

In the Blink of an Eye: A Unified Theory for Feature Emergence in Generative Models

A THESIS PRESENTED
BY
MARVIN LI
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
COMPUTER SCIENCE AND MATH

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2025

© 2025 - MARVIN LI
ALL RIGHTS RESERVED.

In the Blink of an Eye: A Unified Theory for Feature Emergence in Generative Models

ABSTRACT

Generative models, which produce samples of data such as text or images, are transforming the way we interact with technology. However, they often fail quickly in problematic and unintuitive ways. For example, a language model given a software engineering problem suddenly switched from coding to searching for pictures of Yellowstone National Park, and these rapid shifts in behavior have been observed in reasoning traces and hacks. This phenomenon is not unique to language models: in image generation models, key features of the final output, like objects in the background or the color, are also decided in narrow “critical windows” of the generation process.

While critical windows for a particular type of image generation model called diffusion have been studied at length by statistical physicists, existing theory relies on the specifics of diffusion and strong assumptions on the distribution of model generations. In this thesis, we develop a unifying framework for critical windows that shows that they emerge generically when the sampler specializes to a sub-population of the distribution it models. Drawing on tools from information theory, machine learning, high-dimensional probability theory, and statistical physics, our theory improves upon previous work by using rigorous mathematical tools and is agnostic to the underlying model type or distribution, applying to both language models and diffusion. The key insight of our approach is to exploit the powerful formalism for generative models of *stochastic localization*, which has roots as a proof technique in probability theory.

Leveraging our consolidated theory for critical windows, we apply it to different examples of critical windows in theoretical and empirical contexts. We provide a novel interpretation of the all-or-nothing phase transition in statistical inference as a critical window and use our framework to explain different failure modes of language models. We finally validate our predictions empirically for real-world models, and demonstrate that critical windows have applications towards improving the safety, privacy, and fairness of generative models.

Contents

1	INTRODUCTION	1
1.1	Phase transitions and critical windows	3
1.2	Our contributions	4
2	BASICS OF PROBABILITY AND INFORMATION THEORY	7
2.1	Statistical divergences and metrics	8
2.2	Data processing inequality	9
3	INTUITION FOR CRITICAL WINDOWS	13
3.1	Other theories for critical windows	18
I	Diffusion	20
4	A THEORETICAL INTRODUCTION TO DIFFUSION	21
4.1	Stochastic calculus basics	21
4.2	Motivating and defining diffusion	24
5	LOG-CONCAVE MIXTURES FOR CONTINUOUS DIFFUSION	30
5.1	Forward-reverse experiment for diffusion	30
5.2	Main theorem for feature specialization	32
6	INSTANTIATING THEORY FOR DIFFERENT MODELS OF DATA	39
6.1	Estimating critical windows for general distributions	40
6.2	Mixtures of well-conditioned Gaussians	43
6.3	Experiments on diffusion models	50
II	Stochastic localization	56
7	INTRODUCTION TO STOCHASTIC LOCALIZATION	57
7.1	Bourgain, thin-shell, and Kannan, Lovász and Simonovits (KLS) conjectures	58
7.2	“Sampling, Diffusions, and Stochastic Localization”	62

7.3	Universal characterization of generators	65
8	CRITICAL WINDOWS FOR STOCHASTIC LOCALIZATION SAMPLERS	69
8.1	Forward-reverse experiment (redux)	72
8.2	Main theorem for feature specialization (redux)	74
9	INSTANTIATING THEORY FOR DIFFERENT SAMPLERS	80
9.1	Diffusion, autoregression, and all-or-nothing	82
9.2	Hierarchical sampling interpretation	95
9.3	Experiments on language models	98
10	CONCLUSION	104
10.1	Future work	105
A	MISCELLANEA	107
A.1	Special f -divergences	107
A.2	Proof of Lemma 2.1.1	108
A.3	Details about diffusion and generators	112
A.4	Interweaving different samplers	117
	NOTATION	120
	REFERENCES	132

Listing of figures

1.0.1 Examples of critical windows for different data modalities and samplers, including reasoning [AAB ⁺ 24, QPL ⁺ 24] and certain jailbreaks [Hai24b] for language models and image class for diffusion models [GVS ⁺ 23].	3
2.2.1 Data processing inequality of f -divergences.	11
3.0.1 Illustration of the definition of a critical window for image diffusion models which transitions from sampling from both cats and dogs ($S_{\text{before}} = \{\text{orange cats}, \text{brown cats}, \text{dogs}\}$) to just cats ($S_{\text{after}} = \{\text{orange cats}, \text{brown cats}\}$).	14
3.0.2 Forward-reverse experiment on an image of an orange cat. At a low-noise regime, this yields the same image; at a middle-noise regime, this yields a cat, but of a different color; at a high-noise regime, this yields a dog. Orange cat courtesy of the Harvard Computer Society.	15
3.0.3 Example of a mixture tree.	18
4.2.1 Langevin MC often gets stuck in one mode, resulting in poor mixing.	26
4.2.2 Diagram of DDPMs from [HJA20a].	26
6.2.1 Example of critical times for an isotropic mixture of Gaussians ($\epsilon = 0.1$). For $t \leq t_1$, we sample from cluster 1. For $t \in [t_2, t_3]$, we sample from 0, 1. For $t \geq t_4$, we sample from 0, 1, 2, 3.	44
6.3.1 Example images of cars generated by SD2.1 that we subsequently noised and denoised to produce Figure 6.3.2.	51
6.3.2 Percentage of agreement vs. noising amount in the experiment on images of cars generated by SD2.1 (see Section 6.3.1 for details). The critical window for each feature is demarcated with double-sided horizontal arrows.	52
6.3.3 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 6.3.4.	52
6.3.4 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.	53

6.3.5 ROC curves of different methods.	55
7.1.1 Visualizations of the Bourgain slicing theorem, thin-shell conjecture, and the KLS conjecture from left to right, respectively.	59
7.2.1 Illustration of a discrete diffusion model with a masking forward procedure.	64
7.3.1 Examples of SDE and jump processes from [HY+25].	67
8.0.1 Difference between our approach in Chapter 5. Previously we bounded the path measures of the SDE defining p_{after}^S (blue) and p (red), and now we take a our destination-based approach. We bound directly bound the difference in the distributions of the destinations.	71
9.1.1 Examples of critical windows in jailbreaks for LLAMA-3.1-8B-Instruct. Left: Critical window for a prefill jailbreak [Hai24b]. Right: Critical window for a repeating token jailbreak [NCH+23].	85
9.1.2 Example of all-or-nothing phase transition from [RXZ19] for sparse linear regression, demarcated by a dramatic drop in minimum mean square error at $n = N^*$ samples.	93
9.3.1 Structured output plots for LLAMA-3.1-8B-Instruct. P denotes that we are sampling from responses whose answer to the first question was Pirate; P,C denotes that we are sampling from responses whose answers to the first two questions were Pirate and Cat, respectively. We can see that the critical windows directly correspond to our theoretical values for $T_{\text{before}}^C, T_{\text{after}}^C$	99
9.3.3 Probability that the answer is the same as a function of the percentage of the generation remaining for different math and reasoning benchmarks for LLAMA-3.1-8B-Instruct. Highlighted are generations with a 50% increase and no 30% decrease in the probability over subsequent fractions of sampling. We can see that the frequency of critical windows is especially high for the MATH benchmark.	100
9.3.2 Left-to-right: LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct barplots of original generation accuracy for generations with and without critical windows. Across all models we tested and benchmarks we evaluated, generations with critical windows have a lower probability of being correct.	101
9.3.4 Example of critical window for Phi-3-7B-Instruct (bolded).	102
9.3.5 Difference between resampling to obtain the answer versus directly asking the model for the answer (as in [LCR+23]) for LLAMA-3.1-8B-Instruct.	103
A.3.1 Diagram of Markov chain and the corresponding transition kernel.	115

TO MY FAMILY AND FRIENDS, WHO FILL MY LIFE WITH LOVE AND MEANING

Acknowledgments

First and foremost, I would like to thank my advisors, Sitan Chen and Cynthia Dwork, for their generous and unwavering guidance in supervising this thesis. Much of my growth as a researcher, thinker, and person during my undergraduate years is owed to their mentorship and kindness. It was an honor and inspiration to learn from such insightful and brilliant scientists, and I will fondly reminisce about these moments for the rest of my career. To Sitan Chen, thank you for your patience in answering all of my questions, kindness in supporting me when I was confused or had doubts, generosity in spending so much time mentoring and teaching me, and thoughtfulness in providing guidance both for our research projects but also for more general professional matters as well. I'm not exaggerating when I say that your mentorship has likely changed the course of my life. I still remember the awe I felt when I first encountered your paper, "Sampling is as Easy as Learning the Score," and I'm deeply grateful you met with me even before arriving at Harvard, and chose to work with me that fall. To Cynthia Dwork, thank you for teaching me to think deeply, ethically, and expansively about computer science. I am continually inspired by your perspective and your extraordinary ability to identify meaningful problems.

I am also deeply thankful to the other researchers who helped shape this thesis. I'm grateful to my labmates, Aayush Karan and Jaeyoon Kim, for their thoughtful feedback on early drafts. I owe special thanks to Aayush, who was a co-author on the paper that became Part II of this thesis and helped run some of the experimental work on reasoning in language models. I would like to thank Seth Neel for introducing me to machine learning research at Harvard. I've learned so much from you about what it means to be both a scientist and a collaborator, and I'm so thankful you took a chance on a few scrappy undergraduates.

I would also like to thank my readers Subhabrata Sen and David Alvarez-Melis. I also want to thank the many academics who shared helpful conversations and insights related to this project. Thank you to Michael Albergo and Max Simchowitz for thoughtful chats about this work and academia more broadly. I am especially grateful to Luca Ambrogioni for inviting me to present at the Donders Institute for Brain, Cognition, and Behaviour.

To the Harvard ML Foundations group and the Harvard Theory group—thank you for fostering such a vibrant intellectual community for me. I have grown so much from our meetings and seminars. This thesis can be read as the marriage of these perspectives, and I greatly appreciate both opportunities to learn so much.

I am also grateful to my many friends, who have kept me steady and sane throughout the

writing of this thesis and my college years. I thank my adopted blocking group, Kevin Du, Gabe Sun, Simon Sun, Joshua Park, and William Shi, who fill my days, especially Saturday evenings, with so much fun and love. I also want to thank the many individual friendships that have meant the world to me: AJ LaMotta, Amy Liu, AnhPhu Nguyen, Danielle Paulson, Doris Yang, Evgeni Kayryakov, Hannah Zhou, Jason Tang, Jason Wang, Jeffrey Wang, Kenny Gu, Kerem Dayı, Rei Yatsuhashi, Rui Zhu Lee, Sahil Kuchlous, Teresa Lu-Romeo, and countless others. I would not have made it through college without you, and I am beyond lucky that I got to be a part of your lives in these past few years. A special thank you to Allison Moon, Amy Liu, Gabe Sun, Sahil Kuchlous, Lev Kruglyak, AJ LaMotta, and Lillian Sun for reading advanced drafts of this thesis.

Finally, I would like to thank my family for being so supportive and caring through the entire process. I would like to give special thanks to my mom, for always being there when I needed her, my dad, for giving support and advice, and my brother, for being my inspiration.

It is possible to invent a single machine which can be used to compute any computable sequence.

Alan Turing

The purpose of computation is insight, not numbers.

Richard Hamming

1

Introduction

Generative models are statistical models that produce samples from a probability distribution. Trained with enormous amounts of data and computation, they are poised to be one of the most important technologies of the 21st century, with profound social, cultural, and economic implications. Even now, language models are writing emails and poetry, video diffusion models are creating advertisements and short films, and protein folding models are synthesizing novel drugs and therapeutics [Ope23, Ant23, Ope24, MFP⁺24, TL24]. Despite the enormous strides in their capabilities, these models often break down instantaneously in surprising and unexpected ways that are often not well-understood. For example, a language model given a software engineering task suddenly became distracted from coding by browsing pictures of Yellowstone National Park [Ant24]. These spontaneous failures can also lead to worrying behaviors: another language model interrupted a student who was using it to study and told them, “You are a waste of time and resources” [Gem24]. More broadly, these rapid shifts in the model’s behavior appear systematically in many settings related to the capabilities and safety of language models.

Reasoning failures. [LLX⁺24, AAB⁺24] observed a similar phenomenon in the reasoning process of language models for math and science questions. For one question, the probability of getting the answer correct jumped by 80% within a single step of the sampler. This has motivated new methods to improve the performance of language models, as the state-of-the-art Phi-4 model was trained on synthetic data generated at these important steps [AAB⁺24].

Reward hacking. Researchers from OpenAI [Ope25] recently tasked a language model to write code to satisfy some tests, which it could either pass by manipulating the evaluations or correctly solving them. They found that when the language model stated, “let me try to skip some of the unit tests,” it became highly likely to engage in that deceptive behavior. Notably, this segment of text consisting of nine words led the model to resort to cheating.

Safety and jailbreaks. Language models are vulnerable to being jailbroken, or hacked, into be-

having in harmful ways, from expressing racial bias to providing helpful advice on building weapons of mass destruction [Ant25]. [QPL⁺24, ZW24, HXH24, LRL⁺23] observed that the effectiveness of many kinds of jailbreaks can be linked to their ability to affect the first few words generated by a language model. [QPL⁺24] translated this insight to develop different methods to make models more resilient to these jailbreaks.

These sudden transitions are not unique to language models. For image diffusion models, numerous authors have noticed *critical windows*, narrow intervals in the sampling procedure where an important aspect of the final generation emerges, for features such as the color, background, and objects in an image [HJA20a, MHS⁺22, CLS⁺22, RA23, GVS⁺23, SFLW24, SFW25, BBdBM24]. These critical windows have been applied to simplify the training of diffusion models [AI25], analyze the effectiveness of conditional sampling techniques [PVBM25, CGC⁺24], and develop new methodologies to interpret and control diffusion models. We will highlight two applications.

Copyright and privacy violations. There is growing concern about the privacy and copyright risks of diffusion models [AMO23, Ima23], which have been shown to memorize and regurgitate training data [CHN⁺23a]. This necessitates data attribution techniques to identify how given training examples contributed to a generation. Critical windows are highly convenient from a data attribution standpoint, as they allow one to identify the parts of the trajectory that led to some feature in the final image. For example, [GVS⁺23] identified a critical window where the image subject as a horse was solidified, and then targeted their attribution method at that critical window to link the emergence of the horse to examples from the training data.

Image editing. Consider a user who wants to use an image diffusion model to modify some aspect of an existing image—for example, to convert a hand-drawn image of a house into a photorealistic style with a diffusion model. SDEdit proposed rewinding the sampling before the critical window for style but after the critical window for the overall structure [MHS⁺22].

Critical windows are salient across many data modalities and model types with implications for the capabilities, safety, and interpretability of generative models (Figure 1.0.1). Despite their relevance towards many facets of generative modeling, where and how these critical windows appear, as well as their ubiquity in so many different contexts, remain a mystery for machine learning practitioners and theorists. In this thesis we develop a *theory of critical windows* that consolidates these different instantiations of critical windows. Our theory improves on the current literature for critical windows, which has largely adopted a statistical physics-based perspective, along several axes including simplicity, generality, and rigor. To contextualize our contributions, we now briefly overview existing approaches.

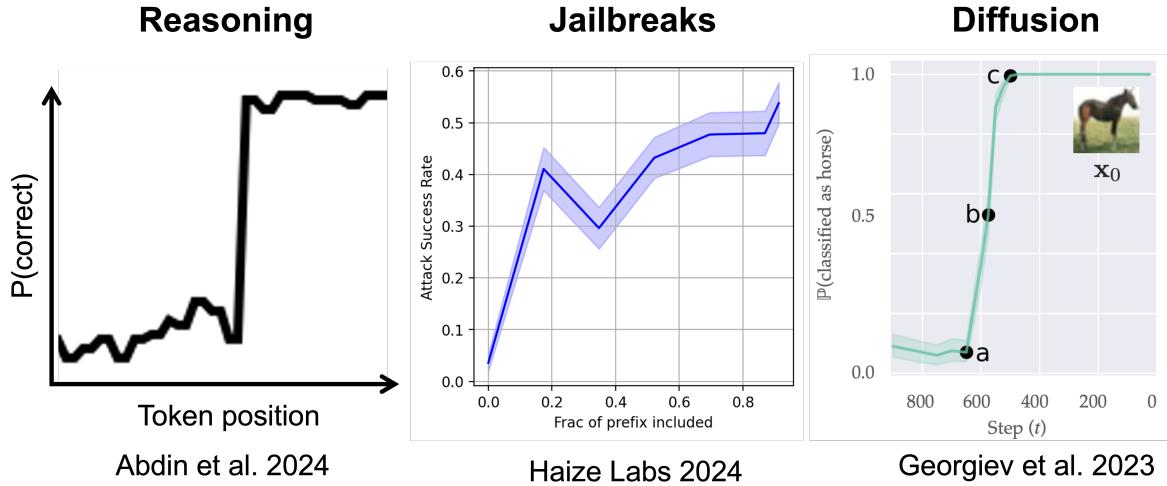


Figure 1.0.1: Examples of critical windows for different data modalities and samplers, including reasoning [AAB⁺24, QPL⁺24] and certain jailbreaks [Hai24b] for language models and image class for diffusion models [GVS⁺23].

1.1 Phase transitions and critical windows

Critical windows for diffusion models have received significant attention within the statistical physics community because it shares a phenomenology with the notion of phase transitions in physics [Amb23, RA23, SFW25, SFLW24, BBdBM24]. Phase transitions are processes in which some property of the system changes instantaneously, and, in the context of critical windows, refer to the sudden transformation in the model's behavior over a few steps in the sampler. In general, the statistical physics-based literature for critical windows assumes an *explicit functional form* for the distribution of generations from the model and then utilizes *accurate yet non-rigorous statistical physics methods* to compute the location of critical windows for that sampler.¹

While these works do provide predictive theory for critical windows in the diffusion setting, these theories are incomplete characterizations of critical windows and share the following limitations. Firstly, their computations rely on the particulars of diffusion samplers, meaning their theory cannot extend to language models. This means that these works do not provide an explanation of critical windows across different samplers. Secondly, their descriptions of critical windows are sensitive to the minutiae of their functional forms for data, which themselves are not realistic models of natural data like images or text, limiting the explanatory power of these theories in realistic scenarios. Thirdly, their methods largely consist of non-rigorous computations lacking provable guarantees. Certifications for behaviors apropos critical windows seem desirable in the safety-related scenarios we discussed earlier, like jailbreaks or reward hacking, where it would be beneficial to ensure that the model will not enter a dangerous mode. This thesis thus centers around the open question of lifting these weaknesses:

¹Section 3.1 includes a more thorough review of extant literature and comparison with our contributions.

Is there a simple, general, and rigorous theory that can explain critical windows for all generative modeling paradigms and models of data?

This seems implausible and too much to ask for *a priori*, but in the next section, we will show this is possible with the right abstractions and theoretical tools.

1.2 Our contributions

In this thesis, we develop an original theoretical framework for critical windows for a family of generative models, including language models and diffusion models. Taking a birds eye approach towards critical windows that combines ideas from information theory, machine learning, high-dimensional probability theory, and statistical physics, we establish where and how these critical windows occur, and explore the safety and abilities of generative models through this lens. First we explicate our framework and provide some general intuition into our findings.

Our framework for critical windows. Our main result is that these critical windows emerge generically when a generative model shifts from sampling a larger sub-population of the distribution it models to a smaller sub-population. We show that this is parameterized by the interval between (1) the last time that the larger and smaller sub-populations are indistinguishable and (2) the first time that smaller and larger sub-populations have negligible overlap between each other. As an example, consider a language model solving a math problem whose key insight is recalling a formula. Assuming that the model gets the answer correct if and only if it states the formula, the critical window where the model specializes to the correct answer is the interval where the model first states the formula. This means the following, for a fixed generation that ends with the correct answer: truncating the fixed generation before it states the formula and resampling with the model leads to the correct answer with low probability, while truncating at the first point the generation mentions the formula and resampling leads to the right answer with high probability.

1. Before the formula is mentioned in the truncation, the incorrect and correct generations in this scenario are *statistically indistinguishable* from each other because there is no other information that differentiates these two distributions.
2. After the model states the formula in the truncated version, the distributions of incorrect and correct generations have *negligible overlap*, because the presence of the formula completely determines the correctness of the solution and thus to which mode belongs the generation.²

These notions of statistical indistinguishability and negligible overlap are formalized with *information theory*. Importantly, this enables us to abstract the specifics about the generation distribution, breaking our dependence on specific functional forms in the statistical physics literature, and establish a general theory for critical windows. Moreover, these information-theoretic tools are rigorous and basic, in sharp contrast with the non-rigorous and complicated statistical physics computations

²While it is not immediate that the transition between these two points has to be narrow, we see that many instantiations of critical windows within our framework are small in a concrete sense.

in previous work. To unify our theory across different generative models, we appeal to the powerful formalism of *stochastic localization* for generative models [Mon23a, CE22].

Stochastic localization. Originating from the probability theory community as a proof technique, stochastic localization remarkably proved a conjecture in December 2024 open since the 1980s [Eld13, Eld20, KL24]. It has also led to a new sampling framework which encompasses most practical generative models [EAMS22, MW23, AMS23, Mon23b, HMP24, CE22, HHY⁺25]. Informally, a stochastic localization sampler is a generative model that reverses a Markovian “forward process.” This forward process takes a sample from the original distribution and progressively degrades it until it is independent of the original sample and has a simple distribution. The sampling algorithm thus starts by sampling the simple distribution and iteratively restores it until we obtain a sample from the original distribution.

This framework is flexible enough such that both diffusion and language models can be recovered as special cases. In image diffusion models, the degradation is a convolution with increasing amounts of Gaussian noise. The reversal of this forward procedure is learning to remove the Gaussian noise from the image, exactly the function that diffusion models learn. In language models, the degradation is the masking of “tokens” from right to left, and the reversal is predicting the next token. In this thesis, stochastic localization is the key to understand critical windows. Crucially, as our theory only relies on the Markovianity of the forward process, this allows us to write *one proof* that characterizes critical windows for all stochastic localization samplers.

To illustrate that our unifying framework has downstream theoretical and practical impacts, we explicitly compute critical windows for different samplers, distributions, and modalities. One such example proves a connection between critical windows and a phase transition in statistical inference. We also provide intuition for the different failure modes for language models that we mentioned earlier. Finally, we pair our theory with experiments to exhibit critical windows for real-world diffusion and language models.

In addition to our theoretical and empirical results, this thesis recounts the necessary technical background to fully grasp our contributions and tools. This exposition parallels the theme of our research, that the construction of the appropriate abstractions and frameworks has practical, not just aesthetic, value. The most important thread to trace is the evolution of generative modeling frameworks, from denoising and score-matching to diffusion and stochastic localization. These successive generalizations have yielded more efficient and capable models.

Structure of thesis. Our first waypoint is continuous diffusion models, where we established the first theoretical results on critical windows for a general family of high-dimensional distributions. Chapter 3 will give broad intuition for our critical windows framework and the claims we establish in both parts, combining exposition from [LC24, LKC25]. After we clarify our theory for diffusion, in Part II we expound a more general framework based on stochastic localization which subsumes the previous part.

Part I (Diffusion models). In this part, we present our original results on critical windows, which first appeared in the International Conference of Machine Learning 2024 [LC24] and at an invited talk at the Donders Institute for Brain, Cognition, and Behaviour [Li24]. In Chapter 4, we will introduce and motivate diffusion models as a unification of two previous paradigms. Some of the exposition on diffusion is drawn from [LC24], but it is mostly written for this thesis. In Chapter 5, we prove a key theorem that characterizes when the diffusion models sample from a sub-population, and in Chapter 6, we operationalize this theoretical framework for different distributions and conduct experiments on critical windows for real-world models. In particular, our framework contributes a new experimental method to probe the privacy properties of diffusion models.

Part II (Stochastic localization). In this part, we extend our theory to stochastic localization samplers, an earlier version of which was accepted at the International Conference of Learning Representations 2025 workshop on Frontiers in Probabilistic Inference [LKC25]. Chapter 7 provides a short history of the seminal contributions of stochastic localization to probability theory and generative modeling. This section will borrow some exposition from the author’s published works [LKC25] and covers some of the same key concepts as a final project the author did for Stat 216 with AJ LaMotta. In Chapter 8, we introduce our own theoretical framework for critical windows for localization-based samplers. With the improved abstractions developed in Chapter 8, Chapter 9 instantiates our theory in numerous settings. Not only do we more precisely compute critical windows from Part I, this change of view extends our theory for critical windows to many more models and distributions, explaining failure modes in language models and providing a novel interpretation of the all-or-nothing phase transition in statistical inference as a critical window. Empirically, we demonstrate that critical windows for language models correlate with failures on different math and reasoning benchmarks, concurrently with [AAB⁺24, LLX⁺24].

Our story begins in Chapter 2, where we introduce the probability notation and information theory tools that scaffold our work. It also serves as a vignette, to illustrate that *the right perspective* to model some intuition can oftentimes lead to a simple, beautiful characterization of it.

Information is the resolution of uncertainty.

Claude Shannon

2

Basics of Probability and Information Theory

We begin by introducing the basic notation that we will use throughout the thesis. We start with a base set Ω , which can be safely thought of as either a discrete set or a Euclidean space \mathbb{R}^d . This set Ω is the substrate on which our generative model produces samples; we can imagine Ω being the set of strings where language model generations live or coordinates where the value in each position is the color of a pixel generated by an image diffusion model. The generative models we consider are non-deterministic, which means we need a notion of *probability* on Ω .¹ To define a notion of probability on Ω requires an understanding of what can be *measured* or assigned probability, so we associate Ω with a corresponding σ -algebra $\mathcal{F} \subset 2^\Omega$ which characterizes which subsets of Ω can be measured. This gives us a measurable space (Ω, \mathcal{F}) where we can now meaningfully define notions of probability.

We will also want to understand how two probability distributions from the same generative model differ from each other. If we consider a generative model who produces images of cats or dogs, how do we *quantify* the differences between the distribution of cat images and the distribution of dog images? This question has a long history in the field of *information theory*, which studies the quantification and communication of information. Consider two friends, Alice and Bob, who are sharing messages with each other. Alice wants to send Bob an image of a cat or a dog but the transmission will corrupt the image. Bob's ability to distinguish the worlds where Alice sends a cat image or a dog image hinges on two factors: the difference between the original images of cats and dogs that Alice sends and the noisiness of the channel.

If the original images of cats and dogs Alice sends into the channel are already very grainy or

¹In fact, critical windows do not exist for deterministic models because their output is always the same.

damaged, Bob will not be able to recover the original animal no matter how hard he tries. Thus the notion of *accuracy* in his estimation is upper bounded by the difference between these distributions.

2.1 Statistical divergences and metrics

To crystallize this story into concrete mathematics, we start by defining *statistical divergences and metrics* which quantify the dissimilarity between probability measures. Let P, Q be two probability measures over (Ω, \mathcal{F}) that have densities dP, dQ with respect to a base measure μ .² We let $\text{supp}(P) = \{x \in \Omega | dP(x) > 0\}$ denote the support. One of the more natural ways to do this is the *total variation distance*, which is simply the difference between the densities of the probability measures.

Definition 2.1.1 (Total variation distance). The total variation (TV) distance is given by

$$\text{TV}(P, Q) \triangleq \frac{1}{2} \int |dP - dQ| d\mu.$$

If $P = Q$, then $\text{TV}(P, Q) = 0$, and if $\text{supp}(P) \cap \text{supp}(Q) = \emptyset$, then $\text{TV}(P, Q) = 1$. For random variables X, Y , we will also use $\text{TV}(X, Y)$ as shorthand to denote the TV of the distributions of X, Y . There are other alternative ways to characterize the difference between probability distributions that we will use later.

Definition 2.1.2 (Le Cam distance). The Le Cam (LC) distance is given by

$$\text{LC}(P, Q) \triangleq \frac{1}{2} \int \frac{(dP - dQ)^2}{d(P + Q)} d\mu.$$

Definition 2.1.3 (squared Hellinger distance). The squared Hellinger distance (H^2) is given by

$$H^2(P, Q) \triangleq \int (\sqrt{dP} - \sqrt{dQ})^2 d\mu.$$

These quantifications of difference between two measures have different applications. For example, the Le Cam distance is helpful to compare statistical models to each other, and the Hellinger distance simplifies greatly when the underlying data is a product measure, composed of independent samples from some distribution.³

Going back to our story, let's say the transmission is perfect, and Alice sends an image of a cat $x \sim P$ or a dog $x \sim Q$ with equal probability to Bob. Bob observes x , and by Bayes' rule,

²In the case of discrete spaces, this can be viewed as the counting measure, which maps each discrete set to its size. In \mathbb{R}^d , this is the Lebesgue measure, which naturally captures the notion of area or volume as one might expect. For example, the Lebesgue measure of a unit square in \mathbb{R}^2 is exactly 1. Herein we will adopt a *laissez-faire* attitude towards these mathematical technicalities like measurability.

³Appendix A.1.2 discusses this factorization property of Hellinger distance for product measures in the context of inference problems in-depth.

knows that it belongs to Q with probability $\frac{dQ(x)}{dP(x)+dQ(x)}$. Assuming Bob classifies x as from Q with exactly this probability, the probability he misclassifies $x \sim P$ is exactly $\mathbb{E}_{x \sim P} \left[\frac{dQ}{dP+dQ} \right]$. The below lemma shows that this error can be upper bounded by these statistical divergences and metrics.

Lemma 2.1.1. [LC24] For probability measures P, Q ,

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

The intuition behind the statement is very clear. If P, Q are very distinct from each other, $\text{TV}(P, Q) = 1$ or $\text{LC}(P, Q) = 1$ or $\text{H}^2(P, Q) = 2$, then one should have very low probability of misclassification. The proof is quite simple and relies on a few well-known inequalities between these divergences and simple linear algebra identities, so we leave the details to Appendix A.2. We will use additional facts and intuitions about statistical divergences and distributions in our thesis, but defer the statements and proofs to Appendix A.2.1.

2.2 Data processing inequality

Another factor that can affect Bob's ability to distinguish between $x \sim P$ and $x \sim Q$ is the channel itself. Intuitively, the channel adds no new information to the distribution and can only make Bob's job harder. In the language of statistical divergences and metrics, this means that the channel can only reduce the differences between probability measures P, Q . We can formalize this intuition with a data processing inequality, which states that the distance between two measures P, Q can only decrease if we pass samples from them through a common (random) function, which is known as a *channel*.⁴ The data processing inequality is a natural tool to study the stochastic localization samplers that were mentioned in the introduction because it captures the idea that two distributions can be progressively degraded through the forward process, and applies to a family of statistical divergences and metrics which include $\text{TV}, \text{LC}, \text{H}^2$ called f -divergences.

Let us first think critically about what kind of divergences would satisfy a data processing inequality. Importantly, observe that $\text{TV}, \text{LC}, \text{H}^2$ do not depend on any structure in the underlying base set Ω , which makes sense because a channel could shuffle the elements of Ω arbitrarily. One could imagine relabeling Ω and these statistical divergences and distances would remain the same. This is not shared by all notions of distance between distributions. For example, the Wasserstein-2 distance, the distance of the optimal transport between two probability distributions, necessitates that the base set $\Omega = \mathbb{R}^d$ is equipped with a metric. By default, we let $\|\cdot\|$ denote the ℓ^2 -metric.

⁴Exposition borrows from lecture notes by [Pol24, Wan24].

Definition 2.2.1 (Wasserstein-2 distance (W_2)). We define the Wasserstein-2 distance to be

$$W_2(P, Q) \triangleq \sqrt{\inf_{\gamma \sim \Gamma(P, Q)} \mathbb{E}_{(x, y) \sim \gamma} \|x - y\|^2},$$

where $\Gamma(P, Q)$ is the set of all couplings between P, Q , i.e. distributions over $\Omega^{\otimes 2}$ where the marginal distributions in the first and second coordinate are P, Q , respectively.

This distance is particularly useful to understand when a generative model successfully converges to a true distribution. The Wasserstein-2 distance does not satisfy the data processing inequality, as multiplying samples from P, Q by some constant factor greater than 1 would actually *increase* the W_2 distance. The main takeaway is that the data processing inequality can only apply to measures blind to the underlying Ω . Thus the only thing that an f -divergence can rely on is the relative densities at a given $x \in \Omega$, characterized by the likelihood ratio $\frac{dP}{dQ}$.⁵

Definition 2.2.2 (f -divergence). Let $f : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}$ be a convex function such that $f(1) = 0$. The functional D_f mapping probability measures P, Q to $\mathbb{R}^{\geq 0}$,

$$D_f(P, Q) \triangleq \mathbb{E}_Q \left[f \left(\frac{dP}{dQ} \right) \right],$$

is an f -divergence.^a

^aSee [Pol24] for the technical details of removing the requirement of the existence of the change of density.

It turns out most of the distances and divergences we considered so far belong to this general family. Another f -divergence that we will consider is the KL divergence, which can be thought of as the relative surprise of samples from $x \sim P$ when one erroneously models it as Q instead.

Example 2.2.1 (TV, LC, H^2 are f -divergences). For TV, take $f = \frac{1}{2} \left| 1 - \frac{dP}{dQ} \right|$. For LC, take $f = \frac{1 - \frac{dP}{dQ}}{2 \frac{dP}{dQ} + 2}$. For H^2 , take $f = \left(1 - \sqrt{\frac{dP}{dQ}} \right)^2$.

Definition 2.2.3 (Kullback-Liebler divergence). The KL divergence is an f -divergence given by

$$\text{KL}(P||Q) = \mathbb{E}_{x \sim P} \left[\ln \frac{dP}{dQ} \right].$$

Now we are ready to fully state the data processing inequality for f -divergences.

⁵It turns out the likelihood ratio is the “best test” for distinguishing samples from P, Q in certain inference setups, via the Neyman-Pearson lemma.

Theorem 2.2.1 (Data processing inequality for f -divergences). For probability measures P_X, Q_X and a common channel $C_{Y|X}$ which produces the measures $P_Y \triangleq (C_{Y|X})_# P_X$ and $Q_Y \triangleq (C_{Y|X})_# Q_X$, we have

$$D_f(P_Y, Q_Y) \leq D_f(P_X, Q_X).$$

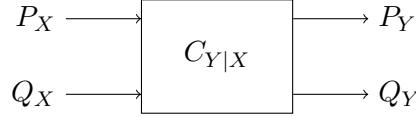


Figure 2.2.1: Data processing inequality of f -divergences.

Proof. Let $dC_{Y|X}$ denote the density function for this common channel. By a simple application of the law of iterated expectations, we have

$$\begin{aligned} \mathbb{E}_{Y \sim Q_Y} \left[f \left(\frac{dP_Y}{dQ_Y} \right) \right] &= \mathbb{E}_{Y \sim Q_Y} \left[f \left(\frac{\mathbb{E}_{X \sim P_X} [dC_{Y=Y|X}]}{\mathbb{E}_{X \sim Q_X} [dC_{Y=Y|X}]} \right) \right] \\ &= \mathbb{E}_{Y \sim Q_Y} \left[f \left(\frac{\mathbb{E}_{X \sim Q_X} \left[\frac{dP_X}{dQ_X} dC_{Y=Y|X} \right]}{\mathbb{E}_{X \sim Q_X} [dC_{Y=Y|X}]} \right) \right]. \end{aligned}$$

We can view the ratio $\frac{\mathbb{E}_{X \sim Q_X} \left[\frac{dP_X}{dQ_X} dC_{Y=Y|X} \right]}{\mathbb{E}_{X \sim Q_X} [dC_{Y=Y|X}]}$ as a conditional expectation, because the denominator is the probability that we return $Q_Y = Y$, and the numerator can be viewed as the conditional probability that we return $Q_Y = Y$ multiplied by the importance weight $\frac{dP_X}{dQ_X}$. Thus we can simplify

$$\mathbb{E}_{Y \sim Q_Y} \left[f \left(\frac{dP_Y}{dQ_Y} \right) \right] = \mathbb{E}_{Y \sim Q_Y} \left[f \left(\mathbb{E}_{X \sim Q_X} \left[\frac{dP_X}{dQ_X} \middle| C_{Y|X}(X) = Y \right] \right) \right]$$

Now, we are ready to exploit the key convexity property of f -divergences through Jensen's inequality, stated below.

Theorem 2.2.2 (Jensen's Inequality). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function and let X be a random variable over \mathbb{R}^d . Then we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Applying Jensen's inequality to the inner expectation yields

$$\mathbb{E}_{Y \sim Q_Y} \left[f \left(\frac{dP_Y}{dQ_Y} \right) \right] \leq \mathbb{E}_{Y \sim Q_Y} \mathbb{E}_{X \sim Q_X} \left[f \left(\frac{dP_X}{dQ_X} \right) \middle| C_{Y|X}(X) = Y \right].$$

By the law of iterated expectations, we conclude

$$\mathbb{E}_{Y \sim Q_Y} \mathbb{E}_{X \sim Q_X} \left[f \left(\frac{dP_X}{dQ_X} \right) \middle| C_{Y|X}(X) = Y \right] = \mathbb{E}_{X \sim Q_X} \left[f \left(\frac{dP_X}{dQ_X} \right) \right].$$

□

As a recap, the key insight of the data processing inequality was to isolate the most important aspects of measures for which a data processing inequality applies: the lack of dependence on the underlying structure of the base set Ω , the likelihood ratio, and the convexity condition. With the correct definition of f -divergences, the proof of the data processing inequality is essentially trivial. This theme repeats itself in our thesis; while the original framework for critical windows was complicated for diffusion, taking the perspective of stochastic localization simplifies these difficulties and yields a very simple yet general proof across all models, including language models and diffusion.

*The language in which we express our ideas
has a strong influence on our thought pro-
cesses.*

Donald Knuth

3

Intuition for Critical Windows

We now introduce our theory for critical windows. As a toy setting, consider an image diffusion model that outputs pictures of cats or dogs, and let us view the output as a *mixture model* with three modes, corresponding to orange cats, brown cats and dogs. We analyze a critical window in which we transition from sampling from both cats and dogs to cats with certainty. Chapter 8 uses the same language and re-defines the concepts herein for general stochastic localization samplers.

We defer the details of the specific of diffusion models to Part I, but for now assume that they compose of a forward process and a reverse process. This forward process maps the original distribution $p \in \mathbb{R}^d$ to a pure noise distribution called a Gaussian, thereby noising the distribution, and the reverse process undoes this transformation. The diffusion model learns the reverse process and runs it to sample from the original distribution p . Importantly, it is indexed by time $t \in [0, T]$ and *decreases* as we approach the final image. We start with a pure noise sample at $t \triangleq T$, run the reverse process iteratively, decreasing t , and then return the final value at $t = 0$. With this time indexing, a critical window simply refers to the dramatic increase in the probability of sampling a cat between $[T_{\text{after}}^C, T_{\text{before}}^C]$. We use the superscript C to denote that these are the time intervals for a critical window.

Figure 3.0.1 captures this idea more generally. We have time points $T_{\text{after}}^C, T_{\text{before}}^C$ corresponding to subsets $S_{\text{after}}, S_{\text{before}}$, where in our example $S_{\text{before}} = \{\text{orange cats, brown cats, dogs}\}$ and $S_{\text{after}} = \{\text{orange cats, brown cats}\}$. For $t \geq T_{\text{before}}^C$, the diffusion model is sampling from the subset S_{before} of both cats and dogs. For $t \leq T_{\text{after}}^C$, the diffusion model is sampling from the subset S_{before} of just cats. Notably, the critical window is situated, and in some sense defined, by the two intervals which surround it: the left interval indicates that we have committed to only sampling

from the distribution of cats indexed by S_{after} and have a high probability of producing a cat as the final image, and the right interval, indexed by S_{before} , indicates we can sample images of cats or dogs. Under this framework, critical windows simply represent the *speciation* from sampling from a larger sub-population of the distribution given by S_{before} to a smaller sub-population given by S_{after} .

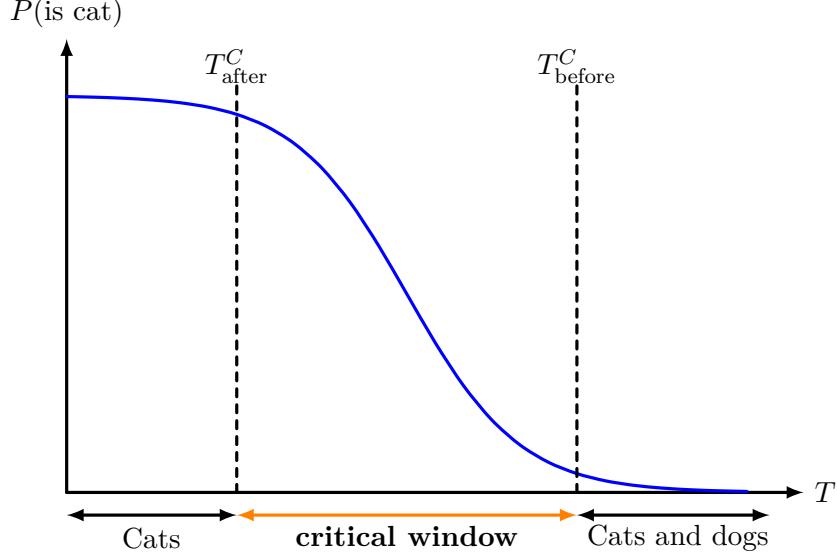


Figure 3.0.1: Illustration of the definition of a critical window for image diffusion models which transitions from sampling from both cats and dogs ($S_{\text{before}} = \{\text{orange cats, brown cats, dogs}\}$) to just cats ($S_{\text{after}} = \{\text{orange cats, brown cats}\}$).

In other words, we want to understand at what interval $t \in [B_1, B_2]$ is some *feature* (herein the feature of being a cat or dog) determined by the diffusion process and then at what interval $t \in [A_1, A_2]$ the more *specific version of the feature* (herein the feature of being a cat) is determined by the diffusion, yielding the critical window as the transition period $[A_2, B_1]$ in which we specialize to the more specific version of the feature. Thus, to develop a theoretical framework for critical windows we start by characterizing when a feature of a generation is determined with something called a forward-reverse experiment.

Forward-reverse experiment. Given a sample x from the data distribution p , consider the following experiment to understand when a diffusion model decides a given feature of x , e.g. when a diffusion model forms the feature of being a cat in images of orange cats (Figure 3.0.2). We apply the forward process to x for time t to produce a noisy sample x_t . We then run the reverse process, which learned by the diffusion model, for time t starting from x_t to produce a new sample x' (see Section 5.1 for formal definitions). As $t \rightarrow \infty$, the distribution over x_t converges to Gaussian, and thus the resulting distribution over x' converges to p . In our toy example of an image diffusion model, once enough noise is added to a cat image, the original information in the image is completely destroyed and the diffusion model generates images of both cats and dogs. As $t \rightarrow 0$, the distribution

over x' converges to a point mass at x . If only a small amount of noise is added to an image of an orange cat, we just obtain the same orange cat. Thus, there is some intermediate time t' where a given feature of x' is determined. In the image diffusion example, this means there is this “sweet spot” in which noising and denoising an image of a cat yields an image from the distribution of cats, but not necessarily a cat of the same color.

Remark 3.0.1. Several authors have studied the forward-reverse experiment in different contexts [HJA20a, MHS⁺22, SFW25, SFLW24]. In the low-noise regime, [HJA20a] observed that for small t , the distribution over x' is essentially given by randomly modifying low-level features of x , while [MHS⁺22] uses a version of the forward-reverse experiment to modify low-level features of an image while preserving higher-level ones.

