Generative Adversarial Networks for Markovian Temporal Dynamics: Stochastic Continuous Data Generation

Sung Woo Park [1] **Dong Wook Shu** [1] **Junseok Kwon** [1]

Abstract

In this paper, we present a novel generative adversarial network (GAN) that can describe Markovian temporal dynamics. To generate stochastic sequential data, we introduce a novel stochastic differential equation-based conditional generator and spatial-temporal constrained discriminator networks. To stabilize the learning dynamics of the min-max type of the GAN objective function, we propose well-posed constraint terms for both networks. We also propose a novel conditional Markov Wasserstein distance to induce a pathwise Wasserstein distance. The experimental results demonstrate that our method outperforms state-of-the-art methods using several different types of data.

1. Introduction

Recently, research has been actively conducted to synthesize realistic dynamical data, which are ubiquitous and natural in real-world scenes. To develop generative methods for time-sequential data, the following important question should be posed: How can we accurately model fake probability distributions to represent *time-varying real distributions*?

To describe the probabilistic sequences of time-sequential data, conventional methods typically adopt model-based approaches that generate discrete-time temporal states. These approaches have widely employed recurrent neural networks to generate samples $(X_1, \cdots, X_n)$, in which the conditional distribution, $p(X_n|X_{k<n})$, is dependent on black-box types of dynamics induced by the recurrent networks. In this paper, we propose an alternative method that utilizes stochastic differential equations (SDEs) as probabilistic models to generate *continuous* time-sequential data, X_t, according to probability distribution $X_t \sim p_t$, which has the following formulation:

$$dX_t^x = f(X_t, t, \theta)dt + \Sigma(X_t, \theta)dW_t, \quad x \sim p_{t=0}, \quad (1)$$

where the first term propagates particles X_t according to drift function $f(\cdot, t, \theta)$ with parameters θ, and the second term imposes randomness on the evolution paths of X_t using Wiener process W_t and diffusion function Σ. Then, the combination of these two terms can explicitly render probabilistic flows of data. Utilizing SDEs for generating time-sequential data in our method has two advantages: it ensures the generation of Markovian temporal dynamics and stochastic/continuous sample paths.

(Model-based Temporal Dynamics → Markovian Temporal Dynamics) Key to generating stochastic sequential data is estimating an accurate conditional density, $p(X_t|X_s)$, which can describe complex dynamics over temporal transitions. For this, recurrent networks (*i.e.*, model-based)(Yoon et al., 2019; Yingzhen & Mandt, 2018) have been used to define the conditional distribution, $p(X_n|X_{<n})$. [1], for accurate time-sequential data generation. In contrast to model-based approaches, our proposed method adopts stochastic dynamics to define a conditional probability by introducing a continuous Markov transition from X_s to X_t, in which the solution to (1), X_t, is inherently a continuous *Markov process*. In particular, let $p(x, t|y, s)$ be a transition kernel[2] that transitions spatial states y to x given a time interval from s to t. Specifically, SDEs in (1) induce a unique form of partial differential equations (PDEs) (*i.e.*, *Fokker-Planck equations*) in a distributional sense, which proposes an evolution rule for probability distribution p_t at temporal state t:

$$\begin{aligned}&\partial_t p(x, t|y, s)\\ &= -\partial_x f(x, t, \theta)p_t(x) + \frac{1}{2}\mathbf{Tr}\left[\partial_x^2 \Sigma(x, \theta) \cdot p_t(x)\right],\end{aligned} \quad (2)$$

where the initial condition is $x = X_0^x \sim p_0$. The mathematical characteristics of PDEs in (2) provide strong advantages in the development of time-sequence generative models.

[1]School of Computer Science and Engineering, Artificial Intelligence Graduate School, Chung-Ang University, Seoul, Korea. Correspondence to: Sung Woo Park <pswkiki@gmail.com>, Dong Wook Shu <seowok@naver.com>, Junseok Kwon <jskwon@cau.ac.kr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]To distinguish between the notations, we use $n, k \in \mathbb{N}^+$ for discrete processes, and $t, s \in \mathbb{R}^+$ for continuous processes.

[2]The transition kernel exists because X_t^x is Markovian given an initial state x.

The solution to this PDE can be more accurately specified than arbitrary solutions in an infinite-dimensional space of probability distributions p_t over time, because any distributional flow $p_s \rightarrow p_t$ from temporal states s to t, induced by the Fokker-Planck equations, is controlled by the parameterized drift and diffusion functions (*i.e.*, $f_t(x, \theta), \Sigma(\cdot, \theta)$). Owing to the specified dynamical formulation of p_t, our method requires a relatively small number of parameters, θ, to search for the optimal solution of probabilistic flows. For example, the proposed stochastic dynamics in (1) only uses approximately $1.5M$ parameters in total, whereas model-based approaches typically require a large number of parameters ($\approx 3.5 \sim 100M$) to search large probability distribution spaces. Moreover, our method, in which the transition dynamics between p_t to p_s implicitly stem from SDEs, does not introduce additional constraint terms on the conditional distribution to relate X_t and X_s.

