PHASE-AWARE TRAINING SCHEDULE SIMPLIFIES LEARNING IN FLOW-BASED GENERATIVE MODELS

Santiago Aranguri *
Courant Institute of Mathematical Sciences
New York University
aranguri@nyu.edu

Francesco Insulla *
Institute of Computational and Mathematical Engineering
Stanford University
franinsu@stanford.edu

ABSTRACT

We analyze the training of a two-layer autoencoder used to parameterize a flow-based generative model for sampling from a high-dimensional Gaussian mixture. Previous work shows that the phase where the relative probability between the modes is learned disappears as the dimension goes to infinity without an appropriate time schedule. We introduce a time dilation that solves this problem. This enables us to characterize the learned velocity field, finding a first phase where the probability of each mode is learned and a second phase where the variance of each mode is learned. We find that the autoencoder representing the velocity field learns to simplify by estimating only the parameters relevant to each phase. Turning to real data, we propose a method that, for a given feature, finds intervals of time where training improves accuracy the most on that feature. Since practitioners take a uniform distribution over training times, our method enables more efficient training. We provide preliminary experiments validating this approach.

1 INTRODUCTION

In recent years, diffusion models have emerged as a powerful technique for learning to sample from high-dimensional distributions Sohl-Dickstein et al. (2015); Song et al. (2021); Song & Ermon (2020); Ho et al. (2020), especially in the context of generating images and recently also for text Lou et al. (2024). The idea lies in learning, from data samples, a velocity field that pushes noisy datapoints to clean datapoints. Despite the remarkable performance of these models, there remain several open questions, including understanding what makes a good noise schedule, which is the focus of this paper.

We consider the problem of training a neural network to learn the velocity field to generate samples from a two-mode Gaussian mixture (GM). This serves as a prototypical example to understand how diffusion models handle learning features at different scales, since the two-mode GM has two scales: the macroscopic scale of the probability of each mode, and the microscopic scale of the variance of each mode.

This problem was previously considered by Cui et al. (2024), but their analysis only handles the *balanced* two-mode GM (i.e. the probability of each mode is exactly $1/2$.) On the other hand, Biroli et al. (2024) assume access to the exact velocity field and find that the phase where the probability of each mode is learned disappears as the dimension of the problem grows.

In this work, we first introduce a noise schedule that makes the phase where this probability is learned not disappear as the dimension goes to infinity. This enables us to extend the analysis of Cui et al. (2024) to the two-mode GM without the balanced assumption. More precisely, our contributions are as follows.

- We give an asymptotic characterization of the learned velocity field for learning to generate the two-mode GM, finding a separation into two phases. We further show that $\Theta_d(1)$ samples are sufficient to learn the velocity field.

*Equal contribution

Published as a workshop paper at DeLTa Workshop (ICLR 2025)

- We show that the neural network representing the velocity field learns to simplify for each phase. In the first phase, it only concerns estimation of the probability of each mode, whereas in the second phase, it concerns estimation of the variance of each mode. This sheds light on the advantage of diffusion models over denoising autoencoders, since the sequential nature of diffusion models shown here allows them to decompose the complexity of the problem.
- We show that the phase transition separating the two phases can be detected from a discontinuity in the Mean Squared Error associated to the learning problem, which suggests a way to find these transitions for general data distributions.
- For real data, this analysis suggests that training more at the times associated with a feature improves accuracy on that feature. In fact, we propose a method that, given a feature, finds an interval of time where more training improves accuracy on that feature the most. We further validate this on the MNIST dataset. We provide the code for the experiments here.

2 Related Works

Phase transitions of generative models in high dimensions. Biroli & Mézard (2023) analyze the Curie-Weiss model and analytically characterize the speciation time, defined as the time after which the mode that the sample will belong to is determined. However, this work is only about sampling. We provide a result for sampling in Proposition 1. Building on this, we give results for learning (i.e. estimating the velocity field through a neural network) which is the main contribution of our paper.

Time-step complexity. Several results give convergence bounds detailing the required time-steps, score accuracy, and/or data distribution regularity to sample accurately. Benton et al. (2024) show that at most $O(d\log^2(1/\delta)/\epsilon^2)$ time steps are required to approximate a distribution corrupted with Gaussian noise of variance δ to within ϵ^2 KL divergence. Chen et al. (2023) study probability flow ODE and obtain $O(\sqrt{d})$ convergence guarantees with a smoothness assumption. An underlying assumption in all these works is that the score or velocity field is learned to a certain accuracy. In the present work, we address this problem in the special case of a Gaussian mixture.

Sample complexity for Gaussian mixtures. Cui et al. (2024) study the learning problem for the Gaussian mixture in high dimensions and demonstrate that $n = \Theta_d(1)$ samples are sufficient under strong symmetry assumptions. Gatmiry et al. (2024) show that quasi-polynomial ($O(d^{\text{poly}(\log(\frac{d+k}{\epsilon}))})$) sample and time complexity is enough for learning k-gaussian mixtures. The data distribution is more general than the one we consider, but we give a $\Theta_d(1)$ sample and time complexity.

3 Background

Data and flow-based generative model. Consider the two-mode Gaussian mixture (GM)

$$\rho = p\mathcal{N}(\mu, \sigma^2 \mathrm{Id}_d) + (1-p)\mathcal{N}(-\mu, \sigma^2 \mathrm{Id}_d) \tag{1}$$

where $p \in (0,1)$ and $\mu \in \mathbb{R}^d$ such that $\|\mu\|^2 = d$ and $\sigma = \Theta_d(1)$. A diffusion model for ρ starts with samples from a simple distribution (say a Gaussian) and sequentially denoises them to get samples from the data. More precisely, consider the stochastic interpolant

$$x_t = \alpha_t x_0 + \beta_t x_1 \tag{2}$$

where $x_0 \sim \mathcal{N}(0, \mathrm{Id}_d)$, $x_1 \sim \rho$, and $\alpha_t, \beta_t : [0,1] \to \mathbb{R}$, $\alpha_0 = 1 = \beta_1$, $\alpha_1 = 0 = \beta_0$. Stochastic interpolants are introduced in Albergo et al. (2023), and they prove that if X_t solves the probability flow ODE

$$\dot{X}_t = b_t(X_t) \qquad \text{with} \qquad b_t(x) = \mathbb{E}[\dot{x}_t | x_t = x] \tag{3}$$

with $X_0 \sim \mathcal{N}(0, \mathrm{Id}_d)$, we then have $X_t \overset{d}{=} x_t$ for $t \in [0,1]$ and hence $X_{t=1} \sim \rho$. We call X_t the flow-based generative model associated to the interpolant I_t.

Since ρ is a Gaussian mixture, the expression for the exact velocity field $b_t(x)$ from equation 3 can be computed exactly. Our goal is to understand how well a neural network can estimate this velocity

field through samples, in the large dimension $d \to \infty$ limit assuming low sample complexity for the data $n = \Theta_d(1)$.

Loss function. To fulfill our goal, we rewrite the velocity field as

$$b_t(x) = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t\right) f(x,t) + \frac{\dot{\alpha}_t}{\alpha_t} x, \tag{4}$$

where $f(x,t) = \mathbb{E}[x_1|x_t = x]$ is called the denoiser since it recovers the datapoint x_1 from a noisy version x_t. The denoiser is characterized as the minimizer of the loss (see Albergo et al. (2023))

$$\mathcal{R}[f] = \int_0^1 \mathbb{E}||f(x_t,t) - x_1||^2 dt. \tag{5}$$

In practice, however, we usually do not have access to the exact data distribution. So we assume we have a dataset $\mathcal{D} = \{x_1^\mu\}_{\mu=1}^n$ where $x_1^\mu \sim_{\text{iid}} \rho$. On the other hand, we have unlimited samples from $x_0 \sim \mathcal{N}(0, \text{Id}_d)$. Hence, to each data sample x_1^μ we can associate several noise samples $x_0^{\mu,\nu}$ with $\nu = 1, \cdots, k$. We then denote $x_t^{\mu,\nu} = \alpha_t x_0^{\mu,\nu} + \beta_t x_1^\mu$. Later in our analysis, we will assume infinitely many noise samples associated to each data sample, so that we can take expectation with respect to the noise distribution. We parameterize the denoiser with a single neural network for each t, which we denote as $f_{\theta_t}(x)$. We get then an empirical version of the loss in equation 5 $\hat{\mathcal{R}}(\{\theta_t\}_{t\in[0,1]}) = \int_0^1 \hat{\mathcal{R}}_t(\theta_t)dt$ where

$$\hat{\mathcal{R}}_t(\theta_t) = \sum_{\mu=1}^n \sum_{\nu=1}^k ||f_{\theta_t}(x_t^{\mu,\nu}) - x_1^\mu||^2 \tag{6}$$

Network architecture. We focus on the case where the neural network parameterizing the denoiser function $f(x,t)$ is a two-layer denoising autoencoder with a trainable skip connection as follows

$$f_{\theta_t}(x) = c_t x + u_t \tanh\left(\frac{w_t \cdot x}{\sqrt{d}} + b_t\right) \tag{7}$$

where $\theta_t = \{c_t, u_t, w_t, b_t\}$; $c_t, b_t \in \mathbb{R}$; and $u_t, w_t \in \mathbb{R}^d$. The structure of this denoising autoencoder is a particular case of the U-Net from Ronneberger et al. (2015) and is motivated by the exact denoiser which can be computed exactly since the data distribution is a Gaussian mixture

$$\mathbb{E}[x_1|x_t = x] = \frac{\beta_t \sigma^2}{\alpha_t^2 + \sigma^2\beta_t^2} x + \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\mu \tanh\left(\frac{\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\mu \cdot x + h\right) \tag{8}$$

where h is such that $e^h/(e^h + e^{-h}) = p$. (See Albergo et al. (2023), Appendix A for the proof.)

We add to the loss regularization terms for w_t and u_t, giving

$$\hat{\mathcal{R}}_t(\theta_t) = \sum_{\mu=1}^n \sum_{\nu=1}^k ||f_{\theta_t}(x_t^{\mu,\nu}) - x_1^\mu||^2 + \frac{\lambda}{2}||u_t||^2 + \frac{\ell}{2}||w_t||^2 \tag{9}$$

Denoting $\hat{\theta}_t$ the minimizer of this loss, we define

$$\hat{b}_t(x) = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t\right) f_{\hat{\theta}_t}(x) + \frac{\dot{\alpha}_t}{\alpha_t} x. \tag{10}$$

Using this velocity field, we then run the probability flow ODE

$$\dot{\hat{X}}_t = \hat{b}_t(\hat{X}_t); \quad \hat{X}_0 \sim \mathcal{N}(0, \text{Id}_d). \tag{11}$$

Our goal is to understand how close $\hat{X}_2$ is to a sample from the Gaussian mixture ρ.

Cui et al. (2024) consider the special case of tied weights $u_t = w_t$ and $b_t = 0$. This is enough to learn to sample from the balanced two-mode GM (i.e. $p = 1/2$) but fails at the two-mode GM for $p \neq 1/2$. This follows because x_0 has an even distribution and their choice of tied weights and no bias yields an odd velocity field which results in an even distribution for x_t. If the weights are untied and the bias is added, the analysis of Cui et al. (2024) still does not work to show that $\hat{X}_1$ has the

correct p for $p \neq 1/2$. This is because the gradients for w_t and b_t vanish as $d \to \infty$ unless special care is given to the small times where a phase transition related to learning the probability between the modes occurs, as will be explained next.

