Adaptive Non-Uniform Timestep Sampling for Accelerating Diffusion Model Training

Myunsoo Kim*, Donghyeon Ki*, Seong-Woong Shim, Byung-Jun Lee

Korea University, Gauss Lab Inc.

Motivation

- We offer a hypothesis explaining why **non-uniform training strategies may offer advantages over the uniform training approach**
- We begin by assuming an independent parameterization across diffusion timesteps, with a independent set of parameters $\theta = \{\theta_t\}_{t=1}^{T}$ for each timestep.
- **Theorem 1** implies that the number of iterations needed for SGD to converge increases in proportion to the gradient's variance.

$$\mathcal{L}_{\text{VLB}}(\theta) = \sum_{t=1}^{T} \mathbb{E}_{x_0,\epsilon}[\mathcal{L}_t(\theta)]$$

$$\mathcal{L}_t(\theta) := c_t \|\epsilon - \epsilon_\theta(x_t)\|^2, \quad c_t := \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}$$

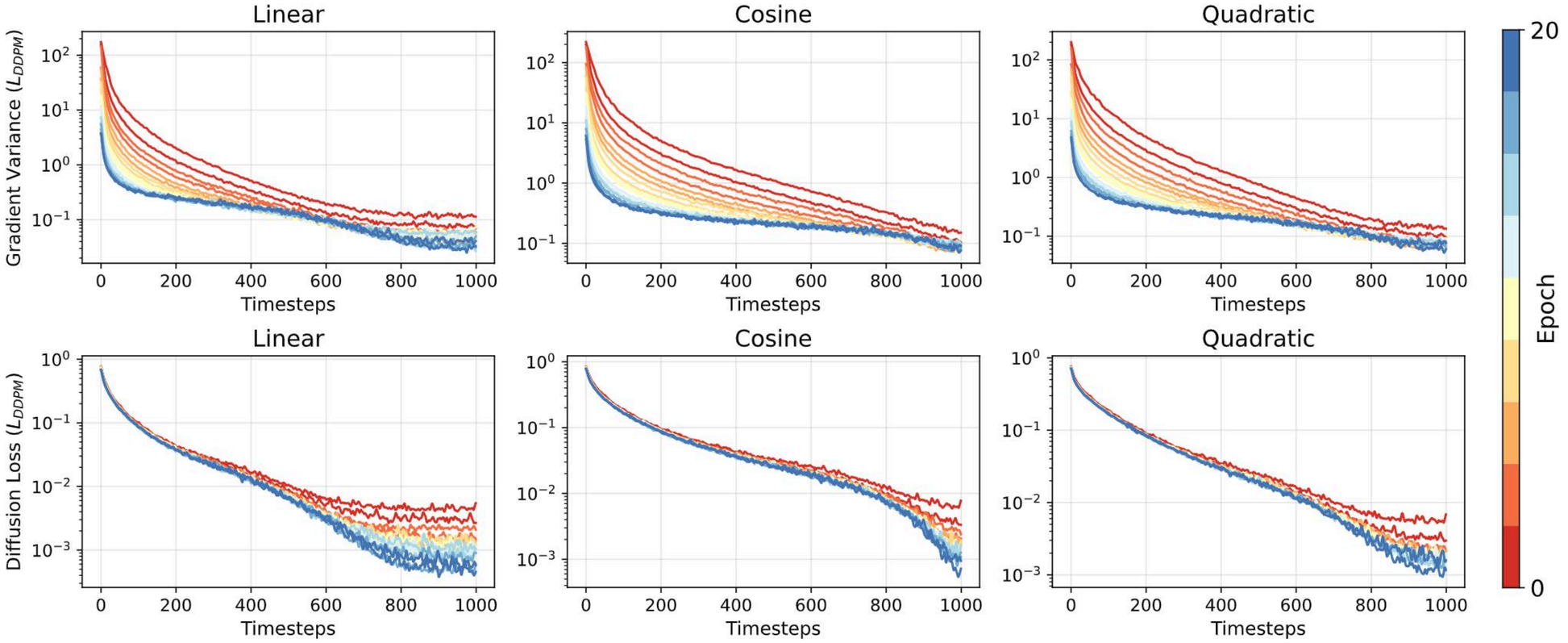
Theorem 1. [Garrigos and Gower [10], informal] *Given a few reasonable assumptions, for every $\varepsilon > 0$, we can guarantee that $\mathbb{E}\left[\mathcal{L}_t(\theta_K) - \inf \mathcal{L}_t\right] \leq \varepsilon$ provided that*

$$\gamma = \mathcal{O}\left(\frac{1}{\sqrt{K}} \left(Var_{x_0,\epsilon}[\nabla_{\theta_t} \mathcal{L}_t(\theta_*)]\right)^{-\frac{1}{2}}\right), \tag{8}$$

$$K \geq \mathcal{O}\left(\frac{1}{\varepsilon^2} Var_{x_0,\epsilon}[\nabla_{\theta_t} \mathcal{L}_t(\theta_*)]\right). \tag{9}$$