(Deterministic/Discrete $\rightarrow$ Stochastic/Continuous Sample Path) Conventional methods typically deal with sequential data in a conditional order, whereby data are represented as d-dimensional state vectors $X(n) = [X_1(n), \cdots, X_d(n)]$. In this setting, sequential data are described by a concatenation of the state vectors, $X(n)$, over T times along *discrete* and *deterministic* sample paths (*i.e.*, $n \in \mathbb{N}^+$, $1 \leq n \leq T$, and $X^T \in \mathbb{R}^{dT}$). Thus, existing methods (Tulyakov et al., 2018; Esteban et al., 2017) cannot describe random changes in state vectors for a particular time interval, because probabilistic information on state vectors over temporal-paths is easily lost owing to the concatenation of state vectors. Given a random latent vector, conventional methods generate samples in a deterministic conditional order and cannot accurately describe stochastic variations over temporal paths. In addition, owing to the characteristics of model-based approaches, continuous states between $X(n-k)$ and $X(n)$ cannot be generated for a finite time interval $k < n \in \mathbb{N}^+$, because generator networks are typically implemented using recurrent network models, inducing discrete-time stochastic processes $X(n)$. In contrast, our proposed method can generate *continuous* samples X_t in a *stochastic* conditional order.

(Contributions & Novelty) Based on the SDEs in (1), our proposed method generates *continuous stochastic* dynamical flows for a given initial probability distribution p_0, which approximates time sequential real-data distributions q_t. In particular, our method aims to minimize the statistical discrepancy (*i.e.*, Wasserstein distance, $\mathcal{W}$) between p_t and q_t:

$$\min_{p_t} \mathcal{W}(p_t, q_t), \quad X_t \sim p_t, \quad \forall 0 \leq t \leq T. \tag{3}$$

To solve (3), we introduce a novel generative adversarial network (GAN) based on a Wasserstein-type of statistical discrepancy. This minimization problem appears to be straightforward, but there are practical and theoretical issues that need to be addressed.

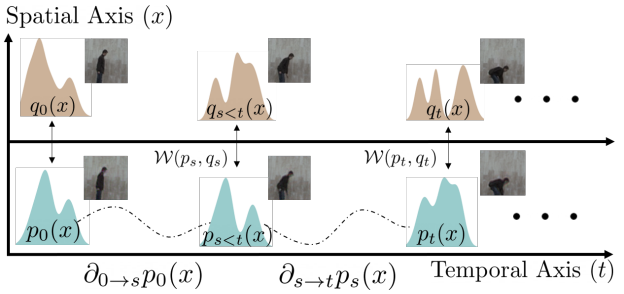

Figure 1. **Conceptual illustration of the proposed method**. The fake distribution, p_t, is learned to minimize the Wasserstein distance, $\mathcal{W}(p_t, q_t)$, according to the real distribution, q_t. The fake distribution p_t is evolved based on the Fokker-Planck equation in (2) with respect to Markov transition density $p(x, t|y, s)$, which propagates distributions $p_0 \rightarrow p_{s<t} \rightarrow p_t$, as shown in the figure.

• The stability and generalization of the proposed generative model is highly dependent on the behavior of the two terms in (1). As stochastic sample X_t is evolved by the stochastic dynamics, drift and diffusion functions without well-posed regularization inevitably induce unstable learning of the min-max types of GAN objectives and over-fitting problems. Thus, well-posed regularization terms should be imposed for both networks to avoid over-parameterization and unstable learning.

• X_t is Markovian, and its conditional dependence according to initial state X_0 is implicitly given. However, the primitive objective function in (3) is fundamentally flawed because the Markov property is not included in the objective; thus, it is unable to explicitly describe the conditional dependence of X_0 and X_t in (3). To analyze the proposed dynamical system rigorously, we propose a Markov-type of Wasserstein distance, which uses the property of Markov process X_t to make the stochastic dependence explicitly.

To solve the aforementioned issues and generate high-quality time-sequential data, we present the following:

1. We present a novel stochastic dynamical GAN (Section 3.2), which can generate stochastic continuous data.
2. We introduce a novel SDE-based conditional generator (Section 3.3) and a spatial-temporal constrained discriminator (Section 3.4) to deal with stochastic sequential data. To stabilize the learning dynamics of the min-max types of GAN objective functions, we propose well-posed regularization terms for both networks.
3. We propose a novel temporal-adaptive Wasserstein distance (Section 3.5) to induce a pathwise distance. The theoretical results support the stabilization of this newly proposed distance.

Fig.1 illustrates the basic idea of our method.

2. Related Work

(Model-based Methods) Time-sequential data were generated using GANs by (Tulyakov et al., 2018). However, this method required two different discriminator networks to classify spatial and temporal dependencies separately, because conditional dependency over temporal paths is not considered in the generative model. To induce accurate temporal dependency, a model-based approach that can approximate conditional dependencies (*e.g.*, RNN) was proposed in (Yingzhen & Mandt, 2018). Recently, causal optimal transport, which imposes constraints on transport plans, was introduced by (Xu et al., 2020) to induce independence of real samples given generated samples.

(Dynamics-based Methods) Dynamical systems (Yildiz et al., 2019) have been proposed, in which continuous probability densities were evolved using the second-order ODEs. However, because of the deterministic characteristics of the sample paths governed by ODEs, they could not accurately generate diverse samples over temporal paths, even though their Bayesian neural network restrictively described the randomness of sample dynamics. In contrast, our method directly implements stochastic models based on SDEs, which can be considered a stochastic version of the dynamical systems in (Yildiz et al., 2019). It should be noted that our method is equivalent to the deterministic model with the first order method proposed by (Yildiz et al., 2019) if we set Σ to 0 in (1). Fig.2 depicts the differences between the SDE- and ODE-based approaches. A dynamical system based on SDEs can stochastically represent data.

