Thesis for the degree
Doctor of Philosophy

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
דוקטור לפילוסופיה

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Dan Mikulincer

מאת
דן מיקולינסר

אוניברסליות במערכות בעלות ממדיות גבוהה
Universality of High-Dimensional Systems

Advisor:
Prof. Ronen Eldan

מנחה:
פרופ' רונן אלדן

June, 2021

תמוז התשפ"א

“Arrakis teaches the attitude of the knife - chopping off what’s incomplete and saying: ’Now, it’s complete because it’s ended here.’”

Irulan Corrino, ”Collected Sayings of Maud’Dib”

Acknowledgments

It is customary to begin the Acknowledgments section by thanking one's advisor, which may sometimes create the impression of simply going through the motions. This is certainly not the case here, and I am genuinely happy to first thank my advisor, Ronen Eldan. It has been one hell of a ride through the entirety of my graduate studies. Ronen, you are the epitome of an Advisor. You were never overbearing but always knew to nudge in the right direction when necessary. I always felt like I could discuss everything with you in equal terms; even subjects I later understood were far deeper and more complex than originally envisioned. Thank you for introducing me to your approach to researching Mathematics, sharing your knowledge and thoughts, and teaching me the invaluable lesson: that sometimes (but only sometimes) it's better to prove what we can rather than we want.

I would also like to express my gratitude to Itai Benjamini, Sébastien Bubeck, Max Fathi, and Bo'az Klartag, whose mentorship was invaluable. I am very fortunate to have had the opportunity to discuss Mathematics with you and have greatly benefited from the interactions. Thank you for your patience and availability, and above all, thank you for your friendship. A special thanks goes to Seb and the AI group in Microsoft Research for hosting me in the summer of 2019.

To my other collaborators and co-authors: Itay Glazer, Yin Tat Lee, Hester Pieters, Tselil Schramm, Yair Shenfeld, and Alex Zhai. I am indebted to you for our shared work. I am honored to have worked beside you and have learned a lot from every one of you.

My time at the Weizmann Institute, albeit long, was an extremely pleasant one. This is mostly due to many good friends and office mates I have had the pleasure to meet over the years. I am very lucky to have met you all. The list is long, without doubt, and I am bound to forget someone. Still, I will make an effort to thank the following (possibly partial) list personally: Fanny Augeri, Daniel Benarroch, Snir Ben Ovadia, Raphael Butez, Tal Cohen, Boaz Elazar, Itay Glazer, Gil Goffer, Uri Grupel, Yotam Hendel, Yael Hitron, Alon Ivtsan, Maya Leshkowitz, Itay Safran, Ary Shaviv, Ofer Shwartz, Yair Shenfeld, and Eliran Subag.

I also wish to thank the extended probability group and its affiliates: Tsachik Gelander, Gady Kozma, Omri Sarig, and Ofer Zeitouni. Thank you for creating a relaxed and inspiring environment for conducting research.

Lastly, this section cannot be complete without thanking my family. To my brother Alon and parents Mario and Debby, thank you for your support during these times and generally in life, and thank you for encouraging me to study something instead of fulfilling my dreams of flipping burgers. To my young daughter Mai, who reminds me every day that there is beauty in simplicity and makes everything worthwhile. And last, but certainly not least, to my wife Mor, who had to live through this crazy period of deadlines, far-off conferences, late-night working, and just me idly gazing at the wall, thinking about some problem while making a supposed conversation. You continue to be the stars in my sky, and I am sure that together we will reach much further than that.

Declaration

I hereby declare that this thesis summarizes my original research, performed under the guidance of my advisor Prof. Ronen Eldan. Other than that I have enjoyed the following collaborations:

- The work presented in Chapter 1 was done in collaboration with Alex Zhai.
- The work presented in Chapter 3 was done in collaboration with Tselil Schramm.
- The work presented in Chapter 6 was done in collaboration with Max Fathi.
- The work presented in Chapter 7 was done in collaboration with Sébastien Bubeck.
- The work presented in Chapter 8 was done in collaboration with Sébastien Bubeck and Yin Tat Lee.
- The work presented in Chapter 9 was done in collaboration with Hester Pieters.

Dan Mikulincer

Contents

Abstract

The main theme explored in this thesis is the interplay between dimension and probabilistic phenomena. The 'curse of dimensionality' is a well-known heuristic that suggests that the complexity of problems should scale exponentially with the dimension. We choose the path of optimism and are interested in identifying unifying and universal structures of high-dimensional distributions that may circumvent the 'curse of dimension.'

In Part I we revisit the classical Central Limit Theorem (CLT) and extend it to higher, and even infinite, dimensions. The main novelty of our results is that the rate of convergence to the normal distribution is explicit. In particular, we show that this rate is typically polynomial in the dimension for a large class of measures. In proving these results, we introduce new methods to establish the validity of normal approximations. These methods are based on a combination of stochastic analysis with Stein's method and optimal transport.

Part II is dedicated to the study of stability properties for several functional inequalities. Loosely speaking, a functional inequality is said to be stable if the following implication holds: Whenever a measure almost saturates the inequality, it must hold that the measure is close, in some sense, to another measure that attains equality. By extending the tools developed in Part I, we identify the normal distribution as the unique equality cases for various functional inequalities and establish dimension-free stability estimates for these inequalities. Thus, we obtain new criteria for normal approximations which are qualitatively different from the CLT.

Finally, in Part III we apply our results to answer questions arising from learning and optimization theory. In the first work, we resolve a long-standing open question concerning the optimal complexity of finding stationary points of non-convex functions. In another work, we study neural networks and introduce several new algorithms to construct efficient networks with minimal size. The last work is dedicated to community detection in geometric random graphs, where we obtain both sufficient and necessary conditions for the possibility of this task.

Due to space and time constraints, several additional works are not included in this thesis. These works revolve around anti-concentration of polynomial with log-concave variables, rapid decay of Fourier coefficients and infinite dimensional generalizations of optimal transport maps.

Introduction

Contents

High-dimensional problems have been the focus of much research in recent years. The results have been applied in various subjects such as theoretical computer science and machine learning. The central theme of this thesis is the investigation of phenomena in high dimensions. A particular case of interest lies in *dimension-free* phenomena, which hold the same in any dimension.

The well-known "curse of dimensionality" phenomenon suggests, informally, that the complexity of some problems scales exponentially with the dimension. Thus, at first glance, it would seem that many high-dimensional problems should be intractable. However, there is balm as well as bitterness, and some high-dimensional systems exhibit universality properties which can make their study actually easier, or at least not harder, than their low-dimensional counterparts. Identifying such properties is often crucial for applications.

A familiar and simple example of this philosophy can be seen in the central limit theorem (CLT). The CLT states that if $(X_i)_{i=1}^n$ are *i.i.d.* random variables, their sum converges to a Gaussian. This can be seen as a statement concerning the space $\mathbb{R}^n$ equipped with a product measure which is the joint law of $(X_i)_{i=1}^n$. Thus, as the dimension increases, in some sense, this measure becomes easier to understand.

A large part of this thesis is dedicated to understanding the theory behind this type of universality phenomena. The over-arching goal is to establish sufficient conditions for such *normal approximations* in different settings. This is achieved by combining tools from different fields, including, among others, ideas stemming from Stein's theory, stochastic calculus, optimal transport, and Malliavin calculus.

The main results in this direction can be classified into two main categories. The first includes quantitative extensions of the central limit theorem to higher and even infinite dimensions. In the second, we establish the standard Gaussian as the stable extremal case for a family functional inequalities. Thus, any measure which comes close to saturating the inequality can be well approximated by the Gaussian. In the course of proving these results, we develop new tools as well as expand upon existing ones.

Another part of this thesis focuses on applying the above and similar results to problems in neighboring fields, such as data science and machine learning. A recurring concept in such applications is the interplay between stochastic processes and learning and optimization algorithms. We show several instances where careful analysis of a carefully chosen stochastic process can lead to information-theoretical lower bounds on various algorithms.

High-Dimensional Central Limit Theorems

The central limit theorem, first proved by Laplace in 1810 ([177]), is one of the most important results in mathematical statistics and probability. The main objects the CLT deals with are sequences X_i of *i.i.d.* random vectors in $\mathbb{R}^d$, and their normalized partial sums $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$. The CLT states that as $n \to \infty$, S_n converges to a Gaussian law, in a suitable metric.

Despite its significance, this formulation of the CLT may be unsuitable for certain applications. For instance, the rate at which convergence happens is also important. This rate tells us how well can a finite sum be approximated by a Gaussian, which is often crucial. In the 1940s Berry and Esseen independently managed to bound from the above the aforementioned rate ([35], [108]) in the one-dimensional case.

However, in modern statistics, we are often interested in cases where the dimension, d, *scales as a function of* n. A natural question arises in this setting: How should d depend on n to ensure that S_n converges to a Gaussian? Berry-Esseen's inequality is inappropriate to answer such a question, and the situation in high-dimensions has attracted much attention over the years ([22, 32, 42, 188, 217, 251]). In the first part of this thesis, we establish quantitative convergence rates with explicit and often optimal dependence on the dimension in various settings.

In Chapter 1, which is based on the paper [106] we have mainly focused on log-concave measures and proved a quantitative entropic CLT. Thus, if $\{X_i\}_{i=1}^n$ are *i.i.d.* log-concave vectors in $\mathbb{R}^d$ and $S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$, then, there exists a Gaussian vector G, such that,

$$\mathrm{Ent}(S_n||G) = O\left(\frac{\mathrm{poly}(d)}{n}\right).$$

Moreover, if X_1 happens to be uniformly log-concave then the bound improves to the optimal,

$$\mathrm{Ent}(S_n||G) = O\left(\frac{d}{n}\right).$$

The above results are the first to give a polynomial dependence on the dimension and apply to the general class of log-concave measures. In particular, the linear dependence for strongly log-concave measures is optimal. Previous results either gave an exponential dependence on the dimension ([44]) or required stronger assumptions such as finiteness of Fisher information

([85]). The chapter also goes beyond the log-concave setting. A similar result is proven for random vectors with bounded support, at the cost of using the weaker quadratic Wasserstein metric $\mathcal{W}_2$.

In the proofs of these results, we've introduced a new method to derive quantitative convergence rates. The technique uses stochastic analysis and is based on carefully chosen processes, which encode entropy. To deal with the quadratic Wasserstein distance, we constructed a high-dimensional counterpart to the process introduced by Eldan in [98].

In another line of work ([180]), on which we base Chapter 2, we have considered the CLT for empirical moment tensors. The main objects of interest are normalized sums of independent copies of $X^{\otimes p}$ for some integer $p \geq 2$. The problem has been studied before ([53, 56, 58, 102, 110, 197]) in different settings, mostly when $p = 2$ and focusing solely on cases where X has independent coordinates.

Since the random vector $X^{\otimes p}$ is constrained to lie in a d-dimensional manifold of the tensor space, standard tools, such as those introduced in Chapter 1 cannot be applied directly, if one wishes to obtain optimal convergence rates. In the chapter we introduce a novel implementation of Stein's method to exploit the latent low-dimensional structure of $X^{\otimes p}$. It is shown that under some regularity assumptions, but with no requirement of independence,

$$\mathcal{W}_2^2(S_n||G) = O\left(\frac{d^{2p-1}}{n}\right).$$

where

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(X_i^{\otimes p} - \mathbb{E}\left[X_i^{\otimes p}\right]\right).$$

We hope that our results may be helpful in similar settings where one considers singular measures which are supported on low-dimensional manifolds of the ambient space.

The final extension of the CLT appears in Chapter 3 and the paper [105] deals with random functions in infinite dimensional spaces. Most examples of central limit theorems deal with random vectors in some finite-dimensional space. To go beyond this setting, consider the $d-1$-dimensional sphere $\mathbb{S}^{d-1}$ and random processes indexed by $\mathbb{S}^{d-1}$, which are essentially random elements taking values in the infinite-dimensional space $L^2(\mathbb{S}^{d-1})$.

One particular case of interest is the random function $F(x) := \psi(G \cdot x)$, for some function $\psi : \mathbb{R} \to \mathbb{R}$ and G a standard Gaussians. In the chapter we study the rate at which $\mathcal{P}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F_i$ converges to a Gaussian process in an L^p transportation metrics. Our approach is to embed the processes into some high-dimensional tensor space. In such spaces, the results obtained in Chapter 2 apply, and we establish several quantitative convergence results, which depend upon the regularity of ψ. Processes like $\mathcal{P}_n$ are ubiquitous in learning theory. They ap-

pear naturally in algorithms involving neural networks (see [139] for example). Understanding the extent to which these processes may be approximated by Gaussian processes can further elucidate the behavior of such algorithms. Prior to this present work there were no quantitative results in the literature.

Stability of Functional Inequalities

The second part of the thesis deals with stability of functional inequalities. Suppose that $H, K : P(\mathbb{R}^d) \to \mathbb{R}$ are two functionals which assign a number to each probability measure, such that, for every $\mu \in P(\mathbb{R}^d)$,

$$H(\mu) \leq K(\mu). \tag{1}$$

If $E = \{\mu \in P(\mathbb{R}^d) | H(\mu) = K(\mu)\}$ is the set of extremizers, then, at the informal level, we say that inequality (1) is stable, if we have the implication:

$$K(\mu) - \varepsilon \leq H(\mu) \implies \mu \text{ is close to a measure in } E.$$

It turns out that for many well-known inequalities the set of extremizers contains only Gaussians. Thus proving that such inequalities are stable yields a criterion for determining when a measure can be approximated by a Gaussian.

