), so that $D^*(x) \approx 1$ on $\mathcal{M}$ and $D^*(x) \approx 0$ on $\mathcal{P}$. In this regime, D^* is approximately locally constant on each support, hence $\nabla_x D^*(x) \approx 0$. Because $V_{\phi^*}(x)/\sigma$ remains finite, $f_{\eta_0}[V_{\phi^*}(x)/\sigma] > 0$ strictly, and Eq. (13) then gives $\nabla_x V_{\phi^*}(x) \approx 0$ locally on $\mathcal{M}$ and $\mathcal{P}$.

We define the deviation $\delta V(x) = V_\phi(x) - V_{\phi^*}(x)$ to quantify the departure from optimality. Since $\nabla_x V_\phi(x) = \nabla_x V_{\phi^*}(x) + \nabla_x \delta V(x)$ and $\nabla_x V_{\phi^*}(x) \approx 0$ in the near-optimal regime established above, the gradient of the discriminator reduces to

$$\nabla_x D(x) = \frac{1}{\sigma} f_{\eta_0} \left(\frac{V_{\phi^*}(x) + \delta V(x)}{\sigma} \right) \nabla_x \delta V(x). \quad (14)$$

We further assume the deviation $|\delta V(x)| < \delta_V$ with its gradient $\|\nabla_x \delta V(x)\|_2 < \epsilon_V$ for constants $\delta_V, \epsilon_V > 0$. In addition, the PDF f_{η_0} is Lipschitz with constant $L_f > 0$, i.e., $|f_{\eta_0}(u) - f_{\eta_0}(v)| \leq L_f |u - v|$. For common distributions, the Lipschitz constants are well established [38]. Specifically, the PDFs f_{η_0} of logistic, Gaussian, and Laplace distributions have Lipschitz constants of $L_f = 1/6\sqrt{3}$, $1/\sqrt{2\pi}e$, and $1/2$, respectively.

Under these conditions, we find that

$$f_{\eta_0} \left(\frac{V_{\phi^*}(x) + \delta V(x)}{\sigma} \right) \leq f_{\eta_0} \left(\frac{V_{\phi^*}(x)}{\sigma} \right) + \frac{L_f \delta_V}{\sigma} \leq B + \frac{L_f \delta_V}{\sigma}, \quad (15)$$

and the norm of the gradient in Eq. (14) satisfies

$$\|\nabla_x D(x)\|_2 \leq \frac{B\epsilon_V}{\sigma} + L_f \frac{\delta_V \epsilon_V}{\sigma^2}. \quad (16)$$

Noting that D^* is locally zero on the support of P_g , the approximation condition $\|D - D^*\| < \epsilon$ with the Sobolev-type norm of Eq. (11) implies [17]

$$|1 - D[G_\theta(z)]| \geq |1 - D^*[G_\theta(z)]| - |D[G_\theta(z)] - D^*[G_\theta(z)]| \geq 1 - \epsilon. \quad (17)$$

Taking the square root of Eq. (12) and substituting Eq. (16) along with the uniform Jacobian bound M , we obtain the bound

$$\|\nabla_\theta \mathcal{L}_G^{\text{sat}}\|_2 \leq \frac{M}{1 - \epsilon} \left(\frac{B\epsilon_V}{\sigma} + L_f \frac{\delta_V \epsilon_V}{\sigma^2} \right). \quad (18)$$

Since δ_V and ϵ_V bound the deviation of V_ϕ , whereas the Sobolev condition $\|D - D^*\| < \epsilon$ is imposed on the discriminator, we now establish the scaling of δ_V and ϵ_V in terms of ϵ . The Sobolev condition implies bounds on both the function value $|D - D^*| < \epsilon$ and its gradient $\|\nabla_x D - \nabla_x D^*\|_2 < \epsilon$. First, applying the mean value theorem to the CDF activation F_η between $V_\phi(x)/\sigma$ and $V_{\phi^*}(x)/\sigma$, together with the bound $f_{\eta_0} \leq B$, yields $|D - D^*| \leq (B/\sigma) |\delta V| \leq (B/\sigma) \delta_V$. This relationship implies that for a given σ , the upper bound δ_V scales linearly with ϵ , i.e., $\delta_V = O(\epsilon)$. Second, since $\nabla_x D^* = 0$ in the near-optimal regime, the gradient restriction simplifies to $\|\nabla_x D\|_2 < \epsilon$. From Eq. (14), due to $\|\nabla_x \delta V\|_2 \leq \epsilon_V$, the gradient deviation satisfies $\|\nabla_x D\|_2 \leq \frac{B}{\sigma} \epsilon_V$. For a given σ , the upper bound ϵ_V also scales linearly with ϵ , yielding $\epsilon_V = O(\epsilon)$. Consequently, the product of these two deviations $\delta_V \epsilon_V = O(\epsilon^2)$ constitutes a higher-order term, which becomes negligible in the limit $D \rightarrow D^*$. Thus, for unit scale $\sigma = 1$, Eq. (18) simplifies to the bound

$$\|\nabla_\theta \mathcal{L}_G^{\text{sat}}\|_2 \leq \frac{M\epsilon}{1 - \epsilon} \quad (19)$$

established in [17].

The gradient bound in Eq. (18) reveals the beneficial role of a small noise scale $\sigma < 1$ in enhancing the training dynamics of GAN optimization. Specifically, the first term $B\epsilon_V/\sigma$ denotes that decreasing σ amplifies the gradient signal by a factor of $1/\sigma$, while the second higher-order term $L_f \delta_V \epsilon_V / \sigma^2$ exhibits even stronger scaling with $1/\sigma^2$. Consequently, an appropriate choice of $\sigma < 1$ can improve the gradient flow to the generator in the regime where the discriminator is close to D^* and the supports of P_r and P_g are largely disjoint. However, it is important to note that this bound has no effect when $\nabla_x V_\phi$ inherently vanishes. Thus, the noise-scale mechanism should be interpreted as a strategy to reduce, rather than completely solve, the vanishing-gradient problem.

B. Gradient amplification of nonsaturating generator loss

When the generator G_θ exhibits suboptimal performance and the discriminator rejects samples from G_θ with high confidence, i.e., $D[G_\theta(z)] \rightarrow 0$, the logarithmic term $\ln[1 - D[G_\theta(z)]]$ in Eq. (9) saturates [16]. Thus, Goodfellow *et al.* [16] proposed an alternative training strategy that employs the nonsaturating generator loss in Eq. (10) during the initial training phase.

However, even with this training strategy indicated in Eq. (10), the generator still suffers from the vanishing-gradient problem during the training. This is because, for the generalized discriminator model in Eq. (6) as the final layer activation, the gradient of the generator loss of Eq. (10) can be expressed as

$$\nabla_\theta \mathcal{L}_G^{\text{ns}} = \mathbb{E}_{P_z} \left[-r \left(\frac{V_\phi(G_\theta)}{\sigma} \right) \frac{1}{\sigma} \frac{\partial V_\phi(G_\theta)}{\partial G_\theta} \frac{\partial G_\theta(z)}{\partial \theta} \right] \quad (20)$$

with the reversed hazard rate $r(u) = f_{\eta_0}(u)/F_{\eta_0}(u)$. Since the discriminator output asymptotically approaches zero, i.e., $D[G_\theta(z)] = F_{\eta_0}\{V_\phi[G_\theta(z)]/\sigma\} \rightarrow 0$, then $V_\phi[G_\theta(z)]$ becomes a large negative value for $\sigma > 0$. In the conventional GAN formulation [16], the discriminator $D(x)$ employs a sigmoid activation, which corresponds to the logistic CDF of Eq. (3) with noise scale $\sigma = 1$, i.e., $F_{\eta_0}(u) = 1/(1 + e^{-u})$. Under this configuration, Eq. (20) becomes

$$\nabla_\theta \mathcal{L}_G^{\text{ns}} = \mathbb{E}_{P_z} \left[\frac{-1}{1 + e^{V_\phi(G_\theta)}} \frac{\partial V_\phi(G_\theta)}{\partial G_\theta} \frac{\partial G_\theta(z)}{\partial \theta} \right] \quad (21)$$

with the term $-1/(1 + e^{V_\phi(G_\theta)})$ approaching -1 . When the generated samples are of poor quality, the discriminator can easily distinguish them from real data, making its value function $V_\phi(G_\theta)$ relatively flat with respect to the generator parameters. This results in $\partial V_\phi(G_\theta)/\partial G_\theta \rightarrow 0$, and the vanishing-gradient problem arises.

However, when the CDF $F_{\eta_0}(u)$ incorporates the scale parameter $\sigma < 1$, the nonsaturating loss in Eq. (20) introduces a gradient scaling mechanism. Specifically, the gradient with respect to the generator parameters is modulated by an overall factor of $r(u)/\sigma$ at $u = V_\phi[G_\theta(z)]/\sigma$. Here it is noted that the internal network derivative $\partial V_\phi(G_\theta)/\partial G_\theta$ is independent of the noise scale σ .

