


STOCHASTIC DIFFERENTIAL EQUATION APPROXIMATIONS OF GENERATIVE ADVERSARIAL NETWORK TRAINING AND ITS LONG-RUN BEHAVIOR

HAOYANG CAO ,* *École Polytechnique*

XIN GUO ,** *University of California, Berkeley*

Abstract

This paper analyzes the training process of generative adversarial networks (GANs) via stochastic differential equations (SDEs). It first establishes SDE approximations for the training of GANs under stochastic gradient algorithms, with precise error bound analysis. It then describes the long-run behavior of GAN training via the invariant measures of its SDE approximations under proper conditions. This work builds a theoretical foundation for GAN training and provides analytical tools to study its evolution and stability.

Keywords: Generative adversarial networks; stochastic gradient algorithm; stochastic differential equation

2020 Mathematics Subject Classification: Primary 68T07; 60H30
Secondary 60F17; 60H10

1. Introduction

Generative adversarial networks (GANs) introduced in [14] are generative models with two competing neural networks: a generator network G and a discriminator network D . The generator network G attempts to fool the discriminator network by converting random noise into sample data, while the discriminator network D tries to identify whether the input sample is fake or true.

After being introduced to the machine learning community, the popularity of GANs has grown exponentially with a wide range of applications, including high-resolution image generation [9, 29], image inpainting [45], image super-resolution [23], visual manipulation [48], text-to-image synthesis [30], video generation [40], semantic segmentation [26], and abstract reasoning diagram generation [21]; in recent years, GANs have attracted a substantial amount of attention in the financial industry for financial time series generation [36, 42, 43, 46], asset pricing [5], market simulation [6, 35], and so on. Despite the empirical success of GANs, there are well-recognized issues in the training of GANs, such as the vanishing gradient when the discriminator significantly outperforms the generator [1], and mode collapse where the

Received 12 February 2022; accepted 26 June 2023.

* Postal address: Centre de Mathématiques Appliquées, École Polytechnique, Route de Saclay, 91128, Palaiseau Cedex, France. Email: haoyang.cao@polytechnique.edu

** Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, CA 94720, USA. Email: xinguo@berkeley.edu

generator cannot recover a multi-modal distribution but only a subset of the modes; this issue is believed to be linked with the gradient exploding [32].

In response to these issues, there has been growing research interest in the theoretical understanding of GAN training. In [3] the authors proposed a novel visualization method for the GAN training process through the gradient vector field of loss functions. In a deterministic GAN training framework, [28] demonstrated that regularization improved the convergence performance of GANs. [7] and [11] analyzed a generic zero-sum minimax game including GANs, and connected the mixed Nash equilibrium of the game with the invariant measure of Langevin dynamics. In addition, various approaches have been proposed for amelioration of the aforementioned issues in GAN training, including different choices of network architectures, loss functions, and regularization. See, for instance, a comprehensive survey on these techniques in [41] and the references therein.

1.1. Our work

This paper focuses on analyzing the training process of GANs via a stochastic differential equation (SDE) approach. It first establishes SDE approximations for the training of GANs under stochastic gradient algorithms (SGAs), with precise error bound analysis. It then describes the long-run behavior of GAN training via the invariant measures of its SDE approximations under proper conditions. This work builds a theoretical foundation for GAN training and provides analytical tools to study its evolution and stability. In particular:

- The SDE approximations characterize precisely the distinction between GANs with alternating update and GANs with simultaneous update, in terms of the interaction between the generator and the discriminator.
- The drift terms in the SDEs show the direction of the parameter evolution; the diffusion terms prescribe the ratio between the batch size and the learning rate in order to modulate the fluctuations of SGAs in GAN training.
- Regularity conditions for the coefficients of the SDEs provide constraints on the growth of the loss function with respect to the model parameters, necessary for avoiding the explosive gradient encountered in the training of GANs; they also explain mathematically some well-known heuristics in GAN training, and confirm the importance of appropriate choices for network depth and of the introduction of gradient clipping and gradient penalty.
- The dissipative property of the training dynamics under the SDE form ensures the existence of the invariant measures, hence the steady states of GAN training in the long run; it underpins the practical tactic of adding a regularization term to the GAN objective to improve the stability of the training.
- Further analysis of the invariant measure for the coupled SDEs gives rise to a fluctuation–dissipation relation (FDR) for GANs. These FDRs reveal the trade-off of the loss landscape between the generator and the discriminator and can be used to schedule the learning rate.

1.2. Related work

Our analysis on the approximation and the long-run behavior of GAN training is inspired by [24] and [25]. The former established the SDE approximation for the parameter evolution

in SGAs applied to pure minimization problems (see also [18] on a similar topic); the latter surveyed the theoretical analysis of deep learning from two perspectives: propagation of chaos through neural networks and the training process of deep learning algorithms. Among other related works on the theoretical understanding of GANs, [13] reviewed the connection between GANs and the dual formulation of optimal transport problems; [27] studied the interplay between the latent distribution and generated distribution in GANs with optimal transport-based loss functions; [7] and [11] focused on the equilibrium of the minimax game and its connection with Langevin dynamics; and [4] studied the connection between GANs and mean-field games. Our focus is the GAN training process: we establish precise error bounds for the SDE approximations, study the long-run behavior of GAN training via the invariant measures of the SDE approximations, and analyze their implications for resolving various challenges in GANs.

1.3. Notation

Throughout this paper, the following notation will be adopted:

- \mathbb{R}^d denotes a d -dimensional Euclidean space, where d may vary from time to time.
- The transpose of a vector $x \in \mathbb{R}^d$ is denoted by x^\top and the transpose of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ is denoted by A^\top .
- Let \mathcal{X} be an arbitrary nonempty subset of \mathbb{R}^d ; the set of k times continuously differentiable functions over some domain \mathcal{X} is denoted by $\mathcal{C}^k(\mathcal{X})$ for any nonnegative integer k . In particular, when $k = 0$, $\mathcal{C}^0(\mathcal{X}) = \mathcal{C}(\mathcal{X})$ denotes the set of continuous functions.
- Let $J = (J_1, \dots, J_d)$ be a d -tuple multi-index of order $|J| = \sum_{i=1}^d J_i$, where $J_i \geq 0$ for all $i = 1, \dots, d$; then the operator ∇^J is $\nabla^J = (\partial_1^{J_1}, \dots, \partial_d^{J_d})$.
- For $p \geq 1$, $\|\cdot\|_p$ denotes the p -norm over \mathbb{R}^d , i.e. $\|x\|_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ for any $x \in \mathbb{R}^d$; $L^p_{\text{loc}}(\mathbb{R}^d)$ denotes the set of functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\int_{\mathcal{X}} |f(x)|^p dx < \infty$ for any compact subset $\mathcal{X} \subset \mathbb{R}^d$.
- Let J be a d -tuple multi-index of order $|J|$. For a function $f \in L^1_{\text{loc}}(\mathbb{R}^d)$, its J th-weak derivative $D^J f \in L^1_{\text{loc}}(\mathbb{R}^d)$ is a function such that, for any smooth and compactly supported test function g , $\int_{\mathbb{R}^d} D^J f(x) g(x) dx = (-1)^{|J|} \int_{\mathbb{R}^d} f(x) \nabla^J g(x) dx$. The Sobolev space $W^{k,p}_{\text{loc}}(\mathbb{R}^d)$ is a set of functions f on \mathbb{R}^d such that, for any d -tuple multi-index J with $|J| \leq k$, $D^J f \in L^p_{\text{loc}}(\mathbb{R}^d)$.
- Fix an arbitrary $\alpha \in \mathbb{N}^+$. $G^\alpha(\mathbb{R}^d)$ denotes a subspace of $\mathcal{C}^\alpha(\mathbb{R}^d; \mathbb{R})$ where, for any $g \in G^\alpha(\mathbb{R}^d)$ and any multi-index J with $|J| = \sum_{i=1}^d J_i \leq \alpha$, there exist $k_1, k_2 \in \mathbb{N}$ such that $\nabla^J g(x) \leq k_1 (1 + \|x\|_2^{2k_2})$ for all $x \in \mathbb{R}^d$. If g is a parametrized function g_β , then $g_\beta \in G^\alpha(\mathbb{R}^d)$ indicates that the choices of constants k_1 and k_2 are uniform over all possible β s.
- Fix an arbitrary $\alpha \in \mathbb{N}^+$. $G^\alpha_w(\mathbb{R}^d)$ denotes a subspace of $W^{\alpha,1}_{\text{loc}}(\mathbb{R}^d)$ where, for any $g \in G^\alpha_w(\mathbb{R}^d)$ and any multi-index J with $|J| = \sum_{i=1}^d J_i \leq \alpha$, there exist $k_1, k_2 \in \mathbb{N}$ such that $D^J g(x) \leq k_1 (1 + \|x\|_2^{2k_2})$ for almost all $x \in \mathbb{R}^d$. If g is a parametrized function g_β , then $g_\beta \in G^\alpha_w(\mathbb{R}^d)$ indicates that the choices of constants k_1 and k_2 are uniform over all possible β s.

2. GAN training

In this section we provide the mathematical setup for GAN training.

2.1. GAN training: Minimax versus maximin

GANs fall into the category of generative models to approximate an unknown probability distribution \mathbb{P}_r . GANs are minimax games between two competing neural networks, the generator G and the discriminator D . The neural network for the generator G maps a latent random variable Z with a known distribution \mathbb{P}_z into the sample space to mimic the true distribution \mathbb{P}_r . Meanwhile, the other neural network for the discriminator D will assign a score between 0 and 1 to an input sample, either a generated sample or a true one. A higher score from the discriminator D indicates that the sample is more likely to be from the true distribution.

Formally, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ be a filtered probability space. Let a measurable space $\mathcal{X} \subset \mathbb{R}^{d_x}$ be the sample space of dimension $d_x \in \mathbb{N}$. Let an \mathcal{X} -valued random variable X denote the random sample, where $X: \Omega \rightarrow \mathcal{X}$ is a measurable function. The unknown probability distribution \mathbb{P}_r is defined as $\mathbb{P}_r = \text{Law}(X)$ such that $\mathbb{P}_r(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$ for any measurable set $A \subset \mathcal{X}$. Similarly, let a measurable space $\mathcal{Z} \subset \mathbb{R}^{d_z}$ be the latent space of dimension $d_z \in \mathbb{N}$. Let a \mathcal{Z} -valued random variable Z denote the latent variable where $Z: \Omega \rightarrow \mathcal{Z}$. The prior distribution \mathbb{P}_z is given by $\mathbb{P}_z = \text{Law}(Z)$ such that $\mathbb{P}_z(Z \in B) = \mathbb{P}(\{\omega \in \Omega : Z(\omega) \in B\})$ for any measurable $B \subset \mathcal{Z}$. Moreover, X and Z are independent, i.e. $\mathbb{P}(\{\omega : X(\omega) \in A, Z(\omega) \in B\}) = \mathbb{P}_r(X \in A)\mathbb{P}_z(Z \in B)$ for any measurable sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Z}$.

In the vanilla GAN framework proposed by [14], the loss function with respect to G and D is given by $L(G, D) = \mathbb{E}_{X \sim \mathbb{P}_r} \log D(X) + \mathbb{E}_{Z \sim \mathbb{P}_z} [\log(1 - D(G(Z)))]$, and the objective is given by a *minimax* problem, $\min_G \max_D L(G, D)$. Under a given G , the concavity of $L(G, D)$ with respect to D follows from the concavity of the functions $\log x$ and $\log(1 - x)$; under a given D , the convexity of $L(G, D)$ with respect to G follows from the linearity of expectation and the pushforward measure $G\#\mathbb{P}_z = \text{Law}(G(Z))$. Therefore, the training loss in vanilla GANs is indeed convex in G and concave in D . In the practical training stage, both G and D become parametrized neural networks G_θ and D_ω , and therefore the working loss function is indeed with respect to the parameter (θ, ω) ,

$$\hat{L}(\theta, \omega) = \mathbb{E}_{X \sim \mathbb{P}_r} \log D_\omega(X) + \mathbb{E}_{Z \sim \mathbb{P}_z} [\log(1 - D_\omega(G_\theta(Z)))].$$

According to the training scheme proposed by [14], in each iteration, ω is updated first followed by the update of θ . This precisely corresponds to the *minimax* formulation of the objective, $\min_\theta \max_\omega \hat{L}(\theta, \omega)$. However, in the practice training stage of GANs, there might be an interchange of training orders between the generator and the discriminator. We should be careful as the interchange implicitly modifies the objective into a *maximin* problem, $\max_\omega \min_\theta \hat{L}(\theta, \omega)$, and hence raises the question of whether these two objectives are equivalent. This question is closely related to the notion of Nash equilibrium in a two-player zero-sum game. According to the original GAN framework, the solution should provide an *upper value* to the corresponding two-player zero-sum game between the generator and the discriminator, i.e. an upper bound for the game value. As pointed out by Sion's theorem (see [34, 39]), a sufficient condition to guarantee equivalence between the two training orders is that the loss function \hat{L} is convex in θ and concave in ω . Though we have seen that the loss function L with respect to G and D satisfies this condition, it is not necessarily true for $\hat{L}(\theta, \omega)$. In fact, [47] points out that these conditions are usually not satisfied with respect to generator and discriminator parameters in common GAN models, and this lack of convexity and/or concavity does create challenges in

the training of GANs. Such challenges motivate us to take a closer look at the evolution of parameters in the training of GANs using mathematical tools. In the following analysis, we will strictly follow the *minimax* formulation and its corresponding training order.