Chapter 4 is devoted to the celebrated Shannon-Stam inequality ([219]). According to the inequality, if X and Y are independent copies of one another and G is a Gaussian with the same covariance as X, then

$$\mathrm{Ent}\left(\frac{X+Y}{\sqrt{2}} || G\right) \leq \mathrm{Ent}(X||G).$$

Moreover, Gaussians are known the only equality cases. One could compare this with the CLT in which we deal with the n-fold convolution of random vectors. In this sense, the stability of this inequality is a more delicate question that asks what happens after a single convolution. In [20], Ball, Barthe, and Naor, gave the first quantitative version of this inequality for one-dimensional random variables. This result was later generalized in [21] for log-concave vectors in any dimension. For X and Y, independent copies of an isotropic log-concave random vector, the result bounds from below the deficit in the inequality in terms of $\mathrm{Ent}(X||G)$.

From the perspective of stability, in [84], Courtade, Fathi and Pananjady gave a similar result, where they removed the restriction that X and Y have identical laws. This was done at the cost of restricting the result to strongly log-concave vectors and the weaker quadratic Wasserstein distance. The paper raised the question of whether an equivalent stability result might hold for general log-concave vectors under the relative entropy distance. The main result of Chapter 4 expands the method introduced in Chapter 1 to give an affirmative answer to this question. The chapter is based on the paper [103].

Chapter 5 is a follow-up to the previous chapter and is based on the paper [179]. In this chapter we explore stability properties of other functional inequalities. Namely, Talagrand's transport entropy inequality [229] and Gross' log-Sobolev inequality [130] which respectively state,

$$\mathcal{W}_2^2(X,G) \leq 2\mathrm{Ent}(X||G) \text{ and } 2\mathrm{Ent}(X||G) \leq \mathrm{I}(X||G).$$

Here, $\mathrm{I}(X||G)$ stands for the relative Fisher information of X. The main result can be stated in the following way. Let X be a log-concave vector on $\mathbb{R}^d$, with a spectral gap c. Then

$$\begin{aligned} 2\mathrm{Ent}(X||G) - \mathcal{W}_2^2(X,G) &\geq c\mathrm{Ent}(X||G) \\ \mathrm{I}(X||G) - 2\mathrm{Ent}(X||G) &\geq c\mathrm{Ent}(X||G). \end{aligned} \tag{2}$$

Stability properties of these inequalities have been studied by various authors (see [39, 81, 101, 113, 116, 155, 162]). The main contribution of [179] was to show that the two inequalities stem from the same general principles and so their stability properties are related. This led to the strengthening of many previously known results, both by considering larger classes of random vectors and by establishing stability bounds in the stronger relative entropy distance.

Moreover, it is well-known that the above inequalities are intimately related to the concentration of measure phenomenon. Since the bounds we obtain are uniform over the class of measures which are log-concave with respect to the Gaussian, we also prove as a corollary an improved concentration bound for convex functions of the Gaussian.

In Chapter 6 we consider a family of functional inequalities, which arise naturally from a recent approach (see [68]) to Stein's method. A matrix valued map τ_X is said to be a Stein kernel for a random vector X, if the following integration by parts formula holds,

$$\mathbb{E}\left[\langle \nabla f(X), X\rangle\right] = \mathbb{E}\left[\langle \tau_X(X), \mathrm{Hess} f(X)\rangle_{HS}\right].$$

The existence of such kernels was explored in a variety of settings (see [85, 112, 194] for examples). In fact, the main ingredient in the proofs given in Chapter 2 was a construction of a Stein kernel for $X^{\otimes p}$.

The seminal paper of Ledoux, Nourdin and Peccati, [161], shows that Stein kernels may be used to bound some known distances. In particular, they prove the following inequality,

$$\mathcal{W}_2^2(X,G) \leq \mathbb{E}\left[\|\tau_X(X) - \mathrm{I_d}\|_{HS}^2\right].$$

where $\mathcal{W}_2^2$ stands for the quadratic Wasserstein distance and G is a standard Gaussian. Since $\mathrm{I_d}$ is a Stein kernel for the standard Gaussian, this may be seen as a stability estimate.

The main goal of this chapter is to go beyond the setting of the standard Gaussian. Based

on the paper [114], we prove an inequality of the following form,

$$\mathcal{W}_2^2(X,Y) \lesssim \mathbb{E}\left[\|\tau_X(X) - \tau_Y(X)\|_{HS}^2\right].$$

where τ_X, τ_Y are Stein kernels for X and Y, respectively.

Our main insight was that Stein kernels may be used to construct stochastic processes with the laws of X and Y as invariant measures. This reduced the above question to show that if two Itô processes have similar coefficients, then their invariant measures must be close. By adding to a line of work concerning the stability of solutions to differential equations, we showed that this is indeed the case under very general conditions. Besides being mathematically aesthetic, such a result says that we may efficiently sample from a given measure as long as we can approximate a Stein kernel to some numerical accuracy.

Applications in Data Science

We now describe the contents of the third part of this thesis in which we apply ideas from high-dimensional geometry to problems which come from learning and optimization theory. The connecting thread between all topics is that we are able to formulate exact quantitative statements with respect to the dimension.

Chapter 7, based on [60], deals with a basic setting in optimization theory. We are given a function $f : \mathbb{R}^d \to \mathbb{R}$ and the goal is to find a point $x \in \mathbb{R}^d$ such that $f(x)$ is approximately minimal. In case f is convex, there is a well-established theory, and much is known ([55]). On the other hand, if f is not convex, then, in general, finding its minimum is not a tractable problem. Instead, as long as the function is differentiable, we content ourselves with finding approximate stationary points. That is, some $x \in \mathbb{R}^d$ for which $\|\nabla f(x)\|$ is small.

Under some minimal regularity assumptions, one can always find a stationary point of f by simply following the direction of steepest descent, i.e., advancing in the direction opposite of f's gradient. In fact, by carefully following this procedure it can be shown that after taking about $\frac{1}{\varepsilon^2}$ steps in these directions one can find a point $x \in \mathbb{R}^d$ such that $\|\nabla f(x)\| < \varepsilon$. This is true in any dimension.

In [237], Vavasis showed that when the dimension, d, is fixed, one can modify the gradient descent procedure and obtain better results. Focusing on the case $d = 2$, Vavasis showed that $\frac{1}{\varepsilon}$ steps suffice for finding an ε-stationary point. However, the best possible rate remained open. Our main contribution in this setting is a new algorithm that improves upon Vavasis' algorithm in any fixed dimension. In particular, when $d = 2$, the algorithm queries the function at most $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ times. We also complement our result by showing that our algorithm has optimal performance. Thus, any other algorithm must make at least $\Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$ queries, in the worst case, to find an approximate stationary point.

We base our algorithm on new ideas, different than gradient descent. Rather than following the trajectory of the gradient, our algorithm preemptively predicts the location of the gradient flow for future times. This allows us to outperform all previously known algorithms. For the lower bound, we construct a distribution over hard instances, inspired by unpredictable random walks, as introduced by Benjamini, Pemantle, and Peres in [31]. We use these random walks to construct families of functions which are hard to optimize.

Chapter 8 is devoted to neural networks, which are a very popular learning model, used in data science. In recent years they have proved successful for a wide range of applications, such as computer vision and natural language processing. The basic idea is to represent arbitrarily complex functions as combinations of many simple ones, called neurons. An example of a simple neural network is a function $N : \mathbb{R}^d \to \mathbb{R}$, of the form,

$$N(x) = \sum_{i=1}^{k} a_i \sigma(w_i \cdot x),$$

where for each $i = 1, \cdots, k$, $a_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$ and σ is a non-linear function. k is also called the width of the network.

The basic procedure of training a neural network consists of receiving some labeled training set $(x_j, y_j)_{j=1}^n$, where $x_j \in \mathbb{R}^d$ and $y_j \in \mathbb{R}$. The goal is then to find a network N, such that for every $j = 1, ..., n$, $N(x_j) \simeq y_j$. We say that such a network fits the training set. It is well known ([88,166]) that any training set may be well approximated by some network with a large enough width. Remarkably, even though this suggests that networks can fit random noise, in many practical settings when $y_j = f(x_j)$ for some function of interest, the obtained network is a good approximation for the function f, even outside of the training set. It is precisely this fact that makes neural networks so popular in practice. However, the theory of this apparently paradoxical phenomenon is poorly understood.

The chapter is based on [57], in which we tried to make a small step towards better understanding the theory of neural networks. Specifically, we tackled the question of how large should k be with respect to the dimension d, and n, so that there even exists a network which fits a training set of size n. We've obtained a nearly optimal bound. Let $\varepsilon > 0$, if (x_i) are in general position, then we show that there exists a neural network N of width $k = O(\frac{n}{d})$, such that

$$\frac{1}{n} \sum_{j=1}^{n} \left(N(x_j) - y_j\right)^2 \leq \varepsilon.$$

Note that $d \cdot k$ is the number of tunable parameters in the neural network. Thus, for general data sets, it must hold that $k \geq \frac{n}{d}$. Although the question was extensively studied in previous years [5, 54, 89, 90, 96, 149, 201, 222], the optimal bound $k \geq \frac{n}{d}$ remained elusive prior to our work. We actually give several different constructions which afford the above bound. Among

these constructions, our main contribution is based on adding complex weights and on an orthogonal decomposition of $L^2(\mathbb{R}^d, \gamma)$. We hope that this construction may prove useful beyond the task of memorization.

In the final Chapter 9 we study the problem of extracting information from large graphs. Problems of this sort may be encountered while researching social networks, gene networks, or (biological) neural networks. A particularly important task in this spirit is to learn a useful representation of the vertices, that is, a mapping from the vertices to some metric space. Consider the following instance of a random geometric graph (RGG). The vertices of the graph $\{X_i\}_{i=1}^d$ are uniformly distributed on the $d-1$-dimensional sphere $\mathbb{S}^{d-1}$ and for $i \neq j$, (X_i, X_j) forms an edge with probability $\varphi(\langle X_i, X_j \rangle)$ for some function $\varphi : [-1, 1] \to [0, 1]$. When the number of vertices remains fixed, and $d \to \infty$, a standard application of the CLT shows that this random model is indistinguishable from an Erdős–Rényi graph. Since the edges are of an Erdős–Rényi graph are independent, we refer to this graph as having no meaningful geometric interpretation.

One can ask many questions in this setting. The chapter is focused on recovering the latent position of the vertices, up to the symmetries of the sphere, when observing a single instance of the graph. We introduce a spectral algorithm that works in every fixed dimension and recovers said positions. In essence, this is a geometric generalization of community detection in the stochastic block model (see [2]). We also provide an impossibility result by proving the first known lower bounds for recovery in this model. The proofs are based on the analysis of spherical random walks and exploit the symmetries of the sphere. The chapter is based on the paper [104].

Preliminaries

Overview of Methods

Here we present some of the recurring methods and tools which appear throughout this thesis. The main goal of this section is to supply the necessary background needed for this thesis and we will often refer to this section in the chapters to come. Other than that, we will also explain how to apply the techniques in some simple scenarios. These applications will later be expanded in the relevant chapters.

A list of common notations and definition is attached, for the convenience of the reader, at the end of this section.

The Föllmer Process

The first method to be presented is based on an entropy minimizing process, known in the literature as the Föllmer process. The high-level idea which underlies the method is to use the

process in order to embed a given measure as the terminal point of some martingale, in the Wiener space. This will induce a coupling between the measure and γ. As will be shown, the process also solves a variational problem, which turns out to yield a representation formula for the relative entropy. Combining these two properties will prove beneficial in the study of functional inequalities, which involve entropy and transportation distances.

The process first appeared in the works of Föllmer ([119,120]). It was later used by Borell in [48] and Lehec in [165] to give simple proofs of various functional inequalities. In this section, we present the relevant details concerning the process. The reader is referred to [100, 106, 165] for further details and a more rigorous treatment.

Throughout this section we fix a centered measure μ on $\mathbb{R}^d$ with a finite second moment matrix and a density f, relative to γ. Consider the Wiener space $C([0,1],\mathbb{R}^d)$ of continuous paths with the Borel sigma-algebra generated by the supremum norm $\|\cdot\|_\infty$. We endow $C([0,1],\mathbb{R}^d)$ with a probability measure P and a process B_t which is a Brownian motion under P. We will denote by ω elements of $C([0,1],\mathbb{R}^d)$ and by $\mathcal{F}_t$ the natural filtration of B_t. Define the measure Q by

$$\frac{dQ}{dP}(\omega) = f(\omega_1).$$

Q is absolutely continuous with respect to P, in which case, a converse to Girsanov's theorem implies that there exists a drift, v_t, adapted to $\mathcal{F}_t$, in the Wiener space, such that the process

$$X_t := B_t + \int_0^t v_s(X_s)ds, \tag{3}$$

has the same law as Q, and that, under Q, X_t is a Brownian motion. In particular, by construction, under P, $X_1 \sim \mu$ and, since $\frac{dQ}{dP}(\omega)$ depends only on ω_1, conditioned on X_1, X_t serves a Gaussian bridge between 0 and X_1. Thus, by the representation formula for Brownian bridges

$$X_t \overset{\text{law}}{=} tX_1 + \sqrt{t(1-t)}G, \tag{4}$$

where G is a standard Gaussian, independent from X_1. We call $v_t(X_t)$ the Föllmer drift and X_t the Föllmer process. As μ and γ are the laws of X_1 and B_1, it is now immediate that

$$\mathrm{Ent}(Q||P) \geq \mathrm{Ent}(\mu||\gamma). \tag{5}$$

A remarkable feature is that, since $\frac{dQ}{dP}$ depends only on the terminal points, the above is actually an equality and $\mathrm{Ent}(Q||P) = \mathrm{Ent}(\mu||\gamma)$. This implies that the drift, v_t, is a martingale (see Lemmas 10 and 11 in [165]).
We now use Girsanov's theorem ([199, Theorem 8.6.3]) to rewrite $\frac{dQ}{dP}$ as an exponential mar-

tingale,

$$\frac{dQ}{dP}(\omega) = \exp\left(-\int_0^1 v_t(\omega)dX_t(\omega) + \frac{1}{2}\int_0^1 \|v_t(\omega)\|_2^2\, dt\right).$$

