d-LINEAR GENERATION ERROR BOUND FOR DISTRIBUTED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent rise of distributed diffusion models has been driven by the explosive growth of data and the increasing demand for data generation. However, distributed diffusion models face unique challenges in resource-constrained environments. Existing approaches lack theoretical support, particularly with respect to generation error in such settings. In this paper, we are the first to derive the generation error bound for distributed diffusion models with arbitrary pruning, not assuming perfect score approximation. By analyzing the convergence of the score estimation model trained with arbitrary pruning in a distributed manner, we highlight the impact of complex factors such as model evolution dynamics and arbitrary pruning on the generation performance. This theoretical generation error bound is linear in the data dimension d, aligning with state-of-the-art results in the single-worker paradigm.

1 INTRODUCTION

Recently, distributed diffusion models have gained significant attention due to the explosive growth of data and the growing interest in data generation (Vora et al., 2024; de Goede et al., 2024). Specifically, in federated settings, diffusion models are trained collaboratively across multiple workers without the need to share personal sensitive data, such as images and audio, directly. This distributed approach enables large-scale data generation while avoiding the privacy risks and practical costs of centralizing data (Tun et al., 2023).

In real-world scenarios, workers typically possess limited computational and communication resources, which significantly hinder the performance (Zhang et al., 2021). When training a parametrized neural network in the reverse process of diffusion models, it would be unaffordable for resource-constrained workers to operate model updates. Some efforts have been made to address this challenge. For example, Li et al. (2024) propose DistriFusion, a method that divides the model input into multiple patches, each assigned to a GPU. This approach tackles the high computational costs involved in generating high-resolution images with diffusion models. Additionally, Lai et al. (2024) introduce an on-demand quantized energy-efficient distributed approach for training diffusion-based models in mobile edge networks. Despite the efforts made by these studies, they primarily focus on improving empirical performance, leaving the theoretical behavior as an open problem.

The theoretical lack in the generation performance of distributed diffusion models is driven by the increased complexity in training dynamics resulting from limited resources. The generation error bound of the diffusion model heavily depends on the loss value of the neural network trained in the reverse process (Benton et al., 2024; Chen et al., 2023). And the coupling between the loss and the gradient often reflects the convergence rate of the neural network (Zhou et al., 2024). As a result, accurately bounding the generation error requires describing the convergence rate of the parameterized neural network trained in the reverse process. However, resource limitations in distributed systems may lead to insufficient training or incomplete transmission of local models (Zhou et al., 2024; Qiao et al., 2023), which exacerbates global error accumulation and complicates the analysis of convergence rates during the reverse process.

In this paper, we provide formal theoretical support for distributed diffusion models in resource-constrained scenarios. To address the performance degradation caused by resource constraints, we

consider introducing pruning operations when training score estimation model in a distributed manner during the reverse process, and using coordinate-aware model aggregation (Zhou et al., 2024) to reduce global error accumulation. To obtain the convergence rate during the reverse process, we utilize the smoothness assumption to measure the inconsistency between the local and global gradients. We also implement a refined treatment of pruning errors and utilize the relationship between iterative model updates to explore their cumulative entanglement. By analyzing the convergence of the score estimation model and exploring the error between local and global training losses, we reveal the impact of complex factors such as the number of communication rounds and the number of workers on the local score estimation error. Using this actual error, rather than the assumed constant error in the single-worker paradigm (Benton et al., 2024; Chen et al., 2023; 2022), we derive the generation error bound for distributed learning diffusion models in resource-constrained scenarios. Specifically, our main contributions can be summarized as follows:

- **To the best of our knowledge, we are the first to incorporate the distributed learning dynamic of the score estimation model during the reverse process into the analysis of the final generation error.** We theoretically assess the discrepancy between the generated sample distribution and the actual distribution for each worker using KL divergence. This generation error bound aligns with the best-known results in the single-worker paradigm (Benton et al., 2024), exhibiting a linear dependence on the data dimension d. Notably, our framework can be seamlessly integrated with the theoretical error bounds of any diffusion model based on the single-worker paradigm under the perfect fractional approximation assumption. This integration ensures that the theoretical error bounds of similar distributed training architectures progress in tandem with advancements in the theoretical error bounds of the single-worker paradigm.
- We also derive convergence bounds for distributed learning of the score estimation model under arbitrary pruning, without relying on the bounded gradient assumption. It shows that the average gradient norm can converge to 0 at a rate of $\mathcal{O}(\frac{1}{\sqrt{\Gamma^* SQ}})$, showing the critical roles of the number of local training steps S and the minimum parallel training degree Γ^* in enhancing convergence efficiency.

2 Related Work

In recent research, diffusion models (Song et al., 2020) have garnered widespread attention due to their remarkable achievements across multiple fields, including computer vision (Harvey et al., 2022), natural language processing (Li et al., 2022), temporal data modeling (Tashiro et al., 2021), and multi-modal learning (Ramesh et al., 2022; Ho et al., 2022). Particularly, some studies highlight that diffusion models not only generate high-quality data but also surpass traditional Generative Adversarial Networks in terms of stability and generation efficiency (Dhariwal & Nichol, 2021).

Recent works have extensively explored the theoretical performance of diffusion models. This provides a robust theoretical foundation for refining the model architecture and optimizing the training process. Initial studies on the convergence of diffusion models often requires restrictive assumptions about the data distribution, such as adherence to a log-Sobolev inequality (Yang & Wibisono, 2022), or results in bounds that are either non-quantitative (Pidstrigach, 2022) or exponential (Block et al., 2020) with respect to the problem parameters. Subsequent research has made significant improvements. For instance, some studies have achieved polynomial convergence rates for diffusion models without restrictive assumptions on the data distribution. Specifically, Chen et al. (2022) obtain polynomial error bounds in total variation (TV) distance, assuming that the score function is Lipschitz. They employ the Girsanov change of measure framework to analyze the discrepancy between the true and approximate reverse processes. Further advances are made by the work (Chen et al., 2023), which develop the Girsanov methodology further and introduce two important theorems: Theorem 2.1 shows that the KL divergence is linear in the data dimension but requires that $\nabla \log q_t$ be Lipschitz; Theorem 2.2 demonstrates that, under an early-stopping setting and with any data distribution having a finite second moment, the error is quadratic in the data dimension. Moreover, Benton et al. (2024) further improve the results under the early-stopping setting described in the work (Chen et al., 2023), achieving the current state-of-the-art error bound that is linear in the data dimension without smoothness assumptions on the data distribution.

However, most current research on diffusion models focuses on a single worker, primarily enhancing empirical performance or exploring theoretical attributes. With the advent of the big data era, dis-

tributed training (Qiao et al., 2023; Yuan et al., 2024) is emerging as a new trend, offering potential solutions to the scalability challenges posed by increasing data volumes. In response, DistriFusion is introduced in the work (Li et al., 2024), a method designed to run diffusion models across multiple devices in parallel, significantly reducing the latency associated with generating individual samples without compromising the quality of the generated images. Despite its practical effectiveness, the theoretical performance of DistriFusion (Li et al., 2024) is not explored. Additionally, Zhao et al. (2023) proposes FedDDA, a data augmentation-based federated learning architecture that utilizes diffusion models to generate data conforming to the global class distribution, thereby alleviating the non-IID data problem. However, theoretical exploration of this approach is also lacking.

In summary, there is a theoretical gap in collaboratively training diffusion models with resource constrains. To the best of our knowledge, we are the first to derive both convergence rate of the collaboratively trained score estimation model and error evaluation of locally generated samples.

3 PRELIMINARIES

3.1 DIFFUSION MODELS

The initial phase of the diffusion model is designed to progressively transform the given data distribution q_0, into a known prior distribution. This is referred to as the forward process, and it can be described using the Ornstein-Uhlenbeck (OU) process via the stochastic differential equation (SDE) (Pedrotti et al., 2023; Benton et al., 2024):

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 \sim q_0 \tag{1}$$

where $(B_t)_{t\in[0,T]}$ denotes a standard Brownian motion on $\mathbb{R}^d$. Equation (1) aligns with a methodology known as Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020), and is also referred to as Variance Preserving SDE in (Song et al., 2020). The OU process is favored for its analytically tractable transition densities, and it holds that $X_t|X_0 \sim \mathcal{N}(X_0e^{-t}, (1-e^{-2t})\boldsymbol{I}_d)$.

We use $q_t(X_t), t \in [0,T]$ to denote the marginals of the forward process and then the reverse process satisfies the SDE:

$$\mathrm{d}X_t = -\{X_t + 2\nabla \log q_t(X_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\tilde{B}_t, \quad X_0 \sim q_0 \tag{2}$$

where $(\tilde{B}_t)_{t\in[0,T]}$ is another standard Brownian motion on $\mathbb{R}^d$. By inverting the time direction t with $T-t$ and setting $X_t = Y_{T-t}$, the reverse process (2) can be transformed to a forward one:

$$\mathrm{d}Y_t = \{Y_t + 2\nabla \log q_{T-t}(Y_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^{'}, \quad Y_0 \sim q_T \tag{3}$$

where $(B_t^{'})_{t\in[0,T]}$ is the standard Brownian motion on $\mathbb{R}^d$. The process $(Y_t)_{t\in[0,T]}$ can thus generate samples from the distribution q_0 by sampling $Y_0 \sim q_T$.

Nevertheless, Benton et al. (2024) pointed out that practical simulation of (3) necessitates overcoming certain challenges, which we also consider in this paper:
(1) Score function estimation: Since the score function $\nabla q_t(X_t)$ is unavailable, it is necessary to learn an estimation $s_\theta(X_t, t)$ of it. Specifically, the goal is to minimize the following loss function:

$$\int_0^T \mathbb{E}_{q_t(X_t)}[\| \nabla \log q_t(X_t) - s_\theta(X_t, t) \|^2]dt \tag{4}$$

While direct computation of (4) poses challenges, numerous score matching techniques (Hyvärinen & Dayan, 2005; Vincent, 2011) offer equivalent objectives that are more tractable. Among these, denoising score matching (Vincent, 2011) is utilized in this paper. Typically, we parameterize the score function $s_\theta(X_t, t)$, where s_θ represents the score, using a neural network with a parameter vector $\theta \in \mathbb{R}^D$. To optimize these parameters, we minimize the loss function through traditional SGD method over θ, effectively training the neural network to accurately estimate the score function based on the input data X_t and time t.
(2) Unknown distribution approximation: Sampling from the distribution q_T is challenging due to the inaccessibility of q_T. Instead, sampling from the standard Gaussian presents a feasible alternative, as the OU process converges exponentially quickly to the standard Gaussian (Bakry et al., 2014; Chen et al., 2023).