3. Stochastic Dynamical GANs

3.1. Mathematical Notations

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ be a complete and filtered probability space on which a d-dimensional Wiener process W_t is defined such that $\{\mathcal{F}_{t\geq 0}\}$ is the natural filtration according to W_t augmented by $\mathbb{P}$-null sets. In this paper, the real and fake probability measures are absolutely continuous with Lebesgue measure $\mathcal{L}$ on data space $\mathbb{R}^d$. In other words, we assume that probability densities p_t and q_t exist such that $d\mathbb{P}_t = p_t(x)d\mathcal{L}(x)$ and $d\mathbb{Q}_t = q_t(x)d\mathcal{L}(x)$. The push-forward operation is defined as $f_{\#}[\mathbb{P}](A) = \mathbb{P}(f^{-1}(A))$ for some set $A \subset \mathbb{R}^d$ with probability measure $\mathbb{P}$ and measurable function f. For simplicity, we will denote $\max(a, b) = a \vee b$ and $[c]_+ = \max(c, 0)$.

3.2. Problem Formulation

The conventional objective of learning WGAN (Arjovsky et al., 2017) is to find a generator network that minimizes the 1-Wasserstein distance between real $\mathbb{Q}$ and fake distributions $\mathbb{P}$ as follows:

$$\begin{aligned}\inf \mathcal{W}_1(\mathbb{P}, \mathbb{Q}) &= \inf_G \sup_D \mathbb{E}_{\mathbb{P}} D(X) - \mathbb{E}_{\mathbb{Q}} D(Y) \\ &= \inf_\theta \sup_\varphi \mathbb{E}_Z \left[D^\varphi \left(G^\theta(Z) \right) \right] - \mathbb{E}_{\mathbb{Q}} D^\varphi(Y),\end{aligned} \tag{4}$$

where G and D denote the generator and the 1-Lipschitzian discriminator network parameterized by θ and φ, respectively. This type of conventional generator takes a Gaussian random variable Z as input and is trained to produce a random variable as output, which is distributed by $\mathbb{P}^\theta$, *i.e.*, $G^\theta(Z) \sim \mathbb{P}^\theta$. Thus, the generated samples, $G(Z)$, represent random objects only for a fixed time.

In contrast to conventional methods, our goal is to find the best temporally evolving fake probability measure, $\mathbb{P}_t^\theta$, which is parameterized by neural networks with parameters θ. In particular, we aim to find a generator $G^\theta(X, t)$ that minimizes Wasserstein distance $\mathcal{W}$ between $\mathbb{Q}_t$ and $\mathbb{P}_t^\theta$ over t time sequences:

$$\begin{aligned}&\min_\theta \int_0^T \mathcal{W}\left(G^\theta_{\#}(\cdot, t)[\mathbb{P}_0], \mathbb{Q}_t \right) T(dt) \\ =&\min_\theta \int_0^T \mathcal{W}\left(\mathbb{P}_t^\theta, \mathbb{Q}_t \right) T(dt),\end{aligned} \tag{5}$$

where $\mathbb{P}_t^\theta = G^\theta_{\#}(\cdot, t)[\mathbb{P}_0]$ denotes the push-forward of $\mathbb{P}_0$ using generator $G^\theta(\cdot, t)$. In (5), time sequence measure T is defined as $T(dt) = \sum_{t_i \in \mathcal{T}} \delta_{t_i}(dt)$ for a strictly ordered set $\mathcal{T} = \{t_i\}$, where $t_i < t_j$ for all $i < j$, with $\max t_i = T$ and Dirac-delta function δ. We denote $|\mathcal{T}|$ as the number of time steps in $\mathcal{T}$. Thus, if we assume $|\mathcal{T}| \to \infty$, it approximates the time sequence entirely on $\mathbb{R}^+$

G^θ in (5) forces our fake stochastic process X_t to follow the real stochastic process $Y_t \sim \mathbb{Q}_t$.

$$\int G^\theta(X_s, t) d\mathbb{P}_s = \int X_t d\mathbb{P}_t \approx \int Y_t d\mathbb{Q}_t, \ \forall t \geq s, \tag{6}$$

where equality holds by the definition of the push-forward operation. We assume that the initial distributions of real and fake data are identical, *i.e.*, $\mathbb{Q}_0 = \mathbb{P}_0$. If we find the best probability measure, $\mathbb{P}_t^\theta$ well describes the real data distribution, $\mathbb{Q}_t$.

3.3. SDE-based Conditional Generator

To generate time-sequential data X_t, given the initial value X_0, we introduce a novel conditional generator $G^\theta : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}^d$. As an equivalent integral formulation of the SDE in (1), we can implement the proposed generator using the following formulation:

$$G^\theta(X_0, t) = X_t = X_0 + \int_0^t f(X_s, s, \theta) ds + \int_0^t \Sigma(X_s, \theta) dW_s, \tag{7}$$

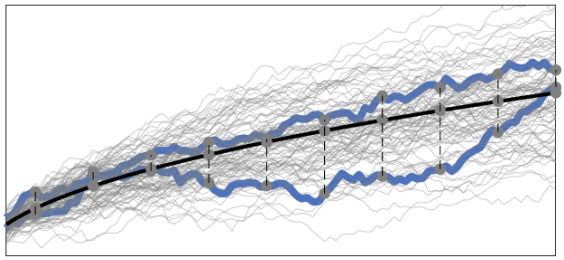

Figure 2. **Contributions of our stochastic diffusion term.** Our sample paths driven by SDEs can be represented stochastically (*i.e.*, blue lines). In contrast, conventional samples driven by ODEs are deterministic over temporal paths (*i.e.*, black lines).

where f denotes a twice differentiable function parameterized by neural network parameters θ, $f : \mathbb{R}^d \times \mathcal{T} \times \mathbb{R}^m \to \mathbb{R}^d$, and W_t denotes a d-dimensional Wiener process. Then, $G^\theta(X_0, t)$ in (7) is interpreted as a conditional generator summed over time t with stochastic noise W_s and temporal conditional code s. In particular, $G^\theta(X_0, t)$ can generate *stochastic sample paths* by gradually transforming X_0 to X_t using the stochastic dynamics defined in (7). If there are infinitesimal time changes $\Delta_s \approx 0$ such that $|t - s| = \Delta_s$, we can rewrite (7) as an approximated recursive formulation.

