# Towards Theoretical Understandings of Self-Consuming Generative Models

**Shi Fu** [1]   **Sen Zhang** [2]   **Yingjie Wang** [3 4]   **Xinmei Tian** [1 5]   **Dacheng Tao** [6]

## Abstract

This paper tackles the emerging challenge of training generative models within a self-consuming loop, wherein successive generations of models are recursively trained on mixtures of real and synthetic data from previous generations. We construct a theoretical framework to rigorously evaluate how this training procedure impacts the data distributions learned by future models, including parametric and non-parametric models. Specifically, we derive bounds on the total variation (TV) distance between the synthetic data distributions produced by future models and the original real data distribution under various mixed training scenarios for diffusion models with a one-hidden-layer neural network score function. Our analysis demonstrates that this distance can be effectively controlled under the condition that mixed training dataset sizes or proportions of real data are large enough. Interestingly, we further unveil a phase transition induced by expanding synthetic data amounts, proving theoretically that while the TV distance exhibits an initial ascent, it declines beyond a threshold point. Finally, we present results for kernel density estimation, delivering nuanced insights such as the impact of mixed data training on error propagation.

## 1. Introduction

With the rapid advancements in deep generative models, synthetic data of all varieties is expanding swiftly. Notably, publicly accessible generative models, such as Stable Diffusion (Rombach et al., 2022) for images and ChatGPT (OpenAI, 2023) for text, have directly enabled the creation and dissemination of synthetic content at scale, thereby accelerating the flow of synthetic data towards the Internet. Consequently, this surge in synthetic data has led to a situation where even existing web-scale datasets are known to contain generated content (Schuhmann et al., 2022). Additionally, the identification of such generated content introduces distinctive technical challenges (Sadasivan et al., 2023; Huschens et al., 2023).

Despite potential risks, synthetic data is also being deliberately leveraged in various applications for several reasons. Firstly, generating synthetic training data offers a more efficient alternative to sourcing real-world samples, and has been shown to improve model performance through data augmentation (Antoniou et al., 2017; Azizi et al., 2023). Secondly, in sensitive domains such as medical imaging, synthetic data enables critical privacy protection (DuMont Schütte et al., 2021). More importantly, the expanding scale of deep generative models necessitates synthetic data, as these models are now trained on web-scale datasets that likely exhaust the supply of readily available real data on the internet (Villalobos et al., 2022). Thus, future generations of deep generative models will inevitably need to confront the presence of synthetic data in their training datasets. Consequently, a self-consuming training loop emerges in which future models are repeatedly trained on synthetic data generated from previous generations.

The study of generative models within the self-consuming loop has attracted substantial attention in current research. Empirical results from Shumailov et al. (2023) and Briesch et al. (2023) suggest that output diversity inevitably decreases after sufficient training generations. Alemohammad et al. (2023) perform experiments under various mixed training scenarios and conclude that injecting real data can mitigate model collapse. Despite these empirical observations, however, theoretical insights are still lacking. Shumailov et al. (2023) and Alemohammad et al. (2023) provide theoretical intuition by analyzing simple Gaussian toy models, but their approach targets intuitive understanding rather than in-depth analysis. Bertrand et al. (2023) further establishes an upper bound on the deviation of the output parameters of the likelihood-based generative model from optimal val-

[1]University of Science and Technology of China, Hefei, China [2]The University of Sydney, Sydney, Australia [3]College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China [4]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, China [5]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China [6]Nanyang Technological University, Singapore. Correspondence to: Shi Fu <fs311@mail.ustc.edu.cn>, Dacheng Tao <dacheng.tao@gmail.com>.

ues. However, a key limitation of Bertrand et al. (2023) is their direct assumption on the upper bounds of optimization errors and statistical errors resulting from finite sampling, rather than providing a rigorous theoretical analysis. Furthermore, their theoretical results are limited to parameter differences when training on mixed datasets comprising real data and synthetic data generated solely from the most recent generative model.

In contrast, our work aims to provide a comprehensive theoretical understanding of how training generative models, such as diffusion models and kernel density estimators, within self-consuming loops on various mixed datasets affects the fidelity of learned data distributions. We analyze this issue by exploring more general and diverse compositions of training data, going beyond simplistic assumptions. Additionally, we tackle the challenge of distributional discrepancy between synthetic and real-world data, transcending the limitations of analyzing model parameter discrepancy. Moreover, to overcome the direct assumptions on the bounds of statistical and optimization errors, we conduct tailored analyses of the training dynamics for specific generative models, including kernel density estimators and simplified diffusion models. Ultimately, this enables us to establish upper bounds on the TV distance. By eschewing simplistic assumptions and undertaking more nuanced analyses, our work offers key insights into the dynamics within this rapidly evolving domain of self-consuming generative modeling. The main contributions of this work include:

1. We propose a theoretical framework to assess the impact of training generative models within self-consuming loops. Specifically, We derive TV distance bounds between the original and the future learned data distributions under various mixed training scenarios for diffusion models with a one-hidden-layer neural network score function.

2. We provide requirements on sample sizes and proportions of real data to control errors. Notably, for the most extreme case of full synthetic data, we demonstrate the necessity of quartic sample growth or incorporating $\Omega((i-1)/i)$ proportions of real data in the final generation to restrict errors, where $i$ denotes the number of training generations.

3. We analyze the dynamics of increasing synthetic data on error propagation, unveiling a phase transition as synthetic data expands while real data remains fixed. Interestingly, before this transition point, more synthetic data impairs performance. However, beyond the threshold, incorporating additional synthetic data enhances performance.

4. We present a theoretical analysis of self-consuming non-parametric generative models, particularly deriving TV distance bounds between future learned synthetic data distributions and the original real data distribution using kernel density estimation and efficient decomposition techniques.
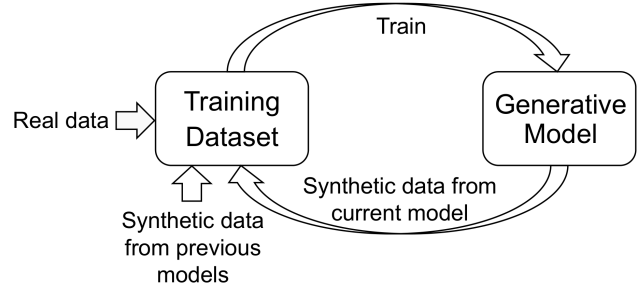


*Figure 1.* Self-consuming training loop: The initial model $\mathcal{G}_0$ is trained on real dataset $\mathcal{D}_0$. In each subsequent generation $i > 0$, $\mathcal{G}_i$ generates samples combined with previous synthetic data and real data into a new training dataset $D_i$. The new model $\mathcal{G}_{i+1}$ is then trained from scratch using $D_i$ for the next generation $i + 1$. This repeats until the maximum generation is reached.

## 2. Related Work

The study of generative models within a self-consuming loop has garnered significant attention recently. Current works primarily analyze this phenomenon from an empirical perspective. Shumailov et al. (2023) observe a degeneration of diversity for variational autoencoders and Gaussian mixture models when a portion of model outputs are recursively reused as inputs. Similarly, Briesch et al. (2023) examines the behavior of language models trained in a self-consuming loop from scratch, finding that while quality and diversity improve over initial generations, their output inevitably becomes less diverse after successive training iterations. Advocating the integration of real data, Alemohammad et al. (2023) performs experiments under various mixed training scenarios, concluding that injecting real data can mitigate model collapse. Martínez et al. (2023a) and Martínez et al. (2023b) further demonstrate that training generative models on web-scale datasets polluted by synthetic samples also corrodes the quality of generated data.

In contrast to the empirical observations, theoretical insights into self-consuming loops remain sparse. Both Shumailov et al. (2023) and Alemohammad et al. (2023) analyze a simple Gaussian toy model to provide theoretical intuition. Bertrand et al. (2023) establishes that the incorporation of real data can enhance stability within self-consuming loops under assumptions of infinite sample sizes and negligible initial model approximation error. Furthermore, they demonstrate, through direct assumptions on the upper bounds of statistical and optimization errors of generative models, the feasibility of achieving stability even with finite sample sizes. Comparatively, our work establishes theoretical bounds on the distributional discrepancy between synthetic and real-world data without relying on assumptions regarding the bounds of statistical and optimization errors.

# 3. Background

## 3.1. Self-Consuming Training of Generative Models

Generative models have advanced in synthesizing realistic data including imagery and text. The resulting synthetic data is widely distributed online and often indistinguishable from genuine content. As generative models evolve, datasets for model training would unintentionally (Schuhmann et al., 2022) or intentionally (Huang et al., 2022) include growing proportions of synthetic data alongside real-world samples. The resulting models, in return, create new content, leading to a cycle where successive generations train on datasets with increasingly synthetic proportions. This is termed as a self-consuming training loop, illustrated in Figure 1.

### 3.1.1. MODEL EVOLUTION IN SELF-CONSUMING LOOP

This paper explores a stochastic process with sequential generations. At generation $i$, we possess a training dataset of $n_i$ samples, $\{x_i^j\}_{j=1}^{n_i}$, independently and identically distributed, drawn from the distribution $p_i$. We denote the empirical distribution inferred from this dataset as $\bar{p}_i$. Notably, $p_0$ symbolizes the original distribution w.r.t. real data. Going from generation $i$ to generation $i+1$, our objective is to estimate distribution $p_i$ using samples $\{x_i^j\}_{j=1}^{n_i}$ via generative model, through parametric models $F_{\theta_{i+1}} : \bar{p}_i \to p_{\theta_{i+1}}$ or non-parametric estimators $\widehat{F}_{i+1} : \bar{p}_i \to \widehat{p}_{i+1}$. Here $p_{\theta_{i+1}}$ indicates a generative model parameterized by $\theta_{i+1}$, approximating $p_i$. For generation $i+1$, we resample training data from $p_{i+1} = \sum_{k=1}^{i+1} \beta_{i+1}^k p_{\theta_k} + \alpha_{i+1} p_0$ for parametric models or $p_{i+1} = \sum_{k=1}^{i+1} \beta_{i+1}^k \widehat{p}_k + \alpha_{i+1} p_0$ for non-parametric models, with non-negative parameters $\alpha_{i+1}$ and $\{\beta_{i+1}^k\}_{k=1}^{i+1}$ summing up to 1. The dataset, sourced from $p_{i+1}$, comprises a mixture of original data at $\alpha_{i+1}$ proportion, data generated by previous generations at $\{\beta_{i+1}^k\}_{k=1}^{i}$ proportions, and current model data at $\beta_{i+1}^{i+1}$ proportion.

Furthermore, the initial generative model is trained on the original dataset from $p_0$. Subsequently, each next generation of models is trained from scratch on a new dataset drawn from the mixed distribution $p_i$. This self-consuming loop repeats until reaching the maximum generation, as shown in Algorithm 1 in the appendix. We then define total variation distance, a metric used to compare probability distributions.

**Definition 3.1** (Total Variation Distance). Given two probability distributions $p$ and $q$ over a multidimensional space $\mathbb{R}^d$, the Total Variation Distance between $p$ and $q$ is:

$$TV(p,q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| \, dx.$$

### 3.1.2. DATA CYCLES

We introduce three different ways of mixing real and synthetic data in self-consuming training loops.

**General Data Cycle**. In this scenario, each model $\mathcal{G}_i$ (for $i \geq 1$) is trained on a mixture of real data and synthetic data from previous models $\{\mathcal{G}_j\}_{j=0}^{i-1}$. The training distribution $p_i$ is a weighted mixture of the original real data distribution $p_0$ and the synthetic data distributions $\{p_{\theta_j}\}_{j=1}^{i}$ from previous models, expressed as $p_i = \sum_{j=1}^{i} \beta_i^j p_{\theta_j} + \alpha_i p_0$. Furthermore, when utilizing non-parametric estimation, the training distribution $p_i$ is a weighted mixture of $p_0$ and the non-parametric synthetic distributions $\{\widehat{p}_j\}_{j=1}^{i}$, expressed as $p_i = \sum_{j=1}^{i} \beta_i^j \widehat{p}_j + \alpha_i p_0$.

**Full Synthetic Data Cycle**. This extreme case explores the use of training datasets that entirely comprise synthetic data recursively generated by the latest model, without real data. At generation $i$, the training distribution is defined either as the synthetic distribution $p_i = p_{\theta_i}$ or the non-parametric synthetic distribution $p_i = \widehat{p}_i$.

**Balanced Data Cycle**. This scenario blends real data from the original distribution and synthetic data from all previous models into a training distribution with equal contributions. This results in $p_i$ expressed as $\frac{1}{i+1}(p_0 + p_{\theta_1} + \cdots + p_{\theta_i})$ or $\frac{1}{i+1}(p_0 + \widehat{p}_1 + \cdots + \widehat{p}_i)$.

## 3.2. Background on Diffusion models

**Forward and Reverse processes.** Given a dataset $\mathcal{D}_x = \{x_i^j\}_{j=1}^{n_i} \subseteq \mathbb{R}^d$ with sample $x_i^j \overset{\text{i.i.d.}}{\sim} p_i(x)$, where $p_i$ denotes the target mixed distribution in the $i$-th generation, the forward diffusion process is:

$$dx_i = f(x_i, t)dt + g(t)dw_t, \quad x_i(0) \sim p_0,$$

where $w_t$ is the standard Wiener process, $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ is the *drift coefficient*, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is the *diffusion coefficient*. The reverse process generates $x_i(0) \sim p_i$ from $x_i(T) \sim p_{i,T}$. The reverse-time SDE (Anderson, 1982) is:

$$dx_i = \left[ f(x_i, t) - g(t)^2 \nabla_{x_i} \log p_{i,t}(x_i) \right] dt + g(t)d\bar{w}_t,$$

where $\bar{w}_t$ is a Wiener process from $T$ to 0, starting with $p_{i,T} \approx \pi$, with $\pi$ a known prior such as Gaussian noise.

**Loss objectives.** The task is to estimate the unknown Stein score function $\nabla_{x_i} \log p_{i,t}(x_i)$ by minimizing weighted denoising score matching objectives:

$$\mathcal{L}(\theta; \lambda(\cdot)) := \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(0) \sim p_{i,0}} \left[ \mathbb{E}_{x_i(t) \sim p_{i,t|0}} \left[ \right. \right. \right.$$
$$\left. \left. \left. \|s_{t,\theta}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t|0}(x_i(t)|x_i(0))\|_2^2 \right] \right] \right], \quad (1)$$

where $\theta^* := \arg\min_\theta \mathcal{L}(\theta; \lambda(\cdot))$ and $\mathcal{U}(0, T)$ denotes the uniform distribution. $\lambda(t) : [0, T] \to \mathbb{R}_+$ is a weighting function. The score function $s_{t,\theta} : \mathbb{R}^d \to \mathbb{R}^d$ can be parameterized as a neural network. The time-dependent

score-matching loss is:

$$
\widetilde{\mathcal{L}}(\theta; \lambda(\cdot)) := \mathbb{E}_{t \sim \mathcal{U}(0,T)} \big[ \lambda(t)
$$
$$
\mathbb{E}_{x_i(t) \sim p_{i,t}} \big[ \| s_{t,\theta}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \|_2^2 \big] \big] .
$$

**Training**. Let $\hat{\mathcal{L}}_{n_i}$ denote the Monte-Carlo estimation of $\mathcal{L}$ defined in Equation 1 on the training dataset. The gradient flows over the empirical loss and the population loss are:

$$
\frac{d}{d\tau} \hat{\theta}_i(\tau) = -\nabla_{\hat{\theta}_i(\tau)} \hat{\mathcal{L}}_{n_i}(\hat{\theta}_i(\tau); \lambda(\cdot)), \quad \hat{\theta}_i(0) := \theta_0,
$$
$$
\frac{d}{d\tau} \theta_i(\tau) = -\nabla_{\theta_i(\tau)} \mathcal{L}_{n_i}(\theta_i(\tau); \lambda(\cdot)), \quad \theta_i(0) := \theta_0.
$$

The learned score functions at training time $\tau$ for SDE time $t$ are $s_{t,\hat{\theta}_i(\tau)}(x_i(t))$ and $s_{t,\theta_i(\tau)}(x_i(t))$, respectively.