(3) Time discretization: Given that equation (3) characterizes a continuous-time process, practical simulation requires the time variable to be discretized. This involves dividing the continuous time into a sequence of discrete points $0 = t_0 < t_1 < t_2 < \cdots < t_K \leq T$. Subsequently, we can initiate the process by sampling $\hat{Y}_0$ from the standard Gaussian and then concentrate on solving the SDE (also known as the exponential integrator Zhang & Chen (2022); De Bortoli (2022); Chen et al. (2023)) for each interval $[t_k, t_{k+1}]$ and $k = 0, \cdots, K-1$:

$$\mathrm{d}\hat{Y}_t = \{\hat{Y}_t + 2s_\theta(\hat{Y}_{t_k}, T - t_k)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\hat{B}_t \tag{5}$$

where $(\hat{B}_t)_{t\in[0,T]}$ is a standard Brownian motion. It allows for the approximation of the continuous-time dynamics of the process within each discrete interval, facilitating the practical simulation of the model. And we denote the marginals of the process (5) by p_ts.

(4) Early stopping requirement: Instead of running (5) to approximate the initial data distribution q_0, we opt to approximate the distribution q_δ as an early-stopping measure (Song et al., 2020). This strategy is deemed acceptable because, for a sufficiently small δ, the discrepancy between q_0 and q_δ remains minimal. It is employed due to the potential for $\nabla \log q_t$ to rapidly increase, or "explode", as time t approaches zero in non-smooth data distributions.

3.2 Distributed Learning with Arbitrary Pruning

In the distributed learning framework, we consider a setup involving N workers and a central server. These workers jointly undertake the task of learning a unified global model characterized by the parameter θ. The objective is to optimize the following function:

$$\min_{\theta\in\mathbb{R}^D} F(\theta) := \frac{1}{N}\sum_{n=1}^{N} \underbrace{\mathbb{E}_{\xi_n\sim\mathcal{D}_n}[f_n(\theta, \xi_n)]}_{:=F_n(\theta)} \tag{6}$$

where $F_n(\theta)$ is a loss function defined on the dataset $\mathcal{D}_n$ based on the worker-n specified $f_n(\theta, \xi_n)$, and ξ_n signifies a data point sampled from the dataset $\mathcal{D}_n$.

In this learning framework , each worker keeps its own local dataset and conducts training operations with arbitrary pruning locally (Zhou et al., 2024). Communication with the server is restricted to the exchange of size-reduced model parameters (or gradients). More specifically, for the q-th round of the process, each worker-n performs model pruning $\theta_{q,n,0} = \theta_q \odot m_{q,n}$ after receiving the latest global model parameter $\theta_q \in \mathbb{R}^D$ from the server, where $m_{q,n} \in \{0,1\}^D$ is a local mask generated based on mask policy P. And then S steps of local training is performed to update pruned model parameters. The update rule can be described as follows:

$$\theta_{q,n,s} = \theta_{q,n,s-1} - \eta\nabla f_n(\theta_{q,n,s-1}, \xi_{n,s-1}) \odot m_{q,n} \tag{7}$$

Here, s ranges from 1 to S, with $\theta_{q,n,0}$ representing the starting parameter for each round of updates at worker-n, and η denotes the local learning rate used for the updates.

Upon completion of a round of training (encompassing S steps) by all workers, the server aggregates all local parameters to form a new global model for the forthcoming round:

$$\theta_{q+1}^{(i)} = \frac{1}{|N_q^{(i)}|}\sum_{n\in N_q^{(i)}} \theta_{q,n,S}, \qquad \text{for each coordinate } i = 1, 2, \cdots, D \tag{8}$$

where $N_q^{(i)} = \{n : m_{q,n}^i = 1\}$ and we denote $\Gamma^* = \min_{q,i} |N_q^{(i)}| \geq 1$.

4 Distributed Learning of Diffusion Models with Arbitrary Pruning

When considering training a diffusion model across multiple workers in a distributed manner, the first objective is to optimize the function described in equation (6) to obtain the score estimation $S_{\theta_Q}(\cdot)$. By employing denoising score matching (Vincent, 2011) (see Appendix B for details), the term $F_n(\theta)$ in equation (6) can be expressed as

$$F_n(\theta) = \sum_{k=0}^{K-1} \gamma_k \mathbb{E}_{X_{n,0}, q(Y_{n,t_k}|X_{n,0})}[\| s_\theta(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}|X_{n,0}) \|^2] \tag{9}$$

where $\gamma_k = t_{k+1} - t_k$ is the length of the k-th discretized time interval, and $q_{n,t}, t \in [0, T]$ denotes the marginal of the forward process of worker-n. Therefore, $X_{n,0}$ is sampled from $q_{n,0}$ by worker-n, and $Y_{n,t_k} = X_{n,T-t_k} \sim q_{n,T-k}$. However, due to the randomness in actual training, such as sampling and noise randomness, we use $f_n(\theta, \xi_n)$ to represent the local loss with randomness during training. Specially, we assume the unbiasedness of $f_n(\theta, \xi_n)$, which is common in distributed scenarios, meaning that $\mathbb{E}[f_n(\theta, \xi_n)] = F_n(\theta)$.

As described in Section 3.2, after completing Q rounds of distributed training (each with S steps), we obtain the score function estimation $s_{\theta_Q}(\cdot)$. Starting from a pure noise state, the noise is gradually transformed into a form that approximates the original data. For each worker n, this process follows Equation (10), which incorporates the worker indicator into Equation (5).

$$\mathrm{d}\widetilde{Y}_{n,t} = \{\widetilde{Y}_{n,t} + 2s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T - t_k)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\widetilde{B}_{n,t} \tag{10}$$

where $(\widetilde{B}_{n,t})_{t\in[0,T]}$ is a standard Brownian motion, and we denote the marginals of the process (10) by $p_{n,t}$s. Specifically, the equation (10) can be solved explicitly by

$$\widetilde{Y}_{n,t_{k+1}} = e^{\gamma_k}\widetilde{Y}_{n,t_k} + 2[e^{\gamma_k} - 1]s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T - t_k) + \sqrt{e^{2\gamma_k} - 1} \cdot \epsilon_{n,k}$$

where $\epsilon_{n,k} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Further details can be found in Appendix C.

To obtain the main theoretical results, we rely on the following core assumptions.

Assumption 1 ***(Lipschitzian gradient)***. *Loss function $F_n(\cdot)$s are with Lipschitzian gradients. i.e., For $\forall\theta, \phi \in \mathbb{R}^D$, it holds that*

$$\| \nabla F_n(\theta) - \nabla F_n(\phi) \| \leq L \| \theta - \phi \|$$

Assumption 2 ***(Pruning-induced Error)***. *For an arbitrary mask $m_{n,q} \in \{0, 1\}^D$ and an arbitrary model $\theta \in \mathbb{R}^D$, we assume that there exists $w^2 \in [0, 1)$:*

$$\| \theta - \theta \odot m_{n,q} \|^2 \leq w^2 \| \theta \|^2$$

Assumption 3 ***(Bounded Variance)***. *For any model θ and sample ξ, there exist $\sigma_1 > 0$ and $\sigma_2 > 0$:*

$$\mathbb{E} \| \nabla f_n(\theta, \xi) - \nabla F_n(\theta) \|^2 \leq \sigma_1^2, \qquad \mathbb{E} \| \nabla F_n(\theta) - \nabla F(\theta) \|^2 \leq \sigma_2^2$$

Assumption 4 ***(Data Distribution)***. *The data distribution $q_{n,0}$ of each worker-n has finite second moments, and is normalized so that $\mathrm{Cov}(q_{n,0}) = \boldsymbol{I}_d$.*

These assumptions (Assumptions 1 to 4) are widely used in studies on diffusion models and distributed learning. Assumption 1 (Lian et al., 2017) is typically employed to ensure the stability and solvability of optimization problems, as it guarantees that the changes in gradients will not increase without bound. Assumption 2 (Zhou et al., 2024) guarantees that pruning operations do not degrade performance beyond a certain threshold, ensuring algorithm robustness. Assumption 3 Lian et al. (2017) restricts the influence of randomness on the optimization process. For Assumption 4, its first part ensures the convergence of the forward process, while the second part simplifies result descriptions, though it is not required for the analysis (Benton et al., 2024).

Building on Assumptions 1-3 mentioned above, we can establish the convergence bound for distributed learning of score estimation with arbitrary pruning.

Theorem 1 *Under Assumptions 1-3, the following convergence result holds for distributed learning of score estimation with arbitrary pruning, provided that the step size η satisfies $0 < \eta \leq \min\{\sqrt{\frac{1}{48S^2L^2Q^2N+8S^2L^2\Gamma^*}}, \frac{3Q^2-2}{36SLQ^2+6SL}, 1\}$ where $\Gamma^* = \min_{q,i} |N_q^{(i)}| \geq 1$:*

$$\frac{1}{Q}\sum_{q=0}^{Q}\mathbb{E} \| \nabla F(\theta_q) \|^2 \leq \frac{6[F(\theta_0) - F(\theta_Q)]}{\eta SQ} + (\frac{w^2}{4\eta^2S^2Q^2} + \frac{w^4}{12\eta^2S^2Q^2} + \frac{Lw^2}{\eta SQ^2})\mathbb{E} \| \theta_0 \|^2$$
$$+ \frac{Nw^2\sigma_1^2}{S(\Gamma^*)^2} + \frac{12\eta LN\sigma_1^2}{(\Gamma^*)^2} + (\frac{1}{2Q^2} + \frac{w^2}{6Q^2} + \frac{2\eta SL}{Q^2})(\sigma_1^2 + \sigma_2^2)$$

Theorem 1 describes the rate at which the average gradient norm converges over all training rounds, which serves one of our main bounds. The term $\frac{F(\theta_0)-F(\theta_Q)}{\eta SQ}$ reflects the impact of iterative updates on the convergence behavior, while the remaining terms capture the combined effects of pruning operations, randomness, and local errors.

