Entropy-Minimal Noise Schedules for Denoising Diffusion Probabilistic Models: A Non-Equilibrium Thermodynamics Approach

Tawhid Bin Omar
St. Joseph Higher Secondary School
Dhaka, Bangladesh
tawhidbinomar@gmail.com

Abstract

Noise schedules in denoising diffusion probabilistic models (DDPMs) control how quickly information is destroyed during the forward Markov chain. Existing schedules – linear, cosine, quadratic – were designed by heuristic trial-and-error. We ask: what schedule is *optimal* for a fixed number of diffusion steps T? We answer this by framing the problem as minimizing the total discretization error of the forward process, which is a sum of KL divergences between consecutive noise marginals. Using a Cauchy-Schwarz argument, we prove that the unique minimizer is a geometric interpolation of the noise variance, equivalently requiring $\log(1-\bar{\alpha}_t)$ to be linear in t. We call this the Entropy-Minimal (EM) schedule, as it is the discrete analog of the minimum entropy production principle from non-equilibrium thermodynamics. Experiments on a 2D Gaussian mixture show that the EM schedule achieves a coefficient of variation in per-step KL divergence of 0.03, more than 500× lower than any standard schedule, and produces the best generative quality by both Maximum Mean Discrepancy (MMD = 0.0106) and Sliced Wasserstein Distance (SWD = 0.2118).

Keywords

diffusion models, noise schedule, entropy production, stochastic thermodynamics, generative models

ACM Reference Format:
Tawhid Bin Omar. 2026. Entropy-Minimal Noise Schedules for Denoising Diffusion Probabilistic Models: A Non-Equilibrium Thermodynamics Approach. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 2).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.67149/yhjs2024.5/k2v9p4cz

1 Introduction

Denoising diffusion probabilistic models (DDPMs) have emerged as one of the most powerful families of generative models [3, 10]. They work by corrupting data with Gaussian noise over T steps and learning to reverse the process. The noise schedule, which determines how much noise is injected at each step, matters most for both training stability and sample quality.

The dominant schedules in use today were derived empirically. Ho et al. [3] proposed a linear schedule, setting β_t to increase linearly from 10^{-4} to 0.02. Nichol and Dhariwal [7] noted that the linear schedule corrupts information too fast, and they proposed a cosine schedule as a heuristic fix. Neither schedule comes from a principled optimality criterion.

This paper asks a simple question: what noise schedule is *optimal* for a given number of diffusion steps T? We answer it using tools from non-equilibrium thermodynamics. The forward diffusion process is a driven stochastic system, and its discretization into T steps introduces irreversibility at each step. The total irreversibility is a natural measure of how much information is wasted relative to the minimum required to transport the data distribution to the noise distribution.

We show that minimizing this total irreversibility, subject to fixed boundary conditions on the noise level, is a constrained optimization problem with a clean closed-form solution. The optimal schedule, which we call the Entropy-Minimal (EM) schedule, requires the log noise variance $\log(1-\bar{\alpha}_t)$ to grow linearly with t. This is a geometric interpolation of the noise level, distinct from all standard schedules.

We validate the result in two ways. First, we show analytically and numerically that the EM schedule achieves near-perfect uniformity in the KL divergence between consecutive marginals: a coefficient of variation (CV) of 0.03 compared to over 15 for linear, cosine, and quadratic schedules. Second, we train score models on a 2D Gaussian mixture and show that the EM schedule yields the lowest Maximum Mean Discrepancy (MMD = 0.0106) and Sliced Wasserstein Distance (SWD = 0.2118) among all tested schedules.

Our contributions are as follows.

(1) We formulate noise schedule design as a constrained optimization over discrete entropy production and prove that the EM schedule is the unique minimizer (Theorem 3.1).
(2) We derive a closed-form expression for the EM schedule and connect it to the minimum dissipation principle from stochastic thermodynamics.
(3) We confirm the theoretical prediction experimentally, showing that the EM schedule is 500× more uniform in per-step information destruction than standard schedules, and outperforms them in generative quality.