**Score networks.** The score function $s_{t,\theta}(x_i)$ is parameterized using the subsequent random feature model:

$$
\frac{1}{m_i} A\sigma(Wx_i + Ue(t)) = \frac{1}{m_i} \sum_{j=1}^{m_i} a_j \sigma(w_j^\top x_i + u_j^\top e(t)).
$$

Here $\sigma$ is ReLU activation, $A = (a_1, \ldots, a_{m_i}) \in \mathbb{R}^{d \times m_i}$ is trainable, $W = (w_1, \ldots, w_{m_i})^\top \in \mathbb{R}^{m_i \times d}$ and $U = (u_1, \ldots, u_{m_i})^\top \in \mathbb{R}^{m_i \times d_e}$ are initially randomized and remain fixed during training. The function $e : \mathbb{R}_{\geq 0} \to \mathbb{R}^{d_e}$ embeds time. Assuming $a_j$, $w_j$ and $u_j$ are i.i.d. from a distribution $\rho$, as $m_i \to \infty$, this approaches:

$$
\bar{s}_{t,\bar{\theta}}(x) = \mathbb{E}_{(w,u) \sim \rho_0} \big[ a(w,u)\sigma(w^\top x + u^\top e(t)) \big],
$$

where $a(w,u)$ is $\frac{1}{\rho_0(w,u)} \int_{\mathbb{R}^d} a\rho(a,w,u)da$ and $\rho_0(w,u)$ is $\int_{\mathbb{R}^d} \rho(a,w,u)da$. By the positive homogeneity of ReLU, we assume $\|w\| + \|u\| \leq 1$. The optimal solution is denoted by $\bar{\theta}^*$ when replacing $s_{t,\theta}(x)$ in loss objectives with $\bar{s}_{t,\bar{\theta}}(x)$.

The kernel $k_{\rho_0}(x, x')$ is defined as $\mathbb{E}_{(w,u) \sim p_0}[\sigma(w^\top x + u^\top e(t))\sigma(w^\top x' + u^\top e(t))]$, and denoted by $\mathcal{H}_{k_{\rho_0}}$, the induced reproducing kernel Hilbert space (RKHS). It follows that $\bar{s}_{t,\bar{\theta}} \in \mathcal{H}_{k_{\rho_0}}$ if the RKHS norm $\|\bar{s}_{t,\bar{\theta}}\|^2_{\mathcal{H}_{k_{\rho_0}}} = \mathbb{E}_{(w,u) \sim \rho_0}[\|a(w,u)\|_2^2]$ is finite. The discrete version is the empirical average $\|s_{t,\theta}\|^2_{\mathcal{H}_{k_{\rho_0}}} = \frac{1}{m} \sum_{j=1}^{m} \|a(w_j, u_j)\|_2^2$.

# 4. Theoretical Results for Diffusion Models within Self-Consuming Loops

In this section, we apply our theoretical framework to diffusion models with a one-hidden-layer neural network score function. The diffusion models have recently gained significant popularity due to their outstanding performance across various applications. Notable examples include DALL·E (Ramesh et al.) and Stable Diffusion (Rombach et al., 2022), both demonstrating notable advancements. However, the

curse of dimensionality renders it challenging to obtain upper bounds on the statistical errors and optimization errors that compound over successive training generations. Inspired by Yang (2022); Yang & Weinan (2022); Li et al. (2023), we conduct a more fine-grained analysis of this error propagation in diffusion models in the self-consuming loops. This enables us to obtain an upper bound on the TV distance between the synthetic data distributions produced by future models and the original real data distribution under various mixed training scenarios.

## 4.1. General Data Cycle: Flexibly Regulating Data Composition Mixture

This framework allows us to flexibly control the compositional mixture between real and synthetic data in the training distribution $p_i$. Our theoretical analysis quantifies the impacts of this adaptive training approach with mixed data on the cumulative error and the fidelity of models in self-consuming loops.

**Theorem 4.1.** *Suppose that $p_i$ is continuously differentiable and has a compact support set, i.e., $\|x\|_\infty$ is uniformly bounded, and there exists a RKHS $\mathcal{H}_{k_{p_0}}$ such that $\bar{s}_{0,\bar{\theta}^*} = \mathbb{E}_{(w,u) \sim \rho_0} \big[ a^*(w,u)\sigma(w^\top x + u^\top e(0)) \big] \in \mathcal{H}_{k_{p_0}}$. Suppose that the initial loss, trainable parameters, the embedding function $e(t)$ and weighting function $\lambda(t)$ are all bounded. Let $n_i$ be the number of training samples obtained from the distribution $p_i = \sum_{j=1}^{i} \beta_i^j p_{\theta_j} + \alpha_i p_0$. Choose $m_i \asymp n_i$ and $\tau_{i+1} \asymp \sqrt{n_i}$[1]. Then, with probability at least $1 - \delta$,*

$$
TV(p_{\theta_{i+1}}, p_0)
$$
$$
\lesssim \sum_{k=0}^{i} A_{i-k} \left( n_{i-k}^{-\frac{1}{4}} \sqrt{d \log \frac{di}{\delta}} + \sqrt{KL(p_{i-k,T} \| \pi)} \right),
$$

*where $A_i = 1$, $A_{i-k} = \sum_{j=i-k+1}^{i} \beta_j^{i-k+1} A_j$ for $1 \leq k \leq i$ and $\lesssim$ hides universal positive constants that depend solely on $T$.*

**Proof sketch of Theorem 4.1.** We begin by leveraging the triangle inequality to decompose the TV distance between $p_{\theta_{i+1}}$ and $p_0$ as follows:

$$
TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0)
$$

Next, we focus on bounding the first term $TV(p_{\theta_{i+1}}, p_i)$. Through an application of Pinsker's inequality, we relate this term to the KL divergence between $p_i$ and $p_{\theta_{i+1}}$. Subsequently, according to Theorem 1 in Song et al. (2021), the KL divergence can be upper bounded by the training loss $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}))$ up to a small error. To dissect the training loss $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}))$, we leverage several decompositions and partition it into multiple constituent

---

[1]We denote $B \asymp \widetilde{B}$ if there are absolute constants $c_1$ and $c_2$ such that $c_1 B \leq \widetilde{B} \leq c_2 B$.

terms: $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$, $\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\widetilde{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$, and $\bar{\widetilde{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$.

For the first term, $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$, we bound the gap through general norm estimates of parameters trained under gradient flow dynamics, as delineated in Li et al. (2023), in conjunction with classical analyses leveraging Rademacher complexity. For the second term, $\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\widetilde{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$, we derive the bound via gradient flow analysis and properties of the RKHS norm. Regarding the third term, $\bar{\widetilde{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$, we establish the bound through properties of the score function space.

By analyzing each constituent term and selecting the model width $m_i \asymp n_i$, optimal early-stopping training time $\tau_{i+1} \asymp \sqrt{n_i}$, in accordance with properties of the underlying data distribution, we arrive at upper bounds for $TV(p_{\theta_{i+1}}, p_i)$. Finally, we substitute this bound back into the original decomposition to obtain the overall bound on $TV(p_{\theta_{i+1}}, p_0)$. We then derive the final result through recursive solving.

*Remark* 4.2. **Comparision with Previous Works.** In the field of self-consuming generative modeling, Shumailov et al. (2023) and Alemohammad et al. (2023) provided foundational insights through their analysis of a simplistic multivariate Gaussian toy model. However, their approach prioritized providing theoretical intuition instead of a detailed and rigorous theoretical analysis. A more pertinent comparison to our research can be drawn with Theorem 2 in Bertrand et al. (2023).

Bertrand et al. (2023) made significant strides by establishing an upper bound on the deviation of generative model output parameters from the optimal parameters, denoted as $\|\theta_{i+1} - \theta^*\|$. This was achieved under specific assumptions on the upper bounds of statistical and optimization errors in generative models, as outlined in their Assumption 3.

Our research diverges from Bertrand et al. (2023)'s work in several key aspects: (1) We eschew the direct assumption of upper bound constraints on statistical and optimization errors. Instead, our analysis delves deeply into the nature of statistical errors by employing kernel density estimation theory and concentration inequalities for a nuanced understanding. (2) We specifically examine optimization errors in generative models, particularly focusing on diffusion models as elaborated in Section 4. This approach allows us to circumvent the need for direct assumptions on errors. (3) Our analysis extends beyond the scope of Bertrand et al. (2023). While they focused on the parameter difference $\|\theta_{i+1} - \theta^*\|$, our research tackles the more complex task of assessing the distributional discrepancy between synthetic and real data, represented as $TV(p_{\theta_{i+1}}, p_0)$, which is inherently more challenging since the optimal estimator does not perfectly mirror the original real data distribution $p_0$. (4) Finally, our assumptions regarding the composition of the

training set are more general and practical compared to those of Bertrand et al. (2023). We consider a training set that incorporates data from all previous generations alongside real data, offering a broader and more realistic foundation for training the generative model. This stands in contrast to their assumption, which limits the training set to real data and synthetic data from only the most recent generation.

In summary, our work extends beyond existing literature by overcoming assumptions on statistical and optimization error bounds, assessing distributional discrepancy between synthetic and real data, and employing more realistic assumptions about the training dataset.

*Remark* 4.3. **Optimal Early-Stopping Strategy**. When we select $m_i \asymp n_i$ and omit the $\sqrt{d \log(d/\delta)}$ term, the bound for $TV(p_{\theta_{i+1}}, p_i)$ in Theorem 4.1 can be expressed as:

$$\sum_{k=0}^{i} A_{i-k} \left( \frac{\tau_{i+1-k}^{3/2}}{n_{i-k}} + \frac{1}{\sqrt{\tau_{i+1-k}}} + \sqrt{KL(p_{i-k,T}\|\pi)} \right).$$

By optimally choosing the early-stopping time $\tau_{i+1} \asymp \sqrt{n_i}$, we arrive at the final result. In essence, our analysis provides that through early stopping, the TV distance can be controlled when retraining diffusion models on mixed datasets over successive generations. The key insight is that by early-stopping the training at the optimal time, we prevent overfitting to the training distribution, thereby controlling the discrepancy between the synthetic data distribution and the original data distribution.

*Remark* 4.4. **Real Data Integration Across Generations.** Previous experimental results suggest that when the proportion of real data is sufficiently large, the error can be effectively controlled (Alemohammad et al., 2023; Shumailov et al., 2023). This remark examines the impact of incorporating real data at each generation on the theoretical outcomes for diffusion models. Specifically, we assume $p_j = \alpha p_0 + (1-\alpha)p_{\theta_j}$ for $1 \le j \le i$ and $0 < \alpha < 1$. Additionally, we assume that the training set sizes and $KL$ terms are of the same order of magnitude across all generations. This setup allows us to derive the following bound:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim$$

$$\left(1 - (1-\alpha)^{i+1}\right) \alpha^{-1} (n^{-\frac{1}{4}} \sqrt{d \log \frac{di}{\delta}} + \sqrt{KL(p_{i,T}\|\pi)}).$$

Based on our theoretical results, we observe that to control the error, the requirement on the proportion of real data added at each generation is more relaxed compared to the approach discussed in Remark 4.7 in the following subsection, where real data is only added in the final generation. Specifically, while Remark 4.7 requires the proportion of real data to increase with the number of training generations, i.e., $\alpha = \Omega(\frac{i-1}{i})$, this analysis shows that if real data is added at each generation, a suitable constant value for

$\alpha$ is sufficient to effectively control the error. Additionally, due to the presence of term $\alpha^{-1}$, the value of $\alpha$ should not be too small, as it may lead to difficulties in controlling the error. For instance, as $\alpha$ tends to 0, employing Taylor expansion, we observe that the error accumulates linearly with the increase of generation $i$.

### 4.2. Full Synthetic Data Cycle: Error Control with Quartically Sampling or High Real Data Ratio

Here we thoroughly analyze the fully synthetic data cycle, wherein each model is trained using synthetic data from the most recent generative model. Although in practical data collection processes, new datasets typically retain portions of original real data even in successive generations, analyzing this extreme synthetic-only training loop provides valuable theoretical insights. Specifically, our theoretical analysis shows that the error can be controlled by increasing the sampling quantity quartically across generations or by injecting sufficient real data in the final generation.

**Corollary 4.5** (Worst Case). *Let $n_i$ represent the number of training samples obtained from the distribution $p_i$ at the $i$-th generation. We define $p_i$ as $p_i = p_{\theta_i}$. Then, with probability at least $1 - \delta$,*

$$TV(p_{\theta_{i+1}}, p_0)$$
$$\lesssim \sum_{k=1}^{i} \left( n_k^{-\frac{1}{4}} \sqrt{d \log \frac{di}{\delta}} + \sqrt{KL(p_{k,T}\|\pi)} \right).$$

*Remark* 4.6. **Controlling Error Accumulation with Quartic Sample Growth.** According to the classical results in Van Handel (2014), since $\pi$ (e.g., the Gaussian density) is log-Sobolev, $KL(p_{i,T}\|\pi)$ is exponentially small in $T$. To better elucidate the behavior of models in the extreme full synthetic case, we first posit that $KL(p_{k,T}\|\pi) = \mathcal{O}(\epsilon^2/i^2)$ for $1 \leq k \leq i$. Consequently, the result of Corollary 4.5 indicates that the number of training samples $n_i$ must grow quartically, specifically as $n_k = \Omega\left((i\sqrt{d}/\epsilon)^4\right)$ for $1 \leq k \leq i$, in order to constrain the TV distance to $\mathcal{O}(\epsilon)$.

Intuitively, in the absence of any grounding from real data, errors can accumulate rapidly as the model at each generation is trained only on synthetic samples generated by the model from the latest generation. Without adequate samples to properly approximate the training distribution, the statistical error and the inherent sample bias accumulate.

To counteract this, each successive generation necessitates an increasingly expansive training set to provide sufficient coverage of the distribution. Our analysis quantifies this requirement, showing a quartic growth in the training samples is imperative to control the error.

*Remark* 4.7. **Integrating Sufficient Real Data in the Final Generation for Error Control**. In addition to the discussion in Remark 4.6, that involves incrementally raising the

sample size through quartic growth across generations, an alternative approach is to increase the proportion of original data in the final generation of training to control the error. To theoretically investigate the role of the real data from the original distribution, we assume $p_i = \alpha p_0 + (1 - \alpha)p_{\theta_i}$, $p_j = p_{\theta_j}$ for $1 \leq j \leq i - 1$ and $KL(p_{j,T}\|\pi) = \mathcal{O}(\epsilon^2/i^2)$ for all $j$. Additionally, we assume that the training set sizes are of the same order of magnitude across all generations, specifically, $n_j = \mathcal{O}(d^2/\epsilon^4)$ for $0 \leq j \leq i$. Under these assumptions, we obtain the bound:

$$TV(p_{\theta_{i+1}}, p_0) = \mathcal{O}((1 - \alpha)i\epsilon).$$

In particular, to ensure $TV(p_{\theta_{i+1}}, p_0) = \mathcal{O}(\epsilon)$, it suffices to have the proportion $\alpha = \Omega(\frac{i-1}{i})$. This analysis reveals that as the number of generations increases, the proportion $\alpha$ of data from the original distribution must be progressively augmented to constrain the error to $\mathcal{O}(\epsilon)$. This suggests that later generational models have greater compounding drift from the original distribution, necessitating a larger fraction of real samples for grounding. By incorporating sufficient proportions of real data in the final generation of training, we can then control the error and avoid the requirement of quartic sample growth.

