arXiv:2605.06831v1 [cs.LG] 7 May 2026

Why DDIM Hallucinates More than DDPM: A Theoretical Analysis of Reverse Dynamics

Muhammad H. Ashiq [*1] **Samanyu Arora** [*1] **Abhinav N. Harish** [1] **Ishaan Kharbanda** [1] **Hung Yun Tseng** [1] **Grigorios G. Chrysos** [1]

Abstract

We theoretically study the hallucination phenomena in two canonical diffusion samplers: the stochastic Denoising Diffusion Probabilistic Model (DDPM) and the deterministic Denoising Diffusion Implicit Model (DDIM). We analyze the reverse ODE (DDIM) and SDE (DDPM) for a Gaussian mixture target, proving that after a critical time τ, (a) DDIM can become stuck on the segment connecting the two nearest modes and (b) DDPM *stochasticity* helps it become unstuck from this region, thus avoiding hallucination. Our empirical validation verifies that DDPM has a significantly lower hallucination rate than DDIM when this region is entered. Building on our observations, we exhibit how using additional stochastic steps can help DDIM avoid hallucinations and offer new insights on how to design improved samplers.

1. Introduction

Despite significant progress, state-of-the-art diffusion models still exhibit hallucinations which do not abide by the structural, semantic constraints present in their training distribution. This includes distorted hands, implausible object boundaries, and temporal inconsistencies in video. Concretely, hallucinations can be characterized by generations which arise when the reverse diffusion trajectory is guided toward regions outside the set of constraints implied by the target distribution.

To address this issue, prior work has focused on a variety of different strategies. In the domain of text generation, Lu et al. (2025) identify a local generation bias in image diffusion models by highlighting their inability to model global context in MNIST digits. Aithal et al. (2024) characterize hallucinations as mode interpolation in Gaussian mixture models and propose a metric for detecting interpolations. Triaridis et al. (2025) explore dynamic guidance as a means of mitigating hallucination. Although these works provide strategies to detect hallucination, none of them explore *why* hallucinations arise by systematically studying the design and mechanics of the reverse diffusion process.

We focus on the distinction between two popular samplers: the stochastic Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) and the deterministic Denoising Diffusion Implicit Models (DDIM) (Song et al., 2021a). These samplers are trained identically; however, they differ in the sampling strategy they employ during the reverse process. DDPM conducts an equal number of forward and reverse steps, matching the introduction of noise in the forward diffusion SDE. DDIM, on the other hand, does *not* include this noise in the reverse process. DDIM introduces step skipping through a non-Markovian reverse process, offering faster inference. We identify that removing the noise in the reverse process contributes significantly to hallucinations. This observation has been made empirically by prior works in the context of counting hallucinations (Fu et al., 2025). However, to the best of our knowledge, a solid theoretical characterization of this phenomenon remains absent. We fill this gap and address hallucination as a *structural artifact* of the reverse process. Concretely, we ask the following question:

Can we formally characterize why DDIM hallucinates more than DDPM?

To answer this question, we study the reverse Probability Flow (PF) ODE (DDIM) and the reverse Variance Preserving (VP) SDE (DDPM) when the data distribution is an N-mode Gaussian mixture[1]. This setup lends itself to rigorous theoretical analysis. Additionally, we study hallucinations formulated as mode interpolations: a phenomenon where a sampled diffusion trajectory ends up at the line connecting modes of the data distribution (Aithal et al., 2024).

[*]Equal contribution [1]University of Wisconsin-Madison, Madison, WI, USA. Correspondence to: Grigorios G. Chrysos <chrysos@wisc.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Under mild regularity conditions, the marginal densities are identical for DDIM and DDPM.

As demonstrated in Figure 1b, our analysis reveals that DDIM/DDPM convergence occurs in 2 phases: (i) an early convergence phase dominated by attraction to the nearest line segment between two data modes and (ii) a late convergence phase governing motion along the line connecting these two modes. We demonstrate that the final steps of a trajectory under the late convergence phase determine the outcome of sampling and are a crucial difference between these two samplers. Specifically, in this final phase, the noise component of DDPM helps it escape regions on this line segment where DDIM gets stuck, thus avoiding hallucinations. Our theory demonstrates that this is the foundational reason that DDPM has an edge over DDIM, *regardless of the number of reverse steps employed*.

These findings rule out several previously held beliefs on DDIM hallucinations. For example, although practitioners often skip DDIM steps and thus incur numerical error when sampling, we demonstrate that this does not explain the hallucination rate gap. We find empirically that the DDIM and DDPM hallucination rate gap persists under the same discretization, and our theoretical results are derived under zero numerical error. Additionally, Aithal et al. (2024) attribute mode interpolation to the smoothing of the score function due to approximation with a neural network. Our results, however, are derived under the *exact score function*, i.e, the expected score on using infinite samples from the distribution. We find that, with the exact score, the mechanism causing mode interpolation can be cleanly characterized.

In summary, our contributions are as follows:

1. We rigorously study the source of DDIM hallucinations as observed in an N-mode Gaussian mixture, demonstrating that after a critical time τ, DDIM trajectories converge to the nearest line segment joining two modes and then can become stuck near the midpoint, thus hallucinating. This is illustrated in Figure 1.

2. We leverage this to provide a theoretical justification for why DDIM hallucinates more than DDPM: the noise of DDPM can help it become unstuck from this hallucination region around the midpoint. This is illustrated in Figure 1c.

3. Empirically, we invalidate that the DDIM hallucination rate gap can be explained by step skipping and demonstrate that adding a few DDPM steps after DDIM converges near the midpoint neighborhood can help the trajectory escape, lowering hallucination rate.

2. Problem Setting

Notation: Bolded lowercase letters $\boldsymbol{x} \in \mathbb{R}^{\varpi}$ denote vectors of dimension ϖ. $[N]$, for scalar $N \in \mathbb{N}$, denotes the integer subset $\{1, 2, ..., N\}$. For a function f, we use $\nabla_{\boldsymbol{x}} f$ to denote its gradient if $f : \mathbb{R}^{\varpi} \to \mathbb{R}$ and its Jacobian if $f : \mathbb{R}^{\varpi} \to \mathbb{R}^{o}$. We denote by $x_t : [0, T] \to \mathbb{R}$ and $\boldsymbol{x}_t : [0, T] \to \mathbb{R}^{\varpi}$ time-dependent scalar-valued and vector-valued functions, respectively, from scalar time T to 0 unless otherwise specified. We use $\dot{x}_t$ to denote the time derivative of x_t and similarly use $\dot{\boldsymbol{x}}_t$ for $\boldsymbol{x}_t$; for all other derivatives, we use the notation $\frac{d}{dx} f(x)$. $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ is used to denote the probability density function of a ϖ-dimensional Gaussian centered at $\boldsymbol{\mu}$ with variance $\sigma^2 \boldsymbol{I}$. We write stochastic differential equations (SDE) as $d\boldsymbol{x}_t = b(t, \boldsymbol{x}_t)dt + \sigma(t, \boldsymbol{x}_t)d\boldsymbol{W}_t$ where $\boldsymbol{W}_t \in \mathbb{R}^{\varpi \times 1}$ is the standard Brownian motion. We provide a full list of notation in Sec. A, and a symbol table in Sec. J.

Diffusion Models: The goal of diffusion models is to learn to transport a simple base distribution p_{base} to a more complicated target distribution p_{data}. Throughout, we consider p_{base} to be the standard Gaussian $\mathcal{N}(0, \boldsymbol{I})$ (Sohl-Dickstein et al., 2015). We first define the DDPM Variance Preserving (VP) forward SDE (Song et al., 2021b):

$$d\boldsymbol{x}_t \;=\; -\tfrac{1}{2}\beta_t\, \boldsymbol{x}_t\, dt \;+\; \sqrt{\beta_t}\, d\boldsymbol{W}_t, \tag{1}$$

where β_t is a noise schedule and $\boldsymbol{W}_t$ is the standard Brownian motion. We then denote $\bar{\alpha}_t := \exp(-\int_0^t \beta(s)ds)$. We consider the standard setting where $\bar{\alpha}_t \in C^2$, $\bar{\alpha}_0 = 1$, and $\bar{\alpha}_T = \gamma$ for some $\gamma \approx 0$. By Anderson's theorem (Anderson, 1982), Eq. (1) can be reversed as:

$$d\boldsymbol{x}_t \;=\; -\frac{1}{2}\beta_t \left(\boldsymbol{x}_t + 2\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right) dt \;+\; \sqrt{\beta_t}\, d\bar{\boldsymbol{W}}_t, \tag{2}$$

for reversed Brownian motion $\bar{\boldsymbol{W}}_t$. Here, the marginals p_t are given by convolving the target distribution p_{data} with a standard Gaussian: $p_t(\boldsymbol{x}) = \int p_{\text{data}}(\boldsymbol{y})\mathcal{N}(\boldsymbol{x}; \sqrt{\bar{\alpha}_t}\boldsymbol{y}, (1 - \bar{\alpha}_t)\boldsymbol{I})d\boldsymbol{y}$. Since access to the true score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ is usually unavailable in practice, one instead learns the score field $s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$. Then, discretizing Eq. (2) with s_θ, using Euler-Maruyama's method, provides a stochastic algorithm to sample from target distribution p_{data}, beginning with a base instance $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$.

For DDIM, we define the (reverse-time) probability flow ordinary differential equation (PF ODE):

$$\dot{\boldsymbol{x}}_t = -\frac{1}{2}\beta_t(\boldsymbol{x}_t + \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)), \tag{3}$$

where the marginals p_t are the same as those in Eq. (2). Notice that Eq. (3) is exactly Eq. (2) without the noise term, up to a constant multiplication of the score. Similarly, one estimates the score field $s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$ and discretizes Eq. (3) with a first-order exponential integrator (Lu et al., 2022) to obtain a *deterministic* sampling method for p_{data} (Song

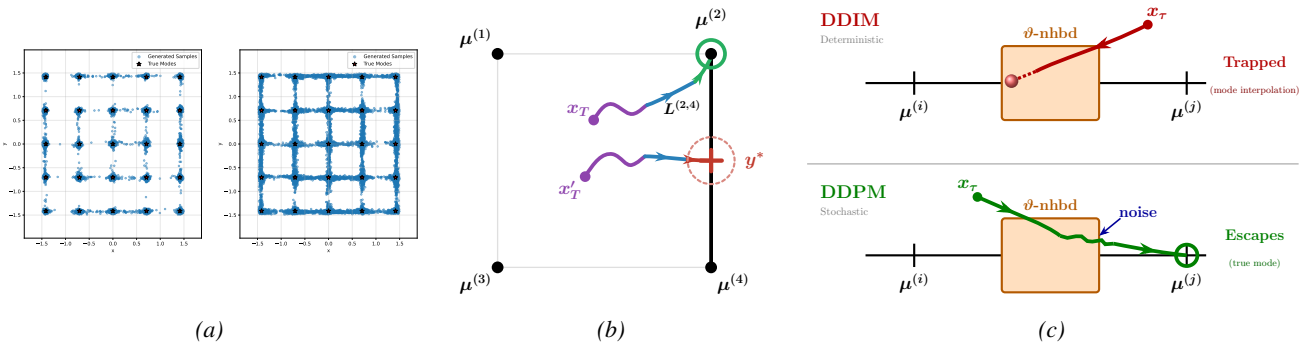


Figure 1. (a) In 100,000 generated samples for a 25-mode Gaussian mixture target, despite using the *same pretrained model*, DDPM (left) hallucinates significantly less than DDIM (right). (b) Towards the beginning of the reverse process, the trajectory selects a line segment to converge to. After that, the trajectory converges rapidly to the nearest line segment: either the true mode or the midpoint neighborhood. (c) Within the midpoint neighborhood, DDIM can get stuck and hallucinate, while DDPM can escape with added noise and avoid hallucination.

et al., 2021a). In the main paper, we theoretically study the dynamics of Eq. (3) and Eq. (2) assuming access to the exact score function. We discuss this regularity condition in Sec. I. We also provide an empirical study under the inexact score function in the main paper experiments (Sec. 6) and theoretically characterize the inexact score function dynamics in Sec. C.

Finally, we consider a Gaussian mixture target distribution of N components, $p_{\text{data}}(\boldsymbol{x}) := \sum_{i=1}^{N} \pi_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(i)}, \sigma^2 \boldsymbol{I})$. Applying linearity of convolution yields that $p_t(\boldsymbol{x}) = \sum_{i=1}^{N} \pi_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_t^{(i)}, \sigma_t^2 \boldsymbol{I})$, where $\boldsymbol{\mu}_t^{(i)} := \sqrt{\bar{\alpha}} \boldsymbol{\mu}^{(i)}$ and $\sigma_t^2 = \sigma^2 \bar{\alpha}_t + (1 - \bar{\alpha}_t)$. We then denote $\tilde{\sigma}_t^2 := \frac{\sigma_t^2}{\bar{\alpha}_t}$ as the effective variance; note that this starts large at the beginning of the reverse process, but becomes σ^2 at the end, i.e, the variance of an individual Gaussian mode. In the main paper, we consider the equal weights case where $\pi_i = \pi_j \; \forall i, j \in [N]$; we extend this to the unequal weights case in Sec. B of the appendix.

3. Related Work

Diffusion Models: Diffusion models generate samples by reversing a noise corruption process, with DDPM (Ho et al., 2020) and its deterministic fewer-step counterpart DDIM (Song et al., 2021a). Score-based generative modeling unifies DDPM-style reverse SDEs and DDIM-style PF ODEs (Song et al., 2021b); recent theory studies the gap between ODE- and SDE- sampling distributions (Deveney et al., 2025), convergence guarantees for DDIM vs. DDPM (Beyler & Bach, 2025), and statistical guarantees on the PF ODE (Chen et al., 2023; Cai & Li, 2025). Recent work also explores differences between the ODE and SDE in terms of performance (Cao et al., 2023) and proposes hybrid samplers (Xu et al., 2023). However, these works do not explicitly study hallucinations. Additionally, instead of studying how ODE/SDE trajectories evolve, they compare statistical distances (e.g., Wasserstein distance (Deveney et al., 2025)). These analyses do not provide insight into *where* and *when* hallucinations arise. To our knowledge, our work is the first to mathematically study the differences in hallucination mechanisms between DDIM and DDPM from the lens of the reverse ODE and SDE.

In our work, we also isolate key times in the reverse process trajectories for DDIM and DDPM. Several recent works isolate key times in the reverse trajectory as well: semantic features emerge in narrow time windows (Li & Chen, 2024) and, given the memorizing score function, there are distinct regimes of feature emergence and memorization (Biroli et al., 2024). Notably, these times and reverse process regimes are obtained under different regularity conditions and differ from ours. Bonnaire et al. (2025) study critical timesteps during diffusion *training* (e.g., the relevant steps of SGD), but this differs from our study of relevant steps in the reverse process.

Aside from quantifying key timesteps in the reverse process, N-mode Gaussian mixtures remain a central controlled testbed for understanding diffusion sampler behavior, with recent theory directly analyzing diffusion guidance (Wu et al., 2024), smoothness (Liang et al., 2025), convergence (Li et al., 2025), memorization (Buchanan et al., 2025), and generalization (Zhang et al., 2026) properties in N-Gaussian-mixture regimes. Our work adds to this thread of research, filling a gap in the mathematical understanding of hallucinations across different types of diffusion inference processes.

Hallucinations in Diffusion Models: A growing line of work defines diffusion hallucinations as "mode interpolations", or sampled instances that are interpolations between modes of the target distribution (Aithal et al., 2024). For example, in our Gaussian mixture setting, modes are the means of the Gaussian components and mode interpolations

are instances which lie on the line segment joining any two modes. This is distinct from "mode collapse" (Zhang et al., 2018; Thanh-Tung & Tran, 2020), which describes a loss of diversity due to collapsing onto a subset of the true modes. In line with these recent works, we use mode interpolation as our working definition of hallucination, finding both empirically and theoretically that DDIM interpolates more often than DDPM. In this work, we assume a *well-separated Gaussian mixture* (Shah et al., 2023). This assumption is natural for any rigorous analysis of mode interpolation hallucinations; if regions of high probability mass around the modes overlap, then there are no mode interpolations since all samples are in these valid regions. Furthermore, while Aithal et al. (2024) attribute mode interpolation solely to the smoothing of the learned diffusion score field, we find that the mechanism underlying mode interpolation can arise even with the exact, non-smoothed score field.

Furthermore, many papers study detection and mitigation of hallucinations (Betti et al., 2026; Chandran.C et al., 2025; Triaridis et al., 2025; Somepalli et al., 2023); in contrast, we study how hallucinations arise specifically due to the choice of inference procedure. This emphasis is further supported by concurrent work showing higher counting hallucination rates for DDIM than DDPM (Fu et al., 2025), distributional differences in ODE- and SDE-based samplers (Deveney et al., 2025; Beyler & Bach, 2025), and empirical evidence that stochastic samplers outperform deterministic ones in scientific generation tasks (Shehata et al., 2025). Additionally, Triaridis et al. (2025) highlight the need to understand the difference in hallucination mechanisms between DDIM and DDPM for future mitigation methods, rather than treating mitigation purely as a post-hoc problem.

4. DDIM Hallucinations

In what follows, we aim to study how DDIM hallucinates. To do so, we first demonstrate that after the variance of p_t becomes sufficiently small, DDIM in Eq. (3) converges to the ε-tube around the nearest i,j-mode segment, which is the line segment joining the two nearest modes $\boldsymbol{\mu}_t^{(i)}$ and $\boldsymbol{\mu}_t^{(j)}$ illustrated in Figure 2. Then, the rest of the DDIM dynamics remain within this tube, parallel to the nearest i,j-mode segment. Thus, if DDIM does not converge directly to one of the modes, then mode interpolation can happen on the i,j-mode segment. Next, we demonstrate that there exists a radius around the midpoint of the nearest i,j-mode segment such that DDIM cannot escape. In Sec. 5, we move to studying DDPM, demonstrating that the noise term in Eq. (2) leads to DDPM escaping the midpoint neighborhood, thus avoiding hallucination.

To begin, we prove that DDIM with Eq. (3) converges to the tube around the nearest i,j-mode segment after sufficient time. Firstly, we define an ε-extended a,b-mode segment for $a,b \in [N]$ as $L_{t,\varepsilon}^{(a,b)} := \{\boldsymbol{\mu}_t^{(b)} + v(\boldsymbol{\mu}_t^{(a)} - \boldsymbol{\mu}_t^{(b)}) : v \in [-\varepsilon, 1+\varepsilon]\}$. When $\varepsilon = 0$, this refers to just the line segment joining the two modes; we denote this by $L_t^{(a,b)}$. Next, we define the orthogonal distance of a trajectory to an ε-extended a,b-mode segment as $d_\perp(\boldsymbol{x}_t, L_{t,\varepsilon}^{(a,b)}) := \inf_{\boldsymbol{y} \in L_{t,\varepsilon}^{(a,b)}} ||\boldsymbol{y} - \boldsymbol{x}_t||_2$ and a ζ-tube around $L_t^{(a,b)}$ as $\text{Tube}_{t,\zeta}^{(a,b)} := \{\boldsymbol{y} \in \mathbb{R}^{\varpi} : \exists \boldsymbol{k} \in L_t^{(a,b)} \text{ s.t. } ||\boldsymbol{y} - \boldsymbol{k}||_2 \leq \zeta\}$. We similarly define the ε-ball around a mode. These key objects are illustrated in Figure 2. We then state our primary assumption.

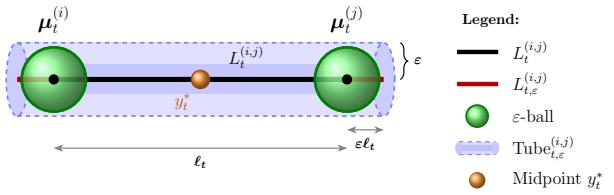


Figure 2. **(1)** In black, we have the line segment $L_t^{(i,j)}$ joining two modes. (2) Together with the red portion, this forms $L_{t,\varepsilon}^{(i,j)}$. (3) We then have the ε-ball surrounding modes i and j. (4) Next, we have $\text{Tube}_{t,\varepsilon}^{(i,j)}$. (5) We also illustrate the midpoint of the line segment $\boldsymbol{y}_t^*$ (where $\boldsymbol{w}_t = 0$), discussed in Prop. 4.7. This provides a high-level description of key objects used throughout Sec. 4, and is not intended to describe the boundary between true and interpolated samples.

Assumption 4.1 (Two-Mode Dominance)**.** *Consider a reverse diffusion process governed by Eq. (3) with Gaussian mixture target distribution p_{data}. Assume that there exists a time $\tau_1 \in (0,T)$ such that for $t \leq \tau_1$ there exists $i,j \in [N]$, $i \neq j$, where, for any $k \notin \{i,j\}$:*

$$||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(k)}||_2^2 \geq \min\left(||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(i)}||_2^2, ||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(j)}||_2^2\right) + \Delta_t,$$

for some margin $\Delta_t \geq 2\sigma_t^2\kappa$ where $\kappa > 1$.

Remark: In the reverse process, the variance of p_t begins large but decays to σ^2 by $t = 0$. Thus, we assume a critical time such that the σ_t is sufficiently small with respect to the distances from the modes. Note that in the case that the trajectory is very close to one particular mode, Asm. 4.1 can hold for several pairs of modes, for example modes i,j and modes i,k. In this case, as we demonstrate later, the instance will simply converge to mode i, thus converging to the line segment between i and any associated mode, including i,j. Otherwise, Asm. 4.1 says that, after sufficient time in the reverse process, the trajectory is nearer to a pair of modes than any other modes. Next, the bound on the margin is saying that this distance is sufficiently large with respect to the variance of the mixture. This well-separation assumption is critical for any analysis of mode interpolations, ensuring that

the rest $N-2$ modes do not have regions of high probability mass which overlap with modes i and j. If this assumption does not hold, a mode interpolation between modes i and j may not be a hallucination and may instead be a high probability instance under p_{data}. Such well-separation conditions are common in diffusion analyses (Shah et al., 2023). We empirically verify this assumption in Figure E.6, finding that it holds for large $\kappa >> 1$.

Given this assumption, we prove our first core result, demonstrating that DDIM converges fast to the ε-tube around the i, j-mode segment after $t < \tau_1$ in Asm. 4.1.

Theorem 4.2 (Convergence to $\text{Tube}_{t,\varepsilon}^{(i,j)}$)**.** *Suppose Asm. 4.1 holds. Then, there exists a time $\tau \leq \tau_1$ such that for all $t \in [0, \tau)$, we have exponential decay of the orthogonal distance to the ε-tube of the nearest i, j-mode segment, i.e.:*

$$d_\perp(\boldsymbol{x}_t, L_{t,\varepsilon}^{(i,j)}) \leq \mathcal{O}\left(d_\perp(\boldsymbol{x}_T, L_{T,\varepsilon}^{(i,j)}) \exp(-u(t)) + \varepsilon\right),$$

where $\varepsilon \in \mathcal{O}(N \exp(-\kappa))$ and $u(t) := \int_t^T \frac{\beta_s}{2\sigma_s^2}\, ds$, which is monotonically increasing as $t \to 0$. This yields that, for any $\phi \in (0, 1)$, with probability at least $1 - 2\exp(-(\varpi)\phi^2/8)$, we have that:

$$d_\perp(\boldsymbol{x}_t, L_{t,\varepsilon}^{(i,j)}) \leq \mathcal{O}\left(\sqrt{(\varpi)(1+\phi)} \exp(-u(t)) + \varepsilon\right).$$

Proof Sketch: Rescale Eq. (3) and then compute the dynamics around the nearest i, j-mode segment. Then, decouple the dynamics into parallel and orthogonal components. Finally, solve a first-order linear ODE for the orthogonal component and show that the parallel component remains restricted to only the ε-extended nearest i, j-mode segment. A full proof is provided in Sec. H.1.

Note that the rate of convergence to $\text{Tube}_{t,\varepsilon}^{(i,j)}$ is damped by the initial distance from the segment, which scales with $\sqrt{\varpi}$ with high probability. Next, note that after τ, if the trajectory enters $\text{Tube}_{t,\varepsilon}^{(i,j)}$, the trajectory stays there.

Corollary 4.3 (Trapping in $\text{Tube}_{t,\varepsilon}^{(i,j)}$)**.** *For any $t \in [0, \tau]$, where τ is defined in Theorem 4.2, if $\boldsymbol{x}_t$ governed by the dynamics of Eq. (3) enters $\text{Tube}_{t,\varepsilon}^{(i,j)}$, where $\varepsilon \in \mathcal{O}(N \exp(-\kappa))$, it stays there.*

As such, once the instance is in $\text{Tube}_{t,\varepsilon}^{(i,j)}$, the only nontrivial component of the dynamics are the parallel dynamics to the nearest i, j-mode segment $L_{t,\varepsilon}^{(i,j)}$ within $\text{Tube}_{t,\varepsilon}^{(i,j)}$. We thus study these parallel dynamics; however, before doing so, we introduce a new assumption.

Assumption 4.4 (Modes Sufficiently Far Apart)**.** *Consider a reverse diffusion process governed by Eq. (3) with Gaussian mixture target distribution p_{data}. Then, consider time $\tau_2 \in (0, T)$, such that for $t \leq \tau_2$:*

$$\ell^2 := ||\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)}||_2^2 \geq 4\kappa\tilde{\sigma}_t^2\varpi,$$

where $\kappa > 1$ from Asm. 4.1 and $\tilde{\sigma}_t^2 = \frac{\sigma_t^2}{\bar{\alpha}_t}$.

Remark: Here, we assume that the effective variance is small enough such that the *distance between the two modes of interest*, $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\mu}^{(j)}$, is larger. Furthermore, ℓ is *not* time-dependent. While κ in Asm. 4.4 could be different, we take it to be the same as Asm. 4.1 for simplicity. Asm. 4.4 scales with dimension ϖ in order to ensure regions of high probability mass do not overlap between modes i and j in high dimensions (Saremi & Hyvärinen, 2019), so that mode interpolations remain a valid definition of hallucination. We empirically verify this assumption in Figure E.6, finding that it holds for large $\kappa >> 1$.

Next, we find that there are stable equilibria near the true modes:

Proposition 4.5 (Mode Stability)**.** *Suppose Asm. 4.1 and Asm. 4.4 hold. Consider $t \leq \hat{\tau}$ where $\hat{\tau} = \min\{\tau, \tau_2\}$. Then, with respect to the parallel dynamics to $L_{t,\varepsilon}^{(i,j)}$ while the trajectory is in $\text{Tube}_{t,\varepsilon}^{(i,j)}$, there exist instantaneous stable equilibria in a ε-ball around modes $\boldsymbol{\mu}_t^{(i)}$ and $\boldsymbol{\mu}_t^{(j)}$, where $\varepsilon \in \mathcal{O}(N \exp(-\kappa))$, for $\kappa \geq 2\log N + \Theta(1)$.*

Proof Sketch: Apply the intermediate value theorem to show existence, then bound the drift derivative to check stability. See Sec. H.3 for a full proof.

Remark: Note that these parallel dynamics are the *exact* parallel dynamics in Eq. (G.50), incorporating an additional term due to the influence of the $N-2$ other modes, which is bounded by ε.

Furthermore, trajectories in a neighborhood of these instantaneous equilibria converge to a *stable trajectory* around the equilibria.

Corollary 4.6 (Tracking Stability)**.** *With the same conditions as Prop. 4.5, there exists a stable trajectory in a neighborhood around $\xi_t^{*,i}$ such that any two trajectories, in that same neighborhood, converge to the stable trajectory. This holds similarly for $\xi_t^{*,j}$.*

Remark: In particular, this means that trajectories which end up near $\xi_t^{*,i}$ stay near $\xi_t^{*,i}$, and same for $\xi_t^{*,j}$.

