Critical windows: non-asymptotic theory for feature emergence in diffusion models

Marvin Li [1] **Sitan Chen** [2]

Abstract

We develop theory to understand an intriguing property of diffusion models for image generation that we term *critical windows*. Empirically, it has been observed that there are narrow time intervals in sampling during which particular features of the final image emerge, e.g. the image class or background color (Ho et al., 2020b; Meng et al., 2022; Choi et al., 2022; Raya & Ambrogioni, 2023; Georgiev et al., 2023; Sclocchi et al., 2024; Biroli et al., 2024). While this is advantageous for interpretability as it implies one can localize properties of the generation to a small segment of the trajectory, it seems at odds with the continuous nature of the diffusion. We propose a formal framework for studying these windows and show that for data coming from a mixture of strongly log-concave densities, these windows can be provably bounded in terms of certain measures of inter- and intra-group separation. We also instantiate these bounds for concrete examples like well-conditioned Gaussian mixtures. Finally, we use our bounds to give a rigorous interpretation of diffusion models as hierarchical samplers that progressively "decide" output features over a discrete sequence of times. We validate our bounds with experiments on synthetic data and show that critical windows may serve as a useful tool for diagnosing fairness and privacy violations in real-world diffusion models.

1. Introduction

Diffusion models currently stand as the predominant approach to generative modeling in audio and image domains (Sohl-Dickstein et al., 2015; Dhariwal & Nichol, 2021; Song et al., 2020; Ho et al., 2020b). At their core is a "forward process" that transforms data into noise, and a learned "reverse process" that progressively undoes this noise, thus generating fresh samples. Recently, a series of works has established rigorous convergence guarantees for diffusion models for arbitrary data distributions (Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2023a). While these results prove that in some sense diffusion models are entirely principled, the generality with which they apply suggests further theory is needed to explain the rich behaviors of diffusion models specific to the *real-world* distributions on which they are trained.

In this work, we focus on a phenomenon that we term *critical windows*. In the context of image generation, it has been observed that there are narrow time intervals along the reverse process during which certain features of the final image are determined, e.g. the class, color, background (Ho et al., 2020b; Meng et al., 2022; Choi et al., 2022; Raya & Ambrogioni, 2023; Georgiev et al., 2023; Sclocchi et al., 2024; Biroli et al., 2024). This suggests that even though the reverse process operates in continuous time, there is a series of discrete "jumps" during the sampling process during which the model "decides" on certain aspects of the output. The existence of these critical windows is highly convenient from an interpretability standpoint, as it lets one zoom in on specific parts of the diffusion model trajectory to understand how some feature of the generated output emerged.

Despite the strong empirical evidence for the existence of critical windows (e.g. the striking Figures 3, B.6, and B.10 from Georgiev et al. (2023) and Figures 1 and 2 from (Sclocchi et al., 2024)), our mathematical understanding of critical windows is very immature. Indeed, from the perspective of prior theory,[1] the different times of the reverse process largely behave as equal-class citizens, outside the realm of very simple toy models of data. We thus ask:

> *Can we **prove** the existence of critical windows in the reverse process for a rich family of data distributions?*

Before stating our theoretical findings, we outline the framework we adopt (see Section 3.2 for a formal treatment). Also,

[1]Harvard College, Cambridge, MA, USA [2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. Correspondence to: Marvin Li <marvinli@college.harvard.edu>, Sitan Chen <sitan@seas.harvard.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]See Section 2 for a discussion of concurrent works.

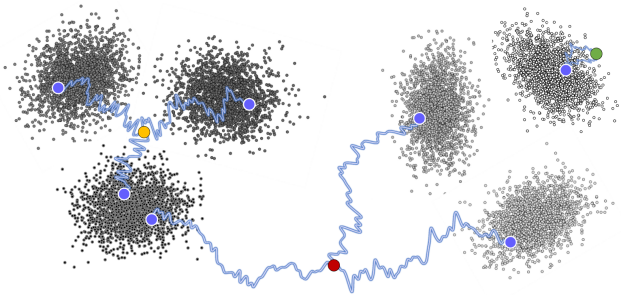

Figure 1. Cartoon depiction of running forward process for time t (to produce one of the red/yellow/green dots corresponding to large/medium/small t) and then running reverse process (trajectories in blue) for time t to sample from some sub-mixture.

as issues of discretization, score error, and the support of the data distribution lying on a lower-dimensional submanifold are orthogonal to this paper, throughout we will conflate the data distribution with the output distribution of the model and assume the reverse process is run in continuous time with perfect score.

1.1. General framework

Our starting point is a setup related to one in Georgiev et al. (2023) and also explored in the concurrent work of (Sclocchi et al., 2024) – see Section 2 for a comparison to these two works. Given a sample x from the data distribution p, consider the following experiment. We run the forward process (see Eq. (1) below) starting from x for intermediate amount of time t to produce a noisy sample x_t. We then run the reverse process (see Eq. (2)) for time t starting from x_t to produce a new sample x' (see Section 3 for formal definitions). Observe that as $t \to \infty$, the distribution over x_t converges to Gaussian, and thus the resulting distribution over x' converges to p. As $t \to 0$, the distribution over x' converges to a point mass at x – in this latter regime, it was empirically observed by (Ho et al., 2020b) that for small t, the distribution over x' is essentially given by randomly modifying low-level features of x.

Critical windows for mixture models. Qualitatively, we can ask for the first, i.e. largest, time t for which samples from the distribution over x' mostly share a certain feature with x. To model this, we consider a data distribution p given by a mixture of sub-populations $p^1, \dots, p^K$. A natural way to quantify whether x' shares a feature with x is then to ask whether the distribution over x' is close to a particular *sub-mixture*. E.g., if x is a cat image, and $S_{\text{targ}} \subset \{1, \dots, K\}$ denotes the sub-populations p^i corresponding to cat images, then one can ask whether there is a critical window of t such that the distribution over x' is close to the sub-mixture given by S_{targ} (see Figure 1).

Finally, rather than reason about specific initial samples from p, we will instead marginalize out the randomness of x so we can reason at a more "population" level. Concretely, we consider x which is drawn from some p^i for $i \in S_{\text{targ}}$, or more generally from some sub-mixture indexed by a subset $S_{\text{init}} \subset S_{\text{targ}}$, and consider the resulting marginal distribution over x', which we denote by $p[S_{\text{init}}^{\langle t \rangle}]$. So if for instance S_{init} denoted the sub-mixture of *brown* cats, then if there is a critical window of times for which $p[S_{\text{init}}^{\langle t \rangle}]$ is close to the sub-mixture given by S_{targ}, then one can interpret these times as the point at which, to generate a brown cat, the diffusion model "decides" its sample will be a cat.

1.2. Our contributions

Our results are threefold: (1) we give a general characterization of the critical window for a rich family of multimodal distributions, (2) we specialize these bounds to specific distribution classes to get closed-form predictions, (3) we use these to prove, under a distributional assumption, that the reverse process is a "hierarchical sampler" that makes a series of discrete feature choices to generate the output.

General characterization of critical window. We consider distributions p which are *mixtures of strongly log-concave distributions* in $\mathbb{R}^d$. In Section 4, we give general bounds on the the critical window at which $p[S_{\text{init}}^{\langle t \rangle}]$ approximates the sub-mixture given by S_{targ} for any choice of S_{init}. These bounds depend on the total variation (TV) distance between sub-populations inside and outside S_{init} and S_{targ} along the forward process. We identify two endpoints (see Eqs. (4) and (5) for formal definitions):

- T_{lower}: the time in the forward process at which the initial sub-mixture indexed by S_{init} and the target sub-mixture indexed by S_{targ} first become close in TV
- T_{upper}: the time in the forward process at which a component in S_{targ} begins to exhibit non-negligible overlap with a component in the rest of the mixture[2]

Theorem 1.1 (Informal, see Theorem 4.5)**.** *Suppose p is a mixture of strongly log-concavedistributions, and let $S_{\text{init}} \subset S_{\text{targ}}$. For any $t \in [T_{\text{lower}}, T_{\text{upper}}]$, if one runs the forward process for time t starting from the sub-mixture given by S_{init}, then runs the reverse process for time t, the result will be close in TV to the sub-mixture given by S_{targ}.*

As we show empirically on synthetic examples (Fig. 3) these bounds can be highly predictive of the true critical windows.

The intuition for this result is that there are two competing effects at work. On the one hand, if t is sufficiently large, then running the forward process for time t starting from either the initial sub-mixture given by S_{init} versus the

[2] *A priori* T_{lower} need not be smaller than T_{upper} In Section 5, we show this holds when S_{targ} corresponds to a "salient" feature.

target sub-mixture given by S_{targ} will give rise to similar distributions. So if we run the reverse process on these, the resulting distributions will remain close, thus motivating our definition of T_{lower}. On the other hand, we want these resulting distributions to be close to the target sub-mixture indexed by S_{targ}. But if t is too large, they will merely be close to p. To avoid this, we need t to be small enough that along the reverse process, the overall score function of p to remains close to the score function of the target sub-mixture. Intuitively, this should happen provided the components of p outside S_{targ} do not overlap much with the ones inside S_{targ} even after running the forward process for time t, thus motivating our definition of T_{upper}.

It may at first glance seem extremely strong that we require that each sub-population form a strongly log-concave component. e.g. in the latent space over which the diffusion operates. We clarify that in our setting, a sub-population might correspond to a *sub-mixture* consisting of multiple such strongly log-concave components. In the example of natural images, we can think of a particular image class as a sub-mixture consisting of neighborhoods around different images in the embedding space.

Concrete estimates for critical times. The endpoints of the critical window in Theorem 1.1 are somewhat abstract. Our second contribution is to provide concrete bounds for these– see Section 5.2 for details. We first consider the general setting of Theorem 1.1 where p is a mixture of strongly log-concave distributions, under the additional assumption that the components would be somewhat close in Wasserstein distance if they were shifted to all have mean zero.

Theorem 1.2 (Informal, see Theorem 5.1)**.** *Suppose p is a mixture of $1/\sigma^2$-strongly log-concave distributions with means $\mu_1, \dots, \mu_K$, and let $S_{\text{init}} \subset S_{\text{targ}}$.*

Suppose that for any $i \in S_{\text{targ}}$ and any $j \notin S_{\text{targ}}$, $\|\mu_i - \mu_j\| \gtrsim \sigma\sqrt{d}$. Then there is an upper bound for T_{lower}, which is dominated by $\ln \max_{i \in S_{\text{init}}, j \in S_{\text{targ}}} \|\mu_i - \mu_j\|$, and there is a lower bound for T_{upper}, which is dominated by $\ln \min_{i \in S_{\text{targ}}, j \notin S_{\text{targ}}} \|\mu_i - \mu_j\|$.

Theorem 1.2 shows that the start time of the critical window scales as the log of the max distance between any component in S_{init} and any component in S_{targ}, whereas the end time scales as the log of the min distance between any component in S_{targ} and any component in $[K] \backslash S_{\text{targ}}$. We can interpret this as saying the following about feature emergence. If S_{targ} corresponds to the part of the data distribution with some particular feature, if that feature is sufficiently *salient* in the sense that typical images with that feature are closer to each other, then $T_{\text{lower}} < T_{\text{upper}}$, and therefore there exists a critical window of times t during which the feature associated to S_{targ} emerges. For latent diffusion models in particular, the manifold of images in latent space becomes highly structured, so there will be "salient" features such that images with the same "salient" feature will be closer together in the latent space. Furthermore, the length of this window, i.e. the amount of time after the features associated to S_{targ} emerge but before other features do, is logarithmic in the ratio between the level of separation between S_{targ} and $[K] \backslash S_{\text{targ}}$, versus the level of separation within S_{targ}. This requirement of salience is also weakened for important applications of this theory, like interpretability. [3]

In Appendix C.2, we specialize the bound in Theorem 1.2 to a sparse coding setting where the means of the components are given by sparse linear combinations of a collection of incoherent "dictionary vectors." In this setting, we show that the endpoints T_{lower} (resp. T_{upper}) have a natural interpretation in terms of the *Hamming distances* between the sparse linear combinations defining the means within S_{init} and S_{targ} (resp. between S_{targ} and $[K] \backslash S_{\text{targ}}$).

Theorem 1.2 is quite general except for one caveat: we must assume that the the components outside of S_{targ} have some level of separation. Note that a $1/\sigma^2$-strongly log-concave distribution in d dimensions will mostly be supported on a thin shell of radius $\sigma\sqrt{d}$ (Kannan et al., 1995), so our assumption essentially amounts to ensuring the balls that these shells enclose, for any component inside S_{targ} and any component outside S_{targ}, do not intersect.

Next, we remove this caveat for mixtures of Gaussians:

Theorem 1.3 (Informal, see Theorem 5.3)**.** *Suppose p is a mixture of K identity-covariance Gaussians in $\mathbb{R}^d$ with means $\mu_1, \dots, \mu_K$, and let $S_{\text{init}} \subset S_{\text{targ}}$. Then there is an upper bound for T_{lower}, which is dominated by $\ln \max_{i \in S_{\text{init}}, j \in S_{\text{targ}}} \|\mu_i - \mu_j\|$, and there is a lower bound for T_{upper}, which is dominated by $\ln \min_{i \in S_{\text{targ}}, j \notin S_{\text{targ}}} \|\mu_i - \mu_j\|$.*

In fact our result extends to general mixtures of Gaussians with sufficiently well-conditioned covariances, see Theorem 5.3. We also explore the dependence on the mixing weights of the different components (see Appendix C.1).

Hierarchical sampling interpretation. Thus far we have focused on a specific target sub-mixture S_{targ}, which would correspond to a specific feature in the generated output. In Section 6, we extend these findings to distributions with a *hierarchy* of features. To model this, we consider Gaussian mixtures with hierarchical clustering structure. This structure ensures the mixture decomposes into well-separated clusters of components such that the separation between clusters exceeds the separation within clusters, and further-

[3] If the target mixture is the same as the initial mixture, $T_{\text{lower}} = 0$ and we only need $T_{\text{upper}} > 0$ to form a critical window. This setting is especially useful for interpretability and data attribution, which usually examines an object x with property p and asks for the largest time for which property p is preserved.

more each cluster recursively satisfies the property of being decomposable into well-separated clusters, etc. This naturally defines a tree, which we call a *mixture tree*, where each node of the tree corresponds to a cluster at some resolution, with the root corresponding to the entire data distribution and the leaves corresponding to the K individual components of the mixture (see Definition 6.1).