Under Q, X_t is a Brownian motion, so

$$\text{Ent}\,(Q||P) = \int_{C([0,1],\mathbb{R}^d)} \ln\left(\frac{dQ}{dP}\right) dQ = \frac{1}{2}\int_0^1 \mathbb{E}\left[\|v_t(X_t)\|_2^2\right] dt,$$

which gives the formula

$$\text{Ent}(\mu||\gamma) = \frac{1}{2}\int_0^1 \mathbb{E}\left[\|v_t(X_t)\|_2^2\right] dt. \tag{6}$$

For simplicity, from now on, we suppress the dependence of v_t on X_t. Combining the above with (5) shows that among all adapted drifts u_t such that $\mu \sim B_1 + \int_0^1 u_t dt$, v_t minimizes the energy in the following sense

$$v_t = \arg\min_{u_t} \frac{1}{2}\int_0^1 \mathbb{E}\left[\|u_t\|_2^2\right] dt. \tag{7}$$

Theorem 12 in [165] capitalizes on the structure of $\frac{dP}{dQ}$ to give an explicit representation of v_t as

$$v_t = \nabla \ln\left(P_{1-t}f(X_t)\right). \tag{8}$$

where P_{1-t} denotes the heat semi-group. Since v_t is a martingale, Itô's formula shows

$$dv_t = \nabla v_t dB_t = \nabla^2 \ln\left(P_{1-t}f(X_t)\right) dB_t.$$

The martingale approach: As is often the case, it turns out that it is easier to work with an equivalent martingale formulation of the Föllmer drift. Consider the Doob martingale $\mathbb{E}\left[X_1|\mathcal{F}_t\right]$. By the martingale representation theorem ([199, Theorem 4.33]) there exists a uniquely-defined, adapted, matrix valued process Γ_t which satisfies

$$\mathbb{E}\left[X_1|\mathcal{F}_t\right] = \int_0^t \Gamma_s dB_s. \tag{9}$$

We claim that

$$v_t = \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s. \tag{10}$$

Indeed, by Fubini's theorem

$$\int_0^1 \Gamma_s dB_s = \int_0^1 \mathrm{I}_d dB_s + \int_0^1 (\Gamma_s - \mathrm{I}_d)\, dB_s = B_1 + \int_0^1 \int_s^1 \frac{\Gamma_s - \mathrm{I}_d}{1-s} dt dB_s$$
$$= B_1 + \int_0^1 \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s dt.$$

For the moment denote $\tilde{v}_t := \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s$. Since v_t is a martingale $v_t - \tilde{v}_t$ is a martingale as well and the above shows that for every $t \in [0,1]$, almost surely,

$$\int_t^1 (v_s - \tilde{v}_s) ds | \mathcal{F}_t = 0.$$

This implies the identity (10). In particular, from (8), Γ_t turns out to be symmetric, which shows, using Itô's formula,

$$2\mathrm{Ent}(\mu||\gamma) = \int_0^1 \mathbb{E}\left[\|v_t\|_2^2\right] dt = \mathrm{Tr} \int_0^1 \int_0^t \frac{\mathbb{E}\left(\Gamma_s - \mathrm{I}_d\right)^2}{(1-s)^2} ds dt = \mathrm{Tr} \int_0^1 \frac{\mathbb{E}\left(\Gamma_t - \mathrm{I}_d\right)^2}{1-t} dt. \tag{11}$$

Also, note that

$$B_1 + \int_0^1 (\Gamma_t - \mathrm{I}_d)\, dB_t = \int_0^1 \Gamma_t dB_t \sim \mu,$$

which implies

$$\mathcal{W}_2^2(\mu, \gamma) \leq \mathrm{Tr} \int_0^1 \mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right] dt. \tag{12}$$

As $X_1 \sim \mu$, from (8) we get

$$\mathrm{Tr} \int_0^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]}{(1-t)^2} dt = \mathbb{E}\left[\|v_1\|_2^2\right] = \int_{\mathbb{R}^d} \|\nabla \ln(f(x))\|_2^2 \, d\mu(x) = \mathrm{I}(\mu||\gamma). \tag{13}$$

Combining (11),(12),(13), we see a very satisfying connection between the log-Sobolev and Talagrand's transport-entropy inequalities (as in (2)), for

$$\mathrm{Tr} \int_0^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]}{(1-t)^2} dt \geq \mathrm{Tr} \int_0^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]}{1-t} dt \geq \mathrm{Tr} \int_0^1 \mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right] dt,$$

implies

$$\mathrm{I}(\mu||\gamma) \geq 2\mathrm{Ent}(\mu||\gamma) \geq \mathcal{W}_2^2(\mu,\gamma).$$

These ideas will be revisited in Chapters 4 and 5, where we will introduce a similar proof for the Shannon-Stam inequality and establish stability estimates and strengthenings for the above inequalities.

Central limit theorems via the Föllmer process: Let us also show how one can prove central limit theorems by using the Föllmer process. Let $\{X_i\}_{i=1}^n$ be *i.i.d.* as μ and let $\{B_t^i\}_{i=1}^n$ be independent Brownian motions with $\{\Gamma_t^i\}_{i=1}^n$ *i.i.d.* copies of Γ_t, the Föllmer martingale associated to μ, predictable with respect to B_t^i.

If $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, then $S_n \overset{law}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_1^i$ where $dY_t^i = \Gamma_t^i dB_t^i$. By standard reasoning about Gaussians random variables, we may write $\sum_{i=1}^n Y_t^i$ as another process $\widetilde{Y}_t$ with $d\widetilde{Y}_t = \sqrt{\frac{1}{n}\sum_{i=1}^n (\Gamma_t^i)^2}\, d\widetilde{B_t}$ and $\widetilde{B_t}$ is some Brownian motion. Denote $\widetilde{\Gamma}_t = \sqrt{\frac{1}{n}\sum_{i=1}^n (\Gamma_t^i)^2}$ and note that

$$\widetilde{Y}_t = \int_0^t \mathbb{E}[\widetilde{\Gamma}_t] d\widetilde{B_t} + \int_0^t \left(\widetilde{\Gamma}_t - \mathbb{E}[\widetilde{\Gamma}_t]\right) d\widetilde{B_t}.$$

Let $\tilde{\gamma} := \int_0^1 \mathbb{E}[\widetilde{\Gamma}_t] d\widetilde{B_t}$. Observe that $\tilde{\gamma}$ follows some (non-standard) Gaussian law. Moreover, there is a natural coupling between $\tilde{\gamma}$ and $\widetilde{Y}_1$, which shows, using Itô's isometry

$$\begin{aligned}\mathcal{W}_2^2\left(\tilde{\gamma}, \widetilde{Y}_1\right) &\leq \mathbb{E}\left[\left(\int_0^1 \left(\widetilde{\Gamma}_t - \mathbb{E}[\widetilde{\Gamma}_t]\right) d\widetilde{B_t}\right)^2\right] = \mathbb{E}\left[\int_0^1 \left(\widetilde{\Gamma}_t - \mathbb{E}[\widetilde{\Gamma}_t]\right)^2 dt\right] \\ &= \int_0^1 \mathrm{Var}\left(\widetilde{\Gamma}_t\right) dt,\end{aligned}$$

where the last equality is due to Fubini's theorem. Recall that $\widetilde{\Gamma}_t = \sqrt{\frac{1}{n}\sum_{i=1}^n (\Gamma_t^i)^2}$. So, the law of large numbers suggests that as $n \to \infty$, $\widetilde{\Gamma}_t \to \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}$ and, in particular, $\mathcal{W}_2^2\left(\tilde{\gamma}, \widetilde{Y}_1\right) \to 0$, which is the central limit theorem. This idea is made both rigorous and quantitative in Chapter 1.

Stein's method via Stein kernels

The second method to appear prominently in this thesis is a somewhat modern manifestation of Stein's method through the so-called Stein kernels. Stein's method is a well-known set of techniques which was developed in order to answer questions related to convergence rates along the CLT. The method was first introduced in [224, 225] as a way to estimate distances to the normal law. Since then, it had found numerous applications in studying the quantitative central limit theorem, also in high-dimensions (see [211] for an overview).

At the heart of the method lies the following observation, sometimes called Stein's lemma ([70, Lemma 2.1]): If $G \sim \gamma_d$, is a standard Gaussian, and $f : \mathbb{R}^d \to \mathbb{R}^d$ is locally-Lipschitz, then a simple application of integration by parts shows,

$$\mathbb{E}\left[\langle G, f(G)\rangle\right] = \mathbb{E}\left[\mathrm{div} f(G)\right]. \tag{14}$$

Moreover, by letting f vary across monomial functions, we can see that (14) completely determines the moments of G. Since the standard Gaussian is determined by its moments, it is the only measure that satisfies (14).

Stein came up with the remarkable idea that this property is robust. Thus, if a measure μ 'approximately satisfies' (14) it should be 'close' to γ_d. Considerable work revolving around Stein's theory focused on making this idea quantitative. A recent approach, pioneered by the seminal paper [68] of Chatterjee, replaced the right hand side of (14) by a more general first-order differential operator, called a Stein kernel. A Stein kernel for a measure μ is defined as measurable matrix valued map $\tau : \mathbb{R}^n \to \mathcal{M}_n(\mathbb{R})$ which satisfies,

$$\int \langle x, f(x)\rangle d\mu(x) = \int \langle \tau(x), Df(x)\rangle_{HS} d\mu(x), \tag{15}$$

where Df stands for the Jacobian matrix of f. Remark that in dimension 1, as long as μ has a density ρ and a finite first moment, it is straightforward to verify that,

$$\tau(x) = \frac{\int_x^\infty y\rho(y)dy}{\rho(y)}, \tag{16}$$

is a Stein kernel for μ. Actually, it is the only function which satisfies (15). In higher dimensions, the situation becomes more complicated. In many cases, there are different constructions of Stein kernels which yield qualitatively different solutions to (15). Not only that, but it also seems like a challenging task to produce a uniform criterion for determining the existence of a Stein kernel. For example, as will be shown in Chapter 2, it is quite common for high-dimensional singular measures to have Stein kernel.

Stein's lemma is the key insight for the use of Stein kernels. According to the Lemma, $\mu = \gamma_d$ if and only if the function $\tau(x) \equiv \mathrm{Id}$. Thus, in some sense, the deviation of τ from the identity measures the distance of μ from the standard Gaussian. Led by this idea we define the Stein discrepancy to the normal distribution as,

$$S(\mu) := \inf_\tau \sqrt{\int \|\tau(x) - \mathrm{Id}\|_{HS}^2 \, d\mu(x)},$$

where the infimum is taken over all Stein kernels of μ. By the above, $S(\mu) = 0$ if and only if $\mu = \gamma$. Thus, while not strictly being a metric, the Stein discrepancy still serves as some

notion of distance to the standard Gaussian. If X is a random vector in $\mathbb{R}^n$, by a slight abuse of notation we will also write $S(X)$ for $S(\mathrm{Law}(X))$.

Decay of Stein's discrepancy along the CLT: Stein kernels exhibit several nice properties which make their analysis tractable for normal approximations. Let $X \sim \mu$, an application of the chain-rule shows that if τ_X is a Stein kernel of X and A is a linear operator with compatible dimensions, a Stein kernel for AX is given by:

$$\tau_{AX}(x) = A\mathbb{E}\left[\tau(X)|AX = x\right]A^T \tag{17}$$

(see [85, Section 3] for some examples). Also, it is not hard to see that, if $\{X_i\}_{i=1}^n$ are *i.i.d.* as μ and $\mathbf{X} = (X_1, ..., X_n)$, then a Stein kernel for $\mathbf{X}$ may be realized as an $nd \times nd$ block diagonal matrix, with each main-diagonal block taken as τ_X. By combining the above constructions we get that, if $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, then

$$\tau_{S_n}(x) = \frac{1}{d}\sum_{i=1}^n \mathbb{E}\left[\tau_X(X_i)|S_n = x\right], \tag{18}$$

is a Stein kernel for S_n.

Now, assume that X is isotropic. By choosing the test function f to be linear in the definition of the Stein kernel we may see that

$$\mathbb{E}\left[\tau_X(X)\right] = \mathrm{Cov}(X) = \mathrm{Id}. \tag{19}$$

Thus, $(\tau_X(X_i) - \mathrm{Id})$ is a centered random variable, and the above observations show,

$$\begin{aligned} S^2(S_n) \leq \mathbb{E}\left[\|\tau_{S_n}(S_n) - \mathrm{Id}\|_{HS}^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\tau_X(X_i) - \mathrm{Id}|S_n\right]\right\|_{HS}^2\right] \\ &\leq \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \tau_X(X_i) - \mathrm{Id}\right\|_{HS}^2\right] = \frac{1}{n}\mathbb{E}\left[\|\tau_X(X) - \mathrm{Id}\|_{HS}^2\right], \end{aligned}$$

where in the second inequality we have applied Jensen's inequality and where we have used the fact that $(\tau_X(X_i) - \mathrm{Id})$ are *i.i.d.* and centered for the last equality. By taking the infimum over all Stein kernels, we get

$$S^2(S_n) \leq \frac{S^2(X)}{n}. \tag{20}$$

Stein's discrepancy as a distance: We now discuss the relations between Stein's discrepancy to other, more classical notions of distance. This notion has recently gained prominence in the study of convergence rates along the high-dimensional central limit theorem (see [85, 112, 196]

for several examples as well as the book [194]). This popularity stems, among others, from the fact that the Stein's discrepancy controls several other known distances.