Separation into phases. Biroli et al. (2024) show that the generative model with the exact velocity field from equation 3 with $\alpha_t = \sqrt{1-t^2}$ and $\beta_t = t$ undergoes a phase transition at the speciation time $t_s = 1/\sqrt{d}$. The speciation time is defined as the time in the generation process after which the mode that the sample will belong to at the end of the process is determined. Their analysis can be extended to show that the speciation time is still $t_s = 1/\sqrt{d}$ if we instead have $\alpha_t = 1 - t$ and $\beta_t = t$ which are the choices in our paper. Since this result is only mentioned as motivation, we will not prove it.

The analysis of Cui et al. (2024) relies on taking the $d \to \infty$ limit and obtaining a limiting ODE. Since $t_s = 1/\sqrt{d}$ goes to zero as $d \to \infty$, their limiting ODE has a singularity at $t = 0$ and the possibility of learning the probability of each mode is lost. This is in essence why the analysis of Cui et al. (2024) can not capture the learning of p for $p \neq 1/2$.

We will dilate time so as to make the speciation time t_s not disappear as $d \to \infty$. More precisely, we define

$$\tau(t) = \begin{cases} \frac{\kappa t}{\sqrt{d}} & \text{if } t \in [0,1] \\ \frac{\kappa}{\sqrt{d}} + \left(1 - \frac{\kappa}{\sqrt{d}}\right)(t-1) & \text{if } t \in [1,2]. \end{cases} \tag{12}$$

This fulfills $\tau(0) = 0, \tau(1) = \kappa/\sqrt{d}$, and $\tau(2) = 1$. We prove next that the generative model from equation 3 with $\alpha_t = 1 - \tau_t$ and $\beta_t = \tau_t$ has two phases: for $t \in [0,1]$ the probability of each mode is estimated, and for $t \in [1,2]$ the variance of each mode is estimated.

Proposition 1. *Let X_t be the solution to the probability flow ODE from equation 3 with $\alpha_t = 1 - \tau_t$ and $\beta_t = \tau_t$ where τ_t is defined in equation 12. Then for $t \in [0,2]$ we have*

$$X_t - \frac{\mu \cdot X_t}{d}\mu \sim \mathcal{N}\left(0, \sigma_t^2 Id_{d-1}\right).$$

where σ_t is characterized below. We further have the following phases

- ***First phase**: For $t \in [0,1]$, we have $\lim_{d\to\infty} \sigma_t = 1$. In addition, $\nu_t = \lim_{d\to\infty} \frac{\mu \cdot X_t}{\sqrt{d}}$ fulfills*
$$\nu_1 \sim p\mathcal{N}(\kappa, 1) + (1-p)\mathcal{N}(-\kappa, 1).$$
- ***Second phase**: We have $\lim_{d\to\infty} \sigma_2 = \sigma$. In addition, $M_t = \lim_{d\to\infty} \frac{\mu \cdot X_t}{d}$ fulfills*
$$M_2 \sim p_\kappa \delta_1 + (1-p_\kappa)\delta_{-1}$$
where p^κ is such that $\lim_{\kappa\to\infty} p_\kappa = p$

See Appendix A for the proof of this Proposition. In Appendix E, we give a generalization of the time dilation formula in equation 12 for a Gaussian mixture with more than two modes.

Without the time dilation, we cannot capture the learning of p for $p \neq 1/2$ because the first phase (where this parameter is learned) disappears as $d \to \infty$. The time dilation will allow us to analyze the phase where p is learned in the $d \to \infty$ limit and hence show that $\hat{X}_2$ recovers p.

We show this in two steps. In Section 4, we characterize the learned parameters of the velocity field in terms of a few projections, called the overlaps. Then, in Section 5, we combine these characterizations with Proposition 1 to show that $\hat{X}_2$ recovers the parameters p and σ^2 of the two-mode Gaussian mixture ρ under appropriate limits.

4 LEARNING

In this section, we will characterize $\hat{\theta}_t$, the minimizer of the loss from equation 9, which is used to parameterize the velocity field that yields $\hat{X}_t$ (see equation 11.) We take $\alpha_t = 1 - \tau_t$ and $\beta_t = \tau_t$ and analyze $\hat{\theta}_t$ in the $d \to \infty$ limit. We first analyze the times $t \in [0,1]$ and then $t \in [1,2]$.

4.1 FIRST PHASE

The interpolant from equation 2 in the first phase reads $x_t^\mu = \left(1 - \frac{\kappa t}{\sqrt{d}}\right) x_0^\mu + \frac{\kappa t}{\sqrt{d}} x_1^\mu$ where $t \in [0,1]$. To characterize $\hat{\theta}_t = \{c_t, u_t, w_t, b_t\}$, we introduce the following overlaps (dropping the dependence on t for notational simplicity.)

$$p_\eta^\mu = \frac{z^\mu \cdot w}{d} \quad \omega = \frac{\mu \cdot w}{d} \quad r = \frac{\|w\|^2}{d} \quad q_\xi^\mu = \frac{x_0^\mu \cdot u}{d} \quad q_\eta^\mu = \frac{z^\mu \cdot u}{d} \quad m = \frac{\mu \cdot u}{d} \quad q = \frac{\|u\|^2}{d}. \tag{13}$$

We now give equations for the overlaps in the asymptotic $d \to \infty$ limit.

Result 1 (Sharp Characterization of Parameters in First Phase)**.** *For any $t \in [0,1]$, the overlaps associated to $\hat{\theta}_t$, the minimizer of the loss from equation 9, satisfy the following in the $d \to \infty$ limit*

$$\begin{aligned}
m &= \frac{n\overline{\phi s}}{\lambda + n\overline{\phi^2}} & c &= q_\xi = p_\eta = 0 \\
q_\eta &= \frac{\sigma}{n} m & q &= m^2 + n q_\eta^2 \\
& & r &= \omega^2 \\
(\lambda + n\overline{\phi^2})\overline{\phi' s} &= n\overline{\phi s}\,\overline{\phi\phi'} \\
\hat{r}(\lambda + n\overline{\phi^2})^2 &= -n((\lambda + n\overline{\phi^2})(\sigma^2 + n)\overline{\phi'' s}\,\overline{\phi s} - n(\sigma^2 + n)\overline{\phi s}^2\,\overline{(\phi\phi')'}) \\
\omega(\ell + \hat{r})(\lambda + n\overline{\phi^2})^2 &= (n\kappa t)(\sigma^2 + n)((\lambda + n\overline{\phi^2})(\overline{\phi'}\,\overline{\phi s}) - n\overline{\phi s}^2\overline{\phi' \phi s})
\end{aligned}$$

Here and in what follows, we denote

$$\overline{y} = \frac{1}{nk} \sum_{\mu=1}^{n} \sum_{\nu=1}^{k} \mathbb{E}_{z^{\mu,\nu}}[y^{\mu,\nu}] = \overline{p}\mathbb{E}_{z^{\mu,\nu}}[y^{\mu,\nu}|s^\mu = 1] + (1 - \overline{p})\mathbb{E}_{z^{\mu,\nu}}[y^{\mu,\nu}|s^\mu = -1].$$

See Appendix B.1 for a heuristic derivation of this result, at the level of rigor of theoretical physics. We next show that the equations for the overlaps simplify in the $n \to \infty$ limit.

Corollary 1 (Parameters given infinite samples)**.** *For any $t \in [0,1]$, taking $d \to \infty$ and then $n \to \infty$ gives the following overlaps*

$$\begin{aligned}
\tanh(b) &= 2\left(p - \tfrac{1}{2}\right), & m &= 1, \\
c = q_\xi = q_\eta = p_\eta &= 0, & \omega &= \kappa t.
\end{aligned}$$

See Appendix B.1.1 for the derivation. Note that the overlaps in the $n \to \infty$ limit do not contain any information about σ^2, showing that the estimation of σ^2 happens completely in the second phase.

We now turn to the Mean Squared Error. Define the scaled train and test MSE of the denoiser as

$$\text{mse}_{\text{train}} = \frac{1}{dnk} \sum_{\mu=1}^{n} \sum_{\nu=1}^{k} ||f_{\theta_t}(x_t^{\mu,\nu}) - x_1^\mu||^2 \qquad \text{mse}_{\text{test}} = \frac{1}{d}\mathbb{E}\left[\|f_{\hat{\theta}_t}(x_t) - x_1\|^2\right].$$

Using the above results we characterize the MSE

Corollary 2. *In the limit of $d \to \infty$,*

$$\begin{aligned}
\textit{mse}_{\textit{train}} &= 1 + \sigma^2 + c^2 + q\overline{\phi^2} - 2\overline{s\phi}(m + \sigma q_\eta - c q_\xi) \\
\textit{mse}_{\textit{test}} &= 1 + \sigma^2 + c^2 + q\overline{\phi^2} - 2\overline{s\phi} m
\end{aligned}$$

For $n \to \infty$, we get $\textit{mse}_{\textit{train}} = \textit{mse}_{\textit{test}} = \sigma^2 + (1 - \overline{\phi s})$.

4.2 SECOND PHASE

We now consider times $t \in [1,2]$ which means we have

$$x_t^\mu = (2 - t)\left(1 - \frac{\kappa}{\sqrt{d}}\right) x_0^\mu + \left(\frac{\kappa}{\sqrt{d}} + \left(1 - \frac{\kappa}{\sqrt{d}}\right)(t - 1)\right) x_1^\mu.$$

Using the same definitions of overlaps as for the first phase, we find closed-form equations for the overlaps in the asymptotic $d \to \infty$ limit, and again find the limit as $n \to \infty$ for the overlaps. See Result 3 in Appendix B.2 for the exact statement and derivation.

