Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics and Computer Science
Cycle: 37
Disciplinary Field (code): FIS/02

Statistical Physics of Generative Diffusion

Advisor: Carlo Lucibello
Co-Advisor: Marc Mézard

PhD Thesis by
Beatrice Achilli
ID number: 3146862

Year: 2026

Abstract

This thesis investigates diffusion models with a statistical physics point of view, focusing on phase transitions and symmetry breaking events.

First, we analyze the reverse diffusion process under the empirical score function for structured data. What we obtain is the description of a dynamical landscape and the characterization of the most important transition times, namely the *memorization* time. We also give a definition of *generalization* in this context, observing that interestingly it always happens after the model starts memorizing.

Then, we give a geometric description, exploring the spectral properties of the score function using tools from random matrix theory. By analyzing the Jacobian spectra of the score, we identify the emergence of geometric phases linked to *spectral gaps*. We also study the phenomenon of *geometric memorization*, demonstrating that it is characterized by a loss of dimensionality where some features of the data are memorized without a full collapse on any individual training point.

Finally, we investigate the *speciation* transition of diffusion models in a case where data are not spatially separated. We obtain a general criterion for the speciation time, as well as its scaling.

This thesis thus provides both theoretical insights and empirical analyses that bridge deep learning and statistical physics, contributing to a deeper understanding of how generative models learn and represent data.

Contents

Chapter 1

Introduction

> Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics. Perhaps it will be wise to approach the subject cautiously.
>
> David Louis Goodstein

1.1 What is Statistical Physics?

Goodstein wrote the book States of Matter, from which this extract is taken, in 1975, and only died in 2024 at the age of 85, so if we follow its advice, we should be safe. But what is, in reality, this scary subject?

Statistical Physics studies the macroscopic properties of a system composed of many microscopic units. When we forget about the details of the single components, collective behaviors appear that are richer than simply the sum of the units. An instance of this is phase transitions. The appearance of such phenomena is bound to having a very large number of units, namely, a high-dimensional regime. Statistical Physics mainly works in the thermodynamic limit, where this number is taken to infinity. In this regime, we can accurately describe phase transitions and obtain sharp phase diagrams.

Statistical Physics has been applied to the study of artificial neural networks since the the '80s. We can recall, for example, Elizabeth Gardner's seminal works Gardner and Derrida [1988] or the recently assigned Nobel prize, John Hopfield Hopfield [1982]. Since neural networks are indeed composed of a large number of neurons, they represent the perfect setting.

1.2 What are Diffusion Models?

In recent years, we have witnessed a sort of Artificial Intelligence (AI) revolution. AI is actually a broad term; the most recent models all fall under the category of Deep Learning (DL). We can talk about DL when the networks we are analyzing are deep, in the sense that they are composed of many layers. In the DL history, one of the main turning points was Krizhevsky et al. [2012], where for the first time a deep convolutional network proved to be better than the standard machine learning methods. Following that, we can identify another breakthrough in the introduction of the transformer architecture Vaswani et al. [2017]: a sequence-to-sequence model that relies on the the attention mechanism. This innovation led directly to models such as BERT, GPT, and later large-scale models like GPT-3, which demonstrated remarkable performance on a wide range of benchmarks. In parallel, multimodal AI systems that process text, images, audio, and other modalities have gained popularity, with models like CLIP and DALL·E showing that transformers can align textual and visual representations.

One of the most innovative aspects in this field is the one of generative models, which synthesize new realistic text, image, audio, and video. In this context, the most prominent kind of models are Diffusion Models. Before their introduction, the most prominent generative models were Generative Adversarial Networks (GANs). Other models, implied for generative purposes, despite not being their prime functionality, are Variational Autoencoders (VAE).

Diffusion Models are a type of generative model inspired by non-equilibrium physics. The main idea is quite simple: we have some data, and we would like to generate new ones that are *similar*. This, of course, can be translated into the problem of approximating the underlying probability distribution of a dataset. To achieve this goal, DMs perform two steps: during the *forward process*, we progressively add noise to the data until we reach a completely random distribution. Then, we reverse this process, and starting from something completely random we are able to obtain samples from the original data distribution. Given this first description, one could think that it is impossible to obtain a signal from pure noise. It is not impossible, but of course there is a trick. During the forward process, one learns how the data are noised, and uses this information to build a drift that guides the reverse process toward the correct samples. Still, this simple idea results in an elegant and very powerful tool.

Of course, what has been described here is only the backbone idea of DMs. In practice, when we refer to DMs as generative models, we are talking about trained neural networks. The role of the neural network is just to approximate the drift term that allows everything

to work. Even if this concept per se is pretty straightforward, it took some time to realize that it could be a viable option, and some more to understand how to do this effectively.

When studying diffusion models, we would like to focus on understanding when and why they work well, and when this does not happen. For this purpose, we hereby introduce two terms that will be used many times in the course of this manuscript: memorization and generalization. Although it can be intuitive what we mean by them, it is a good practice to properly introduce them.

A model is said to **memorize** when it learns the training data too well, including noise or irrelevant details, rather than learning the underlying structure or patterns. In generative diffusion models, memorization can mean the model generates samples that are too close or identical to training examples, rather than synthesizing new, plausible samples that follow the same distribution. It is a phenomenon that has been observed empirically in trained diffusion models Somepalli et al. [2023], Webster [2023], especially when they are overparametrized. Notice that this is linked to overfitting, but the two phenomena do not perfectly correspond. Memorization is a phenomenon that is worth being investigated since its appearance completely defies the purpose of generative diffusion, and it can also lead to privacy issues. Thus, it is important to know the regimes in which it happens in order to better avoid them. When a diffusion model is in the memorization phase, it shares many similarities with dense associative memories.

We say that a model **generalizes** when it captures the essential structure of the data, allowing it to make accurate predictions or generate new samples that are consistent with the overall data distribution and not just the training set. Generalization for DMs refers to the model's ability to sample from the true data manifold without collapsing to a few memorized examples. Since, in principle, diffusion models infer the true data distribution from examples, the reasons why they are so efficient at generalizing it are one of the main discussion points. In Kamb and Ganguli [2025], they suggest that this is due to some inductive biases.

This distinction will be a crucial aspect of our analysis, as we will try to characterize deeply the DM behavior in both regimes.

1.3 Why Statistical Physics of Diffusion Models?

Diffusion models, especially in the context of generative modeling, naturally evoke analogies with thermodynamic processes. Training and sampling procedures in score-based generative models are closely related to Langevin dynamics and stochastic thermodynamics. Furthermore, the statistical physics framework of large deviations offers a powerful

lens through which understanding the behavior of diffusion models, particularly in high dimensions.

Recently, many works have suggested that properties of generative diffusion models could be understood using statistical physics tools, such as symmetry breaking and phase transitions. In particular, Raya and Ambrogioni [2023] were the first to interpret the dynamics of trajectories generated by diffusion models as subject to symmetry breaking. It is quite intuitive to see that, effectively, starting from a Gaussian distributed variable, in the reverse process we pass from a phase where any data point could be generated, to one where only a specific one is selected, so there must be a broken symmetry in between. Biroli and Mézard [2023] analyzes diffusion models in very large dimensions, where symmetry breaking and phase transitions naturally appear. In particular, they investigate the scaling of the number of samples used with respect to their dimensionality in order for the diffusion model to be able to represent enough aspects of their probability distribution. They ask the opposite question: can the diffusion process reproduce the symmetry breaking event of a physical system? This led to the identification of the speciation transition in diffusion models. In Sclocchi et al. [2025], reverting the diffusion process at different times, they notice a phase transition where the probability of remaining in the same class suddenly drops to random chance. This is related to the compositional nature of data; indeed, low-level features of the initial sample can persist and compose the new sample. Another prominent work in the statistical physics of diffusion models is Biroli et al. [2024]. The authors introduce the dynamical phases of diffusion models separated by the speciation and collapse (later memorization) transitions.

The role of statistical physics is to exploit the large dimensional limit to identify general mechanisms, that later we can test on real datasets. We decided to build upon this framework and continue investigating these transitions, adding the complexity of data structure.

1.4 Thesis Overview

This thesis is organized as follows:

We first present all the models that will be used in the following chapters. Chapter 2 is a theoretical chapter where we review all the tools that will be needed throughout the work. We start with a summary of diffusion theory in physics, and then pass to the machine learning scenario of Diffusion Models, recalling the foundational works in this field. Then we introduce the Random Energy Model, a standard model in statistical physics that has been recently applied in the study of associative memory Lucibello and

Mézard [2024] and DMs Biroli et al. [2024], and we also outline this recent line of research of statistical physics applied to DMs. Lastly, we describe the data models that will be used in the other chapters.

In Chapter 3 we study the memorization and generalization capabilities of a Diffusion Model (DM) in the case of structured data defined on a latent manifold. Our analysis considers a reverse process given by the empirical score function as a proxy of the true one, and then precisely characterizes the process in the high-dimensional limit in which both the number of data and their dimension are large, by exploiting a connection with the Random Energy Model (REM). We provide evidence for the existence of an onset time when traps appear in the time-varying potential, although they do not affect typical trajectories. The size of the basins of attraction of such traps is computed at any time. Moreover, we derive the collapse time at which trajectories fall into the basin of one of the training points, implying memorization. An explicit formula for this collapse time is given, proving that the curse of dimensionality issue does not hold for highly structured data, regardless of the non-linearity of the manifold surface. We also prove that collapse coincides with the condensation transition in the REM. Finally, the degree of generalization of DMs is formulated in terms of the Kullback-Leibler divergence between the exact distribution and the one obtained at time t of the reverse process. We show the existence of an additional time, called generalization time, such that the distance between the reverse distribution and the ground-truth is minimal. Counter-intuitively, the best generalization performance is found within the memorization phase of the model. We conclude that the generalization performance of DMs benefits from highly structured data, since.

To complete the framework, in Chapter 4 we analyze the same setting from a geometrical point of view. Instead of considering typical trajectories, we describe what happens at any space point during the diffusion process. This geometrical description is based on the analysis of the spectrum of eigenvalues of the Jacobian of the score function. This Chapter is further divided into two main sections: in the first one, we analyze the true score function, while in the second one we analyze the empirical score.

Regarding the first part, we have used a statistical physics approach to derive the spectral distributions and formulas for the spectral gaps of the Jacobian of the true score function under the manifold hypothesis, and we compare these theoretical predictions with the spectra estimated from trained networks. Our analysis reveals the existence of three distinct qualitative phases during the generative process: a trivial phase, a manifold coverage phase where the diffusion process fits the distribution internal to the manifold, a consolidation phase where the score becomes orthogonal to the manifold and all particles

are projected on the support of the data. This 'division of labor' between different time scales provides an elegant explanation of why generative diffusion models are not affected by the manifold overfitting phenomenon that plagues likelihood-based models, since the internal distribution and the manifold geometry are produced at different time points during generation.

In the second part, we transition into the study of memorization. When studying memorization, it is natural to ask if it is a single transition or if information is lost gradually. In order to study this, we adopt a manifold data model and separate memorization events along the directions spanned by the latent manifold. We refer to this intermediate memorization phenomenon as *geometric memorization*. In the memorization regime, the neural network score of a trained diffusion model approximates the empirical score function. Leveraging on the analogy between the empirical score and the Random Energy Model, we build a theory for geometric memorization based on the analysis of the spectrum of eigenvalues of the Jacobian of the *empirical* score function. What we find is the emergence of a spectral gap, which was not predicted by the exact score theory, and which displays evidence of the progressive loss of dimensionality due to memorization.

Aside from memorization, another interesting transition displayed by DMs is the speciation one. In Chapter 5 we dedicate our attention precisely to the study of the speciation time. In particular, we derive a general criterion for the timescale of this phenomenon in the case of a multimodal distribution without necessarily space separation between the different clusters. The practical example studied is that of a mixture of 1d Ising models at different temperatures.

The more technical aspects of this manuscript will be enclosed in boxes. This is done to maintain readability of the main text while providing all the details where they are needed.

Chapter 2

Preliminaries

> Do the scary thing first, and get scared later.
>
> Lemony Snicket

In this chapter, we are going to review all the theoretical instruments and background upon which we build the following chapters. We present a broad introduction to diffusion models and generative diffusion, revising the most relevant literature. Then we pass to statistical physics of diffusion models, where we review in depth the works that will be useful for this thesis, and introduce the tools, namely the Random Energy Model. Lastly, we describe the data-generating models, the Hidden Manifold Model and 1d Ising with a random field. Although very different, we highlight how these two models help us bring structure into play.

2.1 Diffusions

In this section, we go back to the beginning of the study of diffusion processes and revise the main mathematical tools: Langevin and Fokker-Planck equations. We also introduce two simple diffusion processes, Brownian motion and Orstein-Uhlenbeck, and recap their main features. Interestingly, these two very simple processes have set two paradigms in generative diffusion under the names, respectively, of Variance Exploding and Variance Preserving frameworks.

2.1.1 Langevin diffusion

The motion of a particle suspended in a fluid, such as pollen grains in water, is irregular and seemingly random. This phenomenon, known as *Brownian motion*, was first observed

by Robert Brown in 1827. Brown was investigating the fertilization process in *Clarkia pulchella* when he noticed a rapid oscillatory motion of the microscopic particles within the pollen grains suspended in water under the microscope. Initially, he believed that such motion was a vital activity, and he was only convinced against this when the same phenomena showed up for particles from the Great Sphinx. Albert Einstein provided an explanation in terms of atoms and molecules in 1905. Einstein showed that the erratic movement of the particle arises from countless collisions with the much smaller, faster-moving molecules of the fluid. His equations describing Brownian motion were checked by the experimental work of Jean Baptiste Perrin in 1908, Fig. 2.1, leading him to win the Nobel prize in 1926.

To describe this stochastic behavior more precisely, Paul Langevin[1] proposed a differential equation that incorporates both deterministic and random forces acting on a particle. This equation, now known as the *Langevin equation*, provides a bridge between Newtonian mechanics and stochastic processes.

Consider a particle of mass m moving subject to three types of forces: a deterministic force f, a frictional (viscous) force $-\gamma\dot{x}$, and a stochastic force $\eta(t)$ representing random collisions with surrounding molecules. The *Langevin equation* reads:

$$m\frac{d^2x}{dt^2} = f(x) - \gamma\frac{dx}{dt} + \eta(t). \tag{2.1}$$

Here, $\gamma > 0$ is the friction coefficient, and $\eta(t)$ is a stochastic force typically modeled as Gaussian white noise with mean $\langle\eta(t)\rangle = 0$ and correlation $\langle\eta(t)\eta(t')\rangle = 2\gamma k_B T\,\delta(t-t')$, where k_B is Boltzmann's constant and T is the temperature of the environment. The term $2\gamma k_B T$ ensures that the system reaches thermal equilibrium in the long-time limit, as required by the fluctuation–dissipation theorem. In the following, we will often take $k_B = 1$.

In many systems, particularly those in a high-friction or small-mass regime (e.g., colloidal particles in a fluid), inertial effects are negligible. In this *overdamped limit*, we let $m \to 0$ in (2.1), eliminating the acceleration term:

$$\gamma\frac{dx}{dt} = f(x) + \eta(t). \tag{2.2}$$

We can often assume that f is a conservative force derived from a potential $V(x)$, and thus

[1]Small historical note: he was one of the founders of the Comité de vigilance des intellectuels antifascistes, an anti-fascist organization created after far-right riots in 1934. He was also arrested for being a public opponent of fascism in the 1930s, and he was held under house arrest by the Vichy government for most of World War II.

write it as $f(x) = -\nabla V(x)$. Dividing both sides by γ, we obtain a first-order stochastic differential equation:

$$\frac{dx}{dt} = f(x) + \sqrt{2D}\ \eta(t), \tag{2.3}$$

where we define the *diffusion coefficient* as $D = \frac{k_B T}{\gamma}$ and we have reabsorbed a factor $1/\gamma$ in the definition of $f(x)$. The noise term $\eta(t)$ is now normalized white noise with mean zero and correlation $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. In the following, this type of δ-correlated noise will be referred to as a Gaussian White Noise (GWN).

A more general form of Eq. (2.3) is

$$\frac{dx}{dt} = f(x) + g(x)\ \eta(t) \tag{2.4}$$

where we let the *diffusion* term $g(x)$ be x dependent. $f(x)$ is often referred to as the *drift* coefficient. This can be rewritten as a Itô stochastic differential equation

$$dx = f(x)dt + g(x)dW_t \tag{2.5}$$

Where dW_t is a standard Wiener process, from Norbert Wiener, who gave the first complete and rigorous mathematical analysis in 1923. In physics, it is referred to as Brownian motion. In mathematics, the Wiener process W_t is characterized by four facts

1. $W_0 = 0$
2. W_t is almost surely continuous
3. W_t has independent increments
4. $W_t - W_s \sim \mathcal{N}(0, t - s)$ for $0 \leq s \leq t$.

In this formulation, one could see $\eta(t)$ as the supposed derivative dW_t/dt of the Wiener process. However, this derivative does not exist because the Wiener process is nowhere differentiable, and so the Langevin equation only makes sense if interpreted in a distributional sense.

2.1.2 Fokker-Planck

The Langevin equation describes the stochastic trajectory of an individual particle. However, in many situations we are interested not in a single realization, but in the evolution of the probability distribution $p(x, t)$ of a particle's position over time. This leads to a complementary, macroscopic description via a partial differential equation for $p(x, t)$,

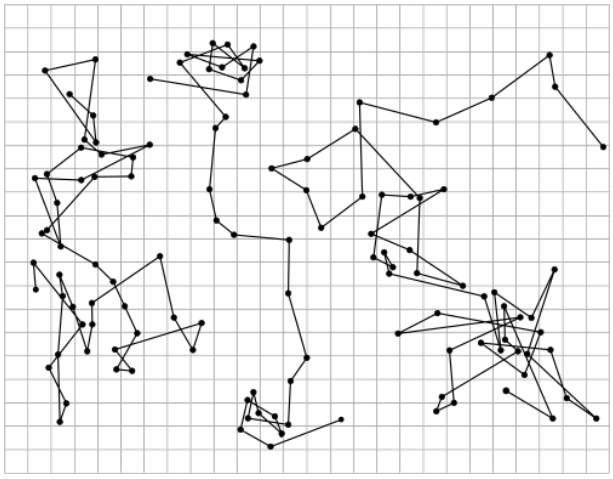


Figure 2.1: From the book of Jean Baptiste Perrin, Les Atomes, three tracings of the motion of particles of line 0.53 μm, as seen under the microscope.

known as the *Fokker-Planck equation*. It is also known as the Kolmogorov forward equation. This equation governs the time evolution of the probability density associated with a stochastic process.

Starting from Eq. (2.5),

$$dx_t = \mu(x_t, t)dt + \sigma(x_t, t)dW_t \tag{2.6}$$

where we explicit the time dependence of drift and diffusion $D(x,t) = \sigma(x,t)^2/2$ coefficients. We can write the corresponding Fokker-Planck equation for the probability density $p(x,t)$ as:

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[\mu(x,t)\ p(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma(x,t)^2\ p(x,t)\right]. \tag{2.7}$$

This equation describes the flow of probability due to both drift and diffusion. The first term on the right-hand side accounts for the deterministic flow, while the second term describes spreading due to randomness.

Brownian Motion (Variance Exploding)

As a first example, we describe the simplest, and historically the first, diffusion process: Brownian motion. Also known as the Wiener process, it corresponds to a free particle undergoing random motion with no drift term and constant diffusion. The Langevin equation thus is

$$m\frac{dv}{dt} = -\gamma v + \eta(t) \tag{2.8}$$

with $v = dx/dt$. The stochastic differential equation for Brownian motion is simply

$$dx = \sqrt{2D}\ dW_t, \tag{2.9}$$

The corresponding Fokker-Planck equation is:

$$\frac{\partial p(x,t)}{\partial t} = D\frac{\partial^2 p(x,t)}{\partial x^2}, \tag{2.10}$$

in which we can identify the classical heat equation.
If the particle starts at $x = 0$, the solution is the Gaussian:

$$p(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right). \tag{2.11}$$

The variance of the distribution grows linearly in time:

$$\mathbb{E}[x^2(t)] = 2Dt. \tag{2.12}$$

This is the reason why, when the Brownian motion SDE is used in a diffusion model, it is referred to as Variance Exploding (VE).

Ornstein-Uhlenbeck Process (Variance Preserving)

The Ornstein-Uhlenbeck (OU) process models a particle subject to a linear restoring force and random noise. It is the simplest example of a *mean-reverting* stochastic process. The Langevin equation for the OU process is:

$$\frac{dx_t}{dt} = -\theta\ x_t + \sigma\ \eta(t) \tag{2.13}$$

or in the standard SDE notation

$$dx = -\theta x\ dt + \sigma\ dW_t, \tag{2.14}$$

where $\theta > 0$ is the strength of the restoring force, and $\sigma > 0$ controls the noise intensity. The corresponding Fokker-Planck equation is:

$$\frac{\partial p(x,t)}{\partial t} = \theta\frac{\partial}{\partial x}\left(x\ p(x,t)\right) + \frac{\sigma^2}{2}\frac{\partial^2 p(x,t)}{\partial x^2}. \tag{2.15}$$

The solution for the initial condition $x = 0$ is again a Gaussian:

$$p(x,t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{x^2}{2\sigma_t^2}\right), \tag{2.16}$$

where the variance evolves as:

$$\sigma_t^2 = \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta t}\right). \tag{2.17}$$

The OU process is an example of a Gaussian process, exactly as Brownian motion. In contrast to Brownian motion, where the drift is absent, here the drift is dependent on the current value of the process: if the current value of the process is less than the mean, the drift will be positive; if the current value of the process is greater than the mean, the drift will be negative. In other words, the mean acts as an equilibrium level for the process (from this, the mean-reverting name).
As $t \to \infty$, the variance reaches the finite value:

$$\lim_{t\to\infty} \sigma_t^2 = \frac{\sigma^2}{2\theta}. \tag{2.18}$$

Thus, the OU process is also known as Variance Preserving (VP): it reaches a stationary distribution, which is Gaussian with zero mean and fixed variance.
We will consider a simpler form for the OU process

$$dx = -x\ dt + \sqrt{2}\ dW_t \tag{2.19}$$

following the standard physics prescription.

2.2 Generative Diffusion

This section is dedicated to the foundations of generative diffusion. Generative diffusion models are a class of probabilistic models that learn to generate complex data (such as images, audio, or text) by simulating a reverse-time stochastic process. These models are built upon the principles of nonequilibrium thermodynamics and stochastic differential equations, bridging ideas from physics and deep learning.

The central idea is to define a *forward process* that gradually destroys structure in the data by adding noise, and then learn a *backward process* that reconstructs the data from noise. When trained effectively, the backward process can generate realistic samples from

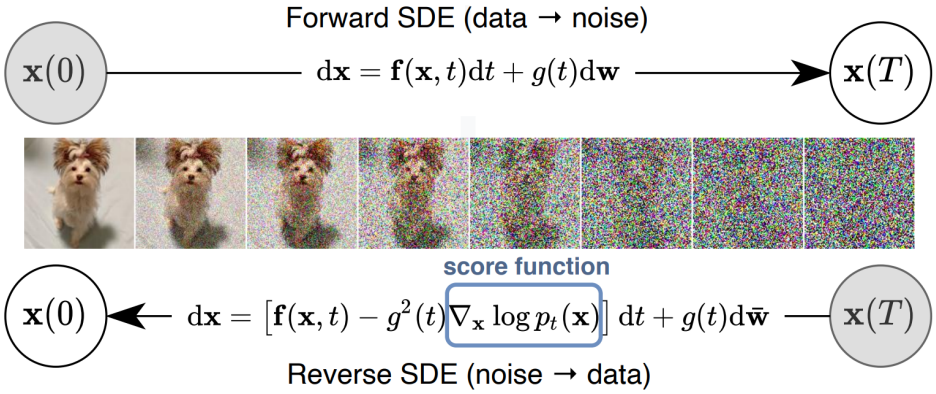


Figure 2.2: Transforming data to a simple noise distribution can be accomplished with a continuous-time SDE. This SDE can be reversed if we know the score of the distribution at each intermediate time step. From Song et al. [2020].

the learned data distribution. This is pictured if Fig. 2.2.

We are particularly interested in this SDE formulation of generative models, Fig.2.3, which was presented in Song et al. [2020], for the clear connection to physics problems. At the end of this section, we also review the most relevant works that led to this formulation in the machine learning community.

2.2.1 Forward process

Assume you have a set of high-dimensional data $x_0 \in \mathbb{R}^d$ that are drawn from a possibly unknown distribution p_0; we can always think about images as an example. We now describe the process through which we corrupt our data, iteratively adding noise, known as the forward process. In the continuous-time limit, the forward process becomes a stochastic differential equation (SDE):

$$dx = f(x, t)\, dt + g(t)\, dW_t. \tag{2.20}$$

If $x \in \mathbb{R}^d$, $f(x, t)$ is a vector-valued function called drift, and $g(t)$ is a scalar called diffusion coefficient, while dW_t denotes standard Brownian motion (or Wiener process). The solution of a stochastic differential equation is a continuous collection of random variables $\{x_t\}_{t\in[0,t_f]}$. These random variables trace stochastic trajectories as the forward time t increases from 0 to the final time t_f. Let us consider $p_t(x)$, the marginal probability density function of x_t. We denote as $p_0(x)$ the original data distribution, which will represent our *target* distribution. Then, $p_t(x)$ is the deformation of our data distribution $p_0(x)$ at time t, obtained by applying the forward-time SDE. You can obtain samples from

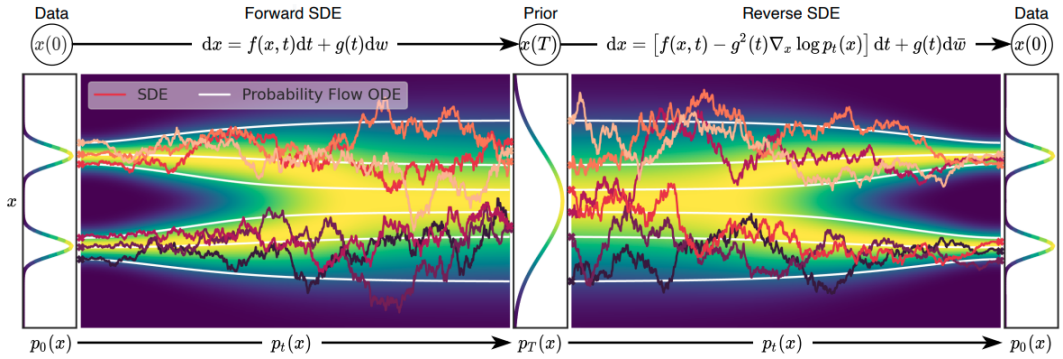


Figure 2.3: Overview of score-based generative modeling through SDEs, from Song et al. [2020].

$p_t(x)$ by sampling x_0 from your data set and then using the forward-time SDE to get a sample x_t.

2.2.2 Reverse Process

Intuitively, the reverse-time dynamics must undo the stochastic diffusion introduced by the forward process. However, because the forward SDE involves both deterministic drift $f(x,t)$ and random diffusion $g(t)dW_t$, the backward process is not simply the same equation with time reversed: diffusion introduces uncertainty that skews the probability flow. Anderson [1982] showed that the time reversal of an SDE can be written as another SDE whose drift includes an additional correction term which captures the instantaneous gradient of the data density. This term effectively steers the diffusion trajectories back toward regions of higher probability density, counteracting the random spreading caused by noise. The closed form of the reverse SDE is given by

$$dx = \left[f(x,t) - g(t)^2 \nabla_x \log p_t(x)\right]\, dt + g(t)\, d\bar{W}_t, \tag{2.21}$$

where $\bar{W}_t$ is a standard Brownian motion in reverse time and $p_t(x)$ is the marginal density of x_t under the forward process. Notice also that dt is a negative infinitesimal time step, since (2.21) needs to be solved backwards in time from $t = t_f$ to $t = 0$.

This result implies that if we can estimate $\nabla_x \log p_t(x)$ at each time t, we can simulate the reverse SDE to generate new samples from the data distribution. This fundamental quantity is known as **score function** of $p_t(x)$. One way to think about this is that it points in the direction where the log-likelihood of the data increases fastest. Let us stress how simple the reverse equation is, while the problem of removing noise over time can seem

very hard. For nice choices of f and g (such as, for example, the variance exploding and variance preserving formalism), p_{t_f} is going to be Gaussian distributed for large enough t_f. In practice, the whole point of diffusion is to make p_{t_f} as simple as possible, so usually we set things up so that $p_{t_f} = N(0, \sigma_{t_f}^2 I)$ for some scalar σ_{t_f}. Then, it suffices to generate Gaussian data x_{t_f} and run the reverse process.

We can use numerical solvers to approximate solutions of Eq. (2.21) and simulate the stochastic process for sample generation. The simplest numerical SDE solver is the Euler-Maruyama method, which discretizes the SDE using finite time steps and Gaussian noise. It is an extension of the Euler method for ordinary differential equations to stochastic differential equations, named after Leonhard Euler and Gisiro Maruyama.

Since, in general, $p_t(x)$ is unknown, estimating the score function is a crucial problem. The success of score-based models relies on the efficient methods designed for this purpose.

2.2.3 Anderson's Theorem

A foundational result in diffusion models is a theorem by Anderson Anderson [1982], which states that most processes defined via a forward time or conventional diffusion equation model have an associated reverse time model, and shows how to build it.

The only sorts of restrictions needed are those that ensure that the Kolmogorov equations for associated probability densities all have unique smooth solutions. They are primarily related to the smoothness and growth properties of the drift and diffusion coefficients of the forward SDE. Sufficient conditions for this include f and g being globally Lipschitz in state and time, being twice continuously differentiable in x, having bounded first-order partial derivatives in x, and their second-order partials not growing faster than $\|x\|^m$ for some $m > 0$. The transition probability density itself should also be sufficiently smooth (twice continuously differentiable in x and continuously differentiable in t). Such restrictions, though hard to translate into requirements on the diffusion and drift quantities, seem nevertheless intrinsic.

The reader is remanded to the paper Anderson [1982] to see the full proof of the result. However, we report here a quick way to see why the reverse SDE has the form of Eq. 2.21. There is a one-to-one correspondence between SDE for x_t and Kolmogorov equations for $p_t(x)$. Consequently, there should be a correspondence also for the reverse SDE and a Kolmogorov equation. In the reversed direction, one may write the Fokker-Planck equation Anderson [1982] for the backward time τ as

$$\partial_\tau \tilde{p}_\tau(x) = -\nabla_x \cdot (x\tilde{p}_\tau(x)) - \Delta_x \tilde{p}_\tau(x). \tag{2.22}$$

where Δ here indicates the Laplacian operator, $\Delta f = \nabla^2 f = \nabla \cdot \nabla f$. By identification with Eq. (2.7), the associated stochastic process is

$$\frac{dx}{d\tau} = x(\tau) + 2\nabla_x \log \tilde{p}_\tau(x) + \eta(\tau). \tag{2.23}$$

Now, noticing that $\tilde{p}_\tau(x) = p_{t_f-\tau}(x)$, and $t_f - \tau = t$, we obtain precisely the expression of the score.

2.2.4 Score function(s)

We have introduced the score function as the gradient of the log-density:

$$s_t(x) = \nabla_x \log p_t(x). \tag{2.24}$$

The score function acts as a guiding field that steers noisy data points toward regions of high data density. It is the continuous analogue of the residual in denoising autoencoders, and it plays a central role in the design of generative diffusion models.

The reverse-time process generates samples from $p_0(x)$ if the score is known. Since, in general, it is not known in practice, it is common to approximate it using a neural network, training a time-dependent score-based model

$$\mathbf{s}_\theta(x, t) \approx \nabla_x \log p_t(x), \tag{2.25}$$

where θ are trainable parameters. Once we have a trained model, it suffices to input pure noise into it; the model will assume it is x_{t_f} for some very large t_f and it will output the corresponding x_0.

In this thesis, we will also encounter situations where we know the true density of our data, and thus we are able to compute the *exact* score function. This can be interesting to see in which regimes the trained score actually reproduces the behavior of the exact one, as we will do in Chapter 4.

Instead, when the true distribution is not known, the best we can do is to approximate it with its empirical counterpart, resulting in what we call the *empirical* score function. The process with the empirical score at $t = 0$ will draw samples from $p_0(x) = \sum_{\mu=1}^{P} \delta(x - \xi^\mu)$, where $\{\xi^\mu\}$ are the data, thus ending up in one of the training points. What is interesting about this approximation is that, in principle, the empirical score has access to the same information as the trained one: the dataset. But we can identify two substantially different behaviours: the empirical score always memorizes the

dataset, while a properly trained model is able to generalize.

2.2.5 Score Matching

Our main goal is to learn score functions by minimizing the expected squared error between the true and estimated score:

$$\mathcal{L}_{\mathrm{SM}}(\theta) = \mathbb{E}_{p_t(x)}\left[\|\nabla_x \log p_t(x) - \mathbf{s}_\theta(x,t)\|^2\right] \tag{2.26}$$

that falls under the category of score matching-problems.

Score matching is an alternative to the maximum likelihood principle suitable for unnormalized probability density models whose partition function is intractable. In score matching, the goal is to make the score function of the model as close as possible to the score function of the data distribution. The explicit score matching objective is given by the Fisher divergence between the two scores

$$J(\theta) = \frac{1}{2}\int p(x)\,\|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2\,dx. \tag{2.27}$$

Minimizing this is still difficult since $p(x)$ is unknown. Hyvärinen and Dayan [2005] showed that the objective can be reformulated using integration by parts

$$J(\theta) = \int \left[\frac{1}{2}\,\|\nabla_x \log p_\theta(x)\|^2 + \Delta_x \log p_\theta(x)\right] p(x)dx, \tag{2.28}$$

which can be estimated directly from data without knowledge of the normalization constant. This is also called implicit score matching.

Vincent [2011] showed that training a denoising autoencoder to predict x_0 from x_t is equivalent to score matching under Gaussian noise. Denoising autoencoders are a simple modification of classical autoencoder neural networks that are trained not to reconstruct their input, but rather to denoise an artificially corrupted version of their input. Consider the joint density $p(x_t, x_0) = p(x_t \mid x_0)p_0(x)$, we define the denoising score matching objective

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{p(x_t,x_0)}\left[\|\mathbf{s}_\theta(x_t,t) - \nabla_{x_t}\log p(x_t \mid x_0)\|^2\right], \tag{2.29}$$

and with the considered Gaussian kernel

$$\nabla_{x_t}\log p(x_t \mid x_0) = \frac{x_0 - x_t}{\sigma_t} \tag{2.30}$$

and given $x_t = x_0 + \sigma_t \epsilon_t$, it suffices indeed to estimate the noise $\hat{\epsilon}_\theta(x, t)$ to learn the score $\hat{s}_\theta(x_t, t) = -\hat{\epsilon}_\theta(x, t)/\sigma_t$. This formulation enables efficient training of score-based models using simple loss functions on noisy inputs.

2.2.6 Review of Generative Diffusion Models

In the following, we present short summaries of the most relevant works about diffusion models in the field of machine learning. Although we will perform all our analysis on the continuous SDE formulation using statistical physics methods, it is important to have in mind how these concepts evolved in the machine learning community.

Deep Unsupervised Learning Using Nonequilibrium Thermodynamics

The seminal paper Sohl-Dickstein et al. [2015] introduced the first full generative diffusion model, inspired by nonequilibrium thermodynamics. They showed that diffusion models work for image synthesis, and they also proposed the forward versus reverse diffusion process, although in a different way from the SDE approach we have described until now. They used a discrete-time Markov chain with a forward diffusion kernel

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I), \tag{2.31}$$

with small variance increments β_t. This is equivalent to a forward time SDE with $dx = (x\sqrt{1-\beta_t})dt + \beta_t dw$. The reverse diffusion kernel is also Gaussian:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{2.32}$$

where you train a neural network to learn the mean and covariance functions. The model is trained by minimizing a variational bound on the negative log-likelihood:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_q \left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})} \right]. \tag{2.33}$$

where $\mathbb{E}_q$ represents expectation taken under the distribution q. This variational objective provides a tractable way to optimize the reverse process. While this model required long diffusion chains and was computationally intensive, it demonstrated that deep unsupervised learning could be formulated using thermodynamic principles. The work set the stage for later improvements in diffusion efficiency and score estimation.

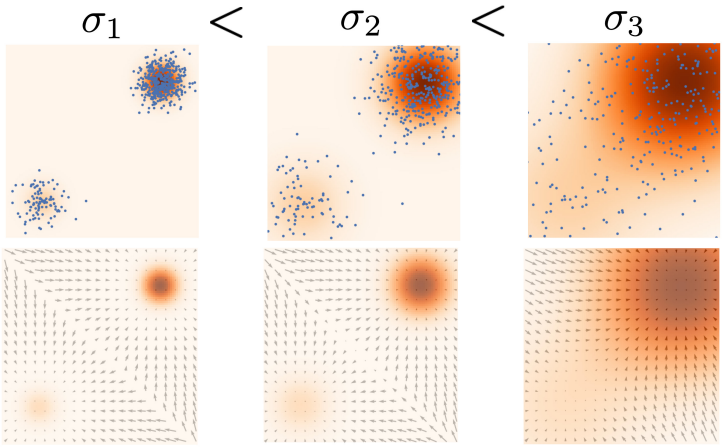


Figure 2.4: Multiple scales of noise to learn the score from Song and Ermon [2019].

Generative Modeling by Estimating Gradients of the Data Distribution

In Song and Ermon [2019], the authors reframed generative modeling as a problem of learning and sampling from the score function across multiple noise levels. The authors proposed a technique called *score-based generative modeling*, which relies entirely on denoising score matching and Langevin dynamics.

One of the known problems of naive score matching is that it does not learn the score well in low-density regions, where presumably there will be less data available. Instead of a forward diffusion model, they defined a family of noise-perturbed data distributions $\{p_t(x)\}_{t\in[0,1]}$, where $x_t = x_0 + \sigma_t\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, with σ_t increasing with t to simulate a continuous noising process. They trained a noise-conditioned neural network (NCSN) $\mathbf{s}_\theta(x, t)$ to approximate $\nabla_x \log p_t(x)$ using denoising score matching, see Fig 2.4. The training objective corresponds to minimizing the Fisher divergence between the true score function $\nabla \log p_t(x)$ and the model prediction $\mathbf{s}_\theta(x, t)$ across all noise levels. Since each noise scale σ_t defines a different perturbed data distribution $p_t(x)$, the total objective integrates these divergences over time, typically with a weighting factor that accounts for the variance or relative importance of each scale. In effect, the model learns to estimate accurate scores over the entire range of noise magnitudes.

This method produced high-quality samples with better mode coverage than GANs and without adversarial training. Importantly, it required no likelihood or variational bound, relying purely on learned gradients of the log-density. This paper laid the theoretical and practical foundation for later work on continuous-time diffusion processes and score-based SDEs.

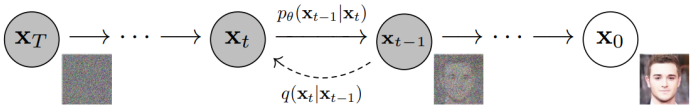


Figure 2.5: The directed graphical model considered in Ho et al. [2020]

Denoising Diffusion Probabilistic Models

In their foundational work Ho et al. [2020], the authors introduced *denoising diffusion probabilistic models* (DDPMs), a class of generative models that significantly improved the efficiency and quality of diffusion-based sampling. The approach is very similar to Sohl-Dickstein et al. [2015], see Fig. 2.5 for the scheme. Indeed, the forward process is a Markov chain defined by the kernel

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{2.34}$$

with a schedule $\{\beta_t\}_{t=1}^T$. By iterating this process, one can express x_t as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{2.35}$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1-\beta_s)$. This is often called the "reparametrization trick".

The reverse process is defined as before

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_t), \tag{2.36}$$

but instead of estimating the reverse mean directly, DDPMs train a neural network $\epsilon_\theta(x_t, t)$ to predict the noise added to x_0, using the following loss:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2.37}$$

then μ_θ is computed from ϵ_θ using the known forward process. This choice greatly simplified training and aligned with denoising score matching theory.

DDPMs achieved state-of-the-art image generation results and demonstrated that score-based diffusion models could match GAN-level sample quality with greater stability and interpretability. This work also highlighted a key insight: minimizing a simple denoising loss is sufficient for training an effective generative model, connecting Vincent [2011] and Sohl-Dickstein et al. [2015] through a unified framework.

Score-Based Generative Modeling through Stochastic Differential Equations

The seminal work Song et al. [2020] unified score-based generative modeling Sohl-Dickstein et al. [2015] and diffusion probabilistic models Ho et al. [2020] under the framework of stochastic differential equations. They showed that both approaches can be described using a continuous-time forward-noising process and a reverse-time generative process governed by learned scores. The framework of diffusion models that we have presented in this section is mainly based on this paper. The authors introduced the forward process modeled by a general Itô SDE (2.20), and the reverse-time SDE is derived using Eq. (2.21). Moreover, they also introduced the score-matching objective. This paper realizes DDPM as a special case, with values of f and g set to

$$dx = -\frac{1}{2}\beta(t)x\ dt + \sqrt{\beta(t)}\ dW \tag{2.38}$$

In addition to stochastic sampling via the reverse SDE, they proposed a deterministic alternative using a probability flow ODE:

$$\frac{dx}{dt} = f(x,t) - \frac{1}{2}g(t)^2\nabla_x \log p_t(x), \tag{2.39}$$

which shares the same marginal distributions as the SDE but allows for exact likelihood computation via the instantaneous change of variables formula. Starting from the Fokker-Planck equation (2.7), that we rewrite here as

$$\frac{\partial p(x,t)}{\partial t} = -\sum_{i=1}^{N}\frac{\partial}{\partial x_i}[f_i(x,t)p(x,t)] + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial^2}{\partial x_i \partial x_j}[D_{ij}(t)p(x,t)]\frac{\partial p_t(x)}{\partial t} \tag{2.40}$$

$$= -\nabla\cdot[f(x,t)p_t(x)] + \nabla\cdot[D(t)\nabla_x p(x,t)] \tag{2.41}$$

if we write

$$\tilde{f}(x,t) = f(x,t) - \frac{1}{2}g(t)^2\nabla_x \log(p_t(x)) \tag{2.42}$$

then

$$\frac{\partial p_t(x)}{\partial t} = -\sum_{i=1}^{N}\frac{\partial}{\partial x_i}[\tilde{f}_i(x,t)p_t(x)] \tag{2.43}$$

has no diffusion term. In fact, if we apply Fokker-Planck to the ODE $dx = \tilde{f}(x,t)dt$, we get the same expression.

This work provided a complete theoretical foundation for score-based generative modeling and demonstrated that it encompasses DDPMs as a special case. It also introduced

efficient sampling algorithms and enabled high-quality image synthesis rivaling GANs.

2.3 Statistical Physics and Diffusion Models

In this section, we revise the main contribution from the field of statistical physics applied to diffusion models. As we have seen in the introduction, many interesting works in this field analyze different phenomena in diffusion models through the lens of statistical physics. Here we go deeper in describing the contribution of two of these papers that constitute an important background for the original work of this thesis. The first paper that started this line is Biroli and Mézard [2023], and the main contributions are summarized in Sec. 2.3.1. In Sec. 2.3.2 we introduce the Random Energy Model, and in Sec. 2.3.3 we review how it was implied in the analysis of diffusion models in Biroli et al. [2024].

Generative diffusion models are associative memory networks

We have talked about memorization in diffusion models, and it is worth noting that *memory* models and their capacity have been studied in physics since Hopfield proposed an associative memory model in 1982 Hopfield [1982]. In recent years, associative memory networks have been generalized to larger capacities Krotov and Hopfield [2016], Demircigil et al. [2017]. Most notably, Ref. Lucibello and Mézard [2024] uses the same REM analogy that we describe for diffusion models, constituting in some sense the inspiration for this line of research. In Ambrogioni [2024], the connection between dense associative memory and diffusion models is consolidated by showing that, when used for storing discrete patterns, the dynamics of generative diffusion models minimize the energy function of continuous modern Hopfield networks.

2.3.1 Generative Diffusion in very large dimensions

The foundational work Ref. Biroli and Mézard [2023] investigates the behavior of generative diffusion models in the high-dimensional limit, employing methods from statistical physics. The two main questions they explore are:

1. How does the number of training samples P need to scale with the system size N in order for the diffusion model to faithfully capture the underlying data distribution?

2. Does varying the scaling of P with N lead to qualitatively different regimes, such as the emergence of spontaneous symmetry breaking in high-dimensional spaces?

They analyze two well-controlled cases: a Gaussian model, in order to answer the first question, and the Curie-Weiss model of ferromagnetism, which exhibits a phase transition, and is, therefore, suited for the second question. They found that the diffusion model generates multivariate Gaussian samples whose covariance coincides with the original one only for $P \gg N^2$. Regarding the second questions, the generative diffusion is able to reconstruct typical configurations belonging to the two states of the Curie-Weiss model as soon as $P \gg 1$, provided the forward process is followed up to times t such that $\sqrt{N}e^{-t} \ll 1$. However, in order to reconstruct the relative weights of these two states, one needs a much larger data base, with a number of points $P \gg N$.

Gaussian data

We report here the full analysis of the Gaussian data case. Suppose that we have P data points drawn from a multivariate normal distribution, $\xi^\mu \sim \mathcal{N}(m, \Sigma)$, $\mu = 1, \dots, P$, with $m \in \mathbb{R}^N$ the vector of means and $\Sigma \in \mathbb{R}^{N \times N}$ the covariance matrix. In this Gaussian context, it is useful to notice

$$\mathcal{N}(x; \xi e^{-t}, \Delta_t) = e^{tN} \mathcal{N}(\xi; x e^t, \Delta_t e^{2t} \mathbb{I}). \tag{2.44}$$

The joint distribution $P_t(x, \xi)$ can then be written as

$$P_t(x, \xi) = \mathcal{N}(\xi; m, \Sigma) e^{tN} \mathcal{N}(\xi; x e^t, \Delta_t e^{2t} \mathbb{I}), \tag{2.45}$$

meaning that averaging over the data distribution amounts to performing a convolution and results in

$$P_t(x) = \mathcal{N}(x; m e^{-t}, \Delta_t + e^{-2t} \Sigma). \tag{2.46}$$

Given that this distribution is Gaussian, the exact score function will necessarily be linear. As a result, one can parametrize the score as

$$\hat{S}_\theta(x, t) = -W_t x - b_t, \tag{2.47}$$

where the parameters are $\theta = \{W, b\}$ and the loss can be computed exactly. Setting $b = m = 0$ for convenience, one has that for a given time

$$\mathcal{L}_t(W_t) = \frac{1}{\Delta_t} \left[\mathrm{Tr}(\mathbb{I} - \Delta_t W_t)^2 + e^{-2t} \mathrm{Tr}(W_t C^e W_t) \right], \tag{2.48}$$

here we have performed the average on the empirical sample, leading to the empirical covariance matrix

$$C^e_{ij} = \sum_{\mu=1}^{P} a^\mu_i a^\mu_j. \tag{2.49}$$

The matrix W_t is finally given by

$$W_t = \left[\Delta_t + e^{-2t} C^e\right]^{-1}, \tag{2.50}$$

which will be close to the exact result $(\Gamma_t)^{-1}$, where $\Gamma_t = \Delta_t + e^{-2t}\Sigma$, only if the empirical correlation matrix is close to Σ.
In the infinite-dimensional limit, the spectral density of that matrix can be studied in order to show that Biroli and Mézard [2023]

- When $\frac{P}{N} \gg 1$, the *spectrum* of C^e matches that of Σ,
- When $\frac{P}{N^2} \gg 1$, the *eigenvectors* of C^e match those of Σ.