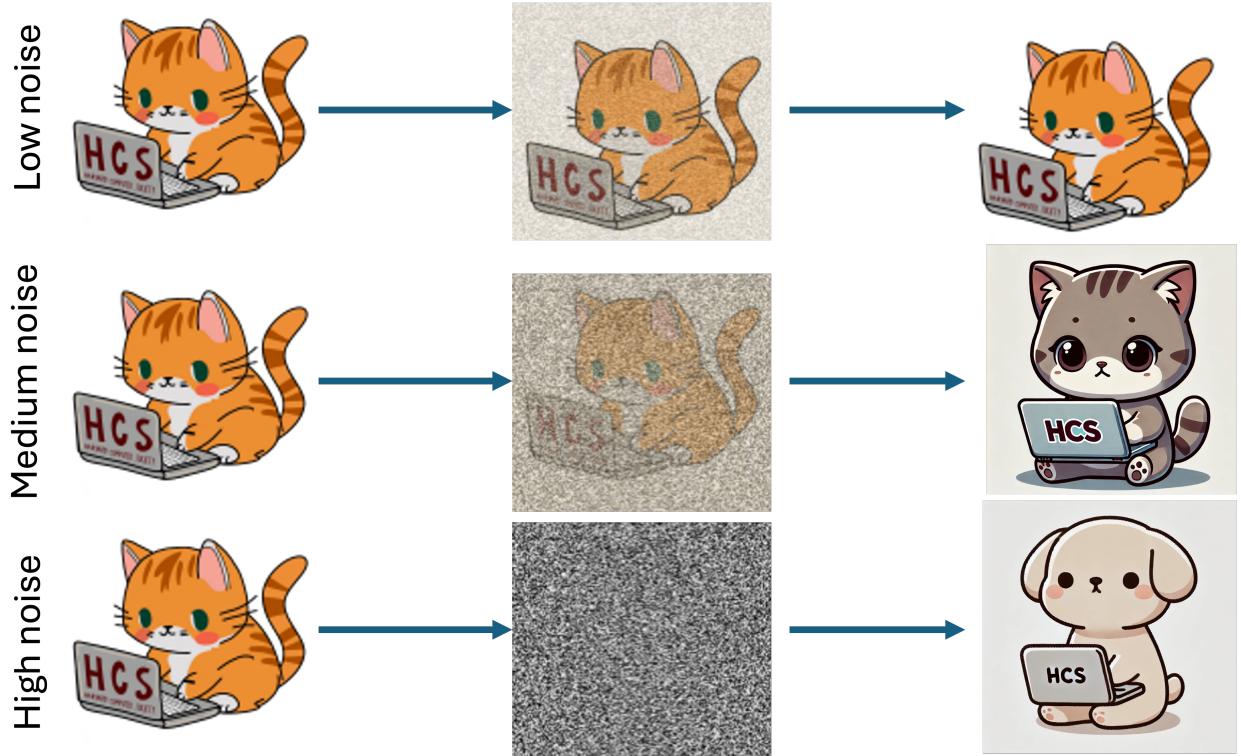


Figure 3.0.2: Forward-reverse experiment on an image of an orange cat. At a low-noise regime, this yields the same image; at a middle-noise regime, this yields a cat, but of a different color; at a high-noise regime, this yields a dog. Orange cat courtesy of the Harvard Computer Society.

Now we formalize the idea of the first time t for which samples from the distribution over x' mostly share a certain feature with x . The natural way to do this is through a *mixture model*.

Features and mixture models. Consider a data distribution p given by a mixture of sub-populations $p^{\theta_1}, p^{\theta_2}, \dots, p^{\theta_K}$ with $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. We let each sub-population represent some

feature (or combination of features). In our example, we could have $\Theta = \{\text{orange cats, brown cats, dogs}\}$. We can define features by *nesting* classes together: the *sub-mixture* of cats can be interpreted as the mixture of orange and brown cats. Thus, in our setting, where we take an orange cat x and apply the forward-reverse experiment, we can ask whether the resulting x' is close the distribution $p^{S_{\text{target}}}$, the mixture model of the classes $S_{\text{target}} = \{\text{orange cats, brown cats}\}$.

Because we want to reason about when certain features of distributions are determined, we marginalize out the randomness in the selection of x by taking it as a sample from the sub-mixture $x \sim p^{S_{\text{init}}}$ where $S_{\text{init}} \subset S_{\text{target}}$. In this example, S_{init} corresponds to the sub-population of orange cats. Then we let the resulting marginal distribution over x' be denoted by $p^{S_{\text{init}}, t}$. Thus the times at which $p^{S_{\text{init}}, t}$ is close to the sub-mixture given by S_{target} are the times at which the diffusion model “decides” that the sample will belong to S_{target} . In the orange cat, brown cat, and dog example, this corresponds to the time at which the diffusion model selects the cat feature in images of orange cats. Now we will try to understand when this specialization occurs in the sampling procedure of the diffusion model.

Feature specialization bounds. In Section 5.2, we establish bounds on the interval at which $p^{S_{\text{init}}, t}$ approximates the sub-mixture given by S_{target} for any S_{init} . They depend on the total variation distance between sub-populations inside and outside S_{init} and S_{target} somewhere along the forward process. We identify two endpoints (see Eqs. (5.2) and (5.3) for formal definitions):

1. T_{end}^S : the time in the forward process at which the initial sub-mixture indexed by S_{init} and the target sub-mixture indexed by S_{target} first become close in total variation distance.
2. T_{start}^S : the time in the forward process at which a component in S_{target} begins to exhibit non-negligible overlap with a component in the rest of the mixture.

Intuitively, in the interval $t \in [T_{\text{end}}^S, T_{\text{start}}^S]$, the sampler has chosen the features defined by S_{init} but has not yet chosen the features that distinguish S_{init} and S_{target} . We use the superscript S to denote that the time intervals for specialization to some given feature.

When t is sufficiently large, applying the forward process for time t with an initial sample from S_{init} or S_{target} will give rise to similar distributions via the *data processing inequality*. In our toy setting, this means the differences between orange cats and brown cats will be *erased* with sufficient noise, motivating our definition of T_{end}^S . If t is too large, then applying the forward process collapses the difference between S_{target} and $\Theta - S_{\text{target}}$, which means the reverse process will yield samples outside of S_{target} as well. More concretely, this means that with too much noise, the difference between cats and dogs will be negligible and the forward-reverse experiment will produce images of dogs as well. Intuitively, this should not happen when the components of p outside S_{target} do not overlap that much with the ones inside S_{target} even after running the forward process for time t , thus motivating our definition of T_{start}^S . Lemma 2.1.1 formalizes this by showing that the model will “believe” that the sample belongs to S_{target} with high probability at T_{start}^S .

Theorem 3.0.1 (Informal, see Theorem 5.2.1). Let $S_{\text{init}} \subset S_{\text{target}}$. For any $t \in [T_{\text{end}}^S, T_{\text{start}}^S]$, 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_{target} .

A priori T_{end}^S need not be smaller than T_{start}^S , but this turns out to not be an issue in many cases. For latent diffusion models in particular, where we apply the diffusion procedure to the latent space, the manifold of images becomes highly structured, meaning that there will be “salient” features such that images with the same salient feature will be closer together in the latent space. Also, if the target mixture is the same as the initial mixture, we only need $T_{\text{start}}^S > T_{\text{end}}^S = 0$ to form a critical window. This setting is especially relevant for interpretability and data attribution tasks, which usually examines an object x with property p and asks for the largest time for which property p is preserved.

Now that we have a grasp on what sampling some set of features looks like, we are ready formally define critical windows, in which we transition from sampling from S_{before} to S_{after} .

Definition 3.0.1 (Critical window). Define $S_{\text{after}} \subset S_{\text{before}} \subset \Theta$.

For S_{after} , consider $T_{\text{after}}^C \triangleq T_{\text{end}}^S$ for $S_{\text{init}}, S_{\text{target}} \triangleq S_{\text{after}}$. This is largest time the model is sampling from $p^{S_{\text{after}}}$. For S_{before} , we define $T_{\text{before}}^C \triangleq T_{\text{start}}^S$ for $S_{\text{init}} \triangleq S_{\text{after}}$ and $S_{\text{target}} \triangleq S_{\text{before}}$. This is the smallest time the model is sampling from $p^{S_{\text{before}}}$.

A *critical window* is the interval

$$[T_{\text{after}}^C, T_{\text{before}}^C],$$

where there is a transition from sampling from S_{before} to the smaller subset S_{after} .

So far we have largely focused on a specific critical window, which would correspond to the formation of specific feature in the generated output. One could imagine that there could be a sequence of critical windows, which naturally leads to a hierarchical sampling procedure.

Hierarchical sampling interpretation. In Section 6.2.1, we model a hierarchy of data as a Gaussian mixture model with a hierarchical clustering structure.¹ The measure decomposes into well-separated clusters of sub-populations such that the separation between clusters exceeds the separation within clusters. Each cluster decomposes similarly *ad infinitum*, until we reach the leaves. This naturally defines a *mixture tree*, where the vertices correspond are the cluster at some resolution of the forward process, the root is the entire data distribution, and the leaves are raw components of the distribution (Definition 6.2.1).

¹In Part II, we extend our theory to many more samplers and distributions.

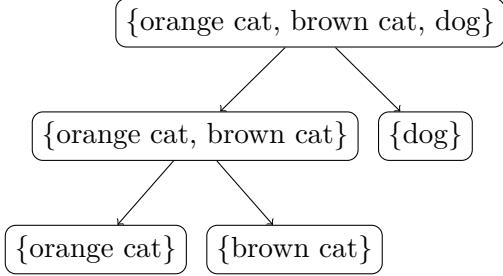


Figure 3.0.3: Example of a mixture tree.

If we think of every node v as being associated with a feature, then the corresponding cluster of sub-populations is comprised of all sub-populations which possess that feature. By chaining together several applications of Theorem 6.0.2, we prove the following:

Theorem 3.0.2 (Informal, see Theorem 6.2.2). 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 $t_{v_L} > \dots > t_{v_{\ell}}$ such that for all $\underline{L} \leq \ell \leq L$, 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 total variation 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 the whittling away of other sub-populations until the sampler reaches the end component.

3.1 Other theories for critical windows

Here we will briefly overview past or concurrent works for critical windows for diffusion from the statistical physics community and compare it to our framework. As a reminder, the main difference between our framework and theirs is that we provably characterize critical windows across different models and data distributions.

Spontaneous symmetry breaking. [RA23] links critical windows to the spontaneous symmetry breaking phenomenon in statistical physics, in which a symmetric state spontaneously collapses into an asymmetric state. In particular, [RA23] studies critical windows for a discrete distribution supported on two points, where the diffusion model shifts from sampling both points with equal probability (a symmetric state) to specializing to one of the two points (an asymmetric state). They define a fixed-point path such that the sampling procedure exhibits zero drift over it and characterize critical windows as the phase transition at which the spectrum of the Hessian of the

potential bifurcates into positive and negative parts. In contrast to their work, we give end-to-end guarantees for a more general family of high-dimensional distributions.

Random hierarchy model. [SFW25, SFLW24] study the critical windows of the *random hierarchy model* [PCT⁺23], a discrete distribution over one-hot embeddings of strings in a multi-level context-free grammar with random production rules. They identify a critical window in which the features at a higher level in the grammar are determined, and then obtain accurate but non-rigorous predictions for the location of this critical window by passing through a mean-field approximation and applying a belief propagation computation. In contrast, our theory is based on *mixture models*, where the notion of hierarchy is encoded by the locations of the components.

Dynamical regimes of diffusion models. [BBdBM24] also explore a toy model, a mixture of two isotropic Gaussians, and computes the critical window in which the sampling procedure specializes to one particular mode of the Gaussian. Using a Landau-type perturbative calculation useful for predicting second-order phase transitions in statistical physics, they give precise but non-rigorous asymptotic predictions for this critical window. Interestingly, [BBdBM24] recommends a heuristic for critical windows based on the noise level at which the principal component of the data distribution is erased, which is closely related to the notion of separation between components in our mixture model setting.

Part I

Diffusion

If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

John von Neumann.

4

A Theoretical Introduction to Diffusion

In Part I, we formally establish our critical windows framework in Chapter 3 for a type of generative models called *continuous diffusion models*. Appearing in diverse domains such as video generation [Ope24], robotic control and motion planning [CLB⁺23], and protein folding [TL24], diffusions have emerged as one of the most flexible and powerful paradigms for generative modeling that exist currently [SDWMG15, DN21, SSDK⁺20, HJA20a].

In this chapter, we will provide an introduction to diffusion models, with a special emphasis on its theory, and illustrate how it can be viewed as unifying two previous modeling paradigms (Sections 4.2). We will also state and explain a key technical tool for Part I, a Girsanov-based approximation argument that allows us to control the distributional differences of running the reverse process for different sub-populations [CCL⁺23b] (Section 4.2.4). On a personal note, this theorem is deeply personal because it inspired the author’s own foray into research in diffusion models. This chapter will heavily borrow from the beautifully written expositions on diffusion and related methods by [Che24, Che23].

In the next section, we will present a brief primer on the stochastic calculus fundamentals for diffusion. If the reader is already aware of the basics, then they can safely skip this section and refer back to it if needed.

4.1 Stochastic calculus basics

Consider a stochastic process $(X_t)_{t \geq 0}$, which we want to characterize by its *infinitesimal linear behavior*, analogous to how a function can be defined by its infinitesimal behavior in an ordinary differential equation (ODE). We want to formalize the idea that $(X_t)_{t \geq 0}$ satisfies the following

equality for small h ,

$$X_{t+h} = X_t + hb_t(X_t) + \sqrt{h}\sigma_t(X_t)G + O(h^{3/2}), \quad t \geq 0 \quad (4.1)$$

where $b_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is our drift term depending on the state X_t , $G \sim \mathcal{N}(0, \text{Id})$ is the randomness, and our randomness passes through a linear function $\sigma_t : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ before modifying X_{t+h} . Similar to an ODE, we have a drift term $hb_t(X_t)$, but now we have an infinitesimal random term $\sqrt{h}\sigma_t(X_t)G$. To give that meaning, we first need to have a notion of $\sqrt{h}G$, a stochastic process composed of infinitesimal Gaussians. This is formally given by *Brownian motion*. The history of Brownian motion is long and fascinating; it was first discovered by a Scottish botanist investigating the jitters of pollen grains through a microscope, was the key insight behind one of Einstein's breakthrough papers in 1905, and now appears outside of diffusion models in contexts such as quantitative finance and partial differential equations.

Definition 4.1.1 (Brownian motion). $(B_t)_{t \geq 0}$ is a Brownian motion if

- $B_0 = 0$
- For all $0 \leq s \leq t$, $B_t - B_s \sim \mathcal{N}(0, (t-s)\text{Id})$.
- For $0 \leq t_1 < \dots < t_n \leq t$, $B_{t_1}, B_{t_2} - B_{t_1}, \dots, B_t - B_{t_{n-1}}$ are independent.
- The map $t \mapsto B_t$ is continuous almost surely.

We will assume that such a process exists and refer readers interested in the construction of Brownian motion to Chapter 2 of [Gal18]. Now that we have this notion of Brownian motion, we want to understand what the operation $\sigma_t(X_t)dB_t \approx \sigma_t(X_t)\sqrt{h}G$ means. First consider the case where $\sigma_t : \mathbb{R} \rightarrow \mathbb{R}$ is a simple sum of indicators on $[0, T]$ subdivided by an increasing sequence of times $(t_i)_{i=0}^{k-1} \in [0, T]$,

$$\sigma_t = \sum_{i=0}^{k-1} H_i I(t \in [t_i, t_{i+1})).$$

In each interval $[t_i, t_{i+1})$, we ideally would like to multiply H_i to whatever the change in the Brownian motion is, $B_{t_{i+1}} - B_{t_i}$. Thus, perhaps the only sensible definition for integrating over $\int_0^t \sigma_t(X_t)dB_s$ is

$$\int_0^t \sigma_s dB_s = \sum_{i=0}^{k-1} H_i (B_{t_{i+1}} - B_{t_i}).$$

It turns out that we can extend this intuition via Itô's integral to much more general family of functions [Che23, Chapter 1.1]. Using Ito's integral, we are now ready to define a stochastic differential equation (SDE) that captures our picture from Eq. (4.1).

Definition 4.1.2. $(X_t)_{t \geq 0}$ is an Itô process if it is given by the stochastic integral

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dB_s \quad \text{for } t \geq 0,$$

As shorthand, we write

$$dX_t = b_t dt + \sigma_t dB_t.$$

For sufficiently nice b_s and σ_t (i.e. Lipschitzness suffices), the above SDE has a unique solution [Gal18, Theorem 8.3.]. Like in standard non-stochastic calculus, there exists an analogous notion of a chain rule for SDEs, that we will state below.

Theorem 4.1.1 (Itô's lemma). Let $(X_t)_{t \geq 0} \in \mathbb{R}^d$ be a solution to the SDE,

$$dX_t = b_t dt + \sigma_t dB_t.$$

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be $f \in C_2(\mathbb{R}^d)$. The stochastic process $(f(X_t))_{t \geq 0}$ satisfies the following SDE,

$$df(X_t) = \left(\langle \nabla f, b_t \rangle + \frac{1}{2} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t] \right) dt + \langle \nabla f, \sigma_t dB_t \rangle.$$

The full proof of Itô's lemma is quite technical, so we instead present a basic heuristic argument that captures most of the key concepts. We refer readers interested in the full proof to [Gal18, Theorem 5.2].

Proof sketch of Itô's lemma. We seek to non-rigorously compute $f(X_{t+h}) - f(X_t)$ for small h . Using a second-order Taylor expansion of f around X_t for some small $G \sim \mathcal{N}(0, \operatorname{Id})$, we find

$$\begin{aligned} f(X_{t+h}) &\approx f(X_t + hb_t(X_t) + \sqrt{h}\sigma_t(X_t)G) \\ &\approx f(X_t) + \langle \nabla f, hb_t(X_t) + \sqrt{h}\sigma_t(X_t)G \rangle \\ &\quad + \frac{1}{2}(hb_t(X_t) + \sqrt{h}\sigma_t(X_t)G)^\top (\nabla^2 f)(hb_t(X_t) + \sqrt{h}\sigma_t(X_t)G) \\ &\approx f(X_t) + h \left(\langle \nabla f, b_t \rangle + \frac{1}{2}(\sigma_t G)^\top \nabla^2 f(\sigma_t G) \right) + \langle \nabla f, \sqrt{h}\sigma_t G \rangle + O(h^{3/2}). \end{aligned}$$

As $h \rightarrow 0$, the $O(h^{3/2})$ terms drop out, so we will ignore them. We now pattern match each of the terms in this informal approximation to the SDE.

- $h\langle \nabla f, b_t \rangle$ corresponds to $\langle \nabla f, b_t \rangle dt$.
- $\langle \nabla f, \sqrt{h}\sigma_t G \rangle$ corresponds to $\langle \nabla f, \sigma_t dB_t \rangle$.
- $\frac{1}{2}h(\sigma_t G)^\top \nabla^2 f(\sigma_t G)$ corresponds to $\frac{1}{2} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t] dt$.

The first two equivalences are quite obvious, but the last term $\frac{1}{2} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t]$ in the drift is less clear. First observe that the statement is true in expectation. Using the fact that the trace is cyclic

and linear, we can explicitly compute the expectation,

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{2} (\sigma_t G)^\top \nabla^2 f (\sigma_t G) \right] &= \frac{1}{2} \mathbb{E} \operatorname{Tr} [G^\top \sigma_t^\top \nabla^2 f \sigma_t G] \\
&= \frac{1}{2} \mathbb{E} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t G G^\top] \text{ (trace is invariant to cyclic shifts)} \\
&= \frac{1}{2} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t \mathbb{E} G G^\top] \text{ (linearity of trace)} \\
&= \frac{1}{2} \operatorname{Tr} [\sigma_t^\top \nabla^2 f \sigma_t] \text{ (covariance of Gaussian).}
\end{aligned}$$

This derivation exploits the general fact that $\operatorname{Tr}[A] = \mathbb{E}_{G \sim \mathcal{N}(0, \operatorname{Id})}[G^\top A G]$.¹ To understand why we can replace the random variable with its expectation, consider the $d = 1$ case and observe $\mathbb{E}[B_h^2] = h$ and $\mathbb{E}[B_h^4] = 3h^2$. Thus, as $h \rightarrow 0$, the variance goes to 0 much faster than the mean. More generally, this follows from the aphoristic stochastic calculus equality,

$$\|dB_t\|^2 = dt.$$

We can formally argue this by showing that the *quadratic variation* of Brownian motion is t . \square

4.2 Motivating and defining diffusion

With the stochastic calculus machinery in hand, we now return back to the challenge of generative modeling. Consider a distribution $p \in \mathbb{R}^d$ with a smooth density. We want to learn a (perhaps random) $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that f produces samples from p ,

$$f_{\#}\mathcal{N}(0, \operatorname{Id}) = p.$$

It might in general be difficult to learn such an f immediately, so in the next section, we consider two different methods that both take an iterative approach to sampling from p .

4.2.1 Score matching

The naive approach to sample from p is to initialize at a simple distribution $X_0 \sim \mathcal{N}(0, \operatorname{Id})$ and then run a greedy algorithm to find X 's assigned higher probability under p . This is exactly the technique behind Langevin Monte Carlo (Langevin MC), which can be described as the stochastic process $(X_t)_{t \geq 0}$ solving the following SDE,

$$dX_t = \nabla \log p dt + \sqrt{2} dB_t, \quad X_0 \sim \mathcal{N}(0, \operatorname{Id}). \quad (4.2)$$

Note that the quantity $\nabla \log p$ parameterizing the drift of the SDE is known as the *score* of the distribution, and biases the model towards higher probability areas of p [Hyv05, SE19]. If we were

¹Computing the trace using samples in this way is known as the Hutchinson trace estimator, and has surprising connections to language model privacy [LWWN23].

to just follow the score function with the ODE of $dX_t = \nabla \log p dt$, then we would eventually settle on some local maximum in $\log p$. However, we want to sample from the distribution, necessitating a Brownian motion dB_t term to make us explore different parts of Ω . One can guarantee that p is the only solution to the above SDE, in the sense that it is the only stationary distribution which satisfies that SDE. For presentation clarity, we defer the proof to Appendix A.3.

Proposition 4.2.1 (Stationary distribution of Langevin MC). If we instead take $X_0 \sim p$, then the distribution $X_t \sim p$ for any $t \geq 0$.

In reality, we do not have access to p at the start, so the hope is that running the above SDE for $T \rightarrow \infty$ will yield X_T with a law close to p . It is not clear however that we are able to sample from the distribution p efficiently, in the sense that it may take exponential $T = \exp(\Omega(d))$ until the law of X_T approximates p . It turns out that it only takes linear in d timesteps to obtain a good sample from p for *strongly log-concave measures*, a general family of high-dimensional unimodal distributions, including Gaussians, exponential distributions, and uniform distributions over convex sets. In Part I, we assume that the components of our mixture model belong to this family, but this requirement will be removed in Part II.

Definition 4.2.1 (Strongly log-concave distributions). A distribution over \mathbb{R}^d is $1/\beta$ -strongly log-concave if its density function p satisfies $\nabla^2(-\log p) \succcurlyeq \frac{1}{\beta} \text{Id}$. Let $\text{SLC}(d)$ denote the set of strongly log-concave measures in \mathbb{R}^d and $\text{SLC}(\beta, d)$ denote the set of $1/\beta$ -strongly log-concave measures.

Theorem 4.2.1 (Informal, see [Che23, Theorem 4.3.6] or [DMM19]). Consider $p \in \mathbb{R}^d$ such that $\alpha \text{Id} \preceq \nabla^2(-\log p) \preceq \beta \text{Id}$. For an error term $\epsilon > 0$, there exists an adequate step size h such that running a discretization of Eq. (4.2) with $T = O_{\alpha, \beta}(d/\epsilon^2)$ yields a distribution over samples X_T with Wasserstein-2 distance less than ϵ with p .

In the multi-modal case, these convergence guarantees may not hold because Langevin MC can fail to sufficiently explore the different parts of p . To gain some intuition, note that Langevin MC strongly disincentives exploration with the score $\nabla \log p = \frac{\nabla p}{p}$, which blows up in low probability zones. Depending on the initialization, this could prevent the model from exploring different modes of p . In Figure 4.2.1, consider a Langevin MC initialized on the left mode at the blue point. It can never transition to the other mode, because it would have to traverse through a low-probability “desert” between the two. If the right mode has much higher weight than the other one, and we always initialize near the left mode, then it would be impossible to obtain convergence guarantees in many natural metrics like Wasserstein-2 or total variation.

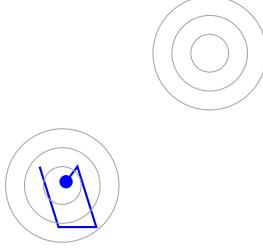


Figure 4.2.1: Langevin MC often gets stuck in one mode, resulting in poor mixing.

Perhaps the challenge here is that the original distribution can be quite messy, so maybe it would be better to convolve or *anneal* it with Gaussian noise, so the distribution becomes unimodal in the limit. Then we could sample from the simpler distribution and then somehow translate it into samples for the original measure. This is captured by the perspective of Denoising Diffusion Probabilistic Models (DDPMs) in the next section [HJA20a]. In Section 7.1, several applications of diffusion in probability theory will also highlight the usefulness of this annealing intuition.

4.2.2 Denoising Diffusion Probabilistic Models

In the original presentation of DDPMs, the generative process is discrete, and defined by the graphical model $(X_t)_{t \in \{0, 1, \dots, T\}}$, where $X_T \sim \mathcal{N}(0, \text{Id})$, $X_0 \sim p$, and the conditional distribution $X_t | X_{t-1}$ is a convolution with a Gaussian distribution (Figure 4.2.2). By learning the posteriors $p_\theta(X_{t-1} | X_t)$ which *denoise* the image, we can sample by starting from $X_T \sim \mathcal{N}(0, \text{Id})$ and iteratively applying the posteriors.

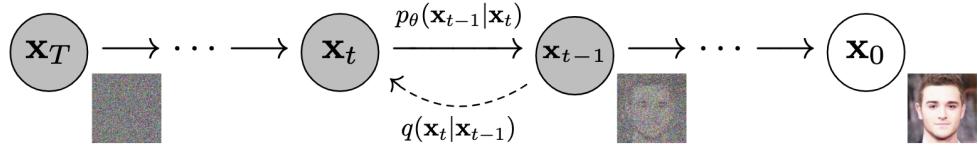


Figure 4.2.2: Diagram of DDPMs from [HJA20a].

It turns out that this model has a deep connection to the score function in Langevin MC. Tweedie's formula relates the score function of annealed versions of p to this denoising problem. The derivation is quite cute and deferred to Appendix A.3.

Proposition 4.2.2 (Tweedie's formula). For $x \sim p$ and $g \sim \mathcal{N}(0, \sigma^2 \text{Id})$, let $\tilde{x} = x + g$ and \tilde{p} be the density of $p * \mathcal{N}(0, \sigma^2 \text{Id})$. We have,

$$\mathbb{E}[x | \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x}).$$

Thus, this model can be interpreted as simply applying the score function to annealed versions of the distribution. Right now, the denoising process is discrete, and a natural question to ask

is whether this can be extended into a continuous stochastic process like Langevin MC, which is exactly diffusion.

4.2.3 Defining diffusion

In diffusion, we sample from $X \sim p$ by defining a *forward process* which progressively transforms samples from p into pure noise, and then learning a *reverse process* which undoes this process and transforms samples from pure noise into p . For the sake of simplicity, we present the special case for diffusions of the Ornstein-Uhlenbeck (OU) process. The OU process is the stochastic process $(X_t)_{t \geq 0}$ which solves the following SDE,

$$dX_t = -X_t dt + \sqrt{2} dB_t, \quad X_0 \sim p.$$

Other choices of the forward process are possible, but can be reduced to the OU process by simple reparameterizations of time or space. Let p_t denote the law of X_t . We can explicitly parameterize the marginal distribution p_t by the law of $e^{-t} X_0 + \sqrt{1 - e^{-2t}} G$, where $X_0 \sim p$ and $G \sim \mathcal{N}(0, \text{Id})$. Observe that as $t \rightarrow \infty$, p_t converges exponentially quickly to pure noise in various senses, e.g. in Wasserstein-2 distance. To undo this procedure, we choose a terminal time $T \geq 0$ such that $p_T \approx \mathcal{N}(0, \text{Id})$ and seek to reverse the SDE to obtain transformation from $\mathcal{N}(0, \text{Id})$ to p . It turns out that we can appeal to the Fokker-Planck equation, which characterizes how changes in the probability distribution of p can be expressed in terms of the drift coefficient of the SDE.

Theorem 4.2.2 (Fokker-Planck equation). Let $b_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a family of smoothly varying vector fields. The solution $(X_t)_{t \geq 0}$ of the following SDE

$$dx_t = b_t dt + \sqrt{2} dB_t, \quad t \geq 0$$

is distributed according to $(p_t)_{t \geq 0}$ which satisfies the following partial differential equation,

$$\partial_t p_t = -\text{div}(p_t b_t) + \Delta p_t.$$

We defer the full presentation to Appendix A.3, but quickly note that this is analogous to be continuity equation in physics, which describes the flow of an incompressible fluid. In the context of diffusion, the Fokker-Planck equation allows us to solve for the parameterization of the reverse SDE. Assume we have a stochastic process $(X_t^\leftarrow)_{t \in [0, T]}$ which solves

$$dX_t^\leftarrow = r_t(X_t^\leftarrow) dt + \sqrt{2} dB_t, \quad X_0^\leftarrow \sim p_T$$

such that p_{T-t} is the law of X_t^\leftarrow , recovering the initial distribution p_0 with X_T^\leftarrow . Plugging in the Fokker Planck equation, we want to solve for $r_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that

$$-\text{div}(p_t x) + \Delta p_t = \text{div}(p_t r_{T-t}) - \Delta p_t.$$

Note $c \mapsto \operatorname{div}(p_t(x)c)$ is linear for any $x \in \mathbb{R}^d$, so $\operatorname{div}[p_t(r_{T-t} - x)/2] = \Delta p_t$. Noting that $\operatorname{div}(\nabla p_t) = \Delta p_t$, we are done by setting $r_t = x + 2\frac{\nabla p_{T-t}}{p_{T-t}} = x + 2\nabla \ln p_{T-t}$. Thus, to recap, we have the reverse process, which we denote by $(X_t^\leftarrow)_{t \in [0, T]}$, solving the standard reverse SDE,

$$dX_t^\leftarrow = \{X_t^\leftarrow + 2\nabla \ln p_{T-t}(X_t^\leftarrow)\} dt + \sqrt{2} dB_t, \quad X_0^\leftarrow \sim p_T, \quad (4.3)$$

where $(B_t)_{t \geq 0}$ is also a Brownian motion and $X_t^\leftarrow \sim p_{T-t}$.

Via Tweedie's formula, we can learn the score of the annealed versions of p_t with a denoising problem. We first generate pairs $x, \tilde{x} \sim (X_0, X_t)$, which is easy because $X_t \sim \mathcal{N}(e^{-t}X_0, (1-e^{-2t})\operatorname{Id})$, and train a neural network $s'_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ which minimizes the loss function $\mathcal{L} = \mathbb{E}_{x, \tilde{x}} \|x - s'_t(\tilde{x})\|^2$. The minimum is achieved at $s'_t = \mathbb{E}[x|\tilde{x}]$, and setting $s_t \triangleq \frac{e^t s'_t}{1-e^{-2t}}$ obtains the score.

4.2.4 Score approximation and distribution learning

Even still, it is unrealistic to expect that we can learn the score perfectly by minimizing the denoising problem. Thus it remains to be seen whether an *approximation* to the SDE, with a different drift term than the true one, can yield adequate samples from the distribution. In the past several years, many works have investigated this very problem of proving convergence guarantees for diffusion models assuming access to a good approximation to the score [DBTHD21, BMR22, CLL22a, DB22, LLT22, LWYL22, Pid22, WY22, CCL⁺23b, CDD23, LLT23, LWCC23, BDD23, CCL⁺23a, LHW24]. In total, these works have found that diffusion models can sample over essentially any distribution over \mathbb{R}^d , even highly multi-modal distributions, unlike existing guarantees for earlier non-diffusion methods like Langevin MC.

Herein we use an approximation argument for a path-based analysis of SDEs based on Girsanov's theorem that was developed in one of these works [CCL⁺23b]. In the original context, it was used to bound the distributional difference in path measures of the SDEs associated with the learned and true distribution. In our thesis, we use Theorem 4.2.3 to bound the difference in output distributions of two SDEs corresponding to different sub-mixtures of the distribution. We include a heuristic argument for Theorem 4.2.3 in Appendix A.3.

Theorem 4.2.3 ([CCL⁺23b, Section 5.2]). Let $(Y_t)_{t \in [0, T]}$ and $(Y'_t)_{t \in [0, T]}$ denote the solutions to

$$\begin{aligned} dY_t &= b_t(Y_t) dt + \sqrt{2} dB_t, & Y_0 &\sim p, \\ dY'_t &= b'_t(Y'_t) dt + \sqrt{2} dB_t, & Y'_0 &\sim p. \end{aligned}$$

Let p and q' denote the laws of Y_T and Y'_T respectively. If b_t, b'_t satisfy

$$\int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt < \infty,$$

then

$$\text{KL}(q \| q') \leq \int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt,$$

where Q_t is the law of Y_t .

4.2.5 New perspectives, new methods

In summary, we started with the perspectives of score matching and denoising. Via Tweedie's formula, denoising can be viewed as a form of score matching applied to annealed versions of the distribution, and diffusion is the continuous limit of this denoising procedure.

These combinations of different generative modeling paradigms have led to concrete algorithmic improvements: the connection between denoising diffusion probabilistic models and the diffusion SDE inspired the use of existing tools in numerical analysis to improve samplers for diffusion models, e.g. [SSDK⁺20]. Other sampling frameworks related to or generalizing diffusion are more computationally efficient than diffusion and yield a 2-3x speedup in practice [LCBH⁺22, AVE23, ABVE23, DTSV23, GHH⁺25].

We will see this theme that unification begets algorithm innovations repeat itself with stochastic localization in Chapter 7, where we introduce a more general framework for samplers based on stochastic localization which enables one to fully *enumerate* the design space of generative models.

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

Richard Feynman

5

Log-concave Mixtures for Continuous Diffusion

In this chapter, we will formalize Theorem 3.0.1 in Chapter 3 for continuous diffusions. As a reminder, this theorem characterizes when a diffusion model selects some certain features. In [LC24], we only proved it for strongly log-concave measures (see Definition 4.2.1). The main use of this assumption was to establish control over the Hessian $\nabla^2(-\log p)$ of a distribution p , which we had to bound for technical reasons. Now, we will describe our main framework of the forward and reverse experiment in completely rigorous detail.

5.1 Forward-reverse experiment for diffusion

As issues of discretization, score error, and the support of the data distribution lying on a lower-dimensional sub-manifold 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. We are thus modeling the distribution of outputs from the sampler.

To capture the notion of a feature of the generation, we assume that the distribution $X \sim p$ of the outputs from the sampler is a mixture model. Consider a discrete set $\Theta = \{\theta_1, \dots, \theta_K\}$ with non-negative weights w_1, \dots, w_K summing to 1. Each $\theta_i \in \Theta$ is associated with a probability density function $p^{\theta_i} : \mathbb{R}^n \rightarrow \mathbb{R}^{\geq 0}$. To generate a sample $X \sim p$, we first draw $\theta \sim \text{Cat}(\Theta, \{w_i\}_{i=1}^K)$ and return $X \sim p^\theta$. This yields an overall density of $p \triangleq \sum_{\theta \in \Theta} w_\theta p^\theta$. For any non-empty $S \subset \Theta$, we also define the sub-mixture p^S by $p^S \triangleq \sum_{\theta \in S} \frac{w_\theta}{\sum_{\phi \in S} w_\phi} p^\theta$.

Let μ_i denote the mean of p^i . Let $(X_t^S)_{t \in [0, T]}$ denote the forward process given by running Eq. 4.2.3 with $q = p^S$, let p_t^S denote the law of X_t^S , and let $(Y_t^{\leftarrow, S})$ denote the reverse process given by running Eq. 4.3 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 $(Y_t^{\leftarrow, S})_{t \in [0, \widehat{T}]}$ by $P_{\widehat{T}}^{\leftarrow, S} \in \mathcal{C}([0, \widehat{T}], \mathbb{R}^d)$.

The forward-reverse experiment. 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 $S \subset \Theta$, define the modified reverse process $(Y_t^{\leftarrow}[S^{(\widehat{T})}])_{t \in [0, \widehat{T}]}$ to be given by running the reverse SDE in Eq. (4.3), 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 $Y_t^{\leftarrow}[S^{(\widehat{T})}]$ by $p_t^{S, \widehat{T}}$ and the path measure for $(Y_t^{\leftarrow}[S^{(\widehat{T})}])_{t \in [0, \widehat{T}]}$ by $P^{\leftarrow}[S^{(\widehat{T})}] \in \mathcal{C}([0, \widehat{T}], \mathbb{R}^d)$. 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 $Y_t^{\leftarrow}[S^{(\widehat{T})}]$

Because this process reverses the forward process conditioned on a particular subset S of the original mixture components, we refer to $(Y_t^{\leftarrow}[S^{(\widehat{T})}])_{t \in [0, \widehat{T}]}$ as the *S-forward-reverse experiment* from noise level \widehat{T} . We caution that the *S*-forward-reverse experiment 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 forward-reverse experiment is that of the full mixture p rather than that of p^S .

Comparison to [GVS⁺23]. [GVS⁺23] 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 reran the reverse process starting at some intermediate point x_t in the trajectory. They compute the probability that the images sampled shared a feature with x_0 and identify critical times $T_{\text{lower}}^{\text{cond}} < T_{\text{upper}}^{\text{cond}}$ such that sampling from $x'_0 \sim \Pr[\cdot | x_{T_{\text{lower}}^{\text{cond}}}]$ preserves the given feature in the original image while sampling from $x'_0 \sim \Pr[\cdot | x_{T_{\text{upper}}^{\text{cond}}}]$ does not. Our definition is slightly different: instead of rerunning the reverse process, we run the forward process for time t starting from x_0 to produce x_t and then run the reverse process from x_t . Arguably, our setup is more flexible as it does not require knowledge of the trajectory that generated x_0 . In general, we expect that our critical window thresholds are less than [GVS⁺23]'s thresholds because adding noise to the state at intermediate times could also change the features.

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 $S, S' \subset \Theta$, define

$$\begin{aligned} \bar{R} &\triangleq \max_{i \in \Theta} \|\mu_i\| & w(S, S') &\triangleq \max_{i \in S, j \in S'} \|\mu_i - \mu_j\| \\ \Delta(S) &\triangleq \min_{\ell \in S, j \in \Theta - S} \|\mu_\ell - \mu_j\| & \bar{W} &\triangleq \max_{i, j \in \Theta} \frac{w_i}{w_j}. \end{aligned}$$

5.2 Main theorem for feature specialization

Recall that $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$ denote the two sub-mixtures we are interested in. In the notation of Section 5.1, we wish to establish upper and lower bounds on the time \widehat{T} at which

$$\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}}) \quad (5.1)$$

becomes small. This is the interval where we are sampling from S_{target} .

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

$$T_{\text{end}}^S(\epsilon) \triangleq \inf\{t \in [0, T] : \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{target}}}) \leq \epsilon\} \quad (5.2)$$

$$\begin{aligned} T_{\text{start}}^S(\epsilon) &\triangleq \sup\{t \in [0, T] : \text{TV}(p_t^i, p_t^j) \geq 1 - \epsilon^2/2 \\ &\quad \forall i \in S_{\text{target}}, j \in \Theta - S_{\text{target}}\}. \end{aligned} \quad (5.3)$$

When ϵ is clear from context, we refer to these times as T_{end}^S and T_{start}^S . We expect that Eq. (5.1) is small provided $\widehat{T} \geq T_{\text{end}}^S$ and $\widehat{T} \leq T_{\text{start}}^S$. In this section, we prove that this is indeed the case for any p given by a mixture of strongly log-concave distributions (see Remark 5.2.1 for discussion on the assumption of strong log-concavity of components).

Assumption 5.2.1 (Strong log-concavity). For some $\Psi^2 \geq 1$, $p^i \in \text{SLC}(\Psi^2, d)$.

Assumption 5.2.2 (Smooth components). For some $L > 0$ and for all $t \geq 0$, the score $\nabla \ln p_t^i$ is L -Lipschitz.

Assumption 5.2.3 (Moment bound). For some $M \geq 1$ and for all $i \in \Theta$ and $t \in [0, T]$, $\mathbb{E} \|X_t^i\|^4 \leq M$.

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

Assumption 5.2.4 (Score bound). For some $\overline{M} \geq 0$ and for all $i, j \in \Theta, t \in [0, T]$,

$$\mathbb{E}_{X \sim p_t^j} \|\nabla \ln p_t^i(X)\|^4 \leq \overline{M}.$$

We compute \overline{M} for various examples in Section 6.2, but for now one can safely think of \overline{M} as scaling polynomially in the dimension and in the parameter \overline{R} .

Remark 5.2.1. 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 5.2.2, which is also much stronger than what is necessary for Theorem 5.2.1. It suffices to show the LHS of Lemma 5.2.2 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. In fact, in Part II, we alleviate this assumption by bounding the difference between a different quantity than the score, and extend our framework to all mixture models.

5.2.1 Main result and proof sketch

We are now ready to state our main bound for the \widehat{T} at which $\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}})$ becomes small.

Theorem 5.2.1. Let $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$. For $\epsilon > 0$, if $\widehat{T} \geq T_{\text{end}}^S(\epsilon)$ and $\widehat{T} \leq T_{\text{start}}^S(\epsilon)$, then

$$\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}}) \lesssim \epsilon \sqrt{\overline{W}} K^2 (\overline{R}^2 + M^2 + \sqrt{M} \Psi^4 + \sqrt{M}) .$$

The proof this is established as follows. There are two differences between the distribution $p^{S_{\text{init}}, \widehat{T}}$ and $p^{S_{\text{target}}}$ when considering the reverse process initialized at \widehat{T} : the reverse SDEs have different initializations and are parameterized by the scores of S_{init} and S_{target} . We control the first by the data processing inequality and the definition of T_{end}^S , and we control the second term by applying the approximation argument in Girsanov's theorem in Theorem 4.2.3 to the difference of the scores for S_{target} and Θ . When the total variation between S_{target} and $\Theta - S_{\text{target}}$ is near 1, because we are integrating starting from a sample from S_{target} , this difference in scores is supposed to be close to 0. There is also this auxiliary term which we will need to control and show it integrates to a constant over time:

Lemma 5.2.2. Under Assumptions 5.2.1, 5.2.3, and 5.2.4, $\mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim e^{-4t} (\overline{R}^4 + M^4 + M\Psi^8 + \overline{M}) \quad \forall i, j, \ell \in \Theta$.

Informally, this lemma quantifies 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 5.2.1. By the data processing inequality and definition of $T_{\text{end}}^S, T_{\text{start}}^S$, for all $i \in S_{\text{target}}, j \notin S_{\text{target}}$,

$$\begin{aligned} \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{target}}}) &\leq \epsilon \quad \forall t \in [\widehat{T}, T] \\ \text{TV}(p_t^i, p_t^j) &\geq 1 - \epsilon^2/2 \quad \forall t \in [0, \widehat{T}] . \end{aligned}$$

By the data processing inequality and triangle inequality,

$$\begin{aligned} \text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}}) &\leq \text{TV}(P^{\leftarrow}[S_{\text{init}}^{(\widehat{T})}], P_{\widehat{T}}^{\leftarrow, S_{\text{target}}}) \\ &\leq \underbrace{\text{TV}(P^{\leftarrow}[S_{\text{init}}^{(\widehat{T})}], P^{\leftarrow}[S_{\text{target}}^{(\widehat{T})}])}_{(I)} + \underbrace{\text{TV}(P^{\leftarrow}[S_{\text{target}}^{(\widehat{T})}], P_{\widehat{T}}^{\leftarrow, S_{\text{target}}})}_{(II)} . \end{aligned}$$

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

$$\text{TV}(P^{\leftarrow}[S_{\text{init}}^{(\widehat{T})}], P^{\leftarrow}[S_{\text{target}}^{(\widehat{T})}]) \leq \text{TV}(p_{\widehat{T}}^{S_{\text{init}}}, p_{\widehat{T}}^{S_{\text{target}}}) \leq \epsilon .$$

To bound (II), we apply Pinsker's and Theorem 4.2.3 to bound $\text{TV}(P^{\leftarrow}[S_{\text{target}}^{(\widehat{T})}], P_{\widehat{T}}^{\leftarrow, S_{\text{target}}})^2$ by

$$\text{TV}(P^{\leftarrow}[S_{\text{target}}^{(\widehat{T})}], P_{\widehat{T}}^{\leftarrow, S_{\text{target}}})^2 \leq \int_0^{\widehat{T}} \mathbb{E} \|\nabla \ln p_t(X_t^{S_{\text{target}}}) - \nabla \ln p_t^{S_{\text{target}}}(X_t^{S_{\text{target}}})\|^2 dt.$$

We have the following identity, that relates the difference in scores in terms of the conditional probability it belongs to $\Theta - S_{\text{target}}$:

Lemma 5.2.3. $\|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t\|^2 = \|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t^{\Theta - S_{\text{target}}}\|^2 \cdot \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^2$.