### 4.3. Balanced Data Cycle: Optimizing Sample Efficiency Across Generations

In this section, we analyze the scenario of balanced data cycle, wherein the training distribution comprises a uniform mixture of the original data distribution and synthetic data distributions from all previous generative models. Moreover, we investigate the sample complexity $\{n_j\}_{j=0}^{i}$ required to control the error over generations. Interestingly, our analysis demonstrates that the requisite number of samples progressively decreases as the number of generations grows.

**Corollary 4.8.** *Define $p_i = \frac{1}{i+1}(p_0 + p_{\theta_1} + p_{\theta_2} + \cdots + p_{\theta_i})$ for $i \geq 1$. Let $n_i$ be the number of training samples obtained from the distribution $p_i$ at the $i$-th generation. With probability at least $1 - \delta$,*

$$TV(p_{\theta_{i+1}}, p_0) \lesssim n_i^{-\frac{1}{4}} \sqrt{d \log \frac{di}{\delta}} + \sqrt{KL(p_{i,T}\|\pi)} +$$
$$\sum_{k=0}^{i-1} \sum_{j=k}^{i-1} \frac{\Gamma(j+2)}{\Gamma(i+2)} \left( n_k^{-\frac{1}{4}} \sqrt{d \log \frac{di}{\delta}} + \sqrt{KL(p_{k,T}\|\pi)} \right),$$

*where the Gamma function $\Gamma(j) = (j - 1)!$ and $j$ is a positive integer.*

*Remark* 4.9. **Diminishing Sample Complexity Over Generations**. In the scenario of balanced data cycle, controlling the total accumulation of errors necessitates sufficiently large training sample sizes for the initial generations of generative models. Intriguingly, as the number of generations increases, the demand for training samples progressively

decreases. We suppose $KL(p_{k,T}||\pi) = \mathcal{O}(\epsilon^2/i^2)$ for all $k$. By selecting $n_k = \mathcal{O}\left(\left(\frac{(i+1)\sqrt{d}}{\epsilon}\sum_{j=k}^{i-1}\frac{\Gamma(j+2)}{\Gamma(i+2)}\right)^4\right)$ for $0 \le k \le i-1$ and $n_i = \mathcal{O}\left(d^2/\epsilon^4\right)$, we obtain:

$$TV(p_{\theta_{i+1}}, p_0) = \mathcal{O}(\epsilon).$$

Critically, the required training samples gradually decrease with increasing generations. In particular, the number of training samples required for the first generation is $n_0 = \mathcal{O}\left(\left(\frac{(i+1)\sqrt{d}}{\epsilon}\sum_{j=0}^{i-1}\frac{\Gamma(j+2)}{\Gamma(i+2)}\right)^4\right)$, while for the $i$-th generation, the required training samples are $n_i = \mathcal{O}\left(\left(\frac{\sqrt{d}}{\epsilon}\right)^4\right)$.

Furthermore, this finding aligns with intuition, as data from earlier generative models is incorporated into training sets earlier, thereby influencing more generations of models and having a greater impact on the cumulative error across the entire process. Therefore, to ensure the fidelity of the initially generated data, more training data is needed to constrain the error.

### 4.4. Phase Transition in Error Dynamics With Synthetic Data Augmentation

While numerous studies have empirically demonstrated the benefit of synthetic data in enhancing model performance (Azizi et al., 2023; Burg et al., 2023), the established error bounds in Section 4.1 to 4.3 for various mixed training scenarios instead raise skepticism regarding whether synthetic data could potentially impair model performance during self-consuming training. The reality appears more nuanced than a simple binary categorization. In this section, we conduct a theoretical analysis by keeping the real data fixed and flexibly varying the numbers of synthetic samples. Interestingly, in the absence of sampling bias, we identify a regime where modest amounts of synthetic data can degrade performance. However, beyond a certain threshold, increasing synthetic data improves model performance.

**Corollary 4.10.** *Define $p_i = \frac{n}{n+m}p_0 + \frac{m}{n+m}p_{\theta_i}$ for $i \ge 1$. Suppose that the $KL$ terms are of the same order of magnitude across all generations. Let $n+m$ be the number of training samples obtained from the distribution $p_i$ at the $i$-th generation. With probability at least $1-\delta$,*

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \left(1 + \frac{m}{n}\right)\left(1 - \left(\frac{m}{n+m}\right)^{i+1}\right)$$
$$\times \left((n+m)^{-\frac{1}{4}}\sqrt{d\log\frac{di}{\delta}} + \sqrt{KL(p_{i,T}||\pi)}\right).$$

*Remark* 4.11. **The Impact of Increasing Synthetic Data on Error Propagation**. There is a tradeoff when fixing the number of training samples $n$ from the original distribution and increasing the number of synthetic samples $m$ generated by the model. On the one hand, augmenting the training set with additional synthetic samples $m$ can reduce the statistical error and estimation error for each generation. This approach is intuitive, as synthetic data transfers knowledge from previously used real data to subsequent generations, thereby enlarging the effective size of the dataset. On the other hand, as $m$ increases, the proportion of original training samples in the mixture distribution $p_i = (n/(n+m))p_0 + (m/(n+m))p_{\theta_i}$ decreases. This exacerbates distribution shift from the original data distribution $p_0$, accumulating errors over successive generations.

Therefore, our results imply a tradeoff between reducing statistical and estimation errors for each generation by adding more synthetic data, and controlling the cumulative effects of distribution shift over successive generations by reducing the proportion of synthetic data. This trade-off is also recognized as the phase transition as evidenced by the experimental findings in Alemohammad et al. (2023). To further investigate this phenomenon, we posit that $m = \lambda n$ where $\lambda > 0$ and disregard the $KL$ term, as it is exponentially small in $T$. Under these assumptions, the result described in Corollary 4.10 is as follows:

$$TV(p_{\theta_{i+1}}, p_0) = \mathcal{O}\left(\frac{1}{n^{1/4}}\frac{(1+\lambda)^{i+1} - \lambda^{i+1}}{(1+\lambda)^{i+\frac{1}{4}}}\right).$$

Our focus is on the impact of synthetic data on the error, which is implicated by the effect of $\lambda$. Consider $f(\lambda, i) = ((1+\lambda)^{i+1} - \lambda^{i+1})/(1+\lambda)^{i+\frac{1}{4}}$. Clearly, as $\lambda$ approaches infinity, by the Taylor's Formula, $f$ converges to $(i+1)(1+\lambda)^{-1/4}$, meaning that $f$ approaches 0. This implies that as we incorporate infinitely abundant synthetic data, the TV distance becomes significantly small.

Notably, in contrast to the experimental results in Alemohammad et al. (2023), where errors continuously decrease with increased $\lambda$ under unbiased sampling beyond Gaussian modeling, our theory indicates that $f$ experiences an initial increase with a rise in $\lambda$ but subsequently decreases after a phase transition point. Crucially, there is no analytical solution expressing the critical value $\lambda'$ at this phase transition point in terms of the number of training generations $i$. Nevertheless, numerical analysis reveals that $\lambda'$ is positively correlated with $i$, increasing as $i$ does.

This aligns with intuition, as the cumulative effects of distribution shift compound over successive generations with larger $i$. To counteract the exacerbated drift from the original distribution for later generative models, it becomes imperative to substantially reduce the statistical and estimation errors for each individual generation. This necessitates further expanding the training set by incorporating more synthetic data, which transfers knowledge from previously utilized real data to subsequent generations. In this way, the phase transition point $\lambda'$ where errors decrease given fixed real data, shifts larger as generations continue.

# 5. Theoretical Results for KDE within Self-Consuming Loops

This section presents our theoretical analyses of self-consuming non-parametric generative models. Specifically, we derive TV distance bounds between synthetic data distributions produced by future models and the original real data distribution under various mixed training scenarios. We introduce kernel density estimation (KDE), a widely utilized non-parametric density estimation method (Devroye & Lugosi, 2001). Let $x_i^1, \cdots, x_i^{n_i}$ be i.i.d. random variables in $\mathbb{R}^d$ with density $p_i$. Then, the kernel estimate is defined by:

$$\widehat{p}_{i+1}(x) = \frac{1}{n_i h_i^d} \sum_{j=1}^{n_i} K\left(\frac{x - x_i^j}{h_i}\right), \qquad (2)$$

where $h_i > 0$ is a bandwidth parameter and $K$ is a kernel.

The following standard notations are used. A vector $\alpha = (\alpha_1, \ldots, \alpha_d)$, composed of nonnegative integers $\alpha_j$ is defined as a multi-index. We denote $|\alpha| = \alpha_1 + \ldots + \alpha_d$, $\alpha! = \alpha_1! \cdots \alpha_d!$ and for a vector $x_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$, we define $x_i^\alpha = x_{i,1}^{\alpha_1} \cdots x_{i,d}^{\alpha_d}$. The partial derivative of the function $p_i$ is represented as $\partial^\alpha p_i = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} p_i$. Additionally, for $s \geq 0$, $W^{s,1}(\mathbb{R}^d)$ is the Sobolev space of functions $p_i$ whose weak ("distributional") partial derivatives $\partial^\alpha p_i$, $|\alpha| \leq s$, are integrable.

The Sobolev assumption is not only very common in KDE (Kroll, 2021; Cleanthous et al., 2019), but also frequently employed in more complex generative models, such as Transformers (Fonseca et al., 2023), GANs (Mroueh et al., 2017), and VAEs (Turinici, 2019).

In the following Lemma 5.1, we provide a theoretical guarantee on controlling the discrepancy between the kernel estimate $\widehat{p}_{i+1}$ and the true underlying distribution $p_i$ in the $i$-th generation. This serves as a key preliminary step before relating $\widehat{p}_{i+1}$ to the distance original data distribution $p_0$. By bounding the gap between $\widehat{p}_{i+1}$ and $p_i$, the lemma establishes a foundation for controlling the accumulation of statistical errors from finite sample estimates under the self-consuming loop framework.

**Lemma 5.1.** *Let $s \geq 1$. Suppose $p_i \in W^{s,1}(\mathbb{R}^d)$, and for some $\epsilon > 0$, $\int |x|^{d+\epsilon} p_i(x)\, dx < \infty$ for all $i$. Let $n_i$ be the number of training samples obtained from distribution $p_i$ in the $i$-th generation. By appropriately selecting the kernel $K$ as defined in Definition G.1 in the appendix, and setting the bandwidth parameter $h_i = n_i^{-1/(2s+2d)}$, the following holds with probability at least $1 - \delta$:*

$$TV(\widehat{p}_{i+1}, p_i) \leq \frac{1}{2} n_i^{-\frac{2s+d}{4s+4d}} (1 + \gamma_{n_i}) \sqrt{\int K^2 \int \sqrt{p_i}}$$

$$+ \frac{1}{2} n_i^{-\frac{s}{2s+2d}} \varphi(s, K, p_i) + n_i^{-\frac{s}{2s+2d}} \sqrt{\frac{1}{2} \left(\int |K|\right)^2 \log \frac{2}{\delta}},$$

*where $\gamma_{n_i} \to 0$ as $n_i \to \infty$ and $\varphi(s, K, p_i)$ is a finite function.*

**Proof sketch of Lemma 5.1.** The proof utilizes the kernel density estimation framework (Holmström & Klemelä, 1992; Devroye & Lugosi, 2001) to bound the TV distance between the non-parametric distribution $\widehat{p}_{i+1}$ estimated from finite samples and the true underlying distribution $p_i$. First, McDiarmid's inequality is applied to show that with high probability, the TV distance between $\widehat{p}_{i+1}$ and $p_i$ is bounded by the expected TV distance $\mathbb{E} \int |\widehat{p}_{i+1} - p_i|$ plus a concentration term that decays as $n_i^{-1/2} h_i^{-d}$, where $h_i$ is the kernel bandwidth. Next, the expected TV distance is decomposed into a bias term $\int |p_i * K_{h_i} - p_i|$ representing the distance between $p_i$ convolved with the kernel $K_{h_i}$ and $p_i$, and a variation term $\mathbb{E} \int |\widehat{p}_{i+1} - p_i * K_{h_i}|$. Using Taylor expansion, properties of the kernel, and Young's inequality, this bias term is shown to decay at the rate $h_i^s$, where $s$ depends on the smoothness of $p_i$. As for the variation term, by applying the Schwarz inequality and Carlson's inequality, this term decays as $(n_i h_i^d)^{-1/2}$. Finally, by optimally selecting the kernel bandwidth $h_i = n_i^{-1/(2s+2d)}$, the bias term, variation term, and concentration term can be balanced, leading to the final TV bound. The key steps involve utilizing kernel density theory and concentration inequalities.

*Remark 5.2.* Regarding the existing work related to Lemma 5.1, Jiang (2017) derives $L_\infty$ density estimation bounds for KDE. On the other hand, Kroll (2021) focuses on the adaptive minimax density estimation problem. As for the $L_1$ error, which is the focus of our paper, Cleanthous et al. (2019), Holmström & Klemelä (1992), and Devroye & Lugosi (2001) establish upper bounds in expectation. In contrast, Lemma 5.1 utilizes concentration inequalities to derive high-probability finite-sample bounds in the $L_1$ norm.

*Remark 5.3.* When $s \ll d$, the rate implies the need for samples that are exponential in the dimension $d$. However, we would like to clarify that this situation arises only for highly non-smooth functions. It is worth noting that in previous works on self-consuming generative models (Bertrand et al., 2023; Alemohammad et al., 2023), the commonly assumed multivariate Gaussian distribution satisfies $s = d$, in which case our rate $O(n_i^{-\frac{s}{2s+2d}}) = O(n_i^{-\frac{1}{4}})$ does not suffer from the curse of dimensionality.

Next, we propose the TV distance bound for KDE within self-consuming loops.

**Theorem 5.4.** *Let $s \geq 1$. we suppose $p_i \in W^{s,1}(\mathbb{R}^d)$, and for some $\epsilon > 0$, $\int |x|^{d+\epsilon} p_i(x)\, dx < \infty$ for all $i$. Let $n_i$ represent the number of training samples obtained from the distribution $p_i$ at the $i$-th generation. We define $p_i$ as $p_i = \beta_i^1 \widehat{p}_1 + \beta_i^2 \widehat{p}_2 + \cdots + \beta_i^i \widehat{p}_i + \alpha_i p_0$. Then, with probability*

*at least* $1 - \delta$,

$$TV(\widehat{p}_{i+1}, p_0)$$
$$\lesssim \sum_{k=0}^{i} A_{i-k} \left( n_{i-k}^{-\frac{s}{2s+2d}} \sqrt{\log(i/\delta)} + n_{i-k}^{-\frac{2s+d}{4s+4d}} \right),$$

*where* $A_i = 1$, $A_{i-k} = \sum_{j=i-k+1}^{i} \beta_j^{i-k+1} A_j$ *for* $1 \leq k \leq i$ *and* $\lesssim$ *hides universal positive constants that depend solely on* $K$, $p_i$, *and* $s$.