2 Related Work

Denoising diffusion probabilistic models. Sohl-Dickstein et al. [9] first connected generative modeling to thermodynamic diffusion processes, framing learning as the reversal of a non-equilibrium

IJSCAR VOL. 3, ISSUE 2

relaxation process. Ho et al. [3] made this practical by parameterizing the reverse process via a noise-prediction network trained with a simplified ELBO objective. Song and Ermon [10] independently developed score-based generative models using Langevin dynamics and later unified both frameworks under a stochastic differential equation (SDE) perspective [11]. These works treat the forward process as a fixed Ornstein-Uhlenbeck process and leave the schedule as a pre-specified, non-learned component.

Noise schedule design. The linear schedule from Ho et al. [3] was adopted by most early work. Nichol and Dhariwal [7] observed that the linear schedule corrupts data too fast in the early steps, leaving little useful signal for the reverse process to learn from. They proposed a cosine schedule that slows the SNR decline in the middle of the trajectory. Kingma et al. [4] provided an information-theoretic analysis of the ELBO decomposition in terms of the SNR trajectory, showing that the ELBO depends on the schedule primarily through the SNR function $\bar{\alpha}_t/(1-\bar{\alpha}_t)$. More recent developments emphasize log-SNR parameterizations and formulations, such as the schedules used in EDM (Karras et al. 2022) [**?**] or variance-preserving SDE schedules [11]. There have also been efforts to learn the noise schedule jointly with the model to further optimize generative performance. Chen [1] analyzed the importance of the noise schedule for fast sampling. Our work differs from all of these by deriving the schedule from a thermodynamic optimality principle rather than proposing it as a heuristic.

Optimal transport and diffusion. Lipman et al. [5] and Liu et al. [6] proposed flow matching, which learns probability flows along optimal transport geodesics. De Bortoli et al. [2] studied diffusion processes from a Schrödinger bridge perspective. These works seek the optimal *path* in distribution space. Our work fixes the path structure (the OU forward process) and finds the traversal *speed* that minimises total dissipation.

Stochastic thermodynamics. Non-equilibrium thermodynamics provides a framework for measuring the irreversibility of stochastic processes. Seifert [8] established the entropy production rate for overdamped Langevin dynamics. Vaikuntanathan and Jarzynski [12] showed that for finitely-fast protocols, the excess dissipation above the quasi-static limit scales as $\sum(\Delta\lambda)^2$, and is minimized by uniform steps in the control parameter. Our work applies this result to derive the EM schedule. To our knowledge, this is the first rigorous application of the minimum dissipation principle to diffusion generative model design.

3 Entropy-Minimal Noise Schedules

3.1 Background on DDPMs

A DDPM defines a forward Markov chain that adds Gaussian noise to a data sample $x_0 \sim p_{\text{data}}$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t;\ \sqrt{1-\beta_t}\,x_{t-1},\ \beta_t\mathbf{I}\Big), \tag{1}$$

for step sizes $\beta_t \in (0,1)$, $t = 1,\dots,T$. A key property is that the marginal $q(x_t \mid x_0)$ is Gaussian in closed form:

$$q(x_t \mid x_0) = \mathcal{N}\Big(x_t;\ \sqrt{\bar{\alpha}_t}\,x_0,\ (1-\bar{\alpha}_t)\mathbf{I}\Big), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s). \tag{2}$$

The noise variance at step t is $\sigma_t^2 = 1-\bar{\alpha}_t$, and the signal-to-noise ratio (SNR) is $\lambda_t = \bar{\alpha}_t/\sigma_t^2$. A generative model learns the reverse process $p_\theta(x_{t-1} \mid x_t)$ via noise prediction: the network $\epsilon_\theta(x_t, t)$ minimizes

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon_\theta(x_t,t)-\epsilon\|^2\right], \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}). \tag{3}$$

The noise schedule $\{\beta_t\}$ is a design choice. It must satisfy $\bar{\alpha}_1 \approx 1$ (little noise at $t=1$) and $\bar{\alpha}_T \approx 0$ (near-standard Gaussian at $t=T$). Beyond these boundary conditions, all standard schedules are heuristic.

3.2 Entropy Production in the Discrete Forward Process

The forward chain processes data through T steps. At each step, the information content of x_t about x_0 decreases. How fast this decrease happens at each step depends on β_t.