Next, we find that there is a neighborhood around the midpoint for which the parallel dynamics are *trapped*. Effectively, if an instance is trapped around the midpoint, this yields mode interpolation.

Proposition 4.7 (Trapping in Midpoint Neighborhood)**.** *Suppose Asm. 4.1 and Asm. 4.4 hold. Let $\hat{\tau} = \min\{\tau, \tau_2\}$. Furthermore, suppose $\pi_i = \pi_j$. Then:*

1. *For $t \le \hat{\tau}$, consider the approximate parallel dynamics in $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$, and denote these dynamics by ξ_t. Then, the point $\boldsymbol{y}_t^* = \frac{\boldsymbol{\mu}^{(i)}+\boldsymbol{\mu}^{(j)}}{2} + \boldsymbol{w}_t$, where $\boldsymbol{w}_t$ is the component of Eq. (3) orthogonal to $L_{t,\varepsilon}^{(i,j)}$, is a hyperbolic (instantaneously) unstable equilibrium with respect to these dynamics, with unstable eigenvalue $\lambda_t := \frac{\ell^2}{4\bar{\sigma}_t^2} - 1$. Note that $\boldsymbol{y}_t^*$ is exactly where $\xi_t = \frac{1}{2}$.*

2. *Suppose $\tau_3 \le \hat{\tau}$. Consider $\xi(\cdot)$ as the approximate parallel dynamics in $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$. Choose a (nondimensionalized) radius $\vartheta \in (0, \frac{1}{2}]$ satisfying $0 < \vartheta < \frac{\underline{\lambda}}{2(1+\bar{\lambda})^2}$, where $\underline{\lambda} := \min_{\{t\in[0,\tau_3]\}} \lambda_t$ and $\bar{\lambda} := \max_{\{t\in[0,\tau_3]\}} \lambda_t$. Let $\Lambda_+ := \bar{\lambda} + 2(1+\bar{\lambda})^2\vartheta$. Then, if $0 < |\xi_{\tau_3} - \frac{1}{2}| < \vartheta \exp(-\Lambda_+\tau_3)$, then:*

$$|\xi_t - \frac{1}{2}| < \vartheta \text{ for every } t \in [0, \tau_3].$$

Equivalently, to possibly exit the midpoint neighborhood of radius ϑ, a necessary condition is:

$$\tau_3 \ge \frac{1}{\Lambda_+} \log\left(\frac{\vartheta}{|\xi_{\tau_3} - \frac{1}{2}|}\right).$$

Proof Sketch: Compute the root of the approximate parallel dynamics and check its derivative's positivity. Compute its unstable Jacobian eigenvalue λ_t. Bound the second derivative of the parallel dynamics and use this to bound the derivative of $|\xi_t - \frac{1}{2}|$; applying Grönwall's and leveraging our choice of ϑ yields the result. A full proof is provided in Sec. H.5.

Discussion: This proposition demonstrates that, for a sufficiently small time budget τ_3, DDIM trajectories cannot leave a ϑ-neighborhood of the midpoint. Even though the midpoint repels the instance along the line as an unstable equilibria, it does not do so sufficiently fast enough to reach the true modes. Additionally, note that if $\boldsymbol{w}_t$ is nonzero, this becomes a hyperbolic saddle: the negative eigenvalues represent attraction to $L_t^{(i,j)}$. This corresponds with the fact that the component of Eq. (3) orthogonal to $L_t^{(i,j)}$ vanishes in sufficient time per Theorem 4.2. We empirically demonstrate this in Figure E.12.

Note that we use the "approximate" parallel dynamics given in Eq. (H.234) to prove this result. The approximate dynamics ignore the ε term induced by the $N-2$ modes. In Prop. B.1, we prove that using the exact parallel dynamics in Eq. (G.50) only shifts the location of the point which trajectories get stuck near. Empirically, as demonstrated in Figure E.12, we find that the location shift is negligible. We provide a comprehensive theoretical study of this in Sec. B. Additionally, the saddle still exists under score estimation error, as demonstrated in Sec. C.

In particular, this defines two regimes of convergence in the diffusion reverse process for the N component Gaussian mixture, which we illustrate in Figure 1: (a) After τ, $\boldsymbol{x}_t$ moves fast towards $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$ (b) If $\boldsymbol{x}_t$ converges outside a ϑ-neighborhood of the midpoint, near one of the true modes $\boldsymbol{\mu}_t^{(i)}$ or $\boldsymbol{\mu}_t^{(j)}$, it converges fast to this true mode. Otherwise, if it converges near the midpoint $\boldsymbol{y}_t^*$, the trajectory gets "stuck" and requires large remaining time τ_3 to leave.

5. DDPM Hallucinations

Recall that DDPM follows the reverse SDE Eq. (2), which shares the same drift as the PF ODE Eq. (3) (up to a constant factor) but includes an additional Brownian forcing term. Consequently, DDPM exhibits the same two qualitative convergence stages as DDIM, described in Sec. 4, which we also verify empirically in Sec. 6. Here we focus on the final stage, when the trajectory is governed by the parallel dynamics within $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$. In this regime, we show that DDPM is substantially less likely than DDIM to terminate in the midpoint neighborhood responsible for mode interpolation hallucinations.

For modes $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\mu}^{(j)}$, we define their bisector hyperplane as $H^{(i,j)} := \left\{\boldsymbol{y} \in \mathbb{R}^{\varpi} : \|\boldsymbol{y} - \boldsymbol{\mu}^{(i)}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{\mu}^{(j)}\|_2^2\right\}$. Inside $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$, Theorem 4.2 implies that after τ, the relevant motion reduces to a one-dimensional displacement from $H^{(i,j)}$ along its unit normal. In what follows, we make this reduction explicit and then bound the probability of ending near the midpoint.

Formally, let $\boldsymbol{u} := \frac{\boldsymbol{\mu}^{(j)}-\boldsymbol{\mu}^{(i)}}{\|\boldsymbol{\mu}^{(j)}-\boldsymbol{\mu}^{(i)}\|_2}$ and define the signed bisector coordinate $A_t := \left\langle \boldsymbol{x}_t - \frac{1}{2}(\boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(j)}), \boldsymbol{u}\right\rangle$ and note that $A_t = 0$ if and only if $\boldsymbol{x}_t \in H^{(i,j)}$, ranging from $-\frac{(\ell+\varepsilon)}{2}$ and $\frac{\ell+\varepsilon}{2}$ along the ϵ-extended i,j-mode segment. Applying Itô's formula to A_t and defining the one-dimensional Brownian motion $dB_t := \langle \boldsymbol{u}, d\bar{\boldsymbol{W}}_t\rangle$ yields:

$$dA_t \;=\; b(A_t, t)\, dt \;+\; \sqrt{2\eta(t)}\, dB_t, \qquad A_T = 0, \quad (4)$$

where $\eta(t) = \frac{1}{2}\beta_t$ and $b(A_t, t) = \left\langle -\frac{1}{2}\beta_t \boldsymbol{x}_t - \beta_t \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t), \boldsymbol{u}\right\rangle$. We then leverage this SDE to bound the probability of DDPM terminating in a ϑ-neighborhood of $\boldsymbol{y}_t^*$.

Proposition 5.1 (DDPM Terminal Midpoint Bound)**.** *Let A_t satisfy Eq. (4) on $t \in [0, \tau_3]$, with $\eta \in L^1([0, \tau_3])$ and $\eta(t) \le \eta_{\max}$ for all $t \in [0, \tau_3]$. Fix $\vartheta > 0$. Assume there exists a measurable function $K : [0, \tau_3] \to [0, \infty)$ such that for every $t \in [0, \tau_3]$ and every $a \in [-2\vartheta, 2\vartheta]$,*

$$|b(a, t)| \le K(t)\, |a|. \quad (5)$$

Assume further there exists $\lambda_{\mathrm{rep}} > 0$ such that for all $t \in$

$[0, \tau_3]$,

$$\begin{aligned} b(a,t) &\le -\lambda_{\text{rep}}\, a \qquad \forall\, a \ge \vartheta, \\ b(a,t) &\ge -\lambda_{\text{rep}}\, a \qquad \forall\, a \le -\vartheta. \end{aligned} \tag{6}$$

Then the terminal midpoint event satisfies

$$\mathbb{P}\big(|A_0| \le \vartheta\big) \le \frac{4}{\pi} \exp\left(- \frac{\pi^2 \int_0^{\tau_3} \eta(s)\, ds}{16\Big(1 + \int_0^{\tau_3} K(s)\, ds\Big)^2 \vartheta^2} \right) + 2 \exp\Big(- \frac{\lambda_{\text{rep}}\, \vartheta^2}{2\eta_{\max}} \Big). \tag{7}$$

Proof Sketch: We upper bound the probability of terminating in the midpoint neighborhood by splitting trajectories into two cases and combining their bounds: (i) those that remain within a 2ϑ-slab around the bisector for the entire reverse-time interval, and (ii) those that exit this slab. On the *confinement event* (i), Eq. (5) implies that staying trapped forces the stochastic martingale term to also remain trapped. Using Dubins & Schwarz (1965), we view this martingale as a time-changed Brownian motion with time horizon given by its quadratic variation $2\int_0^T \eta(s)\, ds$, and a standard Brownian confinement estimate yields the first exponential term. On the *exit event* (ii), to still terminate near the midpoint, the path must leave to distance 2ϑ and later return to within ϑ. The strong Markov property at the exit time reduces this to a return probability from $\pm 2\vartheta$. Under Eq. (6), an exponential supermartingale and optional stopping bound this return, giving the second exponential term. A full proof can be found in Sec. H.6.

Discussion. Prop. 5.1 is the DDPM analogue of the DDIM midpoint persistence result in Prop. 4.7. After trajectories enter $\text{Tube}_{t,\varepsilon}^{(i,j)}$, DDIM's remaining failure mode is to terminate near the bisector, producing mode interpolation samples. This proposition shows that DDPM makes this terminal midpoint event exponentially unlikely, so trajectories in the stage where they are within $\text{Tube}_{t,\varepsilon}^{(i,j)}$ overwhelmingly tend to end near one of the two true modes instead. We validate this empirically in Sec. 6. The assumptions of Prop. 5.1 are further detailed in Sec. F and are verified in Sec. E.7.

6. Experiments

We provide experiments that justify our theoretical assumptions; demonstrate our theoretical results; and validate our subsequent interpretations. We primarily experiment with a Gaussian mixture consisting of 25 modes arranged on a uniform grid in 2 dimensions; we study this dataset with higher dimensions in Sec. E.2. Unless otherwise stated, we use 1000 DDPM steps ($T = 1000$) and 50 DDIM sampling steps with quadratic skipping. Firstly, we show that mode interpolation is a primary source of hallucinations during sampling. We also demonstrate that the high hallucination rate gap between DDIM and DDPM is *not* due to DDIM skipping steps (i.e., PF ODE discretization) in the reverse process. Finally, we demonstrate that DDPM noise helps escape the ϑ-neighborhood around the midpoint, thus avoiding hallucinations, as predicted by our theoretical results. Further experimental details are provided in Sec. F; we also provide additional experiments on images as well as ablations in Sec. E.

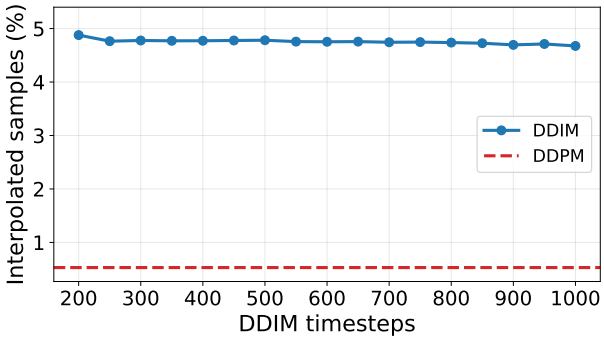


Figure 3. Hallucination rate for varying number of DDIM steps used in the reverse process. Notice that the number of DDIM interpolated samples is consistently larger than that of DDPM. Thus, this invalidates the idea that the gap between DDIM and DDPM hallucination rates arises due to skipping steps.

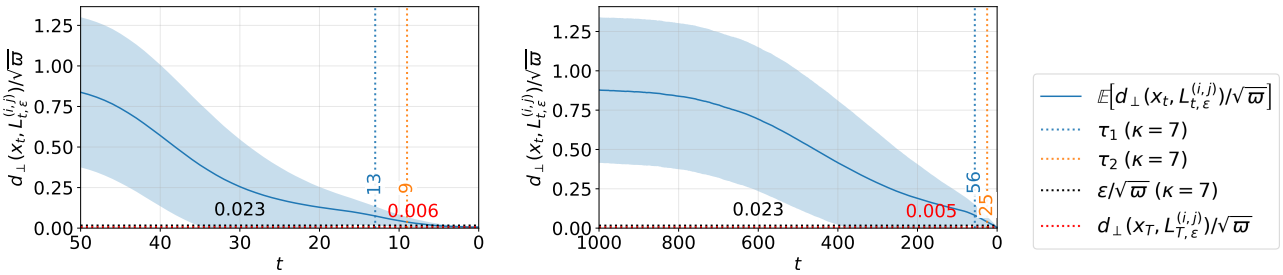
DDIM Hallucinates More than DDPM: In Figure 1a, we find that DDIM hallucinates more than DDPM.

We classify a sample as a *mode interpolation* if it lies more than 5σ from every true mode but within 5σ of a line segment joining two modes. Note that we omit $< 0.01\%$ of instances which land outside 5σ of any line segment; these instances arise due to score error, and we do not focus on them for the purposes of this work. We expand on our choice of threshold and characterization of these samples in Sec. F. Additionally, we reproduce the recent results of Fu et al. (2025) in Sec. E.3, demonstrating that DDIM has higher counting hallucinations than DDPM as well for an image dataset.

Crucially, in Figure 3, we demonstrate that the significantly higher hallucination rate of DDIM is *not due to skipping steps*. Specifically, from coarse ODE discretizations to fine ODE discretizations, the hallucination rate of DDIM remains larger than that of DDPM.

Finally, we observe that for an axis-aligned grid, hallucinations rarely happen on the diagonals joining two modes. This is also observed by Aithal et al. (2024). We provide theoretical justification for this in Sec. D.

Convergence to Nearest i, j-Mode Segment: In Figure 4, we validate Theorem 4.2 empirically, firstly finding that As-



(a) DDIM convergence to the i, j-mode segment. *(b)* DDPM convergence to the i, j-mode segment.

Figure 4. For both DDIM (Figure 4a) and DDPM (Figure 4b), we plot the convergence rate to the nearest i, j-mode segment across 100,000 trajectories, finding that convergence occurs after τ_1 and thus validating Theorem 4.2. Note that i, j change across time in these figures; however, as expected, after τ_1 they become fixed. We plot ε/ϖ as a dotted black line, finding that convergence to $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$ is after τ_2; thus, only the parallel dynamics become relevant at this stage. We also find that DDPM ends up closer to the nearest i, j-mode segment (dotted red line). Note that DDIM has 50 steps due to quadratic skipping. Uncertainty is quantified in the blue region as the 95% confidence interval across trajectories.

sumptions 4.1 and 4.4 hold for large $\kappa = 7$. In Sec. E.1, we find that these assumptions hold for a range of other large κ as well. We then find that DDIM and DDPM converge to the ε-extended nearest i, j-mode segment after τ_1, where $\varepsilon = N\exp(-\kappa)$. Note that this line segment varies with time, and i, j can change over the reverse process. Importantly, to verify Theorem 4.2, we compute the distance (and ε) rescaled by dividing $\sqrt{\varpi}$. We also find that the trajectories pass into the ε-tube after τ_2, empirically verifying that the phenomena described in Prop. 4.7 are relevant as soon as the trajectory enters the ε-tube. These results are for $\varpi = 2$; results for higher dimensions are in Sec. E.2.

DDPM Noise Helps Avoid Hallucinations: In what follows, we aim to demonstrate that Prop. 4.7 and Prop. 5.1 hold for a nontrivial radius ϑ. That is, for $\tau_3 = 9$ (since $\tau_2 = 9$ in Figure 4) remaining steps, DDIM trajectories get stuck within a ϑ-neighborhood of the midpoint $\boldsymbol{y}_t^*$ while DDPM trajectories escape. Specifically, we start DDIM and DDPM (with appropriately adjusted β_t) along $L_t^{(a,b)}$ at time τ_3 and evaluate the hallucination rate after starting at this point. We do so for 100,000 trajectories across all $a, b \in [N]$. Critically, in Figure 5, we find that there exists a $\vartheta(\tau_3) > 0$ such that if a trajectory starts within $\vartheta(\tau_3)$ of the midpoint, it hallucinates. Next, the figure demonstrates that within this radius, DDPM has a significantly lower hallucination rate than DDIM. We thus conclude that the noise term in DDPM helps it escape the ϑ-neighborhood around the midpoint, thus avoiding hallucinations. In Figure 5, we leverage this insight to evaluate the following DDIM-DDPM hybrid procedure: we add z DDPM steps to DDIM after τ_3, and find that this helps avoid hallucinations as well. In Sec. E.4, we demonstrate that our results hold across various choices of τ_3. Note that τ_3 is taken with respect to DDIM steps throughout our experiments.

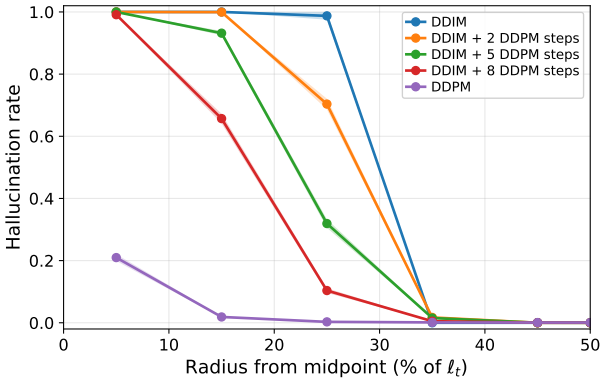


Figure 5. Starting DDIM at $\tau_3 = 9$, we find that for $\vartheta = 0.15\ell_t$, DDIM gets stuck before it can reach the true modes, i.e., it hallucinates, as predicted by Prop. 4.7. Furthermore, DDPM has a lower hallucination rate within this same ϑ. Thus, we conclude that DDPM noise helps escape the ϑ-neighborhood around the midpoint, as predicted by Prop. 5.1. Given this, we find that adding z DDPM steps after starting DDIM at τ_3 helps mitigate hallucination. Uncertainty is quantified by the standard error across mode pairs.

7. Discussion

We present a geometric analysis of hallucinations in diffusion models through the lens of Gaussian mixtures. In particular, after sufficient time, diffusion trajectories undergo two key convergence phases: rapid collapse onto a nearby line segment connecting two modes, followed by dynamics along this line. Next, if a trajectory does not converge to one of the modes, it can get stuck near a point; in the case that Gaussian mixture weights are equal, this is at the midpoint. This provides insight into the core mechanisms which differ between DDIM and DDPM: stochastic noise can help DDPM escape the neighborhood of the midpoint to one of the true modes and avoid hallucination.

Our results thus imply that the noise term is beneficial for reducing mode interpolation hallucinations. Empirically, we find this to be the case as well: adding only a few diffusion steps at the end of the reverse process significantly lowers the hallucination rate. We hope that these results highlight to the community the pitfalls of preferring DDIM over DDPM and inspire future work on designing flow-based generative models which avoid the midpoint neighborhood entirely, thus mitigating hallucinations.

Limitations and Future Work: Our primary analysis assumes a Gaussian mixture target distribution with well-separated modes. While we prove some results under an inexact score in Sec. C, a full characterization of what changes under score error is left to future work. Additionally, Aithal et al. (2024) highlight that mode interpolation can also happen in image latent space. Thus, a natural extension of our theoretical analysis to images is studying latent diffusion (Rombach et al., 2022) with Gaussian mixture latents. Furthermore, we believe that our work provides a first step towards designing adaptive samplers or models which detect when trajectories are stuck and inject DDPM noise accordingly, thus mitigating hallucinations while keeping the computational benefits of DDIM.

Broader Impacts

Hallucinations, generations that violate structural or semantic constraints of the training distribution, are a key obstacle in the adoption of state-of-the-art generative models, like diffusion, in safety critical contexts. Our work contributes to removing this obstacle by giving a mechanistic account of one category of hallucination: mode interpolation, where trajectories terminate between modes of the target distribution. We identify why DDIM, a commonly used diffusion inference procedure, has a higher hallucination rate than its stochastic counterpart, DDPM. Concretely, the noise in the DDPM reverse process can help it get unstuck from a region around the midpoint, where DDIM would otherwise hallucinate. Empirically, we highlight that even mild noise injection can be beneficial, guiding practitioners to safer generative modeling. Altogether, we believe our work has positive societal impact.

References

Aithal, S. K., Maini, P., Lipton, Z. C., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 1982.

Betti, F., Baraldi, L., Baraldi, L., Cucchiara, R., and Sebe, N. Hallucination early detection in diffusion models. *International Journal of Computer Vision (IJCV)*, 2026.

Beyler, E. and Bach, F. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance. *arXiv 2508.03210*, 2025.

Biroli, G., Bonnaire, T., de Bortoli, V., and Mézard, M. Dynamical regimes of diffusion models. In *Nature Communications*, 2024.

Bonnaire, T., Urfin, R., Biroli, G., and Mezard, M. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Buchanan, S., Pai, D., Ma, Y., and Bortoli, V. D. On the edge of memorization in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Cai, C. and Li, G. Minimax optimality of the probability flow ode for diffusion models. *arXiv 2503.09583*, 2025.

Cao, Y., Chen, J., Luo, Y., and Zhou, X. Exploring the optimal choice for generative processes in diffusion models: Ordinary vs stochastic differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chandran.C, B., Anumasa, S., and Liu, D. Laplacian score sharpening for mitigating hallucination in diffusion models. *arXiv 2511.07496*, 2025.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ode is provably fast. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Deveney, T., Stanczuk, J., Kreusser, L., Budd, C., and Schönlieb, C.-B. Closing the ode-sde gap in score-based diffusion models through the fokker-planck equation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2025.

Dubins, L. E. and Schwarz, G. On continuous martingales. *Proceedings of the National Academy of Sciences of the United States of America*, 1965.

Fernandez-Granda, C. Lecture 3: Nyu optimization-based data analysis lecture notes, 2017.

Fu, S., Zhou, J., Chen, Q., Jing, H., Nguyen, H. A., Liu, X., Zeng, Z., Ma, L., Zhang, Q., and Wu, Q. Counting hallucinations in diffusion models. *arXiv 2510.13080*, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Li, G., Cai, C., and Wei, Y. Dimension-free convergence of diffusion models for approximate gaussian mixtures. *arXiv 2504.05300*, 2025.

Li, M. and Chen, S. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

Liang, Y., Shi, Z., Song, Z., and Zhou, Y. Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. *International Conference on Computer Vision (ICCV)*, 2025.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Lu, R., Wang, R., Lyu, K., Jiang, X., Huang, G., and Wang, M. Towards understanding text hallucination of diffusion models via local generation bias. In *International Conference on Learning Representations (ICLR)*, 2025.

Mörters, P. and Peres, Y. *Brownian Motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Saremi, S. and Hyvärinen, A. Neural empirical bayes. *Journal of Machine Learning Research*, 2019.

Shah, K., Chen, S., and Klivans, A. Learning mixtures of gaussians using the DDPM objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shehata, Y., Holzschuh, B., and Thuerey, N. Improved sampling of diffusion models in fluid dynamics with tweedie's formula. In *International Conference on Learning Representations (ICLR)*, 2025.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Thanh-Tung, H. and Tran, T. Catastrophic forgetting and mode collapse in gans. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

Triaridis, K., Graikos, A., Chatziagapi, A., Chrysos, G. G., and Samaras, D. Mitigating diffusion model hallucinations with dynamic guidance. *arXiv 2510.05356*, 2025.

Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *International Conference on Machine Learning (ICML)*, 2024.

Xu, Y., Deng, M., Cheng, X., Tian, Y., Liu, Z., and Jaakkola, T. S. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhang, Z., Li, M., and Yu, J. On the convergence and mode collapse of gan. *SIGGRAPH Asia Technical Briefs*, 2018.

Zhang, Z., Li, X., Li, X., Shi, L., Wu, M., Tao, M., and Qu, Q. Generalization of diffusion models arises with a balanced representation space. In *International Conference on Learning Representations (ICLR)*, 2026.

Contents of the Appendix

- In Sec. A we provide a full list of notation.
- In Sec. B, we provide theoretical results characterizing how our analysis changes if we assume unequally weighted modes (i.e., $\pi_i \neq \pi_j$).
- In Sec. C, we provide theoretical results characterizing how our analysis changes if we assume the learned inexact score, rather than the exact score in our main paper.
- In Sec. D, we provide a theoretical result characterizing why mode interpolations rarely happen on the diagonal lines joining two modes in the axis-aligned grid we use throughout our experiments.
- In Sec. E, we provide additional experiments, including verifying assumptions, empirically evaluating higher dimensional Gaussian mixtures and images, and running ablations on several hyperparameters in our experiments.
- In Sec. F, we provide implementation details for our experiments.
- In Sec. G we provide helpful lemmas, which are then used in our proofs in Sec. H.
- In Sec. I, we discuss the use of the exact score throughout our paper.
- In Sec. J, we provide a table of symbols used throughout our paper.

A. Full Notation

We let $x \in \mathbb{R}$ denote a scalar. Bolded lowercase letters $\boldsymbol{x} \in \mathbb{R}^{\varpi}$ denote vectors of dimension ϖ. Bolded upper case letters $\boldsymbol{X} \in \mathbb{R}^{\varpi \times \varpi}$ denote matrices with ϖ rows and columns, unless dimensionalities are explicitly specified. $[N]$, for scalar $N \in \mathbb{R}$, denotes the integer subset $\{1, 2, ..., N\}$. We use $||\boldsymbol{x}||_2$ to denote the ℓ_2 norm of a vector $\boldsymbol{x}$. We use $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ to denote the inner product between vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, which is the dot product $\boldsymbol{x}^T \boldsymbol{x}$ unless explicitly specified otherwise. For a function f, we use $\nabla_{\boldsymbol{x}} f$ to denote its gradient if $f : \mathbb{R}^{\varpi} \to \mathbb{R}$ and its Jacobian if $f : \mathbb{R}^{\varpi} \to \mathbb{R}^o$. We denote by $x_t : [0, T] \to \mathbb{R}$ and $\boldsymbol{x}_t : [0, T] \to \mathbb{R}^{\varpi}$ time-dependent scalar-valued and vector-valued functions, respectively, from time scalar time T to 0 unless otherwise specified. We use $\dot{x}_t$ to denote the time derivative of x_t and similarly use $\dot{\boldsymbol{x}}_t$ for $\boldsymbol{x}_t$; for all other derivatives, we use the notation $\frac{d}{dx} f(x)$. $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ is used to denote the probability density function of a ϖ-dimensional Gaussian centered at $\boldsymbol{\mu}$ with variance $\sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. We write stochastic differential equations (SDE) as $d\boldsymbol{x}_t = b(t, \boldsymbol{x}_t)dt + \sigma(t, \boldsymbol{x}_t)d\boldsymbol{W}_t$ where $\boldsymbol{W}_t \in \mathbb{R}^{\varpi \times 1}$ is the standard Brownian motion, which is equivalent to the weak form $\boldsymbol{x}_t = \boldsymbol{x}_0 + \int_0^t b(s, \boldsymbol{x}_s)ds + \int_0^t \sigma(s, \boldsymbol{x}_s)d\boldsymbol{W}_t$. We say that, for $f, g : \mathbb{R}^{\varpi} \to \mathbb{R}^o$, $f \in \mathcal{O}(g)$ if $\exists c_1 > 0$ s.t. $f(\boldsymbol{x}) \leq c_1 g(\boldsymbol{x})$ for x sufficiently large; $f \in \Omega(g)$ if $\exists c_2 > 0$ s.t. $f(x) \geq c_2 g(x)$ for x sufficiently large; and $f \in \Theta(g)$ if $f \in \mathcal{O}(g)$ and $f \in \Omega(g)$.