First, by the Kantorovich-Rubinstein duality for $\mathcal{W}_1$ ([240]) it is not hard to show that

$$\mathcal{W}_1(\mu, \gamma) \leq S(\mu). \tag{21}$$

For simplicity of exposition let us focus on the case $d = 1$, although all arguments may be carried out in higher dimensions. Define the Ornstein-Uhlenbeck operator on an appropriate subspace of $L^2(\gamma)$ by,

$$Lf(x) = f'(x) - xf(x),$$

and for a fixed f, consider the Poisson equation,

$$Lh_f(x) = f(x) - \mathbb{E}\left[f(G)\right], \tag{22}$$

where we think about h_f as the unknown function. If $X \sim \mu$ and τ is a Stein kernel for μ, the Kantorovich-Rubinstein duality implies,

$$\begin{aligned}
\mathcal{W}_1(\mu, \gamma) &= \sup_{f \text{ is 1-Lipschitz}} |\mathbb{E}[f(X)] - \mathbb{E}[f(G)]| \\
&= \sup_{f \text{ is 1-Lipschitz}} |\mathbb{E}[Lh_f(X)]| \\
&= \sup_{f \text{ is 1-Lipschitz}} \left|\mathbb{E}[h_f'(X)] - \mathbb{E}[Xh_f(X)]\right| \\
&= \sup_{f \text{ is 1-Lipschitz}} \left|\mathbb{E}[h_f'(X)] - \mathbb{E}[\tau(X)h_f'(X)]\right| \\
&\leq \sqrt{\mathbb{E}[(1-\tau(X))^2]} \sup_{f \text{ is 1-Lipschitz}} \|h_f'\|_\infty = S(\mu) \sup_{f \text{ is 1-Lipschitz}} \|h_f'\|_\infty,
\end{aligned}$$

where in the third equality we have applied (15) to the function h_f. To bound $\|h_f'\|_\infty$ observe that when f is a Lipschitz function it may be verified (see also [194, Proposition 3.5.1]) that the solution to (21) is given by,

$$h_f(x) = -\int_0^\infty \frac{e^{-t}}{\sqrt{1-e^{-2t}}} \mathbb{E}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}G)G\right] dt.$$

In particular, if f is 1-Lipschitz and $|f'(x)| \leq 1$, then

$$|h_f'(x)| = \left|\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} \mathbb{E}\left[f'(e^{-t}x + \sqrt{1-e^{-2t}}G)G\right] dt\right| \leq \sqrt{\mathbb{E}[G^2]} \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} dt = 1.$$

Thus (21) is established.

A more careful analysis and an integration along the Orenstein-Uhlenbeck semigroup produces the more remarkable fact that the same also holds for the quadratic Wasserstein distance ([161, Proposition 3.1]),

$$\mathcal{W}_2^2(\mu,\gamma) \leq S^2(\mu). \tag{23}$$

In fact, by slightly changing the definition of the discrepancy, we may bound the Wasserstein distance of any order. This is the content of Proposition 3.1 in [112] which shows that if τ is a Stein kernel of μ, then,

$$\mathcal{W}_m(\mu,\gamma) \leq C_m \sqrt[m]{\int \|\tau(x) - \mathrm{Id}\|_{HS}^m \, d\mu(x)},$$

where $C_m > 0$ depends only on m. In some cases, one may also compare relative entropy to Stein's discrepancy, which is the content of the so-called HSI inequality from [161]. According to the inequality,

$$\mathrm{Ent}(\mu||\gamma) \leq \frac{1}{2} S^2(\mu) \log\left(1 + \frac{\mathrm{I}(\mu||\gamma)}{S^2(\mu)}\right),$$

where $\mathrm{I}(\mu||\gamma)$ is the (relative) Fisher information of μ. Thus, showing a CLT holds with respect to Stein's discrepancy can imply an entropic CLT, provided that the Fisher information is finite. Unfortunately, verifying that a measure has finite Fisher information is a non-trivial task in high-dimensions (see Section 5 in [161] for further discussion). In Chapter 6 we revisit those estimates. Among others, we give an alternative proof to (23) and also consider extensions of the inequalities to non-Gaussian reference measures.

A quantitative central limit theorem: To demonstrate the usefulness of Stein kernels we now establish a quantitative central limit theorem in dimension 1 for uniformly log-concave measures. Thus, suppose that μ is a isotropic measure on $\mathbb{R}$ with density $e^{-\varphi(x)}$, which satisfies $\varphi(x)'' \geq \sigma$, for some $\sigma > 0$. Suppose further that μ is symmetric, so its density is maximized at the origin. In this case, for $x > 0$, the above condition implies $\varphi(x)' \geq \sigma x + \varphi'(0) = \sigma x$.

Let $x > 0$ and consider the Stein kernel given by (16),

$$\tau(x) = \frac{\int_x^\infty y e^{-\varphi(y)} dy}{e^{-\varphi(x)}} \leq \frac{\frac{1}{\sigma}\int_x^\infty \varphi'(y) e^{-\varphi(y)} dy}{e^{-\varphi(x)}} = \frac{1}{\sigma} \frac{1}{e^{-\varphi(x)}} \Big|_x^\infty e^{-\varphi(y)} = \frac{1}{\sigma}.$$

From symmetry we may conclude $\|\tau\|_\infty \leq \frac{1}{\sigma}$. So,

$$S^2(\mu) = \mathbb{E}_\mu\left[(\tau - 1)^2\right] \leq 2\mathbb{E}_\mu\left[\tau^2\right] + 2 \leq 2\left(\frac{1}{\sigma^2} + 1\right).$$

If $\{X_i\}_{i=1}^n$ are *i.i.d.* as μ and $S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$. Then, from (20) and (23),

$$\mathcal{W}_2^2(S_n, G) \leq S^2(S_n) \leq \frac{S^2(\mu)}{n} \leq \frac{2\left(\frac{1}{\sigma^2}+1\right)}{n}.$$

This shows a Berry-Esseen type bound with convergence rate $\frac{1}{\sqrt{n}}$. In Chapters 2 and 3 we extend this result to much larger classes of measures and higher (even infinite) dimensions.

Notations and Terminology

We now introduce some of the common notations and definitions which are common to most chapters. When relevant and necessary, specific chapters may include further definitions.

We usually work in $\mathbb{R}^d$ equipped with the Euclidean norm, which we denote by $\|\cdot\|_2$ or sometimes just $\|\cdot\|$. For a matrix A we denote its Hilbert-Schmidt norm by $\|A\|_{HS}$. This is the norm induced by the Hilbert-Schmidt inner product $\langle A, B\rangle_{HS} := \mathrm{Tr}\left(AB^T\right)$.

A measure μ on $\mathbb{R}^d$ is said to be log-concave if it is supported on some subspace of $\mathbb{R}^d$ and, relative to the Lebesgue measure of that subspace, it has a density ρ, twice differentiable almost everywhere, for which

$$-\nabla^2 \log(\rho(x)) \succeq 0 \quad \text{for all } x,$$

where ∇^2 denotes the Hessian matrix, in the Alexandrov sense. If in addition there exists an $\sigma > 0$ such that

$$-\nabla^2 \log(\rho(x)) \succeq \sigma \mathrm{I}_d \quad \text{for all } x,$$

we say that μ is σ-uniformly log-concave. The measure μ is called *isotropic* if it is centered and its covariance matrix is the identity, i.e.,

$$\int_{\mathbb{R}^d} x\mu(dx) = 0 \text{ and } \int_{\mathbb{R}^d} x \otimes x \mu(dx) = \mathrm{I}_d.$$

If ν is another measure on $\mathbb{R}^d$ and $m > 0$, the m-Wasserstein's distance between μ and ν is defined by

$$\mathcal{W}_m(\mu, \nu) = \sqrt[m]{\inf_{\pi} \int \|x - y\|_2^m \, d\pi(x, y)},$$

where the infimum is taken over all couplings π, of μ and ν. Another notion of distance is that of relative entropy, which is given by

$$\mathrm{Ent}(\mu||\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu.$$

It is known that relative entropy bounds the quadratic Wasserstein distance to the the Gaussian via Talagrand's transportation-entropy inequality ([229]) as well as controlling the total variation distance through Pinsker's inequality ([86]). We also define the (relative) Fisher information as,

$$\mathrm{I}(\mu||\nu) = \int \left\| \nabla \log \left(\frac{d\mu}{d\nu} \right) \right\|^2 d\mu.$$

γ_d will always stand for the standard normal law on $\mathbb{R}^d$ with density

$$\frac{d\gamma_d}{dx}(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\|x\|_2^2}{2}}.$$

We will sometime omit the subscript, when the dimension is obvious from the context.

Finally, as a convention, we use $C, C', c, c' ...$ to denote absolute positive constants whose value might change between expressions. In case we want to signify that the constant might depend on some parameter a, we will write C_a, C'_a.

PART I

HIGH-DIMENSIONAL CENTRAL LIMIT THEOREMS

"Very simple was my explanation, and plausible enough–as most wrong theories are!"
- The Time Traveler

1

High-Dimensional Central Limit Theorems via Martingale Embeddings

1.1 Introduction

Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. random vectors in $\mathbb{R}^d$. By the central limit theorem, it is well-known that, under mild conditions, the sum $\frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$ converges to a Gaussian. With d fixed, there is an extensive literature showing that the distance from Gaussian under various metrics decays as $\frac{1}{\sqrt{n}}$ as $n \to \infty$, and this is optimal.

However, in high-dimensional settings, it is often the case that the dimension d is not fixed but rather grows with n. It then becomes necessary to understand how the convergence rate depends on dimension, and the optimal dependence here is not well understood. We present a new technique for proving central limit theorems in $\mathbb{R}^d$ that is suitable for establishing quantitative estimates for the convergence rate in the high-dimensional setting. The technique, which is described in more detail in Section 1.1.1 below, is based on pathwise analysis: we first couple the random vector with a Brownian motion via a martingale embedding. This gives rise to a coupling between the sum and a Brownian motion for which we can establish bounds on the concentration of the quadratic variation. We use a multidimensional version of a Skorokhod embedding, inspired by a construction from [98], as a manifestation of the martingale embedding.

Using our method, we prove new bounds on quadratic *transportation* (also known as "Kantorovich" or "Wasserstein") distance in the CLT, and in the case of log-concave distributions, we also give bounds for *entropy* distance. Recall that $\mathcal{W}_2(A, B)$ denotes the quadratic transportation distance between two d-dimensional random vectors A and B. As a first demonstration of our method, we begin with an improvement to the best known convergence rate in the case of bounded random vectors.

Theorem 1.1. *Let X be a random d-dimensional vector. Suppose that $\mathbb{E}[X] = 0$ and $\|X\| \leq \beta$ almost surely for some $\beta > 0$. Let $\Sigma = \mathrm{Cov}(X)$, and let $G \sim \mathcal{N}(0, \Sigma)$ be a Gaussian with covariance Σ. If $\{X^{(i)}\}_{i=1}^n$ are i.i.d copies of X and $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$, then*

$$\mathcal{W}_2(S_n, G) \leq \frac{\beta \sqrt{d} \sqrt{32 + 2\log_2(n)}}{\sqrt{n}}.$$

Theorem 1.1 improves the result from [251] that gives a bound of order $\frac{\beta \sqrt{d} \log n}{\sqrt{n}}$ under the same conditions. It was noted in [251] that when X is supported on a lattice $\beta \mathbb{Z}^d$, then the quantity $\mathcal{W}_2(S_n, G)$ is of order $\frac{\beta \sqrt{d}}{\sqrt{n}}$. Thus, Theorem 1.1 is within a $\sqrt{\log n}$ factor of optimal.

When the distribution of X is isotropic and log-concave, we can improve the bounds guaranteed by Theorem 1.1. In this case, however, a more general bound has already been established in [85], see discussion below.

Theorem 1.2. *Let X be a random d-dimensional vector. Suppose that the distribution of X is log-concave and isotropic. Let $G \sim \mathcal{N}(0, \mathrm{I}_d)$ be a standard Gaussian. If $\{X^{(i)}\}_{i=1}^n$ are i.i.d copies of X and $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$, then there exists a universal constant $C > 0$ such that, if $d \geq 8$,*

$$\mathcal{W}_2(S_n, G) \leq \frac{C d^{1/2 + o_d(1)} \ln(d) \sqrt{\ln(n)}}{\sqrt{n}}.$$

Remark 1.3. We actually prove the slightly stronger bound

$$\mathcal{W}_2(S_n, G) \leq \frac{C \kappa_d \ln(d) \sqrt{d \ln(n)}}{\sqrt{n}},$$

where

$$\kappa_d := \sup_{\substack{\mu \text{ isotropic,} \\ \text{log-concave}}} \left\| \int_{\mathbb{R}^d} x_1 x \otimes x \mu(dx) \right\|_{HS}, \tag{1.1}$$

as defined in [97]. The recent advances towards the KLS conjecture, made by Chen in [72] imply $\kappa_d = O(d^{o_d(1)})$, leading to the bound in Theorem 1.2. If the *thin-shell conjecture* (see [13], as well [40]) is true, then the bound is improved to $\kappa_d = O(\sqrt{\ln(d)})$, which yields

$$\mathcal{W}_2(S_n, G) \leq \frac{C \sqrt{d \ln(d)^3 \ln(n)}}{\sqrt{n}}.$$

By considering, for example, a random vector uniformly distributed on the unit cube, one can see that the above bound is sharp up to the logarithmic factors.

Remark 1.4. To compare with the previous theorem, note that if $\mathrm{Cov}(X) = \mathrm{I}_d$, then $\mathbb{E}\,\|X\|^2 = d$. Thus, in applying Theorem 1.1 we must take $\beta \geq \sqrt{d}$, and the resulting bound is then of order at least $\frac{d\sqrt{\log n}}{\sqrt{n}}$.

Next, we describe our results regarding convergence rate in entropy. As a warm-up, we first use our method to recover the entropic CLT in any fixed dimension. In dimension one this was first established by Barron, [24]. The same methods may also be applied to prove a multidimensional analogue. See [43] for a more quantitative version of the theorem.

Theorem 1.5. *Suppose that* $\mathrm{Ent}\,(X||G) < \infty$. *Then one has*

$$\lim_{n\to\infty} \mathrm{Ent}(S_n||G) = 0.$$

The next result gives the first non-asymptotic convergence rate for the entropic CLT, again under the log-concavity assumption (other non-asymptotic results appear in previous works, notably [85], but require additional assumptions; see below).