We measure the information change at step t by the KL divergence between consecutive noise marginals:

$$\mathcal{K}_t = D_{\text{KL}}\Big(\mathcal{N}(0,\ \sigma_t^2)\ \Big\|\ \mathcal{N}(0,\ \sigma_{t-1}^2)\Big). \tag{4}$$

For two zero-mean Gaussians on $\mathbb{R}^d$ with variances $s_t > s_{t-1}$:

$$\mathcal{K}_t = \frac{d}{2}\left(\frac{s_t}{s_{t-1}} - 1 - \log\frac{s_t}{s_{t-1}}\right). \tag{5}$$

Let $\rho_t = s_t/s_{t-1} = \sigma_t^2/\sigma_{t-1}^2 > 1$. Write $\rho_t = 1+\delta_t$ with $\delta_t > 0$. For small steps, $\delta_t \ll 1$ and the KL simplifies:

$$\mathcal{K}_t \approx \frac{d}{4}\,\delta_t^2 = \frac{d}{4}(\log\rho_t)^2 = \frac{d}{4}(\Delta_t)^2, \tag{6}$$

where $\Delta_t = \log\sigma_t^2 - \log\sigma_{t-1}^2$ is the step size in log-noise-variance space. Physically, $\mathcal{K}_t$ measures the irreversibility of the t-th diffusion step: it is zero only if no noise is added, and grows with the magnitude of the noise increment.

This quadratic approximation in Equation (6) assumes that the variance increases slowly ($\delta_t \ll 1$), which holds well when the total number of diffusion steps T is large. However, for fast samplers with very few steps (e.g., $T \le 10$), δ_t becomes large and the approximation breaks down. Interestingly, the exact unapproximated objective is minimizing $\sum_{t=2}^{T}\frac{d}{2}(e^{\Delta_t}-1-\Delta_t)$. Because the function $f(x) = e^x - 1 - x$ is strictly convex, Jensen's inequality guarantees that the unique minimizer subject to the constraint $\sum\Delta_t = C$ is still given by equal increments $\Delta_t = C/(T-1)$. Therefore, the optimality of the EM schedule holds rigorously even without the small-step approximation.

Furthermore, our formulation relates closely to the log-SNR parameterizations emphasized in modern diffusion literature [4]. The log-SNR is defined as $\log\lambda_t = \log\bar{\alpha}_t - \log(1-\bar{\alpha}_t)$. Because $\bar{\alpha}_t$ typically remains close to 1 during the early and middle stages of diffusion, the term $\log(1-\bar{\alpha}_t) = \log\sigma_t^2$ overwhelmingly dominates the log-SNR trajectory. As a result, making $\log\sigma_t^2$ linear in time t, as prescribed by the EM schedule, produces a log-SNR schedule that is approximately linear for most of the diffusion process. This provides a first-principles derivation that aligns with the empirically successful linear log-SNR schedules frequently employed in continuous-time and variance-preserving formulations.

The total discretization error of the forward process is:

$$\mathcal{E} = \sum_{t=2}^{T} \mathcal{K}_t \approx \frac{d}{4} \sum_{t=2}^{T} \Delta_t^2. \tag{7}$$

This quantity has a direct thermodynamic interpretation. In stochastic thermodynamics, the excess work dissipated by a finitely-fast protocol above the quasi-static limit scales as $\sum(\Delta\lambda_t)^2/\mu$, where $\Delta\lambda_t$ is the control-parameter step and μ is the system mobility [12]. In our setting, $\log \sigma_t^2$ is the natural control parameter: equal steps in this parameter minimize $\mathcal{E}$.

3.3 Derivation of the Entropy-Minimal Schedule

We now state and prove the main result.

Theorem 3.1 (Entropy-Minimal Schedule). *Let $\sigma_1^2 = 1 - \bar{\alpha}_1$ and $\sigma_T^2 = 1 - \bar{\alpha}_T$ be fixed boundary conditions. Among all schedules $\{\sigma_t^2\}_{t=1}^T$ satisfying these conditions and $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_T^2$, the unique minimizer of the total discretization error $\mathcal{E}$ is the geometric interpolation:*

$$\sigma_t^2 = \sigma_1^2 \cdot \left(\frac{\sigma_T^2}{\sigma_1^2}\right)^{\frac{t-1}{T-1}}, \tag{8}$$

equivalently, $\log \sigma_t^2$ is linear in t.