To verify the stability of this gradient amplification across different CDF families, we analyze the asymptotic behavior of the ratio $r(u)$ in the regime where the generator performs poorly, i.e., $u = V_\phi(G_\theta) \rightarrow -\infty$. For the logistic

distribution, the ratio is $r(u) = 1/(1 + e^u)$, which smoothly approaches 1 as $u \rightarrow -\infty$. Similarly, for the Laplace distribution, the PDF and CDF are identical on the negative half-line, i.e., $f_{\eta_0}(u) = F_{\eta_0}(u) = e^u/2$ for $u < 0$. This yields a constant reversed hazard rate of $r(u) = 1$ exactly. Because this ratio remains constant in the saturated regime where the discriminator output approaches zero, the overall gradient amplification factor $r(u)/\sigma$ in Eq. (20) is governed purely by the term $1/\sigma$. In both cases the ratio remains strictly bounded. Therefore, the term $1/\sigma$ provides a stable amplification of the gradient. Conversely, the Gaussian distribution exhibits a different asymptotic behavior. According to Mills ratio [39,40], the ratio $r(u) \sim |u|$ grows without bound as $u \rightarrow -\infty$. The amplification factor $r(u)/\sigma$ is therefore unbounded in principle, and in regimes where $V_\phi(G_\theta)$ takes large negative values, this can manifest as gradient instability during training. This mechanism provides a theoretical explanation for the experimentally observed gradient instability of the Gaussian CDF discussed in Sec. IV C. This asymptotic behavior is in line with the theoretical gradient curves shown in Fig. 2(b), where the Gaussian CDF exhibits sharper gradient growth at large negative input than the logistic and Laplace CDFs. While selecting an appropriate noise scale $\sigma < 1$ can improve the gradient signal available to the generator in the nonsaturating formulation of Eq. (10), the Gaussian CDF may suffer from the gradient-instability risk associated with the unbounded ratio $r(u)$.

C. Local convergence analysis of scaled discriminator

In this subsection we analyze the local convergence properties of the GAN dynamics under the original minimax objective $\mathcal{L}(\theta, \phi, \sigma)$ defined in Eq. (8). This corresponds to the saturating generator loss formulation, and the stability is evaluated near the Nash equilibrium characterized by $P_r = P_g$ and $D(x) = 1/2$ for all x [22,36].

The training dynamics of GANs can be formulated as a two-player differential game, where the discriminator $D(x) = F_{\eta_0}[V_\phi(x)/\sigma]$ maximizes the objective function $\mathcal{L}(\phi, \theta)$ defined in Eq. (8), while the generator minimizes it, i.e., maximizes $-\mathcal{L}(\phi, \theta)$ [22,36]. The corresponding gradient vector field is calculated as [22,36]

$$\mathbf{v}(\phi, \theta) = \begin{bmatrix} \nabla_\phi \mathcal{L}(\phi, \theta) \\ \nabla_\theta [-\mathcal{L}(\phi, \theta)] \end{bmatrix}, \quad (22)$$

with the Jacobian matrix

$$\mathbf{J}(\phi, \theta) = \begin{bmatrix} \nabla_{\phi\phi}^2 \mathcal{L} & \nabla_{\phi\theta}^2 \mathcal{L} \\ -\nabla_{\theta\phi}^2 \mathcal{L} & -\nabla_{\theta\theta}^2 \mathcal{L} \end{bmatrix}. \quad (23)$$

We focus on the top-left block of the Jacobian, namely, the Hessian $\nabla_{\phi\phi}^2 \mathcal{L} = \nabla_{\phi\phi}^2 \mathbb{E}_{P_r} \{\ln[1 - D(G_\theta(\mathbf{z}))]\} + \nabla_{\phi\phi}^2 \mathbb{E}_{P_r} \{\ln D(\mathbf{x})\}$, evaluated near the Nash equilibrium [22,36]. At equilibrium, the first term $\nabla_{\phi\phi}^2 \mathbb{E}_{P_r} \{\ln[1 - D(G_\theta(\mathbf{z}))]\}$ based on the generated data vanishes, while the second term $\nabla_{\phi\phi}^2 \mathbb{E}_{P_r} \{\ln D(\mathbf{x})\}$ derived from real data becomes dominant. Using Eq. (6), it can be

calculated as

$$\nabla_{\phi\phi}^2 \mathcal{L} \approx \mathbb{E}_{P_r} \left[\frac{d}{d\alpha} r(\alpha) \frac{(\nabla_\phi V_\phi)(\nabla_\phi V_\phi)^\top}{\sigma^2} + r(\alpha) \frac{\nabla_{\phi\phi}^2 V_\phi}{\sigma} \right], \quad (24)$$

where $\alpha = V_\phi(x)/\sigma$. For small values of $\sigma \ll 1$, the first term in Eq. (24), scaled by the factor of $1/\sigma^2$, provides the leading contribution to the Hessian. We note that the monotonic decrease of the reversed hazard rate $r(u)$, i.e., its derivative $r'(u) = dr(u)/du < 0$, determines the negative definiteness of the Hessian $\nabla_{\phi\phi}^2 \mathcal{L}$. However, it is important to note that although this negative definiteness provides a crucial stabilizing mechanism for GAN training, it is not a sufficient condition for the overall stability of the adversarial learning dynamics. Within the framework of Mescheder *et al.* [22], local convergence of GAN training around the Nash equilibrium further requires the Jacobian of the coupled gradient vector field to possess eigenvalues with strictly positive real parts. As indicated in Eq. (23), this sufficient condition depends jointly on the discriminator Hessian, the generator Hessian, and the cross-coupling blocks, which together determine the local stability of the GAN optimization dynamics.

For the logistic PDF $f_{\eta_0}(u) = e^{-u}/(1 + e^{-u})^2$ and CDF of Eq. (3) ($\sigma = 1$), the reversed hazard rate $r(u) = 1/(1 + e^u)$ with its derivative

$$r'(u) = \frac{-e^u}{(1 + e^u)^2} < 0. \quad (25)$$

In other words, the logistic CDF of Eq. (3) is also called log-concave [39,40]. For the unit-scale Gaussian distribution with PDF $f_{\eta_0}(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ and CDF of Eq. (4) and noting $f'_{\eta_0}(u) = -u f_{\eta_0}(u)$, the reversed hazard rate $r(u)$ has its derivative as

$$r'(u) = -\frac{f_{\eta_0}(u)[uF_{\eta_0}(u) + f_{\eta_0}(u)]}{F_{\eta_0}^2(u)}. \quad (26)$$

Let $s(u) = [uF_{\eta_0}(u) + f_{\eta_0}(u)]$. It is noted that

$$\lim_{u \rightarrow -\infty} f_{\eta_0}(u) = 0, \quad (27)$$

$$\lim_{u \rightarrow -\infty} s(u) = \lim_{u \rightarrow -\infty} \frac{F_{\eta_0}(u)}{1/u} = 0, \quad (28)$$

and $ds(u)/du = F_{\eta_0}(u) > 0$. Thus, we obtain $s(u) > 0$ and $r'(u) < 0$ in Eq. (26) for strictly log-concave Gaussian CDF [40]. Moreover, for the Laplace PDF $f_{\eta_0}(u) = \frac{1}{2} e^{-|u|}$ and CDF of Eq. (5) ($\sigma = 1$), we find

$$r(u) = \begin{cases} 1, & u < 0, \\ \frac{e^{-u}}{2 - e^{-u}}, & u > 0, \end{cases} \quad (29)$$

$$r'(u) = \begin{cases} 0, & u < 0, \\ -\frac{2e^{-u}}{(2 - e^{-u})^2}, & u > 0, \end{cases} \quad (30)$$

which yields $r'(u) \leq 0$. Combining the two regimes with the continuity $\lim_{u \rightarrow 0^\pm} r(u) = 1$, we conclude that $r(u)$ is nonincreasing on $\mathbb{R}$, so the Laplace CDF is log-concave.

Therefore, based on the analysis above, the Hessian $\nabla_{\phi\phi}^2 \mathcal{L}$ of Eq. (24) becomes negative semidefinite for sufficiently small σ , due to the $1/\sigma^2$ scaling of the leading term. We note that this concerns only the top-left block of the full

Jacobian $\mathbf{J}(\phi, \theta)$ in Eq. (23). The definiteness of this block alone does not, in general, determine the spectrum of the full block matrix, since a complete stability analysis requires additional conditions on the off-diagonal blocks $\nabla_{\phi\theta}^2 \mathcal{L}$ [22,36]. Nevertheless, a negative semidefinite top-left block is a necessary condition for local stability of the discriminator, and the $1/\sigma^2$ scaling indicates that a small σ enhances favorable local curvature properties. In this sense, the scale parameter σ plays a role analogous to an implicit regularizer that influences the local convergence dynamics.

D. Adaptive noise injection

The adaptive optimization of the noise scale σ underlies the proposed method, distinguishing it from fixed-temperature scaling approaches [30,31]. As noted in Sec. II, the objective function $\mathcal{L}(\theta, \phi, \sigma)$ in Eq. (8) depends not only on the generator parameters θ and discriminator parameters ϕ , but also on the noise scale σ within the proposed CDF activations of Eqs. (3)–(5). Rather than treating σ as a fixed hyperparameter, we optimize it jointly with the network parameters during training.

Specifically, σ can be updated via mini-batch gradient ascent, where a minibatch $\mathcal{B}$ refers to a randomly sampled subset of the training data used to compute an approximate gradient at each iteration. Let $\mathcal{L}_{\mathcal{B}}$ denote the objective function evaluated on this minibatch, and the update rule of σ is given by

$$\sigma^t = \sigma^{t-1} + \gamma^t \nabla_{\sigma} \mathcal{L}_{\mathcal{B}}(\theta^{t-1}, \phi^{t-1}, \sigma^{t-1}) \quad (31)$$

with the learning rate $\gamma^t > 0$ at the t th training iteration. In practice, σ can be updated using the Adam optimizer [41], as described in Algorithm 1.

