Extended Abstract Track

The Cosine Schedule is Fisher-Rao-Optimal for Masked Discrete Diffusion Models

Editors: List of editors' names

Abstract

In this work, we study the problem of choosing the discretisation schedule for sampling from masked discrete diffusion models in terms of the information geometry of the induced probability path. Specifically, we show that the optimal schedule under the Fisher-Rao geometry recovers the popularly-used cosine schedule.

Keywords: Discrete Diffusion Models, Information Geometry.

1. Introduction

Modern generative models, such as diffusion and flow-based models (Song et al., 2020; Lipman et al., 2022; Albergo et al., 2023), generate samples via the numerical simulation of dynamical processes. For instance, sampling from a continuous diffusion model requires simulating the time-reversal SDE parametrised by the learnt score function. Therefore, the choice of discretisation schedule is crucial for generating high quality samples.

While the problem of choosing an optimal discretisation schedule has been well-studied in the context of continuous generative models (Watson et al., 2021; Santos and Lin, 2023; Sabour et al., 2024; Williams et al., 2024), there has been growing interest (Park et al., 2024) in this question for discrete diffusion (Austin et al., 2021; Campbell et al., 2022; Shi et al., 2024; Sahoo et al., 2024). Here, the dynamical process underlying this class of models is a continuous-time Markov chain (CTMC) (Del Moral and Penev, 2017) constructed from some forward corruption process of (discrete) data; samples are then generated by simulating a learnt approximation to the time-reversal CTMC. Common choices of the forward corruption process include masked discrete diffusion (Shi et al., 2024; Sahoo et al., 2024) and uniform discrete diffusion (Sahoo et al., 2025).

In this work, we study the choice of discretisation schedule for masked discrete diffusion models from an information-geometric perspective (Amari, 2016). In particular, we show that the probabilistic structure of masked discrete diffusion allows for the closed-form computation of the Fisher-Rao metric of the induced probability path. In turn, this allows us to derive the optimal schedule - in terms of the Fisher-Rao geometry of the path - as the geodesic under this geometry, following the approach of Syed et al. (2022); Williams et al. (2024); Syed et al. (2024). Interestingly, we find that the optimal schedule recovers to the popularly-used cosine schedule (Nichol and Dhariwal, 2021).

2. Background

Masked discrete diffusion Using the notation of Shi et al. (2024), we consider the masked stochastic process $x_t = (x_t^{(1)}, \dots, x_t^{(N)})$ with $t \in [0, 1]$ and $N \in \mathbb{N}$, where each $x_t^{(n)}$

© 2025 .

represents a discrete token in $[m] = \{0, \ldots, m\}$; in particular, we take m to denote the special "masked" state. The forward noising process is then given by:

$$q(x_t|x_0) = \prod_{n=1}^{N} q(x_t^{(n)}|x_0^{(n)}), \quad q(x_t^{(n)}|x_0^{(n)}) = \mathrm{Cat}\left(x_t^{(n)}; \bar{Q}(t)^\top x_0^{(n)}\right),$$

where

$$\bar{Q}(t) = \alpha_t \mathbf{I} + (1-\alpha_t)\mathbf{1}e_m^\top, \quad \alpha_t = \exp\left(-\int_0^t \beta(s)\, ds\right),$$

for some continuous function $\beta : [0,1] \to \mathbb{R}^+$ which ensures that $\alpha_0 = 1$ and $\alpha_1 \approx 0$. We denote the probability path induced by this forward process as $q_t(x_t) = \mathbb{E}_{q_{x_0}(x_0)}[q(x_t|x_0)]$, where q_0 is our initial data distribution.

Information geometry A parametrised collection of probability densities $\mathcal{M} = \{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ can be geometrically studied as a Riemannian manifold (Amari, 2016) by endowing the parameter space Θ with a Riemannian metric $\delta(\theta) \in \mathbb{R}^{d\times d}$ (Lee, 2018). One natural choice is the Fisher-Rao metric $I(\theta) \in \mathbb{R}^{d\times d}$ given by:

$$I(\theta)_{ij} = \mathbb{E}_{X\sim p_\theta}\left[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\frac{\partial}{\partial\theta_j}\log p_\theta(X)\right].$$

We can view the Fisher-Rao metric as characterising the geometry induced by the Kullback–Leibler (KL) divergence from the Taylor expansion (Polyanskiy and Wu, 2025):

$$\mathrm{KL}(p_{\theta_0}||p_{\theta_0+\Delta\theta}) = \frac{1}{2}\Delta\theta^\top I(\theta)\Delta\theta + O(\Delta\theta^3).$$

This framework has been used to study generative models in Das et al. (2023); Santos and Lin (2023); Perone (2024); Williams et al. (2024); Karczewski et al. (2025).

