A Unification of Discrete, Gaussian, and Simplicial Diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

To model discrete sequences such as DNA, proteins, and language using diffusion, practitioners must choose between three major methods: diffusion in discrete space, Gaussian diffusion in Euclidean space, or diffusion on the simplex. Despite their shared goal, these models have disparate algorithms, theoretical structures, and strengths. Ideally we could see each of these models as instances of the same underlying framework, and practitioners could seamlessly transition between the domains to fit their applications. However previous theories have only considered connections in special cases. Here we unify all three methods of discrete diffusion as different parameterizations of the same underlying process: the Wright-Fisher population genetics model. We find simplicial and Gaussian diffusion as two large-population limits. Our theory formally connects the likelihoods and hyperparameters of these models. Finally, we relieve the practitioner of balancing model trade-offs by demonstrating it is possible to train a single model that can perform diffusion in any of these three domains at test time. In a proof of concept result, we show that we can train models on multiple domains at once that are competitive with models trained on any individual domain.

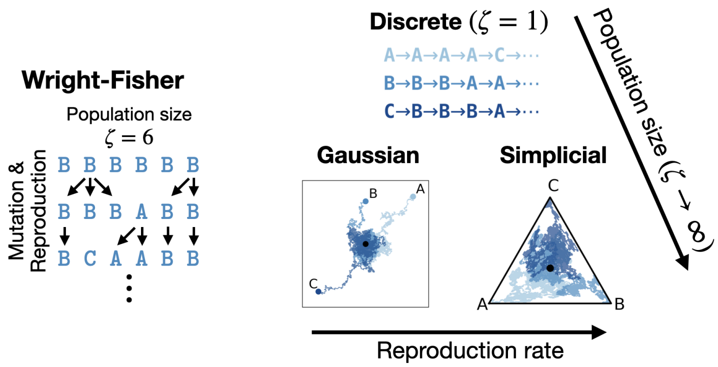
1 Introduction

Practitioners build diffusion models of language, DNA, and proteins to generate high quality sequences conditioned on desirable properties [Sahoo et al., 2024, Sarkar et al., 2024, Alamdari et al., 2023]. These models are used for conditional generation [Wang et al., 2024], optimization [Gruver et al., 2023], and myriad other tasks [Luo et al., 2022, Baron et al., 2025].

A practitioner has three main choices when modeling discrete sequence data with diffusion (Fig. 1b): (1) Discrete diffusion: the most straightforward and natural domain [Campbell et al., 2022]. (2) Gaussian diffusion: a mature field with elaborate sampling and training procedures [Dieleman et al., 2022]. (3) Simplicial diffusion: in theory inherits the continuous algorithms of Gaussian diffusion while working in a "natural" space, but in practice suffers from severe numerical instability issues [Avdeyev et al., 2023].

Unfortunately, there is little work comparing these models, and thus practitioners have minimal practical guidance on model selection. We give an overview of existing theories connecting these models in Appendix A. The gap in theory is particularly evident in light of basic comparison problems which have yet to be solved: (1) **Loss comparisons:** Diffusion models are trained to optimize a lower bound on the likelihood (ELBO). However, despite models achieving similar ELBO values, there is a belief that the "continuous-space likelihood is not directly comparable with discrete-space likelihood" [Avdeyev et al., 2023]. (2) **Hyperparameter comparisons:** Each of these models are specified by hyperparameters with different interpretations, and there is no mechanism to qualitatively compare the assumptions each set of these hyperparameters are making across models.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.



(a) (b)

Figure 1: **Discrete, Gaussian, and Simplicial diffusion for discrete data are unified by Wright-Fisher diffusion.** **(a)** Wright-Fisher diffusion with population size $\zeta = 6$, showing mutation and reproduction processes across generations. **(b)** The three diffusion methods emerge as different limits of Wright-Fisher: discrete diffusion corresponds to $\zeta = 1$, while Gaussian and simplicial diffusion arise as $\zeta \to \infty$ with zero and non-zero reproduction rates.

We address these theoretical and practical challenges by unifying diffusion methods with a process from human population genetics – the Wright-Fisher (WF) model. We formally prove all three methods are instances of WF (Fig. 1). We use this connection to answer the basic theoretical questions of loss and hyperparameter comparison. Then, for the practitioner, we show that a particular parameterization choice – the **sufficient-statistic parameterization** – allows one to train a single model that can perform diffusion on all three domains at test time. We show in a proof of concept that models trained this way can be competitive with models trained on individual domains.

2 Diffusion models for discrete data

We consider modeling a distribution $p(x_0)$ over a discrete space of size B, and extend to sequences of discrete objects in B.2. Our model begins by sampling from distribution $q(x_1)$, and then applies a stochastic process parametrized by θ from time 1 to 0. This produces a trajectory $q_\theta((x_t)_{t=0}^1)$ and we hope to pick θ so that $q_\theta(x_0) \sim p(x_0)$.

Markov processes To generate training data to fit $q_\theta((x_t)_{t=0}^1)$, we take samples $x_0 \sim p(x_0)$ and evolve it according to a Markov process to get a trajectory $p((x_t)_{t=1}^1)$. We can train q_θ on these trajectories by optimizing a negative ELBO

$$\begin{aligned} -\log q_\theta(x_0) &\le -E_{p((x_t)_{t=1}^1|x_0)} \log \frac{q_\theta((x_t)_{t=0}^1)}{p((x_t)_{t=1}^1|x_0)} \\ &= -E_{p((x_t)_{t=1}^1|x_0)} \log \frac{q_\theta((x_t)_{t=0}^1|x_1)}{p((x_t)_{t=1}^1|x_0,x_1)} + \mathrm{KL}(p(x_1|x_0)|q(x_1)). \end{aligned} \tag{1}$$

To make the second term of Eqn. 1 small we need $p(x_1|x_0) \approx q(x_1)$. To do so, we pick an increasing "time dialation" function $\tau : [0,1] \to [0,\infty)$ and simulate x_t so that it has had the Markov process applied to it for time time τ_t (see B.1 for explanation). Picking τ_1 very large, the second term of the ELBO can be made arbitrarily small, so we leave it out of the presentation below.

Matching forward and backward flow q_θ is usually parameterized to take x_t, t and predict the x_0 that generated x_t, that is, approximate $p(x_0 \mid x_t, t)$; we represent this prediction $\tilde{x}_0 = q_\theta(x_0|x_t,t)$ as a vector of probabilities over the B tokens $\sum_b \tilde{x}_{0,b} = 1$. Some rearrangement then allows one to rewrite the first term of Eqn. 1 as an expectation of a term L that can be interpreted as the divergence between the "infinitesimal flow" forward p and backward q_θ at x_t:

$$E_{t\sim \mathrm{Unif}(0,1)} E_{p(x_t|x_0)} L(x_t, t, x_0, \tilde{x}_0).$$

We describe the ELBO algorithms for discrete and Gaussian diffusion, along with the challenges of comparing their losses and hyperparameters in Appendix B.3.

3 Unifying diffusion models

We derive connections between discrete, Gaussian, and simplicial diffusion using a population genetics framework, based on the Wright-Fisher mathematical model of genetic drift [Wright, 1931]. Using this framework we show that Gaussian diffusion can be derived directly from discrete diffusion, enabling previously impossible comparisons between the models. Then in Appendix B.4 and G.5 we derive a unifying connection to simplicial diffusion.

We represent each dimension of a sequence as a population with ζ copies of each letter to get a *sequence of sequences*.

$$\text{ex. for } \zeta = 4, x_0 = \texttt{A|C|C|T} \text{ is represented as } \texttt{AAAA|CCCC|CCCC|TTTT}.$$

Then each letter in each sequence is evolved according to the mutation matrix $\mathcal{L}$ (where $\mathcal{L}_{b_1\to b_2}$ describes the rate at which b_1 mutates to b_2). When $\zeta = 1$ we get discrete diffusion. Next we show that as $\zeta \to \infty$ we get Gaussian diffusion. Below we discuss the one-dimensional case $D = 1$, which can naturally be extended to a multi-dimensional diffusion model.

Algorithm 1 ELBO for ζ discrete diffusion

Sample $t \sim \text{Unif}(0, 1)$
Sample noisy x_t:
Sample $\vec{x}_t \sim \text{Multinomial}(\zeta, x_0^T e^{\tau_t \mathcal{L}})/\zeta$
Predict de-noised X:
Predict $\tilde{x}_0 = q_\theta(x_0 \mid \vec{x}_t, t)$
Compute loss:
$p = \vec{x}_0^T e^{\tau_t \mathcal{L}}$; $q = \tilde{x}_0^T e^{\tau_t \mathcal{L}}$
$L = \sum_{b_1 \neq b_2} \mathcal{L}_{b_2\to b_1} \dot{\tau}_t \zeta \vec{x}_{t,b_1} \mathbb{D}\left(\frac{p_{b_2}}{p_{b_1}} \middle\| \frac{q_{b_2}}{q_{b_1}}\right)$

Representing x_t on the simplex Even though we will ultimately arrive at a Gaussian limit in Euclidean space, we first represent x_t on the simplex. Above x_t was one of B tokens; now it's one of B^ζ sequences of B tokens $x_t = x_t^{(1)} \cdots x_t^{(\zeta)}$. It can be generated as in the $\zeta = 1$ case by sampling each $x_t^{(z)} \sim \text{Categorical}(\vec{x}_0^T e^{\tau_t \mathcal{L}})$. In App. G.1 we note however that the loss and $p(x_0 \mid x_t, t)$ – the target for $q_\theta(x_0 \mid x_t, t)$ – *do not depend on the order* of the letters of x_t. Therefore we can represent x_t as a vector of counts of each letter, or normalize by ζ to get $\sum_b x_{t,b}/\zeta = 1$. In App. G.1 we derive the loss, giving us Alg. 1 – differences to discrete diffusion in Alg. 2 are highlighted in blue.

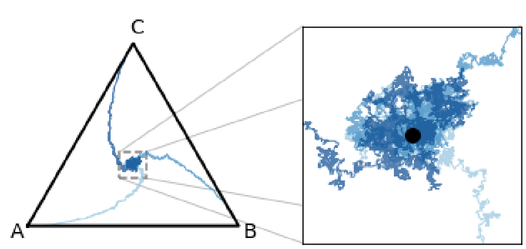


Figure 2: **Discrete diffusion with a large population converges to Gaussian diffusion.** With $\zeta = 1000$, we show example trajectories $(\vec{x}_t)_t$ from A, B, and C that converge to approximate Gaussians near π.

Gaussian limit As $\zeta \to \infty$, trajectories converge quickly to the stationary distribution of $\mathcal{L}$, π, and behave like Gaussians near π because of the central limit theorem (Fig. 2). As $\zeta \to \infty$ we zoom further into the neighbourhood of π where the diffusion occurs, moving from *diffusion on the simplex* to *diffusion in Euclidean space*. Interestingly, we see that in the multi-dimensional case, the relevant Gaussian diffusion can occur in a subspace determined by the spectrum of $\mathcal{L}$.

Theorem 3.1. *(Formal statement and proof in App. G.2) Call λ_1 the largest negative eigenvalue of $\mathcal{L}$ and P_1 the projection onto the corresponding left eigenspace. Without loss of generality, assume $\lambda_1 = 1$. For each ζ pick time dilation $\tau_t^\zeta = \frac{1}{2}\log\left(\zeta e^{-2\tau_t} - \zeta + 1\right)$ and rescale $\vec{x}_t^\zeta = \sqrt{\zeta - (\zeta-1)e^{2\tau_t}}(\vec{x}_t - \pi)/\sqrt{\pi}$. Define the embedding into $\mathbb{R}^{\text{rank}(P_i)}$, $Q_i = \mathfrak{j}_i(\tilde{Q}_i\tilde{Q}_i^T)^{-1/2}\tilde{Q}_i$ where $\tilde{Q}_i = \text{diag}(\pi)^{-1/2}P_i\text{diag}(\pi)^{1/2}$ and $\mathfrak{j}_i$ is any isometry from $\text{Im}(\tilde{Q}_i) \to \mathbb{R}^{\text{rank}(P_i)}$.*

When $\zeta = 1$ we get discrete diffusion: *$\tau_t^\zeta = \tau_t$ and $\vec{x}_t^\zeta$ is only linearly transformed $(\vec{x}_t - \pi)/\sqrt{\pi}$.*

When $\zeta \to \infty$, we get Gaussian diffusion in the first eigenspace. *Only the first eigenspace has signal (in the limit, the component of x_t^ζ in $\text{Ker}Q_1$ is independent of x_0). The paths $(Q_1\vec{x}_t^\zeta)_{t\in(0,1)}$ converge in distribution to paths from Gaussian diffusion with time dilation τ_t and embedding $\text{emb}(x_0) = Q_1(\vec{x}_0/\sqrt{\pi})$. The ELBO in Alg. 1 converges to the ELBO for Gaussian diffusion in Alg. 3.*

3.1 Comparing diffusion models

Loss comparison Gaussian and discrete diffusion are thought to have incomparable likelihoods due to a singularity in the Gaussian diffusion loss (see B.3.1). However, Thm. 3.1 suggests that there is no difference in training a discrete diffusion model with $\zeta = 10^{100}$ and training Gaussian diffusion with Alg. 3 on a computer, suggesting their ELBOs are comparable. Our unification result offers an explanation for why the singularity exists (Appendix C.1), and suggests a practical solution to enable comparison, which we name the "hollow predictor". We weight the output of the neural network by the evidence for each x_0, $q_\theta(x_0 \mid x_t, t) \propto p(x_t \mid x_0, t) q_\theta(x_0)$ where $p(x_t \mid x_0, t)$ automatically handles deciding when x_0 is obvious from x_t's location on the simplex (explanation in C.1). Thus, when this parameterization is used, Gaussian and discrete diffusion likelihoods can in fact be directly compared. In App. G.4 we prove that applying the hollow parametrization removes the singularity at 0 of the Gaussian ELBO.

Hyperparameter comparison Discrete and Gaussian diffusion models are specified by hyperparameters $\mathcal{L}$ and emb with vastly different interpretations (see B.3.1). Thm. 3.1 reveals that embeddings correspond to a corrected fist eigenspace of the mutation matrix, establishing a formal connection between $\mathcal{L}$ and emb (see Fig. 5). The practical implications of this connection are that (1) one can sanity-check their designed $\mathcal{L}$ by checking its induced embeddings, and (2) discrete diffusion offers a richer design space, as one can specify all the interacting eigenspaces of $\mathcal{L}$ rather than just the dominant one, emb.

4 Practical unified diffusion models

We show through a particular parameter choice, the"sufficient-statistic parameterization" (SSP), one can train a single neural network that can perform diffusion on any domain at test time (Fig. 3). Further, the SSP explains the root of the noted "time-invariance" of masking diffusion and extends this property to every diffusion model (C.3).

The goal of a diffusion model is to predict[1] $q_\theta(x_0^d \mid x_t^{-d}, t) \approx p(x_0^d \mid x_t^{-d}, t)$ for all d, t. To do so, one must integrate over the unseen x_0^{-d} weighted by their likelihood of producing the data x_t^{-d}:

$$p(x_0^d \mid x_t^{-d}, t) = \int p(x_0^d \mid x_0^{-d}) dp(x_0^{-d} \mid x_t^{-d}, t).$$

This means that the only way each $x_t^{d'}$ impacts our prediction is through the evidence it gives us about $x_0^{d'}$. We can summarize this "evidence" in the normalized vector[2] $\vec{\phi}(x_t^{d'}, t)_b \propto p(x_t^{d'} \mid t, x_0^{d'} = b)$ (Supp. Fig. 6). A bit of algebra shows that these $\vec{\phi}$'s are sufficient statistics – they contain all relevant information about the diffusion process and t, leaving a regression task that invariant to both.