$$G^\theta(X_s, t) = X_t^\theta \approx X_s + f(X_s, s, \theta)\Delta_s + \Sigma(X_s, \theta)\sqrt{\Delta_s}Z, \tag{8}$$

where small changes in $f(s, \cdot)\Delta_s$ and a d-dimensional Gaussian random variable with small variance $\sqrt{\Delta_s}$ are added to generate X_t^θ at time $t = s + \Delta_s$. As the Gaussian random variable with small variance is added to every sample path to render infinitesimal random changes of objects, probability measures $\mathbb{P}_t$ can be defined at every continuous time $t \geq 0$.

3.4. Spatial-Temporal Constrained Discriminator

To discriminate between time-varying random objects X_t produced by our generator $G^\theta(X_0, t)$ in (7), the proposed discriminator takes an additional random variable t as an input. In particular, our discriminator is defined as a neural network $D^\varphi(X_t, t)$ parameterized by φ, where $D^\varphi : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}$. Similar to the generator, $D^\varphi(X_t, t)$ can be considered a conditional discriminator, in which the conditional code is set to a temporal state t.

To guarantee the stability of discriminators based on the min-max type of GAN objective functions, conventional vanilla GANs typically impose Lipschitz constraints on discriminator networks. For example, WGAN-LP (Petzka et al., 2018) adopts the p-Lipschitzness of discriminators to satisfy assumptions on the Kantorovich duality. In contrast, our discriminator takes two random variables (X_t, t) and needs to classify the time-evolving samples produced by SDE-based generators. Thus, it is quite challenging to make our discriminator more stable than vanilla GANs. Therefore, we introduce two Lipschitz constraints on the spatial ($\mathbb{R}^d$) and temporal ($\mathcal{T}$) domains as follows.

(Spatial Lipschitzness):

$$|D^\varphi(X, \cdot) - D^\varphi(Y, \cdot)| \leq q \|X - Y\|, \ \forall X \neq Y \in \mathbb{R}^d. \tag{9}$$

(Temporal Lipschitzness):

$$|D^\varphi(\cdot, t) - D^\varphi(\cdot, s)| \leq p|t - s|, \ \forall t \neq s \in \mathcal{T}. \tag{10}$$

If our discriminator satisfies the two conditions in (9) and (10), we denote $D^\varphi \in \mathbf{Lip}_p^q$.

To impose spatial-temporal constraints, we propose the following objective for the discriminator network:

$$W(\varphi, p, q) = \mathbb{E}_{X,Y,t,s}\left[\left(\frac{D^\varphi(\cdot, t) - D^\varphi(\cdot, s)}{|t - s|}\right) \vee \left(\frac{D^\varphi(X, \cdot) - D^\varphi(Y, \cdot)}{\|X - Y\|}\right) - p \vee q\right]_+ \tag{11}$$

for any $t \neq s \in \mathcal{T}, X \neq Y \in \mathbb{R}^d$. In (11), spatial-temporal constraints on our discriminator network enable stable learning to reduce undesired temporal perturbation induced by generated stochastic samples X_t.

To investigate the effect of constraint terms explicitly from a probabilistic point of view, we present the following proposition, which demonstrates that the probability for maximal perturbations in our discriminator is bounded by three major factors: spatial-temporal Lipschitz constants (p, q), diffusion terms Σ, and the norm of the Hessian matrix for our discriminator network, $\left\|\nabla^2 D^\varphi\right\|$.

Proposition 1. *(Controlled Stability of the Discriminator) Let $X_s = G^\theta(X_0, s)$ be a stochastic sample generated by (7), where T denotes the maximal element in $T \in \mathcal{T}$. Then, the following probability inequality is satisfied:*

$$\mathbb{P}\left[\sup_{0 \leq s \leq T} \|D^\varphi(X_s, s) - D^\varphi(X_0, 0)\| \geq \epsilon\right] \leq \frac{2}{3\epsilon}\Big\{(p \vee q)\mathbb{E}\|X_T - X_0\| + TC\}, \tag{12}$$

where a numerical constant C is linearly dependent on Σ and $\left\|\nabla^2 D^\varphi\right\|$ (i.e., $C \propto \Sigma, \left\|\nabla^2 D^\varphi\right\|$).

The proof of Proposition 1 can be naturally derived from the martingale property and Itô's lemma of Markov process X_t, governed by the dynamics in (7). According to the Proposition 1, the upper bound in (12) is controlled by the sum of the following three terms:

$$\underbrace{(p \vee q)}_{\text{Lipschitz Constraints}} + \underbrace{T\Sigma}_{\text{Stochastic variance}} + \underbrace{T\left\|\nabla^2 D^\varphi\right\|}_{\text{Hessian norm}}, \tag{13}$$

where Lipschitz constraints (p, q) are not dependent on time variables, and the second term is related to the expressivity