As a result, one needs $P \sim N^2$ samples in order to generate new samples with fully correct statistics.
It is interesting to note that the above derivation also applies to the use of a linear score on arbitrary, possibly non-Gaussian data. We have thus shown that in this case, the optimal linear score will generate a distribution, at the end of the backward process, which is Gaussian, with the empirical mean and covariance of the sample on which the training of the score was done.

Curie-Weiss and Speciation

In the Curie-Weiss model, they highlight the mechanism of symmetry breaking during the inverse diffusion process. Specifically, to accurately reconstruct the relative asymmetry of the two low-temperature states (and thus obtain correct probability weights), a dataset with a number of points much larger than the dimension of each data point is required. This finding underscores the importance of dataset size in high-dimensional generative modeling.
The studied case starts from a Curie-Weiss model coupling some Ising spins $a_i = \pm 1$, at

inverse temperature β, with a small external field h/N that will create a slight asymmetry

$$P_0(a) = \frac{1}{Z_0} \exp\left(\frac{\beta}{2N}\left(\sum_i a_i\right)^2 + \frac{h}{N}\sum_i a_i\right). \tag{2.51}$$

where Z_0 is the partition function of the model.
The distribution $P_0(a)$ can be rewritten by introducing the average magnetization m. Summing over $a_i = \pm 1$, we can deduce the distribution of m, and the saddle point on m gives two dominating magnetizations $\pm m^*$, where m^* is the solution of $m^* = \tanh(\beta m^*)$. In the infinite dimension limit, the distribution will concentrate on the two solutions. As a result, we define the slightly simplified Curie-Weiss model

$$P_0(m) = W_+\delta(m - m^*) + W_-\delta(m + m^*), \tag{2.52}$$

where $W_\pm = \frac{e^{\pm hm^*}}{2\cosh(hm^*)}$ and $P_0(a \mid m) = \frac{1}{Z}\prod_i e^{\beta m a_i}$, which corresponds to the data distribution

$$P_0(a) = W_+ \prod_i \frac{e^{\beta m^* a_i}}{2\cosh(\beta m^*)} + W_- \prod_i \frac{e^{-\beta m^* a_i}}{2\cosh(\beta m^*)}, \tag{2.53}$$

We now place ourselves in the case of the exact score. Recall that we have

$$P_t(x, a) = P_0(a)\ \frac{e^{-\frac{(x - ae^{-t})^2}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}^N}, \tag{2.54}$$

which for the considered data model can be written as

$$P_t(x, a) = W_+ P_t^+(x, a) + W_- P_t^-(x, a), \tag{2.55}$$

where

$$P_t^\pm(x, a) = \left(\prod_i \frac{e^{-\frac{x_i^2 + e^{-2t}}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}}\right)\left(\prod_i e^{a_i\left(\pm\beta m^* + x_i\frac{e^{-t}}{\Delta_t}\right)}\right). \tag{2.56}$$

Let us compute $\langle a\rangle_{t,x}$, the average with respect to $P_t(a \mid x)$, to obtain the score

$$\langle a_i\rangle_{x,t} = \frac{W_+ Q_+(x)\tanh(\beta m^* + x_i\frac{e^{-t}}{\Delta_t}) + W_- Q_-(x)\tanh(-\beta m^* + x_i\frac{e^{-t}}{\Delta_t})}{W_+ Q_+(x) + W_- Q_-(x)}, \tag{2.57}$$

with

$$Q_{\pm}(x) = \prod_j \left(1 \pm m^* \tanh\left(\frac{x_j e^{-t}}{\Delta_t}\right)\right). \tag{2.58}$$

Let us highlight that in the case of the Curie-Weiss, the forward process starts from points $a \in \{\pm 1\}^N$ and, as soon as $t > 0$, the trajectories are in $\mathbb{R}^N$. The score function has, therefore, an argument in $\mathbb{R}^N$ and the backward process (2.2.2) with the exact score must transport points in $\mathbb{R}^N$ to points in the simplex $\{\pm 1\}^N$.
We study the score function and the backward process at large times $t \gg 1$. First, we have that for all j, $\frac{x_j e^{-t}}{\Delta_t} \ll 1$. Thus

$$Q_{\pm}(x) = \prod_j \left(1 \pm m^* x_j \frac{e^{-t}}{\Delta_t}\right) \simeq e^{\pm m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j}. \tag{2.59}$$

Using the mean-field equation on the magnetization $\tanh(\pm\beta m^* + x_i \frac{e^{-t}}{\Delta_t}) \simeq m^*$,

$$\langle a_i \rangle_{x,t} = m^* \frac{W_+ e^{m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j} - W_- e^{-m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j}}{W_+ e^{m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j} + W_- e^{-m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j}}. \tag{2.60}$$

We consider the zero-field case so that $W_+ = W_- = \frac{1}{2}$. The exact score function at large times reads

$$S(x,t) \simeq -\frac{x}{\Delta_t} + m^* \frac{e^{-t}}{\Delta_t} \tanh\left(m^* \frac{e^{-t}}{\Delta_t} \sum_j x_j\right). \tag{2.61}$$

At large times $\Delta_t = 1 - e^{-2t} \simeq 1$ and the score simplifies to

$$S(x,t) \simeq -x + m^* e^{-t} \tanh\left(m^* e^{-t} \sum_j x_j\right). \tag{2.62}$$

This score yields the following backward process for each component x_i of x

$$\frac{dx_i}{d\tau} = -x_i + 2m^* e^{-t} \tanh\left(m^* e^{-t} \sum_j x_j\right) + \eta_i(\tau). \tag{2.63}$$

By summing the backward equation for each component, we obtain an effective equation for the collective variable $\mu = \frac{1}{\sqrt{N}} \sum_j x_j$. We introduce the normalization $1/\sqrt{N}$ so that μ remains of order one when N is large. Indeed, at the beginning of the generative process the x_i are independent Gaussian variables, and the sum of N such terms has a variance that scales as N. Dividing by $\sqrt{N}$ therefore keeps the fluctuations of μ finite in the

large-N limit.

$$\frac{d\mu}{d\tau} = -\mu + 2m^* e^{-t}\sqrt{N}\tanh\left(m^* e^{-t}\sqrt{N}\mu\right) + \eta(\tau) \tag{2.64}$$

$$= -\frac{dV}{d\mu} + \eta(\tau). \tag{2.65}$$

This is a Langevin equation in a time-dependent potential,

$$V(\mu, t) = \frac{\mu^2}{2} - 2\log\cosh\left(\mu e^{-t}\sqrt{N}m^*\right). \tag{2.66}$$

The form of this potential changes qualitatively with time through the factor $e^{-t}\sqrt{N}$. When $\sqrt{N}e^{-t} \ll 1$, the argument of the hyperbolic cosine is small and we can expand $\log\cosh(x) \approx x^2/2$, giving

$$V(\mu, t) \approx \frac{1}{2}\left(1 - (m^*)^2 e^{-2t}N\right)\mu^2, \tag{2.67}$$

so that $V(\mu, t)$ is approximately quadratic and μ fluctuates around zero. In this early regime, the system behaves as if it were in a single harmonic well centered at $\mu = 0$. In contrast, when $\sqrt{N}e^{-t} \gg 1$, the potential develops two symmetric minima corresponding to stable states at nonzero values of μ. Physically, this indicates a spontaneous symmetry breaking: the backward trajectory eventually commits to one of the two wells, which determines the outcome of the generative process.
The crossover between these two regimes occurs when the potential transitions from single-well to double-well. This defines a characteristic or **speciation** time t_s such that

$$\sqrt{N}e^{-t_s} = 1 \quad \Longleftrightarrow \quad t_s = \frac{1}{2}\log N. \tag{2.68}$$

2.3.2 The Random Energy Model

In this section, we present the Random Energy Model following Mezard and Montanari [2009]. We also exploit this model as an occasion to present the celebrated replica method, a pillar of statistical physics of disordered systems introduced by the Nobel prize winner Giorgio Parisi.

The random energy model (REM), introduced by Derrida [1981], is probably the simplest statistical-physics model of a disordered system which exhibits a phase transition. It is not intended to give a realistic description of any specific physical system, but rather

to provide an example in which various concepts and methods can be studied in full detail. Owing to its simplicity, the same mathematical structure reappears in many other contexts.

Consider 2^N spin configurations. The REM is a disordered model: the energy is not a deterministic function, but rather a stochastic process. Assign to each configuration i an i.i.d. energy E_i drawn from a Gaussian law with mean 0 and variance $N/2$, i.e.

$$P(E) = \mathcal{N}\left(0, \frac{N}{2}\right) = \frac{1}{\sqrt{\pi N}} \exp\left(-\frac{E^2}{N}\right), \tag{2.69}$$

even though other distributions can be studied as well. Given an instance of the REM, one assigns to each configuration j a Boltzmann probability $P_\beta(j)$ in the usual way

$$P_\beta(j) = \frac{1}{Z_N} e^{-\beta E_j} \tag{2.70}$$

where the normalization factor Z_N, also known as partition function, at inverse temperature β is

$$Z_N(\beta) = \sum_{i=1}^{2^N} e^{-\beta E_i}. \tag{2.71}$$

We are interested in the properties of a probability distribution, the Boltzmann distribution, which is itself a random object because the energy levels are random variables.

Rigorous solution via large deviations

We start to computing the exact asymptotic solution of the model, without using the replica method. To do so, we shall show that the entropy density as a function of the energy has a deterministic asymptotic limit, that we can compute. Write the energy density $\varepsilon = E/N$. Let $\#(E)$ denote the number of sampled energies falling in $[Ne, N(\varepsilon + \delta\varepsilon)]$. This is a random variable, it depends on the specific draw of the 2^N configuration, but we can compute its mean and variance. By linearity of expectation and the Gaussian tail estimate,

$$\mathbb{E}[\#(E)] \approx \frac{2^N}{\sqrt{\pi N}} e^{-N\varepsilon^2} \, \delta\varepsilon = \frac{1}{\sqrt{\pi N}} e^{N[\log 2 - \varepsilon^2]} \, \delta\varepsilon. \tag{2.72}$$

Since the probability of the energy taken randomly from $P(E)$ to be between e and $\varepsilon + \delta\varepsilon$ is small, this follows a Poisson law, so that the variance is equal to the mean. Define the annealed entropy density $s_{\text{ann}}(\varepsilon) = \log 2 - \varepsilon^2$. A simple Markov bound (or first-moment

method) gives

$$\mathbb{P}\{\#(E) \geq 1\} \leq \mathbb{E}[\#(E)]. \tag{2.73}$$

Hence with high probability as $N \to \infty$ there are no sampled energies outside the band $|\varepsilon| \leq \sqrt{\log 2}$ (where $s_{\text{ann}}(\varepsilon) \geq 0$). Within that band, we have instead an exponential number of configurations. However the entropy density concentrates to a deterministic value, as can be seen from Chebyshev inequality and the fact that the variance is equal to the mean:

$$\mathbb{P}\left(\left|\frac{\#(\varepsilon)}{\mathbb{E}\#(\varepsilon} - 1\right| \geq C\right) \leq \frac{\text{Var}\#(\varepsilon)^2}{C^2(\mathbb{E}\#(\varepsilon))^2} \propto \frac{e^{-Ns_{\text{ann}}(\varepsilon)}}{C^2}. \tag{2.74}$$

Thanks to this very tight concentration, we can write the entropy density as a function of ε

$$s(\varepsilon) = \begin{cases} s_{\text{ann}}(\varepsilon) = \log 2 - \varepsilon^2, & s_{\text{ann}}(\varepsilon) \geq 0, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2.75}$$

We can now compute the partition function. We approximate the sum in (2.71) by an integral over energy densities:

$$Z_N(\beta) = \sum_E \#(E) e^{-\beta E} \approx \int_{-\varepsilon^*}^{\varepsilon^*} e^{N[s(e)-\beta\varepsilon]}\, d\varepsilon. \tag{2.76}$$

Using Laplace method, we approximate it with the maximum. Therefore, the free-entropy density $\phi(\beta)$ is the Legendre transform of $s(e)$:

$$\phi(\beta) = \lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta) = \max_{\varepsilon \in [-\varepsilon^*, \varepsilon^*]} [s(\varepsilon) - \beta\varepsilon]. \tag{2.77}$$

Evaluating the quadratic form $s_{\text{ann}}(\varepsilon) = \log 2 - \varepsilon^2$ there are two situations: the minimum can be reached when the derivative is 0, that is when $\beta = \partial s(\varepsilon) = -2\varepsilon$, but this can only happen when $s(\varepsilon) > 0$. At $\varepsilon = -\sqrt{\log 2}$, $s(\varepsilon)$ reaches zero, with a derivative of $\beta = 2\sqrt{\log 2}$, so this minimum is only valid when $\beta < \beta_c = 2\sqrt{\log 2}$, after which $\varepsilon^* = -\sqrt{\log 2}$. In conclusion, one finds

$$\lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta) = \begin{cases} \log 2 + \dfrac{\beta^2}{4}, & \beta \leq \beta_c, \\ \beta\sqrt{\log 2}, & \beta \geq \beta_c, \end{cases} \qquad \beta_c = 2\sqrt{\log 2}. \tag{2.78}$$

This shows that a phase transition takes place at the inverse critical temperature β_c. The two temperature regimes have distinct qualitative properties:

- In the high-temperature phase, the energy and entropy densities are $u(\beta) = -\beta/2$

and $s(\beta) = \log 2 - \beta^2/4$, respectively. The Boltzmann measure is dominated by configurations with an energy $E_i \approx -N\beta/2$. There is an exponentially large number of configurations having such an energy density and the Boltzmann measure is roughly equidistributed among such configurations. In the infinite-temperature limit $\beta \to 0$ the Boltzmann measure becomes uniform, and one finds, as expected, $u(\beta) \to 0$ and $s \to \log 2$.

- In the low-temperature phase, the thermodynamic potentials are constant: $u(\beta) = -\varepsilon^*$ and $s(\beta) = 0$. The relevant configurations are the ones with the lowest energy density, namely those with $E_i/N \approx -\varepsilon^*$. The Boltzmann measure is dominated by a relatively small set of configurations, which is not exponentially large in N.

The transition is a **condensation** transition: for $\beta > \beta_c$ the Gibbs measure concentrates on the lowest-energy configurations near $\varepsilon = -\sqrt{\log 2}$.

Solution via the replica method

We now move to the computation by the replica method. We would like to compute the thermodynamic potentials, in particular the free-energy density $f(\beta) = -\frac{1}{\beta N} \log Z_N$ since other potentials can be derived from its derivatives. In order to describe typical samples, one has to compute the average of the log-partition function, $\mathbb{E} \log Z_N$. Since this task can be difficult, the idea of the replica method goes as follows: starting from the relation

$$\mathbb{E} \log Z_N = \lim_{n \to 0} \frac{1}{n} \log \mathbb{E} Z_N^n \tag{2.79}$$

we carry out the computation of the moments of the partition function $\mathbb{E} Z_N^n$ as if n was integer. In other words, the new system is formed by n statistically independent copies of the original one, and we shall refer to these copies as **replicas**. Only at the and we perform an analytic continuation recovering the limit $n \to 0$.

In the REM case, consider $n \in \mathbb{N}$ replicas and write the replicated partition function

$$Z_N^n = \prod_{a=1}^{n} \left(\sum_{i=1}^{2^N} e^{-\beta E_i} \right) = \sum_{i_1, \dots, i_n} e^{-\beta(E_{i_1} + \cdots + E_{i_n})} = \sum_{i_1, \dots, i_n} \prod_{j=1}^{2^N} e^{-\beta E_j (\sum_{a=1}^{n} \mathbb{I}(j = i_a))}. \tag{2.80}$$

Using independence across energies and the Gaussian identity $\mathbb{E}[e^{tX}] = e^{\Delta t^2/2}$ for $X \sim$

$\mathcal{N}(0,\Delta)$ (here $\Delta = N/2$), we push the average over disorder inside the sum to obtain

$$\mathbb{E}Z_N^n = \mathbb{E}\sum_{i_1,\dots,i_n}\prod_{j=1}^{2^N} e^{-\beta E_j(\sum_{a=1}^n \mathbb{I}(j=i_a))} = \sum_{i_1,\dots,i_n} e^{\frac{N\beta^2}{4}\sum_{a,b=1}^n \mathbb{I}(i_a=i_b)}. \tag{2.81}$$

This replicated system has several interesting properties. First of all, it is no longer a disordered system: the energy is a deterministic function of the configuration. Second, replicas are no longer statistically independent: they do interact. Given the replicas configurations $(i_1 \dots i_n)$, it is convenient to introduce the $n \times n$ overlap matrix $Q_{ab} = \mathbb{I}(i_a = i_b)$, that takes elements in $\{0,1\}$ respectively if the two replicas (row and column) have different or equal configuration. We shall refer to this matrix as the overlap matrix. We then write the replicated sum over configurations as

$$\mathbb{E}Z_N^n = \sum_Q H(Q) e^{\frac{N\beta^2}{4}\sum_{a,b} Q_{ab}}. \tag{2.82}$$

with combinatorial weights $H(Q)$ counting the number of configurations whose overlap matrix is Q and the sum runs over the symmetric $\{0,1\}$ matrices with ones on the diagonal. Keeping n integer, it is natural to expect that the number of configurations for a given overlap matrix, $H(Q)$, to be exponential so that, $H(Q) = e^{Ns_q(Q)}$, and

$$\mathbb{E}Z_N^n \approx \int dQ\; e^{N\left(\frac{\beta^2}{4}\sum_{a,b} Q_{ab} + s_q(Q)\right)} = \int dQ e^{N\; g(\beta,Q)}. \tag{2.83}$$

The replica-symmetric saddle point

In the large N limit, we expect to be able to perform a Laplace (or *saddle point*) approximation by choosing the structure of the matrix Q, which will dominate the sum. A quite natural *ansatz* is to assume that all replicas are identical, and therefore the system should be invariant under the relabelling of the replicas (permutation symmetry). In this case, called the replica symmetric (RS) one, we only have two choices

1. If $Q_{ab} = 1$ for all a, b, then all the replica are in a single configuration and $g(\beta, Q) = n\beta^2/4 + \log 2$.

2. If instead $Q_{aa} = 1$ and $Q_{ab} = 0$ for all $a \neq b$, then all replicas are in a different configuration, so that $s_q(Q) \approx n \log 2$ and $g(\beta, Q) = n\beta^2/4 + n \log 2$.

At the replica symmetric level, we thus find that the free entropy is given, at all temperature, by

$$f_{\mathrm{RS}}(\beta) = \log 2 + \frac{\beta^2}{4}, \tag{2.84}$$

valid only at high temperature (small β), as it coincides with the rigorous result (2.78) for $\beta \leq \beta_c$. Within the RS framework, there is no way to get the correct solution for $\beta > \beta_c$.

One-step replica symmetry breaking (1RSB)

Since the RS solution leads to the wrong result, Parisi proposed the **replica-symmetry-breaking** scheme, a recursive procedure for defining larger and larger spaces of matrices Q. To perform the first step, called 1RSB, we have to partition the n replicas into n/m groups of size m (eventually $n \to 0$ with $0 < m \leq 1$). Replicas within a group coincide (overlap 1), replicas in different groups are distinct (overlap 0). Thus, we have

$$\mathbb{E}Z^n \simeq e^{N\left(\frac{n}{m}\log 2 + \frac{\beta^2}{4}nm\right)} \tag{2.85}$$

and

$$g(\beta, Q) = \frac{n}{m}\log 2 + \frac{\beta^2}{4}nm \tag{2.86}$$

leads, in the limit $n \to 0$, to

$$f_{\mathrm{1RSB}}(\beta, m) = \frac{\beta^2}{4} + \frac{\log 2}{m}. \tag{2.87}$$

Extremizing with respect to m under the constraint $0 < m \leq 1$ yields

$$m^* = \min\left\{1, \frac{\beta_c}{\beta}\right\}, \qquad \beta_c = 2\sqrt{\log 2}, \tag{2.88}$$

and therefore

$$f_{\mathrm{1RSB}}(\beta, m) = \begin{cases} \log 2 + \dfrac{\beta^2}{4}, & \beta \leq \beta_c \quad (m^* = 1), \\ \beta\sqrt{\log 2}, & \beta \geq \beta_c \quad (m^* = \beta_c/\beta), \end{cases} \tag{2.89}$$

which again reproduces the rigorous two-phase formula (2.78).

Condensation and the participation ratio

The replica method can also be used to shed some light on the type of phase transition happening at β_c. Indeed, we know that for low temperature, the Boltmzann measure

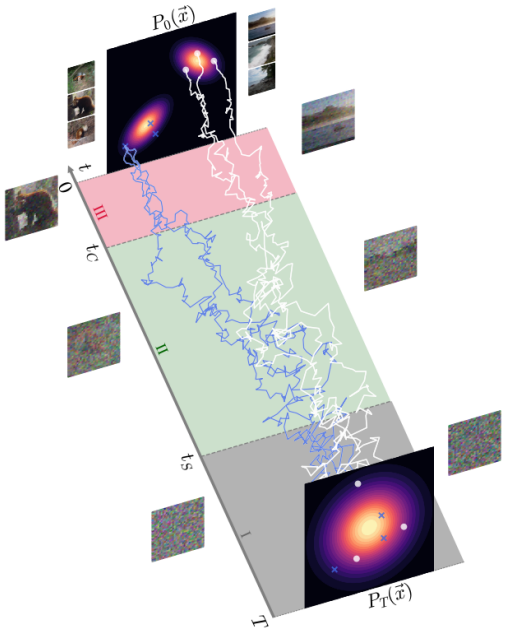


Figure 2.6: Dynamical regimes of diffusion models from Biroli et al. [2024]. In regime I, the trajectories are purely Brownian, in regime II, they commit to one class, and finally, in regime III, they collapse on a single data point.

condensate on the lowest energy level, close to e^*, but we may ask, how many are they really. A sharp way to diagnose condensation is via the participation ratio, defined as

$$Y = \sum_{i=1}^{2^N} \left(\frac{e^{-\beta E_i}}{Z_N} \right)^2 . \tag{2.90}$$

which is is the inverse of the number of configuration that matters in the sum. Replica manipulations (symmetrization and the $n \to 0$ limit) give the quenched expectation

$$\mathbb{E}\, Y = 1 - m^* = \begin{cases} 0, & \beta \le \beta_c, \\ 1 - \dfrac{\beta_c}{\beta}, & \beta \ge \beta_c, \end{cases} \tag{2.91}$$

which rises from 0 and becomes $O(1)$ in the condensed phase, signaling that only $O(1)$ configurations contribute to Z_N.

2.3.3 Dynamical Regimes of Diffusion Models

The seminal paper Biroli et al. [2024] analyzes the behavior of generative diffusion models in high-dimensional spaces with large datasets, under the empirical score function. Their

study reveals three distinct dynamical regimes during the backward diffusion process, represented in Fig. 2.6:

1. **Regime I**: Initially, the generative process resembles pure Brownian motion, with trajectories not yet influenced by the data structure.

2. **Regime II** starts after the **speciation** transition: at a certain time, trajectories begin to commit to a particular data class, a transition related to symmetry breaking in phase transitions. This speciation time can be determined through spectral analysis of the data correlation matrix.

3. **Regime III**: Eventually, trajectories converge to specific training data points, and we have the **collapse** transition, a phenomenon similar to condensation in glassy systems. The collapse time is related to the excess entropy of the data and highlights the challenges posed by the curse of dimensionality.

These findings provide a theoretical framework for understanding the dynamics of diffusion models and their limitations in high-dimensional settings. In the following, we describe how the authors have identified the speciation and collapse transition. The first one is based on a perturbative expansion of the score at large times, while the second one maps the empirical density at time t into a REM. These two mathematical formulations can also be rephrased in terms of the covariance matrix of the data, thus making it possible to derive speciation and collapse time for generic datasets.

The Speciation transition

One can study the speciation transition starting from large times and writing a perturbative expansion of $P_t(x)$ in the parameter e^{-t}

$$P_t(x) = \int d\xi \; P_0(\xi) \frac{e^{-\frac{(x-\xi e^{-t})^2}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}^N} \tag{2.92}$$

$$= \frac{1}{\sqrt{2\pi\Delta_t}^N} e^{-\frac{x^2}{2\Delta_t} + g(x)} \tag{2.93}$$

where $P_0(\xi)$ indicates the data distribution and we define $g(x)$ as

$$g(x) = \log \int d\xi \; P_0(\xi) e^{-\frac{\xi^2 e^{-2t}}{2\Delta_t} + \frac{e^{-t}}{\Delta_t} x\xi} \tag{2.94}$$

$$= \log \int d\xi \; P_{eff}(\xi) e^{\frac{e^{-t}}{\Delta_t} x\xi} \tag{2.95}$$

which can be interpreted as a generating function for the moments of a with respect to the effective distribution $P_{eff}(x) = P_0(\xi) e^{-\frac{\xi^2 e^{-2t}}{2\Delta_t}}$. Expanding at large times, we find

$$g(x) = \frac{e^{-t}}{\Delta_t} \sum_{i=1}^{N} x_i \langle \xi_i \rangle + \frac{e^{-2t}}{2\Delta_t^2} \sum_{i,j=1}^{N} C_{ij} x_i x_j + O((xe^{-t})^3) \tag{2.96}$$

with $C_{ij} = \langle \xi_i \xi_j \rangle - \langle \xi_i \rangle \langle \xi_j \rangle$. With this expansion, we can write

$$\log P_t(x) = const + \frac{e^{-t}}{\Delta_t} \sum_{i=1}^{N} x_i \langle \xi_i \rangle - \frac{1}{2\Delta_t^2} \sum_{ij} M_{ij} x_i x_j \tag{2.97}$$

where $M = \mathbb{I} - e^{-2t} C$. The speciation time is characterized by a change of curvature in this effective potential. Thus, denoting by Λ the largest eigenvalue of C, we can write a general criterion for t_S

$$e^{-2t_S} \Lambda = 1. \tag{2.98}$$

Notice that that at large times $P_{eff}(t) \approx P_0(t)$: in this case the matrix C is the covariance of $P_0(t)$, which can be estimated as the covariance of the data.

The Collapse transition

In this subsection, we present a sketch of the computation of the collapse time t_c using ideas from the Random Energy Model that we have presented in Section 2.3.2. The complete computation can be found in Biroli et al. [2024]. Start from a point of your dataset, say ξ^1. The noisy data at time t can be written as $x(t) = \xi^1 e^{-t} + \sqrt{1 - e^{-2t}} z$, with $z \sim \mathcal{N}(0, I)$. The noisy empirical distribution can be decomposed into two parts

$$P_t^e(x) = \frac{1}{P\sqrt{2\pi\Delta_t}^N} [Z_1 + Z_{2\dots P}], \tag{2.99}$$

with $Z_1 = e^{-\frac{(x-\xi^1 e^{-t})^2}{2\Delta_t}}$ only involving the initial data point and $Z_{2\dots P} = \sum_{\mu=2}^{P} e^{-\frac{(x-\xi^\mu e^{-t})^2}{2\Delta_t}}$ involving the other data points. We refer to Z_1 as the *signal* term, while $Z_{2\dots P}$ can be regarded as the *noise* term. Z_1 concentrates to $e^{-N/2}$ (i.e. $\log Z_1/N \longrightarrow -\frac{1}{2}$). Likewise, $Z_{2\dots P}$ concentrates to $e^{N\phi(t,\alpha)}$, where $\phi(t,\alpha) = \frac{1}{N}\log\sum_{\mu=2}^{P} e^{-\frac{(x-\xi^\mu e^{-t})^2}{2\Delta_t}}$. Thus, $Z_{2\dots P}$ can be seen as the partition function of a Random Energy Model with $P-1$ random energies $\varepsilon^\mu = ((\xi^1-\xi^\mu)e^{-t}+z\sqrt{1-e^{-2t}})^2/(2\Delta_t)$. One can then compare the concentrated versions of these two terms: if $\phi(t,\alpha) < -\frac{1}{2}$ then the $Z_{2\dots P} \ll Z_1$ and the score will attract the trajectory to ξ_1; conversely, if $\phi(t,\alpha) > -\frac{1}{2}$, $Z_{2\dots P} \gg Z_1$ and the trajectories are not attracted to any data points. Hence, the collapse time t_c can be computed with $\phi(t_c(\alpha),\alpha) = -\frac{1}{2}$.

Another criterion for the collapse transition is to compare the entropy of the noisy empirical distribution

$$s(t) = -\frac{1}{N}\int dx\, P_t^e(x)\log P_t^e(x) \tag{2.100}$$

with the entropy associated to P well separated Gaussians with variance $\Delta_t I_N$

$$s^{\text{sep}}(t) = \frac{\log P}{N} + \frac{1}{2} + \frac{1}{2}\log 2\pi\Delta_t. \tag{2.101}$$

We can look at the excess entropy $f(t) = s^{\text{sep}}(t) - s(t)$. At large times, the empirical noisy distribution converges to $\mathcal{N}(0,1_d)$ whose entropy reads $\frac{1}{2}\log 2\pi + \frac{1}{2}$. The excess entropy is thus $f(t) = \frac{\log P}{N} = \alpha$ at large times. On the other hand, at small times, the noisy empirical distribution $P_t^e(x)$ becomes a mixture of well-separeted Gaussian distributions centered on the points of the dataset with variance Δ_t and the excess entropy vanishes $f(t) \longrightarrow 0$. We can also define the collapse time t_c as the time when the excess entropy $f(t)$ vanishes. Since the excess entropy thus goes from 0 to $\alpha = \log P/N$, we see that to have $t_c = \mathcal{O}(1)$, the number of samples needs to scale exponentially with the dimension $P = e^{\alpha N}$. This is a manifestation of the *curse of dimensionality*.

2.4 Data Models

One particular aspect that we would like to highlight is the role of the data structure. Indeed, statistical physics often makes use of independent Gaussian distributed data, which can provide a useful first insight but can also be seen as a limitation. One of the many questions that are driving research in the foundations of machine learning is the impact of the structure of data on learning. Correlations in data are believed to be one of

the most important features that can enhance the learning capabilities of trained models. To address this problem, we will mainly focus on structured data. In the case of manifold data, we show how the presence of a latent dimension can be beneficial and mitigate the curse of dimensionality. The presence of the lower-dimensional manifold also provides us with a tool to gain a geometric perspective on the behavior of DMs. The case of 1d Ising mixture reveals instead the generality of the scaling of the speciation time.

2.4.1 The Hidden Manifold Model

In machine learning, input data usually have a very high dimensionality; we can think, for example, of the large number of pixels that compose an image. This high dimensionality of the inputs is a feature that makes the statistical physics description suitable. However, it is commonly believed that such high-dimensional data have an actual lower intrinsic dimensionality. This idea is often expressed in terms of the *manifold hypothesis* [Bengio et al., 2013]: data live in a much lower-dimensional manifold embedded in a high-dimensional space Peyré [2009], Fefferman et al. [2016]. This is pretty reasonable, since real images only span a portion of all images that could be generated by picking pixels at random, as represented in Fig. 2.7. This hypothesis has guided the development of modern high-dimensional data modeling techniques like GANs and VAEs, as well as dimensionality reduction methods such as PCA and t-SNE. These approaches require knowledge of the data's intrinsic dimension, a critical hyperparameter. In Ref. Goldt et al. [2020], the authors introduce a simple synthetic generative process displaying the idea of the data manifold hypothesis, where data lie on a D-dimensional submanifold of the ambient N-dimensional space, and named it the Hidden Manifold Model (HMM). This generative process has also been investigated in Goldt et al. [2022], Gerace et al. [2020]. According to the HMM, data points $\{\xi^\mu \in \mathbb{R}^N\}_{\mu=1}^P$ are generated as

$$\xi^\mu = g\left(\frac{1}{\sqrt{D}} F z^\mu\right) \tag{2.102}$$

where the latent variables z^μ are Gaussian, $z^\mu \sim \mathcal{N}(0, I_D)$, g is an element-wise non-linearity, and $F \in \mathbb{R}^{N\times D}$ is a projection matrix, that can be taken fixed or random, e.g. with i.i.d. standard Gaussian entries. This projection can also be interpreted as the action of a fully connected neural network with a generic weight matrix F. In this view, we can see that the components of the features ξ^μ acquire non-linear correlations, and furthermore, the embedding dimension of the input N is separated from the intrinsic dimension D.

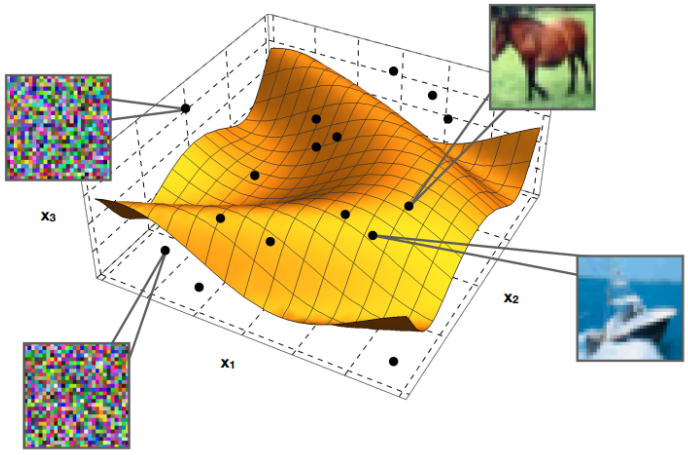


Figure 2.7: Visualization of the manifold hypothesis in the image space from Goldt et al. [2020].

This model can also be regarded as an instance of *Random Feature Learning* where data is randomly projected into a different space before being classified with a linear machine.

We are interested in studying the asymptotic regime where $D, N \to +\infty$, while their ratio $\alpha_D = D/N$ stays finite. One important advantage of this data model is that even though the data are non-Gaussian, in the high-dimensional limit, it can be shown that the joint distribution of outputs on a given pattern, given the set of weights, is asymptotically Gaussian. This result goes under the name of *Gaussian Equivalence Principle* (GEP), and was elevated to the status of theorem in Goldt et al. [2022] under a set of hypotheses.

The HMM will be implied in chapters 3 and 4 of this thesis. We regard it primarily as a data-generating model for structured data, ignoring the connections with learning described above.

2.4.2 Estimating the manifold dimension

Having introduced the manifold hypothesis, one could ask if there have been attempts at estimating the intrinsic dimension of real datasets, and how this task is performed. It goes without saying that the latent manifold is not unequivocally defined. Thinking, for example, of a simple natural image dataset as MNIST, it is clear that images of the digit '1' are quite different from images of '0', so should we consider that they come from different manifolds? Should these manifolds have different latent dimensionality? Or can we model the whole dataset as living in the same low-dimensional space? Here we will try to summarize how these questions have been addressed in the literature.

	Ground Truth	Ours	MLE (m=5)	MLE (m=20)	Local PCA	PPCA
Euclidean Data Manifolds						
10-sphere	10	11	9.61	9.46	11	11
50-sphere	50	51	35.52	34.04	51	51
Spaghetti line	1	1	1.01	1.00	32	98
Image Manifolds						
Squares						
$k = 10$	10	11	8.48	8.17	10	10
$k = 20$	20	22	14.96	14.36	20	20
$k = 100$	100	100	37.69	34.42	78	99
Gaussian blobs						
$k = 10$	10	12	8.88	8.67	10	136
$k = 20$	20	21	16.34	15.75	20	264
$k = 100$	100	98	39.66	35.31	18	985
MNIST	N/A	152	14.12	13.27	38	706

Figure 2.8: Comparison of dimensionality detection methods on various data manifolds from Stanczuk et al. [2024].

Among the traditional methods we can enumerate two main lines: the first one uses Maximum Likelihood Estimation (MLE) Levina and Bickel [2005], and has been recently applied by Pope et al. [2021] in the estimation of the intrinsic dimensionality of datasets such as MNIST, CIFAR and ImageNet, the second one uses local Principal Component Analysis (PCA), or related approaches like probabilistic PCA (PPCA) Minka [2001] or local PCA Fan et al. [2010]. Traditional estimators of intrinsic dimension typically rely on pairwise distances and nearest neighbors, so computing them is prohibitively expensive for large datasets.

A more recent direction, introduced by Stanczuk et al. [2024], leverages trained diffusion models trained to estimate the manifold dimension. The basic idea behind this method is that since diffusion models perform score-matching, they contain information about the gradient of the log-density of the data distribution, and that near the data manifold, such gradient is orthogonal to the manifold itself. Consider a point on the manifold x_0 evolved according to the forward process for a small time t_0 to x_{t_0}. The score vector $s(x_{t_0}, t_0)$ will be almost entirely contained in the Normal Bundle (NB) (the space normal to the manifold at x_{t_0}) and so the rank of a matrix containing score vectors evaluated at K points diffused from x_0 should not exceed the dimension of the NB. They sample $K = 4N$ diffused points at time $t_0 = \varepsilon$, collect them in a matrix S, and calculate the SVD of S, finally estimating the intrinsic dimension as the number of vanishing singular values. The comparison of the estimate of the manifold dimension given by the NB method with respect to the others introduced before is reported in the table in Fig. 2.8.

Another similar estimator that leverages diffusion models has been proposed in Kadkhodaie et al. [2024]: the FLIPD estimator (an acronym rearranged for Fokker-Planck (FP) and Local Intrinsic Dimensionality (LID)). Both NB and FLIPD methods have been shown to correspond closely with the complexity of an image.

2.4.3 Ising model with random field

1d Ising Model

Here we briefly summarize the main points of the analysis of a one-dimensional Ising model, which will be useful when we treat its extension to random fields in the next section. The one-dimensional (1d) Ising model is a classical spin system consisting of N binary spins $\sigma_i \in \{-1, +1\}$ arranged on a linear chain with nearest-neighbor interactions. The Hamiltonian is given by

$$H(\boldsymbol{\sigma}) = -J \sum_{i=1}^{N} \sigma_i \sigma_{i+1} - h \sum_{i=1}^{N} \sigma_i, \tag{2.103}$$

where $J \in \mathbb{R}^+$ is the ferromagnetic coupling constant and $h \in \mathbb{R}$ is an external magnetic field. Periodic boundary conditions $\sigma_{N+1} = \sigma_1$ are often assumed for convenience.
The 1d Ising model can be solved exactly using the transfer matrix method. In zero magnetic field ($h = 0$), the partition function is

$$Z_N = \mathrm{Tr}(T^N), \tag{2.104}$$

where the 2×2 transfer matrix is

$$T = \begin{pmatrix} e^{\beta J} & e^{-\beta J} \\ e^{-\beta J} & e^{\beta J} \end{pmatrix}. \tag{2.105}$$

Its eigenvalues are

$$\lambda_{\pm} = e^{\beta J} \pm e^{-\beta J}, \tag{2.106}$$

so the free energy per spin in the thermodynamic limit is

$$f = -\frac{1}{\beta} \lim_{N \to \infty} \frac{1}{N} \ln Z_N = -\frac{1}{\beta} \ln \left(e^{\beta J} + e^{-\beta J} \right). \tag{2.107}$$

Unlike in higher dimensions, the 1D Ising model does not exhibit a phase transition at finite temperature. The correlation length is finite for any $T > 0$ and diverges only as $T \to 0$. The two-point correlation function decays exponentially:

$$\langle \sigma_i \sigma_{i+r} \rangle \sim e^{-r/\xi}, \tag{2.108}$$

with correlation length

$$\xi = [\ln \coth(\beta J)]^{-1} . \tag{2.109}$$

At zero temperature ($T = 0$), the system becomes fully ordered, with all spins aligned in one of the two ground states.

We present here the replica method applied to an Ising model with random field as it was derived in Weigt and Monasson [1996]. The authors consider the Hamiltonian

$$H = -\sum_{i=1}^{N} (J_i s_i s_{i+1} + h_i s_i) \tag{2.110}$$

with Ising spins $s_i = \pm 1$ and periodic boundary conditions. The coupling constants and the external fields are randomly drawn with the probability distribution $\prod_{i=1}^{N} \mathcal{P}(J_i) p(h_i)$. One can write the 2×2-transfer matrices

$$T_1(J_i, h_i) = \begin{pmatrix} e^{+\beta J_i + \beta h_i} & e^{-\beta J_i + \beta h_i} \\ e^{-\beta J_i - \beta h_i} & e^{+\beta J_i - \beta h_i} \end{pmatrix} \tag{2.111}$$

but in general, they will not be simultaneously diagonalizable. The calculation of the thermodynamic properties, which requires the knowledge of the asymptotic properties of the product $\prod_i T_1(J_i, h_i)$ is rather involved but has been achieved for some particular choices of disorder distribution, using Dyson's method[2]

The authors decided to follow a different route, noticing that, since the disorder is independently distributed from site to site, the free energy can be computed through the knowledge of the replicated and disorder-averaged transfer matrix $T_n = \langle\langle T_1(J,h)^{\otimes n} \rangle\rangle$, where $\langle\langle \cdot \rangle\rangle$ denotes $\int dJdh \cdot \mathcal{P}(J)p(h)(.)$. This $2^n \times 2^n$ matrix is determined by the replicated spins $\{s^a = \pm 1, a = 1, \ldots, n\}$. Then, introduce the 2^n vectors $|a_1, a_2, \ldots, a_\rho\rangle = \bigotimes_a |S^a\rangle$, $a_i \neq a_j \; \forall i \neq j$, with up spins $s^a = +1$ at and only at the sites $a \in \{a_1, \ldots, a_\rho\}$. They constitute an orthonormal basis of the underlying replicated space V_n. The transfer matrix elements are now given by

$$\langle a_1, \ldots, a_\rho | T_n | b_1, \ldots, b_\sigma \rangle = \langle\langle \exp\left(\beta J \sum_{a=1}^{n} s^a s^a + \beta h \sum_{a=1}^{n} r^a \right) \rangle\rangle \tag{2.112}$$

[2]Freeman J. Dyson (1923-2020) was a British theoretical physicist and mathematician whose work spanned quantum electrodynamics, statistical mechanics, solid-state physics, and mathematics. He introduced what is now known as Dyson's method in his 1953 study of disordered one-dimensional spin systems Dyson [1953].

with $\{r^a\}$, $\{s^b\}$ corresponding to $|a_1,\ldots,a_\rho\rangle$, $|b_1,\ldots,b_\sigma\rangle$. These vectors are invariant under replica renumbering, i.e., under all transformations of the permutation group $\mathcal{S}_n$ given by its 2^n-dimensional representation $D(\pi)|a_1,\ldots,a_\rho\rangle = |\pi(a_1),\ldots,\pi(a_\rho)\rangle$, $\forall\pi \in \mathcal{S}_n$. Every irreducible decomposition of this representation D is isomorphic to the most general eigenspace decomposition of V_n. The crucial point in this method is to find such a representation in order to diagonalize the transfer matrix. The number of up spins in the basis vectors of V_n remains invariant under replica permutations. So we find $n+1$ subrepresentations Δ_ρ which are carried by the spaces spanned by $\{|a_1,\ldots,a_\rho\rangle, 1 \le a_1 < \ldots < a_\rho \le n\}$, $\rho = 0,\ldots,n$. But for $\rho \neq 0, n$ these Δ_ρ are still reducible. Using in addition the symmetry $\rho \mapsto n-\rho$, $|\pm\rangle \mapsto |\mp\rangle$, we obtain the complete decomposition of D: $\Delta_0 \cong D_0$, $\Delta_1 \cong D_0 \oplus D_1$, ..., $\Delta_\rho \cong D_0 \oplus D_1 \oplus \ldots \oplus D_{\min(\rho,n-\rho)}$, ..., $\Delta_n \cong D_0$, whose irreducibility has been proven in the book Wigner [1959]. By changing the basis of V_n with respect to this decomposition and taking at first the vectors of all D_0-spaces, then those of all D_1-spaces, and so on, we can block-diagonalize the transfer matrix T_n.

Then we can move to the replica calculation. Replica symmetry (RS) corresponds to the restriction to the first block since all D_0-vectors are invariant under permutations. Due to the representation structure, the RS-transfer matrix has $n+1$ non-degenerate eigenvalues and reads

$$T_n^{(0)}(\sigma,\tau) = \sum_{\mu=\mu_-}^{\mu_+} \binom{\sigma}{\mu}\binom{n-\sigma}{\tau-\mu} \langle\langle \exp\{\beta J(n+4\mu-2\tau-2\sigma)+\beta h(2\sigma-n)\}\rangle\rangle \quad (2.113)$$

where $\mu_- = \max(0,\sigma+\tau-n)$ and $\mu_+ = \min(\sigma,\tau)$ and the indices σ,τ run from 0 to n. The RS site-dependent partition function is given by the iterative description $Z_{i+1}(\tau) = \sum_{\sigma=0}^n T_n^{(0)}(\sigma,\tau) Z_i(\sigma)$. Introducing the generating function $Z_i[x] = \sum_\sigma Z_i(\sigma) x^\sigma$, the latter reads

$$Z_{i+1}[x] = \int_0^\infty dy \langle\langle e^{-\beta h n}(e^{\beta J} + x e^{-\beta J})^n \delta\left(y - f(x)\right)\rangle\rangle Z_i[y] \quad (2.114)$$

where

$$f(x) = e^{2\beta h}\frac{e^{-\beta J} + x e^{+\beta J}}{e^{+\beta J} + x e^{-\beta J}} \; . \quad (2.115)$$

In the thermodynamic limit, we call $\Phi(x)$ the right eigenfunction of $T_{n\to 0}^{(0)}$ which has the (maximal) eigenvalue unitary and an integral normalized to one,

$$\Phi(x) = \int_0^\infty dy \, \langle\langle \delta(x - f(y))\rangle\rangle \Phi(y) \; . \quad (2.116)$$

The free energy density is given by the O(n)-corrections in Eq. (2.114),

$$f = \langle\langle h \rangle\rangle - \frac{1}{\beta} \int_0^\infty dx \, \langle\langle \log(e^{\beta J} + x e^{-\beta J}) \rangle\rangle \Phi(x) \tag{2.117}$$

To be sure that replica symmetry is not violated, we have to check that the eigenvalue unity is not degenerate, i.e., reached in another eigenspace of $T_{n\to 0}$. In the following, we shall therefore analyze the transfer matrix blocks corresponding to the non-trivial representations D_ρ, $\rho \geq 1$. Each has $n + 1 - 2\rho$ different eigenvalues of degeneracy $\binom{n}{\rho} - \binom{n}{\rho-1}$. By ordering the basis vectors according to their permutation properties, one can achieve a further block-diagonalization of these blocks into $\binom{n}{\rho} - \binom{n}{\rho-1}$ identical blocks of size $(n + 1 - 2\rho) \times (n + 1 - 2\rho)$ each one containing every eigenvalue of the D_ρ-block exactly once. The transfer matrix blocks read

$$T_n^{(\rho)}(\sigma, \tau) = \langle\langle (2 \sinh 2\beta J)^\rho \, T_{n-2\rho}^{(0)}(\sigma - \rho, \tau - \rho; J, h) \rangle\rangle \, , \tag{2.118}$$

$\rho \leq \sigma, \tau \leq n - \rho$, where $T_n^{(0)}(\sigma, \tau; J, h)$ is given by Eq. (2.113) without averaging over the quenched disorder. In the limit $n \to 0$ we obtain for every positive number ρ a different eigenvalue equation

$$\lambda^{(\rho)} \; \Phi^{(\rho)}(x) = \int_0^\infty dy \langle\langle \delta(x - f(y)) \; (f'(y))^\rho \rangle\rangle \; \Phi^{(\rho)}(y) \tag{2.119}$$

where we again used the function $f(x)$ defined in Eq. (2.115). For every eigenfunction $\Phi^{(1)}(x)$ of $T_{n\to 0}^{(1)}$ the function $\frac{d}{dx}\Phi^{(1)}(x)$ is an eigenfunction of the RS transfer matrix with the same eigenvalue. Only the largest eigenvalue (equal to one) of $T_{n\to 0}^{(0)}$, corresponding to the density Eq. (2.116), cannot be reached by this procedure. Therefore the largest eigenvalue of $T_{n\to 0}^{(1)}$ equals the second of $T_{n\to 0}^{(0)}$ and so on.