Optimal schedules We note that choosing an discretisation schedule $\mathcal{T} = \{t_i\}_{i=0}^T$ where $0 = t_0 < t_1 < \ldots < t_T = 1$ can be viewed as a reparametrisation of time. Formally, let $\varphi : [0,1] \to [0,1]$ be a strictly increasing, differentiable function such that $\varphi(0) = 0$ and $\varphi(1) = 1$. We say $\mathcal{T}$ is generated by φ if $t_i = \varphi(i/T)$. From an information-geometric perspective, we can view φ as a regular path in the manifold $\mathcal{M} = \{q_t : t \in [0,1]\}$.

In order to provide a notion of optimality, Syed et al. (2022); Williams et al. (2024); Syed et al. (2024) considers placing a metric $\delta(t) \in \mathbb{R}$ on $\mathcal{M}$ so that the geometric length of the path between neighbouring discretisation points $q_{t_i}, q_{t_{i+1}}$ captures a suitable notion of the "cost" from transitioning between q_{t_i} and $q_{t_{i+1}}$[1]. In particular, the length Λ of the entire path is defined in the standard way as

$$\Lambda = \int_0^1 \sqrt{\delta(\varphi(t))}\dot{\varphi}(t)dt.$$

Hence, given some choice of metric on $\mathcal{M}$, the generator φ^* for the optimal schedule can then be naturally defined as the geodesic between q_0 and q_1. By a standard calculation, the form of φ^* is given by Theorem 1 below. See Appendix A.1 for the proof.

1. These works respectively choose metrics in the context of parallel tempering (a MCMC algorithm) and continuous diffusion models.

Extended Abstract Track

Theorem 1 *The optimal schedule under the metric $\delta(t) \in \mathbb{R}$ is generated by φ^* of the form:*

$$\varphi^*(t) = \Lambda^{-1}(\Lambda t) \quad \textit{where} \quad \Lambda(s) = \int_0^s \sqrt{\delta(r)}dr.$$

This formula is due to the fact that geodesics not only minimise the length of the path[2] between q_0 and q_1 but traverse the path at a constant rate. For example, we see that the length between $q_{t_i^*}$ and $q_{t_{i+1}^*}$ equals Λ/T for $t_i^* = \varphi^*(i/T)$. Therefore, the optimal schedule ensures that the cost between discretisation points is equally spaced apart.

Intuition behind the optimal schedule In the context of sampling from a diffusion model, the optimal schedule corresponds to discretisation steps that transition between marginals q_t that have uniform discrepancy between them (as measured by the chosen metric). This avoids taking large steps in the regions where q_t is rapidly changing and small steps where q_t is changing slowly, optimising our computational budget. Further, we note that the Fisher-Rao metric is a sensible, generic choice in this context due to the aforementioned intimate connection with the KL divergence.

3. Main Result

Below we provide a statement of our main result. See Appendix A.2 for the proof.

Theorem 2 *The optimal schedule $\mathcal{T}^* = \{t_i^*\}_{i=0}^T$ where $0 = t_0^* < t_1^* < \ldots < t_N^* = 1$ for a masked discrete diffusion model under the Fisher-Rao metric satisfies:*

$$\alpha_{t_i^*} = \cos^2\left(\frac{i}{N} \cdot \left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_1}\right)\right).$$

The values of t_i^* are given by the formula: $t_i^* = F^{-1}(-\log \alpha_{t_i^*})$ where $F(t) = \int_0^t \beta(s)ds$.

As a trivial consequence of this, we have the following corollary.

Corollary 3 *For a choice of β such that $\alpha_1 = 0$ (e.g. $\beta(s) = \frac{1}{1-s}$ and $\alpha_t = 1 - t$ using the convention that $\exp(-\infty) = 0$), the optimal schedule under the Fisher-Rao metric satisfies:*

$$\alpha_{t_i^*} = \cos^2\left(\frac{i}{T} \cdot \frac{\pi}{2}\right).$$

We see that this schedule coincides with the cosine schedule heuristically introduced in Nichol and Dhariwal (2021) (with $s = 0$) for continuous diffusion models. A similar result was proved by Santos and Lin (2023) which showed for continuous diffusion model, when the data distribution $q_0(x_0)$ is a Dirac measure - i.e. $q_0(x_0) = \delta_x(x_0)$, the cosine schedule is also optimal under the Fisher-Rao geometry. Another related result can be found in Theorem 6 from Chehab et al. (2023) which shows that the Fisher-Rao geodesic between some distributions p_0 and p_1 converges to the arithmetic annealing path with the cosine schedule, as the support of p_0, p_1 become more disjoint. These results provide an interesting theoretical justification for the use of the cosine schedule.