Specially, by tuning the appropriate step size η in Theorem 1, we can directly derive the following result:

Corollary 1 *Under Assumptions 1-3, if the step size η satisfies $\eta = \Theta(\sqrt{\frac{\Gamma^*}{SQ}})$ and $0 < \eta \leq \min\{\sqrt{\frac{1}{48S^2L^2Q^2N+8S^2L^2\Gamma^*}}, \frac{3Q^2-2}{36SLQ^2+6SL}, 1\}$, the local steps S satisfies $S = \Theta(Q)$ and we can set $Q \geq \max\{\Gamma^*, S\}$ to further derive the convergence result of Theorem 1:*

$$\frac{1}{Q}\sum_{q=0}^{Q}\mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 = \mathcal{O}\Big(\frac{6[F(\theta_0)-F(\theta_Q)]}{\sqrt{\Gamma^* SQ}} + (\frac{w^2}{4\Gamma^* SQ} + \frac{w^4}{12\Gamma^* SQ} + \frac{Lw^2}{Q\sqrt{\Gamma^* SQ}})\mathbb{E} \parallel \theta_0 \parallel^2 + \frac{Nw^2\sigma_1^2}{(\Gamma^*)^2 S} + \frac{12LN\sigma_1^2}{\Gamma^*\sqrt{\Gamma^* SQ}} + (\frac{1}{2Q^2} + \frac{w^2}{6Q^2} + \frac{2L}{\sqrt{\Gamma^* SQ}})(\sigma_1^2+\sigma_2^2)\Big)$$

Corollary 1 suggests that with an appropriately chosen step size η and a sufficient number of training rounds Q, the convergence rate of distributed learning for score estimation with arbitrary pruning can be effectively dominated by $\mathcal{O}(\frac{1}{\sqrt{\Gamma^* SQ}})$. Increasing key hyperparameters—such as the number of training rounds Q, the number of local training steps S, and the minimum occurrences Γ^* of any dimension parameter in the local model—results in tighter bounds on the average gradient norm. However, convergence can still be negatively affected by factors such as pruning-induced error, the gradient variance introduced by randomness, and discrepancies between local and global gradients.

When exploring the discrepancy between the distribution of the generated data and the true distribution of the original data, the following assumption is required for traditional single-worker architecture (Benton et al., 2024):

$$\sum_{k=0}^{K-1}\gamma_k\mathbb{E} \parallel s_\theta(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2 \leq \epsilon_{\text{score}}^2$$

However, this assumption may not fully capture the requirements of distributed diffusion model training, as it overlooks the complexity of training score estimation models in practice. By carefully addressing this, we clarify the influence of distributed training dynamics on the generated error bound, as detailed in Corollary 2.

Corollary 2 *Suppose Assumptions 1-4 hold, $T \geq 1$, and there exists a constant $C > 0$, and some $\kappa > 0$ such that for each discretized time point $k = 0, \cdots, K-1$ we have $\gamma_k \leq \kappa \min\{1, T - t_{k+1}\}$. Then under the same settings of η and Q as in Corollary 1, for each worker-n, using the collaboratively learned model θ_Q aforementioned, it yields the following result when approximating the initial data distribution:*

$$KL(q_{n,\delta} \parallel p_{n,t_K}) = \mathcal{O}\Big(F(\theta_0) + (\frac{w^2}{24\sqrt{\Gamma^* SQ}} + \frac{w^4}{72\sqrt{\Gamma^* SQ}} + \frac{Lw^2}{6Q})\mathbb{E} \parallel \theta_0 \parallel^2 + \frac{Nw^2\sigma_1^2}{6(\Gamma^*)^{3/2}} + \frac{2LN\sigma_1^2}{\Gamma^*} + (\frac{1}{12\sqrt{Q}} + \frac{w^2}{36\sqrt{Q}} + \frac{L}{3})(\sigma_1^2+\sigma_2^2) + \parallel F_n(\theta_0) - F(\theta_0) \parallel + \sigma_2 \parallel \theta_Q - \theta_0 \parallel + C(T-\delta) + \kappa dT + \kappa^2 dK + de^{-2T}\Big)$$

In Corollary 2, the term $\parallel F_n(\theta_0) - F(\theta_0) \parallel + \sigma_2 \parallel \theta_Q - \theta_0 \parallel$ captures the local-global error discrepancy. The term $C(T-\delta)$ arises from using denoising score matching to address the discretized form of Equation (4), while $\kappa dT + \kappa^2 dK$ is due to time discretization approximations, and de^{-2T} governs the convergence of the forward process. The remaining terms are interpreted as the global loss associated with θ_Q which results from the distributed learning of score estimation with arbitrary pruning. Corollary 2 highlights how the training dynamics of the score estimation model affect the final generation error. Larger values of Γ^*, S, and Q result in tighter generation error bounds. This further demonstrates that the ideal constant error assumption on score approximation (Benton et al., 2024) is inadequate for practical distributed training scenarios.

Remark 1 ***(Suitable choice of Q, T and K)****. Consider the most extreme case when* $\Gamma^* = 1$*. We introduce* ϵ^2 *to rewrite the KL error in Corollary 2 as KL*$(q_{n,\delta} \parallel p_{n,t_K}) = \mathcal{O}(\epsilon^2 + (\frac{w^2}{24\sqrt{\Gamma^* SQ}} + \frac{w^4}{72\sqrt{\Gamma^* SQ}} + \frac{Lw^2}{6Q})\mathbb{E} \parallel \theta_0 \parallel^2 + (\frac{1}{12\sqrt{Q}} + \frac{w^2}{36\sqrt{Q}})(\sigma_1^2 + \sigma_2^2) + C(T-\delta) + \kappa dT + \kappa^2 dK + de^{-2T})$*. At this point, for* $T \geq 1$*,* $\delta < 1$*,* $K \geq \log(1/\delta)$*, and some* $\kappa = \Theta\left(\frac{T+\log(1/\delta)}{K}\right)$*, if we set* $Q = \Theta(1/\epsilon^4)$*,* $T = \Theta(\min\{\frac{1}{2}\log(\frac{d}{\epsilon^2}), \frac{\epsilon^2}{C}\})$ *and* $K = \Theta(\frac{d(T+\log(1/\delta))^2}{\epsilon^2})$*, we have KL*$(q_{n,\delta} \parallel p_{n,t_K}) = \mathcal{O}(\epsilon^2)$*. In particular, larger* Γ^* *and smaller* w^2 *help tighten the bound on* $\mathcal{O}(\epsilon^2)$*.*

5 Theoretical Guarantee

In this section, we outline the proofs of the main theoretical results, with a focus on Theorem 1 and Corollary 2.

5.1 Proof Sketch of Theorem 1

Utilizing the Lipschitzian gradient assumption, we start the proof by analyzing the change in the loss function during one round as the model transitions from θ_q to θ_{q+1}:

$$\mathbb{E}[F(\theta_{q+1})] - \mathbb{E}[F(\theta_q)] \leq \underbrace{\mathbb{E}\langle \nabla F(\theta_q), \theta_{q+1} - \theta_q \rangle}_{B_1^{(q)}} + \underbrace{\frac{L}{2}\mathbb{E} \parallel \theta_{q+1} - \theta_q \parallel^2}_{B_2^{(q)}} \tag{11}$$

The first challenge in the theoretical analysis is bounding the terms $B_1^{(q)}$ and $B_2^{(q)}$. Based on the local update (7) and the global model aggregation (8), the key to analyzing these terms lies in measuring the inconsistency $B_3^{(q)}$ between the workers' and the server's gradients:

$$\begin{aligned}
B_1^{(q)} &\leq -\frac{S\eta}{2}\mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{1}{2S\eta} B_3^{(q)} \\
B_2^{(q)} &\leq \frac{3NLS\eta^2\sigma_1^2}{2(\Gamma^*)^2} + \frac{3LS^2\eta^2}{2}\mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{3L}{2} B_3^{(q)} \\
B_3^{(q)} &= \sum_{i=1}^{D} \mathbb{E} \parallel \frac{\eta}{|N_q^{(i)}|} \sum_{n \in N_q^{(i)}} \sum_{s=1}^{S} [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)] \parallel^2
\end{aligned}$$

Since each worker trains on its own data, differences in local update direction naturally arise, and multiple local steps further exacerbate these discrepancies. Moreover, the arbitrary pruning operations of local models introduce dimensional inconsistencies in the submodels trained by different workers, necessitating a more refined analysis, which significantly increases the complexity.

Measuring Inconsistency Between the Local and Global Gradients Utilizing the Cauchy-Schwarz inequality and the Lipschitzian gradient assumption, we aim to transform the gradient deviation, represented by $B_3^{(q)}$, into a corresponding deviation in the model parameters:

$$\begin{aligned}
B_3^{(q)} &= \sum_{i=1}^{D} \mathbb{E} \parallel \frac{\eta}{|N_q^{(i)}|} \sum_{n \in N_q^{(i)}} \sum_{s=1}^{S} [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)] \parallel^2 \\
&\leq \sum_{i=1}^{D} \frac{S\eta^2}{|N_q^{(i)}|} \sum_{n \in N_q^{(i)}} \sum_{s=1}^{S} \mathbb{E} \parallel \nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q) \parallel^2 \\
&\leq \frac{S\eta^2}{\Gamma^*} \sum_{n=1}^{N} \sum_{s=1}^{S} \mathbb{E} \parallel \nabla F_n(\theta_{q,n,s-1}) - \nabla F_n(\theta_q) \parallel^2 \\
&\leq \frac{L^2 S\eta^2}{\Gamma^*} \sum_{n=1}^{N} \sum_{s=1}^{S} \mathbb{E} \parallel \theta_{q,n,s-1} - \theta_q \parallel^2
\end{aligned} \tag{12}$$

Note that the term $\mathbb{E} \parallel \theta_{q,n,s-1} - \theta_q \parallel^2$ above satisfies the following inequality:

$$\begin{aligned}\mathbb{E} \parallel \theta_{q,n,s-1} - \theta_q \parallel^2 =& \mathbb{E} \parallel \theta_{q,n,s-1} - \theta_{q,n,0} + \theta_{q,n,0} - \theta_q \parallel^2 \\ \leq& \underbrace{2\mathbb{E} \parallel \theta_{q,n,s-1} - \theta_{q,n,0} \parallel^2}_{B_4^{(q)}} + \underbrace{2\mathbb{E} \parallel \theta_q \odot m_{n,q} - \theta_q \parallel^2}_{B_5^{(q)}}\end{aligned} \tag{13}$$

The $B_4^{(q)}$ term reflects the model evolution caused by local multistep iterative training, while the $B_5^{(q)}$ term represents the error resulting from local arbitrary pruning. Collectively, these two terms lead to the difference between the local model $\theta_{q,n,s-1}$ at any step $s-1$ ($s = 1, \cdots, S$) and the global model θ_q at the beginning of the current round q.