B. Saddle Existence for Unequal Weights and Exact Parallel Dynamics

In Prop. 4.7, we demonstrated that, after sufficient time in the reverse process, the midpoint was an equilibrium with respect to the approximate parallel dynamics, under equal weights of the i, j-mode segment $\pi_i = \pi_j$. Here:

1. Under unequal weights, we demonstrate a sufficient condition for an equilibrium to exist.
2. Under the exact dynamics, we demonstrate that there exists an equilibrium, and that it does not differ greatly from the midpoint. This is also empirically justified by Figure E.12.

Proposition B.1 (Equilibria Location under Unequal Weights and Exact Dynamics)**.** *Suppose Asm. 4.1, Asm. 4.1 and Asm. 4.4 hold and fix $t \leq \hat{\tau}$ (with $\hat{\tau}$ same as in Asm. 4.4). Consider the induced 1D parallel dynamics inside the tube.*

1. ***(Unequal weights).*** *Let $\ell := \|\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)}\|_2$. For the approximate parallel dynamics in Tube$_{t,\varepsilon}^{(i,j)}$, there exists a point parallel to $L_{t,\varepsilon}^{(i,j)}$, which we denote by $\xi_{ij}^\star(t) \in (0, 1)$, which satisfies:*

$$\log\Big(\frac{\xi_{ij}^\star(t)}{1 - \xi_{ij}^\star(t)}\Big) = \log\Big(\frac{\pi_j}{\pi_i}\Big) + \frac{\ell^2}{\tilde{\sigma}_t^2}\Big(\xi_{ij}^\star(t) - \frac{1}{2}\Big). \tag{B.8}$$

   In particular, $\pi_i = \pi_j$ implies $\xi_{ij}^\star(t) = \frac{1}{2}$, and $\pi_j < \pi_i$ implies $\xi_{ij}^\star(t) > \frac{1}{2}$ (and vice versa).

   Moreover,

$$F'_{ij,t}\big(\xi_{ij}^\star(t)\big) = \frac{\ell^2}{\tilde{\sigma}_t^2}\, \xi_{ij}^\star(t)\big(1 - \xi_{ij}^\star(t)\big) - 1, \tag{B.9}$$

   so $\xi_{ij}^\star(t)$ is hyperbolic unstable whenever $F'_{ij,t}(\xi_{ij}^\star(t)) > 0$.

2. ***(Exact Parallel Dynamics).*** *Under the exact parallel dynamics characterized in Eq. (G.50), between the two stable equilibria near modes $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\mu}^{(j)}$ discussed in Prop. 4.5, there exists an equilibrium $\xi_N^\star(t)$ of the exact parallel dynamics Furthermore, for κ sufficiently large, assume there an interval I containing $\xi_{ij}^\star(t)$ such that, for some $m > 0$:*

$$F'_{ij,t}(\xi) \geq m > 0 \qquad \forall\, \xi \in I, \tag{B.10}$$

   where $F_{ij,t}$ is the drift of the approximate parallel dynamics. Then, $\xi_N^\star(t) \in I$ satisfies

$$\big|\xi_N^\star(t) - \xi_{ij}^\star(t)\big| \leq \varepsilon/m. \tag{B.11}$$

   where $\varepsilon \in \mathcal{O}(N \exp(-\kappa))$.

Proof sketch: Item (1) follows by analyzing the approximate parallel dynamics for a single pair of components: the equilibrium condition yields the stated log-odds relation, and the location and slope follow from monotonicity and differentiation of the logistic responsibility. For item (2), under two-mode dominance the exact parallel dynamics are a uniformly small perturbation of the approximate dynamics. Existence of an interior unstable equilibrium then follows by continuity and the intermediate value theorem, and a mean-value argument yields the stated displacement bound. A full proof is given in Sec. H.7.

Discussion: The key contribution of this result is twofold. First, we show that the exact parallel dynamics admit an interior equilibria rather than assuming its existence. Second, we show that this equilibrium is robust: its location does not deviate significantly under unequal mixture weights or when passing from the approximate to the exact dynamics. Item (1) characterizes how unequal weights bias the saddle along the (i, j)-line toward the lower-weight component. Item (2) shows that under two-mode dominance, the exact parallel dynamics differ from the approximate dynamics by only a small perturbation, so the interior saddle persists and remains close to its pairwise location. Finally, since an equilibrium exists between two stable (attracting) ones (which exist and are stable for the exact parallel dynamics per Prop. 4.5), and the exact parallel dynamics are one-dimensional, intuitively this is an equilibrium which repels i.e. it is unstable. As such, it can also admit the confinement in a radius discussed in Prop. 4.7.

C. Results under Inexact Score

The preceding analysis leverages the closed-form score of Gaussian mixtures. In practice, the score $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is approximated by a neural network $s_\theta(\boldsymbol{x}, t)$. In this section, we characterize how score estimation error affects the mechanism established in Prop. 4.7.

Definition C.1 (Score Error Field)**.** *Define the score error field as*

$$\psi(\boldsymbol{x}, t) := s_\theta(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}). \tag{C.12}$$

The learned reverse dynamics become

$$\dot{\boldsymbol{x}}_t = -\frac{1}{2}\beta_t \left(\boldsymbol{x}_t + \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \psi(\boldsymbol{x}_t, t)\right). \tag{C.13}$$

Assumption C.2 (Bounded Score Error)**.** *There exists $\varrho : [0, T] \to \mathbb{R}_{\geq 0}$ such that $\|\psi(\boldsymbol{x}, t)\|_2 \leq \varrho(t)$ for all $\boldsymbol{x} \in \mathrm{Tube}_{t,\epsilon}^{(i,j)}$.*

Assumption C.3 (Lipschitz Score Error)**.** *There exists $L_\psi : [0, T] \to \mathbb{R}_{\geq 0}$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{Tube}_{t,\epsilon}^{(i,j)}$:*

$$\|\psi(\boldsymbol{x}, t) - \psi(\boldsymbol{y}, t)\|_2 \leq L_\psi(t)\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{C.14}$$

Under these assumptions, we establish that (1) the saddle location shifts under score error, and (2) the unstable eigenvalue can decrease, amplifying the slowdown effect.

Proposition C.4 (Saddle Perturbation)**.** *Suppose Asm. 4.1 and Asm. 4.4 hold. Let $\xi_{ij}^*(t)$ denote the unstable equilibrium of the exact parallel dynamics from Prop. B.1, with unstable eigenvalue $\lambda_t = F'_{ij,t}(\xi_{ij}^*(t)) > 0$.*

Then, for $\varrho(t)$ sufficiently small, the perturbed parallel dynamics admit an equilibrium $\xi_\theta^*(t)$ satisfying

$$|\xi_\theta^*(t) - \xi_{ij}^*(t)| \leq \frac{2\varrho(t)}{\ell\lambda_t} + O(\varrho(t)^2), \tag{C.15}$$

where $\ell = \|\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}\|_2$.

Proof Sketch: Under score error ψ, the perturbed parallel drift becomes $\tilde{F}_{N,t}(\xi) = F_{ij,t}(\xi) + e_t(\xi)$, where $|e_t(\xi)| \leq \varrho(t)/\ell$ by Cauchy-Schwarz. Applying the implicit function theorem at the unperturbed equilibrium and Taylor expanding yields the displacement bound. A full proof is provided in Sec. H.8.

Remark:. *The displacement bound reveals a critical sensitivity: when λ_t is small (near the threshold $\ell \approx 2\tilde{\sigma}_t$), even small score errors induce large saddle displacements. This explains why hallucination locations can be unpredictable in trained models.*

Proposition C.5 (Eigenvalue Perturbation)**.** *Suppose Asm. 4.1, Asm. 4.4, Asm. C.2, and Asm. C.3 hold. The perturbed parallel dynamics have eigenvalue at $\xi_\theta^*(t)$:*

$$\lambda_\theta(t) = \lambda_t + \boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi(\boldsymbol{y}(\xi_\theta^*), t)\, \boldsymbol{u} + O(\varrho(t)). \tag{C.16}$$

In particular:

1. *If $\boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi\, \boldsymbol{u} < 0$ (error is "contractive" along $L_t^{(i,j)}$), then $\lambda_\theta(t) < \lambda_t$.*
2. *If $\boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi(\boldsymbol{y}(\xi_\theta^*), t)\, \boldsymbol{u} \le -\lambda_t - C\varrho(t)$ for some constant $C > 0$, then $\lambda_\theta(t) \le 0$, and the perturbed saddle becomes instantaneously stable.*

Proof Sketch: Differentiating the perturbed drift $\tilde{F}_{N,t}(\xi) = F_{ij,t}(\xi) + e_t(\xi)$ and using $e_t'(\xi) = \boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi\, \boldsymbol{u}$ yields the eigenvalue perturbation. A full proof is provided in Sec. H.9.

Remark: (Connection to Prior Work)**.** *Aithal et al. (2024) argue that hallucinations arise from score smoothing due to neural network approximation. Prop. C.5 is complementary: we show that the* directional *effect of score error (specifically, $\boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi\, \boldsymbol{u}$ along the mode-connecting line) determines whether slowdown is amplified or mitigated. Score error can either increase or decrease hallucination propensity depending on its directional structure.*

We conclude by showing that DDPM's noise mechanism provides robustness against eigenvalue perturbation.

Proposition C.6 (DDPM Robustness to Score Error)**.** *Consider the perturbed DDPM dynamics*

$$dA_t = b_\theta(A_t, t)\, dt + \sqrt{2\eta(t)}\, dB_t, \tag{C.17}$$

where $b_\theta(A_t, t) = b(A_t, t) + \tilde{e}_t(A_t)$ with $|\tilde{e}_t(a)| \le \varrho(t)/\ell$.

Define $\varrho_{\max} := \sup_{t\in[0,\hat{\tau}]} \varrho(t)$. Suppose the exact drift satisfies the escape condition in Prop. 5.1 with constant $\lambda_{\text{rep}} > 0$. If

$$\frac{\varrho_{\max}}{\ell} < \lambda_{\text{rep}} \vartheta, \tag{C.18}$$

then the perturbed drift satisfies an escape condition with constant $\tilde{\lambda}_{\text{rep}} := \lambda_{\text{rep}} - \frac{\varrho_{\max}}{\ell\vartheta} > 0$, and the escape bound holds with λ_{rep} replaced by $\tilde{\lambda}_{\text{rep}}$.

Proof: A full proof is provided in Sec. H.10.

Remark:. *DDIM also has an escape mechanism via the unstable eigenvalue $\lambda_t > 0$. However, by Prop. C.5, score error can reduce λ_t or make it negative, converting the saddle into a stable equilibrium. In contrast, DDPM's escape mechanism is additive (noise-driven) rather than relying solely on eigenvalue structure, providing robustness even when score error reduces λ_t.*

D. Interpolation on Mode Diagonals

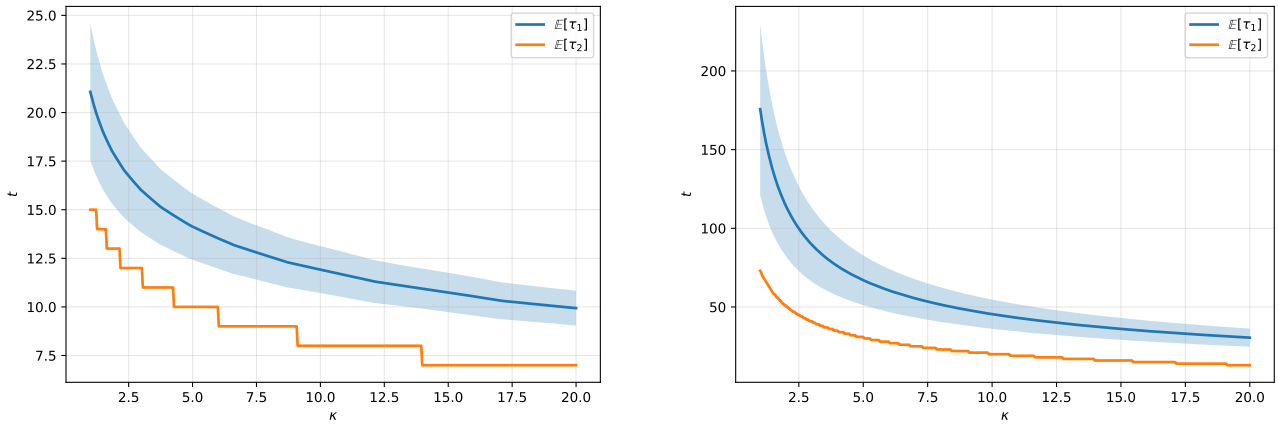
In what follows, we provide a mathematical justification for why trajectories never mode interpolate on diagonals across grids:

Proposition D.1 (Diagonal Avoidance on a Grid)**.** *Suppose Asm. 4.1 holds with the same notation as above, and assume the true modes lie on an axis-aligned rectangular grid. Then, if the (i, j) in Asm. 4.1 corresponds to a diagonal pair of some grid cell, then for every $t \le \tau_1$:*

$$\min\{\tilde{\gamma}_{i,t}, \tilde{\gamma}_{j,t}\} \le \exp(-\kappa), \tag{D.19}$$

, where $\tilde{\gamma}_{i,t}$, $\tilde{\gamma}_{j,t}$ denotes the rescaled responsibilities of modes i, j per Lemma G.2.

Proof Sketch. We can reduce the problem to a single rectangular cell using the implications of Lemma G.3. Suppose, for contradiction, that the dominant pair (i, j) corresponds to a diagonal of this cell and that Eq. (D.19) holds for this pair. Then, using the geometric identity for a rectangle together with Asm. 4.1, we obtain an upper bound on ρ of order $\mathcal{O}(e^{-\kappa})$. A full proof is given in Sec. H.11.



(a) DDIM τ_1 and τ_2 vs. κ.

(b) DDPM τ_1 and τ_2 vs. κ.

Figure E.6. For both DDIM (Figure E.6a) and DDPM (Figure E.6b), we plot how τ_1 and τ_2 change vs. κ over 100,000 sampled trajectories, across 500 equidistant κ increments ranging from 1 to 20. As expected, for large κ, Assumptions 4.1 and 4.4 hold; this yields that ε is small as well. Furthermore, as expected, as κ increases, both of these decrease. Note that τ_2 is the same across all trajectories, since ℓ^2 is the same across all mode i, j pairs, so the standard deviation is 0. Furthermore, $\mathbb{E}[\tau_2]$ is much smoother for DDPM than DDIM, since DDPM has many more sampled timesteps to evaluate τ_2 at; however, since $\mathbb{E}[\tau_1]$ average 100,000 different τ_1, it is smooth across DDPM and DDIM.

Discussion: This proposition is purely geometric, and does not rely on the parallel dynamics. The responsibility of a mode is its probability density proportional to that of all other modes; in particular, if the minimum of $\tilde{\gamma}_{i,t}, \tilde{\gamma}_{j,t}$ is exponentially small, then the trajectory must be very close to the other mode. In particular, by Cor. 4.6, the trajectory *must converge to the other mode*, since it is confined to $\text{Tube}_{t,\varepsilon}^{(i,j)}$. For example, if mode i's responsibility is small, then the trajectory must converge to mode j.

E. Additional Experiments

E.1. Verifying Assumptions: Asm. 4.1 and Asm. 4.4

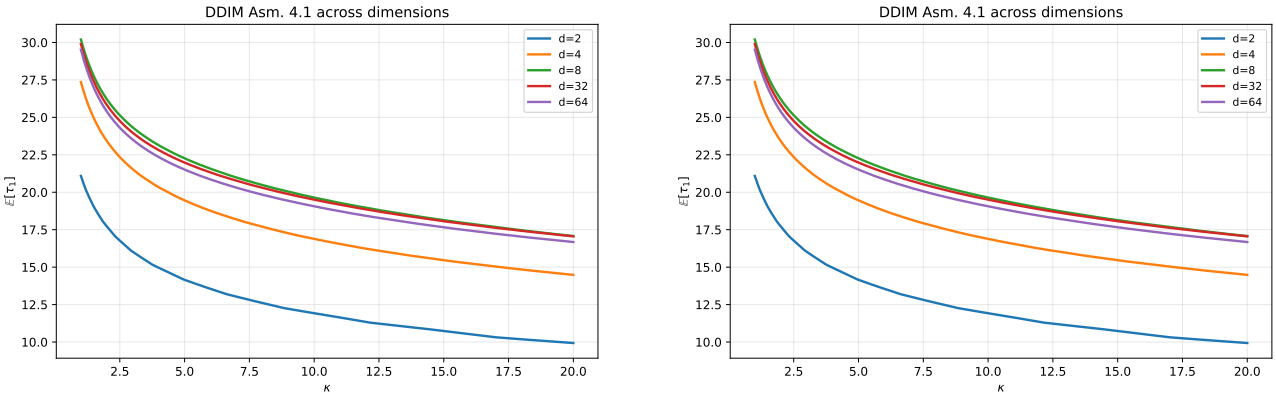
In Figure E.6, for the Gaussian dataset with $\varpi = 2$, we provide results demonstrating that for $\kappa >> 1$, τ_1 and τ_2 exist and are valid. In Sec. E.2, we demonstrate that this holds similarly for higher dimensions. Hence, Asm. 4.1 and Asm. 4.4 hold for $\kappa >> 1$ Note that, consistently, $\tau_2 < \tau_1$; in Figure 4, we show that τ_2 is well before the time that the instance hits the ε-tube, as predicted by Theorem 4.2. Additionally, in u-time per Lemma G.2, Asm. 4.1 and Asm. 4.4 hold similarly.

E.2. Higher Dimensional Gaussian Mixtures

Throughout, we use 25 modes and furthest point sampling, keeping ℓ roughly the same as in the $\varpi = 2$ case. Notably, however, some modes are slightly more or slightly less apart than ℓ, hence adding uncertainty to τ_2 across trajectories as well. In Figures E.7 and E.8, we validate Assumptions 4.1 and 4.4 for $\varpi = 2, 4, 8, 16, 32, 64$. We find that, as dimension increases, τ_1 becomes larger (i.e., nearer to the beginning of the reverse process) and τ_2 becomes smaller (i.e., nearer to the end of the reverse process) for both DDIM and DDPM, hence corresponding to different κ (and thus different ε). Still, our two core assumptions still hold for large κ. In Figure E.9, we valuate Theorem 4.2 for both DDIM and DDPM across the same ϖ increments. For both DDIM and DDPM, we find that exponential convergence still holds. We also find that ε increases for DDIM as the dimension increases. Furthermore, ε increases for DDPM as well, albeit at a much slower rate.

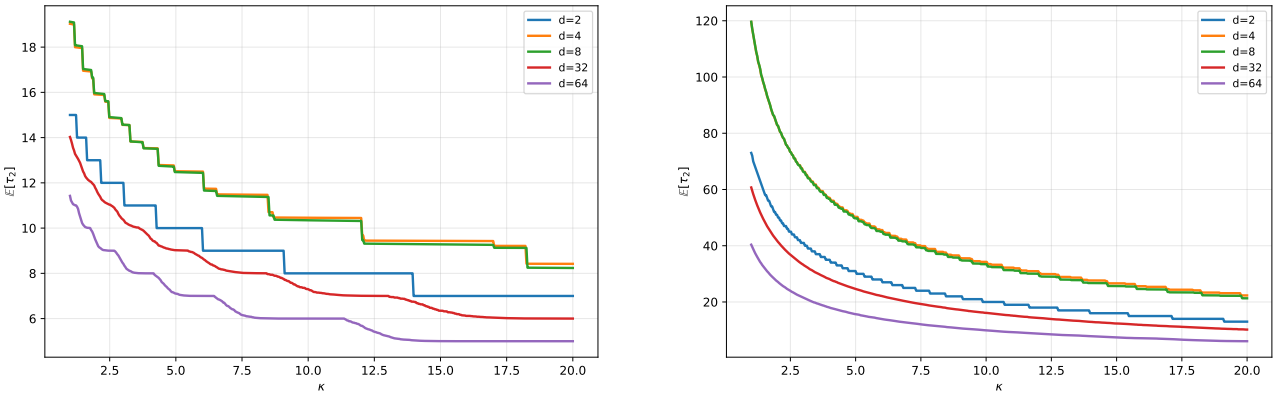
E.3. DDIM Hallucinates More than DDPM For Images

In Figure E.10, we find that for a UNet denoiser pretrained on our 3 triangle dataset described in Sec. F, across various choices of training epochs, DDIM consistently hallucinates more than DDPM.



(a) DDIM *(b)* DDPM

Figure E.7. τ_1 vs. κ for varying dimensions. We find that τ_1 increases as dimension increases across both DDIM and DDPM.



(a) DDIM *(b)* DDPM

Figure E.8. τ_2 vs. κ for varying dimensions. We find that τ_2 decreases as dimension increases across both DDIM and DDPM.

E.4. Ablation on τ_3

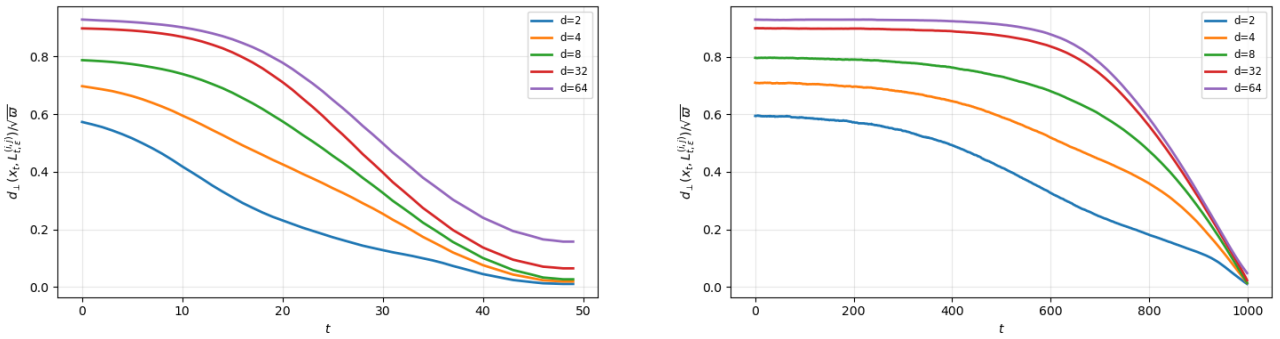
In Figure E.11, we find that as τ_3 increases across samplers, hallucination rate drops, as expected per Prop. 4.7. However, note that a large τ_3 is unrealistic; empirically, as demonstrated in Figure 4, τ_3 is less than 15. Note that all results are in DDIM steps.

E.5. Jacobian Eigenvalues

In Figure E.12, we show that there exists a positive (λ_t) and negative eigenvalue. These two eigenvalues are consistently positive (negative), demonstrating that the midpoint demonstrating that the midpoint is indeed a saddle. Thus, we can conclude that score error and studying the approximate parallel dynamics (rather than the exact parallel dynamics) only has a small effect on the position of the slowdown point.

E.6. Varying Noise Level η in DDIM

In Figure E.13, we provide a first characterization of how different noise levels η (Song et al., 2021a) applied to the DDIM steps affect hallucination rate and escape from the midpoint region (contrasting to adding additional DDPM steps in Figure 5). In line with our results in noise being beneficial for reducing hallucinations, we find that the highest level of noise results in the lowest amount of mode interpolations. We also provide an analogue of Figure 5 for this setting, demonstrating that adding noise to DDIM helps escape the midpoint neighborhood, in line with our result in Prop. 5.1.



(a) DDIM *(b)* DDPM

Figure E.9. Orthogonal distance to ε-extended i, j-mode segment for DDIM. We find that, as dimension increases, ε increases as well. Furthermore, the difference in final distances for DDIM and DDPM become more exacerbated as dimension increases.

E.7. Verifying Assumptions: Prop. 5.1

In Figure E.14, we verify the assumptions on $b(a, t)$ in Prop. 5.1 across 400k samples (10k per each line segment). We find that both assumptions hold, using the exact score presented in our Prop. 5.1, across all 40 line segments for a total of 400,000 samples. Details on how $b(a, t)$, $K(t)$, and λ_{rep} are computed are provided in Sec. F. Note that $\tau_3 = 9$ in DDIM steps corresponds to $\tau_3 = 21$ in DDPM steps.

E.8. Midpoint Trapping under the Exact Score

In Figure E.15, we provide an analogue of Figure 5 in the exact score case to demonstrate that the midpoint trapping mechanism we extract theoretically in Prop. 4.7 and Prop. 5.1 holds, as expected, under the exact score. As discussed in Sec. I, this then becomes useful and predictive in the learned score case as demonstrated by Figure 5 and Tab. I.1.

E.9. Hallucination Rate Ablations on T, σ, and ℓ

In Figure E.16, we demonstrate that the gap in mode interpolation hallucination rate persists across various choices of T, σ, and ℓ. In general, hallucination rates remain roughly stable. As T increases, the gap increases: while discretization error does not explain the gap between DDIM and DDPM hallucinations, DDIM becomes a prohibitively coarse discretization of the PF ODE and thus has a higher hallucination rate. As σ increases, the gap decreases. This is because regions of high probability mass begin to overlap, so feasibly classifying samples as hallucinations becomes rare: this is precisely why Assumptions 4.1 and 4.4 are required, since they ensure well-separated modes where is a useful definition of hallucinations. For similar reasons, as ℓ decreases, the gap decreases.

F. Experimental Details

F.1. Gaussian Mixture

We train our experiments 25-mode gaussian with varying dimensions. Our setup contains 25 modes with an inter-mode separation of 2 with each mode having a standard deviation σ of 0.02. Lengths and standard deviations are then normalized by $2\sqrt{\varpi}$ during sampling. We sample $100,000$ samples uniformly across all modes and train a 3-layer MLP with a hidden size of 64 using a batch size of $10,000$ for $10,000$ epochs. For higher dimensions, modes are sampled from a lattice of 5^{φ}. In higher dimensions ($\varphi > 2$) sampling of the 25 modes is done using farthest point sampling. This is done on set of candidates given by $\min(M, 5^{\varphi})$, with a value of $M = 500,000$ as the maximum number of candidate points. Additionally, in higher dimensions, we use a stronger training pipeline of Fourier features, embedding the log signal-to-noise ratio, and varying depth to ensure approximation of the score.