2.2. SGA for GAN training

Typically, GANs are trained through a stochastic gradient algorithm (SGA). An SGA is applied to a class of optimization problems whose loss function $\Phi(\gamma)$ with respect to the model parameter vector γ can be written as $\Phi(\gamma) = \mathbb{E}_{\mathcal{I}}[\Phi_{\mathcal{I}}(\gamma)]$, where a random variable \mathcal{I} takes values in the index set \mathbb{I} of the data points and, for any $i \in \mathbb{I}$, $\Phi_i(\gamma)$ denotes the loss evaluated at the data point with index i .

Suppose the objective is to minimize $\Phi(\gamma)$ over γ . Applying gradient descent with learning rate $\eta > 0$, at an iteration $k, k = 0, 1, 2, \dots$, the parameter vector is updated by $\gamma_{k+1} = \gamma_k - \eta \nabla \Phi(\gamma_k)$. By the linearity of differentiability and expectation, this update can be written as $\gamma_{k+1} = \gamma_k - \eta \mathbb{E}_{\mathcal{I}}[\nabla \Phi_{\mathcal{I}}(\gamma_k)]$. Under suitable conditions, $\mathbb{E}_{\mathcal{I}}[\nabla \Phi_{\mathcal{I}}(\gamma_k)]$ can be estimated by sample mean

$$\hat{\mathbb{E}}_{\mathcal{B}}[\nabla \Phi_{\mathcal{I}}(\gamma)] = \frac{\sum_{k=1}^B \nabla \Phi_{I_k}(\gamma)}{B},$$

where $\mathcal{B} = \{I_1, \dots, I_B\}$ is a collection of indices with $I_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{I}$, called a minibatch, and $B \ll |\mathbb{I}|$.

Under an SGA, the uncertainty in sampling \mathcal{B} propagates through the training process, making it a stochastic process rather than a deterministic one. This stochasticity motivates us to study a continuous-time approximation for GAN training in the form of SDEs, as will be seen in (SML-SDE) and (ALT-SDE). (See also the connection between stochastic gradient descent and Markov chains in [10]).

Consider GAN training performed on a data set $\mathcal{D} = \{(z_i, x_j)\}_{1 \leq i \leq N, 1 \leq j \leq M}$, where $\{z_i\}_{i=1}^N$ are sampled from \mathbb{P}_z and $\{x_j\}_{j=1}^M$ are real image data following the unknown distribution \mathbb{P}_r . Let $G_{\theta} : \mathcal{Z} \rightarrow \mathcal{X}$ denote the generator parametrized by the neural network with parameter $\theta \in \mathbb{R}^{d_{\theta}}$ of dimension $d_{\theta} \in \mathbb{N}$, and let $D_{\omega} : \mathcal{X} \rightarrow \mathbb{R}^+$ denote the discriminator parametrized by the other neural network with parameter $\omega \in \mathbb{R}^{d_{\omega}}$ of dimension $d_{\omega} \in \mathbb{N}$, where \mathbb{R}^+ denotes the set of nonnegative real numbers. Then the objective of the GAN is to solve the minimax problem

$$\min_{\theta} \max_{\omega} \Phi(\theta, \omega) \tag{1}$$

for some cost function Φ , with Φ of the form

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M J(D_{\omega}(x_j), D_{\omega}(G_{\theta}(z_i)))}{N \cdot M}.$$

For instance, Φ in the vanilla GAN model [14] is given by

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M \log D_{\omega}(x_j) + \log (1 - D_{\omega}(G_{\theta}(z_i)))}{N \cdot M},$$

while Φ in a Wasserstein GAN [2] takes the form

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M D_{\omega}(x_j) - D_{\omega}(G_{\theta}(z_i))}{N \cdot M}.$$

Here, the full gradients of Φ with respect to θ and ω are estimated over a minibatch \mathcal{B} of batch size B . One way of sampling \mathcal{B} is to choose B samples out of a total of $N \cdot M$ samples without

putting back; another is to take B independent and identically distributed (i.i.d.) samples. The analyses for both cases are similar; here we adopt the second sampling scheme.

More precisely, let $\mathcal{B} = \{(z_{j_k}, x_{j_k})\}_{k=1}^B$ be i.i.d. samples from \mathcal{D} . Let g_θ and g_ω be the full gradients of Φ with respect to θ and ω such that

$$\begin{aligned} g_\theta(\theta, \omega) &= \nabla_\theta \Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M g_\theta^{i,j}(\theta, \omega)}{N \cdot M}, \\ g_\omega(\theta, \omega) &= \nabla_\omega \Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M g_\omega^{i,j}(\theta, \omega)}{N \cdot M}. \end{aligned} \tag{2}$$

Here, $g_\theta^{i,j}$ and $g_\omega^{i,j}$ denote $\nabla_\theta J(D_\omega(x_j), D_\omega(G_\theta(z_i)))$ and $\nabla_\omega J(D_\omega(x_j), D_\omega(G_\theta(z_i)))$, respectively, with differential operators defined as $\nabla_\theta := (\partial_{\theta_1} \cdots \partial_{\theta_{d_\theta}})^\top$ and $\nabla_\omega := (\partial_{\omega_1} \cdots \partial_{\omega_{d_\omega}})^\top$. Then, the estimated gradients for g_θ and g_ω corresponding to the minibatch \mathcal{B} are

$$g_\theta^{\mathcal{B}}(\theta, \omega) = \frac{\sum_{k=1}^B g_\theta^{I_k, J_k}(\theta, \omega)}{B}, \quad g_\omega^{\mathcal{B}}(\theta, \omega) = \frac{\sum_{k=1}^B g_\omega^{I_k, J_k}(\theta, \omega)}{B}.$$

Moreover, let $\eta_t^\theta > 0$ and $\eta_t^\omega > 0$ be the learning rates at iteration $t = 0, 1, 2, \dots$ for θ and ω respectively; then, solving the minimax problem (1) with SGA under *alternating parameter update* implies descent of θ along g_θ and ascent of ω along g_ω at each iteration, i.e.,

$$\begin{cases} \omega_{t+1} = \omega_t + \eta_t^\omega g_\omega^{\mathcal{B}}(\theta_t, \omega_t), \\ \theta_{t+1} = \theta_t - \eta_t^\theta g_\theta^{\mathcal{B}}(\theta_t, \omega_{t+1}). \end{cases}$$

Furthermore, within each iteration, the minibatch gradients for θ and ω are calculated on different batches. In order to emphasize this difference, we use $\bar{\mathcal{B}}$ to represent the minibatch for θ and \mathcal{B} for that of ω , with $\bar{\mathcal{B}} \stackrel{\text{i.i.d.}}{\sim} \mathcal{B}$. That is,

$$\begin{cases} \omega_{t+1} = \omega_t + \eta_t^\omega g_\omega^{\mathcal{B}}(\theta_t, \omega_t), \\ \theta_{t+1} = \theta_t - \eta_t^\theta g_\theta^{\bar{\mathcal{B}}}(\theta_t, \omega_{t+1}). \end{cases} \tag{ALT}$$

Some practical training of GANs uses *simultaneous parameter update* between the discriminator and the generator, corresponding to the similar yet subtly different form

$$\begin{cases} \omega_{t+1} = \omega_t + \eta_t^\omega g_\omega^{\mathcal{B}}(\theta_t, \omega_t), \\ \theta_{t+1} = \theta_t - \eta_t^\theta g_\theta^{\mathcal{B}}(\theta_t, \omega_t). \end{cases} \tag{SML}$$

For ease of exposition, we will assume a constant learning rate $\eta_t^\theta = \eta_t^\omega = \eta$ throughout the paper, with η viewed as the time interval between two consecutive parameter updates.

3. Approximation and error bound analysis of GAN training

The randomness in sampling \mathcal{B} (and $\bar{\mathcal{B}}$) brings stochasticity to the GAN training process prescribed by (ALT) and (SML). In this section, we establish their continuous-time approximations and error bounds, where the approximations are in the form of coupled SDEs.

3.1. Approximation

To get an intuition of how the exact expression of SDEs emerges, let us start by some basic properties embedded in the training process. First, let $I : \Omega \rightarrow \{1, \dots, N\}$ and $J : \Omega \rightarrow \{1, \dots, M\}$ denote random indices independently and uniformly distributed respectively; then, according to the definitions of g_θ and g_ω in (2), we have $\mathbb{E}[g_\theta^{I,J}(\theta, \omega)] = g_\theta(\theta, \omega)$ and $\mathbb{E}[g_\omega^{I,J}(\theta, \omega)] = g_\omega(\theta, \omega)$. Denote the correspondence covariance matrices as

$$\Sigma_\theta(\theta, \omega) = \frac{\sum_i \sum_j [g_\theta^{ij}(\theta, \omega) - g_\theta(\theta, \omega)][g_\theta^{ij}(\theta, \omega) - g_\theta(\theta, \omega)]^\top}{N \cdot M},$$

$$\Sigma_\omega(\theta, \omega) = \frac{\sum_i \sum_j [g_\omega^{ij}(\theta, \omega) - g_\omega(\theta, \omega)][g_\omega^{ij}(\theta, \omega) - g_\omega(\theta, \omega)]^\top}{N \cdot M},$$

since the (I_k, J_k) in \mathcal{B} are i.i.d. copies of (I, J) ; then,

$$\mathbb{E}_{\mathcal{B}}[g_\theta^{\mathcal{B}}(\theta, \omega)] = \mathbb{E}\left[\frac{\sum_{k=1}^B g_\theta^{I_k, J_k}(\theta, \omega)}{B}\right] = g_\theta(\theta, \omega),$$

$$\mathbb{E}_{\mathcal{B}}[g_\omega^{\mathcal{B}}(\theta, \omega)] = \mathbb{E}\left[\frac{\sum_{k=1}^B g_\omega^{I_k, J_k}(\theta, \omega)}{B}\right] = g_\omega(\theta, \omega),$$

$$\text{Var}_{\mathcal{B}}(g_\theta^{\mathcal{B}}(\theta, \omega)) = \text{Var}_{\mathcal{B}}\left(\frac{\sum_{k=1}^B g_\theta^{I_k, J_k}(\theta, \omega)}{B}\right) = \frac{1}{B} \Sigma_\theta(\theta, \omega),$$

$$\text{Var}_{\mathcal{B}}(g_\omega^{\mathcal{B}}(\theta, \omega)) = \text{Var}_{\mathcal{B}}\left(\frac{\sum_{k=1}^B g_\omega^{I_k, J_k}(\theta, \omega)}{B}\right) = \frac{1}{B} \Sigma_\omega(\theta, \omega).$$

As the batch size B gets sufficiently large, the classical central limit theorem leads to the following approximation of (ALT):

$$\begin{cases} \omega_{t+1} = \omega_t + \eta g_\omega^{\mathcal{B}}(\theta_t, \omega_t) \approx \omega_t + \eta g_\omega(\theta_t, \omega_t) + \frac{\eta}{\sqrt{B}} \Sigma_\omega^{1/2}(\theta_t, \omega_t) Z_t^1, \\ \theta_{t+1} = \theta_t - \eta g_\theta^{\mathcal{B}}(\theta_t, \omega_{t+1}) \approx \theta_t - \eta g_\theta(\theta_t, \omega_{t+1}) + \frac{\eta}{\sqrt{B}} \Sigma_\theta^{1/2}(\theta_t, \omega_{t+1}) Z_t^2, \end{cases}$$

with independent Gaussian random variables $Z_t^1 \sim N(0, 1 \cdot I_{d_\omega})$ and $Z_t^2 \sim N(0, 1 \cdot I_{d_\theta})$, $t = 0, 1, 2, \dots$. Here, the scalar 1 specifies the time increment $1 = \Delta t = (t + 1) - t$.