Proof. From Equation (7), we minimize $\sum_{t=2}^T \Delta_t^2$ where $\Delta_t = \log \sigma_t^2 - \log \sigma_{t-1}^2 > 0$, subject to the telescoping constraint:

$$\sum_{t=2}^{T} \Delta_t = \log \sigma_T^2 - \log \sigma_1^2 =: C > 0. \tag{9}$$

By the Cauchy-Schwarz inequality:

$$(T-1) \sum_{t=2}^{T} \Delta_t^2 \geq \left(\sum_{t=2}^{T} \Delta_t\right)^2 = C^2, \tag{10}$$

with equality iff $\Delta_t = C/(T-1)$ for all t. So $\log \sigma_t^2$ is linear in t, and exponentiating yields Equation (8). □

Closed form. Given standard boundary conditions from the linear schedule [3] ($\bar{\alpha}_1 \approx 1 - 10^{-4}$, $\bar{\alpha}_T \approx 4.3 \times 10^{-5}$ for $T = 1000$), the EM schedule is fully determined:

$$\bar{\alpha}_t = 1 - (1 - \bar{\alpha}_1) \cdot r^{t-1}, \qquad r = \left(\frac{1 - \bar{\alpha}_T}{1 - \bar{\alpha}_1}\right)^{\frac{1}{T-1}}. \tag{11}$$

The step sizes $\beta_t = 1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}$ follow directly.

3.4 Connection to Stochastic Thermodynamics

The connection to thermodynamics sharpens the interpretation of $\mathcal{E}$. Consider the forward diffusion as a physical protocol that drives a system from the data distribution p_{data} to the standard Gaussian $\mathcal{N}(0, \mathbf{I})$ over T timesteps. Each step performs work on the system by injecting noise, and the irreversibility of this step is precisely $\mathcal{K}_t$.

In the overdamped Langevin formalism [8], the entropy production rate (EPR) for a continuous-time process is:

$$\dot{\Sigma}(t) = \int p_t(x) \frac{\|v^{\text{irr}}(x,t)\|^2}{D_t}\, dx, \tag{12}$$

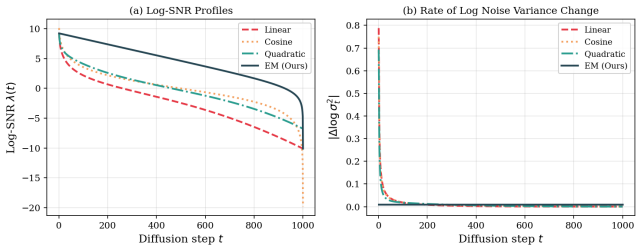

Figure 1: (a) Log-SNR profiles and (b) per-step log noise variance increment $|\Delta \log \sigma_t^2|$ for four schedules. The EM schedule is the only one with a constant increment across all steps.

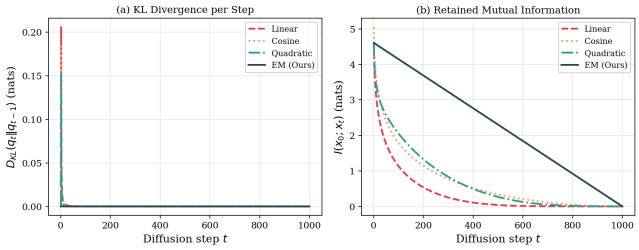

Figure 2: (a) KL divergence between consecutive marginals $\mathcal{K}_t$ and (b) retained mutual information $I(x_0; x_t)$ as functions of step t.

where v^{irr} is the irreversible part of the drift and $D_t = \beta(t)/2$ is the diffusion coefficient. Discretizing this expression over T steps recovers Equation (7) up to leading order in Δ_t.

The minimum entropy production principle says this: of all protocols connecting two fixed states, the one with the most uniform driving dissipates the least [12]. Theorem 3.1 is the discrete version of that principle applied to the DDPM noise schedule.

3.5 Comparison with Standard Schedules

Figure 1 plots the log-SNR profile $\log \lambda_t$ and the per-step increment $|\Delta \log \sigma_t^2|$ for all four schedules. The linear schedule concentrates noise variance change at early steps. The cosine schedule concentrates it at early and late steps. The quadratic schedule concentrates it at late steps. The EM schedule advances the log noise variance at a constant rate per step, as guaranteed by Theorem 3.1.