**Proof sketch of Theorem 5.4**. To establish an upper bound on the TV distance between the non-parametric distribution $\widehat{p}_{i+1}$ of future generative models and the original real data distribution $p_0$, we introduce an effective decomposition for $TV(\widehat{p}_{i+1}, p_0)$:

$$TV(\widehat{p}_{i+1}, p_0) \leq TV(\widehat{p}_{i+1}, p_i) + TV(p_i, p_0).$$

Firstly, the distance between $\widehat{p}_{i+1}$ and $p_i$ is bounded using kernel density estimation theory and concentration inequalities per Lemma 5.1, showing that it decays as $n_i^{-s/(2s+2d)} \sqrt{\log(2/\delta)} + n_i^{-(2s+d)/(4s+4d)}$ with high probability. Secondly, the distance between the mixed distribution $p_i$ and the original distribution $p_0$ is recursively expanded and bounded by a weighted sum of distances between the synthetic data distribution generated by previous generative models and the original distribution. By combining the bounds for the three components, the objective $TV(\widehat{p}_{i+1}, p_0)$ can be upper bounded by a term decaying as $n_i^{-s/(2s+2d)} \sqrt{\log(2/\delta)} + n_i^{-(2s+d)/(4s+4d)}$ plus a weighted sum of errors from previous generations. The final result can be obtained by solving it recursively.

The key steps involve utilizing kernel density estimators to control the error in estimating $p_i$ from finite samples, leveraging the triangle inequality to decompose the overall distance, and recursively bounding the distance between mixed training distributions across generations.

## 6. Conclusion

In this paper, we addressed the emerging challenge of training generative models within a self-consuming loop, where successive generations of models are recursively trained on mixtures of real and synthetic data from previous generations. We constructed a robust theoretical framework to evaluate the impact of this training procedure on the data distributions learned by future models, encompassing both parametric and non-parametric models.

Our work specifically focused on deriving bounds on the TV distance between the synthetic data distributions produced by future models and the original real data distribution under various mixed training scenarios. For diffusion models with a one-hidden-layer neural network score function,

we demonstrated that the TV distance could be effectively controlled by ensuring that mixed training dataset sizes or proportions of real data are sufficiently large. Furthermore, our analysis revealed a phase transition induced by the increasing amounts of synthetic data. We provided theoretical proof that while the TV distance initially increases, it eventually declines beyond a threshold point. This indicates that a balanced approach to incorporating synthetic data can enhance the performance of future generative models.

We also presented a detailed theoretical analysis of self-consuming non-parametric generative models. By employing kernel density estimation and efficient decomposition techniques, we derived TV distance bounds between future learned synthetic data distributions and the original real data distribution. In future work, it would be interesting to extend our theoretical results to biased sampling scenarios.

## Acknowledgements

## Impact Statement

This paper presents theoretical analyses that aim to advance the understanding of training generative models within self-consuming loops, where models are recursively trained on mixtures of real and synthetic data. A potential impact of this paper is guiding practitioners on the standardized use of synthetic data for training generative models. Using synthetic data can also help protect privacy in sensitive domains such as medical imaging. There are no unethical aspects of this paper or anticipated negative effects on society. This theoretical analysis aims to advance the field of machine learning through rigorous study of an emerging challenge.

## References

Adams, R. A. and Fournier, J. J. *Sobolev spaces*. Elsevier, 2003.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Antoniou, A., Storkey, A., and Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Bertrand, Q., Bose, A. J., Duplessis, A., Jiralerspong, M., and Gidel, G. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429*, 2023.

Briesch, M., Sobania, D., and Rothlauf, F. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.

Burg, M. F., Wenzel, F., Zietlow, D., Horn, M., Makansi, O., Locatello, F., and Russell, C. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023.

Cleanthous, G., Georgiadis, A. G., and Porcu, E. Minimax density estimation on sobolev spaces with dominating mixed smoothness. *arXiv preprint arXiv:1906.06835*, 2019.

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

DuMont Schütte, A., Hetzel, J., Gatidis, S., Hepp, T., Dietz, B., Bauer, S., and Schwab, P. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021.

Fonseca, A. H. D. O., Zappala, E., Caro, J. O., and Van Dijk, D. Continuous spatiotemporal transformer. In *International Conference on Machine Learning*, pp. 7343–7365. PMLR, 2023.

Han, J., Hu, R., and Long, J. A class of dimensionality-free metrics for the convergence of empirical measures. *arXiv preprint arXiv:2104.12036*, 196, 2021.

Holmström, L. and Klemelä, J. Asymptotic bounds for the expected l1 error of a multivariate kernel density estimator. *Journal of multivariate analysis*, 42(2):245–266, 1992.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Huschens, M., Briesch, M., Sobania, D., and Rothlauf, F. Do you trust chatgpt?–perceived credibility of human and ai-generated content. *arXiv preprint arXiv:2309.02524*, 2023.

Jiang, H. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pp. 1694–1703. PMLR, 2017.

Kontorovich, A. Concentration in unbounded metric spaces and algorithmic stability. In *International conference on machine learning*, pp. 28–36. PMLR, 2014.

Kroll, M. On density estimation at a fixed point under local differential privacy. 2021.

Li, P., Li, Z., Zhang, H., and Bian, J. On the generalization properties of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ma, C., Wang, Q., et al. A priori estimates of the population risk for residual networks. *arXiv preprint arXiv:1903.02154*, 2019.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023a.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. Towards understanding the interplay of generative artificial intelligence and the internet. *arXiv preprint arXiv:2306.06130*, 2023b.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

OpenAI, O. Gpt-4 technical report. Mar 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models.

*Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021.

Stein, E. M. *Singular integrals and differentiability properties of functions*. Princeton university press, 1970.

Turinici, G. X-ray sobolev variational auto-encoders. 2019.

Van Handel, R. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., and Ho, A. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 2022.

Wu, L. and Su, W. J. The implicit regularization of dynamical stability in stochastic gradient descent. *arXiv preprint arXiv:2305.17490*, 2023.

Yang, H. A mathematical framework for learning probability distributions. *arXiv preprint arXiv:2212.11481*, 2022.

Yang, H. and Weinan, E. Generalization and memorization: The bias potential model. In *Mathematical and Scientific Machine Learning*, pp. 1013–1043. PMLR, 2022.

## A. Self-Consuming Loop of Generative Models

In this section, we present the algorithms describing the self-consuming training loop.

---

**Algorithm 1** Self-Consuming Loop of Generative Models

---

**Input:** generative model $\mathcal{G}$, proportional coefficients $\{\alpha_i\}_{i=1}^I$, $\{\{\beta_i^k\}_{k=1}^i\}_{i=1}^I$.
**Initialize** $\mathcal{D}_0 = \{x_i^j\}_{j=1}^{n_0}$, with $x_i^j \sim p_0$, $p_{\theta_1} = \mathcal{G}_0(\mathcal{D}_0)$.
**for** $i = 1$ **to** $I$ **do**
    $p_i = \beta_i^1 p_{\theta_1} + \cdots + \beta_i^i p_{\theta_i} + \alpha_i p_0$
    $\mathcal{D}_i = \{x_i^j\}_{j=1}^{n_i}$, with $x_i^j \sim p_i$
    $p_{\theta_i} = \mathcal{G}_i(\mathcal{D}_i)$
**end for**
**Return**

---

## B. Auxiliary Lemmas

In this section, we introduce McDiarmid's inequality, a tool that enables us to establish a bound on the probability that a function of multiple independent random variables deviates from its expected value.

**Lemma B.1** (McDiarmid's Inequality). *Consider independent random variables $Z_1, \cdots, Z_n \in \mathcal{Z}$ and a mapping $\phi : \mathcal{Z}^n \to \mathbb{R}$. If, for all $i \in \{1, \cdots, n\}$, and for all $z_1, \cdots, z_n, z_i' \in \mathcal{Z}$, the function $\phi$ satisfies*

$$|\phi(z_1, \cdots, z_{i-1}, z_i, z_{i+1}, \cdots, z_n) - \phi(z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_n)| \le c,$$

*then,*

$$P(|\phi(Z_1, \cdots, Z_n) - \mathbb{E}\phi(Z_1, \ldots, Z_n) \ge t|) \le 2\exp\left(\frac{-2t^2}{nc^2}\right).$$

*Furthermore, for any $\delta \in (0,1)$ the following inequality holds with probability at least $1 - \delta$*

$$|\phi(Z_1, \ldots, Z_n) - \mathbb{E}[\phi(Z_1, \ldots, Z_n)]| \le \frac{c\sqrt{n\log(2/\delta)}}{\sqrt{2}}.$$

## C. Proof of Theorem 4.1

In this section, we present the proof of Theorem 4.1, which establishes TV distance bounds for diffusion models.

**Lemma C.1** (Theorem 1 in Song et al. (2021)). *Let $p_i$ be the data distribution, $\pi$ be a known prior distribution. Then, we have*

$$KL\left(p_i \| p_{\hat{\theta}_{i+1}(\tau)}\right) \le \tilde{\mathcal{L}}\left(\hat{\theta}_{i+1}(\tau); g^2(\cdot)\right) + KL(p_{i,T} \| \pi).$$

**Lemma C.2** (Lemma 4 in Li et al. (2023)). *For any $\tau > 0$ and $\theta, \tilde{\theta}$, we have*

$$\bar{\tilde{\mathcal{L}}}(\bar{\theta}(\tau)) - \bar{\tilde{\mathcal{L}}}(\bar{\theta}) \lesssim \frac{\|\bar{s}_{0,\bar{\theta}_0}\|_{\mathcal{H}}^2 + \|\bar{s}_{0,\bar{\theta}}\|_{\mathcal{H}}^2}{\tau}, \quad \tilde{\mathcal{L}}(\theta(\tau)) - \tilde{\mathcal{L}}(\theta) \lesssim \frac{\|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \|s_{0,\theta}\|_{\mathcal{H}}^2}{\tau}.$$

**Lemma C.3** (Theorem A.5 in (Wu & Su, 2023)). *Consider a function class $\mathcal{F}$ with $\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| \le B$. For any $\delta \in (0,1)$, w.p. at least $1 - \delta$ over the choice of $S = (x_1, x_2, \ldots, x_n)$, we have,*

$$\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_x[f(x)]\right| \lesssim \mathcal{R}_n(\mathcal{F}) + B\sqrt{\frac{\ln(2/\delta)}{n}},$$

*where $\mathcal{R}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.*

**Lemma C.4** (Lemma 5 in Li et al. (2023)). *Suppose that the loss objectives $\tilde{\mathcal{L}}, \tilde{\mathcal{L}}_{n_i}, \bar{\tilde{\mathcal{L}}}, \bar{\tilde{\mathcal{L}}}_{n_i}$ are bounded at the initialization, then for any $\tau > 0$, we have*

$$\|s_{0,\theta(\tau)}\|_{\mathcal{H}}, \|s_{0,\hat{\theta}_i(\tau)}\|_{\mathcal{H}} \lesssim \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i}, \quad \|\bar{s}_{0,\bar{\theta}(\tau)}\|_{\mathcal{H}}, \|\bar{s}_{0,\hat{\bar{\theta}}_i(\tau)}\|_{\mathcal{H}} \lesssim \|\bar{s}_{0,\bar{\theta}_0}\|_{\mathcal{H}} + \sqrt{\tau}.$$

**Lemma C.5** (Lemma 6 in Li et al. (2023))**.** *Suppose that $\|x(0)\|_\infty \leq 1$, and the trainable parameter $a$ and embedding function $e(\cdot)$ are both bounded. Then, given any $\bar{\theta}$, for any $\delta > 0$, with the probability of at least $1 - \delta$, there exists $\theta$ such that*

$$\mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \|s_{t,\theta}(x_i(t)) - \bar{s}_{t,\bar{\theta}}(x_i(t))\|_2^2 \right] \right] \lesssim \frac{\log^2(1/\delta^2)}{m_i} d,$$

*where $\lesssim$ hides universal positive constants only depending on $T$.*

*Proof of Theorem 4.1.* By utilizing the triangle inequality of TV distance, we can deduce the following:

$$TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0).$$

For the second term, $TV(p_i, p_0)$, recalling the definition of the TV distance as $TV(p_i, p_0) = \frac{1}{2} \int |p_i(x) - p_0(x)|\, dx$ and given $p_i(x) = \beta_i^1 p_{\theta_1}(x) + \cdots + \beta_i^i p_{\theta_i}(x) + \alpha_i p_0(x)$, we can deduce:

$$TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + \beta_i^i TV(p_{\theta_i}, p_0) + \beta_i^{i-1} TV(p_{\theta_{i-1}}, p_0) + \cdots + \beta_i^1 TV(p_{\theta_1}, p_0). \tag{3}$$

Next, let's concentrate on the first term, $TV(p_{\theta_{i+1}}, p_i)$. By applying Pinsker's inequality, we obtain:

$$TV(p_{\theta_{i+1}}, p_i) \leq \sqrt{\frac{1}{2} KL(p_i \| p_{\theta_{i+1}})}. \tag{4}$$

Considering the training process of the diffusion model, in the $i + 1$-th generation, the output is denoted as $\theta_{i+1}$, which we also represent as $\hat{\theta}_{i+1}(\tau_{i+1})$, with a total of $n_{i+1}$ training samples and a training time of $\tau_{i+1}$. Then, by employing Lemma C.1, we derive:

$$KL\left(p_i \| p_{\hat{\theta}_{i+1}}(\tau_{i+1})\right) \leq \widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}); g^2(\cdot)) + KL\left(p_{i,T} \| \pi\right). \tag{5}$$

In our approach, we adhere to the framework established by Li et al. (2023) to establish bounds for the term $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}); g^2(\cdot))$. For this purpose, we employ the following decomposition:

$$\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) = \left[ \widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) \right] + \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$$

$$\leq \left[ \widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) \right] + \left[ \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1})) \right] + \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1})). \tag{6}$$

Initially, regarding the final term $\bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$, we can derive the following result by employing Lemma C.2:

$$\bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1})) \lesssim \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*) + \frac{1}{\tau} \left( \|\bar{s}_{0,\bar{\theta}_0}\|_{\mathcal{H}}^2 + \|\bar{s}_{0,\bar{\theta}^*}\|_{\mathcal{H}}^2 \right). \tag{7}$$

For the second term, $\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$, we perform the following decomposition:

$$\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$$

$$\leq \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \|\mathbf{s}_{t,\theta_{i+1}(\tau_{i+1})}(x_i(t)) - \bar{\mathbf{s}}_{t,\bar{\theta}_{i+1}(\tau_{i+1})}(x_i(t))\|^2 \right] \right]$$

$$\lesssim \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \|\mathbf{s}_{t,\theta_{i+1}(\tau_{i+1})}(x_i(t)) - \mathbf{s}_{t,\theta_{i+1}^*}(x_i(t))\|^2 \right] \right]$$

$$+ \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \|\mathbf{s}_{t,\theta_{i+1}^*}(x_i(t)) - \bar{\mathbf{s}}_{t,\bar{\theta}_{i+1}^*}(x_i(t))\|^2 \right] \right]$$

$$+ \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t) \cdot \mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \|\bar{\mathbf{s}}_{t,\bar{\theta}_{i+1}^*}(x_i(t)) - \bar{\mathbf{s}}_{t,\bar{\theta}_{i+1}(\tau_{i+1})}(x_i(t))\|^2 \right] \right]$$

$$=: I_1 + I_2 + I_3, \tag{8}$$

where $\theta_{i+1}^*$ is the Monte Carlo estimator of $\bar{\theta}_{i+1}^*$. According to Lemma C.5, it is derived that, with a probability of at least $1 - \delta$,

$$I_2 \lesssim \frac{\log^2(1/\delta^2)}{m_i} d. \tag{9}$$

From the triangle inequality and Lemma C.2, we can obtain:

$$I_3 \lesssim \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1})) + \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*) \lesssim \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*) + \frac{1}{\tau}\left(\|\bar{s}_{0,\bar{\theta}_0}\|_{\mathcal{H}}^2 + \|\bar{s}_{0,\bar{\theta}^*}\|_{\mathcal{H}}^2\right),$$

and similarly,

$$I_1 \lesssim \widetilde{\mathcal{L}}(\theta_{i+1}^*) + \frac{1}{\tau}\left(\|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \|s_{0,\theta^*}\|_{\mathcal{H}}^2\right).$$