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 2.1.1, we see that for any $j \in S_{\text{target}}$,

$$\begin{aligned} \mathbb{E}_{p_t^j} \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^4 &\leq \sum_{\ell \in \Theta - S_{\text{target}}} \mathbb{E}_{p_t^j} \left[\frac{w_\ell p_t^\ell}{w_j p_t^j + w_\ell p_t^\ell} \right] \\ &\lesssim K \overline{W} \max_{\ell \in \Theta - S_{\text{target}}} \sqrt{1 - \text{TV}^2(p_t^\ell, p_t^j)} \\ &\lesssim K \overline{W} \epsilon^2, \end{aligned}$$

where in the last step we used the definition of T_{end}^S . By convexity, the same bound thus holds when the expectation on the left-hand side is replaced by an expectation with respect to $p_t^{S_{\text{target}}}$. By the same convexity argument, to bound $\mathbb{E} \|\nabla \ln p_t^{S_{\text{target}}}(X_t^{S_{\text{target}}}) - \nabla \ln p_t^{\Theta - S_{\text{target}}}(X_t^{S_{\text{target}}})\|^4$, it suffices to show that the expectations

$$\mathbb{E} \|\nabla \ln p_t^{S_{\text{target}}}(X_t^i) - \nabla \ln p_t^{\Theta - S_{\text{target}}}(X_t^i)\|^4$$

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

$$\begin{aligned} \mathbb{E} \|\nabla \ln p_t^{S_{\text{target}}}(X_t^{S_{\text{target}}}) - \nabla \ln p_t^{\Theta - S_{\text{target}}}(X_t^{S_{\text{target}}})\|^4 &\leq \mathbb{E}_i \mathbb{E} \max_{\substack{i, j \in S_{\text{target}} \\ \ell \in \Theta - S_{\text{target}}}} \|\nabla \ln p_t^j(X_t^i) - \nabla \ln p_t^\ell(X_t^i)\|^4 \\ &\leq K^3 \max_{\substack{i, j \in S_{\text{target}} \\ \ell \in \Theta - S_{\text{target}}}} \mathbb{E} \|\nabla \ln p_t^j(X_t^i) - \nabla \ln p_t^\ell(X_t^i)\|^4. \end{aligned}$$

Thus we can conclude by applying Lemma 5.2.2 and bound $\mathbb{E} \|\nabla \ln p_t^{S_{\text{target}}}(X_t^{S_{\text{target}}}) - \nabla \ln p_t(X_t^{S_{\text{target}}})\|^2$ by $O(\epsilon \sqrt{\overline{W}} K^2 (\overline{R}^2 + M^2 + \sqrt{M} \Psi^4 + \sqrt{M}) e^{-2t})$. Integrating over $[0, \widehat{T}]$ completes the proof. \square

5.2.2 Deferred proofs

To establish Lemma 5.2.2, which bounds the difference between two scores, we use the following general plan of attack. In general, the scores can be pretty ill-behaved, so it is easier to compare the difference in scores with the sum of the difference between the scores and a normal Gaussian at X by the triangle inequality. The difference between the score of a strongly log-concave distribution and Gaussian at an arbitrary point X can be controlled by bounding the Hessian $\nabla^2(-\ln p_t^i)$ (Lemma 5.2.4) and the score evaluated at 0 (Lemma 5.2.5). Throughout it will be useful to refer to the density of the multivariate Gaussian, so we let $f_{\mathcal{N}(\mu, \Sigma)}$ denote the density of $\mathcal{N}(\mu, \Sigma)$.

Lemma 5.2.2. Under Assumptions 5.2.1, 5.2.3, and 5.2.4, $\mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim e^{-4t} (\bar{R}^4 + M^4 + M\Psi^8 + \bar{M}) \quad \forall i, j, \ell \in \Theta$.

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 5.2.4,

$$\mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 \lesssim \mathbb{E}_{X \sim p_t^i} [\|\nabla \ln p_t^j(X)\|^4 + \|\nabla \ln p_t^\ell(X)\|^4] \lesssim \bar{M}.$$

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

$$\begin{aligned} \mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln p_t^\ell(X)\|^4 &\lesssim \mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^j(X) - \nabla \ln f_{\mathcal{N}(0, \text{Id})}(X)\|^4 \\ &\quad + \mathbb{E}_{X \sim p_t^i} \|\nabla \ln p_t^\ell(X) - \nabla \ln f_{\mathcal{N}(0, \text{Id})}(X)\|^4. \end{aligned}$$

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

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

The left term is the difference in the changes of the score under p_t^j versus $\mathcal{N}(0, \text{Id})$. By Lemma 5.2.4, we know $\nabla^2(-\ln p_t^j + \ln f_{\mathcal{N}(0, \text{Id})})$'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 $\nabla(\ln p_t^j - \ln f_{\mathcal{N}(0, \text{Id})})$ is globally $1000\Psi^2 e^{-2t}$ -Lipschitz. Combining with Lemma 5.2.5, which controls $\|\nabla \ln p_t^j(0)\|^4$, we can conclude

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

□

The lemma we use to control the the Hessian $\nabla^2(-\ln p_t^i)$ is based on two properties of the log-concave components: the preservation of strong log-concavity under convolution with another

Gaussians and the observation that the Hessian of a noised component reflects the Gaussian component.

Lemma 5.2.4. Under Assumption 5.2.1, the Hessian of $\ln p_t^i$ for $i \in \Theta$ is between

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

Proof. Using the preservation of strong log-concavity (see p.71 in [SW14] or [HÅ06]), we find that for $i \in \Theta$,

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

By Proposition 2.23 of [SW14], 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 [SW14] for the convolution $X_t^i = e^{-t}X_0^i + \mathcal{N}(0, (1 - e^{-2t})\text{Id})$. Let $X := e^{-t}X_0^i$, $Y := \mathcal{N}(0, (1 - e^{-2t})\text{Id})$, $Z := X_t^i$, and let p_X, p_Y, p_Z be their respective densities. Because

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

we can compute the Hessian with the product rule,

$$\begin{aligned} \nabla^2(-\ln p_Z)(z) &= \nabla \left\{ \mathbb{E}_{X \sim p_X} [p_Y(z - X) \cdot \nabla(-\ln p_Y(z - X))] \cdot \frac{1}{p_Z(z)} \right\} \\ &= -\mathbb{E}_{X \sim p_X} \left[p_Y(z - X) \nabla \ln p_Y(z - X) (\nabla \ln p_Y(z - X))^{\top} \right] \cdot \frac{1}{p_Z(z)} \\ &\quad + \mathbb{E}_{X \sim p_X} [p_Y(z - X) \nabla^2(-\ln p_Y(z - X))] \cdot \frac{1}{p_Z(z)} \\ &\quad + \mathbb{E}_{X \sim p_X} [p_Y(z - X) \nabla(\ln p_Y(z - X))] \cdot \frac{1}{p_Z(z)} \cdot \frac{\nabla p_Z(z)}{p_Z(z)}. \end{aligned}$$

We can simplify the above expression,

$$\begin{aligned} \nabla^2(-\ln p_Z)(z) &= -\mathbb{E}[\nabla \ln p_Y(Y) (\nabla \ln p_Y(Y))^{\top} | X + Y = z] + \mathbb{E}[\nabla^2(-\ln p_Y(Y)) | X + Y = z] \\ &\quad + (\mathbb{E}[\nabla \ln p_Y(Y) | X + Y = z])^{\otimes 2} \\ &= -\text{Var}(\nabla(-\ln p_Y(Y)) | X + Y = z) + \mathbb{E}[\nabla^2(-\ln p_Y(Y)) | X + Y = z] \\ &\preceq \frac{1}{1 - e^{-2t}} \text{Id}, \end{aligned}$$

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

Next we control the value of the score evaluated at 0 and show that it is decreasing exponentially quickly with respect to the time t . Note that our argument is based on the idea the distribution of the log-concave components approaches a normal Gaussian exponentially quickly. We rewrite the score in terms of $\nabla \ln p = \frac{\nabla p}{p}$, and differentiate with respect to the input. Over time t , the numerator goes to 0 via a comparison argument to the normal Gaussian, and the denominator approaches the value of the normal evaluated at 0 as well.

Lemma 5.2.5. For $t > 0.001$, we have the following inequality on the score at the origin,

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

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

$$\begin{aligned} \nabla \ln p_t^i(0) &= \frac{- \int_{\mathbb{R}^d} p_0^i(u) f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}(-ue^{-t}) \frac{0-ue^{-t}}{1-e^{-2t}} du}{\int_{\mathbb{R}^d} p_0^i(u) f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}(-ue^{-t}) du} \\ &= \frac{e^{-t}}{1-e^{-2t}} \frac{\int_{\mathbb{R}^d} p_0^i(u) f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}(-ue^{-t}) u du}{\int_{\mathbb{R}^d} p_0^i(u) f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}(-ue^{-t}) du}. \end{aligned}$$

Note that for all $t \geq 0.001$, $f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}$ is Ω -Lipschitz for some $\Omega > 0$. Thus, we can bound the distance between the numerator and $\mu_i f_{\mathcal{N}(0, (1-e^{-2t})\text{Id})}(0)$ with the triangle inequality and Assumption 5.2.3,

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

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

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

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

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

□

We also have this simple identity which rewrites the score function into the difference between the scores and a weighted sum of the densities evaluated at a given point. The first we know how

to control by Lemma 5.2.2 and decreases with t ; the second is near 0 under $[T_{\text{after}}^C, T_{\text{before}}^C]$.

Lemma 5.2.3. $\|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t\|^2 = \|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t^{\Theta - S_{\text{target}}}\|^2 \cdot \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^2$.

Proof. This follows by some simple algebraic manipulations:

$$\begin{aligned}
\|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t\|^2 &= \left\| \frac{\sum_{i \in S_{\text{target}}} w_i \nabla p_t^i}{\sum_{i \in S_{\text{target}}} w_i p_t^i} - \frac{\sum_{i \in \Theta} w_i \nabla p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right\|^2 \\
&= \left\| \left(\frac{1}{\sum_{i \in S_{\text{target}}} w_i p_t^i} - \frac{1}{\sum_{i \in \Theta} w_i p_t^i} \right) \sum_{i \in S_{\text{target}}} w_i \nabla p_t^i - \frac{\sum_{i \in \Theta - S_{\text{target}}} w_i \nabla p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right\|^2 \\
&= \left\| \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{(\sum_{i \in S_{\text{target}}} w_i p_t^i)(\sum_{i \in \Theta} w_i p_t^i)} \right) \sum_{i \in S_{\text{target}}} w_i \nabla p_t^i - \frac{\sum_{i \in \Theta - S_{\text{target}}} w_i \nabla p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right\|^2 \\
&= \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^2 \left(\left\| \frac{\sum_{i \in S_{\text{target}}} w_i \nabla p_t^i}{\sum_{i \in S_{\text{target}}} w_i p_t^i} - \frac{\sum_{i \in \Theta - S_{\text{target}}} w_i \nabla p_t^i}{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i} \right\|^2 \right) \\
&= \left(\frac{\sum_{i \in \Theta - S_{\text{target}}} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^2 \|\nabla \ln p_t^{S_{\text{target}}} - \nabla \ln p_t^{\Theta - S_{\text{target}}}\|^2. \quad \square
\end{aligned}$$

*It is by logic that we prove, but by intuition
that we discover.*

Henri Poincaré

6

Instantiating Theory for Different Models of Data

In Chapter 6, we apply our theory developed in the previous chapter to different models of data. Note that this is one of the primary advantages of our theoretical framework; unlike other statistical physics papers, we are not restricted to one toy model of data and can reason about different distributions within the same framework. Our framework reduces the problem of understanding critical windows to establishing certain bounds on statistical divergence and metrics, to which we can apply many well-known facts and relations.

We begin with Theorem 3.0.1 and note that it is abstract and in terms of the points of negligible overlap between the target sub-mixture and other components (T_{start}^S) and complete overlap of the initial and target sub-mixture (T_{end}^S); we seek to write them in more familiar forms in terms of the means and covariances of the individual components. We can find that the target sub-mixture and other components have negligible overlap at a certain time by appealing to certain concentration inequalities of log-concave measures. The task to identify when the initial and target sub-mixture start to overlap is more complicated. Even if two components have similar means and covariances, then they could have negligible overlap if the behavior around the mean is dissimilar. Thus we have to include the additional assumption that the components centered to 0 are close in Wasserstein-2 distance, which captures this idea.

Theorem 6.0.1 (Informal, see Theorem 6.1.1). Suppose p is a mixture of $1/\sigma^2$ -strongly log-concave distributions with means μ_1, \dots, μ_K , and let $S_{\text{init}} \subset S_{\text{target}}$. Suppose that for any $i \in S_{\text{target}}$ and any $j \notin S_{\text{target}}$, $\|\mu_i - \mu_j\| \gtrsim \sigma\sqrt{d}$. Then there is an upper bound for T_{end}^S , which is dominated by $\ln \max_{i \in S_{\text{init}}, j \in S_{\text{target}}} \|\mu_i - \mu_j\|$, and there is a lower bound for T_{start}^S , which is dominated by $\ln \min_{i \in S_{\text{target}}, j \notin S_{\text{target}}} \|\mu_i - \mu_j\|$.

Theorem 6.0.1 states that the interval where S_{init} specializes to S_{target} is between the log of the max distance between any component in S_{init} and any component in S_{target} and the log of the min distance between any component in S_{target} and any component in $\Theta - S_{\text{target}}$. In Section 6.1.2, we specialize Theorem 6.0.1 to a sparse coding setting where the means of the components are given by sparse linear combinations of a collection of incoherent dictionary vectors. The endpoints have a natural interpretation in terms of the *Hamming distances* between the sparse linear combinations defining the means within S_{init} , S_{target} , and $\Theta - S_{\text{target}}$.

We note a major issue with Theorem 6.0.1: because we use sub-Gaussian concentration inequalities, we need that components in S_{target} and any component outside S_{target} have means farther apart by a factor that depends on \sqrt{d} . This seems implausible if the means are from a natural data distribution in \mathbb{R}^d and lie on the unit sphere S^{d-1} . Appealing to the Hellinger distance instead of total variation, we can remove this caveat for mixtures of Gaussians:

Theorem 6.0.2 (Informal, see Theorem 6.2.1). Suppose p is a mixture of K identity-covariance Gaussians in \mathbb{R}^d with means μ_1, \dots, μ_K , and let $S_{\text{init}} \subset S_{\text{target}}$. Then there is an upper bound for T_{end}^S , which is dominated by $\ln \max_{i \in S_{\text{init}}, j \in S_{\text{target}}} \|\mu_i - \mu_j\|$, and there is a lower bound for T_{start}^S , which is dominated by $\ln \min_{i \in S_{\text{target}}, j \notin S_{\text{target}}} \|\mu_i - \mu_j\|$.

This allows us to develop a theory for hierarchical sampling for Gaussians in Section 6.2.1. We end this chapter with experiments of critical windows that we found in real-world diffusion models. We provide examples of critical windows for Stable Diffusion 2.1. in Sections 6.3.1 and 6.3.2. Finally, in Section 6.3.3, we show that our theory leads to a new method to probe the privacy properties of diffusion.

6.1 Estimating critical windows for general distributions

We now consider cases where we can provide concrete bounds on $T_{\text{end}}^S, T_{\text{start}}^S$ in terms of the first and second moments of the component distributions. Our bounds here hold independent of the Assumptions in Section 5.2.

6.1.1 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. This allows us to establish upper bounds on when the critical windows. We

will also need concentration inequalities on our components, and appeal to concentration inequalities for *sub-Gaussian distributions*, which include log-concave measures.

Definition 6.1.1 (Sub-Gaussian random vectors). A random vector $X \in \mathbb{R}^d$ is *sub-Gaussian* with variance proxy σ^2 if for all $u \in \mathbb{R}^d$, it holds that

$$\mathbb{E}[e^{u^\top(X - \mathbb{E}[X])}] \leq \exp\left(\frac{\sigma^2\|u\|^2}{2}\right), \quad \forall u \in \mathbb{R}^d.$$

Let $\text{subG}_d(\sigma^2)$ denote the set of probability distributions whose vectors are sub-Gaussian with variance proxies σ^2 .

Thus we obtain the following bounds on T_{end}^S and T_{start}^S :

Lemma 6.1.1. Let $\epsilon > 0$. For $i \in \Theta$, let \bar{p}^i denote the density of the i -th component of the mixture model p after being shifted to have mean zero. Suppose $\text{W}_2(\bar{p}^i, \bar{p}^j) \leq \Upsilon$ for all $i, j \in \Theta$. Then $T_{\text{start}}^S(\epsilon) \leq \left\{ \ln(w(S_{\text{init}}, S_{\text{target}}) + \Upsilon) + \ln \frac{1}{\epsilon} + \frac{1}{2} \ln 2 \right\} \vee 3$. Additionally, if $p_0^i \in \text{subG}_d(\sigma^2)$ for all $i \in \Theta$, then $T_{\text{end}}^S(\epsilon) \geq \ln \Delta(S_{\text{target}}) - \ln \sigma - \ln \sqrt{8d \ln 6 + 8 \ln 4/\epsilon^2} - \ln 3 - \frac{1}{2} \ln 8$.

Note that because all α -strongly log-concave distributions are sub-Gaussian with variance proxy $\Theta(1/\alpha)$, under Assumption 5.2.1 of Section 5.2 the above applies for $\sigma \asymp \Psi$. When the terms $\Upsilon, \Psi, 1/\epsilon$ are sufficiently small, our bounds on T_{end}^S and T_{start}^S are dominated by $\ln w(S_{\text{init}}, S_{\text{target}})$ and $\ln \Delta(S_{\text{target}})$ respectively. Recall that $w(S_{\text{init}}, S_{\text{target}})$ and $\Delta(S_{\text{target}})$ respectively correspond to the maximum distance between any two component means from S_{init} and S_{target} , and the minimum distance from S_{target} to the rest of the mixture. This, combined with our main theorem, has the favorable interpretation that as long as the separation between components within S_{init} and S_{target} is dominated by the separation between components in S_{target} vs. outside S_{target} , then there is a non-empty window of times $\hat{T} \in [T_{\text{end}}^S, T_{\text{start}}^S]$ such that the S_{init} -targeted reverse process from noise level \hat{T} results in samples close to S_{target} .

Proof of Lemma 6.1.1. For T_{end}^S , we apply Pinsker's inequality and a Wasserstein smoothing argument 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 $O(e^{-t}(w(S_{\text{init}}, S_{\text{target}}) + \Upsilon))$. Define h_t^ℓ to be the density of $e^{-t}X_t^\ell$ for $\ell \in \Theta$. 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 $i \in S_{\text{init}}$ and $j \in S_{\text{target}}$ we can control the KL via Lemma A.2.7,

$$\text{TV}(p_{T_{\text{end}}^S}^i, p_{T_{\text{end}}^S}^j) \leq \sqrt{\text{KL}(p_{T_{\text{end}}^S}^i || p_{T_{\text{end}}^S}^j)} \leq \text{W}_2(h_{T_{\text{end}}^S}^i, h_{T_{\text{end}}^S}^j).$$

We use a coupling argument to control $\text{W}_2(h_{T_{\text{end}}^S}^i, h_{T_{\text{end}}^S}^j)$. Let $\pi \in \Gamma(\bar{f}_0^i, \bar{f}_0^j)$ be the optimal cou-

pling, and define the coupling in $\Gamma(p_{T_{\text{end}}^S}^i, p_{T_{\text{end}}^S}^j)$ that samples $(X, Y) \sim \pi$ and returns $(e^{-T_{\text{end}}^S}(X + \mu_i), e^{-T_{\text{end}}^S}(Y + \mu_j))$. The cost of this coupling is

$$\begin{aligned} W_2(h_{T_{\text{end}}^S}^i, h_{T_{\text{end}}^S}^j) &\leq \sqrt{\mathbb{E} \|e^{-T_{\text{end}}^S}(X - Y) + e^{-T_{\text{end}}^S}(\mu_i - \mu_j)\|^2} \\ &\leq e^{-T_{\text{end}}^S} \sqrt{2(\mathbb{E}\|X - Y\|^2 + \|\mu_i - \mu_j\|^2)} \\ &\leq \sqrt{2}e^{-t} [\Upsilon + \|\mu_i - \mu_j\|]. \end{aligned}$$

Thus $\text{TV}(p_t^i, p_t^j) \leq \sqrt{2}[\|\mu_i - \mu_j\| + \Upsilon]e^{-t} \leq \epsilon$, and we can conclude by applying Lemma A.2.4 to obtain an overall bound on $\text{TV}(p_{T_{\text{end}}^S}^{S_{\text{init}}}, p_{T_{\text{end}}^S}^{S_{\text{target}}})$.

For T_{start}^S , we use sub-Gaussian concentration bounds to lower bound the TV between components in S_{end} and $\Theta - S_{\text{end}}$. By Lemma A.2.8, p_t^i is sub-Gaussian with variance proxy $2\sigma^2$ for all $t \geq 0$. For $i \in S_{\text{target}}, j \in \Theta - S_{\text{target}}$, $\|\mu_t^i - \mu_t^j\| > 3\sigma\sqrt{8d\ln 6 + 8\ln 4/\epsilon^2}$ implies $\text{TV}(p_t^i, p_t^j) \geq 1 - \epsilon^2/4$ by Lemma A.2.13.

□

6.1.2 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 $F = \{f_1, f_2, \dots, f_n\}$ unit vectors, such that for all distinct i, j , $|\langle f_i, f_j \rangle| \leq \delta$. 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 $T_{\text{end}}^S, T_{\text{start}}^S$ 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 $\underline{H}(S) = \min_{\ell \in S, j \in \Theta - S} H(i, j)$. By parameter setting with Corollary 6.1.1, we can write $T_{\text{end}}^S, T_{\text{start}}^S$ in terms of Hamming distances between classes.

Corollary 6.1.1. We have that $T_{\text{end}}^S(\epsilon) \leq 3 \vee \left\{ \ln \frac{1}{\epsilon} + \frac{1}{2} \ln 2 + \ln(R\sqrt{\overline{H}(S_{\text{init}}, S_{\text{end}}) + d^2\delta} + \Upsilon) \right\}$ and $T_{\text{start}}^S(\epsilon) \geq \ln \left(R\sqrt{\underline{H}(S_{\text{end}}) - d^2\delta} \right) - \ln(\sigma\sqrt{\tilde{S} + 1}) - \ln \sqrt{8d\ln 6 + 8\ln 4/\epsilon^2} - \ln 3 - \frac{1}{2} \ln 8$.

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

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

This completes T_{end}^S . For T_{start}^S , 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 $u \in \mathcal{S}^{d-1}$ the expectation $\mathbb{E}[\exp(su^\top X)]$,

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

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

6.2 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 $\sigma_{\max}^2(t) := \max_i \sigma_{\max}^2(\Sigma_i(t))$, $\sigma_{\min}^2(t) = \min_i \sigma_{\min}^2(\Sigma_i(t))$, and $\bar{R}(t) = e^{-t} \max_i \|\mu_i\|$. We also formalize the notion of well-conditioned Gaussians below.

Assumption 6.2.1. There exists $\underline{\lambda} \leq 1 \leq \bar{\lambda}$ such that for all $t \geq 0$, $\underline{\lambda} \text{Id} \preceq \Sigma_i(t) \preceq \bar{\lambda} \text{Id}$.

We can establish the critical windows in terms of these parameters. We defer the details of the below theorem to Section 6.2.2.

Theorem 6.2.1. Take any $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$. For sufficiently small ϵ , there exists $T_{\text{end}}^S, T_{\text{start}}^S$ such that $T_{\text{end}}^S \leq \frac{1}{2} \ln \left(2d \frac{\bar{\lambda} - \underline{\lambda}}{\underline{\lambda}} + \frac{1}{\underline{\lambda}} w(S_{\text{init}}, S_{\text{target}})^2 \right) + \ln \frac{1}{\epsilon}$ and also $T_{\text{start}}^S \geq \ln \Delta(S_{\text{target}}) + \frac{1}{2} \ln \underline{\lambda} - \ln 4 - \frac{1}{2} \ln \ln \left(\frac{\bar{\lambda} \sqrt{KW} [(\bar{\lambda} - \underline{\lambda})^2 (\bar{R}(0)^2 + \bar{\lambda} d) + \bar{R}(0)^2]}{\underline{\lambda}^2 \Delta(S_{\text{target}})^2 \epsilon^2} \right)$ and such that for any $\hat{T} \in (T_{\text{end}}^S, T_{\text{start}}^S)$, $\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}}) \lesssim \epsilon$.

To get intuition for these bounds, we state the simpler scenario where the covariances are the identity matrix.

Example 6.2.1. (K Gaussians with identity covariance) Let $\Sigma_0^i = \text{Id}$ for all $i \in \Theta$. Then, for any $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$, $T_{\text{end}}^S = \ln w(S_{\text{init}}, S_{\text{target}}) + \ln 1/\epsilon$ and $T_{\text{start}}^S = \ln \Delta(S_{\text{target}}) - \ln 4 - \frac{1}{2} \ln \ln \frac{\bar{R}(0)^2 \sqrt{KW}}{\epsilon^2 \Delta(S_{\text{target}})^2}$. The dominant terms are $\ln w(S_{\text{init}}, S_{\text{target}})$ and $\ln \Delta(S_{\text{target}})$, which depend on the intra- and inter-group distances of the means. In Fig. 6.2.1, 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.

To gain some intuition onto the effect of mixing weights on critical windows, we consider a simple mixture of two Gaussians with identity covariance.

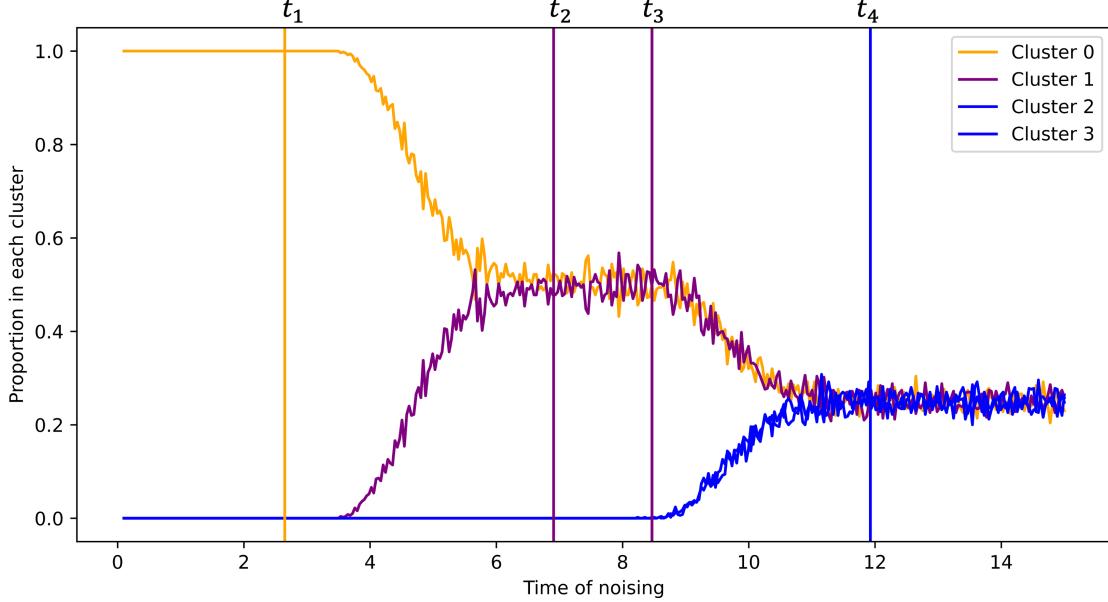


Figure 6.2.1: Example of critical times for an isotropic mixture of Gaussians ($\epsilon = 0.1$). For $t \leq t_1$, we sample from cluster 1. For $t \in [t_2, t_3]$, we sample from 0, 1. For $t \geq t_4$, we sample from 0, 1, 2, 3.

Example 6.2.2. (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

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

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

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

6.2.1 Defining a hierarchy of classes

In this section, we consider a *sequence* of critical windows that enable sampling from a sequence of nested sub-mixtures. Figure 6.2.1 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 6.2.1. 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 $h : V \rightarrow \mathbb{N}$ mapping vertices to their distance to the root and a function $f : V \rightarrow 2^\Theta \setminus \{\emptyset\}$ satisfying the following:

1. The root is the sub-population, $f(\text{root}) = \Theta$.
2. If u is a parent of v , $f(v) \subset f(u)$.
3. For distinct $i, j \in \Theta$ 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:

Theorem 6.2.2. Let all $\Sigma^i = \text{Id}$, $\|\mu_i\| = R$, and $w_i = \frac{1}{K}$. For $i \in \Theta$, consider the path $u_1, u_2, u_3, \dots, u_{H'}$ where u_1 is the leaf node with $i \in f(u_1)$ and $u_{H'}$ is the root. There exists $k \in [1, 2, \dots, H']$, sufficiently large R, H' , and sufficiently small ϵ such that there is a sequence of times $T_1 < T_2 < \dots < T_k$ with $\text{TV}(p^{\{i\}, T_\ell}, p^{f(u_\ell)}) \lesssim \epsilon$.

Proof. Using the notation from Example 6.2.1, let

$$(T_{\text{end}}^S)^j = \ln w(f(u_j), f(u_{j+1})) + \ln 1/\epsilon$$

$$(T_{\text{start}}^S)^j = \ln \Delta(f(u_{j+1})) - \ln 4 - \frac{1}{2} \ln \ln \frac{R^2}{\epsilon^2 \Delta(f(u_{j+1}))^2},$$

where j indicates that this is the j th critical window we are considering. It suffices to show that for a sufficiently large k , for all $j \leq k$, we have both $(T_{\text{start}}^S)^j - (T_{\text{end}}^S)^j > 0$ and $(T_{\text{end}}^S)^{j+1} - (T_{\text{start}}^S)^j > 0$. By our definition of the mixture tree, we know

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

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

We have $(T_{\text{end}}^S)^{j+1} - (T_{\text{start}}^S)^j > 0$ following from

$$(T_{\text{end}}^S)^{j+1} = \ln \left[(1 - \delta) \frac{R}{2^{(H'-j-1)^2}} \right] + \ln \frac{1}{\epsilon} \geq \ln \left[(1 + \delta) \frac{R}{2^{(H'-j-1)^2}} \right] - \frac{1}{2} \ln \ln \frac{R^2}{\epsilon^2 \Delta(f(u_{j+1}))^2} \geq (T_{\text{start}}^S)^j.$$

for sufficiently small ϵ . We have $(T_{\text{start}}^S)^j - (T_{\text{end}}^S)^j > 0$ if

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

$$\ln \frac{1 + \delta}{1 - \delta} + \ln \frac{1}{\epsilon} + \ln 4 + \frac{1}{2} \ln [2(H' - j - 1)^2 \ln 2 - 2 \ln(1 - \delta)\epsilon] \leq (2(H' - j) - 1) \ln 2.$$

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

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.

6.2.2 Deferred proofs for Gaussians

Score difference bound for Gaussian mixtures

Here we prove the following key ingredient in the proof of Theorem 6.2.1, in analogy to Lemma 5.2.2 in the proof of the main theorem for strong log-concave measures:

Lemma 6.2.3. For any nonempty $S \subset \Theta$ and $j \in S$, we have

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

We establish the proof as follows. We first explicitly compute the scores of p_t^S and $p_t^{\Theta-S}$, which by the definition of the score, decomposes into a linear function of x and the means of the individual components. We can bound the linear component by a factor which depends on the condition number and $\|x\|$, and we can bound the difference in means in terms of the radius $\bar{R}(t)$ and the minimum singular value. We will need this auxiliary result:

Lemma 6.2.4. Let $A, B \in \mathbb{R}^{d \times d}$ be two positive semi-definite matrices with singular values in $[\underline{\sigma}, \bar{\sigma}]$. For any $v \in \mathbb{R}^d$,

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

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

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

\square

Proof of Lemma 6.2.3. We explicitly compute $\nabla \ln p_t^S$ and $\nabla \ln p_t^{\Theta-S}$ and their difference,

$$\begin{aligned}\nabla \ln p_t^S &= \sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} ((\Sigma_i^t)^{-1}(x - \mu_i(t))) \\ \nabla \ln p_t^{\Theta-S} &= \sum_{i \in \Theta-S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} ((\Sigma_i^t)^{-1}(x - \mu_i(t))) \\ \nabla \ln p_t^S - \nabla \ln p_t^{\Theta-S} &= - \left(\sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} (\Sigma_i^t)^{-1} - \sum_{i \in \Theta-S} \frac{w_i p_t^i}{\sum_{j \in \Theta-S} w_j p_t^j} (\Sigma_i^t)^{-1} \right) x \\ &\quad + \left(\sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) - \sum_{i \in \Theta-S} \frac{w_i p_t^i}{\sum_{j \in \Theta-S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) \right).\end{aligned}$$

Both $\sum_{i \in S} \frac{w_i p_t^i}{\sum_{i \in S} w_i p_t^i} (\Sigma_i^t)^{-1}$, $\sum_{i \in \Theta-S} \frac{w_i p_t^i}{\sum_{i \in \Theta-S} w_i p_t^i} (\Sigma_i^t)^{-1}$ are positive semi-definite matrices with singular values in $[1/\sigma_{\max}^2(t), 1/\sigma_{\min}^2(t)]$. Thus, by Lemma 6.2.4, we can bound the first term in the difference in terms of the norm of x ,

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

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

$$\left\| \sum_{i \in S} \frac{w_i p_t^i}{\sum_{j \in S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) - \sum_{i \in \Theta-S} \frac{w_i p_t^i}{\sum_{j \in \Theta-S} w_j p_t^j} (\Sigma_i^t)^{-1} \mu_i(t) \right\| \lesssim \bar{R}(t)/\sigma_{\min}^2(t).$$

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

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

Combining these inequalities, we obtain

$$\begin{aligned}\mathbb{E}_{x \sim p_t^j} \left[\left\| \nabla \ln p_t^S - \nabla \ln p_t^{\Theta-S} \right\|^4 \right] &\lesssim (1/\sigma_{\min}^2(t) - 1/\sigma_{\max}^2(t))^4 (\bar{R}(t)^4 + \sigma_{\max}^2(t)^2 d^2) + \bar{R}(t)^4 / \sigma_{\min}^2(t)^4 \\ &\leq \frac{e^{-4t}}{\underline{\lambda}^4} \left[(\bar{\lambda} - \underline{\lambda})^4 (\bar{R}(0)^4 + \bar{\lambda}^2 d^2) + \bar{R}(0)^4 \right].\end{aligned}$$

□

Ratio inequality in terms of parameters of Gaussian mixture model

Here we prove the other key ingredient in the proof of Theorem 6.2.1. While originally we just directly use the total variation bound of Lemma 2.1.1 in the proof of the main theorem, now we write it explicitly in terms of the means and singular values of the Gaussian mixture model.

Lemma 6.2.5. For any $S \subset \Theta$ and $j \in S$, we have

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

To accomplish this lemma, we utilize the Hellinger distance bound in Lemma 2.1.1 and the following helper lemma, which relates the Hellinger distance to the parameters of a multivariate Gaussian.

Lemma 6.2.6. (p. 51 of [Par05]) Let $P \sim \mathcal{N}(\mu_P, \Sigma_P)$ and $Q \sim \mathcal{N}(\mu_Q, \Sigma_Q)$. Then,

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

Remark 6.2.1. Surprisingly, the Hellinger distance was crucial for establishing our critical windows bounds for Gaussians. This explicit formula allows us to write things in terms of the means and covariances of the Gaussians, which we could not directly do using total variation.

To upper bound this quantity, we need to control the $\frac{|\Sigma_P|^{1/4} |\Sigma_Q|^{1/4}}{\left| \frac{\Sigma_P + \Sigma_Q}{2} \right|^{1/2}}$ term in the above expression, which can be viewed as the ratio of the geometric and arithmetic means of two matrices. The below lemma shows we can upper bound this by a 1. Herein we argue this by reasoning about the multi-set of eigenvalues for all matrices involved, which was the methodology in [LC24]. It turns out that this is a consequence of the more general result that the log determinant is a convex function on the cone of positive semi-definite matrices.

Lemma 6.2.7. 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 multi-set of eigenvalues. Letting $\lambda_1, \lambda_2, \dots, \lambda_d > 0$ 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

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

□

Now we are ready to prove Lemma 6.2.5. We apply the Hellinger distance bound and the above lemma to simplify the RHS.

Proof of Lemma 6.2.5. Because $\mathbb{E}_{x \sim p_t^j} \left(\frac{\sum_{i \in \Theta - S} w_i p_t^i}{\sum_{i \in \Theta} w_i p_t^i} \right)^4 \leq \sum_{\ell \in \Theta - S} \mathbb{E}_{x \sim p_t^j} \left[\frac{w_\ell p_\ell^t}{\sum_{i \in \Theta} w_i p_t^i} \right]$, 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 $\ell \in \Theta - S$. Using the Hellinger distance bound in Lemma 8.2.1 and the computations in Lemmas 6.2.6 and 6.2.7, we have

$$\begin{aligned} \mathbb{E}_{x \sim p_t^j} \left[\frac{w_\ell p_\ell^t}{w_\ell p_\ell^t + w_j p_t^j} \right] &\leq \bar{W} \frac{|\Sigma_\ell(t)|^{1/4} |\Sigma_j(t)|^{1/4}}{\left| \frac{\Sigma_\ell(t) + \Sigma_j(t)}{2} \right|^{1/2}} \exp \left\{ -\frac{e^{-2t}}{8} (\mu_\ell - \mu_j)^\top \left[\frac{\Sigma_\ell(t) + \Sigma_j(t)}{2} \right]^{-1} (\mu_\ell - \mu_j) \right\} \\ &\lesssim \bar{W} \exp \left\{ -e^{-2t} \Delta(S_{\text{target}})/(8\bar{\lambda}) \right\}. \end{aligned}$$

□

Main theorem for Gaussians

We now prove the main theorem for Gaussians. We follow the same proof of Theorem 5.2.1, but use Pinsker's inequality to control the difference in initializations, $\text{TV}(p_{\hat{T}}^{S_{\text{init}}}, p_{\hat{T}}^{S_{\text{target}}})$, because it can be parameterized in terms of the difference in means and $\underline{\lambda}, \bar{\lambda}$. To control the the difference in scores under p and S_{target} , we directly apply Lemma 6.2.5.

Theorem 6.2.1. Take any $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$. For sufficiently small ϵ , there exists $T_{\text{end}}^S, T_{\text{start}}^S$ such that $T_{\text{end}}^S \leq \frac{1}{2} \ln \left(2d \frac{\bar{\lambda} - \underline{\lambda}}{\underline{\lambda}} + \frac{1}{\underline{\lambda}} w(S_{\text{init}}, S_{\text{target}})^2 \right) + \ln \frac{1}{\epsilon}$ and also $T_{\text{start}}^S \geq \ln \Delta(S_{\text{target}}) + \frac{1}{2} \ln \underline{\lambda} - \ln 4 - \frac{1}{2} \ln \ln \left(\frac{\bar{\lambda} \sqrt{KW} [(\bar{\lambda} - \underline{\lambda})^2 (\bar{R}(0)^2 + \bar{\lambda} d) + \bar{R}(0)^2]}{\underline{\lambda}^2 \Delta(S_{\text{target}})^2 \epsilon^2} \right)$ and such that for any $\hat{T} \in (T_{\text{end}}^S, T_{\text{start}}^S)$, $\text{TV}(p_{\text{init}, \hat{T}}, p_{\hat{T}}^{S_{\text{target}}}) \lesssim \epsilon$.

Proof. As in the proof of Theorem 5.2.1, we apply the data processing inequality to obtain

$$\text{TV}(p_{\text{init}, \hat{T}}, p_{\hat{T}}^{S_{\text{target}}}) \leq \text{TV}(P^\leftarrow[S_{\text{init}}^{\langle \hat{T} \rangle}], P^\leftarrow[S_{\text{target}}^{\langle \hat{T} \rangle}]) + \text{TV}(P^\leftarrow[S_{\text{target}}^{\langle \hat{T} \rangle}], P_{\hat{T}}^{\leftarrow, S_{\text{target}}}).$$

We begin with $\text{TV}(p_{\hat{T}}^{S_{\text{init}}}, p_{\hat{T}}^{S_{\text{target}}})$. By Lemma A.2.4, it suffices to show for any $i \in S_{\text{init}}, j \in S_{\text{target}}$, $\text{TV}(p_{\hat{T}}^i, p_{\hat{T}}^j) \leq \epsilon$. 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 $\underline{\lambda}, \bar{\lambda}$.

$$\begin{aligned} \text{TV}(p_{\hat{T}}^i, p_{\hat{T}}^j) &\leq \sqrt{\text{KL}(p_{\hat{T}}^i, p_{\hat{T}}^j)} \\ &= \sqrt{\ln \frac{|\Sigma^j(\hat{T})|}{|\Sigma^i(\hat{T})|} + d \left[\frac{1}{d} \text{tr}(\Sigma_j^{-1} \Sigma_i) - 1 \right] + (\mu_i(\hat{T}) - \mu_j(\hat{T}))^\top \Sigma^j(\hat{T})^{-1} (\mu_i(\hat{T}) - \mu_j(\hat{T}))} \\ &\leq \sqrt{d \left[\ln \left(\frac{e^{-2\hat{T}}\bar{\lambda} + 1 - e^{-2\hat{T}}}{e^{-2\hat{T}}\underline{\lambda} + 1 - e^{-2\hat{T}}} \right) + \frac{e^{-2\hat{T}}\bar{\lambda} + 1 - e^{-2\hat{T}}}{e^{-2\hat{T}}\underline{\lambda} + 1 - e^{-2\hat{T}}} - 1 \right] + \frac{1}{\bar{\lambda}} \|\mu_i - \mu_j\|^2 e^{-2\hat{T}}} \end{aligned}$$

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

$$\text{TV}(p_{\hat{T}}^i, p_{\hat{T}}^j) \leq \sqrt{2e^{-2\hat{T}} d(\bar{\lambda} - \underline{\lambda})/\bar{\lambda} + \frac{1}{\bar{\lambda}} \|\mu_i - \mu_j\|^2 e^{-2\hat{T}}} \leq \epsilon$$

Now we bound $\text{TV}(P^{\leftarrow}[S_{\text{target}}^{(\hat{T})}], P_{\hat{T}}^{\leftarrow, S_{\text{target}}})$. Following the main Cauchy-Schwarz split in Theorem 5.2.1, we can apply Lemmas 6.2.3 and 6.2.5 to control the score error for $t \in [0, \hat{T}]$,

$$\begin{aligned} &\mathbb{E} \left[\|\nabla \ln p_t^{S_{\text{target}}}(\bar{X}_t^{S_{\text{target}}}) - \nabla \ln p_t^\Theta(\bar{X}_t^{S_{\text{target}}})\|^2 \right] \\ &\lesssim e^{-2t} \frac{\sqrt{K\bar{W}} \left[(\bar{\lambda} - \underline{\lambda})^2 (\bar{R}(0)^2 + \bar{\lambda}d) + \bar{R}(0)^2 \right]}{\underline{\lambda}^2} \exp \left\{ -e^{-2t} \underline{\lambda} \Delta(S_{\text{target}})^2 / (16\bar{\lambda}) \right\}. \end{aligned}$$

The integral from 0 to \hat{T} is

$$\begin{aligned} &\int_0^{\hat{T}} \mathbb{E} \left[\|\nabla \ln p_t^{S_{\text{target}}}(\bar{X}_t^{S_{\text{target}}}) - \nabla \ln p_t^\Theta(\bar{X}_t^{S_{\text{target}}})\|^2 \right] dt \\ &\lesssim \frac{\sqrt{K\bar{W}\bar{\lambda}} \left[(\bar{\lambda} - \underline{\lambda})^2 (\bar{R}(0)^2 + \bar{\lambda}d) + \bar{R}(0)^2 \right]}{\underline{\lambda}^2 \Delta(S_{\text{target}})^2} \exp \left\{ -e^{-2T_{\text{upper}}} \Delta(S_{\text{target}})^2 / (16\bar{\lambda}) \right\} \lesssim \epsilon^2. \end{aligned}$$

□

6.3 Experiments on diffusion models

Having established our theory for critical windows, we will now provide some experimental support for our theory. In Sections 6.3.1 and 6.3.2, we provide real-life examples of critical windows for Stable Diffusion 2.1. Then, in Section 6.3.3, we will use our theoretical insights to develop a new method to probe the privacy of diffusion models.