If we think of every node v as being associated with a feature, then the corresponding cluster of components is comprised of all sub-populations which possess that feature, in addition to all features associated to nodes on the path from the root to v. By chaining together several applications of Theorem 1.3, we prove the following:

Theorem 1.4 (Informal, see Theorem 6.2)**.** *For a hierarchical mixture of identity-covariance Gaussians with means specified by a mixture tree, for any root-to-leaf path $(v_0, \dots, v_L)$ in the mixture tree, where the leaf v_L corresponds to a component p^i of the mixture, there exists an $\underline{L}$ and a discrete sequence of times $t_{v_{\underline{L}}} > \dots > t_{v_L}$ such that for all $\underline{L} \le \ell \le L$, the distribution if one runs the forward process for time t_{v_ℓ} starting from the sub-mixture given by the node v_L and the reverse process for time t_{v_ℓ}, the result will be close in TV to the sub-mixture given by node v_ℓ.*

This formalizes the intuition that to sample from distributions with this hierarchical structure, the sampler makes a discrete sequence of choices on the features to include. This discrete sequence of choices corresponds to a whittling away of other components from the score until the sampler reaches the end component. Through adding larger scales of noise, contributions to the score from increasingly distant classes are incorporated into the reverse process.

2. Related work

Comparison to (Georgiev et al., 2023). Georgiev et al. (2023) empirically studied a variant of critical windows in the context of data attribution. For a generated image x_0 given by some trajectory $\{x_t\}_{t\in[0,T]}$ of the reverse process, they consider rerunning the reverse process starting at some intermediate point x_t in the trajectory (they refer to this as sampling from the "conditional distribution," which they denote by $p(\cdot \mid x_t)$). They then compute the probability that the images sampled in this fashion share a given feature with x_0 and identify critical times $T^{\text{cond}}_{\text{lower}} < T^{\text{cond}}_{\text{upper}}$ such that sampling from $p(\cdot|x_{T^{\text{cond}}_{\text{lower}}})$ preserves the given feature in the original image while sampling from $p(\cdot|x_{T^{\text{cond}}_{\text{upper}}})$ does not. Our definition is different: instead of rerunning the reverse process, we run the forward process for time t starting from x_0 to produce x^{cond}_t and then run the reverse process from x^{cond}_t to sample from $p(\cdot|x^{\text{cond}}_t)$. Note that in our definition, even after the initial generation $\{x_t\}$ is fixed, there is still randomness in x^{cond}_t. This means that unlike the setting in Georgiev et al. (2023), our setup is meaningful even if the reverse process is deterministic, e.g. based on an ODE. Additionally, our setup is arguably more flexible for data attribution as it does not require knowledge of the *trajectory* $\{x_t\}_{t\in[0,T]}$ that generated x_0. In general, we expect that our critical window thresholds are less than Georgiev et al. (2023)'s thresholds because adding noise to the state at intermediate times could also change the features. We view our theoretical contributions as complementary to their empirical work in rigorously understanding qualitatively similar phenomena and also use CLIP for our own experiments.

Comparison to (Raya & Ambrogioni, 2023; Sclocchi et al., 2024; Biroli et al., 2024). We discuss the relationship between our work and works by (Raya & Ambrogioni, 2023; Sclocchi et al., 2024; Biroli et al., 2024) that also studied the critical window phenomena from a theoretical perspective. Raya & Ambrogioni (2023) analyze the phase transition through the Hessian of the potential and give an end-to-end asymptotic analysis of critical windows for a discrete distribution supported on two points, and some partial results for more general discrete distributions. Sclocchi et al. (2024) also considered running the forward process for some time t starting from a sample and then running the reverse process, which they refer to as "forward-backward experiments." They give accurate but non-rigorous statistical physics-based predictions for critical windows by passing to a mean-field approximation. Biroli et al. (2024) study a mixture of *two* spherical Gaussians, and identify a phase transition that they call "speciation" which roughly corresponds to the critical time. They apply a Landau-type perturbative calculation to give highly precise but non-rigorous asymptotic predictions for the transition time.

In comparison to these works, we give *fully rigorous, end-to-end bounds* on the location of critical windows for a *general family of high-dimensional distributions*. We find it very interesting that critical windows can be understood through such different and complementary theoretical lenses. For example, Biroli et al. (2024) suggest a useful heuristic based on the time at which the noise obscures the principal component of the data distribution and validate this heuristic on real data. In our mixture model setting, this is closely related to the separation between components and thus suggests ties from their theory and numerics to ours.

These works also conduct experiments on real data. We defer a comparison of theirs and our experiments to Appendix A.

Theory for diffusion models. Recently several works have proven convergence guarantees for diffusion models (De Bortoli et al., 2021; Block et al., 2022; Chen et al., 2022; De Bortoli, 2022; Lee et al., 2022; Liu et al., 2022; Pidstrigach, 2022; Wibisono & Yang, 2022; Chen et al.,

2023c;d; Lee et al., 2023; Li et al., 2023a; Benton et al., 2023b; Chen et al., 2023b; Li et al., 2024). Roughly speaking, these results show that diffusion models can sample from essentially any distribution over $\mathbb{R}^d$, assuming access to a sufficiently accurate estimate for the score function. Our work is orthogonal to these results as they focus on showing that diffusion models can be used to sample. In contrast, we take for granted that we have access to a diffusion model that can sample; our focus is on specific properties of the sampling process. That said, there are isolated technical overlaps, for instance the use of path-based analysis via Girsanov's theorem, similar to (Chen et al., 2023c).

Mixtures of Gaussians and score-based methods. Gaussian mixtures have served as a fruitful testbed for the theory of score-based methods. In (Shah et al., 2023), the authors analyzed a gradient-based algorithm for learning the score function for a mixture of spherical Gaussians from samples and connected the training dynamics to existing algorithms for Gaussian mixture learning like EM. In (Cui et al., 2023), the authors gave a precise analysis of the training dynamics and sampling behavior for mixtures of two well-separated Gaussians using tools from statistical physics. Other works have also studied related methods like Langevin Monte Carlo and tempered variants (Koehler & Vuong, 2023; Lee et al., 2018) for learning/sampling from Gaussian mixtures. We do not study the learnability of Gaussian mixtures. Instead, we assume access to the true score and try to understand specific properties of the reverse process.

3. Technical preliminaries

3.1. Probability and diffusion basics

Probability notation. We consider the following divergences and metrics for probability measures. Given distributions P, Q, we use $\mathrm{TV}(P,Q) \triangleq \frac{1}{2}\int |dP - dQ| d\mu$ to denote the *total variation distance*, $\mathrm{H}^2(P,Q) \triangleq \int (\sqrt{dP} - \sqrt{dQ})^2 d\mu$ to denote the squared *Hellinger distance*, and $\mathrm{W}_2(P,Q) \triangleq \sqrt{\inf_{\gamma\sim\Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\|x-y\|^2}$, where $\Gamma(P,Q)$ is the set of all couplings between P, Q, to denote the *Wasserstein-2 distance*. We use the following basic relation among these quantities, a proof of which we include in Appendix B.1 for completeness.

Lemma 3.1. *For probability measures P, Q, $\mathbb{E}_{x\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P+\mathrm{d}Q}\right] \le \frac{1}{2}\sqrt{1-\mathrm{TV}^2(P,Q)}$.*

Let $\mathrm{subG}_d(\sigma^2)$ denote the class of sub-Gaussian random vectors in $\mathbb{R}^d$ with variance proxy σ^2. Let $\mathrm{SLC}(\beta, d)$ denote the set of $1/\beta$-strongly log-concave distributions over $\mathbb{R}^d$.

Diffusion model basics. Let q be a distribution over $\mathbb{R}^d$ with smooth density. In diffusion models, there is a *forward process* which progressively transforms samples from q into pure noise, and a reverse process which undoes this process. For the former, we consider the Ornstein-Uhlenbeck process for simplicity. This is a stochastic process $(X_t)_{t\ge 0}$ given by the stochastic differential equation (SDE)

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad X_0 \sim q\,, \tag{1}$$

where $(B_t)_{t\ge 0}$ is a standard Brownian motion. Given $t \ge 0$, let $q_t \triangleq \mathrm{law}(X_t)$, so as $t \to \infty$, q_t converges exponentially quickly to the standard Gaussian distribution γ^d.

Let $T \ge 0$ denote a choice of terminal time for the forward process. For the reverse process, denoted by $(X_t^{\leftarrow})_{t\in[0,T]}$, we consider the standard reverse SDE given by

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{2}$$

for $X_0^{\leftarrow} \sim q_T$, where here $(B_t)_{t\ge 0}$ is the reversed Brownian motion. The most important property of the reverse process is that q_{T-t} is precisely the law of $X_t^{\leftarrow}$.

Girsanov's theorem. The following is implicit in an approximation argument due to (Chen et al., 2023c) which is applied in conjunction with Girsanov's theorem. This lets us compare the path measures of the solutions to two SDEs with the same initialization:

Theorem 3.2 (Section 5.2 of (Chen et al., 2023c))**.** *Let $(Y_t)_{t\in[0,T]}$ and $(Y_t')_{t\in[0,T]}$ denote the solutions to*

$$\mathrm{d}Y_t = b_t(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y_0 \sim q$$
$$\mathrm{d}Y_t' = b_t'(Y_t')\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y_0' \sim q\,.$$

Let q and q' denote the laws of Y_T and Y_T' respectively. If b_t, b_t' satisfy that $\int_0^T \mathbb{E}_Q\,\|b_t(Y_t) - b_t'(Y_t)\|^2\,\mathrm{d}t < \infty$, then $\mathrm{KL}(q\|q') \le \int_0^T \mathbb{E}_Q\,\|b_t(Y_t) - b_t'(Y_t)\|^2\,\mathrm{d}t$.

3.2. Main framework: noising and denoising mixtures

We will consider data distributions p given by *mixture models*. For component distributions $p^1, \dots, p^K$ over $\mathbb{R}^d$ and mixing weights $w_1, \dots, w_K$ summing to 1, let $p \triangleq \sum_i w_i p^i$. Let μ_i denote the mean of p^i. For any nonempty $S \subset [K]$, we define the *sub-mixture* p^S by $p^S \triangleq \sum_{i\in S} \frac{w_i}{\sum_{j\in S} w_j} p^i$. Let $(X_t^S)_{t\in[0,T]}$ denote the forward process given by running Eq. (1) with $q = p^S$, let p_t^S denote the law of X_t^S, and let $(X_t^{\leftarrow,S})$ denote the reverse process given by running Eq. (2) with $q = p^S$. When $S = \{i\}$, we drop the braces in the superscripts. Given intermediate time $\widehat{T} \in [0,T]$, we denote the path measure for $(X_t^{\leftarrow,S})_{t\in[0,\widehat{T}]}$ by $P_{\widehat{T}}^{\leftarrow,S} \in \mathcal{C}([0,\widehat{T}], \mathbb{R}^d)$.

The targeted reverse process. The central object of study in this work is a modification of the reverse process for the

overall mixture p in which the initialization is changed from p_T to an *intermediate point* in the forward process *for a sub-mixture*. Concretely, given $\widehat{T} \in [0, T]$ and nonempty $S \subset [K]$, define the modified reverse process $(X_t^{\leftarrow}[S^{\langle \widehat{T} \rangle}])_{t \in [0,\widehat{T}]}$ to be given by running the reverse SDE in Eq. (2) with $q = p$, with terminal time $\widehat{T}$ instead of T, and initialized at $p_{\widehat{T}}^S$ instead of $p_{\widehat{T}}$. We denote the law of $X_t^{\leftarrow}[S^{\langle \widehat{T} \rangle}]$ by $p_{T-t}[S^{\langle \widehat{T} \rangle}]$ and the path measure for $(X_t^{\leftarrow}[S^{\langle \widehat{T} \rangle}])_{t \in [0,\widehat{T}]}$ by $P^{\leftarrow}[S^{\langle \widehat{T} \rangle}] \in \mathcal{C}([0, \widehat{T}], \mathbb{R}^d)$. When $t = T$, we omit the subscript in the former.

1. Draw a sample X from the sub-mixture p^S
2. Run forward process for time $\widehat{T}$ from X to produce X'
3. From terminal time $\widehat{T}$, run the reverse process starting from X' for time t to produce $X_t^{\leftarrow}[S^{\langle \widehat{T} \rangle}]$

Because this process reverses the forward process conditioned on a particular subset S of the original mixture components, we refer to $(X_t^{\leftarrow}[S^{\langle \widehat{T} \rangle}])_{t \in [0,\widehat{T}]}$ as the *S-targeted reverse process* from noise level $\widehat{T}$. We caution that the S-targeted reverse process should not be confused with the standard reverse process where the data distribution is taken to be p^S, as the score function being used in the targeted process is that of the full mixture p rather than that of p^S.

Mixture model parameters. We consider the following quantities for a given mixture model, which characterize levels of separation within and across subsets of the mixture. Given $S, S' \subset [K]$, define

$$\overline{R} \triangleq \max_{i \in [K]} \|\mu_i\| \qquad w(S, S') \triangleq \max_{i \in S, j \in S'} \|\mu_i - \mu_j\|$$
$$\Delta(S) \triangleq \min_{\ell \in S, j \in [K]-S} \|\mu_\ell - \mu_j\| \qquad \overline{W} \triangleq \max_{i,j \in [K]} \frac{w_i}{w_j}.$$

Lastly, we characterize the level of imbalance across subpopulations via $\overline{W} \triangleq \max_{i,j \in [K]} \frac{w_i}{w_j}$.

4. Master theorem for critical times

Recall that $S_{\text{init}} \subset S_{\text{targ}} \subset [K]$ denote the two sub-mixtures we are interested in. In the notation of Section 3.2, we wish to establish upper and lower bounds on the time $\widehat{T}$ at which

$$\text{TV}(p[S_{\text{init}}^{\langle \widehat{T} \rangle}], p^{S_{\text{targ}}}) \tag{3}$$

becomes small.

Given error parameter $0 < \epsilon < 1$, define

$$T_{\text{lower}}(\epsilon) \triangleq \inf\{t \in [0, T] : \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{targ}}}) \le \epsilon\} \tag{4}$$
$$T_{\text{upper}}(\epsilon) \triangleq \sup\{t \in [0, T] : \text{TV}(p_t^i, p_t^j) \ge 1 - \epsilon^2/2 \quad \forall i \in S_{\text{targ}}, j \in [K] - S_{\text{targ}}\}. \tag{5}$$

When ϵ is clear from context, we refer to these times as T_{lower} and T_{upper}. Based on the intuition above, we expect that Eq. (3) is small provided $\widehat{T} \ge T_{\text{lower}}$ and $\widehat{T} \le T_{\text{upper}}$. In this section, we prove that this is indeed the case for any p given by a mixture of strongly log-concave distributions.[4]

Assumption 4.1 (Strong log-concavity)**.** For some $\Psi^2 \ge 1$, $p^i \in \text{SLC}(\Psi^2, d)$.

Assumption 4.2 (Smooth components)**.** For some $L > 0$ and for all $t \ge 0$, the score $\nabla \ln p_t^i$ is L-Lipschitz.

Assumption 4.3 (Moment bound)**.** For some $M \ge 1$ and for all $i \in [K]$ and $t \in [0, T]$, $\mathbb{E}\,\|X_t^i\|^4 \le M$.