Write $t + 1 = t + \Delta t$. On one hand, assuming the continuity of the process $\{\omega_t\}_t$ with respect to time t and sending Δt to 0, one intuitive approximation can be easily derived in the following form:

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = \begin{pmatrix} -g_\theta(\Theta_t, \mathcal{W}_t) \\ g_\omega(\Theta_t, \mathcal{W}_t) \end{pmatrix} dt + \sqrt{2\beta^{-1}} \begin{pmatrix} \Sigma_\theta(\Theta_t, \mathcal{W}_t)^{1/2} & 0 \\ 0 & \Sigma_\omega(\Theta_t, \mathcal{W}_t)^{1/2} \end{pmatrix} dW_t, \quad (3)$$

with $\beta = 2B/\eta$ and $\{W_t\}_{t \geq 0}$ being a standard $(d_\theta + d_\omega)$ -dimensional Brownian motion supported by the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$. Let $\{\mathcal{F}_t^W\}_{t \geq 0}$ denote the natural filtration generated by $\{W_t\}_{t \geq 0}$. As a continuous-time approximation for GAN training, SDEs in this rather intuitive form are adopted without justification in some earlier works such as [7] and [11]. Later we will show that (3) is in fact an approximation for GAN training under the simulations update scheme (SML).

On the other hand, the game nature in GANs is demonstrated through the interactions between the generator and the discriminator during the training process; more specifically, the appearance of ω_{t+1} at the update of θ as in (ALT). However, the widely adopted coupled processes (3) do not capture such interactions. One possible approximation for the GAN training process of (ALT) would be

$$\begin{aligned}
 d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = & \begin{bmatrix} -g_\theta(\Theta_t, \mathcal{W}_t) \\ g_\omega(\Theta_t, \mathcal{W}_t) \end{bmatrix} \\
 & + \frac{\eta}{2} \begin{pmatrix} \nabla_\theta g_\theta(\Theta_t, \mathcal{W}_t) & -\nabla_\omega g_\theta(\Theta_t, \mathcal{W}_t) \\ -\nabla_\theta g_\omega(\Theta_t, \mathcal{W}_t) & -\nabla_\omega g_\omega(\Theta_t, \mathcal{W}_t) \end{pmatrix} \begin{pmatrix} -g_\theta(\Theta_t, \mathcal{W}_t) \\ g_\omega(\Theta_t, \mathcal{W}_t) \end{pmatrix} dt \\
 & + \sqrt{2\beta^{-1}} \begin{pmatrix} \Sigma_\theta(\Theta_t, \mathcal{W}_t)^{1/2} & 0 \\ 0 & \Sigma_\omega(\Theta_t, \mathcal{W}_t)^{1/2} \end{pmatrix} dW_t. \tag{4}
 \end{aligned}$$

Equations (3) and (4) can be written in the more compact forms

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = b_0(\Theta_t, \mathcal{W}_t) dt + \sigma(\Theta_t, \mathcal{W}_t) dW_t, \tag{SML-SDE}$$

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = b(\Theta_t, \mathcal{W}_t) dt + \sigma(\Theta_t, \mathcal{W}_t) dW_t, \tag{ALT-SDE}$$

where the drift $b(\theta, \omega) = b_0(\theta, \omega) + \eta b_1(\theta, \omega)$, with

$$b_0(\theta, \omega) = \begin{pmatrix} -g_\theta(\theta, \omega) \\ g_\omega(\theta, \omega) \end{pmatrix}, \tag{5}$$

$$\begin{aligned}
 b_1(\theta, \omega) &= \frac{1}{2} \begin{pmatrix} \nabla_\theta g_\theta(\theta, \omega) & -\nabla_\omega g_\theta(\theta, \omega) \\ -\nabla_\theta g_\omega(\theta, \omega) & -\nabla_\omega g_\omega(\theta, \omega) \end{pmatrix} \begin{pmatrix} -g_\theta(\theta, \omega) \\ g_\omega(\theta, \omega) \end{pmatrix} \\
 &= -\frac{1}{2} \nabla b_0(\theta, \omega) b_0(\theta, \omega) - \begin{pmatrix} \nabla_\omega g_\theta(\theta, \omega) g_\omega(\theta, \omega) \\ 0 \end{pmatrix}, \tag{6}
 \end{aligned}$$

and the volatility $\sigma(\theta, \omega)$ is given by

$$\sigma(\theta, \omega) = \sqrt{2\beta^{-1}} \begin{pmatrix} \Sigma_\theta(\Theta_t, \mathcal{W}_t)^{1/2} & 0 \\ 0 & \Sigma_\omega(\Theta_t, \mathcal{W}_t)^{1/2} \end{pmatrix}. \tag{7}$$

The drift terms in the SDEs, i.e. b_0 in (SML-SDE) and b in (ALT-SDE), show the direction of the parameters' evolution; the diffusion terms σ represent the fluctuations of the learning curves for these parameters. Moreover, the form of the SDEs prescribes β , the ratio between the batch size and the learning rate, in order to modulate the fluctuations of SGAs in GAN training. Even though both (SML-SDE) and (ALT-SDE) are adapted to $\{\mathcal{F}_t^W\}_{t \geq 0}$, the term

$$-\frac{\eta}{2} \begin{pmatrix} \nabla_\omega g_\theta(\theta, \omega) g_\omega(\theta, \omega) \\ 0 \end{pmatrix}$$

in (ALT-SDE) highlights the interaction between the generator and the discriminator in the GAN training process; see Remark 1.

3.2. Error bound for the SDE approximation

We will show that these coupled SDEs are indeed the continuous-time approximations of GAN training processes, with the following error bound analysis. Here the approximations are under the notion of weak approximation as in [24]. More precisely, Theorems 1 and 2 provide conditions under which the evolution of parameters in GANs are within a reasonable distance from the SDE approximation.

Theorem 1. Fix an arbitrary time horizon $\mathcal{T} > 0$, and take the learning rate $\eta \in (0, 1 \wedge \mathcal{T})$ and the number of iterations $\bar{N} = \lfloor \mathcal{T}/\eta \rfloor$. Suppose that

- 1(a) $g_\omega^{i,j}$ is twice continuously differentiable, and $g_\theta^{i,j}$ and $g_\omega^{i,j}$ are Lipschitz, for any $i = 1, \dots, N$ and $j = 1, \dots, M$;
- 1(b) Φ is of $\mathcal{C}^3(\mathbb{R}^{d_\theta+d_\omega})$ and $\Phi \in G_W^4(\mathbb{R}^{d_\theta+d_\omega})$;
- 1(c) $(\nabla_\theta g_\theta)g_\theta, (\nabla_\omega g_\omega)g_\omega, (\nabla_\theta g_\omega)g_\theta$, and $(\nabla_\omega g_\omega)g_\omega$ are all Lipschitz.

Then, $(\Theta_{t_\eta}, \mathcal{W}_{t_\eta})$ as in (ALT-SDE) is a weak approximation of (θ_t, ω_t) as in (ALT) of order 2, i.e. given any initialization $\theta_0 = \theta$ and $\omega_0 = \omega$, for any test function $f \in G^3(\mathbb{R}^{d_\theta+d_\omega})$, we have the estimate

$$\max_{t=1, \dots, \bar{N}} |\mathbb{E}f(\theta_t, \omega_t) - \mathbb{E}f(\Theta_{t_\eta}, \mathcal{W}_{t_\eta})| \leq C\eta^2 \tag{8}$$

for some constant $C \geq 0$; this constant C is independent of the learning rate η but is dependent on the time horizon \mathcal{T} .

Theorem 2. Fix an arbitrary time horizon $\mathcal{T} > 0$, and take the learning rate $\eta \in (0, 1 \wedge \mathcal{T})$ and the number of iterations $\bar{N} = \lfloor \mathcal{T}/\eta \rfloor$. Suppose

- 2(a) $\Phi(\theta, \omega)$ is continuously differentiable and $\Phi \in G_W^{3,1}(\mathbb{R}^{d_\theta+d_\omega})$;
- 2(b) $g_\theta^{i,j}$ and $g_\omega^{i,j}$ are Lipschitz for any $i = 1, \dots, N$ and $j = 1, \dots, M$.

Then, $(\Theta_{t_\eta}, \mathcal{W}_{t_\eta})$ as in (SML-SDE) is a weak approximation of (θ_t, ω_t) as in (SML) of order 1, i.e. given any initialization $\theta_0 = \theta$ and $\omega_0 = \omega$, for any test function $f \in G^2(\mathbb{R}^{d_\theta+d_\omega})$, we have the estimate

$$\max_{t=1, \dots, \bar{N}} |\mathbb{E}f(\theta_t, \omega_t) - \mathbb{E}f(\Theta_{t_\eta}, \mathcal{W}_{t_\eta})| \leq C\eta$$

for some constant $C \geq 0$; this constant C is independent of the learning rate η but is dependent on the time horizon \mathcal{T} .

Theorems 1 and 2 provide SDE approximations for GAN training in practice when we have finite training samples and training iterations, i.e. N, M , and $\bar{N} = \lfloor \mathcal{T}/\eta \rfloor$ being finite and fixed; they also provide error bounds for such approximations, in particular:

$$|\mathbb{E}[f(\theta_{\bar{N}}, \omega_{\bar{N}}) - f(\Theta_{\eta\bar{N}}, \mathcal{W}_{\eta\bar{N}})]| \leq C_1(\mathcal{T})\rho_1(\eta) \quad \text{for all } f \in G^3(\mathbb{R}^{d_\theta+d_\omega}), \tag{9}$$

where $C_1(\mathcal{T})$ is a coefficient depending on the time horizon \mathcal{T} , and $\rho_1(\eta)$ is an appropriate error term such that either $\rho_1(\eta) = \eta^2$ or $\rho_1(\eta) = \eta$. These SDE approximations will enable us to analyze the long-run behavior of GAN training in Section 4 through studying the invariant

measures of SDEs and then control the difference between the training outcome and some equilibrium of the minimax game of GANs.

Remark 1. Modifying the intuitive SDE approximation (SML-SDE) into

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = \left[b_0(\Theta_t, \mathcal{W}_t) - \frac{\eta}{2} \nabla b_0(\Theta_t, \mathcal{W}_t) b_0(\Theta_t, \mathcal{W}_t) \right] dt + \sigma(\Theta_t, \mathcal{W}_t) dW_t \quad (10)$$

and applying similar techniques to the proof of Theorem 1, we can get an $O(\eta^2)$ approximation for (SML). However, comparing (10) and (ALT-SDE), the term

$$-\frac{\eta}{2} \begin{pmatrix} \nabla_{\omega} g_{\theta}(\theta, \omega) g_{\omega}(\theta, \omega) \\ 0 \end{pmatrix}$$

still stands out, which is due to the interactions between the generator and discriminator during training. It implies that the ‘game effect’ between the generator and the discriminator has an impact on the evolution trajectories of the model parameters.

3.3. Proof of Theorem 1

In this section we provide a detailed proof of Theorem 1; the proof of Theorem 2 is a simple analogy and thus omitted. We adapt the approach of [24] to our analysis of GAN training.

3.3.1. Preliminary analysis

One-step difference Recall that under the alternating update scheme and constant learning rate η , the GAN training is given by (ALT).

Let (θ, ω) denote the initial value for (θ_0, ω_0) , and

$$\Delta = \Delta(\theta, \omega) = \begin{pmatrix} \theta_1 - \theta \\ \omega_1 - \omega \end{pmatrix} \quad (11)$$

be the one-step difference. Let $\Delta^{i,j}$ denote the tuple consisting of the i th and j th components of the one-step difference of θ and ω , with $i = 1, \dots, d_{\theta}$ and $j = 1, \dots, d_{\omega}$.

Lemma 1. Assume that $g_{\theta}^{i,j}$ is twice continuously differentiable for any $i = 1, \dots, N$ and $j = 1, \dots, M$.