Theorem 1.6. *Let* X *be a random* d*-dimensional vector. Suppose that the distribution of* X *is log-concave and isotropic. Let* $G \sim \mathcal{N}(0, \mathrm{I}_d)$ *be a standard Gaussian. If* $\{X^{(i)}\}_{i=1}^n$ *are i.i.d copies of* X *and* $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$ *then*

$$\mathrm{Ent}(S_n||G) \leq \frac{Cd^{10}(1 + \mathrm{Ent}(X||G))}{n},$$

for a universal constant $C > 0$.

Our method also yields a different (and typically stronger) bound if the distribution is strongly log-concave.

Theorem 1.7. *Let* X *be a* d*-dimensional random vector with* $\mathbb{E}[X] = 0$ *and* $\mathrm{Cov}(X) = \Sigma$. *Suppose further that* X *is* 1*-uniformly log concave (i.e. it has a probability density* $e^{-\varphi(x)}$ *satisfying* $\nabla^2\varphi \succeq \mathrm{I}_d$*) and that* $\Sigma \succeq \sigma \mathrm{I}_d$ *for some* $\sigma > 0$.

Let $G \sim \mathcal{N}(0, \Sigma)$ *be a Gaussian with the same covariance as* X *and let* $\gamma \sim \mathcal{N}(0, \mathrm{I}_d)$ *be a standard Gaussian. If* $\{X^{(i)}\}_{i=1}^n$ *are i.i.d copies of* X *and* $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$, *then*

$$\mathrm{Ent}(S_n||G) \leq \frac{2\,(d + 2\mathrm{Ent}\,(X||\gamma))}{\sigma^4 n}.$$

Remark 1.8. The theorem can be applied when X is isotropic and σ-uniformly log concave for some $\sigma > 0$. In this case, a change of variables shows that $\sqrt{\sigma}X$ is 1-uniformly log concave

and has $\sigma \mathrm{I}_d$ as a covariance matrix. Since relative entropy to a Gaussian is invariant under affine transformations, if $G \sim \mathcal{N}(0, \mathrm{I}_d)$ is a standard Gaussian, we get

$$\mathrm{Ent}\left(S_n||G\right) = \mathrm{Ent}\left(\sqrt{\sigma}S_n||\sqrt{\sigma}G\right) \leq \frac{2\left(d + 2\mathrm{Ent}\left(\sqrt{\sigma}X||G\right)\right)}{\sigma^4 n}.$$

1.1.1 An informal description of the method

Let B_t be a standard Brownian motion in $\mathbb{R}^d$ with an associated filtration $\mathcal{F}_t$. The following definition will be central to our method:

Definition 1.9. Let X_t be a martingale satisfying $dX_t = \Gamma_t dB_t$ for some adapted process Γ_t taking values in the positive semi-definite cone and let τ be a stopping time. We say that the triplet (X_t, Γ_t, τ) is a martingale embedding of the measure μ if $X_\tau \sim \mu$.

Note that if Γ_t is deterministic, then X_t has a Gaussian law for each t. At the heart of our proof is the following simple idea: Summing up n independent copies of a martingale embedding of μ, we end up with a martingale embedding of μ^{*n} whose associated covariance process has the form $\sqrt{\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2}$. By the law of large numbers, this process is well concentrated and thus the resulting martingale is close to a Brownian motion.

This suggests that it would be useful to couple the sum process $\sum_{i=1}^n X_t^{(i)}$ with the "averaged" process whose covariance is given by $\mathbb{E}\left[\sqrt{\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2}\right]$ (this process is a Brownian motion up to deterministic time change). Controlling the error in the coupling naturally leads to a bound on transportation distance. For relative entropy, we can reformulate the discrepancies in the coupling in terms of a predictable drift and deduce bounds by a judicious application of Girsanov's theorem.

In order to derive quantitative bounds, one needs to construct a martingale embedding in a way that makes the fluctuations of the process Γ_t tractable. The specific choices of Γ_t that we consider are based on a construction introduced in [98]. This construction is also related to the entropy minimizing process used by Föllmer ([119, 120], see also Lehec [165]) and to the stochastic localization which was used in [97]. Such techniques have recently gained prominence and have been used, among other things, to improve known bounds of the KLS conjecture [72,97,163], calculate large deviations of non-linear functions [99] and study tubular neighborhoods of complex varieties [152].

The basic idea underlying the construction of the martingale is a certain measure-valued Markov process driven by a Brownian motion. This process interpolates between a given measure and a delta measure via multiplication by infinitesimal linear functions. The Doob martingale associated to the delta measure (the conditional expectation of the measure, based on the past) will be a martingale embedding for the original measure. This construction is described in detail in Subsection 1.2.3 below.

1.1.2 Related work

Multidimensional central limit theorems have been studied extensively since at least the 1940's [33] (see also [37] and references therein). In particular, the dependence of the convergence rate on the dimension was studied by Nagaev [188], Senatov [217], Götze [128], Bentkus [32], and Chen and Fang [71], among others. These works focused on convergence in probabilities of convex sets. We mention that in dimension 1, the picture is much clearer and that tight estimates are known under various metrics ([35, 41, 42, 108, 209, 210]).

More recently, dependence on dimension in the high-dimensional CLT has also been studied for Wishart matrices (Bubeck and Ganguly [58], Eldan and Mikulincer [102]), maxima of sums of independent random vectors (Chernozhukov, Chetverikov, and Kato [73]), and transportation distance ([251]). As mentioned earlier, Theorem 1.1 is directly comparable to an earlier result from [251], improving on it by a factor of $\sqrt{\log n}$ (see also the earlier work [236]). We refer to [251] for a discussion of how convergence in transportation distance may be related to convergence in probabilities of convex sets.

As mentioned above, Theorem 1.2 is not new, and follows from a result of Courtade, Fathi and Pananjady [85, Theorem 4.1]. Their technique employs Stein's method (see also [47], for a different approach using Stein's method) in a novel way which is also applicable to entropic CLTs (see below). In a subsequent work [112], similar bounds are derived for convergence in the p'th-Wasserstein transportation metric.

Regarding entropic CLTs, it was shown by Barron [24] that convergence occurs as long as the distribution of the summand has finite relative entropy (with respect to the Gaussian). However, establishing explicit rates of convergence does not seem to be a straightforward task. Even in the restricted setting of log-concave distributions, not much is known. One of the only quantitative results is Proposition 4.3 in [85], which gives near optimal convergence, provided that the distribution has finite Fisher information. We do not know of any results prior to Theorem 1.6 which give entropy distance bounds of the form $\frac{\mathrm{poly}(d)}{n}$ to a sum of general log-concave vectors.

A one-dimensional result was established by Artstein, Ball, Barthe, and Naor [17] and independently by Barron and Johnson [145], who showed an optimal $O(1/n)$ convergence rate in relative entropy for distributions having a spectral gap (i.e. satisfying a Poincaré inequality). This was later improved by Bobkov, Chistyakov, and Götze [43, 44], who derive an Edgeworth-type expansion for the entropy distance which also applies to higher dimensions. However, although their estimates contain very precise information as $n \to \infty$, the given error term is only asymptotic in n and no explicit dependence on the measure or on the dimension is given (in fact, the dependence derived from the method seems to be exponential in the dimension d).

A related "entropy jump" bound was proved by Ball and Nguyen [21] for log-concave random vectors in arbitrary dimensions (see also [20]). Essentially, the bound states that for two i.i.d. random vectors X and Y, the relative entropy $\mathrm{Ent}\left(\frac{X+Y}{\sqrt{2}}\middle\|G\right)$ is strictly less than

$\mathrm{Ent}(X||G)$, where the amount is quantified by the spectral gap for the distribution of X. Repeated application gives a bound for entropy of sums of i.i.d. log-concave vectors in any dimension, but the bound is far from optimal. It is not apparent to us whether the method of [21] can be extended to provide quantitative estimates for convergence in the entropic CLT.

1.1.3 Notation

For a positive semi-definite symmetric matrix A we denote by $\sqrt{A}$ the unique positive semi-definite matrix B, for which the relation $B^2 = A$ holds. For symmetric matrices A and B we use $A \preceq B$ to signify that $B - A$ is a positive semi-definite matrix. By $A^\dagger$ we denote the pseudo inverse of A. Put succinctly, this means that in $A^\dagger$ every non-zero eigenvalue of A is inverted. For a random matrix A, we will write $\mathbb{E}\left[A\right]^\dagger$, for the pseudo inverse of its expectation.
If B_t is the standard Brownian motion in $\mathbb{R}^d$ then for any adapted process F_t we denote by $\int_0^t F_s dB_s$, the Itô stochastic integral. We refer by Itô's isometry to the fact

$$\mathbb{E}\left[\left\|\int_0^t F_s dB_s\right\|^2\right] = \int_0^t \mathbb{E}\left[\|F_s\|_{HS}^2\right] ds$$

when F_t is adapted to the natural filtration of B_t.

μ will always stand for a probability measure. To avoid confusion, when integrating with respect to μ, on $\mathbb{R}^d$, we will use the notation $\int \ldots \mu(dx)$. For a measure-valued stochastic process μ_t, the expression $d\mu_t$ refers to the stochastic derivative of the process.

1.2 Obtaining convergence rates from martingale embeddings

Suppose that we are given a measure μ and a corresponding martingale embedding (X_t, Γ_t, τ). The goal of this section is to express bounds for the corresponding CLT convergence rates (of the sum of independent copies of μ-distributed random vectors) in terms of the behavior of the process Γ_t and τ.

Throughout this section we fix a measure μ on $\mathbb{R}^d$ whose expectation is 0, a random vector $X \sim \mu$, and a corresponding Gaussian $G \sim \mathcal{N}(0, \Sigma)$, where $\mathrm{Cov}(X) = \Sigma$. Also, the sequence $\{X^{(i)}\}_{i=1}^\infty$ will denote independent copies of X, and we write $S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X^{(i)}$ for their normalized sum. Finally, we use B_t to denote a standard Brownian motion on $\mathbb{R}^d$ adapted to a filtration $\mathcal{F}_t$.

1.2.1 A bound for Wasserstein-2 distance

The following is our main bound for convergence in Wasserstein distance.

Theorem 1.10. *Let S_n and G be defined as above and let (X_t, Γ_t, τ) be a martingale embedding of μ. Set $\Gamma_t = 0$ for $t > \tau$, then*

$$\mathcal{W}_2^2(S_n, G) \leq \int_0^\infty \min\left(\frac{1}{n}\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^4\right]\mathbb{E}[\Gamma_t^2]^\dagger\right), 4\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right]\right)\right) dt.$$

To illustrate how such a result might be used, let us assume, for simplicity, that $\Gamma_t \prec k\mathrm{I}_d$ almost-surely for some $k > 0$ and that τ has a sub-exponential tail, i.e., there exist positive constants $C, c > 0$ such that for any $t > 0$,

$$\mathbb{P}(\tau > t) \leq Ce^{-ct}. \tag{1.2}$$

Under these assumptions,

$$\begin{aligned}\mathcal{W}_2^2(S_n, G) &\leq \int_0^\infty \min\left(\frac{1}{n}\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^4\right]\mathbb{E}[\Gamma_t^2]^\dagger\right), 4k^2 d\mathbb{P}(\tau > t)\right) dt \\ &\leq dk^2 \int_0^{\frac{\log(n)}{c}} \frac{1}{n} dt + 4Cdk^2 \int_{\frac{\log(n)}{c}}^\infty e^{-ct} dt = \frac{d\log(n)k^2}{cn} + \frac{4Cdk^2}{n}.\end{aligned}$$

Towards the proof, we will need the following technical lemma.

Lemma 1.11. *Let A, B be positive semi-definite matrices with $\ker(A) \subset \ker(B)$. Then,*

$$\mathrm{Tr}\left(\left(\sqrt{A} - \sqrt{B}\right)^2\right) \leq \mathrm{Tr}\left((A-B)^2 A^\dagger\right).$$

Proof. Since A and B are positive semi-definite, $\ker\left(\sqrt{A} + \sqrt{B}\right) \subset \ker\left(\sqrt{A} - \sqrt{B}\right)$.Thus, we have that

$$\begin{aligned}\sqrt{A} - \sqrt{B} &= \left(\sqrt{A} - \sqrt{B}\right)\left(\sqrt{A} + \sqrt{B}\right)\left(\sqrt{A} + \sqrt{B}\right)^\dagger \\ &= \left(A - B + \left[\sqrt{A}, \sqrt{B}\right]\right)\left(\sqrt{A} + \sqrt{B}\right)^\dagger.\end{aligned} \tag{1.3}$$

So,

$$\mathrm{Tr}\left(\left(\sqrt{A} - \sqrt{B}\right)^2\right) = \mathrm{Tr}\left(\left(\left(A - B + \left[\sqrt{A}, \sqrt{B}\right]\right)\left(\sqrt{A} + \sqrt{B}\right)^\dagger\right)^2\right).$$

Note that for any symmetric matrices X and Y, by the Cauchy-Schwartz inequality,

$$\mathrm{Tr}\left((XY)^2\right) \leq \mathrm{Tr}\left(XYXY\right) \leq \sqrt{\mathrm{Tr}\left(XYYX\right) \cdot \mathrm{Tr}\left(YXXY\right)} = \mathrm{Tr}\left(X^2Y^2\right).$$

Applying this to the above equation shows

$$\operatorname{Tr}\left(\left(\sqrt{A}-\sqrt{B}\right)^2\right) \leq \operatorname{Tr}\left(\left(A-B+\left[\sqrt{A},\sqrt{B}\right]\right)^2\left(\left(\sqrt{A}+\sqrt{B}\right)^\dagger\right)^2\right).$$

Note that the commutator $\left[\sqrt{A},\sqrt{B}\right]$ is an anti-symmetric matrix, so that $(A-B)\left[\sqrt{A},\sqrt{B}\right]+\left[\sqrt{A},\sqrt{B}\right](A-B)$ is anti-symmetric as well. Thus, for any symmetric matrix C, we have that

$$\operatorname{Tr}\left(\left((A-B)\left[\sqrt{A},\sqrt{B}\right]+\left[\sqrt{A},\sqrt{B}\right](A-B)\right)C\right)=0.$$