Finally, our bounds will depend on how large the score for any component is over samples from any other component:

Assumption 4.4 (Score bound)**.** For some $\overline{M} \ge 0$ and for all $i, j \in [K]$, $t \in [0, T]$, $\mathbb{E}_{X \sim p_t^j} \|\nabla \ln p_t^i(X)\|^4 \le \overline{M}$.

We compute $\overline{M}$ for various examples in Section 5.2, but for now one can safely think of $\overline{M}$ as scaling polynomially in the dimension and in the parameter $\overline{R}$.

4.1. Main result and proof sketch

We are now ready to state our main bound for the critical time $\widehat{T}$ at which Eq. (3) becomes small.

Theorem 4.5. *Let $S_{\text{init}} \subset S_{\text{targ}} \subset [K]$. For $\epsilon > 0$, if $\widehat{T} \ge T_{\text{lower}}(\epsilon)$ and $\widehat{T} \le T_{\text{upper}}(\epsilon)$, then*

$$\text{TV}(p[S_{\text{init}}^{\langle \widehat{T} \rangle}], p^{S_{\text{targ}}}) \lesssim \epsilon \sqrt{\overline{W}} K^2 \left(\overline{R}^2 + M^2 + \sqrt{M}\Psi^4 + \sqrt{\overline{M}}\right). \tag{6}$$

The proof of Theorem 4.5 relies on the following technical lemma whose proof we defer to Appendix B.2.

Lemma 4.6. *Under Assumptions 4.1, 4.3, and 4.4,* $\mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim e^{-4t}(\overline{R}^4 + M^4 + M\Psi^8 + \overline{M})$ $\forall i, j, \ell \in [K]$.

Informally, this lemma quantities the extent to which the score functions for p^j and p^ℓ become close over the course of the forward process, as measured by an average sample from any other component of the mixture.

Proof of Theorem 4.5. By data processing inequality and definition of $T_{\text{lower}}, T_{\text{upper}}$, for all $i \in S_{\text{targ}}, j \notin S_{\text{targ}}$,

$$\text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{targ}}}) \le \epsilon \qquad \forall t \in [\widehat{T}, T] \tag{7}$$
$$\text{TV}(p_t^i, p_t^j) \ge 1 - \epsilon^2/2 \qquad \forall t \in [0, \widehat{T}], \tag{8}$$

[4] It turns out that the only place where we need strong log-concavity of the components in the mixture is in the rather technical estimate of Lemma 4.6, which is also much stronger than what is necessary for Theorem 4.5. It suffices to show the LHS of Lemma 4.6 integrates to a finite value. While we only prove the bound in that Lemma rigorously for strongly log-concave components, we expect it to hold even for more general families of non-log-concave distributions.

By data processing inequality and triangle inequality,

$$\mathrm{TV}(p[S_{\mathrm{init}}^{\langle\widehat{T}\rangle}], p^{S_{\mathrm{targ}}}) \le \mathrm{TV}(P^{\leftarrow}[S_{\mathrm{init}}^{\langle\widehat{T}\rangle}], P_{\widehat{T}}^{\leftarrow, S_{\mathrm{targ}}})$$
$$\le \underbrace{\mathrm{TV}(P^{\leftarrow}[S_{\mathrm{init}}^{\langle\widehat{T}\rangle}], P^{\leftarrow}[S_{\mathrm{targ}}^{\langle\widehat{T}\rangle}])}_{\text{(I)}} + \underbrace{\mathrm{TV}(P^{\leftarrow}[S_{\mathrm{targ}}^{\langle\widehat{T}\rangle}], P_{\widehat{T}}^{\leftarrow, S_{\mathrm{targ}}})}_{\text{(II)}} .$$

As $P^{\leftarrow}[S_{\mathrm{init}}^{\langle\widehat{T}\rangle}]$ and $P^{\leftarrow}[S_{\mathrm{targ}}^{\langle\widehat{T}\rangle}]$ are the path measures for the solutions to the same SDE with initializations $p_{\widehat{T}}^{S_{\mathrm{init}}}$ and $p_{\widehat{T}}^{S_{\mathrm{targ}}}$ respectively, we can use data processing again to bound (I) via

$$\mathrm{TV}(P^{\leftarrow}[S_{\mathrm{init}}^{\langle\widehat{T}\rangle}], P^{\leftarrow}[S_{\mathrm{targ}}^{\langle\widehat{T}\rangle}]) \le \mathrm{TV}(p_{\widehat{T}}^{S_{\mathrm{init}}}, p_{\widehat{T}}^{S_{\mathrm{targ}}}) \le \epsilon. \quad (9)$$

To bound (II), we apply Pinsker's and Theorem 3.2 to bound $\mathrm{TV}(P^{\leftarrow}[S_{\mathrm{targ}}^{\langle\widehat{T}\rangle}], P_{\widehat{T}}^{\leftarrow, S_{\mathrm{targ}}})^2$ by

$$\int_0^{\widehat{T}} \mathbb{E}\,\|\nabla \ln p_t(X_t^{S_{\mathrm{targ}}}) - \nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}})\|^2 \,\mathrm{d}t\,.$$

We have the following identity (see Appendix B.3 for proof):

Lemma 4.7. $\|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t\|^2 = \|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}\|^2 \cdot \left(\frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i}\right)^2$.

Using this expression, we can invoke Cauchy-Schwarz to separate the two terms that appear on the right-hand side. We bound these two terms in turn. Recalling the definition of $\overline{W}$ and also applying Lemma 3.1, we see that for any $j \in S_{\mathrm{targ}}$,

$$\mathbb{E}_{p_t^j}\Big(\frac{\sum_{i\in[K]\setminus S_{\mathrm{targ}}} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i}\Big)^4 \le \sum_{\ell\in[K]\setminus S_{\mathrm{targ}}} \mathbb{E}_{p_t^j}\Big[\frac{w_\ell p_t^\ell}{w_j p_t^j + w_\ell p_t^\ell}\Big]$$
$$\lesssim K\overline{W} \max_{\ell\in[K]\setminus S_{\mathrm{targ}}} \sqrt{1 - \mathrm{TV}^2(p_t^\ell, p_t^j)} \lesssim K\overline{W}\epsilon^2\,,$$

where in the last step we used Eq. (7). By convexity, the same bound thus holds when the expectation on the left-hand side is replaced by an expectation with respect to $p_t^{S_{\mathrm{targ}}}$.

By the same convexity argument, to bound $\mathbb{E}\,\|\nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}}) - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}})\|^4$, it suffices to show that the expectations

$$\mathbb{E}\,\|\nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^i) - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}(X_t^i)\|^4 \quad (10)$$

for all $i \in [K]$ are bounded. Moreover, the score of a mixture is a weighted average of the scores of the components, $\nabla \ln p_t^{S_{\mathrm{targ}}} = \sum_{i\in S_{\mathrm{targ}}} \frac{w_i p_t^i}{\sum_{j\in S_{\mathrm{targ}}} w_j p_t^j} \nabla \ln p_t^i$. By the triangle inequality, $\|\nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^i) - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}(X_t^i)\|$ is at most the difference between two elements of a weighted score. Thus, we have

$$\mathbb{E}\,\|\nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}}) - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}})\|^4$$
$$\le \mathbb{E}_i\mathbb{E} \max_{\substack{i,j\in S_{\mathrm{targ}}\\ \ell\in[K]-S_{\mathrm{targ}}}} \|\nabla \ln p_t^j(X_t^i) - \nabla \ln p_t^\ell(X_t^i)\|^4$$
$$\le K^3 \max_{\substack{i,j\in S_{\mathrm{targ}}\\ \ell\in[K]-S_{\mathrm{targ}}}} \mathbb{E}\,\|\nabla \ln p_t^j(X_t^i) - \nabla \ln p_t^\ell(X_t^i)\|^4.$$

Thus we can conclude by applying Lemma 4.6 and bound $\mathbb{E}\,\|\nabla \ln p_t^{S_{\mathrm{targ}}}(X_t^{S_{\mathrm{targ}}}) - \nabla \ln p_t(X_t^{S_{\mathrm{targ}}})\|^2$ by $O(\epsilon\sqrt{\overline{W}}K^2(\overline{R}^2 + M^2 + \sqrt{M}\Psi^4 + \sqrt{\overline{M}})e^{-2t})$. Integrating over $[0, \widehat{T}]$ completes the proof. □

5. Instantiating the master theorem

We now consider cases where we can provide concrete bounds on $T_{\mathrm{lower}}, T_{\mathrm{upper}}$. Our bounds here hold independent of the Assumptions in Section 4.

5.1. General mixtures with similar components

We first consider the case where the components of the mixture are "similar" in the sense that if we take any two components and translate them to both have mean zero, then they are moderately close in Wasserstein distance. Here, we obtain the following bounds on T_{lower} and T_{upper}:

Lemma 5.1. *Let $\epsilon > 0$. For $i \in [K]$, let $\overline{p}^i$ denote the density of the i-th component of the mixture model p after being shifted to have mean zero. Suppose $\mathrm{W}_2(\overline{p}^i, \overline{p}^j) \le \Upsilon$ for all $i, j \in [K]$. Then $T_{\mathrm{lower}}(\epsilon) \le \Big\{\ln(w(S_{\mathrm{init}}, S_{\mathrm{targ}}) + \Upsilon) + \ln\frac{1}{\epsilon} + \frac{1}{2}\ln 2\Big\} \vee 3$. Additionally, if $p_0^i \in \mathrm{subG}_d(\sigma^2)$ for all $i \in [K]$, then $T_{\mathrm{upper}}(\epsilon) \ge \ln\Delta(S_{\mathrm{targ}}) - \ln\sigma - \ln\sqrt{8d\ln 6 + 8\ln 4/\epsilon^2} - \ln 3 - \frac{1}{2}\ln 8$.*

Proof sketch of Lemma 5.1, see Appendix B.4. For T_{lower}, we apply Pinsker's inequality and a Wasserstein smoothing to upper bound the TV between components in the initial and target mixture in terms of the Wasserstein-2 distance of the components, which decreases at the rate of $O(e^{-t}(w(S_{\mathrm{init}}, S_{\mathrm{targ}}) + \Upsilon))$. For T_{upper}, we use sub-Gaussian concentration bounds to lower bound the TV between components in S_{end} and $[K] - S_{\mathrm{end}}$. □

Note that because all α-strongly log-concave distributions are sub-Gaussian with variance proxy $\Theta(1/\alpha)$, under Assumption 4.1 of Section 4 the above applies for $\sigma \asymp \Psi$.

When the terms $\Upsilon, \Psi, 1/\epsilon$ are sufficiently small, our bounds on T_{lower} and T_{upper} are dominated by $\ln w(S_{\mathrm{init}}, S_{\mathrm{targ}})$ and $\ln\Delta(S_{\mathrm{targ}})$ respectively. Recall that $w(S_{\mathrm{init}}, S_{\mathrm{targ}})$

and $\Delta(S_{\text{targ}})$ respectively correspond to the maximum distance between any two component means from S_{init} and S_{targ}, and the minimum distance from S_{targ} to the rest of the mixture. This, combined with our master theorem, has the favorable interpretation that as long as the separation between components within S_{init} and S_{targ} is dominated by the separation between components in S_{targ} vs. outside S_{targ}, then there is a non-empty window of times $\widehat{T} \in [T_{\text{lower}}, T_{\text{upper}}]$ such that the S_{init}-targeted reverse process from noise level $\widehat{T}$ results in samples close to S_{targ}.

5.2. Mixtures of well-conditioned Gaussians

We now suppose p is a mixture of Gaussians, with $p^i = \mathcal{N}(\mu_i, \Sigma_i)$. At time $t \geq 0$ in the forward process, if $\mu_i(t) \triangleq e^{-t}\mu_i$ and $\Sigma_i(t) \triangleq e^{-2t}\Sigma_i + (1 - e^{-2t})\text{Id}$, then $p_t^i = \mathcal{N}(\mu_i(t), \Sigma_i(t))$, $p_t = \sum w_i \mathcal{N}(\mu_i(t), \Sigma_i(t))$. We also define $\sigma^2_{\max}(t) := \max_i \sigma^2_{\max}(\Sigma_i(t))$, $\sigma^2_{\min}(t) = \min_i \sigma^2_{\min}(\Sigma_i(t))$, and $\overline{R}(t) = e^{-t}\max_i \|\mu_i\|$.

Assumption 5.2. There exists $\underline{\lambda} \leq 1 \leq \overline{\lambda}$ such that for all $t \geq 0$, $\underline{\lambda} \leq \sigma^2_{\min}(\Sigma_i) \leq \sigma^2_{\max}(\Sigma_i) \leq \overline{\lambda}$. Note that the same bound immediately holds for $\sigma^2_{\min}(t), \sigma^2_{\max}(t)$ as a result.

We can prove analogous bounds to Lemmas 3.1 and 4.6 in terms of these parameters, see Lemmas B.14 and 4.6 in Appendices B.5 and B.6. Using these ingredients, we prove in Appendix B.7 the following:

Theorem 5.3. *Take any $S_{\text{init}} \subset S_{\text{targ}} \subset [K]$. For sufficiently small ϵ, there exists T_{lower} and T_{upper} such that $T_{\text{lower}} \leq \frac{1}{2}\ln\left(2d\frac{\overline{\lambda}-\underline{\lambda}}{\underline{\lambda}} + \frac{1}{\underline{\lambda}}w(S_{\text{init}}, S_{\text{targ}})^2\right) + \ln\frac{1}{\epsilon}$ and also $T_{\text{upper}} \geq \ln\Delta(S_{\text{targ}}) + \frac{1}{2}\ln\underline{\lambda} - \ln 4 - \frac{1}{2}\ln\ln\left(\frac{\overline{\lambda}\sqrt{K\overline{W}}\left[(\overline{\lambda}-\underline{\lambda})^2(\overline{R}(0)^2+\overline{\lambda}d)+\overline{R}(0)^2\right]}{\underline{\lambda}^2\Delta(S_{\text{targ}})^2\epsilon^2}\right)$ and such that for any $\widehat{T} \in (T_{\text{lower}}, T_{\text{upper}})$, $\text{TV}(p[S_{\text{init}}^{\langle\widehat{T}\rangle}], p^{S_{\text{targ}}}) \lesssim \epsilon$.*

To get intuition for the bound, consider the simpler scenario where the covariances are the identity matrix.