Consequently, it can be established that, with a probability of at least $1 - \delta$,

$$\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1})) - \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}(\tau_{i+1}))$$
$$\lesssim \frac{\log^2(1/\delta^2)}{m_i}d + \widetilde{\mathcal{L}}(\theta_{i+1}^*) + \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*) + \frac{1}{\tau}\left(\|\bar{s}_{0,\bar{\theta}_0}\|_{\mathcal{H}}^2 + \|\bar{s}_{0,\bar{\theta}^*}\|_{\mathcal{H}}^2 + \|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \|s_{0,\theta^*}\|_{\mathcal{H}}^2\right). \tag{10}$$

For the first term $\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$, we have:

$$\sqrt{\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}))} - \sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))} \lesssim \left\{\mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{x_i(t)\sim p_{i,t}}\left[\lambda(t)\left\|s_{t,\hat{\theta}_{i+1}(\tau_{i+1})}(x_i(t)) - s_{t,\theta_{i+1}(\tau_{i+1})}(x_i(t))\right\|^2\right]\right\}^{\frac{1}{2}} =$$

$$\left\{\mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{x_i(t)\sim p_{i,t}}\left[\lambda(t)\left\|\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\alpha}_j(\tau_{i+1})\sigma(w_j^\top x_i(t) + u_j^\top e(t)) - \frac{1}{m_i}\sum_{j=1}^{m_i}\alpha_j(\tau_{i+1})\sigma(w_j^\top x_i(t) + u_j^\top e(t))\right\|^2\right]\right\}^{\frac{1}{2}}$$

Furthermore, by the triangle inequality, the Cauchy-Schwartz inequality, and the fact that $\sigma(y) = ReLU(y) \le |y|$ for any $y \in \mathbb{R}$, Hölder's inequality, the positive homogeneity property of the ReLU activation, and the boundedness of the input data, embedding function $e(t)$, and weighting function $\lambda(t)$, we have:

$$\left\|\frac{1}{m_i}\sum_{j=1}^{m_i}(\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1}))\sigma(w_j^\top x_i(t) + u_j^\top e(t))\right\|^2$$
$$\le \frac{1}{m_i{}^2}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\sum_{j=1}^{m_i}|\sigma\left(w_j^\top x_i(t) + u_j^\top e(t)\right)|^2$$
$$\le \frac{2}{m_i{}^2}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\sum_{j=1}^{m_i}(|w_j^\top x_i(t)|^2 + |u_j^\top e(t)|^2)$$
$$\le \frac{2}{m_i{}^2}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\sum_{j=1}^{m_i}(\|w_j\|_1^2\|x_i(t)\|_\infty^2 + \|u_j\|_1^2\|e(t)\|_\infty^2)$$
$$\lesssim \frac{1}{m_i{}^2}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\sum_{j=1}^{m_i}(C_{T,\delta}^2 + C_{T,e}^2), \tag{11}$$

where $C_{T,\delta}$ and $C_{T,e}$ are constants arising from the boundedness of $\|x\|_\infty$ and $e(t)$. Then, we have

$$\sqrt{\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1}))} - \sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))} \lesssim \left[\frac{1}{m_i}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\left(C_{T,\delta}^2 + C_{T,e}^2\right)\right]^{\frac{1}{2}}.$$

$$\lesssim (C_{T,\delta} + C_{T,e})\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2\right]^{\frac{1}{2}}.$$

14

Furthermore, we conclude that:

$$\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$$

$$\lesssim \frac{1}{m_i} \sum_{j=1}^{m_i} \|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2 (C_{T,\delta}^2 + C_{T,e}^2) + \sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))}(C_{T,\delta} + C_{T,e}) \left[ \frac{1}{m_i} \sum_{j=1}^{m_i} \|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2 \right]^{\frac{1}{2}}$$

$$\lesssim \sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*) + \frac{1}{\tau}(\|\mathbf{s}_{0,\theta_0}\|_{\mathcal{H}}^2 + \|\mathbf{s}_{0,\theta^*}\|_{\mathcal{H}}^2)}(C_{T,\delta} + C_{T,e}) \left[ \frac{1}{m_i} \sum_{j=1}^{m_i} \|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2 \right]^{\frac{1}{2}}$$

$$+ \frac{1}{m_i} \sum_{j=1}^{m_i} \|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|^2 (C_{T,\delta}^2 + C_{T,e}^2), \tag{12}$$

where the last equality follows from lemma C.2. We further deduce that:

$$\frac{1}{m_i} \sum_{j=1}^{m_i} \|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|_2^2$$

$$= \frac{1}{m_i} \sum_{j=1}^{m_i} \left\| \int_0^{\tau_{i+1}} \frac{d}{d\tau}(\hat{\alpha}_j(\tau) - \alpha_j(\tau))d\tau \right\|_2^2$$

$$= \frac{1}{m_i} \sum_{j=1}^{m_i} \left\| \int_0^{\tau_{i+1}} \left( \nabla_{\theta_{i+1}^j(\tau)} \tilde{\mathcal{L}}(\theta_{i+1}(\tau)) - \nabla_{\hat{\theta}_{i+1}^j(\tau)} \hat{\tilde{\mathcal{L}}}_{n_i}(\hat{\theta}_{i+1}(\tau)) \right) d\tau \right\|_2^2$$

$$= \frac{1}{m_i^2} \sum_{j=1}^{m_i} \left\| \int_0^{\tau_{i+1}} \left( 2\mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t)\mathbb{E}_{x_i(t) \sim p_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right] \right. \right.$$

$$\left. \left. - 2\mathbb{E}_{t \sim \mathcal{U}(0,T)} \left[ \lambda(t)\mathbb{E}_{x_i(t) \sim \hat{p}_{i,t}} \left[ \left( s_{t,\hat{\theta}_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right] \right) d\tau \right\|_2^2, \tag{13}$$

where $\hat{p}_{i,t}$ denotes the standard empirical distribution of $p_{i,t}$. Note that:

$$\|\theta\|_2^2 = \|\text{vec}(A)\|_2^2/m = \|A\|_F^2/m = \|s_{0,\theta}\|_{\mathcal{H}}^2, \tag{14}$$

By lemma C.4, we get $\|\theta(\tau)\|_2 = \|s_{0,\theta(\tau)}\|_{\mathcal{H}} \lesssim \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m}$. For any $t \in [0, T]$, we define the function space as follows:

$$\mathcal{F}_t := \{f_1(x_i(t); \theta_1(\tau))f_2(x_i(t); \theta_2) : f_1 \in \mathcal{F}_{1,t}, f_2 \in \mathcal{F}_{2,t}\}, \tag{15}$$

where

$$\mathcal{F}_{1,t} := \left\{ s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) : \|\theta_{i+1}(\tau)\|_2 \lesssim \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} \right\},$$

$$\mathcal{F}_{2,t} := \left\{ \sigma(w^\top x_i(t) + u^\top e(t)) : \|w\|_1 + \|u\|_1 \leq 1 \right\}.$$

Subsequently, according to Lemma C.3, for any $\delta \in (0, 1)$, there exists a probability of at least $1 - \delta$ that the dataset

$\mathcal{D}_{x_i} = \{x_i^j\}_{j=1}^{n_i}$ chosen satisfies the following:

$$\mathbb{E}_{x_i(t)\sim p_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right]$$

$$- \mathbb{E}_{x_i(t)\sim \hat{p}_{i,t}} \left[ \left( s_{t,\hat{\theta}_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right]$$

$$\leq \left| \mathbb{E}_{x_i(t)\sim p_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right.$$

$$\left. - \mathbb{E}_{x_i(t)\sim \hat{p}_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right|$$

$$+ \left| \mathbb{E}_{x_i(t)\sim \hat{p}_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right.$$

$$\left. - \mathbb{E}_{x_i(t)\sim \hat{p}_{i,t}} \left[ \left( s_{t,\hat{\theta}_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \right|$$

$$\lesssim \mathcal{R}_{n_i}(\mathcal{F}_t) + \sup_{f\in\mathcal{F}_t, x_i(t)\in[-C_{T,\delta}, C_{T,\delta}]^d} |f(x_i(t))| \sqrt{\frac{\log(2/\delta)}{n_i}}$$

$$+ \mathbb{E}_{x_i(t)\sim \hat{p}_{i,t}} \left[ \left( s_{t,\theta_{i+1}(\tau)}(x_i(t)) - s_{t,\hat{\theta}_{i+1}(\tau)}(x_i(t)) \right) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right] \quad := J_1 + J_2 + J_3, \qquad (16)$$

where $\mathcal{R}_{n_i}(\mathcal{F}_t)$ represents the empirical Rademacher complexity of the function space $\mathcal{F}_t$ on the dataset $\mathcal{D}_{x_i} = \{x_i^j\}_{j=1}^{n_i}$. For the $J_1$ component, as per Lemma A.6 in Wu & Su (2023), we obtain:

$$\mathcal{R}_{n_i}(\mathcal{F}_t) \leq \left( \sup_{f_1\in\mathcal{F}_{1,t}, x_i(t)\in[-C_{T,\delta}, C_{T,\delta}]^d} |f_1(x_i(t))| + \sup_{f_2\in\mathcal{F}_{2,t}, x_i(t)\in[-C_{T,\delta}, C_{T,\delta}]^d} |f_2(x_i(t))| \right) (\mathcal{R}_{n_i}(\mathcal{F}_{1,t}) + \mathcal{R}_{n_i}(\mathcal{F}_{2,t})).$$

$$(17)$$

Note that

$$|\sigma(w^\top x_i(t) + u^\top e(t))| \lesssim \|w\|_1 \|x_i(t)\|_\infty + \|u\|_1 \|e(t)\|_\infty \lesssim C_{T,\delta} + C_{T,e}. \qquad (18)$$

Then we obtain:

$$\left| s_{t,\theta_{i+1}(\tau)}(x_i(t)) \right| = \left| \frac{1}{\sqrt{m_i}} \sum_{j=1}^{m_i} \theta_{i+1}^j(\tau) \sigma(w_j^\top x_i(t) + u_j^\top e(t)) \right|$$

$$\lesssim (C_{T,\delta} + C_{T,e}) \|\theta_{i+1}(\tau)\|_2$$

$$\lesssim (C_{T,\delta} + C_{T,e}) \left( \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} \right), \qquad (19)$$

Subsequently, we define $C'_{T,\delta} := \max_{x_i(t)\in[-C_{T,\delta}, C_{T,\delta}]^d} \left| \nabla_{x_i(t)} \log p_{i,t}(x_i(t)) \right|$. Then, we can deduce that:

$$|f_1(x_i(t))| = |s_{t,\theta_{i+1}(\tau)}(x_i(t)) - \nabla_{x_i(t)} \log p_{i,t}(x_i(t))| \lesssim (C_{T,\delta} + C_{T,e}) \left( \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} \right) + C'_{T,\delta},$$

$$|f_2(x_i(t))| = \sigma(w^\top x_i(t) + u^\top e(t)) \lesssim C_{T,\delta} + C_{T,e}.$$

By substituting the above two equations into inequality 17, we can obtain

$$\mathcal{R}_{n_i}(\mathcal{F}_t) \lesssim (C_{T,\delta} + C_{T,e}) \left( \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} + 1 \right) (\mathcal{R}_{n_i}(\mathcal{F}_{1,t}) + \mathcal{R}_{n_i}(\mathcal{F}_{2,t})). \qquad (20)$$

Define $\mathcal{F}'_{1,t} := \left\{ s_{t,\theta_{i+1}(\tau)}(x_i(t)) : \|\theta_{i+1}(\tau)\|_2 \lesssim \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} \right\}$. According to Lemma 26.6 in (Shalev-Shwartz & Ben-David, 2014), we get $\mathcal{R}_{n_i}(\mathcal{F}_{1,t}) \leq \mathcal{R}_{n_i}(\mathcal{F}'_{1,t})$. Then, by the definition of Rademacher complexity, we obtain:

$$\mathcal{R}_{n_i}(\mathcal{F}'_{1,t}) = \frac{1}{n_i} \mathbb{E}_{\boldsymbol{\xi}} \left[ \sup_{\|\theta_{i+1}(\tau)\|_2 \lesssim \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i}} \left| \sum_{j=1}^{n_i} \xi_j s_{t,\theta_{i+1}(\tau)}(x_i^j(t)) \right| \right]$$

$$\lesssim \frac{1}{n_i} \left( \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} \right) \mathbb{E}_{\boldsymbol{\xi}} \left[ \sup_{\|w\|_1 + \|u\|_1 \leq 1} \left| \sum_{j=1}^{n_i} \xi_j \sigma(w^\top x_i^j(t) + u^\top e(t)) \right| \right].$$

where the $\{\xi_j\}_{j=1}^{n_i}$ are independent random variables with the distribution $\mathbb{P}(\xi_j = 1) = \mathbb{P}(\xi_j = -1) = \frac{1}{2}$. Furthermore, leveraging the principle of symmetry, we derive:

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{n_i}\sup_{\|w\|_1+\|u\|_1\leq 1}\left|\sum_{j=1}^{n_i}\xi_j\sigma(w^\top x_i^j(t) + u^\top e(t))\right|\right] \leq 2\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{n_i}\sup_{\|w\|_1+\|u\|_1\leq 1}\sum_{j=1}^{n_i}\xi_j\sigma(w^\top x_i^j(t) + u^\top e(t))\right]$$
$$= 2\mathcal{R}_{n_i}(\mathcal{F}_{2,t}).$$

Then, we can get $\mathcal{R}_{n_i}(\mathcal{F}_{1,t}) \leq \mathcal{R}_{n_i}(\mathcal{F}_{1,t}') \lesssim (\|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i})\mathcal{R}_{n_i}(\mathcal{F}_{2,t})$. According to Lemma 26.9 (Contraction lemma) and Lemma 26.11 in Shalev-Shwartz & Ben-David (2014), we have

$$\mathcal{R}_{n_i}(\mathcal{F}_{2,t})) \leq (\|x_i(t)\|_\infty + \|e(t)\|_\infty)\sqrt{\frac{2\log(4d)}{n_i}} \lesssim (C_{T,\delta} + C_{T,e})\sqrt{\frac{\log d}{n_i}}.$$

Combining the aforementioned results with Equation 20, we subsequently obtain:

$$J_1 = \mathcal{R}_{n_i}(\mathcal{F}_t)) \lesssim (C_{T,\delta} + C_{T,e})^2(\|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} + 1)^2\sqrt{(\log d)/n_i}. \tag{21}$$

For $J_2$, by equation 18 and 19, we have:

$$|f(x_i(t))| = |(s_{t,\theta_i+1}(\tau)(x_i(t)) - \nabla_{x_i(t)}\log p_{i,t}(x_i(t)))||\sigma(w_j^\top x_i(t) + u_j^\top e(t))|$$
$$\lesssim (C_{T,\delta} + C_{T,e})^2(\|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i}) + C_{T,\delta}'(C_{T,\delta} + C_{T,e}).$$

Thus, we derive:

$$J_2 = \sup_{f\in\mathcal{F}_t, x_i(t)\in[-C_{T,\delta}, C_{T,\delta}]^d}|f(x_i(t))|\sqrt{\frac{\log(2/\delta)}{n_i}} \lesssim (C_{T,\delta} + C_{T,e})^2(\|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i} + 1)\sqrt{\frac{\log(1/\delta)}{n_i}}. \tag{22}$$