Exploring the Cumulative Entanglement of Arbitrary Pruning Operations and Local Multistep Training Local multistep training causes the gradient to cumulatively affect the model update trajectory. Although the common bounded gradient assumption simplifies the analysis, it overlooks the cumulative impact of factors like random sampling noise. Relying solely on Assumption 2 to describe the pruning error neglects the model evolution dynamics, introducing additional non-deterministic dependencies in the final convergence result and making it less intuitive (Zhou et al., 2024). Based on the above considerations, we deal with $B_4^{(q)}$ and $B_5^{(q)}$ as follows:

$$\begin{aligned}B_4^{(q)} =& 2\mathbb{E} \parallel -\eta \sum_{j=0}^{s-2} \nabla f_n(\theta_{q,n,j}, \xi_{n,j}) \odot m_{q,n} \parallel^2 \\ \leq& 2\eta^2(s-1) \sum_{j=1}^{s-1} \mathbb{E} \parallel \nabla f_n(\theta_{q,n,j-1}, \xi_{n,j-1}) - \nabla F_n(\theta_{q,n,j-1}) + \nabla F_n(\theta_{q,n,j-1}) - \nabla F_n(\theta_q) \\ &+ \nabla F_n(\theta_q) - \nabla F(\theta_q) + \nabla F(\theta_q) \parallel^2 \\ \leq& 8\eta^2(s-1)^2(\sigma_1^2 + \sigma_2^2) + 8\eta^2 L^2 (s-1) \sum_{j=1}^{s-1} \mathbb{E} \parallel \theta_{q,n,j-1} - \theta_q \parallel^2 + 8\eta^2(s-1)^2 \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 \\ B_5^{(q)} =& 2\mathbb{E} \parallel \theta_q \odot m_{n,q} - \theta_q \parallel^2 \\ =& 2\mathbb{E} \parallel \sum_{p=0}^{q-1} [(\theta_{p+1} - \theta_p) \odot m_{n,q} - (\theta_{p+1} - \theta_p)] + \theta_0 \odot m_{n,q} - \theta_0 \parallel^2 \\ \leq& 4qw^2 \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + 4w^2 \mathbb{E} \parallel \theta_0 \parallel^2\end{aligned}$$

When bounding the term $B_4^{(q)}$, we avoid the bounded gradient assumption used by Zhou et al. (2024) due to the complexity of the model evolution trajectory in practice. Instead, we utilize the existing Lipschitzian gradient and bounded variance assumptions, also proposed in their work, to derive the bound. Additionally, we have made a more refined treatment of the bound of $B_5^{(q)}$, relaxing it to the scaled sum of the accumulation of $B_2^{(q)}$ over rounds and the norm of the initial model. This treatment makes the final result independent of the average model norm throughout training, improving upon the work of Zhou et al. (2024). This improvement played a key role in the subsequent revelation of the impact of complex factors on the local score estimation error.

Next, we further bound $B_3^{(q)}$ by summing $B_4^{(q)}$ and $B_5^{(q)}$ from $s = 1$ to S as follows:

$$\begin{aligned}B_3^{(q)} \leq& \frac{4\eta^2 S^2 L^2 q w^2 N}{\Gamma^*(1 - 8\eta^2 L^2 S^2)} \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + \frac{4\eta^2 S^2 L^2 w^2 N}{\Gamma^*(1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \theta_0 \parallel^2 \\ &+ \frac{8\eta^4 S^4 L^2 N(\sigma_1^2 + \sigma_2^2)}{\Gamma^*(1 - 8\eta^2 L^2 S^2)} + \frac{8\eta^4 S^4 L^2 N}{\Gamma^*(1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2\end{aligned} \tag{14}$$

By summing $B_1^{(q)}$ and $B_2^{(q)}$ from $q = 0$ to $Q-1$, and substituting $B_3^{(q)}$ into both terms, we can then select an appropriate step size η to obtain the final convergence result.

5.2 Proof Sketch of Corollary 2

According to Corollary 1, we can obtain the following result:

$$F(\theta_Q) = \mathcal{O}\big(F(\theta_0) + \frac{Nw^2\sigma_1^2}{6(\Gamma^*)^{3/2}} + \frac{2LN\sigma_1^2}{\Gamma^*} + (\frac{w^2}{24\sqrt{\Gamma^* SQ}} + \frac{w^4}{72\sqrt{\Gamma^* SQ}} + \frac{Lw^2}{6Q})\mathbb{E} \parallel \theta_0 \parallel^2 + (\frac{1}{12\sqrt{Q}} + \frac{w^2}{36\sqrt{Q}} + \frac{L}{3})(\sigma_1^2 + \sigma_2^2)\big) \tag{15}$$

where $F(\theta_Q) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}|X_{n,0}) \parallel^2$ represents the global loss on the trained score estimation model θ_Q.

Therefore, to relax the constant assumption on the local score estimation error (Benton et al., 2024), which is $\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2 \leq \epsilon_{\text{score}}^2$, we must additionally account for two types of errors: **the loss error introduced by denoising score matching, and the discrepancy between the global loss $F(\theta_Q)$ and the local loss $F_n(\theta_Q)$**.

The former is discussed in detail in Appendix B, and we only list the result, which is

$$\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2$$
$$\leq \sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}|X_{n,0}) \parallel^2 + \sum_{k=0}^{K-1}\gamma_k C = F_n(\theta_Q) + C(T - \delta)$$

where C is a constant. As for the latter, through the constructor $h(t) = \theta_0 + t(\theta_Q - \theta_0)$, it holds that

$$F(\theta_Q) - F(\theta_0) = \int_0^1 \nabla F(h(t))^T(\theta_Q - \theta_0)dt \tag{16}$$

$$F_n(\theta_Q) - F_n(\theta_0) = \int_0^1 \nabla F_n(h(t))^T(\theta_Q - \theta_0)dt \tag{17}$$

By subtracting the two equations, applying the norm and the bounded variance assumption, we obtain

$$\parallel F_n(\theta_Q) - F(\theta_Q) \parallel \leq \parallel F_n(\theta_0) - F(\theta_0) \parallel + \sigma_2 \parallel \theta_Q - \theta_0 \parallel \tag{18}$$

By utilizing the aforementioned inequalities, we derive Corollary 2, which extends the theoretical results of Benton et al. (2024) on diffusion models in the single-worker paradigm to resource-constrained distributed scenarios.

6 Conclusion

In this paper, we provide the first generation error bound for distributed diffusion models, without assuming perfect score approximation. This theoretical bound is linear in the data dimension d, aligning with state-of-the-art results from the single-worker paradigm. Furthermore, it theoretically demonstrates how distributed training dynamics affect generation performance.

While our work enhances theoretical understanding of distributed diffusion models, it also opens up several promising directions for future research. As discussed in Remark 1, larger Γ^* and smaller w^2 help tighten the bound on $\mathcal{O}(\epsilon^2)$. However, finding the optimal balance between a low w^2 and a high Γ^* in model extraction is highly non-trivial. Additionally, resource constraints are only considered when training the score estimation model during the reverse process, but similar constraints may arise in the noise scheduling of the forward process. These open challenges will be explored in future research.

REFERENCES

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Matthijs de Goede, Bart Cox, and Jérémie Decouchant. Training diffusion models with federated learning. *arXiv preprint arXiv:2406.12575*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35: 27953–27965, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Bingkun Lai, Jiayi He, Jiawen Kang, Gaolei Li, Minrui Xu, Shengli Xie, et al. On-demand quantization for green federated generative diffusion in mobile edge networks. *arXiv preprint arXiv:2403.04430*, 2024.

Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Ming-Yu Liu, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. *arXiv preprint arXiv:2402.19481*, 2024.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Jing Qiao, Shikun Shen, Shuzhen Chen, Xiao Zhang, Tian Lan, Xiuzhen Cheng, and Dongxiao Yu. Communication resources limited decentralized learning with privacy guarantee through over-the-air computation. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 201–210, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

Ye Lin Tun, Chu Myaet Thwal, Ji Su Yoon, Sun Moo Kang, Chaoning Zhang, and Choong Seon Hong. Federated learning with diffusion models for privacy-sensitive vision tasks. In *2023 International Conference on Advanced Technologies for Communications (ATC)*, pp. 305–310. IEEE, 2023.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Jayneel Vora, Nader Bouacida, Aditya Krishnan, and Prasant Mohapatra. Feddm: Enhancing communication efficiency and handling data heterogeneity in federated diffusion models. *arXiv preprint arXiv:2407.14730*, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact langevin algorithm and score-based generative models in kl divergence. *arXiv preprint arXiv:2211.01512*, 2022.

Yuan Yuan, Shuzhen Chen, Dongxiao Yu, Zengrui Zhao, Yifei Zou, Lizhen Cui, and Xiuzhen Cheng. Distributed learning for large-scale models at edge with privacy protection. *IEEE Transactions on Computers*, 2024.

Junshan Zhang, Na Li, and Mehmet Dedeoglu. Federated learning over wireless networks: A band-limited coordinated descent approach. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2021.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Zhuang Zhao, Feng Yang, and Guirong Liang. Federated learning based on diffusion model to cope with non-iid data. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 220–231. Springer, 2023.

Hanhan Zhou, Tian Lan, Guru Prasadh Venkataramani, and Wenbo Ding. Every parameter matters: Ensuring the convergence of federated learning with dynamic heterogeneous models reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

A NOTATION TABLE

In Table 1, we summarize the main notations in this paper.