Example 1. *(K Gaussians with identity covariance) Let $\Sigma_0^i = \text{Id}$ for all $i \in [K]$. Then, for any $S_{\text{init}} \subset S_{\text{targ}} \subset [K]$, $T_{\text{lower}} = \ln w(S_{\text{init}}, S_{\text{targ}}) + \ln 1/\epsilon$ and $T_{\text{upper}} = \ln\Delta(S_{\text{targ}}) - \ln 4 - \frac{1}{2}\ln\ln\frac{\overline{R}(0)^2\sqrt{K\overline{W}}}{\epsilon^2\Delta(S_{\text{targ}})^2}$. The dominant terms are $\ln w(S_{\text{init}}, S_{\text{targ}})$ and $\ln\Delta(S_{\text{targ}})$, which depend on the intra- and inter-group distances of the means. In Fig. 3, we plot these critical times and the final membership of the noised then denoised points for a Gaussian mixture. We see that our bounds match real class membership.*

6. Hierarchy of classes

In this section, we consider a *sequence* of critical windows that enable sampling from a sequence of nested submixtures. Figure 3 hints at this idea, that as we noise for longer time periods, we sample from more and more components. Before we continue, it will be useful to formalize our model of a hierarchy of classes as a tree.

Definition 6.1. We define a **mixture tree** as a tuple (T, h, f, R). A tree $T = (V, E)$ of height $H = O(\sqrt{\ln R})$ is associated with a height function $h : V \to \mathbb{N}$ mapping vertices to their distance to the root and a function $f : V \to 2^{[K]}\backslash\{\emptyset\}$ satisfying the following: (1) $f(\text{root}) = [K]$; (2) if u is a parent of v, $f(v) \subset f(u)$; (3) for distinct $i, j \in [K]$ with leaf nodes w, v such that $i \in f(w), j \in f(v)$, if u is the lowest common ancestor of w, v, then $\|\mu_i - \mu_j\| \in (1 \pm \delta)\ln\frac{R}{2^{h(u)^2}}$ with $\delta < 0.01$.

Intuitively, the sequence of increasing critical windows of the noising and denoising process acts as a path up a mixture tree from some leaf. Within each critical window, the noising and denoising process is sampling from every class in the corresponding node in the path to the root. The class means have to be within a constant factor of $\ln\frac{R}{2^{h^2}}$, where h is the height of their lowest common ancestor, to both ensure statistical separation from components outside the target mixture and small statistical distance within the target mixture. To make the critical times more explicit, we consider the setting of a mixture of identity covariance Gaussians (see proof in Appendix B.8):

Theorem 6.2. *Let all $\Sigma^i = \text{Id}$, $\|\mu_i\| = R$, and $w_i = \frac{1}{K}$. For $i \in [K]$, consider the path $u_1, u_2, u_3, \ldots, u_{H'}$ where u_1 is the leaf node with $i \in f(u_1)$ and $u_{H'}$ is the root. There exists $k \in [1, 2, \ldots, H']$, sufficiently large R, H', and sufficiently small ϵ such that there is a sequence of times $T_1 < T_2 < \cdots < T_k$ with $\text{TV}(p[\{i\}^{\langle T_\ell\rangle}], p^{f(u_\ell)}) \lesssim \epsilon$.*

This model also captures the intuition that diffusion models select more substantial features of an image before resolving finer details. When one ascends a tree of sub-mixtures from a leaf to the root through noising, one is essentially adding contributions to the score from more and more components of the mixture. Similarly, when a diffusion model samples from a hierarchy, it can be seen as ignoring negligible components of the mixture from the score until it reaches the end component.

7. Critical windows in Stable Diffusion

In this section, we give an example of a critical window in Stable Diffusion v2.1 (SD2.1) to corroborate our theory. We generated images of cars and chose color, background, and size as our features. We noised and denoised each image for $t = 350$ to 490 time and plotted percentage of feature agreement with the base image vs. time (Figure 5). We defer experimental details to Appendix D. Note the background feature: from time step 480 to 490, the percentage of images with the same background as the original image drops by 25%. The size feature also sees a substantial drop from 470

to 490 by 15%. The agreement for the color also decreases significantly but the drop is much less sharp and occurs between time steps 450 to 470. Our theory for hierarchical sampling suggests that the diffusion model selects the car's size and background before deciding the color. This experiment also implies that critical windows can exist when the target mixture is different from the initial mixture, because noising and denoising to $[450, 470]$ is sampling from the superclass of cars that have the same size and background as the original car but different colors.

8. Applications to fairness and privacy

8.1. Fairness

Generative models can reproduce social biases with their outputs (Luccioni et al., 2023). Do potentially biased features like gender have critical windows? This could help design specific interventions to narrow ranges in the diffusion process to improve image diversity (Raya & Ambrogioni, 2023). We studied outputs of photo portraits of laboratory technician on SD2.1 (Luccioni et al., 2023), sampled 200 images (see Figure 6 for examples), and created an analogous plot of critical times (Figure 2). To determine gender, we against used a CLIP model and tested whether a given image had higher dot product with the prompt appended with "`, male`" or "`, female`". We can see a large drop in agreement between $t = 80$ and $t = 84$, from over 80% to roughly 50%, suggesting a critical window for the gender feature. If the male and female classes are not well-separated at time $t = 80$, then the noising and denoising procedure should result in a more equal mix of images from both classes. This confirms the intuition of our theory that different categories are well-separated before a critical window.

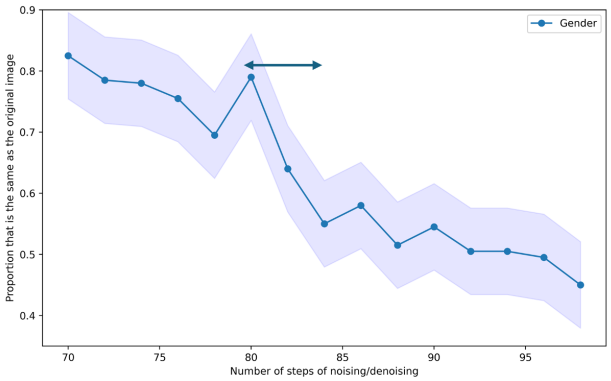

Figure 2. Critical window for gender feature in the experiment on images of laboratory technicians generated by SD2.1. Critical window demarcated with double-sided horizontal arrows.

8.2. A new Membership Inference Attack

Membership Inference Attacks (MIAs) are a class of privacy attacks that try to identify whether a candidate sample belonged to training data (Shokri et al., 2017), and are relevant for diffusion models because of their substantial privacy and copyright risks (Carlini et al., 2023). We present a simple MIA (NoiseDenoise) based on the distance between a candidate and noised and denoised copies (see Appendix E.2 for more experimental details). We applied our attack to a DDPM that was trained on CIFAR-10 in (Duan et al., 2023) and compare it to their methods $\mathrm{SecMI}_{\mathrm{stat}}$ and $\mathrm{SecMI}_{\mathrm{nn}}$. Their attacks exploit a deterministic approximation of the forward and reverse process of a DDPM to estimate the sampling error of a candidate image. Figure 7 and Table 1 show that $\mathrm{SecMI}_{\mathrm{stat}}$ and $\mathrm{SecMI}_{\mathrm{nn}}$ outperform NoiseDenoise, but $11/23$ of the train points NoiseDenoise identifies at FPR $= 0.01$ and $21/140$ of the train points identified at FPR $= 0.05$ are not classified correctly by $\mathrm{SecMI}_{\mathrm{stat}}$ or $\mathrm{SecMI}_{\mathrm{nn}}$ at the same FPR thresholds, suggesting NoiseDenoise can be a complementary approach to these methods.

9. Conclusion

We consider noising and denoising samples from a mixture model and the resulting critical times of this process over which features emerge. We provide theory for the empirical observation from Raya & Ambrogioni (2023), Biroli et al. (2024), Georgiev et al. (2023), and (Sclocchi et al., 2024) that discrete features are decided within short windows in the sampling process. We identified and proved a relationship between these critical windows and statistical distances between components in the initial and target sub-mixture. This same question was studied mathematically in recent and concurrent works (Raya & Ambrogioni, 2023; Sclocchi et al., 2024; Biroli et al., 2024), and our rigorous non-asymptotic bounds for mixture models nicely complement the precise statistical physics-based insights derived in those works. We also present preliminary experiments describing critical windows for features in SD2.1, and demonstrate our framework's value for fairness and privacy.

Limitations and future directions. The most immediate follow-up would be to eliminate the logarithmic dependence on dimension for T_{upper} for more general distributions beyond well-conditioned Gaussians. Another direction is to discover analogues of critical windows for continuous features. Some features, e.g., color, more naturally belong to a continuum rather than discrete bins, but our theorems require strong statistical separation between components inside and outside the target sub-mixture. Furthermore, this work presents exciting empirical opportunities for interpretability, designing better samplers, and data attribution. It would also be interesting to systematically characterize the critical times of features over diverse prompts.

Acknowledgements

SC would like to thank Aravind Gollakota, Adam Klivans, Vasilis Kontonis, Yuanzhi Li, and Kulin Shah for insightful discussions on diffusion models and Gaussian mixtures. ML would like to thank Andrew Campbell and Jason Wang for thoughtful conversations about diffusion models and the applications of this work.

Impact Statement

This paper is largely theoretical in nature but also describes a new membership inference attack against diffusion models. This could potentially impact the privacy of their training data. We hope this paper spurs further research into the fundamental mechanisms behind memorization in diffusion model so that these risks may be limited in the future. Furthermore, the diffusion model that we tested our MIA on was only trained on CIFAR-10 data, limiting the privacy risks of our study.

References

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023a.

Benton, J., Deligiannidis, G., and Doucet, A. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023b.

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

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint 2002.00107*, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023b.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023c. URL https://openreview.net/pdf?id=zyLVMgsZ0U_.

Chen, S., Daras, G., and Dimakis, A. G. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023d.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11462–11471, 2022. doi: 10.1109/CVPR52688.2022.01118.

Cui, H., Krzakala, F., Vanden-Eijnden, E., and Zdeborová, L. Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*, 2023.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion Schrödinger bridge with applications to score-based generative modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709. Curran Associates, Inc., 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Georgiev, K., Vendrow, J., Salman, H., Park, S. M., and Madry, A. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

Henningsson, T. and Åström, K. Log-concave observers. In *17th International Symposium on Mathematical Theory of Networks and Systems, 2006*, 2006. 17th International Symposium on Mathematical Theory of Networks and Systems, 2006 : MTNS 2006 ; Conference date: 24-07-2006 Through 28-07-2006.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020b. URL https://arxiv.org/abs/2006.11239.

Kannan, R., Lovász, L., and Simonovits, M. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13:541–559, 1995.

Koehler, F. and Vuong, T.-D. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.

LeCam, L. *Asymptotic methods in statistical decision theory*. Springer series in statistics. Springer, New York, NY [u.a.], 1986. ISBN 3540963073. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024181773&sourceid=fbw_bibsonomy.

Lee, H., Risteski, A., and Ge, R. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *Advances in neural information processing systems*, 31, 2018.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Li, G., Wei, Y., Chen, Y., and Chi, Y. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023a.

Li, G., Huang, Z., and Wei, Y. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024.

Li, M. and Wang, J. Zero-shot machine-generated image detection using sinks of gradient flows. https://github.com/deep-learning-mit/staging/blob/main/_posts/2023-11-08-detect-image.md, 2023.

Li, M., Wang, J., Wang, J., and Neel, S. MoPe: Model perturbation based privacy attacks on language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13647–13660, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.842. URL https://aclanthology.org/2023.emnlp-main.842.

Liu, X., Wu, L., Ye, M., and Liu, Q. Let us build bridges: understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models, 2023.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Pardo, L. *Statistical Inference Based on Divergence Measures*. CRC Press, Abingdon, 2005. URL https://cds.cern.ch/record/996837.

Pidstrigach, J. Score-based generative models detect manifolds. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35852–35865. Curran Associates, Inc., 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Raya, G. and Ambrogioni, L. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lxGFGMMSVl.

Rigollet, P. and Hutter, J.-C. High-dimensional statistics, 2023.

Saumard, A. and Wellner, J. A. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8(none):

45 – 114, 2014. doi: 10.1214/14-SS107. URL https://doi.org/10.1214/14-SS107.

Sclocchi, A., Favero, A., and Wyart, M. A phase transition in diffusion models reveals the hierarchical nature of data, 2024.

Shah, K., Chen, S., and Klivans, A. Learning mixtures of gaussians using the ddpm objective. *arXiv preprint arXiv:2307.01178*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.41. URL https://doi.org/10.1109/SP.2017.41.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Wibisono, A. and Yang, K. Y. Convergence in KL divergence of the inexact Langevin algorithm with application to score-based generative models. *arXiv preprint 2211.01512*, 2022.

A. Related Works continued

Critical window experiments on real data. Numerous studies have investigated the critical window phenomenon in diffusion models (Ho et al., 2020b; Raya & Ambrogioni, 2023; Biroli et al., 2024; Sclocchi et al., 2024). These papers demonstrate a dramatic jump in the similarity of some feature within a narrow time range, either under our noising and denoising framework (Sclocchi et al., 2024) or the "conditional sampling" framework (Georgiev et al., 2023; Raya & Ambrogioni, 2023; Biroli et al., 2024) described above. The main distinguishing factors between these different experiments are the diffusion models tested, the varying definitions of what a "feature" entails, and the method to determine whether a given image has a certain feature. Raya & Ambrogioni (2023); Georgiev et al. (2023); Sclocchi et al. (2024); Biroli et al. (2024) identify the critical windows of class membership for unconditional diffusion models operating in pixel space that were trained on small, hand-labeled datasets like MNIST or CIFAR-10. Biroli et al. (2024) were able to obtain precise predictions for the critical times for a simple diffusion model trained on two classes. Raya & Ambrogioni (2023); Georgiev et al. (2023); Biroli et al. (2024) trained supervised classifiers to sort image generations into different categories, whereas Sclocchi et al. (2024) employed the hidden layer activations of an ImageNet classifier to define high- and low-level features of an image and computed the cosine similarity of the embeddings of the base and new image. Among all these empirical results, our experimental setup most closely mirrors Figure B.10 of Georgiev et al. (2023); we both experiment with StableDiffusion 2.1, manually inspect the image for potential features, and use CLIP instead of a supervised classifier to label images into different categories. That said, recall from the discussion at the beginning of this section that this paper and Georgiev et al. (2023)'s experiments examine different critical window frameworks (noise and denoise vs. conditional sampling).