Also, since all eigenvalues of anti-symmetric matrices are purely imaginary, the square of such matrices must be negative definite. And again, for any symmetric positive definite matrix C, it holds that $C^{1/2}\left[\sqrt{A},\sqrt{B}\right]^2 C^{1/2}$ is negative definite and $\operatorname{Tr}\left(\left[\sqrt{A},\sqrt{B}\right]^2 C\right)\leq 0$. Using these observations we obtain

$$\operatorname{Tr}\left(\left(A-B+\left[\sqrt{A},\sqrt{B}\right]\right)^2\left(\left(\sqrt{A}+\sqrt{B}\right)^\dagger\right)^2\right) \leq \operatorname{Tr}\left((A-B)^2\left(\left(\sqrt{A}+\sqrt{B}\right)^\dagger\right)^2\right).$$

Finally, if C, X, Y are positive definite matrices with $X \preceq Y$ then $C^{1/2}(Y-X)C^{1/2}$ is positive definite which shows $\operatorname{Tr}(CX)\leq \operatorname{Tr}(CY)$. The assumption $\ker(A)\subset\ker(B)$ implies $\left(\left(\sqrt{A}+\sqrt{B}\right)^\dagger\right)^2 \preceq A^\dagger$, which concludes the claim by

$$\operatorname{Tr}\left((A-B)^2\left(\left(\sqrt{A}+\sqrt{B}\right)^\dagger\right)^2\right) \leq \operatorname{Tr}\left((A-B)^2 A^\dagger\right)$$

□

Proof of Theorem 1.10. Recall that (X_t,Γ_t,τ) is a martingale embedding of μ. Let $\left(X_t^{(i)},\Gamma_t^{(i)},\tau^{(i)}\right)$ be independent copies of the embedding. We can always set $\Gamma_t^{(i)}=0$ whenever $t>\tau^{(i)}$, so that $\int_0^\infty \Gamma_t^{(i)} dB_t^{(i)} \sim \mu$. Define $\tilde{\Gamma}_t = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2}$. Our first goal is to show

$$\mathcal{W}_2^2(G,S_n) \leq \int_0^\infty \mathbb{E}\left[\operatorname{Tr}\left(\left(\tilde{\Gamma}_t-\sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right)\right]dt. \tag{1.4}$$

The theorem will then follow by deriving suitable bounds for $\mathbb{E}\left[\operatorname{Tr}\left(\left(\tilde{\Gamma}_t-\sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right)\right]$ using Lemma 1.11. Consider the sum $\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\infty \Gamma_t^{(i)} dB_t^{(i)}$, which has the same law as S_n. It may

be rewritten as

$$S_n = \int_0^\infty \tilde{\Gamma}_t d\tilde{B}_t,$$

where $d\tilde{B}_t := \frac{1}{\sqrt{n}} \tilde{\Gamma}_t^\dagger \sum_i \Gamma_t^{(i)} dB_t^{(i)}$ is a martingale whose quadratic variation matrix has derivative satisfying

$$\frac{d}{dt}[\tilde{B}]_t = \frac{1}{n} \sum_i \tilde{\Gamma}_t^\dagger \left(\Gamma_t^{(i)}\right)^2 \tilde{\Gamma}_t^\dagger \preceq \mathrm{I}_d. \tag{1.5}$$

(in fact, as long as $\mathbb{R}^d$ is spanned by the images of $\Gamma_t^{(i)}$, this process is a Brownian motion). We may now decompose S_n as

$$S_n = \int_0^\infty \sqrt{\mathbb{E}\left[\tilde{\Gamma}_t^2\right]} d\tilde{B}_t + \int_0^\infty \left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\tilde{\Gamma}_t^2\right]}\right) d\tilde{B}_t. \tag{1.6}$$

Observe that $G := \int_0^\infty \sqrt{\mathbb{E}[\tilde{\Gamma}_t^2]} d\tilde{B}_t$ has a Gaussian law and that $\mathbb{E}[\tilde{\Gamma}_t^2] = \mathbb{E}[\Gamma_t^2]$. By applying Itô's isometry, we may see that G has the "correct" covariance in the sense that

$$\mathrm{Cov}(G) = \mathbb{E}\left[\left(\int_0^\infty \sqrt{\mathbb{E}[\tilde{\Gamma}_t^2]} d\tilde{B}_t\right)^{\otimes 2}\right] = \mathbb{E}\left[\int_0^\infty \Gamma_t^2 dt\right] = \mathbb{E}\left[\left(\int_0^\infty \Gamma_t dB_t\right)^{\otimes 2}\right] = \mathrm{Cov}(X).$$

The decomposition (1.6) induces a natural coupling between G and S_n, which shows, by another application of Itô's isometry, that

$$\begin{aligned}\mathcal{W}_2^2(G, S_n) &\leq \mathbb{E}\left[\left\|\int_0^\infty \left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}[\Gamma_t^2]}\right) d\tilde{B}_t\right\|^2\right] \overset{(1.5)}{\leq} \mathrm{Tr}\left(\mathbb{E}\left[\int_0^\infty \left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}[\Gamma_t^2]}\right)^2 dt\right]\right) \\ &= \int_0^\infty \mathbb{E}\left[\mathrm{Tr}\left(\left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right)\right] dt,\end{aligned}$$

where the last equality is due to Fubini's theorem. Thus, (1.4) is established. Since $\left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2 \preceq 2\left(\tilde{\Gamma}_t^2 + \mathbb{E}[\Gamma_t^2]\right)$, we have

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right]\right) \leq 4\mathrm{Tr}\left(\mathbb{E}[\Gamma_t^2]\right). \tag{1.7}$$

To finish the proof, write $U_t := \frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2$, so that $\tilde{\Gamma}_t = \sqrt{U_t}$. Since Γ_t is positive semi-

definite, it is clear that $\ker\left(\mathbb{E}\left[\Gamma_t^2\right]\right) \subset \ker(U_t)$. By Lemma 1.11,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}\left(\left(\sqrt{U_t}-\sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right)\right] \leq & \mathrm{Tr}\left(\mathbb{E}\left[\left(U_t-\mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right) \\ = & \frac{1}{n^2}\mathrm{Tr}\left(\sum_{i=1}^n \mathbb{E}\left[\left(\left(\Gamma_t^{(i)}\right)^2-\mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right) \\ = & \frac{1}{n}\mathrm{Tr}\left(\left(\mathbb{E}\left[\Gamma_t^4\right]-\mathbb{E}\left[\Gamma_t^2\right]^2\right)\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right) \\ \leq & \frac{1}{n}\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^4\right]\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right),\end{aligned}$$

where we have used the fact $\mathbb{E}\left[\left(\Gamma_t^{(i)}\right)^2\right] = \mathbb{E}\left[\Gamma_t^2\right]$ in the second equality. Combining the last inequality with (1.7) and (1.4) produces the required result.

□

1.2.2 A bound for the relative entropy

As alluded to in the introduction, in order to establish bounds on the relative entropy we will use the existence of a martingale embedding to construct an Itô process whose martingale part has a deterministic quadratic variation. This will allow us to relate the relative entropy to a Gaussian with the norm of the drift term through the use of Girsanov's theorem. As a technicality, we require the stopping time associated to the martingale embedding to be constant. Our main bound for the relative entropy reads,

Theorem 1.12. *Let $(X_t, \Gamma_t, 1)$ be a martingale embedding of μ. Assume that for every $0 \leq t \leq 1$, $\mathbb{E}\left[\Gamma_t\right] \succeq \sigma_t \mathrm{I}_d \gneq 0$ and that Γ_t is invertible a.s. for $t < 1$. Then we have the following inequalities:*

$$\mathrm{Ent}(S_n||G) \leq \frac{1}{n}\int_0^1 \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2-\mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)^2\sigma_t^2}\left(\int_t^1 \sigma_s^{-2}ds\right)dt,$$

and

$$\mathrm{Ent}(S_n||G) \leq \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right]-\mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{(1-t)^2}\left(\int_t^1 \sigma_s^{-2}ds\right)dt,$$

where

$$\tilde{\Gamma}_t = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2}$$

and $\Gamma_t^{(i)}$ are independent copies of Γ_t.

The theorem relies on the following bound, whose proof is postponed to the end of the subsection.

Lemma 1.13. *Let Γ_t be an $\mathcal{F}_t$-adapted matrix-valued processes and let $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be almost surely invertible and locally Lipschitz. Denote $F_t(x) := F(t,x)$ and let X_t, M_t be defined by*

$$X_t = \int_0^t \Gamma_s dB_s \text{ and } M_t = \int_0^t F_s(M_s) dB_s.$$

Define the process Y_t by

$$Y_t = \int_0^t F_s(Y_s)dB_s + \int_0^t \int_0^s \frac{\Gamma_r - F_r(Y_r)}{1-r} dB_r ds.$$

Then,

$$\mathrm{Ent}\left(X_1 || M_1\right) \leq \mathbb{E}\left[\int_0^1 \int_s^1 \left\| F_t^{-1}(Y_t) \frac{\Gamma_s - F_s(Y_s)}{1-s} \right\|_{HS}^2 dtds\right].$$

Note that if the process F_t is deterministic, i.e. it is a constant function, then M_1 has a Gaussian law, so that the lemma can be used to bound the relative entropy of X_1 with respect to a Gaussian.

Proof of Theorem 1.12. Let $(X_t^{(i)}, \Gamma_t^{(i)}, 1)$ be independent copies of the martingale embedding. Consider the sum process $\tilde{X}_t = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_t^{(i)}$, which satisfies $\tilde{X}_t = \int_0^t \tilde{\Gamma}_s d\tilde{B}_s$ where we define, as in the proof of Theorem 1.10,

$$\tilde{\Gamma}_t := \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2} \text{ and } d\tilde{B}_t = \frac{1}{\sqrt{n}} \tilde{\Gamma}_t^{-1} \sum \Gamma_t^{(i)} dB_t^{(i)}.$$

By assumption $\tilde{\Gamma}_t$ is invertible, which makes $\tilde{B}_t$ a Brownian motion. In this case, $(\tilde{X}_t, \tilde{\Gamma}_t, 1)$ is a martingale embedding for the law of S_n. For the first bound, consider the process

$$M_t = \int_0^t \sqrt{\mathbb{E}\left[\Gamma_s^2\right]} d\tilde{B}_s.$$

By Itô's isometry one has $M_1 \sim \mathcal{N}(0, \Sigma)$. Also, by Jensen's inequality

$$\sqrt{\mathbb{E}\left[\Gamma_t^2\right]} \succeq \mathbb{E}\left[\Gamma_t\right] \succeq \sigma_t \mathrm{I}_d.$$

Using this observation and substituting $\sqrt{\mathbb{E}\left[\Gamma_t^2\right]}$ for a constant function F_t in Lemma 1.13 yields,

$$\mathrm{Ent}\left(S_n \| G\right) \leq \int_0^1 \mathbb{E}\left[\left\| \frac{\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}}{1-t} \right\|_{HS}^2\right] \left(\int_t^1 \sigma_s^{-2} ds\right) dt. \tag{1.8}$$

With the use of Lemma 1.11 we obtain

$$\begin{aligned}\mathbb{E}\left\|\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right\|_{HS}^2 &= \mathbb{E}\left[\mathrm{Tr}\left(\left(\tilde{\Gamma}_t - \sqrt{\mathbb{E}\left[\Gamma_t^2\right]}\right)^2\right)\right] \\ &\leq \mathbb{E}\left[\mathrm{Tr}\left(\left(\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2 \mathbb{E}\left[\Gamma_t^2\right]^{-1}\right)\right] \\ &\leq \frac{1}{n\sigma_t^2}\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right].\end{aligned}$$

Plugging the above into (1.8) shows the first bound. To see the second bound, we define a process M_t', which is similar to M_t, and is given by the equation

$$M_t' := \int_0^t \mathbb{E}\left[\tilde{\Gamma}_s\right] d\tilde{B}_s.$$

Let G_n denote a Gaussian which is distributed as M_1'. For any s, we now have the following Cauchy-Schwartz type inequality

$$n\left(\sum_{i=1}^n \left(\Gamma_s^{(i)}\right)^2\right) \succeq \left(\sum_{i=1}^n \Gamma_s^{(i)}\right)^2.$$

Since the square root is monotone with respect to the order on positive definite matrices, this implies

$$\mathbb{E}\left[\tilde{\Gamma}_s\right] \succeq \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \Gamma_s^{(i)}\right] \succeq \sigma_s \mathrm{I}_d.$$

Thus,

$$\begin{aligned}\mathrm{Ent}(S_n||G_n) &\leq \mathbb{E}\left[\int_0^1\int_t^1 \left\|\mathbb{E}\left[\tilde{\Gamma}_s\right]^{-1}\frac{\tilde{\Gamma}_t - \mathbb{E}\left[\tilde{\Gamma}_t\right]}{1-t}\right\|_{HS}^2 dsdt\right] \\ &\leq \int_0^1 \mathbb{E}\left[\left\|\frac{\tilde{\Gamma}_t - \mathbb{E}\left[\tilde{\Gamma}_t\right]}{1-t}\right\|_{HS}^2\right]\left(\int_t^1 \sigma_s^{-2} ds\right) dt \\ &= \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{(1-t)^2}\left(\int_t^1 \sigma_s^{-2} ds\right) dt.\end{aligned}$$

Since $\mathrm{Cov}(G) = \mathrm{Cov}(S_n)$, it is now easy to verify that $\mathrm{Ent}\left(S_n||G\right) \leq \mathrm{Ent}\left(S_n||G_n\right)$, which concludes the proof. □

A key component in the proof of the theorem lies in using the norm of an adapted process

in order to bound the relative entropy. The following lemma embodies this idea. Its proof is based on a straightforward application of Girsanov's theorem. We provide a sketch and refer the reader to [165], where a slightly less general version of this lemma is given, for a more detailed proof.