Table 1: Notations and Descriptions

Notations	Descriptions
T	The total time of noise scheduling
t	The current time of noise scheduling
K	The total number of discretized time interval of noise scheduling
t_k	The k-th discretized time point of noise scheduling, and it holds $0 = t_0 < t_1 < t_2 < \cdots < t_K \leq T$
$X_{n,t}$	The data of worker-n at time t of noise scheduling, such as image data
$Y_{n,t}$	The data of worker-n, which satisfies $Y_{n,t} = X_{n,T-t}$
$q_t, t \in [0,T]$	The marginals of the forward process
d	The dimension of data
$(B_t)_{t\in[0,T]}$	The standard Brownian motion on $\mathbb{R}^d$
$(\tilde{B}_t)_{t\in[0,T]}$	The standard Brownian motion on $\mathbb{R}^d$
$(B_t^{'})_{t\in[0,T]}$	The standard Brownian motion on $\mathbb{R}^d$
$s_\theta(X_t, t)$	The score approximation which can be parameterized by a neural network with a parameter vector $\theta \in \mathbb{R}^D$
D	The dimension of model parameter θ, $\theta \in \mathbb{R}^D$
Q	The total communication round for training the score approximation s_θ
q	The current communication round for training the score approximation s_θ
S	The number of local steps during two communication rounds
N	The total number of workers
$N_q^{(i)}$	The set of workers for which the value of coordinate-i in the mask is non-zero, and $N_q^{(i)} = \{n : m_{q,n}^i = 1\}$
Γ^*	The minimum occurrences of any dimension parameter in the local model, and $\Gamma^* = \min_{q,i} \lvert N_q^{(i)} \rvert \geq 1$
$f_n(\theta_{q,n,s}, \xi_{n,s})$	The loss of worker-n on data sample $\xi_{n,s}$ in the step s of round q
$F_n(\theta)$	The loss function of worker-n, and $F_n(\theta) = \mathbb{E}_{\xi_n \sim \mathcal{D}_n}[f_n(\theta, \xi_n)]$
$m_{q,n}$	The local mask of worker-n generated by mask policy, and $m_{q,n} \in \{0,1\}^D$
η	The step size for training the score approximation s_θ

B EQUIVALENT OBJECTIVE WITH DENOISING SCORE MATCHING

First, we considered the following loss function:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_\theta(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2 \tag{19}$$

where $\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_\theta(Y_{n,t_k}, T - t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2$ can be considered as the time-discretized version of the loss function (4). Since the score function $\nabla \log q_{n,t}(\cdot)$, we alternatively consider a denoising score matching objective, which is derived following:

$$\begin{aligned}
&\mathbb{E} \parallel s_\theta(X_{n,t}, t) - \nabla \log q(X_{n,t}) \parallel^2 \\
=&\mathbb{E} \parallel s_\theta(X_{n,t}, t) \parallel^2 + \mathbb{E} \parallel \nabla \log q(X_{n,t}) \parallel^2 - 2\mathbb{E}\langle s_\theta(X_{n,t}, t), \nabla \log q(X_{n,t})\rangle \\
=&\mathbb{E} \parallel s_\theta(X_{n,t}, t) \parallel^2 + \mathbb{E} \parallel \nabla \log q(X_{n,t}) \parallel^2 - 2\mathbb{E}_{q_{n,0}}\mathbb{E}_{q_{n,t|0}}\langle s_\theta(X_{n,t}, t), \nabla \log q_{n,t|0}(X_{n,t}|X_{n,0})\rangle \\
=&\mathbb{E} \parallel s_\theta(X_{n,t}, t) \parallel^2 + \mathbb{E} \parallel \nabla \log q(X_{n,t}) \parallel^2 + 2\mathbb{E}_{q_{n,0}}\mathbb{E}_{q_{n,t|0}}\langle s_\theta(X_{n,t}, t), \frac{X_{n,t} - e^{-t}X_{n,0}}{1 - e^{-2t}}\rangle \\
=&\mathbb{E} \parallel s_\theta(X_{n,t}, t) + \frac{X_{n,t} - e^{-t}X_{n,0}}{1 - e^{-2t}} \parallel^2 + \mathbb{E} \parallel \nabla \log q(X_{n,t}) \parallel^2 - \frac{d}{1 - e^{-2t}}
\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{E} \parallel s_\theta(X_{n,t}, t) + \frac{X_{n,t} - e^{-t}X_{n,0}}{1-e^{-2t}} \parallel^2 + C_t \\
&=\mathbb{E} \parallel s_\theta(X_{n,t}, t) - \nabla \log q_{n,t|0}(X_{n,t}|X_{n,0}) \parallel^2 + C_t
\end{aligned} \tag{20}
$$

where C_t is a constant independent of θ. Let $C = \max_t C_t$, then it holds that

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_\theta(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2 \\
\leq &\frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\gamma_k \mathbb{E} \parallel s_\theta(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}|X_{n,0}) \parallel^2 + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\gamma_k C \\
= &\frac{1}{N}\sum_{n=1}^{N}\big(F_n(\theta) + C(T-\delta)\big)
\end{aligned} \tag{21}
$$

Therefore, as measures of learning loss, Equations (9) and (19) are equivalent because the only difference between them is a constant.

C Solution to Equation (10)

Consider the Equation (10):

$$
\mathrm{d}\widetilde{Y}_{n,t} = \{\widetilde{Y}_{n,t} + 2s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T-t_k)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\widetilde{B}_{n,t}
$$

And we multiply both sides of the Equation (10) by e^{-t} to get

$$
\mathrm{d}(e^{-t}\widetilde{Y}_{n,t}) = -2s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T-t_k)\}\mathrm{d}(e^{-t}) + \sqrt{2}e^{-t}\mathrm{d}\widetilde{B}_{n,t} \tag{22}
$$

For each time interval $[t_k, t_{k+1}]$, we perform an integration operation to derive the following result:

$$
e^{-t_{k+1}}\widetilde{Y}_{n,t_{k+1}} = e^{-t_k}\widetilde{Y}_{n,t_k} + 2s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T-t_k)\}(e^{-t_k} - e^{-t_{k+1}}) + \sqrt{2}\int_{t_k}^{t_{k+1}} e^{-t}\mathrm{d}\widetilde{B}_{n,t} \tag{23}
$$

And then the following Equation (24) can be derived by multiplying both sides of the Equation (23) by $e^{t_{k+1}}$:

$$
\widetilde{Y}_{n,t_{k+1}} = e^{\gamma_k}\widetilde{Y}_{n,t_k} + 2(e^{\gamma_k} - 1)s_{\theta_Q}(\widetilde{Y}_{n,t_k}, T-t_k) + \sqrt{e^{2\gamma_k} - 1}\epsilon_{n,k} \tag{24}
$$

where $\gamma_k = t_{k+1} - t_k$ and $\epsilon_{n,k} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. And Equation (24) is exactly the solution to Equation (10).

D Proof of Theorem 1

Building on Assumption 1, we can straightforwardly deduce that the function $F(\cdot)$ is also L-smooth, satisfying the following inequality:

$$
\mathbb{E}[F(\theta_{q+1})] - \mathbb{E}[F(\theta_q)] \leq \underbrace{\mathbb{E}\langle \nabla F(\theta_q), \theta_{q+1} - \theta_q\rangle}_{B_1^{(q)}} + \underbrace{\frac{L}{2}\mathbb{E} \parallel \theta_{q+1} - \theta_q \parallel^2}_{B_2^{(q)}} \tag{25}
$$

Now, we consider the situation where $\Gamma^* \geq 1$, and we first discuss the bound of $B_1^{(q)}$:

$$
\begin{aligned}
B_1^{(q)} =&\mathbb{E}\langle \nabla F(\theta_q), \theta_{q+1} - \theta_q\rangle \\
=&\sum_{i=1}^{D}\mathbb{E}\langle \nabla F^{(i)}(\theta_q), \theta_{q+1}^{(i)} - \theta_q^{(i)}\rangle
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{D} \mathbb{E}\langle \nabla F^{(i)}(\theta_q), -\frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta \nabla f_n^{(i)}(\theta_{q,n,s-1}, \xi_{n,s-1})\rangle \\
&= \sum_{i=1}^{D} \mathbb{E}\langle \nabla F^{(i)}(\theta_q), -\frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta \nabla F_n^{(i)}(\theta_{q,n,s-1})\rangle \\
&= -\sum_{i=1}^{D} \mathbb{E}\langle \nabla F^{(i)}(\theta_q), \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)]\rangle \\
&\quad + \sum_{i=1}^{D} \mathbb{E}\langle \nabla F^{(i)}(\theta_q), -\eta S \nabla F^{(i)}(\theta_q)\rangle \\
&\leq \frac{\eta}{2S} \sum_{i=1}^{D} \mathbb{E} \parallel \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)] \parallel^2 \\
&\quad - S\eta \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{S\eta}{2} \sum_{i=1}^{D} \mathbb{E} \parallel \nabla F^{(i)}(\theta_q) \parallel^2 \\
&= -\frac{S\eta}{2} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{\eta}{2S} \sum_{i=1}^{D} \mathbb{E} \parallel \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)] \parallel^2 \\
&= -\frac{S\eta}{2} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{1}{2S\eta} B_3^{(q)} \qquad (26)
\end{aligned}
$$

where

$$
B_3^{(q)} = \sum_{i=1}^{D} \mathbb{E} \parallel \frac{\eta}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} [\nabla F_n^{(i)}(\theta_{q,n,s-1}) - \nabla F_n^{(i)}(\theta_q)] \parallel^2
$$

And we next consider how to bound $B_2^{(q)}$:

$$
\begin{aligned}
B_2^{(q)} &= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel \theta_{q+1}^{(i)} - \theta_q^{(i)} \parallel^2 \\
&= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \theta_{q,n,S}^{(i)} - \theta_q^{(i)} \parallel^2 \\
&= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \big(\theta_{q,n,S-1}^{(i)} - \eta \nabla f_n^{(i)}(\theta_{q,n,S-1}, \xi_{n,S-1}) \cdot m_{q,n}^{(i)}\big) - \theta_q^{(i)} \parallel^2 \\
&= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel -\frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta \nabla f_n^{(i)}(\theta_{q,n,s-1}, \xi_{n,s-1}) + \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \theta_{q,n,0}^{(i)} - \theta_q^{(i)} \parallel^2 \\
&= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel -\frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta \nabla f_n^{(i)}(\theta_{q,n,s-1}, \xi_{n,s-1}) \parallel^2 \\
&= \frac{L}{2} \sum_{i=1}^{D} \mathbb{E} \parallel -\frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} \eta \big(\nabla f_n^{(i)}(\theta_{q,n,s-1}, \xi_{n,s-1}) - \nabla F_n^{(i)}(\theta_{q,n,s-1}) + \nabla F_n^{(i)}(\theta_{q,n,s-1}) \\
&\qquad - \nabla F_n^{(i)}(\theta_q) + \nabla F_n^{(i)}(\theta_q) \parallel^2 \\
&\leq \frac{3L\eta^2}{2} \sum_{i=1}^{D} \mathbb{E} \parallel \frac{1}{|N_q^{(i)}|} \sum_{n\in N_q^{(i)}} \sum_{s=1}^{S} [\nabla f_n^{(i)}(\theta_{q,n,s-1}, \xi_{n,s-1}) - \nabla F_n^{(i)}(\theta_{q,n,s-1})] \parallel^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{3L}{2}\sum_{i=1}^{D}\mathbb{E}\parallel\frac{\eta}{|N_q^{(i)}|}\sum_{n\in N_q^{(i)}}\sum_{s=1}^{S}\left(\nabla F_n^{(i)}(\theta_{q,n,s-1})-\nabla F_n^{(i)}(\theta_q)\right)\parallel^2\\
&+\frac{3L}{2}\sum_{i=1}^{D}\mathbb{E}\parallel\frac{1}{|N_q^{(i)}|}\sum_{n\in N_q^{(i)}}\sum_{s=1}^{S}\eta\nabla F_n^{(i)}(\theta_q)\parallel^2\\
\leq&\frac{3NLS\eta^2\sigma_1^2}{2(\Gamma^*)^2}+\frac{3LS^2\eta^2}{2}\mathbb{E}\parallel\nabla F(\theta_q)\parallel^2+\frac{3L}{2}B_3^{(q)}
\end{aligned}
\tag{27}
$$