B. Deferred proofs

B.1. Proof of Lemma 3.1

Lemma B.1. *For probability measures P, Q, $\mathbb{E}_{x\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P+\mathrm{d}Q}\right] \le \frac{1}{2}\sqrt{1-\mathrm{TV}^2(P,Q)}$.*

Proof. Let $\mathrm{LC}(P,Q) \triangleq \frac{1}{2}\int \frac{(dP-dQ)^2}{d(P+Q)}d\mu$ denote the *Le Cam distance*. It suffices to show

$$\mathbb{E}_{x\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P+\mathrm{d}Q}\right] = \frac{1}{2}(1-\mathrm{LC}(P,Q)) \le \frac{1}{2}(1-\frac{1}{2}\mathrm{H}^2(P,Q)) \le \frac{1}{2}\sqrt{1-\mathrm{TV}^2(P,Q)}.$$

We exhibit the leftmost equality by noting $dPdQ = \frac{1}{4}\left((dP+dQ)^2-(dP-dQ)^2\right)$,

$$\begin{aligned}
\mathbb{E}_{x\sim P}\left[\frac{dQ}{dP+dQ}\right] &= \int \frac{dPdQ}{d(P+Q)} && (11)\\
&= \frac{1}{4}\left[\int \frac{(dP+dQ)^2-(dP-dQ)^2}{d(P+Q)}\right] && (12)\\
&= \frac{1}{4}\left[2-\int \frac{(dP-dQ)^2}{d(P+Q)}\right] && (13)\\
&= \frac{1}{2}\left[1-\mathrm{LC}(P,Q)\right]. && (14)
\end{aligned}$$

The first inequality follows from $\mathrm{LC}(P,Q) \ge \frac{1}{2}\mathrm{H}^2(P,Q)$ (see p.48 in (LeCam, 1986)). The second inequality follows from rearranging $4\mathrm{TV}^2(P,Q) \le \mathrm{H}^2(P,Q)(4-\mathrm{H}^2(P,Q))$ (see p.47 in (LeCam, 1986)) into $1-\frac{1}{2}\mathrm{H}^2(P,Q) \le \sqrt{1-\mathrm{TV}^2(P,Q)}$. □

B.2. Proof of Lemma 4.6

Lemma B.2. *Under Assumption 4.1, the Hessian of $\ln p_t^i$ for $i \in [K]$ is between*

$$\frac{1}{e^{-2t}\Psi^2+1-e^{-2t}}\mathrm{Id} \preceq \nabla^2(-\ln p_t^i) \preceq \frac{1}{1-e^{2t}}\mathrm{Id}. \tag{15}$$

Proof. Using the preservation of strong log-concavity (see p.71 in in (Saumard & Wellner, 2014) or (Henningsson & Åström, 2006)), we find that for $i \in [K]$,

$$p_t^i \in \mathrm{SLC}(e^{-2t}\Psi^2 + (1-e^{-2t}), d).$$

By Proposition 2.23 of (Saumard & Wellner, 2014), this implies $\nabla^2(-\ln p_t^i) \succeq \frac{1}{e^{-2t}\Psi^2+(1-e^{-2t})}$. For the second inequality, we follow the proof of Proposition 7.1. in (Saumard & Wellner, 2014) for the convolution $X_t^i = e^{-t}X_0^i + \mathcal{N}(0,(1-e^{-2t})\mathrm{Id})$. Let $X := e^{-t}X_0^i, Y := \mathcal{N}(0,(1-e^{-2t})\mathrm{Id}), Z := X_t^i$, and let p_X, p_Y, p_Z be their respective densities. Because

$$\nabla(-\ln p_Z)(z) = \frac{-\nabla p_Z(z)}{p_Z(z)} = \mathbb{E}_{X\sim p_X}[p_Y(z-X)\cdot\nabla(-\ln p_Y(z-X))]\cdot\frac{1}{p_Z(z)} = \mathbb{E}[\nabla(-\ln p_Y)(Y)|X+Y=z], \tag{16}$$

we can compute the Hessian with the product rule,

$$\begin{aligned}
\nabla^2(-\ln p_Z)(z) &= \nabla\left\{\mathbb{E}_{X\sim p_X}[p_Y(z-X)\cdot\nabla(-\ln p_Y(z-X))]\cdot\frac{1}{p_Z(z)}\right\} && (17)\\
&= -\mathbb{E}_{X\sim p_X}\left[p_Y(z-X)\nabla\ln p_Y(z-X)(\nabla\ln p_Y(z-X))^\top\right]\cdot\frac{1}{p_Z(z)} && (18)\\
&+ \mathbb{E}_{X\sim p_X}[p_Y(z-X)\nabla^2(-\ln p_Y(z-X))]\cdot\frac{1}{p_Z(z)} && (19)\\
&+ \mathbb{E}_{X\sim p_X}[p_Y(z-X)\nabla(\ln p_Y(z-X))]\cdot\frac{1}{p_Z(z)}\cdot\frac{\nabla p_Z(z)}{p_Z(z)} && (20)\\
&= -\mathbb{E}[\nabla\ln p_Y(Y)(\nabla\ln p_Y(Y))^\top|X+Y=z] + \mathbb{E}[\nabla^2(-\ln p_Y(Y))|X+Y=z] && (21)\\
&+ (\mathbb{E}[\nabla\ln p_Y(Y)|X+Y=z])^{\otimes 2} && (22)\\
&= -\mathrm{Var}(\nabla(-\ln p_Y(Y))|X+Y=z) + \mathbb{E}[\nabla^2(-\ln p_Y(Y))|X+Y=z] && (23)\\
&\preceq \frac{1}{1-e^{-2t}}\mathrm{Id}, && (24)
\end{aligned}$$

where the last line uses $\mathrm{Var}(\nabla(-\ln p_Y(Y))|X+Y=z) \succeq 0$ and $\mathbb{E}[\nabla^2(-\ln p_Y(Y))|X+Y=z] = \frac{1}{1-e^{-2t}}\mathrm{Id}$. □

Lemma B.3. *For $t > 0.001$, we have the following inequality on the score at the origin,*

$$\|\nabla\ln p_t^i(0)\| \lesssim e^{-t}\left[\|\mu_i\| + M\right]. \tag{25}$$

Proof. By the definition of a convolution, we can explicitly compute

$$\nabla\ln p_t^i(0) = \frac{-\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})\frac{0-ue^{-t}}{1-e^{-2t}}du}{\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})du} = \frac{e^{-t}}{1-e^{-2t}}\frac{\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})udu}{\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})du}. \tag{26}$$

Note that for all $t \geq 0.001$, $f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}$ is Ω-Lipschitz for some $\Omega > 0$. Thus, we can bound the distance between the numerator and $\mu_i f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)$ with the triangle inequality and Assumption 4.3,

$$\begin{aligned}
&\left\|\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})udu - \mu_i f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)\right\| && (27)\\
&\leq \int_{\mathbb{R}^d} p_0^i(u)\left\|f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t}) - f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)\right\|\cdot\|u\|du && (28)\\
&\leq \Omega e^{-t}\int_{\mathbb{R}^d} p_0^i(u)\|u\|^2 du \leq \Omega e^{-t}M. && (29)
\end{aligned}$$

The denominator also approaches $f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)$ at the rate of $O(e^{-t})$, and we can express a bound on the distance from $f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)$ in terms of M using Jensen's inequality,

$$\left\|\int_{\mathbb{R}^d} p_0^i(u)\, f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(-ue^{-t})du - f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)\right\| \leq e^{-t}\int_{\mathbb{R}^d} p_0^i(u)\|u\|du \leq e^{-t}M. \tag{30}$$

Thus there exists $0 \le \epsilon_1, \epsilon_2 \le \max(\Omega, 1)$ and $w \in \mathbb{S}^{d-1}$ such that for all $t \ge 0.001$, we have the score bound

$$\left\|\nabla \ln p_t^i(0)\right\| = \frac{e^{-t}}{1-e^{-2t}} \left\| \frac{f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0)\mu_i + \Omega e^{-t} M \epsilon_1 w}{f_{\mathcal{N}(0,(1-e^{-2t})\mathrm{Id})}(0) + e^{-t} M \epsilon_2} \right\| \lesssim e^{-t}\left[\|\mu_i\| + M\right]. \tag{31}$$

□

Lemma B.4. *Under Assumptions 4.1, 4.3, and 4.4,* $\mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim e^{-4t}(\overline{R}^4 + M^4 + M\Psi^8 + \overline{M})\ \ \forall i, j, \ell \in [K]$.

Proof. For $t < 0.001$, we can prove the lemma by directly appealing to the bounded fourth moments of the scores $\nabla \ln q_t(X), \nabla \ln p_t(X)$ by Assumption 4.4,

$$\mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim \mathbb{E}_{X\sim p^i}\left[\|\nabla \ln p^j(X)\|^4 + \|\nabla \ln p^\ell(X)\|^4\right] \lesssim \overline{M}. \tag{32}$$

For $t \ge 0.001$, it suffices to bound the difference with the scores of the standard normal by the triangle inequality,

$$\mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim \mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln f_{\mathcal{N}(0,\mathrm{Id})}(X)\|^4 \tag{33}$$

$$+ \mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^\ell(X) - \nabla \ln f_{\mathcal{N}(0,\mathrm{Id})}(X)\|^4 \tag{34}$$

Both terms with p_t^j, p_t^ℓ are controlled by the same procedure. For j, we can write

$$\mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln f_{\mathcal{N}(0,\mathrm{Id})}(X)\|^4 \lesssim \mathbb{E}_{X\sim p_t^i}\|(\nabla(\ln p_t^j - \ln f_{\mathcal{N}(0,\mathrm{Id})}))(X) - \nabla(\ln p_t^j - \ln f_{\mathcal{N}(0,\mathrm{Id})})(0)\|^4 + \|\nabla \ln p_t^j(0)\|^4.$$

By Lemma B.2,$\nabla^2(-\ln p_t^j + \ln f_{\mathcal{N}(0,\mathrm{Id})})$'s eigenvalues are in $[\frac{e^{-2t}(1-\Psi^2)}{e^{-2t}\Psi^2+1-e^{-2t}}, \frac{e^{-2t}}{1-e^{2t}}] \subset [-\Psi^2 e^{-2t}, 1000e^{-2t}]$. Thus $\nabla \ln p_t^j - \ln f_{\mathcal{N}(0,\mathrm{Id})}$ is globally $1000\Psi^2 e^{-2t}$-Lipschitz. Combining with Lemma B.3, we can conclude

$$\begin{aligned}\mathbb{E}_{X\sim p_t^i}\|\nabla \ln p_t^j(X) - \nabla \ln f_{\mathcal{N}(0,\mathrm{Id})}(X)\|^4 &\lesssim e^{-8t}\left[\mathbb{E}_{X\sim p_t^i}\Psi^8\|X\|^4\right] + e^{-4t}\left[\|\mu_i\|^4 + M^4\right] \\ &\lesssim e^{-4t}\left[\|\mu_j\|^4 + M^4 + M\Psi^8\right]\end{aligned}$$

□

B.3. Proof of Lemma 4.7

Lemma B.5. $\left\|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t\right\|^2 = \left\|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}\right\|^2 \cdot \left(\frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i}\right)^2$.

Proof. This follows by some simple algebraic manipulations:

$$\left\|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t\right\|^2 = \left\| \frac{\sum_{i\in S_{\mathrm{targ}}} w_i \nabla p_t^i}{\sum_{i\in S_{\mathrm{targ}}} w_i p_t^i} - \frac{\sum_{i\in[K]} w_i \nabla p_t^i}{\sum_{i\in[K]} w_i p_t^i} \right\|^2 \tag{35}$$

$$= \left\| \left(\frac{1}{\sum_{i\in S_{\mathrm{targ}}} w_i p_t^i} - \frac{1}{\sum_{i\in[K]} w_i p_t^i} \right) \sum_{i\in S_{\mathrm{targ}}} w_i \nabla p_t^i - \frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i \nabla p_t^i}{\sum_{i\in[K]} w_i p_t^i} \right\|^2 \tag{36}$$

$$= \left\| \left(\frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i}{\left(\sum_{i\in S_{\mathrm{targ}}} w_i p_t^i\right)\left(\sum_{i\in[K]} w_i p_t^i\right)} \right) \sum_{i\in S_{\mathrm{targ}}} w_i \nabla p_t^i - \frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i \nabla p_t^i}{\sum_{i\in[K]} w_i p_t^i} \right\|^2 \tag{37}$$

$$= \left(\frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i} \right)^2 \left(\left\| \frac{\sum_{i\in S_{\mathrm{targ}}} w_i \nabla p_t^i}{\sum_{i\in S_{\mathrm{targ}}} w_i p_t^i} - \frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i \nabla p_t^i}{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i} \right\|^2 \right) \tag{38}$$

$$= \left(\frac{\sum_{i\in[K]-S_{\mathrm{targ}}} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i} \right)^2 \left\|\nabla \ln p_t^{S_{\mathrm{targ}}} - \nabla \ln p_t^{[K]-S_{\mathrm{targ}}}\right\|^2.$$

□

B.4. Proof of Lemma 5.1

Lemma B.6. *Consider mixture $P = \sum_i a_i P_i$ and mixture $Q = \sum_i b_i Q_i$. If $\mathrm{TV}(P_i, Q_j) \le \epsilon$ for all i, j, then*

$$\mathrm{TV}(P, Q) \le \epsilon.$$

Proof. This is a simple application of triangle inequality,

$$\frac{1}{2}\int |\sum_i a_i dP_i - \sum_j b_j dQ_j| \le \frac{1}{2}\sum_i a_i \int |dP_i - \sum_j b_j dQ_j| \le \frac{1}{2}\sum_i a_i \sum_j b_j \int |dP_i - dQ_j| \le \epsilon. \tag{39}$$

□

Lemma B.7. *(Short-time regularization) Convolving with the normal distribution bounds* KL *in terms of W_2,*

$$\mathrm{KL}(p * \mathcal{N}(0, \sigma^2) || q * \mathcal{N}(0, \sigma^2)) \le \frac{1}{2\sigma^2} W_2(p, q)^2$$

Proof. By the joint convexity of KL, it suffices to show for $p = \delta_x$ and $q = \delta_y$. Then,

$$\mathrm{KL}(\mathcal{N}(x, \sigma^2) || \mathcal{N}(y, \sigma^2)) = \frac{\|x - y\|^2}{2\sigma^2}. \tag{40}$$

□

Lemma B.8. *Let $\epsilon > 0$. For $i \in [K]$, let $\overline{p}^i$ denote the density of the i-th component of the mixture model p after being shifted to have mean zero. Suppose $\mathrm{W}_2(\overline{p}^i, \overline{p}^j) \le \Upsilon$ for all $i, j \in [K]$. Then $T_{\mathrm{lower}}(\epsilon) \le \left\{\ln(w(S_{\mathrm{init}}, S_{\mathrm{targ}}) + \Upsilon) + \ln\frac{1}{\epsilon} + \frac{1}{2}\ln 2\right\} \vee 3$. Additionally, if $p_0^i \in \mathrm{subG}_d(\sigma^2)$ for all $i \in [K]$, then $T_{\mathrm{upper}}(\epsilon) \ge \ln \Delta(S_{\mathrm{targ}}) - \ln\sigma - \ln\sqrt{8d\ln 6 + 8\ln 4/\epsilon^2} - \ln 3 - \frac{1}{2}\ln 8$.*