During sampling, we choose $100,000$ sampled freshly from a standard normal distribution $\mathcal{N}(0, 1)$ and run inference on those samples in the reverse process. Furthermore, in $2D$, we eliminate invalid samples by removing those trajectories for which the final distance to the line joining two modes is greater than 5σ. These are those trajectories for which the

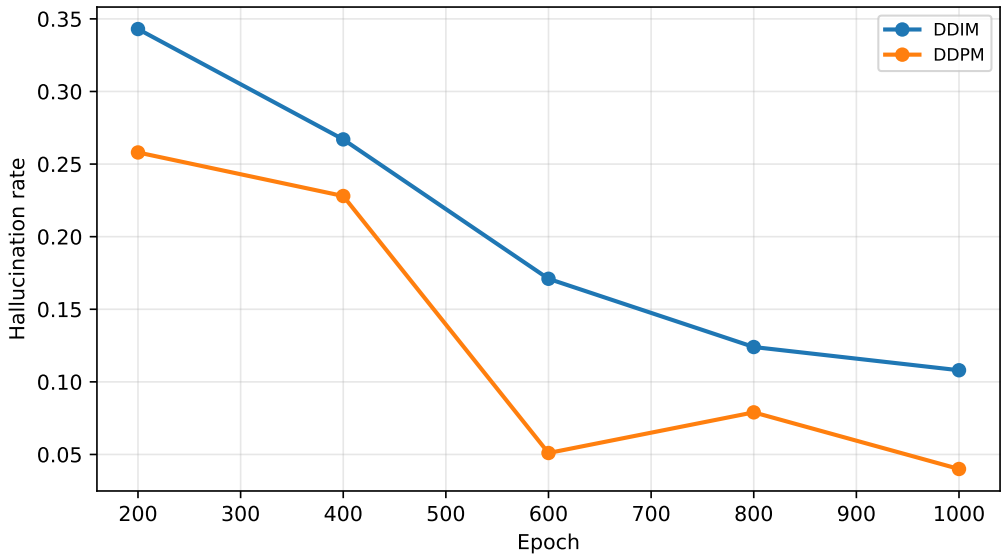


Figure E.10. We compare the image hallucination rate of DDIM and DDPM for 1000 epochs of training using a UNet architecture on the 3 triangle dataset. The evaluation is conducted on 1,000 samples uniformly sampled from $\mathcal{N}(0, 1)$ Note that the hallucination rate of DDPM is consistently lower than DDIM in this case.

outcome is neither a mode interpolation nor a convergence to one of the true modes. In the 2D Gaussian, this amounts to $< 0.01\%$ of the total number of samples used. Mode interpolations in $2D$ are thus samples which are within 5σ of a nearby line but further than 5σ from any mode. A lower threshold would result in false true samples with probability ≈ 1 across 100,000 samples. For higher dimensional Gaussian datasets (> 10), to handle concentration of measure, we use a threshold of $\sigma(\sqrt{\varpi} + 4)$ instead. Additionally, for the analysis of exponential convergence, we find that this threshold is hard to tune, since initial distances scale with $\sqrt{\varpi}$; thus, any instance which is initially further than $\sqrt{\varpi}$ at the beginning of the reverse process is removed. This also amounts to a negligible amount of samples.

F.2. Counting Triangles

We create a dataset of 10,000 images of size 64×64 with 3 triangles placed randomly in the image. The locations are chosen such that the images are non-overlapping. The goal is to faithfully generate images that maintain the number of triangles in the generated image with no constraint on the position of the triangles in the images generated. Example images are provided in Figure F.17.

F.3. Computing Prop. 5.1 Objects

Note that we can write $\boldsymbol{x}_t = \boldsymbol{m} + A_t\boldsymbol{u} + \boldsymbol{z}_t$ where $\langle \boldsymbol{z}_t, \boldsymbol{u}\rangle = 0$ and $||\boldsymbol{z}_t||_2 \leq \varepsilon$, since the trajectory is restricted to the ε-tube. This thus allows us to compute $b(A_t, t)$ with respect to A_t. Note that these require $b(a, t)$ to be negative (positive) for $a \geq \vartheta$ ($a \leq -\vartheta$), since $b(a, t)$ is moving in reverse time in the statement of Prop. 5.1. Next, we check the two assumptions of Prop. 5.1 across all line segments and across 50 intervals of a on each side of the line segment until the endpoints of a. We also sample $\boldsymbol{z}_t$ uniformly in $[-\varepsilon, \varepsilon]$ at each evaluation of a. For the assumption with respect to $K(t)$, the endpoints are $[-2\vartheta, 2\vartheta]$. For the assumption with respect to λ_{rep}, from the proof of Prop. 5.1, note that the assumption with $K(t)$ bounds the probability that the trajectory stays in a 2ϑ-neighborhood and the λ_{rep} assumption bounds the probability that the trajectory returns after exit. Thus, in Sec. E.7, we verify the assumption with λ_{rep} for $|a| \in [\vartheta, a^*]$, where a^* is when return after exit happens 0% of the time and DDPM trajectories interpolate with rate < 0.001 with τ_3 steps remaining. $[\vartheta, a^*]$ is thus the region where return after exit becomes infeasible, which is precisely where the assumption on λ_{rep} must hold. In Sec. E.7, $a^* = 0.280$ where $\vartheta = 0.15$ (nondimensionalized).

F.4. Additional Figures

In Figure F.18, we illustrate the reduction to the DDPM dynamics along the bisector normal which is integral for our theory in Sec. 5.

G. Helpful Lemmas

Lemma G.1. *Given the diffusion marginals as defined in Eq. (2), we can rewrite the score as:*

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = -\frac{1}{\sigma_t^2}\left(\boldsymbol{x}_t - \sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t)\boldsymbol{\mu}_t^{(k)}\right) \tag{G.20}$$

for the responsibilities $\gamma_k(\boldsymbol{x}_t) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I})}{\sum_{k=1}^{N} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I})}$*, where* $\sigma_t^2 = \sigma^2 \bar{\alpha}_t + (1 - \bar{\alpha}_t)$*.*

Proof. Firstly, we compute the gradient of the diffusion marginal as:

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t) &= \sum_{k=1}^{N} \pi_k \nabla_{\boldsymbol{x}_t} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I}) && \text{(G.21)}\\
&= \sum_{k=1}^{N} \pi_k \nabla_{\boldsymbol{x}_t} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I}) \nabla_{\boldsymbol{x}_t}\left(-\frac{||\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(k)}||_2^2}{2\sigma_t^2}\right) && \text{(G.22)}\\
&= \sum_{k=1}^{N} \pi_k \nabla_{\boldsymbol{x}_t} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I}) \left(-\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(k)}}{\sigma_t^2}\right) && \text{(G.23)}
\end{aligned}$$

Then, incorporating Eq. (G.23) into the definition of the score yields:

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) &= \frac{\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t)}{p_t(\boldsymbol{x}_t)} && \text{(G.24)}\\
&= -\frac{1}{\sigma_t^2} \sum_{k=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{(k)}, \sigma_t^2 \boldsymbol{I})}{p_t(\boldsymbol{x}_t)} (\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(k)}) && \text{(G.25)}\\
&= -\frac{1}{\sigma_t^2} \sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{\mu}_t^{(k)}), \text{ definition of } \gamma_k && \text{(G.26)}\\
&= -\frac{1}{\sigma_t^2}\left(\sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t)(\boldsymbol{x}_t) - \sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t)\boldsymbol{\mu}_t^{(k)}\right) && \text{(G.27)}\\
&= -\frac{1}{\sigma_t^2}\left(\boldsymbol{x}_t - \sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t)\boldsymbol{\mu}_t^{(k)}\right), \text{ since } \sum_{k=1}^{N} \gamma_k(\boldsymbol{x}_t) = 1 && \text{(G.28)}
\end{aligned}$$

as desired. □

Lemma G.2. *Let* $\boldsymbol{y}_t = \frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}}$*, where* $\boldsymbol{x}_t$ *is governed by Eq. (3), and let* $\hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) = \sum_{i=1}^{N} \tilde{\gamma}_i(\boldsymbol{y}_t)\boldsymbol{\mu}^{(i)}$ *with new responsibilities:*

$$\tilde{\gamma}_k(\boldsymbol{y}_t) = \frac{\pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(\boldsymbol{k})}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\sum_{m=1}^{N} \pi_m \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(\boldsymbol{m})}, \tilde{\sigma}_t^2 \boldsymbol{I})} \tag{G.29}$$

and $\tilde{\sigma}_t^2 = \frac{\sigma_t^2}{\bar{\alpha}_t}$*.*

Next, define the time-rescaling u *as:*

$$\frac{du}{dt} = \frac{\beta_t}{2\sigma_t^2} \tag{G.30}$$

Then, letting $t = u$ WLOG, we have the rescaled dynamics:

$$\dot{\boldsymbol{y}}_t = \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{y}_t \tag{G.31}$$

Proof. Firstly, let $I(t) = -\int_0^t \beta_s ds$, i.e. $\bar{\alpha}_t = e^{I(t)}$. This yields that:

$$\begin{aligned}
\dot{\bar{\alpha}}_t &= (e^{\dot{I(t)}}) && \text{(G.32)}\\
&= e^{I(t)} \dot{I(t)} && \text{(G.33)}\\
&= e^{I(t)}(-\beta_t) && \text{(G.34)}\\
&= -\beta_t \bar{\alpha}_t && \text{(G.35)}
\end{aligned}$$

Next, by definition of $\boldsymbol{y}_t$, we have that:

$$\begin{aligned}
\dot{\boldsymbol{y}}_t &= \dot{\boldsymbol{x}}_t \bar{\alpha}_t^{-1/2} + \boldsymbol{x}_t \frac{d}{dt} \bar{\alpha}_t^{-1/2} && \text{(G.36)}\\
&= \dot{\boldsymbol{x}}_t \bar{\alpha}_t^{-1/2} + \boldsymbol{x}_t \left(\frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \right) && \text{(G.37)}\\
&= \frac{1}{\sqrt{\bar{\alpha}_t}} \left(-\frac{\beta_t}{2} \left(\boldsymbol{x}_t + \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) \right) \right) + \frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \boldsymbol{x}_t && \text{(G.38)}\\
&= \frac{-\beta_t}{2\sqrt{\bar{\alpha}_t}} \boldsymbol{x}_t - \frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \boldsymbol{x}_t && \text{(G.39)}\\
&= -\frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) && \text{(G.40)}\\
&= -\frac{\beta_t}{2\sqrt{\bar{\alpha}_t}} \left(-\frac{1}{\sigma_t^2} (\boldsymbol{x}_t - \sum_{k=1}^N \gamma_k(\boldsymbol{x}_t) \boldsymbol{\mu}_t^{(k)}) \right), \text{ Lemma G.1} && \text{(G.41)}\\
&= \frac{\beta_t}{2\sigma_t^2 \sqrt{\bar{\alpha}_t}} \left(\sum_{k=1}^N \gamma_k(\boldsymbol{x}_t) \boldsymbol{\mu}_t^{(k)} - \boldsymbol{x}_t \right) && \text{(G.42)}\\
&= \frac{\beta_t}{2\sigma_t^2 \sqrt{\bar{\alpha}_t}} \left(\sqrt{\bar{\alpha}_t} \sum_{k=1}^N \gamma_k(\boldsymbol{x}_t) \boldsymbol{\mu}^{(k)} - \sqrt{\bar{\alpha}_t} \boldsymbol{y}_t \right) && \text{(G.43)}\\
&= \frac{\beta_t}{2\sigma_t^2} \left(\hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{y}_t \right) && \text{(G.44)}
\end{aligned}$$

The chain rule yields the time rescaling:

$$\begin{aligned}
\frac{d\boldsymbol{y}_t}{du} &= \frac{d\boldsymbol{y}_t}{dt} \frac{dt}{du} && \text{(G.45)}\\
&= \left(\frac{\beta_t}{2\sigma_t^2} \left(\hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{y}_t \right) \right) \left(\frac{2\sigma_t^2}{\beta_t} \right) && \text{(G.46)}\\
&= \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{y}_t && \text{(G.47)}
\end{aligned}$$

as desired. □

Remark: This lemma takes the nonautonomous Eq. (3), where the means evolve over time, to dynamics with respect to the *static means*.

Lemma G.3. *Assume Asm. 4.1 holds. Consider τ_1 from Asm. 4.1.. We decompose $\boldsymbol{y}_t$ into parallel and orthogonal components i.e.:*

$$\boldsymbol{y}_t = \boldsymbol{\mu}^{(i)} + \xi_t \ell \boldsymbol{u} + \boldsymbol{w}_t, \tag{G.48}$$

where $\ell = ||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2, \boldsymbol{u} = \frac{\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}}{\ell}$, *and* $\boldsymbol{w}_t$ *satisfies* $\langle \boldsymbol{w}_t, \boldsymbol{u} \rangle = 0$.

Then, for $t \leq \tau_1$, *we have that:*

$$\dot{\boldsymbol{y}} = \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{y}_t \tag{G.49}$$

is equivalent to the dynamical system:

$$\dot{\xi}_t = \tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t + \delta_{\xi_t} \tag{G.50}$$
$$\dot{\boldsymbol{w}}_t = -\boldsymbol{w}_t + \boldsymbol{\delta}_{\boldsymbol{w}_t} \tag{G.51}$$

where:

$$|\delta_\xi|, ||\boldsymbol{\delta}_{\boldsymbol{w}}||_2 \leq \mathcal{O}\left(N \exp(-\kappa)\right) \tag{G.52}$$

Proof. Firstly, we have that:

$$\dot{\boldsymbol{y}}_t = \sum_{k=1}^{N} \tilde{\gamma}_k(\boldsymbol{y}_t)\boldsymbol{\mu}^{(k)} - \boldsymbol{y}_t \tag{G.53}$$

if and only if, expanding with Eq. (G.48):

$$\ell\dot{\xi}_t\boldsymbol{u} + \dot{\boldsymbol{w}}_t = \sum_{k=1}^{N} \tilde{\gamma}_k(\boldsymbol{y}_t)\boldsymbol{\mu}^{(k)} - (\boldsymbol{\mu}^{(i)} + \xi_t\ell\boldsymbol{u} + \boldsymbol{w}_t). \tag{G.54}$$

Next, we have that $\tilde{\gamma}_i(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(i)}) = 0$; $\tilde{\gamma}_j(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}) = \tilde{\gamma}_j(\boldsymbol{y}_t)\ell\boldsymbol{u}$; and $\boldsymbol{E}_t := \sum_{k \notin \{i,j\}} \tilde{\gamma}_k(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)})$, which when incorporated into Eq. (G.54) yields:

$$\ell\dot{\xi}_t\boldsymbol{u} + \dot{\boldsymbol{w}}_t = \tilde{\gamma}_j(\boldsymbol{y}_t)\ell\boldsymbol{u} + \boldsymbol{E}_t - \xi_t\ell\boldsymbol{u} - \boldsymbol{w}_t \tag{G.55}$$
$$= (\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell\boldsymbol{u} - \boldsymbol{w}_t + \boldsymbol{E}_t. \tag{G.56}$$

Next, noting that $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 1$, it follows from Eq. (G.56) that:

$$\langle \ell\dot{\xi}_t\boldsymbol{u}, \boldsymbol{u} \rangle = \langle (\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell\boldsymbol{u} - \boldsymbol{w}_t + \boldsymbol{E}_t, \boldsymbol{u} \rangle \iff \tag{G.57}$$
$$\ell\dot{\xi}_t = (\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell + \langle \boldsymbol{E}_t, \boldsymbol{u} \rangle \iff \tag{G.58}$$
$$\dot{\xi}_t = \tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t + \delta_{\xi_t},\ \delta_{\xi_t} := \frac{1}{\ell}\langle \boldsymbol{E}_t, \boldsymbol{u} \rangle \tag{G.59}$$

yielding the first equation in Eq. (G.51), the parallel dynamics along $L_{t,\varepsilon}^{(i,j)}$. Furthermore, by Eq. (G.58), we have that:

$$\ell\dot{\xi}_t\boldsymbol{u} = (\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell\boldsymbol{u} + \langle \boldsymbol{E}_t, \boldsymbol{u} \rangle\boldsymbol{u} \tag{G.60}$$

Thus, subtracting $\ell\dot{\xi}_t\boldsymbol{u}$ from the RHS of Eq. (G.54) and then plugging in Eq. (G.60) yields:

$$\dot{\boldsymbol{w}}_t = (\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell\boldsymbol{u} - \boldsymbol{w}_t + \boldsymbol{E}_t - ((\tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t)\ell\boldsymbol{u} + \langle \boldsymbol{E}_t, \boldsymbol{u} \rangle\boldsymbol{u}) \tag{G.61}$$
$$= -\boldsymbol{w}_t + \boldsymbol{E}_t - \langle \boldsymbol{E}_t, \boldsymbol{u} \rangle\boldsymbol{u} \tag{G.62}$$

Note that $\langle \boldsymbol{E}_t, \boldsymbol{u}\rangle \boldsymbol{u} = \text{Proj}_{(\boldsymbol{u})}(\boldsymbol{E}_t) =: \text{Proj}_{\parallel}(\boldsymbol{E}_t)$, yielding that $\boldsymbol{E}_t - \langle \boldsymbol{E}_t, \boldsymbol{u}\rangle \boldsymbol{u} =: \text{Proj}_{\perp}(\boldsymbol{E}_t)$. Incorporating this into Eq. (G.62) yields:

$$\dot{\boldsymbol{w}}_t = -\boldsymbol{w}_t + \boldsymbol{\delta}_{\boldsymbol{w}_t},\ \boldsymbol{\delta}_{\boldsymbol{w}_t} := \text{Proj}_{\perp}(\boldsymbol{E}_t). \tag{G.63}$$

This yields the second equation in Eq. (G.51), the orthogonal dynamics with respect to $L_{t,\varepsilon}^{(i,j)}$. To prove the lemma, it thus suffices to bound $|\delta_{\xi_t}|$ and $||\boldsymbol{\delta}_{\boldsymbol{w}_t}||_2$. To do so, we begin by bounding the responsibilities $\tilde{\gamma}_k(\boldsymbol{y}_t)$ for $k \notin \{i,j\}$:

$$\tilde{\gamma}_k(\boldsymbol{y}_t) = \frac{\pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(k)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\sum_{m=1}^N \pi_m \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(m)}, \tilde{\sigma}_t^2 \boldsymbol{I})} \tag{G.64}$$

$$\leq \frac{\pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(k)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\pi_i \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2 \boldsymbol{I}) + \pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})}. \tag{G.65}$$

Let $d^* := \arg\min_{d \in \{i,j\}} ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(d)}||_2$. Note that:

$$\pi_i \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2 \boldsymbol{I}) + \pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I}) \geq \max\{\pi_i \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2 \boldsymbol{I}), \pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})\} \tag{G.66}$$

$$\geq \pi_{d^*} \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(d^*)}, \tilde{\sigma}_t^2 \boldsymbol{I}) \tag{G.67}$$

Incorporating Eq. (G.67) into Eq. (G.65) yields:

$$\tilde{\gamma}_k(\boldsymbol{y}_t) \leq \frac{\pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(k)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\pi_{d^*} \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(d^*)}, \tilde{\sigma}_t^2 \boldsymbol{I})} \tag{G.68}$$

$$= \frac{\pi_k}{\pi_{d^*}} \frac{\exp(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}||_2^2}{2\tilde{\sigma}_t^2})}{\exp(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(d^*)}||_2^2}{2\tilde{\sigma}_t^2})} \tag{G.69}$$

$$= \frac{\pi_k}{\pi_{d^*}} \exp\left(-\frac{1}{2\tilde{\sigma}_t^2}(||\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(d^*)}||_2^2)\right). \tag{G.70}$$

By Asm. 4.1, we have that $||\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}||_2^2 \geq ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(d^*)}||_2^2 + \frac{\Delta_t}{\bar{\alpha}_t}$, which holds if and only if $||\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(d^*)}||_2^2 \geq \frac{\Delta_t}{\bar{\alpha}_t}$. Incorporating this into Eq. (G.70) yields:

$$\tilde{\gamma}_k(\boldsymbol{y}_t) \leq \frac{\pi_k}{\pi_{d^*}} \exp(-\frac{\Delta_t}{2\bar{\alpha}_t \tilde{\sigma}_t^2}) \tag{G.71}$$

$$\leq \frac{\pi_k}{\pi_{d^*}} \exp(-\kappa),\ \text{Asm. 4.1} \tag{G.72}$$

Next, we have that:

$$|\delta_{\xi_t}| = |\frac{1}{\ell}\langle \boldsymbol{E}_t, \boldsymbol{u}\rangle| \tag{G.73}$$

$$\leq \frac{1}{\ell}||\boldsymbol{E}_t||_2 ||\boldsymbol{u}||_2,\ \text{Cauchy-Schwartz} \tag{G.74}$$

$$= \frac{1}{\ell}||\boldsymbol{E}_t||_2 \tag{G.75}$$

$$= \frac{1}{\ell}||\sum_{k \notin \{i,j\}} \tilde{\gamma}_k(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)})||_2 \tag{G.76}$$

$$\leq \frac{\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell} \sum_{k \notin \{i,j\}} \tilde{\gamma}_k(\boldsymbol{y}_t) \tag{G.77}$$

$$\leq \frac{\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell} (N-2) \frac{\pi_k}{\pi_{d^*}} \exp(-\kappa),\ \text{Eq. (G.72)} \tag{G.78}$$

$$\leq \mathcal{O}(N \exp(-\kappa)) \tag{G.79}$$

Similarly, we have that:

$$
\begin{aligned}
||\boldsymbol{\delta}_{\boldsymbol{w}_t}||_2 &\leq ||\text{Proj}_{\perp}(\boldsymbol{E})||_2 && \text{(G.80)}\\
&\leq ||\boldsymbol{E}||_2, \text{ property of orth. projection} && \text{(G.81)}\\
&\leq \sum_{k \notin \{i,j\}} \tilde{\gamma}_k(\boldsymbol{y}_t)||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 && \text{(G.82)}\\
&\leq (N-2) \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 \frac{\pi_i}{\pi_d} \exp(-\kappa), \text{ Eq. (G.72)} && \text{(G.83)}\\
&\leq \mathcal{O}(N \exp(-\kappa)) && \text{(G.84)}
\end{aligned}
$$

as desired.

□

Lemma G.4. *Assume Asm. 4.1 holds. Then, there exists a time $\tau \leq \tau_1$ such that ξ_t, as defined by Eq. (G.51), lies in $[-\epsilon, 1+\epsilon]$, where $\epsilon \in O(N \exp(-\kappa))$, for all time $t \leq \tau$.*

Proof. Recall the definition of an ε-extended line segment in Sec. 4. By enlarging the constant in $\varepsilon \in O(N \exp(-\kappa))$, we may assume that $|\delta_{\xi_t}| \leq \varepsilon/2$ by Lemma G.3. Since $\tilde{\gamma}_j(\boldsymbol{y}_t) \in (0,1)$, if $\xi_t > 1+\varepsilon$ and $r_t := \xi_t - (1+\varepsilon)$, then $\dot{r}_t = \dot{\xi}_t \leq -r_t - \varepsilon/2$; similarly, if $\xi_t < -\varepsilon$ and $r_t := -\varepsilon - \xi_t$, then $\dot{r}_t = -\dot{\xi}_t \leq -r_t - \varepsilon/2$. Thus, for t sufficiently large, r_t vanishes and ξ_t is trapped in $[-\varepsilon, 1+\varepsilon]$. □

Lemma G.5. *For any $\phi \in (0,1)$, with probability at least $1 - 2\exp(-(\varpi)\phi^2/8)$, we have that $||\boldsymbol{w}_T||_2$, where $\boldsymbol{w}_t$ is defined as in Eq. (G.51), satisfies:*

$$
||\boldsymbol{w}_T||_2 \leq \sqrt{(\varpi)(1+\phi)} \tag{G.85}
$$

Proof. Apply standard Chernoff tail bound for $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$, Theorem 2.7 in Fernandez-Granda (2017). □

Lemma G.6 (Brownian confinement in an interval)**.** *Let $(B_t)_{t\geq 0}$ be standard one-dimensional Brownian motion. Then for any $V > 0$ and $a > 0$,*

$$
\mathbb{P}\Big(\sup_{0\leq t\leq V} |B_t| \leq a\Big) \;\leq\; \frac{4}{\pi}\exp\Big(-\frac{\pi^2 V}{8a^2}\Big). \tag{G.86}
$$

Proof. See Eq 7.15 in (Mörters & Peres, 2010) and evaluate at the midpoint keeping the first eigenvalue term. □

H. Proofs

H.1. Proof of Theorem 4.2

Proof. Consider $\tau \leq \tau_1$ such that Lemma G.4 holds, and then fix $t \leq \tau$. Consider the dynamics in Lemma G.2, which are equivalent to those in Eq. (3) and are defined with respect to the rescaled time:

$$
u(t) := \int_t^T \frac{\beta_s}{2\sigma_s^2}\, ds, \tag{H.87}
$$

by the fundamental theorem of calculus. Furthermore, consider the parallel and orthogonal dynamics as defined in Lemma G.3, except parameterized by the $u(t)$ as they were defined originally per Lemma G.3. Note that $u(t)$ is monotonically *decreasing* for forward t, since:

$$
\frac{d}{dt}u(t) = -\frac{\beta_t}{2\sigma_t^2} < 0, \tag{H.88}
$$

by definition of β_t and σ_t^2; in particular, it is monotonically *increasing* at our studied reverse t.

Accordingly, the dynamics in Lemma G.2 are:

$$\frac{d}{du}\xi_{u(t)} = \gamma_j(\boldsymbol{y}_{u(t)}) - \xi_{u(t)} + \delta_{\xi_{u(t)}}, \tag{H.89}$$

$$\frac{d}{du}\boldsymbol{w}_{u(t)} = -\boldsymbol{w}_{u(t)} + \boldsymbol{\delta}_{\boldsymbol{w}_{u(t)}}, \tag{H.90}$$

where $|\delta_{\xi_u}|, \|\boldsymbol{\delta}_{\boldsymbol{w}_u}\|_2 \le \varepsilon$.

Then, by definition of the orthogonal distance (and since, by Lemma G.2, in $\boldsymbol{y}$-space the (i,j)-mode segment is static, i.e. $L_{u(t)} = L_0$ for all t:),

$$d_\perp(\boldsymbol{y}_{u(t)}, L_{0,\varepsilon}^{(i,j)}) = \|\boldsymbol{w}_{u(t)}\|_2, \tag{H.91}$$

since the closest point on the extended line segment is the orthogonal projection. Note that the following inequality follows vacuously if $\|\boldsymbol{w}_{u(t)}\|_2 = 0$, hence we consider $\|\boldsymbol{w}_{u(t)}\|_2 \neq 0$. Define $g(u) := \|\boldsymbol{w}_u\|_2$, where $u := u(t)$.