Proposition 4.1. *(Proof in App. G.3) There is a function F^d, **depending on** $p(x_0)$ **and not on the diffusion process or** t, such that*

$$p(x_0^d \mid x_t^{-d}, t) = F^d(\vec{\phi}(\vec{x}_t^1, t), \ldots, \vec{\phi}(\vec{x}_t^D, t)).$$

Therefore we can parametrize our neural network $q_\theta(x_0^d \mid x_t^{-d}, t) = F_\theta^d(\vec{\phi}(\vec{x}_t^1, t), \ldots, \vec{\phi}(\vec{x}_t^D, t))$ for a neural network F_θ^d that learns the "universal" F^d.

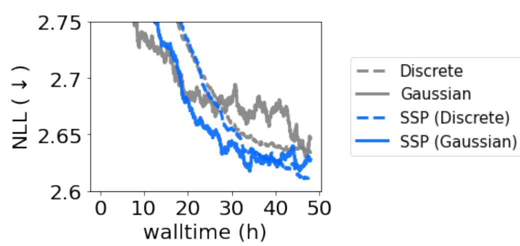


Figure 3: **The sufficient statistic parametrization enables training a single model that can do discrete or Gaussian diffusion.** We used an ESM2 architecture on Uniref50 for 48 hours on a single A100.

Performance of a unified model Lastly, as an initial experiment, we train discrete and Gaussian diffusion models on proteins and compare to a model using the SSP which alternated between discrete and Gaussian training steps. We find that even controlling for compute, the single SSP model is competitive with the single-domain models (Fig. 3).

[1]For the non-hollow parameterization, swap x_t^{-d} with x_t.

[2]Note this only works with diffusion models of discrete data where $\vec{\phi}$ is finite dimensional.

References

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, September 2023.

Alan Nawzad Amin, Nate Gruver, and Andrew Gordon Wilson. Why masking diffusion works: Condition on the jump schedule for improved discrete diffusion. In *Frontiers in Probabilistic Inference: Learning meets Sampling*, April 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Adv. Neural Inf. Process. Syst.*, 34:17981–17993, 2021.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. *arXiv [cs.LG]*, May 2023.

Ethan Baron, Alan Nawzad Amin, Ruben Weitzman, Debora Susan Marks, and Andrew Gordon Wilson. A diffusion model to shrink proteins while maintaining their function. In *The Exploration in AI Today Workshop at ICML 2025*, June 2025.

Richard F Bass. *Stochastic Processes*. Cambridge University Press, October 2011.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, October 2022.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv.org*, 2022.

Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterisation and Convergence*. Probability & Mathematical Statistics S. John Wiley & Sons, Nashville, TN, May 1986.

F Gotze. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2):724–739, 1991.

R C Griffiths. Asymptotic line-of-descent distributions. *J. Math. Biol.*, 21(1):67–75, December 1984.

Nate Gruver, Samuel Don Stanton, Nathan C Frey, Tim G J Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Paul A Jenkins and Dario Spanò. Exact simulation of the Wright–Fisher diffusion. *Ann. Appl. Probab.*, 27(3): 1478–1509, June 2017.

M Kimura. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. U. S. A.*, 41(3):144–150, March 1955.

Bocheng Li, Zhujin Gao, and Linli Xu. Unifying continuous and discrete text diffusion with non-simultaneous diffusion processes. *arXiv [cs.CL]*, May 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *41 st International Conference on Machine Learning*, October 2023.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems 35*. Cold Spring Harbor Laboratory, July 2022.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv [cs.LG]*, June 2024.

Pierre H Richemond, Sander Dieleman, and Arnaud Doucet. Categorical SDEs with simplex diffusion. *arXiv [cs.LG]*, October 2022.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv [cs.CL]*, June 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv [cs.LG]*, June 2025.

Anirban Sarkar, Ziqi Tang, Chris Zhao, and Peter K Koo. Designing DNA with tunable regulatory activity using discrete diffusion. *bioRxiv*, page 2024.05.23.595630, May 2024.

Alexander Shabalin, Viacheslav Meshchaninov, and Dmitry Vetrov. Smoothie: Smoothing diffusion on token embeddings for text generation. *arXiv [cs.CL]*, May 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv [cs.LG]*, June 2024.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. *arXiv [q-bio.BM]*, February 2024.

Charles Stone. Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois J. Math.*, 7(4):638–660, December 1963.

Ushio Sumita, Jun-Ya Gotoh, and Hui Jin. NUMERICAL EXPLORATION OF DYNAMIC BEHAVIOR OF ORNSTEIN-UHLENBECK PROCESSES VIA EHRENFEST PROCESS APPROXIMATION(advanced planning and scheduling for supply chain management). *J. Oper. Res. Soc. Japan*, 49(3):256–278, 2006.

S Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26(2):119–164, October 1984.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. *arXiv [cs.LG]*, October 2024.

Ludwig Winkler, Lorenz Richter, and Manfred Opper. Bridging discrete and continuous state spaces: Exploring the ehrenfest process in time-continuous diffusion models. *arXiv [stat.ML]*, May 2024.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, March 1931. Received January 20, 1930.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv [cs.LG]*, September 2024.

A Related work

To provide context for our work, we give an overview of other theories of unification and diffusion model parameterizations below.

Theories unifying discrete and continuous diffusion There is a long history of deriving continuous limits of discrete processes, the "forward" processes of diffusion models. Groundbreaking work by Stone [1963] derived Gaussian diffusion as a limit to biased one-dimensional random walks. In one of the most celebrated results in mathematical genetics, Kimura [1955] also derived a continuous limit of the Wright-Fisher process with non-zero reproduction. We (1) apply these results to understand and improve diffusion models, (2) also show convergence of the ELBO of diffusion models, and, to our knowledge, (3) derive a new result – the multi-dimensional Gaussian-diffusion limit of Wright-Fisher with zero reproductions – demonstrating previously un-characterized behaviour dependent on the eigenspace of the mutation operator. Results (2) and (3) are what allow us to compare likelihoods and hyperparameters.

Looking at generative diffusion models, Winkler et al. [2024] used the result from [Stone, 1963] (through a citation from Sumita et al. [2006]) to connect the special case of one-dimensional, unbiased discrete diffusion to one-dimensional Gaussian diffusion. They use this observation to heuristically argue, or conjecture, the convergence of the backwards processes as well. Sahoo et al. [2025] suggested that by taking Gaussian diffusion and applying argmax, one recovers discrete diffusion[3]. They used this insight to answer the loss comparison problem by proving that the ELBO of discrete diffusion is always superior to that of continuous diffusion. Unfortunately, this is based on

[3] Interestingly, Stone [1963] also wrote discrete diffusion as the function of an underlying Gaussian diffusion. However the function from Stone [1963] was a path-dependent time-dilation rather than argmax.

a mathematical error (details in App. E): by applying argmax to Gaussian diffusion one does not get a Markov process, a property which was crucial to their proof of the loss comparison question. Li et al. [2025] looked at Gaussian diffusion with a generalized noising strategy; they noted a special case resembled masking diffusion – each token was either fully noised or un-noised. However the training procedure and ELBO of this special case are distinct from standard masking diffusion [Shi et al., 2024].

Parameterizations of discrete diffusion models In diffusion, one uses a neural network to predict the identity of each dimension of the "un-noised" sequence x_0^d given the noised sequence and the time t, $q_\theta(x_0^d \mid x_t, t)$. A number of works suggest superficially distinct, but ultimately equivalent parameterizations [Campbell et al., 2022, Lou et al., 2023].

For discrete diffusion models, Austin et al. [2021] suggested multiplying the output of the neural network by $p(x_t^d \mid x_0^d)$ to "automatically" incorporate the information about the noised token about that particular location. Amin et al. [2025] interpreted this as using a "hollow" predictor as $q_\theta(x_0^d \mid x_t, t) \propto p(x_t^d \mid x_0^d) q_\theta(x_0^d \mid x_t^{-d}, t)$, with the neural network playing the role of $q_\theta(x_0^d \mid x_t^{-d}, t)$. While relegated to the appendix of these works, we show that this choice is crucial for the loss comparison problem when its application is extended to Gaussian and simplicial diffusion.

Zheng et al. [2024], Ou et al. [2024], and Sahoo et al. [2024] noted that for masking diffusion, it is not necessary to pass t to $q_\theta(x_0^d \mid x_t, t)$ – it is "time-invariant". Zheng et al. [2024] suggests this makes masking models a fundamentally different object than other diffusion models: "we reveal that both training and sampling of [masked models] are theoretically free from the time variable, arguably the key signature of diffusion models, and are instead equivalent to masked models." Our sufficient-statistic parameterization shows on the contrary that every diffusion model can be made time-invariant by a choice of parameterization, with masking as a special case.

B Methods

B.1 The time dilation function

To make the second term of Eqn. 1 small we need $p(x_1|x_0) \approx q(x_1)$ which in particular means that $p(x_1|x_0)$ should not strongly depend on x_0. Conveniently, applying a Markov process to x_0 usually leads to $p(x_t|x_0)$ converging to a stationary distribution $p(x_\infty)$ as $t \to \infty$, a good choice for $q(x_1)$. However our t is on the interval $[0, 1]$, not $[0, \infty)$, so we compress $[0, \infty)$ into $[0, 1]$: we pick an increasing "time dialation" function $\tau : [0, 1] \to [0, \infty)$ and simulate x_t so that it has had the Markov process applied to it for time time τ_t. In particular, if τ_1 is very large, $p(x_1|x_0) \approx p(x_\infty) = q(x_1)$.

τ_t is a more convenient parametrization for our presentation than equivalent functions $\beta_t = \dot{\tau}_t, \alpha_t = \exp(-\tau_t)$ in other works [Shi et al., 2024].

B.2 Moving to multiple dimensions

To consider sequences of discrete objects $x_0 = x_0^1 \cdots x_0^D$, we simply apply the Markov process to each position x_0^d independently. Therefore "**Sample noisy** x_t" remains the same, just repeated for every d. As well, the "infinitesimal flow" for each position ends up being independent: the "**Compute ELBO**" step also remains the same, just repeated for every d and then summed across all d. To compute the ELBO therefore, in the "**Predict de-noised** x_0" step we will predict $\tilde{x}_{0,\theta}^d = q_\theta(x_0^d|x_t, t)$ for each d.

B.3 Discrete and Gaussian diffusion

For discrete diffusion, the Markov process is stochastic mutation defined with a rate matrix $\mathcal{L}$ (where $\mathcal{L}_{b_1 \to b_2}$ describes the rate at which b_1 mutates to b_2); the form for L was derived in Campbell et al. [2022]. This gives Alg. 2, where $\vec{x}_0$ is the indicator vector for the token x_0, $\mathbb{D}(\lambda_1||\lambda_2) = \lambda_1 \log \frac{\lambda_1}{\lambda_2} - \lambda_1 + \lambda_2$ is the KL divergence between two Poisson distributions, and $\dot{\tau}_t$ is the derivative of τ_t. For Gaussian diffusion, the Markov process is Brownian motion on embedded vectors $\text{emb}(x_0) \in \mathbb{R}^r$; the form for L was derived in Ho et al. [2020]. This gives Alg. 3.

In summary, getting a stochastic estimate of the ELBO has 3 steps: (1) Sample noisy x_t by simulating the Markov process for time τ_t, (2) Predict de-noised x_0 with $\tilde{x}_{0,\theta}(x_t, t)$, and (3) Compute the particular form of L. The difference between diffusion models lies in the first and third steps.

Algorithm 2 ELBO for discrete diffusion

Sample $t \sim \mathrm{Unif}(0, 1)$
Sample noisy x_t:
Sample $x_t \sim \mathrm{Categorical}(\vec{x}_0^T e^{\tau_t \mathcal{L}})$
Predict de-noised x_0:
Predict $\tilde{x}_0 = q_\theta(x_0|x_t, t)$
Compute ELBO:
$p = \vec{x}_0^T e^{\tau_t \mathcal{L}}, q = \tilde{x}_0^T e^{\tau_t \mathcal{L}}$
$L = \sum_{b \neq x_t} \mathcal{L}_{b \to x_t} \dot{\tau}_t \mathbb{D}\left(\frac{p_b}{p_{x_t}} \middle\| \frac{q_b}{q_{x_t}}\right)$

Algorithm 3 ELBO for Gaussian diffusion

Sample $t \sim \mathrm{Unif}(0, 1)$
Sample noisy x_t:
Set $x_t = e^{-\tau_t}\mathrm{emb}(x_0) + \sqrt{1 - e^{-2\tau_t}} N(0, I)$
Predict de-noised x_0:
Predict $\tilde{x}_0 = q_\theta(x_0|x_t, t)$
Compute ELBO:
$L = \frac{\dot{\tau}_t e^{-2\tau_t}}{(1-e^{-2\tau_t})^2} \|\mathrm{emb}(x_0) - \mathrm{emb}(\tilde{x}_0)\|^2$

B.3.1 Theoretical challenges in discrete and Gaussian diffusion comparison:

Likelihood comparison We would like to compare the likelihoods of discrete and Gaussian diffusion, but these are sometimes infinity. At initialization, $\|\mathrm{emb}(x_0) - \mathrm{emb}(\tilde{x}_0)\|^2$ is roughly a constant, and for the classical choice $\tau_t = -\frac{1}{2}\log(1 - t)$, the square error in Alg. 3 is weighted by $1/2t^2$, so the loss is $\sim \int_0^1 t^{-2} dt = \infty$. To avoid the singularity at small t, one chooses a minimum $t_{\min}$[4]. Formally this is equivalent to estimating an ELBO for $\log p(x_{t_{\min}})$ instead of $\log p(x_0)$. However, $x_{t_{\min}}$ is not a discrete object so $p(x_{t_{\min}})$ is a continuous density, fundamentally a different object than the probability of a discrete object $p(x_0)$. Nevertheless, the values of the ELBO for $\log p(x_{t_{\min}})$ is often close to ELBOs from discrete diffusion models, suggesting they may be comparable.

Hyperparameter comparison Discrete and Gaussian diffusion models are specified by hyperparameters $\mathcal{L}$ and emb with vastly different interpretations. To specify a discrete diffusion model, one must specify a matrix whose entry $\mathcal{L}_{b_1 \to b_2}$ describes the rate at which b_1 mutates to b_2. For proteins for example, this is often specified using the BLOSUM amino acid similarity matrix [Alamdari et al., 2023]. To specify a Gaussian diffusion model, one must specify a embedding function emb that takes the alphabet into Euclidean space $\mathbb{R}^r$ for some r (we write $\mathrm{emb}(\tilde{x}_0)$ as shorthand for $\sum_b \tilde{x}_{0,b}\mathrm{emb}(b)$). This can use pre-trained embeddings [Dieleman et al., 2022] or a variety of other strategies [Shabalin et al., 2025].

B.4 Unifying simplicial diffusion

We now allow our population of ζ to reproduce. At rate ζ we generate ζ "children" which each randomly and uniformly pick a parent; we also allow individuals to continue mutating according to mutation matrix $\mathcal{L}$ so that mutations may be introduced between generations (Fig. 1a). We now ask what happens when $\zeta \to \infty$ by referring to the mathematical genetics literature. One biologically reasonable assumption these works make is a parent-independent mutation rate matrix, that is, $\mathcal{L} = \psi \times (\mathbb{1}\vec{\pi}^T - I)$ for stationary distribution $\vec{\pi}$ and mutation rate $\psi > 0$ (see ex. Tavaré [1984]). Since this does not restrict the design space of simplicial diffusion, which is specified exactly by an intensity parameter ψ and stationary distribution $\vec{\pi}$, we make the same assumption.

The limit of the forward process Kimura [1955] was the first to derive the $\zeta \to \infty$ limit of the stochastic process. Unlike the mutation-only case which zooms in on $\vec{\pi}$, this limiting distribution has paths that travel throughout the simplex (see Fig. 1b. Indeed this limit, often itself called "Wright-Fisher diffusion" is exactly the "Jacobi process" used in simplicial diffusion [Avdeyev et al., 2023]. In higher dimensions, Ethier and Kurtz [1986, Chapter 10] also gives the same result as the construction from Avdeyev et al. [2023].