Corollary 3 (Parameters given inifite samples)**.** *For any $t \in [1, 2]$, taking $d \to \infty$ and then $n \to \infty$ gives the following overlaps, with $\tau = t - 1$,*

$$c = \frac{\tau\sigma^2}{1 + (\sigma^2 - 1)\tau^2} \qquad q_\xi = q_\eta = 0 \qquad m = 1 - c\tau$$

In contrast to the first phase, the parameter p does not appear in the overlaps whereas now σ^2 does. Hence, combining Corollaries 1 and 3 shows that the separation into phases can be learned by the generative model. We also obtain the MSE for the second phase

Corollary 4. *In the limit of $d \to \infty$, we have*

$$\begin{aligned} mse_{train} &= (1 + \sigma^2)(1 - c\tau)^2 + c^2(1 - \tau)^2 + q - 2(1 - c\tau)(\sigma q_\eta + m) + 2c(1 - \tau)q_\xi \\ mse_{test} &= (1 + \sigma^2)(1 - c\tau)^2 + c^2(1 - \tau)^2 + q - 2(1 - c\tau)m \end{aligned}$$

For $n \to \infty$, we get with $\tau = t - 1$ that $mse_{train} = mse_{test} = \sigma^2(1 - c\tau)^2 + c^2(1 - \tau)^2$

In Appendix B.2 we show that combining Corollaries 2 and 4 gives

Corollary 5. *Taking $d \to \infty$ and then $n \to \infty$ gives*

$$mse_{test} = \begin{cases} \sigma^2 + 4p(1 - p) & \text{if } t = 0 \\ \sigma^2 + (1 - \overline{\phi^2}) & \text{if } t \in (0, 1) \\ \sigma^2 & \text{if } t = 1^+ \\ 0 & \text{if } t = 2 \end{cases}$$

If we had not dilated time, in the limit of $d \to \infty$, the mse_{test} would have a jump from $\sigma^2 + 4p(1-p)$ at $t = 0$ to σ^2 at $t = 0^+$. By dilating time, we make a transition between these two values with $\text{mse}_{\text{test}} = \sigma^2 + (1 - \overline{\phi^2})$ for $0 < t < 1$, where $1 - \overline{\phi^2} = 4p(1 - p)$ for $t = 0$ and $1 - \overline{\phi^2}$ goes to 0 exponentially fast as κ grows when $t = 1$. Further, $\overline{\phi^2}$ is continuous in $t \in [0, 1]$. Hence by dilating near the phase transition, we decreased the jump discontinuity of the *mse*.

Remarkably, for generating samples from a general data distribution, this result suggests that the jumps in the *mse* could correspond to phase transitions. Further, this phase transitions could be resolved by dilating near the jump in the *mse*. We leave the study of this conjecture to future work.

5 Generation

Having characterized the parameters $\hat{\theta}_t$, we now show that $\hat{X}_2$ has the right parameters p and σ^2 from the data distribution ρ. Let X_t be the solution to the ODE from equation 3 using the exact denoiser from equation 8. Assume X_t and $\hat{X}_t$ have a shared initial condition $X_{t=0} = \hat{X}_{t=0} \sim \mathcal{N}(0, \mathrm{Id}_d)$. Then $X_t - \hat{X}_t$ fulfills an ODE with initial condition 0 whose velocity field is in the span of u_t and μ.

Result 1 gives that in the first phase $q = m^2 + nq_\eta^2$. This can be explicitly stated as

$$\lim_{d\to\infty} \frac{\|u\|^2}{d} = \lim_{d\to\infty} \left(\frac{\mu \cdot u}{d}\right)^2 + \left(\frac{\eta \cdot u}{d}\right)^2$$

where $\eta = \sigma \sum_{\mu=1}^n z^\mu$. This means that u_t is asymptotically contained in $\text{span}(\mu, \eta)$, in the sense that the projection to the complement of $\text{span}(\mu, \eta)$ has asymptotically vanishing norm, for $t \in [0, 1]$. Similarly, from Result 3, we get $q = m^2 + nq_\xi^2 + nq_\eta^2$, which means that u_t is asymptotically contained in $\text{span}(\mu, \eta, \xi)$ for $t \in [1, 2]$ where $\xi = \sum_\mu s^\mu x_0^\mu$. This means that to show that X_t is close to $\hat{X}_t$, it suffices to bound the projections of $X_t - \hat{X}_t$ onto μ, η, and ξ. In fact, we have the following result (see Appendix C)

Result 2. *Let X_t be the solution of the probability flow ODE from equation 3 using the exact denoiser from equation 8. Let $\hat{X}_t$ be the solution using the learned denoiser. Assume $X_{t=0} = \hat{X}_{t=0} \sim \mathcal{N}(0, Id_d)$. Then for $w \in span(\mu, \eta, \xi)$, with $\|w\|_2 = 1$, we have*

$$\lim_{d\to\infty} \frac{w \cdot (X_2 - \hat{X}_2)}{\sqrt{d}} = O\left(\frac{1}{n}\right).$$

For $w \in span(\mu, \eta, \xi)^{\perp}$, with $\|w\|_2 = 1$, we have

$$\lim_{d\to\infty} \frac{w \cdot (X_2 - \hat{X}_2)}{\sqrt{d}} = 0.$$

Corollary 6 (Parameters p and σ^2 are estimated correctly)**.** *Let $\hat{X}_t$ be the solution of the probability flow ODE from equation 3 using the learned denoiser, starting from $\hat{X}_0 \sim \mathcal{N}(0, Id_d)$. We have*

$$\lim_{\kappa\to\infty} \lim_{n\to\infty} \lim_{d\to\infty} \frac{\mu \cdot \hat{X}_2}{d} \sim p\delta_1 + (1-p)\delta_{-1}.$$

For $w \perp \mu$, with $\|w\|_2 = 1$, we have

$$\lim_{n\to\infty} \lim_{d\to\infty} \frac{w \cdot \hat{X}_2}{\sqrt{d}} \sim \mathcal{N}(0, \sigma^2).$$

We conclude that the distribution generated using the learned denoiser captures both p and σ^2.

6 EXPERIMENTS

6.1 VERIFICATION THAT PARAMETER p IS CAPTURED

To demonstrate the difference between the time dilated and non-dilated interpolants in practice we construct the following simple experiment. We run Gradient Descent with the Adam optimizer Kingma & Ba (2015) to learn the parameters w_t, c_t, u_t, b_t in equation 7 both for $\alpha_t = 1 - t, \beta_t = t$ and the dilated version $\alpha_t = 1 - \tau_t, \beta_t = \tau_t$. The results are shown in Figure 1 (left panel) and suggest time-dilation is required to estimate the probability of each mode. The code for this experiment is available here.

6.2 TRAINING A GIVEN FEATURE ON REAL DATA: MNIST

Recall that in the background we mentioned that the analysis of Biroli et al. (2024) shows that taking $\alpha_t = 1 - t$ and $\beta_t = t$ without any time-dilation gives an speciation time $t_s = 1/\sqrt{d}$. This then means that probability of each mode (given by p) can not be captured as $d \to \infty$. Our analysis then shows that if we dilate time by stretching the interval $[0, \kappa/\sqrt{d}]$ to $[0, 1]$ and the interval $[\kappa/\sqrt{d}, 1]$ to $[1, 2]$, then we get accurate estimation of p.

When training diffusion models in practice, we first sample a batch of times $t_1, \cdots, t_k$ uniformly. We then draw $x_0^\mu \sim \mathcal{N}(0, \mathrm{Id}_d)$, x_1^μ from our data distribution, and form a noisy sample $x_{t^\mu}^\mu = (1-t^\mu)x_0^\mu + t^\mu x_1^\mu$ for $\mu = 1, \cdots, k$. We finally train on the loss $\hat{\mathcal{R}}(\theta) = \sum_{\mu=1}^k ||f_\theta(x_{t^\mu}^\mu, t^\mu) - x_1^\mu||^2$ where we took time as a parameter of the network as it is usually done in practice, as opposed to having a separate network for each time t.

The insight of our analysis is that instead of taking the batch of times uniformly, we can sample more times near the phase transition associated to a given feature, and in this way improve accuracy on that feature.

For a given feature, we can find the times where that feature is learned using the U-Turn method (Sclocchi et al. (2024), Biroli et al. (2024)). Consider a dataset where each sample corresponds to exactly one of finitely many classes. Examples of this are samples of the GM which correspond to one of two modes, or samples of MNIST which correspond to one of ten digits. The U-Turn then consists of starting with a sample from the data, run a backward diffusion model from time $t = 1$

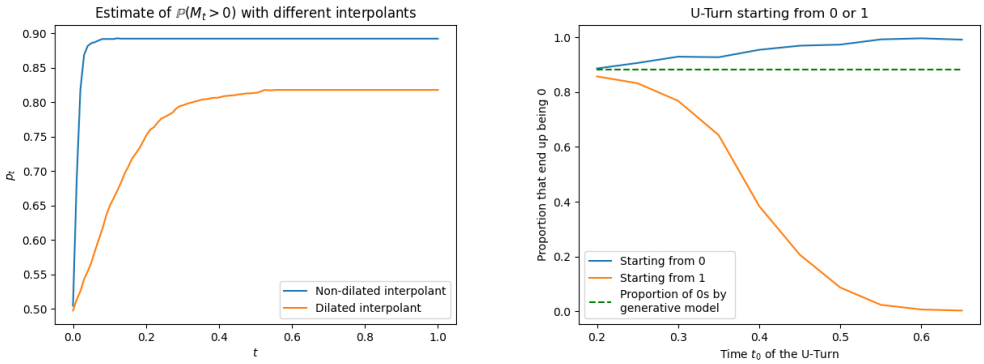


Figure 1: **(Left panel)** We learn the parameters from equation 7 for different choices of interpolant. In all experiments, we take 100 discretization points, train for 5000 epochs, with $n = 128, d = 5000$, and $p = .8$. We then run the probability flow ODE with the learned parameters for $K = 2000$ realizations and estimate $\mathbb{P}(M_t > 0) = p$ with $M_t = \mu \cdot X_t/d$. For the non-dilated interpolant in blue, we use $\alpha_t = 1 - t, \beta_t = t$. We predict the speciation to happen near $t = 1/\sqrt{5000} \approx .014$ as confirmed by the experiment. For the dilated interpolant in orange, we use $\alpha_t = 1 - \tau_t, \beta_t = \tau_t, \kappa = 4$. **(Right panel)** For $t_0 \in [0.2, 0.65]$, we plot the proportion of 0s that we get by doing the U-Turn at time t_0 starting from either 0 or 1 at time $t = 1$. On dashed green, we plot $y = .882$ which is the estimated proportion of 0s that the diffusion model generates starting from noise.

to $t = t_0$, which noises the sample, and then run the forward diffusion model from time $t = t_0$ to $t = 1$ with noise independent from the backward run.

We are then interested in the probability that the sample before the backward and forward passes belongs to the same class as the sample after them. For $t_0 \approx 1$, this probability is close to 1. For $t_0 \approx 0$, this probability is close to the underlying probability of the diffusion model generating a sample of the given class. By running this for different t_0, we can find at what times it is decided to what class the samples belong to. Having found those times, our goal is to have a model that generates samples for each class according to the probabilities that they appear in the dataset. We can then improve the accuracy of the model on this by training on these times.

As a simple example, we train a U-Net (see Appendix D for details) to parameterize the Variance Preserving SDE from Song et al. (2021) to generate either the 0 or 1 digits from MNIST. The dataset we train on consists of 20% 1 digits and 80% 0 digits. We then measure how well is this model in generating samples that represent this asymmetry. The model is trained on approximately 7400 samples for 9 epochs, by sampling times in $[0, 1]$ uniformly as described in the beginning of this section. We then generate 18500 new samples running this model using 1000 discretization steps. Among the 18500 generated samples, 88.2% are digits 0. (For determining this, we used a discriminator with 99.2% accuracy on MNIST, see Appendix D for details.)

We then test our proposed method. First, we determine at what time the digit that the sample represents is decided. We do this with the U-Turn method described above. Note that to do this, we use the model that we already trained. The results are in Figure 1 (right panel). We find that the times important for deciding the digit are early in the generation for $t \in [0.2, 0.6]$ and mostly concentrated on $t \in [0.3, 0.5]$.