Proof of bound on T_{lower} in Lemma 5.1. Define h_t^ℓ to be the density of $e^{-t}X_t^\ell$ for $\ell \in [K]$. We apply Pinsker's inequality and treat the convolution with Gaussian noise in the forward process as a regularization parameter to control KL in terms of the Wasserstein-2 distance. For $i \in S_{\mathrm{init}}$ and $j \in S_{\mathrm{targ}}$ we can control the KL via Lemma B.7,

$$\mathrm{TV}(p^i_{T_{\mathrm{lower}}}, p^j_{T_{\mathrm{lower}}}) \le \sqrt{\mathrm{KL}(p^i_{T_{\mathrm{lower}}} || p^j_{T_{\mathrm{lower}}})} \le \mathrm{W}_2(h^i_{T_{\mathrm{lower}}}, h^j_{T_{\mathrm{lower}}}). \tag{41}$$

We use a coupling argument to control $\mathrm{W}_2(h^i_{T_{\mathrm{lower}}}, h^j_{T_{\mathrm{lower}}})$. Let $\pi \in \Gamma(\overline{f}_0^i, \overline{f}_0^j)$ be the optimal coupling, and define the coupling in $\Gamma(p^i_{T_{\mathrm{lower}}}, p^j_{T_{\mathrm{lower}}})$ that samples $(X, Y) \sim \pi$ and returns $(e^{-T_{\mathrm{lower}}}(X + \mu_i), e^{-T_{\mathrm{lower}}}(Y + \mu_j))$. The cost of this coupling is

$$\mathrm{W}_2(h^i_{T_{\mathrm{lower}}}, h^j_{T_{\mathrm{lower}}}) \le \sqrt{\mathbb{E}\|e^{-T_{\mathrm{lower}}}(X - Y) + e^{-T_{\mathrm{lower}}}(\mu_i - \mu_j)\|^2} \le e^{-T_{\mathrm{lower}}}\sqrt{2\left(\mathbb{E}\|X - Y\|^2 + \|\mu_i - \mu_j\|^2\right)} \tag{42}$$

$$\le \sqrt{2}e^{-t}\left[\Upsilon + \|\mu_i - \mu_j\|\right] \tag{43}$$

Thus $\mathrm{TV}(p_t^i, p_t^j) \le \sqrt{2}\left[\|\mu_i - \mu_j\| + \Upsilon\right]e^{-t} \le \epsilon$, and we can conclude by applying Lemma B.6 to obtain an overall bound on $\mathrm{TV}(p^{S_{\mathrm{init}}}_{T_{\mathrm{lower}}}, p^{S_{\mathrm{targ}}}_{T_{\mathrm{lower}}})$. □

Lemma B.9. *Consider sub-Gaussian random vectors $\{X_i\}_{i=1}^n$ in $\mathbb{R}^d$ with variance proxies $\{\sigma_i^2\}_{i=1}^n$. Let $S = \sum_{i=1}^n \alpha_i X_i$. Then, $S \in \mathrm{subG}_d(\sum_{i=1}^n \alpha_i^2\sigma_i^2)$.*

Proof. This proof is trivial. □

Lemma B.10. *(Theorem 1.19 of (Rigollet & Hutter, 2023)) Let $X \in \mathrm{subG}_d(\sigma^2)$. Then, for any $t \ge 0$,*

$$\mathbb{P}[\|X\| > t] \le 6^d \exp(-t^2/(8\sigma^2)). \tag{44}$$

Lemma B.11. *Consider two random vectors $X, Y \in \mathbb{R}^d$ with probability density functions P_X, P_Y and means μ_X, μ_Y such that $X - \mu_X$ and $Y - \mu_Y$ are sub-Gaussian random vectors with variance proxy σ^2. Let $R = \sigma\sqrt{8d \ln 6 + 8 \ln 1/\epsilon}$. If $\|\mu_X - \mu_Y\| > 2R$ then*

$$\mathrm{TV}(X, Y) \geq 1 - \epsilon$$

Proof. By B.10, $\mathbb{P}(\|X - \mu_X\| \geq R), \mathbb{P}(\|Y - \mu_Y\| \geq R) \leq \epsilon$, and $B_{\leq R}(\mu_X)$ and $B_{\leq R}(\mu_Y)$ are disjoint by definition. Thus,

$$\mathrm{TV}(X, Y) = \frac{1}{2}\int_{\mathbb{R}^d} |dP_X - dP_Y| \geq \frac{1}{2}\int_{B_{\leq R}(\mu_X)} dP_X - dP_Y + \frac{1}{2}\int_{B_{\leq R}(\mu_Y)} dP_Y - dP_X \geq 1 - \epsilon. \tag{45}$$

□

Proof of bound on T_{upper} in Lemma 5.1. By Lemma B.9, p_t^i is sub-Gaussian with variance proxy $2\sigma^2$ for all $t \geq 0$. For $i \in S_{\text{targ}}, j \in [K] - S_{\text{targ}}$, $\|\mu_t^i - \mu_t^j\| > 3\sigma\sqrt{8d \ln 6 + 8 \ln 4/\epsilon^2}$ implies $\mathrm{TV}(p_t^i, p_t^j) \geq 1 - \epsilon^2/4$ by Lemma B.11. □

B.5. Score difference bound for Gaussian mixtures

Here we prove the following key ingredient in the proof of Theorem 5.3, in analogy to Lemma 4.6 in the proof of the master theorem:

Lemma B.12. *For any nonempty $S \subset [K]$ and $j \in S$, we have*

$$\mathbb{E}_{x \sim p_t^j}\left[\left\|\nabla \ln p_t^S - \nabla \ln p_t^{[K]-S}\right\|^4\right] \lesssim \frac{e^{-4t}}{\underline{\lambda}^4}\left[\left(\overline{\lambda} - \underline{\lambda}\right)^4 (\overline{R}(0)^4 + \overline{\lambda}^2 d^2) + \overline{R}(0)^4\right]. \tag{46}$$

To prove this, we need an auxiliary result:

Lemma B.13. *Let $A, B \in \mathbb{R}^{d\times d}$ be two PSD matrices with singular values in $[\underline{\sigma}, \overline{\sigma}]$. For any $v \in \mathbb{R}^d$,*

$$\|(A - B)v\| \leq 2(\overline{\sigma} - \underline{\sigma})\|v\|.$$

Proof. We subtract both Av, Bv by $\underline{\sigma} I$ and apply the triangle inequality,

$$\|(A - B)v\| = \|(A - \underline{\sigma} I)v - (B - \underline{\sigma} I)v\| \leq \|(A - \underline{\sigma} I)v\| + \|(B - \underline{\sigma} I)v\| \leq 2(\overline{\sigma} - \underline{\sigma})\|v\|. \tag{47}$$

□

Proof of Lemma B.12. We explicitly compute $\nabla \ln p_t^S$ and $\nabla \ln p_t^{[K]-S}$ and their difference,

$$\nabla \ln p_t^S = \sum_{i\in S} \frac{w_i p_t^i}{\sum_{j\in S} w_j p_t^j}\left(-(\Sigma_i^t)^{-1}(x - \mu_i(t))\right) \tag{48}$$

$$\nabla \ln p_t^{[K]-S} = \sum_{i\in [K]-S} \frac{w_i p_t^i}{\sum_{j\in S} w_j p_t^j}\left(-(\Sigma_i^t)^{-1}(x - \mu_i(t))\right) \tag{49}$$

$$\nabla \ln p_t^S - \nabla \ln p_t^{[K]-S} = -\left(\sum_{i\in S} \frac{w_i p_t^i}{\sum_{j\in S} w_j p_t^j}(\Sigma_i^t)^{-1} - \sum_{i\in [K]-S} \frac{w_i p_t^i}{\sum_{j\in [K]-S} w_j p_t^j}(\Sigma_i^t)^{-1}\right) x \tag{50}$$

$$+ \left(\sum_{i\in S} \frac{w_i p_t^i}{\sum_{j\in S} w_j p_t^j}(\Sigma_i^t)^{-1}\mu_i(t) - \sum_{i\in [K]-S} \frac{w_i p_t^i}{\sum_{j\in [K]-S} w_j p_t^j}(\Sigma_i^t)^{-1}\mu_i(t)\right). \tag{51}$$

Both $\sum_{i\in S} \frac{w_i p_t^i}{\sum_{i\in S} w_i p_t^i}(\Sigma_i^t)^{-1}, \sum_{i\in [K]-S} \frac{w_i p_t^i}{\sum_{i\in S} w_i p_t^i}(\Sigma_i^t)^{-1}$ are PSD matrices with singular values in $[1/\sigma_{\max}^2(t), 1/\sigma_{\min}^2(t)]$. Thus, by Lemma B.13, we can bound the first term in the difference in terms of the

norm of x,

$$\left\| \left(\sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} (\Sigma_i^t)^{-1} - \sum_{i \in [K]-S} \frac{w_i p_t^i}{\sum_{j \in [K]-S} w_j p_t^j} (\Sigma_i^t)^{-1} \right) x \right\| \leq \left(1/\sigma_{\min}^2(t) - 1/\sigma_{\max}^2(t)\right) \|x\|. \tag{52}$$

By the triangle inequality, we can bound the second term with the singular values as well,

$$\left\| \sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) - \sum_{i \in [K]-S} \frac{w_i p_t^i}{\sum_{j \in [K]-S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) \right\| \lesssim \overline{R}(t)/\sigma_{\min}^2(t). \tag{53}$$

We can decompose $\mathbb{E}_{x \sim p_t^{\{j\}}} \|x\|^4$ into $\overline{R}(t), \sigma_{\max}^2(t), d$ with the triangle inequality,

$$\mathbb{E}_{x \sim p_t^j} \|x\|^4 \lesssim \overline{R}(t)^4 + \sigma_{\max}^2(t)^2 \mathbb{E}_{x \sim p_t^j} \|\Sigma_i(t)^{-1/2}(x - \mu_i(t))\|^4 \lesssim \overline{R}(t)^4 + \sigma_{\max}^2(t)^2 d^2. \tag{54}$$

Combining these inequalities, we obtain

$$\mathbb{E}_{x \sim p_t^j} \left[\left\| \nabla \ln p_t^S - \nabla \ln p_t^{[K]-S} \right\|^4 \right] \lesssim \left(1/\sigma_{\min}^2(t) - 1/\sigma_{\max}^2(t)\right)^4 \left(\overline{R}(t)^4 + \sigma_{\max}^2(t)^2 d^2\right) + \overline{R}(t)^4/\sigma_{\min}^2(t)^4 \tag{55}$$

$$\leq \frac{e^{-4t}}{\underline{\lambda}^4} \left[\left(\overline{\lambda} - \underline{\lambda}\right)^4 \left(\overline{R}(0)^4 + \overline{\lambda}^2 d^2\right) + \overline{R}(0)^4 \right]. \tag{56}$$

□

B.6. Ratio bound for Gaussian mixtures

Here we prove the other key ingredient in the proof of Theorem 5.3, in analogy to Lemma 3.1 in the proof of the master theorem:

Lemma B.14. *For any $S \subset [K]$ and $j \in S$, we have*

$$\mathbb{E}_{x \sim p_t^j} \left(\frac{\sum_{i \in [K]-S} w_i p_t^i}{\sum_{i \in [K]} w_i p_t^i} \right)^4 \lesssim K \overline{W} \exp\left\{ -e^{-2t} \Delta(S_{\text{targ}})^2 / (8\overline{\lambda}) \right\}. \tag{57}$$

We will need the following helper lemmas:

Lemma B.15. *(p. 51 of (Pardo, 2005)) Let $P \sim \mathcal{N}(\mu_P, \Sigma_P)$ and $Q \sim \mathcal{N}(\mu_Q, \Sigma_Q)$. Then,*

$$\mathrm{H}^2(P, Q) = 2 - 2 \frac{|\Sigma_P|^{1/4} |\Sigma_Q|^{1/4}}{\left| \frac{\Sigma_P + \Sigma_Q}{2} \right|^{1/2}} \exp\left\{ -\frac{1}{8} (\mu_P - \mu_Q)^\top \left[\frac{\Sigma_P + \Sigma_Q}{2} \right]^{-1} (\mu_P - \mu_Q) \right\}.$$

Lemma B.16. *For positive semi-definite Σ_i, Σ_j, we have an AM-GM-style inequality for their determinants,*

$$|\Sigma_i| \cdot |\Sigma_j| \leq \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^2.$$

Proof. It suffices to show $1 \leq \left| \frac{1 + \Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2}}{2} \right| \cdot \left| \frac{1 + \Sigma_j^{-1/2} \Sigma_i \Sigma_j^{-1/2}}{2} \right|$. Both $(\Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2})^{-1} = \Sigma_i^{1/2} \Sigma_j^{-1} \Sigma_i^{1/2}$ and $\Sigma_j^{-1/2} \Sigma_i \Sigma_j^{-1/2}$ have the same spectrum and the same algebraic multiplicities. They are also positive semi-definite, which means the geometric multiplicities of their eigenvalues sum to d. Thus, we can conclude that both matrices have the same multiset of eigenvalues. Letting $\lambda_1, \lambda_2, \ldots, \lambda_d > 0$ be the eigenvalues of $(\Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2})^{-1}, \Sigma_j^{-1/2} \Sigma_i \Sigma_j^{-1/2}$, the right-hand side can be bounded by

$$\left| \frac{1 + \Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2}}{2} \right| \cdot \left| \frac{1 + \Sigma_j^{-1/2} \Sigma_i \Sigma_j^{-1/2}}{2} \right| \geq \prod_{i=1}^d \left(\frac{1 + 1/\lambda_i}{2} \right) \left(\frac{1 + \lambda_i}{2} \right) = \prod_{i=1}^d \frac{2 + 1/\lambda_i + \lambda_i}{4} \geq 1.$$

□

Proof of Lemma B.14. Because $\mathbb{E}_{x\sim p_t^j}\left(\frac{\sum_{i\in[K]-S} w_i p_t^i}{\sum_{i\in[K]} w_i p_t^i}\right)^4 \le \sum_{\ell\in[K]-S}\mathbb{E}_{x\sim p_t^j}\left[\frac{w_\ell p_t^\ell}{\sum_{i\in[K]} w_i p_t^i}\right]$, it suffices to bound $\mathbb{E}_{x\sim p_t^j}\left[\frac{w_\ell p_\ell^t}{w_\ell p_\ell^t + w_j p_t^j}\right]$ for any $\ell \in [K]-S$. Using the Hellinger distance bound in Lemma 3.1 and the computations in Lemmas B.15 and B.16, we have

$$\mathbb{E}_{x\sim p_t^j}\left[\frac{w_\ell p_\ell^t}{w_\ell p_\ell^t + w_j p_t^j}\right] \le \overline{W}\frac{|\Sigma_\ell(t)|^{1/4}|\Sigma_j(t)|^{1/4}}{\left|\frac{\Sigma_\ell(t)+\Sigma_j(t)}{2}\right|^{1/2}}\exp\left\{-\frac{e^{-2t}}{8}(\mu_\ell-\mu_j)^\top\left[\frac{\Sigma_\ell(t)+\Sigma_j(t)}{2}\right]^{-1}(\mu_\ell-\mu_j)\right\} \tag{58}$$

$$\lesssim \overline{W}\exp\left\{-e^{-2t}\Delta(S_{\text{targ}})/(8\overline{\lambda})\right\}. \tag{59}$$