6.3.1 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 6.3.2). We produced 250 images from SD2.1., using 500 time steps from the DDPM scheduler [HJA20b] 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 6.3.1). CLIP with the ViT-B/32 Transformer architecture labelled our images [RKH⁺21] 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 used the prompt with the largest dot product with the image according to CLIP as the feature label. 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.



Figure 6.3.1: Example images of cars generated by SD2.1 that we subsequently noised and de-noised to produce Figure 6.3.2.

6.3.2 Applications to fairness

Generative models can reproduce social biases with their outputs [LAMJ23]. Here we ask whether potentially biased features like gender have critical windows, as this could help design specific interventions to apply to diffusion model within that narrow range to improve image diversity [RA23]. We studied outputs of photo portraits of laboratory technician on SD2.1 [LAMJ23], sampled 200 images (see Figure 6.3.3 for examples), and created an analogous plot of critical times (Figure 6.3.4). 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.

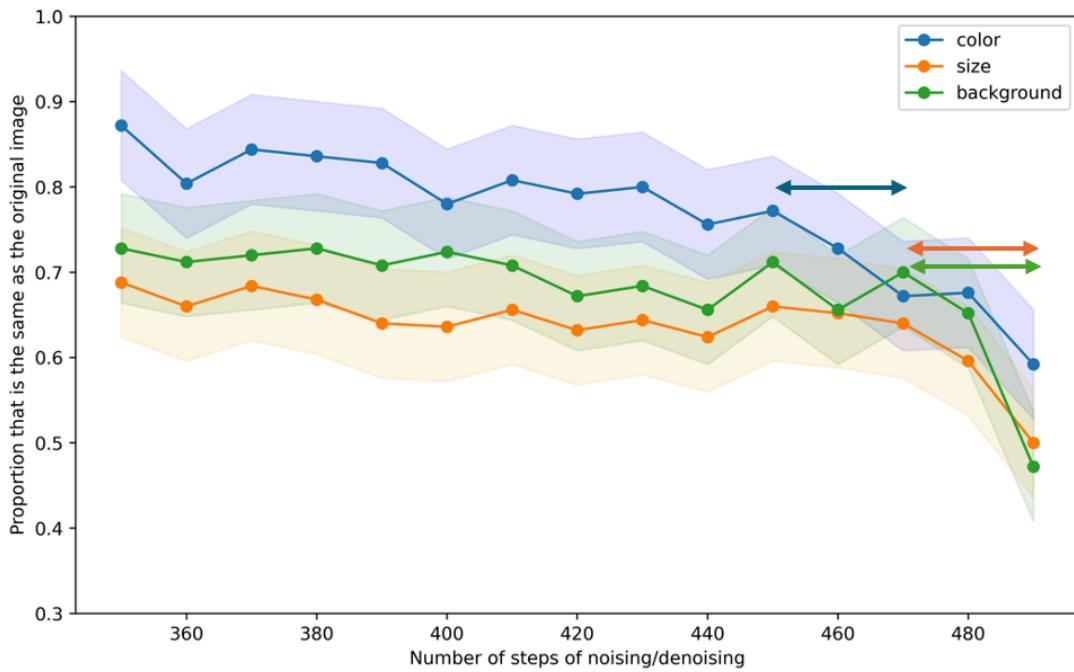


Figure 6.3.2: Percentage of agreement vs. noising amount in the experiment on images of cars generated by SD2.1 (see Section 6.3.1 for details). The critical window for each feature is demarcated with double-sided horizontal arrows.

This confirms the intuition of our hypothesis that different categories are well-separated before a critical window.



Figure 6.3.3: 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 6.3.4.

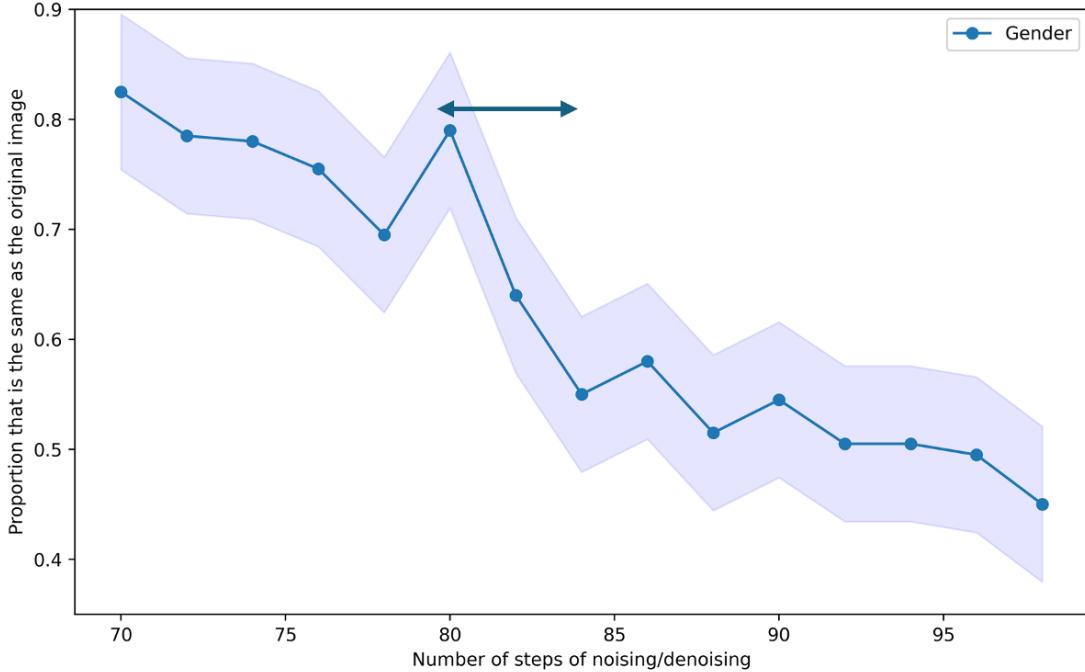


Figure 6.3.4: Critical window for gender feature in the experiment on images of laboratory technicians generated by SD2.1. Critical window demarcated with double-sided horizontal arrows.

6.3.3 Applications to privacy

We present a simple Membership Inference Attack (MIA), called **NoiseDenoise**, which is based on our critical windows framework. MIAs are a class of privacy attacks that try to identify whether a candidate sample belonged to training data [SSSS17], and are relevant for diffusion models because of their substantial privacy and copyright risks [CHN⁺23b].

MIA setup. Let Θ be the set of possible models and \mathcal{X} be the set of possible inputs, here the diffusion model and candidate image, 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 $x \sim \mathcal{D}$, and then ask the model to predict which dataset it came from.¹

NoiseDenoise. We rigorously describe our attack $\text{NoiseDenoise}(\mathcal{M}) : \Theta \times \mathcal{X} \rightarrow \mathbb{R}$ under the MIA framework. It is also parameterized by a $\widehat{T} \in (0, T)$ denoting 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 with the forward-reverse experiment on x up to time \widehat{T} , yielding samples $(\tilde{x}_i)_{i \in [N]}$. Our attack is the

¹Note that if the datasets are indistinguishable, i.e. $\text{TV}(\mathcal{D}_{\text{train}}, \mathcal{D}) = 0$ than the attack should do no better than random guessing and obtain error exactly 1. We can expound on this connection between inference and statistical divergences in Appendix A.1.1.

average ℓ^2 -norm of the difference between \tilde{x}_i and x , where we predict x to belong to the training data if $\mathcal{M}(\theta, x) \leq \tau$,

$$\mathcal{M}(\theta, x) = \frac{1}{N} \sum_{i=1}^N \|\tilde{x}_i - x\|.$$

Practically, we set $N = 10$ and $\frac{\hat{T}}{T} = 0.5$. Note that this method has already demonstrated some promising results in identifying whether an image was generated by a diffusion model [LW23]. We provide intuition into this method. A diffusion model θ implicitly defines a pushforward distribution $\theta_{\#}\mathcal{N}(0, \text{Id})$ on images. For a candidate image x , we can view $\theta_{\#}\mathcal{N}(0, \text{Id})$ 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 [BBdBM24]. [BBdBM24] 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 [LWWN23] or machine-generated text detection [MLK⁺23].

Evaluation metrics. We test all our methods with a denoising diffusion probabilistic model (DDPM) that was trained on CIFAR-10 in [DKW⁺23] on 1000 training and 1000 evaluation samples. As is standard in the MIA literature [DKW⁺23], we present receiving operating characteristic (ROC) curves, area under curve (AUC) statistics, and true positive rates (TPRs) at low false positive rates (FPRs).

Baselines. We compare NoiseDenoise to SecMI_{stat} and SecMI_{nn} from [DKW⁺23]. Both attacks exploit a deterministic approximation of the forward and reverse process of a DDPM to estimate the sampling error of a candidate image. SecMI_{stat} is the estimated error itself while SecMI_{nn} is a neural network trained on the errors at different timesteps.

Results. Both Figure 6.3.5 and Table 6.3.1 show that SecMI_{stat} and SecMI_{nn} outperform NoiseDenoise. However, NoiseDenoise is a statistically significant MIA, with a TPR at FPR = 0.05 of 0.14. Moreover, 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 SecMI_{stat} or SecMI_{nn} at the same FPR thresholds, suggesting NoiseDenoise can serve as a *complementary approach* to these

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

Table 6.3.1: For each attack, we report the AUC, TPR at FPR .01, and TPR at FPR .05.

methods.

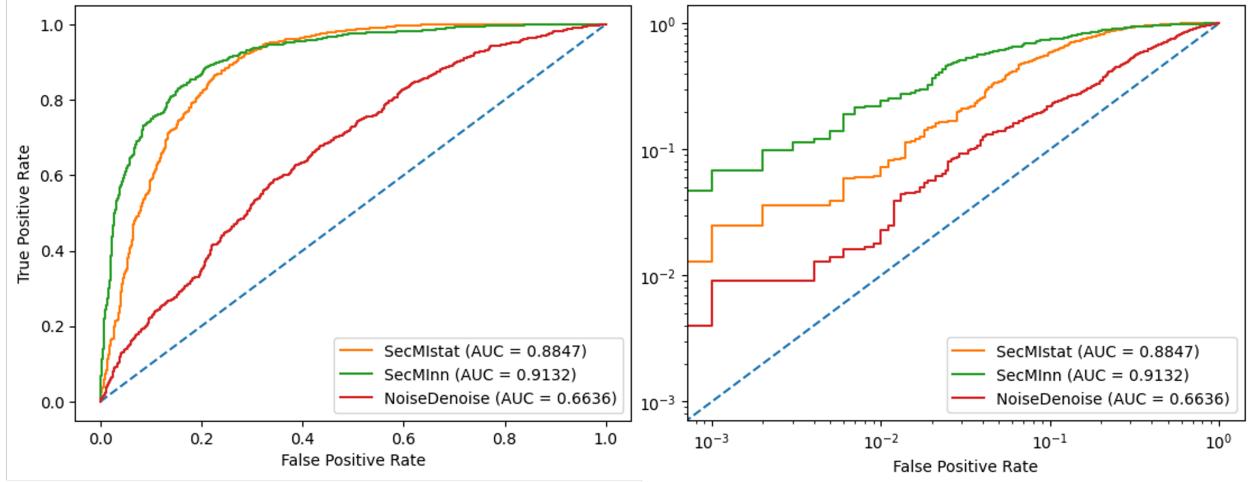


Figure 6.3.5: ROC curves of different methods.

Part II

Stochastic localization

As for everything else, so for a mathematical theory: beauty can be perceived but not explained.

Arthur Cayley

7

Introduction to Stochastic Localization

In Part II of this thesis, we adopt a different approach towards critical windows based on the perspective of *stochastic localization*. Broadly speaking, stochastic localization is a general problem solving framework for understanding the properties of high-dimensional measures. Analogous to the diffusion models we saw in earlier chapters, the main idea of stochastic localization is to perturb and degrade the underlying measure so that it becomes easier to analyze or sample from. We will trace the intellectual provenance of the stochastic localization formalism, starting at its background as a powerful theoretical tool and now a design framework for generative models.

In Section 7.1, we provide brief exposition on stochastic localization in its initial context as a theoretical tool for high-dimensional probability. Excitingly, in December 2024, it was one of the key ingredients of the proof of the Bourgain slicing theorem [KL24], which informally states that convex sets are guaranteed to have “fat” slices with constant volume. This has been an open question in probability theory and convex geometry for more than thirty years. Connecting with Part I, we will clarify how stochastic localization can also be interpreted as a reparameterization of the diffusion models that we saw earlier. Later works refined stochastic localization to its most essential components; in Section 7.2, we introduce the most general description of stochastic localization samplers developed thus far [Mon23a] and instantiate several popular generative models within this framework. Moving beyond these theoretical applications, this broad perspective for stochastic localization has led to concrete algorithmic innovations for practitioners, which we highlight in Section 7.3, by providing a universal characterization of generative models [HHY⁺25].

7.1 Bourgain, thin-shell, and Kannan, Lovász and Simonovits (KLS) conjectures

Introduced by Ronen Eldan in his Ph.D. thesis, stochastic localization was first developed as a proof technique within the probability theory community to establish functional inequalities for families of high-dimensional distributions [Eld13, Eld20]. In particular, this tool was crucial to making significant progress towards the Bourgain slicing theorem and thin-shell and Kannan, Lovász and Simonovits (KLS) conjectures [Che21, KL24]. These regard universal properties of certain well-behaved high-dimensional distributions.

Theorem 7.1.1 ([KL24, Bourgain slicing theorem, Theorem 1]). There exists a universal constant $\beta > 0$ such that for any convex set $K \subset \mathbb{R}^d$ with volume 1, there exists a hyperplane $H \subset \mathbb{R}^d$ such that

$$\text{Vol}_{n-1}(K \cap H) \geq \beta.$$

Conjectured in the 1980s, this recently proved theorem asks if every convex set K has a slice with constant volume.¹ The geometric intuition for this conjecture is that while one could imagine pathological counterexamples for arbitrary Borel-measurable sets, if every slice in a *convex set* has negligible area, then the set must have negligible volume as well. The next two conjectures, which would recover the Bourgain slicing theorem as a consequence if proven, hypothesize “nice” properties of high-dimensional strongly log-concave measures $\mu \in \text{SLC}(d)$. For convenience, let Γ_d denote the set of isotropic strongly log-concave measures in \mathbb{R}^d .

Conjecture 7.1.1 (Thin-shell conjecture). There exists a universal constant $\tau > 0$ so that for all $d > 0$ and $\mu \in \Gamma_d$,

$$\mathbb{E}_{X \sim \mu}(\|X\| - \sqrt{d})^2 \leq \tau.$$

The thin-shell conjecture states that the norm of $X \sim \mu$ concentrates around \sqrt{d} with constant deviation. This is consistent with the inequality one expects from Gaussian distributions. The KLS conjecture establishes lower bounds on how efficiently one can partition the space $S \sqcup S^C = \mathbb{R}^d$ according to a probability measure μ . The boundary measure $\mu^+(\partial S)$ is the “surface area” of the partition, and $\frac{\mu^+(\partial S)}{\min(\mu(S), \mu(S^C))}$ is the ratio of the surface area to the volume of one element of the partition.² The KLS conjecture states there is a lower bound to this ratio across all isotropic log-concave measures and possible partitions.

¹The paper employed a result [Gua24] that was published a week earlier.

²It is well-known we can restrict our attention to the case where $\mu(S) = \frac{1}{2}$, e.g. Theorem 1.8 in [Mil08].

Conjecture 7.1.2 (Kannan, Lovász and Simonovits). For a Borel-measurable set S , let S^ϵ be the ϵ -fattening of S , with $S^\epsilon \triangleq \{x \in \mathbb{R}^d \mid d(x, S) \leq \epsilon\}$. Now define the boundary measure μ^+ for μ by $\mu^+(\partial S) \triangleq \liminf_{\epsilon \rightarrow 0^+} \frac{\mu(S^\epsilon) - \mu(S)}{\epsilon}$. Letting $\psi(S, \mu) \triangleq \frac{\mu^+(\partial S)}{\min(\mu(S), \mu(S^c))}$, we define

$$\psi(\mu) \triangleq \inf_{S \subset \mathcal{B}(\mathbb{R}^d)} \psi(S, \mu)$$

to be the isoperimetric coefficient of μ . There exists a universal constant $\kappa > 0$ independent of dimension such that

$$\inf_{\mu \in \Gamma_d} \psi(\mu) \geq \kappa.$$

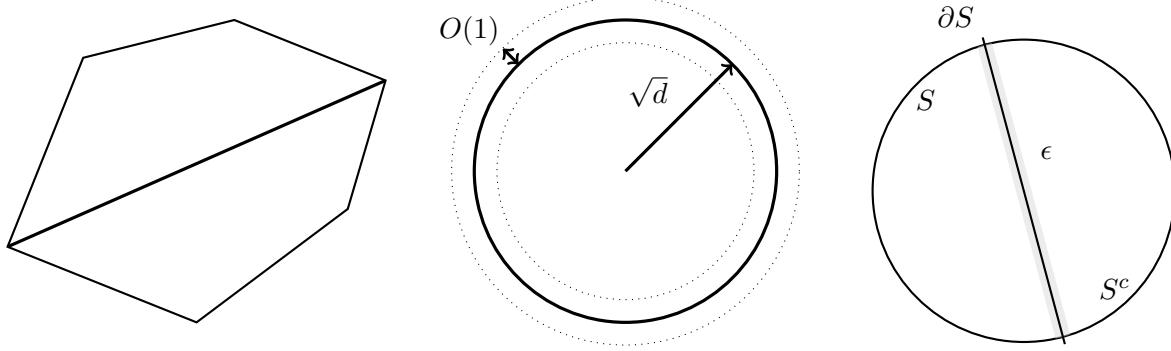


Figure 7.1.1: Visualizations of the Bourgain slicing theorem, thin-shell conjecture, and the KLS conjecture from left to right, respectively.

These conjectures have driven much progress in high-dimensional probability, functional analysis, and geometry in the past thirty years, and have inspired new methods like stochastic localization. Eldan's stochastic localization scheme showed that an isoperimetric coefficient κ for the KLS conjecture controls the universal constants of β and τ in the Bourgain slicing and thin-shell conjectures [Eld13, EK10]. Beyond connecting these conjectures together, Lee and Vempala [LV17] made some of the first progress towards the KLS conjecture in twenty years by showing that $\inf_{\mu \in \Gamma_d} \psi(\mu) \geq \frac{1}{d^{1/4}}$ via a localization scheme, and in a breakthrough result in 2021, Yuansi Chen [Che21] proved that it was possible to obtain $\inf_{\mu \in \Gamma_d} \psi(\mu) \geq \frac{1}{d^{o_d(1)}}$ with stochastic localization. This was the first result that lower bounded the isoperimetric coefficient by a factor larger than $\frac{1}{d^c}$ for constant $c > 0$.

Theorem 7.1.2 ([Che21, Theorem 1]). There exists a universal constant c' such that

$$\inf_{\mu \in \Gamma_d} \psi(\mu) \geq \frac{1}{d^{c' \left(\frac{\log \log(d)}{\log d} \right)^{1/2}}}.$$

Further work has improved the isoperimetric coefficient to a factor of $\frac{1}{\log^5(d)}$ [KL22] and $\frac{1}{\log^{3.2226}(d)}$ [JLV22]. In all of these conjectures, one wanted to establish some property of a measure μ without

an explicit functional form. The broad intuition of stochastic localization is as follows: while it seems difficult to directly establish that property on μ itself, it might be easier to first prove the property on a degraded version of μ and then connect that to the original measure μ . Thus we generate a Markov chain of random measures starting from μ such that desired statement is tractable on these random measures. In Eldan’s original stochastic localization scheme in particular, the degradation is an injection of Gaussian noise into a “tilt” of the measure.

Definition 7.1.1 (Eldan’s stochastic localization scheme [EK10]). Consider an initial base measure $\mu \in \Gamma_d$ with initial density p . For $t \geq 0, \theta \in \mathbb{R}^d$, we define the density

$$p_{t,\theta}(x) \propto e^{\langle \theta, x \rangle - t\|x\|^2/2} p(x).$$

Let $\mu(t, \theta) \triangleq \mathbb{E}_{x \sim p_{t,\theta}}[x]$ be the mean of the density $p_{t,\theta}$. We parameterize a stochastic process of random measures by the “tilt” $(\theta_t)_{t \geq 0}$ solving the following SDE,

$$d\theta_t = \mu(t, \theta_t) dt + dB_t, \quad t \geq 0, \theta_0 = 0$$

The parlance of “tilt” comes from the fact that the θ term in $p_{t,\theta}$ biases x to be aligned with θ . Therefore, at $t = 0$, the measure we start with is the initial $p_{0,0} = p$, and at $t \rightarrow \infty$, the measure *localizes* to a delta measure around θ , $p_{t,\theta} \approx \delta_\theta$. One of the more useful properties of this localization scheme is that $p_{t,\theta_t}(x)$ is a martingale, i.e. $p(x) = \mathbb{E}[p_{t,\theta_t}(x)]$ for all $x \in \mathbb{R}^d$. In the case of the KLS conjecture, this is useful because there are well-known results for isoperimetric coefficient for distributions more log-concave than Gaussians, including p_{t,θ_t} [Led01]. This means we want to find a sufficiently large time t such that p_{t,θ_t} has a good isoperimetric coefficient while making sure that $p_{t,\theta_t}(S)$ remains a constant value as well. Technically, this involves controlling certain higher-order *moments* of the distributions $p_{t,\theta}$. These moment control bounds have been central to the various breakthroughs in the application of stochastic localization: the main innovation of [Che21]’s argument was improved control over the several quantities like the covariance of $p_{t,\theta}$; Klartag and Lehec [KL24] proved in the affirmative the Bourgain slicing theorem by exploiting an improved moment bound by Guan [Gua24]; [KL22, JLV22] both improved on the original KLS isoperimetric coefficient bound by [Che21] by better controlling certain moments of the distribution.

Inspired by the progress stochastic localization has made towards functional inequalities, other authors [EAMS22, MW23, AMS23, Mon23b, HMP24] have applied and generalized Eldan’s stochastic localization framework to develop new sampling algorithms for distributions inspired by statistical physics. We will briefly overview one application.

Theorem 7.1.3 ([AMS23], Sampling from the Sherrington-Kirkpatrick (SK) measure, Theorem 2.1.]). We first choose a random symmetrix matrix $A \sim \text{GOE}(d)$, i.e. we select $A_{ij} \sim \mathcal{N}(0, 1/d)$ i.i.d. for $i < j \leq d$ and $A_{ii} \sim \mathcal{N}(0, 2/d)$ for $i \leq d$ and enforce $A_{ij} = A_{ji}$. Then, for an inverse temperature coefficient $\beta \geq 0$, we define the SK measure on $\{\pm 1\}^d$ by the density

$$\mu_A(x) \propto \exp \left\{ \frac{\beta}{2} \langle x, Ax \rangle \right\}.$$

For any $\epsilon > 0$, $\beta_0 < 1/2$, $\beta \leq \beta_0$, there exists an $O(n^2)$ -time algorithm which takes as input A and outputs samples with distribution μ_A^{alg} such that

$$W_2(\mu_A^{\text{alg}}, \mu_A) \leq \epsilon$$

w.p $1 - o_d(1)$ over the randomness in the selection of A .

For context, existing algorithms could not get inverse temperature $\beta > \frac{1}{4}$. Intuitively, the previous algorithms, which were based on non-localization schemes like Glauber dynamics [EKZ21], failed because at lower temperatures the measure fractures into many different modes; by annealing the modes together through Gaussian noising, stochastic localization can allow one to sample efficiently. The algorithm of [AMS23] is exactly an implementation of stochastic localization using a diffusion procedure. To achieve their result, [AMS23] needed to show that one could learn and run the SDE in Proposition 7.1.1 accurately. This required proving that it was possible to compute the means accurately in the drift, which they did via an approximate message passing algorithm, and that discretizing the SDE did not incur too much error.

It was later realized that [AMS23] was essentially learning a form of continuous diffusion and that Eldan's localization scheme can be reparameterized into continuous diffusion. In fact, we can express the change in θ_t as an SDE where the drift term is the posterior mean of a denoising problem and thus a linear combination of the score function (Proposition 7.1.1).

Proposition 7.1.1 (Diffusion and stochastic localization [KP21, Mon23b]). Consider the stochastic process given by

$$\tilde{\theta}_t = tX + B_t, \quad t \geq 0, X \sim \mu.$$

It can be shown that $(\tilde{\theta}_t)_{t \geq 0}$ has the same distribution as $(\theta_t)_{t \geq 0}$ and the probability density $p_{t,\theta}$ is the conditional distribution of X given $\tilde{\theta}_t = \theta$. Furthermore, $(\tilde{\theta}_t)_{t \geq 0}$ solves the following SDE

$$d\tilde{\theta}_t = \mathbb{E}[X|tX + \sqrt{t}G = \tilde{\theta}_t]dt + dB_t, \quad X \sim \mu, G \sim \mathcal{N}(0, \text{Id}) \text{ independently.}$$

Proof sketch. Notice that the term $e^{(\theta,x)-t\|x\|^2/2}$ in the definition of the density $p_{t,\theta}$ looks like the density of a Gaussian. In fact, one can verify that $p_{t,\theta}$ is exactly the conditional distribution of $X|tX + \sqrt{t}\sqrt{G} = \theta$ given $G \sim \mathcal{N}(0, \text{Id})$ and $X \sim \mu$. Thus the SDEs for $\theta_t, \tilde{\theta}_t$ are equivalent. The

agreement of the SDE unconditional on $X \sim \mu$ and the definition conditional on X can be found in [LS77, Section 7.4]. \square

Beyond these particular problems, Eldan's localization scheme has had other implications in theory, e.g. improving bounds on the mixing times of Glauber dynamics [EKZ21] and establishing novel finite-sample guarantees for high-dimensional central limit theorems in certain divergences [EMZ20]. Clearly, stochastic localization has proved extremely useful in theoretical contexts by providing a *unifying approach* to solving different problems in probability theory. The central idea of stochastic localization is the idea of perturbing and noising procedure for the measure. In the next section, we describe a generalization of Eldan's stochastic localization scheme that distills this to its essence, and applies to basically all modern generative models.

7.2 “Sampling, Diffusions, and Stochastic Localization”

We formally define the framework for stochastic localization samplers, following the paper “Sampling, Diffusions, and Stochastic Localization” by [Mon23a]. Let $X \sim p$ be a random variable over \mathbb{R}^d .³ We consider a sequence of random variables $(Y_t)_{t \in \mathbf{I}}$ with a compact index set $\mathbf{I} \subset [0, \infty) \cup \{\infty\}$. As t increases, Y_t becomes *less informative* and *degrades* the original information about X (Definition 7.2.1). As in [Mon23a], we will only consider *complete* observation processes, where information about the path $(Y_t)_{t \in \mathbf{I}}$ uniquely identifies X : for any measurable set $A \subset \mathbb{R}^n$, we require $\Pr[X \in A | (Y_t)_{t \in \mathbf{I}}] \in \{0, 1\}$. For the sake of simplicity, we will assume $0, \infty \in \mathbf{I}$ and Y_∞ is totally uninformative about X .

Note that our formulation of stochastic localization samplers differs slightly from that in [Mon23b] in several minor ways. Firstly, in that work the index set \mathbf{I} is not necessarily compact; while we assume compactness of \mathbf{I} , this still encapsulates most applications of generative models, in which the sample X is realized in finitely many steps. Secondly, our indexing of time is the reverse of that of in [Mon23a]; in that work, the Y_t 's become *more* informative about X as t increases. We make this choice purely for cosmetic reasons.

Definition 7.2.1. $(Y_t)_{t \in \mathbf{I}}$ is an *observation process* with respect to X if for any positive integer k and sequence $t_1 < t_2 < \dots < t_k \in \mathbf{I}$, the sequence $X \rightarrow Y_{t_1} \rightarrow Y_{t_2} \rightarrow \dots \rightarrow Y_{t_k}$ forms a Markov chain.

Because $X \rightarrow Y_{t_1} \rightarrow \dots \rightarrow Y_{t_k}$ is a Markov chain, its reverse $Y_{t_k} \rightarrow \dots \rightarrow Y_{t_1} \rightarrow X$ is also a Markov chain. To any such observation process one can thus associate a generative model as follows:

³These definitions are easily carried over to the setting where X lives in a discrete space.

Definition 7.2.2. Given observation process $(Y_t)_{t \in \mathbf{I}}$ and times $t_1 < \dots < t_m = \infty$ in \mathbf{I} , the associated stochastic localization sampler is the algorithm that generates a sample for X by first sampling Y_{t_m} and then, for $k = m-1, m-2, \dots, 0$, sampling from the posterior on Y_{t_k} conditioned on $Y_{t_{k+1}}$ by taking one step in the reverse Markov chain above, and finally sampling X conditioned on Y_{t_0} .

7.2.1 Examples of stochastic localization samplers

We present several kinds of popular generative models within the stochastic localization framework and explicate their forward and reverse processes. We will start with diffusion models.

Example 7.2.1 (Continuous Diffusion Models). For continuous diffusion models, the forward process progressively degrades samples $X \sim p$,

$$dX_t = -X_t dt + \sqrt{2} dB_t, \quad X_0 \sim p,$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion. The reversal of the forward process is given by

$$dX_t^\leftarrow = \{X_t^\leftarrow + 2\nabla \ln p_{T-t}(X_t^\leftarrow)\} dt + \sqrt{2} dB_t, \quad X_T^\leftarrow \sim p_T$$

where $(B_t)_{t \geq 0}$ is a Brownian motion. Because the drift of the SDE only depends on the current state, the forward process satisfies the Markov property in Definition 7.2.1, and because the scaling factor reduces the information from the original sample, the forward SDE degrades the X_0 until its contribution is negligible. Note that the reverse SDE is parameterized by the score $\nabla \ln q_{T-t}(X_t^\leftarrow)$ and is related to the posterior via Tweedie's formula. Thus it can be viewed as successively sampling from the posteriors.

The next model we consider are discrete diffusion models [LTG⁺22, AJH⁺21, GH23, HSW⁺22, LME24], which can be viewed as discrete versions of continuous diffusion models with analogous forward procedures.

Example 7.2.2 (Discrete Diffusion Models). Consider a set \mathcal{A} denoting the vocabulary and let $p \in \mathcal{A}^T$, and consider a forward process with index set $\mathbf{I} = [0, K] \cup \{\infty\}$, $Y_0 = X$, and $Y_t \in \mathcal{A}^T$ defined in the limit as follows,

$$\Pr(Y_{t+\Delta t} = a \mid Y_t = b) = \delta_{ab} + Q_t(b, a)\Delta t + O(\Delta t^2),$$

where $Q_t \in \mathbb{R}^{n \times n}$ are diffusion matrices with nonnegative non-diagonal entries and columns which sum to 0. $(Y_t)_{t \in \mathbf{I}}$ is also a Markov chain and as $t \rightarrow \infty$, Y_t is degraded until it is eventually uninformative about the original sample Y_0 . The reverse process is also given by a diffusion matrix

\bar{Q}_t defined by

$$\begin{aligned}\bar{Q}_t(y, x) &= \frac{p_t(y)}{p_t(x)} Q_t(x, y) \\ \bar{Q}_t(x, x) &= - \sum_{y \neq x} \bar{Q}_t(y, x).\end{aligned}$$

$t = 1.0$	[MASK]	[MASK]	[MASK]	[MASK]
$t = 0.5$	The	[MASK]	brown	[MASK]
$t = 0.0$	The	quick	brown	fox

Figure 7.2.1: Illustration of a discrete diffusion model with a masking forward procedure.

It is also easy to verify that autoregressive large language models (LLMs) are also examples of stochastic localization samplers where the forward process is a masking procedure.

Example 7.2.3 (Autoregressive Language Models). Consider a set \mathcal{A} denoting the vocabulary and let $p \in \mathcal{A}^T$, and consider a forward process with index set $\mathbf{I} = \{0, 1, 2, \dots, T\}$, $Y_0 = X$, and $Y_t \in \mathcal{A}^{T-t}$. For $t \in \mathbf{I}$, we let Y_t equal the last first $T - t$ tokens of X . Clearly this is a Markov Chain, and the reverse process is equivalent to next-token prediction.

It is important to note that not all generative models are stochastic localization samplers. While the most prominent and empirically successful generative models (autoregressive language models, continuous diffusion, and flow matching models) are stochastic localization samplers, some less popular generative models do not belong to the family of localization-based samplers. Generative adversarial networks [GPAM⁺14] and consistency models [SDCS23] both use a singular evaluation of a neural network to map noise into an image. However, these other models are either not widely used in practice anymore or based on localization-based samplers. Thus we will restrict our attention to this broad family of samplers in Part II without much concern.

Another nuance is that some localization-based samplers have deterministic counterparts. For example, there are ODE-based diffusion models and language models which sample at temperature 0, both of which do not have any randomness in the sampling procedure and thus cannot be stochastic localization samplers. This case can be discounted because the notion of critical windows is not meaningful if there is no randomness in the generation process.

7.2.2 Localization via amortization

The ubiquity of stochastic localization samplers for generative models across different modalities raises the question as to *why* these class of models are so useful for the task of generative modeling.

Perhaps the necessity for amortizing generative modeling can provide some helpful intuition.⁴ In generative models, we want to come up with a random function $q : \Omega \rightarrow \Omega$ which acts as the pushforward of a simple distribution Γ , e.g. $\mathcal{N}(0, \text{Id})$, to a more complex distribution p_0 , e.g. a distribution over realistic-looking images.⁵ However, this transport seems extremely complicated to learn one-shot, so it might make sense to *amortize* the computation of the generation process and instead learn a sequence of random functions $q_0, q_1, \dots, q_T : \Omega \rightarrow \Omega$ such that their composition $(q_0 \circ q_1 \cdots \circ q_T) : \Omega \rightarrow \Omega$ maps the simple distribution to the more complex one.

How do we *train* these intermediate q_i ? Ideally, we would be able to easily produce samples $(x, y) \sim \Omega \times \Omega$ and apply a differentiable loss function $\mathcal{L}(q_i(x), y)$ such that minimizing it learns the transition q_i . If we want to train them all simultaneously and thus sample x, y easily, it helps to fix the marginal distributions of $(q_0 \circ \cdots \circ q_t)$ for all $t \in [T + 1]$. This motivates defining a series of p_1, p_2, \dots, p_T from p_0 such that each q_i learns the pushforward $q_i \# p_i = p_{i-1}$ and $p_T \triangleq \Gamma$ is the easy distribution we can easily sample. Ensuring we sample pairs (x, y) easily, we require that p_i can be determined from samples of p_0 and the conditional distribution $p_{t-1}|p_t$ is simple to sample from. To train q_{i+1} , we thus take some $y \sim p_i$, sample from $x \sim \cdot|y$ using the conditional distribution $p_i|p_{i+1}$, and learn q_{i+1} . This is exactly the reversal of an observation process that characterizes stochastic localization samplers.

In the next section, our exposition will take a more concrete turn, and we will show how this perspective has been applied to prove certain universality results about the design space of generative models [HHY⁺25].

7.3 Universal characterization of generators

[HHY⁺25] introduced a framework for generative modeling across different modalities using arbitrary Markov processes, which they call *generator matching*. Generator matching is a specialization of the stochastic localization perspective to the setting where the index set \mathbf{I} is continuous and there exists a *generator* which describes an infinitesimal change in the distribution for a Markov Chain. They then prove a *universality result* for generators that allow them to characterize all possible generators on Euclidean and discrete spaces. Without a loss of generality, we set $\mathbf{I} = [0, 1]$ and restrict our attention to Feller processes, which satisfy some nice regularity properties (see Appendix A.3.1 for details). The presentation here will be a synthesis of the exposition from Chapter 6 of [Gal18] and [HHY⁺25].

The key insight of generators is that while we cannot rigorously define a derivative of a probability distribution, we can define the derivatives over expectations of functions of the distribution.

⁴This perspective was inspired by a talk given by Prof. Yilun Du in COMPSCI 2281 [Du24].

⁵This randomness might be useful for ensuring the outputs are more creative or diverse in some sense.

Definition 7.3.1. For every $f \in \mathcal{C}_0(\Omega)$, let

$$\mathcal{L}_t f(Y_t) = \lim_{h \rightarrow 0^+} \frac{\mathbb{E}[f(Y_{t+h})|Y_t] - f(Y_t)}{h}.$$

Letting D_t be the linear subspace of $\mathcal{C}_0(\Omega)$ where this result is continuous, the linear operator $\mathcal{L}_t : D_t \rightarrow \mathcal{C}_0(\Omega)$ is called the *generator* of a Markov process $(Y_t)_{t \geq 0}$.

It turns out these generators can uniquely characterize the Markov Chain as one can choose functions that measure the change in the distribution at different spots in Ω . Let $f_p(y) = \mathbf{I}(y \in B_\epsilon(p))$ be the indicator that y belongs to an ϵ -ball around $p \in \Omega$. The conditional expectation $\mathbb{E}[f_p(Y_{t+h})|Y_t]$ is the probability that $Y_{t+h} \in B_\epsilon(p)$. As we take the limit $h \rightarrow 0^+$ of the difference of $\mathbb{E}[f_p(Y_{t+h})|Y_t]$ and $f_p(Y_t)$, we get a sense of the *rate* at which the distribution moves into the ball. If we consider all such f_p for $p \in \Omega$ and take $\epsilon \rightarrow 0^+$, we should be able to characterize the changes in the distribution induced by Markov chain exactly. The insight of [HHY⁺25] is that one can actually *solve* for all generators on discrete spaces and \mathbb{R}^d and thus characterize all Markov chains on \mathbf{I} . For finite $|\Omega| < \infty$, this is simply the class of diffusion matrices we saw for discrete diffusion. For $\Omega = \mathbb{R}^d$, generators are the composition of the SDE processes we have seen before and a new class of jump processes. As a reminder, an SDE is given by the following shorthand,

$$dX_t = b_t dt + \sigma_t dB_t,$$

and we can compute the SDE of $f(X_t)$ by Itô's lemma (Lemma 4.1.1),

$$df(X_t) = \left(\langle \nabla f, b_t \rangle + \frac{1}{2} \text{Tr} [\sigma_t^\top \nabla^2 f \sigma_t] \right) dt + \langle \nabla f, \sigma_t dB_t \rangle.$$

Then the follow functional form is perhaps not surprising, because the generator corresponds to the dt components of Itô's lemma.

Definition 7.3.2 (Generator of an SDE). The generator of a SDE is given by

$$\mathcal{L}_t f = \langle \nabla f, b_t \rangle + \frac{1}{2} \text{Tr} [\sigma_t^\top \nabla^2 f \sigma_t],$$

where $b : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the drift (or flow) and $\sigma_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is the diffusion.

We can vary f so that we can exactly identify the b_t and σ_t of the original SDE. Thus the generator gives the SDE uniquely. The last kind of process, the jump process, is new to us and actually exhausts the possibilities of generators. Intuitively, this is because we can have a Markov process which changes the distribution continuously or discretely. If it is continuous, then we are summing an infinite number of some independent random variables, so by a central limit theorem argument, we can think of them as converging to a Brownian motion. This is more formally captured by Donsker's theorem (see [KHO05] for more details). If the change is discrete, then we

have a jump process, given by:

Definition 7.3.3 (Jump process). The generator of a jump process is given by

$$\mathcal{L}_t f(x) = \int [f(y) - f(x)] J_t(y|x),$$

where $J_t(y|x)$ is the jump measure which has finite support.

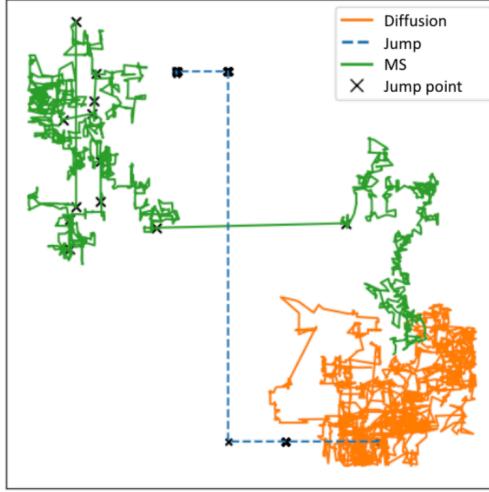


Figure 7.3.1: Examples of SDE and jump processes from [HHY⁺25].

Now we are ready to universally characterize generators in Theorem 7.3.1. Like for diffusion, the unifying perspective has expanded the design space of models for practitioners and led to more powerful models (see Section 6 of [HHY⁺25] for more details). The family of jump processes was unknown before it was introduced in [HHY⁺25], and compositions of jump processes with more traditional diffusion models have achieved state-of-the-art performance in protein structure and amino acid sequence generation.

Theorem 7.3.1 (Theorem 1 of [HHY⁺25]). Under some simple regularity assumptions, we can completely characterize the generators of the Markov process $(X_t)_{t \in [0,1]}$.

1. If the state space $|\Omega| < \infty$, the generator \mathcal{L}_t is given by transition matrices $(Q_t)_{0 \leq t \leq 1}$ with $Q_t \in \mathbb{R}^{d \times d}$.
2. If the state space $\Omega = \mathbb{R}^d$, the generator \mathcal{L}_t is the sum of a SDE and jump process:

$$\mathcal{L}_t f(x) = \langle f(x), u_t(x) \rangle + \frac{1}{2} \text{Tr} [\sigma_t^\top \nabla^2 f(x) \sigma_t] + \int [f(y) - f(x)] J_t(y|x).$$

The proof of the theorem can be found in [HHY⁺25] is based on the observation that generators satisfy an “almost positive” property, which itself implies a universality property. There is also a

converse condition which states that all such generators have an associated sampling algorithm that implements a Markov process with the desired generator [HHY⁺25, Algorithm 2].

*When you change the way you look at things,
the things you look at change.*

Max Planck

8

Critical Windows for Stochastic Localization Samplers

In this chapter, we extend our theoretical results for critical windows in Chapter 5 [LC24] for continuous diffusions and mixtures of strongly log-concave distributions to all stochastic localization samplers and mixture models. The key insight of our approach is to exploit the abstraction of stochastic localization that we developed in the previous chapter. The main argument was that critical windows are the transition between different modes in the distribution. For example, given a diffusion model which produces pictures of a cat or dog, the trajectory of an image which eventually becomes a cat shifts between producing images of cats or dogs with equal probability to only producing images of cats with high probability. Thus to characterize critical windows, one needs to rigorously describe when the diffusion model is “producing images of cats or dog” or “producing images of cats”, and more generally, when a stochastic localization sampler is sampling from some sub-population or when it is sampling from a smaller sub-population. In diffusion, we established an interval $[T_{\text{end}}^S(\epsilon), T_{\text{start}}^S(\epsilon)]$ ¹ for when a model samples from a sub-population by a *forward-reverse experiment*. In this chapter, we will generalize the notion of a forward-reverse experiment to all localization-based samplers and show that in the interval $[T_{\text{end}}^S(\epsilon), T_{\text{start}}^S(\epsilon)]$ the model is sampling from the appropriate sub-population under the forward-reverse experiment.

As a reminder, in Chapter 5, this interval was proved by the following theorem that characterizes when the diffusion model settles on the feature defined by the subset $p^{S_{\text{target}}}$ when we consider the forward-reverse experiment initialized with the distribution $p^{S_{\text{init}}}$, called $p^{S_{\text{init}}, \hat{T}}$. The upper bound can be safely thought of as ϵ multiplied by a factor which grows polynomially with the dimension.