We now train from scratch a model on 7400 samples for 9 epochs as before, except that we do not sample the times uniformly. We instead sample times with probability $1/2$ uniformly in the interval $[0.3, 0.5]$ and with probability $1/2$ uniformly outside that interval. We then generate 18500 new samples with this new model using 1000 discretization steps, and find that 81.0% are 0s. We similarly consider sampling times with probability $1/2$ uniformly in the interval $[0.2, 0.6]$ and with probability $1/2$ outside that interval, generate samples, and find that 81.1% are 0s. This validates our hypothesis in the simple case of MNIST. Although our theoretical analysis is for the probability flow ODE on the two-mode GM data distribution, this example on MNIST shows that the ideas developed here can be useful to the SDE generative models used in practice for real data.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. URL https://arxiv.org/abs/2303.08797.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization, 2024. URL https://arxiv.org/abs/2308.03686.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(9):093402, September 2023. ISSN 1742-5468. doi: 10.1088/1742-5468/acf8ba. URL http://dx.doi.org/10.1088/1742-5468/acf8ba.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024. URL https://arxiv.org/abs/2402.18491.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast, 2023. URL https://arxiv.org/abs/2305.11798.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity, 2024. URL https://arxiv.org/abs/2310.03575.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models, 2024. URL https://arxiv.org/abs/2404.18869.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015. URL https://arxiv.org/abs/1412.6980.

Farley Knight. MNIST Digit Classification Model. https://huggingface.co/farleyknight/mnist-digit-classification-2022-09-04, 2022. Accessed: [September 28, 2024].

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data, 2024. URL https://arxiv.org/abs/2402.16991.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

A PROOF OF PROPOSITION 1

To prove Proposition 1, we will use the following three Lemmas that follow directly from Albergo et al. (2023) (Appendix A)

Lemma 1. *Let $a \sim \rho$ and $z \sim \mathcal{N}(0, Id_d)$. The law of the interpolant $I_t = \alpha_t z + \beta_t a$ coincides with the law of the solution of the probability flow ODE*

$$\dot{X}_t = \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} X_t + \frac{\alpha_t(\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t)}{\alpha_t^2 + \sigma^2 \beta_t^2} \mu \tanh\left(h + \frac{\beta_t \mu \cdot X_t}{\alpha_t^2 + \sigma^2 \beta_t^2}\right), \quad X_0 \sim \mathcal{N}(0, Id_d) \tag{14}$$

where h is such that $e^h/(e^h + e^{-h}) = p$.

Lemma 2. *Let $a \sim p\mathcal{N}(\kappa, 1) + (1-p)\mathcal{N}(-\kappa, 1)$ and $z \sim \mathcal{N}(0, 1)$. The law of the interpolant $I_t = \alpha_t z + \beta_t a$ coincides with the law of the solution of the probability flow ODE. In*

$$\dot{X}_t = \frac{\alpha_t \dot{\alpha}_t + \beta_t \dot{\beta}_t}{\alpha_t^2 + \beta_t^2} X_t + \frac{\kappa \alpha_t(\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t)}{\alpha_t^2 + \beta_t^2} \tanh\left(h + \frac{\kappa \beta_t X_t}{\alpha_t^2 + \beta_t^2}\right), \quad X_0 \sim \mathcal{N}(0, 1) \tag{15}$$

where h is such that $e^h/(e^h + e^{-h}) = p$.

Lemma 3. *Let $a \sim \mathcal{N}(\pm\mu, \sigma^2 Id_d)$ and $z \sim \mathcal{N}(0, Id_d)$. The law of the interpolant $I_t = \alpha_t z + \beta_t a$ coincides with the law of the solution of the probability flow ODE*

$$\dot{X}_t = \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} X_t \pm \frac{\alpha_t(\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t)}{\alpha_t^2 + \sigma^2 \beta_t^2} \mu, \quad X_0 \sim \mathcal{N}(0, Id_d). \tag{16}$$

Proof of Proposition 1. **First phase.** We have $\tau_t = \frac{\kappa t}{\sqrt{d}}$ since $t \in [0, 1]$. Plugging in $\alpha_t = 1 - \tau_t$ and $\beta_t = \tau_t$ into the velocity field from Lemma 1 yields

$$\dot{X}_t = \frac{\kappa}{\sqrt{d}} \left(-X_t + \mu \tanh\left(h + \kappa t \frac{\mu \cdot X_t}{\sqrt{d}}\right)\right) + O\left(\frac{1}{d}\right). \tag{17}$$

We then have, with $\nu_t = \mu \cdot X_t / \sqrt{d}$,

$$\dot{\nu}_t = \kappa \tanh\left(h + \kappa t \nu_t\right) + O\left(\frac{1}{\sqrt{d}}\right). \tag{18}$$

Taking $d \to \infty$ yields the limiting ODE for ν_t. From Lemma 2, we get that this the 1-dimensional velocity field associated to the interpolant $I_t = \sqrt{1-t^2} z + ta$ that transports $z \sim \mathcal{N}(0, 1)$ at $t = 0$ to $a \sim p\mathcal{N}(\kappa, 1) + (1-p)\mathcal{N}(-\kappa, 1)$ at $t = 1$.

Let $X_t^\perp = X_t - \frac{\mu \cdot X_t}{d} \mu$. We have from equation 17

$$\dot{X}_t^\perp = -\frac{\kappa}{\sqrt{d}} X_t^\perp. \tag{19}$$

Since this is a linear ODE with initial condition Gaussian, we have

$$\dot{X}_t^\perp \sim \mathcal{N}\left(0, \sigma_t^2 \mathrm{Id}_{d-1}\right). \tag{20}$$

Further, equation 19 gives $\dot{X}_t^\perp = O(1/\sqrt{d})$ meaning that for $t \in [0, 1]$

$$\lim_{d \to \infty} \sigma_t = 1. \tag{21}$$

Second phase. For $t \in [1, 2]$, we have $\tau_t = \left(1 - \frac{\kappa}{\sqrt{d}}\right)(2t - 1) + \frac{\kappa}{\sqrt{d}}$. Again using Lemma 1, we get

$$\dot{X}_t = \frac{-(2-t) + \sigma^2(t-1)}{(2-t)^2 + \sigma^2(t-1)^2} X_t + \frac{(2-t) \tanh\left(h + \frac{(t-1)\mu \cdot X_t + \kappa \frac{\mu \cdot X_t}{\sqrt{d}}}{(2-t)^2 + \sigma^2(t-1)^2}\right)}{(2-t)^2 + \sigma^2(t-1)^2} \mu + O\left(\frac{1}{\sqrt{d}}\right). \tag{22}$$

Writing $\nu_t = \frac{\mu \cdot X_t}{\sqrt{d}}$, this implies

$$\dot{\nu}_t = \frac{-(2-t)+\sigma^2(t-1)}{(2-t)^2+\sigma^2(t-1)^2}\nu_t + \frac{(2-t)\sqrt{d}\tanh\left(h + \frac{(t-1)\sqrt{d}\nu_t + \kappa\nu_t}{(2-t)^2+\sigma^2(t-1)^2}\right)}{(2-t)^2+\sigma^2(t-1)^2} + O_d(1). \tag{23}$$

Let us calculate the initial condition for ν_t at $t = 1$. Write $a = sm + z$ where $p = \mathbb{P}(s = 1) = 1 - \mathbb{P}(s = -1)$ and $z \sim \mathcal{N}(0, \mathrm{Id}_d)$. Then

$$\frac{\mu \cdot I_{t=1}}{\sqrt{d}} \overset{(d)}{=} Z + \kappa s + O\left(\frac{1}{\sqrt{d}}\right)$$

where $Z \sim \mathcal{N}(0,1)$. This means that for κ large enough, then $|h| < \left|\frac{(t-1)\sqrt{d}\nu_t + \kappa\nu_t}{(2-t)^2+\sigma^2(t-1)^2}\right|$ with high probability. This implies ν_t will not change sign during its trajectory, since whenever $\nu_t = o(\sqrt{d})$, the $\tanh$ term will dominate in equation 23. Hence, the following approximation is valid

$$\tanh\left(h + \frac{(2t-1)\sqrt{d}\nu_t + \kappa\nu_t}{1+(\sigma^2-1)(2t-1)^2}\right) = \tanh\left(\sqrt{d}\nu_t\right) = \mathrm{sgn}(\nu_t) \tag{24}$$

We then use this approximation in the ODEs for X_t to get

$$\dot{X}_t = \frac{-(2-t)+\sigma^2(t-1)}{(2-t)^2+\sigma^2(t-1)^2}X_t + \frac{(2-t)\mathrm{sgn}(\nu_t)}{(2-t)^2+\sigma^2(t-1)^2}\mu + O\left(\frac{1}{\sqrt{d}}\right). \tag{25}$$

Let $M_t = \mu \cdot X_t/d$. We get the induced equation

$$\dot{M}_t = \frac{-(2-t)+\sigma^2(t-1)}{(2-t)^2+\sigma^2(t-1)^2}M_t + \frac{(2-t)\mathrm{sgn}(M_t)}{(2-t)^2+\sigma^2(t-1)^2} + O\left(\frac{1}{\sqrt{d}}\right). \tag{26}$$

From the analysis of the first phase we had

$$\nu_1 \sim p\mathcal{N}(\kappa, 1) + (1-p)\mathcal{N}(-\kappa, 1). \tag{27}$$

We argued above that the sign of ν_t will be preserved for $t \in [1, 2]$ with probability going to 1 as κ tends to ∞. This means that

$$M_2 = p^\kappa \delta_1 + (1-p^\kappa)\delta_{-1} \tag{28}$$

where p^κ is such that $\lim_{\kappa\to\infty} p^\kappa = p$.