Algorithm 1 SD-GANs

Require: Neural networks f, D with initial weights θ_0, φ_0, respectively. Hyperparameters $\lambda = 10^{-3}$ and $p \vee q = 1$.
for $k = 1$ to K (*i.e.*, the total number of training iterations) **do**
1) Generate Markovian temporal samples using SDEs in (7) : $X_t = \int_0^t f ds + \int_0^t \Sigma dW_s$
2) Calculate gradients for the GAN objective of the generator network in (18) with the SC-regularization term in (20):
$\nabla_\theta^t = \nabla_\theta \left[\mathbb{E}_{X_t, \hat{X}_t} V^\lambda \left(\theta^k, X_t, \hat{X}_t \right) + \mathbb{E}_{x \sim \mathbb{P}_0} \mathcal{W}^{\varphi^k} \left(\mathbb{P}_t^{\theta^k} | x, \mathbb{Q}_t \right) \right]$
3) Calculate gradients for the GAN objective of the discriminator network in (18) with the ST-constraint term in (11):
$\nabla_\varphi^t = \nabla_\varphi \left[W(\varphi^k, p, q) - \mathbb{E}_{x \sim \mathbb{P}_0} \mathcal{W}^{\varphi^k} \left(\mathbb{P}_t^{\theta^k} | x, \mathbb{Q}_t \right) \right]$
4) Calculate the temporal expectation of gradients:
$\nabla_\theta^k \leftarrow \mathbb{E}_{t \in \mathcal{T}} \left[\nabla_\theta^t \right], \;\; \nabla_\theta^k \leftarrow \mathbb{E}_{t \in \mathcal{T}} \left[\nabla_\varphi^t \right]$
5) Update the generator and discriminator networks
$\theta^{k+1} = \theta^k - \nabla_\theta^k, \;\; \varphi^{k+1} = \varphi^k - \nabla_\varphi^k$
end for

of our generator. Implementing only the spatial constraint is not sufficient for the complete control of the stability of discriminator outputs, as the probabilistic upper bound is also related to the perturbations of the temporal axis (*i.e.*, q). Thus, without the control term over the temporal axis of $D^\varphi(x,t)$, the proposed dynamical system has the potential risk of failing to learn the GAN objective, while q is not bounded. To stabilize the discriminator network against temporal perturbation, we impose constraints to satisfy $p \vee q = 1$.

3.5. Conditional Markov Wasserstein Distance

Because real and fake samples depend on probability measures with different sample paths, we need to induce a temporal pathwise Wasserstein distance between $\mathbb{P}_t$ and $\mathbb{Q}_t$ for $t \in \mathcal{T}$. To this end, we present a novel conditional Wasserstein distance[3] with fake stochastic samples $\mathbb{P}_t \sim X_t$ started at $\mathbb{P}_0 \sim X_0 = x \in \mathbb{R}^d$, and real samples $\mathbb{Q}_t \sim Y_t$:

$$\mathcal{W}^\varphi(\mathbb{P}_t | x, \mathbb{Q}_t) = M_t D^\varphi(x,0) - \mathbb{E}_{Y_t \sim \mathbb{Q}_t} D^\varphi(Y_t, t), \quad (14)$$

where M_t denotes a *Markov semi-group* and is defined as follows:

$$M_t D^\varphi(x) = \mathbb{E}\left[D^\varphi(X_t, t) | X_0 = x \right]. \quad (15)$$

Then, based on the Markov property, a dual semi-group M_t^* exists such that the following equality holds:

$$\begin{aligned} &\int M_t D^\varphi(x,0) d\mathbb{P}_0 = \int D^\varphi(x,0) d(M_t^* \mathbb{P}_0)(x) \\ &= \int D^\varphi(y,t) p(t,y|s=0,x) p_0(x) d\mathcal{L}(x) \\ &= \int D^\varphi(X_t, t) d\mathbb{P}_t. \end{aligned} \quad (16)$$

The conditional formulation of the Wasserstein distance in (14) reveals the relation between the SDE in (1) and the Fokker-Planck equation in (2), where the transition density according to the solution of the Fokker-Planck equation, $p(\cdot, \cdot | 0, x)$, is explicitly used to define the expectation of the Markov-semi-group in (16).

Using equality in (16), we can obtain $\mathbb{E}_{\mathbb{P}_0 \sim x}[M_s D^\varphi(x,0)]$ = $\mathbb{E}[D^\varphi(X_s, s)]$ for any $s \le t$, which means that the expectation of the proposed conditional Wasserstein distance $\mathcal{W}(\mathbb{P}_t^\theta | x, \mathbb{Q}_t)$ in (14) with respect to the initial distribution is equivalent to a conventional Wasserstein distance.

$$\sup_{D \in \mathbf{Lip}_p^q} \mathbb{E}_{x \sim \mathbb{P}_0} \mathcal{W}\left(\mathbb{P}_s^\theta | x, \mathbb{Q}_s \right) = \mathcal{W}(\mathbb{P}_s^\theta, \mathbb{Q}_s). \quad (17)$$

Using the proposed conditional Markov Wasserstein distance in (14), our objective function $\mathcal{J}$ is defined as follows:

$$\begin{aligned} &\min_\theta \max_\varphi \mathcal{J}(\theta, \varphi) \\ &= \inf_\theta \int_0^t \sup_\varphi \mathbb{E}_{\mathbb{P}_0 \sim x} \mathcal{W}^\varphi \left(\mathbb{P}_s^\theta | x, \mathbb{Q}_s \right) dT(s). \end{aligned} \quad (18)$$

Proposition 2. *Let the objective function in* (18) *be solved by* $(G^\star, D^\star)$ *with parameters* $(\theta^\star, \varphi^\star)$. *Then,* $(G^\star, D^\star)$ *also solves the minimization problem in* (5).

Proposition 2 indicates that we can generate a fake probability measure $\mathbb{P}_t$ to imitate real probability measure $\mathbb{Q}_t$ using the proposed SDE-based generator and our conditional Markov Wasserstein distance. The overall procedure for the proposed method is summarized in Algorithm 1.

In the previous section, we investigated the probabilistic perturbation bounds of the discriminator network according to the bounded ST-constraints. The next proposition demonstrates the effect of the ST-constraints on our conditional Wasserstein distance in (14).