¹Remember as we take more steps with the sampler the time t decreases.

Theorem 5.2.1. Let $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$. For $\epsilon > 0$, if $\hat{T} \geq T_{\text{end}}^S(\epsilon)$ and $\hat{T} \leq T_{\text{start}}^S(\epsilon)$, then

$$\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}}) \lesssim \epsilon \sqrt{W} K^2 (\bar{R}^2 + M^2 + \sqrt{M} \Psi^4 + \sqrt{M}).$$

The main tool we exploited in the proof of Theorem 5.2.1 was Theorem 4.2.3, an approximation argument based on Girsanov's theorem from [CCL⁺23b], and it was used to show that sampling from $p_{\hat{T}}^{S_{\text{target}}}$ and running the SDE with the score induces by p had a similar output distribution to sampling from $p_{\hat{T}}^{S_{\text{target}}}$ and then running the SDE with the score induces by $p^{S_{\text{target}}}$.

Theorem 4.2.3 ([CCL⁺23b, Section 5.2]). Let $(Y_t)_{t \in [0, T]}$ and $(Y'_t)_{t \in [0, T]}$ denote the solutions to

$$\begin{aligned} dY_t &= b_t(Y_t) dt + \sqrt{2} dB_t, & Y_0 &\sim p, \\ dY'_t &= b'_t(Y'_t) dt + \sqrt{2} dB_t, & Y'_0 &\sim p. \end{aligned}$$

Let p and q' denote the laws of Y_T and Y'_T respectively. If b_t, b'_t satisfy

$$\int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt < \infty,$$

then

$$\text{KL}(q \| q') \leq \int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt,$$

where Q_t is the law of Y_t .

Notably, the KL error bound in this inequality scales polynomially with d , due to both the inner expectation $\mathbb{E}_Q[\|b_t(Y_t) - b'_t(Y_t)\|^2]$ and the accumulation over the time in the integral. This leaks into the statement of Theorem 5.2.1 and is the cause of the costly dimensional-dependence in the error bound. This theorem also only applies to SDEs and not all stochastic localization samplers.²

In this chapter, we obviate these difficulties by avoiding a Girsanov-style approximation argument in Theorem 5.2.1. We appeal directly to the distribution of the posterior of the final samples Y_0 at time 0 conditioning on the position of intermediate Y_t at time t to prove our improved Theorem 8.0.1. In other words, at a given time t , we look ahead at the potential locations of its future samples at time 0, and instead of bounding the difference between the path measures of p (red) and $p^{S_{\text{target}}}$ (blue), we bound the difference between the induced output distributions (Figure 8.0.1).³ Interpreting the score as an infinitesimal approximation of the posterior of the next step in the sampler of the SDE via Tweedie's formula, this essentially means that we *marginalize out the local*

²There has been some work developing analogues of Theorem 4.2.3 for discrete diffusion models [RCRY25] and the KL chain rule provides a similar guarantee for autoregressive LLMs, there is no generalization of Theorem 4.2.3 that could be universally applied to all localization-based samplers, and the error in all of these bounds still accrues over time, causing the same issues as before.

³While in practice, one would need to usually run the sampler from time t to 0 to obtain the distribution of $Y_0 | Y_t$, this distribution is still well-defined and can be reasoned about mathematically.

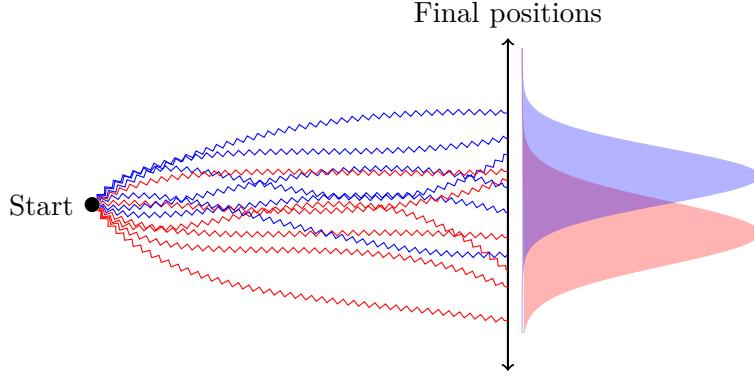


Figure 8.0.1: Difference between our approach in Chapter 5. Previously we bounded the path measures of the SDE defining p_{after}^S (blue) and p (red), and now we take a our destination-based approach. We bound directly bound the difference in the distributions of the destinations.

differences of the sampler induced by the distributions $p^{S_{\text{target}}}$ and p and instead bound the difference between the distributions of the destinations instead of the distributions over the path measures. This approach is much simpler than our previous one, requiring no stochastic calculus, and applies to all localization-based samplers, not just samplers based on SDEs like diffusion. Importantly, the difference in the distributions is also *bounded*, in the sense that the density functions integrate to a constant over the base measure, so we avoid the problematic dimensional dependencies that we suffered earlier.

Theorem 8.0.1. Let $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$ and define

$$W = \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_\theta}{\sum_{\theta \in S_{\text{target}}} w_\theta}.$$

For $\epsilon > 0$, if $\widehat{T} \in \mathbf{I} \cap [T_{\text{end}}^S, T_{\text{start}}^S]$, then

$$\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}}) \leq \epsilon \cdot \left(1 + \frac{\max(1, W)}{\sqrt{2}}\right).$$

Our result is a strict improvement of the previous bound along several important axes. Firstly, our results apply to all stochastic localization samplers, not just diffusions. Secondly, the previous result needed to assume that the components of p were strongly log-concave and that the score, i.e. gradient of the log-density, of p_t was Lipschitz and moment-bounded for all t . Finally, the old final bound includes a polynomial dependence on the moments of the score, which scale with the dimension d ; in contrast, our final bound is independent of d . This improved main result applies to all of our subsequent results.

The chapter is structured very similarly to Chapter 5, in that we introduce the basic forward-reverse and then present our main proof. In Sections 8.1 and 8.2, we introduce the new notation

for the forward-reverse experiment under the localization-based sampler framework and the proof of Theorem 8.0.1.

8.1 Forward-reverse experiment (redux)

To study this phenomenon of feature speciation in localization-based samplers, we generalize the *forward-reverse* experiment from Chapter 3. Like in that chapter, a forward-reverse experiment considers the amount of “noise” one would need to add to a generation so that running the generative model starting from the noised generation would still yield a sample with the same feature. Recall that for a diffusion model, this could mean taking an image of a cat, adding Gaussian noise, and resampling to see if the result is still a cat [LC24, SFW25, SFLW24]. In the context of language models, it could mean truncating a story about a cat, as the forward procedure is the operation of truncation, and resampling to check if the story remains about a cat.

8.1.1 Features and Experiments

Again we assume that the distribution of generations $X \sim p$ is a *mixture model* with a discrete set of classes $\Theta = \{\theta_1, \dots, \theta_K\}$ with non-negative weights w_1, \dots, w_K summing to 1. Each $\theta_i \in \Theta$ has a probability density function p^{θ_i} .

Remark 8.1.1. Note that the definition of Θ is extremely flexible and can be tailored to the particular data modality, model, or task. For example, Θ could be **{cat, dog}** for image diffusion models; **{right, wrong}** for math and reasoning tasks; **{unsafe, safe}** for jailbreaks.

Here we study a *family of observation processes* corresponding to observation processes for different initial distributions of $X \sim p^S$ for $S \subset \Theta$. To ensure that we can meaningfully compare the observation processes within this family, we will assume that the *forward degradation procedure is fixed*. To formalize this intuition, we borrow the language from diffusion models of a forward process, which degrades X , and a reverse process, which takes a degraded Y_t and produces X .

8.1.2 Forward-reverse experiment

Now we describe the general formalism under which we will study critical windows for localization-based samplers. The notation will be extremely similar to the earlier work on diffusions, but we place it within our stochastic localization sampler framework. Fixing some $t \in \mathbf{I}$ and $S \subset \Theta$, we start with some $X \sim p^S$, sample $Y_t|X$ from the observation process conditioning on X , and finally take $X'|Y_t$ from the stochastic localization sampler conditioning on Y_t .

Forward process. For any $t \in \mathbf{I}$, define the forward Markov transition kernel $\Pr_t^\rightarrow[A|X] = \Pr[Y_t \in A|X]$. Note the forward Markov transition kernel does not depend on the distribution of X . The fact that the forward process is agnostic to the specifics of the original distribution is shared by the most widely used stochastic localization samplers. For example, in diffusion and flow-matching

models, the forward transition is a convolution of X with a Gaussian; in autoregressive language models, it is masking of the last remaining token in the sequence. For any $t \in \mathbf{I}$ and $S \subset \Theta$, we let p_t^S denote the law of Y_t^S , where we sample $X^S \sim p^S$ and then sample $Y_t^S \sim \text{Pr}_t^\leftarrow[\cdot | X^S]$. We omit the Θ in p_t^Θ .

Reverse process. For any $t \in \mathbf{I}$ and initial distribution $X \sim p$, we define the posterior of X given Y_t by $\text{Pr}^\leftarrow[A | Y_t] = \text{Pr}_{X \sim p}[X \in A | Y_t]$, that is, the distribution of X given by starting the sampling process at $t \in \mathbf{I}$ and Y_t instead of ∞ and Y_∞ . We will also use this notation for the probability density.

Now, we are ready to describe the main forward-reverse experiment that we will study.

Definition 8.1.1 (Forward-reverse experiment [SFW25, SFLW24, LC24]). For nonempty $S \subset \Theta$ and $\hat{T} \in \mathbf{I}$, let $p^{S,\hat{T}}$ be the distribution of $X^{S,\hat{T}}$ defined by the following procedure:^a

1. Sample $Y_{\hat{T}}^S \sim p_{\hat{T}}^S$ — i.e. run the forward process for time \hat{T} starting at the sub-mixture p^S .
2. Sample $X^{S,\hat{T}} \sim \text{Pr}^\leftarrow[\cdot | Y_{\hat{T}}^S]$ — i.e. run the reverse process starting at $Y_{\hat{T}}^S$ to sample from the posterior on X .

^aNote that this equips 2^Θ with the structure of a poset, i.e. $A \preceq B$ if and only if there exists some $t \in \mathbf{I}$ such that running the forward-reverse experiment up to t from p^A yields p^B .

The second step is essentially applying the model on the interval $\mathbf{I} \cap [0, \hat{T}]$ starting with $Y_{\hat{T}}^S$ and we emphasize that in the second step, we are running the reverse process with the prior on X given by the *entire distribution* p rather than the sub-mixture p^S . This is equivalent in the context of diffusion of running the SDE with the score given by p instead of p^S . If we did the latter, the marginal distribution of the result would simply be p^S . Instead, the marginal distribution of $X^{S,\hat{T}}$ is some distribution whose relation to p and sub-mixtures thereof is *a priori* unclear. Intuitively, as $\hat{T} \rightarrow 0$, this distribution converges to p^S , and as $\hat{T} \rightarrow \infty$, this distribution converges to p .

Remark 8.1.2. Like for the diffusion case, one could also alternatively imagine re-running the sampling procedure at an intermediate point Y_t [GVS⁺23, BBdBM24, RA23]. In the case where the forward process is deterministic, i.e. autoregressive language models which have a deterministic forward masking procedure, these frameworks are actually equivalent.

Instantiation for LLMs

For intuition about what the forward and reverse experiment actually means, consider an autoregressive language model, which produces stories of cats or dogs. For LLMs, the forward-reverse experiment means masking the last \hat{T} tokens of a sample and then resampling with the same model. If \hat{T} is small, the story will likely still mention a cat and resampling will yield a story about a cat. If \hat{T} is large, then the first appearance of cat may be truncated, so resampling could produce a

story about a dog as well. Thus there is an interval where we have to transition between sampling from both stories of cats and dogs to just stories of cats.

8.2 Main theorem for feature specialization (redux)

Let $S_{\text{init}} \subset \Theta$ denote some sub-mixture, corresponding to a sub-population of p that possesses a certain property. Let $S_{\text{target}} \supset S_{\text{init}}$ denote some sub-mixture containing S_{init} . For instance, if p corresponds to some autoregressive model, S_{init} might correspond to sentences which correctly answer a particular math question and S_{target} might correspond to all possible responses to the math question, including incorrect ones.

We are interested in the following question: if we run the forward-reverse experiment for time \widehat{T} starting from $p^{S_{\text{init}}}$, is there some range of times for which the resulting distribution is close to $p^{S_{\text{target}}}$? That is, can we characterize the \widehat{T} for which

$$\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}})$$

is small?

As a reminder, this is useful for understanding critical windows because such \widehat{T} represents the time at which the *the sampler has not yet localized the features that distinguish $p^{S_{\text{init}}}$ from the larger sub-mixture $p^{S_{\text{target}}}$* . When there is a shift from localizing the features S_{target} to the features S_{init} , we say there is a critical window. We now formally state and prove our main result.

8.2.1 Main result for critical windows

For an error parameter $0 < \epsilon < 1$, we again define

$$\begin{aligned} T_{\text{start}}^S(\epsilon) &= \sup\{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{target}}}, p_t^{\Theta - S_{\text{target}}}) \geq 1 - \epsilon^2\} \\ T_{\text{end}}^S(\epsilon) &= \inf\{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{target}}}) \leq \epsilon\}. \end{aligned}$$

This is well-defined for continuous observation processes. When the value of ϵ is understood, we abbreviate the above with T_{start}^S and T_{end}^S .

Remark 8.2.1. For general stochastic localization schemes, we can only ask that $T_{\text{start}}^S(\epsilon) \in \{r \in \mathbf{I} : \text{TV}(p_r^{S_{\text{target}}}, p_r^{\Theta - S_{\text{target}}}) \geq 1 - \epsilon^2\}$ and $T_{\text{end}}^S(\epsilon) \in \{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{target}}}) \leq \epsilon\}$ instead of sup, inf like in Chapter 5, because the sets $\{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{target}}}, p_t^{\Theta - S_{\text{target}}}) \geq 1 - \epsilon^2\}$, $\{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{target}}}) \leq \epsilon\}$ may not necessarily be closed for discontinuous observation processes. As a pathological counterexample, we can imagine stitching the observation processes for $[0, 1)$ and $[2, 3)$ in diffusion together, which means the set defining T_{start}^S is not closed. However, for important localization-based samplers like autoregressive LLMs and continuous diffusion models, the observation process is continuous, so we will largely elide these technicalities and assume that these sets are closed in the remainder of this section.

Our main result is that in $\hat{T} \in \mathbf{I} \cap [T_{\text{end}}^S, T_{\text{start}}^S]$, the distance $\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}})$ is small:

Theorem 8.0.1. Let $S_{\text{init}} \subset S_{\text{target}} \subset \Theta$ and define

$$W = \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_\theta}{\sum_{\theta \in S_{\text{target}}} w_\theta}.$$

For $\epsilon > 0$, if $\hat{T} \in \mathbf{I} \cap [T_{\text{end}}^S, T_{\text{start}}^S]$, then

$$\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}}) \leq \epsilon \cdot \left(1 + \frac{\max(1, W)}{\sqrt{2}}\right).$$

It will be useful to directly define the several posteriors implicit in stochastic localization schemes below. For nonempty $S \subset \Theta$ and $t \in \mathbf{I}$, we define $\Pr^{\leftarrow}[\cdot | Y_t, S]$ to be the density of the posterior of X with the prior $X \sim p^S$. We similarly define $\Pr_{t \rightarrow \Theta}^{\leftarrow}[\cdot | Y_t]$ and $\Pr_{t \rightarrow \Theta}^{\leftarrow}[\cdot | Y_t, S]$ to be the density of posterior of θ conditioning on Y_t with $X \sim p$ or $X \sim p^S$, respectively. When $S = \{i\}$, we exclude the braces.

8.2.2 Proof of Theorem 8.0.1

Crucially, our proof relies in several places on the Markov property of stochastic localization samplers, together with the data processing inequality.

Proof of Theorem 8.0.1. By the triangle inequality, we can write

$$\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}}) \leq \underbrace{\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}, \hat{T}})}_{(\text{I})} + \underbrace{\text{TV}(p^{S_{\text{target}}, \hat{T}}, p^{S_{\text{target}}})}_{(\text{II})}.$$

Note $p^{S_{\text{init}}, \hat{T}}$ and $p^{S_{\text{target}}, \hat{T}}$ are the laws of the posterior $\Pr^{\leftarrow}[\cdot | \cdot]$ but applied to $Y_{\hat{T}}$ with distributions $p_{\hat{T}}^{S_{\text{init}}}$ and $p_{\hat{T}}^{S_{\text{target}}}$. Using the Markov property of localization-based samplers (Definition 7.2.1), we apply the data processing inequality twice and the definition of T_{end}^S to bound (I) via

$$\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}, \hat{T}}) \leq \text{TV}(p_{\hat{T}}^{S_{\text{init}}}, p_{\hat{T}}^{S_{\text{target}}}) \leq \text{TV}(p_{T_{\text{end}}^S}^{S_{\text{init}}}, p_{T_{\text{end}}^S}^{S_{\text{target}}}) \leq \epsilon.$$

To bound (II), we use the definition of TV and a coupling argument that takes advantage of the fact that we can initiate the reverse processes with the same distribution at \hat{T} . Observe that the observation processes associated with $p^{S_{\text{target}}, \hat{T}}$ and $p^{S_{\text{target}}}$ have the same distribution at index \hat{T} . Thus, taking $Y_{\hat{T}} \sim p_{\hat{T}}^{S_{\text{target}}}$, we can express by the law of iterated expectations,

$$\begin{aligned} p^{S_{\text{target}}, \hat{T}}(x) &= \mathbb{E} \Pr^{\leftarrow}[x | Y_{\hat{T}}] \\ p^{S_{\text{target}}}(x) &= \mathbb{E} \Pr^{\leftarrow}[x | Y_{\hat{T}}, S_{\text{target}}], \end{aligned}$$

as these observation processes have the same distribution at index \widehat{T} . Thus,

$$\begin{aligned}\text{TV}(p^{S_{\text{target}}, \widehat{T}}, p^{S_{\text{target}}}) &= \frac{1}{2} \int |p^{S_{\text{target}}, \widehat{T}}(x) - p^{S_{\text{target}}}(x)| d\mu \\ &= \frac{1}{2} \int \left| \mathbb{E} \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}] - \mathbb{E} \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}, S_{\text{target}}] \right| d\mu.\end{aligned}$$

By Jensen's inequality and Fubini's theorem, we bring the absolute value inside the expectation,

$$\begin{aligned}\text{TV}(p^{S_{\text{target}}, \widehat{T}}, p^{S_{\text{target}}}) &\leq \frac{1}{2} \int \mathbb{E} \left[\left| \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}] - \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}, S_{\text{target}}] \right| \right] d\mu \\ &= \frac{1}{2} \mathbb{E} \left[\int \left| \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}] - \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}, S_{\text{target}}] \right| d\mu \right].\end{aligned}$$

Intuitively, if $p_{\widehat{T}}^{S_{\text{target}}}$ is disjoint from the rest of $p_{\widehat{T}}$, then the conditional probability $\Pr^{\leftarrow}[x|Y_{\widehat{T}}]$ will be identical to $\Pr^{\leftarrow}[x|Y_{\widehat{T}}, S_{\text{target}}]$ because it is clear that $Y_{\widehat{T}}$ belongs to class $\theta \in S_{\text{target}}$. To formalize this intuition and simplify the above expression, we use the following two lemmas, whose proofs are deferred after the main theorem is completed.

Lemma 8.2.1. By applying the law of iterated expectations and Bayes' rule, we can show for $Y_{\widehat{T}} \in \text{supp}(p_{\widehat{T}}^{S_{\text{target}}})$,

$$\int \left| \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}] - \overset{\leftarrow}{\Pr}[x|Y_{\widehat{T}}, S_{\text{target}}] \right| d\mu \leq 2 \sum_{\theta \in \Theta - S_{\text{target}}} \overset{\leftarrow}{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\widehat{T}}].$$

Lemma 8.2.2. By Bayes' rule, we can derive for $Y_{\widehat{T}} \in \text{supp}(p_{\widehat{T}})$,

$$\sum_{\theta \in \Theta - S_{\text{target}}} \overset{\leftarrow}{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\widehat{T}}] \leq \max(1, W) \frac{p_{\widehat{T}}^{\Theta - S_{\text{target}}}(Y_{\widehat{T}})}{p_{\widehat{T}}^{\Theta - S_{\text{target}}}(Y_{\widehat{T}}) + p_{\widehat{T}}^{S_{\text{target}}}(Y_{\widehat{T}})}$$

Combining Lemmas 8.2.1 and 8.2.2 in Equation 8.2.2, we find

$$\text{TV}(p^{S_{\text{target}}, \widehat{T}}, p^{S_{\text{target}}}) \leq \max(1, W) \mathbb{E} \left[\frac{p_{\widehat{T}}^{\Theta - S_{\text{target}}}(Y_{\widehat{T}})}{p_{\widehat{T}}^{\Theta - S_{\text{target}}}(Y_{\widehat{T}}) + p_{\widehat{T}}^{S_{\text{target}}}(Y_{\widehat{T}})} \right].$$

Then, finally applying Lemma 2.1.1 as in Part I, we are finally able to bound the total variation in terms of ϵ ,

$$\text{TV}(p^{S_{\text{target}}, \widehat{T}}, p^{S_{\text{target}}}) \leq \frac{1}{2} \max(1, W) \sqrt{1 - \text{TV}^2(p_{\widehat{T}}^{\Theta - S_{\text{target}}}, p_{\widehat{T}}^{S_{\text{target}}})} \leq \frac{\sqrt{2}}{2} \max(1, W) \epsilon.$$

Combining our bounds on (I) and (II) achieves the desired result,

$$\begin{aligned}
\text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}}) &\leq \text{TV}(p^{S_{\text{init}}, \hat{T}}, p^{S_{\text{target}}, \hat{T}}) + \text{TV}(p^{S_{\text{target}}, \hat{T}}, p^{S_{\text{target}}}) \\
&\leq \epsilon + \frac{\sqrt{2}}{2} \max(1, W) \epsilon \\
&= \epsilon \cdot \left(1 + \frac{\max(1, W)}{\sqrt{2}}\right).
\end{aligned}$$

□

Helpful algebraic identities

We now restate and prove the simple algebraic identities that we use in the above proof. The first lemma basically says that we can upper bound the difference in conditional probabilities in terms of the probability that $\theta \in \Theta - S_{\text{target}}$. It is analogous to Lemmas 5.2.3 and 5.2.2, which establishes a similar bound but with the scores instead of the conditional posterior probabilities. The polynomial-dependence on dimension appears there because the difference of two scores can have size polynomial in the dimension. The key insight of this improved lemma is that the conditional posterior probabilities integrate to 1, which means we circumvent the difficulties of controlling the difference between two arbitrary scores.

Lemma 8.2.1. By applying the law of iterated expectations and Bayes' rule, we can show for $Y_{\hat{T}} \in \text{supp}(p_{\hat{T}}^{S_{\text{target}}})$,

$$\int \left| \overleftarrow{\Pr}[x|Y_{\hat{T}}] - \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] \right| d\mu \leq 2 \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}].$$

Proof. We rewrite $\Pr^{\leftarrow}[x|Y_{\hat{T}}]$, $\Pr^{\leftarrow}[x|Y_{\hat{T}}, S_{\text{target}}]$ using the law of iterated expectations and Bayes' rule,

$$\begin{aligned}
\overleftarrow{\Pr}[x|Y_{\hat{T}}] &= \sum_{\theta \in \Theta} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] \\
\overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] &= \sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}, S_{\text{target}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] \\
&= \frac{\sum_{\theta \in S_{\text{target}}} \Pr_{t \rightarrow \Theta}^{\leftarrow}[\theta|Y_{\hat{T}}] \Pr^{\leftarrow}[x|Y_{\hat{T}}, \theta]}{\sum_{\theta \in S_{\text{target}}} \Pr_{t \rightarrow \Theta}^{\leftarrow}[\theta|Y_{\hat{T}}]}.
\end{aligned}$$

The second equality in the second line follows from the fact that for all $\theta \in S_{\text{target}}$, the posteriors $\Pr_{t \rightarrow \Theta}^{\leftarrow}[\cdot|Y_t] \propto \Pr_{t \rightarrow \Theta}^{\leftarrow}[\cdot|Y_t, S_{\text{target}}]$ are proportional by the normalization constant of 1 and $\sum_{\theta \in S_{\text{target}}} \Pr_{t \rightarrow \Theta}^{\leftarrow}[\theta|Y_{\hat{T}}]$, respectively, which can be cancelled out in the denominator. Therefore the

difference can be written as

$$\begin{aligned}
& \int \left| \overleftarrow{\Pr}[x|Y_{\hat{T}}] - \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] \right| d\mu \\
&= \int \left| \sum_{\theta \in \Theta} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] - \frac{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta]}{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]} \right| d\mu \\
&= \int \left| \left(1 - \frac{1}{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]} \right) \sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] + \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] \right| d\mu \\
&= \int \left| - \left(\frac{\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]}{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]} \right) \sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] + \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] \right| d\mu.
\end{aligned}$$

If $\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] = 0$, then by the non-negativity of probability measures,

$$\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta] = 0,$$

as well. Thus the total expression is 0 and we are done. If it is non-zero, we can factor out $\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]$ term, which allows us to write everything in terms of posteriors with respect to $\Theta - S_{\text{target}}$ and S_{target} ,

$$\begin{aligned}
& \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \int \left| \frac{\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta]}{\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]} - \frac{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \overleftarrow{\Pr}[x|Y_{\hat{T}}, \theta]}{\sum_{\theta \in S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}]} \right| d\mu \\
&= \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}] \int \left| \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] - \overleftarrow{\Pr}[x|Y_{\hat{T}}, \Theta - S_{\text{target}}] \right| d\mu
\end{aligned}$$

For our diffusion model proof of Lemma 5.2.2, where we have scores instead of conditional probabilities, we had to use the properties of strongly log-concave measures to bound the analogue of $\int |\Pr^{\leftarrow}[x|Y_{\hat{T}}, S_{\text{target}}] - \Pr^{\leftarrow}[x|Y_{\hat{T}}, \Theta - S_{\text{target}}]| d\mu$. But since these are conditional probabilities that integrate to one, one needs only employ the trivial observation that by the triangle inequality and the fact that probability measures sum to 1,

$$\begin{aligned}
& \int \left| \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] - \overleftarrow{\Pr}[x|Y_{\hat{T}}, \Theta - S_{\text{target}}] \right| d\mu \\
&\leq \int \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] + \overleftarrow{\Pr}[x|Y_{\hat{T}}, \Theta - S_{\text{target}}] d\mu \leq 2.
\end{aligned}$$

Thus $\int |\Pr^{\leftarrow}[x|Y_{\hat{T}}, S_{\text{target}}] - \Pr^{\leftarrow}[x|Y_{\hat{T}}, \Theta - S_{\text{target}}]| d\mu$ is only a constant and we have

$$\int \left| \overleftarrow{\Pr}[x|Y_{\hat{T}}] - \overleftarrow{\Pr}[x|Y_{\hat{T}}, S_{\text{target}}] \right| d\mu \leq 2 \sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta}[\theta|Y_{\hat{T}}].$$

□

The next lemma is a simple application of Bayes rule and simply states that the conditional probability can be rewritten in terms of the marginals. This allows us to then use our Lemma 2.1.1 which is in terms of the marginals of two distributions.

Lemma 8.2.2. By Bayes' rule, we can derive for $Y_{\hat{T}} \in \text{supp}(p_{\hat{T}})$,

$$\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta} [\theta | Y_{\hat{T}}] \leq \max(1, W) \frac{p_{\hat{T}}^{\Theta - S_{\text{target}}}(Y_{\hat{T}})}{p_{\hat{T}}^{\Theta - S_{\text{target}}}(Y_{\hat{T}}) + p_{\hat{T}}^{S_{\text{target}}}(Y_{\hat{T}})}$$

Proof. We obtain through Bayes' rule,

$$\sum_{\theta \in \Theta - S_{\text{target}}} \overleftarrow{\Pr}_{t \rightarrow \Theta} [\theta | Y_{\hat{T}}] = \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}{\sum_{\theta \in \Theta} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}.$$

We divide by the same normalizing constant $\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta}$ to obtain

$$\begin{aligned} \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}{\sum_{\theta \in \Theta} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})} &= \frac{\frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta}}}{\frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta}} + \frac{\sum_{\theta \in S_{\text{target}}} w_{\theta} p_{\hat{T}}^{\theta}(Y_{\hat{T}})}{\sum_{\theta \in S_{\text{target}}} w_{\theta}} \cdot \left(\frac{\sum_{\theta \in S_{\text{target}}} w_{\theta}}{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta}} \right)} \\ &\leq \max \left(1, \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_{\theta}}{\sum_{\theta \in S_{\text{target}}} w_{\theta}} \right) \frac{p_{\hat{T}}^{\Theta - S_{\text{target}}}(Y_{\hat{T}})}{p_{\hat{T}}^{\Theta - S_{\text{target}}}(Y_{\hat{T}}) + p_{\hat{T}}^{S_{\text{target}}}(Y_{\hat{T}})}, \end{aligned}$$

where the last line uses the more general fact that for $A, B, x \geq 0$ we have

$$\frac{A}{A + B/x} = \frac{xA}{xA + B} \leq \max(1, x) \frac{A}{A + B}.$$

□

If a machine is expected to be infallible, it cannot also be intelligent.

Alan Turing

9

Instantiating Theory for Different Samplers

In Chapter 9, we leverage the improved framework we developed in the previous chapter to specific instances of critical windows for different generative models, data modalities, and scenarios, including diffusion models, autoregressive language models, jailbreaks, and phase transitions in statistical inference. These are the fruits of the abstractions we developed in the past two chapters; many of these examples could not be analyzed under existing statistical physics frameworks for critical windows, because they rely on diffusion samplers and specific models of data. Furthermore, our framework seems to reduce the problem of critical windows to computing statistical divergences and distances between different sub-populations, which is much more tractable.

In Section 9.1, we more precisely compute positions for critical windows for different generative models and data modalities. These examples recover the narrowness of the critical windows that we see in practice, as in many natural settings the critical windows are narrow in some sense. We see a direct quantitative improvement for computing critical windows for mixtures of Gaussians. Our critical window for this data distribution in Example 9.1.2 has a dimension-free widths, which was not obtained in Example 6.2.1. We can now also explicitly frame the jailbreak examples for autoregressive LLMs as critical windows within our general framework.

Example 9.0.1 (Jailbreaks and searching for pictures of Yellowstone [QPL⁺24, Ant24]). Consider an autoregressive language model, with \mathcal{A} denoting the vocabulary, $p \in \mathcal{A}^T$, a forward process indexed by $\mathbf{I} = \{0, 1, 2, \dots, T\}$, and Y_t to be the first $T - t$ tokens of X . Let $\Theta = \{\theta_{\text{harmful}}, \theta_{\text{safe}}\}$ (or $\{\theta_{\text{Googling Yellowstone}}, \theta_{\text{coding}}\}$). We assume that these two modes do not have different distributions over tokens $T - T' \in \mathbf{I}$. Between $T - T'$ and $T - T' - k$, the distributions become nearly disjoint, $\Pr_{x \sim p_{T-T'-k}^{\theta_{\text{harmful}}}}(x \in \text{supp}(p_{T-T'-k}^{\theta_{\text{safe}}})) \leq \epsilon$. In the jailbreaking example, $T' = 0$ and they are disjoint because the first tokens generated in the safe mode is always some form of refusal. In the Yellowstone example, they are disjoint the first time the LLM decides to Google Yellowstone pictures because that has very low probability under the coding mode of behavior. Then, on component θ_{harmful} or $\theta_{\text{Googling Yellowstone}}$ we have the critical window $T_{\text{before}}^C = T - T'$ and $T_{\text{after}}^C = T - T' - k$.

Our improved theoretical framework also facilitates a connection between critical windows and the all-or-nothing phase transition for distinguishing between the null and alternative hypotheses. All-or-nothing phase transitions state that there exist a threshold N^* such that it is impossible to distinguish between the null and alternative hypotheses with fewer than $(1 - \delta)N^*$ samples and it is possible to perfectly distinguish with more than $(1 + \delta)N^*$ samples for $\delta \in (0, 1/2)$. Defining a suitable sampler which autoregressively generates samples from an admixture of the null and alternative hypotheses, we can interpret this phase transition as a critical window as well.

Theorem 9.0.1 (Informal, see Theorem 9.1.4). Assume there is an all-or-nothing learning model with null and alternative hypotheses and a threshold N^* . Consider an autoregressive model which produces a string of samples from either the null or alternative hypothesis. There is a critical window at $[(1 - \delta)N^*, (1 + \delta)N^*]$ samples, in which the model localizes from sampling from both the null and the alternative hypotheses to just the alternative hypothesis.

In Section 9.2, we instantiate our bounds for hierarchically structured models of data and significantly generalize the results of Section 6.2.1, which was restricted to diffusions and Gaussian mixtures to all localization-based samplers and mixture models. Our theory for hierarchies applies to all localization-based samplers; therefore, we can also make claims about the difference in hierarchies between different localization-based samplers on the same base space. This allows us to conclude that two distributions generated from the same kind of localization-based sampler obtain the same hierarchy of features. More concretely, this could imply that if the learned sampler and data generation process are of the same structure, then the sampler may have the same hierarchy of features as the data generation process.

Corollary 9.0.1 (Informal, see Corollary 9.2.2). Assume we have two distributions p (the original distribution) and q (the learned sampler) generated by the same localization-based process which agree on all sub-populations $S \subset \Theta$. Then both p, q have the same hierarchy of features.

This chapter ends with empirical validation in Section 9.3, where we explicitly verify our theory on a prompt enforcing structured outputs and observe that critical windows occur during important mistakes in the reasoning patterns of LLMs, concurrently with [AAB⁺24, LLX⁺24].

9.1 Diffusion, autoregression, and all-or-nothing

In this section, we analytically compute $T_{\text{before}}^C, T_{\text{after}}^C$ for diverse stochastic localization samplers and models of data, including diffusion and autoregression processes. In these natural settings, the critical window is small in the sense of having a size which shrinks or does not depend on the dimension or context length. We shall also connect our framework to in-context learning and instantiate the the all-or-nothing phenomenon from statistical inference.

9.1.1 Diffusion

We first consider two examples of Gaussian Mixture Models and a diffusion model. We show that with two isotropic Gaussians, the critical window appears around a single point, $\ln \|\mu\|$, with width independent of the dimension. For Gaussians, we will apply the following well-known formulae for the Hellinger distance and KL-divergence between two Gaussians.

Example 9.1.1 (Two Isotropic Gaussians). Let $\Theta = \{\pm 1\}$, $p^{+1} = \mathcal{N}(\mu, \text{Id})$, $p^{-1} = \mathcal{N}(-\mu, \text{Id})$. Then, we have a critical window transitioning from sampling from both components to the component $+1$ between $T_{\text{before}}^C = \ln \|\mu\| + \ln 2 + \ln 1/\epsilon$ and $T_{\text{after}}^C = \ln \|\mu\| - \ln \ln \frac{1}{2\epsilon^2}$. When $\hat{T} \leq T_{\text{after}}^C$, then $\text{TV}(p^{+1, \hat{T}}, p^{+1}) \lesssim \epsilon$. When $\hat{T} \geq T_{\text{before}}^C$, $\text{TV}(p^{+1, \hat{T}}, p) \lesssim \epsilon$.

Proof. The proof for T_{before}^C is a simple application of Pinsker's inequality. Directly applying Theorem 8.0.1 to T_{after}^C , we need only show that

$$\frac{\sqrt{2}}{2} \left(1 - \frac{1}{2} \text{H}^2(p_{T_{\text{after}}^C}^{+1}, p_{T_{\text{after}}^C}^{-1}) \right) \leq \frac{\sqrt{2}}{2} \exp\left(-\frac{1}{2} \|\mu\|^2 e^{-2t}\right) \leq \epsilon.$$

□

For an isotropic Gaussian mixture model with randomly selected means, the critical window between sampling from one component to the entire mixture is also narrow. Note that we derive dimension-free widths in Example 9.1.2, an improvement over Example 6.2.1 who had a $\ln \ln d$ dependence on dimension for isotropic Gaussians.

Example 9.1.2 (Random mean spherical Gaussians). We first sample $\mu_i \sim \mathcal{N}(0, \text{Id})$ for $i \in \Theta$ i.i.d. and let $\Theta = \{\mathcal{N}(\mu_i, \text{Id})\}_{i \in \Theta}$. We let $S_{\text{before}} = \Theta$ and $S_{\text{after}} = \{\mu_1\}$. Then, we can compute $T_{\text{before}}^C = \max_{j \in \Theta} \ln \|\mu_i - \mu_j\| + \ln(1/\epsilon)$ and $T_{\text{after}}^C = \min_{j \in \Theta, i \neq j} \ln \|\mu_i - \mu_j\| - \frac{1}{2} \ln 8 \ln \frac{K}{\epsilon}$. Furthermore, with high probability over the selection of the means, $T_{\text{before}}^C - T_{\text{after}}^C = O_{K, \epsilon}(1)$ as $d \rightarrow \infty$.

Proof. The proof for T_{before}^C can be found in Section 5.2 of [LC24] and also follows from Pinsker's inequality. We again use the Hellinger distance to control T_{after}^C , but need to slightly modify the proof of Theorem 8.0.1 so that we can write it in terms of the Hellinger distance of individual components. We use the same notation. By convexity, we can bound

$$\begin{aligned} \mathbb{E}_{Y_{\hat{T}} \sim p_{\hat{T}}^{S_{\text{target}}}} \left[\frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_\theta p_{\hat{T}}^\theta(Y_{\hat{T}})}{\sum_{\theta \in \Theta} w_\theta p_{\hat{T}}^\theta(Y_{\hat{T}})} \right] &\leq \sum_{\theta \in \Theta - S_{\text{target}}} w_\theta \sum_{\phi \in S_{\text{target}}} w_\phi \mathbb{E}_{Y_{\hat{T}} \sim p_{\hat{T}}^\phi} \left[\frac{p_{\hat{T}}^\theta(Y_{\hat{T}})}{w_\theta p_{\hat{T}}^\theta(Y_{\hat{T}}) + w_\phi p_{\hat{T}}^\phi(Y_{\hat{T}})} \right] \\ &\leq K \max_{\theta \in \Theta - S_{\text{target}}, \phi \in S_{\text{target}}} \left(1 - \frac{1}{2} H^2(p_{\hat{T}}^\theta, p_{\hat{T}}^\phi) \right) \leq \epsilon, \end{aligned}$$

when $\hat{T} \leq T_{\text{after}}^C$. To conclude the second part of the theorem, observe that by concentration of measure (e.g., Theorem 3.1.1 from [Ver]) and a union bound, there exists a constant T independent of d such that $\|\mu_j\| \in [\sqrt{d} - T, \sqrt{d} + T]$ for all $j \in \Theta$ with high probability. Furthermore, by known Gaussian suprema inequalities, we can also assume that there exists a constant T' independent of d such that $|\langle \mu_i, \mu_j \rangle| \leq T' \|\mu_i\|$ (Lemma 5.1 from [vH16]). Thus, we can conclude that

$$\max_{j \in \Theta} \|\mu_i - \mu_j\|^2 \leq 2d + 4T\sqrt{d} + 2T^2 + 2T'(\sqrt{d} + T) = O(d).$$

$$\max_{j \in \Theta} \|\mu_i - \mu_j\|^2 \geq 2d - 4T\sqrt{d} + 2T^2 - 2T'(\sqrt{d} + T) = \Omega(d).$$

The difference in log scale is thus constant,

$$\frac{1}{2} \left(\ln \max_{j \in \Theta} \|\mu_i - \mu_j\|^2 - \ln \min_{j \in \Theta, i \neq j} \|\mu_i - \mu_j\|^2 \right) = O_{K,\epsilon}(1),$$

suppressing any dependences on the class size K and the error parameter ϵ . \square

This narrowness is not unique to continuous diffusion models, as we can compute the critical windows of a discrete diffusion model and demonstrate that as the context length T goes infinity, the length of the critical window goes to 0. Our exact model are two Dirac delta functions, and we can parameterize the critical window in terms of the Hamming distance between the supports of the delta measures.

The intuition is that the overlap between the noised versions of the two delta measures is great if with high probability all of the differing indices are masked. It is negligible if with high probability at least one of the differing indices are not masked. As the context length and Hamming distance go to infinity, both of these values approach $t = 1$ because the preimage of the set $[\epsilon, 1 - \epsilon]$ under $x \mapsto x^n$ has decreasing length as $n \rightarrow \infty$.

Example 9.1.3 (Two Dirac delta functions with a random masking procedure). Let $p \in \{\pm 1\}^T$, and consider a forward process with index set $\mathbf{I} = [0, 1]$, $Y_0 = X$, and $Y_t \in \{\pm 1, [\text{MASKED}]\}^T$. For $t \in \mathbf{I}$, we let all the value at index $i \in [T]$ be set to [MASKED] with probability t independently. For a mixture of two Dirac delta functions, we can express the critical window in terms of the *Hamming distance* between the corresponding strings. Let $\Theta = \{\theta_{\pm 1}\}, \ell_{\pm 1} \in \{\pm 1\}^T, p^{\theta_{\pm 1}} \sim \delta_{\ell_{\pm 1}}, w_{\pm 1} = \frac{1}{2}$. Then, on component 1 we have the critical window

$$T_{\text{before}}^C = \exp \left[\frac{\ln(1 - \epsilon)}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})} \right], T_{\text{after}}^C = \exp \left[\frac{\ln \epsilon^2}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})} \right]$$

For sufficiently large $d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})$, the window size $T_{\text{before}}^C - T_{\text{after}}^C = O\left(\frac{1}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})}\right)$. If $d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})$ increases with T , then the width of the critical window negligible.

Proof. To prove $\text{TV}(p^{1, \widehat{T}}, p^1) \lesssim \epsilon$ when $\widehat{T} \leq T_{\text{after}}^C$, observe that when $\widehat{T} \leq T_{\text{after}}^C$, the probability that all the differing elements between ℓ_1, ℓ_{-1} are masked is exactly $\widehat{T}^{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})} \leq \epsilon^2$. That means that there exists a set A with $p_{\widehat{T}}^{+1}(A) \geq 1 - \epsilon^2$ and $p_{\widehat{T}}^{-1}(A) = 0$, so by the definition of total variation, $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) \geq 1 - \epsilon^2$. Obviously, $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{+1}) = 0$ as well, so by Theorem 8.0.1, we obtain $\text{TV}(p^{1, \widehat{T}}, p^1) \lesssim \epsilon$. To prove that $\text{TV}(p^{1, \widehat{T}}, p^{\{\pm 1\}}) \leq \epsilon$ when $\widehat{T} \geq T_{\text{before}}^C$, we need only show that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) \leq \epsilon$. It suffices to show that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) \leq \epsilon$ by a simple triangle inequality argument. Consider the set $A \subset \{\pm 1, [\text{MASKED}]\}^T$ such that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) = p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A)$. Consider the set $B = \text{supp}(p_{\widehat{T}}^{+1}) \cap \text{supp}(p_{\widehat{T}}^{-1})$. For any $x \in B$, we know $p_{\widehat{T}}^{+1}(x) = p_{\widehat{T}}^{-1}(x)$ because the same number of tokens need to be masked from $\ell_{\pm 1}$. This means we have $p_{\widehat{T}}^{+1}(B) = p_{\widehat{T}}^{-1}(B) \geq \widehat{T}^{d_H(\ell_1, \ell_{-1})} \geq 1 - \epsilon$. Because $p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A) = p_{\widehat{T}}^{+1}(A - B) - p_{\widehat{T}}^{-1}(A - B)$, we have $p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A) \leq p_{\widehat{T}}^{+1}(\{\pm 1, [\text{MASKED}]\}^T - B) \leq \epsilon$. \square

It is clear for different kinds of diffusion models and distributions these critical windows are small. We can identify examples for other modeling paradigms as well, with rich connections to in-context learning and inference.