2. We note that since the manifold is 1D and we assume φ is a regular path, the length is actually invariant to different choices of φ.

4. Conclusion

We have shown that the optimal schedule under the Fisher-Rao geometry for masked discrete diffusion models recovers the popularly-used cosine schedule; we leave to future work the empirical validation of this schedule choice. Moreover, we hope this spurs further investigation into the information-geometric properties of generative models; for instance, considering optimal schedules under different metric choices. Finally, one limitation of our work is that we only consider the properties of the true probability path q_t. This does not take into account either the errors in the approximation of q_t or the sampling scheme (Kim et al., 2025) used during generation. It is an interesting future direction to study the information geometry of these settings.

Extended Abstract Track

References

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Omar Chehab, Aapo Hyvarinen, and Andrej Risteski. Provable benefits of annealing for estimating normalizing constants: Importance sampling, noise-contrastive estimation, and beyond. *Advances in Neural Information Processing Systems*, 36:45945–45970, 2023.

Ayan Das, Stathi Fotiadis, Anil Batra, Farhang Nabiei, FengTing Liao, Sattar Vakili, Dashan Shiu, and Alberto Bernacchia. Image generation with shortest path diffusion. In *International Conference on Machine Learning*, pages 7009–7024. PMLR, 2023.

Pierre Del Moral and Spiridon Penev. *Stochastic processes: From applications to theory*. Chapman and Hall/CRC, 2017.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. Spacetime geometry of denoising in diffusion models. *arXiv preprint arXiv:2505.17517*, 2025.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*, 2025.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Yong-Hyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, and Yuki Mitsufuji. Jump your steps: Optimizing sampling schedule of discrete diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Christian S. Perone. The geometry of data: the missing metric tensor and the stein score [part ii]. https://blog.christianperone.com/2024/11/the-geometry-of-data-part-ii/, 2024.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025.

Javier E Santos and Yen Ting Lin. Using ornstein-uhlenbeck process to understand denoising diffusion probabilistic model and its noise schedules. *arXiv preprint arXiv:2311.17673*, 2023.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.

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.

Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: a scalable highly parallel mcmc scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, 2022.

Saifuddin Syed, Alexandre Bouchard-Côté, Kevin Chern, and Arnaud Doucet. Optimised annealed sequential monte carlo samplers. *arXiv preprint arXiv:2408.12057*, 2024.

Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. *Advances in Neural Information Processing Systems*, 37: 107960–107983, 2024.

Extended Abstract Track

Appendix A. Proofs

A.1. Proof of Theorem 1

Proof We recall from standard Riemannian geometry the definition of the energy $E(\varphi)$ of the path φ:

$$E(\varphi) = \int_0^1 \delta(\varphi(s))\dot{\varphi}(s)^2 ds.$$

We note that while the length in our context is invariant for all suitable choices of φ, the energy is not invariant - i.e. the energy depends on the rate at which we traverse the path. It is a standard result that the geodesic will minimise the energy functional; we will reproduce this proof for completeness.

By the Cauchy–Schwarz inequality applied to functions in the space $L^2([0,1])$, we have:

$$\begin{aligned}
\Lambda^2 = \left(\int_0^1 \sqrt{\delta(\varphi(s))}\dot{\varphi}(s)ds\right)^2 &= \left\langle 1, \sqrt{\delta(\varphi(s))}\dot{\varphi}(s)\right\rangle^2_{L^2([0,1])} \\
&\leq ||1||^2_{L^2([0,1])} \cdot ||\sqrt{\delta(\varphi(s))}\dot{\varphi}(s)||^2_{L^2([0,1])} \\
&= \int_0^1 \delta(\varphi(t))\dot{\varphi}(s)^2 ds \\
&= E(\varphi).
\end{aligned}$$

It is a standard result that this inequality becomes an equality if and only if 1 and $\sqrt{\delta(\varphi(s))}\dot{\varphi}(s)$ are linearly dependent. Therefore, taking $\sqrt{\delta(\varphi^*(s))}\dot{\varphi}^*(s) = C$ for some constant $C > 0$, we have

$$Ct = \int_0^t \sqrt{\delta(\varphi^*(s))}\dot{\varphi}^*(s)ds = \int_0^{\varphi^*(t)} \sqrt{\delta(r)}dr = \Lambda(\varphi^*(t)),$$

where $C = \Lambda$ from taking $t = 1$ and the second equality follows from the substitution $r = \varphi^*(t)$. Moreover, this implies that

$$\varphi^*(t) = \Lambda^{-1}(\Lambda t).$$

We see by construction that φ^* is both a length minimiser and travels at a constant rate, hence, we can conclude that φ^* is a geodesic. ■