Proposition 3. *(Controlled Stability of the Wasserstein distance) Let us define the spatial-temporal gradient operator as* $\tilde{\nabla}_{x,t} = \nabla_x + \partial_t$. *Then, the expectation norm of the spatial-temporal gradient for conditional distance* (14)*, is*

[3] $\mathcal{W}(\mathbb{P}_t | x, \mathbb{Q}_t)$ is not a true distance on $\mathcal{P}_p(\mathbb{R}^d)$. Instead, it is a real-valued measurable function $\mathcal{W}(\mathbb{P}_t | x, \mathbb{Q}_t) : \mathbb{R}^d \to \mathbb{R}$.

bounded as follows:

$$\mathbb{E}_{x,t}\left[\left\|\tilde{\nabla}_{x,t}\mathcal{W}^{\varphi}(\mathbb{P}_t|x,\mathbb{Q}_t)\right\|\right] \leq C + (p\vee q)\frac{1+\kappa-e^{-\kappa T}}{\kappa} \tag{19}$$

for some numerical constants $\kappa, C > 0$.

The conditional formulation of the Wasserstein distance in (14) is vital for calculating the Kantorovich dual formulation (*i.e.*, the min-max GAN objective), and the stability of this term should be ensured during training. Stabilized training can be induced as the gradient is being forced to lie within the proposed ST-constraints.

4. Generalization to Out-of-Distribution Data

Let us consider a tuple $(\mathbb{Q}_t, \hat{\mathbb{Q}}_t)$ where $\mathbb{Q}_t$ and $\hat{\mathbb{Q}}_t$ indicate probability measures for training and out-of-distribution (OOD) datasets, respectively. Let us assume that generator $G^\theta(\cdot, t)$ is trained to represent $\mathbb{Q}_t$ by solving the objective in (18). In this section, we aim to answer the following important question: *Can our generator* $G^\theta(\hat{x}, t)$ *preserve the dynamics of generated samples with different initial states* $\hat{x} \sim \hat{\mathbb{Q}}_0$*?*

In other words, we want to investigate the effect of the Wasserstein distance between $\mathbb{P}_t$ and the newly propagated distribution $\hat{\mathbb{P}}_t$ given a different initial state $\hat{x} \sim \hat{\mathbb{Q}}_0$ (*i.e.*, out-of-distribution), $\mathcal{W}(\mathbb{P}_t, \hat{\mathbb{P}}_t)$. In particular, if initial state distribution $\hat{\mathbb{Q}}_0$ is arbitrary, then the distance is not controlled during the test time; thus, we cannot ensure that the generated images are semantically consistent with the original samples.

In this section, to suppress the semantic inconsistency during the test time, we introduce a novel *stochastic contraction* (SC)-constraint term on the generator network as follows.

$$V^\lambda(\theta, x, \hat{x}) = \mathbb{E}_{\alpha\sim U}\left[(x-\hat{x})^T\left[\nabla f|_{\alpha x+(1-\alpha)y} + \lambda I\right](\hat{x}-x)\right]_+, \tag{20}$$

where α is sampled from uniform distribution U on $[0, 1]$, and $\lambda > 0$ is a hyperparameter. The following proposition shows that semantic inconsistency is controlled if the generator minimizes V^λ:

Proposition 4. *Let* V^λ *be the function defined above, and* $x, \hat{x}$ *be two initial states such that* $\hat{\mathbb{P}}_t = h_\#[\mathbb{P}_t]$*. Assume that the generator satisfies the following constraint term:*

$$\min_\theta \mathbb{E}_{X_t\sim\mathbb{P}_t, \hat{X}_t\sim\hat{\mathbb{P}}_t} V^\lambda(\theta, X_t, \hat{X}_t) = 0. \tag{21}$$

Then, the following inequality holds:

$$\mathcal{W}_2(\mathbb{P}_t, \hat{\mathbb{P}}_t) \leq \sqrt{K\lambda^{-1} + e^{-2\lambda T}\|h\|_{\mathbf{L}_2(\mathbb{P})}}, \tag{22}$$

where $\|\cdot\|_{\mathbf{L}_2(\mathbb{P})}$ *denotes* L_2*-norm over probability measure* $\mathbb{P}$ *for some numerical constant* $K > 0$.

The proposition is a direct consequence of Theorem 2 (Pham et al., 2009). Moreover, we can obtain the super-martingale inequality:

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}}\left\|X_t - \hat{X}_t\right\|^2 \geq \epsilon\right) \leq K\lambda^{-1}e^{-\lambda T}\left\|X_0 - \hat{X}_0\right\|^2. \tag{23}$$

Both inequalities in (22) (in terms of distributional metric) and (23) (in terms of probabilistic concentration) indicate that generated samples $\hat{X}_t$ with different initial states are forced to be similar to X_t with the bounded exponential ratio.

We trained our generator network G^θ to minimize the regularization term in (21) for semantic consistency. For the numerical investigation of the generator representation with out-of-distribution initial states, we provide three examples.

(1) Noise Robustness. Let us assume that $h(x) = x + Z$, where $Z \sim \mathcal{N}(0, I_d)$ makes an initial state distorted by Gaussian random noise. Let $T = 1$ and $K = \lambda$ for convenience. In this case, $\mathbb{E}[\|h - Id\|^2] = \mathbb{E}\|Z\|^2 = d$, and the upper bound in (22) is induced as $\sqrt{1 + e^{-2\lambda t}d}$, which shows that although the initial distribution $\hat{\mathbb{Q}}_0^y$ is perturbed by Gaussian noise, the Wasserstein distance is not widely distorted. By considering the bounded Wasserstein distance, one can expect that generated samples $\hat{\mathbb{P}}_t$ will be similar to clean images $\mathbb{P}_t$ for a large t. Fig.4 (second row) shows the generated samples from the initial state with injected Gaussian noise. Artifacts are gradually removed as t increases, which demonstrates the effectiveness of the proposed regularizer. Contrary to our method, the generated samples from the ODE-based method in Fig.4 (first row) show that the injected noise in the initial state is still propagated by the dynamics.