Similarly, in the case of $J_3$, we have:

$$J_3 = \mathbb{E}_{x_i(t)\sim\hat{p}_{i,t}}\left[\left(s_{t,\theta_{i+1}(\tau)}(x_i(t)) - s_{t,\hat{\theta}_{i+1}(\tau)}(x_i(t))\right)\sigma(w_j^\top x_i(t) + u_j^\top e(t))\right]$$
$$\lesssim (C_{T,\delta} + C_{T,e})^2(\|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\tau/m_i}). \tag{23}$$

By integrating Equation 13 with Equation 16, we derive the following result:

$$\frac{1}{m_i}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|_2^2 \lesssim \frac{1}{m_i^2}\sum_{i=1}^{m_i}\left\|\int_0^{\tau_{i+1}}\mathbb{E}_{t\sim u(0,T)}[\lambda(t)(J_1 + J_2 + J_3)\mathbf{1}_d]d\tau\right\|_2^2$$
$$\lesssim (J_1 + J_2 + J_3)^2\tau_{i+1}^2\frac{d}{m_i}.$$

Upon substituting Equations 21, 22, and 23 into the aforementioned equation, we subsequently obtain:

$$\frac{1}{m_i}\sum_{j=1}^{m_i}\|\hat{\alpha}_j(\tau_{i+1}) - \alpha_j(\tau_{i+1})\|_2^2 \lesssim \tau_{i+1}^2\frac{d}{m_i}(C_{T,\delta} + C_{T,e})^4\left((\|s_{0,\theta_0}\|_{\mathcal{H}}^4 + \frac{\tau_{i+1}^2}{m_i^2} + 1)\frac{\log(d/\delta)}{n_i} + \|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \frac{\tau_{i+1}}{m_i}\right).$$

By inserting the previously discussed equation into Equation 12, we then obtain:

$$\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) - \widetilde{\mathcal{L}}(\theta_{i+1}(\tau_{i+1}))$$

$$\lesssim \tau_{i+1}\sqrt{\frac{d}{m_i}}\left(\sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*)} + \frac{1}{\sqrt{\tau}}(\|s_{0,\theta_0}\|_{\mathcal{H}} + \|s_{0,\theta_*}\|_{\mathcal{H}})\right)\left((\|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \frac{\tau_{i+1}}{m_i} + 1)\sqrt{\frac{\log(d/\delta)}{n_i}} + \|s_{0,\theta_0}\|_{\mathcal{H}} + \sqrt{\frac{\tau_{i+1}}{m_i}}\right)$$

$$+ \tau_{i+1}^2\frac{d}{m_i}\left((\|s_{0,\theta_0}\|_{\mathcal{H}}^4 + \frac{\tau_{i+1}^2}{m_i{}^2} + 1)\frac{\log(d/\delta)}{n_i} + \|s_{0,\theta_0}\|_{\mathcal{H}}^2 + \frac{\tau_{i+1}}{m_i}\right)$$

$$\lesssim \tau_{i+1}\sqrt{\frac{d}{m_i}}\left((\frac{\|A_0\|_F^2}{m_i} + \frac{\tau_{i+1}}{m_i} + 1)\sqrt{\frac{\log(d/\delta)}{n_i}} + \frac{\|A_0\|_F}{\sqrt{m_i}} + \sqrt{\frac{\tau_{i+1}}{m_i}}\right)\left(\sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*)} + \frac{\|A_0\|_F + \|A_*\|_F}{\sqrt{m_i}}\right)$$

$$+ \tau_{i+1}\sqrt{\frac{d}{m_i}}\left((\frac{\|A_0\|_F^2}{m_i} + \frac{\tau_{i+1}}{m_i} + 1)\sqrt{\frac{\log(d/\delta)}{n_i}} + \frac{\|A_0\|_F}{\sqrt{m_i}} + \sqrt{\frac{\tau_{i+1}}{m_i}}\right))$$

$$\lesssim \tau_{i+1}\sqrt{\frac{d}{m_i}}\sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*)} + \frac{\tau_{i+1}^2 d\log(d/\delta)}{m_i n_i} + \frac{\tau_{i+1}^4 d\log(d/\delta)}{m_i{}^3 n_i} + \frac{\tau_{i+1}^3 d}{m_i{}^2}. \tag{24}$$

By incorporating the above equation along with Equations 7 and 10 into Equation 6, we consequently obtain:

$$\widetilde{\mathcal{L}}(\hat{\theta}_{i+1}(\tau_{i+1})) \lesssim \frac{\tau_{i+1}^2 d\log(d/\delta)}{mn_i} + \frac{\tau_{i+1}^4 d\log(d/\delta)}{m^3 n_i} + \frac{\tau_{i+1}^3 d}{m_i{}^2} + \frac{\log^2(1/\delta^2)}{m_i}d + \widetilde{\mathcal{L}}(\theta_{i+1}^*) + \bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*) + \frac{1}{\tau_{i+1}}. \tag{25}$$

Through the integration of the aforementioned equation with Equations 5 and 4, we consequently arrive at the following result:

$$TV(p_{\theta_{i+1}}, p_i)$$
$$\lesssim \sqrt{\frac{\tau_{i+1}^2 d\log(d/\delta)}{m_i n_i}} + \sqrt{\frac{\tau_{i+1}^4 d\log(d/\delta)}{m_i{}^3 n_i}} + \sqrt{\frac{\tau_{i+1}^3 d}{m_i{}^2}} + \sqrt{\frac{1}{\tau_{i+1}}} + \sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*)} + \sqrt{\bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*)} + \sqrt{KL(p_{i,T}\|\pi)}.$$

Let $m_i \asymp n_i$. Upon choosing $\tau_{i+1} \asymp n_i^{1/2}$, we omit terms $\sqrt{\widetilde{\mathcal{L}}(\theta_{i+1}^*)}$ and $\sqrt{\bar{\bar{\mathcal{L}}}(\bar{\theta}_{i+1}^*)}$. Consequently, we obtain the following result:

$$TV(p_{\theta_{i+1}}, p_i) \lesssim \frac{(d\log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T}\|\pi)}. \tag{26}$$

Upon substituting the above equation into Equation 3, we obtain:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim (d\log(d/\delta)^{\frac{1}{2}})/n_i^{\frac{1}{4}} + \sqrt{KL(p_{i,T}\|\pi)} + \beta_i^i TV(p_{\theta_i}, p_0) + \cdots + \beta_i^1 TV(p_{\theta_1}, p_0). \tag{27}$$

Analogous to the proof analysis process of Theorem 5.4, we can derive the following through recursive methods, with probability at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \sum_{k=0}^{i} A_{i-k}\left(n_{i-k}^{-\frac{1}{4}}\sqrt{d\log\frac{di}{\delta}} + \sqrt{KL(p_{i-k,T}\|\pi)}\right),$$

where $A_i = 1$, $A_{i-k} = \sum_{j=i-k+1}^{i} \beta_j^{i-k+1} A_j$ for $1 \le k \le i$. The proof is completed.

$$\square$$

# D. Proof of Corollary 4.5

In this section, we present the proof of Corollary 4.5, which considers the most extreme case of full synthetic data.

*Proof of Corollary 4.5.* Similar to the proof process of Theorem 4.1, we first decompose using the triangle inequality:

$$TV(p_{\theta_{i+1}}, p_0) \le TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0). \tag{28}$$

Utilizing inequality 26, we can establish the following with a probability of at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_i) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)},$$

Given that the definition of $p_i$ in the full synthetic data cycle is $p_i = p_{\theta_i}$, we can deduce the following with a probability of at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)} + TV(p_i, p_0).$$

$$= \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)} + TV(p_{\theta_i}, p_0)$$

Since the initial term $TV(p_{\theta_1}, p_0)$ can be expressed as follows:

$$TV(p_{\theta_1}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_0^{1/4}} + \sqrt{KL(p_{0,T} \| \pi)}.$$

Consequently, by solving recursively, we can obtain the following result, with probability at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \sum_{k=1}^{i} \left( \frac{(d \log(di/\delta))^{1/2}}{n_k^{1/4}} + \sqrt{KL(p_{k,T} \| \pi)} \right).$$

The proof is completed. $\square$

## E. Proof of Corollary 4.8

In this section, we present the proof of Corollary 4.8, which examines the scenario of balanced data cycle.

*Proof of Corollary 4.8.* Utilizing the triangle inequality of the TV distance, we can deduce the following:

$$TV(p_{\theta_{i+1}}, p_0) \le TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0). \tag{29}$$

By employing inequality 26, we can establish the following with a probability of at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)} + TV(p_i, p_0). \tag{30}$$

Recalling the definition of TV distance, we initially expound on the term $TV(p_i, p_0)$ as follows:

$$TV(p_i, p_0) = \frac{1}{2} \int |p_i(x) - p_0(x)| \, dx.$$

$$\le \frac{1}{2} \int |\frac{1}{i+1}(p_0(x) + p_{\theta_1}(x) + p_{\theta_2}(x) + \cdots + p_{\theta_i}(x)) - p_0(x)| \, dx$$

$$\le \frac{1}{2} \frac{1}{i+1} \int |p_{\theta_1}(x) - p_0(x)| + |p_{\theta_2}(x) - p_0(x)| + \cdots + |p_{\theta_i}(x) - p_0(x)| \, dx$$

$$\le \frac{1}{i+1} \sum_{j=1}^{i} TV(p_{\theta_j}, p_0). \tag{31}$$

Substituting the aforementioned inequality into Inequality 30, we obtain:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T}\|\pi)} + \frac{1}{i+1} \sum_{j=1}^{i} TV(p_{\theta_j}, p_0).$$

Specify the function $f(n_i)$ in the following manner: $f(n_i) = \frac{(d \log(d/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T}\|\pi)}$. As a result, we derive the subsequent outcome:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim f(n_i) + \frac{1}{i+1} TV(p_{\theta_i}, p_0) + \frac{1}{i+1} TV(p_{\theta_{i-1}}, p_0) + \cdots + \frac{1}{i+1} TV(p_{\theta_1}, p_0). \tag{32}$$

From the aforementioned expression, we can additionally infer:

$$TV(p_{\theta_i}, p_0) \lesssim f(n_{i-1}) + \frac{1}{i} TV(p_{\theta_{i-1}}, p_0) + \frac{1}{i} TV(p_{\theta_{i-2}}, p_0) + \cdots + \frac{1}{i} TV(p_{\theta_1}, p_0).$$

By multiplying both sides of the above equation by $\frac{1}{i+1}$, we obtain the following result:

$$\frac{1}{i+1} TV(p_{\theta_i}, p_0) \lesssim \frac{1}{i+1} f(n_{i-1}) + \frac{1}{i+1} \frac{1}{i} TV(p_{\theta_{i-1}}, p_0) + \cdots + \frac{1}{i+1} \frac{1}{i} TV(p_{\theta_1}, p_0).$$

Plugging the above inequality into inequality 32, we obtain:

$$TV(p_{\theta_{i+1}}, p_0)$$
$$\lesssim f(n_i) + \frac{1}{i+1} f(n_{i-1}) + (\frac{1}{i+1} + \frac{1}{i+1} \frac{1}{i}) TV(p_{\theta_{i-1}}, p_0) + \cdots + (\frac{1}{i+1} + \frac{1}{i+1} \frac{1}{i}) TV(p_{\theta_1}, p_0). \tag{33}$$

Note that $TV(p_{\theta_1}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{n_0^{1/4}} + \sqrt{KL(p_{0,T}\|\pi)}$, By solving recursively,

$$TV(p_{\theta_{i+1}}, p_0) \lesssim f(n_i) + \frac{1}{i+1} f(n_{i-1}) + (\frac{1}{i+1} + \frac{1}{i+1} \frac{1}{i}) f(n_{i-2}) + \cdots +$$
$$(\frac{1}{i+1} + \frac{1}{i+1} \frac{1}{i} + \cdots + \frac{1}{i+1} \frac{1}{i} \cdots \frac{1}{2}) f(n_0). \tag{34}$$

Thus, we have, with probability at least $1 - \delta$,

$$TV(p_{\theta_{i+1}}, p_0)$$
$$\lesssim \frac{(d \log(di/\delta))^{1/2}}{n_i^{1/4}} + \sqrt{KL(p_{i,T}\|\pi)} + \sum_{k=0}^{i-1} \sum_{j=k}^{i-1} \frac{\Gamma(j+2)}{\Gamma(i+2)} \left( \frac{(d \log(di/\delta))^{1/2}}{n_k^{1/4}} + \sqrt{KL(p_{k,T}\|\pi)} \right),$$

where the Gamma function $\Gamma(j) = (j-1)!$ and $j$ is a positive integer. The proof is complete. $\qquad \square$

## F. Proof of Corollary 4.10

In this section, we present the proof of Corollary 4.10, which analyzes the phase transition phenomena in error dynamics when increasing synthetic data while keeping real data fixed.

*Proof of Corollary 5.* By utilizing the triangle inequality of the TV distance, we can derive the following:

$$TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0). \tag{35}$$

By employing inequality 26, we can establish the subsequent statement with a probability of at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{(n+m)^{1/4}} + \sqrt{KL(p_{i,T}\|\pi)} + TV(p_i, p_0). \tag{36}$$

Recalling the definition of the total variation distance, let's first elaborate on the expression $TV(p_i, p_0)$ as follows:

$$TV(p_i, p_0) = \frac{1}{2} \int |p_i(x) - p_0(x)| \, dx.$$

$$\leq \frac{1}{2} \int |\frac{n}{n+m} p_0(x) + \frac{m}{n+m} p_{\theta_i}(x) - p_0(x)| \, dx$$

$$\leq \frac{1}{2} \frac{m}{n+m} \int |p_{\theta_i}(x) - p_0(x)| \, dx$$

$$\leq \frac{m}{n+m} TV(p_{\theta_i}, p_0). \tag{37}$$

Substituting the aforementioned inequality into Inequality 36, we obtain:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{(n+m)^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)} + \frac{m}{n+m} TV(p_{\theta_i}, p_0).$$

Note that $TV(p_{\theta_1}, p_0) \lesssim \frac{(d \log(d/\delta))^{1/2}}{(n+m)^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)}$. Similar to the proof of Corollary 4.8, we can solve recursively, with probability at least $1 - \delta$,

$$TV(p_{\theta_{i+1}}, p_0) \leq \left(1 + \frac{m}{n}\right) \left(1 - (\frac{m}{n+m})^{i+1}\right) \left(\frac{(d \log(di/\delta))^{1/2}}{(n+m)^{1/4}} + \sqrt{KL(p_{i,T} \| \pi)}\right).$$

The proof is completed. $\qquad\square$

## G. Proof of Lemma 5.1

In this section, we present the proof of Lemma 5.1. The proof utilizes the kernel density estimation framework along with McDiarmid's inequality. We first define the class $s$ kernel as follows:

**Definition G.1.** Let $s \geq 1$. A class $s$ kernel is a Borel measurable function $K$ which satisfies

- K is symmetric, i.e., $K(-x) = K(x)$, $x \in \mathbb{R}^d$.
- $\int K = 1$.
- $\int x^\alpha K(x) \, dx = 0$ for $1 \leq |\alpha| \leq s - 1$.
- $\int |x^\alpha| |K(x)| \, dx < \infty$ for $|\alpha| = s$.
- $\int (1 + \|x\|^{d+\epsilon}) K(x)^2 \, dx < \infty$ for some $\epsilon > 0$