[4]Most discrete diffusion models also have a singularity at $t \to 0^+$, requiring one to specify a $t_{\min}$ [Campbell et al., 2022, Lou et al., 2023]. This is not the case for "schedule-conditioned" models, including masking, partially explaining its popularity [Amin et al., 2025, Shi et al., 2024].

The limit of the ELBO We add to these results by also deriving the limit of the discrete diffusion ELBO. Remarkably, we get the "score-matching" objective of Avdeyev et al. [2023] scaled by $\dot{\tau}_t/2$. This justifies its use as an ELBO while Avdeyev et al. [2023] only recognized it as a stable training objective.

Theorem B.1. *(Proof in App. G.5) As $\zeta \to \infty$, the discrete diffusion objective in Alg. 2 converges to the teal quantity from Alg. 4.*

Algorithm 4 ELBO for simplicial diffusion. Our changes to Avdeyev et al. [2023] are coloured.

Sample $t \sim \mathrm{Unif}(0, 1)$
Sample noisy x_t:
Sample $m \sim A(\psi, \tau_t)$ with Alg. 5; if $\tau_t < 0.05$, use Alg. 6
Sample $\vec{x}_t \sim \mathrm{Dirichlet}(\psi\vec{\pi} + m\vec{x}_0)$.
Predict de-noised x_0:
Predict $\tilde{x}_0 \propto q_\theta(x_0 \mid x_t, t)$
Compute ELBO:
Compute $\vec{s}(\vec{x}_t \mid x_0, t) = \nabla_{x_t} \log p(x_t|x_0, t)$ with Eqn. 2; if $\tau_t < 0.05$, use Eqn. 3
$L = \frac{\dot{\tau}_t}{2} \|\vec{s}(\vec{x}_t \mid x_0, t) - \vec{s}(\tilde{x}_t \mid x_0, t)\|^2_{\mathrm{diag}(\vec{x}_t) - \vec{x}_t\vec{x}_t^T}$

Sampling noisy x_t Avdeyev et al. [2023] and Richemond et al. [2022] suggested sampling x_t by costly and approximate simulation from a stochastic differential equation. Instead, the suggestively titled paper "Exact simulation of the Wright-Fisher diffusion" [Jenkins and Spanò, 2017] gives a simple formula for the marginals x_t (blue in Alg. 4). The algorithm samples $\vec{x}_t$ from a Dirichlet that is centred at the stationary mutation distribution $\vec{\pi}$ when $m = 0$ and becomes more concentrated around the signal x_0 when m is larger. m itself is an integer sampled from a distribution $A(\psi, \tau_t)$ that represents, going back in time τ_t, how many ancestors the population descend from – it is small when τ_t is large, when everyone descended from a handful of individuals from far back in time. Indeed Stark et al. [2024] suggested a Dirichlet distribution as a natural noising distribution for x_t for flow matching – we see this intuition applies without having to do away with diffusion altogether.

Low t behaviour Both the simulation of $A(\psi, \tau_t)$ and the calculation of the gradients $\nabla_{x_t} \log p(x_t \mid x_0)$ involves an infinite series [Tavaré, 1984]. Luckily the terms converge extremely fast – at square exponential rate. This is not true however at low t, leading to the well known instability of simplicial diffusion [Avdeyev et al., 2023, Richemond et al., 2022]. This instability is also well known in the mathematical genetics literature, with Griffiths [1984] emphatically stating that using the infinite series at low t "produces nonsense from a computer."

The solution at low t is to replace the series approximation, which gets worse with lower t, with a central limit approximation for $A(\psi, \tau_t)$ [Griffiths, 1984, Jenkins and Spanò, 2017] that improves with lower t (purple in Alg. 4); this is analogous to how reflected diffusion models were made stable despite their own infinite series expansion with the same problem Luo et al. [2022]. We picked the $\tau_t < 0.05$ threshold as recommended by Jenkins and Spanò [2017]. In App. F.1 we describe how to use this approximation to also stabilize the loss computation.

C Discussion of Theoretical Results

C.1 Enabling loss comparison

Fig. 2 suggests why the limiting Gaussian ELBO is infinite: paths from $\vec{x}_t$ have two phases, a nearly deterministic phase where no information about x_0 has been lost (Fig 2 left), and a random phase (Fig 2 right). Diffusion models reversing these paths should therefore go through a random phase, until $p(x_0 \mid x_t, t)$ becomes obvious, and then trace a deterministic path back to x_0. However, at initialization, x_0 is "never obvious" to the neural network $q_\theta(x_0 \mid x_t, t)$, leading to mismatches to the deterministic paths (Fig. 4 "Random"). As ζ gets larger, the paths get more deterministic

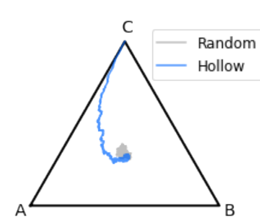


Figure 4: **The hollow parameterization leads to realistic reverse path samples.** $\zeta = 300$.

and our choice of τ_t^ζ "smooshes" the deterministic phase towards $t = 0$, **causing the singularity in the limit**.

The practical solution is therefore simple – weight the output of the neural network by the evidence for each x_0, $q_\theta(x_0 \mid x_t, t) \propto p(x_t \mid x_0, t) q_\theta(x_0)$ where $p(x_t \mid x_0, t)$ "automatically handles" deciding when x_0 is obvious (Fig. 4 "Hollow").In App. G.4 we prove that applying the hollow parametrization removes the singularity at 0 of the Gaussian ELBO.

This was suggested by Austin et al. [2021] to improve discrete diffusion models, but here we show that important for building Gaussian diffusion models with formally comparable likelihoods as well[5]. Amin et al. [2025] showed that in higher dimensions this becomes equivalent to using the **"hollow" predictor**[6].

C.2 Embedding hyperparameter comparison

Hyperparameter comparison Thm. 3.1 gives us a formula for an embedding function emb determined by the slowest-decaying directions in $\mathcal{L}$. Remarkably, this connection accommodates Gaussian diffusion in different dimensions $\mathbb{R}^r$: r is simply determined by the dimension of the dominant eigenspace of $\mathcal{L}$. In Fig. 5 we show the top eigenspace of the BLOSUM matrix, often used to build stochastic processes for amino acids, seeing that it clusters similar amino acids together.

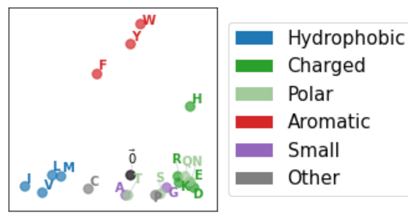


Figure 5: **Euclidean embeddings of amino acids from BLOSUM $\mathcal{L}$.** With the BLOSUM discrete diffusion process inspired by [Alamdari et al., 2023], we extract $\mathrm{emb}(x_0)$ from Thm. 3.1 for all amino acids.

C.3 Time-invariant diffusion models

Masking diffusion is celebrated for its "time-invariance" [Zheng et al., 2024, Sahoo et al., 2024]: its optimal $q_\theta(x_0^d \mid x_t^{-d}, t)$ does not depend on time. This theoretically connects it with masked language models and practically means that one does not need to engineer neural networks of two variables, both x_t and t. Our SSP allows us to make any diffusion model time-invariant.

Time-invariance is a function of parameterization: Masking is time-invariant due to a choice of parametrization. To see this, imagine applying a time-dependent rotation to each x_t^d; we are essentially performing the same diffusion but now must also pass t to q_θ so it can "undo" the transformation. The $\vec{\phi}$ can be thought of as automatically transforming x_t so F^d is independent of time in any diffusion model.

Masking uses SSP: Indeed the SSP of masking diffusion, $\vec{\phi}(x_t^d, t) = \delta_{x_t}$ if $x_t \neq \mathrm{mask}$ and $\vec{\phi}(x_t^d, t) = [\frac{1}{B}, \ldots, \frac{1}{B}]$ otherwise, is exactly the canonical parametrization. Thus the time-invariance of masking isn't special – rather masking's most convenient parametrization happens to be the SSP.

[5] Note this hollow parametrization is specific to our setting of Gaussian diffusion for *discrete data* where there are only finitely many possible x_0.

[6] Note this does not require a change of architecture – $q_\theta(x_0^d | x_t^{-d}, t)$ can be a function of all of x_t but must learn to disregard x_t^d.

D Supplementary Figures

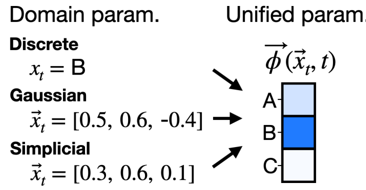


Figure 6: **The sufficient statistic parameterization represents $\vec{x}_t$ from all diffusion models in the same space.**

E Mathematical error in Sahoo et al. (2025)

In Theorem 3.1, Sahoo et al. [2025] shows that the ELBO of a discrete diffusion model is always tighter than that of a Gaussian diffusion model. In its proof, with w_t from Gaussian diffusion, $z_t = \text{argmax}(w_t)$, and $x = z_0 = w_0$, they state "Since the transition $z_t \to z_s$ is Markov, we get: $q(z_s \mid w_t, z_t, x) = q(z_s \mid z_t, x)$". Putting aside the correctness of this statement, it is clear that the proof as stated requires the Markov property of $(z_t)_t$.

The way the Markov property is shown is as follows. They first define a discrete diffusion model, let's call this $(\tilde{z}_t)_t$, such that $\tilde{z}_0$ comes from the data distribution and $\tilde{z}$ evolves with respect to a uniform forward process with rate parameter $\beta(t)$ chosen such that the marginals match $p(z_t|z_0) = p(\tilde{z}_t|\tilde{z}_0)$. In Eqn. 29 they compute $\frac{d}{dt}p(z_t|z_0)$ and in Eqn. 32 they compute $\frac{d}{dt}p(\tilde{z}_t|\tilde{z}_0)$ for all starting points and show they are identical. After noting the equivalence of equations 29 and 32, they state "This pmf and the ODE are the unique signatures of a Uniform-state discrete diffusion process (Lou et al., 2023; Schiff et al., 2025)." and from this conclude that the path distributions of $(\tilde{z}_t)_t$ and $(z_t)_t$ are equivalent, and in particular, that $(z_t)_t$ is Markov[7].

However, despite a similar result for Markov chains (two Markov processes with identical semi-groups are equivalent), $p(z_t|z_0) = p(\tilde{z}_t|\tilde{z}_0)$ and $\frac{d}{dt}p(z_t|z_0) = \frac{d}{dt}p(\tilde{z}_t|\tilde{z}_0)$ for all starting points is not enough to conclude the identity of the path distributions $p((z_t)_t|z_0) = p((\tilde{z}_t)_t|\tilde{z}_0)$. First note that $\frac{d}{dt}p(z_t|z_0) = \frac{d}{dt}p(\tilde{z}_t|\tilde{z}_0)$ is not an independent condition: it follows from $p(z_t|z_0) = p(\tilde{z}_t|\tilde{z}_0)$. Next consider this counter example:

- $\tilde{z}_0 = 1$ and $(\tilde{z}_t)_t$ evolves by switching sign with rate 1. Therefore $p(\tilde{z}_t = 0) = 1 - \frac{1}{2}e^{-2t}$.
- $z_0 = 1$ and $(z_t)_t$ has a 50% chance to stay at 0 forever and a 50% chance to swap sign at time $-\frac{1}{2}\log U$ for a $U \sim \text{Uniform}$ and never again. Therefore $p(\tilde{z}_t = 1) = \frac{1}{2}(1 + p(-\frac{1}{2}\log \text{Uniform} > t)) = 1 - \frac{1}{2}e^{-2t}$.
- When $z_0 = -1$ or $\tilde{z}_0 = -1$, then swap signs.

We have $p(z_t|z_0) = p(\tilde{z}_t|z_0)$ for all z_0 and therefore $\frac{d}{dt}p(z_t|z_0) = \frac{d}{dt}p(\tilde{z}_t|z_0)$ but clearly $p((z_t)_t) \neq p((\tilde{z}_t)_t)$.

Simple computer simulations indeed show that $p((\text{argmax}(w_t))_t)$ and $p((\tilde{z}_t)_t)$ are different. We show this in Fig. 7. Indeed a statistical test applied to these simulations shows $p((\text{argmax}(w_t))_t) \neq p((\tilde{z}_t)_t)$: a Mann-Whitney test shows that the paths of the argmax of Gaussian diffusion have more transitions that those of discrete diffusion with $p < 10^{-300}$.

[7] This interpretation of the text was confirmed in personal communication with the first author of Sahoo et al. [2025]

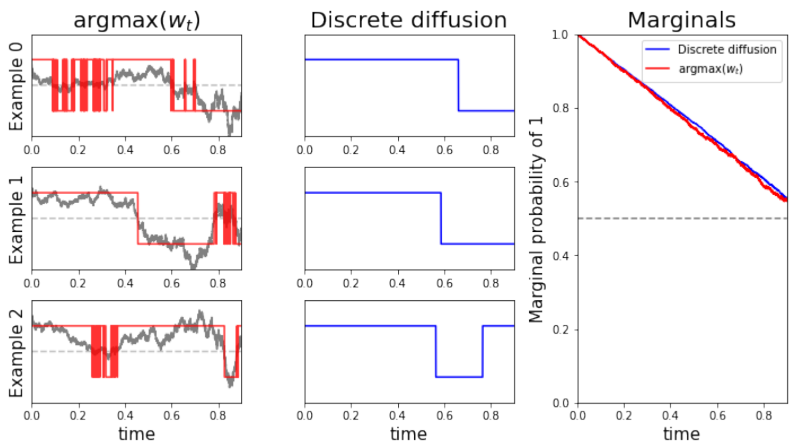


Figure 7: **The argmax of Gaussian diffusion appears different from discrete diffusion in simulation, despite having the same marginals.** We compare example paths of $p((\text{argmax}(w_t))_t)$ (left, red; we show Gaussian diffusion w_t in grey), $p((\tilde{z}_t)_t)$ for uniform discrete diffusion (centre, blue), and their empirical marginals over 10'000 simulations (right); we simulate using a grid size of 0.0001. Note the two processes have the same marginals but their paths appear different; in particular, whenever w_t is near 0, $(\text{argmax}(w_t))_t$ undergoes a very large number of transitions in a small time[8].

F Wright-Fisher sampling and score calculations

Note, just like App. G.1, we can deal with $\vec{x}_t$ rather than the actual sequences x_t. We now discuss how to sample and calculate the functions $\vec{s}(\vec{v} \mid x_0)$.

Sample noisy x_t We've discussed the algorithm from Jenkins and Spanò [2017] in the main text. We now present their algorithm for sampling from $A(\psi, \tau_t)$.