As detailed in Algorithm 1, the discriminator parameters ϕ and the noise scale σ are updated jointly by maximizing the objective $\mathcal{L}$, while the generator parameters θ are updated by minimizing the nonsaturating loss $\mathcal{L}_G^{\text{ns}}$ in Eq. (10). It is noted that the noise-scale parameter σ must be strictly positive throughout GAN training. In practice, we parameterize $\sigma = \exp(\tilde{\sigma})$ and use $\tilde{\sigma} = \ln \sigma$ as the learnable parameter in Algorithm 1, with gradients computed via the chain rule $\partial \mathcal{L} / \partial \tilde{\sigma} = \sigma \partial \mathcal{L} / \partial \sigma$. The notation $\mathcal{L}_{\mathcal{B}}$ in Algorithm 1 denotes the empirical estimate of $\mathcal{L}$ computed on a minibatch by replacing the expectations $\mathbb{E}_{P_r}(\cdot)$ and $\mathbb{E}_{P_z}(\cdot)$ with their corresponding empirical averages.

IV. EXPERIMENTAL RESULTS

Before presenting the experimental results, we briefly clarify the distinction between training and inference of a GAN. During training, the generator G_{θ} and the discriminator D are jointly optimized via the minimax objective in Eq. (8), as summarized in Algorithm 1. Once training is completed, the generation is performed by the trained generator, i.e., a latent noise vector $\mathbf{z} \sim P_z$ is mapped through G_{θ} to produce a synthetic sample $G_{\theta}(\mathbf{z})$, while the discriminator D is not involved. Accordingly, all generated samples shown in the following subsections are obtained by the trained generator.

TABLE I. Comparison of generation quality metrics on the Gaussian mixture dataset for $p = 100$ and $p = 1000$ samples per mode. Coverage and precision are evaluated with a distance threshold $\varepsilon = 0.1$.

Activation function	$p = 100$		$p = 1000$	
	C_{ε}	P_{ε}	C_{ε}	P_{ε}
Sigmoid ($\sigma = 1$)	0.014	0.179	0.463	0.908
Logistic CDF	0.878	0.821	0.651	0.982
Gaussian CDF	0.718	0.895	0.697	0.992
Laplace CDF	0.772	0.871	0.734	0.995

A. Gaussian mixture dataset

To demonstrate the theoretical results of the proposed GAN model in Fig. 1, we employ a synthetic dataset consisting of a two-dimensional Gaussian mixture with n modes uniformly distributed along the unit circle. For each mode $k \in \{0, 1, \dots, n-1\}$, we set the mean vector as $\boldsymbol{\mu}_k = [\sin(2\pi k/n), \cos(2\pi k/n)]^T$ and the covariance matrix as $\boldsymbol{\Sigma}_k = 0.0125 \mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Using $n = 5$ Gaussian modes, we first consider a sparse setting with $p = 100$ samples per mode, yielding 500 real data points in total. The latent noise vector $\mathbf{z} \in \mathbb{R}^2$ is drawn from a standard bivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. To obtain a centered and normalized dataset $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2$, the empirical mean is subtracted in each dimension, and the data are subsequently scaled by the global maximum magnitude. This sparse configuration is deliberately chosen so that the supports of the real and generated distributions remain largely disjoint, which is precisely the regime under which the vanishing-gradient problem is pronounced. [17]. This setting therefore serves as a rigorous test for evaluating the effectiveness of the proposed method.

The generator network is a fully connected feedforward neural network with three hidden layers of 16 ReLU units each, followed by a linear output layer. It maps the latent noise vector $\mathbf{z} \in \mathbb{R}^2$ to the generated sample $G_{\theta}(\mathbf{z}) \in \mathbb{R}^2$. The discriminator processes either the real input $\mathbf{x}$ or the generated input $G_{\theta}(\mathbf{z})$ through three hidden layers and yields a scalar output via the final-layer activation F_{η} defined in Eq. (2). Both networks are trained with the Adam optimizer [41] using a learning rate of 5×10^{-5} , a minibatch size of 32, and the nonsaturating generator loss $\mathcal{L}_G^{\text{ns}}$ defined in Eq. (10), for a total of 10 000 training epochs (source code is provided in GitHub repository [42]). To quantitatively evaluate the generation quality, we employ two complementary metrics: coverage (C_{ε}) and precision (P_{ε}). Here C_{ε} measures the fraction of real samples that have at least one generated sample within distance $\varepsilon = 0.1$, and P_{ε} denotes the fraction of generated samples that lie within distance ε of at least one real sample. For both C_{ε} and P_{ε} , higher values indicate better generator performance.

As shown in Fig. 3(a), when the discriminator employs the sigmoid activation (i.e., the logistic CDF with fixed $\sigma = 1$), the GAN exhibits mode collapse, with a coverage of only 0.014 and a precision of 0.179 (Table I), and cannot successfully learn the target modes indicated by filled circles.

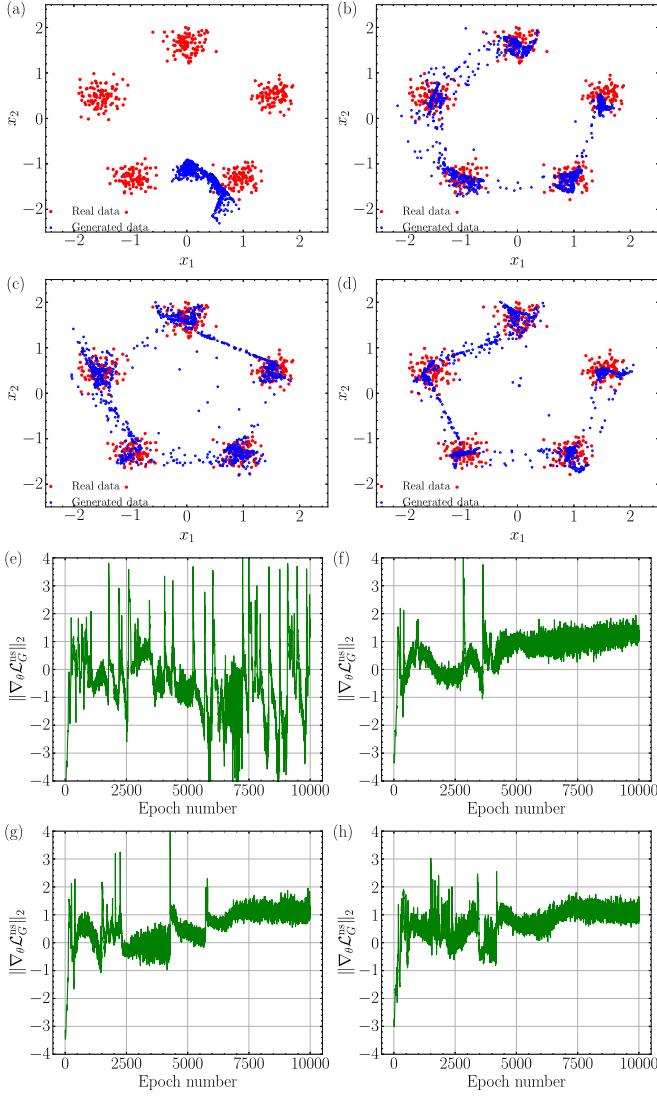


FIG. 3. Generated samples $G_\theta(\mathbf{z})$ (blue circles) and real data (red dots) from a GAN trained on a five-mode Gaussian mixture dataset ($p = 100$ per mode), with different final-layer activations: (a) sigmoid, (b) logistic CDF, (c) Gaussian CDF, and (d) Laplace CDF. The corresponding gradient norms $\|\nabla_\theta \mathcal{L}_G^{\text{ns}}\|_2$ during training are shown in (e)–(h).

The corresponding gradient norm $\|\nabla_\theta \mathcal{L}_G\|_2$ of the generator during training is illustrated in Fig. 3(e), where both vanishing and exploding gradients [17] are observed, particularly after 5000 training epochs. In contrast, when the discriminator employs the logistic, Gaussian, or Laplace CDF activations, the GAN successfully learns all five target modes, as shown in Figs. 3(b)–3(d). Among these, the GAN with the logistic CDF activation achieves the highest coverage ($C_\varepsilon = 0.878$), while those with the Gaussian and Laplace CDFs also perform well ($C_\varepsilon = 0.718$ and 0.772 , respectively). All three models attain comparable precision ($P_\varepsilon > 0.82$), indicating that the generated samples lie within the vicinity of the real data distribution. It is worth noting that a natural trade-off can occur between these two metrics. For instance, as observed for the logistic CDF with ($p = 100$) in Table I, the model successfully captures a wider variety of modes (yield-

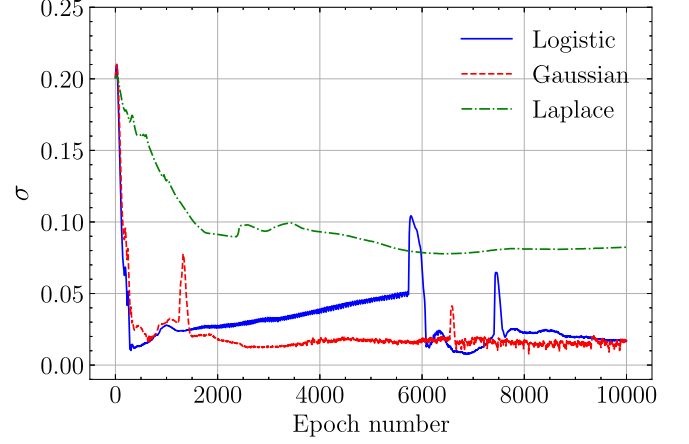


FIG. 4. Learning curves of the noise-scale parameter σ for the logistic, Gaussian, and Laplace CDF activations on the Gaussian mixture dataset with $p = 100$ samples per mode. Other parameters are the same as in Fig. 3.

ing the highest coverage). However, this broader exploration occasionally causes generated samples to deviate from the high-density regions of the real data distribution, resulting in a lower precision. The corresponding gradient norms in Figs. 3(f)–3(h) remain stable throughout training, indicating that the adaptive noise scale effectively provides informative gradient signals.