Chapter 3

Generative Diffusion under the Manifold Hypothesis

What is presented in this chapter is based on the paper Achilli et al. [2025]. Here, we provide a detailed theoretical analysis of generative diffusion models when the data are sampled from a low-dimensional, possibly nonlinear, manifold using random energy and replica techniques.

When using the empirical score function as an approximation of the true one, we highlight the presence of two dynamical phase transitions when simulating the reverse process with time t going from $+\infty$ to 0.

1. The first one, at time t_o, is called the onset transition. It is when basins of attraction arise in correspondence to most data points, but they are not large enough to affect typical trajectories.

2. The second, at time $t_c < t_o$, is called collapse transition Biroli et al. [2024]. It corresponds to typical diffused particles being trapped in the potential well of one of the data points, with no chance of escaping it for the rest of the evolution. These last results are consistent with the recent analysis of Ref. George et al. [2025].

For generically distributed data points, we show that the collapse transition corresponds to the condensation transition in the REM. Moreover, we show that for $t > t_c$ the empirical score is close to the true score.

Finally, in Section 3.3 we analyze the problem of generalization in DMs driven by the empirical score, using two approaches. We first compute the optimal stopping time t_g, which is the time at which the KL divergence between the diffused empirical distribution and the target distribution is minimal. We use the REM formalism again to compute

this stopping time t_g. We find that it is always located in the condensed phase, i.e., $t_g < t_c$, a phenomenon that has been observed recently in the related framework of kernel approximations to large-dimensional densities Biroli and Mézard [2024]. The optimal time t_g is found to shrink to zero values faster than t_c when the latent dimension of the data decreases. In a second approach, we combine results obtained via REM formalism with random matrix computations (as performed in Ventura et al. [2025]), in order to deduce an empirical generalization criterion for DMs sampling before memorization. In the whole chapter, we analyze diffusion models driven by the empirical score function. The main objective is to understand how data structure can impact memorization and generalization; thus, we adopt the Hidden Manifold Model, described in Sec. 2.4.1. The diffusion paradigm we adopt is the variance-exploding one.

3.1 The Random Energy Model formalism

In order to compute the main quantities that characterize Diffusion Models (DMs), we introduce the tools needed to solve a generic REM, following Lucibello and Mézard [2024].

Let us consider $P = e^{\alpha N}$ (or equivalently $P = e^{\alpha N} - 1$) i.i.d. energy levels $\varepsilon^\mu \sim p(\varepsilon \,|\, \omega)$, where we extend the typical REM setting allowing for a common source of quenched disorder $\omega \sim p_\omega$. The goal is to compute the average asymptotic free energy of the system, defined by

$$\phi_\alpha(\lambda) = \lim_{N\to\infty} \frac{1}{\lambda N} \mathbb{E} \log \sum_\mu e^{\lambda N \varepsilon^\mu} \tag{3.1}$$

We shall assume that the probability distribution of the energy levels is such that, with probability one over the choice of ω when $N \to \infty$, the cumulant generating function has a well-defined limit: $\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}_{\varepsilon\,|\omega}\, e^{\lambda N \varepsilon}$ exists, and the distribution over the choices of ω concentrates around its mean. Then we define the typical cumulant generating function and its Legendre transform:

$$\zeta(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\omega \log \mathbb{E}_{\varepsilon\,|\omega}\, e^{\lambda N \varepsilon}, \tag{3.2}$$

$$s(\varepsilon) = \sup_\lambda \; \varepsilon\, \lambda - \zeta(\lambda). \tag{3.3}$$

The total entropy of the system is $\Sigma(\epsilon) = \alpha - s(\varepsilon)$. Depending on the value of $\Sigma(\epsilon)$, the REM displays a separation into two thermodynamic phases: an *uncondensed* phase where the system can *populate* an exponential number of energy levels, at lower values

of λ; a *condensed* phase where the system is able to populate a unique energy state, at higher values of λ.

Let us define the quantities $\varepsilon_*(\alpha)$ and $\lambda_*(\alpha)$ respectively as the maximum value of the energy levels in the uncondensed phase, obtained as the largest root of $\Sigma(\varepsilon_*) = 0$, and the condensation threshold. Notice that we are seeking the maximum energy, by definition of the free-energy function in Eq. (3.1). In the uncondensed phase, i.e. when $\lambda < \lambda_*(\alpha)$, the dominating energy level $\tilde{\varepsilon}(\lambda)$ is obtained as the stationary point of $\lambda\,\varepsilon - s(\varepsilon)$, and by the Legendre transform definition of $\zeta(\lambda)$ this is equivalent to $\tilde{\varepsilon}(\lambda) = \zeta'(\lambda)$. The entropy of the dominating state can be rewritten as $\Sigma(\tilde{\varepsilon}(\lambda)) = \alpha - s(\tilde{\varepsilon}(\lambda)) = \alpha + \zeta(\lambda) - \lambda\zeta'(\lambda)$, so the condensation threshold $\lambda_*(\alpha)$ is obtained from the condensation condition

$$\alpha + \zeta(\lambda_*) - \lambda_*\zeta'(\lambda_*) = 0. \tag{3.4}$$

Finally, the free energy is given by

$$\phi_\alpha(\lambda) = \begin{cases} \frac{\alpha+\zeta(\lambda)}{\lambda} & \lambda < \lambda_*(\alpha), \\ \varepsilon_*(\alpha) & \lambda \geq \lambda_*(\alpha). \end{cases} \tag{3.5}$$

3.2 Memorization in Generative Diffusion

We here analyze the memorization phenomenology in generative diffusion when the model is trained on structured data. Other works in the literature that employ similar data models, such as Boffi et al. [2025], while studies contained in Leigh Ross et al. [2025], Chen et al. [2023], Pidstrigach [2022], Stanczuk et al. [2024] study the effect of a latent data dimensionality on generative diffusion. Let us recall the expression of the empirical density at time t in the variance exploding formalism

$$p_t^{emp}(x) = \frac{1}{P\sqrt{2\pi t}^N} \sum_{\mu=1}^{P} e^{-\frac{1}{2t}\|x-\xi^\mu\|^2}. \tag{3.6}$$

We will hereby use three expressions that all refer to the same dynamic process: *collapse*, *condensation*, and *memorization*. The first two idioms, which derive from the REM terminology, will be proved to coincide in this framework, due to the typicality of the stochastic trajectories involved (see Achilli et al. [2024] for a case where this equivalence does not hold); the third concept, i.e. memorization, is more widely employed in the literature and we will use it as an umbrella term for the first two. Following Biroli et al.

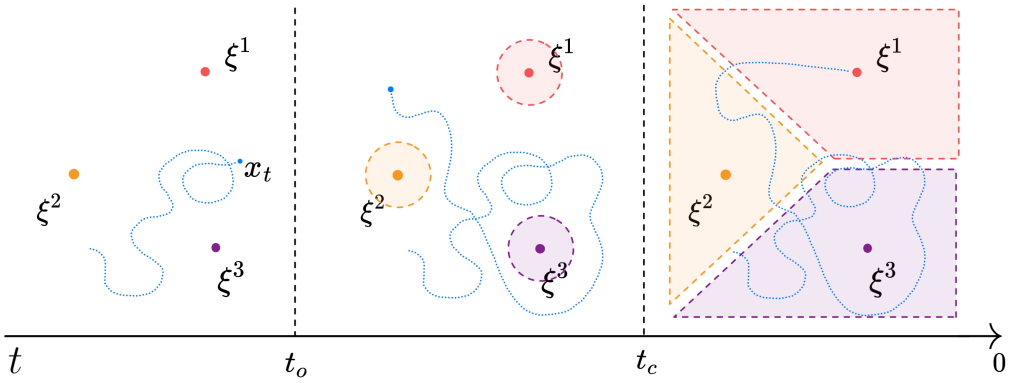


Figure 3.1: Pictorial representation of the phases identified in the reverse process (from large times to $t = 0$) driven by the empirical score. The evolution of a typical trajectory is represented by a dotted blue line. For $t < t_o$, data points form a basin around them, but the typical trajectory is not affected by this. For $t_o < t < t_c$, the basins of attraction of the data points cover the whole space, trajectories cannot escape them once inside, and eventually fall into the data points at time $t = 0$.

[2024], we are treating the attraction of the diffusive trajectories by the data points in terms of the collapse phase transition occurring in an effective REM. We find two main dynamical events occurring in time:

1. The appearance of attractors with finite basins of attraction in the diffusion at time $t = t_o$. We call this time *onset time*, and it consists of the moment when training data become attractive, yet without influencing the typical diffusive trajectory of the model. This is also the time where data typically become local maxima of the mixture of gaussian in Eq. (3.6).

2. The *collapse* of the typical diffusive trajectory on the training data points, occurring at time $t = t_c < t_o$.

Fig. 3.1 provides for a sketch of the phase separation described above.

3.2.1 Collapse Time

Here we first recap the collapse condition for diffusion models as it was introduced in Biroli et al. [2024], that we have described in Section 2.3.3, and then proceed to compute it for our data-generating model. If we start the forward diffusion process from one of the data points, e.g. ξ^1, then the typical trajectory is $x_t = \xi^1 + \omega\sqrt{t}$, with $\omega \sim \mathcal{N}(0, I_N)$. We want to see at which time t_c the term $\mu = 1$ dominates the summation in the measure,

which for our choice of x_t takes the form

$$p_t^{emp}(x) = \frac{1}{P\sqrt{2\pi t}^N}\left(e^{-\frac{\|\omega\|^2}{2}} + \sum_{\mu\geq 2} e^{-\frac{1}{2t}\|(\xi^1-\xi^\mu)+\omega\sqrt{t}\|^2}\right) \tag{3.7}$$

$$= \frac{1}{P\sqrt{2\pi t}^N}\left(Z_1 + Z_{2,\dots,P}\right). \tag{3.8}$$

In the limit of $P, N \to \infty$ with $\alpha = \frac{\log P}{N}$ fixed, we find $Z_1 \simeq e^{-N/2}$, while $\frac{1}{N}\log Z_{2,\dots,P}$ concentrates around ϕ_t, with

$$\phi_t = \lim_{N\to\infty}\frac{1}{N}\log\sum_{\mu\geq 2} e^{-\frac{1}{2t}\|(\xi^1-\xi^\mu)+\omega\sqrt{t}\|^2}. \tag{3.9}$$

One can make a signal-to-noise argument by comparing the concentrated versions of Z_1 and $Z_{2,..,P}$. This approach leads to the so-called *collapse* criterion, also used in Biroli et al. [2024], Lucibello and Mézard [2024]. This criterion requires

$$\alpha + \zeta_{t_c}(1) = -\frac{1}{2}. \tag{3.10}$$

Since now the noise in the process is played by the factor λ/t. As noticeable from Eq. (3.10), in this problem we are imposing $\lambda^* = 1$ to compute the time t_c at which collapse occurs. For given ξ^1, ϕ_t is minus the average free energy density of a REM, $\phi_t = \lim_{N\to\infty}\frac{1}{N}\log\sum_{\mu\geq 2} e^{\epsilon_\mu}$ with $P-1$ energy levels $\epsilon_\mu = -\frac{1}{2t}\|(\xi^1-\xi^\mu)+\omega\sqrt{t}\|^2$.

We then need to find the cumulant generating function for the energy levels

$$\zeta_t(\lambda) = \lim_{N\to+\infty}\frac{1}{N}\log\mathbb{E}_\epsilon e^{\lambda\epsilon} \tag{3.11}$$

$$= \lim_{N\to+\infty}\frac{1}{N}\mathbb{E}_{\xi^1,\omega}\log\mathbb{E}_{\xi^2} e^{-\frac{\lambda}{2t}\|(\xi^1-\xi^\mu)+\omega\sqrt{t}\|^2}. \tag{3.12}$$

3.2.2 Collapse Time for Homogeneous Gaussian Data

In order to investigate the scaling of t_c with respect to the control parameters, let us simplify even more the data model and assume that D dimensions have variance $\sigma_i^2 = \sigma^2$ and $N - D$ have variance $\sigma_i^2 = 0$. We have

$$\zeta_t(\lambda) = -\frac{1}{2}\alpha_D\log(1+\frac{\lambda}{t}\sigma^2) - \frac{\lambda}{2}\alpha_D\frac{t+\sigma^2}{t+\lambda\sigma^2} - \frac{\lambda}{2}(1-\alpha_D). \tag{3.13}$$

We can find the collapse time from the condition in Eq. (3.10) which implies

$$-\alpha_D \log(1 + \frac{\sigma^2}{\alpha_D t}) - 1 + 2\alpha = -1 \tag{3.14}$$

The solution is

$$t_c = \frac{\sigma^2 N/D}{e^{2\log P/D} - 1}. \tag{3.15}$$

The collapse time depends on the manifold dimension and the number of hidden points. The so-called *curse of dimensionality*, i.e. the need for a number of training data points that scales exponentially in the visible dimension of the data space Yarotsky [2017], Cybenko [1989], has been mitigated by the fact that we have an effective dimensionality for the data. Indeed, if in general one needs a number of data that scales exponentially with the system size $P = e^{\alpha N}$ in order to have a collapse time of $O(1)$, here we only need $P = e^{\beta D}$.

If we consider the limit of $D \ll \log P$ and $D \ll N$ we have

$$t_c \approx \frac{\sigma^2}{2}\frac{N}{D} e^{-\frac{2\log P}{D}}, \tag{3.16}$$

which goes to zero fast.

Variance Preserving Case

Here we report the same derivation in the variance-preserving scenario, for comparison with estimates obtained by the unstructured case in Biroli et al. [2024]: the main difference, which is conserved in the variance-exploding model, lies in the substitution of the visible dimension N with the latent one D in the exponent contained in t_c.

If indeed we consider the variance preserving framework where $x = \xi^1 e^{-t} + \omega\sqrt{\Delta_t}$, $\Delta_t = 1 - e^{-2t}$, the cumulant generating function reads

$$\zeta_t(\lambda) = -\frac{1}{2}\alpha_D \log(1 + \frac{\lambda e^{-2t}}{\Delta_t}\sigma^2) - \frac{\lambda}{2}\alpha_D \frac{\Delta_t + \sigma^2 e^{-2t}}{\Delta_t + \lambda\sigma^2 e^{-2t}} - \frac{\lambda}{2}(1 - \alpha_D) \tag{3.17}$$

The solution of Eq. (3.10) is

$$t_c = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{e^{2\alpha/\alpha_D} - 1}\right) \tag{3.18}$$

$$= \frac{1}{2}\log\left(1 + \frac{\sigma^2}{e^{\frac{2\log P}{D}} - 1}\right). \tag{3.19}$$

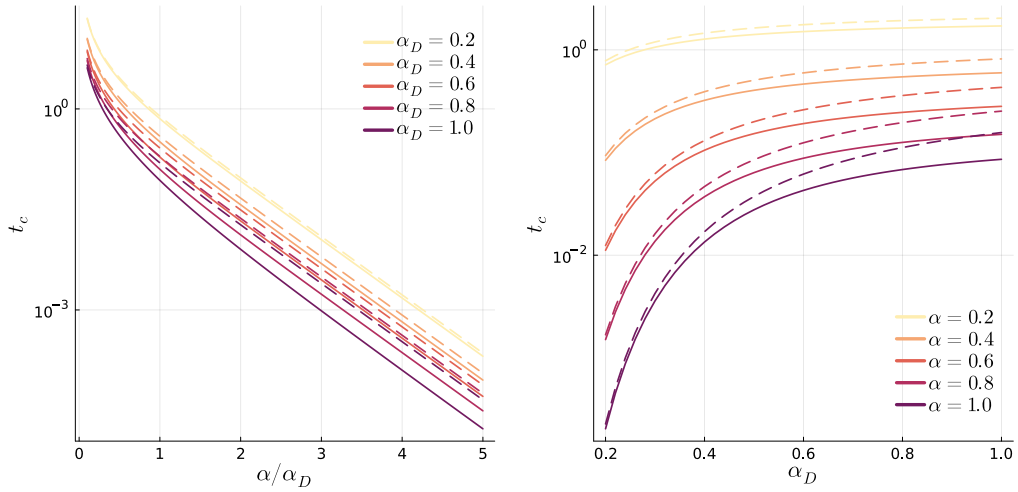


Figure 3.2: Semi-logarithmic plots of t_c in the linear manifold case (solid) compared to the homogeneous Gaussian case (dashed) for different values of α_D (**Left**) and α (**Right**).

Such expression for the collapse time has been also recently found by George et al. [2025], and it can be easily compared with the one found in Biroli et al. [2024] for $\alpha_D = 1$, i.e. the homogeneous Gaussian case.

3.2.3 Collapse for Manifold Data

If we assume that the data points come from a linear manifold, $\xi^\mu = \frac{1}{\sqrt{D}} F z^\mu$, then Eq. (3.12) becomes

$$\zeta_t(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{F,z^1,\omega} \log \mathbb{E}_{z^2} e^{-\frac{\lambda}{2t}\left\| \left(Fz^2 - Fz^1\right) + \omega\sqrt{t}\right\|^2}. \tag{3.20}$$

We derive the collapse equation for this case in Section 3.2.3, that then we solve numerically. In Fig. 3.2 we show how t_c scales with the ratio α/α_D, i.e. $\log P/D$. These curves are compared with the Gaussian expression for t_c contained in Eq. (3.15). It is straightforward to notice that the slopes of the curves are the same for $\alpha \gg \alpha_D$, meaning that even in the linear manifold case we observe the same exponential scaling with $\log P/D$ obtained for the homogeneous Gaussian scenario. Fixing α, which here corresponds to fixing the number of data points, we see that the collapse time decreases with the hidden dimensionality D. Moreover, collapse occurs earlier in the reversed process when the number of data points is smaller.

Let us now consider a non-linear manifold for the data points, i.e. $\xi^\mu = g\left(\frac{1}{\sqrt{D}} F z^\mu\right)$.

In this case Eq. (3.12) assumes the following expression

$$\zeta_t(\lambda) = \lim_{N\to+\infty} \frac{1}{N}\mathbb{E}_{z^1,F,\omega} \log \mathbb{E}_{z^2} e^{-\frac{\lambda}{2t}\|\left(g\left(\frac{1}{\sqrt{D}}Fz^1\right)-g\left(\frac{1}{\sqrt{D}}Fz^2\right)\right)+\omega\sqrt{t}\|^2}. \tag{3.21}$$

This function can be computed using the replica method, as shown in Sec. 3.2.3. We find an expression for ζ_t in the RS approximation, i.e.

$$\zeta_t(\lambda; q_d, q_0, m, \hat{q}_d, \hat{q}_0, \hat{m}) = -\alpha_D m\hat{m} - \frac{\alpha_D}{2}(q_d\hat{q}_d - q_0\hat{q}_0) + \alpha_D G_S(\hat{q}_d, \hat{q}_0, \hat{m}) + G_E(\lambda, t; q_d, q_0, m), \tag{3.22}$$

with

$$G_S(\hat{q}_d, \hat{q}_0, \hat{m}) = -\frac{1}{2}\log\left(1 - \hat{q}_d + \hat{q}_0\right) + \frac{1}{2}\frac{\hat{m}^2 + \hat{q}_0}{1 - \hat{q}_d + \hat{q}_0}, \tag{3.23}$$

and

$$G_E(\lambda, t; q_d, q_0, m) = \int D\omega \int D\gamma \int Du^0 \log\left(\int Du\; e^{-\frac{\lambda}{2t}\left(g\left(u^0\right)-g\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}\gamma\right)+\sqrt{t}\omega\right)^2}\right). \tag{3.24}$$

Then we solve the saddle point equations (which depend on the choice of the non-linearity) to obtain the typical value of ζ_t. At this point, the collapse condition is solved numerically, and the scaling of the collapse time can be compared to the one found for linear manifolds. Fig. 3.3 depicts the instance of $g(x) = \tanh(x)$. As one can notice, curves for the non-linear and linear cases show the same qualitative behavior, displaying the same type of scaling as a function fo α/α_D and α_D when the number of data are fixed.

Computation of the Generating function: Linear case

Here we report some calculations needed to obtain the generating function in the linear manifold case. All the results are summarized in Sec. 3.2.3, while this one summarizes

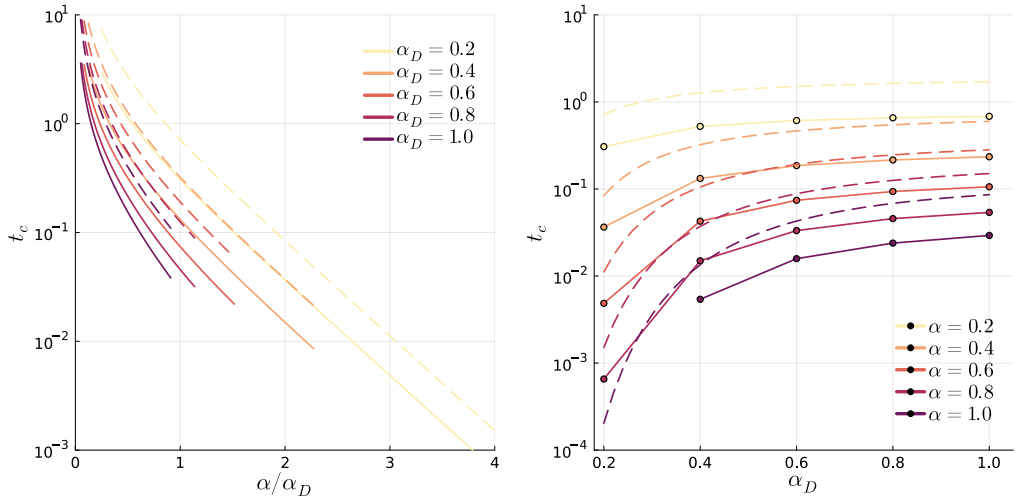


Figure 3.3: Semi-logarithmic plots of t_c in the hidden manifold case (solid) with tanh activation compared to the linear manifold case (dashed) for different values of α_D (**Left**) and α (**Right**).

the techniques. In the variance exploding case we have

$$\zeta_t(\lambda) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^1,\omega} \log \mathbb{E}_{z^2} e^{-\frac{\lambda}{2t}\|\left(\frac{F}{\sqrt{D}}z^2 - \frac{F}{\sqrt{D}}z^1\right)+\omega\sqrt{t}\|^2} \tag{3.25}$$

$$= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^1,\omega} \log \int \frac{dz^2}{2\pi}\; e^{-\frac{1}{2}z^2(I+\frac{\lambda}{t}\frac{F^TF}{D})z^2 + \frac{\lambda}{t}z^2(\frac{F^TF}{D}z^1 - \frac{F^T}{\sqrt{D}}\omega\sqrt{t}) - \frac{\lambda}{2t}\|-\frac{F}{\sqrt{D}}z^1+\omega\sqrt{t}\|^2} \tag{3.26}$$

$$\begin{aligned} &= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^1,\omega}\Bigg[-\frac{1}{2}\log\det(I+\frac{\lambda}{t}\frac{F^TF}{D}) \\ &+ \frac{1}{2}\frac{\lambda^2}{t^2}(\frac{F^TF}{D}z^1 - \frac{F^T}{\sqrt{D}}\omega\sqrt{t})^T(I+\frac{\lambda}{t}\frac{F^TF}{D})^{-1}(\frac{F^TF}{D}z^1 - \frac{F^T}{\sqrt{D}}\omega\sqrt{t}) \\ &- \frac{\lambda}{2t}\|-\frac{F}{\sqrt{D}}z^1+\omega\sqrt{t}\|^2\Bigg] \end{aligned} \tag{3.27}$$

$$\begin{aligned} &= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^1,\omega}\Bigg[-\frac{1}{2}\log\det(I+\frac{\lambda}{t}\frac{F^TF}{D}) \\ &+ \frac{\lambda^2}{2t^2}(\frac{F}{\sqrt{D}}z^1)^T\frac{F}{\sqrt{D}}(I+\frac{\lambda}{t}\frac{F^TF}{D})^{-1}\frac{F^T}{\sqrt{D}}(\frac{F}{\sqrt{D}}z^1) \\ &- \frac{\lambda}{2t}\|-\frac{F}{\sqrt{D}}z^1\|^2 + \frac{\lambda^2}{2t^2}(\frac{F^T}{\sqrt{D}}\omega\sqrt{t})^T(I+\frac{\lambda}{t}\frac{F^TF}{D})^{-1}(\frac{F^T}{\sqrt{D}}\omega\sqrt{t}) - \frac{\lambda}{2t}\|\omega\sqrt{t}\|^2\Bigg] \end{aligned} \tag{3.28}$$

Now with a rotation we can position in the basis of the eigenvectors of $\frac{F^\top F}{N}$, with eigenvalues σ_k^2

$$= \lim_{N\to\infty} \frac{1}{N}\sum_i^N -\frac{\lambda}{2} + \frac{1}{N}\sum_k^D \Bigg[-\frac{1}{2}\log(1+\frac{\lambda}{\alpha_D t}\sigma_k^2) + \frac{\lambda^2}{2\alpha_D^2 t^2}(\frac{\sigma_k^4}{1+\frac{\lambda}{\alpha_D t}\sigma_k^2})$$
$$- \frac{\lambda}{2\alpha_D t}\sigma_k^2 + \frac{\lambda^2}{2\alpha_D t^2}(\frac{t\sigma_k^2}{1+\frac{\lambda}{\alpha_D t}\sigma_k^2})\Bigg] \tag{3.29}$$
$$= -\frac{\lambda}{2} + \lim_{N\to\infty}\frac{1}{N}\sum_k \Bigg[-\frac{1}{2}\log(1+\frac{\lambda}{\alpha_D t}\sigma_k^2) + \frac{\lambda^2}{2\alpha_D t}(\frac{\sigma_k^4}{\alpha_D t+\lambda\sigma_k^2})$$
$$- \frac{\lambda}{2\alpha_D t}\sigma_k^2 + \frac{\lambda^2}{2t}(\frac{t\sigma_k^2}{\alpha_D t+\lambda\sigma_k^2})\Bigg] \tag{3.30}$$
$$= -\frac{\lambda}{2} + \lim_{N\to\infty}\frac{1}{N}\sum_k \left[-\frac{1}{2}\log(1+\frac{\lambda}{\alpha_D t}\sigma_k^2)\right] \tag{3.31}$$

Here we have assumed that $\alpha_D < 1$. Taking the limit $N \to \infty$ the sum becomes an integration over the distribution ν of σ^2, which is the bulk of a Marchenko-Pastur distribution

$$\zeta_t(\lambda) = -\frac{\lambda}{2} - \frac{\alpha_D}{2}\int \nu_{\alpha_D}(d\sigma^2)\log\left(1+\frac{\lambda\sigma^2}{\alpha_D t}\right) \tag{3.32}$$

with

$$d\nu_\gamma(x) = \frac{1}{2\pi}\frac{\sqrt{(\gamma^+-x)(\gamma^--x)}}{\gamma x}\mathbb{I}\left(x\in[\gamma^-,\gamma^+]\right) \tag{3.33}$$
$$\gamma^\pm = (1\pm\sqrt{\gamma})^2 \tag{3.34}$$

If we compute everything at $\lambda = 1$ this becomes

$$\zeta_t(1) = -\frac{1}{2}\int \nu_{\alpha_D}(d\sigma^2)\left[\log(1+\frac{\sigma^2}{\alpha_D t})\right] - \frac{1}{2} \tag{3.35}$$

Taking the derivative

$$\zeta_t'(\lambda) = -\frac{\alpha_D}{2}\int \nu_{\alpha_D}(d\sigma^2)\frac{\sigma^2}{\alpha_D t+\lambda\sigma^2} \tag{3.36}$$
$$\zeta_t'(1) = -\frac{1}{2} \tag{3.37}$$

We can also use replica theory, which will be necessary in the non-linear case, and compare the results. The replicated ζ_t reads

$$\mathbb{E}\mathcal{Z}^n = \mathbb{E}_{F,\omega}\mathbb{E}_{z^{0:n}}\ e^{-\frac{\lambda}{2t}\sum_{a'}\|\frac{Fz^0}{\sqrt{D}}-\frac{Fz^{a'}}{\sqrt{D}}\|^2-\frac{\lambda\omega}{t}\sum_{a'}\left(\frac{Fz^0}{\sqrt{D}}-\frac{Fz^{a'}}{\sqrt{D}}\right)-\frac{\lambda}{2}\|\omega\|^2} \tag{3.38}$$

$$= \mathbb{E}_{F,\omega}\mathbb{E}_{z^{0:n}}\ e^{-\frac{\lambda}{2t}\sum_{a'}\|\frac{Fz^0}{\sqrt{D}}-\frac{Fz^{a'}}{\sqrt{D}}\|^2-\frac{\lambda\omega}{t}\sum_{a'}\left(\frac{Fz^0}{\sqrt{D}}-\frac{Fz^{a'}}{\sqrt{D}}\right)-\frac{\lambda}{2}\|\omega\|^2} \tag{3.39}$$

$$= \mathbb{E}_{F,\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_{i,a'}\left(u_i^0-u_i^{a'}\right)^2-\frac{\lambda}{2}\sum_{i,a'}\omega_i(u_i^0-u_i^{a'})-\frac{\lambda}{2}\sum_i\omega_i^2}$$
$$\times\ e^{-i\sum_{i,a=0}^n\hat{u}_i^a u_i^a+\sum_a\frac{i}{\sqrt{D}}\sum_{ik}\hat{u}_i^a F_{ik}z_k^a} \tag{3.40}$$

$$= \mathbb{E}_{\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_{i,a'}\left(u_i^0-u_i^{a'}\right)^2-\frac{\lambda}{2}\sum_{i,a'}\omega_i(u_i^0-u_i^{a'})-\frac{\lambda}{2}\sum_i\omega_i^2}$$
$$\times\ e^{-i\sum_{i,a=0}^n\hat{u}_i^a u_i^a-\frac{1}{2D}\sum_{ik}\left(\sum_a\hat{u}_i^a z_k^a\right)^2} \tag{3.41}$$

$$= \mathbb{E}_{\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_{i,a'}\left(u_i^0-u_i^{a'}\right)^2-\frac{\lambda}{2}\sum_{i,a'}\omega_i(u_i^0-u_i^{a'})-\frac{\lambda}{2}\sum_i\omega_i^2}$$
$$\times\ e^{-i\sum_{i,a=0}^n\hat{u}_i^a u_i^a-\frac{1}{2D}\sum_{ab}\left(\sum_i\hat{u}_i^a\hat{u}_i^b\right)\left(\sum_k z_k^a z_k^b\right)} \tag{3.42}$$

$$= \int dq\, d\hat{q}\ e^{nN\phi_\lambda(q,\hat{q})} \tag{3.43}$$

with the overlaps defined as

$$q_{ab} = \frac{1}{D}\sum_k z_k^a z_k^b \tag{3.44}$$

so that we can write the replicated action

$$\zeta_t(\lambda,t;q,\hat{q}) = -\frac{1}{2n}\frac{D}{N}\sum_{ab=0}^n q_{ab}\hat{q}_{ab} + \frac{D}{N}G_S(\hat{q}) + G_E(\lambda,t;q) \tag{3.45}$$

with

$$G_S(\hat{q}) = \frac{1}{n}\log\mathbb{E}_{z^{0:n}}\ e^{\frac{1}{2}\sum_{ab}\hat{q}_{ab}z^a z^b} \tag{3.46}$$

$$G_E(\lambda,t;q) = \frac{1}{n}\log\int D\omega\int\prod_{a=0}^n\frac{d\hat{u}_a du_a}{2\pi}e^{-\frac{\lambda}{2t}\sum_{a'}(u^0-u^{a'}+\omega\sqrt{t})^2-i\sum_{a=0}^n\hat{u}^a u^a-\frac{1}{2}\sum_{ab}\hat{u}^a\hat{u}^b q_{ab}} \tag{3.47}$$

Using the replica symmetric ansatz

$$q_{ab} = \begin{pmatrix} 1 & m & \dots & m \\ m & q_d & & q_0 \\ \vdots & & \ddots & \\ m & q_0 & & q_d \end{pmatrix}; \quad \hat{q}_{ab} = \begin{pmatrix} 0 & \hat{m} & \dots & \hat{m} \\ \hat{m} & \hat{q}_d & & \hat{q}_0 \\ \vdots & & \ddots & \\ \hat{m} & \hat{q}_0 & & \hat{q}_d \end{pmatrix} \tag{3.48}$$

we find

$$\zeta_t(\lambda; q_d, q_0, m, \hat{q}_d, \hat{q}_0, \hat{m}) = -\alpha_D m\hat{m} - \frac{\alpha_D}{2}(q_d\hat{q}_d - q_0\hat{q}_0) + \alpha_D G_S(\hat{q}_d, \hat{q}_0, \hat{m}) + G_E(\lambda, t; q_d, q_0, m) \tag{3.49}$$

with

$$G_S(\hat{q}_d, \hat{q}_0, \hat{m}) = -\frac{1}{2}\log\left(1 - \hat{q}_d + \hat{q}_0\right) + \frac{1}{2}\frac{\hat{m}^2 + \hat{q}_0}{1 - \hat{q}_d + \hat{q}_0} \tag{3.50}$$

and for the energetic term

$$G_E = \int D\omega \int D\gamma \int Du^0 \log \int Du\ e^{-\frac{\lambda}{2t}\left(u^0 - \sqrt{q_d - q_0}u - mu^0 + \sqrt{q_0 - m^2}\gamma\right)^2} \tag{3.51}$$

$$\times\, e^{-\frac{\lambda\omega}{\sqrt{t}}\left(u^0 - \sqrt{q_d - q_0}u - mu^0 + \sqrt{q_0 - m^2}\gamma\right) - \frac{\lambda}{2}\omega^2} \tag{3.52}$$

$$\begin{aligned} &= \int D\omega \int D\gamma \int Du^0 \Big[-\frac{\lambda}{2t}\left((1-m)u^0 + \sqrt{q_0 - m^2}\gamma\right)^2 \\ &- \frac{\lambda\omega}{\sqrt{t}}\left((1-m)u^0 + \sqrt{q_0 - m^2}\gamma\right) - \frac{\lambda}{2}\omega^2 - \frac{1}{2}\log\left(1 + \frac{\lambda(q_d - q_0)}{t}\right) \\ &+ \frac{1}{2}\left(1 + \frac{\lambda(q_d - q_0)}{t}\right)^{-1}\left(\frac{\lambda}{t}\sqrt{q_d - q_0}\left((1-m)u^0 + \sqrt{q_0 - m^2}\gamma\right) + \frac{\lambda\omega}{t}\sqrt{q_d - q_0}\right)^2 \Big] \end{aligned} \tag{3.53}$$

$$= -\frac{1}{2}\left(\lambda + \log\left(1 + \frac{\lambda(q_d - q_0)}{t}\right) + \frac{\lambda}{t}(1 - 2m + q_0) - \frac{\lambda^2(q_d - q_0)(1 - 2m + q_0 + t)}{t^2\left(1 + \frac{\lambda(q_d - q_0)}{t}\right)}\right) \tag{3.54}$$

In order to compute the typical value of $\zeta_t(\lambda)$ to employ for solving the collapse condition,

we need to derive the following saddle point equations

$$\frac{\partial \zeta_t}{\partial \hat{q}_d} = 0 \qquad q_d = \frac{1}{1-\hat{q}_d+\hat{q}_0} + \frac{\hat{m}^2+\hat{q}_0}{\left(1-\hat{q}_d+\hat{q}_0\right)^2} \tag{3.55}$$

$$\frac{\partial \zeta_t}{\partial \hat{q}_0} = 0 \qquad q_0 = \frac{1}{1-\hat{q}_d+\hat{q}_0} + \frac{\hat{m}^2+\hat{q}_d-1}{\left(1-\hat{q}_d+\hat{q}_0\right)^2} \tag{3.56}$$

$$\frac{\partial \zeta_t}{\partial \hat{m}} = 0 \qquad m = \frac{\hat{m}}{1-\hat{q}_d+\hat{q}_0} \tag{3.57}$$

$$\frac{\partial \zeta_t}{\partial q_d} = 0 \qquad \hat{q}_d = \frac{2}{\alpha_D}\left(-\frac{1}{2}\right)\frac{\lambda\left(t+(-1-2m+q_0+t)\lambda\right)}{\left(t+(q-q_0)\lambda\right)^2} \tag{3.58}$$

$$\frac{\partial \zeta_t}{\partial q_0} = 0 \qquad \hat{q}_0 = -\frac{2}{\alpha_D}\left(-\frac{1}{2}\right)\frac{(1-2m+q_0+t)\lambda^2}{\left(t+(q-q_0)\lambda\right)^2} \tag{3.59}$$

$$\frac{\partial \zeta_t}{\partial m} = 0 \qquad \hat{m} = \frac{1}{\alpha_D}\left(-\frac{1}{2}\right)\left(-\frac{2\lambda}{t}+\frac{2(q-q_0)\lambda^2}{t^2\left(1+\frac{(q-q_0)\lambda}{t}\right)}\right). \tag{3.60}$$

By solving these saddle point equations we recover perfect agreement with Eq. (3.35).

Computation of the Generating function: Non-linear case

In the non-linear case, the replicated partition function reads

$$\mathbb{E}\mathcal{Z}^n = \mathbb{E}_{F,\omega}\mathbb{E}_{z^{0:n}}\; e^{-\frac{\lambda}{2t}\sum_{a'}\|g\left(\frac{Fz^0}{\sqrt{D}}\right)-g\left(\frac{Fz^{a'}}{\sqrt{D}}\right)+\omega\sqrt{t}\|^2} \tag{3.61}$$

$$= \mathbb{E}_{F,\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_i\sum_{a'}\left(g\left(u_i^0\right)-g\left(u_i^{a'}\right)\right)+\omega_i\sqrt{t})^2}e^{-i\sum_i\sum_{a=0}^n\hat{u}_i^a u_i^a+\sum_a\frac{i}{\sqrt{D}}\sum_{ik}\hat{u}_i^a F_{ik}z_k^a} \tag{3.62}$$

$$= \mathbb{E}_{\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_i\sum_{a'}\left(g\left(u_i^0\right)-g\left(u_i^{a'}\right)+\omega_i\sqrt{t}\right)^2}e^{-i\sum_i\sum_{a=0}^n\hat{u}_i^a u_i^a-\frac{1}{2D}\sum_{ik}\left(\sum_a\hat{u}_i^a z_k^a\right)^2} \tag{3.63}$$

$$= \mathbb{E}_{\omega}\mathbb{E}_{z^{0:n}}\int\frac{d\hat{u}du}{2\pi}e^{-\frac{\lambda}{2t}\sum_i\sum_{a'}\left(g\left(u_i^0\right)-g\left(u_i^{a'}\right)+\omega_i\sqrt{t}\right)^2}e^{-i\sum_i\sum_{a=0}^n\hat{u}_i^a u_i^a-\frac{1}{2D}\sum_{ab}\left(\sum_i\hat{u}_i^a\hat{u}_i^b\right)\left(\sum_k z_k^a z_k^b\right)} \tag{3.64}$$

$$= \int dq\, d\hat{q}\; e^{nN\phi_\lambda(q,\hat{q})} \tag{3.65}$$

with the overlaps defined as

$$q_{ab} = \frac{1}{D}\sum_k z_k^a z_k^b \tag{3.66}$$

so that we can write the replicated action

$$\zeta_t(q,\hat{q}) = -\frac{1}{2n}\frac{D}{N}\sum_{ab=0}^{n} q_{ab}\hat{q}_{ab} + \frac{D}{N}G_S(\hat{q}) + G_E(q) \tag{3.67}$$

with

$$G_S = \frac{1}{n}\log \mathbb{E}_{z^{0:n}}\ e^{\frac{1}{2}\sum_{ab}\hat{q}_{ab}z^a z^b} \tag{3.68}$$

$$G_E = \frac{1}{n}\log \int D\omega \int \prod_{a=0}^{n}\frac{d\hat{u}_a du_a}{2\pi} e^{-\frac{\lambda}{2t}\sum_{a'}(g\left(u^0\right)-g\left(u^{a'}\right)+\omega\sqrt{t})^2 - i\sum_{a=0}^{n}\hat{u}^a u^a - \frac{1}{2}\sum_{ab}\hat{u}^a\hat{u}^b q_{ab}} \tag{3.69}$$

We invoke the replica symmetric ansatz as performed in the linear case (see Sec 3.2.3) and obtain the same expression for $\zeta_t(q_d, q_0, m, \hat{q}_d, \hat{q}_0, \hat{m})$ with a different energetic term, due to the non-linearity, that reads

$$G_E = \frac{1}{n}\log \int D\omega \int \prod_{a=0}^{n}\frac{d\hat{u}_a du_a}{2\pi} e^{-\frac{\lambda}{2t}\sum_{a'}(g\left(u^0\right)-g\left(u^{a'}\right)+\omega\sqrt{t})^2 - i\sum_{a=0}^{n}\hat{u}^a u^a - \frac{1}{2}\sum_{ab}\hat{u}^a\hat{u}^b q_{ab}} \tag{3.70}$$

$$= \frac{1}{n}\log \int D\omega \int \frac{du^0 d\hat{u}^0}{2\pi}\prod_{a'=1}^{n}\frac{du^{a'}d\hat{u}^{a'}}{2\pi}\ e^{-\frac{\lambda}{2t}\sum_{a'}(g\left(u^0\right)-g\left(u^{a'}\right)+\omega\sqrt{t})^2}$$
$$\times\ e^{-\sum_{a'} i\hat{u}^{a'}u^{a'} - i\hat{u}^0 u^0 - \frac{1}{2}\left(\hat{u}^0\right)^2 - m\hat{u}^0\sum_{a'}\hat{u}^{a'} - \frac{1}{2}(q_d-q_0)\sum_{a'}\left(\hat{u}^{a'}\right)^2 - \frac{1}{2}q_0\left(\sum_{a'}\hat{u}^{a'}\right)^2} \tag{3.71}$$

$$= \frac{1}{n}\log \int D\omega \int \frac{du^0}{\sqrt{2\pi}}\prod_{a'=1}^{n}\frac{du^{a'}d\hat{u}^{a'}}{2\pi}\ e^{\frac{1}{2}\left(m\sum_{a'}\hat{u}^{a'}+iu^0\right)^2} e^{-\frac{\lambda}{2t}\sum_{a'}(g\left(u^0\right)-g\left(u^{a'}\right)+\omega\sqrt{t})^2}$$
$$\times\ e^{-\sum_{a'} i\hat{u}^{a'}u^{a'} - \frac{1}{2}(q_d-q_0)\sum_{a'}\left(\hat{u}^{a'}\right)^2 - \frac{1}{2}q_0\left(\sum_{a'}\hat{u}^{a'}\right)^2} \tag{3.72}$$

$$= \frac{1}{n}\log \int D\omega \int D\gamma \int \frac{du^0}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(u^0\right)^2}\int\prod_{a'=1}^{n}\frac{du^{a'}d\hat{u}^{a'}}{2\pi}\ e^{-\frac{\lambda}{2t}\sum_{a'}(g\left(u^0\right)-g\left(u^{a'}\right)+\omega\sqrt{t})^2}$$
$$\times\ e^{-i\sum_{a'}\hat{u}^{a'}\left(u^{a'}-mu_0+\sqrt{q_0-m^2}\gamma\right) - \frac{1}{2}(q_d-q_0)\sum_{a'}\left(\hat{u}^{a'}\right)^2} \tag{3.73}$$

$$= \frac{1}{n}\log \int D\omega \int D\gamma \int Du^0 \left(\int \frac{du}{\sqrt{2\pi}}\ e^{-\frac{\lambda}{2t}(g\left(u^0\right)-g(u)+\omega\sqrt{t})^2}\frac{1}{\sqrt{q_d-q_0}}\right.$$
$$\left.\times\ e^{-\frac{1}{2(q_d-q_0)}\left(u-mu^0+\sqrt{q_0-m^2}\gamma\right)^2}\right)^n \tag{3.74}$$

$$= \int D\omega \int D\gamma \int Du^0 \log\left(\int Du\ e^{-\frac{\lambda}{2t}\left(g\left(u^0\right)-g\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}\gamma\right)+\sqrt{t}\omega\right)^2}\right). \tag{3.75}$$

We can then take derivatives to obtain the saddle point equations, which will of course depend on the choice of the non-linearity g, and solve them numerically, to obtain the typical $\zeta_t(\lambda)$. At this point one can solve the collapse condition and recover the memorization time as in Fig. 3.3.

Equivalence between Collapse and Condensation

In Eq. (3.10) we have introduced a criterion for collapse time. In Sec. 2.3.2 we have also discussed the condensation threshold for the REM which, in the context of DMs reads

$$\alpha + \zeta_{t_{cond}}(1) - \zeta'_{t_{cond}}(1) = 0. \tag{3.76}$$

In order to establish that the condensation and collapse phenomena happen at the same time, $t_c = t_{cond}$, we would therefore need to prove that

$$\zeta'_{t_c}(1) = -\frac{1}{2}. \tag{3.77}$$

This is indeed what we find for a typical trajectory as a consequence of the Nishimori condition.

We consider a typical data to be a diffused version of one of the starting training points

$$x_t = \xi^1 + \sqrt{t}\omega. \tag{3.78}$$

Notice that here we use the variance exploding diffusion process for homogeneity with the rest of the paper, but this analysis does not depend on the diffusion protocol, as long as we consider a typical point.

We now write $\zeta_t(\lambda)$ as

$$\zeta_t(\lambda) = \mathbb{E}_{\xi^1}\mathbb{E}_{p(x_t|\xi^1)} \log \mathbb{E}_{\xi} p_\lambda(x_t|\xi) \tag{3.79}$$

where the data points come form a prior distribution, $\xi^1, \xi \sim p(\xi)$, and the likelihood has the form

$$p_\lambda(x_t|\xi) \propto e^{-\frac{\lambda}{2t}\|x_t - \xi\|^2}. \tag{3.80}$$

Then we compute $\zeta_t'(\lambda)$ taking the derivative

$$\partial_\lambda \log \mathbb{E}_\xi p_\lambda(x_t|\xi) = \frac{\int -\frac{\|x_t-\xi\|^2}{2t} p_\lambda(x_t|\xi)p(\xi)\,d\xi}{\int p_\lambda(x_t|\xi)p(\xi)\,d\xi} \tag{3.81}$$

$$= \int -\frac{\|x_t-\xi\|^2}{2t} p_\lambda(\xi|x_t)\,d\xi \tag{3.82}$$

so we can write this quantity as an average with respect to the posterior distribution $p(\xi|x_t)$, which we will indicate with $\langle\cdot\rangle_{\xi|x}$. Substituting $\lambda = 1$ and applying the Nishimori condition Nishimori [1980] we finally obtain

$$\zeta_t'(1) = \mathbb{E}_{\xi^1}\left[\mathbb{E}_{x|\xi^1}\left[\langle -\frac{\|x_t-\xi\|^2}{2t}\rangle_{\xi|x}\right]\right] \tag{3.83}$$

$$= \mathbb{E}_{\xi^1}\left[\mathbb{E}_{x|\xi^1}\left[-\frac{\|x_t-\xi^1\|^2}{2t}\right]\right] \tag{3.84}$$

$$= -\frac{1}{2}. \tag{3.85}$$

3.2.4 Onset Time and Basins of Attraction

The goal of this section is to compute the onset time t_o, i.e. the time at which data points start to become attractors in the diffusion potential. Since the computation does not average over the typical positions sampled by the reverse process, even if data become locally attractive, we find that they do not influence the typical trajectories until t_c. This aspect is the main difference between the onset time and the *speciation* time computed by Biroli and Mézard [2023]: while the former is intrinsic in the dataset itself, the latter depends on the structure of the data points as divided in multiple classes and it does affect the direction of the diffusion in the ambient space.