As in the first phase, we let $X_t^\perp = X_t - \frac{\mu \cdot X_t}{d}\mu$. We have from equation 25 that

$$\dot{X}_t^\perp = \frac{-(2-t)+\sigma^2(t-1)}{(2-t)^2+\sigma^2(t-1)^2}X_t^\perp + O\left(\frac{1}{\sqrt{d}}\right). \tag{29}$$

Since this is a linear ODE with initial condition Gaussian, we have

$$\dot{X}_t^\perp \sim \mathcal{N}\left(0, \sigma_t^2 \mathrm{Id}_{d-1}\right). \tag{30}$$

Under the change of variables $t(s) = s + 1$ we get that the ODE becomes

$$\dot{X}_s^\perp = \frac{-(1-s)+\sigma^2 s}{(1-s)^2+\sigma^2 s^2}X_s^\perp. \tag{31}$$

By taking one coordinate $i \in \{1, \cdots, d-1\}$ of $X_s^\perp$ we get from Lemma 3 that this is the velocity field associated with the interpolant $I_s = \sqrt{1-s^2}z + sa$ where $z \sim \mathcal{N}(0,1)$ is transported to $a \sim \mathcal{N}(0, \sigma^2)$ as desired. □

B Derivations of learning results

B.1 Derivation of First Phase

Let $t \in [0, 1]$ so that

$$x_t^\mu = \left(1 - \frac{\kappa t}{\sqrt{d}}\right)x_0^\mu + \frac{\kappa t}{\sqrt{d}}x_1^\mu$$

Consider a denoiser parametrized as

$$f(x) = cx + u \tanh\left(b + \frac{w \cdot x}{\sqrt{d}}\right)$$

We introduce the following overlaps which we assume to be of order 1 in d

$$p_\eta^\mu = s^\mu \frac{z^\mu \cdot w}{d}, \quad \omega = \frac{\mu \cdot w}{d}, \quad r = \frac{\|w\|^2}{d} \quad q_\xi^\mu = s^\mu \frac{x_0^\mu \cdot u}{d}, \quad q_\eta^\mu = s^\mu \frac{z^\mu \cdot u}{d}, \quad m = \frac{\mu \cdot u}{d}, \quad q = \frac{\|u\|^2}{d}$$

We note that

$$\phi^\mu = f(x_t^\mu) = \tanh\left(b + \kappa t \sigma s^\mu p_\eta^\mu + \kappa t \omega s^\mu + \sqrt{r} Z^\mu + o_d(1)\right)$$

where $Z^\mu \sim \mathcal{N}(0,1)$. We now compute the loss

$$\begin{aligned}
\frac{1}{d}\sum_\mu \|x_1^\mu - f(x_t^\mu)\|^2 &= \frac{1}{d}\sum_\mu \left\| x_1^\mu - c((1-\kappa t/\sqrt{d})x_0^\mu + (\kappa t/\sqrt{d})x_1^\mu) - u\phi^\mu \right\|^2 \\
&= \frac{1}{d}\sum_\mu \|x_1^\mu - cx_0^\mu - u\phi^\mu\|^2 + o_d(1) \\
&= \sum_\mu 1 + \sigma^2 + c^2 + \frac{\|u\|^2}{d}(\phi^\mu)^2 - 2\left((\mu s^\mu + \sigma z^\mu)(1 - c\kappa t/\sqrt{d}) - cx_0^\mu\right) \cdot \frac{u}{d}\phi^\mu + o_d(1) \\
&= \sum_\mu 1 + \sigma^2 + c^2 + q(\phi^\mu)^2 - 2(m + \sigma q_\eta^\mu - cq_\xi^\mu)s^\mu \phi^\mu + o_d(1)
\end{aligned}$$

We follow the same style of calculation as Cui et al. (2024) to compute the partion function. First we write the partition function

$$\begin{aligned}
\mathcal{Z} &= \int d\theta\, e^{-\frac{\gamma}{2}\hat{R}_t(\theta)} \\
&= \int dcdudwdb\, e^{-\frac{\gamma d}{2}\left(\sum_\mu 1+\sigma^2+c^2+\frac{\|u\|^2}{d}(\phi^\mu)^2 - 2(\frac{u\cdot\mu}{d} + \sigma\frac{u\cdot z^\mu}{d} - c\frac{u\cdot x_0^\mu}{d})s^\mu\phi^\mu\right) - \frac{\gamma\lambda}{2}\|u\|^2 - \frac{\gamma\ell}{2}\|w\|^2}
\end{aligned}$$

Next we introduce overlaps into the integral and their conjugates by Dirac-Fourier, which we will denote as the vectors ζ and $\hat{\zeta}$ to simplify notation, and rearrange to integrate u, w

$$\begin{aligned}
\mathcal{Z} = &\int dcdb\, d\zeta d\hat{\zeta}\, e^{d\left(\frac{1}{2}\hat{q}q + \hat{m}m + \frac{1}{2}\hat{r}r + \hat{\omega}\omega + \sum_{\mu=1}^n (q_\xi^\mu \hat{q}_\xi^\mu + q_\eta^\mu \hat{q}_\eta^\mu + p_\eta^\mu \hat{p}_\eta^\mu)) - \frac{\gamma}{2}\sum_\mu (1+\sigma^2+c^2+q(\phi^\mu)^2 - 2(m+\sigma q_\eta^\mu - cq_\xi^\mu)s^\mu\phi^\mu)\right)} \\
&\int due^{-\frac{\hat{q}+\gamma\lambda}{2}\|u\|^2 - u\cdot\left(\hat{m}\mu + \sum_{\mu=1}^n \hat{q}_\xi x_0^\mu + \hat{q}_\eta z^\mu\right)} \int dwe^{-\frac{\hat{r}+\gamma\ell}{2}\|w\|^2 - w\cdot\left(\hat{\omega}\mu + \sum_{\mu=1}^n \hat{p}_\eta z^\mu\right)}
\end{aligned}$$

Next we evaluate the u, w integrals to get

$$\begin{aligned}
&e^{d\left(\frac{1}{2}\log(\hat{q}+\gamma\lambda) + \frac{1}{2}\log(\hat{r}+\gamma\ell) + \frac{1}{2(\hat{q}+\gamma\lambda)}\frac{1}{d}\|\hat{m}\mu + \sum_{\mu=1}^n \hat{q}_\xi^\mu x_0^\mu + \hat{q}_\eta^\mu z^\mu\|^2 + \frac{1}{2(\hat{r}+\gamma\ell)}\frac{1}{d}\|\hat{\omega}\mu + \sum_{\mu=1}^n \hat{p}_\eta^\mu z^\mu\|^2\right)} \\
&= e^{d\left(\frac{1}{2}\log(\hat{q}+\gamma\lambda) + \frac{1}{2}\log(\hat{r}+\gamma\ell) + \frac{1}{2(\hat{q}+\gamma\lambda)}(\hat{m}^2 + \sum_{\mu=1}^n (\hat{q}_\xi^\mu)^2 + (\hat{q}_\eta^\mu)^2) + \frac{1}{2(\hat{r}+\gamma\ell)}(\hat{\omega}^2 + \sum_{\mu=1}^n (\hat{p}_\eta^\mu)^2) + O(1/\sqrt{d})\right)}
\end{aligned}$$

We now assume a sample-symmetry ansatz on the overlaps which means that $q_\xi^\mu = q_\xi$ for every μ are all equal, and the same for $\hat{q}_\xi^\mu, q_\eta^\mu, \hat{q}_\eta^\mu, p_\eta^\mu, \hat{p}_\eta^\mu$. We then take $d \to \infty$, rescale all conjugates with γ, change all conjugates signs except for $\hat{q}$ and $\hat{r}$ for cleaner equations, and take $\gamma \to \infty$. This gives us the following effective field (log partition function)

$$\begin{aligned}
\log \mathcal{Z}(\mathcal{D}) = \text{extr}\Bigg\{ &-\frac{n}{2}\left(-2(\sigma q_\eta + m - cq_\xi)\overline{\phi s} + c^2 + q\overline{\phi^2}\right) \\
&+ \frac{q\hat{q}}{2} + \frac{r\hat{r}}{2} - m\hat{m} - \omega\hat{\omega} - n(q_\xi\hat{q}_\xi + q_\eta\hat{q}_\eta + p_\eta\hat{p}_\eta) \\
&+ \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2)}{2(\lambda + \hat{q})} + \frac{\hat{\omega}^2 + n\hat{p}_\eta^2}{2(\ell + \hat{r})}\Bigg\}
\end{aligned}$$

Taking gradients we get the following saddle-point equations

$$
\begin{cases}
(\sigma q_\eta + m - c q_\xi)\overline{\phi' s} - q\overline{\phi\phi'} = 0 \\
c = -\overline{\phi s} q_\xi \\
p_\eta = \frac{\hat{p}_\eta}{\ell + \hat{r}} \\
\omega = \frac{\hat{\omega}}{\ell + \hat{r}} \\
r = \frac{\hat{\omega}^2 + n\hat{p}_\eta^2}{(\ell+\hat{r})^2} = \omega^2 + n p_\eta^2 \\
q_\xi = \frac{\hat{q}_\xi}{\lambda + \hat{q}} \\
q_\eta = \frac{\hat{q}_\eta}{\lambda + \hat{q}} \\
m = \frac{\hat{m}}{\lambda + \hat{q}} \\
q = \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2)}{(\lambda+\hat{q})^2} = m^2 + n q_\xi^2 + n q_\eta^2
\end{cases}
\qquad
\begin{cases}
\hat{p}_\eta = (\kappa t)((\sigma q_\eta + m - c q_\xi)\overline{\phi' s} - q\overline{\phi'\phi}) = 0 \\
\hat{\omega} = (n\kappa t)((\sigma q_\eta + m - c q_\xi)\overline{\phi'} - q\overline{\phi'\phi s}) \\
\hat{r} = -n((\sigma q_\eta + m - c q_\xi)\overline{Z\phi' s} - q\overline{Z\phi\phi'})/\sqrt{r} \\
\hat{q}_\xi = -c\overline{\phi s} \\
\hat{q}_\eta = \sigma\overline{\phi s} \\
\hat{m} = n\overline{\phi s} \\
\hat{q} = n\overline{\phi^2}
\end{cases}
$$

Combining the equations for c and q_ξ we get that $c = c\overline{\phi s}^2/(\lambda + n\overline{\phi^2})$. We now argue that $c = 0$ almost surely, since otherwise $\overline{\phi} = 0$ on a non-zero measure, implying $b = \omega = 0$, which then results in all overlaps being zero, giving a suboptimal log partition function. This can be seen more explicitly by noting that the log partition function is zero for $c \neq 0$, but for $c = 0$ it is instead

$$
\log \mathcal{Z}(\mathcal{D}) = \text{extr}_{\omega,b} \left\{ n \frac{\overline{\phi s}^2 (\sigma^2 + n)}{2\left(\lambda + n\overline{\phi^2}\right)} - \frac{1}{2}\ell\omega^2 \right\}
$$

which has positive values for example at $\omega = 0, b \neq 0$. The above formulation is also useful for solving for the overlaps numerically.

B.1.1 Argument for Corollary 1

We now focus on $n \to \infty$ and on verifying that $\omega = \kappa t, b = \tanh^{-1}(\overline{s})$ is a solution. We will need the following preliminary claims.