Lemma 1.14. *Let $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be almost surely invertible and locally Lipschitz. Denote $F_t(x) := F(t,x)$ and let $M_t = \int_0^t F_s(M_s)dB_s$. For u_t, an adapted process, set $Y_t := \int_0^t F_s(Y_s)dB_s + \int_0^t u_s ds$. Then*

$$\mathrm{Ent}\left(Y_1||M_1\right) \leq \frac{1}{2}\int_0^1 \mathbb{E}\left[\left\|F_t^{-1}(Y_t)u_t\right\|^2\right] dt.$$

Proof. Since M_t is an Itô diffusion, by Girsanov's theorem ([199, Theorem 8.6.5]), the density of $\{Y_t\}_{t\in[0,1]}$ with respect to that of $\{M_t\}_{t\in[0,1]}$ on the space of paths is given by

$$\mathcal{E} := \exp\left(-\int_0^1 F_t(Y_t)^{-1}u_t dB_t - \frac{1}{2}\int_0^1 \left\|F_t(Y_t)^{-1}u_t\right\|^2 dt\right).$$

If f is the density of Y_1 with respect to M_1, this implies

$$1 = \mathbb{E}\left[f(Y_1)\mathcal{E}\right].$$

By Jensen's inequality

$$0 = \ln\left(\mathbb{E}\left[f(Y_1)\mathcal{E}\right]\right) \geq \mathbb{E}\left[\ln\left(f(Y_1)\mathcal{E}\right)\right] = \mathbb{E}\left[\ln(f(Y_1))\right] + \mathbb{E}\left[\ln(\mathcal{E})\right].$$

But,

$$\mathbb{E}\left[\ln(\mathcal{E})\right] = -\frac{1}{2}\int_0^1 \mathbb{E}\left[\left\|F_t^{-1}(Y_t)u_t\right\|^2\right] dt,$$

and

$$\mathbb{E}\left[\ln(f(Y_1))\right] = \mathrm{Ent}(Y_1||M_1),$$

which concludes the proof. □

The proof of Lemma 1.13 now amounts to invoking the above bound with a suitable construction of the drift process u_t.

Proof of Lemma 1.13. By defintion of the process Y_t, we have the following equality

$$Y_1 = \int_0^1 F_t(Y_t)dB_t + \int_0^1 \int_0^t \frac{\Gamma_s - F_s(Y_s)}{1-s} dB_s dt = \int_0^1 F_t(Y_t)dB_t + \int_0^1 (\Gamma_t - F_t(Y_t))\, dB_t = X_1, \tag{1.9}$$

where we have used Fubini's theorem in the penultimate equality. Now, consider the adapted process

$$u_t = \int_0^t \frac{\Gamma_s - F_s(Y_s)}{1-s} dB_s,$$

so that,

$$dY_t = F_t(Y_t)dB_t + u_t dt.$$

Applying Lemma 1.14 and using Itô's isometry, we get

$$\begin{aligned}
\mathrm{Ent}(X_1||M_1) &\leq \int_0^1 \mathbb{E}\left[\left\|F_t^{-1}(Y_t)u_t\right\|^2\right] dt = \int_0^1 \mathbb{E}\left[\left\|\int_0^t F_t^{-1}(Y_t)\frac{\Gamma_s - F_s(Y_s)}{1-s} dB_s\right\|^2\right] dt \\
&= \mathbb{E}\left[\int_0^1 \int_0^t \left\|F_t^{-1}(Y_t)\frac{\Gamma_s - F_s(Y_s)}{1-s}\right\|_{HS}^2 ds dt\right] \\
&= \mathbb{E}\left[\int_0^1 \int_s^1 \left\|F_t(Y_t)^{-1}\frac{\Gamma_s - F_s(Y_s)}{1-s}\right\|_{HS}^2 dt ds\right],
\end{aligned}$$

where last equality follows from another use of Fubini's theorem. □

1.2.3 A stochastic construction

In this section we introduce the main construction used in our proofs, a martingale process which meets the assumptions of Theorems 1.10 and 1.12. The construction in the next proposition is based on the Skorokhod embedding described in [98]. Most of the calculations in this subsection are very similar to what is done in [98], except that we allow some inhomogeneity in the quadratic variation according to the function C_t below. In particular, C_t will be a symmetric matrix almost surely, and we will denote the space of $d \times d$ *symmetric* matrices by Sym_d.

Proposition 1.15. *Let μ be a probability measure on $\mathbb{R}^d$ with smooth density and bounded support. For a probability measure-valued process μ_t, let*

$$a_t = \int_{\mathbb{R}^d} x\mu_t(dx), \quad A_t = \int_{\mathbb{R}^d} (x - a_t)^{\otimes 2}\mu_t(dx)$$

denote its mean and covariance.

Let $C : \mathbb{R} \times \mathrm{Sym}_d \to \mathrm{Sym}_d$ be a continuous function. Then, we can construct μ_t so that the following properties hold:

1. *$\mu_0 = \mu$,*
2. *a_t is a stochastic process satisfying $da_t = A_t C(t, A_t^\dagger) dB_t$, where B_t is a standard Brownian motion on $\mathbb{R}^d$, and*
3. *For any continuous and bounded $\varphi : \mathbb{R}^d \to \mathbb{R}$, $\int_{\mathbb{R}^d} \varphi(x)\mu_t(dx)$ is a martingale.*

Remark 1.16. We will be mainly interested in situations where μ_t converges almost surely to a point mass in finite time. In this case, we obtain a martingale embedding $(a_t, A_t C(t, A_t^\dagger), \tau)$ for μ, where τ is the first time that μ_t becomes a point mass.

In the sequel, we abbreviate $C_t := C(t, A_t^\dagger)$. We first give an informal description of how $\mu_{t+\varepsilon}$ is constructed from μ_t for $\varepsilon \to 0$. Consider a stochastic process $\{X_s\}_{0 \le s \le 1}$ in which we first sample $X_1 \sim \mu_t$ and then set

$$X_s = (1-s)a_t + sX_1 + C_t^{-1} B_s,$$

where B_s is a standard Brownian bridge. We can write $X_\varepsilon = a_t + \sqrt{\varepsilon} C_t^{-1} Z$, where Z is close to a standard Gaussian. We then take $\mu_{t+\varepsilon}$ to be the conditional distribution of X_1 given X_ε. This immediately ensures that property 3 holds and that a_t is a martingale.

It remains to see why property 2 holds. A direct calculation with conditioned Brownian bridges gives a first-order approximation

$$\begin{aligned}
\mu_{t+\varepsilon}(dx) &\propto e^{-\frac{1}{2}(\sqrt{\varepsilon} C_t^{-1} Z - \varepsilon(x-a_t))^T C_t^2 (\sqrt{\varepsilon} C_t^{-1} Z - \varepsilon(x-a_t))} \mu_t(dx) \\
&\propto e^{\sqrt{\varepsilon}\langle C_t Z, x - a_t\rangle + O(\varepsilon)} \mu_t(dx) \\
&\approx (1 + \sqrt{\varepsilon}\langle C_t Z, x - a_t\rangle) \mu_t(dx).
\end{aligned}$$

Then, to highest order, we have

$$a_{t+\varepsilon} - a_t \approx \sqrt{\varepsilon} \int_{\mathbb{R}^d} \langle C_t Z, x - a_t\rangle (x - a_t)\, \mu_t(dx) = \sqrt{\varepsilon} A_t C_t Z,$$

which translates into property 2 as $\varepsilon \to 0$.

Observe that the procedure outlined above yields measures μ_t that have densities which are proportional to the original density μ times a Gaussian density. (This applies at least when A_t is non-degenerate; something similar also holds when A_t is degenerate, as we will see shortly.) Let us now perform the construction formally. We will proceed by iterating the following preliminary construction, which handles the case when A_t remains non-degenerate.

Lemma 1.17. *Let μ be a measure on $\mathbb{R}^d$ with smooth density and bounded support, and let $C : \mathbb{R} \times \mathrm{Sym}_d \to \mathrm{Sym}_d$ be a continuous map. Then, there is a measure-valued process μ_t and*

a stopping time T such that μ_t satisfies the properties in Proposition 1.15 for $t < T$ and the affine hull of the support of μ_T has dimension strictly less than d. Moreover, if μ_T is considered as a measure on this affine hull, it has a smooth density.

Proof. We will construct a $(\mathbb{R}^d \times \mathrm{Sym}_d)$-valued stochastic process $(c_t, \tilde{\Sigma}_t)$ started at $(c_0, \tilde{\Sigma}_0) = (0, \mathrm{I}_d)$. Let us write

$$Q_t(x) = \frac{1}{2}\left\langle x - c_t, \tilde{\Sigma}_t^{-1}(x - c_t)\right\rangle,$$

and let $\tilde{\mu}$ be the probability measure satisfying $\frac{d\tilde{\mu}}{d\mu}(x) \propto e^{\frac{1}{2}\|x\|^2}$. We will then take μ_t to be $\mu_t(dx) = F_t(x)\tilde{\mu}(dx)$, where

$$F_t(x) = \frac{1}{Z_t}e^{-Q_t(x)}, \qquad Z_t = \int_{\mathbb{R}^d} e^{-Q_t(x)}\tilde{\mu}(dx).$$

Note that since $\tilde{\Sigma}_0 = \mathrm{I}_d$, we have $\mu_0 = \mu$.[1]

In order to specify the process, it remains to construct $(c_t, \tilde{\Sigma}_t)$. We take it to be the solution to the SDE

$$dc_t = \tilde{\Sigma}_t C_t dB_t + \tilde{\Sigma}_t C_t^2(a_t - c_t)dt, \qquad d\tilde{\Sigma}_t = -\tilde{\Sigma}_t C_t^2 \tilde{\Sigma}_t dt.$$

Note that the coefficients of this SDE are continuous functions of $(c_t, \tilde{\Sigma}_t)$ so long as $\tilde{\Sigma}_t \succ 0$. By standard existence and uniqueness results, this SDE has a unique solution up to a stopping time T (possibly $T = \infty$), at which point A_t (and hence $\tilde{\Sigma}_t$) becomes degenerate. Observe that, for every t, $\tilde{\Sigma}_t \preceq \mathrm{I}_d$ and so, the matrix process is continuous on the interval $[0, T]$.

By a limiting procedure, it is easy to see that μ_T has a smooth density when considered as a measure on the affine hull of its support. (Indeed, its density is proportional to the conditional density of $\tilde{\mu}$ times a Gaussian density.) It remains to verify that μ_t is a martingale and $da_t = A_t C_t dB_t$.

By direct calculation, we have

$$\begin{aligned}
d(\tilde{\Sigma}_t^{-1}) &= C_t^2 dt \\
d(\tilde{\Sigma}_t^{-1} c_t) &= C_t^2 c_t dt + C_t^2(a_t - c_t)dt + C_t dB_t \\
&= C_t^2 a_t dt + C_t dB_t \\
dQ_t(x) &= \left\langle x, \left(\frac{1}{2}C_t^2 x - C_t^2 a_t\right)dt - C_t dB_t \right\rangle \\
d(e^{-Q_t(x)}) &= -e^{-Q_t(x)}dQ_t(x) + \frac{1}{2}e^{-Q_t(x)}d[Q_t(x)] \\
&= e^{-Q_t(x)}\left\langle x, C_t dB_t + C_t^2 a_t dt\right\rangle
\end{aligned}$$

[1] Conceptually, one can replace all instances of $\tilde{\mu}$ with μ if we think of the initial value $\tilde{\Sigma}_0$ as being an "infinite" multiple of identity. However, to avoid issues with infinities, we have expressed things in terms of $\tilde{\mu}$ instead.

Integrating against $\tilde{\mu}(dx)$, we obtain

$$\begin{aligned} dZ_t &= Z_t \left\langle a_t, C_t dB_t + C_t^2 a_t dt \right\rangle \\ dZ_t^{-1} &= -\frac{1}{Z_t^2} dZ_t + \frac{1}{Z_t^3} d[Z_t] = \frac{1}{Z_t} \langle a_t, -C_t dB_t \rangle \\ dF_t(x) &= e^{-Q_t(x)} dZ_t^{-1} + Z_t^{-1} d(e^{-Q_t(x)}) + d[Z_t^{-1}, e^{-Q_t(x)}] \\ &= F_t(x) \cdot \langle x - a_t, C_t dB_t \rangle. \end{aligned}$$

Thus, $F_t(x)$ is a martingale for each fixed x, and furthermore,

$$da_t = d \int_{\mathbb{R}^d} x \mu_t(dx) = \int_{\mathbb{R}^d} x d\mu_t(dx) = \int_{\mathbb{R}^d} x(x - a_t) C_t \mu_t(dx) dB_t = A_t C_t dB_t.$$

□

Proof of Proposition 1.15. We use the process given by Lemma 1.17, which yields a stopping time T_1 and a measure μ_{T_1} with a strictly lower-dimensional support. If μ_T is a point mass, then we set $\mu_t = \mu_T$ for all $t \geq T$.

Otherwise, by the smoothness properties of μ_{T_1} guaranteed by Lemma 1.17, we can recursively apply Lemma 1.17 again on μ_{T_1} conditioned on the affine hull of its support. Repeating this procedure at most d times gives us the desired process. □

1.2.4 Properties of the construction

We record here various formulas pertaining to the quantities a_t, A_t, and μ_t constructed in Proposition 1.15.