Define $g(u) := \|\boldsymbol{w}_u\|_2$. In the first case that $g(u) \neq 0$, we have

$$g(u)^2 = \boldsymbol{w}_u^\top \boldsymbol{w}_u \quad \Longrightarrow \quad \frac{d}{du}g(u)^2 = \frac{d}{du}\big(\boldsymbol{w}_u^\top \boldsymbol{w}_u\big) = 2\,\boldsymbol{w}_u^\top \frac{d\boldsymbol{w}_u}{du}. \tag{H.92}$$

Using the orthogonal dynamics $\frac{d\boldsymbol{w}_u}{du} = -\boldsymbol{w}_u + \boldsymbol{\delta}_{\boldsymbol{w}_u}$ and $\|\boldsymbol{\delta}_{\boldsymbol{w}_u}\|_2 \le \varepsilon$, we obtain

$$\frac{d}{du}g(u)^2 = 2\,\boldsymbol{w}_u^\top(-\boldsymbol{w}_u + \boldsymbol{\delta}_{\boldsymbol{w}_u}) = -2\|\boldsymbol{w}_u\|_2^2 + 2\,\boldsymbol{w}_u^\top \boldsymbol{\delta}_{\boldsymbol{w}_u} \tag{H.93}$$

$$\le -2g(u)^2 + 2\|\boldsymbol{w}_u\|_2\|\boldsymbol{\delta}_{\boldsymbol{w}_u}\|_2 \le -2g(u)^2 + 2\varepsilon g(u). \tag{H.94}$$

Since $\frac{d}{du}g(u)^2 = 2g(u)g'(u)$ for $g(u) \neq 0$, dividing by $2g(u)$ yields

$$g'(u) \le -g(u) + \varepsilon. \tag{H.95}$$

In the other case that $g(u) = 0$, the inequality holds trivially. Multiplying by e^u then yields:

$$\frac{d}{du}(e^u g(u)) \le \varepsilon e^u. \tag{H.96}$$

Integrating from 0 to u then gives:

$$e^u g(u) - g(0) \le \varepsilon(e^u - 1), \tag{H.97}$$

hence:

$$g(u) \le e^{-u}g(0) + \varepsilon(1 - e^{-u}) \le e^{-u}g(0) + \varepsilon. \tag{H.98}$$

In particular, denoting $z : u(z) = 0$ to mean the time $z \in [0, T]$ s.t. $u(z) = 0$, we have that:

$$\|\boldsymbol{w}_{u(t)}\|_2 \le e^{-u(t)}\|\boldsymbol{w}_{z:u(z)=0}\|_2 + \varepsilon(1 - e^{-u(t)}) \quad \le \quad e^{-u(t)}\|\boldsymbol{w}_T\|_2 + \varepsilon. \tag{H.99}$$

In particular, by definition of u, $z = T$, yielding:

$$\|\boldsymbol{w}_{u(t)}\|_2 \le e^{-u(t)}\|\boldsymbol{w}_T\|_2 + \varepsilon(1 - e^{-u(t)}) \quad \le \quad e^{-u(t)}\|\boldsymbol{w}_T\|_2 + \varepsilon. \tag{H.100}$$

By definition of $\boldsymbol{y}_{u(t)}$ in Lemma G.2, we have that $d_\perp(\boldsymbol{x}_t, L_{t,\varepsilon}^{(i,j)}) = \sqrt{\bar{\alpha}_t}\, d_\perp(\boldsymbol{y}_{u(t)}, L_{0,\varepsilon}^{(i,j)})$. Applying this to Eq. (H.100),

we obtain that:

$$
\begin{aligned}
d_\perp(\boldsymbol{x}_t, L_{t,\varepsilon}^{(i,j)}) &= \sqrt{\bar{\alpha}_t}\, d_\perp(\boldsymbol{y}_{u(t)}, L_{0,\varepsilon}^{(i,j)}) && \text{(H.101)}\\
&= \sqrt{\bar{\alpha}_t}\, \|\boldsymbol{w}_{u(t)}\|_2 && \text{(H.102)}\\
&\le \sqrt{\bar{\alpha}_t}\Big(e^{-u(t)}\|\boldsymbol{w}_T\|_2 + \varepsilon\Big) && \text{(H.103)}\\
&= \sqrt{\bar{\alpha}_t}\exp(-u(t)) d_\perp(\boldsymbol{y}_T, L_{T,\varepsilon}^{(i,j)}) + \sqrt{\bar{\alpha}_t}\varepsilon && \text{(H.104)}\\
&= \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_T}}\exp(-u(t)) d_\perp(\boldsymbol{x}_T, L_{T,\varepsilon}^{(i,j)}) + \sqrt{\bar{\alpha}_t}\varepsilon && \text{(H.105)}\\
&= \sqrt{\frac{\bar{\alpha}_t}{\gamma}}\exp(-u(t)) d_\perp(\boldsymbol{x}_T, L_{T,\varepsilon}^{(i,j)}) + \sqrt{\bar{\alpha}_t}\varepsilon, && \text{(H.106)}\\
&\le \mathcal{O}(\exp(-u(t)) d_\perp(\boldsymbol{x}_T, L_{T,\varepsilon}^{(i,j)}) + \varepsilon) && \text{(H.107)}
\end{aligned}
$$

by definition of $\bar{\alpha}_T$ and $u(t)$. To obtain the result without the initial distance, note that $||\boldsymbol{w}_T||_2 \le ||\boldsymbol{y}_T||_2 \le \frac{\sqrt{\varpi(1+\phi)}}{\sqrt{\bar{\alpha}_T}} = \frac{\sqrt{\varpi(1+\phi)}}{\sqrt{\gamma}}$ since $\boldsymbol{w}_T$ is a component of $\boldsymbol{y}_T$, and by applying Lemma G.5; note that the extra $\sqrt{\gamma}$ in the denominator is again absorbed by the $\mathcal{O}$. Finally, since t was arbitrary, this holds for all $t \le \tau$ as desired.

□

H.2. Proof of Corollary 4.3

Proof. From the proof of Theorem 4.2, we have that the dynamics of $||\boldsymbol{w}_u||_2$ satisfy the differential inequality:

$$
\frac{d}{du}||\boldsymbol{w}_u||_2 \le -||\boldsymbol{w}_u||_2 + \varepsilon \tag{H.108}
$$

At the boundary of the tube, we thus have:

$$
\frac{d}{du}||\boldsymbol{w}_u||_2\Big|_{||\boldsymbol{w}_u||_2=\varepsilon} \le -\varepsilon + \varepsilon = 0. \tag{H.109}
$$

In particular, if $||\boldsymbol{w}_u||_2 \le \varepsilon$, it cannot cross the boundary to the outside, because doing so would require a strictly positive derivative. Note that since $\sqrt{\bar{\alpha}_u}$ and its reciprocal are uniformly bounded on $[T-\tau_1, T]$, the rescaling only changes the tube radius by a multiplicative constant which we absorb into ε. Thus, together with Lemma G.4, the ε-tube around the line segment in Theorem 4.2 is an invariant set, and for reverse $t \le \tau$, if a trajectory enters this set, it stays there. □

H.3. Proof of Proposition 4.5

Proof. We begin by showing existence of an instantaneously stable equilibria $\xi_t^{(i),*}$ in the ε-ball around $\boldsymbol{\mu}_i$, without loss of generality. The proof is symmetric for $\boldsymbol{\mu}_j$.

By Lemma G.3, we have the parallel dynamics:

$$
\dot{\xi}_t = \tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t + \delta_{\xi_t}, \tag{H.110}
$$

where:

$$
\boldsymbol{y}_t = \boldsymbol{\mu}^{(i)} + \xi_t \ell \boldsymbol{u} + \boldsymbol{w}_t, \tag{H.111}
$$

where $\ell = ||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2,\ \boldsymbol{u} = \frac{\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}}{\ell}$, and $\boldsymbol{w}_t$ satisfies $\langle \boldsymbol{w}_t, \boldsymbol{u}\rangle = 0$. Furthermore, since the trajectory is in the ε-tube around $L_t^{(i,j)}$, we have that $||\boldsymbol{w}_t||_2 \le \varepsilon$.

Fix $\xi := \xi_t$ for $t \leq \hat{\tau}$ arbitrary. Then, we have that:

$$
\begin{aligned}
||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2 &= ||\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(i)} + \xi\ell\boldsymbol{u} + \boldsymbol{w}_t||_2^2 && \text{(H.112)}\\
&= ||\xi\ell\boldsymbol{u} + \boldsymbol{w}_t||_2^2 && \text{(H.113)}\\
&= ||\xi\ell\boldsymbol{u}||_2^2 + ||\boldsymbol{w}_t||_2^2 + 2\langle\xi\ell\boldsymbol{u}, \boldsymbol{w}_t\rangle && \text{(H.114)}\\
&= \xi^2\ell^2||\boldsymbol{u}||_2^2 + ||\boldsymbol{w}_t||_2^2,\ \langle\boldsymbol{u}, \boldsymbol{w}_t\rangle = 0 && \text{(H.115)}\\
&= \xi^2\ell^2||\boldsymbol{u}||_2^2 + ||\boldsymbol{w}_t||_2^2,\ ||\boldsymbol{u}||_2 = 1. && \text{(H.116)}
\end{aligned}
$$

Next, we have that:

$$
\begin{aligned}
||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 &= ||\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)} + \xi\ell\boldsymbol{u} + \boldsymbol{w}_t||_2^2 && \text{(H.117)}\\
&= ||\boldsymbol{\mu}^{(i)} - (\boldsymbol{\mu}^{(i)} + \ell\boldsymbol{u}) + \xi\ell\boldsymbol{u} + \boldsymbol{w}_t||_2^2 && \text{(H.118)}\\
&= ||(\xi - 1)\ell\boldsymbol{u} + \boldsymbol{w}_t||_2^2 && \text{(H.119)}\\
&= ||(\xi - 1)\ell||_2^2 + ||\boldsymbol{w}_t||_2^2 + 2\langle(\xi - 1)\ell\boldsymbol{u}, \boldsymbol{w}_t\rangle && \text{(H.120)}\\
&= (1 - \xi)^2\ell^2 + ||\boldsymbol{w}_t||_2^2, && \text{(H.121)}
\end{aligned}
$$

by the same token, since $||\boldsymbol{u}||_2^2 = 1$ and $\langle\boldsymbol{u}, \boldsymbol{w}_t\rangle = 0$.

Incorporating Eq. (H.116) and Eq. (H.121), we have that:

$$
\begin{aligned}
||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2 &= (1 - \xi)^2\ell^2 + ||\boldsymbol{w}_t||_2^2 - (\xi^2\ell^2 + ||\boldsymbol{w}_t||_2^2) && \text{(H.122)}\\
&= (1 - \xi)^2\ell^2 - \xi^2\ell^2 && \text{(H.123)}\\
&= ((1 - \xi)^2 - \xi^2)\ell^2 && \text{(H.124)}\\
&= (1 - 2\xi)\ell^2. && \text{(H.125)}
\end{aligned}
$$

We then bound the responsibility of mode j as:

$$
\begin{aligned}
\tilde{\gamma}_j(\boldsymbol{y}_t) &= \frac{\pi_j\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2\boldsymbol{I})}{\sum_{m=1}^N \pi_m\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(m)}, \tilde{\sigma}_t^2\boldsymbol{I})} && \text{(H.126)}\\
&\leq \frac{\pi_j\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2\boldsymbol{I})}{\pi_i\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2\boldsymbol{I}) + \pi_j\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2\boldsymbol{I})} && \text{(H.127)}\\
&= \frac{\pi_j \exp\left(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2}{2\tilde{\sigma}_t^2}\right)}{\pi_i \exp\left(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2}{2\tilde{\sigma}_t^2}\right) + \pi_j \exp\left(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2}{2\tilde{\sigma}_t^2}\right)} && \text{(H.128)}\\
&= \frac{1}{1 + \frac{\pi_i}{\pi_j}(\exp\left(\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2}{2\tilde{\sigma}_t^2}\right)}, \text{ dividing by numerator} && \text{(H.129)}\\
&\leq \frac{\pi_j}{\pi_i} \exp\left(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2}{2\tilde{\sigma}_t^2}\right) && \text{(H.130)}\\
&\leq \frac{\pi_j}{\pi_i} \exp\left(-\frac{(1 - 2\xi)\ell^2}{2\tilde{\sigma}_t^2}\right) && \text{(H.131)}
\end{aligned}
$$

Next, suppose that $\xi \in [-a, a]$ for $a \in \mathcal{O}(N\exp(-\kappa))$ satisfying $a \leq \frac{1}{4}$, to be chosen later. Then, $1 - 2a \geq \frac{1}{2}$, and thus:

$$\varepsilon_{\tilde{\gamma}} = \sup_{\xi \in [-a,a]} \tilde{\gamma}_j(\boldsymbol{y}_t) \tag{H.132}$$
$$\leq \frac{\pi_j}{\pi_i} \exp\left(-\frac{(1-2a)\ell^2}{2\tilde{\sigma}_t^2}\right) \tag{H.133}$$
$$\leq \frac{\pi_j}{\pi_i} \exp\left(-\frac{(1-2a)\ell^2}{2\tilde{\sigma}_t^2}\right) \tag{H.134}$$
$$\leq \frac{\pi_j}{\pi_i} \exp\left(-(1-2a)2\kappa\right), \text{ Asm. 4.4} \tag{H.135}$$
$$\in \mathcal{O}(\exp(-\kappa)),\ 1-2a \geq \frac{1}{2}. \tag{H.136}$$

Then, by the above, $\varepsilon_{\tilde{\gamma}} \in \mathcal{O}(N\exp(-\kappa))$. Furthermore, by Lemma G.3, we have that $\varepsilon_{\delta_\xi} := \sup_{\xi\in[-a,a]} |\delta_\xi| \leq \sup_{\xi\in[-a,1+a]} |\delta_\xi| \in \mathcal{O}(N\exp(-\kappa))$. Thus, let $a := 2(\varepsilon_{\delta_\xi} + \varepsilon_{\tilde{\gamma}})$, which satisfies $a \in \mathcal{O}(N\exp(-\kappa))$. Since $a \in \mathcal{O}(N\exp(-\kappa))$ there exists a $C_a > 0$ s.t. $a \leq C_a N\exp(-\kappa)$ for N sufficiently large; choosing $\kappa \geq \log N + \log(4C_a)$ ensures that $a \leq \frac{1}{4}$.

We then evaluate the behavior of ξ at the endpoints. At $\xi = -a$, we have that:

$$F_t(-a) = \tilde{\gamma}_j(\boldsymbol{y}_t(-a)) + a + \delta_{-a} \tag{H.137}$$
$$\geq a + \delta_{-a},\ \tilde{\gamma}_j \geq 0 \tag{H.138}$$
$$\geq a - |\delta_a|,\ \ x \geq -|x| \text{ for } x \in \mathbb{R} \tag{H.139}$$
$$\geq a - \varepsilon_{\delta_a}, \delta_a \leq \varepsilon_{\delta_a} \tag{H.140}$$
$$= 2\varepsilon_{\delta_a} + 2\varepsilon_{\tilde{\gamma}} - \varepsilon_{\delta_a} \tag{H.141}$$
$$\geq \varepsilon_{\delta_a} + \varepsilon_{\tilde{\gamma}} \tag{H.142}$$
$$> 0. \tag{H.143}$$

Next, at $\xi = a$, we have that:

$$F_t(a) = \tilde{\gamma}_j(\boldsymbol{y}_t(a)) - a + \delta_a \tag{H.144}$$
$$\leq \varepsilon_{\tilde{\gamma}} - a + \varepsilon_{\delta_a},\ \tilde{\gamma}_j \leq \varepsilon_{\tilde{\gamma}} \text{ and } \delta_a \leq \varepsilon_{\delta_a} \tag{H.145}$$
$$= \varepsilon_{\tilde{\gamma}} - 2\varepsilon_{\delta_a} - 2\varepsilon_{\tilde{\gamma}} + \varepsilon_{\delta_a} \tag{H.146}$$
$$= -(\varepsilon_{\delta_a} + \varepsilon_{\tilde{\gamma}}) \tag{H.147}$$
$$< 0. \tag{H.148}$$

Since F_t is continuous in ξ, since it is a composition of continuous functions (responsibilities are compositions of continuous Gaussian p.d.f.s, and hence continuous), by the intermediate value theorem there exists a $\xi_t^{*,i} \in (-a,a)$ s.t. $F_t(\xi_t^{*,i}) = 0$. Furthermore:

$$||\boldsymbol{y}_t(\xi_t^{*,i}) - \boldsymbol{\mu}^{(i)}||_2 \leq |\xi_t^{*,i}|\ell + ||\boldsymbol{w}||_2, \text{ triangle inequality} \tag{H.149}$$
$$\leq a\ell + \varepsilon \tag{H.150}$$
$$\in \mathcal{O}(N\exp(-\kappa)), \tag{H.151}$$

i.e. there exists an instantaneous equilibria of Eq. (G.50) in a ε-ball of $\boldsymbol{\mu}^{(i)}$, where $\varepsilon \in \mathcal{O}(N\exp(-\kappa))$. Since ξ was arbitrary, this holds for all $t \leq \hat{\tau}$.

Next, we show the (uniform) instantaneous stability of $\xi_t^{*,i}$, without loss of generality; the proof is symmetric for $\xi_t^{*,j}$. Fix $t \leq \hat{\tau}$. Then, we have that:

$$\frac{d}{d\xi_t}F_t(\xi_t^{*,i}) = \left(\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{d}{d\xi_t}\delta_{\xi_t} - 1\right)|_{\xi_t=\xi_t^{*,i}}. \tag{H.152}$$

We demonstrate that $F_t'(\xi_t^{*,i}) < 0$. Fix time $t \leq \hat{\tau}$. We have that:

$$\nabla_{\boldsymbol{y}_t}\tilde{\gamma}_j(\boldsymbol{y}_t) = \frac{\tilde{\gamma}_j(\boldsymbol{y}_t)}{\tilde{\sigma}_t^2}(\boldsymbol{\mu}^{(j)} - \hat{\boldsymbol{\mu}}) \tag{H.153}$$

where $\hat{\boldsymbol{\mu}}$ is the same as that in Lemma G.2. Hence, we have that:

$$\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t) = \langle \nabla_{\boldsymbol{y}_t}\tilde{\gamma}_j(\boldsymbol{y}_t), \frac{d\boldsymbol{y}_t}{d\xi_t}\rangle, \text{ chain rule} \tag{H.154}$$

$$= \frac{\ell\tilde{\gamma}_j(\boldsymbol{y}_t)}{\tilde{\sigma}_t^2}\langle \boldsymbol{\mu}^{(j)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y})), \boldsymbol{u}\rangle \text{ Eq. (H.153)} \tag{H.155}$$

To bound the inner product, note that, by linearity of the inner product:

$$\langle \boldsymbol{\mu}^{(j)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t), \boldsymbol{u}\rangle = \langle \boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}, \boldsymbol{u}\rangle + \langle \boldsymbol{\mu}^{(i)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t), \boldsymbol{u}\rangle \tag{H.156}$$

This implies:

$$|\langle \boldsymbol{\mu}^{(j)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t), \boldsymbol{u}\rangle| \leq |\langle \boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}, \boldsymbol{u}\rangle| + |\langle \boldsymbol{\mu}^{(i)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t), \boldsymbol{u}\rangle|, \text{ triangle inequality} \tag{H.157}$$

$$\leq ||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2||\boldsymbol{u}||_2 + ||\boldsymbol{\mu}^{(i)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t)||_2||\boldsymbol{u}||_2, \text{ Cauchy-Schwartz} \tag{H.158}$$

$$\leq \ell + ||\sum_{k\in[N]}\tilde{\gamma}_k(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2, \ ||\boldsymbol{u}||_2 = 1 \tag{H.159}$$

$$\leq \ell + \tilde{\gamma}_j(\boldsymbol{y}_t)||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2 + ||\sum_{k\in[N]}\tilde{\gamma}_k(\boldsymbol{y}_t)(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2, \text{triangle inequality} \tag{H.160}$$

$$\leq \ell + \tilde{\gamma}_j\ell + \max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t) \tag{H.161}$$

$$\leq 2\ell + \max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t) \tag{H.162}$$

Incorporating Eq. (H.162) into Eq. (H.155) yields:

$$|\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t)| \leq \frac{2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\ell}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t)\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t) \tag{H.163}$$

$$\leq \frac{2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t)\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t). \tag{H.164}$$

Recall that $\xi_t^{*,i} \in (-a, a)$ where $a \leq \frac{1}{4}$, and thus we have $(1-2a) \geq \frac{1}{2}$. Let $x := \frac{\ell^2}{2\tilde{\sigma}_t^2}$. We then have, by Eq. (H.135), that:

$$\tilde{\gamma}_j(\boldsymbol{y}_t) \leq \frac{\pi_j}{\pi_i}(\exp(-(1-2a)x) \leq \frac{\pi_j}{\pi_i}\exp(-x/2). \tag{H.165}$$

Thus:

$$\frac{2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t) = 4x\tilde{\gamma}_j(\boldsymbol{y}_t) \leq 4\frac{\pi_j}{\pi_i}x\exp(-x/2). \tag{H.166}$$

Furthermore, by Asm. 4.4, we have that $\frac{\ell^2}{2\tilde{\sigma}_t^2} \geq 2\kappa$, i.e. $x \geq 2\kappa$. Therefore, we have that $x\exp(-x/2) \leq \sup_{x\geq 2\kappa} x\exp(-x/2) = (2\kappa)\exp(-2\kappa/2) = 2\kappa\exp(-\kappa)$. Then, incorporating this and Eq. (H.166) into Eq. (H.164), we have that:

$$\sup_{\xi\in[-a,a]} |\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t)| \leq \sup_{\xi\in[-a,a]} \left(\frac{2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t)\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t)\right) \tag{H.167}$$

$$\leq 2\sup_{\xi\in[-a,a]}\sup_{\frac{\ell^2}{2\tilde{\sigma}_t^2}\geq\kappa}\frac{\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t) + \max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 \sup_{\xi\in[-a,a]}\sup_{\frac{\ell^2}{2\tilde{\sigma}_t^2}\geq\kappa}\frac{\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j(\boldsymbol{y}_t)\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t) \tag{H.168}$$

$$\leq 8\frac{\pi_j}{\pi_i}\kappa\exp(-\kappa) + \max_{k\notin\{i,j\}} 2||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\kappa\exp(-\kappa)\sum_{k\notin\{i,j\}}\tilde{\gamma}_k(\boldsymbol{y}_t) \tag{H.169}$$

$$\leq 8\frac{\pi_j}{\pi_i}\kappa\exp(-\kappa) + \max_{k\notin\{i,j\}} 2||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\kappa\exp(-\kappa)(N-2)\max_{k\notin\{i,j\}}\frac{\pi_k}{\pi_{d^*}}\exp(-\kappa), \text{ Eq. (G.72).} \tag{H.170}$$

We next bound the derivative of δ_{ξ_t} for $\xi_t \in (-a, a)$ arbitrary.

We have that:

$$\frac{d}{d\xi_t}\delta_{\xi_t} = \frac{1}{\ell}\sum_{k\notin\{i,j\}}\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)\langle\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}, \boldsymbol{u}\rangle \tag{H.171}$$

which yields that:

$$|\frac{d}{d\xi_t}\delta_{\xi_t}| \leq \frac{1}{\ell}\sum_{k\notin\{i,j\}}|\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)||\langle\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}, \boldsymbol{u}\rangle|, \text{ triangle inequality} \tag{H.172}$$

$$\leq \frac{1}{\ell}\sum_{k\notin\{i,j\}}|\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)|||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2||\boldsymbol{u}||_2 \tag{H.173}$$

$$\leq \frac{\max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell}(N-2)\max_{k\notin\{i,j\}}|\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)|. \tag{H.174}$$

We then have that:

$$|\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)| = |\langle\nabla_{\boldsymbol{y}_t}\tilde{\gamma}_k(\boldsymbol{y}_t), \frac{d\boldsymbol{y}_t}{d\xi_t}\rangle|, \text{ chain rule} \tag{H.175}$$

$$= \frac{\ell\tilde{\gamma}_k(\boldsymbol{y}_t)}{{\tilde{\sigma}_t}^2}|\langle\boldsymbol{\mu}^{(k)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t), \boldsymbol{u}\rangle|, \tag{H.176}$$

$$\leq \frac{\ell\tilde{\gamma}_k(\boldsymbol{y}_t)}{{\tilde{\sigma}_t}^2}||\boldsymbol{\mu}^{(k)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t)||_2, \text{ Cauchy-Schwartz, } ||\boldsymbol{u}||_2 = 1. \tag{H.177}$$

$$\leq \frac{\ell}{\tilde{\sigma}_t^2}\frac{\pi_k}{\pi_{d^*}}\exp(-\kappa)||\boldsymbol{\mu}^{(k)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t)||_2, \text{ Eq. (G.72)} \tag{H.178}$$

Then, we have that:

$$||\boldsymbol{\mu}^{(k)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t)||_2 = ||\boldsymbol{\mu}^{(k)} - \hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(i)}||_2 \tag{H.179}$$

$$\leq ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + ||\hat{\boldsymbol{\mu}}(\boldsymbol{y}_t) - \boldsymbol{\mu}^{(i)}||_2 \tag{H.180}$$

$$\leq \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + ||\sum_{m \in [N]} \tilde{\gamma}_m(\boldsymbol{y}_t)\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2 \tag{H.181}$$

$$\leq \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \tilde{\gamma}_j(\boldsymbol{y}_t)||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2 + ||\sum_{m \notin \{i,j\}} \tilde{\gamma}_m(\boldsymbol{y}_t)\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2 \tag{H.182}$$

$$\leq \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\exp(-\frac{(1-2\xi)\ell^2}{2\tilde{\sigma}_t^2})||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2 + ||\sum_{m \notin \{i,j\}} \tilde{\gamma}_m(\boldsymbol{y}_t)\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2, \text{ Eq. (H.131)} \tag{H.183}$$

$$\leq \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\ell\exp(-\kappa) + ||\sum_{m \notin \{i,j\}} \tilde{\gamma}_m(\boldsymbol{y}_t)\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2, \ \xi_t \in (-a,a), 1-2a \geq \frac{1}{2} \tag{H.184}$$

$$\leq \max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\ell\exp(-\kappa) + \max_{m \notin \{i,j\}} ||\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2(N-2)\exp(-\kappa), \text{ Eq. (G.72)}. \tag{H.185}$$

Incorporating Eq. (H.185) into Eq. (H.178) yields:

$$|\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)| \leq |\sup_{\xi_t \in (-a,a)} \frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)| \tag{H.186}$$

$$\leq |\sup_{\xi_t \in (-a,a)} \frac{\ell}{\tilde{\sigma}_t^2}\frac{\pi_k}{\pi_{d^*}}(\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2)| \tag{H.187}$$

$$\leq \frac{\ell}{\tilde{\sigma}_t^2}\frac{\pi_k}{\pi_{d^*}}\exp(-\kappa)(\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\ell\exp(-\kappa) \tag{H.188}$$

$$+ \max_{m \notin \{i,j\}} ||\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2(N-2)\exp(-\kappa)) \tag{H.189}$$

Then, incorporating Eq. (H.189) into Eq. (H.174) yields:

$$|\frac{d}{d\xi_t}\delta_{\xi_t}| \leq \frac{\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell}(N-2)\max_{k \notin \{i,j\}} |\frac{d}{d\xi_t}\tilde{\gamma}_k(\boldsymbol{y}_t)| \tag{H.190}$$

$$\leq \frac{\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell}(N-2)\frac{\ell}{\tilde{\sigma}_t^2}\max_{k \notin \{i,j\}} \frac{\pi_k}{\pi_{d^*}}\exp(-\kappa)(\max_{k \notin \{i,j\}} ||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\ell\exp(-\kappa) \tag{H.191}$$

$$+ \max_{m \notin \{i,j\}} ||\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2(N-2)\exp(-\kappa)) \tag{H.192}$$

Incorporating Eq. (H.170) and Eq. (H.192) into Eq. (H.152) yields:

$$\frac{d}{d\xi_t}F_t(\xi_t^{*,i}) \leq |\frac{d}{d\xi_t}F_t(\xi_t^{*,i})| \tag{H.193}$$

$$= \left|\left(\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{d}{d\xi_t}\delta_{\xi_t} - 1\right)|_{\xi_t=\xi_t^{*,i}}\right| \tag{H.194}$$

$$\leq \sup_{\xi_t\in(-a,a)} \left|\left(\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t) + \frac{d}{d\xi_t}\delta_{\xi_t} - 1\right)\right| \tag{H.195}$$

$$\leq \sup_{\xi_t\in(-a,a)} |\frac{d}{d\xi_t}\tilde{\gamma}_j(\boldsymbol{y}_t)| + \sup_{\xi_t\in(-a,a)} |\frac{d}{d\xi_t}\delta_{\xi_t}| - 1 \tag{H.196}$$

$$\leq 8\frac{\pi_j}{\pi_i}\kappa\exp(-\kappa) + \max_{k\notin\{i,j\}} 2||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2\kappa\exp(-\kappa)(N-2)\max_{k\notin\{i,j\}}\frac{\pi_k}{\pi_{d^*}}\exp(-\kappa) \tag{H.197}$$

$$+ \frac{\max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2}{\ell}(N-2)\frac{\ell}{\sigma^2}\max_{k\notin\{i,j\}}\frac{\pi_k}{\pi_{d^*}}\exp(-\kappa)(\max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2 + \frac{\pi_j}{\pi_i}\ell\exp(-\kappa) \tag{H.198}$$

$$+ \max_{m\notin\{i,j\}}||\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}^{(i)}||_2(N-2)\exp(-\kappa)) - 1 \tag{H.199}$$

since $\frac{1}{\tilde{\sigma}_t^2} \leq \frac{1}{\min_{t\in[t,\hat{\tau})}\tilde{\sigma}_t^2} = \frac{1}{\sigma^2}$.