Algorithm 5 Exact sampling from ancestral process $A(\psi, \tau_t)$

Define coefficients: $c^{\psi}_{km} = \frac{(2k+\psi-1)(\psi)_{(k-1)}}{m!(k-m)!}$ for $k \geq m$
Define PMF: $q^{\psi}_m(\tau_t) = \sum_{k=m}^{\infty}(-1)^{k-m} c^{\psi}_{km} e^{-k(k+\psi-1)\tau_t/2}$
Sample $U \sim \text{Uniform}[0, 1]$
Initialize $M \leftarrow 0$
Compute initial bounds: $S^- \leftarrow 0$, $S^+ \leftarrow q^{\psi}_0(\tau_t)$
while not ($S^- > U$ or $S^+ < U$) **do**
    Find K_M such that $c^{\psi}_{(K_M+1)M} e^{-(K_M+1)(K_M+\psi)\tau_t/2} < c^{\psi}_{K_M M} e^{-K_M(K_M+\psi-1)\tau_t/2}$
    Update lower bound: $S^- \leftarrow S^- + \sum_{i=0}^{\lfloor K_M/2 \rfloor}(-1)^i c^{\psi}_{(M+2i)M} e^{-(M+2i)(M+2i+\psi-1)\tau_t/2}$
    Update upper bound: $S^+ \leftarrow S^- + \sum_{i=0}^{\lfloor (K_M-1)/2 \rfloor}(-1)^i c^{\psi}_{(M+2i+1)M} e^{-(M+2i+1)(M+2i+\psi)\tau_t/2}$
    if $S^- > U$ **then**
        return $m = M$
    else if $S^+ < U$ **then**
        $M \leftarrow M + 1$
    end if
end while

Compute loss We present a formula for $\vec{s}(\vec{v} \mid x_0, t) = \nabla \log p(\vec{x}_t \mid x_0, t)|_{\vec{v}}$ to enable computation of the loss. Avdeyev et al. [2023] computed these scores using a previously determined result with

[8]Indeed, noting the self-similarity of Brownian motion, one can show that, conditioned on $w_t = 0$, with probability 1 $(z_t)_t$ makes infinitely many transitions in the interval $[t, t + \epsilon)$ for any $\epsilon > 0$. The probability of infinitely many transitions in a bounded interval for discrete diffusion however is 0.

$B = 2$ then generalizing to higher dimensions with their stick-breaking procedure and a change of variables. We are instead able to derive it directly from first principles.

There are two infinite series which will be important,

$$G_\psi(t, x_0, \vec{x}_t) = 1 + \sum_{k=1}^{\infty} (-1)^k a_k^\psi(t, \pi_{x_0}, \vec{x}_{t,x_0})$$
$$F_\psi(t, x_0, \vec{x}_t) = 1 + \sum_{k=1}^{\infty} (-1)^k b_k^\psi(t, \pi_{x_0}, \vec{x}_{t,x_0})$$

where

$$a_k^\psi(t, \pi_{x_0}, \vec{x}_{t,x_0}) = e^{-\frac{k(k+\psi-1)t}{2}} \frac{(2k+\psi-1)(\psi)_{(k-1)}}{k!} {}_2F_1(-k, \psi+k-1; \psi\pi_{x_0}; \vec{x}_{t,x_0})$$
$$b_k^\psi(t, \pi_{x_0}, \vec{x}_{t,x_0}) = e^{-\frac{k(k+\psi+1)t}{2}} \frac{(\psi)_{(k)}}{k!} \frac{(2k+\psi+1)(\psi+k)}{(\psi+1)\psi} {}_2F_1(-k, \psi+k+1; \psi\pi_{x_0}+1; \vec{x}_{t,x_0})$$

where ${}_2F_1$ is the hypergeometric function. Although these look complicated, in practice, most terms in the numerators and denominator of a and b nearly cancel to 1, and, when t is not too small, $e^{-k(k+\psi+1)t/2}$ decays extremely quickly.

Using the results in Tavaré [1984] we compute $\vec{s}(\vec{v} \mid x_0)$ in terms of these series. Since we're only interested in differences for calculating the ELBO, $\vec{s}(\vec{v} \mid x_0, t) - \vec{s}(\vec{v} \mid \tilde{x}_0, t)$ we ignore constants not depending on x_0.

Proposition F.1. *(Proof in App. G.6)*

$$p(\vec{x}_t \mid x_0, t) = \text{Dirichlet}(\pi\psi)(\vec{x}_t) G_\psi(\tau_t, x_0, \vec{v}).$$

For $\vec{c}(\vec{v}) = \nabla \log \text{Dirichlet}(\pi\psi)(\vec{x}_t)$ *which does not depend on* x_0,

$$\vec{s}(\vec{v} \mid x_0, t) = \vec{c}(\vec{v}) + \vec{x}_0 w(x_0, \vec{v}) \tag{2}$$

where

$$w(x_0, \vec{v}) = \frac{e^{-\psi\tau_t/2}(\psi+1)}{\pi(x_0)} \frac{F_\psi(\tau_t, x_0, \vec{v})}{G_\psi(\tau_t, x_0, \vec{v})}.$$

Note with the hollow parameterization, calling $\vec{w}_b = w(b)$, we get

$$\vec{s}(\vec{v} \mid \tilde{x}_0, t) = \vec{c}(\vec{v}) + \frac{e^{-\psi\tau_t/2}(\psi+1)}{\pi(x_0)} \frac{\sum_b \tilde{x}_{0,b} F_\psi(\tau_t, b, \vec{v})}{\sum_b \tilde{x}_{0,b} G_\psi(\tau_t, b, \vec{v})}.$$

F.1 Low time regimen

When t is small, sampling from $A(\psi, \tau_t)$ or calculating G_ψ, F_ψ become unstable. Griffiths [1984] suggested a Gaussian approximation for $A(\psi, \tau_t)$ which we will also use for deriving stable approximations of $\vec{s}(\vec{v} \mid x_0, t)$.

Sample noisy x_t We copy the following from Jenkins and Spanò [2017].

Algorithm 6 Sampling from ancestral process $A(\psi, \tau_t)$ - Low t approximation

1: Set $\beta \leftarrow \frac{1}{2}(\psi - 1)\tau_t$
2: **if** $\beta \neq 0$ **then**
3: Set $\eta \leftarrow \beta e^{1-\beta}$
4: Set $\mu \leftarrow \frac{2\eta}{\tau_t}$
5: Set $\sigma^2 \leftarrow \frac{2\eta}{\tau_t} \left(\frac{\eta+\beta}{1-e^{-2\beta}}\right)^2 \left(1 + \frac{\eta}{\eta+\beta} - 2\eta\right) \beta^{-2}$
6: **else**
7: Set $\mu \leftarrow \frac{2}{\tau_t}$
8: Set $\sigma^2 \leftarrow \frac{2}{3\tau_t}$
9: **end if**
10: Sample $Z \sim \mathcal{N}(\mu, \sigma^2)$
11: **return** $m = \max(0, \lfloor Z + 0.5 \rfloor)$ ▷ Round to nearest non-negative integer

Compute loss The loss in this regimen, even with the Griffiths approximation, becomes intractable; instead we use the Griffiths approximation to simply bound the loss.

When t is small, x_0 is almost always $b^* = \text{argmax}_b \vec{x}_{t,b}$. We therefore set $\tilde{x}_0 = \delta_{b^*}$. $x_0 \neq b^*$ is so rare we only aim to find a loose bound. Calling $\vec{v} = \vec{x}_t$ we bound the loss by

$$\begin{aligned} L \leq & \frac{\dot{\tau}_t}{2}(\|\vec{s}(\vec{v} \mid x_0, t) - \vec{c}(\vec{v})\|_{\text{Diag}(\vec{v})-\vec{v}\vec{v}^T} + \|\vec{s}(\vec{v} \mid b^*, t) - \vec{c}(\vec{v})\|_{\text{Diag}(\vec{v})-\vec{v}\vec{v}^T})^2 \\ = & \frac{\dot{\tau}_t}{2}(w(x_0, \vec{v})\sqrt{\vec{v}_{x_0}} + w(b^*, \vec{v})\sqrt{\vec{v}_{b^*}})^2. \end{aligned}$$

In the next proposition we give an alternate formula for $w(x_0, \vec{v})$ which will allow us to Griffith's approximation and a saddle point approximation to estimate $w(b^*, \vec{v})$. It will also allow us to bound $w(x_0, \vec{v})$. To our knowledge, this strategy is original.

Proposition F.2.

$$w(x_0, \vec{v}) = \vec{v}_{x_0}^{-1}\tilde{\mathbb{E}}_{\vec{v}_{x_0}} m_t$$

where $\tilde{\mathbb{E}}$ is over the weighted, normalized distribution $p(A(\psi, \tau_t) = m_t)\frac{(\psi)_{(m_t)}}{(\psi\pi_{x_0})_{(m_t)}}\vec{v}_{x_0}^{m_t}$.

Proof. Simple inspection of first expression of the proof of Prop. F.1. □

For $w(b^*, \vec{v})$, when t is small and $\vec{v}_{b^*}$ is not small, we derive a saddle point approximation to $\tilde{\mathbb{E}}_{\vec{v}_{x_0}} m_t$ in Eqn. 3 below. $\vec{v}_{x_0}$ is small, the saddle point approximation fails. However, if we're only interested in getting a bound, we can bound $\tilde{\mathbb{E}}_{\vec{v}_{x_0}} m_t \leq \tilde{\mathbb{E}}_{\vec{v}_{b^*}} m_t$ which can then be estimated using the saddle point approximation; this is our strategy for $w(x_0, \vec{v})$. Therefore, we get

$$L \leq 2\dot{\tau}_t \vec{v}_{x_0}^{-1}(\tilde{\mathbb{E}}_{\vec{v}_{b^*}} m_t)^2.$$

Saddle point approximation Let's take the Griffiths approx as t becomes small, so $w_t \sim N(\mu, \sigma)$ where μ, σ are form Alg. 6. Let's use Stirling to approximate

$$(\psi)_{(m_t)} = \frac{(\psi + m_t - 1)!}{\Gamma(\psi - 1)} \approx \frac{\sqrt{2\pi(\psi + m_t - 1)}}{\Gamma(\psi - 1)} \left(\frac{(\psi + m_t - 1)}{e}\right)^{(\psi + m_t - 1)}$$

so

$$\begin{aligned} \frac{(\psi)_{(m_t)}}{(\psi\pi_{x_0})_{(m_t)}} \approx & \frac{\Gamma(\psi\pi_{x_0} - 1)}{\Gamma(\psi - 1)} e^{-(1-\pi_{x_0})\psi} \\ & \times \left(1 + \frac{(1-\pi_{x_0})\psi}{\psi\pi_{x_0} + m_t - 1}\right)^{(\psi\pi_{x_0} + m_t - 1) + 1/2} (\psi + m_t - 1)^{(1-\pi_{x_0})\psi} \\ \approx & \frac{\Gamma(\psi\pi_{x_0} - 1)}{\Gamma(\psi - 1)} e^{\pi_{x_0}\psi} \\ & \times \exp\left(\frac{(1-\pi_{x_0})\psi}{2(\psi\pi_{x_0} + m_t - 1)}\right)(\psi + m_t - 1)^{(1-\pi_{x_0})\psi}. \end{aligned}$$

We take a saddle point approximation of $\tilde{\mathbb{E}}_{x_{t,x_0}} m_t$, i.e. take its value as the maximizer of the approximate log likelihood

$$\begin{aligned} & C - \frac{1}{2\sigma^2}(m_t - \mu)^2 + \frac{(1-\pi_{x_0})\psi}{2(\psi\pi_{x_0} + m_t - 1)} \\ & + (1-\pi_{x_0})\psi\log(\psi + m_t - 1) + m_t\log(\vec{v}_{x_0}) + O(1/m_t). \end{aligned}$$

Forgetting the reciprocal terms, the most naive approximation therefore is

$$\vec{w} \approx \vec{v}_{x_0}^{-1}\left(\mu + \sigma^2 \log \vec{v}_{x_0}\right).$$

Taking into account only the larger reciprocal term, you get a slightly more accurate approximation,

$$\vec{w} \approx \vec{v}_{x_0}^{-1}\left((\tilde{\mu} - (\psi - 1)) + \sqrt{(\tilde{\mu} + (\psi - 1))^2 + 4(1-\pi_{x_0})\psi\sigma^2}\right)/2 \quad (3)$$

where $\tilde{\mu} = \mu + \sigma^2 \log \vec{v}_{x_0}$ is the naive approximation.

Note the second approximation becomes the first when the "perturbation" $4(1 - \pi(x_0))\psi\sigma^2$ is small. Noting $m_t \sim t^{-1}$, the first approximation has relative error roughly $O(t)$ while the second has relative error roughly $O(t^2)$;

When t is extremely small and $\vec{v}_{x_0}$ is large, then

$$\vec{w}_b \approx 2t^{-1}\vec{v}_{x_0}^{-1}\left(1 + \frac{1}{3}\log \vec{v}_{x_0}\right) \approx 2t^{-1}.$$

This is a good approximation when $\log \vec{v}_{x_0}$ is large (say $\vec{v}_{x_0} > 0.5$) but can fail otherwise – it can even give negative numbers!

G Theoretical results

G.1 Mutation population discrete diffusion loss

In this appendix we derive Alg. 1 by showing it is equivalent to Alg. 2. Namely, we assume $D = 1$ and x_t is a sequence of length ζ and show

- **Predict de-noised x_0:** the target of $q_\theta(x_0 \mid x_t, t)$, $p(x_0 \mid x_t, t)$, only depends on the vectorized $\vec{x}_t$.
- **Compute loss:** $L = \sum_{x' \neq x_t} \mathcal{L}_{x' \to x_t} \dot{\tau}_t \mathbb{D}\left(\frac{p(x'|x_0,t)}{p(x_t|x_0,t)} \middle\| \frac{p(x'|\tilde{x}_0,t)}{p(x_t|\tilde{x}_0,t)}\right)$ is equivalent to the form in Alg 1.

Given prediction and loss computation only depend on $\vec{x}_t$, we can also replace sampling x_t with just sampling $\vec{x}_t \sim \mathrm{Mult}(\zeta, \vec{x}_0^T e^{\tau_t \mathcal{L}})/\zeta$, giving Alg. 1.

Predict de-noised x_0 Simply note

$$\begin{aligned} p(x_0 \mid x_t, t) \propto & p(x_0) p(x_t \mid x_0, t) \\ =& p(x_0) \prod_{z=0}^{\zeta} (\vec{x}_0^T e^{\tau_t \mathcal{L}})_{x_t^{(z)}} \\ =& p(x_0) \prod_{b=1}^{B} (\vec{x}_0^T e^{\tau_t \mathcal{L}})_b^{\zeta \vec{x}_{t,b}}. \end{aligned}$$

Compute loss For sequences $x \neq x$ of length ζ which differ in exactly one position, say $x^{(z)} = b \neq b' = x'^{(z)}$, then $\mathcal{L}_{x \to x'} = \mathcal{L}_{b \to b'}$ and for every x_0

$$\frac{p(x' \mid x_0, t)}{p(x \mid x_0, t)} = \frac{\vec{x}_0 e^{\tau_t \mathcal{L}} \vec{b}'}{\vec{x}_0 e^{\tau_t \mathcal{L}} \vec{b}}.$$

If x, x' differ in more than one position, then $\mathcal{L}_{x \to x'} = 0$. Call $x_t^{[z,b]}$ a sequence which has all the same letters as x_t except has b in position z. Then calling $\vec{p} = \vec{x}_0^T e^{\tau_t \mathcal{L}}$ and $\vec{q} = \tilde{x}_0^T e^{\tau_t \mathcal{L}}$,

$$\begin{aligned} L =& \sum_{x' \neq x_t} \mathcal{L}_{x' \to x_t} \dot{\tau}_t \mathbb{D}\left(\frac{p(x' \mid x_0, t)}{p(x_t \mid x_0, t)} \middle\| \frac{p(x' \mid \tilde{x}_0, t)}{p(x_t \mid \tilde{x}_0, t)}\right) \\ =& \sum_{z=0}^{\zeta} \sum_{b' \neq x_t^{(z)}} \mathcal{L}_{b' \to x_t^{(z)}} \dot{\tau}_t \mathbb{D}\left(\frac{\vec{p}_{b'}}{\vec{p}_{x_t^{(z)}}} \middle\| \frac{\vec{q}_{b'}}{\vec{q}_{x_t^{(z)}}}\right) \\ =& \sum_{b} \#\{z \mid x_t^{(z)} = b\} \sum_{b' \neq b} \mathcal{L}_{b' \to b} \dot{\tau}_t \mathbb{D}\left(\frac{\vec{p}_{b'}}{\vec{p}_b} \middle\| \frac{\vec{q}_{b'}}{\vec{q}_b}\right) \\ =& \sum_{b' \neq b} \mathcal{L}_{b' \to b} \dot{\tau}_t \zeta \vec{x}_{t,b} \mathbb{D}\left(\frac{\vec{p}_{b'}}{\vec{p}_b} \middle\| \frac{\vec{q}_{b'}}{\vec{q}_b}\right). \end{aligned}$$

G.2 Proof of Gaussian convergence

Our formal statement of the theorem adds some mild positivity assumptions for τ, π and P_1 which are satisfied by any reasonable choice of τ and almost every choice of $\mathcal{L}$. It is also more specific about the limiting behaviour of $\vec{x}_t^\zeta$ in non-dominant eigenspaces: we also limit to Gaussian diffusion, but with meaningless embeddings sampled from random Gaussian vectors independent of x_0.