Motivation

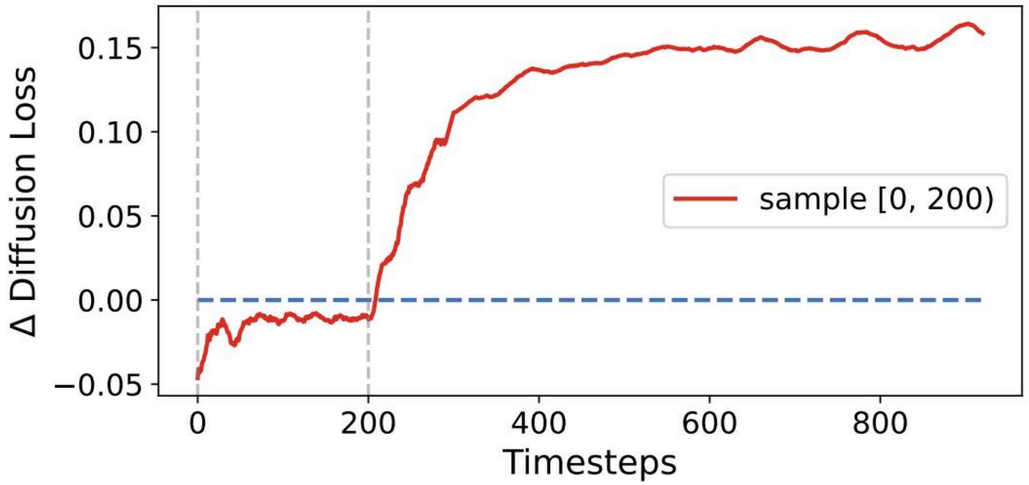
- As gradient variance at optimal parameters serves as a lower bound on the number of gradient iterations required for convergence (Theorem 1), early timesteps may bottleneck diffusion training under uniform sampling.



Interdependence of $\mathcal{L}_t$

- However, without the assumption of independence across subproblems $\mathcal{L}_t$ based on a specific parameterization, the optimality of such a sampler is not guaranteed.

- The figure reveals that the loss increase in untrained timesteps is significantly greater than the loss reduction in the trained timesteps, highlighting the strong interdependence between subproblems



	0.2M	0.4M	0.6M	0.8M	1.0M
Variance	21.3	16.0	13.5	12.1	11.3
Ours	**3.99**	**3.21**	**3.10**	**2.94**	**2.94**

Table 4. Performance (FID) comparison on the CIFAR-10 dataset between our method and the approach that samples timesteps proportional to gradient variance. We reported the lowest FID score achieved for each training duration. The algorithm achieving the lowest FID score is highlighted in bold.

Method

- We propose a non-uniform timestep sampling strategy for diffusion model training, utilizing a timestep sampler π

$$\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}_t,\ t \sim \pi(\cdot|x_0),\ \mathcal{L}_t = c_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2$$