Define $W_{i,j} := \frac{\pi_j}{\pi_i}$, $M := \max_{k\notin\{i,j\}}||\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(i)}||_2$, and $R := \max_{k\notin\{i,j\}} W_{k,d^*}$. We then must choose κ s.t.

$$8W_{i,j}\kappa\exp(-\kappa) + 2M\kappa\exp(-\kappa)(N-2)R\exp(-\kappa) \tag{H.200}$$

$$+\frac{M(N-2)R}{\sigma^2}(M\exp(-\kappa) + W_{i,j}\ell\exp(-2\kappa) + M(N-2)\exp(-2\kappa)) < 1 \tag{H.201}$$

Denote:

$$T_1(\kappa) = 8W_{i,j}\kappa\exp(-\kappa), \tag{H.202}$$

$$T_2(\kappa) = 2M\kappa\exp(-\kappa)(N-2)R\exp(-\kappa), \tag{H.203}$$

and,

$$T_3(\kappa) = \frac{M(N-2)R}{\sigma^2}(M\exp(-\kappa) + W_{i,j}\ell\exp(-2\kappa) + M(N-2)\exp(-2\kappa)) \tag{H.204}$$

We first bound T_1:

$$T_1(\kappa) = 8W_{i,j}\kappa\exp(-\kappa) \tag{H.205}$$

$$\leq 8W_{i,j}\exp(-\kappa/2) < \frac{1}{3} \tag{H.206}$$

$$\iff \exp(-\kappa/2) < \frac{1}{24W_{i,j}} \tag{H.207}$$

$$\iff \kappa > 2\log(24W_{i,j}) \tag{H.208}$$

$$\in 2\log N + \Theta(1) \tag{H.209}$$

Next, we bound T_2. Since $\kappa > 1$, we have that $\kappa \exp(-\kappa) \leq \exp(-1)$, hence $\kappa \exp(-2\kappa) = \exp(-\kappa)(\kappa \exp(-\kappa)) \leq \exp(-1)\exp(-\kappa)$. Thus:

$$T_2(\kappa) = 2M\kappa \exp(-2\kappa)(N-2)R \tag{H.210}$$
$$\leq 2M \exp(-1)\exp(-\kappa)(N-2)R < \frac{1}{3} \tag{H.211}$$
$$\iff \exp(-\kappa) < \frac{\exp(1)}{6MR(N-2)} \tag{H.212}$$
$$\iff \kappa > \log(6\exp(-1)MR(N-2)) \tag{H.213}$$
$$\in \log N + \Theta(1) \tag{H.214}$$

Finally, we bound T_3. Since the inverse exponential is monotonically decreasing, $\exp(-2\kappa) \leq \exp(-\kappa)$. Hence, $M\exp(-\kappa) + W_{i,j}\ell(\exp(-2\kappa) + M(n_2)\exp(-2\kappa) \leq (M + W_{i,j}\ell + M(N-2))\exp(-\kappa)$. Define the constant:

$$A := \frac{M(N-2)R}{\sigma^2}(M + W_{i,j}\ell + M(N-2)) \tag{H.215}$$

Then:

$$T_3(\kappa) = \frac{M(N-2)R}{\sigma^2}(M\exp(-\kappa) + W_{i,j}\ell\exp(-2\kappa) + M(N-2)\exp(-2\kappa)) \tag{H.216}$$
$$\leq A\exp(-\kappa) < \frac{1}{3} \tag{H.217}$$
$$\iff \exp(-\kappa) < \frac{1}{3A} \tag{H.218}$$
$$\iff \kappa > \log(3A) \tag{H.219}$$
$$\iff \kappa > \log(3\frac{M(N-2)R}{\sigma^2}(M + W_{i,j}\ell + M(N-2))) \tag{H.220}$$
$$\in 2\log N + \Theta(1) \tag{H.221}$$

Furthermore, $\kappa \geq \log N + \log(4C_a) \in 2\log N + \Theta(1)$. Thus, the maximum over all these constraints is also in $2\log N + \Theta(1)$, as desired. Finally, note that $t \in [0, \hat{\tau})$ was arbitrary, and the bound of the derivative of $F_t(\xi_t)$ evaluated at the equilibria does not depend on time, so $\xi_t^{*,i}$ is uniformly instantaneously stable for all $t \in [0, \hat{\tau})$, as desired. □

H.4. Proof of Corollary 4.6

Proof. Fix $\hat{\tau} = \min\{\tau, \tau_2\}$ as in Prop. 4.7. By Prop. 4.5, there exists an instantaneously stable equilibrium $\xi_t^{*,i}$ of Eq. (G.50) within a ε-ball of mode $\mu^{(i)}$, for all $t < \hat{\tau}$. Moreover, the proof of Prop. 4.5 shows that the stability is *uniform* over $t \in [0, \hat{\tau}]$, in the sense that there exists a constant $m_0 > 0$ (independent of t) such that

$$F_t'(\xi_t^{*,i}) \leq -m_0 \qquad \forall t \in [0, \hat{\tau}]. \tag{H.222}$$

Since $\tilde{\gamma}_j$ is a smooth composition of Gaussian p.d.f.s and Δ_{ξ_t} is differentiable in ξ (as used throughout the Prop. 4.5 proof), we have that F_t is C^1 in ξ for each fixed t. Since $F_t(\xi_t^{*,i}) = 0$ and $\partial_\xi F_t(\xi_t^{*,i}) \leq -m_0 < 0$, the implicit function theorem implies that the equilibria can be chosen as a continuous branch $t \mapsto \xi_t^{*,i}$ on $[0, \hat{\tau}]$.

Hence, by continuity of F_t' in ξ for each fixed $t \in [0, \hat{\tau}]$ there exists a radius $\varkappa_t > 0$ such that

$$F_t'(\xi) \leq -\frac{m_0}{2} \qquad \forall \xi \in [\xi_t^{*,i} - \varkappa_t,\ \xi_t^{*,i} + \varkappa_t]. \tag{H.223}$$

By joint continuity in (t, ξ) and compactness of $[0, \hat{\tau}]$, the radii may be chosen so that $\varkappa := \inf_{t \in [0,\hat{\tau}]} \varkappa_t > 0$. Recall from the proof of Prop. 4.5 that, for $a := 2(\varepsilon_{\delta_\xi} + \varepsilon_{\tilde{\gamma}})$, we have $F_t(-a) > 0, F_t(a) < 0, |\xi_t^{*,i}| \leq a, \forall t \in [0, \hat{\tau}]$. Since

$a \in O(N\exp(-\kappa))$, there exists $C > 0$ such that $a \le CN\exp(-\kappa)$. Thus, to ensure $2a \le \varkappa$, it suffices to take $\kappa \ge \log\big(\frac{2CN}{\varkappa}\big)$,, which is satisfied by the condition $\kappa \ge 2\log N + \Theta(1)$. Therefore, $[-a, a] \subseteq [\xi_t^{*,i} - \varkappa_t, \xi_t^{*,i} + \varkappa_t],\ \forall t \in [0, \hat{\tau}]$.

Let $\xi_t^{(1)}$ and $\xi_t^{(2)}$ be two solutions of Eq. (G.50) with $\xi_{t_0}^{(1)}, \xi_{t_0}^{(2)} \in [-a, a]$. Since $F_t(-a) > 0$ and $F_t(a) < 0$ uniformly in t, the interval $[-a, a]$ is invariant. Hence, for every $t < t_0$,

$$\xi_t^{(1)},\ \xi_t^{(2)} \in [-a, a] \subseteq [\xi_t^{*,i} - \varkappa_t,\ \xi_t^{*,i} + \varkappa_t]. \tag{H.224}$$

Define $\Delta_t := \xi_t^{(1)} - \xi_t^{(2)}$. Then Δ_t is C^1 in t, and:

$$\dot{\Delta}_t \ = \ F_t(\xi_t^{(1)}) - F_t(\xi_t^{(2)}). \tag{H.225}$$

Now, fix any $t < t_0 \le \hat{\tau}$. By the mean value theorem applied to the C^1 function $F_t(\cdot)$, there exists θ_t lying between $\xi_t^{(1)}$ and $\xi_t^{(2)}$ such that:

$$F_t(\xi_t^{(1)}) - F_t(\xi_t^{(2)}) \ = \ F_t'(\theta_t)\,(\xi_t^{(1)} - \xi_t^{(2)}) \ = \ F_t'(\theta_t)\,\Delta_t. \tag{H.226}$$

By Eq. (H.224), we have $\theta_t \in [\xi_t^{*,i} - \varkappa_t, \xi_t^{*,i} + \varkappa_t]$, so Eq. (H.223) yields $F_t'(\theta_t) \le -m_0/2$. Incorporating this into Eq. (H.225) and Eq. (H.226) yields:

$$\dot{\Delta}_t \ = \ F_t'(\theta_t)\,\Delta_t, \qquad \text{with } F_t'(\theta_t) \le -\frac{m_0}{2}. \tag{H.227}$$

Next, consider the absolutely continuous function $t \mapsto |\Delta_t|$. For t such that $\Delta_t \ne 0$,

$$\frac{d}{dt}|\Delta_t| \ = \ \mathrm{sign}(\Delta_t)\,\dot{\Delta}_t \ = \ \mathrm{sign}(\Delta_t)\,F_t'(\theta_t)\,\Delta_t \ = \ F_t'(\theta_t)\,|\Delta_t| \ \le \ -\frac{m_0}{2}\,|\Delta_t|. \tag{H.228}$$

Integrating Eq. (H.228) and applying Grönwall's inequality yields for all $t < \hat{\tau}$:

$$|\xi_t^{(1)} - \xi_t^{(2)}| \ = \ |\Delta_t| \ \le \ \exp\Big(-\frac{m_0}{2}\big(u(t) - u(t_0)\big)\Big)\,|\Delta_{t_0}| \ = \ \exp\Big(-\frac{m_0}{2}\big(u(t) - u(t_0)\big)\Big)\,|\xi_{t_0}^{(1)} - \xi_{t_0}^{(2)}|. \tag{H.229}$$

Taking $t_0 = \hat{\tau}$ and $\xi^{(2)} = \bar{\xi}$, where $\bar{\xi}_{\hat{\tau}} = 0$, gives a solution $\bar{\xi}_t \in [-a, a]$ s.t. for all $t \in [0, \hat{\tau}]$, every trajectory in $[-a, a]$ contracts exponentially fast to $\bar{\xi}_t$ by Eq. (H.229). Note that this follows symmetrically in a neighborhood of $\xi_t^{*,j}$.

□

H.5. Proof of Proposition 4.7

Before stating the proof, we firstly recall the parallel dynamics from Lemma G.3:

$$\dot{\xi}_t = \tilde{\gamma}_j(\boldsymbol{y}_t) - \xi_t + \delta_{\xi_t}, \tag{H.230}$$

where:

$$|\delta_\xi|, ||\boldsymbol{\delta_w}||_2 \le \varepsilon \in \mathcal{O}\left(N\exp(-\kappa)\right), \tag{H.231}$$

for:

$$\boldsymbol{y}_t = \boldsymbol{\mu}^{(i)} + \xi_t \ell \boldsymbol{u} + \boldsymbol{w}_t, \tag{H.232}$$

where $\ell = ||\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}||_2, \boldsymbol{u} = \frac{\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}}{\ell}$, and $\boldsymbol{w}_t$ satisfies $\langle \boldsymbol{w}_t, \boldsymbol{u} \rangle = 0$, and we also consider the responsibility of mode j:

$$\tilde{\gamma}_j(\boldsymbol{y}_t) = \frac{\pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\sum_{m \in [N]} \pi_m \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(m)}, \tilde{\sigma}_t^2 \boldsymbol{I})}. \tag{H.233}$$

We simplify the parallel dynamics by dropping the term bounded by ε, and additionally consider only the dynamics with respect to the two-mode responsibility, i.e. we consider:

$$\dot{\xi}_t = \tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t) - \xi_t, \tag{H.234}$$

where:

$$\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t) := \frac{\pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\pi_i \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2 \boldsymbol{I}) + \pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})}. \tag{H.235}$$

Throughout the statement and the proof, we work in the rescaled forward time u defined in Lemma G.2 and relabel this rescaled time by t for simplicity. Thus, the value ξ_{τ_3} denotes the initial value of the approximate parallel dynamics over the remaining time interval of length τ_3. Then, we prove the result:

Proof. **(1)**: Since the instance is within the ε-tube around $L_t^{(i,j)}$, we have that $||\boldsymbol{w}_t|| \leq \varepsilon$, just as we do Prop. 4.5. Thus, by the same token as in the proof of Prop. 4.5, we have that:

$$||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2 = \xi_t^2 \ell^2 + ||\boldsymbol{w}_t||_2^2, \tag{H.236}$$

and:

$$||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 = ||(\xi_t - 1)\ell \boldsymbol{u} + \boldsymbol{w}_t||_2^2 = (1 - \xi_t)^2 \ell^2 + ||\boldsymbol{w}_t||_2^2, \tag{H.237}$$

yielding:

$$||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2 = (1 - 2\xi_t)\ell^2. \tag{H.238}$$

From Eq. (H.235), we have that:

$$\frac{\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)}{1 - \tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)} = \frac{\pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(j)}, \tilde{\sigma}_t^2 \boldsymbol{I})}{\pi_i \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}^{(i)}, \tilde{\sigma}_t^2 \boldsymbol{I})} \tag{H.239}$$

$$= \frac{\pi_j}{\pi_i} \exp(-\frac{||\boldsymbol{y}_t - \boldsymbol{\mu}^{(j)}||_2^2 - ||\boldsymbol{y}_t - \boldsymbol{\mu}^{(i)}||_2^2}{2\tilde{\sigma}_t^2}) \tag{H.240}$$

$$= \frac{\pi_j}{\pi_i} \exp(\frac{\ell^2}{\tilde{\sigma}_t^2}(\xi - \frac{1}{2})), \text{ Eq. (H.238)} \tag{H.241}$$

Since $\pi_i = \pi_j$, at $\xi = \frac{1}{2}$, Eq. (H.241) evaluates to:

$$\frac{\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)}{1 - \tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)}\Big|_{\xi_t = 1/2} = 1 \tag{H.242}$$

yielding:

$$\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)_{\xi_t = 1/2} = 1/2 \tag{H.243}$$

Define $F_t(\xi_t) = \tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t) - \xi_t$. Then, $F_t(1/2) = 0$, so $\xi^* = \frac{1}{2}$ is an (instantaneous) equilibrium of Eq. (H.234).

The corresponding point in $\mathbb{R}^d$ is thus:

$$\boldsymbol{y}_t^* := \boldsymbol{y}_t\big|_{\xi=1/2} \tag{H.244}$$
$$= \boldsymbol{\mu}^{(i)} + \frac{1}{2}\ell\boldsymbol{u} + \boldsymbol{w}_t \tag{H.245}$$
$$= \frac{\boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(j)}}{2} + \boldsymbol{w}_t. \tag{H.246}$$

Next, we demonstrate that this is a hyperbolic unstable (instantaneous) equilibria. We have that:

$$\frac{d}{d\xi_t}F_t(\xi_t) = \frac{d}{d\xi_t}\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t) - 1 \tag{H.247}$$
$$= \frac{\ell^2}{\tilde{\sigma}_t^2}\tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)(1 - \tilde{\gamma}_j^{(i,j)}(\boldsymbol{y}_t)) - 1, \tag{H.248}$$

We thus have that:

$$\frac{d}{d\xi_t}F_t(\xi_t)\big|_{\xi_t=1/2} = \frac{\ell^2}{\tilde{\sigma}_t^2}\frac{1}{2}\frac{1}{2} - 1 = \frac{\ell^2}{4\tilde{\sigma}_t^2} - 1. \tag{H.249}$$

By Asm. 4.4, we thus have that:

$$\lambda_t := \frac{\ell^2}{4\tilde{\sigma}_t^2} - 1 \geq \kappa - 1 > 0 \tag{H.250}$$

since $\kappa > 1$. Since λ_t has no complex components, $\xi_t = 1/2$ thus corresponds to a hyperbolic unstable (instantaneous) equilibria.

(2): Consider $t \in [0, \tau_3]$ and recall the approximate parallel dynamics Eq. (H.234). Note that $\underline{\lambda}$ and $\bar{\lambda}$ exist by the extreme value theorem, since they are defined on a compact domain with respect to continuous λ_t (since $\tilde{\sigma}_t^2$ is continuous). Within $\text{Tube}_{\varepsilon,t}^{(i,j)}$, by Eq. (H.241), we have the identity:

$$\frac{\tilde{\gamma}_j^{(i,j)}(\xi_t)}{1 - \tilde{\gamma}_j^{(i,j)}(\xi_t)} = \frac{\pi_j}{\pi_i}\exp\left(\frac{\ell^2}{\tilde{\sigma}_t^2}\left(\xi_t - \frac{1}{2}\right)\right). \tag{H.251}$$

Noting that $\pi_i = \pi_j$, denoting $a_t := \frac{\ell^2}{\tilde{\sigma}_t^2}$ and solving the log-odds equation obtains the logistic form:

$$\tilde{\gamma}_j^{(i,j)} = \frac{1}{1 + \exp(-a_t(\xi_t - \frac{1}{2}))} =: \varsigma\left(a_t\left(\xi_t - \frac{1}{2}\right)\right), \tag{H.252}$$

where $\varsigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Since $\varsigma'(z) = \varsigma(z)(1 - \varsigma(z))$ and $\varsigma''(z) = \varsigma'(z)(1 - 2\varsigma(z))$, and since $0 \leq \varsigma \leq 1$ implies $0 \leq \varsigma' \leq \frac{1}{4}$ and $|1 - 2\varsigma| \leq 1$, we have $|\varsigma''(z)| \leq \frac{1}{4}$. Hence,

$$|F_t''(\xi_t)| = \left|\tilde{\gamma}_j^{(i,j)\prime\prime}(\xi_t)\right| \leq \frac{a_t^2}{4}. \tag{H.253}$$

Next, for an a to be chosen, consider $0 < |a| \leq \vartheta$. Since $F_t \in C^2$, Taylor's theorem with Lagrange remainder at $\xi = \frac{1}{2}$ gives:

$$F_t\left(\frac{1}{2} + a\right) = F_t\left(\frac{1}{2}\right) + F_t'\left(\frac{1}{2}\right)a + \frac{1}{2}F_t''\left(\frac{1}{2} + ra\right)a^2 \tag{H.254}$$

for some $r \in (0, 1)$. Using $F_t(\frac{1}{2}) = 0$ and $F_t'(\frac{1}{2}) = \lambda_t$, we obtain

$$F_t\left(\frac{1}{2} + a\right) = \lambda_t a + \frac{1}{2}F_t''\left(\frac{1}{2} + ra\right)a^2. \tag{H.255}$$

By Eq. (H.253),

$$\left|F_t\left(\frac{1}{2}+a\right)-\lambda_t a\right| \le \frac{1}{2}\cdot\frac{a_t^2}{4}|a|^2 = \frac{a_t^2}{8}|a|^2. \tag{H.256}$$

Dividing by $|a|$ yields

$$\left|\frac{F_t(\frac{1}{2}+a)}{a}-\lambda_t\right| \le \frac{a_t^2}{8}|a| \le \frac{a_t^2}{8}\vartheta, \tag{H.257}$$

and thus

$$\lambda_t - \frac{a_t^2}{8}\vartheta \le \frac{F_t(\frac{1}{2}+a)}{a} \le \lambda_t + \frac{a_t^2}{8}\vartheta. \tag{H.258}$$

Since $\lambda_t = \frac{\ell^2}{4\tilde{\sigma}_t^2}-1$, we have $a_t = \frac{\ell^2}{\tilde{\sigma}_t^2} = 4(1+\lambda_t)$, hence $\frac{a_t^2}{8} = 2(1+\lambda_t)^2$. Plugging into Eq. (H.258) gives

$$\frac{F_t(\frac{1}{2}+a)}{a} \le \lambda_t + 2(1+\lambda_t)^2\vartheta \le \bar{\lambda} + 2(1+\bar{\lambda})^2\vartheta =: \Lambda_+, \tag{H.259}$$

and similarly

$$\frac{F_t(\frac{1}{2}+a)}{a} \ge \lambda_t - 2(1+\lambda_t)^2\vartheta \ge \underline{\lambda} - 2(1+\bar{\lambda})^2\vartheta. \tag{H.260}$$

By our choice of radius ϑ in the statement of Prop. 4.7, we have $\underline{\lambda} - 2(1+\bar{\lambda})^2\vartheta > 0$, hence $\frac{F_t(\frac{1}{2}+a)}{a} > 0$, so

$$\operatorname{sign}\big(F_t\big(\tfrac{1}{2}+a\big)\big) = \operatorname{sign}(a) \qquad \text{whenever } 0 < |a| \le \vartheta,$$

uniformly for all $t \in [0, \tau_3]$.

Now fix a trajectory that stays in the neighborhood $0 < |\xi_t - \frac{1}{2}| \le \vartheta$ and define

$$A_t := \xi_t - \frac{1}{2}.$$

Then $A_t \neq 0$ and, using the sign property above,

$$\frac{d}{dt}|A_t| = \operatorname{sign}(A_t)\dot{A}_t = \operatorname{sign}(A_t)\dot{\xi}_t = \operatorname{sign}(A_t)F_t\big(\tfrac{1}{2}+A_t\big) = \big|F_t\big(\tfrac{1}{2}+A_t\big)\big|. \tag{H.261}$$

Using Eq. (H.259) with $a = A_t$ yields

$$\big|F_t\big(\tfrac{1}{2}+A_t\big)\big| \le \Lambda_+|A_t|,$$

and hence

$$\frac{d}{dt}|A_t| \le \Lambda_+|A_t| \qquad \text{whenever } 0 < |A_t| \le \vartheta. \tag{H.262}$$

Next, we show by contradiction that a trajectory ξ_t satisfying $0 < |\xi_{\tau_3} - \frac{1}{2}| < \vartheta\exp(-\Lambda_+\tau_3)$ never exits the ϑ-neighborhood on $[0, \tau_3]$. Define $A_t = \xi_t - \frac{1}{2}$ and relabel s as t. Let ϱ be the first exit time from the neighborhood $|A_t| < \vartheta$, i.e.

$$\varrho := \inf\{t \in [0, \tau_3] : |A_t| \ge \vartheta\}, \tag{H.263}$$

with the convention that $\inf\emptyset = +\infty$. Suppose for contradiction that $\varrho \le \tau_3$. Then $|A_t| < \vartheta$ for all $t \in [0, \varrho)$ and $|A_\varrho| = \vartheta$ by continuity. On $[0, \varrho)$, Eq. (H.262) holds, so Grönwall's inequality gives

$$|A_t| \le |A_{\tau_3}|\exp(\Lambda_+ t) \qquad \forall t \in [0, \varrho). \tag{H.264}$$

Taking $t \to \varrho$ in Eq. (H.264) and using continuity of $|A_t|$ yields:

$$|A_\varrho| \le |A_{\tau_3}|\exp(\Lambda_+\varrho) \le |A_{\tau_3}|\exp(\Lambda_+\tau_3) < \vartheta, \tag{H.265}$$

since $|A_{\tau_3}| = \left|\xi_{\tau_3} - \frac{1}{2}\right| < \vartheta \exp(-\Lambda_+ \tau_3)$. This contradicts $|A_\varrho| = \vartheta$. Hence, $\varrho = +\infty$, and therefore $|A_t| < \vartheta$ for every $t \in [0, \tau_3]$. This thus yields:

$$\left|\xi_t - \frac{1}{2}\right| < \vartheta \quad \text{for every } t \in [0, \tau_3], \tag{H.266}$$

which is the persistence result.

Finally, by contrapositive, if exit does occur by time 0, necessarily:

$$\left|\xi_{\tau_3} - \frac{1}{2}\right| \geq \vartheta \exp(-\Lambda_+ \tau_3), \tag{H.267}$$

and thus a necessary condition for exit is:

$$\tau_3 \geq \frac{1}{\Lambda_+} \log\left(\frac{\vartheta}{|\xi_{\tau_3} - \frac{1}{2}|}\right), \tag{H.268}$$

as desired. □

H.6. Proof of Proposition 5.1

Proof. Set $T := \tau_3$. Under the time relabeling $\theta := T - t$, the proposition's event $\{|A_0| \leq \vartheta\}$ corresponds to $\{|A_T| \leq \vartheta\}$; $A_T = 0$ in Eq. (4) corresponds to $A_0 = 0$; and the assumption with λ_{rep} is negated in the resulting forward-time process, which we analyze below.

Define the continuous martingale

$$M_t := \int_0^t \sqrt{2\eta(s)}\, dB_s, \qquad t \in [0, T]. \tag{H.269}$$

Integrating Eq. (4) yields the decomposition

$$A_t = \int_0^t b(A_s, s)\, ds + M_t, \qquad t \in [0, T]. \tag{H.270}$$

We first bound the confinement event

$$E_{2\vartheta} := \Big\{ \sup_{0 \leq t \leq T} |A_t| \leq 2\vartheta \Big\}. \tag{H.271}$$

On $E_{2\vartheta}$ we have $|A_s| \leq 2\vartheta$ for all $s \in [0, T]$, hence by the local drift bound,

$$|b(A_s, s)| \leq K(s)\, |A_s| \leq 2K(s)\vartheta, \qquad s \in [0, T]. \tag{H.272}$$

Therefore, for each $t \in [0, T]$ and on $E_{2\vartheta}$,

$$\Big| \int_0^t b(A_s, s)\, ds \Big| \leq \int_0^t |b(A_s, s)|\, ds \leq 2\vartheta \int_0^T K(s)\, ds. \tag{H.273}$$

Combining Eqs. (H.270) and (H.273), for all $t \in [0, T]$ and on $E_{2\vartheta}$,

$$|M_t| = \Big| A_t - \int_0^t b(A_s, s)\, ds \Big| \leq |A_t| + \Big| \int_0^t b(A_s, s)\, ds \Big| \leq 2\vartheta \Big(1 + \int_0^T K(s)\, ds\Big). \tag{H.274}$$

Hence,

$$E_{2\vartheta} \subseteq \Big\{ \sup_{0 \leq t \leq T} |M_t| \leq 2\Big(1 + \int_0^T K(s)\, ds\Big)\vartheta \Big\}. \tag{H.275}$$

We next identify the quadratic variation of M. Fix $t \in [0, T]$ and let $\pi = \{0 = t_0 < t_1 < \cdots < t_n = t\}$ be a partition of $[0, t]$ with mesh

$$|\pi| := \max_{0 \leq k \leq n-1} (t_{k+1} - t_k). \tag{H.276}$$

Write increments

$$\Delta M_k := M_{t_{k+1}} - M_{t_k} = \int_{t_k}^{t_{k+1}} \sqrt{2\eta(s)}\, dB_s. \tag{H.277}$$

Define

$$Q(\pi) := \sum_{k=0}^{n-1} (\Delta M_k)^2. \tag{H.278}$$

By Itô isometry applied to Eq. (H.277),

$$\mathbb{E}\big[(\Delta M_k)^2\big] = \int_{t_k}^{t_{k+1}} 2\eta(s)\, ds =: v_k. \tag{H.279}$$

Since ΔM_k is Gaussian with mean 0 and variance v_k,

$$\mathbb{E}\big[(\Delta M_k)^4\big] = 3v_k^2, \qquad \mathrm{Var}\big((\Delta M_k)^2\big) = 2v_k^2. \tag{H.280}$$

Using independence of Brownian increments,

$$\mathbb{E}[Q(\pi)] = \sum_{k=0}^{n-1} v_k = \int_0^t 2\eta(s)\, ds, \tag{H.281}$$

and

$$\mathrm{Var}(Q(\pi)) = 2\sum_{k=0}^{n-1} v_k^2 \le 2\Big(\max_{0\le k\le n-1} v_k\Big) \sum_{k=0}^{n-1} v_k. \tag{H.282}$$

Since $\eta \in L^1([0,T])$, absolute continuity of the integral implies $\max_k v_k \to 0$ as $|\pi| \to 0$. Hence,

$$\mathbb{E}\Big[\big(Q(\pi) - \int_0^t 2\eta(s)\, ds\big)^2\Big] = \mathrm{Var}(Q(\pi)) \to 0 \qquad \text{as } |\pi| \to 0. \tag{H.283}$$

Therefore $Q(\pi) \to \int_0^t 2\eta(s)\, ds$ in probability. In particular, the quadratic variation

$$[M]_t := \lim_{|\pi|\to 0} Q(\pi) \tag{H.284}$$

satisfies

$$[M]_t = \int_0^t 2\eta(s)\, ds, \qquad t \in [0,T]. \tag{H.285}$$

Next, by Dambis–Dubins–Schwarz (Dubins & Schwarz, 1965), there exists a standard Brownian motion $(W_r)_{r\ge 0}$ such that

$$M_t = W_{[M]_t}, \qquad t \in [0,T]. \tag{H.286}$$

Since $[M]_t$ is nondecreasing,

$$\sup_{0\le t\le T} |M_t| = \sup_{0\le r\le [M]_T} |W_r|, \qquad [M]_T = 2\int_0^T \eta(s)\, ds. \tag{H.287}$$

Combining Eqs. (H.275) and (H.287) gives

$$\mathbb{P}(E_{2\vartheta}) \le \mathbb{P}\Big(\sup_{0\le r\le [M]_T} |W_r| \le 2\Big(1 + \int_0^T K(s)\, ds\Big)\vartheta\Big). \tag{H.288}$$

Applying Lemma G.6 with

$$V = [M]_T = 2\int_0^T \eta(s)\, ds, \qquad a = 2\Big(1 + \int_0^T K(s)\, ds\Big)\vartheta, \tag{H.289}$$

yields the confinement bound used in Eq. (7).