(2) Fashion-MNIST. Let $y \sim \hat{\mathbb{Q}}_0^y$ be an initial state measure, in which random variable y indicates the samples from Fashion-MNIST. Fig.5 shows that the proposed method learns dynamic transitions of image rotations, although the initial state is sampled from unseen datasets. Specifically, the global structure of each object is preserved, and we can observe the image dynamics (*e.g.*, rotation).

(3) Temporal Interpolation. Let us assume that the objective function in (18) is defined on discrete time intervals $\mathcal{T} = \{t_0, \cdots, t_T\}$, where real data distribution $Y_t \sim q_t$ indicates the sequence of video data for a finite time $t \in \mathcal{T}$. As mentioned in Section 1, the proposed dynamical system generates a stochastic process X_t over a continuous temporal path $t \in \mathbb{R}^+$. Thus, the following question can be asked: *Can we generate stochastic samples given another time set* $\hat{\mathcal{T}}$ *with the generator originally trained on* $\mathcal{T}$*?*

To answer this question, we define a new time set $\hat{\mathcal{T}}_i$ such that $\mathcal{T} = \hat{\mathcal{T}}_0 \subset \hat{\mathcal{T}}_1 \cdots \subset \hat{\mathcal{T}}_L$, where $\hat{\mathcal{T}}_i$ is defined recursively as follows: $\hat{\mathcal{T}}_{i+1} = \left\{t_0^i, \frac{t_0^i+t_1^i}{2}, t_1^i, \frac{t_1^i+t_2^i}{2}, \cdots, t_T^i\right\}$,

Figure 3. **(Temporal interpolation)** To learn GANs, we only used samples from $X_{t\in\mathcal{T}}$ (highlighted in blue). Owing to the continuity of the proposed method, we can also generate every finite unseen stochastic sample $X_{t\in\hat{\mathcal{T}}\setminus\mathcal{T}}$ (highlighted in red).

Figure 4. **(Noise Robustness).** For the generative model trained on the Rot-MNIST dataset, the Gaussian noise is added to the initial state distribution. The upper and bottom figures show generated samples from ODE^{2}VAE (Yildiz et al., 2019) and SD-GAN, respectively.

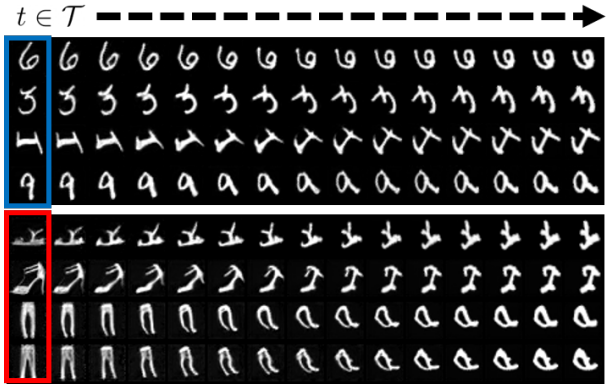

Figure 5. **(OOD sample generation)** For trained generator G^θ with the Rot-MNIST dataset, the upper figure indicates generated samples X_t^θ with an initial measure on the test dataset (blue box). We switched the initial measure to $y \sim \hat{\mathbb{Q}}_0^y$, where y indicates samples from Fashion-MNIST (red box).

and $\hat{T}_i = \{t_0^i \cdots t_T^i\}$. Thus, a total number of time intervals of $\hat{T}_i$ is 2^i-times larger than the original time set $\mathcal{T}$. The generated images, highlighted in blue in Fig.3 indicate that the images are sampled from the original temporal states $\mathcal{T}$, $\{G^\theta(\cdot,t)\}_{t\in\mathcal{T}}$ for learned generator G^θ. The images highlighted in red indicate temporally interpolated samples $\{G^\theta(\cdot,t)\}_{t\in\hat{\mathcal{T}}_3\setminus\mathcal{T}}$. As shown in Fig.3, samples can be temporally interpolated via smooth 2D image transitions, which verifies that our generator network can learn the stochastic dynamics of images with a relatively small number of temporal states.

Figure 6. **Generated samples for the Sprite Animation dataset.**

5. Experiments

5.1. Implementation Details

(Generator Network) Conventional generative models (Tulyakov et al., 2018; Xu et al., 2020; Yingzhen & Mandt, 2018) for time-sequential data typically use generic LSTM layers to encode conditional temporal transitions for a sequence of frames. In contrast, our generator network consists of two functions (*i.e.*, the drift function f and diffusion function Σ in (7)) with general convolutional layers. The drift function was implemented as a four-layer convolutional network, where temporal information t was fused into each convolutional layer through adaptive instance normalization. The diffusion function was implemented as a one-layer convolutional network. We implemented the numerical SDE-solver presented in (Li et al., 2020) to simulate the stochastic processes using the aforementioned two functions.