The first moment is given by

$$\mathbb{E}[\Delta^{i,j}] = \eta \begin{pmatrix} -g_{\theta}(\theta, \omega)_i \\ g_{\omega}(\theta, \omega)_j \end{pmatrix} + \eta^2 \begin{pmatrix} \{-\nabla_{\omega}[g_{\theta}(\theta, \omega)_i]\}^{\top} g_{\omega}(\theta, \omega) \\ 0 \end{pmatrix} + O(\eta^3).$$

The second moment is given by

$$\mathbb{E}[\Delta^{i,j} (\Delta^{k,l})^{\top}] = \eta^2 \left[\frac{1}{B} \begin{pmatrix} \Sigma_{\theta}(\theta, \omega)_{i,k} & 0 \\ 0 & \Sigma_{\omega}(\theta, \omega)_{j,l} \end{pmatrix} + \begin{pmatrix} -g_{\theta}(\theta, \omega)_i \\ g_{\omega}(\theta, \omega)_j \end{pmatrix} \begin{pmatrix} -g_{\theta}(\theta, \omega)_k \\ g_{\omega}(\theta, \omega)_l \end{pmatrix}^{\top} \right] + O(\eta^3),$$

where $\Sigma_{\theta}(\theta, \omega)_{i,k}$ and $\Sigma_{\omega}(\theta, \omega)_{j,l}$ denote the elements at positions (i,k) and (j,l) of the matrices $\Sigma_{\theta}(\theta, \omega)$ and $\Sigma_{\omega}(\theta, \omega)$, respectively.

The third moments are all of order $O(\eta^3)$.

Proof. By a second-order Taylor expansion, we have

$$\Delta(\theta, \omega) = \eta \begin{pmatrix} -g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega) \\ g_{\omega}^{\mathcal{B}}(\theta, \omega) \end{pmatrix} + \eta^2 \begin{pmatrix} -\nabla_{\omega} g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega) g_{\omega}^{\mathcal{B}}(\theta, \omega) \\ 0 \end{pmatrix} + O(\eta^3).$$

Then,

$$\begin{aligned} \Delta^{i,j}(\theta, \omega) &= \eta \begin{pmatrix} -g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_i \\ g_{\omega}^{\mathcal{B}}(\theta, \omega)_j \end{pmatrix} + \eta^2 \begin{pmatrix} \{-\nabla_{\omega} [g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_i]\}^{\top} g_{\omega}^{\mathcal{B}}(\theta, \omega) \\ 0 \end{pmatrix} + O(\eta^3), \\ \Delta^{i,j}(\theta, \omega) [\Delta^{k,l}(\theta, \omega)]^{\top} &= \eta^2 \begin{pmatrix} g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_i g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_k - g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_i g_{\omega}^{\mathcal{B}}(\theta, \omega)_l \\ -g_{\theta}^{\tilde{\mathcal{B}}}(\theta, \omega)_k g_{\omega}^{\mathcal{B}}(\theta, \omega)_j - g_{\omega}^{\mathcal{B}}(\theta, \omega)_j g_{\omega}^{\mathcal{B}}(\theta, \omega)_l \end{pmatrix} + O(\eta^3), \end{aligned}$$

and higher-order polynomials are of order $O(\eta^3)$. Notice that $\tilde{\mathcal{B}} \perp \mathcal{B}$ and recall the definition of Σ_{θ} and Σ_{ω} . The conclusion follows. \square

Now, for (ALT-SDE) with the same initialization as (11), define the corresponding one-step difference:

$$\tilde{\Delta} = \tilde{\Delta}(\theta, \omega) = \begin{pmatrix} \Theta_{1 \times \eta} - \theta \\ \mathcal{W}_{1 \times \eta} - \omega \end{pmatrix}.$$

Let $\tilde{\Delta}_k$ be the k th component of $\tilde{\Delta}$, $k = 1, \dots, d_{\theta} + d_{\omega}$, and $\tilde{\Delta}^{i,j}$ be the tuple consisting of the i th and j th components of the one-step difference of Θ and \mathcal{W} , with $i = 1, \dots, d_{\theta}$ and $j = 1, \dots, d_{\omega}$.

Lemma 2. Suppose b_0, b_1 , and σ given by (5), (6), and (7) are from $\mathcal{C}^3(\mathbb{R}^{d_{\theta} + d_{\omega}})$ such that, for any multi-index J of order $|J| \leq 3$, there exist $k_1, k_2 \in \mathbb{N}$ satisfying

$$\max \{ |\nabla^J b_0(\theta, \omega)|, |\nabla^J b_1(\theta, \omega)|, |\nabla^J \sigma(\theta, \omega)| \} \leq k_1 \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2^{2k_2} \right)$$

and they are all Lipschitz.

The first moment is given by

$$\mathbb{E}[\tilde{\Delta}^{i,j}] = \eta \begin{pmatrix} -g_{\theta}(\theta, \omega)_i \\ g_{\omega}(\theta, \omega)_j \end{pmatrix} + \eta^2 \begin{pmatrix} \{-\nabla_{\omega} [g_{\theta}(\theta, \omega)_i]\}^{\top} g_{\omega}(\theta, \omega) \\ 0 \end{pmatrix} + O(\eta^3).$$

The second moment is given by

$$\mathbb{E}[\tilde{\Delta}^{i,j} (\tilde{\Delta}^{k,l})^{\top}] = \eta^2 \left[\frac{1}{B} \begin{pmatrix} \Sigma_{\theta}(\theta, \omega)_{i,k} & 0 \\ 0 & \Sigma_{\omega}(\theta, \omega)_{j,l} \end{pmatrix} + \begin{pmatrix} -g_{\theta}(\theta, \omega)_i \\ g_{\omega}(\theta, \omega)_j \end{pmatrix} \begin{pmatrix} -g_{\theta}(\theta, \omega)_k \\ g_{\omega}(\theta, \omega)_l \end{pmatrix}^{\top} \right] + O(\eta^3).$$

The third moments are all of order $O(\eta^3)$.

Proof. Let $\psi : \mathbb{R}^{d_\theta+d_\omega} \rightarrow \mathbb{R}$ be any smooth test function. Under the dynamic (ALT-SDE), define the following operators:

$$\begin{aligned} \mathcal{L}_1\psi(\theta, \omega) &= b_0(\theta, \omega)^\top \nabla \psi(\theta, \omega), \\ \mathcal{L}_2\psi(\theta, \omega) &= b_1(\theta, \omega)^\top \nabla \psi(\theta, \omega), \\ \mathcal{L}_3\psi(\theta, \omega) &= \frac{1}{2} \text{Tr}(\sigma(\theta, \omega)\sigma(\theta, \omega)^\top \nabla^2 \psi(\theta, \omega)). \end{aligned}$$

Applying Itô’s formula to $\psi(\Theta_t, \mathcal{W}_t)$, $\mathcal{L}_i\psi(\Theta_t, \mathcal{W}_t)$ for $i = 1, 2, 3$, and $\mathcal{L}_1^2\psi(\Theta_t, \mathcal{W}_t)$, we have

$$\begin{aligned} \psi(\Theta_\eta, \mathcal{W}_\eta) &= \psi(\theta, \omega) + \int_0^\eta (\mathcal{L}_1 + \eta\mathcal{L}_2 + \mathcal{L}_3)\psi(\Theta_t, \mathcal{W}_t) dt \\ &\quad + \int_0^\eta [\nabla \psi(\Theta_t, \mathcal{W}_t)]^\top \sigma(\Theta_t, \mathcal{W}_t) dW_t \\ &= \psi(\theta, \omega) + \eta(\mathcal{L}_1 + \mathcal{L}_3)\psi(\theta, \omega) + \eta^2\left(\frac{1}{2}\mathcal{L}_1^2 + \mathcal{L}_2\right)\psi(\theta, \omega) \\ &\quad + \int_0^\eta \int_0^t \int_0^s \mathcal{L}_1^3\psi(\Theta_u, \mathcal{W}_u) du ds dt \\ &\quad + \int_0^\eta \int_0^t (\mathcal{L}_3\mathcal{L}_1 + \mathcal{L}_1\mathcal{L}_3 + \mathcal{L}_3^2)\psi(\Theta_s, \mathcal{W}_s) ds dt \\ &\quad + \eta \int_0^\eta \int_0^t (\mathcal{L}_2\mathcal{L}_1 + \mathcal{L}_1\mathcal{L}_2 + \mathcal{L}_3\mathcal{L}_2 + \mathcal{L}_2\mathcal{L}_3)\psi(\Theta_s, \mathcal{W}_s) ds dt \\ &\quad + \eta^2 \int_0^\eta \int_0^t \mathcal{L}_2^2\psi(\Theta_s, \mathcal{W}_s) ds dt \end{aligned} \tag{12}$$

$+ M_\eta,$

where M_η denotes the remaining martingale term with mean zero. Given the regularity conditions of b_0 , b_1 , and σ , [20, Theorem 9, Section 2.5] implies that (12) is of order $O(\eta^3)$. Therefore,

$$\mathbb{E}[\psi(\Theta_\eta, \mathcal{W}_\eta) \mid \Theta_0 = \theta, \mathcal{W}_0 = \omega] = \psi(\theta, \omega) + \eta(\mathcal{L}_1 + \mathcal{L}_3)\psi(\theta, \omega) + \eta^2\left(\frac{1}{2}\mathcal{L}_1^2 + \mathcal{L}_2\right)\psi(\theta, \omega).$$

Take $\psi(\Theta_\eta, \mathcal{W}_\eta)$ as $\tilde{\Delta}_i$, $\tilde{\Delta}_i\tilde{\Delta}_j$, and $\tilde{\Delta}_i\tilde{\Delta}_j\tilde{\Delta}_k$ for arbitrary indices $i, j, k = 1, \dots, d_\theta + d_\omega$, and the conclusion follows. □

Estimate of moments. Next, we bound the moments of GAN parameters under (ALT).

Lemma 3. Fix an arbitrary time horizon $\mathcal{T} > 0$ and take the learning rate $\eta \in (0, 1 \wedge \mathcal{T})$ and the number of iterations $\bar{N} = \lfloor \mathcal{T}/\eta \rfloor$. Suppose that $g_\theta^{i,j}$ and $g_\omega^{i,j}$ are all Lipschitz, i.e. there exists $L > 0$ such that

$$\max_{i,j} \{ |g_\theta^{i,j}(\theta, \omega)|, |g_\omega^{i,j}(\theta, \omega)| \} \leq L \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \right).$$

Then, for any $m \in \mathbb{N}$,

$$\max_{t=1, \dots, \bar{N}} \mathbb{E} \left[\left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^m \right]$$

is uniformly bounded, independent of η .

Proof. Throughout the proof, the positive constants C and C' may vary from line to line. The Lipschitz assumption suggests that

$$\max \{ |g_\theta^{\mathcal{B}}(\theta, \omega)|, |g_\omega^{\mathcal{B}}(\theta, \omega)| \} \leq L \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \right).$$

For any $k = 1, \dots, m$,

$$\max \{ |g_\theta^{\mathcal{B}}(\theta, \omega)|^k, |g_\omega^{\mathcal{B}}(\theta, \omega)|^k \} \leq L \cdot k \binom{k}{\lfloor k/2 \rfloor} \cdot \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \right)^k$$

and

$$\left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2^k + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2^m \leq 2 \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \right)^m.$$

For any $t = 0, \dots, \bar{N} - 1$,

$$\begin{aligned} \left\| \begin{pmatrix} \theta_{t+1} \\ \omega_{t+1} \end{pmatrix} \right\|_2^m &\leq \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^m + \sum_{k=1}^m \binom{m}{k} \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^{m-k} \eta^k \left\| \begin{pmatrix} -g_\theta^{\mathcal{B}}(\theta_t, \omega_t) \\ g_\omega^{\mathcal{B}}(\theta_t, \omega_t) \end{pmatrix} \right\|_2^k \\ &\leq \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^m + C\eta \sum_{k=1}^m \binom{m}{k} \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^{m-k} \left(1 + \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2 \right)^m \\ &\leq (1 + C\eta) \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^m + C'\eta. \end{aligned}$$

Write

$$a_t^m = \left\| \begin{pmatrix} \theta_t \\ \omega_t \end{pmatrix} \right\|_2^m.$$

Then $a_{t+1}^m \leq (1 + C\eta)a_t^m + C'\eta$, which leads to

$$\begin{aligned} a_t^m &\leq (1 + C\eta)^t \left(a_0^m + \frac{C'}{C} \right) - \frac{C'}{C} \\ &\leq (1 + C\eta)^{\mathcal{T}/\eta} \left(a_0^m + \frac{C'}{C} \right) - \frac{C'}{C} \\ &\leq e^{C\mathcal{T}} \left(a_0^m + \frac{C'}{C} \right) - \frac{C'}{C}. \end{aligned}$$

The conclusion follows. □

Mollification Notice that in Theorem 1 (and Theorem 2) the condition about the differentiability of the loss function Φ is in the weak sense. For ease of analysis, we adopt the following mollification, given in [12].