The learning curves of the noise-scale parameter σ for the three CDF activations, all initialized at $\sigma^0 = 0.2$, are plotted in Fig. 4. As shown in Fig. 4, σ first decreases rapidly and then converges to a locally optimal value below unity for all three CDF activations. Specifically, the logistic and Gaussian CDFs converge to values on the order of 10^{-2} , while the Laplace CDF converges to a slightly larger value of 0.08. According to the theoretical gradient bound in Eq. (18), a noise scale $\sigma < 1$ amplifies the gradient signals. The bound involves theoretical quantities such as the generator Jacobian bound M , the Sobolev neighborhood radius ϵ , and the logit deviations ϵ_V , δ_V , which serve as analytical tools to characterize the gradient amplification mechanism but are not used to be computed in practice. The initial value σ^0 is therefore chosen heuristically. In practice, the choice of σ^0 is not critical as long as $\sigma^0 < 1$, since the adaptive update rule in Eq. (31) can automatically drive σ toward its locally optimal value during training, as shown in Fig. 4.

We also carry out the experiment with $p = 1000$ samples per mode (5000 total). The generated samples and the gradient norms for the dense-data case are shown in Fig. 5, and the corresponding coverage and precision metrics are reported as the $p = 1000$ block of Table I. With the denser dataset, all three CDF activations attain high precision $P_\varepsilon > 0.98$ (see Table I), and the intermodal artifacts observed in Figs. 3(a)–3(d) become less evident in the case of $p = 1000$, as shown in Figs. 5(a)–5(d). The corresponding stable gradient norms are plotted in Figs. 5(f)–5(h). While the adaptive noise scale σ provides gradient flows and prevents mode collapse, all three CDF activations exhibit reduced coverage compared to the sparse setting. This reduction arises from the increased difficulty of the evaluation itself: when $p = 1000$, the true

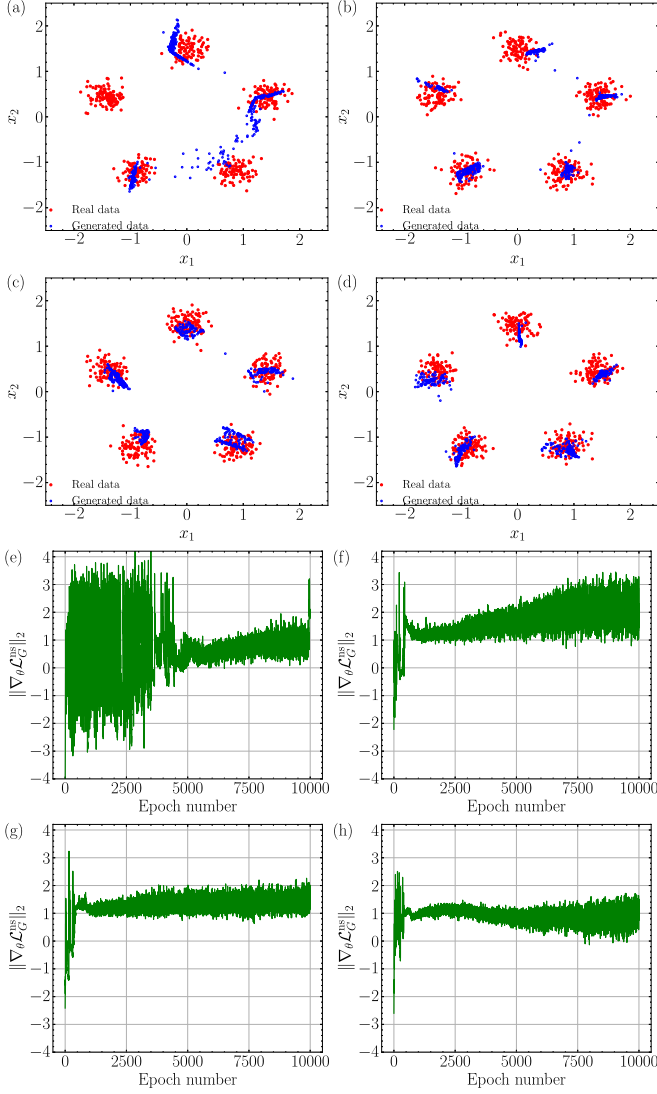


FIG. 5. Generated samples $G_\theta(\mathbf{z})$ (blue circles) and real data (red dots) using (a) sigmoid, (b) logistic CDF, (c) Gaussian CDF, and (d) Laplace CDF. For clarity, 500 of the 5000 generated and real samples are randomly selected and plotted for the case of $p = 1000$ per mode. The corresponding gradient norms $\|\nabla_\theta \mathcal{L}_G^{\text{ns}}\|_2$ during training are shown in (e)–(h).

distribution spreads deeper into the Gaussian tails, and a fixed threshold $\varepsilon = 0.1$ covers a smaller fraction of these peripheral samples. We emphasize that C_ε and P_ε are meaningful only for comparisons under the same threshold ε and the same sample density p .

B. MNIST dataset

We also apply the logistic, Gaussian, and Laplace CDF activations to the discriminator to learn the MNIST dataset [43]. The generator transforms a 100-dimensional uniformly distributed latent noise vector $\mathbf{z}$ into a 784-dimensional output vector, which is then reshaped into a 28×28 grayscale image. The generator network consists of four fully connected layers with hidden dimensions of 256, 512, and 1024, respectively. Each hidden layer passes through the leaky ReLU with a

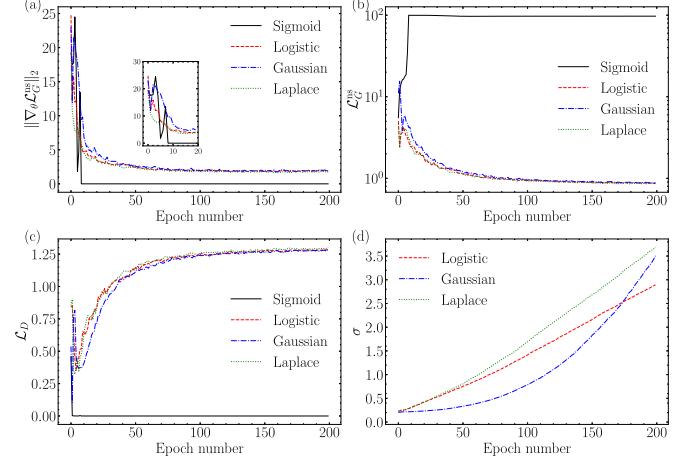


FIG. 6. Training dynamics on MNIST for GANs with the sigmoid, logistic, Gaussian, and Laplace CDFs. (a) Generator gradient norm $\|\nabla_\theta \mathcal{L}_G^{\text{ns}}\|_2$, (b) nonsaturating generator loss $\mathcal{L}_G^{\text{ns}}$, (c) discriminator loss $\mathcal{L}_D$, and (d) the noise scale σ with the initial value of $\sigma^0 = 0.2$ versus the epoch number.

negative slope coefficient of 0.2, while the final layer of the generator uses a tanh activation to produce outputs normalized to the range $[-1, 1]$. The discriminator has layers of size 1024, 512, and 256, and also employs leaky ReLU activations along with dropout layers with a probability of 0.3. The discriminator is trained using both real and fake (generated) samples, and the generator is optimized to maximize the probability of the discriminator misclassifying generated samples as real by the nonsaturating loss indicated in Eq. (10). Both networks are trained using the Adam optimizer with a learning rate of 2×10^{-4} . The input images are normalized to $[-1, 1]$, and a batch size of 100 is used for 200 epochs in all experiments (source codes in our provided GitHub repository). To quantitatively assess the degree of mode collapse, we employ a convolutional classifier to recognize the unique class labels represented in the generated samples. This classifier is trained on the MNIST training set from random initialization, without any pretraining on external datasets such as ImageNet. It achieves a test accuracy exceeding 99%, and its network architecture is provided in our GitHub repository.

As shown in Fig. 6(a), the sigmoid GAN exhibits vanishing generator gradients within the first few epochs, whereas GANs with the three CDFs have nontrivial gradients during an early exploration phase before gradually decaying. Therefore, the discriminator of the sigmoid GAN rapidly becomes a near-perfect classifier, with $D(\mathbf{x}_{\text{real}}) \approx 1$ and $D[G(\mathbf{z})] \approx 0$. Consequently, this causes the nonsaturating generator loss $\mathcal{L}_G^{\text{ns}} = -\ln D[G(\mathbf{z})]$ to diverge (truncated to 100 for visualization). In contrast, GANs with the three CDFs converge to an adversarial equilibrium with $D(\mathbf{x}) \approx 1/2$ for both real and generated samples, yielding the theoretical Nash values $\mathcal{L}_D = 2 \ln 2$ and $\mathcal{L}_G^{\text{ns}} = \ln 2$. Figures 6(b) and 6(c) show that the experimental losses approach these limits, i.e., $\mathcal{L}_D \approx 1.26$ – 1.32 and $\mathcal{L}_G^{\text{ns}} \approx 0.8$, indicating that the discriminator and generator settle into a near-balanced adversarial equilibrium. Finally, the learnable noise scale σ in Fig. 6(d) grows monotonically from $\sigma^0 = 0.2$. We argue that small σ in the early exploration



FIG. 7. Generated samples from a GAN trained on the MNIST dataset using different activation functions in the final layer of the discriminator: (a) sigmoid, (b) logistic CDF of Eq. (3), (c) Gaussian CDF of Eq. (4), and (d) Laplace CDF of Eq. (5).

phase amplifies the gradient flow, while its subsequent increase smooths the discriminator response and stabilizes the adversarial dynamics.