Lemma 4. *Let $\phi^\mu = \tanh(b + \kappa t \omega s^\mu + \omega Z^\mu)$. If $\omega = \kappa t$ and $b = \tanh^{-1}(\overline{s})$ then $\overline{\phi - s} = 0$*

Proof. Let $\phi^\mu_\pm = \phi^\mu|_{s^\mu = \pm 1}$ and $\phi_0 = \phi^\mu|_{s^\mu = 0}$. Then

$$
\begin{aligned}
\overline{\phi - s} &= \overline{p}\overline{(\phi_+ - 1)} + (1 - \overline{p})\overline{(\phi_- + 1)} \\
&= \int dz\, e^{-\frac{z^2}{2}} \left\{ \overline{p}(\phi_+ - 1) + (1 - \overline{p})(\phi_- + 1) \right\} \\
&= \int dz\, e^{-\frac{z^2 + (\kappa t)^2}{2}} \left\{ e^{\kappa t z}\overline{p}(\phi_0 - 1) + e^{-\kappa t z}(1 - \overline{p})(\phi_0 + 1) \right\} \qquad z \to z \mp \kappa t
\end{aligned}
$$

Finally note that the integrad is zero for all z if

$$
\phi_0 = \frac{e^{\kappa t z}\overline{p} - e^{-\kappa t z}(1 - \overline{p})}{e^{\kappa t z}\overline{p} + (1 - \overline{p})e^{-\kappa t z}} = \tanh\left(\tanh^{-1}(\overline{s}) + \kappa t z\right)
$$

which occurs for $\omega = \kappa t$ and $b = \tanh^{-1}(\overline{s})$. □

Corollary 7. *Let $\phi^\mu = \tanh(b + \kappa t \omega s^\mu + \omega Z^\mu)$. If $\omega = \kappa t$ and $b = \tanh^{-1}(\overline{s})$ then for any function g where $\overline{g(Z \pm \kappa t)(\phi \mp 1)}$ exist we have*

$$
\overline{g(Z + s\kappa t)(\phi - s)} = 0
$$

In particular,

$$
\overline{\phi - s} = \overline{\phi(\phi - s)} = \overline{\phi'(\phi - s)} = \overline{(Z + s\kappa t)\phi'(\phi - s)} = 0
$$

Solving for q_η, m, q_ξ, q and plugging into the equation for b we get

$$(\lambda + n\overline{\phi^2})\,\overline{\phi' s} = n\overline{\phi s}\,\overline{\phi\phi'}$$

Taking $n \to \infty$, to leading order in n the equality becomes $(\overline{\phi^2})(\overline{\phi' s}) - (\overline{\phi'\phi})(\overline{\phi s}) = 0$ which holds by Corollary 7.

Using the independence of s^μ, Z^μ and taking the limit of infinitely many Z^μ, we can use Stein's lemma to rewrite the $\hat{r}$ equation as

$$\hat{r} = -n((\sigma q_\eta + m - c q_\xi)\overline{\phi'' s} - q\overline{(\phi\phi')'})$$

Plugging in q_η, m, q_ξ, q in gives

$$\hat{r}(\lambda + n\overline{\phi^2})^2 = -n((\lambda + n\overline{\phi^2})(\sigma^2 + n)\overline{\phi'' s}\,\overline{\phi s} - n(\sigma^2 + n)\overline{\phi s}^2\,\overline{(\phi\phi')'})$$

Plugging q_η, m, q_ξ, q into equations for ω and $\hat{\omega}$ gives

$$\omega(\ell + \hat{r})(\lambda + n\overline{\phi^2})^2 = (n\kappa t)((\lambda + n\overline{\phi^2})(\sigma^2 + n)(\overline{\phi'}\,\overline{\phi s}) - n(\sigma^2 + n)\overline{\phi s}^2\overline{\phi' \phi s})$$

For large n we get

$$\begin{aligned}
-\omega((\overline{\phi^2})(\overline{\phi'' s})(\overline{\phi s}) - (\overline{\phi s}^2)(\overline{(\phi\phi')'}) &= (\kappa t)((\overline{\phi^2})(\overline{\phi'})(\overline{\phi s})) - (\overline{\phi s}^2)(\overline{\phi'\phi s}))\\
\omega &= (\kappa t)\frac{((\overline{\phi'}) - (\overline{\phi'\phi s}))}{-((\overline{\phi'' s}) - (\overline{(\phi\phi')'}))}\\
&= (\kappa t)\frac{\overline{(\phi - \frac{1}{2}\phi^2 s)'}}{-\overline{(\phi - \frac{1}{2}\phi^2 s)'' s}}
\end{aligned}$$

Finally note that

$$\overline{(\phi - \tfrac{1}{2}\phi^2 s)'} + \overline{(\phi - \tfrac{1}{2}\phi^2 s)'' s} = -\frac{1}{2\kappa t}\overline{[(\phi - s)^2]'\,(s\kappa t + Z)} = 0$$

applying Corollary 7.

B.2 Derivation of Second Phase

Analogously to Result 1, we obtain the following result for the second phase.

Result 3 (Sharp Characterization of Parameters in Second Phase)**.** *For any $t \in [1, 2]$, in the $d \to \infty$ limit, the parameters minimizing the loss from equation 9 satisfy the following equations*

$$q_\xi = \frac{c(1-\tau)}{\lambda + n} \qquad m = \frac{n(1 - c\tau)}{\lambda + n}$$

$$q_\eta = \frac{\sigma(1 - c\tau)}{\lambda + n} \qquad q = m^2 + n q_\xi^2 + n\sigma^2 q_\eta^2$$

$$c = \frac{\tau\left((1+\sigma^2)(\lambda + n) - (\sigma + n)\right)}{(\lambda + n)((1-\tau^2) + (1+\sigma^2)\tau^2) + ((1-\tau)^2 - \tau^2(\sigma + n))}$$

where $\tau = t - 1$.

We proceed to demonstrate this result. Consider times $t \in [1, 2]$ which means we have

$$x_t^\mu = (2 - t)\left(1 - \frac{\kappa}{\sqrt{d}}\right)x_0^\mu + \left(\frac{\kappa}{\sqrt{d}} + \left(1 - \frac{\kappa}{\sqrt{d}}\right)(t - 1)\right)x_1^\mu.$$

We change variables to $\tau = t - 1$ and consider $\tau \in [0, 1]$ so that

$$x_\tau^\mu = (1 - \tau)\left(1 - \frac{\kappa}{\sqrt{d}}\right)x_0^\mu + \left(\frac{\kappa}{\sqrt{d}} + \left(1 - \frac{\kappa}{\sqrt{d}}\right)\tau\right)x_1^\mu.$$

We compute the loss for a single datapoint, defining $\nu^\mu = s^\mu \phi\left(w \cdot x_\tau^\mu + b\right)$

$$
\begin{aligned}
&\frac{1}{d}\left\| x_1^\mu - c\left((1-\tau)\left(1-\frac{\kappa}{\sqrt{d}}\right)x_0^\mu + \left(\frac{\kappa}{\sqrt{d}} + \left(1-\frac{\kappa}{\sqrt{d}}\right)\tau\right)x_1^\mu\right) - u\phi\left(w\cdot x_\tau^\mu + b\right)\right\|^2 \\
&= \frac{1}{d}\left\| x_1^\mu - c((1-\tau)x_0^\mu + \tau x_1^\mu) - u s^\mu \nu^\mu \right\|^2 + o_d(1) \\
&= \frac{1}{d}\left\| (1-c\tau)(\sigma z^\mu + s^\mu \mu) - c(1-\tau)x_0^\mu - u s^\mu \nu^\mu \right\|^2 + o_d(1) \\
&= (1+\sigma^2)(1-c\tau)^2 + c^2(1-\tau)^2 + \frac{\|u\|^2}{d} - \frac{2 s^\mu \nu^\mu}{d} u \cdot ((1-c\tau)(\sigma z^\mu + s^\mu \mu) - c(1-\tau)x_0^\mu) + o_d(1) \\
&= (1+\sigma^2)(1-c\tau)^2 + c^2(1-\tau)^2 + q - 2\nu^\mu(1-c\tau)(\sigma q_\eta^\mu + m) + 2\nu^\mu c(1-\tau)q_\xi^\mu + o_d(1)
\end{aligned}
$$

where we defined the overlaps

$$
q = \frac{\|u\|^2}{d}, \quad q_\eta^\mu = s^\mu \frac{u\cdot z^\mu}{d}, \quad q_\xi^\mu = s^\mu \frac{u \cdot x_0^\mu}{d}, \quad m = \frac{u\cdot \mu}{d}
$$

$$
p_\eta^\mu = s^\mu \frac{w \cdot z^\mu}{d}, \quad p_\xi^\mu = s^\mu \frac{w\cdot x_0^\mu}{d}, \quad \omega = \frac{w\cdot \mu}{d}.
$$

We also have

$$
\begin{aligned}
\nu^\mu = \phi\left(s^\mu w \cdot x_\tau^\mu + s^\mu b\right) &= \tanh\left((1-\tau)s^\mu w\cdot x_0^\mu + \left(\tau + \frac{\kappa}{\sqrt{d}}\right)s^\mu w \cdot (\sigma z^\mu + s^\mu \mu) + s^\mu b\right) \\
&= \tanh\left(d\left((1-\tau)p_\xi^\mu + \left(\tau + \frac{\kappa}{\sqrt{d}}\right)(\sigma p_\eta^\mu + \omega)\right) + s^\mu b\right) \\
&\asymp \text{sign}\left(\sqrt{d}\left((1-\tau)p_\xi^\mu + \tau(\sigma p_\eta^\mu + \omega)\right) + \kappa(\sigma p_\eta^\mu + \omega)\right)
\end{aligned}
$$

This gives the following

$$
\begin{aligned}
\log \mathcal{Z}(\mathcal{D}) = \text{extr}\Bigg\{ &-\frac{n}{2}\left((1+\sigma^2)(1-c\tau)^2 + c^2(1-\tau)^2 + q - 2(1-c\tau)(\sigma q_\eta + m)\overline{\nu} + 2c(1-\tau)q_\xi \overline{\nu}\right) \\
&+ \frac{q\hat{q}}{2} - m\hat{m} - n(q_\xi \hat{q}_\xi + q_\eta \hat{q}_\eta) + \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2)}{2(\lambda + \hat{q})}\Bigg\}
\end{aligned}
$$

Taking gradients we get the following saddle-point equations wherefrom Result 3 is obtained

$$
\begin{cases}
q_\xi = \frac{\hat{q}_\xi}{\lambda+\hat{q}} = \frac{c(1-\tau)\overline{\nu}}{\lambda+n} \\
q_\eta = \frac{\hat{q}_\eta}{\lambda+\hat{q}} = \frac{(1-c\tau)\overline{\nu}}{\lambda+n} \\
m = \frac{\hat{m}}{\lambda+\hat{q}} = \frac{n(1-c\tau)\overline{\nu}}{\lambda+n} \\
q = \frac{\hat{m}^2 + n(\hat{q}_\xi^2 + \hat{q}_\eta^2)}{(\lambda+\hat{q})^2} = m^2 + n q_\xi^2 + n\sigma^2 q_\eta^2 \\
c = \frac{(1+\sigma^2)\tau - \tau(\sigma q_\eta + m)\overline{\nu} - (1-\tau)q_\xi \overline{\nu}}{(1-\tau)^2 + (1+\sigma^2)\tau^2}
\end{cases}
\qquad
\begin{cases}
\hat{q}_\xi = c(1-\tau)\overline{\nu} \\
\hat{q}_\eta = \sigma(1-c\tau)\overline{\nu} \\
\hat{m} = n(1-c\tau)\overline{\nu} \\
\hat{q} = n
\end{cases}
$$

$$
c = \frac{(1+\sigma^2)\tau(\lambda+n) - \overline{\nu}^2\tau(\sigma+n)}{(\lambda+n)((1-\tau^2) + (1+\sigma^2)\tau^2) + \overline{\nu}^2\left((1-\tau)^2 - \tau^2(\sigma+n)\right)}
$$

Corollary 3 simply follows from taking the $n \to \infty$ limit of this equations.

Lastly, we now argue that we can take $\overline{\nu} = 1$ without loss of generality. If we assume a sample symmetric ansatz for $p_\eta^\mu = p_\eta, p_\xi^\mu = p_\xi$, then $\overline{\nu}$ can either be ± 1 depending on the sign of argument. Noting that q, c are unchanged while $q_\eta, q_\xi, m, \hat{q}_\eta, \hat{q}_\xi, \hat{m}$ flip sign, we can conclude that the log partition function is invariant so $\overline{\nu} = 1$.