A.2. Proof of Theorem 2

Proof Due to the structure of the probability path induced by masked discrete diffusion models, we can derive the following analytic form for $q_t(x_t)$:

$$q_t(x_t) = (1-\alpha_t)^{n_m(x_t)}\alpha_t^{N-n_m(x_t)}\mathbb{E}_{q(x_0)}\left[\mathbf{1}(x_t^{(i)} = x_0^{(i)} : x_t^{(i)} \neq m)\right],$$

and by a similar argument, $\partial_t q_t(x_t)$ has the form:

$$\begin{aligned}
\partial_t q_t(x_t) &= \partial_t \mathbb{E}_{q(x_0)}\left[\prod_{n=1}^N q(x_t^{(i)}|x_0^{(i)})\right] \\
&= \partial_t (1-\alpha_t)^{n_m(x_t)}\alpha_t^{N-n_m(x_t)}\mathbb{E}_{q(x_0)}\left[\mathbf{1}(x_t^{(i)} = x_0^{(i)} : x_t^{(i)} \neq m)\right],
\end{aligned}$$

where we define $n_m(x_t)$ as the number of masked tokens m present in x_t.

Hence, the Fisher score has the form:

$$\begin{aligned}\partial_t \log q_t(x_t) &= \partial_t \log \left((1-\alpha_t)^{n_m(x_t)} \alpha_t^{N-n_m(x_t)} \right) \\ &= -\frac{n_m(x_t)\dot{\alpha}_t}{1-\alpha_t} + (N - n_m(x_t))\frac{\dot{\alpha}_t}{\alpha_t}.\end{aligned}$$

To compute the Fisher-Rao metric $I(t) \in \mathbb{R}$, we can use the standard result:

$$I(t) = \mathrm{Var}_{x_t \sim q_t} \left(\partial_t \log q_t(x_t) \right),$$

with the fact that $n_m(x_t)$ follows a $\mathrm{Bin}(N, 1-\alpha_t)$ distribution when $x_t \sim q_t$ (since the probability of $x_t^{(i)}$ being a masked token m is Bernoulli and we assume a factorised forward process). This allows us to conclude

$$\begin{aligned}I(t) = \mathrm{Var}_{x_t \sim q_t}(\partial_t \log q_t(x_t)) &= \frac{N\alpha_t(1-\alpha_t)\dot{\alpha}_t^2}{\alpha_t^2(1-\alpha_t)^2} \\ &= \frac{N\dot{\alpha}_t^2}{\alpha_t(1-\alpha_t)}.\end{aligned}$$

In order to apply Theorem 1, we compute:

$$\begin{aligned}\Lambda(t) = \int_0^t \sqrt{\frac{N\dot{\alpha}_t^2}{\alpha_t(1-\alpha_t)}}dt = -\int_1^{\alpha_t} \sqrt{\frac{N}{y(1-y)}}dy &= 2\sqrt{N}[\arcsin\sqrt{y}]_{\alpha_t}^1 \\ &= 2\sqrt{N}\left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_t}\right),\end{aligned}$$

where for the first equality, we use the substitution $y = \alpha_t$ and $\frac{|\dot{\alpha}_t|}{\dot{\alpha}_t} = -1$ since α_t is constructed to be strictly decreasing, and for the second equality, we use a standard integral identity.

Furthermore, to find the generator φ^* for the optimal schedule, we use the identity $\Lambda(\varphi^*(t)) = \Lambda t$ from Theorem 1 to show

$$\begin{aligned}2\sqrt{N}\left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_{\varphi^*(t)}}\right) &= 2\sqrt{N}t\left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_1}\right) \\ \implies \alpha_{\varphi^*(t)} &= \sin^2\left(\frac{\pi}{2} - t\left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_1}\right)\right) \\ &= \cos^2\left(t \cdot \left(\frac{\pi}{2} - \arcsin\sqrt{\alpha_1}\right)\right),\end{aligned}$$

where the last equality uses the trigonometric identity $\sin(a+b) = \sin a \cos b + \sin b \cos a$. We conclude the main result from substituting the values $t = i/T$ into φ^*. Finally, the formula for t_i^* is a simple consequence of the fact that $t \mapsto \alpha_t$ is an injective map.

■