We now bound the terminal event. Let $(\mathcal{F}_t)_{t\in[0,T]}$ be the natural filtration of (A_t). Define

$$H_\vartheta := \{|A_T| \le \vartheta\}.$$

Then

$$\mathbb{P}(H_\vartheta) = \mathbb{P}(H_\vartheta \cap E_{2\vartheta}) + \mathbb{P}(H_\vartheta \cap E_{2\vartheta}^c) \le \mathbb{P}(E_{2\vartheta}) + \mathbb{P}(H_\vartheta \cap E_{2\vartheta}^c). \tag{H.290}$$

Next, let

$$\tau := \inf\{t \in [0,T] : |A_t| \ge 2\vartheta\}. \tag{H.291}$$

By convention, $\tau = +\infty$ if no such time exists. By continuity,

$$E_{2\vartheta} = \{\tau > T\}, \qquad E_{2\vartheta}^c = \{\tau \le T\}.$$

Hence

$$\mathbb{P}(H_\vartheta \cap E_{2\vartheta}^c) = \mathbb{P}(|A_T| \le \vartheta,\ \tau \le T). \tag{H.292}$$

Using conditional expectation at τ and the strong Markov property,

$$\begin{aligned}\mathbb{P}(|A_T| \le \vartheta,\ \tau \le T) &= \mathbb{E}\Big[\mathbf{1}_{\{\tau\le T\}}\, \mathbb{P}_{\tau,A_\tau}(|A_T| \le \vartheta)\Big] \\ &\le \sup_{s\in[0,T]} \sup_{a\in\{\pm 2\vartheta\}} \mathbb{P}_{s,a}(|A_T| \le \vartheta).\end{aligned} \tag{H.293}$$

Fix $s \in [0,T]$ and consider $A_s = 2\vartheta$. Define the first return time

$$\sigma := \inf\{t \in [s,T] : A_t = \vartheta\}. \tag{H.294}$$

By continuity,

$$\mathbb{P}_{s,2\vartheta}(|A_T| \le \vartheta) \le \mathbb{P}_{s,2\vartheta}(\sigma \le T). \tag{H.295}$$

By Eq. (6), for all $t \in [0,T]$,

$$b(a,t) \ge \lambda_{\text{rep}}\, a \qquad \forall a \ge \vartheta, \tag{H.296}$$

and recall $\eta(t) \le \eta_{\max}$ on $[0,T]$. For $r > 0$ define $Z_t := e^{-rA_t}$. Applying Itô's formula on $[s, \sigma \wedge T]$ gives

$$dZ_t = Z_t\big(-r\, b(A_t,t) + r^2\eta(t)\big)\, dt - r Z_t \sqrt{2\eta(t)}\, dB_t. \tag{H.297}$$

On $[s,\sigma)$, using (H.296) and $\eta(t) \le \eta_{\max}$,

$$-r\, b(A_t,t) + r^2\eta(t) \le -r(\lambda_{\text{rep}}\vartheta) + r^2\eta_{\max}. \tag{H.298}$$

Choose

$$r := \frac{\lambda_{\text{rep}}\vartheta}{2\eta_{\max}}. \tag{H.299}$$

Then

$$-r(\lambda_{\text{rep}}\vartheta) + r^2\eta_{\max} = -\frac{\lambda_{\text{rep}}^2\vartheta^2}{4\eta_{\max}} < 0, \tag{H.300}$$

so $(Z_{t\wedge\sigma})_{t\in[s,T]}$ is a supermartingale. By optional stopping,

$$\mathbb{E}[Z_{\sigma\wedge T}] \le Z_s = e^{-r\cdot 2\vartheta}. \tag{H.301}$$

On $\{\sigma \le T\}$ we have $Z_{\sigma\wedge T} = e^{-r\vartheta}$, hence

$$\mathbb{P}_{s,2\vartheta}(\sigma \le T) \le e^{-r\vartheta} = \exp\Big(-\frac{\lambda_{\text{rep}}\vartheta^2}{2\eta_{\max}}\Big). \tag{H.302}$$

An identical argument with $\widetilde{Z}_t := e^{rA_t}$ yields the same bound starting from $A_s = -2\vartheta$. Therefore,

$$\sup_{s\in[0,T]} \sup_{a\in\{\pm 2\vartheta\}} \mathbb{P}_{s,a}(|A_T| \le \vartheta) \le 2\exp\Big(-\frac{\lambda_{\mathrm{rep}}\vartheta^2}{2\eta_{\max}}\Big). \tag{H.303}$$

Plugging this into (H.293) and then into (H.290) yields

$$\mathbb{P}(H_\vartheta) \le \mathbb{P}(E_{2\vartheta}) + 2\exp\Big(-\frac{\lambda_{\mathrm{rep}}\vartheta^2}{2\eta_{\max}}\Big). \tag{H.304}$$

Finally, substitute the bound on $\mathbb{P}(E_{2\vartheta})$ from Eq. (H.288) (via Lemma G.6) into Eq. (H.304) to obtain Eq. (7). □

H.7. Proof of Proposition B.1

Proof. **Unequal Weights**

We consider Eq. (H.234) and $t \le \hat{\tau}$, where $\hat{\tau}$ is defined in Asm. 4.4. By Eq. (H.241), we have that:

$$\begin{aligned}\frac{\tilde{\gamma}_{j,t}^{(ij)}(\xi)}{1-\tilde{\gamma}_{j,t}^{(ij)}(\xi)} &= \frac{\pi_j}{\pi_i}\exp\Big(-\frac{\|\boldsymbol{y}(\xi)-\boldsymbol{\mu}^{(j)}\|_2^2 - \|\boldsymbol{y}(\xi)-\boldsymbol{\mu}^{(i)}\|_2^2}{2\tilde{\sigma}_t^2}\Big)\\ &= \frac{\pi_j}{\pi_i}\exp\Big(\frac{\ell^2}{\tilde{\sigma}_t^2}\Big(\xi-\frac{1}{2}\Big)\Big),\end{aligned} \tag{H.305}$$

Define the pairwise parallel drift

$$F_{ij,t}(\xi) := \tilde{\gamma}_{j,t}^{(ij)}(\xi) - \xi. \tag{H.306}$$

Let $\xi_{ij}^\star(t) \in (0,1)$ be an interior equilibrium, i.e. $F_{ij,t}(\xi_{ij}^\star(t)) = 0$, so $\tilde{\gamma}_{j,t}^{(ij)}(\xi_{ij}^\star(t)) = \xi_{ij}^\star(t)$. Taking log-odds in Eq. (H.305) and substituting $\tilde{\gamma}_{j,t}^{(ij)}(\xi_{ij}^\star(t)) = \xi_{ij}^\star(t)$ yields

$$\log\Big(\frac{\xi_{ij}^\star(t)}{1-\xi_{ij}^\star(t)}\Big) = \log\Big(\frac{\pi_j}{\pi_i}\Big) + \frac{\ell^2}{\tilde{\sigma}_t^2}\Big(\xi_{ij}^\star(t)-\frac{1}{2}\Big), \tag{H.307}$$

which is Eq. (B.8). In particular, if $\pi_i = \pi_j$ then $\xi_{ij}^\star(t) = \frac{1}{2}$. If $\pi_j < \pi_i$, then $\log(\pi_j/\pi_i) < 0$ and Eq. (H.307) implies $\xi_{ij}^\star(t) > \frac{1}{2}$ (and similarly $\pi_j > \pi_i$ implies $\xi_{ij}^\star(t) < \frac{1}{2}$).

Next, differentiating Eq. (H.305) yields:

$$\tilde{\gamma}_{j,t}^{(ij)\,\prime}(\xi) = \frac{\ell^2}{\tilde{\sigma}_t^2}\,\tilde{\gamma}_{j,t}^{(ij)}(\xi)\big(1-\tilde{\gamma}_{j,t}^{(ij)}(\xi)\big), \tag{H.308}$$

so

$$F_{ij,t}'(\xi) = \tilde{\gamma}_{j,t}^{(ij)\,\prime}(\xi) - 1. \tag{H.309}$$

Evaluating at $\xi_{ij}^\star(t)$ (where $\tilde{\gamma}_{j,t}^{(ij)}(\xi_{ij}^\star(t)) = \xi_{ij}^\star(t)$) yields

$$F_{ij,t}'\big(\xi_{ij}^\star(t)\big) = \frac{\ell^2}{\tilde{\sigma}_t^2}\,\xi_{ij}^\star(t)\big(1-\xi_{ij}^\star(t)\big) - 1, \tag{H.310}$$

which is Eq. (B.9). Thus $\xi_{ij}^\star(t)$ is hyperbolic unstable whenever $F_{ij,t}'(\xi_{ij}^\star(t)) > 0$.

Exact Parallel Dynamics By Lemma G.3, the exact parallel dynamics in Eq. (G.50) differs from Eq. (H.234) by an error term δ_{ξ_t}. By Lemma G.4, we can thus write:

$$F_{N,t}(\xi) = F_{ij,t}(\xi) + e_t(\xi), \qquad \sup_{\xi\in[-\varepsilon,1+\varepsilon]} |e_t(\xi)| \le \varepsilon, \qquad \varepsilon \in \mathcal{O}(N\exp(-\kappa)). \tag{H.311}$$

We first show existence of an interior equilibrium for $F_{N,t}$, the exact parallel dynamics in Eq. (G.50). We denote $F_{ij,t}$ as the drift of the approximate parallel dynamics in Eq. (H.234). By Prop. 4.5, the exact parallel dynamics has instantaneously stable equilibria near $\xi = 0$ and $\xi = 1$, denoted $\xi_t^{*,i}$ and $\xi_t^{*,j}$, respectively. Instantaneous stability implies that the drift points toward each equilibrium. Hence we can pick points ξ_L, ξ_R with

$$\xi_t^{*,i} < \xi_t^{(L)} < \xi_t^{(R)} < \xi_t^{*,j}, \qquad F_{N,t}(\xi_L) < 0, \qquad F_{N,t}(\xi_R) > 0. \tag{H.312}$$

(For example, take ξ_L slightly to the right of $\xi_t^{*,i}$ and ξ_R slightly to the left of $\xi_t^{*,j}$.) Since $F_{N,t}$ is continuous in ξ, the intermediate value theorem applied to Eq. (H.312) yields an interior root $\xi_N^\star(t) \in (\xi_t^{(L)}, \xi_t^{(R)})$ with $F_{N,t}(\xi_N^\star(t)) = 0$.

Assume there exists an interval I containing $\xi_{ij}^\star(t)$ such that

$$F'_{ij,t}(\xi) \geq m > 0 \qquad \forall \xi \in I, \tag{H.313}$$

which is Eq. (B.10). Define

$$\rho := \sup\Big\{r > 0 : \ [\xi_{ij}^\star(t) - r, \ \xi_{ij}^\star(t) + r] \subset I\Big\}. \tag{H.314}$$

Since $\rho > 0$, recalling that $\varepsilon \leq C_1 \exp(-\kappa)$, we choose κ sufficiently large such that:

$$\frac{\varepsilon}{m} < \rho. \tag{H.315}$$

Now set $\xi_\pm := \xi_{ij}^\star(t) \pm \varepsilon/m$, so that Eq. (H.315) implies $[\xi_-, \xi_+] \subset I$. Since $F_{ij,t}(\xi_{ij}^\star(t)) = 0$ and $F'_{ij,t} \geq m$ on $[\xi_-, \xi_+] \subset I$, the mean value theorem gives

$$F_{ij,t}(\xi_+) \geq \varepsilon, \qquad F_{ij,t}(\xi_-) \leq -\varepsilon, \tag{H.316}$$

Combining Eqs. (H.311) and (H.316) gives

$$F_{N,t}(\xi_+) \geq \varepsilon - \varepsilon \geq 0, \qquad F_{N,t}(\xi_-) \leq -\varepsilon + \varepsilon \leq 0.$$

By continuity, there exists a root $\xi_N^\star(t) \in [\xi_-, \xi_+]$, and therefore

$$\left|\xi_N^\star(t) - \xi_{ij}^\star(t)\right| \leq \frac{\varepsilon}{m}, \tag{H.317}$$

which is Eq. (B.11).

This completes the proof. □

H.8. Proof of Prop. C.4

Proof. The exact parallel drift along $L_t^{(i,j)}$ is $F_{ij,t}(\xi) = \tilde{\gamma}_j^{(ij)}(\boldsymbol{y}(\xi)) - \xi$, where $\boldsymbol{y}(\xi) = \boldsymbol{\mu}^{(i)} + \xi\ell\boldsymbol{u}$ and $\boldsymbol{u} = (\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)})/\ell$. Under score error ψ, the perturbed PF ODE (C.13) induces perturbed parallel dynamics. Projecting onto $\boldsymbol{u}$, the drift becomes

$$\tilde{F}_{N,t}(\xi) = F_{ij,t}(\xi) + e_t(\xi), \quad \text{where} \quad e_t(\xi) := \frac{1}{\ell}\langle \psi(\boldsymbol{y}(\xi), t), \boldsymbol{u}\rangle. \tag{H.318}$$

By Cauchy-Schwarz and Asm. C.2,

$$|e_t(\xi)| \leq \frac{\|\psi(\boldsymbol{y}(\xi), t)\|_2 \cdot \|\boldsymbol{u}\|_2}{\ell} \leq \frac{\varrho(t)}{\ell}. \tag{H.319}$$

Define $G : [0, T] \times [0, 1] \to \mathbb{R}$ by $G(t, \xi) := \tilde{F}_{N,t}(\xi)$. At the unperturbed equilibrium with $\varrho = 0$, we have $G(t, \xi_{ij}^*(t))|_{\varrho=0} = F_{ij,t}(\xi_{ij}^*(t)) = 0$. The partial derivative satisfies

$$\left.\frac{\partial G}{\partial \xi}\right|_{\xi=\xi_{ij}^*(t)} = F'_{ij,t}(\xi_{ij}^*(t)) + e'_t(\xi_{ij}^*(t)) = \lambda_t + O(L_\psi(t)), \tag{H.320}$$

where $|e'_t(\xi)| = |\boldsymbol{u}^\top \nabla_{\boldsymbol{x}} \psi \, \boldsymbol{u}| \leq L_\psi(t)$ by Asm. C.3. For $L_\psi(t) \leq \lambda_t/2$, we have $\partial G/\partial \xi \geq \lambda_t/2 > 0$, so by the implicit function theorem there exists a unique $\xi_\theta^*(t)$ near $\xi_{ij}^*(t)$ with $G(t, \xi_\theta^*(t)) = 0$.

Taylor expanding $G(t,\xi)$ around $\xi^*_{ij}(t)$:

$$0 = G(t,\xi^*_\theta(t)) = \underbrace{F_{ij,t}(\xi^*_{ij}(t))}_{=0} + e_t(\xi^*_{ij}(t)) + (\lambda_t + O(L_\psi(t)))(\xi^*_\theta - \xi^*_{ij}) + O(|\xi^*_\theta - \xi^*_{ij}|^2). \tag{H.321}$$

Rearranging and using (H.319):

$$|\xi^*_\theta(t) - \xi^*_{ij}(t)| \le \frac{2|e_t(\xi^*_{ij}(t))|}{\lambda_t} + O(\varrho(t)^2) \le \frac{2\varrho(t)}{\ell\lambda_t} + O(\varrho(t)^2). \tag{H.322}$$

□

H.9. Proof of Prop. C.5

Proof. The perturbed parallel drift is $\tilde{F}_{N,t}(\xi) = F_{ij,t}(\xi) + e_t(\xi)$, with derivative

$$\tilde{F}'_{N,t}(\xi) = F'_{ij,t}(\xi) + e'_t(\xi). \tag{H.323}$$

Since $e_t(\xi) = \frac{1}{\ell}\langle \psi(\boldsymbol{y}(\xi),t), \boldsymbol{u}\rangle$ and $\frac{d\boldsymbol{y}}{d\xi} = \ell\boldsymbol{u}$,

$$e'_t(\xi) = \frac{1}{\ell}\langle \nabla_{\boldsymbol{x}}\psi(\boldsymbol{y}(\xi),t)\cdot(\ell\boldsymbol{u}), \boldsymbol{u}\rangle = \boldsymbol{u}^\top \nabla_{\boldsymbol{x}}\psi(\boldsymbol{y}(\xi),t)\,\boldsymbol{u}. \tag{H.324}$$

At the perturbed equilibrium $\xi^*_\theta(t)$:

$$\begin{aligned}
\lambda_\theta(t) := \tilde{F}'_{N,t}(\xi^*_\theta(t)) &= F'_{ij,t}(\xi^*_\theta(t)) + \boldsymbol{u}^\top \nabla_{\boldsymbol{x}}\psi(\boldsymbol{y}(\xi^*_\theta),t)\,\boldsymbol{u} && \text{(H.325)}\\
&= F'_{ij,t}(\xi^*_{ij}(t)) + O(|\xi^*_\theta - \xi^*_{ij}|) + \boldsymbol{u}^\top \nabla_{\boldsymbol{x}}\psi(\boldsymbol{y}(\xi^*_\theta),t)\,\boldsymbol{u} && \text{(H.326)}\\
&= \lambda_t + \boldsymbol{u}^\top \nabla_{\boldsymbol{x}}\psi(\boldsymbol{y}(\xi^*_\theta),t)\,\boldsymbol{u} + O(\varrho(t)), && \text{(H.327)}
\end{aligned}$$

where the second line uses Lipschitz continuity of $F'_{ij,t}(\xi)$: by the mean value theorem,

$$|F'_{ij,t}(\xi^*_\theta) - F'_{ij,t}(\xi^*_{ij})| \le \sup_{\xi\in[0,1]} |F''_{ij,t}(\xi)| \cdot |\xi^*_\theta - \xi^*_{ij}| = O(|\xi^*_\theta - \xi^*_{ij}|), \tag{H.328}$$

and the last line uses Prop. C.4. The two items follow from the sign of $\boldsymbol{u}^\top \nabla_{\boldsymbol{x}}\psi\,\boldsymbol{u}$. □

H.10. Proof of Prop. C.6

Proof. For $a \in [\vartheta, 2\vartheta]$, the exact drift satisfies $b(a,t) \ge \lambda_{\text{rep}} a \ge \lambda_{\text{rep}}\vartheta$. The perturbed drift satisfies

$$b_\theta(a,t) = b(a,t) + \tilde{e}_t(a) \ge \lambda_{\text{rep}} a - |\tilde{e}_t(a)| \ge \lambda_{\text{rep}} a - \frac{\varrho(t)}{\ell} \ge \lambda_{\text{rep}} a - \frac{\varrho_{\max}}{\ell}. \tag{H.329}$$

Under the condition $\varrho_{\max}/\ell < \lambda_{\text{rep}}\vartheta$, for $a \in [\vartheta, 2\vartheta]$:

$$b_\theta(a,t) \ge \left(\lambda_{\text{rep}} - \frac{\varrho_{\max}}{\ell a}\right) a \ge \left(\lambda_{\text{rep}} - \frac{\varrho_{\max}}{\ell\vartheta}\right) a = \tilde{\lambda}_{\text{rep}}\, a, \tag{H.330}$$

where $\tilde{\lambda}_{\text{rep}} > 0$ by assumption. The symmetric argument holds for $a \in [-2\vartheta, -\vartheta]$, so the escape mechanism of Prop. 5.1 applies with λ_{rep} replaced by $\tilde{\lambda}_{\text{rep}}$. □

H.11. Proof of Proposition D.1

Proof. Fix any $t \le \tau_1$ and write $\boldsymbol{p} := \boldsymbol{y}_t$ for the state at time t. We first prove the claim for a single rectangular cell.

Consider a single grid cell with corners (true modes) $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \boldsymbol{\mu}^{(3)}, \boldsymbol{\mu}^{(4)}$, where the diagonal pairs are $(1,4)$ and $(2,3)$. Define $d_a^2 := \|\boldsymbol{\mu}^{(a)} - \boldsymbol{p}\|_2^2$. For any rectangle, one has the identity

$$d_1^2 + d_4^2 = d_2^2 + d_3^2. \tag{H.331}$$

(For example, place the rectangle at $(0,0),(L,0),(0,H),(L,H)$ and expand the four squared distances to verify the equality for all $\boldsymbol{p}$.)

Assume the dominant pair (i,j) from Asm. 4.1 corresponds to the diagonal $(1,4)$ of this cell. Let

$$m := \min\{d_1^2, d_4^2\}, \qquad \delta := |d_1^2 - d_4^2|.$$

Then (wlog, say $d_1^2 = m$ and $d_4^2 = m+\delta$) we have

$$\frac{d_1^2 + d_4^2}{2} = m + \frac{\delta}{2}. \tag{H.332}$$

Also, using $\min\{a,b\} \le (a+b)/2$ and Eq. (H.331),

$$\min\{d_2^2, d_3^2\} \le \frac{d_2^2 + d_3^2}{2} = \frac{d_1^2 + d_4^2}{2} = m + \frac{\delta}{2}. \tag{H.333}$$

By Asm. 4.1 for the pair $(1,4)$, every non-dominant mode $k \notin \{1,4\}$ satisfies

$$d_k^2 \ge m + \Delta_t.$$

In particular,

$$d_2^2 \ge m + \Delta_t, \qquad d_3^2 \ge m + \Delta_t. \tag{H.334}$$

Combining Eq. (H.334) with Eq. (H.333) yields

$$m + \Delta_t \le m + \frac{\delta}{2} \quad \Longrightarrow \quad \Delta_t \le \frac{\delta}{2}. \tag{H.335}$$

Let $\tilde{\gamma}_{a,t}$ denote the (rescaled) responsibilities as in Lemma G.2; for brevity write $\gamma_a := \tilde{\gamma}_{a,t}$. By Eq. (H.241), for any a, b, we have that:

$$\frac{\gamma_a}{\gamma_b} = \frac{\pi_a}{\pi_b} \exp\left(\frac{d_b^2 - d_a^2}{2\tilde{\sigma}_t^2}\right).$$

Applying this with $(a,b) = (1,4)$ and taking logs gives

$$d_4^2 - d_1^2 = 2\tilde{\sigma}_t^2 \left(\log\frac{\gamma_1}{\gamma_4} - \log\frac{\pi_1}{\pi_4}\right),$$

hence

$$\delta = |d_4^2 - d_1^2| \le 2\tilde{\sigma}_t^2 \left(\left|\log\frac{\gamma_1}{\gamma_4}\right| + \left|\log\frac{\pi_1}{\pi_4}\right|\right). \tag{H.336}$$

Plug Eq. (H.336) into Eq. (H.335):

$$\Delta_t \le \tilde{\sigma}_t^2 \left(\left|\log\frac{\gamma_1}{\gamma_4}\right| + \left|\log\frac{\pi_1}{\pi_4}\right|\right). \tag{H.337}$$

By Asm. 4.1, $\Delta_t \ge 2\tilde{\sigma}_t^2\kappa$, so

$$2\kappa \le \left|\log\frac{\gamma_1}{\gamma_4}\right| + \left|\log\frac{\pi_1}{\pi_4}\right|.$$

Exponentiating and using $\exp(|\log x|) = \max\{x, 1/x\}$ yields

$$e^{2\kappa} \le \max\left\{\frac{\gamma_1}{\gamma_4}, \frac{\gamma_4}{\gamma_1}\right\} \cdot \max\left\{\frac{\pi_1}{\pi_4}, \frac{\pi_4}{\pi_1}\right\}. \tag{H.338}$$

Define $\rho_t := \min\{\gamma_1, \gamma_4\}$. Since $\gamma_a \in (0,1)$ and $\gamma_1 + \gamma_4 \le 1$, we have $\max\{\gamma_1, \gamma_4\} \le 1 - \rho_t$, hence

$$\max\left\{\frac{\gamma_1}{\gamma_4}, \frac{\gamma_4}{\gamma_1}\right\} = \frac{\max\{\gamma_1, \gamma_4\}}{\min\{\gamma_1, \gamma_4\}} \le \frac{1-\rho_t}{\rho_t}. \tag{H.339}$$

Let $C_\pi := \max\{\pi_1/\pi_4, \pi_4/\pi_1\}$. Combining Eqs. (H.338) and (H.339) gives

$$e^{2\kappa} \le C_\pi \frac{1-\rho_t}{\rho_t} \quad \Longrightarrow \quad \rho_t \le \frac{C_\pi}{C_\pi + e^{2\kappa}} \le C_\pi e^{-2\kappa}.$$

Since $\rho_t = \min\{\gamma_1, \gamma_4\}$ and $\gamma_a = \tilde{\gamma}_{a,t}$, we have shown

$$\min\{\tilde{\gamma}_{1,t}, \tilde{\gamma}_{4,t}\} \le C\pi e^{-2\kappa} \qquad \forall\, t \le \tau_1.$$

This proves the desired $\mathcal{O}(e^{-2\kappa})$ bound for a single diagonal pair.

Next, we extend this to the rest of the cell. If (i, j) is a diagonal pair of some grid cell, let (k, ℓ) be the other two corners of that cell. Then the rectangle identity Eq. (H.331) holds for the quadruple (i, k, ℓ, j) for the same point $\boldsymbol{p} = \boldsymbol{y}_t$, and Asm. 4.1 for (i, j) implies $d_k^2, d_\ell^2 \ge m + \Delta_t$ just as in Eq. (H.334). Repeating the above argument with the relabeled indices gives $\min\{\tilde{\gamma}_{i,t}, \tilde{\gamma}_{j,t}\} \le C_{\pi,ij} e^{-2\kappa}$ where $C_{\pi,ij} := \max\{\pi_i/\pi_j, \pi_j/\pi_i\}$. Lastly, since the cell chosen was arbitrary, this holds for all (i,j) in the grid, and we are done. □

I. Use of the Exact Score

In what follows, we detail why we use the exact score in our theoretical analysis and how this connects to the realistic learned score setting which we aim to study. Since the marginals under the exact score agree for the ODE and SDE as discussed in Sec. 2, the hallucination rate for the ODE and SDE agree under the exact score (in particular, it is 0). However, we can still leverage the ODE/SDE under the exact score to obtain a *mechanism* which becomes useful and predictive in the learned score setting. In particular, assuming the exact score as regularity, we show that $\boldsymbol{x}_t$ under the ODE can get stuck (Prop. 4.7) while $\boldsymbol{x}_t$ under the SDE can escape (Prop. 5.1). We study the exact score scenario in our main paper because it provides the cleanest presentation of this vulnerability. Our experiments then demonstrate that the trapping region in Prop. 4.7 characterizes hallucination differences between DDIM and DDPM in the realistic learned score scenario (e.g., in Figure 5).