Let us interpret the embedding Q_1. In the case that $\mathcal{L}$ is doubly stochastic, or reversible, $\pi = [\frac{1}{B}, \dots, \frac{1}{B}]$ and $\mathcal{L}$ is symmetric; in this case $Q_1 = \mathrm{j}_1 P_1$ is just the orthogonal projection onto the dominant eigenspace. In the more general case that $\mathcal{L}$ satisfies detailed balance, $(\mathrm{diag}(\pi)^{1/2}\mathcal{L}\mathrm{diag}(\pi)^{-1/2})_{ij} = \sqrt{\frac{\pi_i}{\pi_j}}\mathcal{L}_{ij}$ is symmetric so $\tilde{Q}_i$ is the orthogonal projection onto the dominant eigenspace of the "symmetrized" generator. In more general cases, we don't get a symmetrized operator or an orthogonal projection $\tilde{Q}_i$, so we must "correct" for this with the adjustment $(\tilde{Q}_i\tilde{Q}_i^T)^{-1/2}\tilde{Q}_i$.

Theorem G.1. *(Formal statement and proof of Thm. 3.1) Call $-\lambda_1 > -\lambda_2 > \dots$ the negative eigenvalues of $\mathcal{L}$ and $P_1, P_2, \dots$ the projections onto the corresponding left eigen-space. Without loss of generality, assume $\lambda_1 = 1$. Assume $\dot{\tau}_t$ is bounded on every compact interval of $(0, 1)$, $\pi_b > 0$ and $P_1\vec{b} \neq 0$ for all b and $P_1\vec{b} \neq P_1\vec{b}'$ for any $b \neq b'$. For each ζ pick time dilation $\tau_t^\zeta = \frac{1}{2}\log\left(\zeta e^{2\tau_t} - \zeta + 1\right)$ and rescale $\vec{x}_t^\zeta = \sqrt{\zeta - (\zeta - 1)e^{-2\tau_t}}(\vec{x}_t - \pi)/\sqrt{\pi}$. Define the embedding into $\mathbb{R}^{\mathrm{rank}(P_i)}$, $Q_i = \mathrm{j}_i(\tilde{Q}_i\tilde{Q}_i^T)^{-1/2}\tilde{Q}_i$ where $\tilde{Q}_i = \mathrm{diag}(\pi)^{-1/2}P_i\mathrm{diag}(\pi)^{1/2}$ and j_i is any isometry from $\mathrm{Im}(\tilde{Q}_i) \to \mathbb{R}^{\mathrm{rank}(P_i)}$.*

Fix an x_0.

- *(Path convergence) Call $(\vec{z}_t)_{t=0}^1$ the paths with $\vec{z}_0 = Q_1(\vec{x}_0/\sqrt{\pi})$ evolving under the Ornstein-Uhlenbeck process*

$$d\tilde{z}_\tau = -\tilde{z}_\tau d\tau + \sqrt{2}dW_\tau$$

  for a Brownian motion $(W_\tau)_{\tau=0}^\infty$ and call $\vec{z}_t = \tilde{z}_{\tau_t}$. Then $(Q_1\vec{x}_t^\zeta)_{t\in(0,1)}$ converges in distribution to $(\vec{z}_t)_{t\in(0,1)}$ in the sense of Lem. G.7.

- *(Convergence of non-dominant directions) The component of $\vec{x}_t^\zeta$ in $\mathrm{Ker}Q_1$ is $\sum_{i>1}\tilde{Q}_i\vec{x}_t^\zeta$. Each component $(Q_i\vec{x}_t^\zeta)_t$ also converges to a Gaussian diffusion independent of $\vec{x}_0$ with modified time-dilation and scaling: call $(\vec{z}_t)_{t=0}^1$ the paths with $\vec{z}_0 \sim \mathcal{N}(0, I)$ independent of x_0 evolving, forward and backward on $(-\infty, \infty)$, under the stationary Ornstein-Uhlenbeck process*

$$d\tilde{z}_\tau = -\tilde{z}_\tau d\tau + \sqrt{2}dW_\tau$$

  for a Brownian motion $(W_\tau)_{\tau=0}^\infty$ and call $\vec{z}_t = \tilde{z}_{\tau_t^{(i)}}$ where $\tau_t^{(i)} = \frac{\lambda_i}{2}\log(e^{2\tau_t} - 1)$. Then

$$((1 - e^{-2\tau_t})^{-1/2}Q_i\vec{x}_t^\zeta)_{t\in(0,1)} \rightsquigarrow (\vec{z}_t)_{t\in(0,1)}.$$

- *Call the ELBO in Alg. 1*

$$L(\vec{x}_t^\zeta, t, \vec{x}_0, \tilde{x}_0) = \sum_{b_1\neq b_2}\mathcal{L}_{b_2\to b_1}\dot{\tau}_t^\zeta\zeta\vec{x}_{t,b_1}(\vec{x}_t^\zeta)\mathbb{D}\left(\frac{p_{b_2}}{p_{b_1}}\middle\|\frac{q_{b_2}}{q_{b_1}}\right)$$

  where $\vec{x}_{t,b_1}(\vec{v})$ is the inverse of the transform from $\vec{x}_{t,b_1}$ to $\vec{x}_{t,b_1}^\zeta$. Then, for all $\vec{v}, t, \vec{x}_0, \tilde{x}_0$

$$L(\vec{v}, t, \vec{x}_0, \tilde{x}_0) \to \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2}\left\|\mathrm{emb}(x_0) - \mathrm{emb}(\tilde{x}_0)\right\|^2,$$

  the ELBO in Alg. 3, which, in particular, is independent of the value of $\vec{v}$.

Proof. We prove the convergence of paths using Lem. G.7 which makes use of standard techniques. We break the proof up into four sections: the first three verify the conditions of Lem. G.7 and the last shows the convergence of the ELBO.

Part 1. Convergence of Marginals: Note

$$\vec{z}_t \sim e^{-\tau_t}\vec{z}_0 + \sqrt{1-e^{-2\tau_t}}\mathcal{N}(0,I).$$

We want to prove convergence to this quantity. Note, writing Mult for a multinomial distribution,

$$\begin{aligned}\vec{x}_t^\zeta \sim & \frac{\sqrt{\zeta-(\zeta-1)e^{-2\tau_t}}}{\zeta}\left(\mathrm{Mult}(\zeta, \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}) - \zeta\vec{\pi}\right)/\sqrt{\vec{\pi}}\\ =&(1+o(1))\sqrt{1-e^{-2\tau_t}}\left(\zeta^{-1/2}(\mathrm{Mult}(\zeta, \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}) - \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}) + \zeta^{1/2}(\vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}} - \vec{\pi})\right)/\sqrt{\vec{\pi}}.\end{aligned}$$

The second term is

$$\begin{aligned}\sqrt{1-e^{-2\tau_t}}\zeta^{1/2}(\vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}} - \vec{\pi}) =& \sqrt{1-e^{-2\tau_t}}\sum_i \zeta^{1/2}e^{-\lambda_i\tau_t^\zeta}P_i\vec{x}_0\\ =&\sum_i\left(\frac{\zeta(1-e^{-2\tau_t})}{(\zeta(e^{2\tau_t}-1)+1)^{\lambda_i}}\right)^{1/2}P_i\vec{x}_0\\ \to& e^{-\tau_t}P_1\vec{x}_0.\end{aligned}$$

For the first term, we need a "uniform" central limit theorem as the underlying distribution changes with ζ because of $\vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}$. Lem. G.8 shows that $\zeta^{-1/2}(\mathrm{Mult}(\zeta, \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}) - \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}})$ approaches $\mathcal{N}(0, \mathrm{diag}(\vec{p}_t) - \vec{p}_t\vec{p}_t^T)$ for $\vec{p}_t = \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}$, which itself approaches $\vec{\pi}$ as $\tau_t^\zeta \to \infty$. Therefore the first term, divided by $\sqrt{\vec{\pi}}$ approaches

$$\sqrt{1-e^{-2\tau_t}}\mathcal{N}\left(0, I - \sqrt{\vec{\pi}}\sqrt{\vec{\pi}}^T\right).$$

Note $\tilde{Q}_i\sqrt{\vec{\pi}} = \sqrt{\vec{\pi}}^{-1}P_i\vec{\pi} = 0$ for each i and, for $i>1$, $\tilde{Q}_i(P_1\vec{x}_0/\sqrt{\vec{\pi}}) = \sqrt{\vec{\pi}}^{-1}P_iP_1\vec{x}_0 = 0$. Therefore, as deired,

$$Q_1 x_t^\zeta \rightsquigarrow \sqrt{1-e^{-2\tau_t}}\mathcal{N}(0,I) + e^{-\tau_t}\mathrm{emb}(x_0),$$

and for $i>1$,

$$(1-e^{-2\tau_t})^{-1/2}Q_i x_t^\zeta \rightsquigarrow \mathcal{N}(0,I).$$

Part 2. Local uniform convergence of conditionals: Note

$$\vec{z}_t|\vec{z}_s \sim e^{-(\tau_t-\tau_s)}\vec{z}_s + \sqrt{1-e^{-2(\tau_t-\tau_s)}}\mathcal{N}(0,I).$$

We want to prove convergence to this quantity. Note

$$\vec{x}_t|\vec{x}_s \sim \sum_b \mathrm{Mult}(\zeta\vec{x}_{s,b}, \vec{b}^T e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}})/\zeta$$

where $\vec{x}_t = \sqrt{\pi}\circ\vec{x}_t^\zeta/\sqrt{\zeta-(\zeta-1)e^{-2\tau_t}} + \pi$ are the "unscaled" versions of the vector and $\vec{x}_s$ is similar. It will be convenient below to extend this definition to $\vec{x}_s^\zeta$ for which $\zeta\vec{x}_{s,b}$ are not integers, but which still satisfy $\sum_b\sqrt{\pi_b}\vec{x}_{t,b}^\zeta = 0$. To do so, we just round $\zeta\vec{x}_{s,b}$ down to $\lfloor\zeta\vec{x}_{s,b}\rfloor$.

Fix $\vec{v}$. We now show $\vec{x}_t^\zeta|\vec{x}_s^\zeta = \vec{v} \rightsquigarrow \vec{z}_t|\vec{z}_s = \vec{v}$; a very similar argument also shows $\vec{x}_t^\zeta \rightsquigarrow \vec{z}_t$. Call $\vec{x}^\zeta$ a variable distributed as $\vec{x}_t^\zeta|\vec{x}_s^\zeta = \vec{v}$, so, calling

$$w_t^\zeta = \frac{\sqrt{\zeta-(\zeta-1)e^{-2\tau_t}}}{\zeta},$$

$$N_{s,b}^\zeta = \sqrt{\pi_b}\vec{v}_b/w_s^\zeta + \zeta\pi_b,$$

$$C_{t,b}^\zeta \sim \mathrm{Mult}\left(\left\lfloor N_{s,b}^\zeta\right\rfloor, \vec{b}^T e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}}\right) \text{ independent across } b,$$

then

$$\begin{aligned}\vec{x}_t^\zeta \sim & w_t^\zeta\left(\sum_b C_{t,b}^\zeta - \zeta\pi\right)/\sqrt{\pi}\\ =& w_t^\zeta\left(\sum_b\left[(C_{t,b}^\zeta - N_{s,b}^\zeta\vec{b}^T e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}}) + N_{s,b}^\zeta(\vec{b}^T e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}} - \pi)\right]\right)/\sqrt{\pi}\end{aligned}$$

noting $\sum_b p^\zeta_{t,b} = \zeta$. This is exactly the "noise, signal" breakdown we had in the proof sketch.

For the signal (second term), first note

$$\sum_b \pi_b(\vec{b}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}} - \vec{\pi}) = \vec{\pi}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}} - \vec{\pi} = 0,$$

so, ignoring the π term in $N^\zeta_{s,b}$ the second term is

$$\begin{aligned}
\frac{w_t}{w_s}\left(\sum_b \sqrt{\pi_b}\vec{v}_b(\vec{b}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}} - \vec{\pi})\right)/\sqrt{\pi} =& \frac{w_t}{w_s}\left((\sqrt{\pi}\circ\vec{v})^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}}\right)/\sqrt{\pi} \\
=&(1+o(1))\sum_i \left(\frac{1-e^{-2\tau_s}}{1-e^{-2\tau_t}}\right)^{(\lambda_i-1)/2} e^{-\lambda_i(\tau_t-\tau_s)}\tilde{Q}_i\vec{v}.
\end{aligned}$$

For the first term, we again apply Lem. G.8, noting $N^\zeta_{s,b} = (1+o(1))\zeta\pi_b$ to get

$$\begin{aligned}
&\sum_b w_t(C^\zeta_{t,b} - N^\zeta_{s,b}\vec{b}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}})/\sqrt{\vec{\pi}} \\
&\quad\rightsquigarrow\sqrt{1-e^{-2\tau_t}}\sum_b \sqrt{\pi_b}\mathcal{N}\left(0, \mathrm{diag}(\vec{b}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}}) - e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}^T}\vec{b}\vec{b}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}}\right)/\sqrt{\vec{\pi}} \\
&\quad=\sqrt{1-e^{-2\tau_t}}\mathcal{N}\left(0, \mathrm{diag}(\vec{\pi}^T e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}}) - e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}^T}\mathrm{diag}(\vec{\pi})e^{(\tau^\zeta_t - \tau^\zeta_s)\mathcal{L}}\right)/\sqrt{\vec{\pi}} \\
&\quad=\sqrt{1-e^{-2\tau_t}}\mathcal{N}\left(0, \mathrm{diag}(\vec{\pi}) - (\sum_i e^{-\lambda_i(\tau^\zeta_t - \tau^\zeta_s)}P_i)\mathrm{diag}(\vec{\pi})(\sum_i e^{-\lambda_i(\tau^\zeta_t - \tau^\zeta_s)}P_i^T)\right)/\sqrt{\vec{\pi}} \\
&\quad=\sqrt{1-e^{-2\tau_t}}\mathcal{N}\left(0, I - (\sum_i e^{-\lambda_i(\tau^\zeta_t - \tau^\zeta_s)}\tilde{Q}_i)(\sum_i e^{-\lambda_i(\tau^\zeta_t - \tau^\zeta_s)}\tilde{Q}_i^T)\right).
\end{aligned}$$

Therefore, as desired,

$$\begin{aligned}
Q_1\vec{x}^\zeta_t \mid \vec{x}^\zeta_s =& \vec{v} \sim e^{-(\tau_t-\tau_s)}Q_1\vec{v} + \sqrt{(1-e^{-2\tau_t})\left(1 - \frac{1-e^{-2\tau_s}}{1-e^{-2\tau_t}}e^{-2(\tau_t-\tau_s)}\right)}\mathcal{N}(0,I) \\
=& \vec{v} \sim e^{-(\tau_t-\tau_s)}Q_1\vec{v} + \sqrt{1-e^{-2(\tau_t-\tau_s)}}\mathcal{N}(0,I)
\end{aligned}$$

and similarly

$$\begin{aligned}
(1-e^{-2\tau_t})^{-1/2}Q_i\vec{x}^\zeta_t \mid \vec{x}^\zeta_s = \vec{v} \sim& \left(\frac{1-e^{-2\tau_s}}{1-e^{-2\tau_t}}\right)^{\lambda_i/2} e^{-\lambda_i(\tau_t-\tau_s)}((1-e^{-2\tau_2})^{-1/2}Q_i\vec{v}) \\
&+ \sqrt{1 - \left(\frac{1-e^{-2\tau_s}}{1-e^{-2\tau_t}}\right)^{\lambda_i} e^{-2\lambda_i(\tau_t-\tau_s)}}\mathcal{N}(0,I) \\
=& e^{-(\tau^{(i)}_t - \tau^{(i)}_s)}((1-e^{-2\tau_2})^{-1/2}Q_i\vec{v}) \\
&+ \sqrt{1-e^{-2(\tau^{(i)}_t - \tau^{(i)}_s)}}\mathcal{N}(0,I)
\end{aligned}$$

Finally, convergence is clearly uniform for nearby $\vec{v}$ using the uniformity of Lem. G.8.