The onset time can be computed setting $x_t = \xi^1$ and checking when the corresponding collapse condition is satisfied. Such, condition, that is analogous to the one in Eq. (3.10), consists in requiring the relative REM free energy equal to zero, i.e. $\phi_{t_o} = 0$. As done for the condensation time, let us first compute the collapse time in the simple homogeneous Gaussian setting, where D variances are equal to σ^2 and the remaining ones are null. The moment-generating function of the relative REM is

$$\zeta_t(\lambda) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\xi^1} \log \mathbb{E}_\xi e^{-\frac{\lambda}{2t}\|\xi^1-\xi\|^2} = -\frac{\alpha_D}{2}\left(\log\left(1+\frac{\lambda\sigma^2}{\alpha_D t}\right) + \frac{\lambda\sigma^2}{\alpha_D t + \lambda\sigma^2}\right). \tag{3.86}$$

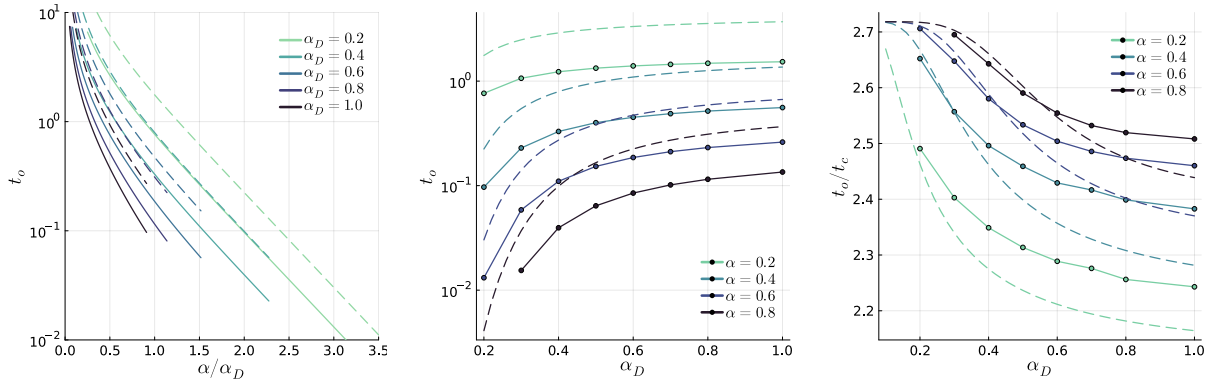


Figure 3.4: (**Left**) Onset time t_o as a function of α/α_D in semi-log scale in the hidden manifold case (solid) with tanh activation compared to the linear manifold case (dashed); (**Center**) t_o as a function of α_D for fixed α in semi-log scale for tanh activation; (**Right**) comparison of the onset time t_o with the collapse time t_c as a function of α_D when α is fixed and $g = \tanh$.

In analogy with the collapse condition in Eq. (3.10), the on-set time condition must be

$$\zeta_{t_o}(1) + \alpha = 0, \tag{3.87}$$

which reads

$$\log\left(1 + \frac{\sigma^2}{\alpha_D t_o}\right) + \frac{\sigma^2}{\sigma^2 + \alpha_D t_o} - \frac{2\alpha}{\alpha_D} = 0. \tag{3.88}$$

The same calculation in then performed in the case of manifold structured data for different choices of the g function. The computation is carried out by means of the replica method and it is reported later in this Section. Results are reported in Fig. 3.4. The left panel in the figure shows the onset time as a function of the ratio α/α_D, suggesting that t_o behaves similarly to the condensation time t_c. The right panel shows how t_o/t_c increases when the data are more structured (i.e. when α_D decreases). Surprisingly, this quantity also reaches a constant value when α is fixed and $\alpha_D \to 0$, in the linear case. Due to numerical limitations in computing the saddle point equations we could not observe the same trend in the non-linear scenario. Yet, the good qualitative agreement between the linear and non-linear cases at higher values of α_D suggests that the non-linear model might reach the same plateau. This particular behavior of the onset time might be attributable to the exponentially large size of the basins of attraction of the data points.

Let us now consider a more general case where $x_t = \xi^1 + \omega\sqrt{R}$ where $\omega \sim \mathcal{N}(0, I_N)$

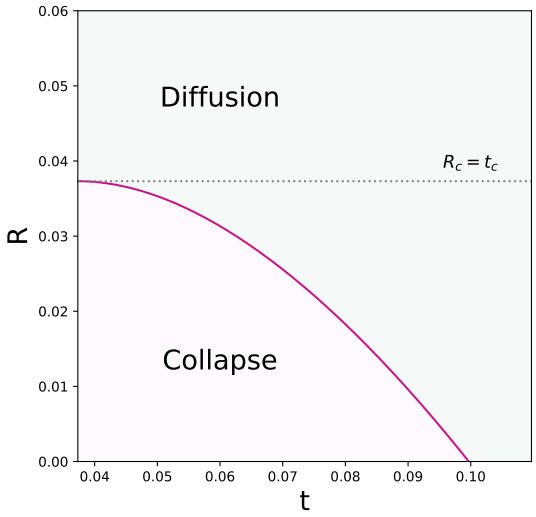


Figure 3.5: The violet line gives the radius R of the basin of attraction around one data point as a function of time. All particles at a distance smaller than $R(t)$ collapse to the same data point, while the ones at larger distances do not. The basin of attraction appears at $t = t_o$ and it is maximal at $t = t_c$, where the value of R equals the diffusion noise. Typical trajectories are trapped into the basin of attraction at times smaller than t_c. The parameters are $\sigma^2 = 1, \alpha = 1, \alpha_D = 0.5$.

and R is an arbitrary positive real value. Then one can repeat the calculation for the homogeneous Gaussian framework and obtain

$$\zeta_{t,R}(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\xi^1,\omega} \log \mathbb{E}_{\xi} e^{-\frac{\lambda}{2t}\|(\xi^1-\xi)+\omega\sqrt{R}\|^2} \tag{3.89}$$

$$= -\frac{1}{2}\left(\alpha_D \log\left(1 + \frac{\lambda\sigma^2}{\alpha_D t}\right) + \frac{\alpha_D \lambda\sigma^2}{t}\frac{(t-\lambda R)}{\alpha_D t + \lambda\sigma^2} + \frac{\lambda R}{t}\right). \tag{3.90}$$

Note that this expression for ζ coincides with Eq. (3.13) when $R = t$ and with Eq. (3.86) when $R = 0$. The new collapse condition for $R(t)$ is given by

$$\zeta_{t,R_c}(1) + \alpha = -\frac{R_c}{2t}. \tag{3.91}$$

The value of R_c when $t \in [t_c, t_o]$ represents the main distance at which particles would start feeling the attraction to the data point ξ^1, i.e. the particle is in the basin of attraction of the pattern if $R < R_c$. Fig. 3.5 reports the size of the basins of attraction as a function of the time for one realization of $\sigma^2, \alpha, \alpha_D$. The radius R starts assuming non-zero values at $t = t_o$ and equals the noise of stochastic process $R_c = t_c$ when $t = t_c$. When $t \in [0, t_c]$ each possible trajectory (both typical and non-typical) has collapsed in one of the basins, by definition of collapse time.

Computation of the Generating function: Linear case

As explained in Section 2.3.2, we need to compute the cumulant generating function as

$$\zeta_t(\lambda) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^0} \log \mathbb{E}_z e^{-\frac{\lambda}{2t}\|\left(\frac{Fz}{\sqrt{D}}-\frac{Fz^0}{\sqrt{D}}\right)\|^2} \tag{3.92}$$

$$= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^0} \log \int \frac{dz}{\sqrt{2\pi}}\ e^{-\frac{1}{2}z(I_D+\frac{\lambda}{t}\frac{F^TF}{D})z+\frac{\lambda}{t}z(\frac{F^TF}{D}z^0)-\frac{\lambda}{2t}\|-\frac{Fz^0}{\sqrt{D}}\|^2} \tag{3.93}$$

$$= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z^0}\Bigg[-\frac{1}{2}\log\det\left(I_D+\frac{\lambda}{t}\frac{F^TF}{D}\right)$$
$$+\frac{1}{2}\frac{\lambda^2}{t^2}\left(\frac{F^TF}{D}z^0\right)^T\left(I_D+\frac{\lambda}{t}\frac{F^TF}{D}\right)^{-1}\left(\frac{F^TF}{D}z^0\right)-\frac{\lambda}{2t}\|\frac{Fz^0}{\sqrt{D}}\|^2\Bigg]. \tag{3.94}$$

Now with a rotation we can position in the basis of the eigenvectors of $\frac{F^\top F}{N}$, with eigenvalues σ_k^2

$$= \lim_{N\to\infty} \frac{1}{N}\sum_k^D \left[-\frac{1}{2}\log\left(1+\frac{\lambda}{\alpha_D t}\sigma_k^2\right)+\frac{\lambda^2}{2\alpha_D^2 t^2}\left(\frac{\sigma_k^4}{1+\frac{\lambda}{\alpha_D t}\sigma_k^2}\right)-\frac{\lambda}{2\alpha_D t}\sigma_k^2\right] \tag{3.95}$$

$$= \lim_{N\to\infty} \frac{1}{N}\sum_k \left[-\frac{1}{2}\log\left(1+\frac{\lambda}{\alpha_D t}\sigma_k^2\right)-\frac{\lambda}{2}\frac{\sigma_k^2}{\alpha_D t+\lambda\sigma_k^2}\right]. \tag{3.96}$$

Here we have assumed that $\alpha_D < 1$. Replacing with the law ν for the bulk of the Marchenko-Pastur distribution we have

$$\zeta_t(\lambda) = -\frac{\alpha_D}{2}\int \nu_{\alpha_D}(d\sigma^2)\left[\log\left(1+\frac{\lambda\sigma^2}{\alpha_D t}\right)+\frac{\lambda\sigma^2}{\alpha_D t+\lambda\sigma^2}\right]. \tag{3.97}$$

This expression of ζ_t at $\lambda = 1$ is then used to obtain $\phi(\alpha, t)$.

Computation of the Generating function: Non-linear case

In case of non-linear functions that define the manifold, we are going to employ the replica method to compute the REM free-energy, as we performed for the condensation

time. First we need to compute the cumulant generating function as

$$\zeta_t(\lambda) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_x \log \mathbb{E}_\xi[e^{-\frac{\lambda}{2t}\|x-\xi\|^2}] \tag{3.98}$$

$$= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z_0} \log \mathbb{E}_z[e^{-\frac{\lambda}{2t}\|g(\frac{Fz_0}{\sqrt{D}})-g(\frac{Fz}{\sqrt{D}})\|^2}] \tag{3.99}$$

$$= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{F,z_a}[e^{-\frac{\lambda}{2t}\sum_{a'}\|g(\frac{Fz_0}{\sqrt{D}})-g(\frac{Fz_{a'}}{\sqrt{D}})\|^2}]. \tag{3.100}$$

Using the replica symmetric ansatz we obtain

$$\zeta_t(\lambda; q_d, q_0, m, \hat{q}_d, \hat{q}_0, \hat{m}) = -\alpha_D m\hat{m} - \frac{\alpha_D}{2}(q_d\hat{q}_d - q_0\hat{q}_0) + \alpha_D G_S(\hat{q}_d, \hat{q}_0, \hat{m}) + G_E(\lambda, t; q_d, q_0, m) \tag{3.101}$$

with

$$G_S(\hat{q}_d, \hat{q}_0, \hat{m}) = -\frac{1}{2}\log(1 - \hat{q}_d + \hat{q}_0) + \frac{1}{2}\frac{\hat{m}^2 + \hat{q}_0}{1 - \hat{q}_d + \hat{q}_0} \tag{3.102}$$

and for the energetic term

$$G_E(\lambda, t; q_d, q_0, m) = \int D\gamma \int Du^0 \log\left(\int Du e^{-\frac{\lambda}{2t}\left(g(u^0) - g\left(\sqrt{q_d - q_0}u + mu^0 - \sqrt{q_0 - m^2}\gamma\right)\right)^2}\right). \tag{3.103}$$

Then one can solve the saddle point equation, which will depend on the choice of the non-linearity g, and obtain $\zeta_t(\lambda)$ at the fixed point.

3.3 Generalization in Generative Diffusion

In this section, we compute the optimal time t_g such that the empirical probability distribution of a DM better fits the target distribution. The degree of generalization of a DM driven by its empirical score can be quantified in terms of the Kullback-Leibler (KL) divergence between the empirical probability distribution of the model and the distribution of the data points on the manifold. We first show that the true score and the empirical one do not differ, in the large volume limit, above the collapse transition. Secondly, we calculate t_g for different choices of α and α_D, showing that this time is always contained within the condensed phase of the auxiliary REM, i.e. the memorization phase of the DM. A similar effect has been found when seeking the best kernel to approximate probability densities from large-dimensional data: the optimal kernel width is found in the condensed phase Biroli and Mézard [2024]. This is no coincidence: in generative diffusion, the effective probability distribution p_t^{emp} is a sum of Gaussian kernels centered on the

data points. Finally, since the computation of t_g relies on the presence of collapse over the training set, which is not always encountered in real-world applications of generative diffusion, we propose an alternative criterion to define generalization in DMs.

3.3.1 True vs Empirical Distribution

The Kullback-Leibler (KL) divergence between the true and empirical distribution is

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\, D_{KL}[p_t(x)\,|\,p^{emp}_{t,\mathcal{D}}(x)] = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\left[\int dx_t\; p_t(x)\,\log\, p_t(x) - \int dx\; p_t(x)\,\log\, p^{emp}_{t,\mathcal{D}}(x)\right] \tag{3.104}$$

In the uncondensed phase we can exploit the fact that the annealed approximation holds, combined with $\mathbb{E}_{\mathcal{D}}\left[p^{emp}_{t,\mathcal{D}}(x)\right] = p_t(x)$ to obtain

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\, D_{KL}[p_t(x)\,|\,p^{emp}_{t,\mathcal{D}}(x)] = \begin{cases} 0 & \text{uncondensed phase} \\ \varepsilon^*(t,\alpha) - \alpha - \frac{1}{2}\log(2\pi t) - H_t & \text{condensed phase} \end{cases} \tag{3.105}$$

with $\varepsilon^*(t,\alpha) = -\lim_{N\to\infty}\mathbb{E}_{x,\mathcal{D}}\,\frac{1}{2Nt}\|x_t - \xi^*(x_t,\mathcal{D})\|^2$ and ξ^* being the nearest neighbor to x among the data points, while H_t is an additional time dependent term. The divergence between the empirical and true scores starting from t_c is represented in the bi-dimensional plot contained in Fig. 3.6 for one explanatory diffusion experiment, and it is validated by Fig. 3.7 relative to a further analysis of generalization.

3.3.2 Generalization Time: Generalizing while Collapsing

We would like to understand if there is a time at which the empirical score function points towards the original data manifold and not directly to the data points. To study this, we compute the KL divergence between the target distribution, i.e. p_0, and the empirical distribution at time t, and then minimize it to find the *generalization* time.

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}} D_{KL}[p_0\,|\,p^{emp}_{t,\mathcal{D}}] = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\left[\int dx\; p_0(x)\log p_0(x) - \int dx\; p_0(x)\,\log\, p^{emp}_{t,\mathcal{D}}(x)\right]. \tag{3.106}$$

The second term can be computed using the REM formalism (see Sec. 3.3.2) as

$$\tilde{D}_{KL}[p_0|p^{emp}_t] = -\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\int dx\, p_0(x)\,\log\, p^{emp}_{t,\mathcal{D}}(x) \simeq -\phi_{t,\alpha}(1)+\alpha+\frac{1}{2}\log(2\pi t). \tag{3.107}$$

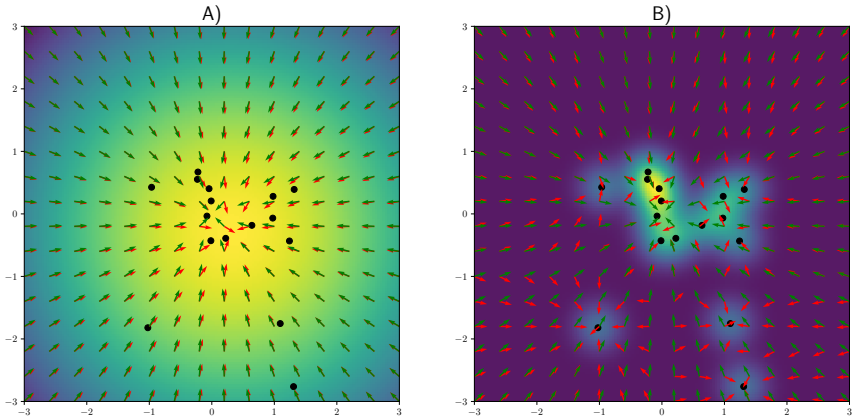


Figure 3.6: In these Figures, at two given times, the heat map indicates the empirical sampling distribution of a DM, red arrows represent the empirical score while green ones represent the exact score. The black dots denote individual data points. Panel A depicts such quantities at a $t > t_c$ and panel B displays the typical scenario at a $t < t_c$. The score transitions from a phase where its direction is dominated by the expectation (i.e. the exact score) to a phase where its orientation is mostly determined by the individual data points.

We show the behavior of the KL divergence for data from a hidden manifold model with tanh non-linearity in Fig. 3.7. Interestingly, the time t_g where the discrepancy between p_0 and p_t^{emp} reaches a minimum is always smaller than the corresponding collapse time (reported as a dashed line in the figure): the best generalization of the DM is reached inside the condensation phase, while the diffusive trajectory is trapped into the basin of attraction of the closest data point. It is also worth to notice that

$$\lim_{\alpha\to\infty} \tilde{D}_{KL}[p_0|p_t^{emp}] = \tilde{D}_{KL}[p_0|p_t]$$

where $p_t(x)$ is the exact probability distribution of the diffusive process. This quantity is represented by the line onto which all the curves in Fig. 3.7 collapse, i.e. the black dashed line in the figure: the computation in Eq. (3.105) is validated by the fact that curves start diverging from the asymptotic line exactly at $t = t_c(\alpha)$. Moreover, Fig. 3.8 (Center) displays that t_g decreases with α_D when α is fixed, while Fig. 3.8 (Right) shows that the ratio t_g/t_c vanishes when $\alpha_D \to 0$. This result means that t_g goes to zero faster than the collapse time t_c. We can thus conclude that a high structure of the data helps the empirical-score-driven diffusion model for two reasons:

- Both t_c and t_g are pushed towards $t = 0$ when $\alpha_D \to 0$ but the generalization time is moving faster towards smaller times. Since t_g represents the best stopping time to sample along the reverse process from the point of view of the KL divergence,

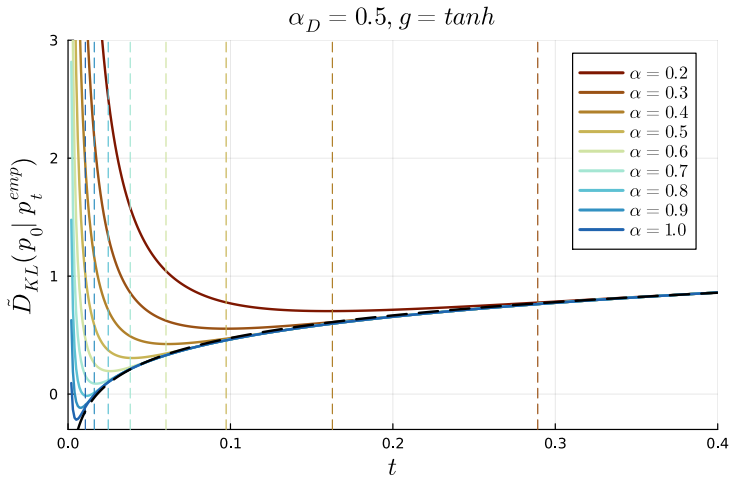


Figure 3.7: Time-dependent component of the KL divergence between target distribution and empirical distribution as a function of time t and for different values of α. The data are generated from a HMM with tanh activation and aspect ratio $\alpha_D = 0.5$. We report with colored dashed lines the condensation time t_c at the corresponding value of α, and with the black dashed line the limit $\alpha \to +\infty$.

one sees that this optimal time occurs after the condensation threshold and much closer to $t = 0$, when the true memorization occurs.

- The generalization time occurs inside the memorization phase, i.e.

$$0 < t_g < t_c \quad \forall \alpha, \alpha_D,$$

and the Kullback-Leibler distance between p_0 and p_t^{emp} is a monotonic function in $t \in [t_g, t_c]$ i.e.

$$D_{KL}[p_0 | p_{t_c}^{emp}] > D_{KL}[p_0 | p_{t_g}^{emp}] > 0. \tag{3.108}$$

Since the empirical model tends to the exact one when $\alpha_D \to 0$ i.e.

$$\lim_{\alpha_D \to 0} \lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\mathcal{D}} D_{KL}[p_0 \,|\, p_{t_c,\mathcal{D}}^{emp}] = 0, \tag{3.109}$$

then we must have

$$\lim_{\alpha_D \to 0} \lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\mathcal{D}} D_{KL}[p_0 \,|\, p_{t_g,\mathcal{D}}^{emp}] = 0, \tag{3.110}$$

which means that the degree of generalization of the DM improves when data are more structured.

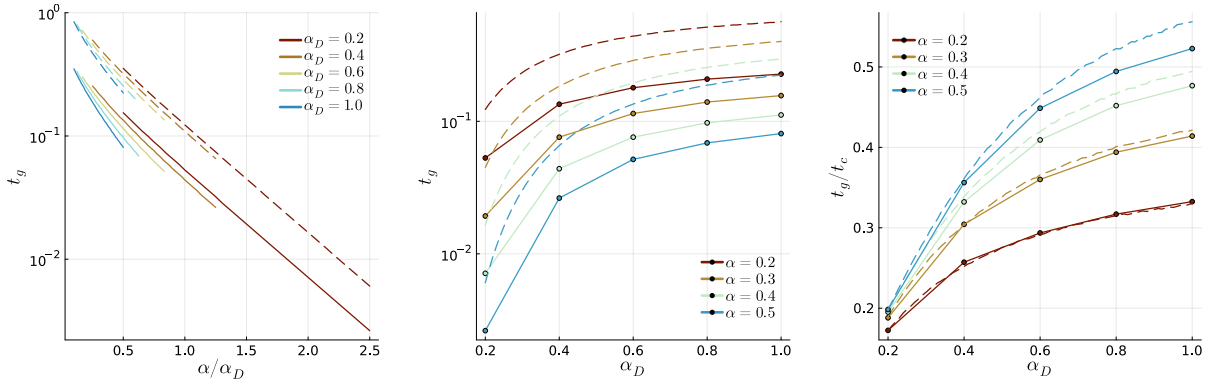


Figure 3.8: (**Left**) Generalization time t_g as a function of α/α_D in semi-log scale for tanh (solid) and linear (dashed) activation; (**Center**) t_g as a function of α_D for fixed α in semi-log scale for tanh (solid) and linear (dashed) activation; (**Right**) comparison of the generalization time t_g with the collapse time t_c as a function of α_D when α is fixed for tanh (solid) and linear (dashed) activation.

Computation of the KL-Divergence

The Kullback-Leibler (KL) divergence is a type of statistical distance between two probability density functions. Given the two distributions $p_0(x)$, namely the ground-truth distribution of the data, and $p^{emp}_{t,\mathcal{D}}(x)$, namely the empirical distribution of the data according to the model, the full KL divergence between these two functions assumes the following expression

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}D_{KL}\left[p_0|p^{emp}_{t,\mathcal{D}}\right] = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\left[\int dx p_0(x)\log p_0(x) - \int dx p_0(x)\log p_{t,\mathcal{D}}(x)\right] \tag{3.111}$$

$$= -s_0 + \tilde{D}_{KL}\left[p_0|p^{emp}_t\right], \tag{3.112}$$

where s_0 is the entropy of the p_0 distribution and $\tilde{D}_{KL}$ is the only time-dependent component of the KL divergence. Since we are studying a data-model where $p_0(x)$ is defined on a support having a lower dimensionality with respect to the N-dimensional data-space, we expect the entropy s_0 to diverge. This issue might be controlled by adding some noise to either the latent data points z^μ or the features in F, but we will not engage into this analysis. Nevertheless, for studying the dependence on t we can compute the $\tilde{D}_{KL}$ function in order to find the *generalization time* t_g at which the distance between the two distribution is minimal. We derive below $\tilde{D}_{KL}$ in both the linear and non-linear manifold cases by expressing this quantity in terms of time-dependent free-energy function in the

REM formalism.
The time-dependent part of the KL divergence is given by is given by

$$\tilde{D}_{KL}[p_0|p_t^{emp}] = -\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mathcal{D}}\int dx\ p_0(x)\ \log\ p_{t,\mathcal{D}}^{emp}(x), \tag{3.113}$$

with $x \sim g(\frac{Fz}{\sqrt{D}})$, $z \sim \mathcal{N}(0, I_D)$, $F \in \mathbb{R}^{N\times D}$, and the empirical score reads

$$\log p_{t,\mathcal{D}}^{emp}(x) = \log\frac{1}{P\sqrt{2\pi t}^N}\sum_{\mu=1}^{P} e^{-\frac{1}{2t}\|x-\xi^\mu\|^2} \simeq N\left[\Phi_t(x) - \alpha - \frac{1}{2}\log(2\pi t)\right], \tag{3.114}$$

where

$$\Phi_t(x) = \frac{1}{N}\log\sum_{\mu=1}^{P} e^{-\frac{1}{2t}\|x-\xi^\mu\|,^2} \tag{3.115}$$

is again minus the free energy density of a REM. For $P, N \to \infty$ with $\alpha = \log P/N$ this concentrates to

$$\phi(\alpha, t) = \lim_{N\to\infty}\mathbb{E}_{x\sim p_0}[\Phi_t(x)], \tag{3.116}$$

and to know this limit we need to compute the large deviation function.
In case of a linear manifold we can compute $\tilde{D}_{KL}$ in terms of the free-energy of a REM, as in Eq. (3.107). The computation of the free-energy function coincides with the one performed in Sec. 3.2.4.
In case of non-linear functions that define the manifold, we are going to employ the replica method to compute the REM free-energy, as we performed for the condensation time. The computation coincides with the one performed in Sec. 3.2.4.

3.3.3 Generalization Condition: Generalizing before Collapsing

We now propose a more empirical definition of generalization for DMs. The main idea consists of sampling configurations from the data manifold before the model enters its memorization phase. The current definition of generalization is supported by the common routine used in generative modeling consisting in early-stopping the stochastic sampling process Li et al. [2023], Yang and E [2021], with the aim of improving the quality of the examples. Consistently with Li et al. [2023], our analysis shows that we need a polynomial number of training data points to obtain generalization without falling into memorization.

Let us consider the exact score function measured from a dataset embedded in a linear manifold: we have proved in Sec. 3.3.1 that the true score coincides with the empirical

one for $t > t_c$. The argument around the linear manifold can be extended to a non-linear one by observing that the interesting phenomenology in DMs occur at very small times (mainly due to the data structure, see Sec. 3.3), where the amplitude of the stochastic noise $\sqrt{t}$ is much smaller than the manifold curvature. A more extensive dissertation about this aspect can be found in Ventura et al. [2025]. When F is a random matrix with i.i.d. standard Gaussian entries, $F^\top F/D$ is a Wishart matrix and its eigenvalues satisfy the Marchenko-Pastur distribution. As showed in Ventura et al. [2025], the Jacobian of the empirical score function before condensation is given by

$$J_t = \frac{1}{t} F \left[I_D + \frac{1}{t} F^\top F \right]^{-1} F^\top - I_N, \tag{3.117}$$

where we have re-absorbed the $1/D$ factor for the sake of clarity. Therefore, the spectrum of the eigenvalues of J_t can be derived by a propagation of the spectrum of $F^\top F$ and it is

$$\rho_t(r) = (1-\alpha_m)\,\delta\,(r+1)\,\theta\left[\alpha_D^{-1} - 1\right] - \frac{\alpha_D}{2\pi}\frac{1}{r(1+r)}\sqrt{(r_+ - r)\,(r - r_-)}\,\theta\left[(r_+ - r)\,(r - r_-)\right], \tag{3.118}$$

with $r_\pm(t) = -\frac{t}{\left(1 \pm \frac{1}{\sqrt{\alpha_D}}\right)^2 + t}$. The first term in $\rho_t(r)$ is a spike in $r = -1$ with mass equal to $(1 - \alpha_D)$, the second term is a bulk of mass α_D, ranging in $[r_-(t), r_+(t)]$, and moving from $r = -1$ towards $r = 0$.

The structure and dynamics of the eigenspectrum, composed by moving bulks of non-zeros eigenvalues towards the origin, suggest the presence of an evolving latent manifold. Vanishing eigenvalues are relative to tangent directions to the manifold, while non-zero ones must be associated to orthogonal directions. The score function, in fact, always points orthogonally towards the evolving manifold, since it is *projecting* diffusive trajectories onto it. The final part of this transformation of the spectrum, described in detail in Ventura et al. [2025], represents the consolidation of the target manifold, and it is represented by the last bulk being absorbed by the $r = 0$ spike.

We are now interested in evaluating the width of the gap forming between $r = -1$ and $r = r_-(t)$, i.e. the gap separating the last moving bulk and the $r = -1$ spike. We know that such gap is progressively closing when $t \to 0^+$, because the spectrum must be formed of two spikes, one of mass $(1 - \alpha_D)$ in $r = -1$ and one of mass α_D in $r = 0$. We can hence find the approximate time at which the score function points towards the true target manifold by imposing such gap to equal a quantity $\delta \approx 1$. We call such time t_g^{RMT} and it is given by

$$t_g^{RMT}(\delta) = \left(1 - \frac{1}{\sqrt{\alpha_D}}\right)^2 \left(\frac{1-\delta}{\delta}\right). \tag{3.119}$$

Let us compute the condition such that the score is sufficiently orthogonal to the manifold (i.e. the model generates examples that live on the data manifold) and it has not collapsed yet. Such condition reads

$$t_c \leq t_g^{RMT}(\delta). \tag{3.120}$$

Let us assume to be in the $D \ll \log P$ and $D \ll N$ regime where t_c is given by Eq. (3.16). Moreover, we choose $\delta = 1 - \epsilon$ with small $\epsilon > 0$. Hence, condition (3.120) reads

$$t_c \approx \frac{1}{2\alpha_D} e^{-\frac{2\alpha}{\alpha_D}} \leq \frac{1}{2}\left(1 - \frac{1}{\sqrt{\alpha_D}}\right)^2 \left(\frac{1-\Delta}{\Delta}\right) \simeq \frac{\epsilon}{2\alpha_D}, \tag{3.121}$$

where we employed the fact that $\left(1 - \alpha_D^{-1/2}\right)^2 \simeq \alpha_D^{-1}$, when $\alpha_D \ll 1$. Eq. (3.121) thus becomes $e^{\frac{2\alpha}{\alpha_D}} \geq \epsilon^{-1}$. As a consequence, the minimum amount of data points such that the generalization condition (3.120) is satisfied, must scale as

$$P_{\min} = \epsilon^{-\frac{D}{2}}, \tag{3.122}$$

which surprisingly is a function of the dimension of the manifold rather than the ambient space.

3.4 Conclusions

We have extensively analyzed the memorization and generalization performance of a DM driven by the empirical score function, that is, the score corresponding to the noised empirical distribution, as a proxy of true or learned scores. Our main contribution is the extension the REM framework introduced by Refs. Lucibello and Mézard [2024], Biroli et al. [2024] to the case of structured data living on a hidden manifold. Our study sheds light on the role of the manifold structure in learning the ground-truth distribution underneath the training set.

Firstly, we find that empirical-score-driven DMs can both memorize and generalize a set of data points at different times. We highlighted a rich sequence of dynamical phases occurring during the reverse diffusion process that starts from $t = t_f \gg 1$ and reaches $t = 0$:

- $t_c < t \leq t_o$: diffusive trajectories explore a diffusion potential which is now multi-stable, since data points have become local minima surrounded by basins of attraction that grow while time decreases. The typical stochastic path of the system is

not trapped in one of the basins, without showing any trace of memorization.

- $t_g \leq t \leq t_c$: the diffusive trajectory is now trapped in the basin of attraction, and the empirical score function points towards the closest data point. At the same time, the trajectory is also approaching the hidden data manifold. The highest proximity between the empirical distribution of the states sampled by diffusion and the ground-truth distribution of the data points is reached at $t = t_g$. This time can be interpreted as the optimal stopping time for sampling.

- $0 < t < t_g$: the quality of the sampled examples now deteriorates until full memorization is reached at $t = 0$.

Note that the so-called *speciation time* studied in Biroli et al. [2024], Ambrogioni [2025], understood as the time when the diffusive potential undergoes a spontaneous symmetry breaking into multiple ergodic components that are representative of the data classes, has not been analyzed in our paper since our data model does not have clear class separation. We refer the reader to George et al. [2025] for the study of the speciation time under the manifold hypothesis.

Surprisingly, the best degree of generalization is reached inside the memorization phase of the model, while the score function drives the model towards the closest attractor. The dynamical picture of the DM reported above is deformed by the presence of structure in the data, as it emerged from our analysis. Specifically, when α is fixed and $\alpha_D \to 0$:

1. Even though the onset time exponentially decreases, the distance between t_o and the condensation time increases until reaching a constant plateau.

2. The collapse time t_c shrinks towards $t = 0$, and the empirical-score-drive DM tends to the exact model, hence reducing the volume of the memorization phase of the model. This result is consistent with the very recent result obtained by George et al. [2025] in the matter of variance-preserving DMs.

3. The generalization time t_g also moves towards $t = 0$, yet faster than t_c.

In light of point (3) we conclude that DMs benefit from highly structured data, even when α_D has not completely vanished, since the model can be basically stopped at $t \simeq 0$ and obtain a good degree of generalization, as one would obtain through a neural-network-trained model.

As an alternative to this definition of generalization, we use a combination of the REM formalism and Random Matrix Theory (RMT) to provide the reader with the minimal

number of training data point to build the empirical score function in such a way that the DM is capable of sampling from the manifold with a minimal KL divergence. We find that the size of the data set needs to scale exponentially with the latent dimension of the data, instead of the visible dimension, mitigating the curse of dimensionality that affects learning in generative models Yarotsky [2017], Cybenko [1989].

Summary

The main points highlighted in this chapter are:

- We obtained the **memorization time** for **structured data** and found that it shows a mitigated curse of dimensionality
- We defined **generalization** for the process driven by the **empirical score**, and saw that it always happen after memorization

Chapter 4

Geometric perspective on generative diffusion

This chapter contains contributions coming from two works: Ventura et al. [2025] and Achilli et al. [2024]. For this reason, it can be formally divided into two parts:

- The geometric phases of generative diffusion
- Losing dimensions

that nonetheless share a geometrical approach to the study of DMs and an application to manifold data, thus providing a complete overview of the problem.

The geometric phases of generative diffusion

In this first part, we uncover the geometric phases of generative diffusion. Generative diffusion models synthesize images through a stochastic dynamical denoising process. Experimental and theoretical arguments suggest that different features, such as frequency modes and class labels, are generated at different times during the process. For example, it has been shown that separation between isolated classes, as in the case of mixture of Gaussian models, happens at critical phase transition points of spontaneous symmetry breaking (speciation events) [Biroli et al., 2024]. It is also well known that subspaces corresponding to different frequency modes emerge at different times of diffusion [Kingma and Gao, 2024]. This idea has been recently refined by [Kadkhodaie et al., 2024], who showed that diffusion models give rise to a local decomposition of the image manifold into a basis of geometry-adaptive harmonic basis functions. These decomposition phenomena

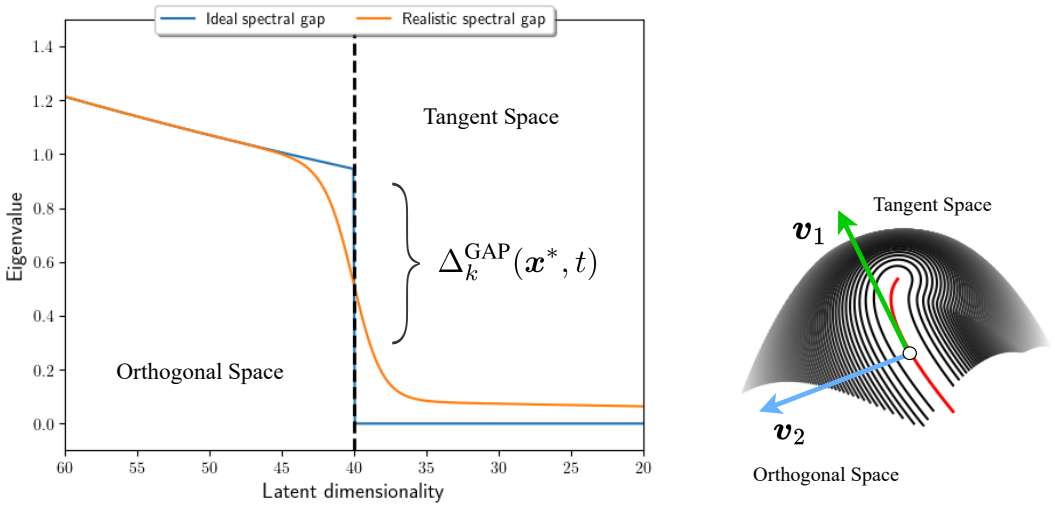


Figure 4.1: (a) Visualization of the gaps in the spectrum of the (negative) Jacobian of the score for data supported on a latent manifold. Blue line: idealized spectrum of distribution with uniform internal density; Orange line: spectrum of a more realistic distribution. (b) Sketch of the local structure of the data manifold with tangent and orthogonal components of the score function.

cannot be directly explained in terms of critical phase transitions, as they are fundamentally linear processes. In this chapter, we will provide a precise theoretical analysis of the separation of subspaces for data defined on low-dimensional linear manifolds.

Our main contributions are: I) an in-depth theoretical random-matrix analysis of the distribution of Jacobian spectra in diffusion models on linear manifolds and II) a detailed experimental analysis of Jacobian spectra extracted from trained networks on linear manifolds and on image datasets. The analysis of these spectra is important as it provides a detailed picture of the latent geometry that guides the generative diffusion process. We show that the linear theory predicts several phenomena that we observed in trained networks. Based on our result, we divide the generative process into three qualitatively different phases: **trivial phase**, **manifold coverage phase**, and **manifold consolidation phase**. Using these concepts, we provide a concise explanation of why diffusion models can avoid the *manifold overfitting* pathology that characterizes likelihood-based generative models [Loaiza-Ganem et al., 2022].

4.1 Dynamic latent manifolds and spectral gaps

The manifold hypothesis states that the distribution on natural data, such as images and sound recordings, is supported on a $D-$dimensional manifold $\mathcal{M}$ embedded in a larger

Euclidean ambient space $\mathbb{R}^N$. For more details and references, we refer the reader to the introductory Section 2.4.1. While a data probability distribution supported on a $D < M$ manifold $\mathcal{M}$ cannot be expressed using a proper density function, we loosely define such density as

$$p_0(\boldsymbol{x}) = \delta_{\mathcal{M}}(\boldsymbol{x})\, \rho(\boldsymbol{x}) \ , \tag{4.1}$$

where $\delta_{\mathcal{M}}$ is the Dirac function for the manifold and such that $\int_{\mathbb{R}^d} \delta_{\mathcal{M}}(\boldsymbol{x}) \bullet d\boldsymbol{x} = \int_{\mathcal{M}} \bullet\, d\boldsymbol{x}$. We call $\rho(\boldsymbol{x})$ the *internal density*, that is, the density restricted to the manifold. The density $p_0(\boldsymbol{x})$ is zero outside the manifold and diverges on the manifold.

Consider a generative diffusion model with $p_0(\boldsymbol{x})$ defined on a D-dimensional manifold $\mathcal{M}$ according to Eq. (4.1). In the course of the diffusion process, we can define a time-dependent locus of points

$$\mathcal{M}_t = \{\boldsymbol{x}^* \mid \tilde{s}_{\mathcal{M}}(\boldsymbol{x}^*, t) = 0, \text{ with } J_{\mathcal{M}}(\boldsymbol{x}^*, t) \text{ n.s.d.}\} \ , \tag{4.2}$$

that we name **stable latent set** of the process. In Eq. 4.2 we have used $\mathcal{M} \equiv \mathcal{M}_0$. The negative semi-definiteness (n.s.d.) is a stability condition on the Jacobian matrix $J_{\mathcal{M}}(\boldsymbol{x}, t)$ of the support score $\tilde{s}_{\mathcal{M}}(\boldsymbol{x}^*, t)$, defined as the score function obtained from the uniform data distribution $\tilde{p}_0(\boldsymbol{x}) = \frac{1}{|\mathcal{M}|}\delta_{\mathcal{M}}(\boldsymbol{x})$. Due to the noise, the diffusing particles typically explore shells of a radius that concentrates on $\sqrt{t}$ around each point of the latent stable set. For a small perturbation $\boldsymbol{p}$ around a point $\boldsymbol{x^*}$ on the latent manifold at time t, the score function is well approximated by its linearization:

$$s(\boldsymbol{p}, t) \approx J(\boldsymbol{x}^*, t)\ \boldsymbol{p} = -\sum\nolimits_j (\boldsymbol{v}_j \cdot \boldsymbol{p})\, \lambda_j(\boldsymbol{x}^*, t) \boldsymbol{v}_j \ , \tag{4.3}$$

where $J(\boldsymbol{x}^*, t)$ is the Jacobian of the score and the $\boldsymbol{v}_j$ and $\lambda_j(\boldsymbol{x}^*, t)$ are respectively the j-th eigenvector and the associated eigenvalue of $-J(\boldsymbol{x}^*, t)$. The spectrum of eigenvalues provides detailed information concerning the local geometry of the stable latent set. Perturbations aligned with the tangent space of $\mathcal{M}_t$ correspond to small eigenvalues, while orthogonal perturbations correspond to high eigenvalues, as the score tends to push the stochastic dynamics towards its fixed points. Therefore, since at $t \approx 0$ the score is orthogonal to the manifold, we can estimate the dimensionality of the manifold from the location of a drop (i.e., a sharp change) in the sorted spectrum of eigenvalues [Stanczuk et al., 2024]. This is visualized in Fig. 4.1, panels (a) and (b). This drop corresponds exactly to a gap (i.e., a separation) in the eigenvalue spectrum; in the following, we will refer to both as gaps.

At large times, instead, the score function can point in any direction, so what happens

in between? In the following, we describe how the manifold structure emerges with time from the diffusion process.

4.1.1 Subspaces and intermediate gaps

Consider the situation where the internal density $\rho_{\text{int}}(\boldsymbol{x})$ is not locally flat around a point $\boldsymbol{x}^* \in \mathcal{M}$. In this case, at a finite time t, the actual score function does not vanish on the latent stable set $\mathcal{M}_t$ as there is a gradient of the log-density along the tangent directions. This implies that the spectrum of tangent eigenvalues can have a series of sub-gaps with separate different tangent subspaces with different *local variance*. In image generation tasks, these subspaces are often associated with different frequency modes, as noted in [Kingma and Gao, 2024]. Consequently, we can quantify the sensitivity to the internal density at time t by studying the statistics and temporal evolution of intermediate gaps

$$\Delta_k^{\text{GAP}}(\boldsymbol{x}^*, t) = \lambda_{k+1}(\boldsymbol{x}^*, t) - \lambda_k(\boldsymbol{x}^*, t) \ , \tag{4.4}$$

where the indices k depend on the dimensionality of the subspaces. Note, however, that under realistic data distributions it is unlikely to find sharp intermediate discontinuities, since each subspace will have a different eigenvalue, resulting in a smooth gradient.

4.2 Phenomenology of generative diffusion on manifolds

This section contains an intuitive picture that follows from our theoretical results on linear models, which we will fully outline in the next section. The theory considers the case of a linear manifold with a Gaussian internal distribution. A linear-manifold data model is made of a set of points

$$\boldsymbol{\xi}^\mu = F\boldsymbol{z}^\mu, \tag{4.5}$$

where F and $\boldsymbol{z}^\mu$ have been introduced in Section 2.4.1. While only linear models are theoretically tractable, we conjecture that their phenomenology captures the main features of subspace separation in the tangent space of curved manifolds (see Sec. 4.5.1 for further details). We validated the theory using networks trained on both linear data and highly non-linear data such as natural images (see Sections 4.5 and 4.6). Based on the dynamics of the spectral gaps, we found that the generative dynamics of $\boldsymbol{x}_t$ according to the backward equation can be separated into three distinct phases. The phase separation does not

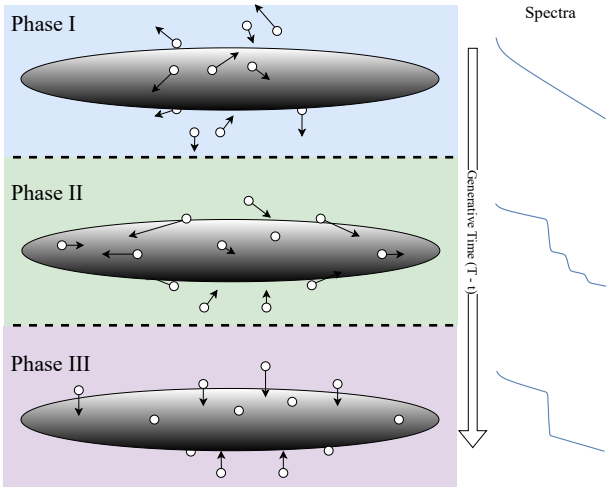


Figure 4.2: Representation of the proposed phases with a sketch of the corresponding Jacobian spectrum on the side.

correspond to singularities, as there are cross-over events, not genuine phase transitions. During all our analysis, we will exclusively work with *eigenvalues* since, in the linear manifold model, the Jacobian of the true score is symmetric. The same phenomenology is nevertheless fully appreciable when using the *singular values* in our experiments with neural approximations of the score.

Phase I: The trivial phase

In the *trivial phase*, the diffusing particle moves according to the noise distribution without strong biases towards the manifold directions. In this dynamic regime, the stable latent set $\mathcal{M}_t$ is a single point surrounded by an isotropic quadratic well of potential. The spectral gaps are not visible, and all eigenvalues have approximately the same value due to the isotropy of the noise distribution. This trivial phase is analogous to the initial phases described in [Raya and Ambrogioni, 2023] and [Biroli et al., 2024].

Phase II: Manifold coverage

The *manifold coverage phase* begins with the opening of the first of a series of spectral gaps corresponding to local subspaces. In this phase, different subspaces with different variances can therefore be identified by intermediate gaps in the spectra, as sketched in Fig. 4.2. When the intermediate gaps are opened, the diffusing particles spread across the manifold directions according to their relative variances. In other words, during this regime of generative diffusion, the process fits the distribution of the data internal to the manifold.

We assume a low-rank covariance $\Sigma = FF^\top$ for the data distribution. In terms of random matrix theory, the gap-forming phenomenology has two distinct processes: the emergence of intermediate gaps (i.e., steps in the dimensionality plot) between separated bulks of the spectrum, and the opening of a final gap that allows one to infer the dimensionality of the full manifold. Our analysis gives us the time scale at which such intermediate gaps are maximally opened, i.e.

$$t_{\max}^{(k)} = \sqrt{\gamma_+(\sigma_k)\,\gamma_-(\sigma_{k+1})}, \tag{4.6}$$

where $\gamma_-(\sigma_{k+1})$ and $\gamma_+(\sigma_k)$ are specific eigenvalues of Σ (see Fig. 4.4) associated with two hierarchically consecutive variances (see Sec. 4.4.2 for an exhaustive analysis). In most cases, when $\sigma_{k+1}^2 \ll \sigma_k^2$, the dependence on the two variances is $\mathcal{O}(\sigma_k \cdot \sigma_{k+1})$. This is the timescale where the score is maximally sensitive to the relative variance of the two subspaces, which guides the particles toward the correct internal distribution.