□

B.7. Proof of Theorem 5.3

Theorem 5.3. *Take any $S_{\text{init}} \subset S_{\text{targ}} \subset [K]$. For sufficiently small ϵ, there exists T_{lower} and T_{upper} such that $T_{\text{lower}} \le \frac{1}{2}\ln\left(2d\frac{\overline{\lambda}-\underline{\lambda}}{\underline{\lambda}} + \frac{1}{\underline{\lambda}}w(S_{\text{init}},S_{\text{targ}})^2\right) + \ln\frac{1}{\epsilon}$ and also $T_{\text{upper}} \ge \ln\Delta(S_{\text{targ}}) + \frac{1}{2}\ln\underline{\lambda} - \ln 4 - \frac{1}{2}\ln\ln\left(\frac{\overline{\lambda}\sqrt{K\overline{W}}\left[(\overline{\lambda}-\underline{\lambda})^2(\overline{R}(0)^2+\overline{\lambda}d)+\overline{R}(0)^2\right]}{\underline{\lambda}^2\Delta(S_{\text{targ}})^2\epsilon^2}\right)$ and such that for any $\widehat{T}\in(T_{\text{lower}},T_{\text{upper}})$, $\text{TV}(p[S_{\text{init}}^{\langle\widehat{T}\rangle}], p^{S_{\text{targ}}}) \lesssim \epsilon$.*

Proof. As in the proof of Theorem 4.5, we apply the data processing inequality to obtain

$$\text{TV}(p[S_{\text{init}}^{\langle\widehat{T}\rangle}], p^{S_{\text{targ}}}) \le \text{TV}(P^{\leftarrow}[S_{\text{init}}^{\langle\widehat{T}\rangle}], P^{\leftarrow}[S_{\text{targ}}^{\langle\widehat{T}\rangle}]) + \text{TV}(P^{\leftarrow}[S_{\text{targ}}^{\langle\widehat{T}\rangle}], P_{\widehat{T}}^{\leftarrow,S_{\text{targ}}}). \tag{60}$$

We begin with $\text{TV}(p_{\widehat{T}}^{S_{\text{init}}}, p_{\widehat{T}}^{S_{\text{targ}}})$. By Lemma B.6, it suffices to show for any $i\in S_{\text{init}}, j\in S_{\text{targ}}$, $\text{TV}(p_{\widehat{T}}^i, p_{\widehat{T}}^j) \le \epsilon$. To control this quantity, we use Pinsker's inequality to write in terms of KL and the KL formula for two Gaussians, and further bound the determinant and trace in terms of $\underline{\lambda}, \overline{\lambda}$.

$$\text{TV}(p_{\widehat{T}}^i, p_{\widehat{T}}^j) \le \sqrt{\text{KL}(p_{\widehat{T}}^i, p_{\widehat{T}}^j)} \tag{61}$$

$$= \sqrt{\ln\frac{|\Sigma^j(\widehat{T})|}{|\Sigma^i(\widehat{T})|} + d\left[\frac{1}{d}\text{tr}(\Sigma_j^{-1}\Sigma_i) - 1\right] + (\mu_i(\widehat{T})-\mu_j(\widehat{T}))^\top\Sigma^j(\widehat{T})^{-1}(\mu_i(\widehat{T})-\mu_j(\widehat{T}))} \tag{62}$$

$$\le \sqrt{d\left[\ln\left(\frac{e^{-2\widehat{T}}\overline{\lambda}+1-e^{-2\widehat{T}}}{e^{-2\widehat{T}}\underline{\lambda}+1-e^{-2\widehat{T}}}\right) + \frac{e^{-2\widehat{T}}\overline{\lambda}+1-e^{-2\widehat{T}}}{e^{-2\widehat{T}}\underline{\lambda}+1-e^{-2\widehat{T}}} - 1\right] + \frac{1}{\underline{\lambda}}\|\mu_i-\mu_j\|^2 e^{-2\widehat{T}}} \tag{63}$$

We now use the inequality $\ln(x)\le x-1$ and note $\frac{e^{-2t}\overline{\lambda}+1-e^{-2t}}{e^{-2t}\underline{\lambda}+1-e^{-2t}} - 1 \le e^{-2t}\frac{\overline{\lambda}-\underline{\lambda}}{\underline{\lambda}}$,

$$\text{TV}(p_{\widehat{T}}^i, p_{\widehat{T}}^j) \le \sqrt{2e^{-2\widehat{T}}d(\overline{\lambda}-\underline{\lambda})/\underline{\lambda} + \frac{1}{\underline{\lambda}}\|\mu_i-\mu_j\|^2 e^{-2\widehat{T}}} \le \epsilon \tag{64}$$

Now we bound $\text{TV}(P^{\leftarrow}[S_{\text{targ}}^{\langle\widehat{T}\rangle}], P_{\widehat{T}}^{\leftarrow,S_{\text{targ}}})$. Following the main Cauchy-Schwarz split in Theorem 4.5, we can apply Lemmas B.12 and B.14 to control the score error for $t\in[0,\widehat{T}]$,

$$\mathbb{E}\left[\|\nabla\ln p_t^{S_{\text{targ}}}(\overline{X}_t^{S_{\text{targ}}}) - \nabla\ln p_t^{[K]}(\overline{X}_t^{S_{\text{targ}}})\|^2\right] \tag{65}$$

$$\lesssim e^{-2t}\frac{\sqrt{K\overline{W}}\left[(\overline{\lambda}-\underline{\lambda})^2(\overline{R}(0)^2+\overline{\lambda}d)+\overline{R}(0)^2\right]}{\underline{\lambda}^2}\exp\left\{-e^{-2t}\underline{\lambda}\Delta(S_{\text{targ}})^2/(16\overline{\lambda})\right\}. \tag{66}$$

The integral from 0 to $\widehat{T}$ is

$$\int_0^{\widehat{T}}\mathbb{E}\left[\|\nabla\ln p_t^{S_{\text{targ}}}(\overline{X}_t^{S_{\text{targ}}}) - \nabla\ln p_t^{[K]}(\overline{X}_t^{S_{\text{targ}}})\|^2\right]dt \tag{67}$$

$$\lesssim \frac{\sqrt{K\overline{W}\overline{\lambda}}\left[(\overline{\lambda}-\underline{\lambda})^2(\overline{R}(0)^2+\overline{\lambda}d)+\overline{R}(0)^2\right]}{\underline{\lambda}^2\Delta(S_{\text{targ}})^2}\exp\left\{-e^{-2T_{\text{upper}}}\Delta(S_{\text{targ}})^2/(16\overline{\lambda})\right\} \lesssim \epsilon^2. \tag{68}$$

□

B.8. Proof of Theorem 6.2

Theorem 6.2. *Let all $\Sigma^i = \mathrm{Id}$, $\|\mu_i\| = R$, and $w_i = \frac{1}{K}$. For $i \in [K]$, consider the path $u_1, u_2, u_3, \ldots, u_{H'}$ where u_1 is the leaf node with $i \in f(u_1)$ and $u_{H'}$ is the root. There exists $k \in [1, 2, \ldots, H']$, sufficiently large R, H', and sufficiently small ϵ such that there is a sequence of times $T_1 < T_2 < \cdots < T_k$ with $\mathrm{TV}(p[\{i\}^{\langle T_\ell \rangle}], p^{f(u_\ell)}) \lesssim \epsilon$.*

Proof. Using the notation from Example 1, let

$$T^j_{\text{lower}} = \ln w(f(u_j), f(u_{j+1})) + \ln 1/\epsilon \tag{69}$$

$$T^j_{\text{upper}} = \ln \Delta(f(u_{j+1})) - \ln 4 - \frac{1}{2} \ln \ln \frac{R^2}{\epsilon^2 \Delta(f(u_{j+1}))^2}. \tag{70}$$

It suffices to show that for a sufficiently large k, for all $j \le k$, we have both $T^j_{\text{upper}} - T^j_{\text{lower}} > 0$ and $T^{j+1}_{\text{lower}} - T^j_{\text{upper}} > 0$. By our definition of the mixture tree, we know

$$w(f(u_j), f(u_{j+1})) \in \left[(1-\delta)\frac{R}{2^{(H'-j)^2}}, (1+\delta)\frac{R}{2^{(H'-j)^2}}\right]$$
$$\Delta(f(u_{j+1})) \in \left[(1-\delta)\frac{R}{2^{(H'-j-1)^2}}, (1+\delta)\frac{R}{2^{(H'-j-1)^2}}\right].$$

$T^{j+1}_{\text{lower}} - T^j_{\text{upper}} > 0$ follows from

$$T^{j+1}_{\text{lower}} = \ln\left[(1-\delta)\frac{R}{2^{(H'-j-1)^2}}\right] + \ln\frac{1}{\epsilon} \ge \ln\left[(1+\delta)\frac{R}{2^{(H'-j-1)^2}}\right] - \frac{1}{2}\ln\ln\frac{R^2}{\epsilon^2\Delta(f(u_{j+1}))^2} \ge T^j_{\text{upper}}.$$

for sufficiently small ϵ. We have $T^j_{\text{upper}} - T^j_{\text{lower}} > 0$ if

$$\ln\left[(1+\delta)\frac{R}{2^{(H'-j)^2}}\right] + \ln(1/\epsilon) \le \ln\left[(1-\delta)\frac{R}{2^{(H'-j-1)^2}}\right] - \ln 4 - \frac{1}{2}\ln\ln\left[\frac{R^2}{\epsilon^2\left((1-\delta)\frac{R}{2^{(H'-j-1)^2}}\right)^2}\right] \tag{71}$$

$$\ln\frac{1+\delta}{1-\delta} + \ln\frac{1}{\epsilon} + \ln 4 + \frac{1}{2}\ln\left[2(H'-j-1)^2\ln 2 - 2\ln(1-\delta)\epsilon\right] \le (2(H'-j)-1)\ln 2. \tag{72}$$

This is true for sufficiently small j and large H'. □

B.9. Critical windows experiment for Gaussians

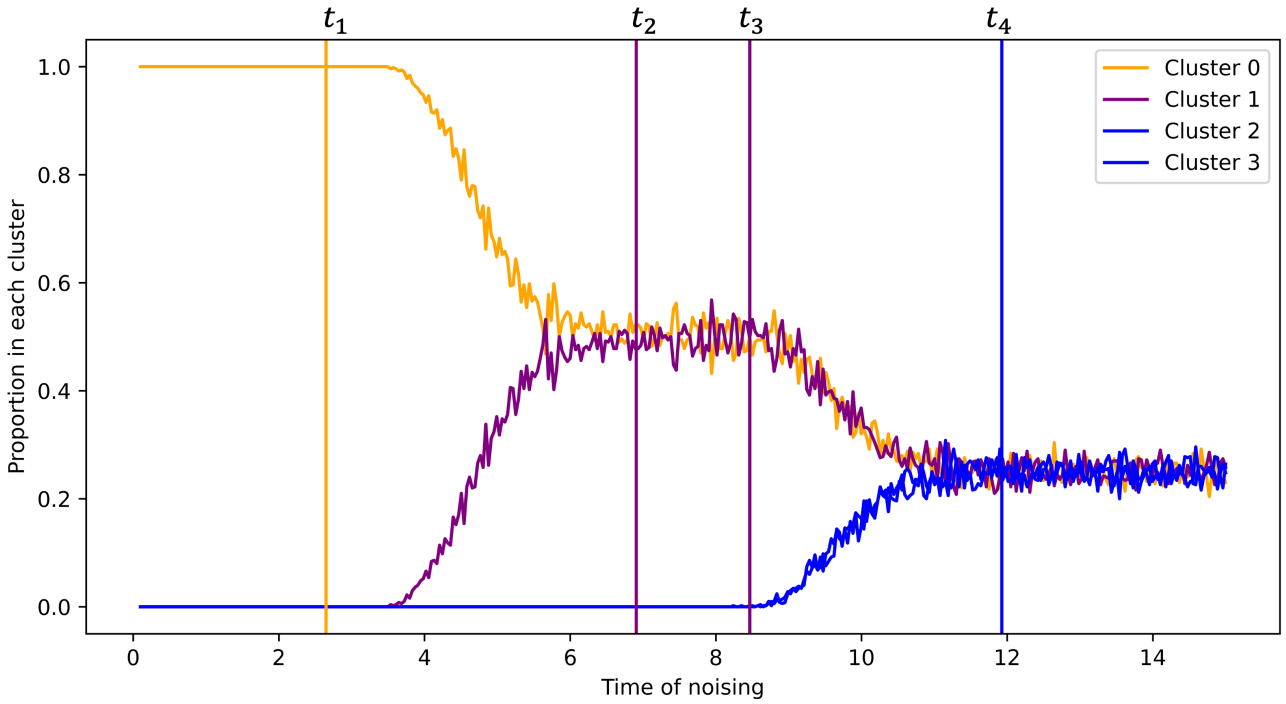

Figure 3. Example of critical times and proportion in each cluster as a function of noise timesteps. A point belongs to a cluster if its distance to the cluster mean is ≤ 5. All clusters have identity covariance; cluster 0 has mean (-15100); cluster 1 has mean (-14900); cluster 2 has mean (14900); cluster 3 has mean (15100). We compute the thresholds t_1, t_2, t_3, t_4 with $\epsilon = 0.1$ with the formulae from Example 1. By noising for $t \leq t_1$, we only sample from cluster 1. By noising for $t \in [t_2, t_3]$, we sample from clusters $0, 1$. By noising for $t \geq t_4$, we sample from all clusters.

C. Other concrete calculations for critical windows

Here we include two additional calculations, one for Gaussian mixtures with imbalanced mixing weights, and one in a dictionary learning setting, that were deferred to the appendix due to space constraints.

C.1. Dependence of $T_{\text{upper}}, T_{\text{lower}}$ on mixing weights

Consider the scenario of two Gaussians with identity covariance.

Example 2. *(Two Gaussians with identity covariance) Let $K = 2$, $p_0^1 = \mathcal{N}(\mu, \text{Id})$, $p_0^2 = \mathcal{N}(-\mu, \text{Id})$. Then, focusing on component 1 we have*

$$T_{\text{one}} = \ln \|\mu\| - \ln 2 - \frac{1}{2} \ln \ln \frac{\sqrt{2w_2/w_1}}{4\epsilon^2} \tag{73}$$

$$T_{\text{all}} = \ln \|\mu\| + \ln 2 + \ln 1/\epsilon \tag{74}$$

When $\widehat{T} \leq T_{\text{one}}$, then $\text{TV}(p[\{1\}^{\langle \widehat{T} \rangle}], p^{\{1\}}) \lesssim \epsilon$. When $\widehat{T} \geq T_{\text{all}}$, $\text{TV}(p[\{1\}^{\langle \widehat{T} \rangle}], p^{\{1,2\}}) \lesssim \epsilon$. We can see that as w_2 increases, the cutoff T_{one} becomes smaller, though the amount by which it decreases only scales at $O(\ln \ln w_2/w_1)$.