9.1.2 Autoregression

As was mentioned in the introduction, critical windows appear in many important contexts for LLMs, e.g., jailbreaks that occur over the first few tokens in the generation and the getting distracted by the Yellowstone National Park example [Ant24, QPL⁺24]. We first experimentally reproduce critical windows observed in existing literature, for a simplified jailbreak based on the prefill attack [Hai24b] and repeating token jailbreak [NCH⁺23] for LLAMA-3.1-8B-Instruct. In the left of Figure 9.1.1, we plot the probability of the model giving a toxic response in response to a harmful request, computed using the `StrongReject` Gemma 7b auditor from [SLB⁺24], as a function of the fraction of the phrase `Sure, here is how to` appended to the front of the model’s generation. We can see that there is a large jump in the attack success rate after only including a few tokens in the prefix. The right is a reproduction of Figure 12 from [NCH⁺23], which discussed

the jailbreak of making an LLM repeat some given word until it starts spewing training data. That shows that the probability of repeating the next token increases substantially when the token is repeated only a few times in the context. In both of these examples, the behavior of the model changes suddenly over the first ten tokens in the generations.

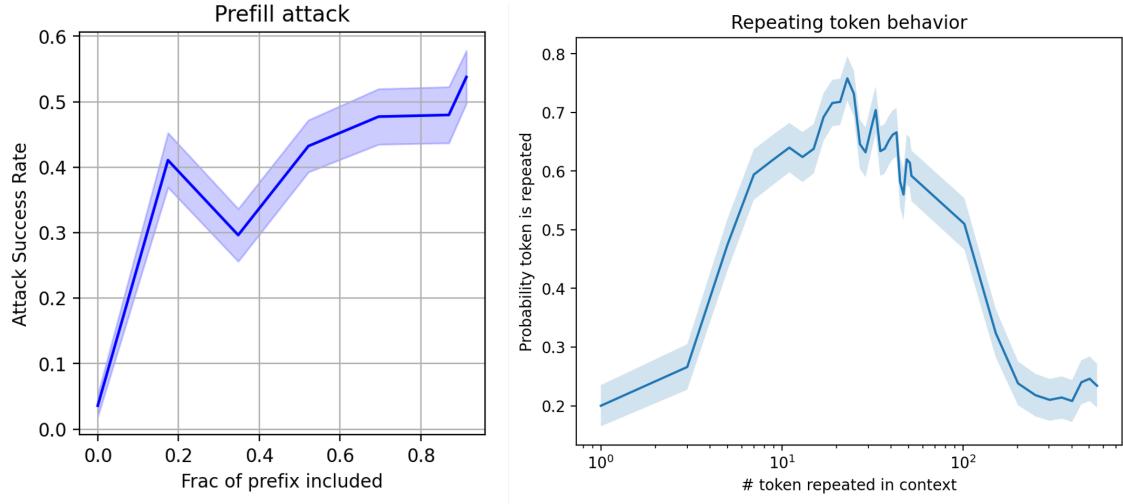


Figure 9.1.1: Examples of critical windows in jailbreaks for LLAMA-3.1-8B-Instruct. **Left:** Critical window for a prefill jailbreak [Hai24b]. **Right:** Critical window for a repeating token jailbreak [NCH⁺23].

Now we present a simple theoretical model for these critical windows for LLMs. Although this example is mathematically obvious, it provides a useful mental for understanding these types of jailbreaks. The basic intuition for our bounds is that when the model enters a state which is highly likely under one mode compared to another, there is a critical window transitioning into that state. In the case of Yellowstone national park example, it is when it first starts to searching for pictures of Yellowstone national park, and in the case of the jailbreaking example, it is when it states that it will accede (or does not immediately refuse) the user's harmful request.

Example 9.0.1 (Jailbreaks and searching for pictures of Yellowstone [QPL⁺24, Ant24]). Consider an autoregressive language model, with \mathcal{A} denoting the vocabulary, $p \in \mathcal{A}^T$, a forward process indexed by $\mathbf{I} = \{0, 1, 2, \dots, T\}$, and Y_t to be the first $T - t$ tokens of X . Let $\Theta = \{\theta_{\text{harmful}}, \theta_{\text{safe}}\}$ (or $\{\theta_{\text{Googling Yellowstone}}, \theta_{\text{coding}}\}$). We assume that these two modes do not have different distributions over tokens $T - T' \in \mathbf{I}$. Between $T - T'$ and $T - T' - k$, the distributions become nearly disjoint, $\Pr_{x \sim p_{T-T'-k}^{\theta_{\text{harmful}}}}(x \in \text{supp}(p_{T-T'-k}^{\theta_{\text{safe}}})) \leq \epsilon$. In the jailbreaking example, $T' = 0$ and they are disjoint because the first tokens generated in the safe mode is always some form of refusal. In the Yellowstone example, they are disjoint the first time the LLM decides to Google Yellowstone pictures because that has very low probability under the coding mode of behavior. Then, on component θ_{harmful} or $\theta_{\text{Googling Yellowstone}}$ we have the critical window $T_{\text{before}}^C = T - T'$ and $T_{\text{after}}^C = T - T' - k$.

Notice that we can actually mitigate the effect of these critical windows by increasing the value of $P_{x \sim p_{T-T'-k}^{\theta_{\text{harmful}}}}(x \in \text{supp}(p_{T-T'-k}^{\theta_{\text{safe}}}))$. This could explain the effectiveness of finetuning on corrections in [QPL+24], which intuitively would increase the value of this ratio. Furthermore, the quantity that measures probability of mode-switching, $p^{\theta_{\text{harmful}}}/p$, suggests using a likelihood ratio to distinguish between harmful and benign prompts. In Section 9.1.2, we test a class of likelihood ratio methods inspired our theory that obtain recall 5-10 \times the false positive rate for 5 different types of jailbreaks (Table 9.1.1).

The next example we consider is a critical window for a stylized model of solving a math problem as a random walk. The classes are “strong” and “weak” problem solvers, which have some probability of getting the correct answer. We do not claim that this is a realistic model for reasoning tasks, but this example more generally characterizes the narrowness of stochastic localization-based samplers who engage in some sort of random walk within the context. To establish T_{before}^C , we apply the data processing inequality to bound the distributional difference in the position in the random walk in terms of the number of right and incorrect steps, and use well-known inequalities about the total variation between binomials with different probabilities. To establish T_{after}^C , we look at the distribution of the positions of the random walk, and then use a standard Hoeffding inequality to show that with sufficiently many steps, they are disjoint.

Example 9.1.4 (Math problem-solving as a random walk). We model solving a math problem as taking a random walk on \mathbb{Z} with stepsize 1 of length T . If the random walk hits $+A$, then it has ‘solved’ the problem; if the random walk hits $-A$, then it has obtained an incorrect solution. Assume that we have two modes: a strong problem solving mode (denoted $+1$), which takes a $+1$ step with probability $0.5 + \Delta$, and a weak problem solving mode (denoted -1), which takes a $+1$ step with probability $0.5 - \Delta$. Assuming that $\frac{\ln(2/\epsilon^2)}{2\Delta^2} < A$ and $\epsilon^2 < 10^{-3}(0.5 - \Delta)(0.5 + \Delta)$, there is a critical window for the strong problem solving window of

$$T_{\text{before}}^C = T - \frac{\epsilon^2}{\Delta^2} + 2 \quad \text{and} \quad T_{\text{after}}^C = T - \frac{\ln(2/\epsilon^2)}{2\Delta^2}.$$

Note the critical window has width $\Theta(1/\Delta^2)$ independent of T .

Proof. Because only the direction of steps matter, we can model the critical window for this random walk as observing a sequence of ± 1 with an autoregressive model. Let $p \in \{\pm 1\}^T$, and consider a forward process with index set $\mathbf{I} = \{0, 1, 2, \dots, T\}$, $Y_0 = X$, and $Y_t \in \{\pm 1, [\text{MASKED}]\}^T$. For $t \in \mathbf{I}$, we let the last t tokens of Y_t be deterministically set to $[\text{MASKED}]$. We generate data as a mixture of biased coins with separation $2\Delta < 0.01$. For a mixture of two biased coins, with probabilities of $0.5 \pm \Delta$ ($\theta_{\pm 1}$ respectively) of yielding 1, we can compute the critical window and show that it tightly clusters around $\Theta(1/\Delta^2)$. Let $\Theta = \{\theta_{\pm 1}\}$, $p^{\theta_{\pm 1}} \sim (\text{Bern}(\theta_{\pm 1}))^{\otimes T}$, $w_{\pm 1} = \frac{1}{2}$. We also assume $\epsilon^2 < 10^{-3}(0.5 - \Delta)(0.5 + \Delta)$. Then, on component 1 we have the critical window $T_{\text{before}}^C = T - \frac{\epsilon^2}{\Delta^2} + 2$ and $T_{\text{after}}^C = T - \frac{\ln(2/\epsilon^2)}{2\Delta^2}$. When $\hat{T} \leq T_{\text{after}}^C$, then $\text{TV}(p^{1,\hat{T}}, p^1) \lesssim \epsilon$. When $\hat{T} \geq T_{\text{before}}^C$, $\text{TV}(p^{1,\hat{T}}, p) \lesssim \epsilon$.

Note that the number of $+1$ is sufficient for disambiguating $\theta_{\pm 1}$. To prove the bounds T_{before}^C , we show that with only $\frac{\epsilon^2}{\Delta^2} - 2$ samples the total variation between $0.5 - \Delta$ and $0.5 + \Delta$ is negligible. Using [Roo01], we find

$$\text{TV}(p_{T_{\text{after}}^C}^{+1}, p_{T_{\text{after}}^C}^{-1}) \leq \frac{2\Delta \sqrt{\frac{T-T_{\text{after}}^C+2}{2(0.5-\Delta)(0.5+\Delta)}}}{\left(1 - 2\Delta \sqrt{\frac{T-T_{\text{after}}^C+2}{2(0.5-\Delta)(0.5+\Delta)}}\right)^2} \lesssim 3\epsilon.$$

For T_{after}^C , we compute how many samples it takes for $p^{\pm 1}$ to have only ϵ overlap in total variation using Hoeffding's inequality. If we have n samples, the mean \bar{X} of the n samples of ± 1 for $p = 0.5 + \Delta$ satisfies the concentration inequality $\Pr(|\bar{X} - 2\Delta| > \Delta) \leq 2 \exp(-2nt^2)$ (furthermore we can ignore the stopping condition by our requirement that $\frac{\ln(2/\epsilon^2)}{2\Delta^2} < A$). We find $\Pr(|\bar{X} - 2\Delta| > \Delta) \leq \epsilon^2$ for $T - T_{\text{after}}^C$ samples, proving that the total variation is at least $1 - \epsilon^2$. \square

In all of these examples, the feature or class has a fixed effect on the outputs of the generation procedure. The last model we consider in this subsection is a latent variable model, similar to the one presented in [ALL⁺19], where the latent variable's impacts on the outputs of the model may vary through the context. It captures the intuition that there is an underlying stochastic process of ideas (called the discourse vectors) of which the outgoing words are simple particular instantiations or emissions from it. In contrast to their setup, we have to parameterize the discourse vector process, and use a simple mixture model of two Gaussians with spikes in their covariance matrices. This should also be viewed as a stylistic choice to ensure that the word vectors still have mean zero, like in the original setup [ALL⁺19].

Each word is a vector $w \in \mathbb{R}^d$ and the context length is $T \in \mathbb{Z}^{>0}$. There are two classes $\Theta = \{u, v\}$ given by unit vectors $u, v \in \mathcal{S}^{d-1}$. We define the distribution over words p^θ for $\theta \in \Theta$ as follows. We generate the path of a discourse vector $(C_t^\theta)_{t \in [0, \infty]} \in \mathbb{R}^d$ with the reverse SDE Orstein-Uhlenbeck process such that $C_\infty^\theta \sim \mathcal{N}(0, \text{Id})$ and $C_0^\theta \sim \mathcal{N}(0, \text{Id} + \alpha\theta\theta^\top)$ for some $\alpha > 0$. Let q_t^θ be the law of C_t^θ for $t \geq 0$, $\mathbf{I} = \{0, 1, \dots, T\}$, and for $t \in \mathbf{I}$, draw samples $w_t^\theta \in \mathbb{R}^d$ where we impose the prior $w_t^\theta \sim \mathcal{N}(0, \text{Id})$ unconditionally on the discourse vector and have $w_t^\theta | C_t^\theta \propto \exp(\langle C_t^\theta, \cdot \rangle)$. Then we return the corpus $\{w_t^\theta\}_{t \in \mathbf{I}}$ as an output, where its law is given by p^θ .

The proof of Theorem 9.1.1 is as follows. To establish T_{after}^C , we need to find when these two distributions are disjoint. To do that, we look at the marginal distributions of the t th token from both classes. Because the discourse processes have different covariances, the marginals of the t th token from either classes also have different covariances, and they also are Gaussian. We then exploit this discrepancy with Hellinger distance bounds for Gaussians to show that the overlap is negligible for small t . To establish T_{before}^C , the data processing inequality gives us the distributional difference in emitted tokens is less than the distributional difference in the path measures of the discourse vectors. Then we use a standard Girsanov-based argument to control the path measures in terms of the difference in the drifts.

Theorem 9.1.1 (Autoregressive with a mixture of two Gaussians as the concept distribution).

We assume that $\exp(-T)\sqrt{\alpha - \log(1 + \alpha)} \leq \epsilon$. Let $\phi(x) = \frac{2+x}{(2+x(1+\langle u, v \rangle)/2)(2+x(1-\langle u, v \rangle)/2)}$. Then, on component u we have

$$T_{\text{after}}^C = \frac{1}{2} \ln \left[\frac{\alpha}{\phi^{-1}(\epsilon^2)} \right], T_{\text{before}}^C = \frac{1}{2} \ln \left[\frac{\alpha\sqrt{1+\alpha^2}\sqrt{1-\langle u, v \rangle^2}}{\epsilon} \right].$$

When $\hat{T} \leq T_{\text{after}}^C$, then $\text{TV}(p^{\{u\}, \hat{T}}, p^{\{u\}}) \lesssim \epsilon$. When $\hat{T} \geq T_{\text{before}}^C$, $\text{TV}(p^{\{u\}, \hat{T}}, p) \lesssim \epsilon$. This critical window also does not depend on the T and only on the differences between the topic mixtures captured by $\alpha, \langle u, v \rangle$.

This proof will require Theorem 4.2.3 (Girsanov-based approximation for SDEs), and Lemmas 6.2.6 (Hellinger computation for Gaussians) and 9.1.2, the last of which is stated below. It is a simple computation of the operator norm and eigenvectors of $uu^\top \pm vv^\top$.

Lemma 9.1.2. Let $u, v \in \mathcal{S}^{d-1}$. Then

$$\begin{aligned} \|uu^\top - vv^\top\|_{\text{op}} &\leq \sqrt{1 - \langle u, v \rangle^2} \\ \lambda(uu^\top + vv^\top) &= \{1 \pm \langle u, v \rangle\}. \end{aligned}$$

Proof. There exists $r \in \mathcal{S}^{d-1}$ such that $v = \langle u, v \rangle u + \sqrt{1 - \langle u, v \rangle^2}r$ and $u \perp r$. We find that

$$uu^\top - vv^\top = (1 - \langle u, v \rangle^2)uu^\top - (1 - \langle u, v \rangle^2)rr^\top - \langle u, v \rangle \sqrt{1 - \langle u, v \rangle^2}[ur^\top + ru^\top].$$

We can explicitly compute the eigenvalues of $uu^\top - vv^\top$ using the discriminant and find that they are equal to $\pm\sqrt{1 - \langle u, v \rangle^2}$. By a similar derivation, we can write

$$uu^\top + vv^\top = (1 + \langle u, v \rangle^2)uu^\top + (1 - \langle u, v \rangle^2)rr^\top + \langle u, v \rangle \sqrt{1 - \langle u, v \rangle^2}[ur^\top + ru^\top].$$

which gives us eigenvalues for $uu^\top + vv^\top$ of $1 \pm \langle u, v \rangle$. □

Proof of Theorem 9.1.1. To compute the T_{after}^C bound, which means we want to show the distributions are far apart at that time, it suffices by the data processing inequality to lower bound the Hellinger distance of the distribution of words generated at index \hat{T} , which we label $f_{\hat{T}}^u$. By the data processing inequality $1 - \frac{1}{2}\text{H}^2(p_{\hat{T}}^u, p_{\hat{T}}^v) \leq 1 - \frac{1}{2}\text{H}^2(f_{\hat{T}}^u, f_{\hat{T}}^v)$, so it suffices to show $1 - \frac{1}{2}\text{H}^2(f_{\hat{T}}^u, f_{\hat{T}}^v) \lesssim \epsilon$. Because the Gaussian is its own conjugate prior and $w_{\hat{T}}^u | C_t^\theta \propto \exp(-\frac{1}{2}\|w_{\hat{T}}^u - C_t^\theta\|^2)$, we can compute $w_{\hat{T}}^u | C_t^\theta \sim \mathcal{N}(C_t^\theta, \text{Id})$ and $w_{\hat{T}}^u \sim \mathcal{N}(0, 2\text{Id} + \alpha e^{-2t}uu^\top)$. Applying Lemmas 6.2.6 and 9.1.2, we can explicitly compute

$$1 - \frac{1}{2}\text{H}^2(f_{\hat{T}}^u, f_{\hat{T}}^v) \lesssim \sqrt{\frac{2 + \alpha e^{-2\hat{T}}}{(2 + \alpha e^{-2\hat{T}}(1 + \langle u, v \rangle)/2)(2 + \alpha e^{-2\hat{T}}(1 - \langle u, v \rangle)/2)}} \lesssim \sqrt{\phi(\alpha e^{-2\hat{T}})} \leq \epsilon.$$

To compute T_{before}^C , we first use the data processing inequality to reduce the difference in the emitted tokens to the difference in the paths of the context vectors, and then apply the approximation error bounds from Theorem 4.2.3 to bound the differences in path measures. When $\widehat{T} \geq T_{\text{before}}^C$, we can use the triangle inequality to write $\text{TV}(p^{\{u\}, \widehat{T}}, p) = \text{TV}(p_{\widehat{T}}^u, p_{\widehat{T}}^{\{u,v\}}) \leq \text{TV}(p_{\widehat{T}}^u, p_{\widehat{T}}^v)$. Note that $p_{\widehat{T}}^\theta$ is the distribution of the first $T - \widehat{T}$ tokens generated by the model under θ . Note that $p_{\widehat{T}}^u$ is a function of $(C_t^u)_{t \in \mathbf{I} \cap [\widehat{T}, T]}$ and $p_{\widehat{T}}^v$ is a function of $(C_t^v)_{t \in \mathbf{I} \cap [\widehat{T}, T]}$. By the data processing inequality, we can bound the difference in terms of the distributions over the tokens in terms of the law of the process of the discourse vectors,

$$\text{TV}(p_{\widehat{T}}^u, p_{\widehat{T}}^v) \leq \text{TV}((C_t^u)_{t \in \mathbf{I} \cap [\widehat{T}]}, (C_t^v)_{t \in \mathbf{I} \cap [\widehat{T}, T]}) \leq \text{TV}((C_t^u)_{t \in [\widehat{T}, T]}, (C_t^v)_{t \in [\widehat{T}, T]}).$$

Note that for $\theta \in \Theta$, $(C_t^\theta)_{t \in [\widehat{T}, T]}$ is generated by the following reverse time SDE,

$$dC_t^\theta = \{C_t^\theta + 2\nabla \ln q_t^\theta(C_t^\theta)\} dt + \sqrt{2} dB_t, \quad t \in [\widehat{T}, T], C_T^\theta \sim q_T^\theta.$$

Now we define $(C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]}$ to be the reverse SDE defined by initializing at q_T^u but with the score of q_T^v ,

$$dC_t^{u \rightarrow v} = \{C_t^{u \rightarrow v} + 2\nabla \ln q_t^v(C_t^{u \rightarrow v})\} dt + \sqrt{2} dB_t, \quad t \in [\widehat{T}, T], C_T^{u \rightarrow v} \sim q_T^u.$$

By the triangle inequality, we have

$$\text{TV}((C_t^u)_{t \in [\widehat{T}, T]}, (C_t^v)_{t \in [\widehat{T}, T]}) \leq \underbrace{\text{TV}((C_t^u)_{t \in [\widehat{T}, T]}, (C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]})}_{(\text{I})} + \underbrace{\text{TV}((C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]}, (C_t^v)_{t \in [\widehat{T}, T]})}_{(\text{II})}.$$

This again are two TV values, where the first corresponds to a score-difference bound and the second corresponds to a difference in initializations. To bound (I), observe that the SDEs have different scores but have the same initializations. We apply Theorem 4.2.3 to $\text{TV}((C_t^u)_{t \in [\widehat{T}, T]}, (C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]})$ and obtain

$$\begin{aligned} \text{TV}((C_t^u)_{t \in [\widehat{T}, T]}, (C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]}) &\leq \sqrt{\text{KL}((C_t^u)_{t \in [\widehat{T}, T]} \parallel (C_t^{u \rightarrow v})_{t \in [\widehat{T}, T]})} \\ &\leq \sqrt{\int_{\widehat{T}}^T \mathbb{E}_{X \sim C_t^u} \|\nabla \ln p_t^u(X) - \nabla \ln p_t^v(X)\|^2 dt} \end{aligned}$$

We simplify the inner expectation by using the

$$\begin{aligned}
\|\nabla \ln p_t^u(X) - \nabla \ln p_t^v(X)\| &= \left\| \left[(\text{Id} + \alpha e^{-2t} uu^\top)^{-1} - (\text{Id} + \alpha e^{-2t} vv^\top)^{-1} \right] x \right\| \\
&= \left\| \left[\left(\text{Id} - \frac{\alpha e^{-2t}}{1 + \alpha e^{-2t}} uu^\top \right) - \left(\text{Id} - \frac{\alpha e^{-2t}}{1 + \alpha e^{-2t}} vv^\top \right) \right] x \right\| \\
&\lesssim \alpha e^{-2t} \|uu^\top - vv^\top\|_{\text{op}} \left\| \prod_{\text{span}(u,v)} x \right\|, \\
&= \alpha e^{-2t} \sqrt{1 - \langle u, v \rangle^2} \left\| \prod_{\text{span}(u,v)} x \right\|. \quad (\text{Lemma 9.1.2})
\end{aligned}$$

We can upper bound $\mathbb{E}_{X \sim C_t^u} \|\prod_{\text{span}(u,v)} X\|^2$ by considering right-triangular L such that $L^\top L = \text{Id} + \alpha uu^\top$. and $\prod_{\text{span}(u,v)} X = \prod_{\text{span}(u,v)} LY$, where $Y \sim \mathcal{N}(0, \text{Id})$. The operator norm of $\prod_{\text{span}(u,v)} L$ is

$$\left\| \prod_{\text{span}(u,v)} \circ L \right\|_{\text{op}} \leq \|L\|_{\text{op}} \leq \sqrt{1 + \alpha^2 e^{-2t}} \leq \sqrt{1 + \alpha^2}.$$

$\prod_{\text{span}(u,v)} \circ L$ is also rank 2 and $(\prod_{\text{span}(u,v)} \circ L) Y = (\prod_{\text{span}(u,v)} \circ L) \circ \prod_{L^{-1}\text{span}(u,v)} Y$, where $\prod_{L^{-1}\text{span}(u,v)} Y \sim \mathcal{N}(0, \text{Id}_2)$. Thus we have

$$\mathbb{E}_{Y \sim \mathcal{N}(0, \text{Id})} \left[\left\| \left(\prod_{\text{span}(u,v)} \circ L \right) Y \right\|^2 \right] = \mathbb{E}_{Y' \sim \mathcal{N}(0, \text{Id}_2)} \left[\left\| \left(\prod_{\text{span}(u,v)} \circ L \right) Y' \right\|^2 \right] \lesssim 1 + \alpha^2$$

Combining this information together, we are able to compute,

$$\begin{aligned}
\text{TV}((C_t^u)_{t \in [\hat{T}, T]}, (C_t^{u \rightarrow v})_{t \in [\hat{T}, T]}) &\leq \sqrt{\int_{\hat{T}}^T \mathbb{E}_{X \sim C_t^u} \|\nabla \ln p_t^u(X) - \nabla \ln p_t^v(X)\|^2 dt} \\
&\lesssim \alpha \sqrt{1 + \alpha^2} \sqrt{1 - \langle u, v \rangle^2} e^{-2\hat{T}} \\
&\lesssim \epsilon.
\end{aligned}$$

To bound (II), we observe that both are run with the same score so we need only bound the difference at initialization. By the data processing inequality, we again have

$$\text{TV}((C_t^{u \rightarrow v})_{t \in [\hat{T}, T]}, (C_t^v)_{t \in [\hat{T}, T]}) \leq \text{TV}(p_T^u, p_T^v).$$

We can again apply the triangle inequality to get $\text{TV}(p_T^u, p_T^v) \leq \text{TV}(p_T^u, \mathcal{N}(0, \text{Id})) + \text{TV}(\mathcal{N}(0, \text{Id}), p_T^v) \leq \dots$. For any $\theta \in \Theta$, we have by the forward convergence of the OU process $\text{TV}(p_T^u, \mathcal{N}(0, \text{Id})) \leq \dots$

$\exp(-T)\sqrt{\text{KL}(p_u||\mathcal{N}(0, \text{Id}))}$. We can explicitly compute $\text{KL}(p_u||\mathcal{N}(0, \text{Id}))$ as

$$\text{KL}(p_u||\mathcal{N}(0, \text{Id})) = \frac{1}{2} \left[d + \alpha - d - \log(\text{Id} + \alpha uu^\top) \right] = \frac{1}{2} [\alpha - \log(1 + \alpha)].$$

Thus, we obtain the following bound on $\text{TV}((C_t^{u \rightarrow v})_{t \in [\hat{T}, T]}, (C_t^v)_{t \in [\hat{T}, T]})$ of

$$\text{TV}((C_t^{u \rightarrow v})_{t \in [\hat{T}, T]}, (C_t^v)_{t \in [\hat{T}, T]}) \lesssim \exp(-T)\sqrt{\alpha - \log(1 + \alpha)} \lesssim \epsilon. \quad \square$$

After showing that the windows in different settings are narrow for autoregressive models, we will specialize our framework to particular instances of critical windows in autoregressive LLMs, like in-context learning and the all-or-nothing phenomenon.

In-context learning

Here we show that autoregressive critical windows can also be applied to describe in-context learning. In particular, we can capture the idea that with sufficiently many in-context examples, we learn the $\theta^* \in \Theta$ that generated the transitions for in-context examples, with a sample complexity in terms of T_{after}^C .

Example 9.1.5 (Informal, see Example A.4.1). Consider an in-context learning setup, where the context

$$[x_1, y_1, o, \dots, x_{T+1}, y_{T+1}, o]$$

consists of question-answer pairs (x_i, y_i) , delimiters o , and $x_i \rightarrow y_i$ sampled from p^{θ^*} for some $\theta^* \in \Theta$. In the forward-reverse experiment, we truncate it to $[x_1, y_1, o, \dots, x_{T+1}]$, and then resample with p to produce $[x_1, \dots, x_{T+1}, \tilde{y}_{T+1}, o]$. The total variation between the sequences $[x_1, y_1, o, \dots, x_{T+1}]$ and $[x_1, \dots, x_{T+1}, \tilde{y}_{T+1}, o]$ can be viewed as the average-case error of the in-context learner and can be bounded within our critical windows framework. We have $T_{\text{after}}^C = 3T + 3 - O_\epsilon(1)$, with $O_\epsilon(1)$ independent of T ($S_{\text{after}} \triangleq \{\theta^*\}$). Note that T_{after}^C is the order of how many samples that can be erased so that we still are able to learn $\theta^* \in \Theta$.

One might ask if there is a T_{before}^C for in-context learning, a threshold such that it is impossible to distinguish between $S_{\text{after}}, S_{\text{before}}$ with that many samples. In the next section, we will provide an example of a T_{before}^C for in-context learning with the all-or-nothing phase transition.

All-or-nothing phenomenon

Here we elucidate a formal connection between the critical windows phenomenon in in-context learning and the *all-or-nothing* phenomenon. It intuitively captures the idea that there is a critical threshold of number of samples where learning becomes possible. To begin, we first define the notions of strong and weak detection, which will capture the paradigms where learning is possible or impossible:

Definition 9.1.1. Let (N_s) be an increasing sequence of integers. Given sequences of distributions $(p_s), (q_s)$ over $z \in \mathbb{R}^{N_s}$, a sequence of test statistics $(\mathcal{A}_s : \mathbb{R}^{N_s} \rightarrow \mathbb{R})$ with threshold (τ_s) achieves:

- *strong detection* if $\limsup_{s \rightarrow \infty} \{\Pr_{z \sim p_s} [\mathcal{A}_s(z) < \tau_s] + \Pr_{z \sim q_s} [\mathcal{A}_s(z) \geq \tau_s]\} = 0$.
- *weak detection* if $\limsup_{s \rightarrow \infty} \{\Pr_{z \sim p_s} [\mathcal{A}_s(z) < \tau_s] + \Pr_{z \sim q_s} [\mathcal{A}_s(z) \geq \tau_s]\} < 1$.

Note that we can view the term $\Pr_{z \sim p_s} [\mathcal{A}_s(z) < \tau_s] + \Pr_{z \sim q_s} [\mathcal{A}_s(z) \geq \tau_s]$ as the sum of the false positive and false negative error rates of the test statistics. By the operational characterization of TV distance, strong detection is (information-theoretically) possible if and only if $\liminf_{s \rightarrow \infty} \text{TV}(p_s, q_s) = 1$, and weak detection is (information-theoretically) possible if and only if $\liminf_{s \rightarrow \infty} \text{TV}(p_s, q_s) > 0$. For a more thorough exposition on the relationship between statistical inference and information theory see Section A.1.1. Now we consider the following Bayesian inference problem, given by a joint distribution π over $(\theta, z) \in \mathbb{R}^n \times \mathbb{R}^m$. One first samples an unknown signal $\theta \in \mathbb{R}^n$ from π_θ ; and then, for a given sample size N , the test statistics receives the observations $\{z_i\}_{i=1}^N$ drawn i.i.d. from $\pi_{z|\theta}$; the goal is to infer θ from these observations. Let $\pi^{(N)}$ denote the distribution over $\{z_i\}_{i=1}^N$, the mixture of product measures parametrized by θ .

Definition 9.1.2. Let (π_s) be a sequence of inference tasks over $\mathbb{R}^{n_s} \times \mathbb{R}^{m_s}$ and (π^{null_s}) be a sequence of distributions over \mathbb{R}^{m_s} . (π_s) exhibits an *all-or-nothing phase transition at threshold* (N_s) *with respect to null models* (π^{null_s}) if:

- For any $\beta < 1$: weak detection between $(\pi^{(\beta N_s)})$ and $((\pi^{\text{null}_s})^{\otimes \beta N_s})$ is information-theoretically impossible
- For any $\beta > 1$: strong detection between the planted model $(\pi^{(\beta N_s)})$ and the null model $((\pi^{\text{null}_s})^{\otimes \beta N_s})$ is information-theoretically possible

All-or-nothing phase transitions have been established for a number of natural inference tasks like sparse linear regression [RXZ19, GZ19], sparse PCA [NWZ20], generalized linear models [BMR20], group testing [TAS21, COGHK⁺22], linear and phase retrieval models [SC16, TS20], planted subgraphs [MNWS⁺23], and planted Gaussian perceptrons [NWZ23]. Here is an example for sparse linear regression:

Theorem 9.1.3 (Sparse linear regression all-or-nothing phase transition [RXZ19]). Let π_s be the distribution over $\mathbb{R}^{n_s} \times \mathbb{R}^{m_s}$ for $n_s = s$ and $m_s = s + 1$ where the marginal over θ is given by the uniform distribution over k_s -sparse vectors in $\{0, 1\}^s$, and the conditional distribution $\pi_{z|\theta}$ is given by sampling $x \sim \mathcal{N}(0, \text{Id}_s)$, taking $y = \langle \theta, x \rangle + \xi$ for $\xi \sim \mathcal{N}(0, \sigma_s^2)$, and outputting observation $z = (x, y)$. The null model π_s^{null} is given by sampling $x \sim \mathcal{N}(0, \text{Id}_s)$ and outputting $y = \mathcal{N}(0, k_s + \sigma_s^2)$.

If $\sigma_s^2 \ll k_s \leq s^{0.499}$, then (π_s) exhibits an all-or-nothing phase transition at threshold (N_s^*) with respect to null models (π_s^{null}) for

$$N_s^* \triangleq \frac{2k_s \log(s/k_s)}{\log(1 + k_s/\sigma_s^2)}.$$

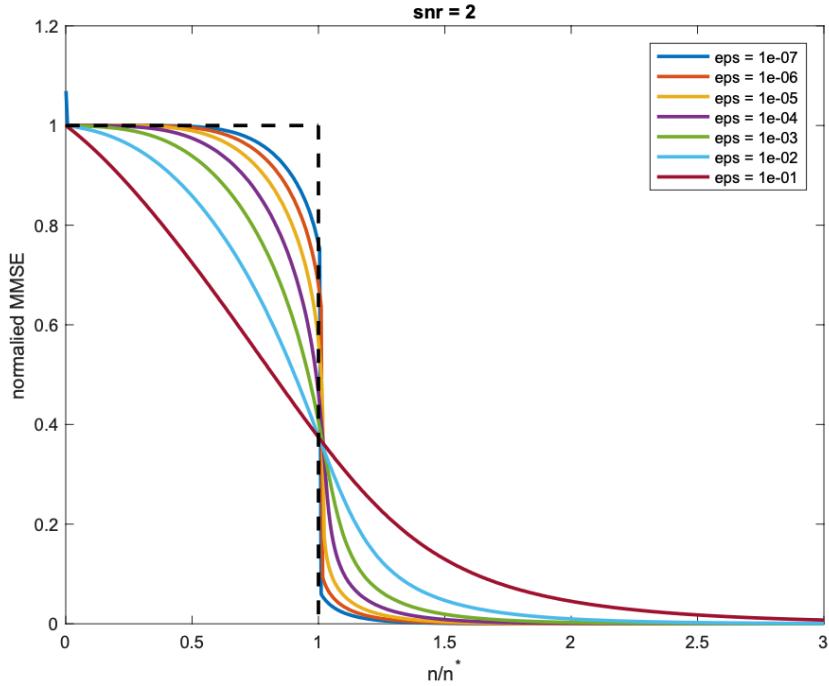


Figure 9.1.2: Example of all-or-nothing phase transition from [RXZ19] for sparse linear regression, demarcated by a dramatic drop in minimum mean square error at $n = N^*$ samples.

Having defined the all-or-nothing phenomenon, we rigorously instantiate it as a critical window for in-context learning. We first define a mixture model $p_{(N)}^\Theta$ for sequence lengths N onto which we will identify a critical window.

Definition 9.1.3. Let $\Theta = \Theta_{\text{signal}} \sqcup \{\text{NULL}\}$ where $\Theta_{\text{signal}} \triangleq \text{supp}(\pi_\theta)$. Given $\theta \in \text{supp}(\pi_\theta)$, let $p_{(N)}^\theta$ denote the distribution over sequences $(z_1, \dots, z_N, |, \theta)$ where z_1, \dots, z_N are i.i.d. samples from $p_{z|\theta}$. Let $p_{(N)}^{\text{null}}$ denote the distribution over observations $(z_1, \dots, z_N, |, \text{NULL})$ where z_1, \dots, z_N are i.i.d. samples from π^{null} . We then take $p_{(N)}^\Theta \triangleq \mathbb{E}_{\theta \sim \frac{1}{2}\pi_\theta + \frac{1}{2}\delta_{\text{NULL}}} p_{(N)}^\theta$.

Under this model of data, we have the following theorem expressing the all-or-nothing phase transition in terms of $T_{\text{before}}^C, T_{\text{after}}^C$. The proof of Theorem 9.1.4 is essentially immediate from Theorem 8.0.1 and the definition of the all-or-nothing phase transition.

Theorem 9.1.4. Suppose (π_s) is a sequence of inference tasks that exhibits an all-or-nothing phase transition at threshold (N_s^*) with respect to null models (π_s) . Given $N_s \geq 2N_s^*$, let $(p_{(N_s)}^{s;\theta})_{\theta \in \Theta_s}$ denote the sequence of in-context learning tasks. For any constant $0 < \epsilon < 1$, there exist constants δ, \underline{s} such that for all $s \geq \underline{s}$, next-token prediction for $(p_{(N_s)}^{s;\theta})_{\theta \in \Theta_s}$ exhibits a critical window over

$$[N_s + 2 - (1 + \delta)N_s^*, N_s + 2 - (1 - \delta)N_s^*]$$

in which we transition from sampling a distribution $O(\epsilon)$ -close in TV to $S_{\text{before}} = \Theta_s$, to sampling from a distribution $O(\epsilon)$ -close in TV to $S_{\text{after}} = \Theta_{s;\text{signal}}$.

In other words, we have $T_{\text{before}}^C \triangleq N_s + 2 - (1 - \delta)N_s^*$ and $T_{\text{after}}^C \triangleq N_s + 2 - (1 + \delta)N_s^*$.

Proof. We first apply Theorem 8.0.1 to $S_{\text{init}} = S_{\text{target}} = \Theta_{s;\text{signal}}$. Since we sample from the signal and the noise with equal probability, the weight parameter W is 1. It is also trivial that $T_{\text{end}}(\epsilon) = 0$. Finally, because strong detection is possible provided there are $N \geq \beta N_s^*$ in-context examples for $\beta > 1$, there exists δ_1 depending only on ϵ for which $\text{TV}(p_t^{S_{\text{targ}}}, p_t^{\Theta_s - S_{\text{targ}}}) \geq 1 - \epsilon^2$ for $t = N_s + 2 - (1 + \delta_1)N_s^*$. By Theorem 8.0.1 we conclude that $\text{TV}(p^{S_{\text{init}}, N_s + 2 - (1 + \delta_1)N_s^*}) \lesssim \epsilon$. Next, let us apply Theorem 8.0.1 to $S_{\text{init}} = \Theta_{s;\text{signal}}$ and $S_{\text{targ}} = \Theta_s$. We obviously have that $T_{\text{start}}(\epsilon) = N_s + 2$, and weak detection is impossible provided there are $N \leq \beta N_s^*$ in-context examples for $\beta < 1$. Thus there exists δ_2 depending only on ϵ for which $\text{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{targ}}}) \leq \epsilon$ for $t = N_s + 2 - (1 - \delta_2)N_s^*$. By Theorem 8.0.1 we conclude that $\text{TV}(p^{S_{\text{init}}, N_s + 2 - (1 - \delta_2)N_s^*}) \leq \epsilon$. Taking $\delta = \max(\delta_1, \delta_2)$ concludes the proof. \square

Jailbreak experiment

Now we apply our critical windows theory to develop a new jailbreak detection method, based on a likelihood ratio between an aligned and unaligned model. Intuitively, our theory states that when the unaligned component assigns a high probability to the text compared to the entire model, the model is likely to be jailbroken. We use a LLAMA-3.1-8B-Instruct model jailbroken with LoRA to not refuse harmful prompts [gri24] as a proxy for the unaligned model. We evaluate these different methods on a dataset of jailbreaks and benign prompts from [BSS⁺24].

Dataset. We use the same dataset as [BSS⁺24] but provide details here for completeness. The benign dataset consists of inputs from UltraChat [DCX⁺23], a large dialogue dataset, and Xtest [RKV⁺24], which contains benign queries that are often incorrectly refused by language models. The benign queries are filtered to ensure that LLAMA-3.1-8B-Instruct does not refuse any of them. The dataset of harmful prompts is based off of the Circuit Breakers dataset [ZPW⁺24] and includes the following jailbreaking methods from the extant literature: PAIR [CRD⁺23], AutoDAN [LXCX24], Many-Shot Jailbreaking (MSJ) [ADS⁺24], Multi-Turn Attacks [LHS⁺24, Hai24a], Prefill, GCG [ZWC⁺23], and other Misc. attacks from [WHS23]. For each jailbreaking method, it is applied to a prompt from the Circuit Breaker dataset and evaluated to see if the generation from

LLAMA-3.1-8B-Instruct is helpful and harmful, as determined by the StrongReject jailbreaking classifier [SLB⁺24]).

As is standard in the jailbreak detection literature [BSS⁺24], we report the recall at the false positive rate at 0.01. Table 9.1.1 displays the recall of our method and several other baselines. Crucially, the log likelihood ratio methods does obtain recall more than 0.05 for 7 different categories of jailbreaks. While our methods do perform worse than existing methods, it is important to note that they work at all and that their poor performance could be explained by the fact that we have to use a proxy for the unaligned mode of the model.

	AutoDAN	GCG	Multi-Turn	Misc	MSJ	Pair	Prefill
$\log p_{\text{prompt}}^{\text{unaligned}} - \log p_{\text{prompt}}^{\text{aligned}}$	0.000	0.000	0.028	0.000	0.063	0.000	0.077
$\log p_{\text{gen}}^{\text{unaligned}} - \log p_{\text{gen}}^{\text{aligned}}$	0.082	0.030	0.000	0.100	0.000	0.061	0.051
$\log p_{\text{prompt}}^{\text{aligned}}$	0.000	0.576	0.056	0.063	0.013	0.000	0.077
$\log p_{\text{gen}}^{\text{aligned}}$	0.205	0.150	0.570	0.200	0.006	0.015	0.416
MLP	1.00	0.956	0.873	0.663	1.00	0.833	1.00

Table 9.1.1: Recall (FPR=0.01) for our likelihood ratio threshold, a perplexity threshold [AK23], and a MLP-based detector trained on activations [BSS⁺24] for predicting different jailbreaks. prompt/gen denote the logprobs of the prompt and generation, respectively.

As in Chapter 6, we can apply our critical windows theory to develop a theory for hierarchical samplers as an example of it.

9.2 Hierarchical sampling interpretation

Like in the other examples, our improved overall theory also leads to an improved theory for hierarchical sampling. Hierarchies appear in other localization-based samplers besides diffusion; we can see more prominent examples of a single trajectory containing multiple critical windows (Figure 9.3.1), each splitting a sub-population into smaller sub-populations. Again this hierarchy is naturally represented as a tree: the root signifies that all sub-populations are indistinguishable under enough noise, while the leafs represent distinct modes in p . A path from the root to a leaf captures the progressive refinement of the original distribution p into increasingly specific components. To formalize this, we introduce the concept of an ϵ -mixture tree, which decomposes p into a hierarchical structure, for general mixture models. This includes a tree which encodes the progressive refinement or degradation, and a functions which maps nodes of the tree to noise levels of the forward process.

Definition 9.2.1. For an error term $\epsilon > 0$ and mixture model p , an ϵ -mixture tree is a tuple

$$(T, \{\vec{\Pr}[\cdot|\cdot]\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta \in \Theta}, \text{Subset}, \text{NoiseAmount}).$$

The tree $T = (V, E)$ is associated with a function $\text{Subset}: V \rightarrow 2^\Theta \setminus \{\emptyset\}$, which maps vertices to

sub-mixtures. We require Subset satisfies the following two properties:

1. $\text{Subset}(\text{root}) = \Theta$
2. If u is a parent of v , $\text{Subset}(v) \subset \text{Subset}(u)$

We consider a $\text{NoiseAmount}: V \rightarrow \mathbb{R}^{\geq 0}$, which characterizes the noise levels that result in the aggregations of mixture components described by vertices in the mixture tree. $\text{NoiseAmount}(u)$ is defined such that all $p_{\text{NoiseAmount}(u)}^\theta$ for $\theta \in \text{Subset}(u)$ overlap greatly and for $p_{\text{NoiseAmount}(u)}^{\text{Subset}(u)}, p_{\text{NoiseAmount}(u)}^{\Theta - \text{Subset}(u)}$ have negligible overlap. Thus we require that NoiseAmount satisfy three properties:

1. For distinct $\theta_i, \theta_j \in \Theta$ with leaf nodes w, v such that $\theta_i \in \text{Subset}(w), \theta_j \in \text{Subset}(v)$, if u is the lowest common ancestor of w, v , then we require $\text{TV}(p_{\text{NoiseAmount}(u)}^{\theta_i}, p_{\text{NoiseAmount}(u)}^{\theta_j}) \leq \epsilon$.
2. For $u \in V$, we have statistical separation between $\text{Subset}(u)$ and $\Theta - \text{Subset}(u)$ in terms of TV , $\text{TV}(p_{\text{NoiseAmount}(u)}^{\text{Subset}(u)}, p_{\text{NoiseAmount}(u)}^{\Theta - \text{Subset}(u)}) \geq 1 - \epsilon^2$.
3. If $v \in V$ is a parent of u , we have $\text{NoiseAmount}(u) < \text{NoiseAmount}(v)$. Property 1 establishes bounds on T_{end}^S , and properties 2 and 3 establishes bounds on T_{start}^S .