Phase III: Manifold consolidation

Finally, the *manifold consolidation phase* is characterized by the asymptotic closure of the intermediate gaps and the sharpening of the total manifold gap, indicating the full dimensionality of $\mathcal{M}$. In this final regime, the score assumes the form

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \simeq \frac{1}{t}\left[\Pi - I_N\right]\boldsymbol{x}. \tag{4.7}$$

where $\Pi = F(F^\top F)^{-1}F^\top$ is the projection matrix over the manifold. The component of the score orthogonal to the manifold diverges proportionally to t^{-1}, while the tangent components converge to a constant and become therefore negligible in this regime. This results in the consolidation of the gap corresponding to the manifold dimensionality m and to the (relative) closure of the intermediate gaps. Therefore, in this final phase the dynamics of the model simply projects the particles into the manifold $\mathcal{M}_t \to \mathcal{M}$. In the generative modeling literature, this phenomenon is also known as *manifold overfitting* as the terms corresponding to the internal distribution are negligible [Loaiza-Ganem et al., 2022]. In the next section, we comment on how these three phases can explain why diffusion models are not affected by this phenomenon, namely why we claim that the manifold is *consolidated* rather than *overfitted*.

4.2.1 The geometric phases and manifold overfitting

The probability density of data defined on a manifold is a spiked object $\delta_{\mathcal{M}}(\boldsymbol{x})\rho(\boldsymbol{x})$, where the Dirac-delta $\delta_{\mathcal{M}}(\boldsymbol{x})$ determines the manifold and $\rho(\boldsymbol{x})$ determines its internal density. Likelihood-based generative models are defined by a highly parameterized likelihood function $f(\boldsymbol{x};\boldsymbol{\theta})$, whose parameters are trained by minimizing the loss

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E} f(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{x} \sim p_0(\boldsymbol{x}) \ , \tag{4.8}$$

which maximizes the probability of the data given the model. This maximum likelihood loss is minimized if $f(\boldsymbol{x};\boldsymbol{\theta}) = p_0(\boldsymbol{x})$. A trained likelihood-based model can only fit the true density by having it diverge to infinity on the manifold. Such a divergence makes it impossible to correctly model the internal density $\rho(\boldsymbol{x})$. More problematically, the optimization problem becomes almost insensitive to the internal density $\rho(\boldsymbol{x})$. This phenomenon is called *manifold overfitting* [Loaiza-Ganem et al., 2022], since the trained model fits the manifold while ignoring its internal density, resulting in poor generation.

Our analysis suggests that the temporal dynamics of generative diffusion models overcome this limitation because, for intermediate values of t, the score is still sensitive to the density internal to the manifold, which can be identified through the differences in the tangent singular values. During this manifold coverage phase, the score directs the dispersion of the particles according to these differences, with higher singular values resulting in larger 'opposing force' from the score, which results in smaller displacements of the generated samples along these directions. For t tending to zero, these differences are suppressed due to the divergence of the likelihood, which results in a score function that is orthogonal to the manifold and that is insensitive to $\rho(\boldsymbol{x})$. However, at this stage of generative diffusion, the internal dispersion of the particles has already been affected by the previous coverage phase, and therefore, the manifold overfitting of the score does not negatively affect generation. Instead, the consolidation phase plays the important role of projecting the particles to the support of the data.

4.3 Theoretical analysis of the spectral gaps in linear diffusion models

In this section, we provide our main theoretical results concerning the spectral distribution for random linear subspaces and the relative spectral gaps formulas. We start by reviewing diffusion with data supported on linear manifolds, where the exact score function can be

computed.

4.3.1 Linear manifolds

Normally, the distribution $p_0(\boldsymbol{x})$ is unknown. It is, however, interesting to investigate a tractable special case where the distribution is a multivariate Gaussian defined as in Eq. 4.5 where $F \in \mathbb{R}^{N\times D}$ is an arbitrary projection matrix that implicitly defines the structure of the latent manifold. and $\boldsymbol{z}^\mu \sim \mathcal{N}(0, I_D)$ the latent space vector. In this setting, the distribution can be explicitly written as $p_0(\mathbf{x}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I_D)}\delta(\boldsymbol{x} - F\,\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; 0, FF^\top)$. Therefore, the density of the process at a given time t is again Gaussian and can be computed from

$$p_t(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}} \frac{1}{\sqrt{(2\pi t)^d}}\, e^{-\frac{1}{2t}\|\mathbf{x}-F\mathbf{z}\|^2}. \tag{4.9}$$

While linear manifolds are very simple when compared with real data, they still exhibit a rich and non-trivial phenomenology that elucidates several universal phenomena of diffusion under the manifold hypothesis. In fact, these linear models capture the structure of tangent spaces of smooth manifolds (see Sec. 4.5.1).

The score function of the linear model is solvable analytically since we only have to perform Gaussian integrals, from which we obtain a quadratic form in $\boldsymbol{x}$ that we can rewrite as

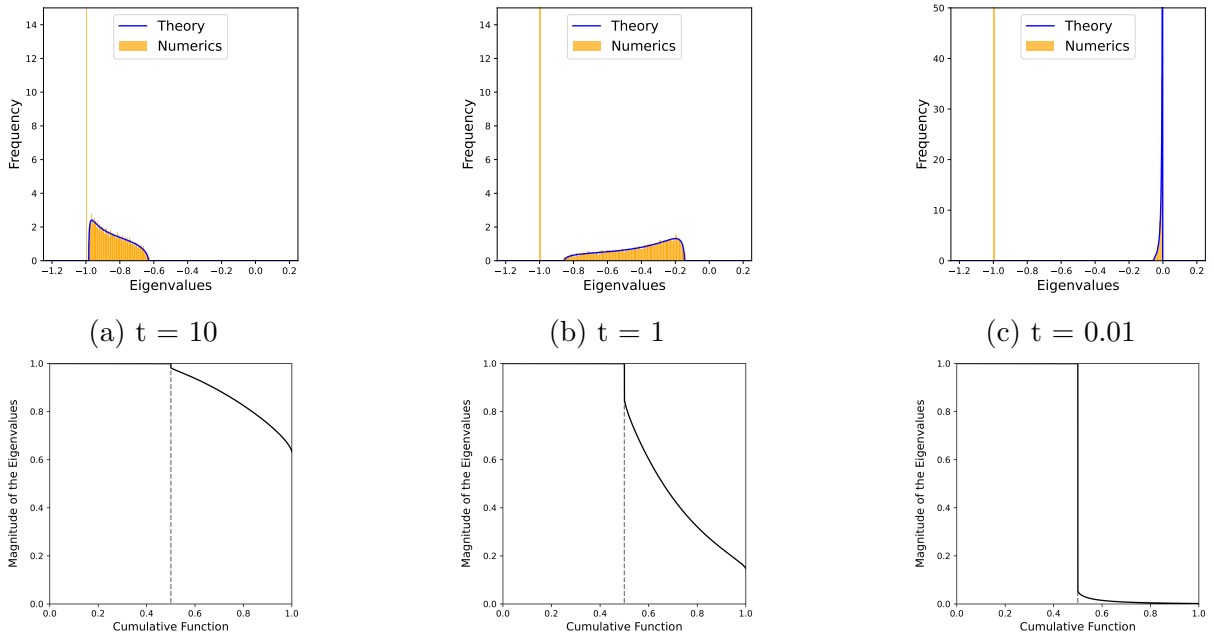
$$\log p_t(\boldsymbol{x}) = \frac{1}{2t}\boldsymbol{x}^\top J_t \boldsymbol{x} + const. \tag{4.10}$$

where the constant does not depend on $\boldsymbol{x}$ and

$$J_t = \frac{1}{t}F\left[I_N + \frac{1}{t}F^\top F\right]^{-1} F^\top - I_N. \tag{4.11}$$

The score function is thus derived as $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = \frac{1}{t}J_t\boldsymbol{x}$. It is then useful to analyze the spectrum of the matrix J_t, since J_t is proportional to the Jacobian of the score function. In fact, since the gradient of the score is orthogonal to the manifold sufficiently close to it, the number of null eigenvalues of J_t will correspond to the manifold dimension, and we should expect to see a drop in the spectrum.

In the following, we provide an outline of our theoretical results on the distribution of spectral gaps in the matrix J_t under random linear manifolds. This choice reflects the fact that the distribution and support of the data are usually not known in advance, and it is therefore important to quantify the statistical variability induced by this uncertainty. We will consider two different distributions for the random projection matrices F: an isotropic Gaussian case, and a multiple-variance one. To ensure tractability, we perform



(a) t = 10 (b) t = 1 (c) t = 0.01

(d) t = 10 (e) t = 1 (f) t = 0.01

Figure 4.3: Spectrum of the eigenvalues of J_t and drop in the dimensionality of the data-manifold estimated from theory in the single-variance case, with $\alpha_D = 0.5, \sigma^2 = 1$. Numerical data are generated with $N = 100$ and collected over 100 realizations of the F matrix.

the analysis in the limit of large N (visible) and D (latent) dimensions while keeping the ratio $\alpha_D = D/N$ constant.

4.3.2 The isotropic case

If the elements of the projection matrix F are sampled as $F_{ij} \sim \mathcal{N}(0, \sigma^2/D)$, we are able to derive analytically the full expression of the distribution of the eigenvalues of J_t. It is given by a simple transformation of the distribution of the eigenvalues of $F^\top F$, which is known to be the Marchenko-Pastur distribution reported in Sec. 4.4.1.

In Fig. 4.3 we show the shape of the spectrum at different times. The bulk of the distribution, inherited from the density of the eigenvalues of $F^\top F$, gradually shifts from left to right in the support. By measuring the cumulative function of the spectrum, one can isolate a drop in the effective dimensionality of the manifold, as also plotted in Fig. 4.3. The step is present at any time in the process, and it is implied by the gap between the left bound of the bulk and the spike in -1. The width of this gap evolves in

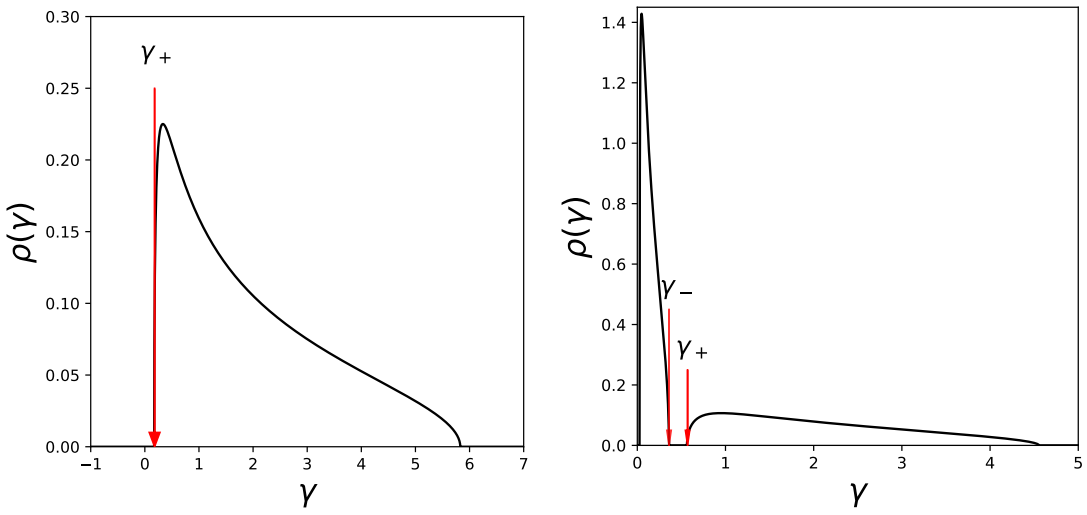


Figure 4.4: Spectrum of the eigenvalues of $F^\top F$ as obtained from random matrix theory with eigenvalues γ_- and γ_+ indicated by red arrows. Control parameters are chosen to be: single variance case (left) $\alpha_D = 0.5$, $\sigma^2 = 1$, double variance case (right) $\alpha_D = 0.5$, $f = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 0.1$.

time according to

$$\Delta_{\text{fin}}^{\text{GAP}}(t;\sigma) = \frac{\sigma^2(1+\alpha_D^{-1/2})^2}{t+\sigma^2(1+\alpha_D^{-1/2})^2}. \tag{4.12}$$

If we name $\gamma_+(\sigma)$ the left bound eigenvalue of the bulk in the spectrum of $F^\top F$ (see Fig. 4.4 left panel), one can recover a more general expression for the gap, being

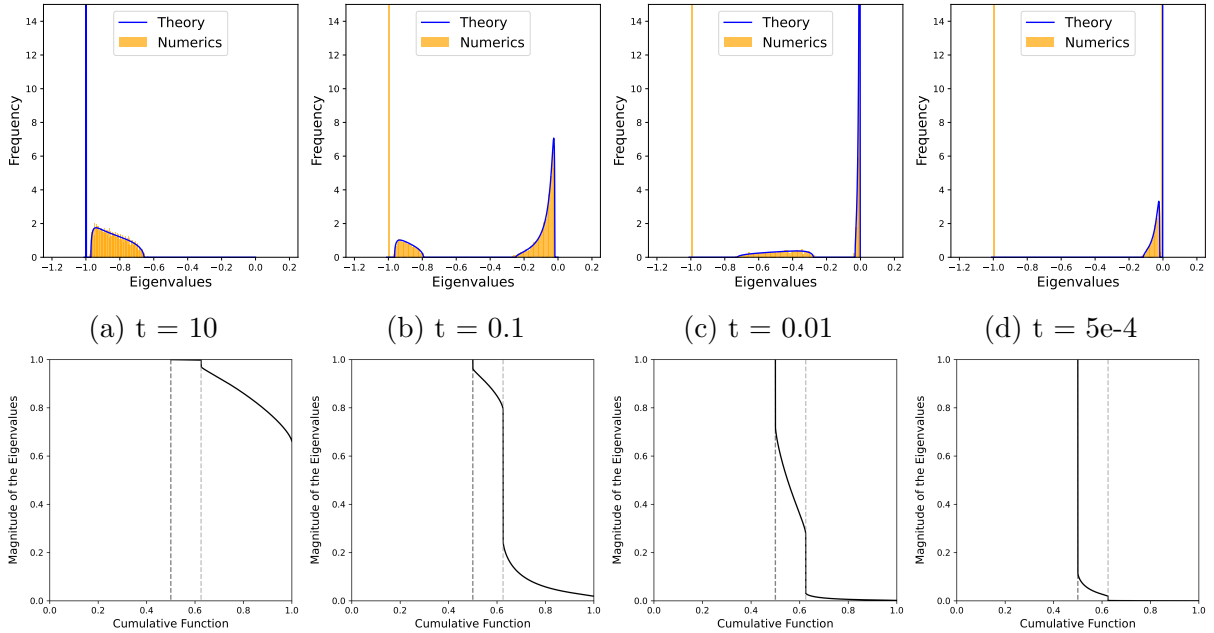
$$\frac{\gamma_+(\sigma)}{t+\gamma_+(\sigma)} = \Delta. \tag{4.13}$$

Hence, we can resolve the gap at a scale Δ at the time

$$t_{\text{in}} = \gamma_+(\sigma)\left(\frac{1-\Delta}{\Delta}\right). \tag{4.14}$$

4.3.3 Intermediate gaps and subspaces with different variances

Another relevant case for our study is the one where we consider a manifold having multiple subspaces with different variances. Here we will focus on the instance of two distinct variances. This scenario is reproduced by considering a number $f \cdot D$ of columns of F to have elements Gaussian distributed with zero mean and variance σ_1^2/D, and the remaining $(1-f)\cdot D$ columns with elements from a Gaussian with variance σ_2^2/D. The



(a) t = 10 (b) t = 0.1 (c) t = 0.01 (d) t = 5e-4

(e) t = 10 (f) t = 0.1 (g) t = 0.01 (h) t = 5e-4

Figure 4.5: Spectrum of the eigenvalues of J_t and drop in the dimensionality of the data-manifold estimated from theory in the double-variance case, with $\alpha_D = 0.5, \sigma_1^2 = 1, \sigma_2^2 = 0.01$, $f = 0.75$. Numerical data are generated with $N = 100$ and collected over 100 realizations of the F matrix.

spectrum of J_t can also be computed in this case, as explained in Sec. 4.4.2. The density function of the eigenvalues shows a transient behavior of the spectrum in the form of an intermediate drop in the estimated dimensionality of the hidden data manifold. This behavior is reported in Fig. 4.5. Even though the expression of the spectral density doesn't have an explicit analytical form and has to be computed numerically, one can adopt a special assumption on the behavior of the density of the eigenvalues of $F^\top F$ to estimate the typical times at which the intermediate drop occurs. Generally speaking, the spectrum of $F^\top F$ can be composed of two separated bulks, as observable in Fig. 4.4 right panel. This happens when σ_2^2 and σ_1^2 are significantly different. In analogy with the single variance scenario, we name γ_+ the left bound of the bulk associated with higher eigenvalues, i.e., with the higher variance, and γ_- the right bound of the bulk associated with smaller eigenvalues, i.e., smaller variance. Most commonly, $\gamma_- = \gamma_-(\sigma_2)$ and $\gamma_+ = \gamma_+(\sigma_1)$. In this case, the gap-forming phenomenology provides for two distinct processes: the emergence of intermediate gaps (i.e., steps in the dimensionality plot) between separated bulks of the spectrum, the opening of a final gap that allows one to infer the dimensionality of the full manifold. The width of the intermediate gap between two bulks can be obtained

from Eq. (4.23) as

$$\Delta^{\mathrm{GAP}}_{\mathrm{inter}}(t;\sigma_1,\sigma_2) = \frac{t}{t+\gamma_-(\sigma_2)} - \frac{t}{t+\gamma_+(\sigma_1)}. \tag{4.15}$$

By imposing $\Delta^{\mathrm{GAP}}_{\mathrm{inter}}(t;\sigma_1,\sigma_2) = \Delta$ one finds the following quadratic form

$$\Delta t^2 + \big[(\Delta-1)\gamma_- + (\Delta+1)\gamma_+\big]t + \Delta\gamma_-\gamma_+ = 0. \tag{4.16}$$

Considering $\Delta \ll 1$ and $\gamma_+ \ll \gamma_-$, the opening time for the intermediate gap can be found by

$$t_{\mathrm{in}}(\Delta) \simeq \Delta^{-1}\gamma_-(\sigma_1), \tag{4.17}$$

that is a reference time at which the gap becomes visible. On the other hand, by assuming the closure time to be close to zero, it can be obtained as

$$t_{\mathrm{fin}}(\Delta) \simeq \Delta\gamma_+(\sigma_2). \tag{4.18}$$

Furthermore, the time at which the gap is maximum in width, and so maximally visible, is located in between t_{in} and t_{fin}. This is the most important time scale for the problem, it is obtained by imposing $\partial\Delta^{\mathrm{GAP}}/\partial t = 0$ and it measures

$$t_{\mathrm{max}} = \sqrt{\gamma_-(\sigma_1)\gamma_+(\sigma_2)}. \tag{4.19}$$

Indeed, when $\sigma_1^2 \gg \sigma_2^2$ the total spectrum can be approximated by a mixture of two separated Marchenko-Pastur distributions, with variances σ_1^2 and σ_2^2, and parameters α_D and γ to be rescaled with respect to f and $(1-f)$. This approximation becomes exact under a slight modification of F, which does not imply any loss of the quality of the description. Now the relevant quantities for the gap become

$$\Delta^{\mathrm{GAP}}_{\mathrm{inter}}(t;\sigma_1,\sigma_2) = \frac{t\left[f\sigma_1^2(1-\sqrt{\frac{1}{f\alpha_D}})^2 - (1-f)\sigma_2^2(1+\sqrt{\frac{1}{(1-f)\alpha_D}})^2\right]}{\left[t+(1-f)\sigma_2^2(1+\sqrt{\frac{1}{(1-f)\alpha_D}})^2\right]\left[t+f\sigma_1^2(1-\sqrt{\frac{1}{f\alpha_D}})^2\right]} \tag{4.20}$$

$$t_{\mathrm{in}}(\Delta) \simeq \Delta^{-1} f\left(1-\sqrt{\frac{1}{f\alpha_D}}\right)^2\sigma_1^2, \qquad t_{\mathrm{fin}}(\Delta) \simeq \Delta(1-f)\left(1+\sqrt{\frac{1}{(1-f)\alpha_D}}\right)^2\sigma_2^2, \tag{4.21}$$

$$t_{\mathrm{max}} = \sqrt{f(1-f)}\left(1-\sqrt{\frac{1}{f\alpha_D}}\right)\left(1+\sqrt{\frac{1}{(1-f)\alpha_D}}\right)\sigma_1\cdot\sigma_2. \tag{4.22}$$

This same analysis can be extended to the more general case where the spectral density

is known to be formed by different detached bulks, associated with hierarchically smaller variances of the data. The evolution of the intermediate gaps in a double-variance diffusion model is reported in Fig. 4.5: notice that $t_{\max} = \mathcal{O}(\sigma_2)$ is consistent with Fig. 4.5b and 4.5f, where the gap was found to be maximum in width. It is worth noting that subspaces with higher variances are the first ones to be explored by diffusion and to be learned by the model. This point suggests that the model is sensitive to the parameters of the probability distribution on the manifold, as recently suggested by other works in the literature.

4.4 Analytical derivation of the Spectrum of J_t

In the previous section, we have described at a high level the theoretical analysis that brought out the phenomenology introduced in Sec. 4.2. This is a technical section, where we report in depth the details of the derivation of the analytical results, namely, of the spectrum of the Jacobian of the score under the linear manifold hypothesis. We consider two cases: the one where the projection matrix F has all entries from a Gaussian distribution with the same variance, and the more interesting case in which F projects on two linear subspaces with different variances and dimensions.

4.4.1 Single variance scenario

We want to compute the spectrum of the matrix in (4.11). Let us first consider the case in which F is an $N \times D$ matrix with Gaussian entries, and call γ the eigenvalues of $FF^\top$. The function that gives the eigenvalues r of J_t as function of γ is

$$r_j = \frac{1}{t}\frac{\gamma_j}{1+\frac{1}{t}\gamma_j} - 1 = -\frac{t}{t+\gamma_j} \tag{4.23}$$

Thus, knowing that the distribution of γ is Marchenko-Pastur, we can obtain the distribution of r

$$\rho_t(r) = -\frac{\alpha_D}{2\pi}\frac{1}{r(1+r)}\sqrt{(r_+ - r)(r - r_-)} + (1-\alpha_D)\,\delta(r+1)\,\theta\left(\alpha_D^{-1} - 1\right) \tag{4.24}$$

for $r \in [r_-(t), r_+(t)]$, with $r_\pm(t) = -\frac{t}{\left(1\pm\frac{1}{\sqrt{\alpha_D}}\right)^2 + t}$.
One could ask whether the bulk of J_t separates from $r = -1$ at a discrete time. This separation corresponds to a drop in the histogram of eigenvalues. According to Eq. (4.24), the bulk is always separated from the spike at finite time t, because $(1+\alpha_D^{-1/2})^2 + t$ for

every t, and the width of the gap is given by $\Delta^{\text{GAP}}(t) = r_-(t) + 1$

$$\Delta^{\text{GAP}}(t) = \frac{(1+\alpha_D^{-1/2})^2}{t + (1+\alpha_D^{-1/2})^2}. \tag{4.25}$$

With respect to the starting spectrum of $F^\top F$, this condition reads

$$\frac{\gamma_+}{t+\gamma_+} = \Delta \tag{4.26}$$

so the time when we see the drop at a scale Δ is $t = \gamma_+^2 \frac{(1-\Delta)}{\Delta}$.

4.4.2 Double variance scenario

We want to compute the spectrum of J_t when $F_{i\mu} \sim \mathcal{N}(0, \sigma_1^2)$ for $\mu < fD/2$ and $F_{i\mu} \sim \mathcal{N}(0, \sigma_2^2)$ for $\mu > (1-f)D/2$, with $f \in [0,1]$. We use the replica method to compute the spectrum of $A = \frac{1}{D} F F^\top$, then with a transform we obtain the spectrum of J_t. In order to obtain the spectrum, we need to compute the expectation of the resolvent of A in the $N \to +\infty$ limit, and to do this, we will rely on the replica method

$$\mathbb{E}\left[g_A(z)\right] = -\frac{2}{N}\frac{\partial}{\partial z}\mathbb{E}\left[\log \frac{1}{\sqrt{\det\left(zI_N - A\right)}}\right] \tag{4.27}$$

$$= -\frac{2}{N}\frac{\partial}{\partial z}\lim_{n\to 0}\mathbb{E}\left[\frac{Z^n - 1}{n}\right] \tag{4.28}$$

with

$$Z^n = \det\left(zI_N - A\right)^{-n/2} \tag{4.29}$$

$$= \int \prod_{a=1}^{n}\prod_{i=1}^{N} \frac{d\phi_i^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_{a=1}^n \sum_{i,j=1}^N \phi_i^a \left(z\delta_{ij} - \frac{1}{m}\sum_\mu F_{i\mu}F_{j\mu}\right)\phi_j^a} \tag{4.30}$$

and taking the expectation

$$\mathbb{E}\left[Z^n\right] = \int \prod_{a,i} \frac{d\phi_i^a}{\sqrt{2\pi}}\, e^{-\frac{z}{2}\sum_a \sum_i (\phi_i^a)^2} \mathbb{E}\left[e^{\frac{1}{2D}\sum_a \sum_\mu (\sum_i \phi_i^a F_{i\mu})^2}\right] \tag{4.31}$$

$$= \int \prod_{a,\mu} \frac{d\eta_\mu^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_a \sum_\mu (\eta_\mu^a)^2} \int \prod_{a,i} \frac{d\phi_i^a}{\sqrt{2\pi}}\, e^{-\frac{z}{2}\sum_a \sum_i (\phi_i^a)^2} \prod_\mu \mathbb{E}\left[e^{\frac{1}{\sqrt{D}}\sum_a \left(\sum_i \phi_i^a F_{i\mu}\right)\eta_\mu^a}\right] \tag{4.32}$$

where, in the last step, we have used the independence of the rows of F and applied a Hubbard-Stratonovic transform. We can separate the product over μ and integrate over the distribution of F

$$\mathbb{E}\left[Z^n\right] = \int \prod_{a,\mu} \frac{d\eta_\mu^a}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{a,\mu}(\eta_\mu^a)^2} \int \prod_{a,i} \frac{d\phi_i^a}{\sqrt{2\pi}} e^{-\frac{z}{2}\sum_{a,i}(\phi_i^a)^2} \tag{4.33}$$

$$\times \prod_{\mu=1}^{fD-1} \mathbb{E}\left[e^{\frac{1}{2\sqrt{D}}\sum_a\left(\sum_i \phi_i^a F_{i\mu}\right)\eta_\mu^a}\right] \prod_{\mu=fD}^{D} \mathbb{E}\left[e^{\frac{1}{2\sqrt{D}}\sum_a\left(\sum_i \phi_i^a F_{i\mu}\right)\eta_\mu^a}\right] \tag{4.34}$$

$$= \int \prod_{a,\mu} \frac{d\eta_\mu^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_{a,\mu}(\eta_\mu^a)^2} \int \prod_{a,i} \frac{d\phi_i^a}{\sqrt{2\pi}} e^{-\frac{z}{2}\sum_{a,i}(\phi_i^a)^2} \tag{4.35}$$

$$\times\, e^{\frac{\sigma_1^2}{2D}\sum_i\sum_{\mu<fD}\left(\sum_a \phi_i^a\eta_\mu^a\right)^2 + \frac{\sigma_2^2}{2D}\sum_i\sum_{\mu\geq fD}\left(\sum_a \phi_i^a\eta_\mu^a\right)^2} \tag{4.36}$$

$$= \int \prod_{a,\mu} \frac{d\eta_\mu^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_{a,\mu}(\eta_\mu^a)^2} \int \prod_{a,i} \frac{d\phi_i^a}{\sqrt{2\pi}}\, e^{-\frac{z}{2}\sum_{a,i}(\phi_i^a)^2}) \tag{4.37}$$

$$\times\, e^{\frac{\sigma_1^2}{2D}\sum_{ab}\left(\sum_i \phi_i^a\phi_i^b\right)\left(\sum_{\mu<fD}\eta_\mu^a\eta_\mu^b\right) + \frac{\sigma_2^2}{2D}\sum_{ab}\left(\sum_i \phi_i^a\phi_i^b\right)\left(\sum_{\mu>fD}\eta_\mu^a\eta_\mu^b\right)}. \tag{4.38}$$

Introducing $q_{ab} = \frac{1}{N}\sum_i \phi_i^a\phi_i^b$

$$\mathbb{E}\left[Z^n\right] = \int \prod_{a,b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi} \int \prod_{ai} \frac{d\phi_i^a}{\sqrt{2\pi}}\, e^{-\sum_{ab}\frac{1}{2}\hat{q}_{ab}\left(dq_{ab}-\sum_i \phi_i^a\phi_i^b\right) - \frac{z}{2}\sum_a\sum_i(\phi_i^a)^2} \tag{4.39}$$

$$\times \left[\int \prod_{a=1}^{n} \frac{d\eta^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_a(\eta^a)^2 + \frac{\sigma_1^2}{2\alpha_D}\sum_{ab} q_{ab}\eta^a\eta^b}\right]^{fD} \tag{4.40}$$

$$\times \left[\int \prod_{a=1}^{n} \frac{d\eta^a}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_a(\eta^a)^2 + \frac{\sigma_2^2}{2\alpha_D}\sum_{ab} q_{ab}\eta^a\eta^b}\right]^{(1-f)D} \tag{4.41}$$

$$= \int \prod_{a,b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\, e^{nd\Phi(q,\hat{q})} \tag{4.42}$$

with

$$\Phi(q,\hat{q}) = -\frac{1}{2n}\sum_{a,b} q_{ab}\hat{q}_{ab} + G_S(\hat{q}) + f\alpha_D G_E(q,\sigma_1) + (1-f)\alpha_D G_E(q,\sigma_2) \tag{4.43}$$

where

$$G_S(\hat{q}) = \frac{1}{n} \log \int \prod_{a=1}^{n} \frac{d\phi^a}{\sqrt{2\pi}} \, e^{-\frac{z}{2}\sum_a (\phi^a)^2 + \frac{1}{2}\sum_{ab} \hat{q}_{ab}\phi^a\phi^b} \tag{4.44}$$

$$G_E(q, \sigma) = \frac{1}{n} \log \int \prod_{a=1}^{n} \frac{d\eta^a}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\sum_a (\eta^a)^2 + \frac{\sigma^2}{2\alpha_D}\sum_{ab} q_{ab}\eta^a\eta^b} \tag{4.45}$$

Using the replica symmetric ansatz $q_{ab} = \delta_{ab} q$, $\hat{q}_{ab} = -\delta_{ab}\hat{q}$

$$G_S(\hat{q}) = -\frac{1}{2}\log(z + \hat{q}) \tag{4.46}$$

$$G_E(q, \sigma) = -\frac{1}{2}\log(1 - \frac{\sigma^2 q}{\alpha_D}). \tag{4.47}$$

Putting all together, we have

$$\Phi(z) = \frac{1}{2}\hat{q}q - \frac{1}{2}\log(z + \hat{q}) - f\frac{\alpha_D}{2}\log\left(1 - \frac{\sigma_1^2 q}{\alpha_D}\right) - (1 - f)\frac{\alpha_D}{2}\log\left(1 - \frac{\sigma_2^2 q}{\alpha_D}\right). \tag{4.48}$$

The integral can be evaluated by the saddle point method

$$q = \frac{1}{z + \hat{q}} \tag{4.49}$$

$$\hat{q} = -f\frac{\alpha_D \sigma_1^2}{\alpha_D - \sigma_1^2 q} - (1 - f)\frac{\alpha_D \sigma_2^2}{\alpha_D - \sigma_2^2 q}. \tag{4.50}$$

We can find the Stieltjes transform

$$\mathbb{E}[g_A(z)] = -2\alpha_D \frac{\partial \Phi(z)}{\partial z} \tag{4.51}$$

$$= \alpha_D q^*(z) \tag{4.52}$$

where q^* is found by solving the saddle point equation

$$zq^3 + q^2\left(\alpha_D - 1 - \frac{z\alpha_D}{\sigma_1^2} - \frac{z\alpha_D}{\sigma_2^2}\right) + \\ + q\left(\frac{\alpha_D^2}{\sigma_1^2\sigma_2^2}(z - f\sigma_1^2 - (1 - f)\sigma_2^2) + \frac{\alpha_D}{\sigma_1^2} + \frac{\alpha_D}{\sigma_2^2}\right) - \frac{\alpha_D^2}{\sigma_1^2\sigma_2^2} = 0. \tag{4.53}$$

The asymptotic distribution of eigenvalues can be obtained from the Stieltjes transform

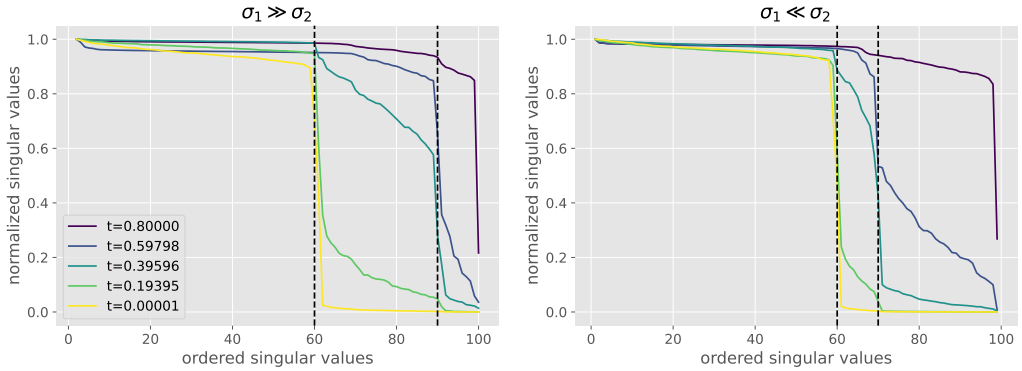


Figure 4.6: Ordered singular values obtained with the trained score model, for different variances on the subspaces. Data are generated according to the linear-manifold model with $N = 100$ and $D = 40$ and $\sigma_1^2 = 1$, $\sigma_2^2 = 0.01$, $f = 0.75$; Center: $\sigma_1^2 = 0.01$, $\sigma_2^2 = 1$, $f = 0.75$. The neural network is trained as prescribed in Appendix A and spectra are measured according to Appendix B.

as

$$\rho(\gamma) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\left(g_A(\gamma - i\epsilon)\right) \tag{4.54}$$

$$= \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\Phi'(\gamma - i\epsilon) \tag{4.55}$$

$$= \frac{1}{\pi} \alpha_D \lim_{\epsilon \to 0^+} \mathrm{Im}[q^*] \tag{4.56}$$

Once the density of the eigenvalues is computed, one can perform the same change of variables described in Sec. 4.4.1 for the case of single variance, and obtain the density $\rho_t(r)$ for the eigenvalues of J_t.

Fig. 4.5 reports the evolution in time of the spectral density, as well as its cumulative function $f = 0.75$. The cumulative function has been used to estimate the formation of the intermediate gaps to compare with the experiments for the estimation of the data manifold dimension.

4.5 Experiments with synthetic linear datasets

We first measure the spectrum of the singular values of the Jacobian of a score function trained through a neural network on a linear manifold data-model generated by two variances σ_1^2, σ_2^2 as described in Sec. 4.3.3. Results are reported in Fig. 4.6. The opening of the gaps is consistent with the theory for the exact score: an intermediate gap associated with the subspace with higher variance first opens; subsequently, the gap relative to the

lower variance subspace, which here corresponds to the final gap, opens. We can infer the dimensions of the subspaces by subtracting the location of the dashed vertical lines from N. We underline the fact that higher-variance subspaces are learned first by repeating the experiment after swapping the values of the variances. Eventually, Fig. 4.7 reports the same experiments where variances are uniformly generated in the interval $[10^{-2}, 1]$: it is evident that the N intermediate gaps are now a continuous line, as it is expected to be in more realistic natural datasets.

We will now compare the gaps computed analytically with ones obtained from real neural networks trained on the same linear data model. The results of such a comparison are presented in Fig. 4.8, and they show a good agreement between the ordered distribution of the singular values obtained through empirical methods and the relative analytical counterpart, computed through the replica method. The opening of the predicted intermediate gaps signals the right dimension of the linear subspaces, as verified from the experiments. One can notice from the figure that the analytical profile shows the shape of a sharp step between the zero value along the x-axis and the first appearing gap: this shape is related to the Dirac-delta spike that the spectrum of the eigenvalues presents at -1 (see Sec. 4.4.2 for details about the spectrum); on the other hand, the numerical profile looks different in the same region, and this behavior is associated with the absence of the spike in the distribution of the singular values, that leaves room to a separated bulk from the other ones. This evident discrepancy between theory and experiment is probably due to the final configuration of the trained neural network and leaves space for

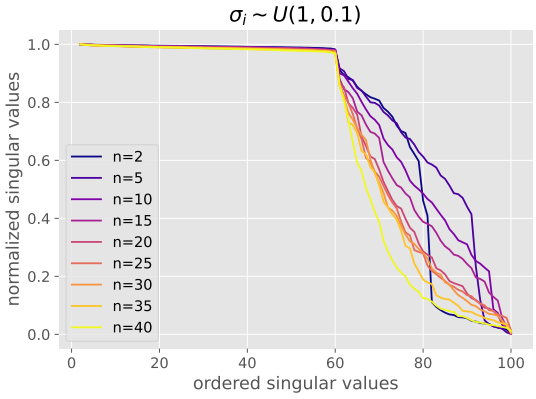


Figure 4.7: Ordered singular values obtained with the trained score model, for different variances on the subspaces. Data are generated according to the linear-manifold model with $N = 100$ and $D = 40$ and a progressive number n of variances sampled uniformly between 10^{-2} and 1, each one assigned to a fraction $f = 1/n$ of matrix columns. The neural network is trained as prescribed in Appendix A and spectra are measured according to Appendix B.

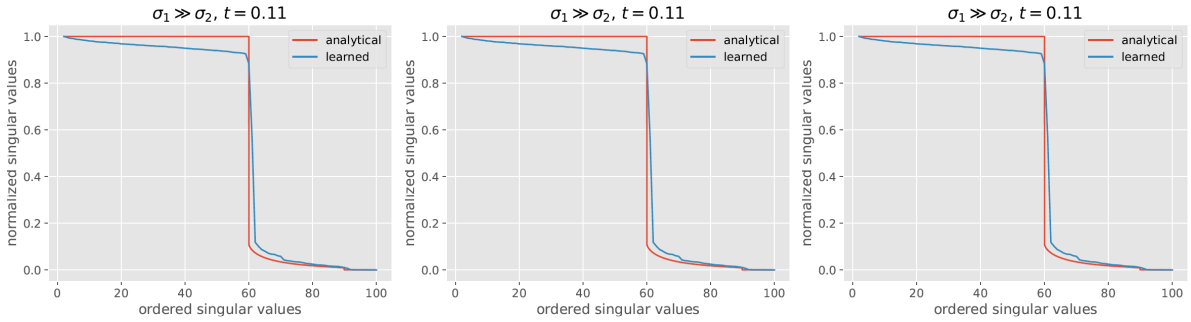


Figure 4.8: Comparison between spectra obtained with the trained score model and with the numerical analysis, for different variances on the subspaces. Data are generated according to the linear-manifold model with $N = 100$ and $D = 40$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.01$, $f = 0.75$; from left to right, the spectra are evaluated at time $t \approx 0.45$, $t \approx 0.3$, $t \approx 0.11$. The neural network is trained as prescribed in Appendix A and spectra are measured according to Appendix B.

further investigations.

4.5.1 Remarks on the linear manifold model hypothesis

In order to apply the replica method from statistical physics and compute the distribution of the eigenvalues of the Jacobian of the score function, we have constrained ourselves to the simpler case of a linear manifold, i.e., $g(\boldsymbol{x}) = \boldsymbol{x}$ (see Section 4.3.1). Indeed, this might sound as a limitation to the reproducibility of our results to more realistic data-sets, e.g., natural images. However, we claim that, for ordinary diffusive diffusion models in the variance exploding setup, the theory developed for the linear model still applies to non-linear manifold instances, due to the following reasons:

1. The diffusive trajectories at large t, where the *trivial* phase occurs, are sampled by a probability distribution that is smooth, due to the Gaussian kernel implied by the reverse diffusion process. As a consequence, the stable latent set defined in Eq. (4.2) will be approximately linear.

2. The diffusive trajectories at small t, where the *manifold coverage* and *manifold consolidation* phases occur, explore a region contained into a ball of radius proportional to $\sqrt{t}$, that is supposed to be smaller than then the inverse local curvature of the manifold.

We now compare the distribution of the singular values of the Jacobian obtained from the linear manifold model and the non-linear one. Specifically, we will repeat the plot in Fig. 4.6 with a toy dataset generated as $\boldsymbol{\xi}^\mu = F\tilde{\boldsymbol{z}}^\mu$, with $\tilde{\boldsymbol{z}}^\mu = \boldsymbol{z}^\mu / \|\boldsymbol{z}^\mu\|$. As a consequence,

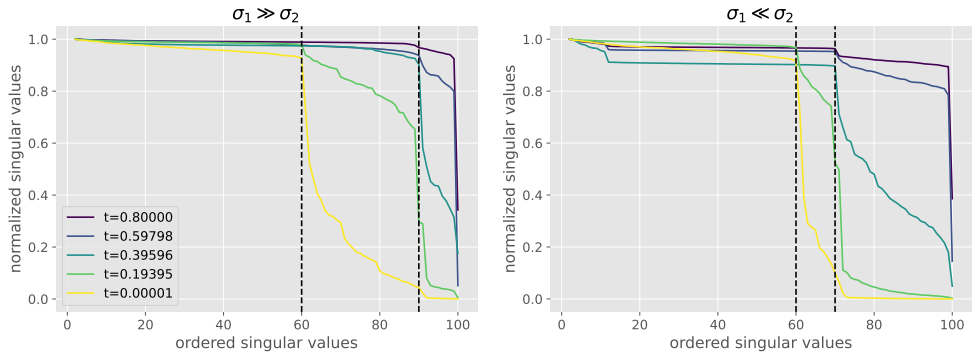


Figure 4.9: Ordered singular value spectrum estimated experimentally from the analysis of a data-set living on a non-linear manifold built according to Sec.4.5.1, where parameters are chosen as in Fig. 4.6. The neural network is trained as prescribed in Appendix A, and spectra are measured according to Appendix B.

the new data will now live on a $(N-1)$-dimensional ellipsoid inscribed in the original N-dimensional hyperplane. The results are reported in Fig. 4.9 and they show no evident discrepancy between the linear and non-linear case, as predicted by our argument. The only small difference consists of a slight delay in the time of the gap phenomenology that is present in the non-linear manifold data, which impedes the final gap from fully opening at the smallest observable time.

4.6 Experiments with natural image datasets

While our theoretical analysis is limited to linear random manifold models, several qualitative aspects of their phenomenology can be observed in networks trained on natural images. Fig. 4.10 shows the temporal evolution of the spectrum estimated numerically from the Jacobian of models trained on MNIST, Cifar10, and CelebA. Details about the training process are provided in Appendix A and B. In these experiments, we can recognize the three geometric phases of diffusion described above:

Trivial phase: at large times (i.e., from $t = 200$ to 100) the ordered spectrum of the singular values appears flat, suggesting the diffusive motion to be Brownian in the ambient space.

Manifold coverage phase: at intermediate times, the spectrum shows a clear trace of multiple simultaneous opening gaps. The shape of the curves is, however, different from the controlled scenario showed in Section 4.3, due to two different reasons: similar latent variances associated to different latent dimensions imply a smoothing of the curve, as displayed by Fig. 4.7; the local eigenspace of the data is complex and hard to model

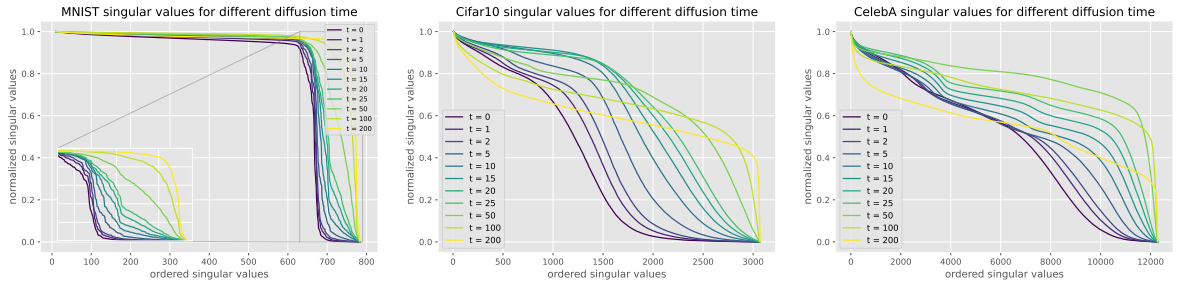


Figure 4.10: Jacobian spectra of diffusion models trained on MNIST, Cifar10, and CelebA. The neural network is trained as prescribed in Appendix A and spectra are measured according to Appendix B.

microscopically: for instance, the pixellated appearance of Cifar10 images might explain the scarce emergence of the gaps, while the larger gap structure showed by CelebA might be due to correlations among the latent variances.

Manifold consolidation phase: at small times (below $t = 2$) we finally see only the full manifold gap open and progressively sharpening, in the sense that many singular values become exactly zero. The spectra of this phase are the only ones analyzed in Stanczuk et al. [2024] and used to estimate the manifold dimensionality. For instance, we see from Fig. 4.10 (left panel) that, at the end of the reverse process, the network trained on MNIST shows a latent dimension $D \sim 100$ that is coherent with Stanczuk et al. [2024].

We conclude that, although our analysis focuses on the local structure of the data manifold (or the stable latent set in diffusion time), it is effectively supported by experiments on real-world complex datasets.

4.7 Conclusions

We investigate the latent geometry of generative diffusion models under the manifold hypothesis. For this purpose, we analyze the spectrum of eigenvalues (and singular values) of the Jacobian of the score function, whose discontinuities (gaps) reveal the presence and dimensionality of distinct sub-manifolds. Using a statistical physics approach, we derive the spectral distributions and formulas for the spectral gaps, and we compare these theoretical predictions with the spectra estimated from trained networks. Our analysis reveals the existence of three distinct qualitative phases during the generative process:

1. trivial phase;

2. manifold coverage phase, where the diffusion process fits the distribution internal to the manifold;

3. consolidation phase where the score becomes orthogonal to the manifold and all particles are projected on the support of the data.

This 'division of labor' between different time scales provides an elegant explanation of why generative diffusion models are not affected by the manifold overfitting phenomenon that plagues likelihood-based models, since the internal distribution and the manifold geometry are produced at different time points during generation.

Losing Dimensions

That's me in the corner
That's me in the spotlight, losing my ...

R.E.M.

In this second part, we study geometric memorization. Our theoretical and experimental findings indicate that different tangent subspaces are lost due to memorization effects at different critical times and dataset sizes, which depend on the local variance of the data along their directions. Perhaps counterintuitively, we find that, under some conditions, subspaces of higher variance are lost first due to memorization effects. This leads to a selective loss of dimensionality, where some prominent features of the data are memorized without a full collapse on any individual training point. We validate our theory with a comprehensive set of experiments on networks trained both in image datasets and on linear manifolds, which result in a remarkable qualitative agreement with the theoretical predictions.