Definition 1. (*Mollifier.*) Define the function $v : \mathbb{R}^{d_\theta+d_\omega} \rightarrow \mathbb{R}$,

$$v(u) = \begin{cases} C \exp \{-1/(\|u\|_2^2 - 1)\}, & \|u\|_2 < 1, \\ 0, & \|u\|_2 \geq 1, \end{cases}$$

such that $\int_{\mathbb{R}^{d_\theta+d_\omega}} v(u) \, du = 1$. For any $\varepsilon > 0$, define

$$v^\varepsilon(u) = \frac{1}{\varepsilon^{d_\theta+d_\omega}} v\left(\frac{u}{\varepsilon}\right).$$

Note that the mollifier $v \in C^\infty(\mathbb{R}^{d_\theta+d_\omega})$ and, for any $\varepsilon > 0$, $\text{supp}(v^\varepsilon) = B_\varepsilon(0)$, where $B_\varepsilon(0)$ denotes the ε ball around the origin in the Euclidean space $\mathbb{R}^{d_\theta+d_\omega}$.

Definition 2. (*Mollification.*) Let $f \in \mathcal{L}_{\text{loc}}^1(\mathbb{R}^{d_\theta+d_\omega})$ be any locally integrable function. For any $\varepsilon > 0$, define $f^\varepsilon = v^\varepsilon * f$ such that

$$f^\varepsilon(u) = \int_{\mathbb{R}^{d_\theta+d_\omega}} v^\varepsilon(u-v)f(v) \, dv = \int_{\mathbb{R}^{d_\theta+d_\omega}} v^\varepsilon(v)f(u-v) \, dv.$$

By a simple change of variables and integration by parts, we can derive that, for any multi-index J , $\nabla^J f^\varepsilon = v^\varepsilon * [D^J f]$. Here we quote some well-known results about this mollification from [12, Theorem 7, Appendix C.4].

Lemma 4.

- (i) $f^\varepsilon \in C^\infty(\mathbb{R}^{d_\theta+d_\omega})$.
- (ii) $f^\varepsilon \rightarrow f$ almost everywhere as $\varepsilon \rightarrow 0$.
- (iii) If $f \in \mathcal{C}(\mathbb{R}^{d_\theta+d_\omega})$, then $f^\varepsilon \rightarrow f$ uniformly on compact subsets of $\mathbb{R}^{d_\theta+d_\omega}$.
- (iv) If $f \in \mathcal{L}_{\text{loc}}^p(\mathbb{R}^{d_\theta+d_\omega})$ for some $1 \leq p < \infty$, then $f^\varepsilon \rightarrow f$ in $\mathcal{L}_{\text{loc}}^p(\mathbb{R}^{d_\theta+d_\omega})$.

To give a convergence rate for the pointwise convergence in Lemma 4, we have the following lemma.

Lemma 5. Assume $f \in W_{\text{loc}}^{1,1}(\mathbb{R}^{d_\theta+d_\omega})$ and there exist k_1, k_2 such that $|Df(u)| \leq k_1(1 + \|u\|_2^{2k_2})$; then, for any $u \in \mathbb{R}^{d_\theta+d_\omega}$, there exists $\rho : \mathbb{R}^+ \rightarrow \mathbb{R}$ such that $\lim_{\varepsilon \rightarrow 0} \rho(\varepsilon) = 0$ and $|f^\varepsilon(u) - f(u)| \leq \rho(\varepsilon)$.

Proof.

$$\begin{aligned} |f^\varepsilon(u) - f(u)| &= \left| \int_{B_\varepsilon(0)} v^\varepsilon(v)[f(u-v) - f(u)] \, dv \right| \\ &= \left| \int_{B_\varepsilon(0)} v^\varepsilon(v) \int_0^1 [Df(u-hv)^\top v] \, dh \, dv \right| \\ &\leq \varepsilon \int_{B_\varepsilon(0)} v^\varepsilon(v) \int_0^1 |Df(u-hv)| \, dh \, dv. \end{aligned}$$

Since there exist k_1, k_2 such that $|Df(u)| \leq k_1(1 + \|u\|_2^{2k_2})$,

$$\begin{aligned} |f^\varepsilon(u) - f(u)| &\leq \varepsilon \int_{B_\varepsilon(0)} v^\varepsilon(v) \int_0^1 \left[k_1(1 + \|u - hv\|_2^{2k_2}) \right] dh dv \\ &\leq \varepsilon \int_{B_\varepsilon(0)} v^\varepsilon(v) \int_0^1 \left[k_1(1 + \|u\|_2^{2k_2} + h^{2k_2} \|v\|_2^{2k_2}) \right] dh dv \\ &\leq \varepsilon \int_{B_\varepsilon(0)} v^\varepsilon(v) \left[k_1(1 + \|u\|_2^{2k_2}) + \frac{k_1}{2k_2 + 1} \|v\|_2^{2k_2} \right] dv \\ &\leq \varepsilon \left[k_1(1 + \|u\|_2^{2k_2}) \right] + \frac{k_1}{2k_2 + 1} \varepsilon^{2k_2+1}. \end{aligned}$$

Let $\rho(\varepsilon) = \varepsilon \left[k_1(1 + \|u\|_2^{2k_2}) \right] + (k_1/(2k_2 + 1))\varepsilon^{2k_2+1}$. Then $\rho(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$. □

It is also straightforward to see that the mollification preserves Lipschitz conditions. Consider the following SDE under component-wise mollification of coefficients:

$$d \begin{pmatrix} \Theta_t^\varepsilon \\ \mathcal{W}_t^\varepsilon \end{pmatrix} = [b_0^\varepsilon(\Theta_t^\varepsilon, \mathcal{W}_t^\varepsilon) dt + \eta b_1^\varepsilon(\Theta_t^\varepsilon, \mathcal{W}_t^\varepsilon)] + \sigma^\varepsilon(\Theta_t^\varepsilon, \mathcal{W}_t^\varepsilon) dW_t. \tag{SDE-MLF}$$

Lemma 6. Assume b_0, b_1 , and σ are all Lipschitz. Then

$$\mathbb{E} \left[\max_{t=1, \dots, \tilde{N}} \left\| \begin{pmatrix} \Theta_{t\eta}^\varepsilon \\ \mathcal{W}_{t\eta}^\varepsilon \end{pmatrix} - \begin{pmatrix} \Theta_{t\eta} \\ \mathcal{W}_{t\eta} \end{pmatrix} \right\|_2^2 \right] \xrightarrow{\varepsilon \rightarrow 0} 0,$$

where

$$\begin{pmatrix} \Theta_{t\eta}^\varepsilon \\ \mathcal{W}_{t\eta}^\varepsilon \end{pmatrix} \text{ is given by (SDE-MLF),} \quad \begin{pmatrix} \Theta_{t\eta} \\ \mathcal{W}_{t\eta} \end{pmatrix} \text{ is given by (ALT-SDE).}$$

Proof. With Lemma 5, the conclusion follows from [20, Theorem 9, Section 2.5]. □

3.3.2. *Remaining proof* Given the conditions of Theorem 1 and the fact that mollification preserves Lipschitz conditions, $b_0^\varepsilon, b_1^\varepsilon$, and σ^ε inherit regularity conditions from Theorem 1. Therefore, the conclusion from Lemma 2 holds. Lemma 2 holds. Lemmas 1, 2, 3, and 5 verify the condition in [24, Theorem 3]. Therefore, for any test function $f \in \mathcal{C}^3(\mathbb{R}^{d_\theta+d_\omega})$ such that, for any multi-index J with $|J| \leq 3$, there exist $k_1, k_2 \in \mathbb{N}$ satisfying

$$|\nabla^J f(\theta, \omega)| \leq k_1 \left(1 + \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2^{2k_2} \right),$$

we have the weak approximation given by (8), where (θ_t, ω_t) and $(\Theta_{t\eta}, \mathcal{W}_{t\eta})$ are given by (ALT) and (SDE-MLF), respectively, and ρ is given as in Lemma 5.

Finally, taking ε to 0, Lemma 6 and the explicit form of ρ lead to the conclusion.

4. The long-run behavior of GAN training via SDE invariant measures

In this section we study the long-run behavior of GAN training and discuss some of the implications of the technical assumptions as well as the steady state.

4.1. Long-run behavior of GAN training

In addition to the evolution of parameters in GANs, the long-run behavior of GAN training can be estimated from the SDEs (ALT-SDE) and (SML-SDE). This limiting behavior is characterized by their invariant measures. Recall the following definition of invariant measures in [8].

Definition 3. A probability measure $\mu^* \in \mathcal{P}(\mathbb{R}^{d_\theta+d_\omega})$ is called an invariant measure for a stochastic process $\{(\Theta_t, \mathcal{W}_t)^\top\}_{t \geq 0}$ if, for any measurable bounded function f and $t \geq 0$,

$$\int \mathbb{E}[f(\Theta_t, \mathcal{W}_t) \mid \Theta_0 = \theta, \mathcal{W}_0 = \omega] \mu^*(d\theta, d\omega) = \int f(\theta, \omega) \mu^*(d\theta, d\omega).$$

Remark 2. Intuitively, an invariant measure μ^* in the context of GAN training describes the joint probability distribution of the generator and discriminator parameters $(\Theta^*, \mathcal{W}^*)$ in equilibrium. For instance, if the training process converges to the unique minimax point (θ^*, ω^*) for $\min_\theta \max_\omega \Phi(\theta, \omega)$, the invariant measure is the Dirac mass at (θ^*, ω^*) .

Moreover, the invariant measure μ^* and the marginal distribution of Θ^* characterize the generated distribution $\text{Law}(G_{\Theta^*}(Z))$, necessary for producing synthesized data and for evaluating the performance of the GAN model through metrics such as inception score and Fréchet inception distance. (See [16, 32] for more details on these metrics.)

Finally, as emphasized in Section 2 GANs are minimax games. From a game perspective, the probability distribution of Θ^* conditioning on the discriminator parameter \mathcal{W}^* , denoted by the $\text{Law}(\Theta^* \mid \mathcal{W}^*)$, corresponds to the *mixed strategies* adopted by the generator; likewise, the probability distribution of \mathcal{W}^* conditioning on the generator parameter Θ^* , denoted by $\text{Law}(\mathcal{W}^* \mid \Theta^*)$, characterizes the *mixed strategies* adopted by the discriminator.

Recall that the SDE approximation (ALT-SDE) for the GAN training process is given by

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = b(\Theta_t, \mathcal{W}_t) dt + \sigma(\Theta_t, \mathcal{W}_t) dW_t,$$

where the drift coefficient is given by $b(\theta, \omega) = b_0(\theta, \omega) + \eta b_1(\theta, \omega)$ with

$$\begin{aligned} b_0(\theta, \omega) &= \begin{pmatrix} -g_\theta(\theta, \omega) \\ g_\omega(\theta, \omega) \end{pmatrix}, \\ b_1(\theta, \omega) &= \frac{1}{2} \begin{pmatrix} \nabla_\theta g_\theta(\theta, \omega) & -\nabla_\omega g_\theta(\theta, \omega) \\ -\nabla_\theta g_\omega(\theta, \omega) & -\nabla_\omega g_\omega(\theta, \omega) \end{pmatrix} \begin{pmatrix} -g_\theta(\theta, \omega) \\ g_\omega(\theta, \omega) \end{pmatrix} \\ &= -\frac{1}{2} \nabla b_0(\theta, \omega) b_0(\theta, \omega) - \begin{pmatrix} \nabla_\omega g_\theta(\theta, \omega) g_\omega(\theta, \omega) \\ 0 \end{pmatrix}, \end{aligned}$$

and the diffusion coefficient is given by

$$\sigma(\theta, \omega) = \sqrt{2\beta^{-1}} \begin{pmatrix} \Sigma_\theta(\Theta_t, \mathcal{W}_t)^{1/2} & 0 \\ 0 & \Sigma_\omega(\Theta_t, \mathcal{W}_t)^{1/2} \end{pmatrix}.$$

Note that (ALT-SDE) depends on the first- and second-order derivatives of the training loss with respect to the generator and the discriminator parameters.