Part 3. Tightness: Pick $s < t \in (0,1)$.

$$\mathbb{E}\|\vec{x}^\zeta_t - \vec{x}^\zeta_s\|^2 = \mathbb{E}\|\mathbb{E}[\vec{x}^\zeta_t|\vec{x}^\zeta_s] - \vec{x}^\zeta_s\|^2 + \mathbb{E}\|\vec{x}^\zeta_t - \mathbb{E}[\vec{x}^\zeta_t|\vec{x}^\zeta_s]\|^2$$

The first term has, for each x_0,

$$
\begin{aligned}
\mathbb{E}\|\mathbb{E}[\vec{x}_t^\zeta|\vec{x}_s^\zeta]-\vec{x}_s^\zeta\|^2 =&\mathbb{E}\|w_t(\vec{x}_s e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}}-\vec{\pi})/\sqrt{\vec{\pi}}-\vec{x}_s^\zeta\|^2\\
=&\frac{1}{\min_b \pi_b}\mathbb{E}\|w_t(\vec{x}_s e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}}-\vec{x}_s)-(w_s-w_t)(\vec{x}_s-\vec{\pi})\|^2\\
\leq&\frac{1}{\min_b \pi_b}\mathbb{E}\left(|w_t|\|\vec{x}_s e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}}-\vec{x}_s\|+|w_s-w_t|\|\vec{x}_s-\vec{\pi}\|\right)^2\\
=&\frac{1}{\min_b \pi_b}\mathbb{E}\left(|w_t|\|(\vec{x}_s-\vec{\pi})^T(I-e^{(\tau_t^\zeta-\tau_s^\zeta)\mathcal{L}})\|+|w_s-w_t|\|\vec{x}_s-\vec{\pi}\|\right)^2\\
\leq&\frac{1}{\min_b \pi_b}\mathbb{E}\left(|w_t|(1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B})\|\vec{x}_s-\vec{\pi}\|+|w_s-w_t|\|\vec{x}_s-\vec{\pi}\|\right)^2\\
=&\frac{1}{\min_b \pi_b}\left(|w_t|(1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B})+|w_s-w_t|\right)^2\mathbb{E}\|\vec{x}_s-\vec{\pi}\|^2\\
\leq&\frac{\zeta}{\min_b \pi_b}\left((1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B})+|1-\frac{w_s}{w_t}|\right)^2\\
&\times\left(\mathbb{E}\mathrm{TrCov}(\mathrm{Mult}(\zeta,\vec{x}_0^T e^{\tau_s^\zeta)\mathcal{L}}/\zeta)+\|\vec{x}_0^T e^{\tau_s^\zeta\mathcal{L}}-\vec{\pi}\|^2\right)\\
\leq&\frac{1}{\min_b \pi_b}\left((1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B})+|1-\frac{w_s}{w_t}|\right)^2(1+\zeta e^{-2\tau_s^\zeta})\\
\leq&\frac{1}{\min_b \pi_b}\left((1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B})+|1-\frac{w_s}{w_t}|\right)^2\left(1+\frac{1}{e^{2\tau_s}-1}\right)
\end{aligned}
$$

Now,

$$
\begin{aligned}
1-e^{-(\tau_t^\zeta-\tau_s^\zeta)\lambda_B}=&1-e^{-2\lambda_B(\tau_t-\tau_s)}\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)^{\lambda_B/2}\\
\leq&1-e^{-2\lambda_B(\tau_t-\tau_s)}\\
&+1-\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)^{\lambda_B/2}.
\end{aligned}
$$

When $|\tau_s-\tau_t|<1/4\lambda_B$

$$1-e^{-2\lambda_B(\tau_t-\tau_s)}\leq 4\lambda_B(\tau_t-\tau_s)\leq 4\lambda_B|t-s|\sup_{u\in[s,t]}\dot{\tau}_u.$$

Next note that if $\alpha\geq 1$, $x\mapsto 1-x^\alpha$ has decreasing derivative, from 0 to $-\alpha$ on the interval $x\in[0,1]$, so, it is dominated on this interval by $\alpha(1-x)$. If $\zeta>1$,

$$
\begin{aligned}
1-\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)^{\lambda_B/2}\leq&1-\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)^{1\vee(\lambda_B/2)}\\
\leq&(1\vee(\lambda_B/2))\left(1-\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)\right)\\
\leq&(1\vee(\lambda_B/2))\left(\frac{(e^{-2\tau_s}-e^{-2\tau_t})(1-\zeta^{-1})}{1-e^{-2\tau_t}}\right)\\
\leq&\frac{1\vee(\lambda_B/2)e^{-2\tau_s}}{1-e^{-2\tau_t}}\left(1-e^{-2(\tau_t-\tau_s)}\right)\\
\leq&\frac{4\vee(2\lambda_B)e^{-2\tau_s}}{1-e^{-2\tau_t}}|t-s|\sup_{u\in[s,t]}\dot{\tau}_u
\end{aligned}
$$

Finally

$$1-\frac{w_s}{w_t}=1-\left(\frac{1-e^{-2\tau_s}(1-\zeta^{-1})}{1-e^{-2\tau_t}(1-\zeta^{-1})}\right)^{1/2}.$$

which is similar to above.

The second term has

$$
\begin{aligned}
\mathbb{E}\|\vec{x}_t^\zeta - \mathbb{E}[\vec{x}_t^\zeta|\vec{x}_s^\zeta]\|^2 \leq & \frac{2\zeta}{\min_b \pi_b} \sum_b \mathbb{E}\mathrm{TrCov}(\mathrm{Mult}(\zeta \vec{x}_{s,b}, \vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}})/\zeta \mid \vec{x}_t^\zeta) \\
= & \frac{2}{\min_b \pi_b} \sum_b \mathbb{E}\vec{x}_{s,b}^\zeta \sum_{b'} (\vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}} \vec{b}')(1 - \vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}} \vec{b}') \\
\leq & \frac{2}{\min_b \pi_b} \left(\sum_{b \neq b'} \vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}} \vec{b}' + \sum_b (1 - \vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}} \vec{b}) \right) \\
= & \frac{4}{\min_b \pi_b} \sum_b (1 - \vec{b}^T e^{(\tau_t^\zeta - \tau_s^\zeta)\mathcal{L}} \vec{b}) \\
\leq & \frac{4B}{\min_b \pi_b} (1 - e^{-(\tau_t^\zeta - \tau_s^\zeta)\lambda_B})
\end{aligned}
$$

which is bounded similar to the first term.

Part 4. Convergence of the ELBO: Define $p = \vec{x}_0^T e^{\tau_t^\zeta \mathcal{L}}$; $q = \tilde{x}_0^T e^{\tau_t^\zeta \mathcal{L}}$. We've shown above that

$$
p = \vec{\pi} + \sqrt{\frac{1}{\zeta(e^{2\tau_t} - 1)}} P_1 \vec{x}_0 + o(\zeta^{-1/2})
$$

so

$$
\frac{p_{b_2}}{p_{b_1}} = \frac{\pi_{b_2}}{\pi_{b_1}} + \frac{1}{\pi_{b_1}} \sqrt{\frac{1}{\zeta(e^{2\tau_t} - 1)}} \left(\vec{b}_2 - \frac{\pi_{b_2}}{\pi_{b_1}} \vec{b}_1 \right)^T P_1 \vec{x}_0 + o(\zeta^{-1/2})
$$

and similar for q. Using a second-order Taylor expansion on $\mathbb{D}$, we get

$$
\mathbb{D}\left(\frac{p_{b_2}}{p_{b_1}} \middle\| \frac{q_{b_2}}{q_{b_1}} \right) = \frac{1}{2} \frac{\pi_{b_1}}{\pi_{b_2}} \frac{1}{\pi_{b_1}^2 \zeta(e^{2\tau_t} - 1)} \left(\left(\vec{b}_2 - \frac{\pi_{b_2}}{\pi_{b_1}} \vec{b}_1 \right)^T P_1(\vec{x}_0 - \tilde{x}_0) \right)^2 + o(\zeta^{-1}).
$$

Next note $\dot{\tau}_t^\zeta = \dot{\tau}_t \frac{e^{2\tau_t}}{e^{2\tau_t} - 1} + o(1)$. Finally note

$$
\vec{x}_t(\vec{v}) = \sqrt{\pi} \circ \vec{v} / \sqrt{\zeta - (\zeta - 1)e^{-2\tau_t}} + \pi = \pi + o(1).
$$

Putting this together, we get

$$
\begin{aligned}
& L(\vec{v}, t, \vec{x}_0, \tilde{x}_0) \\
& = \sum_{b_1 \neq b_2} \mathcal{L}_{b_2 \to b_1} \dot{\tau}_t^\zeta \zeta \vec{x}_{t,b_1}(\vec{v}) \mathbb{D}\left(\frac{p_{b_2}}{p_{b_1}} \middle\| \frac{q_{b_2}}{q_{b_1}} \right) \\
& = \dot{\tau}_t \sum_{b_1 \neq b_2} \mathcal{L}_{b_2 \to b_1} \frac{e^{2\tau_t}}{e^{2\tau_t} - 1} \pi_{b_1} \frac{1}{2\pi_{b_2}\pi_{b_1}} \frac{1}{(e^{2\tau_t} - 1)} \left(\left(\vec{b}_2 - \frac{\pi_{b_2}}{\pi_{b_1}} \vec{b}_1 \right)^T P_1(\vec{x}_0 - \tilde{x}_0) \right)^2 + o(1) \\
& = \frac{\dot{\tau}_t e^{2\tau_t}}{2(e^{2\tau_t} - 1)^2} \sum_{b_1 \neq b_2} \mathcal{L}_{b_2 \to b_1} \left(\left(\vec{b}_2 - \sqrt{\frac{\pi_{b_2}}{\pi_{b_1}}} \vec{b}_1 \right)^T \tilde{Q}_1 \left((\vec{x}_0 - \tilde{x}_0)/\sqrt{\vec{\pi}} \right) \right)^2 + o(1) \\
& = \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} \left\| \tilde{Q}_1 \left((\vec{x}_0 - \tilde{x}_0)/\sqrt{\vec{\pi}} \right) \right\|_\Sigma^2 + o(1)
\end{aligned}
$$

where

$$
\Sigma = \frac{1}{2} \sum_{b_1 \neq b_2} \mathcal{L}_{b_2 \to b_1} \left(\vec{b}_2 - \sqrt{\frac{\pi_{b_2}}{\pi_{b_1}}} \vec{b}_1 \right) \left(\vec{b}_2 - \sqrt{\frac{\pi_{b_2}}{\pi_{b_1}}} \vec{b}_1 \right)^T.
$$

To solve Σ, we note

$$\begin{aligned}
\sum_{b_1\neq b_2}\mathcal{L}_{b_2\to b_1}\vec{b}_2\vec{b}_2^T &= \sum_{b_2}\vec{b}_2\vec{b}_2^T\sum_{b_1\neq b_2}\mathcal{L}_{b_2\to b_1} = -\sum_{b_2}\vec{b}_2\vec{b}_2^T\mathcal{L}_{b_2,b_2}\\
\sum_{b_1\neq b_2}\frac{\pi_{b_2}}{\pi_{b_1}}\mathcal{L}_{b_2\to b_1}\vec{b}_2\vec{b}_2^T &= \sum_{b_1}\vec{b}_1\vec{b}_1^T\sum_{b_2\neq b_1}\frac{\pi_{b_2}}{\pi_{b_1}}\mathcal{L}_{b_2\to b_1} = -\sum_{b_1}\vec{b}_1\vec{b}_1^T\mathcal{L}_{b_1,b_1}\\
\sum_{b_1\neq b_2}\sqrt{\frac{\pi_{b_2}}{\pi_{b_1}}}\mathcal{L}_{b_2\to b_1}\vec{b}_2\vec{b}_1^T &= \mathrm{diag}(\sqrt{\vec{\pi}})\left(\mathcal{L}-\mathrm{diag}\mathcal{L}\right)\mathrm{diag}(1/\sqrt{\vec{\pi}})\\
\sum_{b_1\neq b_2}\sqrt{\frac{\pi_{b_2}}{\pi_{b_1}}}\mathcal{L}_{b_2\to b_1}\vec{b}_1\vec{b}_2^T &= (\mathrm{diag}(\sqrt{\vec{\pi}})\left(\mathcal{L}-\mathrm{diag}\mathcal{L}\right)\mathrm{diag}(1/\sqrt{\vec{\pi}}))^T.
\end{aligned}$$

So,

$$\Sigma = -\frac{1}{2}\mathrm{diag}(\sqrt{\vec{\pi}})\mathcal{L}\mathrm{diag}(1/\sqrt{\vec{\pi}}) - \frac{1}{2}(\mathrm{diag}(\sqrt{\vec{\pi}})\mathcal{L}\mathrm{diag}(1/\sqrt{\vec{\pi}}))^T.$$

In particular, since $\tilde{Q}_1^T\mathrm{diag}(\sqrt{\vec{\pi}})\mathcal{L}\mathrm{diag}(1/\sqrt{\vec{\pi}}) = -\tilde{Q}_1^T$,

$$\tilde{Q}_1^T\Sigma\tilde{Q}_1 = \tilde{Q}_1^T\tilde{Q}_1 = Q_1^TQ_1^T.$$

This gives us

$$L(\vec{v},t,\vec{x}_0,\tilde{x}_0) \to \frac{\dot{\tau}_te^{-2\tau_t}}{(1-e^{-2\tau_t})^2}\left\|\mathrm{emb}(x_0)-\mathrm{emb}(\tilde{x}_0)\right\|^2.$$

□

G.3 Proof of sufficient statistics

Proposition G.2. *(Proof of Prop. 4.1) There is a function F^d, **depending on** $p(x_0)$ **and not on the diffusion process or** t**,** such that*

$$p(x_0^d \mid x_t^{-d}, t) = F^d(\vec{\phi}(\vec{x}_t^1, t), \dots, \vec{\phi}(\vec{x}_t^D, t)).$$

Proof.

$$\begin{aligned}
p(x_0^d \mid x_t^{-d}) =& \int p(x_0^d \mid x_0^{-d})dp(x_0^{-d} \mid x_t^{-d})\\
=&\frac{1}{p(x_t^{-d})}\int p(x_0^d \mid x_0^{-d})p(x_t^{-d} \mid x_0^{-d})dp(x_0^{-d})\\
=&\frac{1}{p(x_t^{-d})}\int p(x_0^d \mid x_0^{-d})\prod_{d'\neq d}p(x_t^{d'} \mid x_0^{d'})dp(x_0^{-d})\\
=&\frac{\prod_{d'\neq d}\sum_b p(x_t^{d'} \mid x_0^{d'}=b)}{p(x_t^{-d})}\int p(x_0^d \mid x_0^{-d})\prod_{d'\neq d}\frac{p(x_t^{d'} \mid x_0^{d'})}{\sum_b p(x_t^{d'} \mid x_0^{d'}=b)}dp(x_0^{-d})\\
=&E_{p(x_0^{-d})}\left(p(x_0^d|x_0^{-d})\prod_{d'\neq d}\vec{\phi}(x_t^{d'})_{x_0^{d'}}\right)/E_{p(x_0^{-d})}\left(\prod_{d'\neq d}\vec{\phi}(x_t^{d'})_{x_0^{d'}}\right),
\end{aligned}$$

□

G.4 Hollow parameterization solves Gaussian ELBO singularity

Here we show that the hollow parametrization introduced above resolves the singularity of the Gaussian ELBO in Alg. 3 at $t \to 0^+$. Before going into the proof, let us give some intuition. Assume, x_0^d were distributed uniformly and independently. Then

$$p(x_0^d \mid x_t, t) \propto p(x_t^d \mid x_0^d, t)p(x_0^d \mid x_t^{-d}, t),$$

where x_t^{-d} includes all positions but d. However

$$p(x_0^d \mid x_t^{-d}, t) = \int p(x_0^d \mid x_0^{-d}) dp(x_0^{-d} \mid x_t^{-d}, t) = \text{Uniform}.$$

Therefore, we get $p(x_0^d \mid x_t, t) \propto p(x_t^d \mid x_0^d, t)$. At initialization, we can say our neural network $q_\theta(x_0^d \mid x_t^{-d}, t) \approx \text{Uniform}$, so,

$$q_\theta(x_0^d \mid x_t, t) \approx p(x_0^d \mid x_t, t).$$

Therefore, **the hollow parametrization initializes the diffusion model near a uniform, site-wise independent model.** The proof below involves a lot of algebra, but the basic intuition for why we should not see singularities is that by initializing at a *valid* diffusion model, we get comparable ELBOs.