4.8 Background on Geometric Memorization

In the machine learning field, a lot of attention has been devoted to the study of memorization in generative diffusion. In general, we say that a model memorizes when it generates samples from the training set, instead of new ones from the same distribution. Pidstrigach [2022] was the first to show that DMs are capable of learning low-dimensional structure in $\mathbb{R}^N$ and that this manifold learning capability is a driver of memorization: a model capable of learning 0-dimensional manifolds can memorize the training data. While the first definition of memorization is about reproducing data points, the phenomenon can be further more complex. Somepalli et al. [2023] noticed that some features, as foreground or background, present in the dataset can be reproduced without copying the entire image, in a phenomenon that they call reconstructive memory. Webster [2023] refers to similar instances as template verbatim. These phenomena have been categorized under the term *reconstructive* memory: the model can replicate "objects" that are are semantically equivalent to their source object without being pixel-wise identical. In a recent paper, Leigh Ross et al. [2025] and proposed the *manifold memorization hypothesis*, where they analyze memorization in terms of the relationship between the dimensionalities of the true data manifold and the manifold learned by the model. This categorizes memorized

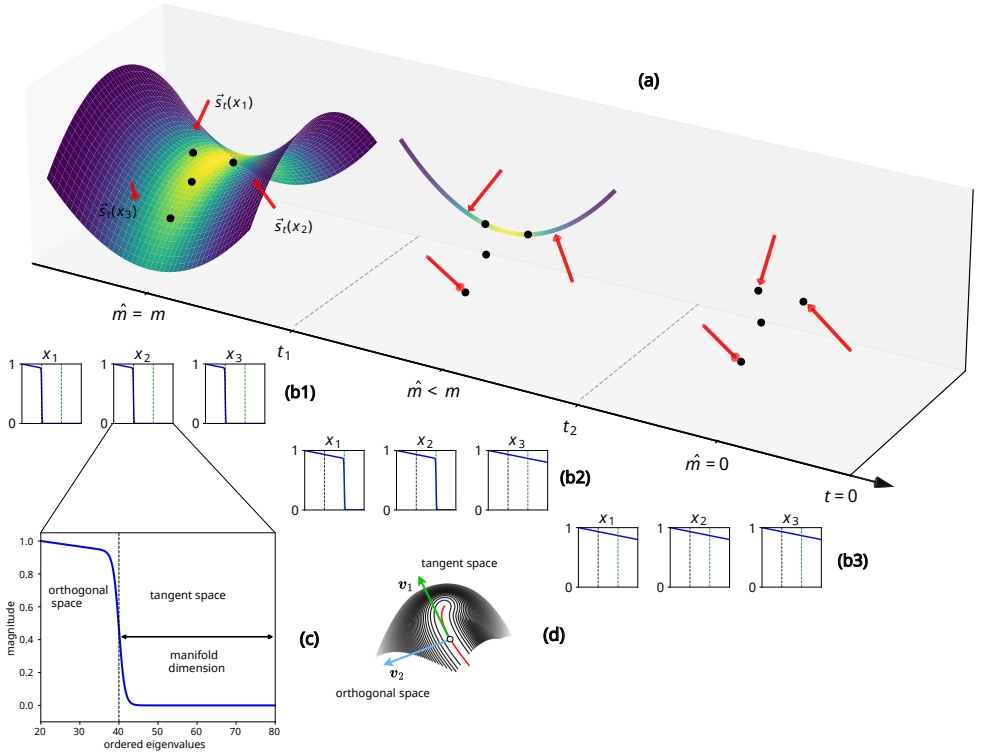


Figure 4.11: Pictorial representation of the geometric memorization phenomenon. Panel (a): data-points lie on a D-dimensional manifold in the ambient space of dimension N, with $D < N$. The target distribution is a 2-variate Gaussian density function with diagonal covariance matrix and variances $\sigma_1^2 > \sigma_2^2$, having the manifold as support. We represent four data points: two being fully aligned with the direction having larger dispersion σ_1^2, the remaining two being fully aligned with the orthogonal direction. When $t > t_1$ the score function vector field locally projects the sampling process on the manifold (as described in Stanczuk et al. [2024], Ventura et al. [2025]). We expect a neural network trained on a very large training set to maintain this behavior until very small times t_1, ensuring generalization. When $t_1 > t > t_2$ the model starts memorizing fractions of the manifold: points aligned with the larger variance are fully memorized and become point-wise attractors in the vector field; the points aligned with the smaller variance sub-manifold are not yet memorized, and the surrounding score function vectors are orthogonal to the sub-manifold of dimensions $\hat{D} < D$. When $t < t_2$ all points are attractors in the vector field, and the original D-dimensional manifold gets shattered into 0-dimensional sub-manifolds, i.e. disconnected points in the ambient space. Panels (b): ordered eigenspectrum measured in positions $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ according to the Normal Bundle (NB) method described in Section 4.1. When $t > t_1$ spectral gaps opened on the black dashed line, signal an estimated manifold dimension $\hat{D} = D$ all over in space. When $t_1 > t > t_2$ the method applied to $\boldsymbol{x_1}, \boldsymbol{x}_2$ estimates $\hat{D} < D$ with spectral gaps opened on the green dashed line. The method in $\boldsymbol{x}_3$ estimates a 0-dimensional manifold, i.e. the closest data point is fully memorized, because it lies on the larger variance sub-manifold. When $t < t_2$ spectral gaps estimate $\hat{D} = 0$ all over in space. Panel (c): detailed interpretation of the way the NB method estimates the manifold latent dimension. Panel (d): visualization of the orthogonal and tangent subspaces with respect to a latent manifold.

data into two types: overfitting-driven memorization and memorization driven by the underlying data distribution.

In all these works, there is the common idea that memorization can happen separately in different subspaces/directions of the ambient space of the data. Here we try to give a theoretical basis to this conjecture, proving that indeed the phenomenon of memorization is bound to the geometry of the data.

4.8.1 Generative Diffusion Models and Memorization

Diffusion models fund their functioning on the properties of stochastic dynamic processes. Let us consider a Brownian process where an ensemble of particles $\boldsymbol{x}_0$ starts from an initial distribution $p_0(\boldsymbol{x})$ and evolves through the stochastic differential equation

$$\mathrm{d}\boldsymbol{x}_t = \mathrm{d}\boldsymbol{W}_t \tag{4.57}$$

where $\mathrm{d}\boldsymbol{W}_t$ is a standard Wiener process. In generative modeling applications, $\boldsymbol{x}_0$ is usually chosen to be the distribution of data such as natural images. Eq. (4.57) is solved by the formal expression:

$$p(\boldsymbol{x}_t, t) = \mathbb{E}_{\boldsymbol{x}_0 \sim p_0}\left[\frac{1}{(2\pi t)^{N/2}} e^{-\frac{\|\boldsymbol{x}_t - \boldsymbol{x}_0\|_2^2}{2t}}\right], \tag{4.58}$$

where N is the dimensionality of the ambient space. The *target distribution* $p_0(\boldsymbol{x})$ can be recovered by using a reversed diffusion process [Anderson, 1982]. At large time t_f, we start from a sample $\boldsymbol{x}_{t_f} \sim \mathcal{N}(0, t_f I_d)$, and let it evolve through the backward process defined by

$$\mathrm{d}\boldsymbol{x}_t = -\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{W}_t \tag{4.59}$$

backward from t_f to $t_0 = 0$. In the machine learning literature, the term $s(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is commonly referred to as the score function. These formulas are given according to what is known as the variance-exploding framework in the generative diffusion literature. However, all results can be ported directly to the variance-preserving (i.e. Ornstein–Uhlenbeck) case.

Given a training set $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^P\}$ sampled from $p_0(\boldsymbol{x})$, we can obtain a neural network approximation of the score function by training a noise-prediction network $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ (parameterized by $\boldsymbol{\theta}$) using the empirical denoising score matching objective [Hyvärinen and Dayan, 2005, Vincent, 2011, Ho et al., 2020]. The network $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is trained to predict the added noise $\boldsymbol{\epsilon}$ from a noisy state $\boldsymbol{x}_t = \boldsymbol{x}_0 + \sqrt{t}\boldsymbol{\epsilon}$. The estimated score is then given by

$\hat{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = -\frac{\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x},t)}{\sqrt{t}}$. Learning $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}$ instead of $\hat{s}_{\boldsymbol{\theta}}$ has the advantage of maintaining finite the output of the network for small $t \to 0$, where we known that the magnitude of the score becomes infinite outside of the support of the data.

Another approach consists of making an educated guess over $p_0(\boldsymbol{x})$ by modeling it as the empirical distribution of the data, i.e.

$$p_0(\boldsymbol{x}) = \frac{1}{P} \sum_{\mu=1}^{P} \delta\left(\boldsymbol{x} - \boldsymbol{\xi}^{\mu}\right). \tag{4.60}$$

At this point, from Eq. (4.58), $p(\boldsymbol{x}_t, t)$ is a mixture of Gaussians that can be studied analytically at any time t. Specifically, it is now implied that all diffusive trajectories must collapse in one of the spikes contained in Eq. (4.60), i.e., the system must memorize the data points at $t = 0$.

As we have already done in chapter 3, one can address memorization using a statistical mechanics approach. Specifically, one can use the Random Energy Model formalism to determine when the backward diffusion process is captured by the basin of attraction of one of the training points, signaling memorization. This analysis, however, does not take into account the effect of the latent manifold supporting the target distribution, as well as the local heterogeneities in the distribution of the data features. We conjecture that these geometric aspects might affect the memorization transition, implying different memorization times along different directions in the ambient space.

4.8.2 The Spectral Gap Analysis

In the previous part of this chapter, we have analyzed the spectrum of the Jacobian of the score function to investigate learning properties of the diffusion model. The reader is remanded to Sec. 4.1 for a more detailed introduction to the topic. The scope of this spectral gap analysis is to probe the geometry of the diffusive trajectory and the way it evolves in time. From this procedure, it is possible to infer whether the trajectory is constrained to sample along certain specific directions in the ambient space due, for instance, to latent structures emerging in the sampling space.

Consider a given point $\boldsymbol{x}^*$ in the ambient space, and analyze the effect of adding a small perturbation vector $\boldsymbol{p}$ with magnitude in the order of $\sqrt{t}$:

$$s(\boldsymbol{x}^* + \boldsymbol{p}, t) \approx J(\boldsymbol{x}^*, t)\ \boldsymbol{p}\ , \tag{4.61}$$

where $J(\boldsymbol{x}^*, t)$ is the Jacobian of the score function. Performing a spectral decomposition

of $J(\boldsymbol{x}^*, t)$, we can realize that perturbations aligned to the tangent space correspond to zero eigenvalues of the Jacobian. This was the founding idea implied by Stanczuk et al. [2024] to estimate the latent manifold dimension using diffusion models. In the first part of this chapter (and in the corresponding paper Ventura et al. [2025]), the analysis of the spectrum of the Jacobian of the score is used to determine a series of sub-gaps that open corresponding to sub-spaces of the stable latent set. In particular, sub-spaces associated with higher local variance are mapped to sub-gaps that open earlier. This sheds some light on the way diffusion models learn the distribution of data supported on manifolds, but does not investigate the problem of memorization. Indeed, we have analyzed the spectrum of the Jacobian of the *true* score function, the one we can obtain knowing the true density of the data. This has been shown to align well with the phenomenology of trained diffusion models. However, if one wants to study memorization, we should compare instead a theory for the Jacobian of the *empirical* score with a trained model in a memorization phase. This is precisely what we do here in Sec. 4.9. In this case, one considers the *smoothed Jacobian matrix* $J(\boldsymbol{x}, t)$, whose elements are defined as

$$J_{ij}(\boldsymbol{x}, t) = \left[s_i(\boldsymbol{x} + \sqrt{t}\boldsymbol{e}_j, t) - s_i(\boldsymbol{x}, t) \right] / \sqrt{t} \ , \tag{4.62}$$

where $\boldsymbol{e}_j$ is a vector in an orthonormal basis set, which converges to the exact Jacobian of the score for $t \to 0$.

4.8.3 Data model

For our theoretical analysis, we are going to adopt a slight simplification of the hidden manifold model described in Sec. 2.4.1. In particular, we will generate data points to have coordinates $\boldsymbol{\xi}_i^\mu \sim \mathcal{N}(0, \sigma_i^2)$ with $\sigma_i^2 > 0$ when $i = 1, .., D$ and $\sigma_i^2 = 0$ when $i = D+1, .., N$. This case corresponds to choosing a linear activation function in the hidden manifold model and subsequently aligning the ambient space with the eigenvectors of the matrix $\frac{1}{\sqrt{D}} F^\top F$. The choice of the linear manifold is moved by computational reasons, has already been motivated in Sec. 4.5.1. This model is particularly realistic at small diffusion times, where diffusive trajectories are very close to the manifold, as explained in [Stanczuk et al., 2024, Ventura et al., 2025].

4.9 Theory of geometric memorization

In this section, we present a theory of geometric memorization based on the heuristic analysis of the spectrum of the smoothed Jacobian. First, we adopt the analogy of the empirical density with the Random Energy Model introduced in the statistical physics literature Lucibello and Mézard [2024], Biroli et al. [2024], Achilli et al. [2025] to compute a *geometric* memorization time. Then, this result is used to obtain an approximate formula for the empirical score and its Jacobian. Lastly, we derive a formula for the eigenvalues that depends explicitly on the directions in the ambient space.

4.9.1 Geometric memorization time

In Chapter 3 we have derived the collapse time for manifold data, and we have shown that collapse and condensation coincide for a typical trajectory. We now adopt a geometric approach and aim to derive these times at a generic point $\boldsymbol{x}$ in $\mathbb{R}^N$.

The statistical behavior of the empirical score can be analyzed in the large P limit by interpreting Eq. (4.67) as proportional to the partition function of a Random Energy Model (REM) [Biroli et al., 2024, Lucibello and Mézard, 2024], which offers a simple model of disordered thermodynamic systems. We have introduced the REM in Sec. 2.3.2 and applied it to the problem of deriving the collapse time in Sec. 3.2. In summary, each energy level ε_μ is associated with a data point $\boldsymbol{\xi}^\mu$ in the training set and its energy depends on its Euclidean distance with the current state $\boldsymbol{x}_t$ [Ambrogioni, 2025], with the energy given by

$$\varepsilon_\mu(\boldsymbol{x}) = -\frac{1}{2}\|\boldsymbol{\xi}^\mu\|^2 + \boldsymbol{x}\cdot\boldsymbol{\xi}^\mu \tag{4.63}$$

which leads to the partition function

$$Z_P(\boldsymbol{x}, t) = \sum_{\mu=1}^{P} e^{-\frac{1}{t}\varepsilon_\mu(\boldsymbol{x})} \tag{4.64}$$

where the time parameter t is analogous to the temperature of the system, which can be used to express the weights as a Boltzmann distribution: $w_\mu(\boldsymbol{x}, t) = \frac{1}{Z_P(\boldsymbol{x},t)} e^{-\frac{1}{t}\varepsilon_\mu}$. Since the empirical score is a Boltzmann average according to Eq. (4.64), studying its fluctuation under the random sampling of the data allows us to quantify the deviations from the exact score due to memorization effects. In our case, the energy levels are distributed according to

$$p(\varepsilon; \boldsymbol{x}) = \int_{\mathbb{R}^n} \delta\left(\varepsilon + \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \boldsymbol{x}\cdot\boldsymbol{\xi}\right) \mathrm{d}p_0(\boldsymbol{\xi}) \tag{4.65}$$

For small values of t and large dataset sizes, the empirical score can be shown to be self-averaging, meaning that it is insensitive to the specific sampling of the training points, resulting in generalization of the underlying distribution. More formally, from the physical theory of REMs [Derrida, 1981], we know that, at the asymptotic limit of $N \to \infty$, the statistical system specified by Eq. (4.67) undergoes a phase transition that separates a self-averaging high-temperature regime from a *condensation* regime where Boltzmann averages depend on a small (i.e., sub-exponential) fraction of energy levels [Mézard et al., 2009]. In Sec. 4.9.2, we show that, for N much smaller than P, the condensation time for linear manifold data is, to a leading exponential order, equal to

$$t_c(\boldsymbol{x}) = \sqrt{\frac{N}{2 \log P} \left(\frac{r_{4,\sigma}}{2} + \omega^2(\boldsymbol{x}) \right)} \,, \tag{4.66}$$

which demarcates the diffusion time when the empirical score becomes susceptible to fluctuations introduced by the random sampling of the dataset. In the formula, the term $r_{4,\sigma} = \sum_{i=1}^{N} \sigma_i^4 / N$ captures the fluctuations in the norm of the data, while the directional quantity $\omega^2(\boldsymbol{x}) = \sum_{i=1}^{N} x_i^2 \sigma_i^2 / N$ is the *variance density* along the direction $\boldsymbol{x}$. If each dimension has equal variance σ^2, the directional variance density is just σ^2, which implies that the critical condensation time depends linearly on the dimensionality but only logarithmically on the number of data points. This implies that in the isotropic case, in order to avoid condensation, an exponential number of data points is needed. However, if only α_D dimensions have non-zero variance, it is straightforward to see that the exponential dependency will scale with α_D instead of D. More generally, the exponential scaling depends on the total variance $D\ \omega(\boldsymbol{x})$, which implies that it is realistic to learn high-dimensional spaces as far as most of these dimensions have vanishing variance. As we shall see, the balance between these two quantities plays a crucial role in geometric memorization effects.

The derivation of this geometric condensation time allows us to write an approximate expression for the empirical score. The empirical distribution at time t in the variance exploding framework is

$$p_t^P(\boldsymbol{x}) = \frac{1}{P\sqrt{(2\pi t)^d}} \sum_{\mu=1}^{P} e^{-\frac{\|\boldsymbol{x} - \boldsymbol{\xi}^\mu\|^2}{2t}} \,. \tag{4.67}$$

From the empirical distribution, we can write down the empirical score:

$$\nabla \log p_t^P(\boldsymbol{x}) = \sum_{\mu=1}^{P} w_\mu(\boldsymbol{x}, t) \, \frac{\boldsymbol{\xi}^\mu - \boldsymbol{x}}{t} \, , \tag{4.68}$$

where the weight

$$w_\mu(\boldsymbol{x}, t) = \frac{p(\boldsymbol{\xi}^\mu \mid \boldsymbol{x})}{\sum_{\nu=1}^{P} p(\boldsymbol{\xi}^\nu \mid \boldsymbol{x})} \tag{4.69}$$

is the posterior probability of the pattern $\boldsymbol{\xi}^\mu$ given the noisy state $\boldsymbol{x}$, where the possible states are restricted to the empirical set. This estimator is consistent, meaning that its bias approaches the true score for $P \to \infty$.

The random sampling of the dataset introduces statistical fluctuations that we can quantify by considering the estimator variance, which for large P can be approximated as

$$\mathrm{Var}\big[\nabla \log p_t^P(\boldsymbol{x})\big] \approx \frac{\mathrm{Var}(\boldsymbol{x}_0 \mid \boldsymbol{x})}{\mathbb{E}\Big[\tilde{P}_t(\boldsymbol{x})\Big]}, \tag{4.70}$$

where $\mathrm{Var}(\boldsymbol{x}_0 \mid \boldsymbol{x})$ is the true posterior variance and

$$\tilde{P}_t(\boldsymbol{x}) = \left(\sum_{\mu=1}^{P} w_\mu^2(\boldsymbol{x}, t) \right)^{-1} \leq P \tag{4.71}$$

is the effective number of data points used to estimate the score. When $t \to 0$, we always have that $\tilde{P}_t(\boldsymbol{x}) \to 1$, because the empirical score always fully memorizes in this limit. However, the empirical score exhibits generalization when the expected value is larger than the standard deviation induced by $\tilde{P}_t(\boldsymbol{x})$. Note that $\tilde{P}_t(\boldsymbol{x})$ is a function of the state $\boldsymbol{x}$ and that, consequently, the fluctuations in the empirical score depend on the location $\boldsymbol{x}$. This property is fundamental in our analysis of geometric memorization.

From standard REM calculations (see Sec. 4.9.3), we can express the effective number of data points used to estimate the score at $\boldsymbol{x}$ at time t as

$$\tilde{P}_t(\boldsymbol{x}) = \min\left(P, \frac{t}{1 - t_c^{-1}(\boldsymbol{x})} \right) \, . \tag{4.72}$$

where we introduced the minimum operator heuristically to account for the finite size of the system. The exact asymptotic theory is recovered for $P \to \infty$. Note that, since these quantities scale to the leading exponential order, they are therefore neglected quantities that scale sub-exponentially in P.

4.9.2 Condensation time for positional REM

For simplicity, we will perform the analysis for coordinate-aligned linear manifolds. Consider N-dimensional normally distributed vector-valued data $\boldsymbol{\xi}^\mu$ where each component ξ_k^μ follows a centered normal distribution with variance σ_k^2. In the linear manifold case, the number $N-D$ of these variances is equal to zero, meaning that the distribution spans a D-dimensional linear manifold. The number of data points is taken to be exponential in the size of the ambient space, i.e. $P = \exp(\alpha N)$, with $\alpha > 0$. Notice that σ_k^2 corresponds to the eigenvalues of $F^\top F$ in the random projection model, and we assume we have changed the coordinate system. Let us take a fixed $\boldsymbol{x}$. Hence, in the variance exploding framework, we have

$$p_t(\boldsymbol{x}) = \frac{1}{P\sqrt{2\pi t}^N} \sum_{\mu=0}^{P} e^{-\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{\xi}^\mu\|^2} \tag{4.73}$$

$$= \frac{1}{P\sqrt{2\pi t}^N} e^{-\frac{\|\boldsymbol{x}\|^2}{2t}} \sum_{\mu=0}^{P} \exp\left(-\frac{1}{2t}\|\boldsymbol{\xi}^\mu\|^2 + \frac{1}{t}\boldsymbol{x}\boldsymbol{\xi}^\mu\right). \tag{4.74}$$

It is useful, at this point, to recover the Random Energy Model (REM), which we have introduced in its original form in Sec. 2.3.2, and already applied to diffusion models in Chapter 3 following [Biroli et al., 2024]. The REM consists of a collection of energy levels $\{\varepsilon_\mu\}_{\mu \le P}$ that interact with an external heat-bath at an inverse temperature β. The energy levels are random variables generated from a probability density function $p(\varepsilon \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ can be some parameters of the model and source of disorder for the system. The thermodynamics of the model shows a condensation phase at a critical temperature β_c that shares similarities with glassy transitions in spin-glass models. Condensation, in turn, is analogous to memorization in diffusion models. The main thermodynamic quantities, such as the condensation temperature, can be fully recovered starting from the *partition function* of the system, given by

$$Z_P(\beta) = \sum_{\mu=1}^{P} e^{-\beta\varepsilon_\mu}. \tag{4.75}$$

We can now map our diffusion model into a REM by redefining

$$\beta(t) = 1/t, \tag{4.76}$$

and

$$\varepsilon_\mu(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\xi}^\mu\|^2. \tag{4.77}$$

We call this model *positional* REM, because the occurrence of condensation will depend on a position in the N-dimensional Euclidean space. Standard REM calculations are now performed to compute the free energy of the model and then the condensation time. The moment generating function of the energies is

$$\begin{aligned}
\zeta(\lambda) &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}_{\boldsymbol{\xi}} e^{-\frac{\lambda}{2t}\|\boldsymbol{\xi}\|^2 + \frac{\lambda}{t}\boldsymbol{x}\boldsymbol{\xi}} && (4.78)\\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log \int \frac{dy_i}{\sqrt{2\pi\sigma_i^2}} \exp -\frac{y_i^2}{2}\left(\frac{1}{\sigma_i^2} + \frac{\lambda}{t}\right) + \frac{\lambda}{t} x_i y_i && (4.79)\\
&= \lim_{N\to\infty} \frac{1}{N} \left[-\frac{1}{2} \sum_{i=1}^{N} \log\left(1 + \lambda\frac{\sigma_i^2}{t}\right) + \frac{\lambda^2}{2t^2} \sum_{i=1}^{N} \frac{x_i^2\sigma_i^2}{1 + \lambda\frac{\sigma_i^2}{t}} \right] && (4.80)
\end{aligned}$$

The derivative of the zeta function is

$$\zeta'(\lambda) = \lim_{N\to\infty} \frac{1}{N} \left[-\frac{1}{2t} \sum_i \frac{\sigma_i^2}{1 + \lambda\frac{\sigma_i^2}{t}} + \frac{\lambda}{t^2} \sum_i \frac{x_i^2\sigma_i^2}{1 + \lambda\frac{\sigma_i^2}{t}} - \frac{\lambda^2}{2t^3} \sum_i \frac{x_i^2\sigma_i^4}{(1 + \lambda\frac{\sigma_i^2}{t})^2} \right]. \tag{4.81}$$

At large times, $\zeta(\lambda)$ and $\zeta'(\lambda)$ become respectively

$$\zeta(\lambda) = -\frac{\lambda}{2t} r_{2,\sigma} + \frac{\lambda^2}{4t^2} r_{4,\sigma} + \frac{\lambda^2}{2t^2}\omega^2(\boldsymbol{x}), \tag{4.82}$$

$$\zeta'(\lambda) = -\frac{\lambda}{2t} r_{2,\sigma} + \frac{\lambda^2}{2t^2} r_{4,\sigma} + \frac{\lambda^2}{t^2}\omega^2(\boldsymbol{x}). \tag{4.83}$$

Where

$$\begin{aligned}
r_{2,\sigma} &= \lim_{N\to\infty} \frac{1}{N} \sum_i \sigma_i^2 && (4.84)\\
r_{4,\sigma} &= \lim_{N\to\infty} \frac{1}{N} \sum_i (\sigma_i^2)^2 && (4.85)\\
\omega^2(\boldsymbol{x}) &= \lim_{N\to\infty} \frac{1}{N} \sum_i (x_i)^2\sigma_i^2. && (4.86)
\end{aligned}$$

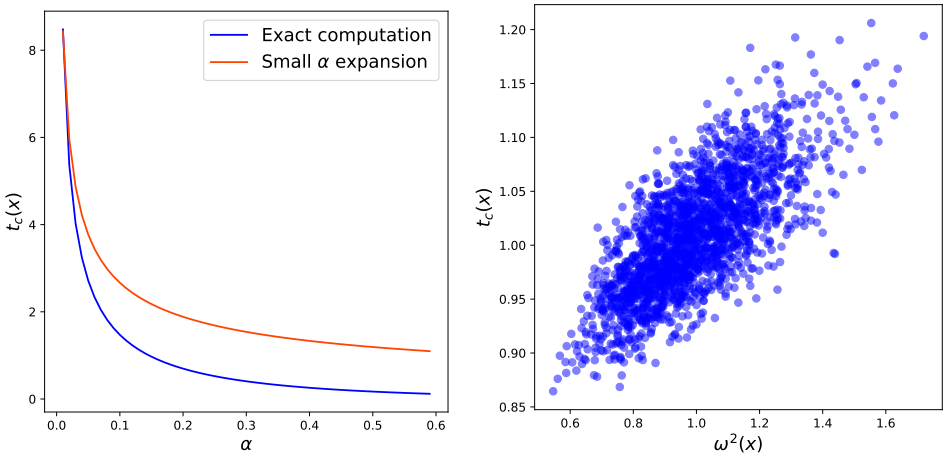


Figure 4.12: Condensation time as a function of position x computed according to the REM calculations. Left: for one position $\boldsymbol{x}$ in an ambient space of dimension $N = 100$ and one matrix F of dimensions 100×50 (with $D = 50$ dimension of the latent space), we show the comparison between the exact calculation of the positional condensation time and the approximated version that is fully explicit in the directional variance $\omega^2(\boldsymbol{x})$. Both $\boldsymbol{x}$ and F are generated according to a Gaussian process with zero mean and unitary variance. Right: we generate 2000 random positions $\boldsymbol{x}$ around the origin of the ambient space of dimension $N = 100$; the latent space dimension is $D = 50$ and $\alpha = 0.15$; we show the dependence of the exact positional condensation time as a function of $\omega^2(\boldsymbol{x})$, showing a qualitatively similar behavior with respect to the approximated expression of t_c.

The condition for the condensation time is $\alpha + \zeta(1) - \zeta'(1) = 0$, from which we obtain

$$t_c(\boldsymbol{x}) = \sqrt{\frac{\frac{r_{4,\sigma}}{2} + \omega^2(\boldsymbol{x})}{2\alpha}}. \tag{4.87}$$

As is clear from the formula, this time depends on the variance $\omega^2(\boldsymbol{x})$ along the direction of $\boldsymbol{x}$. This implies that, when $\boldsymbol{x}$ is aligned to a linear sub-manifold with higher variance, condensation around this state will happen earlier, leading to a decrease in the estimated commonality of the latent manifold. Fig. 4.12 shows a comparison between the exact approach for computing $t_c(x)$ (i.e. using Eqs. (4.78), (4.81)) and the small α expansion (i.e. Eq. (4.66)), showing a good qualitative agreement between the two quantities at all values of α. The right panel of the same figure also displays a strong dependence of the exactly computed condensation time.

4.9.3 Participation ratio

In the condensation phase, the empirical score is dominated by the (quenched) fluctuations in the data distribution. First, we can introduce the participation ratio

$$Y(\beta, \boldsymbol{x}) = \frac{Z(2\beta, \boldsymbol{x})}{Z(\beta, \boldsymbol{x})^2} \ . \tag{4.88}$$

This thermodynamic quantity can be roughly interpreted as the inverse of the number of energy levels with non-vanishing weights. In the condensation phase, this will be a finite number, while it becomes infinite in the high temperature phase. In the thermodynamic limit and for $\beta(t) \geq \beta_c(t)$, the participation ratio of our REM model is given by

$$\mathbb{E}[Y(\beta, \boldsymbol{x})] = 1 - \frac{\beta_c(t, \boldsymbol{x})}{\beta(t)} \ , \tag{4.89}$$

which implies that the number of data points that contribute to the score function at $\boldsymbol{x}$ is

$$\tilde{P} = e^{\alpha \tilde{N}(\beta, \boldsymbol{x})} = 1/Y(\beta, \boldsymbol{x}) = \frac{\beta(t)}{\beta(t) - \beta_c(t, \boldsymbol{x})} \ . \tag{4.90}$$

Note that this number tends to one for $\beta(t) \to \infty$, meaning that in the low-time limit the score depends on a single pattern.

4.9.4 Spectral analysis of the empirical Jacobian

The large P analysis outlined in the previous sections gives us a description of the fluctuations in the empirical score as a function of the state $\boldsymbol{x}$. The spatial dependency of these fluctuations ultimately depends on the data distribution $p_0(\boldsymbol{x})$, which outlines a rich geometric landscape that interacts in a complex way with the spatial variations in the exact score $\nabla \log p_t(\boldsymbol{x})$.

To study the effect of these spatially non-homogeneous random fluctuations in the spectrum, we start from an approximate formula for the empirical score obtained by restricting the average to only $\tilde{P}_t(\boldsymbol{x})$ *active samples*:

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \approx \frac{1}{\tilde{P}_t(\boldsymbol{x})} \sum_{\mu=1}^{\tilde{P}_t(\boldsymbol{x})} \frac{\boldsymbol{\xi}^\mu - \boldsymbol{x}}{t} \tag{4.91}$$

where the active samples $\boldsymbol{\xi}^\mu$ are sampled from the posterior distribution $p(\boldsymbol{x}_0 \mid \boldsymbol{x}; t)$. If we

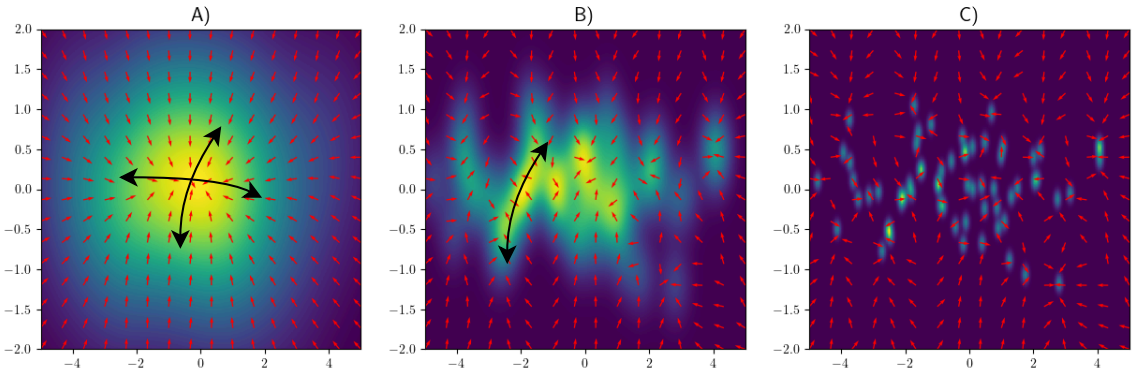


Figure 4.13: Visualization of the geometric memorization, implying the typical dimensionality loss phenomenon. Manifold subspaces with higher variance are lost due to 'condensation' (i.e., memorization). Panels A, B, and C show the score estimated from a bivariate distribution with unequal variances for $t = 1$, $t = 10^{-1}$, and $t = 10^{-2}$, respectively. The red arrows show the empirical score, while the heat-map visualizes the sampling probability distribution.

consider the data model introduced in 2.4.1 when g is the identity, i.e., a linear manifold model, Eq. (4.91) follows a Gaussian distribution, since the posterior is itself Gaussian. In the following, for the sake of simplicity, we will assume that F is a diagonal $N \times N$ with diagonal entries σ_k, with $\sigma_k = 0$ for $N - D$ indices. This corresponds to a rotation to the basis of eigenvectors of $F^\top F$. If we assume that the fluctuations in the score are uncorrelated for a separation of the order of $\sqrt{t}$, we can quantify the statistical variability of the (smoothed) Jacobian in Eq. (4.62) through the formula

$$J_{ij}(t) \sim \mathcal{N}\left(-\delta_{ij}\left(t+\sigma_i^2\right)^{-1}, \frac{\sigma_i^2}{t\left(t+\sigma_i^2\right)}\left[\phi(t,\mathbf{0}) + \phi(t, \boldsymbol{e}_j\sqrt{t})\right]\right), \tag{4.92}$$

where we defined the function $\phi(t, \boldsymbol{x}) = \max\left(1/P, t^{-1} - t_c^{-1}(\boldsymbol{x})\right)$. The expected value of this expression is just the Jacobian of the exact score, which determines the opening of the spectral gaps as explained in the first part of this chapter. On the other hand, in this model, gaps can close due to the variance term. We can see this phenomenon qualitatively by considering the singular values spectrum of the expected value of Eq. (4.92):

$$\bar{s}_i = \sqrt{\frac{1}{\left(t+\sigma_i^2\right)^2} + \sum_{k=1}^{d} \frac{\sigma_k^2}{t^2\left(t+\sigma_k^2\right)^2}\left[\phi(t,\mathbf{0}) + \phi(t,\boldsymbol{e}_i\sqrt{t})\right]^2}. \tag{4.93}$$

Remember that we see a gap in the sorted spectrum if there is a large difference between

two consecutive sorted singular values s_k and s_{k+1}. This gap can disappear mainly for two reasons:

1. The first one is that $\phi(t, \boldsymbol{e}_k\sqrt{t})$ is larger than $\phi(t, \boldsymbol{e}_{k+1}\sqrt{t})$. This case is directional, as it depends on the direction of perturbations $\boldsymbol{e}_k$, and $\boldsymbol{e}_{k+1}$ and it leads to the selective suppression of a particular subspace.

2. The second one instead is non-directional: it induces a synchronous suppression of all gaps and leads to complete memorization. It happens if the contribution of these variance terms makes the contribution of the expected value negligible

We are more interested in the first case. This phenomenon of selective memorization is visualized in Fig. 4.13 for a two-dimensional distribution. For linear Gaussian, the closing times are determined by the critical time $t_c^{-1}(\boldsymbol{x})$, which, has we have seen in the previous section, is itself depends on the constant term $r_{4,\sigma} = \sum_{i=1}^{N} \sigma_i^4/N$ and on the *directional* term $\omega^2(\boldsymbol{x}) = \sum_{i=1}^{N} x_i^2\sigma_i^2/N$. This latter term is proportional to the variance along the subspace spanned by $\boldsymbol{x}$ and plays a crucial role in determining the differential disappearance of different subspaces at different times. Perhaps counter-intuitively, the subspace spanned by $\boldsymbol{e}_k$ is more vulnerable to memorization when $\omega^2(\boldsymbol{e}_k)$ is large. Therefore, subspaces that are more prominent in the distribution of the data and that emerge earlier during the diffusion process are also more vulnerable to memorization in the later stages of diffusion. This corresponds to the form of feature memorization suggested in [Leigh Ross et al., 2025].

Figs. 4.14 report the eigenvalues sampled from the distribution in Eq. (4.92) ordered from the largest to the smallest in magnitude, to reproduce the typical plots obtained by the improved NB method. Figures show drops in magnitude at different eigenvalue numbers, proving the evolution of the local latent dimensionality of the underneath manifold, i.e. a coherent deformation of the score function field in the ambient space. The latent dimensionality decreases in diffusion time, as predicted by the theory.

On the other hand, Figs. 4.14 also report the ordered eigenvalues yet computed through Eq. (4.93). The curves are consistent with the behavior of the empirical eigenvalues plotted in Figs. and also the experimental curves obtained through the improved NB method.

4.9.5 Full derivation of the empirical Jacobian spectrum

We can relate this random energy analysis to the spectra of Jacobian eigenvalues using a heuristic argument. In the linear manifold example, the Jacobian of the true score

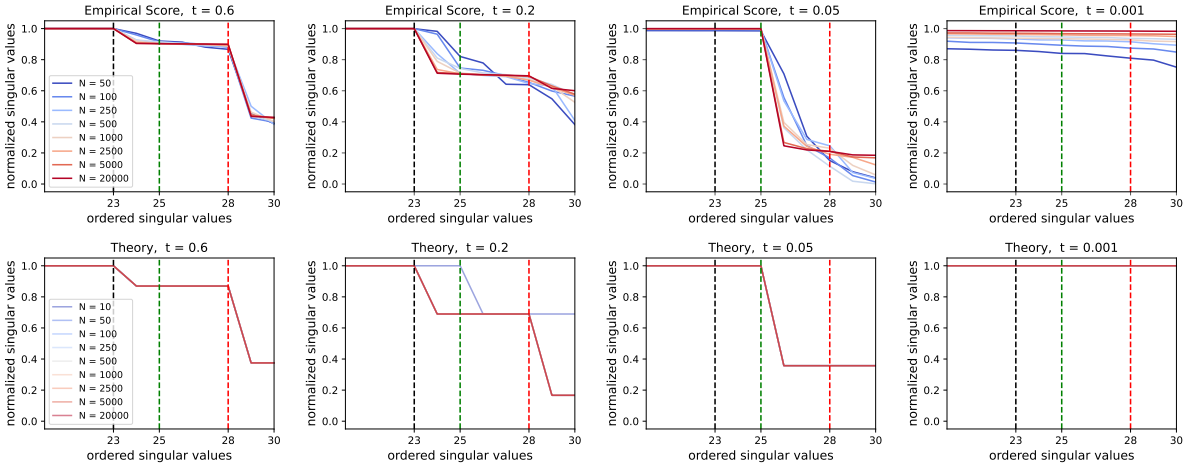


Figure 4.14: Ordered singular values of the Jacobian of the empirical score function in the case of the linear data model, sampled from Eq. (4.92) (first row) and computed from Eq. (4.93) (second row). Choice of the parameters: $N = 30$, $D = 7$ with a subspace associated to a variance $\sigma_1^2 = 1$ of dimension $D_1 = 2$ and another subspace with variance $\sigma_2^2 = 0.3$ and dimension $D_2 = 5$. Different lines are associated to different sizes of the training set P, as reported in the legend. Measures have been averaged over 30 realizations of the experiment.

function at $t = 0$ is diagonal with eigenvalues equal to $-1/\sigma_k^2$. This results in spectral gaps when different sub-spaces have different variances. For a finite value of the inverse temperature $\beta(t) = 1/t$, the eigenvalues are $-1/\sigma_k^2 - \beta$. After the critical condensation time, the empirical score gives a good approximation of the true score.
In this phase, the score is dominated by approximately $e^{\alpha\tilde{N}(\beta,\boldsymbol{x})} = \frac{\beta(t)}{\beta(t)-\beta_c(t,\boldsymbol{x})}$, leading to the expression

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \approx \frac{\beta}{e^{\alpha\tilde{N}(\beta,\boldsymbol{x})}} \sum_{\mu=1}^{e^{\alpha\tilde{N}(\beta,\boldsymbol{x})}} (\boldsymbol{\xi}^\mu - \boldsymbol{x}) \tag{4.94}$$

where $\boldsymbol{\xi}^\mu \sim p(\boldsymbol{\xi}^\mu \mid \boldsymbol{x}, \beta) \propto e^{-\boldsymbol{\xi}^T\left(\Lambda^{-1}+\beta I_N\right)\boldsymbol{\xi}/2+\beta\boldsymbol{x}\cdot\boldsymbol{\xi}}$. Therefore, the empirical score approximately follows the distribution

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \sim \mathcal{N}\left(-M(\beta)\boldsymbol{x}, \beta(\Lambda^{-1} + \beta I_N)^{-1} \max(0, \beta - \beta_c(\boldsymbol{x}))\right) \ . \tag{4.95}$$

where $M(\beta) = \beta(\Lambda^{-1} + \beta I_N)^{-1}\Lambda^{-1}$ and Λ being the diagonal matrix collecting the variances σ_k^2 and we used the fact that $e^{\alpha\tilde{N}(\beta,\boldsymbol{x})} = \beta/(\beta - \beta_c(\boldsymbol{x}))$. The minimum in the formula is due to the fact that, for $\beta < \beta_c$, an exponentially large number of patterns participate in the estimation of the score, which leads to a complete suppression of the variance. On

the other hand, the variance of the empirical score estimator diverges for $\beta \to \infty$. In fact, during condensation, the fluctuations in the random sampling of the data points are not suppressed due to the small number of non-vanishing weights.
We can finally estimate the distribution of the eigenvalues estimated from the empirical Jacobian matrix. Let us set ourselves on $\boldsymbol{x} = \mathbf{0}$ and perturb along the directions of the eigenvectors of $F^\top F$. We estimate the elements of the Jacobian of the score function with respect to the orthogonal direction $\boldsymbol{e}_j$ using a perturbative approach, i.e.

$$J_{ij}(\beta) \approx \sqrt{\beta}\left(\partial_{x_i} \log p_t(\boldsymbol{e}_j/\sqrt{\beta}) - \partial_{x_i} \log p_t(\mathbf{0})\right). \tag{4.96}$$

Using Eq. (4.95), we can then write an approximate distribution for the elements of the Jacobian as

$$J_{ij}(\beta) \sim \mathcal{N}\left(-\beta\delta_{ij}\left(1+\beta\sigma_i^2\right)^{-1}, \beta^2\left(\sigma_i^{-2}+\beta\right)^{-1}\left[\phi(\beta,\mathbf{0}) + \phi(\beta,\boldsymbol{e}_j/\sqrt{\beta})\right]\right) \ . \tag{4.97}$$

where we assumed that the fluctuations in $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{e}_j/\sqrt{\beta})$ are independent from the fluctuations in $\nabla_{\boldsymbol{x}} \log p_t(\mathbf{0})$ and $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{e}_k/\sqrt{\beta})$ for all ks. In this expression, we introduced the function

$$\phi(\beta, \boldsymbol{x}) = \max\left(0, \beta - \beta_c(\boldsymbol{x})\right) \ . \tag{4.98}$$

We can now recover the singular values of $J(\beta)$ as minus the square roots of the eigenvalues of $J(\beta)^\top J(\beta)$. In general, the matrix $J(\beta)^\top J(\beta)$ can have a complex spectral distribution. An approximate formula for the singular values of $J(\beta)$ is

$$s_i \approx -\sqrt{\beta^2(1+\beta\sigma_i^2)^{-2} + \beta^4 \sum_{k=1}^{N}\left(\sigma_k^{-2}+\beta\right)^{-2}\left[\phi(\beta,\mathbf{0}) + \phi(\beta,\boldsymbol{e}_i/\sqrt{\beta})\right]^2}. \tag{4.99}$$

To obtain this formula, we write J as

$$J = A + B \tag{4.100}$$

where A is a diagonal matrix corresponding to the mean of Eq. (4.97), while B corresponds to the variance. Therefore, $J^\top J$ becomes

$$J^\top J = A^\top A + A^\top B + B^\top A + B^\top B. \tag{4.101}$$

This expression is dominated by the two symmetric terms, so we can write

$$J^\top J \approx A^\top A + B^\top B. \tag{4.102}$$

Then, the term $A^\top A = A^2$ is, of course, still diagonal, while the term $B^\top B$ is diagonally dominant. Calling $C = \sum_{ik} B_{ik} B_{ik}$, we can approximate the singular values as $\sqrt{A^2 + C^2}$, obtaining Eq. (4.99). Note, however, that the distribution of the spectrum does not concentrate exactly to Eq. 4.99 in the large P. Nevertheless, Eq. 4.99 gives an accurate picture of the qualitative behavior, as shown in Sec. 4.10.2.
These results also show that in some regimes Eq. 4.99 is more in agreement with the numerical empirical score than the correct spectrum of Eq. 4.97, which is likely due to the fact that Eq. 4.97 overestimates the fluctuations by ignoring the correlations of the score at different points.

4.10 Experiments

In this section, we test the theory of geometric memorization with simulations and experiments with trained neural networks. The main experimental results have already been summarized in the previous section, in the description of Fig. 4.16. Here we expand the experimental settings, providing more evidence to support the theoretical analysis.

4.10.1 Diffusion networks trained on linear manifold data

We first perform an experiment on a controlled environment, i.e. by training a deep network on data generated by a simple model. Data have been generated from the synthetic manifold model described in Section 4.8.3. We have considered a linear manifold model with $N = 30$, $D = 7$, and, in particular, the two subspaces that make the manifold have dimension $D_1 = 2$ with variance $\sigma_1^2 = 1.0$ and $D_2 = 5$ with variance $\sigma_2^2 = 0.3$. We have subsequently trained the neural network on synthetic datasets with an increasing size and applied the improved NB method described in 4.1 to estimate the ordered eigenspectrum of the score function Jacobian around the position $\boldsymbol{x} = 0$ (which we recall to be part of the latent data manifold). The results are reported in Fig. 4.15, where the latent dimension of the subspace spanned by the largest variance σ_1^2 can be estimated by the position of the red dashed line, the dimension of the subspace spanned by the smallest variance σ_2^2 can be extracted by the green dashed line, and eventually the full latent dimension of the manifold is associated to the black dashed line. The left panel in the Figure shows a net-

work trained on a large dataset, which is capable to generalize even at small times. Such model is driven by the true score function that be geometrically reconstructed through the theory developed in Ventura et al. [2025]. This network first estimates the smaller latent dimension D_2, relative to the sub-manifold spanned by the larger variance σ_2^2, by opening a gap on the red dashed line, and then opens a last gap on the line relative to the true latent manifold dimension D. On the other hand, an overfitting model does not manage to geometrically estimate the correct latent dimensionality of the true manifold, because it starts memorizing the data, driven by the empirical score function. This scenario can be appreciated from both central and right panels in Fig. 4.15. Specifically, the model represented in the central panel first reproduces the generalizing behavior showed by the left panel, but then opens a gap on the green dashed line at smaller times. This gap is not predicted by a theory based on the true score as the one in Ventura et al. [2025]. On the other hand it can predicted by a theory based on the empirical score function, and it is a signature of the progressive reduction of the support of the diffusive trajectory that we name geometry memorization. The right panel shows a even more extreme memorization instance, where the model does not manage to open the full manifold dimension gap and instead clearly opens the smaller dimensionality gap. Both the central and right panels would show a flattening of the curves, signaling a zero-dimensional manifold, for very small times, that have not been included in the plot.