Theorem 3. Assume the following conditions hold:

3(a) both b and σ are bounded and smooth, and have bounded derivatives of any order;

3(b) there exist some positive real numbers r and M_0 such that, for any $(\theta \ \omega)^\top \in \mathbb{R}^{d_\theta+d_\omega}$,

$$(\theta \ \omega)b(\theta, \omega) \leq -r \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \quad \text{if} \quad \left\| \begin{pmatrix} \theta \\ \omega \end{pmatrix} \right\|_2 \geq M_0;$$

3(c) A is uniformly elliptic, i.e. there exists $l > 0$ such that

$$\text{for any } \begin{pmatrix} \theta \\ \omega \end{pmatrix}, \begin{pmatrix} \theta' \\ \omega' \end{pmatrix} \in \mathbb{R}^{d_\theta+d_\omega}, \quad (\theta' \ \omega')^\top \sigma(\theta, \omega) \sigma(\theta, \omega)^\top \begin{pmatrix} \theta' \\ \omega' \end{pmatrix} \geq l \left\| \begin{pmatrix} \theta' \\ \omega' \end{pmatrix} \right\|_2^2.$$

Then (ALT-SDE) admits a unique invariant measure μ^* with an exponential convergence rate of the joint distribution of $(\Theta_t, \mathcal{W}_t)$ towards μ^* as $t \rightarrow \infty$.

Similar results hold for the invariant measure of (SML-SDE) with b replaced by b_0 .

Proof. In order to prove Theorem 3, we construct an appropriate Lyapunov function to characterize the long-term behavior for the SDE (ALT-SDE); the associated Lyapunov condition leads to the existence of an invariant measure for the dynamics of the parameters. We highlight this technique since it can be used in the analysis of broader classes of dynamical systems, for both stochastic and deterministic cases; see, for instance, [22]. Consider the function $V : [0, \infty) \times \mathbb{R}^{d_\theta+d_\omega} \rightarrow \mathbb{R}$ given by $V(t, u) = \exp\{\delta t + \varepsilon \|u\|_2\}$ for all $u \in \mathbb{R}^{d_\theta+d_\omega}$, where the parameters $\delta, \varepsilon > 0$ will be determined later. Note that V is a smooth function, and

$$\lim_{\|u\|_2 \rightarrow \infty} \inf_{t \geq 0} V(t, u) = +\infty \tag{13}$$

for any fixed $\delta, \varepsilon > 0$. Under (ALT-SDE), applying Itô’s formula to V gives

$$\begin{aligned} dV(t, \Theta_t, \mathcal{W}_t) = & V(t, \Theta_t, \mathcal{W}_t) \left[\varepsilon \frac{(\Theta_t \ \mathcal{W}_t)b(\Theta_t, \mathcal{W}_t)}{\|(\Theta_t \ \mathcal{W}_t)^\top\|_2} + \delta \right. \\ & + \frac{1}{2} \text{Tr} \left(\sigma(\Theta_t, \mathcal{W}_t) \sigma(\Theta_t, \mathcal{W}_t)^\top \right. \\ & \left. \left. \times \left\{ \frac{\varepsilon \|(\Theta_t \ \mathcal{W}_t)^\top\|_2^2 I + (\varepsilon^2 \|(\Theta_t \ \mathcal{W}_t)^\top\|_2 - \varepsilon) (\Theta_t \ \mathcal{W}_t)^\top (\Theta_t \ \mathcal{W}_t)}{\|(\Theta_t \ \mathcal{W}_t)^\top\|_2^3} \right\} \right) \right] dt \\ & + \varepsilon V(t, \Theta_t, \mathcal{W}_t) \frac{(\Theta_t \ \mathcal{W}_t) \sigma(\Theta_t, \mathcal{W}_t)}{\|(\Theta_t \ \mathcal{W}_t)^\top\|_2} dW_t. \end{aligned}$$

Define the Lyapunov operator

$$\mathcal{L}V(t, u) = V(t, u) \left[\varepsilon \frac{u^\top b(u)}{\|u\|_2} + \delta + \frac{1}{2} \text{Tr} \left(\sigma(u) \sigma(u)^\top \frac{\varepsilon \|u\|_2^2 I + (\varepsilon^2 \|u\|_2 - \varepsilon) uu^\top}{\|u\|_2^3} \right) \right].$$

Given the boundedness of σ , i.e. there exists $K > 0$ such that $\|\sigma\|_F \leq K$, and the dissipative property given by condition 3(b), i.e. there exist $r, M_0 > 0$ such that, for any $u \in \mathbb{R}^{d_\theta+d_\omega}$ with $\|u\|_2 > M_0, u^\top b(u) \leq -r\|u\|_2$, we have

$$\begin{aligned} \mathcal{L}V(t, u) &\leq V(t, u) \left[\delta - r\varepsilon + \frac{1}{2} \left(\varepsilon \frac{\|\sigma\|_F^2}{\|u\|_2} + \varepsilon^2 \|\sigma\|_F^2 \right) \right] \\ &\leq V(t, u) \left[\delta + \frac{K^2\varepsilon^2}{2} - \left(r - \frac{K^2}{2\|u\|_2} \right) \varepsilon \right]. \end{aligned}$$

Now take

$$M > \max \left\{ \frac{K^2}{2r}, M_0 \right\}, \quad 0 < \varepsilon < \frac{2r}{K^2} - \frac{1}{M}, \quad \delta = -\frac{1}{2} \left[\frac{K^2\varepsilon^2}{2} + \left(\frac{K^2}{2M} - r \right) \varepsilon \right] > 0;$$

then, for any $\|u\|_2 > M, \mathcal{L}V(t, u) \leq -\delta V(t, u)$. Therefore,

$$\lim_{\|u\|_2 \rightarrow \infty} \inf_{t \geq 0} \mathcal{L}V(t, u) = -\infty. \tag{14}$$

Following [19, Theorem 3.7], (13) and (14) ensure the existence of an invariant measure μ^* for (ALT-SDE). By the uniform elliptic condition 3(c), uniqueness follows from [17, Theorem 2.3]. Following from the main result in [38], the mixing coefficient

$$\beta(s) := \sup_t \mathbb{E} \left[\text{TV}_{B \in \sigma(X_u, t+s \leq u < \infty)} \left| \mathbb{P}(B \mid \sigma(X_u, 0 \leq u \leq t)) - \mathbb{P}(B) \mid X_0 = x \right. \right], \tag{15}$$

decays exponentially as $s \rightarrow \infty$. For a Borel measurable set $C \subset \mathbb{R}^{d_\theta+d_\omega}$,

$$\left| \mathbb{E} \left[\mathbf{1} \{ (\Theta_s, \mathcal{W}_s) \in C \mid (\Theta_0, \mathcal{W}_0) = (\theta, \omega) \} \right] - \mathbb{E} \left[\mathbf{1} \{ (\Theta_s, \mathcal{W}_s) \in C \} \right] \right| \leq \beta(s)$$

for all $s > 0$. Since any bounded and measurable function f can be approximated by simple functions, the usual argument from indicator functions to simple functions implies that

$$\left| \mathbb{E} [f(\Theta_s, \mathcal{W}_s) \mid (\Theta_0, \mathcal{W}_0) = (\theta, \omega)] - \mathbb{E}_{\mu^*} [f(\Theta, \mathcal{W})] \right| \leq C\beta(s)$$

for all $s > 0$, for some constant $C > 0$. Let ν be an arbitrary initial distribution of $(\Theta_0, \mathcal{W}_0)$. Then we have

$$\left| \mathbb{E}_{(\Theta_0, \mathcal{W}_0) \sim \nu} f(\Theta_s, \mathcal{W}_s) - \mathbb{E}_{\mu^*} f(\Theta, \mathcal{W}) \right| < C\beta(s)$$

for all $s > 0$. The conclusion follows. □

Theorem 3, together with Theorems 1 and 2, help to control the distance between the training output after $\bar{N} = \lfloor T/\eta \rfloor$ iterations and the mixed-strategy equilibrium $(\Theta^*, \mathcal{W}^*) \sim \mu^*$ in a sense that

$$\left| \mathbb{E} [f(\theta_{\bar{N}}, \omega_{\bar{N}})] - \mathbb{E} [f(\Theta^*, \mathcal{W}^*)] \right| \leq C_1(T)\rho_1(\eta) + C_2\beta(T) \tag{16}$$

for any bounded measurable function $f \in G^3(\mathbb{R}^{d_\theta+d_\omega})$, where C_1 and ρ_1 are as in (9), C_2 is some positive constant, and β is as in (15).

4.2. Discussion of the assumptions

4.2.1. *Implications of the technical assumptions on GAN training.* The assumptions 1(a)–(c), 2(a) and (b), and 3(a) for the regularity conditions of the drift, the volatility, and the derivatives of the loss function Φ , are more than mathematical convenience. They are essential constraints on the growth of the loss function with respect to the model parameters, necessary for avoiding the explosive gradient encountered in the training of GANs.

Moreover, these conditions put restrictions on the gradients of the objective functions with respect to the parameters. By the chain rule, it requires both careful choices of network structures and particular forms of the loss function Φ .

In terms of proper neural network architectures, let us take an example of a network with one hidden layer. Let $f : \mathcal{X} \subset \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ be such that

$$f(x; W^h, w^o) = \sigma_o(w^o \cdot \sigma_h(W^h x)) = \sigma_o\left(\sum_{i=1}^h w_i^o \sigma_h\left(\sum_{j=1}^{d_x} W_{i,j}^h x_j\right)\right).$$

Here, h is the width of the hidden layer, $W^h \in \mathbb{R}^{h \times d_x}$ and $w^o \in \mathbb{R}^h$ are the weight matrix and vector for the hidden and output layers respectively, and $\sigma_h : \mathbb{R} \rightarrow \mathbb{R}$ and $\sigma_o : \mathbb{R} \rightarrow \mathbb{R}$ are the activation functions for the hidden and output layers. Then, taking partial derivatives with respect to the weights yields

$$\begin{aligned} \partial_{W_{i,j}^h} f(x; W^h, w^o) &= \sigma_o' \left(\sum_{i=1}^h w_i^o \sigma_h \left(\sum_{j=1}^{d_x} W_{i,j}^h x_j \right) \right) \cdot w_i^o \cdot \sigma_h' \left(\sum_{j=1}^{d_x} W_{i,j}^h x_j \right) \cdot x_j, \\ \partial_{w_i^o} f(x; W^h, w^o) &= \sigma_o' \left(\sum_{i=1}^h w_i^o \sigma_h \left(\sum_{j=1}^{d_x} W_{i,j}^h x_j \right) \right) \cdot \sigma_h \left(\sum_{j=1}^{d_x} W_{i,j}^h x_j \right), \end{aligned}$$

from which we can see that the regularity conditions, especially the growth of the loss function with respect to the model parameters (i.e. assumptions 1(a)–(c) and 2(a)–(b)) rely on the regularity and the boundedness of the activation functions and the width and depth of the network, as well as the magnitudes of the parameters and data. Therefore, assumptions 1(a)–(c), 2(a) and (b), and 3(a) explain mathematically some well-known practices in GAN training such as introducing various forms of gradient penalties; see, for instance, [15, 37]. See also [33] for a combination of competition and gradient penalty to stabilize GAN training. It is worth noticing that apart from affecting the stability of GAN training, the regularity of the network can also affect the sample complexity of GANs and this phenomenon has been studied in [27] for a class of GANs with optimal transport-based loss functions.