The MNIST dataset contains 70 000 images (60 000 for training and 10 000 for testing), distributed roughly equally among the 10 digit classes (0–9) [43]. As illustrated in Fig. 7(a), when the final layer of the discriminator employs the sigmoid function, the vanishing-gradient problem arises and all generated images are classified as the digit 3, indicating a complete collapse of mode diversity. In contrast, when the final layer of the discriminator employs the logistic, Gaussian, or Laplace CDFs, the classifier identifies all 10 digit classes (0–9) in the generated samples, as depicted in Figs. 7(b), 7(c), and 7(d). The corresponding classwise sample counts are listed in Table II and plotted as histograms in Fig. 8. The logistic and Laplace CDF activations yield distributions that fluctuate closely around the balanced reference of 100 samples per class. By contrast, the GAN with the Gaussian CDF activation produces a strongly imbalanced distribution, for example, the digit 3 alone accounting for 367 out of 1000 generated samples. This indicates that, although the Gaussian

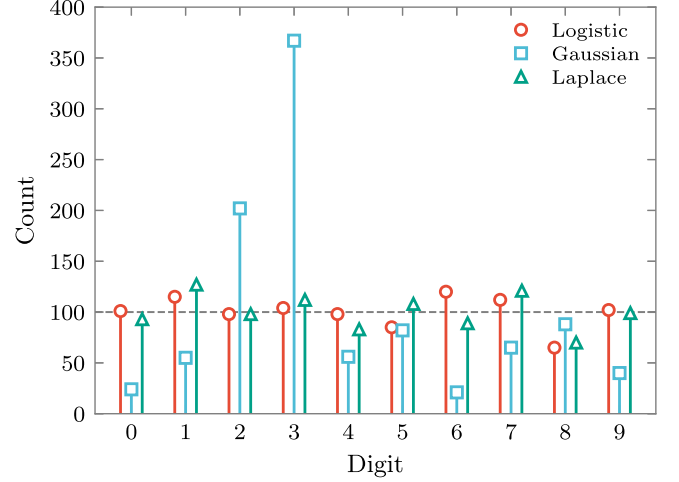


FIG. 8. Classwise counts of 1000 generated digits for GANs with logistic, Gaussian, and Laplace CDF discriminator activations on MNIST. The dashed line at 100 indicates the perfectly balanced reference (1000 samples evenly distributed over 10 classes).

CDF avoids the mode collapse, it still suffers from a partial mode imbalance.

Furthermore, we also employ the inception score (IS) [18] and the Fréchet inception distance (FID) [44] to assess the quality of images generated by the proposed GAN model on the MNIST dataset. The IS measure is calculated as [18]

$$\text{IS} = e^{(\mathbb{E}_{\mathbf{x}}[\text{D}_{\text{KL}}(p(y|\mathbf{x}) \| p(y))])}, \quad (32)$$

where $p(y|\mathbf{x})$ represents the class prediction distribution from a pre-trained classifier for a single generated image $\mathbf{x}$, $p(y) = \int p(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ denotes the marginal class distribution across all generated samples, and the Kullback-Leibler divergence D_{KL} between distributions $p(y|\mathbf{x})$ and $p(y)$ is defined as $\text{D}_{\text{KL}}(p(y|\mathbf{x}) \| p(y)) = \sum_y p(y|\mathbf{x}) \ln[p(y|\mathbf{x})/p(y)]$. The IS measure evaluates both the quality and the diversity of generated samples, wherein higher scores correspond to better generative capability. Since the computation of IS requires dividing the generated samples into multiple subsets, IS is calculated for each subset, and the mean and standard deviation are then given. The $\pm$ symbols in the following tables denote these corresponding statistical measures. Here

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (33)$$

quantifies the statistical similarity between the distributions of real and generated images in a feature space [44]. It models these distributions as multivariate Gaussian with means and covariances (μ_r, Σ_r) for real data and (μ_g, Σ_g) for generated data. A lower FID corresponds to a closer match between the real and generated distributions, thereby indicating better generation quality. We note that IS and FID are dependent on the specific classifier used for their computation and have known limitations. In particular, IS may not fully capture mode collapse [45]. These metrics should therefore be interpreted as heuristic indicators of generation quality rather than definitive

TABLE II. Classwise sample counts of 1000 generated digits for GANs with different discriminator activations.

Activation Function	Generated digit counts									
	0	1	2	3	4	5	6	7	8	9
Sigmoid	All classified as digit 3 (mode collapse)									
Logistic CDF	101	115	98	104	98	85	120	112	65	102
Gaussian CDF	24	55	202	367	56	82	21	65	88	40
Laplace CDF	93	127	98	112	83	108	89	121	70	99

TABLE III. Comparison of generated samples under different discriminator activation functions.

Activation function	IS	FID	Mode	KL _{mode}
Sigmoid	1.0211 ± 0.0032	0.0581	1/10	2.3026
Logistic CDF	9.1365 ± 0.1504	0.0005	10/10	0.0022
Gaussian CDF	5.2918 ± 0.4560	0.0024	10/10	0.3946
Laplace CDF	9.1116 ± 0.1586	0.0003	10/10	0.0139

measures. In addition, the metric

$$\text{KL}_{\text{mode}} = \sum_{y=0}^9 p(y) \ln[p(y)/(1/10)] \quad (34)$$

is computed between the marginal class distribution $p(y)$ of 1000 generated digits with the uniform reference $1/10$. A smaller KL_{mode} indicates a more balanced class coverage of the generator, while the upper bound $\ln 10 \approx 2.3026$ corresponds to complete mode collapse onto a single class. Note that KL_{mode} differs from the D_{KL} in Eq. (32) that measures per-sample class-prediction confidence relative to the marginal.

For the quantitative evaluation below, we compute IS and FID using the standard Inception v3 network pretrained on ImageNet [18,44,46]. We emphasize that this Inception v3 network is used only for the IS/FID computation, and the MNIST convolutional classifier described above, trained from scratch on MNIST, is used exclusively for the class-distribution analysis in Table II and Fig. 8. As listed in Table III, the IS, FID, and KL_{mode} metrics of generated samples under different discriminator activation functions show that the logistic and Laplace CDF activations outperform the Gaussian CDF activation in terms of image quality, diversity, and mode balance. In particular, the logistic and Laplace CDFs yield smaller KL_{mode} values (0.0022 and 0.0139, respectively) than the Gaussian CDF (0.3946). The sigmoid GAN attains the theoretical upper bound $\text{KL}_{\text{mode}} \approx 2.3026$, corresponding to complete collapse onto a single mode. This performance comparison indicates the gradient characteristics illustrated in Fig. 2(b). The logistic and Laplace CDFs have bounded Mills ratios and therefore yield smoothly bounded generator gradients with respect to the input. In contrast, as illustrated in Fig. 6(a), the unbounded Mills ratio of the Gaussian CDF leads to large gradients during the first 20 epochs, behaving similarly to the sigmoid GAN. We attribute the relatively imbalanced mode distribution of the Gaussian CDF in Table II to this transient gradient instability. Thus, the choice of the CDF form, in particular the boundedness of its Mills ratio, plays an important role in GAN model performance.

In order to investigate the learning dynamics of the noise-scale parameter σ during training, we initialized σ in the logistic CDF of Eq. (3) with two different values of $\sigma^0 = 0.2$ and $\sigma^0 = 1.0$, as illustrated in Fig. 9. After 200 epochs, the corresponding performance metrics achieve an IS of 9.1365, a FID of 0.0005, and a highly balanced KL_{mode} of 0.0022 for $\sigma^0 = 0.2$, while $\sigma^0 = 1.0$ yields an IS of 8.9121 and a FID of 0.0002 with a small $\text{KL}_{\text{mode}} = 0.0349$. As shown in Fig. 9, for the case of the initial value $\sigma^0 = 1.0$, the GAN

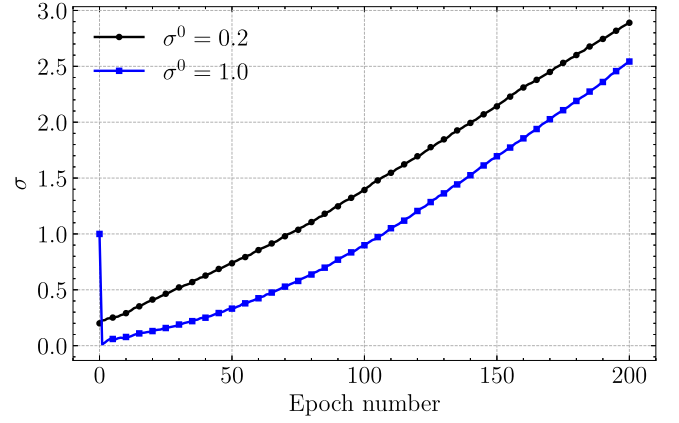


FIG. 9. Learning curves of the noise-scale parameter σ in the logistic CDF defined in Eq. (3) for two initial values of $\sigma^0 = 0.2$ and $\sigma^0 = 1.0$ on the MNIST dataset.

model rapidly reduces σ to the order of 10^{-2} during the first update step. Moreover, σ updates below unity during epochs 0–70 (or up to 100), which corresponds to the above analysis that smaller σ values enhance gradient flow and benefit the training of GAN model. However, as the training proceeds, σ gradually increases above unity, as shown in Fig. 9. This transition indicates that the GAN training process has two distinct optimization regimes. During the early exploration phase, small σ values provide gradient amplification to overcome the vanishing-gradient problem. In contrast, during the later training phase, larger σ values broaden the transition region of the CDF activation [18], thereby stabilizing the training dynamics.