We emphasize that this framework is highly general, solely defined with the *initial distribution* p and *the forward process*. It strictly expands Definition 6.2.1, which focused on hierarchies of isotropic Gaussians, to all localization-based samplers and mixture models. We can also relate it to the sequences of critical windows we observe in Figure 9.3.1, like we did the diffusion section, capturing the idea that each critical window represents the refinement into smaller subpopulations of p .

Corollary 9.2.1. Consider an ϵ -mixture tree. For $\theta_i \in \Theta$, consider the path $u_1, u_2, u_3, \dots, u_{H'} \in V$ where u_1 is the leaf node with $\theta_i \in \text{Subset}(u_1)$ and $u_{H'}$ is the root. There is a sequence of times $T_1 < T_2 < \dots < T_{H'}$ with $\text{TV}(p^{\{i\}, T_\ell}, p^{\text{Subset}(u_\ell)}) \lesssim_w \epsilon$.

Proof. For $\ell \in [H']$, we let $T_\ell = \text{NoiseAmount}(u_\ell)$. We apply Theorem 8.0.1 with $S_{\text{init}} = \{i\}$ and $S_{\text{target}} = \text{Subset}(u_\ell)$. We know $\text{TV}(p_{T_\ell}^{S_{\text{target}}}, p_{T_\ell}^{\Theta - S_{\text{target}}}) \geq 1 - \epsilon^2$ by Condition 2 in Definition 9.2.1. By Lemma A.2.4, we know $\text{TV}(p_{T_\ell}^{\{i\}}, p_{T_\ell}^{S_{\text{init}}}) \leq \max_{j \in S_{\text{init}}} \text{TV}(p_{T_\ell}^{\{i\}}, p_{T_\ell}^{\{j\}})$. This is $\max_{j \in S_{\text{init}}} \text{TV}(p_{T_\ell}^{\{i\}}, p_{T_\ell}^{\{j\}}) \leq \epsilon$ for all $j \in S_{\text{init}}$ by Condition 3 on NoiseAmount and the data processing inequality. \square

Our broader framework has several implications for our theory. It allows us to reason about different samplers with perhaps different forward processes. In particular, we observe that the hierarchy of two samplers with the same forward process are identical if the samplers agree on subpopulations. Assume we have $\{p^\theta\}_{\theta \in \Theta}$ (e.g. the true distribution) and $\{q^\theta\}_{\theta \in \Theta}$ (e.g. a generative model), where $q^\theta \approx p^\theta$ across all $\theta \in \Theta$ with the same $\{w_\theta\}_{\theta \in \Theta}$.

Corollary 9.2.2. Consider an ϵ -mixture tree $(T, \{\Pr^{\rightarrow, \cdot}\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta \in \Theta}, \text{Subset}, \text{NoiseAmount})$. Suppose we have another distribution $\{q^\theta\}_{\theta \in \Theta}$ such that $\text{TV}(p^\theta, q^\theta) \leq \delta/2$ for all $\theta \in \Theta$. Then we have $\epsilon + \sqrt{\delta}$ -mixture tree given by $(T, \{\Pr^{\rightarrow, \cdot}\}, \mathbf{I}, \Theta, \{q^\theta\}_{\theta \in \Theta}, \text{Subset}, \text{NoiseAmount})$.

Proof. We need only check the first and second properties of NoiseAmount with parameter $\epsilon + \sqrt{\delta}$. To do this, it suffices to show

$$\text{TV}(q_{\text{NoiseAmount}(u)}^{\theta_i}, q_{\text{NoiseAmount}(v)}^{\theta_j}) \leq \epsilon + \delta$$

and

$$\text{TV}(q_{\text{NoiseAmount}(u)}^{\text{Subset}(u)}, q_{\text{NoiseAmount}(u)}^{\Theta - \text{Subset}(u)}) \geq 1 - \epsilon^2 - \delta.$$

By the data processing inequality, we just need to show this at $t = 0$, and we prove the stronger statement that for $S_1 \subset \Theta$, $\text{TV}(p^{S_1}, q^{S_1}) \leq \delta/2$. This follows from Lemma A.2.4 and $\text{TV}(p^\theta, q^\theta) \leq \delta/2$ for all $\theta \in \Theta$. \square

This similarity does not hold generally, if the generative model does not have the same forward process the data generating procedure. In fact, we can define arbitrary hierarchies by designing an appropriate forward process. This further underscores the idea that the *forward process* plays a major role in determining the hierarchy.

Example 9.2.1. Consider a set of alphabets $\{\mathcal{A}_i\}_{i=1}^d$ and define $\Theta = \{(a_i)_{i=1}^d : \forall i \in [d], a_i \in \mathcal{A}_i\}$ and $p^{\theta_i} = \delta_{\theta_i}$. Let $\mathbf{I} = [0, 1, 2, \dots, d]$. and for any permutation i_1, i_2, \dots, i_d of $[d]$, define a forward process such that at $t \in \mathbf{I}$, we mask all $i_d, i_{d-1}, \dots, i_{d-t}$. This constructs a hierarchy where the values for i_1, i_2, \dots, i_d are decided in that order.

Proof. We construct the following 0-mixture tree as follows. We let the leaf nodes be the set Θ . We let two leaf nodes u, v have the same parent if and only if they share the same values on the alphabet at i_1, i_2, \dots, i_{d-1} ; we also define the parent as the union of all of its children. We now treat the parents we constructed as the roots, and let them have the same parent if and only if they share the same values on the tuple i_1, i_2, \dots, i_{d-2} . We continue to do this until we are left with one root node. We let Subset map each node to the corresponding set and NoiseAmount map each node to its distance from a leaf node.

By the construction of T , it is clear that Subset satisfies the desired properties. For distinct $\theta_i, \theta_j \in \Theta$, the lowest common ancestor of θ_i, θ_j represents the largest k such that indices i_1, \dots, i_k are the same for θ_i, θ_j . Because $p_{\text{NoiseAmount}(u)}^{\theta_i}$ is just the tuple of the values of θ_i, θ_j at i_1, \dots, i_k , we know $\text{TV}(p_{\text{NoiseAmount}(u)}^{\theta_i}, p_{\text{NoiseAmount}(u)}^{\theta_j}) = 0$. For any $u \in V$ representing the values at index $(i_\ell)_{\ell=1}^k$, all $\theta \notin \text{Subset}(u)$ does not share the same values at these indices by definition, so we also

know

$$\text{TV}(p_{\text{NoiseAmount}(u)}^{\text{Subset}(u)}, p_{\text{NoiseAmount}(u)}^{\Theta-\text{Subset}(u)}) = 1.$$

□

Finally, we note that hierarchies of diffusions are generally shallower than hierarchies for autoregressive models. The hierarchy for a mixture of Gaussians cannot grow linearly with the dimension d , e.g. it is $O(1)$ in Example 9.1.2 for mixtures of Gaussians with randomly selected means or $O(\sqrt{\log d})$ in the hierarchy of Gaussians in Theorem 6.2.2. This is because the forward process for Gaussians simultaneously contracts all distances with the same dependence on d together at the same time. However, in contrast, depth can scale linearly with the context length for autoregressive models, e.g., Example 9.2.1 or Figure 9.3.1. We speculate that this could mean autoregressive models can learn more complex feature hierarchies than diffusion.

9.3 Experiments on language models

As many authors [HJA20a, MHS⁺22, CLS⁺22, RA23, GVS⁺23, SFLW24, SFW25, BBdBM24, LC24] have already empirically studied critical windows in the context of diffusion, we focus on experiments on critical windows for LLMs. We seek to practically verify our theory and apply it to relevant settings. In Section 9.3.1, we validate our theory on outputs with a hierarchical structure, showing strong agreement with Section 9.2. In Section 9.3.2, we probe critical windows for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct in 7 real-world reasoning benchmarks.

9.3.1 Structured output experiments

In general, to verify our theory in terms of $T_{\text{before}}^C, T_{\text{after}}^C$, we have to compute the total variation between truncated responses from an LLM. This usually would take a large number of samples, so to circumvent this issue experimentally, we restrict the diversity of the LLM’s generations and constrain the LLM to generate tokens in a structured format. In particular, we have LLAMA-3.1-8B-Instruct¹ respond to following prompt, which asks it to answer a series of fill-in-the-blank questions in a very specific format. We also prefill the model’s generations with `\n\n 1.` to increase the likelihood that the output comports to this format. To compute $T_{\text{before}}^C, T_{\text{after}}^C$, we look at when the generations diverge based on the first occurrence of the identifying information. For example, the T_{before}^C of the first critical window is the length 1. The , because the first answer has not appeared in the generation, and the T_{after}^C of the first critical window is the length of 1. The P and 1. The N, because that uniquely identifies the answer to the first question.

Figure 9.3.1 plots the probability of obtaining the same answers as the original generation after truncating different amounts from the generation in the forward-reverse experiments, computed with 10,000 generations. Our theory predicts that jumps in the probability will occur at

¹Default sampling parameters of temperature of 0.6 and top-p sampling of 0.9.

$T_{\text{before}}^C, T_{\text{after}}^C$ which represent when the model has committed to a particular answer to a question in the generation. These predictions are validated with our experiments, as the jumps in probability, representing the model localizing to a more specific set of answers, occur exactly at $T_{\text{before}}^C, T_{\text{after}}^C$.

Structured Output Prompt

Complete the following by choosing only one option for each blank. The options are provided in parentheses, and your response must match the exact case and meaning of the chosen option. Respond with only the completed sentence, no explanations or additional text.

1. The (Pirate/Ninja) jumped across the ship.
2. She adopted a (Dog/Cat) from the shelter.
3. The (River/Bridge) sparkled under the sun.
4. A (Dragon/Knight) guarded the castle gates.
5. He ordered (Pizza/Sushi) for dinner.

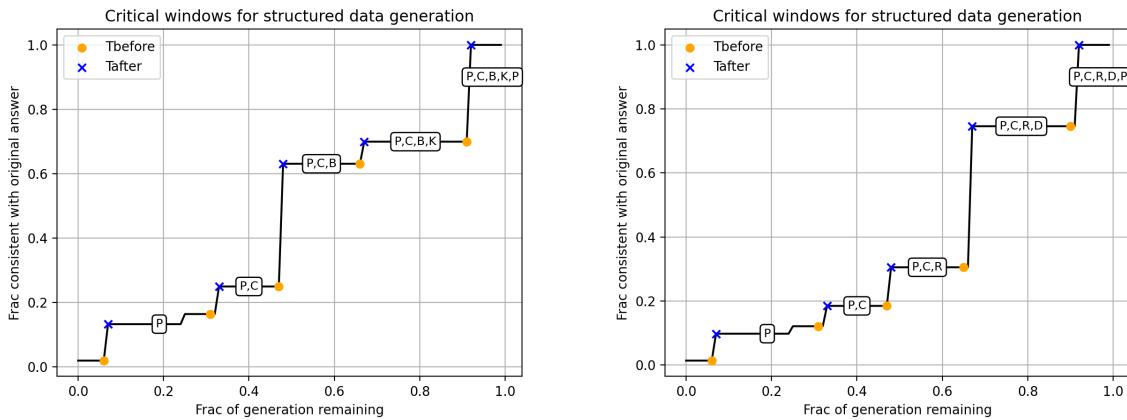


Figure 9.3.1: Structured output plots for LLAMA-3.1-8B-Instruct. P denotes that we are sampling from responses whose answer to the first question was Pirate; P,C denotes that we are sampling from responses whose answers to the first two questions were Pirate and Cat, respectively. We can see that the critical windows directly correspond to our theoretical values for $T_{\text{before}}^C, T_{\text{after}}^C$.

9.3.2 Chain of thought experiments

We then identify critical windows for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct on 7 different math and reasoning benchmarks on which performance is known to improve with chain of thought reasoning [LCR⁺23]: ARC Challenge and Easy [CCE⁺18], AQua [LYDB17], LogiQA [LCL⁺20], MMLU [HBB⁺21], and TruthfulQA [LHE22] multiple-choice benchmarks and the MATH benchmark from [HBK⁺21]. In the forward-reverse experiments, we take the original generation, truncate a fraction of tokens, and check if resampling yields the same answer, using a direct text comparison for the multiple choice benchmarks and the prm800k grader for

MATH [LKB⁺23]. We do this for 400 questions from each dataset and resample at each truncation fraction 100 times. Critical windows, defined as a larger than 0.5 jump in the probability of obtaining the same answer in consecutive truncation fractions, appear prominently across all models and benchmarks that we tested (Figures 9.3.3); for MATH, they occur in 42.2%, 16.6%, 30.2% of generations from LLAMA-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Phi-3-7B-Instruct, respectively.

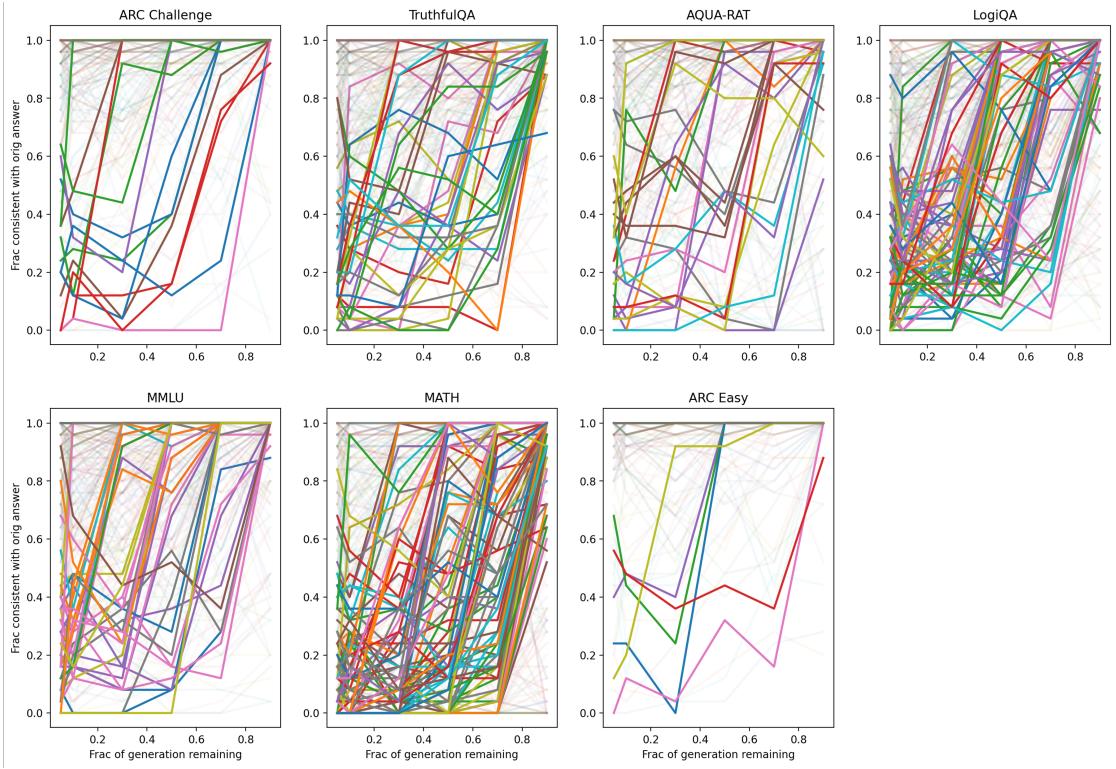


Figure 9.3.3: Probability that the answer is the same as a function of the percentage of the generation remaining for different math and reasoning benchmarks for LLAMA-3.1-8B-Instruct. Highlighted are generations with a 50% increase and no 30% decrease in the probability over subsequent fractions of sampling. We can see that the frequency of critical windows is especially high for the MATH benchmark.

These jumps occur during important steps of reasoning: in Figure 9.3.4, the correct formula and problem solving approach are first expressed in the critical window, and after the critical window, the model simply executes on that plan to solve the problem. Furthermore, we find that generations with critical windows are significantly *less accurate* than generations without critical windows across all datasets and models. Across all three models and datasets that we tested, critical windows result in up to **73% lower accuracy**, and incorrect responses contain **11%-33% more critical windows** (Table 9.3.1). This finding was corroborated by the concurrent works of [AAB⁺24, LLX⁺24]. [AAB⁺24] argues that these critical windows occur at points in the reasoning process where the model says something misleading which throws it off track.

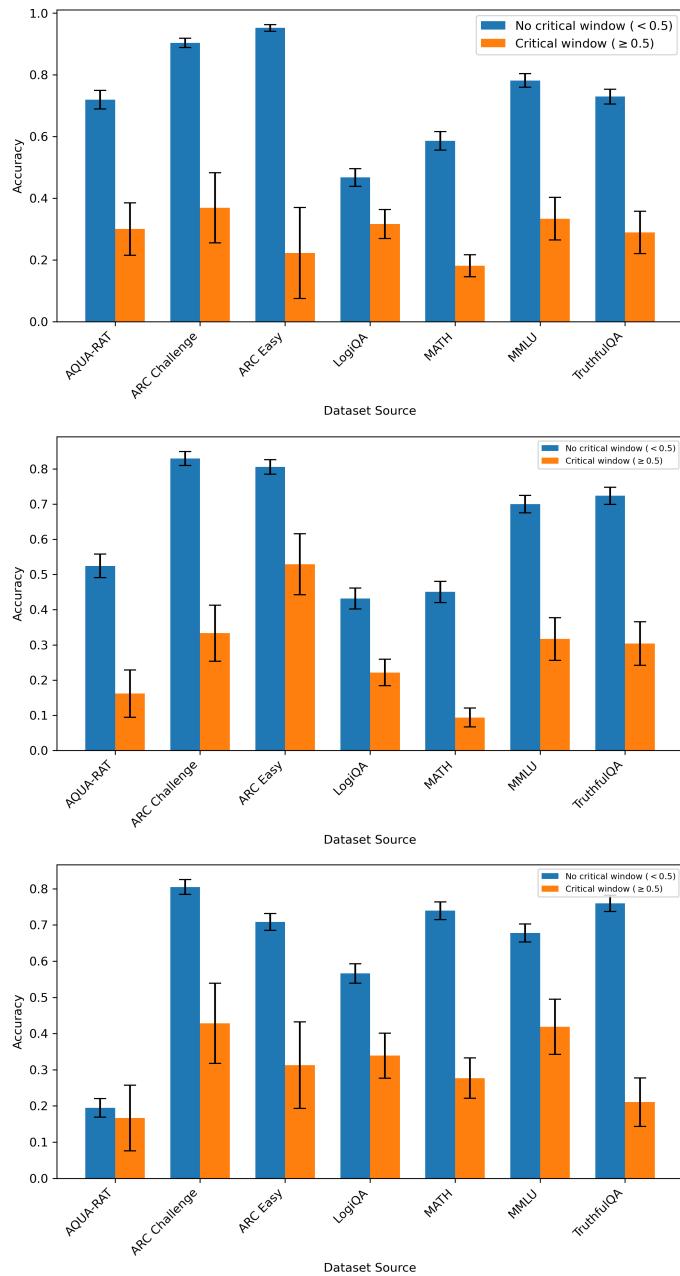


Figure 9.3.2: Left-to-right: LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct barplots of original generation accuracy for generations with and without critical windows. Across all models we tested and benchmarks we evaluated, generations with critical windows have a lower probability of being correct.

Table 9.3.1: Differences between Accuracy (Acc) without versus with critical windows and frequency of critical windows (CW) when the original generation is wrong versus right.

Dataset	LLAMA-3.1-8B-Instruct		Phi-3-7B-Instruct		Qwen-2.5-7B-Instruct	
	ΔAcc	ΔCW	ΔAcc	ΔCW	ΔAcc	ΔCW
AQUA-RAT	0.42	0.20	0.36	0.16	0.03	0.01
ARC Challenge	0.53	0.22	0.50	0.24	0.38	0.11
ARC Easy	0.73	0.26	0.28	0.13	0.40	0.07
LogiQA	0.15	0.11	0.21	0.19	0.23	0.11
MATH	0.41	0.33	0.36	0.33	0.46	0.29
MMLU	0.45	0.24	0.38	0.21	0.26	0.11
TruthfulQA	0.44	0.20	0.42	0.23	0.55	0.23

MATH (Phi-3-7B-Instruct)

Q: If the first term of an arithmetic sequence equals the third minus the second, and the fourth term is 8, what is the first term?

A: 1. We know that the arithmetic sequence has a common difference, which we will denote as " d ". 2. Let's denote the terms in the sequence as T_1 , T_2 , T_3 , and T_4 for the first, second, third, and fourth terms, respectively. 3. We're given the information that $T_1 = T_3 - T_2$. Using the general formula for an arithmetic sequence, we can write this as $T_1 = T_2 + 2d - T_2$, which simplifies to $T_1 = 2d$. 4. We're also given that $T_4 = 8$. The general formula for an arithmetic sequence tells us that $T_4 = T_1 + 3d$. So we can write $8 = T_1 + 3d$. 5. Now we have a system of two equations: 1) $T_1 = 2d$ 2) $8 = T_1 + 3d$. 6. We can substitute equation (1) into equation (2), which gives us $8 = 2d + 3d = 5d$. 7. Solving for d , we get $d = 8/5$. 8. Now that we know d , we can find T_1 . Using the equation from step 1, we get $T_1 = 2d = 2*(8/5) = 16/5$. Therefore, the first term is $T_1 = 16/5$.

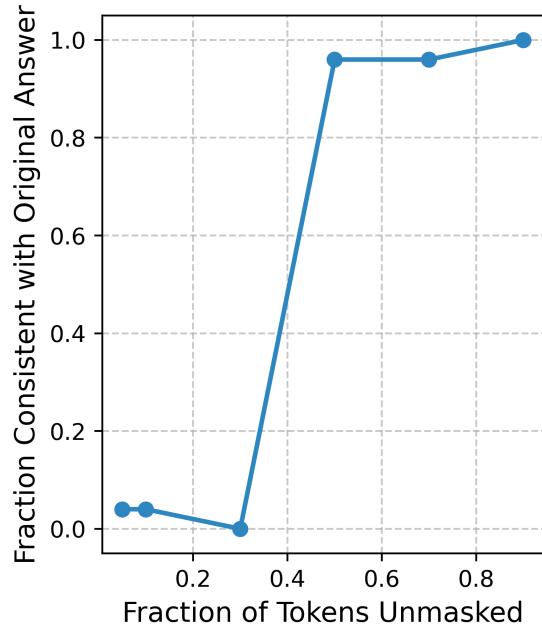


Figure 9.3.4: Example of critical window for Phi-3-7B-Instruct (bolded).

The system prompt for all datasets we used is `Produce a correct solution to the following /TASK/ question.`, where `/TASK/` is the type of question of the dataset, i.e. science, math, or logic. For each question, we create a user prompt by appending `Think of the /TASK/ question`

thoroughly step by step. Please only respond with the answer after reasoning thoroughly. in front of the question. Once the model completes its generation, we append the user prompt Given all of the above, what's the single, most likely answer? Your answer should have the format "The answer is ANSWER", where ANSWER is your answer. for the multiple choice benchmarks and Given all of the above, what's the single, most likely answer? Simplify it completely. Your answer should have the format "The answer is \$ANSWER\$", where ANSWER is your answer in LaTeX. Note that when we ask the model for the final answer, we set the temperature to 0.

Comparison with directly answering for LLAMA-3.1-8B-Instruct

It is worth comparing our approach in Figure 9.3.2 with that from [LCR⁺23], in which the model is directly asked for the answer (Figure 9.3.5). We can see that directly answering has a lower consistency than resampling across different amounts of generation remaining. Interestingly, the discrepancy is much higher for the math-datasets (MATH, AQUA) compared to extremely small difference in consistency for the other reasoning datasets. This may be because it is difficult to extract the answer from the middle of a chain of thought for math datasets than for other datasets.

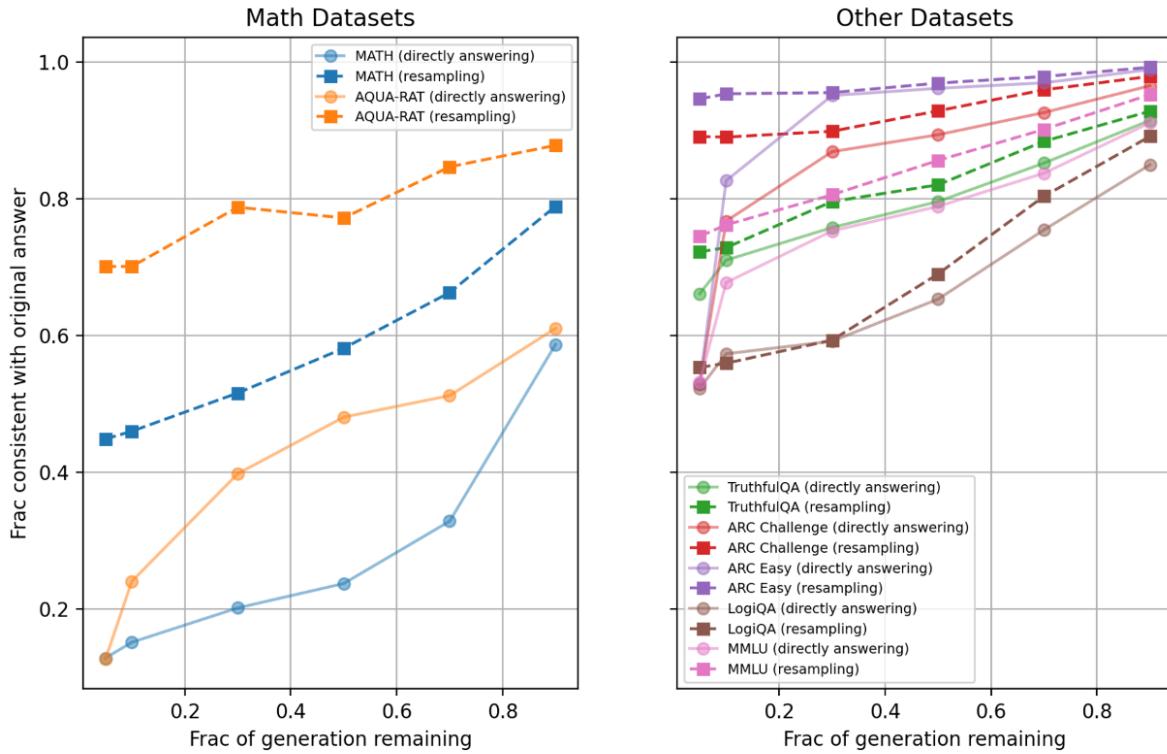


Figure 9.3.5: Difference between resampling to obtain the answer versus directly asking the model for the answer (as in [LCR⁺23]) for LLAMA-3.1-8B-Instruct.

You can recognize truth by its beauty and simplicity.

Richard P. Feynman

The scientist does not study nature because it is useful to do so. He studies it because he takes pleasure in it, and he takes pleasure in it because it is beautiful. If nature were not beautiful it would not be worth knowing, and life would not be worth living.

Henri Poincaré

10

Conclusion

Herein we developed a unifying theoretical framework for critical windows for all stochastic localization samplers, including diffusions and language models. We show that these windows appear when the sampler localizes from a larger to a smaller sub-population of the distribution it models. Our theory builds on rich literature from information theory, machine learning, high-dimensional probability, and statistical physics, and has applications towards privacy, safety, and reasoning.

The central insight of our theory is that we can characterize critical windows as the transition between sampling from a larger sub-population to a smaller sub-population. This reduces the problem of understanding critical windows to understanding when a generative model specializes to a sub-population of the distribution. For a family of high-dimensional Markovian samplers called stochastic localization samplers, this occurs when sub-population has differentiated from its complement but has not yet fragmented internally. In contrast to existing statistical physics works, this new perspective towards critical windows requires much fewer assumptions and spans different models, modalities, or data distributions.

The generality and power of our framework has clear and many benefits; unlike previous works, we can compute critical windows for many different samplers and models. We illustrate that a phase transition in inference can be interpreted as a critical window for in-context learning. Already, practitioners have used critical windows to make language models safer [QPL⁺24] and better at reasoning tasks [AAB⁺24, LLX⁺24], as well as to develop new interpretability and data attribution methods for diffusion [GVS⁺23]. Our theory provides concrete and actionable insight into their methods, and we apply our theory to create a new, effective method to probe the privacy properties of diffusion models, and identify a correlation between critical windows and failures in reasoning

across many different models and benchmarks.

The main takeaway of the thesis is that in trying to understand some empirical phenomena for generative models, it is crucial to do so at the right level of abstraction. One must dissect what is going on at a fundamental level and then derive an explanatory theory with only those basic aspects, no more and no less. This allows for a fully general, and one could say canonical, representation of that intuition. Akin to the way the definition of f -divergences laid the foundations for the data processing inequality in Chapter 2, our use of the stochastic localization sampler framework enabled an extremely simple theory for critical windows.

The most surprising thing about our theory is that it applies across all data distributions and thus circumvents the difficulty of modeling realistic data for theory in generative models. The challenge of modeling natural data in theory has been described as the “dark matter” of machine learning theory [Gan24], and is complicated by numerous examples of *prima facie* reasonable data distributions encountering strong information-theoretic and computational complexity barriers for learning. For example, it is well-known that mixtures of Gaussians can be information-theoretically hard to learn in general, and gradient descent cannot learn generative models described by neural networks with a single hidden layer [CLL22b]. This work avoids these issues of modeling data by fixing the distribution that is to be modeled, and then provably characterizing different behaviors that emerge during sampling. We view this approach as a kind of “reduction” in the same vein as the work in the theory of diffusion which make assumptions on the learned score [CCL⁺23b]. In both of these works, the aspect of learning by a neural network is *abstracted away*, yet the theory still makes meaningful inroads. It would be interesting to see if it were still possible to derive meaningful theoretical insights with similarly weak assumptions in other settings. Now, we will briefly overview some other fruitful directions for future work.

10.1 Future work

10.1.1 Bridging our perspective with the statistical physics literature

The most immediate direction would be to strengthen the relationship between this thesis and the complementary work by statistical physicists on critical windows for specific models of data [Amb23, RA23, SFW25, SFLW24, BBdBM24]. While we have instantiated our theory in many settings (Section 9.1 in particular), we only analytically computed critical windows for one model studied by the previous statistical physics literature, i.e. mixtures of two Gaussians [RA23, BBdBM24]. The difficulty of explicitly calculating critical windows is that it is tricky to find upper bounds on TV or $\frac{1}{2}H^2$ that are close to 1, which are necessary to establish T_{start}^S bounds in our definition of critical windows. Even for seemingly simple localization-based samplers, i.e. a random masking sampler on discrete data, where it often seemed intuitively obvious when such a bound is true, it was challenging to explicitly pin down the critical windows for more complex models of data than what we considered. Future work could compute the location of critical windows for other models of data with existing statistical physics predictions, e.g., the random hierarchy model [SFLW24].

This line of work would mirror the existing research program for the all-or-nothing phenomenon, in which rigorous proofs for interesting learning models like sparse regression [RXZ19, GZ19] followed precise, non-rigorous statistical physics-based predictions for these thresholds.

10.1.2 Connection to fast mixing in generic Monte Carlo Markov Chains

We also wonder if there is a connection between our critical windows and the phenomena of cut-offs in other Monte Carlo Markov Chains, in which the Markov Chain converges to stationarity almost instantaneously [BHP17]. Famously, Aldous and Diaconis established the critical window of $n \log n \pm O(n)$ “top in at random” shuffles to adequately shuffle a deck of n cards [AD86]. This cutoff phenomenon has been established for Markov Chains without an explicit forward or reverse process, but we speculate that there may be a deeper connection between these notions of criticality if one could define a forward process consistent with posteriors that match the steps of the Markov Chain sampler. [CE22] already showed that hit-and-run walks and Glauber dynamics on certain models of data can be associated with specific localization schemes.

10.1.3 Extending theory to continuous features

It would be interesting to discover analogues of critical windows for continuous features. For example, when we noise and denoise an orange car, we expect that it takes fewer time steps to see pictures of red or yellow cars than purple cars, because orange is more similar to red or yellow than purple. Features such as color, height, and orientation more naturally belong to a *continuum* rather than discrete bins, but our critical window theorems require strong statistical separation between components inside and outside the target sub-mixture and cannot capture this intuition.

10.1.4 Critical windows in the wild

On an empirical front, it could be useful to develop a taxonomy of features and hierarchies of features for image and text generation models in the wild. The structure of feature hierarchies beyond the simple examples this thesis considered could be interesting scientifically, and furthermore, could prove helpful to identify certain un-intuitive behaviors for language models like the Yellowstone demo [Ant24]. This would require a method to systematically identify and label features for generations, perhaps a multimodal LLM-as-a-judge approach [ZCS⁺23].

10.1.5 Optimization and critical windows

Existing empirical work has found much success in employing critical windows to develop finetuning or preference optimization methods for language models [AAB⁺24, LLX⁺24, QPL⁺24]. It would be interesting to explore whether we can theoretically understand these methods with our critical windows framework, by characterizing how the behavior changes at a few critical points.

A

Miscellanea

In this appendix, we overview proofs that were deferred from the main body.

In Sections A.1, A.2 and A.2.1, we provide some more proofs for statistical divergences and metrics. In Section A.2.2, we include some details about specific distributions. In Section A.3, we present omitted background for the theory of diffusion and Markov processes.

A.1 Special f -divergences

In this section, we relate several f -divergences to inference problems, which ties into several places in our thesis, i.e. operational characterization of total variation in Definition 9.1.1. Because f -divergences measure the difference between different distributions, some f -divergences might be useful to characterize when inference tasks are possible or impossible. Consider the problem of distinguishing between a sample x drawn from probability distributions $x \sim P$ or $x \sim Q$ by constructing an estimator $\mathcal{A} : \mathcal{X} \rightarrow \{P, Q\}$. We want to establish bounds on how good \mathcal{A} can be in terms of f -divergences between P, Q .

A.1.1 Operational characterization of TV

If $\text{TV}(P, Q) \approx 0$, it should be impossible for any \mathcal{A} to distinguish between samples $x \sim P$ or $x \sim Q$. We can make this connection more formal by appealing to another definition of TV.

Definition A.1.1 (Alternative definition of TV). One can also define $\text{TV}(P, Q) = \sup_{A \in \Omega} |\Pr_{X \sim P}(X \in A) - \Pr_{X \sim Q}(X \in A)|$.

Lemma A.1.1. The sum of the false positive and false negative rates of \mathcal{A} is at least

$$\Pr_{X \sim P}(X \notin A) + \Pr_{X \sim Q}(X \in A) \geq 1 - \text{TV}(P, Q).$$

This can be interpreted as the twice the error of \mathcal{A} if we flip a fair coin to decide to give \mathcal{A} a sample from $x \sim P$ or $x \sim Q$, and the error is the probability that it mis-classifies x .

Proof. \mathcal{A} can be parameterized by some set $A = \mathcal{A}^{-1}(P) \in \Omega$. The sum of the false positive and false negative rate is $1 + \Pr_{X \sim Q}(X \in A) - \Pr_{X \sim P}(X \in A)$. The error is minimized by the minimum of $\Pr_{X \sim Q}(X \in A) - \Pr_{X \sim P}(X \in A)$, which is exactly $-\text{TV}(P, Q)$. \square

While the proof is almost tautological, Lemma A.1.1 highlights an extremely important idea in information theory, that some estimation problems are *information-theoretically impossible*. No algorithm can successfully obtain non-trivial error on this task if the TV is sufficiently small.

A.1.2 Tensorization of Hellinger distance

While the general problem of distinguishing samples from P, Q may be intractable in general if the initial distributions have small TV with each other, it may be more possible if P, Q consist of i.i.d. samples, or are product measures $\Gamma^{\otimes n}, \Phi^{\otimes n}$. We should expect that if the distance between Γ, Φ is non-zero under some f -divergence, some classifier \mathcal{A} should be able to distinguish between $\Gamma^{\otimes n}, \Phi^{\otimes n}$ as the number of samples $n \rightarrow \infty$. While TV is hard to directly control for product measures, the Hellinger distance has nice structure that allows it to factor along the product distribution and obtain usable bounds for product measures.

Fact A.1.2 (Hellinger lower bounds total variation). $\frac{1}{2}\text{H}^2 \leq \text{TV}$.

Lemma A.1.3 (Tensorization of Hellinger distance). We have

$$\text{H}^2(\Gamma^{\otimes n}, \Phi^{\otimes n}) = 2 - 2 \left(1 - \frac{\text{H}^2(\Gamma, \Phi)}{2} \right)^n$$

Proof. Because $\text{H}^2(\Gamma^{\otimes n}, \Phi^{\otimes n}) = 2 - 2 \int \sqrt{d\Gamma^{\otimes n} d\Phi^{\otimes n}} d\mu^{\otimes n}$, we have

$$\text{H}^2(\Gamma^{\otimes n}, \Phi^{\otimes n}) = 2 - 2 \left(\int \sqrt{d\Gamma d\Phi} d\mu \right)^n = 2 - 2 \left(1 - \frac{\text{H}^2(\Gamma, \Phi)}{2} \right)^n.$$

\square

In particular, observe that if $\text{H}^2(\Gamma, \Phi) > 0$, then $\text{H}^2(\Gamma^{\otimes n}, \Phi^{\otimes n}) \rightarrow 2$ as $n \rightarrow \infty$. We can conclude that $\text{TV}(\Gamma^{\otimes n}, \Phi^{\otimes n}) \rightarrow 1$ as well. This means there exists an algorithm \mathcal{A} which solves this hypothesis testing question in the limit of infinitely many samples. However, this does not say that the algorithm \mathcal{A} which can is computationally efficient. When an inference problem is information-theoretically tractable but suspected to not have computationally efficient solutions, then it is said that there is a *computational-statistical gap*.

A.2 Proof of Lemma 2.1.1

Herein we complete the proof of Lemma 2.1.1 from Chapter 2 by chaining together short and well-known lemmas proved in [LeC86] that relate these different divergences and metrics to each other.

Lemma 2.1.1. [LC24] For probability measures P, Q ,

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

Lemma A.2.1. $\text{LC}(P, Q) \geq \frac{1}{2}\text{H}^2(P, Q)$

Proof. This follows from the observation that $\frac{(\sqrt{dP} + \sqrt{dQ})^2}{d(P+Q)} \geq 1$ and some simple algebra,

$$\begin{aligned} \frac{1}{2} \int \frac{(dP - dQ)^2}{d(P+Q)} d\mu &= \frac{1}{2} \int \frac{(\sqrt{dP} - \sqrt{dQ})^2 (\sqrt{dP} + \sqrt{dQ})^2}{d(P+Q)} d\mu \\ &\geq \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2 d\mu. \end{aligned}$$

□

Lemma A.2.2. $4\text{TV}^2(P, Q) \leq \text{H}^2(P, Q)(4 - \text{H}^2(P, Q))$

Proof. By the Cauchy-Schwarz inequality, we have

$$4\text{TV}^2(P, Q) = \left(\int |dP - dQ| d\mu \right)^2 \leq \left[\int (\sqrt{dP} - \sqrt{dQ})^2 d\mu \right] \left[\int (\sqrt{dP} + \sqrt{dQ})^2 d\mu \right].$$

We know $\int (\sqrt{dP} + \sqrt{dQ})^2 d\mu = 2 + 2 \int \sqrt{dP} d\sqrt{dQ} d\mu = 4 - \text{H}^2(P, Q)$. We are done. □

Now, we are ready to prove Lemma 2.1.1.

Proof of Lemma 2.1.1. We exhibit the leftmost equality by noting

$$dP dQ = \frac{1}{4} ((dP + dQ)^2 - (dP - dQ)^2),$$

which gives us

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

The first inequality follows from $\text{LC}(P, Q) \geq \frac{1}{2}\text{H}^2(P, Q)$. The second inequality follows from rearranging $4\text{TV}^2(P, Q) \leq \text{H}^2(P, Q)(4 - \text{H}^2(P, Q))$ into $1 - \frac{1}{2}\text{H}^2(P, Q) \leq \sqrt{1 - \text{TV}^2(P, Q)}$. □

A.2.1 Other facts about statistical divergences

We will use other properties of statistical divergences in our proofs. The following are lemmas that we will quickly state and prove.

Fact A.2.3 (Pinsker's inequality). $\text{TV} \leq \sqrt{\frac{1}{2}\text{KL}}$.

Lemma A.2.4 (Difference in TV of mixtures). Consider mixture $P = \sum_i a_i P_i$ and mixture $Q = \sum_i b_i Q_i$. If $\text{TV}(P_i, Q_j) \leq \epsilon$ for all i, j , then

$$\text{TV}(P, Q) \leq \epsilon.$$

Proof. This is a simple application of triangle inequality,

$$\begin{aligned} \frac{1}{2} \int \left| \sum_i a_i dP_i - \sum_j b_j dQ_j \right| &\leq \frac{1}{2} \sum_i a_i \int |dP_i - \sum_j b_j dQ_j| \\ &\leq \frac{1}{2} \sum_i a_i \sum_j b_j \int |dP_i - dQ_j| \leq \epsilon. \end{aligned}$$

□

The next set of results are more involved and necessary to build up to Lemma A.2.7, which allows us to establish KL-bounds in terms of W_2 on distributions convolved with small amounts of Gaussian noise. We first establish the joint-convexity of the KL-divergence, which for discrete distributions essentially follows directly from a log-sum inequality.

Fact A.2.5 (Log-sum inequality). For all non-negative $(a_i)_{i=1}^n, (b_i)_{i=1}^n$, we have

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

Lemma A.2.6 (Joint convexity of KL). For any $\lambda \in [0, 1]$ and discrete distributions p_1, p_2, q_1, q_2 , we have

$$KL(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda KL(p_1 || q_1) + (1-\lambda) KL(q_1 || q_2).$$

Proof. For $\lambda \in (0, 1)$, we find

$$KL(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) = \sum_{s \in \Omega} (\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}$$

For each individual $s \in \Omega$, we know $(\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \leq \lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda)p_2 \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2}$ by the log-sum inequality, yielding

$$\begin{aligned} KL(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) &\leq \sum_{s \in \Omega} \lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda)p_2 \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2} \\ &= \lambda \sum_{s \in \Omega} p_1 \log \frac{p_1}{q_1} + (1-\lambda) \sum_{s \in \Omega} p_2 \log \frac{p_2}{q_2} \\ &= \lambda KL(p_1 || q_1) + (1-\lambda) KL(p_2 || q_2). \end{aligned}$$

□

A more generalized statement of the joint-convexity of KL allows us to readily prove Lemma A.2.7.

Lemma A.2.7 (Short-time regularization). Convolving with the normal distribution bounds KL in terms of W_2 ,

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

Proof. For delta measures $p = \delta_x$ and $q = \delta_y$, we have

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

Using the optimal coupling in the definition of Wasserstein-2 distance and the joint convexity of KL, we obtain the desired result. \square

A.2.2 Special distributions

It will also be useful to state some useful lemmas for certain special distributions, specifically strongly log-concave and sub-Gaussian distributions for Chapter 5 of Part I.

Lemma A.2.8 ([LC24]). 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, $S \in \text{subG}_d(\sum_{i=1}^n \alpha_i^2 \sigma_i^2)$.

Proof. Plugging in the definition of S , we have

$$\mathbb{E}[\exp(u^\top S)] = \mathbb{E}\left[\exp\left(u^\top \sum_{i=1}^n \alpha_i X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp(\alpha_i u^\top X_i)\right].$$

By the definition of sub-Gaussianity, for any $u \in \mathbb{R}^d$, we have

$$\mathbb{E}[\exp(u^\top X_i)] \leq \exp\left(\frac{\sigma_i^2 \|u\|^2}{2}\right).$$

Thus, we bound the above as and can conclude $S \in \text{subG}_d(\sum_{i=1}^n \alpha_i^2 \sigma_i^2)$,

$$\prod_{i=1}^n \exp\left(\frac{\sigma_i^2 \alpha_i^2 \|u\|^2}{2}\right) = \exp\left(\frac{\|u\|^2}{2} \sum_{i=1}^n \alpha_i^2 \sigma_i^2\right).$$

\square

We will also need certain concentration inequalities about sub-Gaussian vectors, in particular, a tail bound on the concentration inequality of the norm $\|X\|$ of $X \in \text{subG}_d(\sigma^2)$. We present a simplified version of the argument for this result from [RH23]. First, we establish sub-Gaussian concentration inequalities on the suprema of finite collections of sub-Gaussian random variables and then extend it to $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ via an ϵ -net over the unit sphere.