In terms of choices of loss functions, the objective function of the vanilla GANs in [14] is given by

$$l(\theta, \omega) = \mathbb{E}_{X \sim \mathbb{P}_r} [\log D_\omega(X)] + \mathbb{E}_{Z \sim \mathbb{P}_z} [\log (1 - D_\omega(G_\theta(Z)))].$$

Taking partial derivatives with respect to θ and ω , we see that

$$\begin{aligned} \nabla_\theta l(\theta, \omega) &= -\mathbb{E}_{Z \sim \mathbb{P}_z} \left[\frac{1}{1 - D_\omega(G_\theta(Z))} \mathbf{J}_\theta^G(Z) \nabla_x D_\omega(G_\theta(Z)) \right], \\ \nabla_\omega l(\theta, \omega) &= \mathbb{E}_{X \sim \mathbb{P}_r} \left[\frac{1}{D_\omega(X)} \nabla_\omega D_\omega(X) \right] - \mathbb{E}_{Z \sim \mathbb{P}_z} \left[\frac{1}{1 - D_\omega(G_\theta(Z))} \nabla_\omega D_\omega(G_\theta(Z)) \right], \end{aligned}$$

where \mathbf{J}_θ^G denotes the Jacobian matrix of G_θ with respect to θ , and ∇_x denotes the gradient operator over a (parametrized) function with respect to its input variable. [1] analyzed the difficulties of stabilizing GAN training under the above loss function due to the lack of proper regularity conditions, and proposed a possible remedy by an alternative Wasserstein distance which enjoys better smoothness conditions.

4.2.2. *Verifiability of assumptions in Theorems 1, 2, and 3* The assumptions from these theorems can be summarized into the three categories specified below. For some of the assumptions, there are available choices of GAN structures for a wide range of applications where these assumptions can be verified easily; some are consistent with certain choices of regularization applied in the training procedures of GANs; others are more subtle.

On the smoothness and the boundedness of drift and volatility Take the example of Wasserstein GANs (WGANs) for image processing. Given that sample data in image processing problems are supported on a compact domain, assumptions 1(a)–(c), 2(a) and (b), and 3(a) are easily satisfied with proper choices of prior distribution and activation function: first, the prior distribution \mathbb{P}_z such as the uniform distribution is naturally compactly supported; next, take $D_\omega = \tanh(\omega \cdot x)$, $G_\theta(z) = \tanh(\theta \cdot z)$, and the objective function

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M D_\omega(x_j) - D_\omega(G_\theta(z_i))}{N \cdot M}.$$

Then, assumptions 1(a)–(c), 2(a) and (b), and 3(a) are guaranteed by the boundedness of the data $\{(z_i, z_j)\}_{1 \leq i \leq N, 1 \leq j \leq M}$ and the very structure of the activation function:

$$\psi(y) = \tanh y = \frac{e^y - e^{-y}}{e^y + e^{-y}} = 1 - \frac{2}{e^{2y} + 1} \in (-1, 1).$$

More precisely, the first- and second-order derivatives of ψ are

$$\psi'(y) = \frac{4}{(e^y + e^{-y})^2} \in (0, 1], \quad \psi''(y) = -8 \frac{e^y - e^{-y}}{(e^y + e^{-y})^3} = -2\psi(y)\psi'(y) \in (-2, 2);$$

any higher-order derivatives can be written as functions of $\psi(\cdot)$ and $\psi'(\cdot)$ and are therefore bounded.

On the dissipative property The dissipative property specified by 3(b) essentially prevents the evolution of the parameters from being driven to infinity. The weight clipping technique in WGANs, for instance, is consistent with this assumption.

On the elliptic condition Compared with the above two categories of assumptions, the uniform ellipticity condition 3(c) is intrinsically rooted in the stochasticity brought by the sampling procedures of stochastic gradient algorithms in general, i.e. the *microscopic fluctuation from the noise of SGAs*, instead of the *macroscopic loss landscape of GANs*. Recall from Section 2 that a cost function of the form

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M J(\theta, \omega; x_j, z_i)}{N \cdot M}$$

naturally induces a random variable $g(\theta, \omega; X, Z) = (\nabla_\theta J(\theta, \omega; X, Z), \nabla_\omega J(\theta, \omega; X, Z))$ with mean $(g_\theta(\theta, \omega), g_\omega(\theta, \omega))$, where (X, Z) follows the empirical distribution given by the dataset \mathcal{D} . The elliptic condition is essentially equivalent to the random variable $g(\theta, \omega; X, Z)$ being

non-degenerate, and the smallest eigenvalue of its covariance matrix, $\underline{\sigma}(\text{Cov}(g(\theta, \omega; X, Z)))$, being bounded away from 0. Note that this condition cannot be guaranteed by adding parameter regularizations as in the case of the dissipative property, since parameter regularizations only change the drift term b . For suitable choices of loss function Φ such that $\underline{\sigma}(\text{Cov}(g(\theta, \omega; X, Z)))$ is indeed bounded away from 0, control of the training outcome (16) holds; otherwise, we could consider a perturbed SDE approximation,

$$d \begin{pmatrix} \Theta_t^\lambda \\ \mathcal{W}_t^\lambda \end{pmatrix} = b(\Theta_t^\lambda, \mathcal{W}_t^\lambda) dt + [\sigma(\Theta_t^\lambda, \mathcal{W}_t^\lambda) + \lambda I] dW_t, \quad \begin{pmatrix} \Theta_0^\lambda \\ \mathcal{W}_0^\lambda \end{pmatrix} = \begin{pmatrix} \Theta_0 \\ \mathcal{W}_0 \end{pmatrix},$$

for sufficiently small $\lambda > 0$. Under condition 3(a), Itô isometry and Gronwall’s inequality give the error bound

$$\mathbb{E} \left\| \begin{pmatrix} \Theta_t^\lambda \\ \mathcal{W}_t^\lambda \end{pmatrix} - \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} \right\|_2 \leq \alpha \lambda, \quad t > 0,$$

for some positive coefficient $\alpha = \alpha(t)$ depending on time t . We can still control the distance between the training outcome and the perturbed equilibrium by

$$|\mathbb{E}[f(\theta_{\bar{N}}, \omega_{\bar{N}})] - \mathbb{E}[f(\Theta^{\lambda,*}, \mathcal{W}^{\lambda,*})]| \leq C_1(\mathcal{T})\rho_1(\eta) + C_2\beta(\mathcal{T}) + C_3(\mathcal{T})\lambda^k$$

for any bounded measurable function $f \in G^3(\mathbb{R}^{d_\theta+d_\omega})$, where C_1, C_2, ρ_1 , and β are as in (16), C_3 is some positive coefficient depending on \mathcal{T} , and $k \in \mathbb{N}$ is some positive integer.

4.3. Dynamics of training loss and FDR

We can further analyze the dynamics of the training loss based on the SDE approximations and derive a fluctuation–dissipation relation (FDR) for the GAN training.

To see this, let $\mu = \{\mu_t\}_{t \geq 0}$ be the flow of probability measures for $\{(\Theta_t, \mathcal{W}_t)^\top\}_{t \geq 0}$ given by (ALT-SDE). Then, applying Itô’s formula to the smooth function Φ (see [31, Section 4.1.8]) gives the following dynamics of training loss:

$$\Phi(\Theta_t, \mathcal{W}_t) = \Phi(\Theta_s, \mathcal{W}_s) + \int_s^t \mathcal{A}\Phi(\Theta_r, \mathcal{W}_r) dr + \int_s^t \sigma(\Theta_r, \mathcal{W}_r)^\top \nabla \Phi(\Theta_r, \mathcal{W}_r) dW_r, \quad (17)$$

where

$$\mathcal{A}f(\theta, \omega) = b(\theta, \omega)^\top \nabla f(\theta, \omega) + \frac{1}{2} \text{Tr}(\sigma(\theta, \omega) \sigma(\theta, \omega)^\top \nabla^2 f(\theta, \omega)) \quad (18)$$

is the infinitesimal generator for (ALT-SDE) on any given test function $f : \mathbb{R}^{d_\theta+d_\omega} \rightarrow \mathbb{R}$.

The existence of the unique invariant measure μ^* for (ALT-SDE) implies the convergence of $\{(\Theta_t, \mathcal{W}_t)^\top\}_{t \geq 0}$ in (ALT-SDE) to some $(\Theta^*, \mathcal{W}^*)^\top \sim \mu^*$ as $t \rightarrow \infty$. By Definition 3 of the invariant measure and (17), we have $\mathbb{E}_{\mu^*}[\mathcal{A}\Phi(\Theta^*, \mathcal{W}^*)] = 0$. Applying the operator (18) over the loss function ϕ yields

$$\begin{aligned} \mathcal{A}\Phi(\theta, \omega) &= b_0(\theta, \omega)^\top \nabla \Phi(\theta, \omega) + \eta b_1(\theta, \omega)^\top \nabla \Phi(\theta, \omega) + \frac{1}{2} \text{Tr}(\sigma(\theta, \omega) \sigma(\theta, \omega)^\top \nabla^2 \Phi(\theta, \omega)) \\ &= -\|\nabla_\theta \Phi(\theta, \omega)\|_2^2 + \|\nabla_\omega \Phi(\theta, \omega)\|_2^2 \\ &\quad - \frac{\eta}{2} [\nabla_\theta \Phi(\theta, \omega)^\top \nabla_\theta^2 \Phi(\theta, \omega) \nabla_\theta \Phi(\theta, \omega) + \nabla_\omega \Phi(\theta, \omega)^\top \nabla_\omega^2 \Phi(\theta, \omega) \nabla_\omega \Phi(\theta, \omega)] \\ &\quad + \beta^{-1} \text{Tr}(\Sigma_\theta(\theta, \omega) \nabla_\theta^2 \Phi(\theta, \omega) + \Sigma_\omega(\theta, \omega) \nabla_\omega^2 \Phi(\theta, \omega)). \end{aligned}$$

Based on the evolution of the loss function (17), convergence to the invariant measure μ^* leads to the following FRD for GAN training.

Theorem 4. Assume the existence of an invariant measure μ^* for (ALT-SDE); then

$$\begin{aligned} & \mathbb{E}_{\mu^*} [\|\nabla_{\theta} \Phi(\Theta^*, \mathcal{W}^*)\|_2^2 - \|\nabla_{\omega} \Phi(\Theta^*, \mathcal{W}^*)\|_2^2] \\ &= \beta^{-1} \mathbb{E}_{\mu^*} [\text{Tr}(\Sigma_{\theta}(\Theta^*, \mathcal{W}^*) \nabla_{\theta}^2 \Phi(\Theta^*, \mathcal{W}^*) + \Sigma_{\omega}(\Theta^*, \mathcal{W}^*) \nabla_{\omega}^2 \Phi(\Theta^*, \mathcal{W}^*))] \\ &\quad - \frac{\eta}{2} \mathbb{E}_{\mu^*} [\nabla_{\theta} \Phi(\Theta^*, \mathcal{W}^*)^{\top} \nabla_{\theta}^2 \Phi(\Theta^*, \mathcal{W}^*) \nabla_{\theta} \Phi(\Theta^*, \mathcal{W}^*) \\ &\quad\quad + \nabla_{\omega} \Phi(\Theta^*, \mathcal{W}^*)^{\top} \nabla_{\omega}^2 \Phi(\Theta^*, \mathcal{W}^*) \nabla_{\omega} \Phi(\Theta^*, \mathcal{W}^*)]. \end{aligned} \tag{FDR1}$$

The corresponding FDR for the simultaneous update case of (SML-SDE) is

$$\begin{aligned} & \mathbb{E}_{\mu^*} [\|\nabla_{\theta} \Phi(\Theta^*, \mathcal{W}^*)\|_2^2 - \|\nabla_{\omega} \Phi(\Theta^*, \mathcal{W}^*)\|_2^2] \\ &= \beta^{-1} \mathbb{E}_{\mu^*} [\text{Tr}(\Sigma_{\theta}(\Theta^*, \mathcal{W}^*) \nabla_{\theta}^2 \Phi(\Theta^*, \mathcal{W}^*) + \Sigma_{\omega}(\Theta^*, \mathcal{W}^*) \nabla_{\omega}^2 \Phi(\Theta^*, \mathcal{W}^*))]. \end{aligned}$$

Remark 3. This FDR relation in GANs connects the microscopic fluctuation from the noise of SGAs with the macroscopic dissipation phenomena related to the loss function. In particular, the quantity $\text{Tr}(\Sigma_{\theta} \nabla_{\theta}^2 \Phi + \Sigma_{\omega} \nabla_{\omega}^2 \Phi)$ links the covariance matrices Σ_{θ} and Σ_{ω} from SGAs with the loss landscape of Φ , and reveals the trade-off of the loss landscape between the generator and the discriminator.

Note that this FDR relation for GAN training is analogous to that for the stochastic gradient descent algorithm on a pure minimization problem in [25, 44].