Therefore, discussing the bound of $B_3^{(q)}$ will help us explore $B_1^{(q)}$ and $B_2^{(q)}$:

$$
\begin{aligned}
B_3^{(q)}=&\sum_{i=1}^{D}\mathbb{E}\parallel\frac{\eta}{|N_q^{(i)}|}\sum_{n\in N_q^{(i)}}\sum_{s=1}^{S}[\nabla F_n^{(i)}(\theta_{q,n,s-1})-\nabla F_n^{(i)}(\theta_q)]\parallel^2\\
\leq&\sum_{i=1}^{D}\frac{S\eta^2}{|N_q^{(i)}|}\sum_{n\in N_q^{(i)}}\sum_{s=1}^{S}\mathbb{E}\parallel\nabla F_n^{(i)}(\theta_{q,n,s-1})-\nabla F_n^{(i)}(\theta_q)\parallel^2\\
\leq&\frac{S\eta^2}{\Gamma^*}\sum_{n=1}^{N}\sum_{s=1}^{S}\mathbb{E}\parallel\nabla F_n(\theta_{q,n,s-1})-\nabla F_n(\theta_q)\parallel^2\\
\leq&\frac{L^2S\eta^2}{\Gamma^*}\sum_{n=1}^{N}\sum_{s=1}^{S}\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_q\parallel^2
\end{aligned}
\tag{28}
$$

And it holds that

$$
\begin{aligned}
\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_q\parallel^2\leq&2\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_{q,n,0}\parallel^2+2\mathbb{E}\parallel\theta_{q,n,0}-\theta_q\parallel^2\\
=&\underbrace{2\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_{q,n,0}\parallel^2}_{B_4^{(q)}}+\underbrace{2\mathbb{E}\parallel\theta_q\odot m_{n,q}-\theta_q\parallel^2}_{B_5^{(q)}}
\end{aligned}
$$

We bound $B_4^{(q)}$ and $B_5^{(q)}$ separately:

$$
\begin{aligned}
B_4^{(q)}=&2\mathbb{E}\parallel-\eta\sum_{j=0}^{s-2}\nabla f_n(\theta_{q,n,j},\xi_{n,j})\odot m_{q,n}\parallel^2\\
\leq&2\eta^2(s-1)\sum_{j=1}^{s-1}\mathbb{E}\parallel\nabla f_n(\theta_{q,n,j-1},\xi_{n,j-1})-\nabla F_n(\theta_{q,n,j-1})+\nabla F_n(\theta_{q,n,j-1})-\nabla F_n(\theta_q)\\
&+\nabla F_n(\theta_q)-\nabla F(\theta_q)+\nabla F(\theta_q)\parallel^2\\
\leq&8\eta^2(s-1)^2(\sigma_1^2+\sigma_2^2)+8\eta^2L^2(s-1)\sum_{j=1}^{s-1}\mathbb{E}\parallel\theta_{q,n,j-1}-\theta_q\parallel^2+8\eta^2(s-1)^2\mathbb{E}\parallel\nabla F(\theta_q)\parallel^2\\
B_5^{(q)}\leq&4q\sum_{p=0}^{q-1}\mathbb{E}\parallel(\theta_{p+1}-\theta_p)\odot m_{n,q}-(\theta_{p+1}-\theta_p)\parallel^2+4\mathbb{E}\parallel\theta_0\odot m_{n,q}-\theta_0\parallel^2\\
\leq&4qw^2\sum_{p=0}^{q-1}\mathbb{E}\parallel\theta_{p+1}-\theta_p\parallel^2+4w^2\mathbb{E}\parallel\theta_0\parallel^2
\end{aligned}
$$

Then $\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_q\parallel^2$ can be bounded as

$$
\mathbb{E}\parallel\theta_{q,n,s-1}-\theta_q\parallel^2\leq4qw^2\sum_{p=0}^{q-1}\mathbb{E}\parallel\theta_{p+1}-\theta_p\parallel^2+4w^2\mathbb{E}\parallel\theta_0\parallel^2+8\eta^2(s-1)^2(\sigma_1^2+\sigma_2^2)
$$

$$+ 8\eta^2 L^2 (s-1) \sum_{j=1}^{s-1} \mathbb{E} \parallel \theta_{q,n,j-1} - \theta_q \parallel^2 + 8\eta^2 (s-1)^2 \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 \tag{29}$$

We get the following result by summing (29) from $s = 1$ to S:

$$\begin{aligned}
&(1 - 8\eta^2 L^2 S^2) \sum_{s=1}^{S} \mathbb{E} \parallel \theta_{q,n,s-1} - \theta_q \parallel^2 \\
\leq & 4Sqw^2 \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + 4Sw^2 \mathbb{E} \parallel \theta_0 \parallel^2 + 8\eta^2 S^3 (\sigma_1^2 + \sigma_2^2) + 8\eta^2 S^3 \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2
\end{aligned} \tag{30}$$

Substitute (30) into (28) to get the following bound of $B_3^{(q)}$ for $\eta < \min\{\sqrt{\frac{1}{8L^2S^2}}, 1\}$:

$$\begin{aligned}
B_3^{(q)} \leq & \frac{L^2 S \eta^2}{\Gamma^*} \sum_{n=1}^{N} \sum_{s=1}^{S} \mathbb{E} \parallel \theta_{q,n,s-1} - \theta_q \parallel^2 \\
\leq & \frac{L^2 S \eta^2}{\Gamma^*} \sum_{n=1}^{N} \Big(\frac{4Sqw^2}{1 - 8\eta^2 L^2 S^2} \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + \frac{4Sw^2}{1 - 8\eta^2 L^2 S^2} \mathbb{E} \parallel \theta_0 \parallel^2 + \\
& \frac{8\eta^2 S^3 (\sigma_1^2 + \sigma_2^2)}{1 - 8\eta^2 L^2 S^2} + \frac{8\eta^2 S^3}{1 - 8\eta^2 L^2 S^2} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 \Big) \\
\leq & \frac{4\eta^2 S^2 L^2 q w^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + \frac{4\eta^2 S^2 L^2 w^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \theta_0 \parallel^2 + \\
& \frac{8\eta^4 S^4 L^2 N (\sigma_1^2 + \sigma_2^2)}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} + \frac{8\eta^4 S^4 L^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2
\end{aligned} \tag{31}$$

Substitute the bound of $B_3^{(q)}$ (29) into (25) to get the following bound of $B_2^{(q)}$:

$$\begin{aligned}
B_2^{(q)} = & \frac{L}{2} \mathbb{E} \parallel \theta_{q+1} - \theta_q \parallel^2 \\
\leq & \frac{3NLS\eta^2 \sigma_1^2}{2(\Gamma^*)^2} + \frac{3LS^2\eta^2}{2} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{3L}{2} B_3^{(q)} \\
\leq & \frac{3NLS\eta^2 \sigma_1^2}{2(\Gamma^*)^2} + \frac{3LS^2\eta^2}{2} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + \frac{6\eta^2 S^2 L^3 q w^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \sum_{p=0}^{q-1} \mathbb{E} \parallel \theta_{p+1} - \theta_p \parallel^2 + \\
& \frac{6\eta^2 S^2 L^3 w^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \theta_0 \parallel^2 + \frac{12\eta^4 S^4 L^3 N (\sigma_1^2 + \sigma_2^2)}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} + \frac{12\eta^4 S^4 L^3 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2
\end{aligned} \tag{32}$$

Summing (32) from $q = 0$ to $Q - 1$:

$$\begin{aligned}
& [1 - \frac{12\eta^2 S^2 L^2 Q^2 w^2 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)}] \cdot \frac{L}{2} \sum_{q=0}^{Q-1} \mathbb{E} \parallel \theta_{q+1} - \theta_q \parallel^2 \\
\leq & \frac{3\eta^2 SLQN\sigma_1^2}{2(\Gamma^*)^2} + \frac{6\eta^2 S^2 L^3 Q N w^2}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} \mathbb{E} \parallel \theta_0 \parallel^2 + \frac{12\eta^4 S^4 L^3 Q N (\sigma_1^2 + \sigma_2^2)}{\Gamma^* (1 - 8\eta^2 L^2 S^2)} + \\
& [\frac{3\eta^2 S^2 L}{2} + \frac{12\eta^4 S^4 L^3 N}{\Gamma^* (1 - 8\eta^2 L^2 S^2)}] \sum_{q=0}^{Q-1} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2
\end{aligned} \tag{33}$$

If

$$\frac{12\eta^2 S^2 L^2 Q^2 w^2 N}{\Gamma^*(1-8\eta^2 L^2 S^2)} \leq \frac{w^2}{4} \leq \frac{1}{4} \Rightarrow \eta \leq \sqrt{\frac{\Gamma^*}{48 S^2 L^2 Q^2 N + 8 S^2 L^2 \Gamma^*}}$$

$$\text{and} \quad H_0 = 1 - \frac{12\eta^2 S^2 L^2 Q^2 w^2 N}{\Gamma^*(1-8\eta^2 L^2 S^2)} \geq \frac{3}{4} \Rightarrow \frac{1}{H_0} \leq \frac{4}{3},$$

then (33) can be written as

$$\begin{aligned}
&\frac{L}{2}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \theta_{q+1}-\theta_q \parallel^2 \\
\leq&\frac{2\eta^2 SLQN\sigma_1^2}{(\Gamma^*)^2} + \frac{8\eta^2 S^2 L^3 QNw^2}{\Gamma^*(1-8\eta^2 L^2 S^2)}\mathbb{E}\parallel \theta_0 \parallel^2 + \frac{16\eta^4 S^4 L^3 QN(\sigma_1^2+\sigma_2^2)}{\Gamma^*(1-8\eta^2 L^2 S^2)} + \\
&[2\eta^2 S^2 L + \frac{16\eta^4 S^4 L^3 N}{\Gamma^*(1-8\eta^2 L^2 S^2)}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 \\
\leq&\frac{2\eta^2 SLQN\sigma_1^2}{(\Gamma^*)^2} + \frac{\eta^2 S^2 L(\sigma_1^2+\sigma_2^2)}{3Q} + \frac{Lw^2}{6Q}\mathbb{E}\parallel \theta_0 \parallel^2 + [2\eta^2 S^2 L + \frac{\eta^2 S^2 L}{3Q^2}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2
\end{aligned} \tag{34}$$