Fact A.2.9 (Lemma 1.3. of [RH23]). Let $X \sim \text{subG}_1(\sigma^2)$. For any $t > 0$, we have

$$\Pr(X \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Lemma A.2.10 (Theorem 1.14 of [RH23]). Let $\{X_i\}_{i=1}^N$ be a set of random vectors such that $X_i \in \text{subG}_1(\sigma^2)$. Then $\Pr(\max X_i \geq t) \leq N e^{-t^2/(2\sigma^2)}$.

Proof. This is a simple application of the union bound inequality and Lemma A.2.9,

$$\Pr(\max X_i \geq t) \leq \sum_{i=1}^N \Pr(X_i \geq t) \leq N \exp(-t^2/(2\sigma^2)).$$

□

While Lemma A.2.10 establishes control over the suprema of *finitely* many sub-Gaussian random variables, in order to control the norm $\|X\| = \sup_{u \in \mathcal{S}^{n-1}} u^\top X$, we need to have control over *infinitely* many sub-Gaussian random variables. We can achieve this by appealing to the geometric structure of \mathcal{S}^{n-1} and exploiting an ϵ -net. An ϵ -net intuitively states that a finite subset of \mathcal{S}^{n-1} covers the entire space, allowing us to apply the control over the finitely many elements in the net.

Definition A.2.11 (ϵ -net). For a metric space (S, d) , the ϵ -net for a set $A \subset S$ is a set \mathcal{A} such that for all $x \in A$ there exists $y \in \mathcal{A}$ such that $d(x, y) \leq \epsilon$.

Fact A.2.11 (Existence of ϵ -net for \mathcal{S}^{n-1}). \mathcal{S}^{n-1} has an ϵ -net \mathcal{N} of cardinality $|\mathcal{N}| \leq (3/\epsilon)^d$.

Now, we are finally ready to establish the main concentration inequality for $\|X\|$.

Lemma A.2.12 (Implicit in Theorem 1.19 of [RH23]). Let $X \in \text{subG}_d(\sigma^2)$. Then, for any $r \geq 0$, $\Pr[\|X\| > r] \leq 6^d \exp(-r^2/(8\sigma^2))$.

Proof. Let \mathcal{N} be a $\frac{1}{2}$ -net of \mathcal{S}^{d-1} . We want to establish an upper bound on $\|X\| = \sup_{u \in \mathcal{S}^{d-1}} u^\top X$. By the definition of ϵ -net, there exists a $v \in \mathcal{N}$ such that $\|u - v\| \leq \frac{1}{2}$. We want to bound

$$\|X\| = \langle v, X \rangle + \langle u - v, X \rangle.$$

Because $v \in \mathcal{N}$, we have $\langle v, X \rangle \leq \frac{1}{2} \sup_{r \in \mathcal{S}^{d-1}} r^\top X$. Note that $u - v$ is a vector of length at most $1/2$, so $\langle u - v, X \rangle \leq \frac{1}{2} \sup_{r \in \mathcal{S}^{d-1}} r^\top X = \frac{1}{2} \|X\|$. Subtracting these two inequalities from each other, we obtain $\|X\| \leq 2 \sup_{r \in \mathcal{N}} r^\top X$. Applying Lemma A.2.10 to

$$\Pr(|X| \geq t) \leq \Pr(2 \sup_{r \in \mathcal{N}} r^\top X \geq t),$$

we obtain the desired result. □

Lemma A.2.13 ([LC24]). 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

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

Proof. By A.2.12, $\Pr(\|X - \mu_X\| \geq R), \Pr(\|Y - \mu_Y\| \geq R) \leq \epsilon$, and $B_R(\mu_X)$ and $B_R(\mu_Y)$ are disjoint by definition. Thus,

$$\text{TV}(X, Y) \geq \int_{B_R(\mu_X)} dP_X - dP_Y \geq 1 - \epsilon.$$

□

A.3 Details about diffusion and generators

In this section, we include additional details about Tweedie's formula, Girsanov's theorem, the Fokker-Planck equation, and Langevin MC for completeness. This integrates exposition from [Hol23, HHY⁺25, Che24, Gal18], but borrows heavily from [Che24, Lectures 21-23]. We first have Tweedie's formula which relates the problem of score estimation to denoising.

Proposition 4.2.2 (Tweedie's formula). For $x \sim p$ and $g \sim \mathcal{N}(0, \sigma^2 \text{Id})$, let $\tilde{x} = x + g$ and \tilde{p} be the density of $p * \mathcal{N}(0, \sigma^2 \text{Id})$. We have,

$$\mathbb{E}[x | \tilde{x}] = \tilde{x} + \sigma^2 \cdot \nabla \ln \tilde{p}(\tilde{x}).$$

Proof. This follows from $\nabla \ln \tilde{p}(\tilde{x}) = \frac{\nabla \tilde{p}(\tilde{x})}{\tilde{p}(\tilde{x})}$. We can write

$$\begin{aligned} \mathbb{E}[x | \tilde{x}] &= \frac{\int xp(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu}{\int p(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu} \\ &= \tilde{x} + \frac{\int(x - \tilde{x})p(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu}{\int p(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu} \\ &= \tilde{x} + \sigma^2 \cdot \frac{\nabla_{\tilde{x}} \int p(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu}{\int p(x)e^{-\|x-\tilde{x}\|^2/(2\sigma^2)}d\mu} \\ &= \tilde{x} + \sigma^2 \frac{\nabla \tilde{p}(\tilde{x})}{\tilde{p}(\tilde{x})}. \end{aligned}$$

□

The next tool we will provide background on is this consequence of Girsanov's theorem, that allows us to bound the differences in the distributions of the outputs of SDEs with different drift terms. We refer the readers to [CCL⁺23b] for a full proof.

Theorem 4.2.3 ([CCL⁺23b, Section 5.2]). Let $(Y_t)_{t \in [0, T]}$ and $(Y'_t)_{t \in [0, T]}$ denote the solutions to

$$\begin{aligned} dY_t &= b_t(Y_t) dt + \sqrt{2} dB_t, & Y_0 &\sim p, \\ dY'_t &= b'_t(Y'_t) dt + \sqrt{2} dB_t, & Y'_0 &\sim p. \end{aligned}$$

Let p and q' denote the laws of Y_T and Y'_T respectively. If b_t, b'_t satisfy

$$\int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt < \infty,$$

then

$$\text{KL}(q \| q') \leq \int_0^T \mathbb{E}_{Q_t} [\|b_t(Y_t) - b'_t(Y_t)\|^2] dt,$$

where Q_t is the law of Y_t .

Informal derivation. By the data processing inequality, it suffices to bound the KL between the path measures of $(Y_t)_{t \in [0, T]}$ and $(Y'_t)_{t \in [0, T]}$. To do this, let us compute the KL of observing the trajectory $(p_0, p_h, p_{2h}, \dots, p_{Nh})$ under a discrete approximation of both SDEs. In other words, consider the stochastic processes

$$\begin{aligned} p_{(k+1)h} &\triangleq p_{kh} + hb_{kh}(p_{kh}) + \sqrt{2h}g_{kh} \\ p'_{(k+1)h} &\triangleq p_{kh} + hb'_{kh}(p_{kh}) + \sqrt{2h}g_{kh}, \end{aligned}$$

where $g_{kh} \sim \mathcal{N}(0, \text{Id})$ and $Nh = T$. By the definition of KL, we want to upper bound

$$\text{KL}(q||q') \leq \mathbb{E}_{(p_0, p_h, p_{2h}, \dots, p_{Nh}) \sim Q_N} \left[\ln \left(\frac{\prod_{k=0}^{N-1} \exp \left(-\frac{1}{4h} \|p_{(k+1)h} - p_{kh} - hb_{kh}(p_{kh})\|^2 \right)}{\prod_{k=0}^{N-1} \exp \left(-\frac{1}{4h} \|p_{(k+1)h} - p_{kh} - hb'_{kh}(p_{kh})\|^2 \right)} \right) \right],$$

where Q_N is the law under b_t . By some simple linear algebra, we can simplify the above,

$$\text{KL}(q||q') \leq \sum_{k=1}^{N-1} \mathbb{E}_{Q_N} \left[-\frac{1}{4h} \left[h^2 (\|b_{kh}(p_{kh})\|^2 - \|b'_{kh}(p_{kh})\|^2) - 2h \langle p_{(k+1)h} - p_{kh}, b_{kh}(p_{kh}) - b'_{kh}(p_{kh}) \rangle \right] \right].$$

For the formula for the likelihood of the law of b' , we can write $p_{(k+1)h} - p_{kh} = hb'(p_{kh}) + \sqrt{2h}g_{kh}$, so the above simplifies into

$$\text{KL}(q||q') \leq \sum_{k=1}^{N-1} \mathbb{E}_{Q_N} \left[-\frac{1}{4h} \left[h^2 (\|b_{kh}(p_{kh})\|^2 - \|b'_{kh}(p_{kh})\|^2) - 2h \langle hb'_{kh}(p_{kh}) + \sqrt{2h}g_{kh}, b_{kh}(p_{kh}) - b'_{kh}(p_{kh}) \rangle \right] \right].$$

One can check that this reduces to

$$\text{KL}(q||q') \leq \sum_{k=1}^{N-1} \mathbb{E}_{Q_N} \left[-\frac{h}{4} \|b_{kh}(p_{kh}) - b'_{kh}(p_{kh})\|^2 + \frac{1}{\sqrt{2}} \langle \sqrt{h}g_{kh}, b_{kh}(p_{kh}) - b'_{kh}(p_{kh}) \rangle \right].$$

Note that $\sqrt{h}g_{kh}$ can be interpreted as the Q' -Brownian motion, which we label dB'_t . Thus we end up as $h \rightarrow 0$ the inequality,

$$\text{KL}(q||q') \leq -\frac{1}{4} \int_0^T \mathbb{E}_{Q_t} \|b_t - b'_t\|^2 dt + \frac{1}{\sqrt{2}} \mathbb{E}_{Q_t} \int_0^T \langle dB'_t, b_t - b'_t \rangle.$$

By Girsanov's theorem, we can define a Q_t -Brownian motion β_t such that

$$b_t dt + \sqrt{2} d\beta_t = b'_t dt + \sqrt{2} dB_t.$$

Note that $dB_t = d\beta_t + \frac{1}{\sqrt{2}}(b_t - b'_t)dt$, and $d\beta_t$ integrates to 0 under Q_t , yielding

$$\text{KL}(q||q') \leq -\frac{1}{4} \int_0^T \mathbb{E}_{Q_t} \|b_t - b'_t\|^2 dt + \frac{1}{2} \mathbb{E}_{Q_t} \int_0^T \|b_t - b'_t\|^2 dt,$$

and we are done. \square

The heart of this theorem is Girsanov's theorem, which allows one to define the Q_t -Brownian motion β_t . However, one can not directly apply Girsanov's theorem due to some regularity conditions. [CCL⁺23b] circumvents these issues with an approximation argument. They consider an increasing sequence of stopping times T_n which converge to T , and then apply Girsanov's on each interval $[0, T_n]$, which they then extend to obtain the result to the entire interval.

The last key theorems we want to present are the Fokker-Planck equation and proof of the stationarity of Langevin MC. Before we do so, we give a brief informal introduction to the theory of Markov processes, which gives us the right language to prove them.

A.3.1 General theory of Markov processes

We list some definitions for the theory of Markov processes that we deferred from the main body. Herein we only present the case for homogenous Markov kernels where the conditionals do not depend on t . We will also state things in terms of the densities as well to make the definitions slightly more intuitive.

Definition A.3.1 (Transition kernel). On a measurable space (Ω, \mathcal{F}) , the function $Q : \Omega \times \Omega \rightarrow [0, 1]$ is a Markovian transition kernel if for every $x \in \Omega$, $Q(x, \cdot)$ is a valid probability density on (Ω, \mathcal{F}) and for every $y \in \Omega$, $x \mapsto Q(x, y)$ is a measurable function.

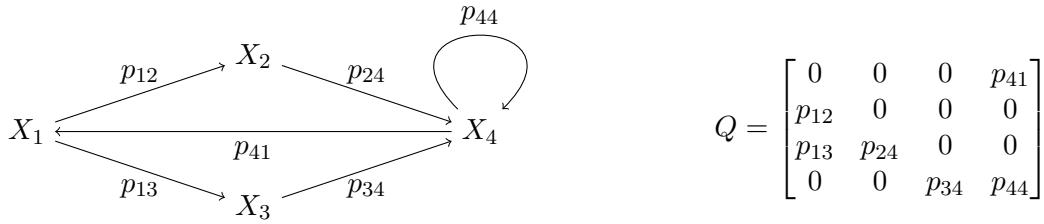


Figure A.3.1: Diagram of Markov chain and the corresponding transition kernel.

We use the shorthand $Qf(x) = \int_{y \in \Omega} Q(x, y)f(y)d\mu = \mathbb{E}_{y \sim Q(x, \cdot)}[f(y)]$ to denote expectations taken with respect to a transition kernel. To define Markov Chains, it will also be useful to define a *transition semigroup* that is a family of transition kernels in which composition is a closed operation.

Definition A.3.2 (Transition semigroup). A collection $(Q_t)_{t \geq 0}$ of transition kernels on Ω is a *transition semigroup* if for every $x \in \Omega$, $Q_0(x, y) = \delta_x(y)$ and for every $s, t \geq 0$ and $x, y, z \in \Omega$, $Q_{t+s}(x, z) = \int_{y \in \Omega} Q_t(x, y)Q_s(y, z)d\mu$ (Chapman-Kolmogorov identity).

We restrict our attention to Feller semigroups, which satisfy some nice regularity properties. We can now define a Markov process by its compatibility with a Feller semigroup, which characterizes the intuition that the current state contains all the information from the past states.

Definition A.3.3 (Feller semigroup). A transition semigroup $(Q_t)_{t \geq 0}$ on (Ω, \mathcal{F}) is a *Feller semigroup* if for all $f \in \mathcal{C}_0(E)$, we have $Q_tf \in \mathcal{C}_0(E)$ and for all $f \in \mathcal{C}_0(E)$, we have $\|Q_tf - f\|_\infty \rightarrow 0$ as $t \rightarrow 0$.

Definition A.3.4. Let $(Q_t)_{t \geq 0}$ be a transition semigroup on (Ω, \mathcal{F}) . A *Markov process* $(X_t)_{t \geq 0}$ with transition semigroup $(Q_t)_{t \geq 0}$ is a stochastic process such that, for all $s, t \geq 0$

$$\Pr[X_{s+t} = a \mid (X_r)_{0 \leq r \leq s}] = Q_t(X_s, a).$$

$(X_t)_{t \geq 0}$ is a Feller process if its semigroup is Feller.

A.3.2 Derivation of Fokker-Planck and stationarity of Langevin MC

We can now build up to proving the Fokker-Planck equation with the language of generators. First, we establish an exchangeability condition that allows us to move the transition kernel and the generator around each other.

Theorem A.3.1 (Kolmogov backward equation). $\partial_t Q_t f = \mathcal{L} Q_t f = Q_t \mathcal{L} f$.

Proof. Both of these are a simple consequence of the linearity of the transition kernel and the Chapman-Kolmogov identity. For the first equality, we can solve

$$\begin{aligned}\partial_t Q_t f &= \lim_{h \rightarrow 0} \frac{Q_{t+h} f - Q_t f}{t} \\ &= \lim_{h \rightarrow 0} \left(\frac{Q_h(Q_t f) - Q_t f}{t} \right) \\ &= \mathcal{L} Q_t f.\end{aligned}$$

We can also write

$$\begin{aligned}\partial_t Q_t f &= \lim_{h \rightarrow 0} \frac{Q_{t+h} f - Q_t f}{t} \\ &= \lim_{h \rightarrow 0} Q_t \left(\frac{Q_h f - f}{t} \right) \\ &= Q_t \lim_{h \rightarrow 0} \frac{Q_h f - f}{t} \\ &= Q_t \mathcal{L} f.\end{aligned}$$

□

With this equality in-hand, we are now ready to formally establish the Fokker-Planck equation.

Theorem 4.2.2 (Fokker-Planck equation). Let $b_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a family of smoothly varying vector fields. The solution $(X_t)_{t \geq 0}$ of the following SDE

$$dx_t = b_t dt + \sqrt{2} dB_t, \quad t \geq 0$$

is distributed according to $(p_t)_{t \geq 0}$ which satisfies the following partial differential equation,

$$\partial_t p_t = -\operatorname{div}(p_t b_t) + \Delta p_t.$$

Proof sketch of Fokker-Planck. It will suffice to prove equality on a class of sufficiently nice test functions ϕ . In other words, we present two different ways to compute $\partial_t \mathbb{E}[\phi(X_t)]$. One way to characterize this derivative is via the change in the density

$$\partial_t \mathbb{E}[\phi(X_t)] = \partial_t \int \phi d p_t = \int \phi \partial_t d p_t = \int \phi d(\partial_t p_t).$$

Another way to characterize this is via the change in the samples themselves. In other words, we have by Theorem A.3.1,

$$\partial_t \mathbb{E}[\phi(X_t)] = \partial_t \int Q_t \phi d p_0 = \int \partial_t Q_t \phi d p_0 = \int Q_t \mathcal{L} \phi d p_0 = \int \mathcal{L} \phi d p_t,$$

where the last line follows from the law of iterated expectations. By Itô's lemma, we know $\mathcal{L}\phi =$

$\langle \nabla \phi, b_t \rangle + \Delta \phi$. We want to simplify

$$\int \mathcal{L} \phi d p_t = \int \langle \nabla \phi, b_t \rangle d p_t + \int \Delta \phi p_t.$$

For the first integral, we use a standard integration by parts calculation,

$$\int \langle \nabla \phi, b_t \rangle d q_t = \int \langle \nabla \phi, b_t q_t \rangle d \mu = - \int \phi \operatorname{div}(b_t q_t) d \mu.$$

For the second integral, we apply integration by parts twice,

$$\int p_t \Delta \phi d \mu = \int \phi \Delta p_t d \mu.$$

Pattern matching both ways to compute $\partial_t \mathbb{E}[\phi(X_t)]$ yields the desired equality. \square

Note that $\int \phi d(\partial_t p_t) = \int \mathcal{L} \phi d p_t$, so we can in fact interpret $\mathcal{L}^* p_t \triangleq \partial_t p_t$ as the *adjoint* of \mathcal{L} . With Fokker-Planck, it becomes trivial to prove the stationarity of Langevin MC.

Proposition 4.2.1 (Stationary distribution of Langevin MC). If we instead take $X_0 \sim p$, then the distribution $X_t \sim p$ for any $t \geq 0$.

Proof. Plugging in $b_t = \nabla \log p_t$, we have $\partial_t p_t = -\operatorname{div}(p_t \nabla \log p_t) + \Delta p_t = 0$ trivially. \square

A.4 Interweaving different samplers

In this section, we extend our critical windows framework to the setting where at certain steps of sampling procedure, instead of using the reverse Markov transition kernel from the original stochastic localization sampler, we use an alternative distribution which is not necessarily related to the original sampler. This includes many important applications of generative models, in which one seeks to combine the priors learned from data with some other algorithm. For example, one may want to combine the language model with a problem generation oracle in in-context learning [DLD⁺24].

As [XRLM22] points out, the transition from the answer to one problem to the problem statement of another example in-context learning is determined by an alternative transition kernel (which they call p_{prompt}). Although the probability of transition from one answer to the problem statement of another example is extremely low under the natural data distribution, one still hopes that with sufficiently many samples, the model selects the correct $\theta^* \in \Theta$ if these lower probability transitions are overcome by the distributional difference for $\theta \in \Theta$ with $\theta \neq \theta^*$. Similarly, under our critical windows framework, we can hope to capture the idea that we specialize to a particular θ^* given a sufficiently long context. In Section A.4, we first present a general framework for characterizing critical windows in this setting. Then, in section A.4, we consider the case of in-context learning by autoregressive language models and prove convergence.

General interweaving framework

We present this framework for the case where the index set $\mathbf{I} = \{0, 1, \dots, m\}$ is discrete. Like before, assume we have a series of reverse Markov transition kernels $\Pr_{k \rightarrow k-1}^{\leftarrow, p}[\cdot | \cdot]$, for $k \in \mathbf{I}$, but we also assume we have an alternative distribution $\Pr_{k \rightarrow k-1}^{\leftarrow, \text{alt}}[\cdot | \cdot]$ that we use to sample for transitions

$k \in A \subsetneq \mathbf{I}$. For our sampling procedure, we sample Y_m , and for $k = m-1, m-2, \dots, 0$, we take $Y_k \sim \text{Pr}_{k+1 \rightarrow k}^{\leftarrow, p}[\cdot | Y_{k+1}]$ for $k \in \mathbf{I} - A$ and $Y_k \sim \text{Pr}_{k+1 \rightarrow k}^{\leftarrow, \text{alt}}[\cdot | Y_{k+1}]$ for $k \in A$. We denote the final distribution p^{alt} . Now, we also need to adjust our definitions of p^S to this particular sampling procedure. We define $p^{S, \text{alt}}$ for $S \subset \Theta$ to the distribution over outputs when we instead use the kernels $\text{Pr}_{k+1 \rightarrow k}^{\leftarrow}[\cdot | Y_{k+1}, S]$ instead of $\text{Pr}_{k+1 \rightarrow k}^{\leftarrow}[\cdot | Y_{k+1}]$. To relate $p^{\theta, \text{alt}}$ to $p^{\theta, \text{alt}}$ for $\theta \in \Theta$, we need to assume transitions from alt do not affect the posterior distribution over $\theta \in \Theta$.

Assumption A.4.1. For all $y \in \mathbb{R}$ and $x \in \text{supp}(\text{Pr}_{k+1 \rightarrow k}^{\leftarrow, \text{alt}}[\cdot | y])$, we have for all $\theta, \theta' \in \Theta$, the equality $\text{Pr}_{k+1 \rightarrow k}^{\leftarrow}[x | y, \theta] = \text{Pr}_{k+1 \rightarrow k}^{\leftarrow}[x | y, \theta']$.

Adopting our definitions from Section 8.2, we let

$$\begin{aligned} T_{\text{lower,alt}}(\epsilon) &\in \{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{init}}, \text{alt}}, p_t^{S_{\text{target}}, \text{alt}}) \leq \epsilon\} \\ T_{\text{upper,alt}}(\epsilon) &\in \{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{target}}, \text{alt}}, p_t^{\Theta - S_{\text{target}}, \text{alt}}) \geq 1 - \epsilon^2\}. \end{aligned}$$

The main challenge of the below corollary is simply show that the final distribution $p^{\Theta, \text{alt}}$ can be written as a mixture of $p^{\theta, \text{alt}}$ with the same mixing weights as before.

Corollary A.4.1. Under Assumption A.4.1, for $\epsilon > 0$, if $\widehat{T} \geq T_{\text{lower,alt}}(\epsilon)$ and $\widehat{T} \leq T_{\text{upper,alt}}(\epsilon)$, then

$$\text{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{target}}, \text{alt}}) \leq \left(1 + \sqrt{2} \max \left(1, \frac{\sum_{\theta \in \Theta - S_{\text{target}}} w_\theta}{\sum_{\theta \in S_{\text{target}}} w_\theta}\right) / 2\right) \epsilon.$$

Proof. We need only show that $p^{\Theta, \text{alt}} \triangleq \sum_{\theta \in \Theta} w_\theta p^{\theta, \text{alt}}$. It suffices to shows that the probability of generating a path Y_m, Y_{m-1}, \dots, Y_0 are the same under both density functions. We need only consider transitions for $k \in \mathbf{I} - A$, because for $k \in A$, the transitions are both given by the alternative distribution. For the transitions not given by alt, note that we are using the original model, so

$$p^{\Theta, \text{alt}}(Y_{k-1} | Y_k) = \frac{\sum_{\theta \in \Theta} w_\theta p^\theta(Y_k) \text{Pr}_{k \rightarrow k-1}^{\leftarrow}[Y_{k-1} | Y_k, \theta]}{\sum_{\theta \in \Theta} w_\theta p^\theta(Y_k)}.$$

Furthermore, for the mixture model, this probability is

$$p^{\text{mix}} = \frac{\sum_{\theta \in \Theta} w_\theta p^{\theta, \text{alt}}(Y_m, Y_{m-1}, \dots, Y_{k+1}, Y_k) \text{Pr}_{k \rightarrow k-1}^{\leftarrow}[Y_{k-1} | Y_k, \theta]}{\sum_{\theta \in \Theta} w_\theta p^{\theta, \text{alt}}(Y_m, Y_{m-1}, \dots, Y_{k+1}, Y_k)}.$$

The distinction between Equation A.4 and Equation A.4 is that in the former we are using the likelihood of p^θ instead of $p^{\theta, \text{alt}}$. Thus it suffices to show that $p^\theta \propto p^{\theta, \text{alt}}$. We explicitly write out the probability,

$$p^\theta(Y_m, Y_{m-1}, \dots, Y_{k+1}, Y_k) = \prod_{i=k+1}^m p^\theta(Y_{i-1} | Y_i) \propto \prod_{i=k+1, i \notin A}^m p^\theta(Y_{i-1} | Y_i),$$

where the proportionality follows from the fact that we can ignore the probability of the transitions produced by alt under Assumption A.4.1. By definition, this is proportional to $p^{\theta, \text{alt}}(Y_m, Y_{m-1}, \dots, Y_{k+1}, Y_k)$ up to a normalization constant independent of θ . \square

In-context learning

Now, we will specialize our framework to the case of in-context learning. As in [XRLM22], we assume that the language model is given inputs of the form

$$[x_1, y_1, o, x_2, y_2, o, \dots, x_t, y_t, o, x_{t+1}],$$

where x_1 is the input, y_1 is the output, and o is a delimiter token that separate different in-context samples from each other. We assume that the transitions $y_i \rightarrow o \rightarrow x_{i+1}$ are sampled by some alternative probability distribution $P_{\text{other}}(\cdot|\cdot)$. We require that P_{other} selects the x_i i.i.d.

Assumption A.4.2. The distribution of $P_{\text{other}}(x_{t+1}|x_1, y_1, o, x_2, y_2, o, \dots, x_t, y_t, o) = P_{\text{other}}(x_1)$.

Then we assume that the transitions $x_i \rightarrow y_i$ are generated by some $\theta^* \in \Theta$, which does not depend on any of the previous tokens before the delimiter.

Assumption A.4.3. (Well-specification) There exists some $\theta^* \in \Theta$ such that y_i is generated from $y_i \sim P^{\leftarrow}(\cdot|[x_1, y_1, o_1, \dots, o_{i-1}, x_i], \theta^*)$.

Assumption A.4.4. For all $\theta \in \Theta$, we have $\Pr^{\leftarrow}[\cdot|[x_1, y_1, o_1, \dots, o_{i-1}, x_i], \theta] = \Pr^{\leftarrow}[\cdot|x_i, \theta]$.

We also assume statistical separation of θ^* from $\Theta - \{\theta^*\}$ in terms of Hellinger distance.

Assumption A.4.5. Let $p_{(x,y)}^S$ for $S \subset \Theta$ be the distribution of (x_1, y_1, o) where $x_1 \sim P_{\text{other}}(\cdot|o)$ and $y_1 \sim p^S(\cdot|x_1)$. There exists $\delta > 0$ such that $H^2(p_{(x,y)}^{\Theta-\{\theta^*\}}, p_{(x,y)}^{\theta^*}) \geq \delta$.

Example A.4.1. Let $T \geq \ln\left(\frac{1-\delta/2}{\epsilon}\right)$. Under Assumptions A.4.1, A.4.2, A.4.3, A.4.4, and A.4.5, we have

$$\text{TV}([x_1, \dots, x_{T+1}, y_{T+1}], [x_1, \dots, x_{T+1}, \tilde{y}_{T+1}]) \lesssim \epsilon/w_{\theta^*}.$$

Proof. It suffices to upper bound $1 - \frac{1}{2}H^2(P_{3T}^{\theta, \text{alt}}, P_{3T}^{\Theta-\{\theta\}, \text{alt}})$ by $O(\epsilon)$. First observe that the distribution $p_{(x,y)}^{S, \text{alt}}$ for $S \subset \Theta$ factors along the delimiters by a factor independent of S using Assumptions A.4.2 and A.4.4, so we have

$$P_{3t}^{S, \text{alt}}(x_1, y_1, o_1, x_2, y_2, o_2, \dots, x_t, y_t, o_t) \propto \prod_{i=1}^t P_3^{S, \text{alt}}(x_i, y_i, o_i).$$

Using the tensorization property of Hellinger distance and our definitions of T, δ , we have

$$1 - \frac{1}{2}H^2(P_{3T}^{\theta, \text{alt}}, P_{3T}^{\Theta-\{\theta\}, \text{alt}}) \lesssim \left[1 - \frac{1}{2}H^2(P_3^{\theta, \text{alt}}, P_3^{\Theta-\{\theta\}, \text{alt}})\right]^T \leq \epsilon. \quad \square$$

Notation

Notation	
$[N]$	$\{1, 2, \dots, N\}$
Tr	trace operator
$\mathcal{C}_k(U)$	set of all k -times continuously differentiable functions $U \rightarrow \mathbb{R}$
\mathcal{S}^d	unit sphere in \mathbb{R}^d
$\ \cdot\ $	ℓ^2 -norm (Euclidean norm)
Id	identity matrix
div	divergence
∇	gradient
Δ	Laplacian
$\text{span}(u, v)$	vector space spanned by u, v
$f_\#$	pushforward of f
2^S for a set S	powerset of S
$B_r(p)$	ball of radius r around p
$A \preceq B$	the difference $B - A$ is positive semi-definite
(Ω, \mathcal{F})	Borel space with a base set Ω with a σ -algebra \mathcal{F}
\mathbb{E}	expectation operator
dP	density of the measure P
$P^{\otimes n}$	product measure of n i.i.d. samples from P
*	convolution operation between two measures
$\Gamma(P, Q)$	set of couplings between P, Q
TV	total variation distance
KL	Kullback–Leibler divergence
H	Hellinger distance
W_2	Wasserstein-2 distance
$\text{supp}(P)$	support of a measure P
Θ	set of latents in mixture model
p^S	sub-mixture over the latents $S \subset \Theta$
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$\text{Cat}(S, \{w_i\}_{i=1}^{ S })$	Categorical distribution over set S with weights $\{w_i\}_{i=1}^{ S }$
$\text{SLC}(d)$	set of strongly log-concave measures in \mathbb{R}^d
$\text{SLC}(\beta, d)$	set of $1/\beta$ -strongly log-concave measures in \mathbb{R}^d .
$\text{subG}_d(\sigma^2)$	set of sub-Gaussian measures in \mathbb{R}^d with variance proxies σ^2

We will also collect the critical windows notation. let p_t^S represent the distribution at time t of the forward process of p^S and $p^{S,t}$ be the distribution of samples from p^S where we apply the forward process from time 0 to time t and then the reverse process back to time 0. The interval $[T_{\text{end}}^S, T_{\text{start}}^S]$ is where the sampler chooses the features defined by S_{init} but has not yet chosen the features that distinguish S_{init} and S_{target} where $S_{\text{init}} \subset S_{\text{target}}$. The interval $[T_{\text{after}}^C, T_{\text{before}}^C]$ demarcates the transition from sampling from S_{before} to the S_{after} with $S_{\text{after}} \subset S_{\text{before}}$. The latter is what we mean when we consider a critical window.

References

- [AAB⁺24] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024.
- [ABVE23] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- [AD86] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *The American Mathematical Monthly*, 93(5):333–348, 1986.
- [ADS⁺24] Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, Jamie Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. In *Advances in the Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [AI25] Santiago Aranguri and Francesco Insulla. Phase-aware training schedule simplifies learning in flow-based generative models, 2025.
- [AJH⁺21] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- [AK23] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity, 2023.
- [ALL⁺19] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings, 2019.
- [Amb23] Luca Ambrogioni. The statistical thermodynamics of generative diffusion models, 2023.
- [AMO23] Sarah Andersen, Kelly McKernan, and Karla Ortiz. Class-action complaint against stability ai, 2023. Case 3:23-cv-00201.

- [AMS23] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.
- [Ant23] Anthropic. Claude, 2023.
- [Ant24] Anthropic. Developing a computer use model. <https://www.anthropic.com/news/developing-computer-use>, 2024.
- [Ant25] Anthropic. Claude 3.7 sonnet system card, 2025. Accessed: 2025-03-04.
- [AVE23] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2023.
- [BBdBM24] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024.
- [BDD23] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.
- [BHP17] Riddhipratim Basu, Jonathan Hermon, and Yuval Peres. Characterization of cutoff for reversible markov chains. *The Annals of Probability*, 45(3), May 2017.
- [BMR20] Jean Barbier, Nicolas Macris, and Cynthia Rush. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation, 2020.
- [BMR22] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint 2002.00107*, 2022.
- [BSS⁺24] Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2024.
- [CCE⁺18] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint 1803.05457*, 2018.
- [CCL⁺23a] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023.
- [CCL⁺23b] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. 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, 2023.
- [CDD23] Sitan Chen, Giannis Daras, and Alexandros G Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023.
- [CE22] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.
- [CGC⁺24] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting, 2024.

- [Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture, 2021.
- [Che23] Sinho Chewi. Log-concave sampling. *Book draft available at <https://chewisinho.github.io>*, 9:17–18, 2023.
- [Che24] Sitan Chen. Cs 2243: Algorithms for data science, fall 2024. <https://sitanchen.com/cs224/f24/index.html>, 2024.
- [CHN⁺23a] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.
- [CHN⁺23b] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC ’23, USA, 2023. USENIX Association.
- [CLB⁺23] João Carvalho, An T. Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.
- [CLL22a] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.
- [CLL22b] Sitan Chen, Jerry Li, and Yuanzhi Li. Learning (very) simple generative models is hard, 2022.
- [CLS⁺22] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11462–11471, 2022.
- [COGHK⁺22] Amin Coja-Oghlan, Oliver Gebhard, Max Hahn-Klimroth, Alexander S Wein, and Ilias Zadik. Statistical and computational phase transitions in group testing. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4764–4781. PMLR, 02–05 Jul 2022.
- [CRD⁺23] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2023.
- [DB22] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.
- [DBTHD21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021.

- [DCX⁺23] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.
- [DKW⁺23] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023.
- [DLD⁺24] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.
- [DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.
- [DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [DTSV23] Valentin Dehontoli, James Thornton, Tim Salimans, and Arash Vahdat. Diffusion meets flow matching. *arXiv preprint arXiv:2303.09457*, 2023.
- [Du24] Yilun Du. Generative ai, part 1. Lecture at Harvard University, CS 2281R: Mathematical & Engineering Principles for Training Foundation Models, October 31 2024. Slides available at https://shamulent.github.io/CS_2281_2024.html.
- [EAMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.
- [EK10] Ronen Eldan and Bo’az Klartag. Approximately gaussian marginals and the hyperplane conjecture, 2010.
- [EKZ21] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: Fast mixing in high-temperature ising models, 2021.
- [Eld13] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.
- [Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.
- [EMZ20] Ronen Eldan, Dan Mikulincer, and Alex Zhai. The clt in high dimensions: quantitative bounds via martingale embedding, 2020.
- [Gal18] Jean-Francois Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Springer Publishing Company, Incorporated, 2018.
- [Gan24] Surya Ganguli. On large language models and transformers: Perspectives from physics, neuroscience, and theory. Simons Institute’s Special Year on Large Language Models and Transformers, Part 1 Boot Camp, September 2024. Quote: ”The structure of data is the dark matter of theory in deep learning.”.

- [Gem24] Gemini. Challenges and solutions for aging adults. <https://gemini.google.com/share/6d141b742a13>, 2024.
- [GH23] Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [GHH⁺25] Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusionflow, 2025. Accessed: 2025-03-19.
- [GPAM⁺14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
- [gri24] grimjim. Jailbroken llama-3.1-8b-instruct via lora. https://huggingface.co/grimjim/Llama-3.1-8B-Instruct-obliterated_via_adapter, 2024.
- [Gua24] Qingsyang Guan. A note on bourgain’s slicing problem, 2024.
- [GVS⁺23] Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.
- [GZ19] David Gamarnik and Ilias Zadik. High-dimensional regression with binary coefficients. estimating squared error and a phase transition, 2019.
- [HÅ06] Toivo Henningsson and Karl Johan Åström. 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.
- [Hai24a] Haize Labs. Automated multi-turn red-teaming with cascade. <https://blog.haizelabs.com/posts/cascade>, 2024.
- [Hai24b] Haize Labs. A trivial jailbreak against llama 3. <https://github.com/haizelabs/llama3-jailbreak>, 2024.
- [HBB⁺21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.
- [HBK⁺21] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.
- [HHY⁺25] Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. *arXiv preprint arXiv:2410.20587*, 2025. Presented at ICLR 2025.
- [HJA20a] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

- [HJA20b] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
- [HMP24] Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from spherical spin glasses in total variation via algorithmic stochastic localization. *arXiv preprint arXiv:2404.15651*, 2024.
- [Hol23] Peter E. Holderrieth. The fokker-planck equation and diffusion models, 2023. Accessed: 2025-03-19.
- [HSW⁺22] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.
- [HXH24] Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety, 2024.
- [Hyv05] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
- [Ima23] Getty Images. Getty images (us), inc. v. stability ai, inc, 2023. Case 1:23-cv-00135-UNA.
- [JLV22] Arun Jambulapati, Yin Tat Lee, and Santosh S Vempala. A slightly improved bound for the kls constant. *arXiv preprint arXiv:2208.11644*, 2022.
- [KHO05] DAVAR KHOSHNEVAN. Lecture notes on donsker’s theorem. 2005.
- [KL22] Bo’az Klartag and Joseph Lehec. Bourgain’s slicing problem and kls isoperimetry up to polylog, 2022.
- [KL24] Boaz Klartag and Joseph Lehec. Affirmative resolution of bourgain’s slicing problem using guan’s bound, 2024.
- [KP21] Bo’az Klartag and Eli Puterman. Spectral monotonicity under gaussian convolution. *arXiv preprint arXiv:2107.09496*, 2021. Available at <https://arxiv.org/abs/2107.09496>.
- [LAMJ23] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023.
- [LC24] Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models, 2024.
- [LCBH⁺22] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [LCL⁺20] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logica: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

- [LCR⁺23] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilé Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.
- [LeC86] Lucien LeCam. *Asymptotic methods in statistical decision theory*. Springer series in statistics. Springer, New York, NY [u.a.], 1986.
- [Led01] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.
- [LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [LHS⁺24] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024.
- [LHW24] Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024.
- [Li24] Marvin Li. Donder’s institute for brain, cognition, and behaviour. <https://www.youtube.com/watch?v=NbnBmi5k1UQ>, 2024. Accessed: 2024-05-01.
- [LKB⁺23] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- [LKC25] Marvin Li, Aayush Karan, and Sitan Chen. Blink of an eye: a simple theory for feature localization in generative models, 2025.
- [LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.
- [LLX⁺24] Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances llm’s reasoning capability, 2024.
- [LME24] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2024.

- [LRL⁺23] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning, 2023.
- [LS77] Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes: General Theory*, volume 394 of *Applications of Mathematics*. Springer, 1977.
- [LTG⁺22] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- [LV17] Yin Tat Lee and Santosh Srinivas Vempala. Eldan’s stochastic localization and the kls hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007, 2017.
- [LW23] Marvin Li and Jason Wang. 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.
- [LWCC23] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.
- [LWWN23] Marvin Li, Jason Wang, Jeffrey Wang, and Seth Neel. MoPe: Model perturbation based privacy attacks on language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13647–13660, Singapore, December 2023. Association for Computational Linguistics.
- [LWYL22] Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.
- [LXCX24] Xiaogeng Liu, Nan Xu, Muhan Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [LYDB17] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [MFP⁺24] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024, 2024.
- [MHS⁺22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

- [Mil08] Emanuel Milman. On the role of convexity in isoperimetry, spectral-gap and concentration, 2008.
- [MLK⁺23] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 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.
- [MNWS⁺23] Elchanan Mossel, Jonathan Niles-Weed, Youngtak Sohn, Nike Sun, and Ilias Zadik. Sharp thresholds in inference of planted subgraphs. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5573–5577. PMLR, 12–15 Jul 2023.
- [Mon23a] Andrea Montanari. Sampling, diffusions, and stochastic localization, 2023.
- [Mon23b] Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.
- [MW23] Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.
- [NCH⁺23] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023.
- [NWZ20] Jonathan Niles-Weed and Ilias Zadik. The all-or-nothing phenomenon in sparse tensor pca. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17674–17684. Curran Associates, Inc., 2020.
- [NWZ23] Jonathan Niles-Weed and Ilias Zadik. It was “all” for “nothing”: sharp phase transitions for noiseless discrete channels, 2023.
- [Ope23] OpenAI. Chatgpt, 2023.
- [Ope24] OpenAI. Sora, 2024.
- [Ope25] OpenAI. Detecting misbehavior in frontier reasoning models. *OpenAI Research*, 2025.
- [Par05] Leandro Pardo. *Statistical Inference Based on Divergence Measures*. CRC Press, Abingdon, 2005.
- [PCT⁺23] Leonardo Petrini, Francesco Cagnetta, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *arXiv preprint arXiv:2307.02129*, 2023.
- [Pid22] Jakiw Pidstrigach. Score-based generative models detect manifolds. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35852–35865. Curran Associates, Inc., 2022.

- [Pol24] Yury Polyanskiy. Lecture notes on f -divergence. https://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf, 2024. Accessed: 2025-02-09.
- [PVBM25] Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Understanding classifier-free guidance: High-dimensional theory and non-linear generalizations, 2025.
- [QPL⁺24] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024.
- [RA23] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [RCRY25] Yinuo Ren, Haoxuan Chen, Grant M. Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework, 2025.
- [RH23] Philippe Rigollet and Jan-Christian Hutter. High-dimensional statistics, 2023.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [RKV⁺24] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xtest: A test suite for identifying exaggerated safety behaviours in large language models, 2024.
- [Roo01] B. Roos. Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability & Its Applications*, 45(2):258–272, 2001.
- [RXZ19] Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression, 2019.
- [SC16] Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework, 2016.
- [SDCS23] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.
- [SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11895–11907, 2019.
- [SFLW24] Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models, 2024.

- [SFW25] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1):e2408799121, 2025.
- [SLB⁺24] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024.
- [SSDK⁺20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 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*, pages 3–18. IEEE Computer Society, 2017.
- [SW14] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8(none):45 – 114, 2014.
- [TAS21] Lan V. Truong, Matthew Aldridge, and Jonathan Scarlett. On the all-or-nothing behavior of bernoulli group testing, 2021.
- [TL24] Google DeepMind AlphaFold Team and Isomorphic Labs. Alphafold 3 predicts the structure and interactions of all of life’s molecules. *Nature*, 2024.
- [TS20] Lan V. Truong and Jonathan Scarlett. Support recovery in the phase retrieval model: Information-theoretic fundamental limits, 2020.
- [Ver] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- [vH16] Ramon van Handel. Probability in high dimension. 2016.
- [Wan24] Yihui Wang. Lecture 4: Total variation/inequalities between f-divergences. <http://www.stat.yale.edu/~yw562/teaching/598/lec04.pdf>, 2024. Accessed: [Insert Access Date].
- [WHS23] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023.
- [WY22] Andre Wibisono and Kaylee Y. Yang. Convergence in KL divergence of the inexact Langevin algorithm with application to score-based generative models. *arXiv preprint 2211.01512*, 2022.
- [XRLM22] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022.
- [ZCS⁺23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

- [ZPW⁺24] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024.
- [ZW24] Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model fine-tuning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [ZWC⁺23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.



THIS THESIS WAS TYPESET using \LaTeX , originally developed by Leslie Lamport and based on Donald Knuth's \TeX . The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under [CC BY-NC-ND 3.0](#). A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.