The characterizations of the learned parameters can be used to evaluate the MSE as a function of t, which we now describe, in the limit of $d \to \infty$ and then $n \to \infty$. For the first and second phase we obtain

$$\text{mse}_{\text{train}} = \text{mse}_{\text{test}} = \begin{cases} \sigma^2 + (1 - \overline{\phi^2}) & t \in [0,1] \\ \sigma^2(1-c\tau)^2 + c^2(1-\tau)^2 & t \in [1,2] \end{cases}$$

At $t = 0$, $\phi = \tanh(b) = 2(p - 1/2)$ hence the MSE is $\sigma^2 + 4p(1-p)$. At $t = 1$ we have $c = 0$ hence the MSE is σ^2, while at $t = 2$ we have $c = 1$ so the MSE is 0.

C ARGUMENTS FOR GENERATION

Combining equations 4 and 8 gives the exact velocity field

$$b_t(x) = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t\right)\left(\frac{\beta_t \sigma^2}{\alpha_t^2 + \sigma^2\beta_t^2}x + \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\mu \tanh\left(\frac{\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\mu \cdot x + h\right)\right) + \frac{\dot{\alpha}_t}{\alpha_t}x. \tag{32}$$

where $\alpha_t = 1 - \tau_t$ and $\beta_t = \tau_t$ with τ_t from equation 12. Let $\hat{\theta}_t$ denote any overlap from the first phase (see equation 13) in the limit of $d \to \infty$ but for finite n, where θ_t denotes the corresponding overlap with $d \to \infty$ and then $n \to \infty$. From Results 1 and 3 and their Corollaries 1 and 3, we have that $|\hat{\theta}_t - \theta_t| = O_n(1/n)$ for all overlaps.

Since $X_t - \hat{X}_t$ is contained in span(u_t, η) which is in turn contained in span(μ, η, ξ), it suffices to show that, after dividing by d, the projections of $X_t - \hat{X}_t$ onto μ, η, and ξ are $O(1/n)$ to show that $\frac{1}{d}\|X_t - \hat{X}_t\|$ is $O(1/n)$.

C.1 ARGUMENT FOR RESULT 2

First, we note that as described in the paragraph above the statement of Result 2, we have that since in the first phase $q = m^2 + nq_\eta^2$ from Result 1 we get for $t \in [0,1]$

$$\lim_{d\to\infty} \frac{\|u_t\|^2}{d} = \lim_{d\to\infty} \left(\frac{\mu \cdot u_t}{d}\right)^2 + \left(\frac{\eta \cdot u_t}{d}\right)^2$$

also since $q = m^2 + nq_\xi^2 + nq_\eta^2$ in the second phase, we get that for $t \in [1,2]$

$$\lim_{d\to\infty} \frac{\|u_t\|^2}{d} = \lim_{d\to\infty} \left(\frac{\mu \cdot u_t}{d}\right)^2 + \left(\frac{\eta \cdot u_t}{d}\right)^2 + \left(\frac{\xi \cdot u_t}{d}\right)^2$$

where $\eta = \sigma \sum_{\mu=1}^n z^\mu$ and $\xi = \sum_\mu s^\mu x_0^\mu$ which implies that for any $w \in$ span$(\mu, \eta, \xi)^\perp$ with $\|w\|_2 = 1$ we have

$$\lim_{d\to\infty} \frac{w \cdot (X_2 - \hat{X}_2)}{\sqrt{d}} = 0.$$

C.1.1 FIRST PHASE

We focus on $t \in [0,1]$ and define

$$\begin{aligned} &\epsilon_t^m = \frac{1}{\sqrt{d}}\mu \cdot (X_t - \hat{X}_t), \quad \epsilon_t^\eta = \frac{1}{\sigma^2 n\sqrt{d}}\eta \cdot (X_t - \hat{X}_t), \\ &\delta_t = \frac{\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}, \quad \gamma_t = \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}, \\ &M_t = \frac{\mu \cdot X_t}{\sqrt{d}}, \quad Q_t^\eta = \frac{\eta \cdot X_t}{\sigma^2 n\sqrt{d}}. \end{aligned}$$

We have

$$
\begin{aligned}
\dot{\epsilon}_t^m &= \frac{1}{\sqrt{d}}\mu \cdot (\dot{X}_t - \dot{\hat{X}}_t) \\
&= \frac{1}{\sqrt{d}}\mu \cdot (b_t(X_t) - \hat{b}_t(\hat{X}_t)) \\
&= \frac{1}{\sqrt{d}}\mu \cdot \left(\dot{\beta} - \frac{\dot{\alpha}}{\alpha}\beta\right)\Big(\sigma^2\delta_t(X_t - \hat{X}_t) + (c_t - \sigma^2\delta_t)X_t + (\gamma_t\mu - u_t)\tanh(\delta_t\mu \cdot X_t + h) \\
&\quad + u_t\left(\tanh(\delta_t\mu \cdot X_t + h) - \tanh(w_t \cdot X_t + b_t)\right)\Big) + \frac{\dot{\alpha}}{\alpha}\frac{1}{\sqrt{d}}\mu \cdot \left(X_t - \hat{X}_t\right) \\
&= \left(\dot{\beta} - \frac{\dot{\alpha}}{\alpha}\beta\right)\Big(\sigma^2\delta_t\epsilon_t^m + (c_t - \sigma^2\delta_t)M_t + \sqrt{d}(\gamma_t - m_t)\tanh(\delta_t\mu \cdot X_t + h) \\
&\quad + \sqrt{d}m_t\left(\tanh(\delta_t\mu \cdot X_t + h) - \tanh(w_t \cdot \hat{X}_t/\sqrt{d} + b_t)\right)\Big) + \frac{\dot{\alpha}}{\alpha}\epsilon_t^m \\
&= \kappa\left(1 - t\frac{\dot{\alpha}}{\alpha}\right)\Big(\frac{\delta_t}{\sqrt{d}}\epsilon_t^m + (c_t - \sigma^2\delta_t)\frac{M_t}{\sqrt{d}} + (\gamma_t - m_t)\tanh(\delta_t\mu \cdot X_t + h) \\
&\quad + m_t\left(\tanh(\delta_t\mu \cdot X_t + h) - \tanh(w_t \cdot \hat{X}_t/\sqrt{d} + b_t)\right)\Big) + \frac{\dot{\alpha}}{\alpha}\epsilon_t^m
\end{aligned}
$$

We now focus on the $\tanh$

$$
\begin{aligned}
&\left|\tanh(\delta_t\mu \cdot X_t + h) - \tanh(w_t \cdot \hat{X}_t/\sqrt{d} + b_t)\right| \\
&\leq \left|\left(\delta_t\mu - \frac{w_t}{\sqrt{d}}\right) \cdot X_t\right| + \left|\frac{w_t}{\sqrt{d}} \cdot (X_t - \hat{X}_t)\right| + |h - b_t| \\
&\leq \left|\left(\delta_t\mu - \frac{w_t}{\sqrt{d}}\right)\left(\frac{\mu\mu^T}{d} + \frac{\eta\eta^T}{\sigma^2 n^2 d}\right)X_t\right| + \left|\frac{w_t}{\sqrt{d}}\left(\frac{\mu\mu^T}{d} + \frac{\eta\eta^T}{\sigma^2 n^2 d}\right)(X_t - \hat{X}_t)\right| + |h - b_t| \\
&\leq |\left(\sqrt{d}\delta_t - \omega_t\right)M_t| + |\left(\delta_t Z - p_t^\eta\right)Q_t| + |\omega_t\epsilon_t^m| + |p_t^\eta\epsilon_t^\eta| + |h - b_t| \\
&\leq \omega_t|\epsilon_t^m| + O\left(\frac{1}{n}\right) + O\left(\frac{1}{\sqrt{d}}\right).
\end{aligned}
$$

Coming back to the ODE for $\dot{\epsilon}_t^m$, we get with high probability

$$
|\dot{\epsilon}_t^m| = \kappa\gamma_t\omega_t|\epsilon_t^m| + O\left(\frac{1}{n}\right) + O\left(\frac{1}{\sqrt{d}}\right).
$$

Since $\kappa\gamma_t\omega_t = \Theta_{n,d}(1)$ for $t \in [0, 1]$, we get that with high probability

$$
\epsilon_{t=1}^m = O\left(\frac{1}{n}\right) + O\left(\frac{1}{\sqrt{d}}\right).
$$

By performing a similar computation for the ODE for ϵ_t^η, we get that with high probability

$$
\epsilon_{t=1}^\eta = O\left(\frac{1}{n}\right) + O\left(\frac{1}{\sqrt{d}}\right).
$$

C.1.2 SECOND PHASE

We now turn to $t \in [1, 2]$ and define

$$
\zeta_t^m = \frac{1}{d}\mu \cdot (X_t - \hat{X}_t), \quad \zeta_t^\eta = \frac{1}{\sigma^2 nd}\eta \cdot (X_t - \hat{X}_t) \quad \zeta_t^\xi = \frac{1}{nd}\xi \cdot (X_t - \hat{X}_t).
$$

With high probability, we have the following ODEs hold

$$\frac{d}{dt}\zeta^m = \left(\beta(t)c_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\zeta^m + O\left(\frac{1}{n}\right), \tag{33}$$

$$\frac{d}{dt}\zeta^\eta = \left(\beta(t)c_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\zeta^\eta + O\left(\frac{1}{n}\right), \tag{34}$$

$$\frac{d}{dt}\zeta^\xi = \left(\beta(t)c_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1 - c_t\beta(t))\right)\zeta^\xi + O\left(\frac{1}{n}\right). \tag{35}$$

from the initial condition $\zeta_1^m, \zeta_1^\eta = O(\frac{1}{\sqrt{d}}) + O(\frac{1}{n}), \zeta_1^\xi = 0$. This yields

$$\zeta_2^m, \zeta_2^\eta, \zeta_2^\xi = O(\tfrac{1}{n}) + O(\tfrac{1}{\sqrt{d}})$$

C.2 Argument for Corollary 6

By Proposition 1, we know that

$$\lim_{\kappa\to\infty}\lim_{d\to\infty}\frac{\mu\cdot X_2}{d} \sim p\delta_1 + (1-p)\delta_{-1}.$$

By Result 2, we get that

$$\lim_{n\to\infty}\lim_{d\to\infty}\frac{\mu\cdot(\hat{X}_2 - X_2)}{d} = 0.$$

Combining the last two equations gives the first claim from the Corollary.

Fix $w \perp \mu$, $\|w\| = 1$. Again by Proposition 1, we have that

$$\lim_{d\to\infty}\frac{w\cdot X_2}{\sqrt{d}} \sim \mathcal{N}(0,\sigma^2).$$

Also, Result 2 gives that

$$\lim_{n\to\infty}\lim_{d\to\infty}\frac{w\cdot(\hat{X}_2 - X_2)}{\sqrt{d}} = 0.$$

which combined with the previous equation gives the second claim from the Corollary.