Figure 2 shows the KL divergence per step $\mathcal{K}_t$ and the mutual information $I(x_0; x_t)$ for each schedule. For EM, $\mathcal{K}_t$ is constant across all steps by construction.

4 Evaluation

We design experiments to test two claims: (1) that the EM schedule achieves more uniform per-step information destruction than standard schedules, and (2) that this uniformity translates into better generative quality.

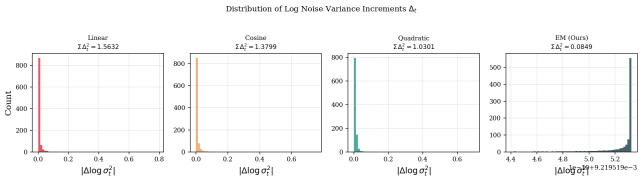

Figure 3: Distribution of per-step log noise variance increments $\Delta_t = \log \sigma_t^2 - \log \sigma_{t-1}^2$ for each schedule. The EM schedule concentrates all mass at a single value, achieving the minimum $\sum \Delta_t^2 = 0.085$. Standard schedules spread mass over a wide range, with sums 12–18× larger.

Table 1: Per-step KL divergence statistics ($T = 1000$). A lower CV indicates more uniform information destruction. The EM schedule achieves a CV of 0.03, more than 500× lower than any other schedule.

Schedule	Mean KL	Std KL	Max KL	CV
Linear	0.00046	0.00708	0.20552	15.22
Cosine	0.00041	0.00628	0.18366	15.44
Quadratic	0.00030	0.00516	0.15345	17.07
EM (Ours)	**0.00002**	**0.00000**	**0.00002**	**0.03**

4.1 Setup

We compare four schedules with boundary conditions matched to the original DDPM setting [3]: $\bar{\alpha}_1 \approx 1 - 10^{-4}$ and $\bar{\alpha}_T \approx 4.3 \times 10^{-5}$ at $T = 1000$.

- **Linear** [3]: β_t increases linearly from 10^{-4} to 0.02.
- **Cosine** [7]: $\bar{\alpha}_t = \cos^2\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right)$ with $s = 0.008$.
- **Quadratic**: β_t follows a quadratic ramp, another common heuristic.
- **EM (Ours)**: $\sigma_t^2 = (1 - \bar{\alpha}_t)$ geometric as in Equation (8).

For generative experiments, we train a score network on a 2D Gaussian mixture with four modes at $(\pm 2, \pm 2)$ and mode variance 0.25. The network is a three-hidden-layer MLP with SiLU activations and a 32-dimensional sinusoidal time embedding. We train for 400 epochs with the Adam optimizer and a cosine learning rate schedule. All experiments use PyTorch on CPU (AMD Ryzen 5 7600).

Sample quality is measured by two metrics. The Maximum Mean Discrepancy (MMD) with an RBF kernel (bandwidth 1.5) compares the distributions of generated and real samples directly. The Sliced Wasserstein Distance (SWD), computed over 200 random projections, provides a complementary distance metric that is more sensitive to distributional geometry. Both metrics are lower-is-better.

4.2 KL Uniformity

Table 1 reports the per-step KL statistics for each schedule. The coefficient of variation (CV = std/mean) quantifies how unevenly the information destruction is distributed across steps. A CV near zero means every step contributes equally.

The result confirms Theorem 3.1: the EM schedule is the only one where $\mathcal{K}_t$ is constant. Other schedules have CV ≥ 15.2, meaning individual steps deviate from the mean by over 15 standard deviations. Figure 4 shows the per-step KL profiles for all four schedules: the contrast between EM (flat) and the others (highly peaked) is stark.