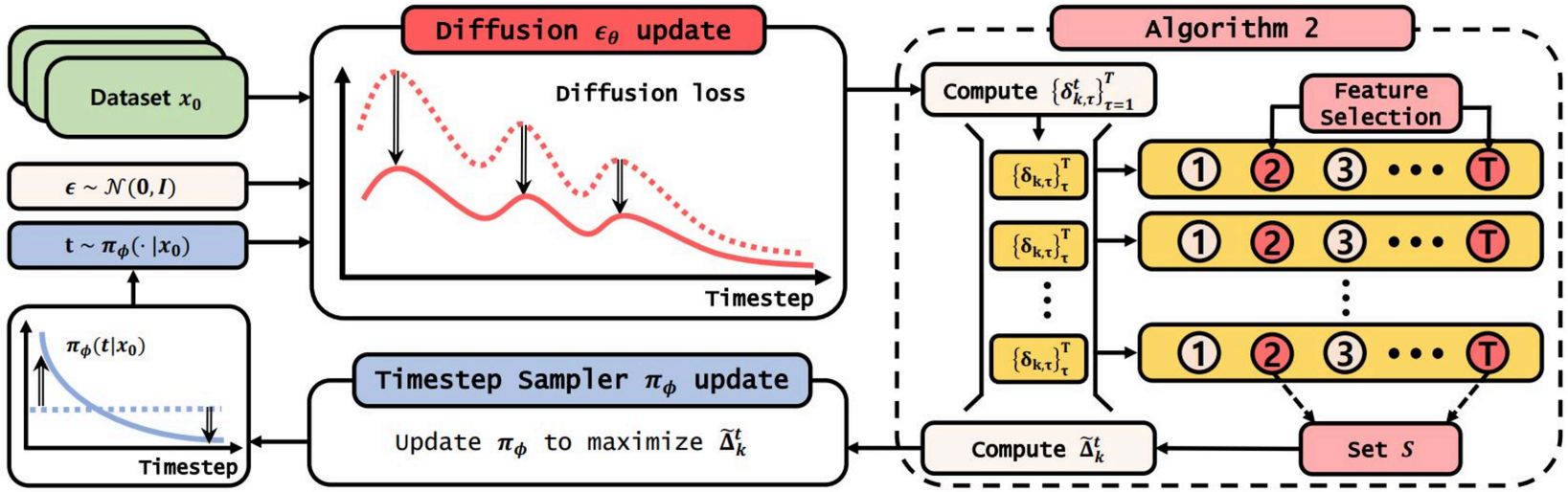
- Our goal is to train the timestep sampler π to minimize the original objective, the variational lower bound $\mathcal{L}_{\text{VLB}}$. To achieve this, we seek to optimize the sampler π_{ϕ_k} to sample timesteps t that yield the greatest reduction of $\mathcal{L}_{\text{VLB}}$ at each iteration k.

$$\phi_k = \arg\max_{\phi_k} \mathbb{E}_{x_0,\epsilon,t\sim\pi_{\phi_k}(\cdot|x_0)} \left[\Delta_k^t\right], \quad \text{where}$$

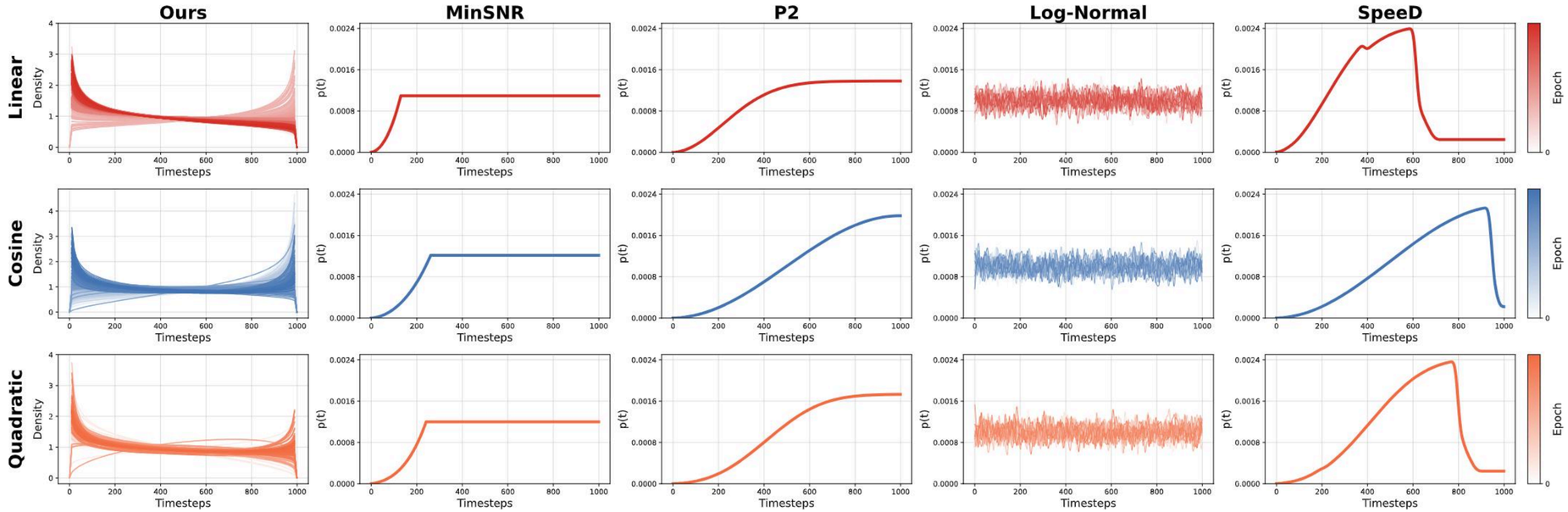
$$\Delta_k^t = \frac{1}{T}\sum_{\tau=1}^{T} \delta_{k,\tau}^t = \frac{1}{T}\sum_{\tau=1}^{T} \mathcal{L}_\tau(\theta_k) - \mathcal{L}_\tau\left(\theta_k - \gamma \nabla_{\theta_k} \mathcal{L}_t\right).$$