C. CIFAR-10

The above experiments demonstrate that the learnable noise-scale parameter σ can improve the gradient flow in the designed GANs, where the sigmoid GAN suffers from gradient vanishing. A natural question is whether the same mechanism remains useful in the deep convolutional GAN (DCGAN) architecture that is less affected by the vanishing-gradient problem [47]. To investigate this, we further apply the proposed CDF activations to a DCGAN trained on the CIFAR-10 dataset [48], which consists of 60 000 natural color images of size 32×32 (50 000 for training and 10 000 for testing).

As shown in Fig. 10, both the generator and the discriminator in the designed DCGAN are implemented as convolutional neural networks, which exploit spatial locality and translation equivariance inherent in image data [47]. The final activation function of the discriminator employs either the sigmoid function or one of the CDF activations defined in Eqs. (3)–(5). The DCGAN is trained with a minibatch size of 128 using the Adam optimizer for both networks with a learning rate of 2×10^{-4} and 50 epochs. The latent noise vector $\mathbf{z}$ is sampled from a 100-dimensional standard Gaussian distribution. Weights are initialized from a Gaussian distribution with mean 0 and standard deviation 0.02 for convolutional layers, and mean 1 and standard deviation 0.02 for batch-normalization

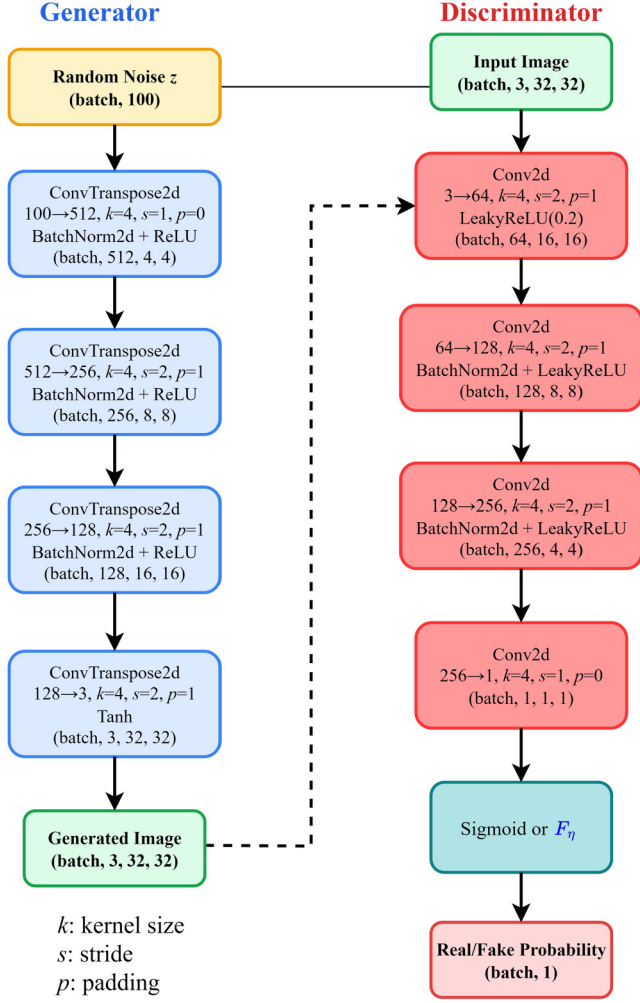


FIG. 10. Diagram of the DCGAN model. The generator maps a batch of 100-dimensional latent noise vectors z through a two-dimensional convolutional network to produce a three-channel 32×32 synthetic image. The discriminator is composed of a two-dimensional convolutional network, and its final layer employs either a sigmoid or a CDF activation F_η .

layers, while biases are initialized to zero (source code is provided in GitHub repository).

For the quantitative evaluation on CIFAR-10 [48], we compute IS and FID with the same Inception v3 network [18,44,46] described in Sec. IV B. In addition, we train a ResNet-18 classifier on CIFAR-10, achieving a test accuracy of 94.3%, which is subsequently used to classify generated images of the DCGAN. The corresponding results are summarized in Table IV, and the IS, FID, and KL_{mode} values lie within a narrow band ($\text{FID} \approx 79\text{--}90$, $\text{IS} \approx 5.2\text{--}5.8$, and $\text{KL}_{\text{mode}} \approx 0.06\text{--}0.10$). For brevity, Fig. 11 shows representative generated samples for the sigmoid and Laplace CDF activations. These comparisons indicate that, for the DCGAN architecture considered here, the convolutional structure and batch normalization already provide stable gradient flow to the generator, so that the CDF-based activations yield no improvement over the conventional sigmoid. Furthermore, we generate 1000 samples for DCGANs with different activations and predict their class labels using the trained ResNet-18

TABLE IV. Comparison of different activation functions in the final layer of the discriminator of a DCGAN trained on CIFAR-10.

Activation function	IS	FID	Mode	KL_{mode}
Sigmoid	5.7526 ± 0.2631	80.8773	10/10	0.0696
Logistic CDF	5.3784 ± 0.3085	82.6740	10/10	0.0737
Gaussian CDF	5.2168 ± 0.2575	90.8796	10/10	0.1049
Laplace CDF	5.5909 ± 0.3124	79.6230	10/10	0.0672

classifier. The resulting classwise counts, illustrated in Fig. 12, show that all four configurations cover the 10 classes without mode collapse, and the class distributions are similar across the four discriminator activations. This finding also indicates that the proposed framework is particularly effective when vanishing gradients are present, but may offer less benefit when gradient flow has already been stabilized through the network architecture itself.

Figure 13(a) shows the learning curves of the noise-scale parameter σ over 50 training epochs for the three CDF activations. The DCGANs with the logistic and Laplace CDF activations exhibit a monotonic decrease of σ from the initial value 0.8, whereas the DCGAN with the Gaussian CDF activation yields a nonmonotonic curve. The corresponding gradient norms $\|\nabla_\theta \mathcal{L}_G^{\text{ns}}\|_2$ for the sigmoid, logistic CDF, and Laplace CDF activations remain stable throughout training, while the Gaussian CDF activation exhibits a vanishing-gradient phase before epoch 25. This dynamic leads to the comparatively poor metrics of the Gaussian CDF activation in Table IV (e.g., $\text{KL}_{\text{mode}} = 0.1049$). We attribute this behavior to the Gaussian-tail property discussed in Sec. III B. In the DCGAN architecture considered here, the convolutional layers combined with batch normalization drive the discriminator to high confidence on generated samples in the early epochs, i.e., $V_\phi(G_\theta) \rightarrow +\infty$, where the ratio $r(u) = f_{\eta_0}/F_{\eta_0}$ decays rapidly and the generator gradient is suppressed. By contrast, in the MNIST setting shown in Fig. 6(a), the relevant regime is $V_\phi(G_\theta) \rightarrow -\infty$, where the same ratio is unbounded and produces large gradient norms.

D. Comparison results on MNIST

The theoretical results in Sec. III and experimental results in Sec. IV show that the proposed adaptive noise scale is

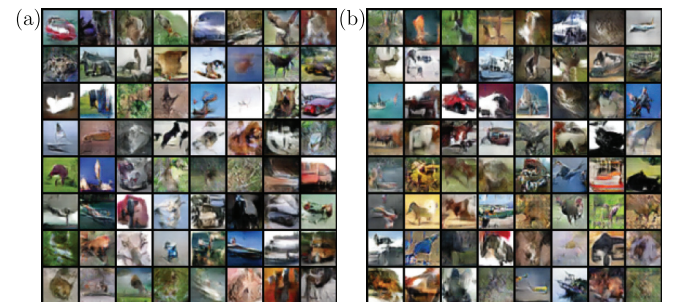


FIG. 11. Representative generated image samples from a DCGAN trained on CIFAR-10, using (a) the sigmoid function and (b) the Laplace CDF as the final activation of the discriminator.

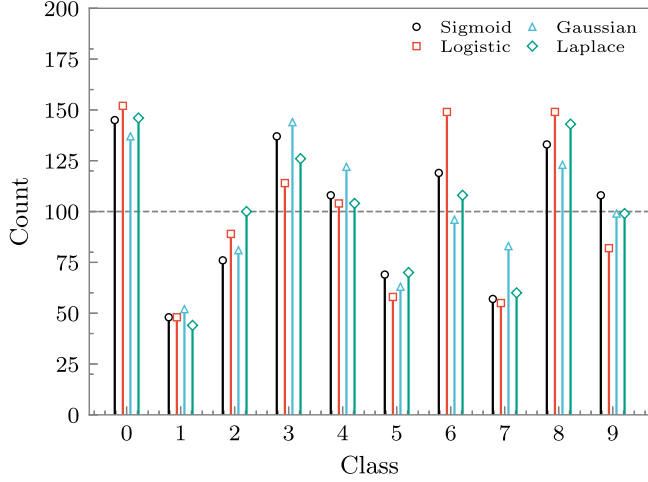


FIG. 12. Classwise counts of 1000 generated CIFAR-10 images for DCGANs with different activations. Class labels (0-9) correspond to airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, respectively, and are predicted by a ResNet-18 classifier trained on CIFAR-10. The dashed line at 100 marks the balanced reference.

effective for GANs subject to the vanishing-gradient problem. However, it is well known that noise injection or regularization during GAN training [16–18,23–25] can also yield improvements. To clarify the role of the proposed method, we perform an ablation study of the GAN model training on MNIST dataset, where the vanishing-gradient regime of the sigmoid GAN is evident (see Sec. IV B).