C.2. Sparse dictionary example

Now we consider a dictionary learning setting, in which classes are described by subsets of nearly-orthogonal feature vectors. Consider a set of $F = \{f_1, f_2, \ldots, f_n\}$ unit vectors, such that for all distinct i, j, $\text{cov}(f_i, f_j) \leq \delta$. Fix some large $R = \Omega(d)$. Consider the families of random variables $\mathcal{Y}_\ell = \{Y \in \mathbb{R}^\ell : \mathbb{E}[Y] = 0, Y \in \text{subG}_\ell(\sigma^2)\}$. We define scalar random variables $Y_{S,i} \in \mathcal{Y}_1$ for $S \subset F$ and $i \in [n]$, that represent the scaling along each feature vector, and $Y_S \in \mathcal{Y}_d$, which represents variation not along the features. Classes are subsets $S \subset F$ of cardinality $|S| \leq \tilde{S}$, such that a sample

$X \sim p_t^S$ has the distribution of $\sum_{i\in S}(Y_{S,i}+R)f_i + Y_S$. We let the Wasserstein-2 distance between any centered classes be less than Υ. We can characterize the $T_{\rm lower}, T_{\rm upper}$ in terms of the Hamming distances H between classes. We define $\overline{H}(S,S') := \max_{i\in S, j\in S'} H(i,j)$ and $\underline{H}(S) = \min_{\ell\in S, j\in[K]-S} H(i,j)$. By parameter setting with Corollary 5.1, we can write $T_{\rm lower}, T_{\rm upper}$ in terms of Hamming distances between classes.

Corollary C.1. *We have that* $T_{\rm lower}(\epsilon) \le 3 \vee \left\{\ln\frac{1}{\epsilon} + \frac{1}{2}\ln 2 + \ln(R\sqrt{\overline{H}(S_{\rm init}, S_{\rm end}) + d^2\delta} + \Upsilon)\right\}$ *and* $T_{\rm upper}(\epsilon) \ge \ln\left(R\sqrt{\underline{H}(S_{\rm end}) - d^2\delta}\right) - \ln(\sigma\sqrt{\tilde{S}+1}) - \ln\sqrt{8d\ln 6 + 8\ln 4/\epsilon^2} - \ln 3 - \frac{1}{2}\ln 8$.

Proof. We show that $\|\mu_i - \mu_j\|$ is only slightly differs from a constant factor from the Hamming distance,

$$\|\mu_i - \mu_j\|^2 = R^2\left\|\sum_{\ell\in i\setminus j} f_\ell - \sum_{\ell\in j\setminus i} f_\ell\right\|^2 \in \left[R^2(H(i,j) - d^2\delta), R^2(H(i,j) + d^2\delta)\right]$$

This completes $T_{\rm lower}$. For $T_{\rm upper}$, we also need to upper bound the variance proxies for each component. Letting $X \sim \sum_{i\in S} Y_{S,i} f_i + Y_S$, we can compute for all $u \in \mathbb{S}^{d-1}$ the expectation $\mathbb{E}[\exp(su^\top X)]$,

$$\begin{aligned}
\mathbb{E}[\exp(su^\top X)] = \mathbb{E}[\exp(su^\top X)] &= \mathbb{E}\left(\exp\left(su^\top Y_S\right)\right)\prod_{i\in S}\mathbb{E}\left(\exp\left(su^\top f_i Y_i\right)\right) \\
&\le \exp\left(s^2\sigma^2/2\right)\prod_{i\in S}\exp\left(s^2\sigma^2(u^\top f_i)^2/2\right) \\
&\le \exp\left(\frac{s^2\sigma^2(|S|+1)}{2}\right) \le \exp\left(\frac{s^2\sigma^2(\tilde{S}+1)}{2}\right).
\end{aligned}$$

Thus $X \in \mathrm{subG}_d(\sigma(\tilde{S}+1))$. □

D. Critical windows in Stable Diffusion V2.1

Here we include experimental details for the car critical window experiment, which was moved to the appendix because of space constraints. We produced 250 images from SD2.1., using 500 time steps from the DDPM scheduler (Ho et al., 2020a) and the prompt "`Color splash wide photo of a car in the middle of empty street, detailed, highly realistic, brightly colored car, black and white background.`" (see Figure 4). We used the CLIP with the ViT-B/32 Transformer architecture to label our images (Radford et al., 2021) according to the subject matter of their background ("`car in a city/on a road/in a field`"), color intensity ("`black or white/pale colored/brightly colored car`"), and size ("`big/medium/small car`"). We chose the prompt with the largest dot product with the image according to CLIP as the feature label.

Figure 4. Example images of cars generated by SD2.1 that we subsequently noised and denoised to produce Figure 5.

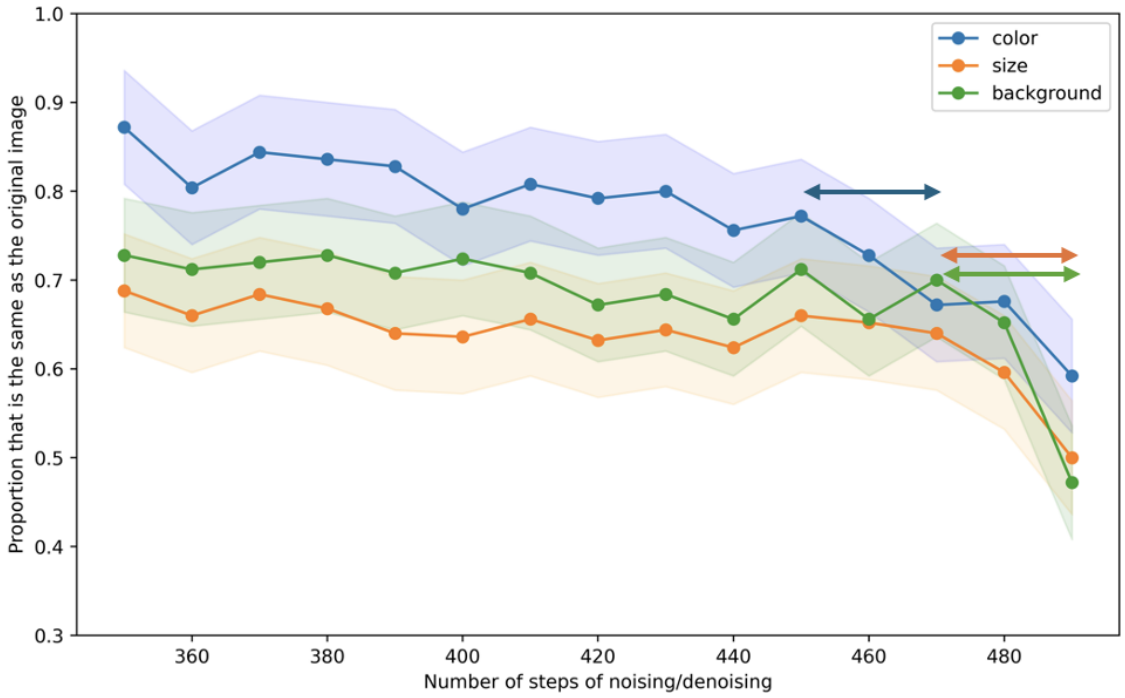

Figure 5. Percentage of agreement vs. noising amount in the experiment on images of cars generated by SD2.1 (see Section 7 for details). The critical window for each feature is demarcated with double-sided horizontal arrows.

E. Applications to Fairness and Privacy

Here we include more results from our fairness and privacy experiments, which were moved to the appendix because of space constraints.

E.1. Fairness

Figure 6. Example images generated by SD2.1 from the prompt "`Photo portrait of a laboratory technician`," that we subsequently noised and denoised for 100 timesteps to produce Figure 2.

E.2. Membership Inference Attack (MIA)

Method. Let Θ be the set of possible models and $\mathcal{X}$ be the set of possible inputs, herein the diffusion model and candidate image distribution, respectively. Let $\mathcal{D}_{\text{train}}$ be the training data and $\mathcal{D}$ be the distribution from which the training data was drawn. To evaluate a MIA, we sample with probability $\frac{1}{2}$ some $x \sim \mathcal{D}_{\text{train}}$ and otherwise sample $x \sim \mathcal{D}$. We rigorously describe our attack NoiseDenoise$(\mathcal{M}) : \Theta \times \mathcal{X} \to \mathbb{R}$. For a diffusion model $\theta \in \Theta$, let $q(x_t|x_{t-1})$ denote the T-step forward process and $p_\theta(x_{t-1}|x_t)$ denote the learned denoising process. Let $\underline{T} \in (0, T)$ denote the number of noising steps of our attack and N the number of samples of our attack. For $x \in \mathcal{X}$, we generate N samples $\tilde{x}^i_{\underline{T}} \sim q(x_{\underline{T}}|x_0 = x)$ and $\tilde{x}^i_0 \sim p_\theta(x_0|x_T = \tilde{x}^i_{\underline{T}})$ for $i \in [N]$. Our attack is the average L2 difference between $\tilde{x}^i_0$ and x for $i \in [N]$, and we predict x to belong to the training data if $\mathcal{M}(\theta, x) \leq \tau$,

$$\mathcal{M}(\theta, x) = \frac{1}{N} \sum_{i \in [N]} \|\tilde{x}^i_0 - x\|_2. \tag{75}$$

Note that this method has already demonstrated some promising results in identifying whether an image was generated by a diffusion model (Li & Wang, 2023). We present a conceptual explanation of our attack as follows. A diffusion model θ implicitly defines a pushforward distribution $\theta_*\gamma^d$ on images. For a candidate image x, we can view $\theta_*\gamma^d$ as a mixture of a ball around x, i.e. some $B_R(x)$ with $R > 0$, and the remainder of the distribution. Within a ball $B_R(x)$, we expect diffusion models to typically place more of the mass close to x when $x \in \mathcal{D}_{\text{train}}$ because training data have smaller losses. Thus we have greater separation from the remainder of the distribution for training data, and based on our theoretical framework, we can noise and denoise $x \in \mathcal{D}_{\text{train}}$ for more time steps than $x \notin \mathcal{D}_{\text{train}}$ and obtain samples close to x.

Our justification is similar to the logic characterizing diffusion model memorization in the independent and concurrent work of Biroli et al. (2024). (Biroli et al., 2024) considers the volume of neighborhoods around training data to identify critical times in their "collapse" regime, while we relate the size of these neighborhoods to our critical window theorems and develop these intuitions into a MIA. Additionally, this technique can be viewed as the diffusion model analogue of language model methods which perturb the inputs as part of MIAs (Li et al., 2023b) or machine-generated text detection (Mitchell et al., 2023).

Resutls. We tested our attack on a DDPM that was trained on CIFAR-10 in (Duan et al., 2023) and we compare it to their methods $\text{SecMI}_{\text{stat}}$ and SecMI_{nn}. Both their attacks exploit a deterministic approximation of the forward and reverse

process of a DDPM to estimate the sampling error of a candidate image. $\text{SecMI}_{\text{stat}}$ is the error itself while SecMI_{nn} is a neural network trained on the errors at different timesteps. We set $N = 10$ and $\hat{T} = 50$ (with $T = 100$), and compare all methods with 1000 training data samples and 1000 CIFAR-10 held-out samples. As in (Duan et al., 2023), we present ROC curves, AUC statistics, and TPRs at low FPRs of all MIAs, see Figure 7 and Table 1 in Appendix E.2.

As in (Duan et al., 2023), we present ROC curves, AUC statistics, and TPRs at low FPRs of all MIAs, see Figure 7 and Table 1. Both Figure 7 and Table 1 show that $\text{SecMI}_{\text{stat}}$ and SecMI_{nn} outperform NoiseDenoise. However, 11 of 23 of the train points NoiseDenoise identifies at FPR $= 0.01$ and 21 of 140 of the train points identified at FPR $= 0.05$ are not classified correctly by $\text{SecMI}_{\text{stat}}$ or SecMI_{nn} at the same FPR thresholds, suggesting NoiseDenoise can serve as a *complementary approach* to these methods.

Method	AUC	$\text{TPR}_{.01}$	$\text{TPR}_{.05}$
NoiseDenoise	.6636	.023	.14
$\text{SecMI}_{\text{stat}}$	.8847	.073	.344
SecMI_{nn}	.9132	.245	.609

Table 1. For each attack, we report the AUC, TPR at FPR .01, and TPR at FPR .05.

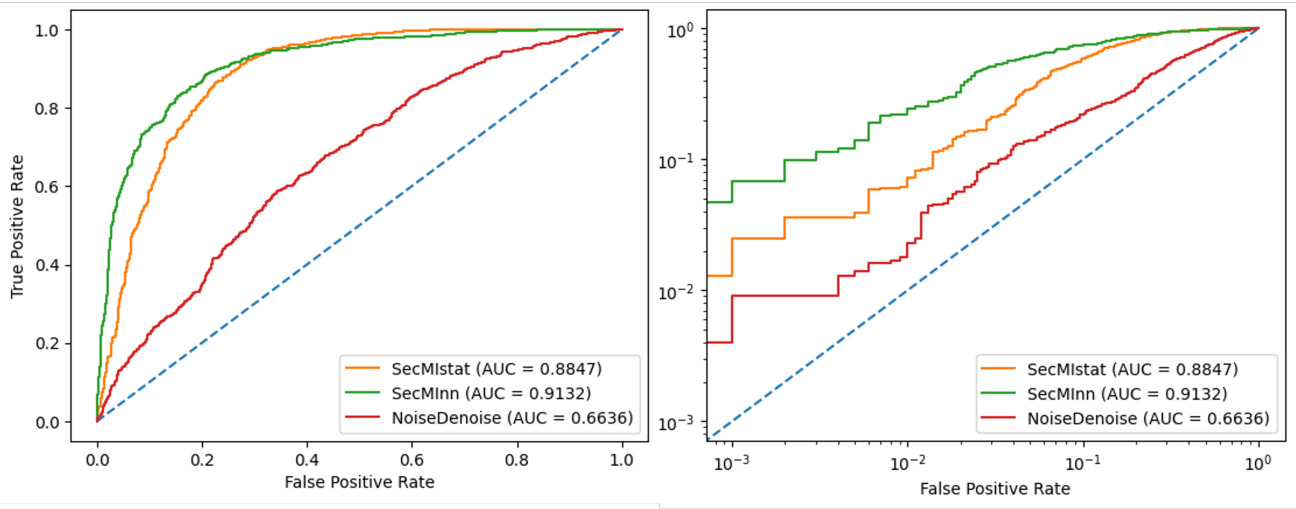

Figure 7. ROC curves of different methods.