We substitute the bound of $B_3^{(q)}$ (31) into (26) and sum from $q=0$ to $Q-1$ to get the following result:

$$\begin{aligned}
&\sum_{q=0}^{Q-1}\mathbb{E}\langle \nabla F(\theta_q), \theta_{q+1}-\theta_q\rangle \\
\leq& -\frac{S\eta}{2}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 + \frac{1}{2S\eta}\sum_{q=0}^{Q-1} B_3^{(q)} \\
\leq& -\frac{\eta S}{2}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 + \frac{2\eta S L^2 Q^2 N w^2}{\Gamma^*(1-8\eta^2 L^2 S^2)}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \theta_{q+1}-\theta_q \parallel^2 + \\
&\frac{2\eta S L^2 NQw^2}{\Gamma^*(1-8\eta^2 L^2 S^2)}\mathbb{E}\parallel \theta_0 \parallel^2 + \frac{4\eta^3 S^3 L^2 QN(\sigma_1^2+\sigma_2^2)}{\Gamma^*(1-8\eta^2 L^2 S^2)} + \frac{4\eta^3 S^3 L^2 N}{\Gamma^*(1-8\eta^2 L^2 S^2)}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 \\
\leq&[-\frac{\eta S}{2} + \frac{\eta S}{12Q^2}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 + \frac{\eta S(\sigma_1^2+\sigma_2^2)}{12Q} + \frac{w^2}{12\eta SL}\cdot\frac{L}{2}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \theta_{q+1}-\theta_q \parallel^2 \\
&+ \frac{w^2}{24\eta SQ}\mathbb{E}\parallel \theta_0 \parallel^2
\end{aligned} \tag{35}$$

Summing (25) from $q=0$ to $Q-1$ and substituting (34) and (35) to it, we get

$$\begin{aligned}
&F(\theta_Q) - F(\theta_0) \\
\leq&[-\frac{\eta S}{2} + \frac{\eta S}{12Q^2}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 + \frac{\eta S(\sigma_1^2+\sigma_2^2)}{12Q} + [\frac{w^2}{12\eta SL}+1]\cdot\frac{L}{2}\sum_{q=0}^{Q-1}\mathbb{E}\parallel \theta_{q+1}-\theta_q \parallel^2 \\
&+ \frac{w^2}{24\eta SQ}\mathbb{E}\parallel \theta_0 \parallel^2 \\
\leq&[-\frac{\eta S}{2} + \frac{\eta S}{12Q^2}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2 + [\frac{w^2}{12\eta SL}+1]\cdot\big(\frac{2\eta^2 SLQN\sigma_1^2}{(\Gamma^*)^2} + \frac{\eta^2 S^2 L(\sigma_1^2+\sigma_2^2)}{3Q} + \\
&\frac{Lw^2}{6Q}\mathbb{E}\parallel \theta_0 \parallel^2 + [2\eta^2 S^2 L + \frac{\eta^2 S^2 L}{3Q^2}]\sum_{q=0}^{Q-1}\mathbb{E}\parallel \nabla F(\theta_q) \parallel^2\big) + \frac{\eta S(\sigma_1^2+\sigma_2^2)}{12Q} + \frac{w^2}{24\eta SQ}\mathbb{E}\parallel \theta_0 \parallel^2
\end{aligned}$$

$$\leq H_1 \sum_{q=0}^{Q-1} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 + H_2 \mathbb{E} \parallel \theta_0 \parallel^2 + H_3(\sigma_1^2 + \sigma_2^2) + \frac{\eta Q N w^2 \sigma_1^2}{6(\Gamma^*)^2} + \frac{2\eta^2 SLQN\sigma_1^2}{(\Gamma^*)^2} \tag{36}$$

where

$$H_1 = -\frac{\eta S}{2} + \frac{\eta S}{12Q^2} + \frac{\eta S w^2}{6} + \frac{\eta S w^2}{36Q^2} + 2\eta^2 S^2 L + \frac{\eta^2 S^2 L}{3Q^2}$$
$$H_2 = \frac{w^2}{24\eta SQ} + \frac{w^4}{72\eta SQ} + \frac{Lw^2}{6Q}$$
$$H_3 = \frac{\eta S}{12Q} + \frac{\eta S w^2}{36Q} + \frac{\eta^2 S^2 L}{3Q}$$

When $0 < \eta \leq \min\{\sqrt{\frac{1}{48S^2L^2Q^2N + 8S^2L^2\Gamma^*}}, \frac{3Q^2-2}{36SLQ^2+6SL}, 1\}$, it holds that $H_1 \leq -\frac{\eta S}{6}$.

Then (36) can be rewritten as

$$\frac{1}{6Q}\sum_{q=0}^{Q} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 \leq \frac{F(\theta_0) - F(\theta_Q)}{\eta SQ} + (\frac{w^2}{24\eta^2 S^2 Q^2} + \frac{w^4}{72\eta^2 S^2 Q^2} + \frac{Lw^2}{6\eta SQ^2})\mathbb{E} \parallel \theta_0 \parallel^2$$
$$+ \frac{Nw^2\sigma_1^2}{6S(\Gamma^*)^2} + \frac{2\eta LN\sigma_1^2}{(\Gamma^*)^2} + (\frac{1}{12Q^2} + \frac{w^2}{36Q^2} + \frac{\eta SL}{3Q^2})(\sigma_1^2 + \sigma_2^2)$$

which completes the proof of Theorem 1.

E Proof of Corollary 1

If $0 < \eta \leq \min\{\sqrt{\frac{1}{48S^2L^2Q^2N + 8S^2L^2\Gamma^*}}, \frac{3Q^2-2}{36SLQ^2+6SL}, 1\}$, $\eta = \Theta(\sqrt{\frac{\Gamma^*}{SQ}})$, and $S = \Theta(Q)$, we can set $Q \geq \max\{\Gamma^*, S\}$ to derive Corollary 1 from Theorem 1:

$$\frac{1}{Q}\sum_{q=0}^{Q} \mathbb{E} \parallel \nabla F(\theta_q) \parallel^2 = \mathcal{O}\big(\frac{6[F(\theta_0) - F(\theta_Q)]}{\sqrt{\Gamma^* SQ}} + (\frac{w^2}{4\Gamma^* SQ} + \frac{w^4}{12\Gamma^* SQ} + \frac{Lw^2}{Q\sqrt{\Gamma^* SQ}})\mathbb{E} \parallel \theta_0 \parallel^2$$
$$+ \frac{Nw^2\sigma_1^2}{(\Gamma^*)^2 S} + \frac{12LN\sigma_1^2}{\Gamma^*\sqrt{\Gamma^* SQ}} + (\frac{1}{2Q^2} + \frac{w^2}{6Q^2} + \frac{2L}{\sqrt{\Gamma^* SQ}})(\sigma_1^2 + \sigma_2^2)\big)$$

where $\frac{1}{\sqrt{\Gamma^* SQ}}$ dominates the convergence rate.

F Proof of Corollary 2

According to Corollary 1, we can obtain the following result:

$$F(\theta_Q) = \mathcal{O}\big(F(\theta_0) + \frac{Nw^2\sigma_1^2}{6(\Gamma^*)^{3/2}} + \frac{2LN\sigma_1^2}{\Gamma^*} + (\frac{w^2}{24\sqrt{\Gamma^* SQ}} + \frac{w^4}{72\sqrt{\Gamma^* SQ}} + \frac{Lw^2}{6Q})\mathbb{E} \parallel \theta_0 \parallel^2$$
$$+ (\frac{1}{12\sqrt{Q}} + \frac{w^2}{36\sqrt{Q}} + \frac{L}{3})(\sigma_1^2 + \sigma_2^2)\big) \tag{37}$$

Now we need to bound the discrepancy between local and global errors $\parallel F(\theta_Q) - F_n(\theta_Q) \parallel$. Consider function $h(t) = \theta_0 + t(\theta_Q - \theta_0)$, then it holds that

$$F(\theta_Q) - F(\theta_0) = \int_0^1 \nabla F(h(t))^T (\theta_Q - \theta_0) dt \tag{38}$$

$$F_n(\theta_Q) - F_n(\theta_0) = \int_0^1 \nabla F_n(h(t))^T (\theta_Q - \theta_0) dt \tag{39}$$

Subtract the two equations and take the norm to get

$$\parallel F_n(\theta_Q) - F(\theta_Q) \parallel \leq \parallel F_n(\theta_0) - F(\theta_0) \parallel + \sigma_2 \parallel \theta_Q - \theta_0 \parallel \tag{40}$$

Then based on (21), (37) and (40), we can describe the score estimation error as

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}) \parallel^2 \\
\leq &\sum_{k=0}^{K-1} \gamma_k \mathbb{E} \parallel s_{\theta_Q}(Y_{n,t_k}, T-t_k) - \nabla \log q(Y_{n,t_k}|X_{n,0}) \parallel^2 + \sum_{k=0}^{K-1} \gamma_k C \\
=&F_n(\theta_Q) + C(T-\delta) \\
\leq&F(\theta_Q) + \parallel F_n(\theta_Q) - F(\theta_Q) \parallel + C(T-\delta) \\
=&\mathcal{O}\big(F(\theta_0) + (\frac{w^2}{24\sqrt{\Gamma^* SQ}} + \frac{w^4}{72\sqrt{\Gamma^* SQ}} + \frac{Lw^2}{6Q})\mathbb{E} \parallel \theta_0 \parallel^2 + \frac{Nw^2\sigma_1^2}{6(\Gamma^*)^{3/2}} + \frac{2LN\sigma_1^2}{\Gamma^*} \\
&+ (\frac{1}{12\sqrt{Q}} + \frac{w^2}{36\sqrt{Q}} + \frac{L}{3})(\sigma_1^2 + \sigma_2^2) + \parallel F_n(\theta_0) - F(\theta_0) \parallel + \sigma_2 \parallel \theta_Q - \theta_0 \parallel + C(T-\delta)\big)
\end{aligned} \tag{41}
$$

And according to the Theorem 1 in the work Benton et al. (2024), the Corollary 2 holds.