We consider seven variants under the identical sigmoid GAN architecture employed in Sec. IV B, with the baseline sigmoid GAN marked as case (1). Other GAN variants are (2) injecting instance noise (Gaussian noise with scale 0.1) into the discriminator input [18], (3) the one-sided label smoothing [18] (Gaussian noise with scale 0.9), (4) a discriminator dropout rate of 0.5, (5) injecting instance noise (Gaussian noise with scale 0.1) into the leaky ReLUs of the discriminator, the logistic CDF with a fixed noise scale (6) $\sigma^* = 0.2$ and (7) $\sigma^* = 0.5$ (similar to the fixed-temperature scheme of Hinton *et al.* [30]), and (8) the proposed logistic CDF with adaptive σ initialized at $\sigma^0 = 0.2$. Variants (2)–(5) represent standard noise regularization strategies for stabilizing GANs, while variants (6) and (7) mimic the effect of a static CDF-based activation. The performance metrics are summarized in

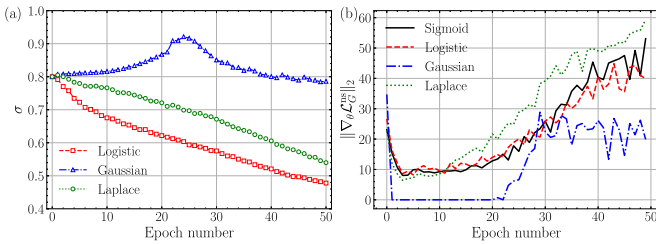


FIG. 13. (a) Learning curves of the noise-scale parameter σ in the logistic, Gaussian, and Laplace CDF discriminator activations of the DCGAN on CIFAR-10, all initialized at $\sigma^0 = 0.8$. (b) The corresponding gradient norms $\|\nabla_{\theta} \mathcal{L}_G^{ns}\|_2$ during training.

TABLE V. Comparison of the discriminator with logistic CDF and the noise regularization on MNIST, under identical sigmoid GAN architecture.

Variant	IS	FID	Modes	KL _{mode}
(1) Sigmoid	1.0211 ± 0.0032	0.0581	1/10	2.3026
(2) Instance noise	8.8886 ± 0.1506	0.0003	10/10	0.0135
(3) Label smoothing	1.0000 ± 0.0000	0.7569	1/10	2.3026
(4) Dropout 0.5	8.7331 ± 0.2884	0.0004	10/10	0.0096
(5) Activation noise	9.0241 ± 0.1624	0.0004	10/10	0.0127
(6) Logistic ($\sigma = 0.2$)	8.9174 ± 0.2306	0.0005	10/10	0.0124
(7) Logistic ($\sigma = 0.5$)	8.8793 ± 0.4073	0.0003	10/10	0.0208
(8) Logistic ($\sigma^0 = 0.2$)	9.1365 ± 0.1504	0.0005	10/10	0.0022

Table V. The generator gradient norms are shown in Fig. 14, and the classwise distributions of 1000 generated samples for the six variants without mode collapse are plotted in Fig. 15.

It is seen in Fig. 14 that, besides the sigmoid GAN, the label-smoothing variant also has the vanishing-gradient problem and the mode collapse effect occurs. Conversely, variants of instance noise, dropout, and activation noise restore mode diversity. These regularization mechanism of adding noise to the input or to the hidden activations prevent $V_{\phi}(\mathbf{x})$ and $V_{\phi}[G_{\theta}(\mathbf{z})]$ from diverging to $\pm\infty$ and thus keep the sigmoid away from its saturation regime where $\partial V_{\phi}/\partial G_{\theta} \rightarrow 0$. Based on the scale amplification mechanism of $\sigma < 1$, the fixed-scale variants also achieve better performances, e.g., $IS \approx 8.8$ – 9.0 and 10/10 mode coverage. This validates that a fixed $\sigma < 1$ can amplify the generator gradient by $1/\sigma$, which is similar to the fixed-temperature scheme of Hinton *et al.* [30]. Most importantly, the proposed logistic CDF-based GAN model attains the highest $IS = 9.1365$, and the most balanced classwise distribution with $KL_{mode} = 0.0022$, as shown in Fig. 15. We emphasize that these regularization techniques, which directly inject noise or apply dropout, are mechanistically orthogonal to our approach. The proposed CDF activation acts only on the final layer of the discriminator and analytically modifies the generator gradient in Eqs. (18)

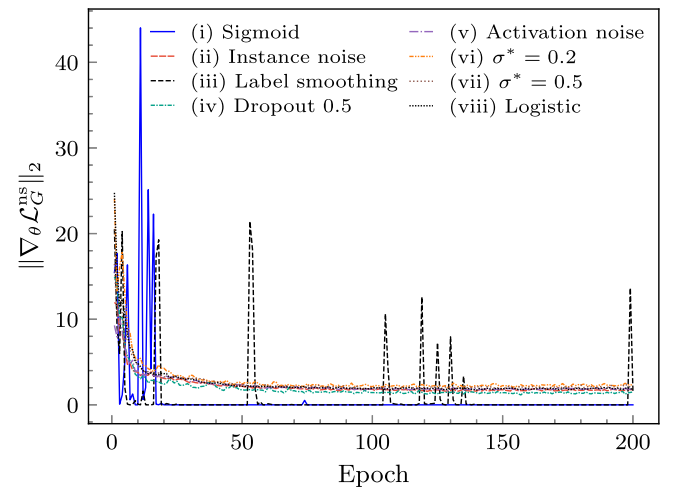


FIG. 14. Generator gradient norm $\|\nabla_{\theta} \mathcal{L}_G^{ns}\|_2$ for different variants under identical sigmoid GAN architecture.

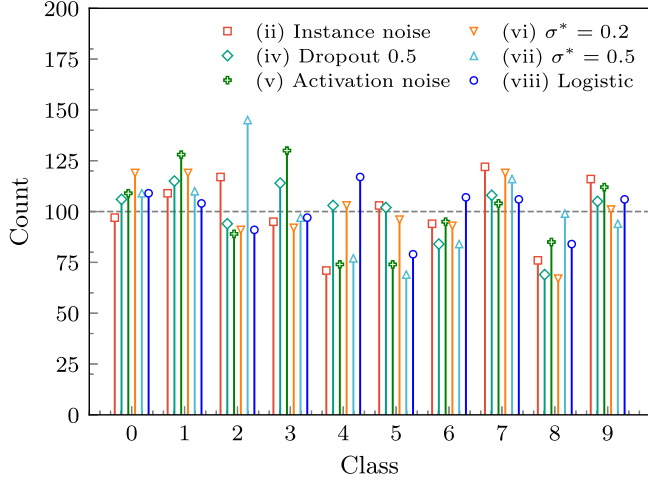


FIG. 15. Classwise counts of 1000 generated digits for six variants without mode collapse. The dashed line at 100 marks the balanced reference.

and (20), thereby making the proposed method easy to integrate into existing GAN architectures. Combining it with conventional regularization techniques [18,30] is worth investigating in future work.

V. CONCLUSION

In this work we generalize the final-layer activation of the GAN discriminator by modeling it as the CDF of a zero-mean random variable, and introduce a learnable noise-scale parameter σ to modulate the gradient dynamics. Theoretical analysis indicates that, for logistic, Gaussian, and Laplace CDFs, noise-scale values below unity can provide more informative gradient signals to the generator when the discriminator operates near its optimum, rather than definitively resolving vanishing gradients across all regimes. Experiments on a two-dimensional Gaussian mixture and MNIST demonstrate that the proposed CDF-based activation constitutes an effective and practical mechanism for stabilizing GAN training. However, the experiments on CIFAR-10 with a DCGAN architecture show that the benefit diminishes when batch normalization and convolutional layers already stabilize gradient flow. This indicates that the proposed framework is most useful in regimes where vanishing gradients are the primary limitation of training.

An ablation study on MNIST further compares the proposed adaptive CDF activation with conventional regularization strategies, including instance noise, one-sided label smoothing, increased dropout, and activation noise, as well as with the CDF with fixed noise scales analogous to temperature scaling. The results show that the proposed method achieves superior performance metrics while requiring only minimal modification of the discriminator, namely, tuning the final-layer activation. Since the proposed CDF activation only replaces the output activation of discriminators, it is mechanistically orthogonal to noise regularization [18,30] and

Lipschitz methods [26–29]. The combination of these strategies remain a promising direction for future work.

We emphasize that the effectiveness of the framework depends on both the choice of CDF family and the dataset. Our analysis indicates that CDFs with bounded reversed hazard rates in the negative tail, such as the logistic and Laplace CDFs, are better suited to this formulation than the Gaussian CDF. These observations motivate the design of CDF families for specific data distributions as a direction for future work. In addition, the adaptive role of the noise-scale parameter can be interpreted as exhibiting two regimes: gradient amplification at small σ and output smoothing at large σ . Both effects might coexist within the proposed GAN governed by the training dynamics of the model. A more unified characterization of this interplay, as well as its potential connection to stochastic-resonance phenomena in neural systems, remains an open problem.