Method

- Following an update from θ_k to θ_{k+1}, we compute the set $\{\delta_{k,\tau}^t\}_{\tau=1}^T$ based on θ_k and θ_{k+1},and add it to a queue.
- A feature selection method is then applied to identify $|S|$ timesteps that best explains Δ_k^t from this queue.
- Then a timestep sampler π_ϕ is trained to maximize Δ_k^t, which samples timesteps t that is expected to achieve the largest reduction in the diffusion loss.

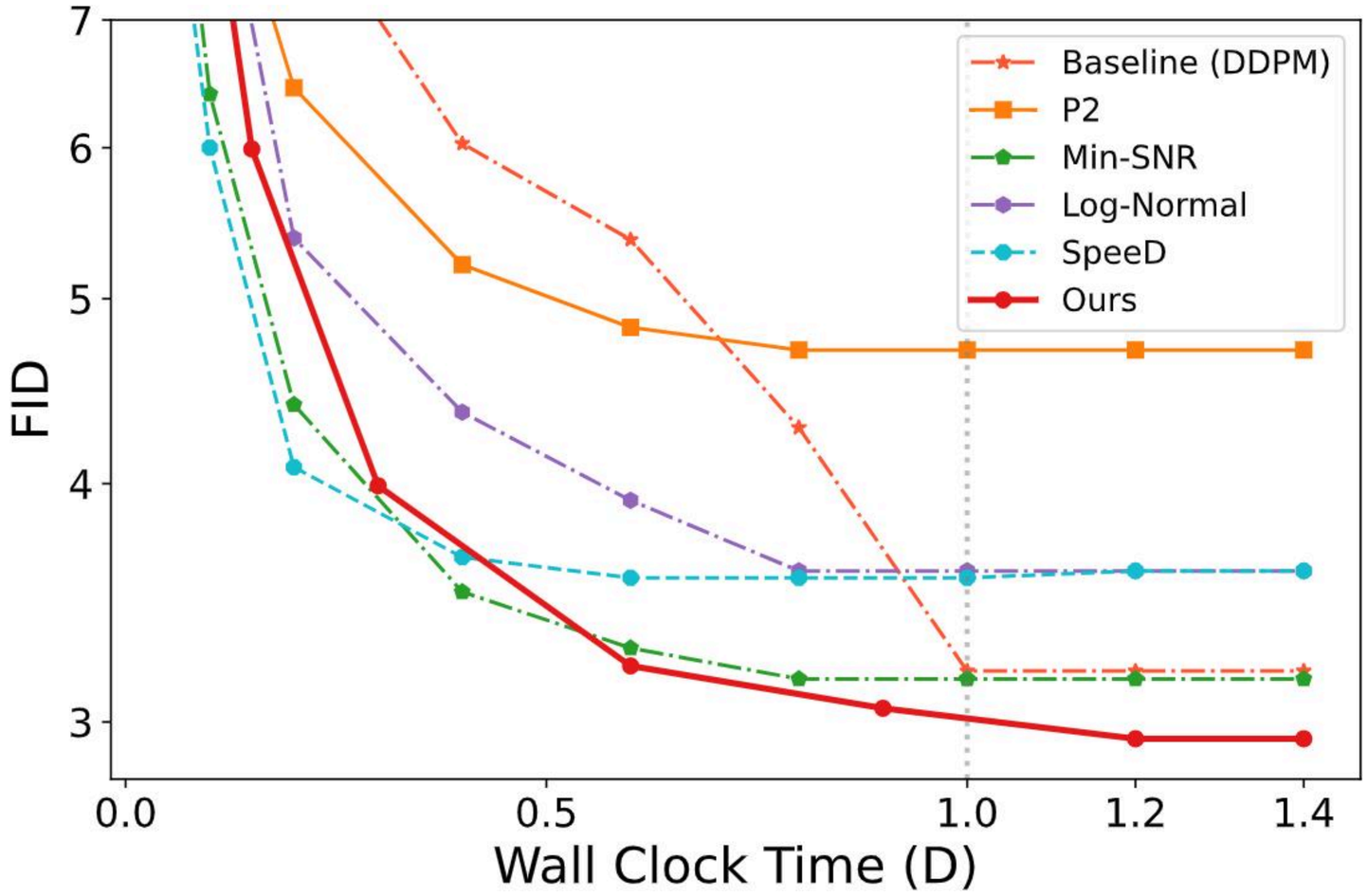


Experiments

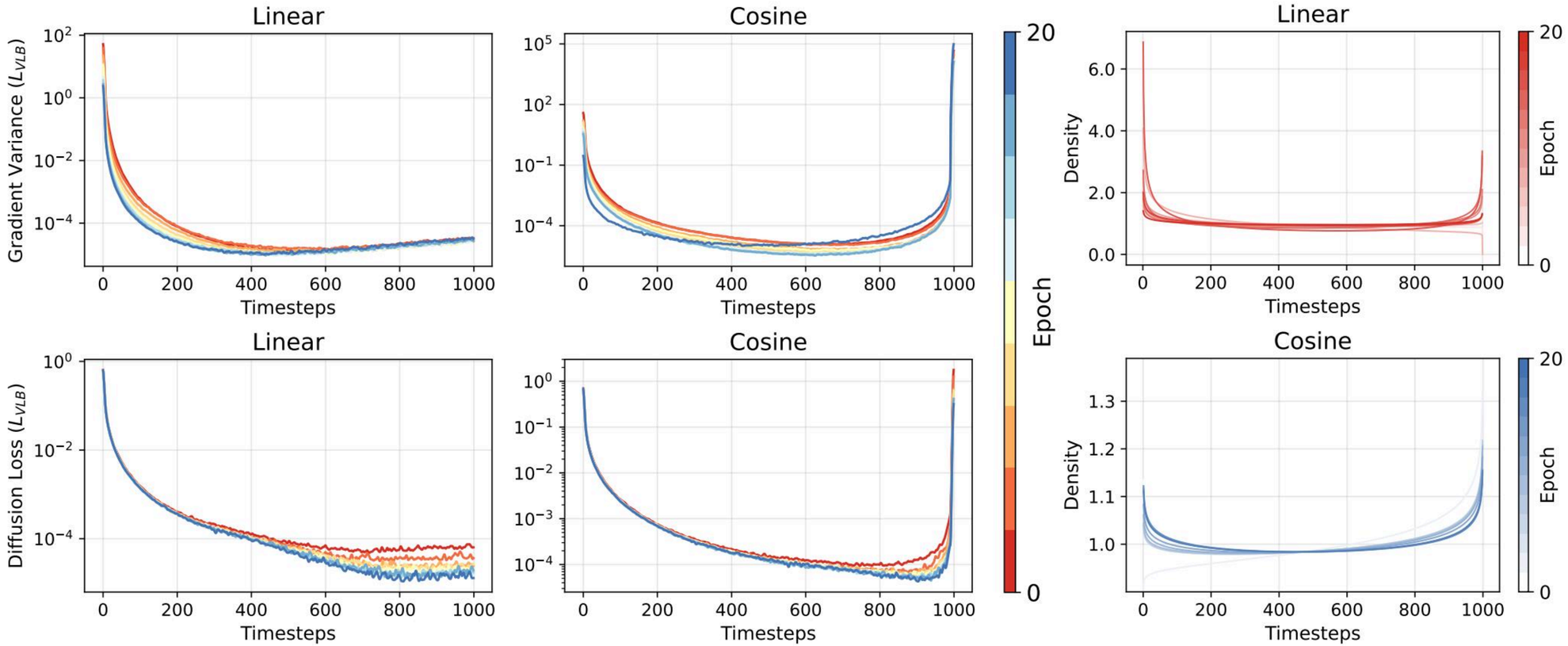


	Linear					Cosine			Quadratic			
	0.2M	0.4M	0.6M	0.8M	1.0M	0.2M	0.4M	0.6M	0.2M	0.4M	0.6M	0.8M
DDPM [14]	8.16	6.03	5.37	4.28	3.19	3.74	3.14	3.14	4.04	4.32	3.01	3.01
P2 [7]	6.45	5.21	4.83	4.70	4.70	5.64	5.21	5.04	4.93	4.06	3.74	3.66
Min-SNR [11]	4.40	3.51	3.28	3.16	3.16	**3.51**	3.07	3.07	4.76	4.17	4.17	4.17
Log-normal [17]	5.38	4.36	3.92	3.60	3.60	4.55	3.89	3.78	4.78	3.81	3.61	3.51
SpeeD [31]	4.08	3.66	3.57	3.57	3.57	4.12	3.57	4.57	3.99	3.54	3.52	3.52
Ours	**3.99**	**3.21**	**3.10**	**2.94**	**2.94**	3.87	**3.06**	**3.00**	**3.96**	**3.07**	**2.95**	**2.95**

Experiments



Experiments



Thank you