*Proof of lemma 5.1.* Revisiting the definition of the kernel density estimation $\widehat{p}_i$, we derive the following:

$$\widehat{p}_{i+1}(x) = \frac{1}{n_i h_i^d} \sum_{j=1}^{n_i} K\left(\frac{x - x_j}{h_i}\right),$$

where $K$ belongs to the class $s$. Define $K_{h_i}(u) = (1/h_i^d) K(u/h_i)$. Our objective is to establish a bound for $TV(\widehat{p}_{i+1}, p_i) = \frac{1}{2} \int |\widehat{p}_{i+1}(x) - p_i(x)| \, dx$. Consider two sets: $\{x_1, \cdots, x_{n_i}\}$ and $\{x_1', \cdots, x_{n_i}'\}$, where $x_j' = x_j$ for all $j$ except when $j = t$. As a result, we derive:

$$\left| \frac{1}{2} \int |\widehat{p}_{i+1}(x; x_1, \ldots, x_n) - p_i(x)| \, dx - \frac{1}{2} \int |\widehat{p}_{i+1}(x; x_1', \ldots, x_n') - p_i(x)| \, dx \right|$$

$$\leq \frac{1}{2} \int |\widehat{p}_{i+1}(x; x_1, \ldots, x_n) - \widehat{p}_{i+1}(x; x_1', \ldots, x_n')| \, dx$$

$$\leq \frac{1}{2n_i} \int |K_{h_i}(x - x_t) - K_{h_i}(x - x_t')| \, dx$$

$$\leq \frac{1}{n_i h_i^d} \int |K|. \tag{38}$$

Since $\frac{1}{2}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx$ is $\frac{\int|K|}{n_ih_i^d}$-Lipschitz under the Hamming metric, the application of McDiarmid's Inequality (lemma B.1) yields the ensuing result with a probability of at least $1-\delta$:

$$\left|\frac{1}{2}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx-\mathbb{E}\,\frac{1}{2}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx\right|\leq\sqrt{\frac{(\int|K|)^2}{2n_ih_i^{2d}}\log\frac{2}{\delta}}.$$

Then, we obtain:

$$TV(\widehat{p}_i,p_i)=\frac{1}{2}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx\leq\frac{1}{2}\mathbb{E}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx+\sqrt{\frac{(\int|K|)^2}{2n_ih_i^{2d}}\log\frac{2}{\delta}}. \tag{39}$$

Next, our objective is to establish a bound for $\mathbb{E}\int|\widehat{p}_{i+1}(x)-p_i(x)|\,dx$. For simplicity, we omit the $(\cdot)$ and $dx$ in our notation. Through decomposition, we can derive the following results:

$$\mathbb{E}\int|\widehat{p}_{i+1}-p_i|\leq\int|p_i*K_{h_i}-p_i|+\mathbb{E}\int|\widehat{p}_{i+1}-p_i*K_{h_i}|, \tag{40}$$

where $p_i*K_{h_i}(x)=\int p_i(y)K_{h_i}(x-y)\,dy$. The terms on the right-hand side will be called the bias and variation terms of the error. Assume first that $p_i\in C_0^\infty(\mathbb{R}^d)$, where $C_0^\infty(\mathbb{R}^d)$ denotes the space of infinitely differentiable functions with compact support. By Taylor's theorem,

$$p_i(x+y)-p_i(x)=\sum_{j=1}^{s-1}\sum_{|\alpha|=j}\frac{1}{\alpha!}y^\alpha\partial^\alpha p_i(x)+\sum_{|\alpha|=s}\frac{s!}{\alpha!}\int_0^1\frac{(1-t)^{s-1}}{(s-1)!}y^\alpha\partial^\alpha p_i(x+ty)dt. \tag{41}$$

Since $K$ is a symmetric kernel, we have $\int K(z)\,dz=1$ and $\int zK(z)\,dz=0$. Thus, for $x\in\mathbb{R}^d$, we derive:

$$p_i*K(x)-p_i(x)=\int(p_i(y)-p_i(x))K(x-y)\,dy$$
$$=\int(p_i(x-y)-p_i(x))K(y)\,dy$$
$$=\int(p_i(x+y)-p_i(x))K(y)\,dy. \tag{42}$$

Substituting Equation 41 into Equation 42 and taking into account that $\int x^\alpha K(x)\,dx=0$ for $1\leq|\alpha|\leq s-1$, we derive:

$$p_i*K(x)-p_i(x)=\sum_{|\alpha|=s}\frac{s!}{\alpha!}\int\int_0^1\frac{(1-t)^{s-1}}{(s-1)!}y^\alpha\partial^\alpha p_i(x+ty)K(y)\,dt\,dy. \tag{43}$$

The integrals with respect to $y$ exist due to the application of Fubini's theorem and the introduction of new variables $\eta=-ty$ and $\tau=t^{-1}$. This holds true for $|\alpha|=s$,

$$\int\int_0^1\frac{(1-t)^{s-1}}{(s-1)!}|y^\alpha|\,|\partial^\alpha p_i(x+ty)|\,|K(y)|\,dt\,dy$$
$$=\int_0^1\int\frac{(1-t)^{s-1}}{(s-1)!}|y^\alpha|\,|\partial^\alpha p_i(x+ty)|\,|K(y)|\,dy\,dt$$
$$=\int_0^1\int\frac{(1-t)^{s-1}}{(s-1)!}t^{-s-d}|\eta^\alpha|\,|\partial^\alpha p_i(x-\eta)|\,|K(\frac{\eta}{t})|\,d\eta\,dt$$
$$=\int|\partial^\alpha p_i(x-\eta)|\int_1^\infty\frac{(\tau-1)^{s-1}}{(s-1)!}\tau^{d-1}|\eta^\alpha|\,|K(\tau\eta)|\,d\tau\,d\eta.$$

Since the integral $\int|x^\alpha K(x)|dx$ is finite for $|\alpha|=s$, the last integral is also finite. By repeating the same steps, we can conclude:

$$\int\int_0^1\frac{(1-t)^{s-1}}{(s-1)!}y^\alpha\partial^\alpha p_i(x+ty)K(y)\,dt\,dy=\int\partial^\alpha p_i(x-\eta)\,(-1)^s\int_1^\infty\frac{(\tau-1)^{s-1}}{(s-1)!}\tau^{d-1}\eta^\alpha\,K(\tau\eta)\,d\tau\,d\eta.$$

By defining the associated kernel $L^\alpha(x)$ as $(-1)^{|\alpha|} \int_1^\infty \frac{(t-1)^{|\alpha|-1}}{(|\alpha|-1)!} t^{d-1} x^\alpha K(tx) \, dt$, we can then obtain:

$$\int \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} y^\alpha \partial^\alpha p_i(x+ty) K(y) \, dt \, dy = \partial^\alpha p_i * L^\alpha(x). \tag{44}$$

Plugging Equation 44 into the Equation 43, then, we obtain:

$$p_i * K(x) - p_i(x) = \sum_{|\alpha|=s} \frac{s!}{\alpha!} \partial^\alpha p_i * L^\alpha(x). \tag{45}$$

When $p_i \in W^{s,1}(\mathbb{R}^d)$, there are functions $p_{i,n} \in C_0^\infty(\mathbb{R}^d)$, $n = 0, 1, \ldots$, such that $\partial^\alpha p_{i,n} \to \partial^\alpha p_i$ in $L^1(\mathbb{R}^d)$ for $|\alpha| \le s$ (Adams & Fournier, 2003). Since Equation 45 holds for each $p_{i,n}$, it also holds for $p_i$ a.e. on $\mathbb{R}^d$. Furthermore, for the kernel $K_{h_i}$, applying Young's inequality, we obtain:

$$\int |p_i * K_{h_i} - p_i| \le h_i^s \sum_{|\alpha|=s} \frac{s!}{\alpha!} \int |\partial^\alpha p_i| \int |L^\alpha| = h_i^s \varphi(s, K, p_i). \tag{46}$$

Given that $p_i \in W^{s,1}(\mathbb{R}^d)$ and $\int |x^\alpha| |K(x)| \, dx < \infty$ for $|\alpha| = s$, it follows that $\varphi(s, K, p_i) = \sum_{|\alpha|=s} \frac{s!}{\alpha!} \int |\partial^\alpha p_i| \int |L^\alpha|$ is finite. Thus, the bias tends to zero at least at the rate $h_i^s$. Next, we address the variation term $\mathbb{E} \int |\widehat{p}_{i+1} - p_i * K_{h_i}|$. By the Schwarz inequality, we obtain:

$$\mathbb{E} \int |\widehat{p}_{i+1} - p_i * K_{h_i}| \le (n_i h_i^d)^{-\frac{1}{2}} \int \sqrt{p_i * (K^2)_{h_i}}.$$

Let $Q = K^2 / \int K^2$, and by utilizing the inequality $\sqrt{p_i * Q_{h_i}} \le \sqrt{p_i} + \sqrt{|p_i - p_i * Q_{h_i}|}$, we obtain:

$$\mathbb{E} \int |\widehat{p}_{i+1} - p_i * K_{h_i}| \le (n_i h_i^d)^{-\frac{1}{2}} \left( 1 + \frac{\int \sqrt{|p_i - p_i * Q_{h_i}|}}{\int \sqrt{p_i}} \right) \sqrt{\int K^2} \int \sqrt{p_i}.$$

Let $\gamma(h_i) = \int \sqrt{|p_i - p_i * Q_{h_i}|} / \int \sqrt{p_i}$. Then, our objective is to show $\gamma(h_i) \to 0$ as $h_i \to 0^+$. First, we apply Carlson's inequality to get

$$\int \sqrt{|p_i - p_i * Q_{h_i}|} \le C \left( \int |p_i - p_i * Q_{h_i}| \right)^{\epsilon/2(\epsilon+d)} \times \left( \int \|x\|^{d+\epsilon} |p_i(x) - p_i * Q_{h_i}(x)| \, dx \right)^{d/2(\epsilon+d)},$$

where $C$ is a constant. Since $\int |p_i - p_i * Q_{h_i}|$ tends to zero as $h_i \to 0^+$ [(Stein, 1970), Chapt. III], our task is to demonstrate that the second integral remains bounded. Thus, we have

$$\int \|x\|^{d+\epsilon} |p_i(x) - p_i * Q_{h_i}(x)| \, dx \le \int \|x\|^{d+\epsilon} p_i(x) \, dx + \int \|x\|^{d+\epsilon} p_i * Q_{h_i}(x) \, dx.$$

According to the assumption, we have $\int \|x\|^{d+\epsilon} p_i(x) \, dx$ is finite. For the second integral, we utilize $\xi = x - y$, $\|\xi + y\|^{d+\epsilon} \le 2^{d+\epsilon-1} (\|\xi\|^{d+\epsilon} + \|y\|^{d+\epsilon})$, and the fact that $\int p_i = \int Q_{h_i} = 1$, which implies that:

$$\int \|x\|^{d+\epsilon} p_i * Q_{h_i}(x) \, dx = \int \left( \int \|x\|^{d+\epsilon} p_i(x-y) \, dx \right) Q_{h_i}(y) \, dy$$

$$\le 2^{d+\epsilon-1} \int_{\mathbb{R}^d} \|\xi\|^{d+\epsilon} p_i(\xi) \, d\xi + 2^{d+\epsilon-1} \int \|y\|^{d+\epsilon} Q_{h_i}(y) \, dy.$$

Since $\int (1 + \|x\|^{d+\epsilon}) K(x)^2 \, dx < \infty$ , the above integral is finite. Thus, we have

$$\mathbb{E} \int |\widehat{p}_{i+1} - p_i * K_{h_i}| \le (n_i h_i^d)^{-\frac{1}{2}} (1 + \gamma(h_i)) \sqrt{\int K^2} \int \sqrt{p_i}, \tag{47}$$

where $\gamma(h_i) \to 0$ as $h_i \to 0^+$. Combining the inequality 46 for the bias term and inequality 47 for the variation term into the inequality 40, we obtain:

$$\mathbb{E} \int |\widehat{p}_{i+1} - p_i| \leq h_i^s \varphi(s, K, p_i) + (n_i h_i^d)^{-\frac{1}{2}}(1 + \gamma(h_i))\sqrt{\int K^2} \int \sqrt{p_i}. \tag{48}$$

Plugging the above inequality into the inequality 39, then, we obtain:

$$TV(\widehat{p}_{i+1}, p_i) \leq (n_i h_i^{2d})^{-\frac{1}{2}}\sqrt{\frac{1}{2}(\int |K|)^2 \log \frac{2}{\delta}} + \frac{1}{2}h_i^s \varphi(s, K, p_i) + \frac{1}{2}(n_i h_i^d)^{-\frac{1}{2}}(1 + \gamma(h_i))\sqrt{\int K^2}\int\sqrt{p_i}.$$

By the choice of $h_i = n_i^{-\frac{1}{2s+2d}}$, then, we obtain:

$$TV(\widehat{p}_{i+1}, p_i) \leq n_i^{-\frac{s}{2s+2d}}\sqrt{\frac{1}{2}(\int |K|)^2 \log \frac{2}{\delta}} + \frac{1}{2}n_i^{-\frac{s}{2s+2d}}\varphi(s, K, p_i) + \frac{1}{2}n_i^{-\frac{2s+d}{4s+4d}}(1 + \gamma_{n_i})\sqrt{\int K^2}\int\sqrt{p_i}. \tag{49}$$

Where $\gamma_{n_i} \to 0$ as $n_i \to \infty$ and $\varphi(s, K, p_i)$ is a finite function. The proof is completed. $\qquad\square$

## H. Proof of Theorem 5.4

In this section, we present the proof of Theorem 5.4, which targets the scenario of the general data cycle.