Finally, our framework opens several promising directions for future work. Coupling the proposed CDF-based activations with GAN variants such as Style-GAN [23], which uses explicit noise injection at an optimal intensity, may clarify the interplay between implicit and explicit noise-based stabilization strategies. We also note that the CDF-based activations in Eq. (2) and one-sided linear units (e.g., the input multiplied by the logistic or Gaussian CDF, as in SiLU and GELU [8,14]) constitute a class of activation functions with an implicit noise-injection interpretation through the noise-scale parameter. This approach may also extend to other generative frameworks. For instance, the proposed CDF activation can be incorporated into the final layer of the discriminator in hybrid variational autoencoders [49], yielding comparable benefits in gradient propagation and training stability. For diffusion models [50], the bounded CDF parametrization is not immediately applicable, since score networks predict unbounded score functions and are trained via score matching rather than a binary classification objective. A natural extension is the use of one-sided linear units of the form $u F_{\eta_0}(u/\sigma)$ (e.g., SiLU and GELU [8,14]), which are unbounded and contain the learnable noise scale σ . Incorporating these activations in score networks achieves the unbounded output required by score matching while utilizing the same learnable noise-scale mechanism analyzed in this work. A deep investigation of this adaptive, unbounded activation in diffusion models remains an open direction for future research.

ACKNOWLEDGMENT

D.A. gratefully acknowledges support from the Australian Research Council (DP240103404 and FL240100217).

DATA AVAILABILITY

The PyTorch source code is publicly available at [42]. The MNIST and CIFAR-10 datasets used in this study are publicly available in Refs. [43,48].

- [1] C. M. Bishop, Training with noise is equivalent to Tikhonov regularization, *Neural Comput.* **7**, 108 (1995).
- [2] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**, 1929 (2014).
- [3] M. Ferianc, O. Bohdal, T. Hospedales, and M. Rodrigues, Impact of noise on calibration and generalisation of neural networks, in *The Fortieth International Conference on Machine Learning, Workshop on Spurious Correlations, Invariance and Stability, Honolulu, Hawaii*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (ICML, 2023), pp. 1–7.
- [4] S. Bai, F. Duan, F. Chapeau-Blondeau, and D. Abbott, Generalization of stochastic-resonance-based threshold networks with Tikhonov regularization, *Phys. Rev. E* **106**, L012101 (2022).
- [5] R. Benzi, A. Sutera, and A. Vulpiani, The mechanism of stochastic resonance, *J. Phys. A: Math. Gen.* **14**, L453 (1981).
- [6] N. G. Stocks, Suprathreshold stochastic resonance in multilevel threshold systems, *Phys. Rev. Lett.* **84**, 2310 (2000).
- [7] Z. Liao, Z. Shi, M. S. Sarker, and H. Tabata, Robust QRS detection based on simulated degenerate optical parametric oscillator-assisted neural network, *Heliyon* **10**, e28903 (2024).
- [8] F. Duan, F. Chapeau-Blondeau, and D. Abbott, Optimized injection of noise in activation functions to improve generalization of neural networks, *Chaos Solitons Fractals* **178**, 114363 (2024).
- [9] B. Kosko, K. Audhkhasi, and O. Osoba, Noise can speed back-propagation learning and deep bidirectional pretraining, *Neural Netw.* **129**, 359 (2020).
- [10] S. Ikemoto, F. Dallalibera, and K. Hosoda, Noise-modulated neural networks as an application of stochastic resonance, *Neurocomputing* **277**, 29 (2018).
- [11] W. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**, 115 (1943).
- [12] Z. Xu, S. Lu, Y. Kang, and J. Jiang, Intelligent fault classification exploration inspired by suprathreshold stochastic resonance, *IEEE Trans. Instrum. Meas.* **74**, 1 (2025).
- [13] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, edited by J. Fürnkranz and T. Joachims (Omni Press, Haifa, Israel, 2010), pp. 807–814.
- [14] D. Hendrycks and K. Gimpel, Gaussian error linear units (GELUs), [arXiv:1606.08415](https://arxiv.org/abs/1606.08415).
- [15] Y. Ren, F. Duan, F. Chapeau-Blondeau, and D. Abbott, Self-gating stochastic-resonance-based autoencoder for unsupervised learning, *Phys. Rev. E* **110**, 014107 (2024).
- [16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (MIT Press, Cambridge, MA, 2014), pp. 2672–2680.
- [17] M. Arjovsky and L. Bottou, Towards principled methods for training generative adversarial networks, [arXiv:1701.04862](https://arxiv.org/abs/1701.04862).
- [18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, Improved techniques for training GANs, in *Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain*, edited by D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, and I. Guyon (Curran Associates Inc., Red Hook, NY, 2016), pp. 2234–2242.
- [19] H. Thanh-Tung and T. Tran, Catastrophic forgetting and mode collapse in GANs, in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Glasgow, 2020), pp. 1–10.
- [20] D. Saxena and J. Cao, Generative adversarial networks (GANs): Challenges, solutions, and future directions, *ACM Comput. Surv.* **54**, 1 (2022).
- [21] M. Wiatrak, S. V. Albrecht, and A. Nystrom, Stabilizing generative adversarial networks: A survey, [arXiv:1910.00927](https://arxiv.org/abs/1910.00927).
- [22] L. Mescheder, A. Geiger, and S. Nowozin, Which training methods for GANs do actually converge? in *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, Proceedings of Machine Learning Research, Vol. 80 (PMLR, Stockholm, 2018), pp. 3481–3490.
- [23] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by L. Davis, P. Torr, and S. Zhu (IEEE/CVF, Long Beach, CA, 2019), pp. 4401–4410.
- [24] R. Feng, D. Zhao, and Z. Zha, Understanding noise injection in GANs, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, Proceedings of Machine Learning Research, Vol. 139 (PMLR, Cambridge, MA, 2021), pp. 3284–3293.
- [25] S. H. Hong and J. W. Jeong, Dynamic noise injection for facial expression recognition in-the-wild, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by J. Hoffman and J. Sivic (IEEE/CVF, Vancouver, Canada, 2023), pp. 5709–5715.
- [26] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein generative adversarial networks, in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, edited by D. Precup and Y. W. Teh, Proceedings of Machine Learning Research, Vol. 70 (PMLR, Sydney, 2017), pp. 214–223.
- [27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, Improved training of Wasserstein GANs, in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, edited by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Curran Associates, Long Beach, CA, 2017), pp. 5767–5777.
- [28] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, On convergence and stability of GANs, [arXiv:1705.07215](https://arxiv.org/abs/1705.07215).
- [29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, Spectral normalization for generative adversarial networks, in *6th International Conference on Learning Representations (ICLR)*, edited by I. Murray, M. Ranzato, and O. Vinyals (OpenReview.net, Vancouver, Canada, 2018).
- [30] G. Hinton, O. Vinyals, and J. Dean, Distilling the knowledge in a neural network, [arXiv:1503.02531](https://arxiv.org/abs/1503.02531).
- [31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, On calibration of modern neural networks, in *International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh (PMLR, Sydney, 2017), Vol. 70, pp. 1321–1330.
- [32] J. Zhang, Y. Dai, X. Yu, M. Harandi, N. Barnes, and R. Hartley, Uncertainty-aware deep calibrated salient object detection, [arXiv:2012.06020](https://arxiv.org/abs/2012.06020).

- [33] N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson, Tempered sigmoid activations for deep learning with differential privacy, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI Press, Palo Alto, CA, 2021), Vol. 35, pp. 9047–9055.
- [34] D. Page, K. Lin, and J. E. Wulsin, Do better backbones lead to better architectures? in *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, edited by A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, Lecture Notes in Computer Science Vol. 12346 (Springer, Glasgow, 2020), pp. 577–592.
- [35] H. Asatryan, H. Gottschalk, M. Lippert, and M. Rottmann, A convenient infinite dimensional framework for generative adversarial learning, *Electron. J. Stat.* **17**, 391 (2023).
- [36] V. Nagarajan and J. Z. Kolter, Gradient descent GAN optimization is locally stable, in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, edited by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Long Beach, CA, 2017), pp. 5585–5595.
- [37] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, Generalization and equilibrium in generative adversarial nets (GANs), in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, edited by D. Precup and Y. W. Teh, Proceedings of Machine Learning Research, Vol. 70 (PMLR, Sydney, 2017), pp. 224–232.
- [38] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
- [39] W. Feller, *An Introduction to Probability Theory and its Applications*, 3rd ed. (John Wiley & Sons, New York, 1968), Vol. I.
- [40] Á. Baricz, Mills’ ratio: Monotonicity patterns and functional inequalities, *J. Math. Anal. Appl.* **340**, 1362 (2008).
- [41] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations (ICLR) 2015, Poster Track*, edited by Y. Bengio and Y. LeCun (ICLR, San Diego, CA, 2015), pp. 1–13.
- [42] F. Duan, Source code, 2026, <https://github.com/resonance-dfb/learning-noise-injection>.
- [43] Y. LeCun, C. Cortes, and C. J. C. Burges, The MNIST database of handwritten digits (1998), <https://yann.lecun.org/exdb/mnist/>.
- [44] 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 *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Curran Associates, Long Beach, CA, 2017), pp. 6626–6637.
- [45] S. Barratt and R. Sharma, A note on the inception score, [arXiv:1801.01973](https://arxiv.org/abs/1801.01973).
- [46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Las Vegas, NV, 2016), pp. 2818–2826.
- [47] A. Radford, L. Metz, and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, [arXiv:1511.06434](https://arxiv.org/abs/1511.06434).
- [48] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report TR-2009, University of Toronto, 2009.
- [49] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *2nd International Conference on Learning Representations (ICLR 2014)*, edited by Y. Bengio and Y. LeCun (OpenReview.net, Banff, AB, Canada, 2014).
- [50] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, Proceedings of Machine Learning Research, Vol. 37 (PMLR, Lille, France, 2015), pp. 2256–2265.