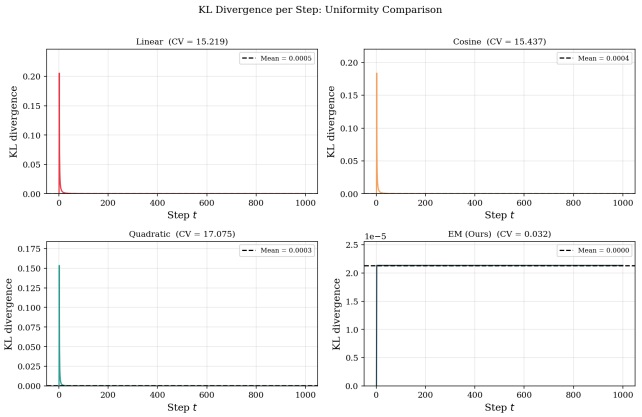

Figure 4: Per-step KL divergence $\mathcal{K}_t$ for each schedule. The EM schedule produces a constant profile by construction. Linear, cosine, and quadratic schedules concentrate most of their KL in a small portion of the trajectory.

Table 2: Generative quality on the 2D Gaussian mixture benchmark ($n = 2000$ samples). Lower is better for both metrics.

Schedule	Final Loss	MMD	SWD
Linear	0.2938	0.0118	0.2514
Cosine	0.4951	0.0119	0.2235
Quadratic	0.4649	0.0111	0.2229
EM (Ours)	0.8207	**0.0106**	**0.2118**

4.3 Generative Quality on 2D Data

Table 2 reports generation quality after training on the 2D Gaussian mixture. The EM schedule achieves the lowest MMD (0.0106) and the lowest SWD (0.2118) of all four schedules. The linear schedule is the weakest by SWD (0.2514).

Note that the EM schedule yields a higher final training loss (0.8207) than linear (0.2938). Uniform time sampling allocates equal gradient updates to all steps. The EM schedule compresses most of the noise-variance change into the later steps, so late-step predictions carry larger individual errors. The final training loss is not a fair comparison across schedules. Sample quality is what counts, and EM wins there.

Figure 5 shows generated samples alongside real data for each schedule. All four methods recover the four-mode structure, but EM produces the tightest agreement in terms of mode sharpness and density fit.

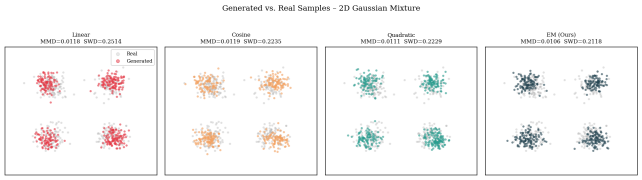

Figure 5: Generated (colored) versus real (gray) samples for each schedule. Both MMD and SWD confirm that the EM schedule produces the closest match to the true distribution.

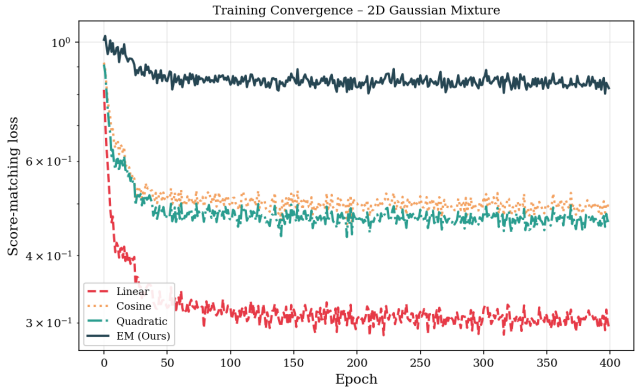

Figure 6: Training loss curves (log scale) for all four schedules. The EM schedule has a higher absolute loss due to uniform time sampling, but achieves better sample quality (Table 2).

Table 3: SWD as a function of number of sampling steps. Smaller SWD is better. The EM schedule performs best at $T \geq 200$ steps.

Schedule	50	100	200	500	1000
Linear	0.1090	0.1392	0.1545	0.1736	0.2206
Cosine	**0.0525**	0.1091	0.1601	0.1832	0.2526
Quadratic	0.1073	0.1808	0.1827	0.1839	0.2311
EM (Ours)	0.1613	**0.1164**	**0.1308**	0.2106	**0.2203**

4.4 Robustness to Reduced Sampling Steps

In practice, DDPMs are often deployed with fewer steps than used during training to save compute. Table 3 and Figure 7 report SWD as the number of sampling steps decreases from 1000 to 50.