Again we assume $D = 1$ for simplicity as results are straightforward to generalize to higher D.

Proposition G.3. *Assume* emb *is injective and* τ_t *is increasing and differentiable. Define*

$$L = \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} \|\text{emb}(x_0) - \text{emb}(\tilde{x}_0)\|^2,$$

and the normalized vectors $\vec{\phi}(x_t, t) \propto p(x_t \mid x_0, t)$. *For* $\tilde{x}_0$ *build using the hollow predictor* $\tilde{x}_0 = \vec{\phi}(x_t, t) \circ \vec{q} / \vec{\phi}(x_t, t)^T \circ \vec{q}$ *for a vector* t *bounded away from* 0 *and* ∞*,*

$$0 < c = \min_b \vec{q}_b \leq \max_b \vec{q}_b < C < \infty,$$

we have

$$\mathbb{E}_{t, x_0, x_t} L < \infty.$$

Proof. Note first

$$\|\text{emb}(x_0) - \text{emb}(\tilde{x}_0)\|^2 \leq \|\text{emb}\|^2 \|\vec{x}_0 - \tilde{x}_0\|$$

and, simplifying $\vec{\phi} = \vec{\phi}(x_t, t)$,

$$\mathbb{E}_{x_0 \mid x_t} \|\vec{x}_0 - \tilde{x}_0\| = \|\vec{\phi} \circ \vec{p} / \vec{\phi}^T \vec{p} - \vec{\phi} \circ \vec{q} / \vec{\phi}^T \vec{q}\|$$

for $\vec{p}_b = p(x_0)$.

Call $b = \text{argmax}_{b'} \vec{\phi}_{b'}$, so

$$\begin{aligned}
\|\vec{\phi} \circ \vec{p} / \vec{\phi}^T \vec{p} - \vec{\phi} \circ \vec{q} / \vec{\phi}^T \vec{q}\| &\leq \left(\frac{\vec{\phi}_b \vec{p}_b}{\vec{\phi}^T \vec{p}} - \frac{\vec{\phi}_b \vec{q}_b}{\vec{\phi}^T \vec{q}} \right)^2 + (1 - \vec{\phi}_b)^2 \sum_{b' \neq b} \left(\frac{\vec{p}_{b'}}{\vec{\phi}^T \vec{p}} - \frac{\vec{q}_{b'}}{\vec{\phi}^T \vec{q}} \right)^2 \\
&= \left(\frac{1}{1 + \sum_{b' \neq b} \frac{\vec{\phi}_{b'} \vec{p}_{b'}}{\vec{\phi}_b \vec{p}_b}} - \frac{1}{1 + \sum_{b' \neq b} \frac{\vec{\phi}_{b'} \vec{q}_{b'}}{\vec{\phi}_b \vec{q}_b}} \right)^2 \\
&\quad + \left(\frac{C}{c} \right)^2 B (1 - \phi_b)^2 \\
&\leq \left(1 - \frac{1}{1 + \frac{CB(1 - \phi_b)}{c}} \right)^2 \\
&\leq \left(\frac{CB}{c} \right)^2 (1 - \phi_b)^2 + \left(\frac{C}{c} \right)^2 B (1 - \phi_b)^2.
\end{aligned}$$

We've therefore bounded $\mathbb{E}_{t, x_0, x_t} L$ above by some constant times $\mathbb{E}_{t, x_t} \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} (1 - \max_b \vec{\phi}_b)^2$. Note without the hollow parameterization, we wouldn't have the $(1 - \max_b \vec{\phi}_b)^2$ term; we now show this becomes small very fast as $t \to 0$ (because x_0 becomes "obvious" from x_t), cancelling out the singularity.

Next note, calling $b = \text{argmin}_{b'} \|\text{emb}(b') - \vec{x}_t\|$,

$$(1 - \max_b \vec{\phi}_b)^2 = \left(1 - \frac{1}{1 + \sum_{b' \neq b} \exp(-\frac{1}{2(1-e^{-2\tau_t})^2}(\|\text{emb}(b') - \vec{x}_t\|^2 - \|\text{emb}(b) - \vec{x}_t\|^2))}\right)^2$$
$$\leq \sum_{b' \neq b} \exp\left(-\frac{1}{2(1 - e^{-2\tau_t})}(\|\text{emb}(b') - \vec{x}_t\|^2 - \|\text{emb}(b) - \vec{x}_t\|^2)\right),$$

which is only large if $\vec{x}_t$ is roughly equidistant to two potential x_0. Call $\epsilon = \min_{b \neq b'} \|\text{emb}(b) - \text{emb}(b')\|/4$, so, if $\min_{b'} \|\text{emb}(b') - \vec{x}_t\| < \epsilon$ then, by the triangle inequality

$$\begin{aligned}\|\text{emb}(b') - \vec{x}_t\|^2 - \|\text{emb}(b) - \vec{x}_t\|^2 \geq& (\|\text{emb}(b) - \text{emb}(b')\| - \|\text{emb}(b) - \vec{x}_t\|)^2 \\ &- \|\text{emb}(b) - \vec{x}_t\|^2 \\ =& \|\text{emb}(b) - \text{emb}(b')\| \\ &- 2\|\text{emb}(b) - \text{emb}(b')\|\|\text{emb}(b) - \vec{x}_t\| \\ \geq& 16\epsilon^2 - 8\epsilon^2 = 8\epsilon^2.\end{aligned}$$

Therefore, $\mathbb{E}_{t,x_t} \frac{\dot{\tau}_t e^{-2\tau_t}}{(1-e^{-2\tau_t})^2}(1 - \max_b \vec{\phi}_b)^2$ is bounded by

$$B\mathbb{E}_t \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} \left(\exp\left(-\frac{4\epsilon^2}{(1 - e^{-2\tau_t})}\right) + p(\min_{b'} \|\text{emb}(b') - \vec{x}_t\| \geq \epsilon)\right).$$

To deal with the first term, perform a change of variables $u = (1 - e^{-2\tau_t})^{-1}$, giving

$$\mathbb{E}_t \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} \exp\left(-\frac{4\epsilon^2}{(1 - e^{-2\tau_t})}\right) = \frac{1}{2}\int_0^\infty du \exp(-4\epsilon^2 u) < \infty.$$

For the second term, note

$$\begin{aligned}p(\min_{b'} \|\text{emb}(b') - \vec{x}_t\| \geq \epsilon) \leq& \sum_b p(x_0 = b)p(\|\mathcal{N}(0, (1 - e^{-2\tau_t})I_{r\times r})\| > \epsilon) \\ =& p(\chi_r^2/\epsilon^2 > 1/(1 - e^{-2\tau_t}))\end{aligned}$$

where χ_r^2 is a chi-squared distribution with r degrees of freedom. Finally, by the same change of variables u as above, we get

$$\mathbb{E}_t \frac{\dot{\tau}_t e^{-2\tau_t}}{(1 - e^{-2\tau_t})^2} p(\min_{b'} \|\text{emb}(b') - \vec{x}_t\| \geq \epsilon) = \frac{1}{2}\int_0^\infty dup(\chi_r^2/\epsilon^2 > u) = \mathbb{E}\chi_r^2/\epsilon^2 < \infty.$$

□

G.5 Proof of Wright-Fisher convergence

Formal definitions Define $\Delta^B \subset \mathbb{R}^B$ be the simplex, i.e. the set of non-negative vectors with components summing to 1. Let $(\vec{x}_t^\zeta)_{t=0}^1$ be a stochastic process on $(\frac{1}{\zeta}\mathbb{Z}^B) \cap \Delta^B$ with $\vec{x}_0^\zeta = \vec{x}_0$ evolving with respect to $\mathcal{L}^{\text{mut}} + \zeta\mathcal{L}^{\text{wf}}$ where

$$\mathcal{L}^{\text{wf}}_{\vec{x}^\zeta \to \vec{x}'^\zeta} = \frac{\zeta!}{\prod_b \zeta\vec{x}'^\zeta_b!} \prod_b (\vec{x}_b^\zeta)^{\zeta\vec{x}'^\zeta_b} = \text{Mult}(\zeta, \vec{x}^\zeta)(\zeta\vec{x}'^\zeta),$$

and, if $\vec{x}^\zeta, \vec{x}'^\zeta$ differ by one count $b \to b'$,

$$\mathcal{L}^{\text{mut}}_{\vec{x}^\zeta \to \vec{x}'^\zeta} = (\theta(\mathbb{1}\vec{\pi}^T - I))_{b,b'} = \theta\vec{\pi}_{b'}$$

otherwise it's 0. Let $(\vec{z}_t)_t$ be a continuous Wright-Fisher process, that is, $\vec{z}_t = \vec{x}_0$ and

$$d\vec{z}_t = \mathcal{L}^{\text{mut}T}\vec{z}_t dt + \text{diag}\left(\sqrt{\vec{z}_t}\right)\left(I - \sqrt{\vec{z}_t}\sqrt{\vec{z}_t}^T\right)d\vec{W}_t$$

where $(W_t)_t$ is a Brownian motion.

Convergence of the forward process We have convergence of the forward processes from previous literature.

Theorem G.4. *(Thm 1.1 Ethier and Kurtz [1986, Chapter 10]) Assume $\mathcal{L} = \psi \times (\mathbb{1}\vec{\pi}^T - I)$. In the topology of convergence of compact sets, $(\vec{x}_t^\zeta)_{t\in[0,1)} \rightsquigarrow (\vec{z}_t)_{t\in[0,1)}$.*

Note when $B = 2$, $(\vec{z}_t)_t$ is distributed as the Jacobi process described in Avdeyev et al. [2023]. One can easily check in $B > 2$, the stick-breaking procedure of Avdeyev et al. [2023] also leads to a continuous Wright-Fisher process.

Convergence of the ELBO Call $\vec{s}(\vec{v} \mid x_0) = \nabla \log p(z_t|x_0, t)|_{z_t=\vec{v}}$, and $\vec{s}(\vec{v} \mid \tilde{x}_0, t) = \sum_b \tilde{x}_{0,b}\vec{s}(\vec{v} \mid b)$.

Theorem G.5. *Call the ELBO in Alg. 2*

$$L(\vec{x}^\zeta, t, x_0, \tilde{x}_0) = \sum_{\vec{x}_t'^\zeta \neq \vec{x}_t^\zeta} (\zeta\mathcal{L}^{\mathrm{wf}}_{\vec{x}_t'^\zeta \to \vec{x}_t^\zeta} + \mathcal{L}^{\mathrm{mut}}_{\vec{x}_t'^\zeta \to \vec{x}_t^\zeta})\dot{\tau}_t\mathbb{D}\left(\frac{p(\vec{x}_t'^\zeta \mid x_0, t)}{p(\vec{x}_t^\zeta \mid x_0, t)}\middle|\middle|\frac{p(\vec{x}_t'^\zeta \mid \tilde{x}_0, t)}{p(\vec{x}_t^\zeta \mid \tilde{x}_0, t)}\right).$$

Then

$$L(\vec{v}, t, x_0, \tilde{x}_0) \to \frac{\dot{\tau}_t}{2}\|\vec{s}(\vec{v} \mid x_0, t) - \vec{s}(\vec{v} \mid \tilde{x}_0, t)\|^2_{\mathrm{diag}\vec{v} - \vec{v}\vec{v}^T}$$

Proof. (We provide an **informal** argument) For $\vec{x}_t'^\zeta \approx \vec{x}_t^\zeta$, we can approximate

$$\frac{p(\vec{x}_t'^\zeta \mid x_0, t)}{p(\vec{x}_t^\zeta \mid x_0, t)} \approx 1 + \vec{s}(\vec{v} \mid x_0, t)^T(\vec{x}_t'^\zeta - \vec{x}_t^\zeta).$$

Therefore,

$$\begin{aligned}\mathbb{D}\left(\frac{p(\vec{x}_t'^\zeta \mid x_0, t)}{p(\vec{x}_t^\zeta \mid x_0, t)}\middle|\middle|\frac{p(\vec{x}_t'^\zeta \mid \tilde{x}_0, t)}{p(\vec{x}_t^\zeta \mid \tilde{x}_0, t)}\right) \approx& \frac{1}{2}\left(\vec{s}(\vec{x}_t^\zeta \mid x_0, t) - \vec{s}(\vec{x}_t^\zeta \mid \tilde{x}_0, t))^T(\vec{x}_t'^\zeta - \vec{x}_t^\zeta)\right)^2 \\ =& \frac{1}{2}\|\vec{s}(\vec{x}_t^\zeta \mid x_0, t) - \vec{s}(\vec{x}_t^\zeta \mid \tilde{x}_0, t)\|^2_{(\vec{x}_t'^\zeta - \vec{x}_t^\zeta)(\vec{x}_t'^\zeta - \vec{x}_t^\zeta)^T}.\end{aligned}$$

Therefore,

$$L(\vec{x}^\zeta, x_0, \tilde{x}_0, t) \to \frac{\dot{\tau}_t}{2}\|\vec{s}(\vec{v} \mid x_0, t) - \vec{s}(\vec{v} \mid \tilde{x}_0, t)\|^2_\Sigma$$

where

$$\Sigma = \sum_{\vec{x}_t'^\zeta \neq \vec{x}_t^\zeta} (\zeta\mathcal{L}^{\mathrm{wf}}_{\vec{x}_t'^\zeta \to \vec{x}_t^\zeta} + \mathcal{L}^{\mathrm{mut}}_{\vec{x}_t'^\zeta \to \vec{x}_t^\zeta})(\vec{x}_t'^\zeta - \vec{x}_t^\zeta)(\vec{x}_t'^\zeta - \vec{x}_t^\zeta)^T.$$

Note, by the Stirling approximation, for $\vec{x}^\zeta \neq \vec{x}'^\zeta$

$$\mathcal{L}^{\mathrm{wf}}_{\vec{x}'^\zeta \to \vec{x}^\zeta} = (1 + O((\zeta \min_b \vec{x}_b^\zeta)^{-1}))(\prod_b \vec{x}_b^\zeta)^{-1/2}(2\pi\zeta)^{-(B-1)/2}e^{-\zeta\mathrm{KL}(\vec{x}^\zeta||\vec{x}'^\zeta)}.$$