4.10.2 Comparing Experiments with the Theory

Fig. 4.16 reports the main results from the analysis of geometric memorization on the synthetic manifold model described in Section 4.8.3. At this point we compare the time evolution of the score function Jacobian computed through three different manners around the position $\boldsymbol{x} = 0$:

- Our theory based on mapping the empirical score on a statistical mechanics model, as described in Sec. 4.9.
- A neural network model trained on synthetic manifold data.
- A numerical simulation of generative diffusion drive by the empirical score defined in Eq. 4.92 in Sec. 4.9.

Starting from large diffusion times, Fig. 4.16 shows the first gap opening on the red dashed line, signaling that the score function is fitting the m_1-dimensional sub-manifold with larger variance σ_1^2 coherently with the theory presented in Ventura et al. [2025].

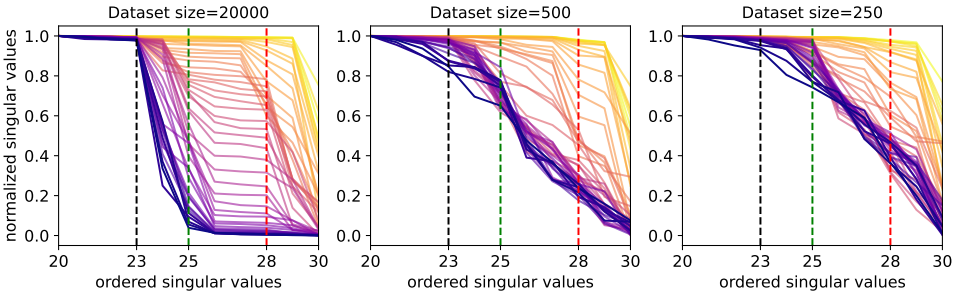


Figure 4.15: The evolution in diffusion time of the ordered singular values of the Jacobian of the score function estimated by a Deep Neural Network trained a linear manifold model. The parameters for the model are $N = 30$, $D = 7$, $\log(P)/N = 0.23$ with subspaces associated to variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.3$ with dimensions $D_1 = 2$ and $D_2 = 5$ respectively. Lighter colors are associated to larger times in the color map. Different panels have been realized by training a deep network on growing sizes of the training dataset and applying the improved NB method in $\boldsymbol{x} = 0$. In the Left panel the size of the data-set is sufficiently large to allow the network to detect the true latent dimensionality of the data manifold. In the Central panel the network first starts to reproduce the true dimensionality of the manifold and then enters the geometric memorization phase, opening a gap on the green line, signaling a disruption of such manifold and the estimation of a lower manifold dimension. In the Right panel the network skips the opening of the final gap signaling the true latent manifold dimensions and directly enters the memorization regime, as predicted by our theory. This toy experiment reproduces the results obtained on real-world data and reported in Section 4.10.3.

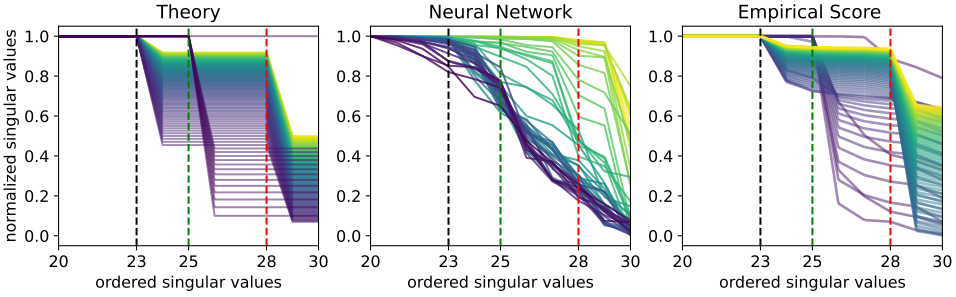


Figure 4.16: The evolution in time of the ordered singular values of the Jacobian of the score function for a linear manifold model. The parameters for the model are $N = 30$, $D = 7$, $\log(P)/N = 0.23$ with subspaces associated to variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.3$ with dimensions $D_1 = 2$ and $D_2 = 5$ respectively. Lighter colors are associated to larger times in the color map. Left: heuristic theoretical prediction in the memorization phase according to Eq. (4.93). Center: singular values obtained by the improved NB method in $\boldsymbol{x} = 0$. after training a Neural Network over a synthetic dataset of $P = 500$ points. Right: singular values obtained by the numerical measure of the Jacobian of the empirical score function (as described in Sec. 4.9), evaluated from a synthetic dataset of $N = 10^3$ points.

Subsequently, a gap on the black dashed line appears, suggesting that the score function has reconstructed the full D-dimensional manifold. Eventually, this gap closes and leaves room to a new intermediate gap opened on the green dashed-line, associated to a D_2-dimensional sub-manifold. The opening of the last gap is a trace of geometric memorization which cannot occur in a true score-driven diffusion model. It suggests that the subspace of higher variance has been memorized, thus it does not count anymore in the manifold dimension. These dimensions have been effectively *lost*, i.e. removed by the subspace spanned by diffusion at small times. Yet it is predicted by our theory based on the spectral analysis of the Jacobian of the empirical score function. This phenomenology is ubiquitous to all the three pictures reported in Fig. 4.16.

4.10.3 Experimental evidence of Geometric Memorization

In this section we provide experimental evidence of geometric memorization in generative diffusion. Such observations have motivated the theory developed in Sec. 4.9.

Let us consider Diffusion Models trained on a series of sub-datasets extracted from MNIST, Cifar10, Fashion-MNIST, CelebA-HQ and LSUN-Churches. For each dataset size, we fix a small diffusion time $t = 10^{-5}$, estimate the latent dimensionality around the position $\boldsymbol{x} = 0$ according to the Normal Bundle analysis described in Section 4.1, and study how this dimensionality varies with dataset size. Full details of the dimensionality estimation procedure and training setup are provided in Appendix B. The average dimensionality as a function of the dataset size is reported in Fig. 4.17. We observe the following progression:

I. When the dataset size is very large, i.e. of order $\sim 10^4$, the latent dimensionality of the data manifold remains stable: the sample complexity is sufficient, and the network successfully captures the underlying data structure.

II. As the dataset size decreases in the interval $[10^3 \div 10^4]$, the latent dimensionality gradually declines: the network begins to overfit, but the manifold dimensionality does not collapse abruptly.

III. For even smaller datasets, the latent dimensionality has approached zero, indicating that the network-fitted score function concentrates on individual data points, effectively reducing the manifold to zero-dimensional objects.

Previous theoretical and numerical studies of memorization in diffusion models [Biroli et al., 2024, Achilli et al., 2025, Lucibello and Mézard, 2024, Pham et al., 2025] have shown

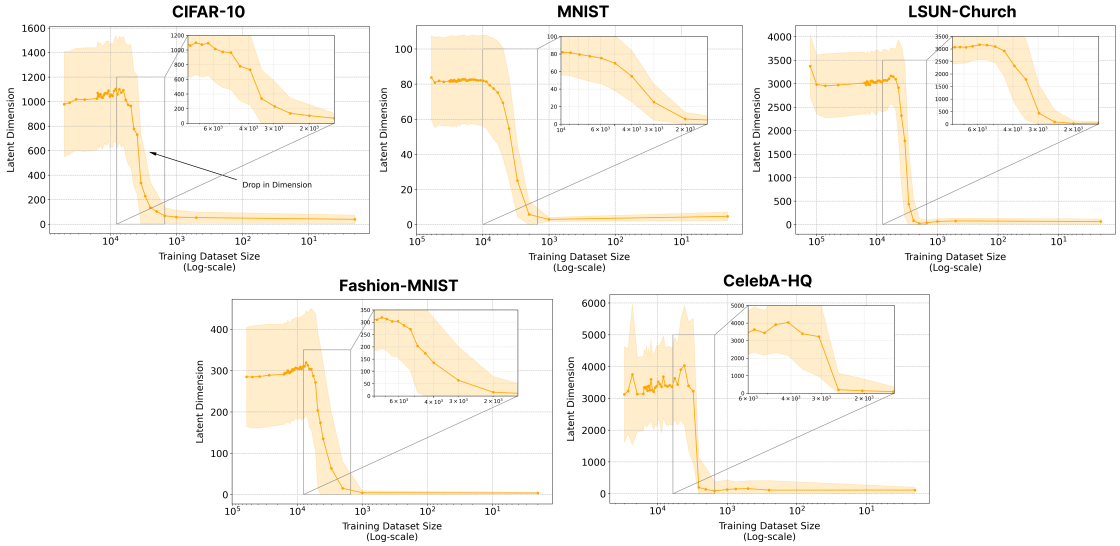


Figure 4.17: Latent manifold dimensionality estimated on deep networks trained on natural image datasets at $t = 10^{-5}$. The estimated dimensionality tends to smoothly decrease when the dataset size is smaller, suggesting an underlying phenomenon of geometric memorization. Latent dimensionality has been estimated according to Appendix B.

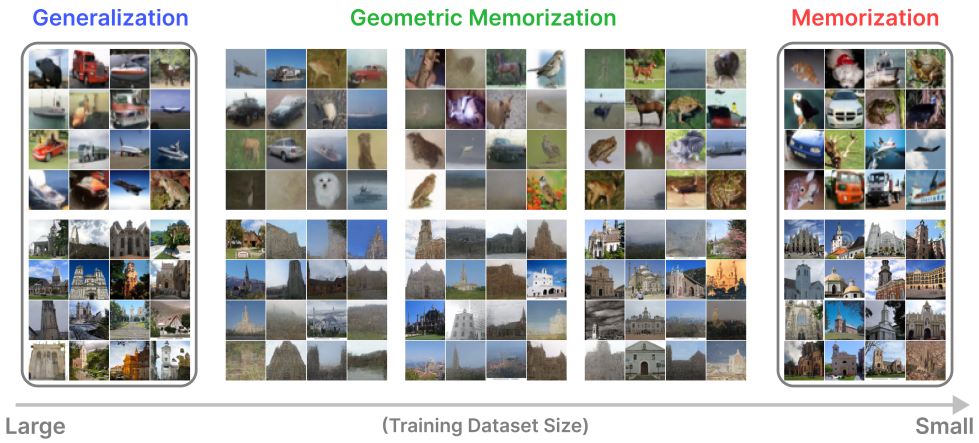


Figure 4.18: Randomly chosen images that have been generated by diffusion models trained as in Fig. 4.17. Training datasets are: Cifar10 (top row) and LSUN-Churches (bottom row). The dataset size decreases from left to right: in the generalization phase, the score function fits the exact one, and images are coherent with a high saturation; during geometric memorization, images look clearly foggy, with a low saturation; when the model memorizes images return to be clear and well saturated since they coincide with existing examples. We conjecture that the reduction in saturation is correlated with the dimensional reduction of the image latent manifold.

that the score function ceases to point towards the data manifold after a characteristic memorization time, which depends on both the properties of the target distribution and the size of the dataset. A natural question is whether, beyond this time, the score field begins orienting towards individual data points abruptly or progressively.

Our results reveal a universal phenomenology across different datasets: memorization unfolds gradually rather than occurring as a sharp transition. In our experiments, the measurement time for dimensionality is fixed for each dataset size. If memorization were abrupt, we would observe a sudden drop in Fig. 4.17, which does not occur. Instead, the latent dimensionality decreases smoothly across a range $[10^3 \div 10^4]$ in the dataset sizes, suggesting that the diffusion process progressively loses degrees of freedom until it ultimately collapses onto individual data points.

Fig. 4.18 visually represents examples generated by neural networks trained with different dataset sizes, as in Fig. 4.17. For large datasets, the model generalizes, generating new images that are perfectly coherent with the training examples. For intermediate sizes of the dataset, geometric memorization occurs, and generated images experience a drop in saturation: pixel values are more similar to each other, and shapes look foggy. We conjecture that this effect might be correlated with a decrease in the dimensionality of the latent manifold that is connected to a reduction in the number of relevant Fourier modes of the pictures. For small datasets instead, models are fully memorizing, generated images are data points, and saturation is fully restored. We also report that this visual effect does not evidently appear when we train the model on human faces, as those contained in CelebA-HQ dataset. Further analysis of the link between loss of latent dimensionality and the Fourier decomposition of images will be the object of further investigations.

4.11 Conclusions

In this second part, we present experimental evidence and a theoretical analysis of a phenomenon that has so far been rarely observed or discussed in the context of generative diffusion, which we call *geometric memorization*. Specifically, we show that data memorization in diffusion models is not a sudden process but a gradual one, deeply intertwined with the structure of the manifold that encodes the correlations and symmetries of a dataset.

Recent studies [Ventura et al., 2025], that we have described in the first part of this chapter, have shown that generative models trained on structured data progressively reconstruct both the target distribution and the geometry of the low-dimensional manifold that underlies it. The model first fits the subspaces populated by configurations where

features have higher variance and then gradually incorporates dimensions associated with smaller variances. In this way, the model reconstructs large-scale structures of the data before refining the smaller details. We demonstrate that, when a model overfits the data, it enters a phase in which this manifold is progressively destroyed, following a very similar sequence of steps, but in reverse. The system does not memorize entire data points all at once; instead, it first freezes the features with higher variance and then the finer details, gradually *losing dimensions* until the vector field of the score function becomes fully aligned with the data. To this end, we introduce a toy data model that can be analyzed within the framework of statistical mechanics. Following the approach of [Biroli et al., 2024, Ambrogioni, 2025, Achilli et al., 2025], we first map an empirical score-driven diffusion model onto a Random Energy Model, compute the relevant thermodynamic quantities of the system, and then use these to estimate the statistics of the eigenvalues of the Jacobian of the empirical score at small times.

We emphasize that our analysis, ultimately based on the Jacobian of the score function and thus on the local gradient of the diffusion potential, is a static and geometric study of generative diffusion. In contrast, prior works such as [Biroli and Mézard, 2023, Biroli et al., 2024] focus on dynamic descriptions, tracking temporal evolution along typical stochastic trajectories. We therefore conjecture that our geometric findings are deeply connected and simultaneously complementary to these dynamic aspects, since the score function governs diffusion, particularly at small times when stochastic noise is minimal. To sum up, while [Wang et al., 2024, Ventura et al., 2025, Leigh Ross et al., 2025] have highlighted the importance of manifold structure in the generalization behavior of generative diffusion models, our work is the first to explain how the manifold hypothesis influences memorization, suggesting a new way of conceiving data overfitting in these types of machine learning models.

Summary

Let us summarize our contribution to the problem in a few salient points, in the first part of the chapter:

- We compute analytically the spectrum of the Jacobian of the true score function for linear manifold data.
- We use the previous result to describe **geometric phases** in the learning process of diffusion models, which provide an explanation of why they are free from the

manifold overfitting phenomenon.

While in the second part:

- We give experimental evidence of a new phenomenon occurring in diffusion models, i.e., **geometric memorization**. In the same context, we attempt to identify the specific setting of the model that allows such observations.
- We provide a full theoretical explanation of geometric memorization. Our analysis, based on tools from statistical mechanics of disordered systems, successfully predicts the experimental results from neural-network-trained models.

Chapter 5

General theory of speciation in multiphase probability distributions

In this chapter, we present preliminary results of a joint work with Marco Benedetti, Giulio Biroli, and Marc Mézard.

A recent research line aims to give theoretical foundations of Diffusion Models (DMs) for high dimensional data by using tools from statistical physics. In particular, Biroli et al. [2024] has identified three dynamical phases in DMs. The separation between the first regime, where backward trajectories are purely noisy, and the second one, features of the data distribution start to emerge, is the called speciation time. For data coming from a mixture of two Gaussians with different means, they have obtained a speciation time $t_S = \frac{1}{2} \log N$. However, there is no current theory for the speciation transition in the case of data coming from classes which are not spatially separated.

In this chapter we extend the treatment of speciation to a more general setting. We first carefully define the properties of a target probability distribution composed of many classes. In particular, the probability to assign the class given a typical data is 1 for the correct class and 0 otherwise. Once we add noise with the diffusive process it is clear that this property no longer holds. Thus, we introduce a criterion to identify the speciation time based on the probability of attributing the data to the correct class. The quantity of interest will be the free entropy difference between classes. The typical size of fluctuations of this quantity is $O(1/\sqrt{N})$, but we will show that, in absence of first moments separation, it also depends on time as e^{-2t}, meaning that on logarithmic time scale $t \propto \log N$ the actual scaling is of order $O(1/N)$. This is precisely the scale at which clusters become mixed, i.e. the time scale of speciation.

This general criterion recovers the previous theory for the speciation transition. It

also allows us to study new cases where we do not have separation of first moments in different classes. Our prototypical models in this setting will be a mixture of 1d Ising models with different temperatures and a mixture of Gaussians with zero means and different covariances. Interestingly, we are able to obtain analytical expressions for the 1d Ising mixture by mapping this problem into a 1d Ising model with random field, following the replica calculation in [Weigt and Monasson, 1996, Lucibello et al., 2014].

5.1 Bayes attribution and pure densities mixtures

We are interested in studying probability distributions whose samples can be labeled as belonging to one among a number R of well distinguished classes. To model this, we will assume that $P(a) = \sum_{r=1}^{R} w_r P_r(a)$, where w_r are non-negative weights with $\sum_r w_r = 1$, and $P_r(a)$ represent the different classes. Each sample will be assigned to a class on the basis of the *Bayesian attribution to component*: given a sample a drawn from $P(a)$, we compute

$$P(s \mid a) = \frac{P(s,a)}{P(a)} = \frac{w_s P_s(a)}{P(a)}, \tag{5.1}$$

and we assign it to class $\operatorname{argmax} P(s \mid a)$. Given $P(a)$, there are many ways to decompose it in classes. To address this ambiguity, we shall restrict our analysis to *Proper Density Decompositions*, requiring that, as N increases, the Bayesian Classifier is able to attribute a typical noise-corrupted sample from $P(a)$ with certainty to one of the components of $P(a)$, for any small but finite amount of noise. Formally, let $\tilde{a}_r = a_r + \eta \mathcal{N}(0,1)$ be the random variable obtained by adding i.i.d. Gaussian noise to each feature of a_r, sampled from $P_r(a_r)$. We will say that $P(a) = \sum_{r=1}^{R} w_r P_r(a)$ forms a *Proper Density Decomposition* of $P(a)$ if, for any r and $s \neq r$, there exists $\eta = O_N(1)$ and $\epsilon(N) = o_N(1)$ such that $P(s|\tilde{a}_r) \leq \epsilon(N)$ with high probability over the statistics of $\tilde{a}_r$.

When $a_r \sim P_r$, the numbers $P_s(a_r)$, $s = 1, ..., n$ are random variables, and they can be expressed as $P_s(a) = e^{N f_s(a,N)}$, where $f_s(a,N) = (1/N) \log P_s(a)$. If $N f_s(a_s, N) \gg N f_r(a_s, N)$ with high probability over the sampling of a_s for any $r \neq s$ as $N \to \infty$, we have a *Proper Density Decomposition*. If $P(a) = \sum_{r=1}^{R} w_r P_r(a)$ is a *Proper Density Decomposition* and the distribution of $f_s(a_r, N)$ concentrates at large N towards its mean, will say that forms a *Pure Density Decomposition*. In that case, one can write

$$P_s(a_r) = e^{N f_s^r + o(N)}, \qquad f_s^r = \frac{1}{N} \langle \log P_s(a_r) \rangle_{a_r}, \tag{5.2}$$

where the average is taken over the distribution of diffused samples that originate from

cluster r, and the $o(N)$ contribution encapsulates the dependence on the specific realization of the disorder a_r (see Biroli and Mézard [2024] for more details).

5.2 Speciation

Symmetry breaking events in diffusion models have been first introduced by Raya and Ambrogioni [2023] and Biroli et al. [2024], where they were named speciation transitions. They are characteristic timescales of the backwards process, describing a very general, and target distribution independent phenomenology. They indicate sharp transitions in the qualitative behavior of trajectories during the backwards process.

Consider, as an example, an image dataset comprising photos of cats, dogs, eagles and seagulls. At any instant of the backwards process, one can try and guess which of the four types of animals will be represented in the final image, at $t = 0$. At the beginning of the backwards process, the image is just Gaussian noise, and it is impossible to attribute it to any of the four categories. As backwards diffusion takes place, features from the target distribution gradually emerge, and one is able to guess the final image better than random. Over time, the prediction for a given single trajectory will change, even back and forth, between classes, but well established time windows exist, during which such fluctuations happen only between a subset of the classes. In our example, there will be a time window during which the prediction fluctuates between cat and dog, but never to a bird, or conversely, between seagull and eagle, but never to a quadruped. The boundaries of such time windows are called *speciation times*. After each speciation, backwards trajectories become committed to a smaller subset of classes in the dataset. Conversely, during the forward diffusion process, these transitions mark moments when it becomes hard to guess which class generated the image, now corrupted by increasing amounts of noise.

5.2.1 A general criterion for speciation

Given a Proper Density Decomposition $P(a) = \sum_{r=1}^{R} w_r P_r(a)$, as defined in Sec.5.1, we need a rigorous notion of attribution to a component of a forward-diffused data point x. The most natural way is generalizing *Bayesian attribution to component* to the noise corrupted sample, namely we compute

$$P(s \mid x, t) = \frac{P(s, x; t)}{P(x, t)} = \frac{w_s P_s(x, t)}{P(x, t)}, \tag{5.3}$$

where $P(x,t) = \sum_s w_s P_s(x,t)$ and

$$P_s(x,t) = \int d^N a \; P_s(a) \exp\left(-\frac{(x - ae^{-t})^2}{2\Delta_t}\right), \tag{5.4}$$

and we assign it to class $\operatorname{argmax} P(s \mid x,t)$. Given r and $s \neq r$, if $P(r|a) \gg P(s|a)$ with high probability over the statistics of a sampled from $P_r(a)$, we say that r and s are well distinguishable classes at time t. Otherwise, if $P(r|a)$ and $P(s|a)$ are of the same order of magnitude, we say that class r is merged with class s at time t. When x is obtained by diffusing a sample originated from cluster r, each of the $P_s(x,t)$ is reminiscent of the partition function of a disordered system. The disorder is represented by the external field x. It is then natural to write

$$P_s(x,t) = e^{N f_s^r(t) + \delta f_s^r(x,t)}, \qquad f_s^r(t) = \frac{1}{N} \langle \log P_s(x,t) \rangle_r \tag{5.5}$$

where the average is taken over the distribution of diffused samples that originate from cluster r, and the $f_s^r(x,t) = o(N)$ contribution encapsulates the dependence on the specific realization of the disorder x. Component r can be reliably identified by the Bayesian Classifier as the origin of the trajectory as long as

$$N f_s^s(t) + \delta f_s^s(x,t) \gg N f_s^r(t) + \delta f_s^r(x,t) \qquad \forall s \neq r \tag{5.6}$$

with high probability over the statistics of x. When this condition is not met, we shall say that cluster r is merged at time t with cluster s. Hence, two things happen at once: the Bayes Classifier starts assigning finite probability to more than one class, and with finite probability $\operatorname{argmax} P(s \mid x,t)$ misattributes the origin of the trajectory. From (5.6), one can see that merging occurs when the difference between average free energies becomes comparable with their fluctuations.

$$|f_r^r(t) - f_s^r(t)| \simeq \sqrt{\operatorname{Var}\left[\frac{1}{N}\log P_r(x,t) - \frac{1}{N}\log P_s(x,t)\right]}. \tag{5.7}$$

As we will see, on the timescale of merging, the difference between the average free energies becomes $O_N(1)$ with high probability.

5.2.2 Scaling of speciation time

The criterion in Eq. 5.7 can be used to derive the scaling the first speciation time encountered during the backward diffusion, i.e. the largest speciation time. Explicitly, the average free entropy difference reads

$$f_r^r(t) - f_s^r(t) = \frac{1}{N}\left[\int dx P_r(x,t)\log P_r(x,t) - \int dx P_r(x,t)\log P_s(x,t)\right]. \tag{5.8}$$

Notice that this can be seen ad the Kullback-Leibler divergence between the components of the mixture. At large forward times one can approximate $P_r(x,t)$ by expanding in e^{-2t} its exact expression. The result is a Gaussian distribution (see e.g. Biroli et al. [2024]) with mean μ_s and variance Σ_s. Then, the asymptotic average free entropy difference is a Kullback-Leibler divergence between Gaussians

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_r,\Sigma_r)\,\|\,\mathcal{N}(\mu_s,\Sigma_s)) = a_{r,s}(e^{-2t}+e^{-4t}) + C_{r,s}e^{-4t} + e^{-4t}S + o(e^{-4t}). \tag{5.9}$$

The values of μ_r, Σ_r, $a_{r,s}$ and $C_{r,s}$ are fully determined by the first and second moments of the $P_s(x,t=0)$ distributions (see 5.2.3 for details). In particular, $a_{r,s} = \sum_i(\langle a_i\rangle_r - \langle a_i\rangle_s)^2/N$, which is zero if the components are not characterized by different means. Leveraging the Gaussian approximation for $P_r(x,t)$ at large t, one can also approximate the right hand side of Eq. (5.7):

$$\mathrm{Var}\left[\frac{1}{N}\log P_r(\mathbf{x},t) - \frac{1}{N}\log P_s(\mathbf{x},t)\right] = \frac{a_{r,s}}{N}(e^{-2t}+2e^{-4t}) + \frac{C_{r,s}}{N}e^{-4t} + o(e^{-4t}). \tag{5.10}$$

In this large t regime, leveraging Eqs. (5.9), (5.10), (5.7), the criterion for speciation reads

1. If $a_{r,s} \neq 0$, i.e. there is separation of first moments,

$$a_{r,s}e^{-2t} \simeq \sqrt{\frac{a_{r,s}}{N}}e^{-t} \implies t \sim \frac{1}{2}\log N. \tag{5.11}$$

   This recovers the speciation time scaling obtained in Biroli et al. [2024].

2. If $a_{r,s} = 0$,

$$e^{-4t}C_{r,s} \sim \sqrt{\frac{2}{N}}e^{-2t}C_{r,s}^{1/2} \implies t \sim \frac{1}{4}\log N. \tag{5.12}$$

   This case extends previous literature on speciation time to cases where the class distribution do not have any first moment.

Notice that in both cases, on the timescale of speciations, the free energy differences scale

as $f_r^r(t) - f_s^r(t) \simeq 1/N$.

5.2.3 Detailed large time analysis

At large forward times one can obtain $P_r(x,t)$ by expanding in e^{-2t} its exact expression. The result is a Gaussian distribution (see e.g. Biroli et al. [2024])

$$\log P_r(x,t) = \text{const} + \frac{e^{-t}}{\Delta_t} \sum_{i=1}^{N} x_i \langle a_i \rangle_r - \frac{1}{2\Delta_t} \sum_{i,j=1}^{N} x_i \, M_{r,ij} \, x_j + O\left((xe^{-t})^3\right), \tag{5.13}$$

where

$$M_{r,ij} = \delta_{ij} - e^{-2t} \left(\langle a_i a_j \rangle_r - \langle a_i \rangle_r \langle a_j \rangle_r \right). \tag{5.14}$$

and $\Delta_t = 1 - e^{-2t}$, so when we expand in e^{-t} we write

$$\frac{1}{\Delta_t} = \frac{1}{1-e^{-2t}} = 1 + e^{-2t} + e^{-4t} + O(e^{-4t}) \qquad \frac{1}{\Delta_t^2} = \frac{1}{(1-e^{-2t})^2} = 1 + 2e^{-2t} + 3e^{-4t} + O(e^{-4t}) \tag{5.15}$$

Completing the square in (5.13) yields a Gaussian with

$$\mu_r(t) = \frac{e^{-t}}{\Delta_t} M_r^{-1} \langle a \rangle_r = e^{-t} \left(1 + e^{-2t} + e^{-4t}\right) \left[I + e^{-2t} C + O(e^{-4t}) \right] \langle a \rangle_r \tag{5.16}$$

$$= e^{-t} \left[\langle a \rangle_r + e^{-2t} \left(\langle a \rangle_r + C_r \langle a \rangle_r \right) + O(e^{-4t}) \right], \tag{5.17}$$

$$\Sigma_r(t) = \frac{1}{\Delta_t} \Delta_t M_r^{-1} = I + e^{-2t} C_r + O(e^{-4t}), \tag{5.18}$$

where $C_{r,ij} = \langle a_i a_j \rangle_r - \langle a_i \rangle_r \langle a_j \rangle_r$. Then, the average free entropy difference is a Kullback-Leibler divergence between Gaussians

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_r, \Sigma_r) \,\|\, \mathcal{N}(\mu_s, \Sigma_s)) = \frac{1}{2} \left[\mathrm{Tr}(\Sigma_s^{-1} \Sigma_r) + (\mu_s - \mu_r)^\top \Sigma_s^{-1} (\mu_s - \mu_r) - N - \log \frac{\det \Sigma_r}{\det \Sigma_s} \right]. \tag{5.19}$$

Define

$$\Delta_{rs} a_i = \langle a_i \rangle_r - \langle a_i \rangle_s, \tag{5.20}$$

$$\Delta_{rs} C_{ij} = \left(\langle a_i a_j \rangle - \langle a_i \rangle \langle a_j \rangle \right)_r - \left(\langle a_i a_j \rangle - \langle a_i \rangle \langle a_j \rangle \right)_s, \tag{5.21}$$

then using the Gaussian KL formula and expanding up to $O(e^{-4t})$ gives

$$\frac{1}{N} D_{\mathrm{KL}}\left(P_r(\cdot,t) \,\|\, P_s(\cdot,t)\right) = \frac{1}{2N}\frac{e^{-2t}}{\Delta_t}\sum_i (\Delta_{rs} a_i)^2 + \frac{e^{-4t}}{4N}\frac{1}{\Delta_t^2}\sum_{i,j}\left(\Delta_{rs} C_{ij}\right)^2 \tag{5.22}$$

$$+\frac{e^{-4t}}{2N}\frac{1}{\Delta_t}\left[2\sum_i \Delta_{rs} a_i\left(\sum_j C_{s,ij}\langle a_j\rangle_s - \sum_j C_{r,ij}\langle a_j\rangle_r\right) - \sum_{i,j}\Delta_{rs} a_i C_{s,ij}\Delta_{rs} a_j\right] + O(e^{-4t}). \tag{5.23}$$

Now expand Δ_t and retain terms through $O(e^{-4t})$

$$\frac{1}{N} D_{\mathrm{KL}} = \frac{1}{2N}\left(e^{-2t}+e^{-4t}\right)\sum_i (\Delta_{rs} a_i)^2 + \frac{e^{-4t}}{4N}\sum_{i,j}\left(\Delta_{rs} C_{ij}\right)^2 \tag{5.24}$$

$$+\frac{e^{-4t}}{2N}\left[2\sum_i \Delta_{rs} a_i\left(\sum_j C_{ij}^{(s)}\langle a_j\rangle_s - \sum_j C_{ij}^{(r)}\langle a_j\rangle_r\right) - \sum_{i,j}\Delta_{rs} a_i C_{ij}^{(s)}\Delta_{rs} a_j\right] + O(e^{-4t}) \tag{5.25}$$

$$\simeq a_{r,s}(e^{-2t}+e^{-4t}) + C_{r,s}e^{-4t} + e^{-4t}S, \tag{5.26}$$

where

$$a_{r,s} = \frac{\sum_i (\Delta_{rs} a_i)^2}{N}, \qquad C_{r,s} = \frac{\sum_{i,j}\left(\Delta_{rs} C_{ij}\right)^2}{N}, \tag{5.27}$$

$$S = \frac{2\sum_i \Delta_{rs} a_i\left(\sum_j C_{s,ij}\langle a_j\rangle_s - \sum_j C_{r,ij}\langle a_j\rangle_r\right) - \sum_{i,j}\Delta_{rs} a_i C_{s,ij}\Delta_{rs} a_j}{2N} \tag{5.28}$$

Leveraging the Gaussian approximation for $P_r(x,t)$ at large t, one can also approximate $\mathrm{Var}[\frac{1}{N}\log P_s(x,t) - \frac{1}{N}\log P_r(x,t)]$. Specifically:

$$\frac{1}{N}\log P_r(x,t) - \frac{1}{N}\log P_s(x,t) = \frac{1}{N}\left[\frac{e^{-t}}{\Delta_t}\sum_i x_i\,\Delta_{rs} a_i + \frac{e^{-2t}}{2\Delta_t}\sum_{i,j} x_i\,\Delta_{rs} C_{ij}\,x_j\right] + O\left((xe^{-t})^3\right). \tag{5.29}$$

The dominant contribution to the variance comes from considering $x_i \simeq z_i$ with $z_i \overset{\text{i.i.d.}}{\sim}$

$\mathcal{N}(0,1)$. The linear and quadratic parts are uncorrelated for a centered Gaussian, so

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{N}\log P_s(x,t)-\frac{1}{N}\log P_r(x,t)\right] &= \mathrm{Var}\left[\frac{e^{-t}}{N\Delta_t}\sum_i \Delta_{rs}a_i z_i\right] \\ &\quad + \mathrm{Var}\left[\frac{e^{-2t}}{2N\Delta_t}\sum_{i,j}\Delta_{rs}C_{ij}z_i z_j\right] + O(e^{-6t}) \qquad (5.30)\\ &= \frac{e^{-2t}}{N^2}\frac{1}{\Delta_t^2}\sum_i(\Delta_{rs}a_i)^2 + \frac{e^{-4t}}{4N^2}\frac{1}{\Delta_t^2}\sum_{i,j}(\Delta_{rs}C_{ij})^2 + O(e^{-6t}). \qquad (5.31)\end{aligned}$$

Expanding also Δ_t and keeping through $O(e^{-4t})$,

$$\mathrm{Var}\left[\frac{1}{N}\log P_r(\mathbf{x},t)-\frac{1}{N}\log P_s(\mathbf{x},t)\right] \simeq \left(e^{-2t}+2e^{-4t}\right)\frac{1}{N^2}\sum_i(\Delta_{rs}a_i)^2 + \frac{e^{-4t}}{4N^2}\sum_{i,j}\left(\Delta_{rs}C_{ij}\right)^2 \qquad (5.32)$$

$$\simeq \frac{a_{r,s}}{N}(e^{-2t}+2e^{-4t}) + \frac{C_{r,s}}{N}e^{-4t}. \qquad (5.33)$$

In this large t regime, the criterion for speciation then reads

1. If $a_{r,s}\neq 0$,
$$a_{r,s}e^{-2t} \simeq \sqrt{\frac{a_{r,s}}{N}}e^{-t} \implies t\sim\frac{1}{2}\log N. \qquad (5.34)$$
This recovers the speciation time scaling obtained in Biroli et al. [2024].

2. If $a_{r,s}=0$,
$$e^{-4t}C_{r,s}\sim\sqrt{\frac{2}{N}}e^{-2t}C_{r,s}^{1/2} \implies t\sim\frac{1}{4}\log N. \qquad (5.35)$$
This case extends previous literature on speciation time to cases where the class distribution do not have any first moment.

Notice that, on this timescale, $\mathrm{Var}\left[1/N\log P_r(x,t)-1/N\log P_s(x,t)\right]=O(\frac{1}{N^2})$: this is due to a combination of factors: a free entropy variance has a natural scaling of $1/N$, at fixed time. But speciation happens on a timescale which is logarithmic in N, which contributes an additional $1/N$ factor to the variance. Hence, an equivalent definition of speciation time is $f_r^r(t)-f_s^r(t)\simeq 1/N$.

5.3 Toy models for speciation time

In this section, we illustrate with examples the speciation time scaling obtained in Sec. 5.2.2. We first recap the case with different first moments as it was derived in Biroli et al. [2024] for a mixture of two Gaussians with separate means. Then we proceed to the case without first moments separation analyzing again a mixture of two Gaussians both centered in zero with different variances. We verify that speciation time scaling for these target distributions can be retrieved on the basis of a more explicit transition in the shape of an effective potential.

5.3.1 Gaussian Mixture with different means

Consider a balanced mixture of two multivariate Gaussians with means $\pm m$ and the same isotropic variance σ^2

$$P_t(x) = \frac{1}{2}\frac{1}{(2\pi\Gamma_t)^{N/2}} e^{-\frac{\|x-me^{-t}\|^2}{2\Gamma_t}} + \frac{1}{2}\frac{1}{(2\pi\Gamma_t)^{N/2}} e^{-\frac{\|x+me^{-t}\|^2}{2\Gamma_t}}, \tag{5.36}$$

where where $\Gamma_t = \sigma^2 e^{-2t} + (1 - e^{-2t})$, and assume that in large dimensions they are well separated, so $\|m\|^2 = N\tilde{\mu}^2$. Thus, this setting enters the case $a_{r,s} \neq 0$, for which our criterion for predicts a speciation time $t \sim 1/2 \log N$ (see Eq. (5.11)).

This result was previously obtained in Biroli et al. [2024] by analyzing the backward diffusion potential. The score function reads

$$S_t(x) = -\frac{x}{\Gamma_t} + m\frac{e^-t}{\Gamma_t} \tanh\left(x \cdot m\frac{e^-t}{\Gamma_t}\right). \tag{5.37}$$

The reverse-time backwards diffusion process for $x_t \in \mathbb{R}^N$ is the Orstein-Uhlenbeck process:

$$dx = (x + 2S_t(x))\, dt + \sqrt{2} dW_t \tag{5.38}$$

where dW_t is standard Brownian motion. Introducing the overlap $q(t) = \frac{1}{\sqrt{N}} m \cdot x_t$, we can obtain a closed backward stochastic equation, defined by the potential

$$V(q,t) = \frac{1}{2}q^2 - 2\tilde{\mu}^2 \log\cosh\left(qe^{-t}\sqrt{N}\right) \tag{5.39}$$

This potential shows a transition: it is quadratic for large times, then at $t = \frac{1}{2}\log N$ it develops a double well structure.

5.3.2 Gaussian Mixture with different variances

In this section, we study the case of a 2-Gaussians mixture in $\mathbb{R}^N$, both centered in 0, with different isotropic variances σ_1^2 and σ_2^2. This setting reproduces the case of $a_{r,s} = 0$, for which the scaling of speciation time predicted by our criterion is $t \sim 1/4 \log N$, reported in Eq. 5.12. The distribution of the mixture at time t is

$$P_t(x) = \frac{1}{2}\frac{1}{(2\pi\Gamma_1)^{N/2}} e^{-\frac{\|x\|^2}{2\Gamma_1}} + \frac{1}{2}\frac{1}{(2\pi\Gamma_2)^{N/2}} e^{-\frac{\|x\|^2}{2\Gamma_2}}, \tag{5.40}$$

where $\Gamma_{1,2}(t) = \sigma_{1,2}^2 e^{-2t} + (1 - e^{-2t})$. For large N, the target probability density concentrates on thin shells of different radii depending on the two variances. We then choose $\sigma_1^2 = 1 - \delta$ and $\sigma_2^2 = 1 + \delta$, so $\Gamma_1(t) = 1 - \delta e^{-2t}$ and $\Gamma_2(t) = 1 + \delta e^{-2t}$, and tune the parameter δ in order to have two well distinct radii.

The exact score function for this model is $S_t(x) = -\lambda(\|x\|, t)x$, where

$$\lambda(\|x\|, t) = \frac{\Gamma_1^{-N/2} e^{-\|x\|^2/(2\Gamma_1)}/\Gamma_1 + \Gamma_2^{-N/2} e^{-\|x\|^2/(2\Gamma_2)}/\Gamma_2}{\Gamma_1^{-N/2} e^{-\|x\|^2/(2\Gamma_1)} + \Gamma_2^{-N/2} e^{-\|x\|^2/(2\Gamma_2)}}, \tag{5.41}$$

see Sec. 5.3.3 for details. The reverse-time backwards diffusion process for $x_t \in \mathbb{R}^N$ is again the Orstein-Uhlenbeck process.

Introduce the scalar variable $r_t = \frac{\|x_t\|^2}{N}$, the reverse SDE for the radial coordinate reads

$$dr = \left[2(r+1) - 4r\lambda(\sqrt{Nr}, t)\right] dt + 2\sqrt{\frac{2r}{N}} dB_t = -\frac{\partial U_t(r)}{\partial r} dt + 2\sqrt{\frac{2r}{N}} dB_t, \tag{5.42}$$

where we have defined the effective potential $U_t(r)$ associated with the deterministic part of the SDE as minus the integral of the drift.

$$U_t(r) = -\int_0^r \left[2(s+1) - 4s\lambda(\sqrt{Ns}, t)\right] ds \tag{5.43}$$

In Fig. 5.1 left panel, we can see how the shape of the potential evolves in time. It is clearly visible a symmetry-breaking phenomenon, from which we can estimate the speciation time for this model. Indeed, we identify the transition as the time where the curvature vanishes in $r = 1$: $\frac{\partial^2 U_t(r=1)}{\partial r^2} = 0$. Thus, we can derive a scaling for the speciation time by imposing that the derivative of $U_t(r)$ is zero at $r = 1$, which translates into $4 - 4\lambda(\sqrt{N}, t_s) = 0$, or more simply

$$\lambda(\sqrt{N}, t_s) = 1. \tag{5.44}$$

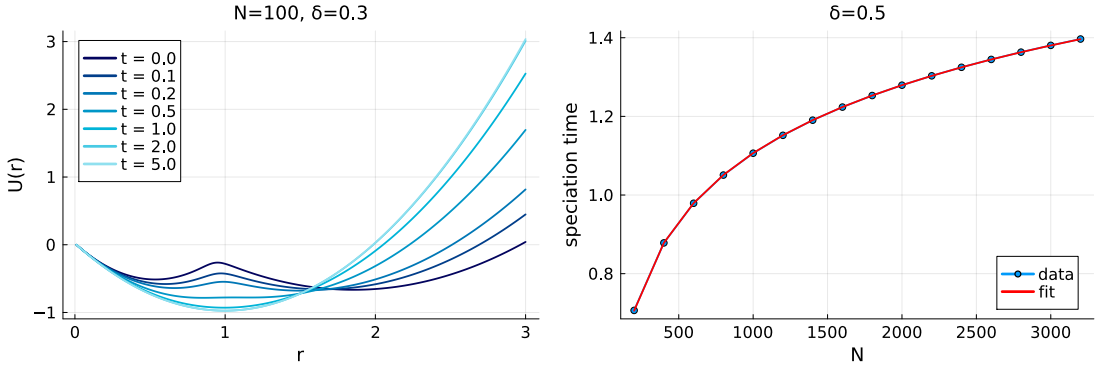


Figure 5.1: (Left) Potential of the reverse SDE as a function of r for different times. It is clearly noticeable a change in shape, from a single well for large t to a double well for small t. We identify the speciation time in correspondence with the change in curvature in $r = 1$. (Right) Speciation time for different dimensions obtained numerically and fitted curve with a logarithmic model $a \log t + b$ with parameters $a = 0.249091$, $b = -0.613988$.

In Fig. 5.1 right panel we report the speciation time obtained numerically by solving Eq. (5.44) for different N, and we fit the curve with a logarithmic model, recovering the desired scaling $t_s \approx \frac{1}{4} \log N$.

5.3.3 Score function and radial SDE

To compute the exact score function, defined as $S_t(x) = \nabla_x \log p_t(x)$, one can define the weights:

$$w_i(x) = \Gamma_i^{-N/2} e^{-\frac{\|x\|^2}{2\Gamma_i}} \tag{5.45}$$

and call

$$m_i(x) = \frac{w_i(x)}{w_1(x) + w_2(x)}. \tag{5.46}$$

Then, the exact score is:

$$S_t(x) = -x \left(\frac{m_1(x)}{\Gamma_1} + \frac{m_2(x)}{\Gamma_2} \right) \tag{5.47}$$

$$= -\lambda(\|x\|, t) x \tag{5.48}$$

where

$$\lambda(\|x\|, t) = \frac{\Gamma_1^{-N/2} e^{-\|x\|^2/(2\Gamma_1)}/\Gamma_1 + \Gamma_2^{-N/2} e^{-\|x\|^2/(2\Gamma_2)}/\Gamma_2}{\Gamma_1^{-N/2} e^{-\|x\|^2/(2\Gamma_1)} + \Gamma_2^{-N/2} e^{-\|x\|^2/(2\Gamma_2)}} \tag{5.49}$$

We consider the Orstein-Uhlenbeck reverse-time diffusion process for $x_t \in \mathbb{R}^N$:

$$dx = (x + 2S_t(x))\, dt + \sqrt{2} dW_t \tag{5.50}$$

where dW_t is standard Brownian motion. Our goal is to obtain an equation for the speciation time, similarly to what was done in Biroli et al. [2024] for the case of a mixture of Gaussians centered respectively in $\pm m$, with $\|m\|^2 = N\mu$. Define the scalar variable:

$$r_t = \frac{\|x_t\|^2}{N} \tag{5.51}$$

Using Itô's lemma:

$$dr = \frac{2}{N} x^\top dx + \frac{1}{N} \mathrm{Tr}(dx\ dx^\top) \tag{5.52}$$

From the SDE for x_t we see $dx\ dx^\top = 2I dt$, so

$$dr = \frac{2}{d} x^\top (x + 2S_t(x)) dt + 2dt + \frac{2\sqrt{2}}{N} x^\top dW_t \tag{5.53}$$

$$= [2r + 4\alpha(t, x) + 2]\, dt + \frac{2\sqrt{2}}{N} x^\top dW_t \tag{5.54}$$

where $\alpha(t, x) = \frac{1}{N} x^\top S_t(x)$.
Using our expression for the score $S_t(x) = -\lambda(\|x\|, t)x$, we get:

$$\alpha(r, t) = -r\lambda(\sqrt{Nr}, t) \tag{5.55}$$

so the drift becomes:

$$2r + 4\alpha(r, t) + 2 = 2(r + 1) - 4r\lambda(\sqrt{Nr}, t) \tag{5.56}$$

The noise term can be rewritten as

$$\frac{2\sqrt{2}}{N} x^\top dW_t \approx 2\sqrt{\frac{2r}{N}} dB_t \tag{5.57}$$

which is negligible for large N.
Putting all together, we find

$$dr = \left[2(r + 1) - 4r\lambda(\sqrt{Nr}, t)\right] dt + 2\sqrt{\frac{2r}{N}} dB_t \tag{5.58}$$

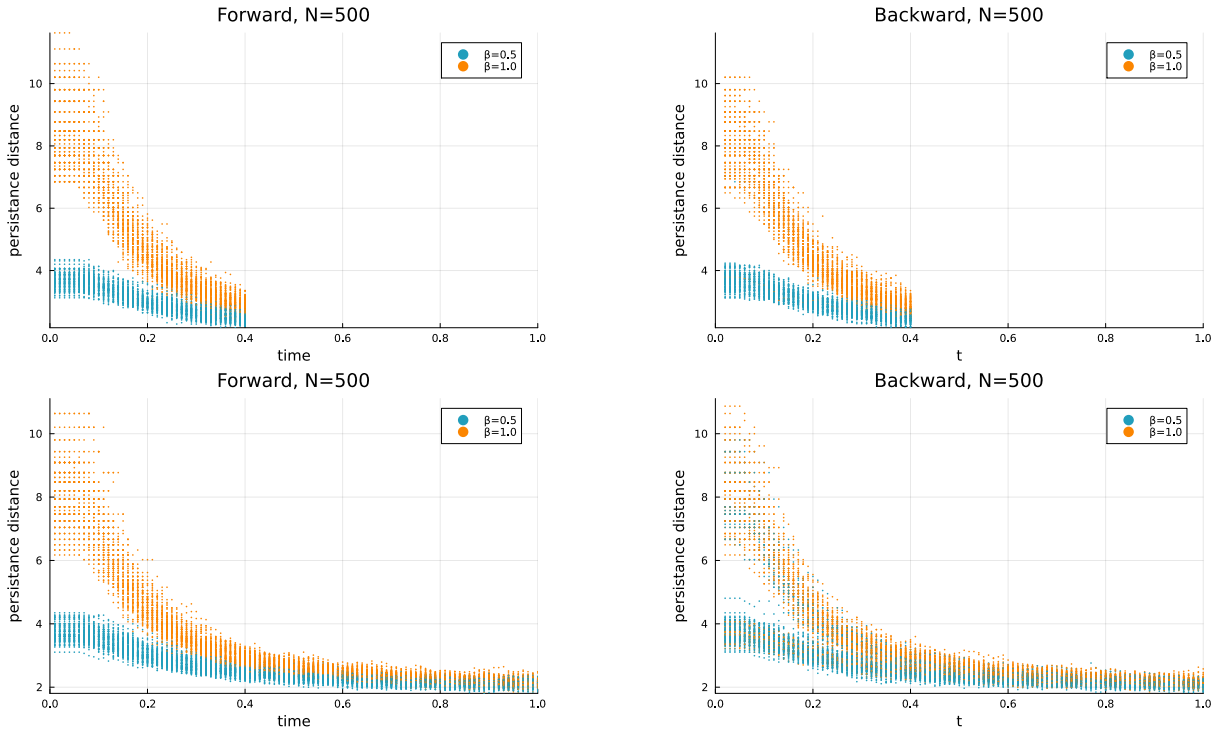


Figure 5.2: Visualization of speciation: U-turn experiments for 500 trajectories (dots) from the 1d Ising mixture with $\beta_1 = 0.5$ and $\beta_2 = 1.0$. Even if there is no spatial separation, we can clearly identify the two classes corresponding to the different temperatures and recover the speciation phenomenology in the persistence distance.