Further analysis of the invariant measure can lead to a different type of FDR that will be practically useful for learning rate scheduling. Indeed, applying Itô’s formula to the squared norm of the parameters $\|(\Theta_t, \mathcal{W}_t)^{\top}\|_2^2$ shows the following dynamics:

$$d \left\| \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} \right\|_2^2 = 2 \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix}^{\top} d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} + \text{Tr}(\sigma(\Theta_t, \mathcal{W}_t) \sigma(\Theta_t, \mathcal{W}_t)^{\top}) dt.$$

Theorem 5. Assume the existence of an invariant measure μ^* for (SML-SDE); then

$$\begin{aligned} & \mathbb{E}_{\mu^*} [\Theta^{*,T} \nabla_{\theta} \Phi(\Theta^*, \mathcal{W}^*) - \mathcal{W}^{*,T} \nabla_{\omega} \Phi(\Theta^*, \mathcal{W}^*)] \\ &= \beta^{-1} \mathbb{E}_{\mu^*} [\text{Tr}(\Sigma_{\theta}(\Theta^*, \mathcal{W}^*) + \Sigma_{\omega}(\Theta^*, \mathcal{W}^*))]. \end{aligned} \tag{FDR2}$$

Given the infinitesimal generator for (ALT-SDE), Theorems 4 and 5 follow from direct computations.

Remark 4. (Scheduling of learning rate.) Notice that the quantities in (FDR2), including the parameters (θ, ω) and first-order derivatives of the loss function g_{θ} , g_{ω} , g_{θ}^{ij} , and g_{ω}^{ij} , are computationally inexpensive. Therefore, (FDR2) enables customized scheduling of learning rate, instead of predetermined scheduling ones such as Adam or RMSprop optimizer.

For instance, recall that $g_\theta^{\mathcal{B}}$ and $g_\omega^{\mathcal{B}}$ are respectively unbiased estimators for g_θ and g_ω , and

$$\hat{\Sigma}_\theta(\theta, \omega) = \frac{\sum_{k=1}^B [g_\theta^{I_k, J_k}(\theta, \omega) - g_\theta^{\mathcal{B}}(\theta, \omega)][g_\theta^{I_k, J_k}(\theta, \omega) - g_\theta^{\mathcal{B}}(\theta, \omega)]^\top}{B - 1},$$

$$\hat{\Sigma}_\omega(\theta, \omega) = \frac{\sum_{k=1}^B [g_\omega^{I_k, J_k}(\theta, \omega) - g_\omega^{\mathcal{B}}(\theta, \omega)][g_\omega^{I_k, J_k}(\theta, \omega) - g_\omega^{\mathcal{B}}(\theta, \omega)]^\top}{B - 1}$$

are respectively unbiased estimators of $\Sigma_\theta(\theta, \omega)$ and $\Sigma_\omega(\theta, \omega)$. Now, in order to improve GAN training with simultaneous update, we can introduce two tunable parameters $\varepsilon > 0$ and $\delta > 0$ to have the following scheduling:

$$\text{if } \left| \frac{\Theta^\top g_\theta^{\mathcal{B}}(\Theta_t, \mathcal{W}_t) - \mathcal{W}_t^\top g_\omega^{\mathcal{B}}(\Theta_t, \mathcal{W}_t)}{\beta^{-1} \text{Tr}(\hat{\Sigma}_\theta(\Theta_t, \mathcal{W}_t) + \hat{\Sigma}_\omega(\Theta_t, \mathcal{W}_t))} - 1 \right| < \varepsilon, \text{ then update } \eta \text{ by } (1 - \delta)\eta.$$

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

- [1] ARJOVSKY, M. AND BOTTOU, L. (2017). Towards principled methods for training generative adversarial networks. In *Proc. 15th Int. Conf. Learning Representations*.
- [2] ARJOVSKY, M., CHINTALA, S. AND BOTTOU, L. (2017). Wasserstein generative adversarial networks. *Proc. Mach. Learn. Res.* **70**, 214–223.
- [3] BERARD, H., GIDEL, G., ALMAHAIRI, A., VINCENT, P. AND LACOSTE-JULIEN, S. (2020). A closer look at the optimization landscape of generative adversarial networks. In *Proc. Int. Conf. Learning Representations*.
- [4] CAO, H., GUO, X. AND LAURIÈRE, M. (2020). Connecting GANs, MFGs, and OT. Preprint, arXiv:2002.04112.
- [5] CHEN, L., PELGER, M. AND ZHU, J. (2023). Deep learning in asset pricing. *Management Science*.
- [6] COLETTA, A., PRATA, M., CONTI, M., MERCANTI, E., BARTOLINI, N., MOULIN, A., VYETRENKO, S. AND BALCH, T. (2021). Towards realistic market simulations: A generative adversarial networks approach. In *Proc. 2nd ACM Int. Conf. AI in Finance*.
- [7] CONFORTI, G., KAZEYKINA, A. AND REN, Z. (2023). Game on random environment, mean-field Langevin system, and neural networks. *Math. Operat. Res.* **48**, 78–99.
- [8] DA PRATO, G. (2006). *An Introduction to Infinite-Dimensional Analysis*. Springer, New York.
- [9] DENTON, E. L., CHINTALA, S., SZLAM, A. AND FERGUS, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. *Adv. Neural Inf. Proc. Sys.* **28**, 1486–1494.
- [10] DIEULEVEUT, A., DURMUS, A. AND BACH, F. (2020). Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Ann. Statist.* **48**, 1348–1382.
- [11] DOMINGO-ENRICH, C., JELASSI, S., MENSCH, A., ROTSKOFF, G. AND BRUNA, J. (2020). A mean-field analysis of two-player zero-sum games. *Adv. Neural Inf. Proc. Sys.* **33**, 20215–20226.
- [12] EVANS, L. C. (1998). *Partial Differential Equations*, vol. **19**. American Mathematical Society, Providence, RI.
- [13] GENEVAY, A., PEYRÉ, G. AND CUTURI, M. (2017). GAN and VAE from an optimal transport point of view. Preprint, arXiv:1706.01807.
- [14] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. AND BENGIO, Y. (2014). Generative adversarial nets. *Adv. Neural Inf. Proc. Sys.* **27**, 2672–2680.
- [15] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V. AND COURVILLE, A. (2017). Improved training of Wasserstein GANs. *Adv. Neural Inf. Proc. Sys.* **31**, 5767–5777.

- [16] HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B. AND HOCHREITER, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Adv. Neural Inf. Proc. Sys.* **31**, 6626–6637.
- [17] HONG, J. AND WANG, X. (2019). Invariant measures for stochastic differential equations. In *Invariant Measures for Stochastic Nonlinear Schrödinger Equations: Numerical Approximations and Symplectic Structures*. Springer, Singapore, pp. 31–61.
- [18] HU, W., LI, C. J., LI, L. AND LIU, J.-G. (2019). On the diffusion approximation of nonconvex stochastic gradient descent. *Ann. Math. Sci. Appl.* **4**, 3–32.
- [19] KHASMINSKII, R. (2011). *Stochastic Stability of Differential Equations* 2nd edn, vol. **66**. Springer, New York.
- [20] KRYLOV, N. V. (2008). *Controlled Diffusion Processes*. Springer, New York.
- [21] KULHARIA, V., GHOSH, A., MUKERJEE, A., NAMBOODIRI, V. AND BANSAL, M. (2017). Contextual RNN-GANs for abstract reasoning diagram generation. In *Proc. 31st AAAI Conf. Artificial Intelligence*, pp. 1382–1388.
- [22] LABORDE, M. AND OBERMAN, A. (2020). A Lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case. In *Proc. 23rd Int. Conf. Artificial Intelligence Statist.*, pp. 602–612.
- [23] LEDIG, C., THEIS, L., HUSZÁR, F., CABALLERO, J., CUNNINGHAM, A., ACOSTA, A., AITKEN, A., TEJANI, A., TOTZ, J., WANG, Z. *et al.* (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4681–4690.
- [24] LI, Q., TAI, C. AND E, W. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *J. Mach. Learn. Res.* **20**, 1–47.
- [25] LIU, G.-H. AND THEODOROU, E. A. (2019). Deep learning theory review: An optimal control and dynamical systems perspective. Preprint, arXiv:1908.10920.
- [26] LUC, P., COUPRIE, C., CHINTALA, S. AND VERBEEK, J. (2016). Semantic segmentation using adversarial networks. In *Proc. NIPS Workshop Adversarial Training*.
- [27] LUISE, G., PONTIL, M. AND CILIBERTO, C. (2020). Generalization properties of optimal transport GANs with latent distribution learning. Preprint, arXiv:2007.14641.
- [28] MESCHERER, L., GEIGER, A. AND NOWOZIN, S. (2018). Which training methods for GANs do actually converge? In *Proc. Int. Conf. Machine Learning*, pp. 3481–3490.
- [29] RADFORD, A., METZ, L. AND CHINTALA, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. 4th Int. Conf. Learning Representations*.
- [30] REED, S., AKATA, Z., YAN, X., LOGESWARAN, L., SCHIELE, B. AND LEE, H. (2016). Generative adversarial text to image synthesis. In *Proc. 33rd Int. Conf. Machine Learning*, pp. 1060–1069.
- [31] ROGERS, L. C. G. AND WILLIAMS, D. (2000). *Diffusions, Markov Processes and Martingales. Volume 2: Itô Calculus*. Cambridge University Press.
- [32] SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A. AND CHEN, X. (2016). Improved techniques for training GANs. In *Proc. 30th Int. Conf. Neural Inf. Proc. Syst.*, pp. 2234–2242.
- [33] SCHAEFER, F., ZHENG, H. AND ANANDKUMAR, A. (2020). Implicit competitive regularization in GANs. *Proc. Mach. Learn. Res.* **119**, 8533–8544.
- [34] SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8**, 171–176.
- [35] STORCHAN, V., VYETRENKO, S. AND BALCH, T. (2021). Learning who is in the market from time series: Market participant discovery through adversarial calibration of multi-agent simulators. Preprint, arXiv:2108.00664.
- [36] TAKAHASHI, S., CHEN, Y. AND TANAKA-ISHII, K. (2019). Modeling financial time-series with generative adversarial networks. *Physica A* **527**, 121261.
- [37] THANH-TUNG, H., TRAN, T. AND VENKATESH, S. (2019). Improving generalization and stability of generative adversarial networks. In *Proc. Int. Conf. Learning Representations*.
- [38] VERETENNIKOV, A. Y. (1988). Bounds for the mixing rate in the theory of stochastic equations. *Theory Prob. Appl.* **32**, 273–281.
- [39] VON NEUMANN, J. (1959). On the theory of games of strategy. *Contrib. Theory Games* **4**, 13–42.
- [40] VONDRICK, C., PIRSIAVASH, H. AND TORRALBA, A. (2016). Generating videos with scene dynamics. In *Proc. 30th Conf. Neural Inf. Proc. Syst.*, pp. 613–621.
- [41] WIATRAC, M., ALBRECHT, S. V. AND NYSTROM, A. (2019). Stabilizing generative adversarial networks: A survey. Preprint, arXiv:1910.00927.
- [42] WIESE, M., BAI, L., WOOD, B., MORGAN, J. P. AND BUEHLER, H. (2019). Deep hedging: Learning to simulate equity option markets. Preprint, arXiv:1911.01700.
- [43] WIESE, M., KNOBLOCH, R., KORN, R. AND KRETSCHMER, P. (2020). Quant GANs: Deep generation of financial time series. *Quant. Finance* **20**, 1–22.
- [44] YAIDA, S. (2019). Fluctuation–dissipation relations for stochastic gradient descent. In *Proc. Int. Conf. Learning Representations*.

- [45] YEH, R. A., CHEN, C., YIAN LIM, T., SCHWING, A. G., HASEGAWA-JOHNSON, M. AND DO, M. N. (2017). Semantic image inpainting with deep generative models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 5485–5493.
- [46] ZHANG, K., ZHONG, G., DONG, J., WANG, S. AND WANG, Y. (2019). Stock market prediction based on generative adversarial network. *Procedia Comp. Sci.* **147**, 400–406.
- [47] ZHU, B., JIAO, J. AND TSE, D. (2020). Deconstructing generative adversarial networks. *IEEE Trans. Inf. Theory* **66**, 7155–7179.
- [48] ZHU, J.-Y., KRÄHENBÜHL, P., SHECHTMAN, E. AND EFROS, A. A. (2016). Generative visual manipulation on the natural image manifold. In *Proc. Eur. Conf. Computer Vision*, pp. 597–613.