*Proof of Theorem 5.4.* By leveraging the triangle inequality of TV distance, we can deduce the following:

$$TV(\widehat{p}_{i+1}, p_0) \leq TV(\widehat{p}_{i+1}, p_i) + TV(p_i, p_0). \tag{50}$$

Revisiting the definition of TV distance, we initially elucidate the term $TV(p_i, p_0)$ as follows:

$$TV(p_i, p_0) = \frac{1}{2}\int |p_i(x) - p_0(x)|dx$$
$$\leq \frac{1}{2}\int |\beta_i^1 \widehat{p}_1(x) + \beta_i^2 \widehat{p}_2(x) + \cdots + \beta_i^i \widehat{p}_i(x) + \alpha_i p_0(x) - p_0(x)| \; dx.$$

Given that $\beta_i^1 + \beta_i^2 + \cdots + \beta_i^i + \alpha_i = 1$, we can deduce the following:

$$TV(p_i, p_0) \leq \frac{1}{2}\int |\beta_i^1 \widehat{p}_1(x) + \beta_i^2 \widehat{p}_2(x) + \cdots + \beta_i^i \widehat{p}_i(x) - (\beta_i^1 + \beta_i^2 + \cdots + \beta_i^i)p_0(x)| \; dx$$
$$\leq \frac{1}{2}\int \beta_i^1 |\widehat{p}_1(x) - p_0(x)| + \beta_i^2 |\widehat{p}_2(x) - p_0(x)| + \cdots + \beta_i^i |\widehat{p}_i(x) - p_0(x)| \; dx$$
$$= \beta_i^1 TV(\widehat{p}_1, p_0) + \beta_i^2 TV(\widehat{p}_2, p_0) + \cdots + \beta_i^i TV(\widehat{p}_i, p_0). \tag{51}$$

By employing Lemma 5.1, and considering that $\gamma_{n_i} \to 0$ as $n_i \to \infty$, along with $\varphi(s, K, p_i)$ being a finite function, we can derive the following with a probability of at least $1 - \delta$:

$$TV(\widehat{p}_{i+1}, p_i) \leq n_i^{-\frac{s}{2s+2d}}\sqrt{\frac{1}{2}(\int |K|)^2 \log \frac{2}{\delta}} + \frac{1}{2}n_i^{-\frac{s}{2s+2d}}\varphi(s, K, p_i) + \frac{1}{2}n_i^{-\frac{2s+d}{4s+4d}}(1 + \gamma_{n_i})\sqrt{\int K^2}\int\sqrt{p_i}$$
$$\lesssim n_i^{-\frac{s}{2s+2d}}\sqrt{\log(2/\delta)} + n_i^{-\frac{2s+d}{4s+4d}}, \tag{52}$$

where $\lesssim$ conceals universal positive constants that depend solely on $K$, $p_i$ and $s$. By incorporating Inequality 51 and 52 into the Inequality 50, we can consequently derive the following with a probability of at least $1 - \delta$:

$$TV(\widehat{p}_{i+1}, p_0) \lesssim n_i^{-\frac{s}{2s+2d}}\sqrt{\log(2/\delta)} + n_i^{-\frac{2s+d}{4s+4d}} + \beta_i^i TV(\widehat{p}_i, p_0) + \beta_i^{i-1}TV(\widehat{p}_{i-1}, p_0) + \cdots + \beta_i^1 TV(\widehat{p}_1, p_0).$$

Define the function $f(n_i)$ as follows: $f(n_i) = n_i^{-\frac{s}{2s+2d}} \sqrt{\log(2/\delta)} + n_i^{-\frac{2s+d}{4s+4d}}$. Consequently, we obtain the ensuing result:

$$TV(\widehat{p}_{i+1}, p_0) \lesssim f(n_i) + \beta_i^i TV(\widehat{p}_i, p_0) + \beta_i^{i-1} TV(\widehat{p}_{i-1}, p_0) + \cdots + \beta_i^1 TV(\widehat{p}_1, p_0). \tag{53}$$

From the above expression, we can further deduce:

$$TV(\widehat{p}_i, p_0) \lesssim f(n_{i-1}) + \beta_{i-1}^{i-1} TV(\widehat{p}_{i-1}, p_0) + \beta_{i-1}^{i-2} TV(\widehat{p}_{i-2}, p_0) + \cdots + \beta_{i-1}^1 TV(\widehat{p}_1, p_0).$$

By multiplying both sides of the above equation by $\beta_i^i$, we obtain the following result:

$$\beta_i^i TV(\widehat{p}_i, p_0) \lesssim \beta_i^i f(n_{i-1}) + \beta_i^i \beta_{i-1}^{i-1} TV(\widehat{p}_{i-1}, p_0) + \beta_i^i \beta_{i-1}^{i-2} TV(\widehat{p}_{i-2}, p_0) + \cdots + \beta_i^i \beta_{i-1}^1 TV(\widehat{p}_1, p_0).$$

Plugging the above inequality into inequality 53, we obtain:

$$TV(\widehat{p}_{i+1}, p_0) \lesssim f(n_i) + \beta_i^i f(n_{i-1}) + (\beta_i^{i-1} + \beta_i^i \beta_{i-1}^{i-1}) TV(\widehat{p}_{i-1}, p_0) + \cdots + (\beta_i^1 + \beta_i^i \beta_{i-1}^1) TV(\widehat{p}_1, p_0).$$

Define the coefficient preceding the term $f(n_i)$ as $A_i$. Consequently, we have $A_i = 1$ and $A_{i-1} = \beta_i^i$. By adopting a procedure analogous to the one described above, we can derive the following result:

$$A_{i-2} = \beta_i^{i-1} + \beta_i^i \beta_{i-1}^{i-1} = \beta_i^{i-1} A_i + \beta_{i-1}^{i-1} A_{i-1},$$
$$A_{i-3} = \beta_i^{i-2} + \beta_i^i \beta_{i-1}^{i-2} + (\beta_i^{i-1} + \beta_i^i \beta_{i-1}^{i-1})\beta_{i-2}^{i-2} = \beta_i^{i-2} A_i + \beta_{i-1}^{i-2} A_{i-1} + \beta_{i-2}^{i-2} A_{i-2}.$$

Based on the aforementioned discussion, we can draw the following conclusion:

$$A_{i-t} = \beta_i^{i-t+1} A_i + \beta_{i-1}^{i-t+1} A_{i-1} + \cdots + \beta_{i-t+1}^{i-t+1} A_{i-t+1} = \sum_{j=i-t+1}^{i} \beta_j^{i-t+1} A_j, \quad 1 \le t \le i. \tag{54}$$

Furthermore, for the initial term $TV(\widehat{p}_1, p_0)$, we have:

$$TV(\widehat{p}_1, p_0) \lesssim n_0^{-\frac{s}{2s+2d}} \sqrt{\log(2/\delta)} + n_0^{-\frac{2s+d}{4s+4d}}$$
$$= f(n_0). \tag{55}$$

By synthesizing Inequalities 53, 54, and 55, we arrive at the following result, with probability at least $1 - \delta$:

$$TV(\widehat{p}_{i+1}, p_0)$$
$$\lesssim \sum_{k=0}^{i} A_{i-k} \left( n_{i-k}^{-\frac{s}{2s+2d}} \sqrt{\log(i/\delta)} + n_{i-k}^{-\frac{2s+d}{4s+4d}} \right),$$

$$\tag{56}$$

where $A_i = 1, A_{i-k} = \sum_{j=i-k+1}^{i} \beta_j^{i-k+1} A_j$ for $1 \le k \le i$. $\qquad\qquad\square$

# I. Extensions

In this section, we present some extensions of our analyses.

## I.1. Extension to Normalizing Flows

For Normalizing flows, we follow the setting in Yang (2022). The theorem is extended as follows:

**Theorem I.1.** *Assuming the second moment of $p_i$ is finite for all $i$, we set the base distribution to be the unit Gaussian $\mathcal{N}$. Let the velocity field be modeled by $V$, which belongs to $\mathcal{H}(\mathbb{R}^{d+1}, \mathbb{R}^d)$, a space of functions representable as $f_{\mathbf{a}}(\mathbf{x}) = \mathbb{E}_{\rho(\mathbf{w},b)}[\mathbf{a}(\mathbf{w}, b)\,\sigma(\mathbf{w} \cdot \mathbf{x} + b)]$, where $\rho \in P(\mathbb{R}^{d+1})$ is a fixed parameter distribution and $\mathbf{a} \in L^2(\rho, \mathbb{R}^d)$ is a parameter function. Let $G_{V_i}$ denote the flow map defined as $G_{V_i}(x_i(0)) = x_i(1)$, $\frac{d}{d\tau}x_i(\tau) = V_i(x_i(\tau), \tau)$, and let the reverse-time flow map for $\tau \in [0, 1]$ be defined as $F_{V_i}(x_i(1), \tau) = x_i(\tau)$, $\frac{d}{d\tau}x_i(\tau) = V_i(x_i(\tau), \tau)$. Assuming all the flow-induced*

*norm of optimal $V_i^*$ satisfies $\|V_i^*\|_{\mathcal{F}} = \exp\|V_i^*\|_{\mathcal{H}} \leq R$. Let $n_i$ be the number of training samples obtained from the distribution $p_i$. Then with probability at least $1 - \delta$, we establish that:*

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \sum_{k=0}^{i} A_{i-k}\left(n_{i-k}^{-\frac{1}{4}} R\sqrt{1+R^2}\log^{\frac{1}{4}}\frac{i}{\delta}\right),$$

*where $A_i = 1$, $A_{i-k} = \sum_{j=i-k+1}^{i}\beta_j^{i-k+1}A_j$ for $1 \leq k \leq i$.*

*Proof of Theorem I.1.* Using the triangle inequality for the TV distance, we can infer the following:

$$TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + TV(p_i, p_0).$$

For the second term, $TV(p_i, p_0)$, considering the definition of TV distance as $\frac{1}{2}\int|p_i(x) - p_0(x)|\,dx$, and given $p_i(x) = \beta_i^1 p_{\theta_1}(x) + \cdots + \beta_i^i p_{\theta_i}(x) + \alpha_i p_0(x)$, we can infer:

$$TV(p_{\theta_{i+1}}, p_0) \leq TV(p_{\theta_{i+1}}, p_i) + \beta_i^i TV(p_{\theta_i}, p_0) + \beta_i^{i-1}TV(p_{\theta_{i-1}}, p_0) + \cdots + \beta_i^1 TV(p_{\theta_1}, p_0). \tag{57}$$

Next, let's focus on the first term, $TV(p_{\theta_{i+1}}, p_i)$. Utilizing Pinsker's inequality, we derive:

$$TV(p_{\theta_{i+1}}, p_i) \leq \sqrt{\frac{1}{2}KL(p_i\|p_{\theta_{i+1}})}. \tag{58}$$

Following the framework established by Yang (2022), we establish bounds for the $KL$ term. Let $L$ be the population loss defined as:

$$L(V_i) = \iint_0^1 \text{Tr}\big[\nabla V_i(x_i(\tau), \tau)\big]d\tau + \frac{1}{2}\|x_i(0)\|^2 dp_i(x_i(1))$$
$$\mathbf{x}_i(\tau) := G_\tau(G^{-1}(x_i(1))) \tag{59}$$

In addition, let $L^{(n_i)}$ and $V_i^{(n_i)}$ be the corresponding empirical loss and the output. It follows that:

$$L\left(V_i^{(n_i)}\right) \leq L^{(n_i)}\left(V_i^{(n_i)}\right) + \sup_{\|V_i\|_{\mathcal{F}}\leq R} L(V_i) - L^{(n_i)}(V_i)$$
$$\leq L^{(n_i)}(V_i^*) + \sup_{\|V_i\|_{\mathcal{F}}\leq R} L(V_i) - L^{(n_i)}(V_i)$$
$$\leq L(V_i^*) + 2\sup_{\|V_i\|_{\mathcal{F}}\leq R} L(V_i) - L^{(n_i)}(V_i)$$

Then, we obtain the following:

$$\sup_{\|V_i\|_{\mathcal{F}}\leq R} L(V_i) - L^{(n_i)}(V_i)$$
$$\leq \sup_{\|V_i\|_{\mathcal{F}}\leq R} \iint_0^1 \text{Tr}\left[\nabla V_i\left(F_{V_i}\left(x_i(\tau), \tau\right), \tau\right)\right]d\tau d\left(p_i - \overline{p}_i\right)(x)$$
$$+ \sup_{\|V_i\|_{\mathcal{F}}\leq R} \iint_0^1 \frac{1}{2}\|F_{V_i}(x, 1)\|^2 d\left(p_i - \overline{p}_i\right)(x) \tag{60}$$

Let $A$ and $B$ represent the two terms as random variables. Employing the techniques elucidated in [Ma et al. (2019), Theorem 2.11] and [Han et al. (2021), Theorem 3.3] to bound the Rademacher complexity of flow-induced functions, we can obtain:

$$\mathbb{E}[A] \lesssim \frac{R}{\sqrt{n_i}}\mathbb{E}\left[\max_{1\leq j\leq n_i}\left\|x_i^j\right\|\right] \lesssim \frac{R^2}{\sqrt{n_i}}$$
$$\mathbb{E}[B] \lesssim \frac{R^2}{\sqrt{n_i}}\mathbb{E}\left[\max_{1\leq j\leq n_i}\left\|x_i^j\right\|^2\right] \lesssim \frac{R^4}{\sqrt{n_i}}$$

Furthermore, concerning the variances $A - \mathbb{E}[A]$ and $B - \mathbb{E}[B]$, we can employ the extension of McDiarmid's inequality to sub-Gaussian random variables (Kontorovich, 2014), to demonstrate that, with a probability of $1 - \delta$,

$$A - \mathbb{E}[A] \lesssim \frac{R^2 \sqrt{\log 1/\delta}}{\sqrt{n_i}}, \quad B - \mathbb{E}[B] \lesssim \frac{R^4 \sqrt{\log 1/\delta}}{\sqrt{n_i}}$$

Combining these inequalities, we induce that:

$$KL(p_i \| p_{\theta_{i+1}}) \lesssim \frac{R^2(1 + R^2)\sqrt{\log 1/\delta}}{\sqrt{n_i}}$$

Utilizing Pinsker's inequality, we derive:

$$TV(p_{\theta_{i+1}}, p_i) \lesssim n_i^{-\frac{1}{4}} R \sqrt{1 + R^2} \log^{\frac{1}{4}} \frac{1}{\delta}. \tag{61}$$

Upon substituting the above equation into Equation 57, we obtain:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim n_i^{-\frac{1}{4}} R \sqrt{1 + R^2} \log^{\frac{1}{4}} \frac{1}{\delta} + \beta_i^i TV(p_{\theta_i}, p_0) + \cdots + \beta_i^1 TV(p_{\theta_1}, p_0). \tag{62}$$

Analogous to the proof analysis process of Theorem 5.4, we can derive the following through recursive methods, with probability at least $1 - \delta$:

$$TV(p_{\theta_{i+1}}, p_0) \lesssim \sum_{k=0}^{i} A_{i-k} \left( n_{i-k}^{-\frac{1}{4}} R \sqrt{1 + R^2} \log^{\frac{1}{4}} \frac{i}{\delta} \right),$$

where $A_i = 1$, $A_{i-k} = \sum_{j=i-k+1}^{i} \beta_j^{i-k+1} A_j$ for $1 \leq k \leq i$.

$\square$

*Remark* I.2. **Extension to Transformer Models.** Exploring the theoretical extension of this work to transformer models trained within a self-consumption loop presents significant challenges. Firstly, characterizing the generalization error of transformers at each generation is essential. It's worth noting that this task is particularly challenging due to the complexity of the training data, which consists of a mixture of data from different distributions. Furthermore, delving into how this error compounds with each generation poses an even more complex scenario. Investigating this intricate situation remains a subject for future work, requiring further exploration.

## J. Experiments

In this section, we present some experimental results. Specifically, we trained a diffusion model on the MNIST dataset. Consistent with previous works (Alemohammad et al., 2023; Bertrand et al., 2023), we employed the FID score as a metric to evaluate model performance. We trained multiple generations under two scenarios: 1) a 1:1 ratio of real and synthetic data, and 2) fully synthetic data. This allowed us to investigate the impact of the number of training generations, the total number of training samples, and the presence of real data on model performance, as detailed in the table below:

| Dataset | gen01 | gen02 | gen03 | gen04 |
|---------|-------|-------|-------|-------|
| syn10k  | 120.99 | 156.77 | 165.04 | 180.23 |
| syn15k  | 55.46 | 137.45 | 151.95 | 157.62 |
| syn20k  | 24.44 | 46.78 | 61.14 | 78.77 |
| mix10k  | 58.94 | 117.73 | 137.62 | 136.39 |
| mix15k  | 55.69 | 66.93 | 69.63 | 75.91 |
| mix20k  | 20.05 | 31.92 | 37.36 | 36.13 |

We observed that our experimental findings closely aligned with the theoretical results derived in the paper. When the number of training samples and the mix in the training set remained constant, the model performance deteriorated as the number of training generations increased. However, increasing the number of training samples and incorporating real data both led to improved model performance.

In addition, we fixed the amount of real data at 9K samples and varied the number of synthetic samples to 13K, 15K, 17K, and 20K. The experimental results are as follows:

| Dataset | syn 13k | syn 15k | syn 17k | syn 20k |
|---------|---------|---------|---------|---------|
| **real 9k** | 74.95 | 87.15 | 33.50 | 28.39 |

These findings corroborate the theoretical predictions in Remark 4.11, where we indicate that the TV distance first exhibits an ascent and then declines beyond a threshold point as synthetic data expands while real data remains fixed.