Therefore, calling $Z(\vec{x}_t^\zeta) = \{\vec{v} \in \mathbb{Z}^B/\zeta \mid \sum_b \vec{v}_b = 0, \vec{x}_t^\zeta + \vec{v}_b \geq 0\ \forall b\}$, we can analyze the first term as

$$\begin{aligned}\sum_{v \in Z(\vec{x}_t^\zeta)} \zeta\mathcal{L}^{\mathrm{wf}}_{\vec{x}^\zeta + \vec{v} \to \vec{x}^\zeta}\vec{v}\vec{v}^T \approx& \sum_{v \in Z(\vec{x}_t^\zeta)} (\prod_b \vec{x}_b^\zeta)^{-1/2}(2\pi\zeta)^{-(B-3)/2}e^{-\zeta\mathrm{KL}(\vec{x}^\zeta||\vec{x}^\zeta + \vec{v})}\vec{v}\vec{v}^T \\ \approx& \sum_{v \in Z(\vec{x}_t^\zeta)} (\prod_b \vec{x}_b^\zeta)^{-1/2}(2\pi\zeta)^{-(B-1)/2}\zeta e^{-\zeta\frac{1}{2}\|\vec{v}\|^2_{\mathrm{diag}(\vec{x}_t^\zeta)^{-1}}}\vec{v}\vec{v}^T \\ \approx& \int_{\sum_b \vec{v}=0} d\vec{v}|(\mathrm{diag}(\vec{x}_t^\zeta) - \vec{x}_t^\zeta\vec{x}_t^{\zeta T})/\zeta|_+^{-1/2}(2\pi)^{-(B-1)/2}\zeta \\ &\times e^{-\zeta\frac{1}{2}\|\vec{v}\|^2_{(\mathrm{diag}(\vec{x}_t^\zeta) - \vec{x}_t^\zeta\vec{x}_t^{\zeta T})^{-1}}}\vec{v}\vec{v}^T \\ =& \mathrm{diag}(\vec{x}_t^\zeta) - \vec{x}_t^\zeta\vec{x}_t^{\zeta T}.\end{aligned}$$

On the other hand,

$$\sum_{v\in Z(\vec{x}_t^\zeta)} \mathcal{L}^{\text{mut}}_{\vec{x}^\zeta+\vec{v}\to\vec{x}^\zeta}\vec{v}\vec{v}^T = \zeta\times(B-1)\times O(1/\zeta^2) = o(1).$$

□

G.6 Wright-Fisher loss calculations

See the discussion above Prop. F.1 for definitions.

Proposition G.6. *(Proof of Prop. F.1)*

$$p(\vec{x}_t \mid x_0, t) = \text{Dirichlet}(\pi\psi)(\vec{x}_t)G_\psi(\tau_t, x_0, \vec{v}).$$

For $\vec{c}(\vec{v}) = \nabla \log \text{Dirichlet}(\pi\psi)(\vec{x}_t)$ *which does not depend on* x_0*,*

$$\vec{s} = \vec{s}(\vec{v} \mid x_0, t) = \vec{c}(\vec{v}) + \vec{x}_0 w(x_0)$$

where

$$w(x_0) = \frac{e^{-\psi\tau_t/2}(\psi+1)}{\pi(x_0)}\frac{F_\psi(\tau_t, x_0, \vec{v})}{G_\psi(\tau_t, x_0, \vec{v})}.$$

Proof. Say $m_t \sim A(\psi, \tau_t)$, so

$$\begin{aligned}
p(\vec{x}_t \mid x_0, t) =& E_{m_t}\text{Dirichlet}(\psi\pi + m_t x_0)(\vec{x}_t)\\
=& \prod_{b\neq x_0}\vec{x}_{t,b}^{\psi\pi_b-1}E_{m_t}\frac{\Gamma(\psi+m_t)}{\Gamma(\psi\pi_{x_0}+m_t)\prod_{b\neq x_0}\Gamma(\psi\pi_b)}\vec{x}_{t,x_0}^{\psi\pi_{x_0}+m_t-1}\\
=& \frac{\Gamma(\psi)}{\prod_{b\in\mathcal{B}}\Gamma(\psi\pi_b)}\prod_{b\in\mathcal{B}}\vec{x}_{t,b}^{\psi\pi_b-1}E_{m_t}\frac{\Gamma(\psi\pi(x_0))\Gamma(\psi+m_t)}{\Gamma(\psi)\Gamma(\psi\pi_{x_0}+m_t)}\vec{x}_{t,x_0}^{m_t}\\
=& \text{Dirichlet}(\psi\pi)(\vec{x}_t)E_{m_t}\frac{(\psi)_{(m_t)}}{(\psi\pi(x_0))_{(m_t)}}\vec{x}_{t,x_0}^{m_t}.
\end{aligned}$$

From Eqn. 5.2 of Tavaré [1984], we have

$$p(m_t = j) = \sum_{k=j}^{\infty} e^{-k(k+\psi-1)\tau_t/2}(-1)^k(-1)^j\frac{(2k+\psi-1)(j+\psi)_{(k-1)}}{j!(k-j)!}.$$

He wrote, in Eqn. A5,

$$\begin{aligned}
&\sum_{j=1}^{\infty}x^j p(m_t=j)\\
&=\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}(-1)^k(2k+\psi-1)\sum_{j=1}^{k}\frac{x^j}{j!}\frac{(j+\psi)_{(k-1)}}{(k-j)!(-1)^j}\\
&\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}(-1)^k(2k+\psi-1)\sum_{j=1}^{k}\frac{x^j}{j!}\frac{(\psi)_{(j+k-1)}(-k)_{(j)}}{k!\psi_{(j)}}\\
&\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}\frac{(-1)^k(2k+\psi-1)(\psi)_{(k-1)}}{k!}\sum_{j=1}^{k}\frac{x^j}{j!}\frac{(\psi+k-1)_{(j)}(-k)_{(j)}}{\psi_{(j)}}.
\end{aligned}$$

The last sum is then written as ${}_2F_1(-k, \psi+k-1; \psi; x) - 1$ for the hyper-geometric function ${}_2F_1$. A very simple extension gives us

$$\begin{aligned}
\sum_{j=1}^{\infty}\frac{(\psi)_{(j)}}{(\psi\pi_{x_0})_{(j)}}x^j p(m_t=j) = \sum_{k=1}^{\infty}&e^{-k(k+\psi-1)t/2}\frac{(-1)^k(2k+\psi-1)(\psi)_{(k-1)}}{k!}\\
&\times\left({}_2F_1(-k, \psi+k-1; \psi\pi_{x_0}; x) - 1\right).
\end{aligned}$$

Including the $j=0$ term, by Eqn 5.3 of Tavaré [1984], cancels out the -1 in the brackets above, so our expectation

$$
\begin{aligned}
E_{m_t}\frac{(\psi)_{(m_t)}}{(\psi\pi_{x_0})_{(m_t)}}\vec{x}_{t,x_0}^{m_t} =&1+\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}\frac{(-1)^k(2k+\psi-1)(\psi)_{(k-1)}}{k!}\\
&\times {}_2F_1(-k,\psi+k-1;\psi\pi_{x_0};\vec{x}_{t,x_0})\\
=&G_\psi(t,x_0,\vec{x}_t).
\end{aligned}
$$

Finally, using identities of the hypergeometric function,

$$
\begin{aligned}
\nabla_{\vec{x}_{t,x_0}}G_\psi(t,x_0,\vec{x}_t)=&\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}\frac{(-1)^k(2k+\psi-1)(\psi)_{(k-1)}}{k!}\frac{-k(\psi+k-1)}{\psi\pi_{x_0}}\\
&\times {}_2F_1(-k+1,\psi+k;\psi\pi_{x_0}+1;\vec{x}_{t,x_0})\\
=&\frac{1}{\psi\pi_{x_0}}\sum_{k=1}^{\infty}e^{-k(k+\psi-1)\tau_t/2}\frac{(-1)^{k-1}(2k+\psi-1)(\psi+k-1)(\psi)_{(k-1)}}{(k-1)!}\\
&\times {}_2F_1(-k+1,\psi+k;\psi\pi_{x_0}+1;\vec{x}_{t,x_0})\\
=&\frac{1}{\psi\pi_{x_0}}\sum_{k=0}^{\infty}e^{-(k+1)(k+\psi)\tau_t/2}\frac{(-1)^k(2k+\psi+1)(\psi+k)(\psi)_{(k)}}{k!}\\
&\times {}_2F_1(-k,\psi+k+1;\psi\pi_{x_0}+1;\vec{x}_{t,x_0})\\
=&\frac{e^{-\psi t/2}(\psi+1)}{\pi_{x_0}}\sum_{k=0}^{\infty}e^{-k(k+\psi+1)\tau_t/2}\frac{(-1)^k(\psi)_{(k)}}{k!}\frac{(2k+\psi+1)(\psi+k)}{(\psi+1)\psi}\\
&\times {}_2F_1(-k,\psi+k+1;\psi\pi_{x_0}+1;\vec{x}_{t,x_0})\\
=:&\frac{e^{-\psi t/2}(\psi+1)}{\pi_{x_0}}F_\psi(t,x_0,\vec{x}_t).
\end{aligned}
$$

□

G.7 Lemmas

Our first lemma establishes conditions for convergence of paths using standard techniques inspired by arguments used throughout Ethier and Kurtz [1986] or Bass [2011] for example.

Lemma G.7. *Say $(\vec{x}_t^\zeta)_{t\in(0,1)}$ are Markov processes on $\mathbb{R}^r$ for $\zeta=1,2,\dots$ and $(\vec{z}_t)_{t\in(0,1)}$ is another Markov process on $\mathbb{R}^r$. Say the following conditions are satisfied*

1. *(Convergence of marginals) $\vec{x}_t^\zeta \rightsquigarrow \vec{z}_t$ for each t.*
2. *(Local uniform convergence of conditionals) Conditional distributions exist such that for each $\vec{v}\in\mathbb{R}^r$, $s<t$, and bounded compactly supported measurable function f, there is an $\epsilon>0$, such that*
$$\sup_{\|\vec{w}-\vec{v}\|<\epsilon}|\mathbb{E}_{\vec{x}_t^\zeta|\vec{x}_s^\zeta=\vec{w}}f-\mathbb{E}_{\vec{z}_t|\vec{z}_s=\vec{w}}f|\to 0.$$
3. *(Tightness) For every $[a,b]\subset(0,1)$, there are $\beta,\theta,M>0$ such that for all $s,t\in[a,b]$, $\sup_{\zeta>M}\mathbb{E}\|\vec{x}_s^\zeta-\vec{x}_t^\zeta\|^\beta<C(s-t)^\theta$.*

Then, with the topology of convergence on compact sets[9]*, the paths converge in distribution*

$$(\vec{x}_t^\zeta)_{t\in(0,1)}\rightsquigarrow(\vec{z}_t)_{t\in(0,1)}.$$

Proof. Pick a compact set $[a,b]\subset(0,1)$. We show $(\vec{x}_t^\zeta)_{t\in[a,b]}\rightsquigarrow(\vec{z}_t)_{t\in[a,b]}$. Say $(\vec{x}_t^{\zeta_m})_{t\in[a,b]}$ is a subsequence which doesn't enter a neighbourhood of $(\vec{z}_t)_{t\in[a,b]}$; we'll now show a contradiction.

[9] This is a standard topology for these results. See for example Thm 1.1 of Ethier and Kurtz [1986, Chapter 10].

By Prokhorov's theorem, since it's tight by Assumption 3 and Thm. 8.8 of Ethier and Kurtz [1986, Chapter 3], it has a subsequence which converges to a process $(\vec{y}_t)_{t\in[a,b]}$. As we'll show below, for every set $a \leq t_1 < t_2 < \cdots < t_m \leq b$, $(\vec{y}_t)_{t\in\{t_i\}_{i=1}^m} = (\vec{z}_t)_{t\in\{t_i\}_{i=1}^m}$. This must mean $(\vec{y}_t)_t = (\vec{z}_t)_t$ by the Kolmogorov extension theorem, a contradiction.

What remains is to show, for $a \leq t_1 < t_2 < \cdots < t_m \leq b$, $(\vec{x}_t^\zeta)_{t\in\{t_i\}_{i=1}^m} \rightsquigarrow (\vec{z}_t)_{t\in\{t_i\}_{i=1}^m}$. It is sufficient to prove that for any $t_1 < \cdots < t_m$ and compactly supported continuous function on $\mathbb{R}^r$, h,

$$Eh(\vec{x}_1^\zeta, \ldots, \vec{x}_m^\zeta) \to Eh(\vec{z}_1, \ldots, \vec{z}_m). \tag{4}$$

By the Stone-Weierstrass theorem, each such h can be arbitrarily well approximated by product of m univariate functions, so it is sufficient to consider $h(\vec{z}_1, \ldots, \vec{z}_m) = \prod_{i=1}^m h_i(\vec{z}_i)$. Finally, by the Markov property,

$$\mathbb{E}h(\vec{x}_1^\zeta, \ldots, \vec{x}_m^\zeta) = \mathbb{E}_{\vec{x}_1^\zeta|\vec{x}_0^\zeta} h_1(\vec{x}_1^\zeta)\mathbb{E}_{\vec{x}_2^\zeta|\vec{x}_1^\zeta} h_2(\vec{x}_2^\zeta)\cdots\mathbb{E}_{\vec{x}_m^\zeta|\vec{x}_{m-1}^\zeta} h_m(\vec{x}_m^\zeta).$$

We can call $\tilde{h}_{m-1}^\zeta(\vec{x}_{m-1}^\zeta) = h_m(\vec{x}_{m-1}^\zeta)E_{\vec{x}_m^\zeta|\vec{x}_{m-1}^\zeta} h_m(\vec{x}_m^\zeta)$. By Assumption 2 $\tilde{h}_{m-1}^\zeta(\vec{x}_{m-1}^\zeta)$ converges uniformly to $h_m(\vec{x}_{m-1}^\zeta)E_{\vec{z}_m|\vec{z}_{m-1}=\vec{x}_{m-1}^\zeta} h_m(\vec{z}_m)$, a bounded function with compact support. Therefore, to prove Eqn. 4 it is sufficient to show the result replacing h with $h_1 \times h_2 \times \cdots \times h_{m-2} \times \tilde{h}_{m-1}$. By induction, we reach $h = \tilde{h}_1$for which we get Eqn. 4 by Assumption 1. □

Our next Lemma is a non-asymptotic bound on the convergence of multinomials to Normal distributions. It states that as long as $\zeta \to \infty$ and the probabilities don't get too low, we can bound the expectation of a function by $O(\zeta^{-1/2})$.

Lemma G.8. *Let $Y_\zeta \sim$ Mult$(\zeta, \vec{p})$ for probability vector $\vec{p} \in \mathbb{R}^B$ with $\min_i p_i \geq c > 0$. Call $Z_\zeta = \zeta^{-1/2}(Y_\zeta - \zeta p)$. For any bounded measurable function f,*

$$|\mathbb{E}f(Z_\zeta) - \mathbb{E}f(Z)| = o_{c,B,f}(1)$$

where $Z \sim \mathcal{N}(0, \mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^T)$ and the rate of decay $o_{c,B,f}(1)$ only depends on c, B, and f.

Proof. For every ϵ, pick a compactly supported C^∞ function g_ϵ such that $\|g_\epsilon - f\|_\infty < \epsilon/2$, so

$$|\mathbb{E}f(Z_\zeta) - \mathbb{E}f(Z)| = \epsilon + |\mathbb{E}g_\epsilon(Z_\zeta) - \mathbb{E}g_\epsilon(Z)| = \epsilon + o_{c,B,g_\epsilon}(1)$$

by Thm 1.3 of Gotze [1991]. □