5.4 1d Ising mixture

As a second example, we consider a mixture of 1d Ising models at different inverse temperatures

$$P(\sigma) = \sum_r w_r \frac{1}{Z(\beta_r)} e^{\beta_r \sum_i \sigma_i \sigma_{i+1}}. \tag{5.59}$$

Notice that, also in this case, the components are not spacially well separated, meaning that the Euclidean distance between samples that come from the same component of the mixture is not more likely to be small than the one between samples from different classes. Nonetheless, even without spatial separation, we can still have speciation: one can clearly see a separation between samples coming from different clusters, measuring their correlation length or persistence distance, see Fig. 5.2. This example is instructive since it allows the analytical prediction of all speciation times, even in the case of an arbitrary number of components. If the number of components is $n > 2$, our method pinpoints a number of speciation transitions, of which only the first would have been otherwise computable.

The Bayes attribution to component r of a forward diffused sample x_t reads

$$P(r \mid x_t) = \frac{P(r) \int P(x_t \mid s) P(s \mid r) ds}{P(x_t)} = \frac{\frac{w_r}{Z(\beta_r)} \int ds \; e^{\sum_i \frac{e^{-t}}{\Delta t} x_i^t s_i + \beta_r \sum_i s_i s_{i+1}}}{\sum_c \frac{w_c}{Z(\beta_c)} \int ds \; e^{\sum_i \frac{e^{-t}}{\Delta t} x_i^t s_i + \beta_c \sum_i s_i s_{i+1}}} \tag{5.60}$$

The numerator is, in this case, the partition function of a 1d Ising spin system, at temperature β_r subject to a (random) external field x_t. According to our criterion, speciations are marked by average differences between pairs of free energies

$$f^r(x,t) = \frac{1}{N} \log \left(\sum_s e^{\beta_r \sum_{i=1}^{N-1} s_i s_{i+1} + \frac{e^{-t}}{\Delta_t} \sum_{i=1}^{N-1} x_i s_i} \right) \tag{5.61}$$

becoming comparable with with their fluctuations, when x_t is generated by sampling a data point a according to one of the P_{β_s}, and diffusing it forward in time. One can obtain an exact expression for the average free energies making use of replica computations, following the approach reported in Weigt and Monasson [1996]. The main difference is that in our case the individual features of the disorder $(x_t)_i$ are correlated, not i.i.d. variables. The analytical computation of the average free entropy is reported in Sec. 5.4.2. When it comes to the variance term, we approximate it with its scaling, which is the general one for pure density decompositions

$$\mathrm{Var}\left[\frac{1}{N} \log P_r(\mathbf{x}, t) - \frac{1}{N} \log P_s(\mathbf{x}, t) \right] \sim \frac{e^{-4t}}{N}, \tag{5.62}$$

where the $1/N$ accounts for the standard scaling of free energy fluctuations with system size, and the e^{-4t} accounts for the e^{-t} scaling of the size of the external fields in Eq. (5.61). The merging time between component r and s is then marked by

$$|f_r^r(t) - f_s^r(t)| \simeq \frac{e^{-2t}}{\sqrt{N}}. \tag{5.63}$$

In the next sections we will present numerical evidence that the speciation times predicted according to our Bayes attribution criterion indeed match the dynamical ones, defined experimentally in terms of U-turn experiments.

5.4.1 Exact Score

The exact score function can be obtained from

$$\mathcal{S}(x,t) = -\frac{x - e^{-t}\langle s\rangle_x}{\Delta_t} \tag{5.64}$$

where the average of s under the tilted measure reads

$$\langle s_i\rangle_x = \int ds\ s_i P(s \mid x) = \int ds\ s_i \frac{P(s,x_t)}{P(x_t)} \tag{5.65}$$

In the Ising mixture case, one has

$$P(s,x_t) = P_0(s)\frac{e^{-\frac{(x-se^{-t})^2}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}^N} = \frac{1}{\sqrt{2\pi\Delta_t}^N} e^{-\frac{||x_t - se^{-t}||^2}{2\Delta t}} \left(\sum_r w_r \frac{e^{\beta_r \sum_i s_i s_{i+1}}}{Z_r}\right) \tag{5.66}$$

with $Z_r = 2(2\cosh\beta_r)^{N-1}$, leading to

$$\langle s_i\rangle_x = \int ds\ s_i\ P(s \mid x) = \frac{\int ds\ s_i\ e^{\sum_i \frac{e^{-t}}{\Delta t} x_i^t s_i} \sum_r w_r \frac{e^{\beta_r \sum_i s_i s_{i+1}}}{Z_r}}{\int ds\ e^{\sum_i \frac{e^{-t}}{\Delta t} x_i^t s_i} \sum_r w_r \frac{e^{\beta_r \sum_i s_i s_{i+1}}}{Z_r}}. \tag{5.67}$$

Both the trace in the denominator and the average in the numerator can be computed through the transfer matrix method, for each value of x. For every β_r, the terms in the denominator are standard partition functions

$$B(\beta) = \int ds\ \frac{e^{\beta\sum_i s_i s_{i+1} + \frac{e^{-t}}{\Delta_t}\sum_i x_i s_i}}{Z(\beta)} = \mathrm{Tr}\prod_{i=1}^N \frac{T_i^\beta}{z_\beta} \tag{5.68}$$

where

$$T_i^\beta = \begin{pmatrix} e^{\beta + \frac{e^{-t}}{\Delta_t}x_i} & e^{-\beta - \frac{e^{-t}}{\Delta_t}x_i} \\ e^{-\beta + \frac{e^{-t}}{\Delta_t}x_i} & e^{\beta - \frac{e^{-t}}{\Delta_t}x_i} \end{pmatrix}$$

and $z_\beta = e^\beta + e^{-\beta}$. Instead, terms in the numerator will have the form

$$A(\beta)_j = \int ds\ s_j \frac{e^{\beta\sum_i s_i s_{i+1} + \frac{e^{-t}}{\Delta_t}\sum_i x_i s_i}}{Z(\beta)} = \sum_{s_j} s_j \mathrm{Tr}\left(\left(\prod_{i=j}^N \frac{T_i^\beta}{z_\beta}\right)\left(\prod_{i=1}^{j-1}\frac{T_i^\beta}{z_\beta}\right)\right). \tag{5.69}$$

Finally

$$\langle s \rangle_x = \frac{\sum_r w_r A(\beta_r)}{\sum_r w_r B(\beta_r)}. \tag{5.70}$$

5.4.2 Computation of analytical Free Entropy

In this section, we compute analytically Eq. (5.61) following the approach reported in Weigt and Monasson [1996] that we have outlined in Sec. 5.4. The main difference is that in our case, the noise is partially correlated with the data. Indeed, if we take on from Eq. 5.61, for x evolved according to the forward process $x_t = \xi_i e^{-t} + z_i\sqrt{\Delta_t}$ this reads

$$F_r(x,t) = \frac{1}{N}\log\left(\sum_{\{s\}} e^{\beta_r \sum_{i=1}^{N-1} s_i s_{i+1} + \frac{e^{-t}}{\Delta_t}\sum_{i=1}^{N-1}\left(\xi_i e^{-t} + z_i\sqrt{\Delta_t}\right)s_i}\right). \tag{5.71}$$

Now we call $\gamma = \frac{e^{-t}}{\sqrt{\Delta_t}}$, and assume $P(\xi) = \frac{1}{Z(\beta_0)} e^{\beta_0 \sum_i \xi_i \xi_{i+1}}$

$$F = \frac{1}{N}\log\sum_{\{s\}} e^{\beta \sum_i s_i s_{i+i} + \gamma^2 \sum_i \xi_i s_i + \gamma \sum_i z_i s_i} \tag{5.72}$$

and in order to compute this we introduce n replicas s^a, $a = 1, \ldots, n$, and also count the starting point ξ as the $n+1$ replica s^0

$$\bar{Z}^n = \frac{1}{Z(\beta_0)}\sum_{s^a} e^{\beta_0 \sum_i s_i^0 s_{i+i}^0 + \beta \sum_{i,a} s_i^a s_{i+i}^a + \gamma^2 \sum_i s_i^0 \sum_a s_i^a} \langle\langle e^{\gamma \sum_i z_i \sum_a s_i^a}\rangle\rangle_z \tag{5.73}$$

In our case, to describe the replica symmetric subspace, we introduce vectors $|s_0, a_1 \ldots a_p\rangle$, where we have p spins up at indices a_j. Then, we take the sum of such vectors

$$||s^0, p\rangle\rangle = \sum_{a_1 < \ldots < a_p} |s_0, a_1 \ldots a_p\rangle \tag{5.74}$$

and the collection of these vectors $||s^0, p\rangle\rangle_{s^0=\pm 1, p=1,\ldots,n}$ gives the RS subspace of dimension $2(n+1)$. Then, we look at the replicated transfer matrix $T_{2^{n+1}\times 2^{n+1}}$ projected onto this subspace

$$\langle\langle \sigma^0, q||T||s^0, p\rangle\rangle = e^{\beta_0 \sigma^0 s^0} e^{\gamma^2 \sigma^0 (2q-n)} \langle\langle e^{\gamma z (2q-n)}\rangle\rangle_z \sum_{r=r_{min}}^{r_{max}} \binom{q}{r}\binom{n-q}{p-r} e^{\beta(4r-2q-2p+n)} \tag{5.75}$$

with $r_{min} = \max(0, p+q-n)$ and $r_{max} = \min(p,q)$. Then the site-dependent partition function is given by

$$Z_{i+1}(s^0, p) = \sum_{\sigma^0=\pm 1} \sum_{q=0}^{n} T(\sigma^0, q; s^0, p) Z_i(\sigma^0, q). \tag{5.76}$$

We can write this on a function basis, similarly to the Fourier transform

$$Z_{i+1}^{s^0}[x] = \sum_{p=0}^{n} Z_{i+1}(s^0, p) x^p \tag{5.77}$$

$$= \sum_{p=0}^{n} x^p \sum_{\sigma^0=\pm 1} \sum_{q=0}^{n} Z_i(\sigma^0, q) \sum_{r=r_{min}}^{r_{max}} e^{\beta_0 \sigma^0 s^0} e^{\gamma^2 \sigma^0 (2q-n)} \langle\langle e^{\gamma z (2q-n)} \rangle\rangle_z \binom{q}{r} \binom{n-q}{p-r} e^{\beta(4r-2q-2p+n)} \tag{5.78}$$

$$= \sum_{q,r,r'} x^{r+r'} \sum_{\sigma^0=\pm 1} Z_i(\sigma^0, q) \binom{q}{r} \binom{n-q}{r'} e^{\beta(2r-2q-2r'+n)} e^{\beta_0 \sigma^0 s^0} e^{\gamma^2 \sigma^0 (2q-n)} \langle\langle e^{\gamma z (2q-n)} \rangle\rangle_z \tag{5.79}$$

$$= \sum_{q} \sum_{\sigma^0=\pm 1} Z_i(\sigma^0, q) \langle\langle e^{\gamma z (2q-n)} \rangle\rangle_z e^{\beta(n-2q)} e^{\beta_0 \sigma^0 s^0} e^{\gamma^2 \sigma^0 (2q-n)} (1 + x e^{2\beta})^q (1 + x e^{-2\beta})^{n-q} \tag{5.80}$$

$$= e^{\beta n} (1 + x e^{-2\beta})^n \sum_{q} \sum_{\sigma^0=\pm 1} Z_i(\sigma^0, q) \langle\langle \left(e^{2\gamma z} \frac{1 + x e^{-2\beta}}{1 + x e^{2\beta}} e^{2\gamma^2 \sigma^0} \right)^q e^{-n\gamma z} \rangle\rangle_z e^{\beta_0 \sigma^0 s^0} e^{-\gamma^2 \sigma^0 n} \tag{5.81}$$

$$= e^{\beta n} (1 + x e^{-2\beta})^n \sum_{\sigma^0=\pm 1} e^{\beta_0 \sigma^0 s^0} \langle\langle Z_i^{\sigma^0}[y = f^{\sigma^0}(x, z)] e^{-n\gamma z} \rangle\rangle_z e^{-\gamma^2 \sigma^0 n} \tag{5.82}$$

where

$$f^{\sigma^0}(x, z) = e^{2\gamma z} \frac{e^{-\beta} + x e^{\beta}}{e^{\beta} + x e^{-\beta}} e^{2\gamma^2 \sigma^0}. \tag{5.83}$$

With integral notation

$$Z_{i+1}^{s^0}[x] = \sum_{\sigma^0=\pm 1} e^{\beta_0 \sigma^0 s^0} \int_0^{\infty} dy \; e^{\beta n} (1 + x e^{-2\beta})^n e^{-\gamma^2 \sigma^0 n} \langle\langle \delta \left(y - f^{\sigma^0}(x, z) \right) e^{-n\gamma z} \rangle\rangle_z Z_i^{\sigma^0}[y] \tag{5.84}$$

$$= \sum_{\sigma^0=\pm 1} e^{\beta_0 \sigma^0 s^0} \int_0^{\infty} dy \; K_n^{\sigma^0}(x, y) Z_i^{\sigma^0}[y] \tag{5.85}$$

with the definition of the kernel

$$K_n^{\sigma^0}(x,y) = e^{\beta n}(1+xe^{-2\beta})^n e^{-\gamma^2\sigma^0 n}\langle\langle\delta\left(y - f^{\sigma^0}(x,z)\right)e^{-n\gamma z}\rangle\rangle_z. \tag{5.86}$$

For $n = 0$, this becomes

$$K_0^{\sigma^0}(x,y) = \langle\langle\delta\left(y - f^{\sigma^0}(x,z)\right)\rangle\rangle_z \tag{5.87}$$

which is very similar to the Monasson-Weigt kernel, but with a slightly different f function which depends from σ^0. We call $\Psi(x) = \begin{pmatrix} \psi^+(x) \\ \psi^-(x) \end{pmatrix}$ the right eigenvector of the largest eigenvalue of

$$\begin{pmatrix} \frac{e^{\beta_0}}{2\cosh\beta_0}K^+(x,y) & \frac{e^{-\beta_0}}{2\cosh\beta_0}K^-(x,y) \\ \frac{e^{-\beta_0}}{2\cosh\beta_0}K^+(x,y) & \frac{e^{\beta_0}}{2\cosh\beta_0}K^-(x,y) \end{pmatrix}, \tag{5.88}$$

where the factor $2\cosh\beta_0$ is needed for normalization, and $\Phi(x)$ for the left eigenvector. It is actually easy to see that $\Psi(x) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. If we expand the kernel at linear order in n, we find

$$K_n^{\sigma^0}(x,y) \approx \langle\langle\delta\left(y - f^{\sigma^0}(x,z)\right)\rangle\rangle_z + n\langle\langle\delta\left(y - f^{\sigma^0}(x,z)\right)\left[-\gamma z - \gamma^2\sigma^0 + \beta + \log(1+bx)\right]\rangle\rangle_z \tag{5.89}$$

$$= K_0^{\sigma^0}(x,y) + n\delta K^{\sigma^0}(x,y) \tag{5.90}$$

so we can write the maximum eigenvalue to the linear order as $\lambda = 1 + kn$, with

$$k = \int dx\,dy\ \Phi(x)\delta K(x,y)\Psi(y) \tag{5.91}$$

$$= \int dx\ \phi^+(x)\left(\frac{e^{\beta_0}}{2\cosh\beta_0}\delta K^+(x) + \frac{e^{-\beta_0}}{2\cosh\beta_0}\delta K^+(x)\right) \tag{5.92}$$

$$+ \int dx\ \phi^-(x)\left(\frac{e^{-\beta_0}}{2\cosh\beta_0}\delta K^-(x) + \frac{e^{\beta_0}}{2\cosh\beta_0}\delta K^-(x)\right) \tag{5.93}$$

$$= \frac{1}{2}\int dx\ [\phi^+(x)\delta K^+(x) + \phi^-(x)\delta K^-(x)] \tag{5.94}$$

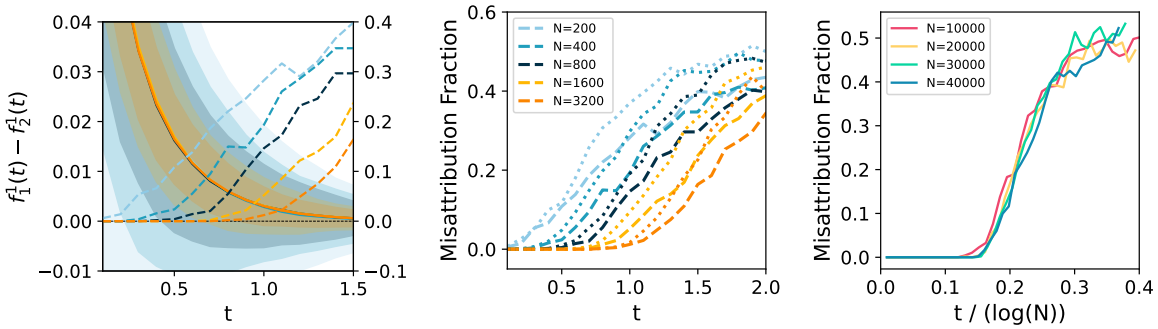


Figure 5.3: (Left) Solid lines show the average free entropy difference for different number of spins N during the forward process. The shading is at 3σ. With dashed lines, misattribution fraction during the forward process. The misattribution starts to rise when 0 is in the confidence interval. (Center) With dotted lines, misattribution fraction at the end of the backward process after a U-turn at the corresponding time. Despite being two different quantities, their behavior is strongly correlated. (Right)Misattribution fraction during the forward process for large N, with time rescaled by $\log N$. The collapse of the curves shows the common scaling of speciation time.

where $\delta K^{\sigma^0}(x) = \beta - \gamma^2\sigma^0 + \log(1 + xb)$. Passing to the exponential variables

$$k = \frac{\left[\beta - \gamma^2 + \int dt\ \log(1 + be^t)\hat{\phi}^+(t)\right] + \left[\beta + \gamma^2 + \int dt\ \log(1 + be^t)\hat{\phi}^-(t)\right]}{2} \tag{5.95}$$

and $\hat{\phi}^{\sigma^0}$ can be found by iteration.

5.4.3 Comparison with simulations

We shall first verify that the Bayes Classifier starts misattributing trajectories at a timescale in the forward process coherent with the one predicted in Eq. (5.63), with the free energies computed in Sec. 5.4.2. Secondly, we shall verify that the speciation times computed by means of the Bayes Classifier indeed capture the dynamic features of the backwards diffusion process, namely the fact that, after speciation, backwards trajectories are committed to a subset of classes, and do not erratically oscillate between all classes. This will be done by comparing analytical results with so called *U-turn experiments*, conducted as follows. Generate an initial data a according to the distribution P_r corresponding to cluster r. Then let it evolve up to some intermediate time t, call x the point reached at this time. Then reverse the process, doing a "U-turn": start the backward process from time t, starting from x. Denote by b the point which has been reached at the end of the backward process, $t' = t$. If one uses ideal score, b is a generic

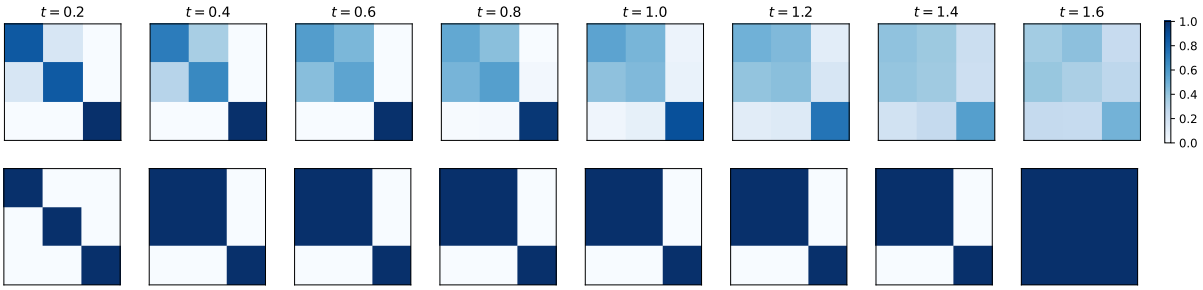


Figure 5.4: Attribution matrices at the end of the backward process for increasing U-turn times computed numerically with the transfer matrix method and compared with the analytical ones. Particles are drawn from a mixture of 3-Ising mixture with $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\beta_3 = 1.0$.

point distributed according to $P = \sum_s w_s P_s$. Because this is a pure density decomposition, with high probability we can assert with certainty what is the index s of the cluster where one finds the point b, by means of Bayes attribution. Repeating this experiment many times, one can monitor the probability that b is found in cluster r knowing that the initial condition a was in cluster a.

5.4.4 2-Ising Mixture

We choose two clusters with $\beta_1 = 0.5$ and $\beta_2 = 1.0$ and generate 1000 independent samples randomly taken from each cluster ($w_1 = w_2 = 0.5$). In Figure 5.3 we report the evolution of $f_1^1 - f_2^1$ during the forward process (solid lines) for different number of spins. The lines obtained from the numerical experiment and from the analytical computation match perfectly. The numerical experiment is nonetheless needed to compute the fluctuations, represented in the plot with the shaded areas, for which we do not have an analytical counterpart. We compute the probability of attributing one particle to the correct cluster during the forward process and plot with dashed lines the misattribution fraction. Finally, we compare it with the misattribution fraction that we find at the end of the backward process once we do a U-turn at the corresponding time. One can notice that when fluctuations of the free entropies difference cross zero, both misattribution fractions start to rise.

5.4.5 n-Ising Mixture

We now consider a homogeneous mixture of 8 Ising chains with hierarchically generated inverse temperatures. For each particle generated from one of the temperatures in the mixture, we perform a U-turn experiment at different times, and compute at the end of the

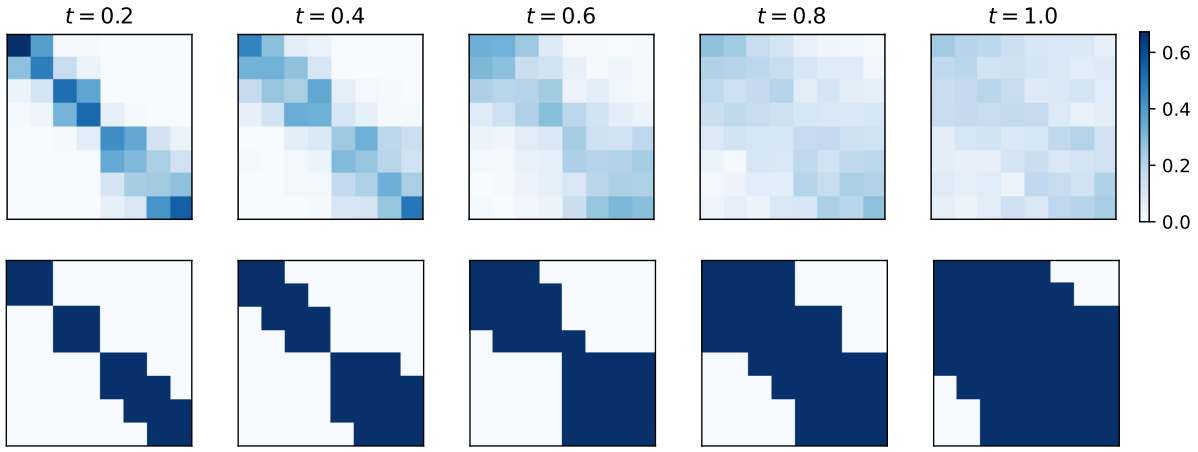


Figure 5.5: Attribution matrices at the end of the backward process for increasing U-turn times computed numerically with the transfer matrix method and compared with the analytical ones. Particles are drawn from a mixture of 8 hierarchically generated temperatures.

backward process the probability of attributing it to any temperature. This results in an attribution matrix that of course for small t_U is almost diagonal, and for larger times show a progressively increasing block structure that reveals several speciation transitions. In Figure 5.5 we show the attribution matrices obtained numerically on the left, and compare them with their analytical counterpart. We can first of all notice a good agreement between attribution matrices obtained analytically and numerically for all the U-turn times considered. As time increases a block structure emerges in the matrices. This can be explained with subsequent speciation events, where trajectories commit to the direction of a range of temperatures, and only afterwards to a specific temperature.

5.5 Conclusions

In this work, we have extended the theoretical understanding of speciation transitions in generative diffusion models to the case of data originating from mixture distributions, where components are not well spatially separated. We have done so by introducing a new criterion for speciation, based on the analysis of the Bayes cluster attribution during the forward process.

As a practical example, we applied the technique to mixtures of 1d Ising models. We derived explicit expressions for the mean and variance of the free entropy difference between clusters as a function of time. Our analysis reveals that both the mean and the standard deviation of this difference decay exponentially with time, and that the

critical timescale for speciation is logarithmic in the system size, $t_S \sim \frac{1}{4} \log N$. We have validated our theoretical predictions with numerical experiments, demonstrating excellent agreement between analytical and empirical results for the onset of misattribution between clusters. Our approach provides a principled criterion for speciation that is applicable to a wide range of mixture models.

These findings contribute to the growing theoretical foundation of diffusion models and highlight the universality of the speciation phenomenon in high-dimensional generative processes.

Summary

In this chapter, the main contributions are:

- We introduced a **new criterion** for **speciation**, based on Bayesian attribution to cluster and the difference of free entropies for different clusters
- We obtain a **scaling** for the speciation time for non spatially separated data
- We test this new criterion on a mixture of 1d Ising model, for which we derive analytical and numerical results
- We verify the speciation scaling for a mixture of Gaussians with the same mean but different variances

Chapter 6

Conclusions and Future Perspectives

In this thesis, we applied statistical physics to the study of diffusion models. Although statistical physics has been previously used to study deep learning, the analysis of generative diffusion in high dimensions is relatively new and raises many open questions. Here, we have investigated the dynamical phases of diffusion models in the presence of structured data. Modeling correlated data has given us the opportunity to advance our understanding of how these models act on real data. Let us briefly recap the main results presented in this work.

In Chapter 3, we focused on the memorization transition for data lying on a low-dimensional manifold within the ambient space. We found that the intrinsic manifold dimension regulates the scaling of the collapse time at which this transition occurs. Even in regimes where diffusion models fit the empirical distribution, the presence of structure in the data causes them to memorize later than one might expect. Moreover, we have shown that within this memorization phase, the empirical distribution is closest to the target distribution; thus, we can say that the model exhibits its best generalization ability while memorizing.

In Chapter 4, we adopted a geometric perspective. Introducing structure in the data naturally leads to geometric properties. Using again the Hidden Manifold Model for data generation, we described the geometric phases through which diffusion models fit the data distribution on the manifold. This behavior is reflected in the spectral density of the Jacobian of the score function, which displays a series of time-evolving sub-gaps associated with the different subspaces of the manifold. This spectral analysis can also be applied to the Jacobian of the empirical score function. In this setting, we observed the emergence of geometric memorization: the memorization process is not a trivial all-or-nothing phenomenon but happens gradually, manifesting as a progressive loss of dimensionality

in the subspaces associated with larger variances.

In Chapter 5, we turned our attention to the speciation transition. We proposed a general criterion for speciation that recovers previous results and extends them to a broader class of data distributions.

Many open questions remain about why diffusion models are so effective. For example, regarding memorization, one could further investigate to what extent it is simply a form of overfitting, or whether it represents a fundamentally different phenomenon. Understanding this distinction could shed light on why diffusion models generalize so well despite their high dimensionality and overparameterization. Bonnaire et al. [2025] introduced a random-feature score (RFS) model to analyze memorization in diffusion models, characterizing it as the late-time separation between training and test loss. They identified two dynamical regimes: an early generalization phase and a later memorization phase. Montanari and Urbani [2025], in an independent line of work on regression and feature learning, used dynamical mean-field theory (DMFT) to demonstrate a clear separation of timescales between feature learning and overfitting. Since diffusion models essentially solve a regression problem, these two observations could be related. Investigating the dynamics of an RFS diffusion model could therefore help to better characterize memorization in this setting and possibly reveal universal mechanisms underlying learning in high dimensions.

Another line of research concerns Classifier-Free Guidance (CFG) Ho and Salimans [2022]. CFG is a technique designed to improve the quality and controllability of generated samples. Instead of directly using the conditional score function to sample from a given class, it prescribes using a convex combination of the conditional and unconditional scores, determined by a guidance parameter. The main question regarding this practice is whether it introduces distortions with respect to the conditioned distribution. Pavasovic et al. [2025] proved that there is no distortion in high dimensions, provided the number of classes is finite and partitions the space. However, what happens when we consider continuous guidance, or when the number of classes grows exponentially with the data dimension? A deeper theoretical understanding of CFG could clarify why it empirically improves sample quality and control, and might lead to the design of optimal guidance schedules or alternative sampling strategies.

All these open questions deserve further attention. With the tools presented in this thesis, we have laid the groundwork for such investigations, providing both conceptual and methodological foundations for a statistical-physics theory of modern generative modeling.

Bibliography

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv:2410.08727*, 2024.

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, jul 2025. doi: 10.1088/1742-5468/ade136.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 5(26):381, 2024.

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking, and critical instability. *Entropy*, 27(3):291, 2025.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Giulio Biroli and Marc Mézard. Kernel density estimators in large dimensions. *arXiv:2408.05807*, 2024.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(9):093402, 2023.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Nicholas Boffi, Arthur Jacot, Stephen Tu, and Ingvar Ziemann. Shallow diffusion networks provably learn hidden low-dimensional structure. *International Conference on Learning Representations*, 2025.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training, 2025.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2:303–314, 1989.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613–2626, 1981. ISSN 0163-1829. doi: 10.1103/PhysRevB.24.2613.

Freeman J. Dyson. The dynamics of a disordered linear chain. *Physical Review*, 92(6): 1331–1338, 1953.

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis. *arXiv:1002.2050*, 2010.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, jan 1988. doi: 10.1088/0305-4470/21/1/031.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Analysis of diffusion models for manifold data. *arXiv:2502.04339*, 2025.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv:2204.03458*, 2022.

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.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv:2412.20292*, 2025.

D. Kingma and R. Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Krizhevsky, V. Nair, and G. Hinton. The cifar-10 dataset. 55(5):2, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Dmitry Krotov and John Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 2016.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. *International Conference on Learning Representations*, 2025.

Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2005.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 2097–2127. Curran Associates, Inc., 2023.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

G. Loaiza-Ganem, B. L. R. Ross, C Cresswell J., and A.L. Caterini. Diagnosing and fixing manifold overfitting in deep generative models. *Transactions on Machine Learning Research*, 2022.

Carlo Lucibello and Marc Mézard. The Exponential Capacity of Dense Associative Memories. *Physical Review Letters*, 132:077301, 2024.

Carlo Lucibello, Flaviano Morone, and Tommaso Rizzo. One-dimensional disordered ising models by replica and cavity methods. *Physical Review E*, 90(1), July 2014. ISSN 1550-2376. doi: 10.1103/physreve.90.012140.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Marc Mézard, Andrea Montanari, and Marc Mézard. *Information, physics, and computation*. Oxford Univ. Press, 2009.

Thomas Minka. Automatic choice of dimensionality for pca. *Technical Report 514, MIT Media Lab Vision and Modeling Group*, 02 2001.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks, 2025.

Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *J. Phys. C: Solid State Phys.*, 13:4071–4076, 1980.

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Classifier-free guidance: From high-dimensional analysis to generalized guidance forms, 2025.

G. Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.

B. Pham, G. Raya, M. Negri, M. J. Zaki, L. Ambrogioni, and D. Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv:2104.08894*, 2021.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Neural Information Processing Systems*. Conference on Neural Information Processing Systems, 2023.

T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv:1701.05517*, 2017.

Antonio Sclocchi, Alessandro Favero, and Mathieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1):e2408799121, 2025.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. ICML, 2015.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*. ICLR, 2020.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. *International Conference on Learning Representations*, 2025.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00142.

P. Wang, H. Zhang, Z. Zhang, S. Chen, H. Ma, and Q. Qu. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv:2409.02426.*, 2024.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv:2305.08694*, 2023.

M Weigt and R Monasson. Replica structure of one-dimensional disordered ising models. *Europhysics Letters (EPL)*, 36(3):209–214, October 1996. ISSN 1286-4854. doi: 10.1209/epl/i1996-00212-8.

Eugene Paul Wigner. *Group theory and its applications to the quantum mechanics of atomic spectra*. Academic Press, 1959.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint 1708.07747*, 2017.

Hongkang Yang and Weinan E. Generalization error of gan from the discriminator's perspective. *Research in the Mathematical Sciences*, 9, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Math. Control Signals Systems*, 94:103–114, 2017.

F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Appendix A

Network training and Model Architecture Details

Dataset	Image Size	Latent Dim.	Channel Mult.	Param. Count	Batch size	Iterations
Cifar10	32	128	(1, 2, 2, 2)	35.7M	128	500,000
MNIST	28	128	(1, 2, 2)	24.5M	128	400,000
CelebA	64	64	(1, 1, 2, 2, 4, 4)	27.4M	64	800,000

Table A.1: Table displaying both model and training configurations for each dataset.

A.1 Geometric phases

For the image datasets, we used the diffusion setting in [Ho et al., 2020]. We use the variance scheduler with $\beta_{\text{min}} = 10^{-4}$ and $\beta_{\text{min}} = 2 \times 10^{-2}$, $T = 1000$ time steps, and score model backbone (PixelCNN++ [Salimans et al., 2017]). Furthermore, for each of the datasets, we adjusted the partameters to account for the different complexity (see Table A.1). For each data-set, the number of used training data-points amounts to the full set of data available.

For the linear models, we used a Variance Exploding continuous score model trained with 2M steps (batch size 128). The model had a Residual architecture with size 128 hidden channels in each layer, two residual blocks comprised by two linear layers with SiLu. In this case, the number of used training-data amounts to $2 \cdot 10^5$ synthetic examples generated according to Section 4.3.1.

For all experiments we primarily utilized NVIDIA Tesla V100 GPUs with 32 GB of memory.

A.2 Losing dimensions

Table A.2: Table displaying both model and training configurations for each dataset.

Dataset	Image Size	Latent Dim.	Channel Mult.	Param. Count	Batch size	Iterations
Cifar10	32	128	(1, 2, 2, 2)	35.7M	128	500,000
Mnist	28	128	(1, 2, 2)	24.5M	128	400,000
Fashion-Mnist	28	128	(1, 2, 2)	24.5M	128	400,000
CelebA-HQ	64	64	(1, 1, 2, 2, 4, 4)	27.4M	64	800,000
Lsun-Church	64	96	(1, 1, 2, 2, 4, 4)	61.7M	64	800,000

For our toy models, we train a Variance Exploding continuous score model with 2M training steps with batch size 128. We use a Residual Multi Layer Perceptron with hidden size of 128, with two residual blocks. Each block is composed by two linear layers with SiLU activation.

For the image models, we follow the diffusion setting in [Ho et al., 2020]. We kept the variance scheduler, where $\beta_{\text{min}} = 10^{-4}$ and $\beta_{\text{min}} = 2 \times 10^{-2}$, the time steps $T = 1000$, and the score model backbone (PixelCNN++ [Salimans et al., 2017]) the same. In addition, for each of the datasets, we varied the model's channel multipliers, latent dimension, batch size, and training iterations to account for the complexity of the dataset and our available computing resources; see Table (A.2). For context, we primarily utilized NVIDIA Tesla V100 GPUs with 32 GB of memory for the training of our models.

For what concerns the data sizes that we chose for our experiments, we used the pretrained DDPM models in [Ho et al., 2020, Pham et al., 2025]: for Cifar10 [Krizhevsky et al., 2014], Mnist [Deng, 2012], Fashion-Mnist [Xiao et al., 2017], and Lsun-Church [Yu et al., 2015] datasets. Specifically, these models were trained using $M = 38$ different data split sizes with the goal of finely observing the memorization-to-generalization transition, which also allows us a comprehensive view of the reduction in the manifold size. For each dataset, the smallest model was trained on the training set of $N_1 = 2$ data points while the largest model was trained on the entire original training set $N_M = N$.

The selection of these data split sizes relied on spotting a point in which the memorization rate of the model plateaus and another point where its generalization rate increases. Then, linear spacing of 30 points was used between these two points. For example, for CelebA-HQ [Liu et al., 2015] models which we have to trained, these two points are located at 1000 and 16000 data sizes. We used linearly spacing of 30 points between these two points; while for points outside of the transition, we used linearly spacing of 5 points: from 2 to 1000 and 16000 to the full dataset size N. For Cifar10, Mnist, and Fashion-Mnist, center-crop and resizing were not used. However, images of the CelebA-HQ and

Lsun-Church datasets were both center-cropped and down-sampled to 64×64 resolution. Finally, we trained our CelebA-HQ models without random flipping and utilized the exponential moving average version of the trained models for our analyses, where we set the decay value to 0.9999 during training. Please refer to Tables (A.2)-(A.3) for additional details.

Table A.3: A table showing the critical points A and B of the memorization-generalization transition for each dataset.

	Training Data Size		
Dataset	Point A	Point B	Total
Mnist	4000	32000	60000
Cifar10	2000	16000	50000
CelebA-HQ	2000	16000	30,000
Lsun-Church	2000	16000	126227
Fashion-Mnist	4000	16000	60000

Appendix B

Measuring the intrinsic dimension of the data manifold

One method used to geometrically estimate the intrinsic dimension of the data manifold is an improved version of Normal Bundle (NB) method used in [Stanczuk et al., 2024]: the score function is measured across d orthogonal directions in the vicinity of the manifold and ordered as the columns of a squared matrix S; the singular values of the matrix S are computed and collected; the intrinsic dimension of the manifold is given by the $d - \ker(S)$, with the kernel is estimated directly from the spectrum of the singular values of S. The algorithm for the improved NB method is described in Section B and the spectral analysis is reported in Section B. On the other hand, we propose an alternative estimation method, and we describe it more technically in Section B. The procedure starts with extracting the singular values of Jacobian of the score function, as performed in the previous method described above. At this point we order the values by their magnitude and we compute the absolute value of the second derivative of the singular values, selecting the first bigger value with respect to the median multiplied by a threshold factor. We further discard the initial singular value as it tends to be large, resulting in instabilities. The selected singular value signals the formation of a drop in the ordered eigenspectrum, suggesting the beginning of the tangent subspace. Similarly to the previous method, by computing the dimension of the tangent subspace we obtain the best estimate for the latent manifold dimension. We found this method to be more robust than the one proposed in [Stanczuk et al., 2024], especially for high dimensional datasets where there is no sharp drop in the spectrum of the singular values.

Computing the Singular Values of the Jacobian of the score

In order to compute the singular values of the Jacobian of the score function, we make use of the prescription described in [Stanczuk et al., 2024] improved through a choice of orthogonal perturbations instead of purely random ones. The procedure is reported in Algorithm 1. For linear models and MNIST models we used a symmetrized version, that we call *central difference* method, which we empirically found to be more stable, reported in Algorithm 2. In the algorithms, the forward process $\mathcal{F}$ represents forward processes typically used in diffusion models, such as variance exploding and variance preserving noise schedules. Specifically, in the case of the variance exploding schedule, employed by our analysis, the forward process takes the form $\mathcal{F}(x_0, \epsilon, t) = x_0 + t\epsilon$.

Algorithm 1 Estimate singular values at x_0

Require: s_θ (trained score model), t_0 (sampling time), forward process $\mathcal{F}$

Sample $x_0 \sim p_0(x)$ from the data set
$d \leftarrow \dim(x_0)$
$S \leftarrow$ empty matrix
for $i = 1, \dots, d$ **do**
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    $x_{t_0}^{(i)} \leftarrow \mathcal{F}(x_0, \epsilon, t_0)$ ▷ perturbation
end for
$(x_{t_0}^{(i)})_{i=1}^d \leftarrow (\tilde{x}_{t_0}^{(i)})_{i=1}^d$ ▷ Orthogonalize the perturbations
for $i = 1, \dots, d$ **do**
    Append $s_\theta(\tilde{x}_{t_0}^{(i)}, t_0)$ as a new column to S
end for
$(s_i)_{i=1}^d,\ (v_i)_{i=1}^d,\ (w_i)_{i=1}^d \leftarrow \mathrm{SVD}(S)$

Spectral Analysis

Once we have extracted the collected the singular values of the Jacobian of the score function throughout Algorithms 1 and 2 we can divide their magnitudes by the highest one and order them from the highest to the slowest. The ordered eigenspectrum plot allows to visualize whether the diffusion model is constrained to sample from a lower dimensional manifold and to infer geometric insights about such manifold. Specifically, drops in magnitude at a certain singular value will display the presence of a gap in the spectrum, which is a signature of a net separation between the tangent and orthogonal subspaces with respect to a manifold. The local dimension of the manifold can be inferred by counting the number of singular values associated to a vanishing magnitude, i.e. relative to the right side of the drop. Figure B.1 depicts the time evolution of the ordered

Algorithm 2 Estimate singular values at x_0 with central difference

Require: s_θ (trained score model), t_0 (sampling time), forward process $\mathcal{F}$

Sample $x_0 \sim p_0(x)$ from the data set
$d \leftarrow \dim(x_0)$
$S \leftarrow$ empty matrix
for $i = 1, \dots, d$ **do**
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    $x_{t_0}^{+(i)} \leftarrow \mathcal{F}(x_0, \epsilon, t_0)$ ▷ right perturbation
    $x_{t_0}^{-(i)} \leftarrow \mathcal{F}(x_0, -\epsilon, t_0)$ ▷ left perturbation
end for
$(x_{t_0}^{+(i)}, x_{t_0}^{-(i)})_{i=1}^d \leftarrow (\tilde{x}_{t_0}^{+(i)}, \tilde{x}_{t_0}^{-(i)})_{i=1}^d$ ▷ Orthogonalize the perturbations
for $i = 1, \dots, d$ **do**
    Append $\dfrac{s_\theta(\tilde{x}_{t_0}^{+(i)}, t_0) - s_\theta(\tilde{x}_{t_0}^{-(i)}, t_0)}{2}$ as a new column to S ▷ central difference
end for
$(s_i)_{i=1}^d,\ (v_i)_{i=1}^d,\ (w_i)_{i=1}^d \leftarrow \mathrm{SVD}(S)$

eigenspectrum, estimated by the neural network model trained with dataset of different sizes.

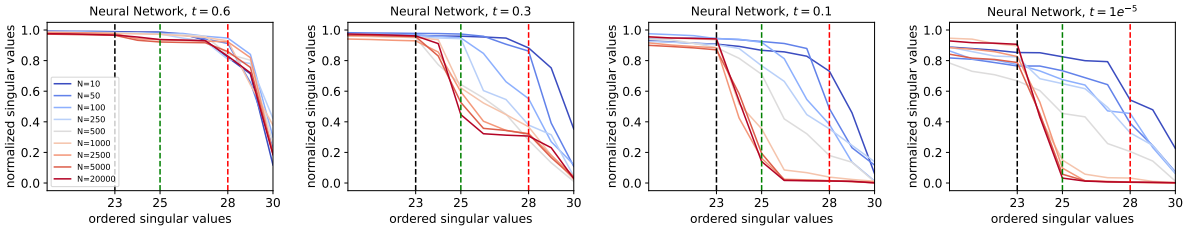


Figure B.1: Ordered singular values of the Jacobian of the score estimated by a neural network trained on a linear data model. The parameters for the model are $d = 30$, $m = 7$ with a subspace associated to a variance $\sigma_1^2 = 1$ of dimension $m_1 = 2$ and another subspace with variance $\sigma_2^2 = 0.3$ and dimension $m_2 = 5$. Different lines are associated to different sizes of the training set N, as reported in the legend.

Estimating the local latent dimension in real datasets

In this section we report Algorithm (3) that we used to find the local latent manifold dimension given the singular values. For Mnist, we used $\bar{d} = 100$, $c = 20$; for Cifar10, we used $\bar{d} = 1000$, $c = 10$; for CelebA-HQ, we used $\bar{d} = 1500$, $c = 10$; and for Lsun-Church, we used $\bar{d} = 2000$, $c = 20$. For all these datasets, we employed the same $t_0 = 2$ corresponding to the diffusion index in DDPM. With the exception of CelebA-HQ, which we used Algorithm (1), we primarily used the central-difference version explained

in Algorithm (2). We report an example of second derivative used in the method in Fig. (B.2).

Algorithm 3 Estimate intrinsic manifold dimension at x_0

Require: Singular values $(s_i)_{i=1}^{d}$ from Alg. 1 or 2, diffusion time t, data dimension d, threshold c, starting index $\bar{d}$

$d^2_{\text{svd}} \leftarrow \left|\frac{d^2}{ds^2} s_t[\bar{d}:]\right|$ ▷ second derivative tail spectrum

$m \leftarrow \text{median}(d^2_{\text{svd}})$

$n \leftarrow \#\{i \mid d^2_{\text{svd},i} > c \cdot m\}$ ▷ indices above threshold

$k \leftarrow d - n + \bar{d}$

return estimated intrinsic manifold dimension k

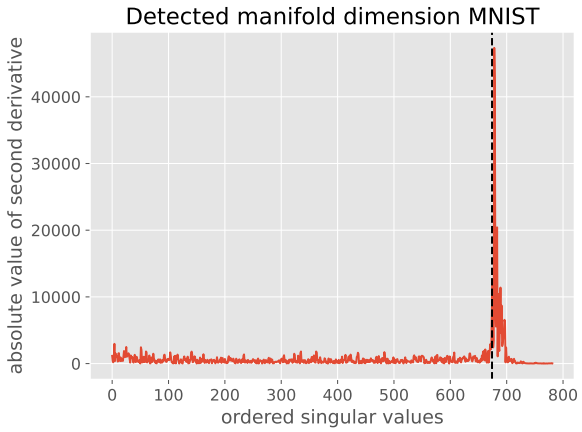


Figure B.2: The figure shows the absolute values of the second derivatives computed with algorithm 3, at a small diffusion time when the model is trained with a large batch of the MNIST dataset. The local dimension of the manifold can be estimated by counting the number of singular values on the right of the dashed line.