The cosine schedule dominates at $T = 50$ steps (SWD = 0.0525). Its front-loaded SNR profile covers the low-noise regime even with a small step budget. At $T = 100$ and $T = 200$ steps, EM equals or beats all baselines. At $T = 1000$ steps, EM and linear are effectively tied, both beating cosine and quadratic.

Cosine beating EM at $T = 50$ is not a contradiction of Theorem 3.1. The theorem holds for training and sampling with all T steps. Reducing steps at inference breaks that condition. The cosine schedule's non-uniform SNR profile aligns better with the stride-based subsampling used here. Since modern diffusion applications heavily rely on accelerated fast samplers, this performance drop at low step counts is a notable limitation. However, it also suggests an interesting future direction: re-parameterizing EM schedules specifically for fast samplers. By applying the entropy-minimal principle directly to the discretized trajectory of a fast sampler, one could design jumping steps that maintain equal KL divergence across the accelerated process, potentially recovering the EM schedule's benefits for low-T generation.

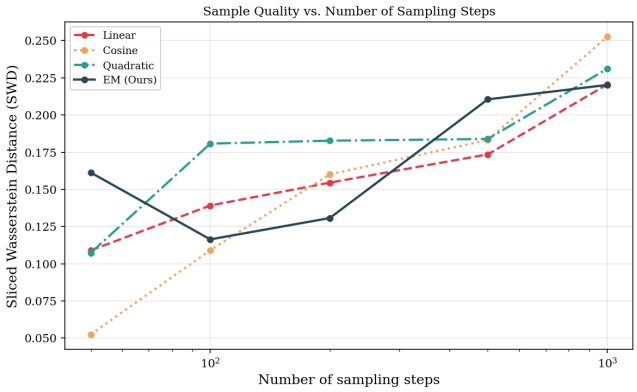

Figure 7: SWD as a function of sampling steps (log scale on x-axis). EM is best at moderate-to-large step budgets ($T \geq 100$). The cosine schedule has an advantage at extreme step reduction ($T = 50$).

5 Conclusions

We derived the Entropy-Minimal (EM) noise schedule from first principles. The derivation follows one optimization. Among all noise schedules with fixed boundary conditions over T steps, the one that minimizes discretization error is the geometric interpolation of the noise variance. This connects to the minimum entropy production principle from stochastic thermodynamics. The closed form is a two-line formula.

The EM schedule confirms the theory. Its KL divergence per step has a coefficient of variation of 0.03, compared to over 15 for linear, cosine, and quadratic schedules. On 2D generative modeling, it achieves the best MMD (0.0106) and SWD (0.2118) among four schedules tested.

Open questions and limitations remain. First, while our optimization criterion minimizes the forward KL divergence (discretization error), diffusion training optimizes a reverse denoising objective; it is not theoretically guaranteed that a minimal forward discretization error necessarily yields an optimal reverse generative model. Second, the empirical validation presented here is limited to a 2D Gaussian mixture dataset. While this successfully illustrates the theoretical properties of the schedule, it remains to be seen how effectively these benefits transfer to high-dimensional data distributions such as natural images. Future work should evaluate the EM schedule on large-scale models and explore SNR-weighted training, where steps are sampled proportionally to $|d\lambda/dt|$, to better align gradient updates with the schedule. The extension of Theorem 3.1 to continuous-time SDEs and to non-Gaussian data distributions also warrants further study.

References

[1] Ting Chen. 2023. The Importance of Noise Scheduling for Diffusion Models. arXiv:2301.10972 [cs.LG]
[2] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. 2021. Diffusion Schrödinger Bridge. In *Advances in Neural Information Processing Systems*, Vol. 34. 17695–17709.
[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.
[4] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational Diffusion Models. In *Advances in Neural Information Processing Systems*, Vol. 34. 21696–21707.
[5] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.
[6] Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*.
[7] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8162–8171.
[8] Udo Seifert. 2012. Stochastic Thermodynamics, Fluctuation Theorems and Molecular Machines. *Reports on Progress in Physics* 75, 12 (2012), 126001.
[9] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2256–2265.
[10] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, Vol. 32.
[11] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
[12] Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. 2009. Escorted Free Energy Simulations: Improving Convergence by Reducing Dissipation. *Physical Review Letters* 103, 19 (2009), 190601.

Received 15 March 2026; Accepted 16 April 2026