To make this mathematically precise, we have $p_t^{\text{DDIM,exact}} = p_t^{\text{DDPM,exact}}$. Let $H := \{\text{mode interpolation occurs}\}$. Then, we have $\Pr^{\text{DDIM,exact}}(H) = \Pr^{\text{DDPM,exact}}(H)$. However, we are interested in explaining the observed differences in hallucination rates in the realistic, practical learned score setting where $\Pr^{\text{DDIM,learned}}(H) \gg \Pr^{\text{DDPM,learned}}(H)$ (Figure 1). To bridge this gap, we must look beyond unconditional marginals p_t and examine the *conditional trajectory dynamics* $p(\boldsymbol{x}_0 \mid \boldsymbol{x}_{\tau_3})$, which differ entirely for DDIM and DDPM. We find that we can extract a difference in DDIM (Prop. 4.7) and DDPM (Prop. 5.1) under the exact score which uncovers the trapping mechanism driving $\Pr^{\text{DDIM,learned}}(H) \gg \Pr^{\text{DDPM,learned}}(H)$. Concretely, let $M_{\vartheta,\tau_3} := \{\text{trajectory enters } \vartheta\text{-midpoint neighborhood during final } \tau_3 \text{ steps}\}$. Then:

$$\Pr(H) = \Pr(H \mid M_{\vartheta,\tau_3}) \Pr(M_{\vartheta,\tau_3}) + \Pr(H \mid M^c_{\vartheta,\tau_3}) \Pr(M^c_{\vartheta,\tau_3}) \tag{I.340}$$

The contributions of our exact score theory in Prop. 4.7 and Prop. 5.1 are twofold:

1. The midpoint neighborhood and M_{ϑ,τ_3} are driving the differences in hallucination rates between DDIM and DDPM, i.e. $\Pr(H \mid M^c_{\vartheta,\tau_3})$ is small and the first term dominates so that $\Pr(H) \approx \Pr(H \mid M_{\vartheta,\tau_3}) \Pr(M_{\vartheta,\tau_3})$. This is done in Prop. 4.7.

2. Demonstrate that $\Pr^{\text{DDIM,exact}}(H \mid M_{\vartheta,\tau_3}) \gg \Pr^{\text{DDPM,exact}}(H \mid M_{\vartheta,\tau_3})$. This is done in Prop. 5.1.

These results arise due to the differences in (conditional) dynamics, even though the marginals are the same under the exact score. The exact score assumption in our theory allows us to demonstrate this cleanly. Still, $\Pr^{\text{DDIM,exact}}(M_{\vartheta,\tau_3}) = \Pr^{\text{DDPM,exact}}(M_{\vartheta,\tau_3}) = 0$.

But, in the practical learned score setting where $\Pr^{\text{learned}}(M_{\vartheta,\tau_3}) \neq 0$ for both samplers, the fundamental difference $\Pr^{\text{DDIM,learned}}(H \mid M_{\vartheta,\tau_3}) \gg \Pr^{\text{DDPM,learned}}(H \mid M_{\vartheta,\tau_3})$, which persists in the learned score case but is uncovered via our exact score analysis, helps explain $\Pr^{\text{DDIM,learned}}(H) \gg \Pr^{\text{DDPM,learned}}(H)$. Empirically, in Tab. I.1, we find that the first term in Eq. (I.340) remains dominant while $\Pr^{\text{DDIM,learned}}(H \mid M_{\vartheta,\tau_3}) \gg \Pr^{\text{DDPM,learned}}(H \mid M_{\vartheta,\tau_3})$.

sampler	$P(H)$	$P(H \mid M)$	$P(H \mid M^c)$	$P(M)$	$P(H \mid M)P(M)$	$P(H \mid M^c)P(M^c)$
DDIM	0.078	0.735	0.006	0.099	0.073	0.005
DDPM	0.005	0.103	0.001	0.042	0.004	0.001

Table I.1. Hallucination probabilities in the learned score case, averaged across 100k trajectories, with $\tau_3 = 11$ and $\vartheta = 0.35\ell_t$. Here, H denotes the event that the final generated sample is a mode interpolation, and $M := M_{\vartheta,\tau_3}$ denotes the event that the trajectory enters the ϑ-midpoint neighborhood of the nearest i, j-mode segment during the final τ_3 steps. We observe that $\Pr^{\mathrm{DDIM}}(H \mid M) \gg \Pr^{\mathrm{DDPM}}(H \mid M)$, while $\Pr(H \mid M^c)$ remains small for both samplers. Thus, the dominant contribution comes from the first decomposition term $\Pr(H \mid M)\Pr(M)$, supporting the conclusion that midpoint-neighborhood entry is the key event relevant for hallucinations in the learned score case.

J. Symbol Table

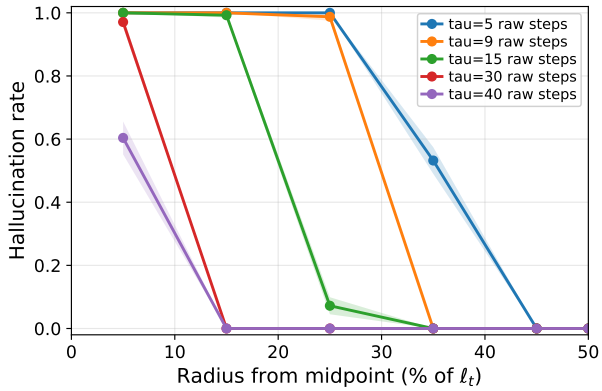
Symbols	
ϖ	Dimension of all vectors throughout.
p_{base}	The distribution which the diffusion sampling process starts from. Throughout, we take this to be $\mathcal{N}(0, \boldsymbol{I})$, the standard Gaussian.
p_{data}	The distribution which the diffusion process ends at and aims to sample from. Throughout, we take this to be a N-component Gaussian mixture.
$\boldsymbol{x}_t$	A trajectory satisfying the PF ODE given in Eq. (3) for DDIM or Eq. (2) for DDPM, given in reverse time from $t = T$ to $t = 0$ unless otherwise specified.
$\beta(t)$	Diffusion noise schedule.
$\boldsymbol{W}_t$	Standard ϖ-dimensional Brownian motion
$\bar{\alpha}_t$	Amount of signal from data distribution retained in forward or reverse process. Given by $\exp(-\int_0^t \beta(s)\, ds)$
p_t	Marginal distribution of diffusion process instance $\boldsymbol{x}_t$ at time t; equivalent in forward and reverse process.
$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$	Exact score function of marginal p_t.
$\boldsymbol{I}$	Identity matrix.
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$	Probability density function of Gaussian distribution centered at $\boldsymbol{\mu}$ with isotropic covariance matrix $\sigma^2 \boldsymbol{I}$.
$\bar{\boldsymbol{W}}_t$	Standard Brownian motion in reverse time.
$s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$	Learned score function.
π_i	Coefficient of Gaussian mixture.
N	Number of components in Gaussian mixture.
$\boldsymbol{\mu}^{(i)}$	ith mode of Gaussian mixture, $i \in [N]$.
σ^2	Variance of Gaussian mixture, i.e. each component has covariance $\sigma^2 \boldsymbol{I}$.
$\boldsymbol{\mu}_t^{(i)}$	Time-dependent modes of p_t given Gaussian mixture target.
σ_t^2	Time-dependent variance of p_t given Gaussian target.
$\tilde{\sigma}_t^2$	Rescaled time-dependent variance of p_t.
$L^{(i,j)}$	Line segment joining mode i and mode j.
$L_t^{(i,j)}$	Line segment joining time-dependent mode i and time-dependent mode j.
$L_{t,\varepsilon}^{(i,j)}$	ε-extended line segment joining time-dependent mode i and time-dependent mode j.
$d_\perp$	Orthogonal distance between trajectory and (nearby) line segment.
$\text{Tube}_{t,\zeta}^{(i,j)}$	Tube around $L_t^{(i,j)}$ of radius ζ. Throughout, $\zeta = \varepsilon$, unless specified otherwise.

Symbols

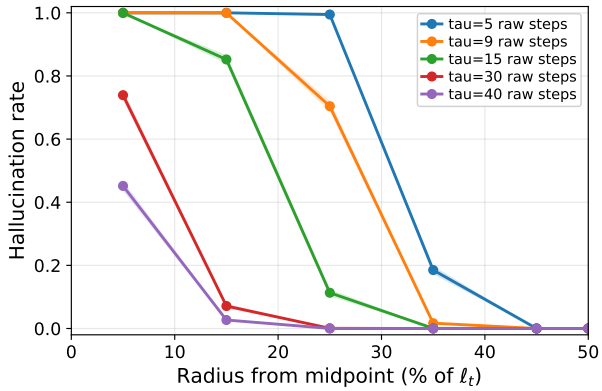
Δ_t	Time-dependent margin between modes. Must be sufficiently large after time in reverse process w.r.t. $\tilde{\sigma}_t^2$.
κ	Margin coefficient, defining gap between $\tilde{\sigma}_t^2$ and Δ_t. Also used for margin w.r.t. ℓ^2.
τ_1	Critical time for Asm. 4.1 to hold.
τ	Time which is less than τ_1; critical time for exponential convergence in Theorem 4.2 to hold.
τ_2	Critical time for Asm. 4.4 to hold. Empirically strictly less than τ_1.
$\hat{\tau}$	Minimum of τ and τ_2. Critical time for Prop. 4.5 to hold.
τ_3	Critical time less than $\hat{\tau}$ for Prop. 4.7 to hold.
ϕ	Used to define probabilities in concentration inequalities throughout.
ε	Size by which nearby line segment is extended; radius of containment tube; and radius of ball around modes of mixture where stable equilibria exist.
ξ_t	Used to denote parallel dynamics to $L_t^{(i,j)}$. ξ^* is an equilibrium of some kind, stated specifically in each proposition in which it appears.
$\boldsymbol{w}_t$	Used to denote orthogonal dynamics to $L_t^{(i,j)}$.
$\boldsymbol{y}_t^*$	Unstable equilibria at midpoint along $L_t^{(i,j)}$ (w.r.t. approximate dynamics)
$\boldsymbol{y}_t$	Used for time-rescaled dynamics in Lemma G.2, where modes become fixed and effective variance σ_t^2 is rescaled to $\tilde{\sigma}_t^2$.
λ_t	Unstable eigenvalue on midpoint, or unstable equilibria on $L_t^{(i,j)}$ more broadly.
ϑ	Radius from midpoint for which, given τ_3 time remaining.
$\tilde{\lambda}_t$	Integral of λ_t from time t_{ν_1} to time t_{ν_2}.
ℓ^2	Squared distance between (fixed, time-independent) modes.
$H^{(i,j)}$	Bisector hyperplane to $L^{(i,j)}$, i.e. the plane orthogonal to the line which passes through the midpoint of the line.
$\boldsymbol{u}$	The coordinate normal to $H^{(i,j)}$.
A_t	One dimensional dynamics of distance from midpoint along $\boldsymbol{u}$, with respect to DDPM.
$b(A_t, t)$	Deterministic dynamics (drift) of A_t.
π	Without a subscript, defines the usual scalar π.
$K(t)$	Controls how fast $b(A_t, t)$ is.
$\eta(t)$	$1/2\beta(t)$.
δ	In the main paper, used to define the terminal confinement events. Used to define residual terms due to $N-2$ components in appendix proofs.

Symbols

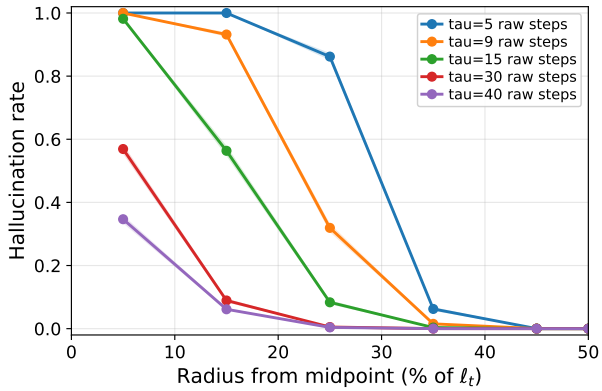
H_δ	Event that at the end of the reverse process, trajectory stays near the bisector hyperplane.
$E_{2\delta}$	Event that at throughout the reverse process, one enters the midpoint neighborhood.
λ_{rep}	Coefficient which controls speed of drift away from bisector hyperplane.
γ_k	Gaussian responsibilities under Eq. (3).
$\tilde{\gamma}_k$	Gaussian responsibilities under time-rescaled dynamics in Lemma G.2.
$\hat{\boldsymbol{\mu}}^{(i)}$	Convex combination of static modes w.r.t. time-rescaped Gaussian responsibilities.
$I(t)$	Integral of $\beta(s)$ without exponential
δ_ξ	Effect of $N-2$ modes on parallel dynamics to nearby $L_t^{(i,j)}$.
$\boldsymbol{E}_t$	Convex combination of distances from mode i, given time-rescaled modes, excluding modes i and j.
$\text{Proj}_{\|\|}(\boldsymbol{v})$	Projection onto vector parallel to $\boldsymbol{v}$.
$\text{Proj}_{\perp}(\boldsymbol{v})$	Projection onto vector orthogonal to $\boldsymbol{v}$.
$\boldsymbol{\delta_w}$	Effect of $N-2$ modes on orthogonal dynamics to nearby $L_t^{(i,j)}$.
d^*	Mode which for which the trajectory has minimum distance from one of the two selected (nearby) modes i and j.
$\mathcal{O}$, Ω, Θ	Used for standard asymptotic notation.
M	Defines maximum non-selected mode distance.
R	Defines maximum Gaussian mixture weights w.r.t. d^*, excluding π_i, π_j.
θ	Used to rescale reverse time into forward time, i.e. $\theta = T - t$.
M, Q	For martingale M, Q is its quadratic variation.



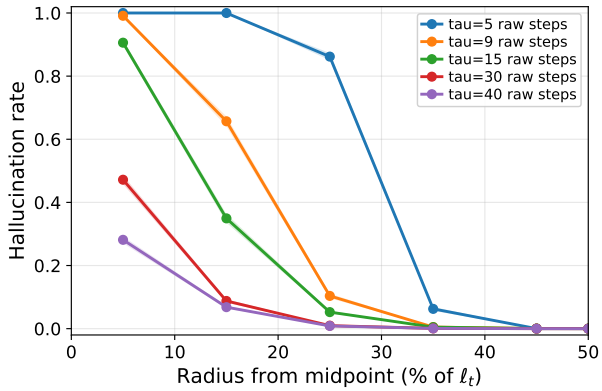
(a) DDIM



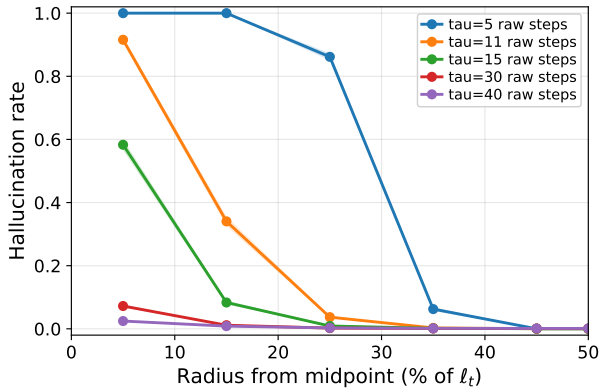
(b) DDIM + $z = 2$ DDPM



(c) DDIM + $z = 5$ DDPM



(d) DDIM + $z = 8$ DDPM



(e) DDPM

Figure E.11. Ablation on τ_3 across various DDIM-DDPM hybrid samplers, starting from the midpoint neighborhood and evaluating hallucination rate. Uncertainty is quantified by the standard error across mode pairs.

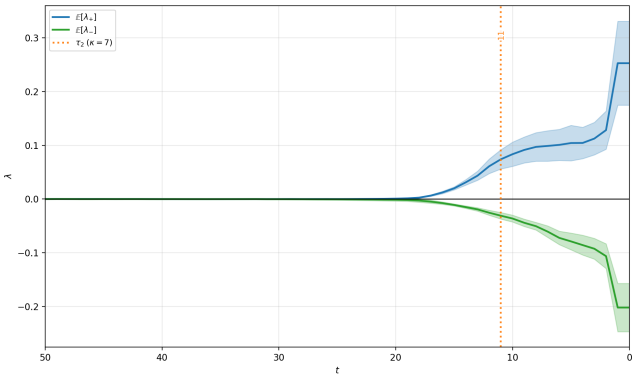
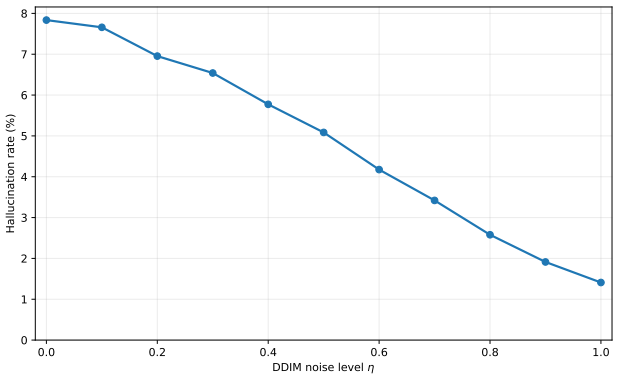
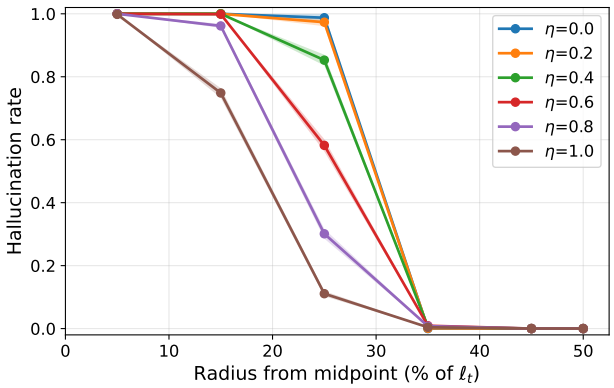


Figure E.12. Average DDIM Jacobian eigenvalues evaluated at the midpoint joining mode segments, λ_t, grouped into positive (blue) and negative (green) groups, during the reverse process. A vertical dotted orange line represents τ_2. We find that there exists both a positive and corresponding negative eigenvalue after time τ_2. Notably, there were no eigenvalues excluded from this plot; all midpoints had one positive eigenvalue and one negative eigenvalue. In particular, the midpoint remains a hyperbolic saddle equilibria with appropriate eigenvalues. This indicates that despite score error and our study of the approximate dynamics, the slowdown region does not greatly shift for a well-trained diffusion model. Finally, note that the Jacobian eigenvalues are identical (up to numerical error) for DDIM and DDPM, since Eq. (3) and Eq. (2) have the same Jacobian.

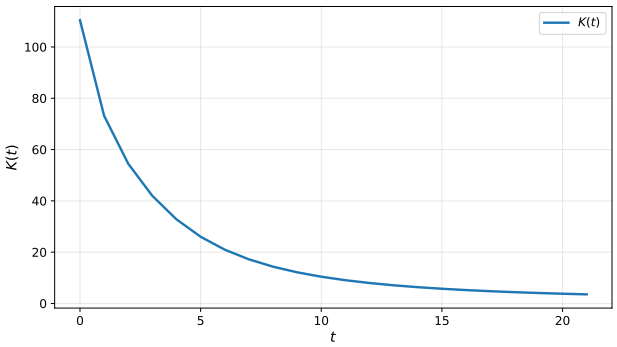


(a) Change in hallucination rate while varying DDIM noise level η

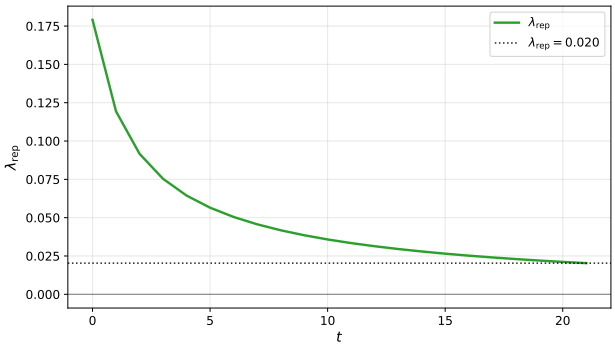


(b) Change in midpoint neighborhood escape while varying DDIM noise level η

Figure E.13. Increasing DDIM noise level η decreases mode interpolation hallucination rate (Figure E.13a) while helping DDIM trajectories escape the midpoint trapping region (Figure E.13b), demonstrating that SDE noise helps with reducing hallucinations, in line with our theoretical results.



(a) $K(t)$ in Prop. 5.1



(b) λ_{rep} in Prop. 5.1

Figure E.14. $K(t)$ (Figure E.14a) and λ_{rep} (Figure E.14b) computed across a total of 400k samples (10k samples per line segment) under the exact score. We find that $K(t)$ exists and $\lambda_{\text{rep}} > 0$ as desired. Details on how to compute these are included in Sec. F.

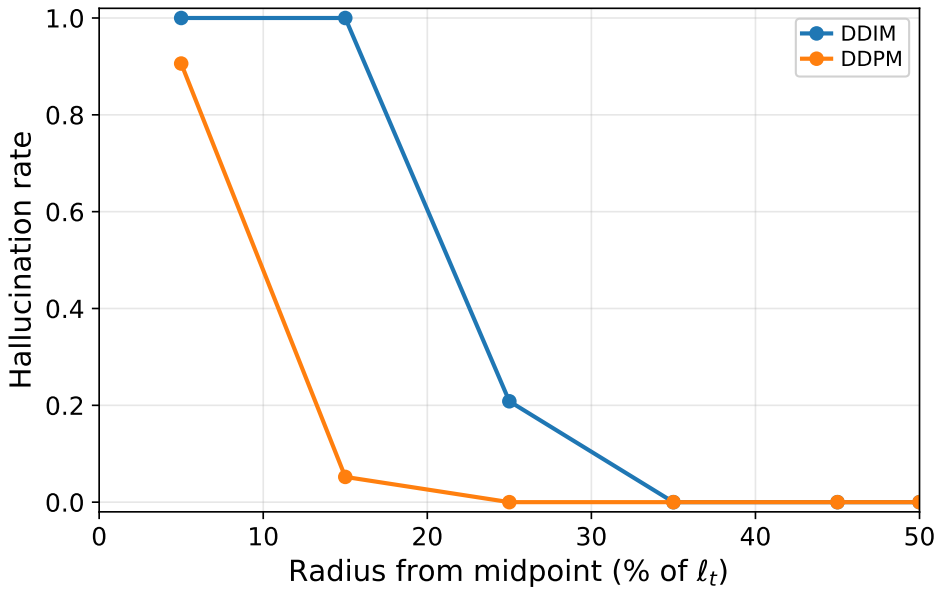


Figure E.15. Exact score analogue of Figure 5. We find that at $\vartheta = 0.15\ell_t$ and $\tau_3 = 3$, DDIM gets stuck in the midpoint neighborhood while DDPM escapes, confirming our theoretical results in Prop. 4.7 and Prop. 5.1.

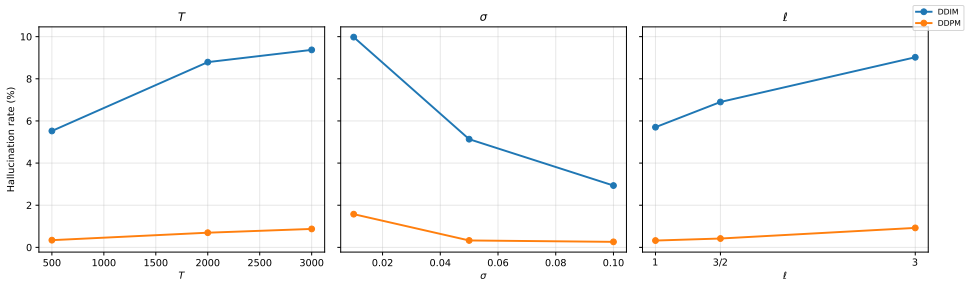


Figure E.16. Ablations on hallucination rate gap between DDIM and DDPM across various choices of σ, T, and ℓ. We find that the hallucination rate gap between DDIM and DDPM persists across several different hyperparameter configurations.

Figure F.17. Samples chosen from the dataset of $10,000$ elements containing 3 triangles. The positions are chosen such that the triangles are non-overlapping.

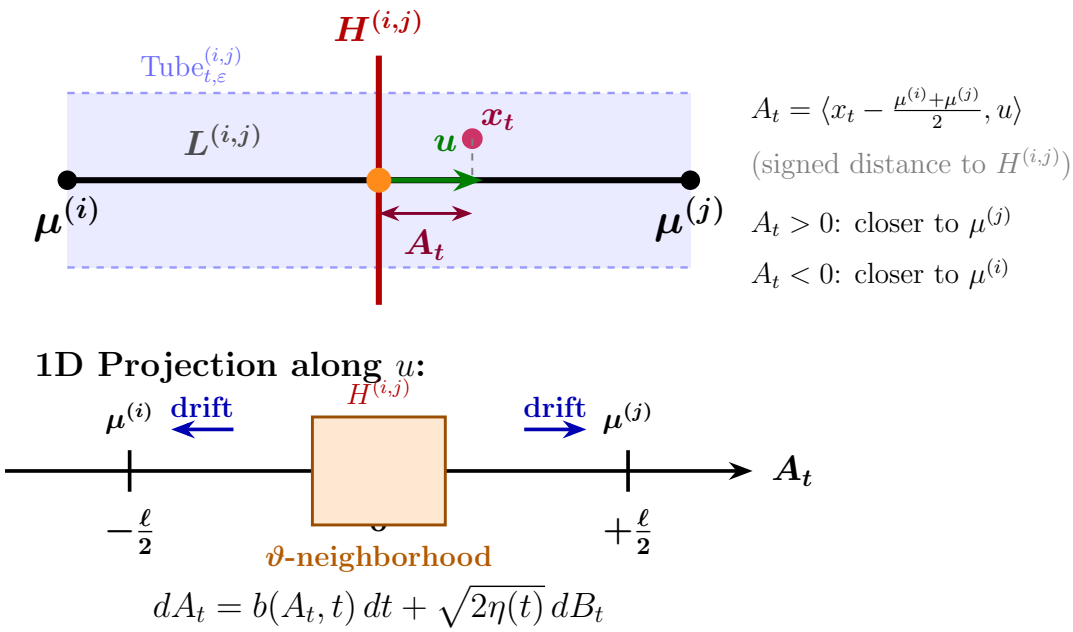


Figure F.18. Reduction to 1D dynamics along the bisector normal. **Top**: Inside $\mathrm{Tube}_{t,\varepsilon}^{(i,j)}$, the bisector hyperplane $H^{(i,j)}$ (red) is orthogonal to $L^{(i,j)}$. The signed 1D coordinate $A_t = \langle x_t - \frac{\mu^{(i)}+\mu^{(j)}}{2}, u\rangle$ measures displacement from $H^{(i,j)}$ along the unit normal u. **Bottom**: The 1D reduction with modes at $\pm\frac{\ell}{2}$, the ϑ-neighborhood around the midpoint, and unstable drift directions away from the bisector.