D Experimental details

The model used for the MNIST experiment consists of a U-Net architecture (Ronneberger et al. (2015)), consisting of four downsampling and four upsampling blocks with two layers per block and output channels of 128, 128, 256, and 512, respectively. Attention mechanisms are integrated into the third downsampling block and the second upsampling block to enhance feature representation at multiple scales. The training of the denoiser is described in the main text. We then use this denoiser to estimate the score and run the Variance Preserving SDE (see equation (11) in Song et al. (2021).)

For the discriminative model, we use the MNIST digit classification model by Knight (2022) available on Hugging Face which achieves an accuracy of 99.1% on MNIST classification.

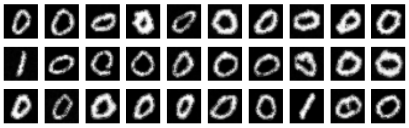
As a sanity check, we show non-cherry-picked samples generated by the three models we considered in Figure 2.

E General time dilation formula

In this section, we generalize the time dilation formula from equation 12 for a Gaussian mixture with more than two modes. Although the arguments in Results 1 and 2 only hold for the two-mode GM, the fact that a more general time dilation formula exists suggests that these results could be

(a) Training times with prob. $1/2$ on $[.2, .6]$

(b) Training times with prob. $1/2$ on $[.3, .5]$



(c) Training times uniform on $[0, 1]$

Figure 2: Non-cherry-picked samples from the three generative models considered. **(a)** Samples from the VP SDE, where the times for training are drawn with probability $1/2$ uniformly from $[.2, .6]$ and with probability $1/2$ uniformly outside. **(b)** Same as left panel except that with probability $1/2$ training times are sampled from $[.3, .5]$. **(c)** Samples from the VP SDE with training times that are uniform in $[0, 1]$.

extended to the GM with more than two modes.

Consider $\mu = \sum_{i=1}^{m} p_i \mathcal{N}\left(r_i, \sigma^2 \mathrm{I}\right)$ where $r_i \in \mathbb{R}^d$ and $|r_i|$ goes to infinity with d,but m, p_i, σ^2 are constant with respect to d. If X_t is the generative model associated with the interpolant $I_t = (1-t)z + ta$ where $z \sim \mathcal{N}(0, \mathrm{I})$ and $a \sim \mu$ (as we do in equation equation 3 in the main text) then X_t estimates p_i at times of the order $1/|r_i|$. We show this in Proposition 2 below by arguing that it is only at times of order $1/|r_i|$ that the denoiser associated to $r_i \cdot X_t/|r_i|$ is nontrivial. Hence, to estimate p_i we require a time dilation τ_t such that there exists a and b with $b - a = \Theta_d(1)$ where

$$\tau_t = \Theta_d\left(\frac{1}{|r_i|}\right) \text{ for } \; t \in [a, b]. \tag{36}$$

We specify next a time dilation that for every i would ensure that Equation 36 is fulfilled. Assume $|r_1| \leq |r_2| \leq \cdots \leq |r_m|$, let $n = m + 1$ and let $\kappa > 0$. Then

$$\tau_t = \begin{cases} \frac{\kappa n t}{|r_m|} & \text{if } t \in [0, 1/n] \\ \frac{\kappa(nt-1)}{|r_{m-1}|} + \frac{\kappa}{|r_m|} & \text{if } t \in [1/n, 2/n] \\ \cdots & \\ \frac{\kappa(nt-(m-1))}{|r_1|} + \kappa\left(\frac{1}{|r_2|} + \cdots + \frac{1}{|r_m|}\right) & \text{if } t \in [(m-1)/n, m/n] \\ \left(1 - \kappa\left(\frac{1}{|r_1|} + \cdots + \frac{1}{|r_m|}\right)\right) t + \kappa\left(\frac{1}{|r_1|} + \cdots + \frac{1}{|r_m|}\right) & \text{if } t \in [m/n, 1] \end{cases} \tag{37}$$

Then we have that p_i is learned when $t \in [(m-i)/n, (m-i+1)/n]$ and the σ^2 will be learned when $t \in [m/n, 1]$, giving rise to $m+1$ different phases. In the special case of $|r_i| = |r_{i+1}|$, both p_i and p_{i+1} will already be learned in $[(m-i)/n, (m-i+1)/n]$ so that the phase on the interval $[(m-i+1)/n, (m-i+2)/n]$ is unnecessary. Taking this consideration into account when using the general formula in equation 37 for the two-mode GM gives the time dilation formula from equation 12. The only difference is that the time dilation here maps $[0, 1]$ to $[0, 1]$ and the one in equation 12 maps $[0, 1]$ to $[0, 2]$.

Proposition 2. *Let $\mu = \sum_{i=1}^{m} p_i \mathcal{N}\left(r_i, \sigma^2 I\right)$ where $r_i \in \mathbb{R}^d$ and $|r_i| = \omega_d(1)$. Consider the interpolant $I_t = (1-t)z + ta$ where $z \sim \mathcal{N}(0, Id)$ and $a \sim \mu$. Let X_t be the generative model associated to I_t as in equation equation 3. Then X_t learns the p_i at times $\Theta_d\left(1/|r_i|\right)$.*

Proof. Fix i. Let $m_t = r_i \cdot I_t/|r_i|$. We have $m_t \overset{d}{=} (1-t)Z + t|r_i|m$ where $Z \sim \mathcal{N}(0,1)$ and $m = r_i \cdot a/|r_i|^2 = \Theta_d(1)$. Let $\nu_t = r_i \cdot X_t/|r_i|$. By Lemma 5, ν_t obeys the self-consistent ODE

$$\dot{\nu}_t = \frac{\nu_t}{t} - \frac{\eta_t(\nu_t)}{t} \tag{38}$$

where $\eta_t(\nu)$ is the denoiser for ν_t

$$\eta_t(\nu) = \mathbb{E}[Z|m_t = \nu] = \mathbb{E}[Z|(1-t)Z + t|r_i|m = \nu].$$

By Lemma 6, since $|r_i| = \omega_d(1)$, the only times where this denoiser is nontrivial are $t = \Theta_d\left(1/|r_i|\right)$. We note that to estimate p_i we need to estimate ν_t, which requires spending a constant length of time in the nontrivial times of the ODE in equation 38, which are the nontrivial times for the denoiser. Indeed, p_i is learned on that interval, and if the length of that interval goes to 0 as d goes to infinity, we cannot estimate p_i.

□

Lemma 5. *Let $\mu = \sum_{i=1}^m p_i \mathcal{N}\left(r_i, \sigma^2 I\right)$ where $r_i \in \mathbb{R}^d$. Consider the interpolant $I_t = (1-t)z + ta$ where $z \sim \mathcal{N}(0, Id)$ and $a \sim \mu$. Let X_t be the generative model associated to I_t from Lemma 5. Fix i and let $m_t = r_i \cdot I_t/|r_i|$ and $\nu_t = r_i \cdot X_t/|r_i|$. Then with $\eta_t(\nu) = \mathbb{E}[Z|m_t = \nu]$ we have*

$$\dot{\nu}_t = \frac{\nu_t}{t} - \frac{\eta_t(\nu_t)}{t}$$

Proof. We have from Appendix A, Albergo et al. (2023) that the velocity field $b_t(x)$ associated with $I_t = (1-t)z + ta$ where $a \sim \sum_{i=1}^m p_i \mathcal{N}\left(r_i, \mathrm{Id}\right)$ can be written explicitly as

$$\begin{aligned}
\frac{\sum_{i=1}^m p_i \left(r_i + \frac{\dot{c}_t}{2c_t}(x - tr_i)\right) \mathcal{N}(x \mid tr_i, c_t \mathrm{I})}{\sum_{i=1}^m p_i \mathcal{N}(x \mid tr_i, c_t \mathrm{I})} &= \frac{\sum_{i=1}^m p_i \left(r_i + \frac{\dot{c}_t}{2c_t}(x - tr_i)\right) e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}}{\sum_{i=1}^m p_i e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}} \\
&= \frac{\dot{c}_t}{2c_t} x + \frac{\sum_{i=1}^m p_i \left(1 - \frac{\dot{c}_t}{2c_t} t\right) e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}}{\sum_{i=1}^m p_i e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}} r_i
\end{aligned}$$

where $c_t = (1-t)^2 + t^2$. The denoiser $\eta_t(x) = \mathbb{E}[z|I_t = x]$ is

$$\eta_t(x) = x - tb_t(x) = \left(1 - \frac{\dot{c}_t}{2c_t}\right) x - \frac{\sum_{i=1}^m p_i \left(t - \frac{\dot{c}_t}{2c_t} t^2\right) e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}}{\sum_{i=1}^m p_i e^{\frac{2tr_i \cdot x - t^2|r_i|^2}{2((1-t)^2+t^2)}}} r_i. \tag{39}$$

Fix i and let $m_t = r_i \cdot I_t/|r_i|$ and $\nu_t = r_i \cdot X_t/|r_i|$. Since $\dot{X}_t = b(X_t)$, we get that

$$\dot{\nu}_t = \frac{r_i \cdot b(X_t)}{|r_i|} = \frac{\nu_t}{t} - \frac{1}{t}\frac{r_i \cdot \eta_t(X_t)}{|r_i|} = \frac{\nu_t}{t} - \frac{\eta_t(\nu_t)}{t},$$

where the denoiser for the ν_t is defined as $\eta_t(\nu) = \mathbb{E}[Z|m_t = \nu]$. The last step in the displayed equality follows since from equation 39 we get that $r_i \cdot \eta_t(x)/|r_i|$ depends on x only through ν_t. □

Lemma 6. *Let $Z \sim \mathcal{N}(0,1)$ and $M \sim \mu$. Then for fixed $\gamma > 0$ we have that as $d \to \infty$*

$$\begin{aligned}
\mathbb{E}[Z|Z + d^{-\gamma} M = x] &\to x \\
\mathbb{E}[Z|Z + d^{\gamma} M = x] &\to \mathbb{E}[Z] = 0
\end{aligned}$$

Proof. Let $f_{Z,X}(z,x)$ be the joint density of Z and $X = Z + d^{-\gamma}M$ and $f_{Z,M}(z,m)$ the joint density of Z and M. We note that $f_{Z,X}(z,x) = f_{Z,M}(z, d^{\gamma}(x-z)) = f_Z(z)f_M(d^{\gamma}(x-z))$

$$\begin{aligned}
\mathbb{E}[Z|Z + d^{-\gamma}M = x] &= \frac{\int z f_{Z,X}(z,x)dz}{\int f_{Z,X}(z,x)dz} \\
&= \frac{\int z f_Z(z) f_M(d^{\gamma}(x-z))dz}{\int f_Z(z) f_M(d^{\gamma}(x-z))dz} \\
&= \frac{\int z f_Z(z) d^{\gamma} f_M(d^{\gamma}(x-z))dz}{\int f_Z(z) d^{\gamma} f_M(d^{\gamma}(x-z))dz} \\
&\to x
\end{aligned}$$

where the last step follows since $d^{\gamma} f_M(d^{\gamma} z)$ is an approximation to the identity. The other limit follows similarly. □