(Discriminator Network) To encode the temporal information of the time-sequential data, input data were augmented using the temporal bases of the Hilbert space $\mathcal{H}(\mathcal{L}(t))$. For example, our discriminator network takes an input in the form of $\tilde{X} = (X, \Phi^1, \Phi^2)$, where $\Phi^l(t) = \sin\left(\frac{l\pi t}{4}\right)$. Thus, in the case of 2D image data, temporal-augmented inputs have dimensions of $(C+2)\times H \times W$. We designed the discriminator network, which has a similar architecture to that of PatchGAN (Isola et al., 2017). However, the convolutional filters were smaller than those used in PatchGAN. Please refer to the supplementary materials for detailed architectures and hyperparameters settings.

Table 1. Image quality evaluation using various datasets. The best results are boldfaced.

	Sprites Animation		Human Actions		Rot-MNIST	
Methods	FID	KID	FID	KID	FID	KID
MoCoGAN, $\#3.5M$	2.65	2.81	2.42	1.95	–	–
DisVAE, $\#162M$	1.48	1.76	3.08	3.26	–	–
ODE^{2}VAE, $\#1.3M$	0.77	0.63	3.52	3.86	0.44	0.25
COT-GAN,	0.81	0.79	1.65	1.01	–	–
SD-GAN, $\#1.5M$	**0.35**	**0.37**	**0.88**	**0.83**	**0.31**	**0.22**

Figure 7. **(Human action videos)** The top, middle, and bottom rows show real data samples, generated samples from SD-GAN, and samples from MoCoGAN, respectively.

5.2. Qualitative Evaluation

(Rot-MNIST) Similar to (Yildiz et al., 2019), we conducted experiments on the MNIST dataset with temporal dynamics for image rotation. For example, the dataset is represented as $q_t \sim Y_t = \mathbf{R}_{\theta_t} Y_0$ for rotation matrix $\mathbf{R}_{\theta_t}$ with angle θ_t with $\{\theta_t\} = \{\theta_0 = 0, \cdots, \theta_T = \frac{\pi}{2}\}$. The total length of the sequences was set to $T = 16$. As shown in Fig.5, the samples generated by our generator network exhibited smooth changes in each digit, and the image styles were not fixed during the transformation.

(Human Actions) We trained our method on a human action video dataset (Gorelick et al., 2007). For data prepossessing, we followed the settings presented in (Tulyakov et al., 2018). In particular, we normalized each video to have $T = 16$ sequences as the maximal length, because different videos had inconsistent lengths. 72 videos were used to train the GANs. Each frame of the videos had dimensions of $3 \times 64 \times 64$. As shown in Fig.7, the proposed method produced smooth and realistic transitions for the frames, whereas previous methods (Tulyakov et al., 2018) produced blurry artifacts in frames.

(LPC-Sprite Animations) We trained our method on the LPC-Sprite dataset with animated cartoon characters, in which the visual styles for clothing and hairstyles can be controlled. The dataset was obtained from an open-source project page[4]. For more challenging tasks, we chose approximately $13K$ unique characters, which is a larger number than the original settings in (Yingzhen & Mandt, 2018) (*i.e.*, $1K$). Each generated data had $T = 8$ sequences of frames with dimensions of $3 \times 64 \times 64$. Fig.6 shows that the generated samples appear realistic and represent various temporal dynamics.

5.3. Quantitative Evaluation

To evaluate time sequential 2D image data, we used the Fréchet Inception distance (FID) (Heusel et al., 2017) and Kernel Inception distance (KID) (Bińkowski et al., 2018) as evaluation metrics. To evaluate the time sequential data with these metrics, we estimated the scores using 5000 generated images. In Table 4, the FID and KID scores were multiplied by 10^{-2} and 10^{1}, respectively. For comparison, we chose state-of-the-art temporal data generation methods, which are MocoGAN (Tulyakov et al., 2018), DisVAE (Yingzhen & Mandt, 2018), ODE^{2}VAE (Yildiz et al., 2019), and COT-GAN (Xu et al., 2020). Because compared methods had different network architectures, we reported the number of trainable learning parameters for the generative model for a fair comparison. As shown in Table 4, our method (**SD-GAN**) considerably outperformed other state-of-the-art methods in terms of both the FID and KID.

[4] http://lpc.opengameart.org

6. Conclusion

We presented a novel GAN describing the Markovian temporal dynamics. ST-constraints are suggested to stabilize the learning dynamics induced by the SDE-based generator network with respect to the controlled Fokker-Planck equation. The theoretical results reveal that the proposed ST-constraints stabilize the learning dynamics. To guarantee the generalization of the OOD dataset, an SC-regularization term is proposed, which induces a bounded Wasserstein distance of the generated samples with different initial states. The experimental results show that the proposed method produces a realistic and smooth transition between frames.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang university)).

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Esteban, C., Hyland, S. L., and Rätsch, G. Real-valued (medical) time series generation with recurrent conditional GANs, 2017.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition*, 2017.

Li, X., Wong, T.-K. L., Chen, R. T. Q., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Petzka, H., Fischer, A., and Lukovnikov, D. On the regularization of wasserstein GANs. In *International Conference on Learning Representations*, 2018.

Pham, Q.-C., Tabareau, N., and Slotine, J.-J. A contraction theory approach to stochastic incremental stability. *IEEE Transactions on Automatic Control*, 54(4):816–820, 2009.

Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. MocoGAN: Decomposing motion and content for video generation. In *Computer Vision and Pattern Recognition*, 2018.

Xu, T., Wenliang, L. K., Munn, M., and Acciaio, B. COT-GAN: generating sequential data via causal optimal transport. In *Advances in Neural Information Processing System*, 2020.

Yildiz, C., Heinonen, M., and Lahdesmaki, H. ODE2VAE: Deep generative second order odes with bayesian neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Yingzhen, L. and Mandt, S. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, 2018.

Yoon, J., Jarrett, D., and van der Schaar, M. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2019.