G Experiments

G.1 Experimental Setup

We conduct experiments using the Cifar-10 (Krizhevsky et al., 2009) SVHN (Netzer et al., 2011), and Fashion-MNIST (Xiao et al., 2017) datasets. To simulate a distributed learning scenario, we partition the training data among 10 workers. As described in Section 3.1, DDPM (Ho et al., 2020) can be viewed as a special case of our work, so we consider its distributed version (known as FedDM (Vora et al., 2024)) under resource-constrained conditions. In the experiments, we mainly consider two pruning techniques: Random Pruning (R) and Top-k Pruning (T) based on model weight. In particular, in order to explore the heterogeneity of pruning policy caused by resource differences among workers, we set for different pruning levels named F (Full), L (Large), M (Medium) and S (Small):

- **F:** All workers with full model;
- **L:** 80% workers with full model, and 20% workers with 75% model parameters;
- **M:** 60% workers with full model, 20% workers with 80% model parameters, and 20% workers with 75% model parameters;
- **S:** 60% workers with full model, and 40% workers with 75% model parameters.

We utilize multiple metrics to evaluate the performance of distributed training diffusion models with different pruning levels: Training loss is used to assess the convergence for distributed learning of score estimation. Additionally, the Inception Score (IS) and Fréchet Inception Distance (FID) are employed to evaluate the quality of data generation.

In the training stage of obtaining a score estimation, we use the U-Net backbone containing residual blocks (Tun et al., 2023). And we use the following settings unless otherwise stated: The number of communication rounds Q is set as 300, the local training steps S are configured as 5 epochs for Cifar-10 and 2 epochs for both SVHN and Fashion-MNIST, and the step size η is 0.0001.

All the experiments s are implemented in PyTorch 2.5.1, Python 3.12, Cuda 12.1. And we run them on a Cloud Server with Intel(R) Xeon(R) Platinum 8358P CPU and total 10 RTX 3090 GPUs in Ubuntu 22.04.

G.2 Model convergence for distributed learning of score estimation

We assess the convergence for distributed learning of score estimation on the above three datasets, using Random (R) and Top-k (T) pruning techniques. Specifically, we establish four pruning levels (F, L, M, and S) to observe the effects on convergence behavior. This series of experiments

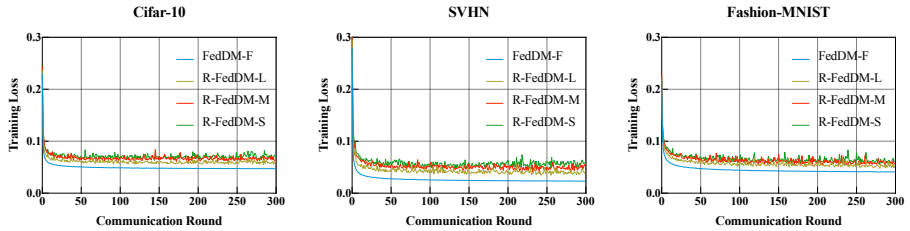

Figure 1: Training loss of FedDM under the random pruning with different pruning levels

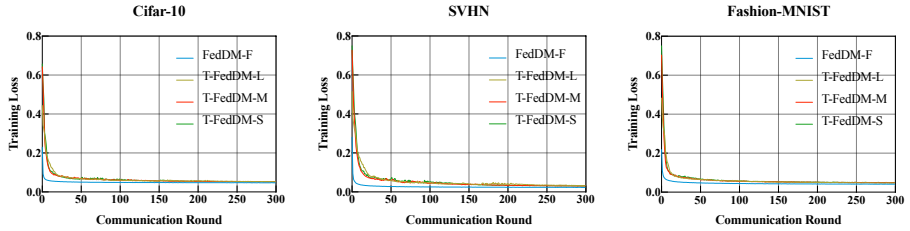

Figure 2: Training loss of FedDM under the Top-k pruning with different pruning levels

is designed to systematically evaluate how various levels of model sparsity influence the training dynamics.

Figures 1 and 2 illustrate the impact of different pruning strategies and pruning levels on the convergence rate of the distributed training diffusion model across three datasets. Overall, the training loss in all settings is effectively reduced as the number of communication rounds increases, verifying the effectiveness of the coordinate-wise aggregation method. Under both pruning strategies, as the degree of pruning increases (denoted by F, L, M, S), the training loss requires more communication rounds to decrease effectively, and the total reduction diminishes. This is because the reduced model introduces additional errors, which slows the convergence rate to a certain extent.

G.3 Data Generation Quality

We assess the performance of distributed training DDPM (known as FedDM) with different pruning levels on the above three datasets. Specifically, we establish four pruning levels (F, L, M, and S) and utilize two indicators, IS and FID, to observe and compare the average data generation quality.

As shown in Table 2, the experimental results demonstrate that pruning significantly impacts the performance of diffusion models in distributed learning, with the effects closely related to the pruning strategy, dataset complexity, and model heterogeneity. On complex datasets such as CIFAR-10 and SVHN, the full model (FedDM-F) achieves the best performance, while increased pruning levels

Table 2: IS and FID comparison of FedDM with different pruning levels.

Method	Cifar-10		SVHN		Fashion-MNIST	
	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)
FedDM-F	4.59 ± 0.13	73.73	2.79 ± 0.04	163.36	3.58 ± 0.08	87.59
R-FedDM-L	3.95 ± 0.12	103.59	2.76 ± 0.04	93.78	3.47 ± 0.04	53.70
R-FedDM-M	4.01 ± 0.08	104.53	2.60 ± 0.04	127.47	3.32 ± 0.08	52.31
R-FedDM-S	3.60 ± 0.07	111.21	2.53 ± 0.05	120.57	3.46 ± 0.07	49.94
T-FedDM-L	4.39 ± 0.08	83.75	2.72 ± 0.04	157.19	3.59 ± 0.07	87.85
T-FedDM-M	4.54 ± 0.10	80.42	2.55 ± 0.05	146.27	3.54 ± 0.06	100.69
T-FedDM-S	4.31 ± 0.13	84.98	2.51 ± 0.06	193.84	3.63 ± 0.07	109.83

lead to a substantial decline in the quality of random pruning (R-FedDM), as indicated by decreased IS scores and increased FID values, particularly at high pruning levels (e.g., S). In contrast, Top-k pruning (T-FedDM) better preserves model performance by retaining critical parameters, resulting in smaller increases in FID and performance closer to the full model, especially at moderate pruning levels (e.g., M). For simpler datasets like Fashion-MNIST, where the data distribution is less complex, pruning has a relatively smaller impact, and the performance difference between random pruning and Top-k pruning is minimal. Additionally, on Fashion-MNIST, higher pruning levels unexpectedly improve FID values. This phenomenon can be attributed to the lower capacity requirements of simple data distributions, where high pruning reduces redundant parameters, acting as a regularization effect to prevent overfitting, thus smoothing the generated distribution and making it closer to the real distribution. Model heterogeneity introduced by pruning is another critical factor affecting global performance, with random pruning more likely to cause aggregation errors, while Top-k pruning alleviates this issue to some extent. Overall, Top-k pruning proves more advantageous for complex datasets, while random pruning is better suited for resource-constrained scenarios involving simpler tasks. Future work can focus on optimizing pruning strategies and aggregation algorithms to further balance model efficiency and performance across various data distributions and task requirements.

H SOME ADDITIONAL DISCUSSION

Relaxed Assumptions and Improved Convergence Result: In deriving the convergence rate for training the score estimation model in a distributed manner, our proof builds on the work of Zhou et al., with the following key differences: 1) We eliminate their reliance on the bounded gradient assumption by modeling the iteration relationship. 2) By carefully handling the pruning error, we achieve the ultimate goal of gradient descent-based methods, allowing the final average gradient norm to converge to 0. Compared to their result, which converges only to a scaled version of $\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\parallel\theta_q\parallel^2$, our approach transforms the uncertain dependency in the convergence result into a deterministic one. 3) We achieve a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{\Gamma^* SQ}})$ by adjusting parameters such as the step size η, improving upon their result of $\mathcal{O}(\frac{1}{\sqrt{Q}})$.

Error Bound Refinement and Controllable Convergence: Directly using our analytical framework to integrate the theoretical results of Zhou et al. (2024) (Theorem 1 in their paper) with the single-worker diffusion model generation error bound, we obtain the following error bound:

$$\mathrm{KL}(q_{n,\delta}\parallel p_{n,t_K}) = \mathcal{O}\big(F(\theta_0) + \frac{3LN(\sigma_1^2+\sigma_2^2)}{2S(\Gamma)^2} + \frac{L^2NG}{2\Gamma\sqrt{Q}} + \frac{3L^2w^2N\sqrt{Q}}{\Gamma^*}\cdot\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\parallel\theta_q\parallel^2 + \parallel F_n(\theta_0)-F(\theta_0)\parallel + \sigma_2\parallel\theta_Q-\theta_0\parallel + C(T-\delta) + \kappa dT + \kappa^2 dK + de^{-2T}\big)$$

The above error bound includes an uncertainty term $\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\parallel\theta_q\parallel^2$, which prevents the bound from being tightened by adjusting Q. This limitation restricts their ability to improve the error bound in collaborative training. In contrast, our approach eliminates this uncertainty by leveraging the model iteration relationship, transforming it into a deterministic dependency. We also show that the error bound can be effectively tightened by adjusting Q, as discussed in our Remark 1. This offers a clear advantage over their results.

Unified Analytical Framework to Bridge Diffusion Models and Distributed Learning: We propose a novel framework that bridges these two areas of diffusion models and distributed learning, providing the first unified approach to connect their theoretical foundations. Specifically, we propose a simple yet effective analytical approach based on function construction (Lines 460-465) to bridge the theoretical error bounds between distributed diffusion model training and single-worker diffusion model training. Notably, this analytical approach is applicable to any generation error bound obtained under the assumption on perfect score approximation in a single-worker paradigm. We chose to integrate with the work of Benton et al. (2024) as their results represent the current state-of-the-art results in a single-worker paradigm. In fact, as long as the theoretical generation error bound in the single-worker mode based on the perfect score assumption is developed into a better result, our analytical framework allows for an immediate extension to the corresponding distributed training error bound.

Limitations and Future Work: There are still some limitations in our work, which inspire some future research directions. As discussed in Remark 1, larger Γ^* and smaller w^2 help tighten the bound on $\mathcal{O}(\epsilon^2)$. However, finding the optimal balance between a low w^2 and a high Γ^* in model extraction remains a challenging task. Therefore, it is necessary to design a suitable pruning strategy according to the specific task to balance model performance and resource consumption. Additionally, resource constraints are only considered when training the score estimation model during the reverse process. However, noise schedule during the forward process may still encounter similar constraints, which we will leave for the future.