Proposition 1.18. *Let μ, C_t, and μ_t be as in Proposition 1.15. Then, there is a Sym_d-valued process $\{\Sigma_t\}_{t>0}$ satisfying the following:*

- *For all t, there is an affine subspace $L = L_t \subset \mathbb{R}^d$ and a Gaussian measure γ_t on $\mathbb{R}^d$, supported on L, with covariance Σ_t such that μ_t is absolutely continuous with respect to γ_t, and*
$$\frac{d\mu_t}{d\gamma_t}(x) \propto \mu(x), \ \forall x \in L.$$
- *Σ_t is continuous and for almost every t obeys the differential equation*
$$\frac{d}{dt}\Sigma_t = -\Sigma_t C_t^2 \Sigma_t.$$
- $\lim_{t \to 0^+} \Sigma_t^{-1} = 0$.

Proof. For $1 \leq k \leq d$, let T_k denote the first time the measure μ_t is supported in a $(d-k)$-dimensional affine subspace, and denote by L_t the affine hall of the support of μ_t. We will define Σ_t inductively for each interval $[T_{k-1}, T_k]$. Recall from the proof of Proposition 1.15 that μ_t is constructed by iteratively applying Lemma 1.17 to affine subspaces of decreasing dimension $d, d-1, d-2, \ldots, 1$. Let $\tilde{\Sigma}_{k,t}$ denote the quantity $\tilde{\Sigma}_t$, from the k-th application of Lemma 1.17, so that $\tilde{\Sigma}_{k,t}$ is a linear operator on the subspace L_{T_k}.

For the base case $0 < t \leq T_1$, take $\Sigma_t = (\tilde{\Sigma}_{0,t}^{-1} - \mathrm{I}_d)^{-1}$. A straightforward calculation shows that over this time interval, $\frac{d\mu_t}{d\mu}$ is proportional to the density of a Gaussian with covariance Σ_t. Note that since $\tilde{\Sigma}_{0,0}^{-1} = \mathrm{I}_d$, we also have $\lim_{t \to 0^+} \Sigma_t^{-1} = 0$.

Now suppose that Σ_t has been defined up until time T_k; we will extend it to time T_{k+1}. Let L_k denote the affine hull of the support of μ_{T_k}, so that $\dim(L_k) = d - k$ (if $\dim(L_k) < d - k$, then we simply have $T_{k+1} = T_k$). Then, for $0 \leq t \leq T_{k+1} - T_k$, we may set

$$\Sigma_{T_k + t} := \left(\tilde{\Sigma}_{k,t}^{-1} + \Sigma_{T_k}^{-1} - \mathrm{I}_d \right)^{-1},$$

where the quantities involved are matrices over the subspace parallel to L_k but may also be regarded as degenerate bilinear forms in the ambient space $\mathbb{R}^d$. First, observe that continuity of the processes $\tilde{\Sigma}_{k,t}$ implies the same for Σ_t. Once again, a straightforward calculation shows that for $T_k \leq t < T_{k+1}$, $\frac{d\mu_t}{d\mu}$ is proportional to the density of a Gaussian with covariance Σ_t, where we view μ_t and μ as densities on L_k (for μ, we take its conditional density on L_k).

It remains only to show that Σ_t satisfies the required differential equation. From our construction, we see that Σ_t always takes the form $\left(\tilde{\Sigma}_t^{-1} - H \right)^{-1}$, where $H \preceq \mathrm{I}_d$ and

$$\frac{d}{dt} \tilde{\Sigma}_t = -\tilde{\Sigma}_t C_t^2 \tilde{\Sigma}_t.$$

Then, we have

$$\begin{aligned}
\frac{d}{dt} \Sigma_t &= -\left(\tilde{\Sigma}_t^{-1} - H \right)^{-1} \left(\frac{d}{dt} \tilde{\Sigma}_t^{-1} \right) \left(\tilde{\Sigma}_t^{-1} - H \right)^{-1} \\
&= -\Sigma_t \left(-\tilde{\Sigma}_t^{-1} \left(\frac{d}{dt} \tilde{\Sigma}_t \right) \tilde{\Sigma}_t^{-1} \right) \Sigma_t \\
&= -\Sigma_t C_t^2 \Sigma_t,
\end{aligned}$$

as desired. ∎

Proposition 1.19. $dA_t = \int_{\mathbb{R}^d} (x - a_t)^{\otimes 3} \mu_t(dx) C_t dB_t - A_t C_t^2 A_t dt$

Proof. We consider the Doob decomposition of $A_t = M_t + E_t$, where M_t is a local martingale and E_t is a process of bounded variation. By the previous two propositions and the definition

of A_t, we have on one hand

$$dA_t = d\int_{\mathbb{R}^d} x^{\otimes 2}\mu_t(dx) - da_t^{\otimes 2} = d\int_{\mathbb{R}^d} x^{\otimes 2}\mu_t(dx) - a_t \otimes da_t - da_t \otimes a_t - A_t C_t^2 A_t dt.$$

Clearly the first 3 terms are local martingales, which shows, by the uniqueness of the Doob decomposition, $dE_t = -A_t C_t^2 A_t dt$. On the other hand, one may also rewrite the above as

$$\begin{aligned} dA_t =& d\int_{\mathbb{R}^d} (x-a_t)^{\otimes 2}\mu_t(dx) = \int_{\mathbb{R}^d} d\left((x-a_t)^{\otimes 2}\mu_t(dx)\right) \\ =& -\int_{\mathbb{R}^d} da_t \otimes (x-a_t)\mu_t(dx) - \int_{\mathbb{R}^d} (x-a_t)\otimes da_t \mu_t(dx) + \int_{\mathbb{R}^d} (x-a_t)^{\otimes 2} d\mu_t(dx) \\ & - 2\int_{\mathbb{R}^d} (x-a_t)\otimes d[a_t, \mu_t(dx)]_t + \int_{\mathbb{R}^d} d[a_t,a_t]_t \mu_t(dx). \end{aligned}$$

Note that the first 2 terms are equal to 0, since, by definition of a_t,

$$\int_{\mathbb{R}^d} da_t \otimes (x-a_t)\mu_t(dx) = da_t \otimes \int_{\mathbb{R}^d} (x-a_t)\mu_t(dx) = 0.$$

Also, the last 2 terms are clearly of bounded variation, which shows

$$dM_t = \int_{\mathbb{R}^d} (x-a_t)^{\otimes 2} d\mu_t(dx) = \int_{\mathbb{R}^d} (x-a_t)^{\otimes 3} C_t \mu_t(dx) dB_t.$$

□

Define the stopping time $\tau = \inf\{t | A_t = 0\}$. Then, at time τ, μ_τ is just a delta mass located at a_τ and $\mu_s = \mu_\tau$ for every $s \geq \tau$. A crucial is observation is

Proposition 1.20. *Suppose that there exists constants $t_0 \geq 0$ and $c > 0$ such that a.s. one of the following happens*

1. *for every $t_0 < t < \tau$, $\mathrm{Tr}\left(A_t C_t^2 A_t\right) > c$,*
2. *$\int_0^{t_0} \lambda_{\min}\left(C_t^2\right) dt = \infty$, where $\lambda_{\min}\left(C_t^2\right)$ is the minimal eigenvalue of C_t^2,*

then τ is finite a.s. and in the second case $\tau \leq t_0$. Moreover, if τ is finite a.s. then a_τ has the law of μ.

Proof. Consider the process $R_t = A_t + \int_0^t A_s C_s^2 A_s ds$. For the first case, the previous proposition shows that the real-valued process $\mathrm{Tr}\left(R_t\right)$ a positive local martingale; hence, a supermartingale. By the martingale convergence theorem $\mathrm{Tr}\left(R_t\right)$ converges to a limit almost surely.

By our assumption, if $\tau = \infty$ then

$$\int_0^\infty \mathrm{Tr}(A_t C_t^2 A_t)dt \geq \int_{t_0}^\infty \mathrm{Tr}(A_t C_t^2 A_t)dt \geq \int_{t_0}^\infty c dt = \infty.$$

This would imply that $\lim_{t\to\infty} \mathrm{Tr}(A_t) = -\infty$ which clearly cannot happen.

For the second case, under the event $\{\tau > t_0\}$, by continuity of the process A_t there exists $a > 0$ such that for every $t \in [0, t_0]$, there is a unit vector $v_t \in \mathbb{R}^d$ for which $\langle v_t, A_t v_t\rangle \geq a$. We then have,

$$\int_0^{t_0} \mathrm{Tr}(A_t C_t^2 A_t)dt \geq \int_0^{t_0} \langle A_t v_t, C_t^2 A_t v_t\rangle dt \geq a^2 \int_0^{t_0} \lambda_{\min}(C_t^2)dt = \infty,$$

which implies $\lim_{t\to t_0} \mathrm{Tr}(A_t) = -\infty$. Again, this cannot happen and so $\mathbb{P}(\tau > t_0) = 0$.

To understand the law of a_τ, let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be any continuous bounded function. By Property 3 of Proposition 1.15 $\int_{\mathbb{R}^d} \varphi(x)\mu_t(dx)$ is a martingale. We claim that it is bounded. Indeed, observe that since μ_t is a probability measure for every t, then

$$\int_{\mathbb{R}^d} \varphi(x)\mu_t(dx) \leq \max_x |\varphi(x)|.$$

τ is finite a.s., so by the optional stopping theorem for continuous time martingales ([199] Theorem 7.2.4)

$$\mathbb{E}\left[\int_{\mathbb{R}^d} \varphi(x)\mu_\tau(dx)\right] = \int_{\mathbb{R}^d} \varphi(x)\mu(dx).$$

Since μ_τ is a delta mass, we have that $\int_{\mathbb{R}^d} \varphi(x)\mu_\tau(dx) = \varphi(a_\tau)$ which finishes the proof. □

We finish the section with an important property of the process A_t.

Proposition 1.21. *The rank of A_t is monotonic decreasing in t, and $\ker(A_t) \subset \ker(A_s)$ for $t \leq s$.*

Proof. To see that $\mathrm{rank}(A_t)$ is indeed monotonic decreasing, let v_0 be such that $A_{t_0} v_0 = 0$ for some $t_0 > 0$, we will show that for any $t \geq t_0$, $A_t v_0 = 0$. In a similar fashion to Proposition 1.20, we define the process $\langle v_0, A_t v_0\rangle + \int_0^t \langle v_0, A_s C_s^2 A_s v_0\rangle ds$, which is, using Proposition 1.19, a positive local martingale and so a super-martingale. This then implies that $\langle v_0, A_t v_0\rangle$ is itself a positive super-martingale. Since $\langle v_0, A_{t_0} v_0\rangle = 0$, we have that for any $t \geq t_0$, $\langle v_0, A_t v_0\rangle = 0$

as well.

□

1.3 Convergence rates in transportation distance

1.3.1 The case of bounded random vectors: proof of Theorem 1.1

In this subsection we fix a measure μ on $\mathbb{R}^d$ and a random vector $X \sim \mu$ with the assumption that $\|X\| \leq \beta$ almost surely for some $\beta > 0$. We also assume that $\mathbb{E}[X] = 0$.

We define the martingale process a_t along with the stopping time τ as in Section 1.2.3, where we take $C_t = A_t^\dagger$, so that $a_t = \int_0^t A_s A_s^\dagger dB_s$. We denote $P_t := A_t A_t^\dagger$, and remark that since A_t is symmetric, P_t is a projection matrix. As such, we have that for any $t < \tau$, $\mathrm{Tr}(P_t) \geq 1$. By Proposition 1.20, a_τ has the law μ.

In light of the remark following Theorem 1.10, our first objective is to understand the expectation of τ.

Lemma 1.22. *Under the boundedness assumption $\|X\| \leq \beta$, we have $\mathbb{E}[\tau] \leq \beta^2$.*

Proof. Let $H_t = \|a_t\|^2$. By Itô's formula and since P_t is a projection matrix,

$$dH_t = 2\langle a_t, P_t dB_t\rangle + \mathrm{Tr}(P_t)\, dt = 2\langle a_t, P_t dB_t\rangle + \mathrm{rank}(P_t)\, dt.$$

So, $\frac{d}{dt}\mathbb{E}[H_t] = \mathbb{E}[\mathrm{rank}(P_t)]$. Since $\mathbb{E}[H_\infty] \leq \beta^2$,

$$\beta^2 \geq \mathbb{E}[H_\infty] - \mathbb{E}[H_0] = \int_0^\infty \mathbb{E}[\mathrm{rank}(P_t)]\, dt \geq \int_0^\infty \mathbb{P}(\tau > t)\, dt = \mathbb{E}[\tau].$$

□

The above claim gives bounds on the expectation of τ, however in order to use Theorem 1.10, we need bounds for its tail behaviour in the sense of (1.2). To this end, we can use a bootstrap argument and invoke the above lemma with the measure μ_t in place of μ, recalling that $X_\infty|\mathcal{F}_t \sim \mu_t$ and noting that $\|X_\infty|\mathcal{F}_t\| \leq \beta$ almost surely. Therefore, we can consider the conditioned stopping time $\tau|\mathcal{F}_t - t$ and get that

$$\mathbb{E}[\tau|\mathcal{F}_t] \leq t + \beta^2.$$

The following lemma will make this precise.

Lemma 1.23. *Suppose that, for the stopping time τ, it holds that for every $t > 0$, $\mathbb{E}[\tau|\mathcal{F}_t] \leq t + \beta^2$ a.s., then*

$$\forall i \in \mathbb{N}, \ \ \mathbb{P}\left(\tau \geq i \cdot 2\beta^2\right) \leq \frac{1}{2^i}. \tag{1.10}$$

Proof. Denote $t_i = i \cdot 2\beta^2$. Since μ_t is Markovian, and by the law of total probability, for any $i \in \mathbb{N}$ we have the relation

$$\mathbb{P}(\tau \geq t_{i+1}) \leq \mathbb{P}(\tau > t_i) \operatorname*{ess\,sup}_{\mu_{t_i}} \left(\mathbb{P}\left(\tau - t_i \geq 2\beta^2 | \mathcal{F}_{t_i}\right)\right),$$

where the essential supremum is taken over all possible states of μ_{t_i}. Using Markov's inequality, we almost surely have

$$\mathbb{P}\left(\tau - t_i \geq 2\beta^2 | \mathcal{F}_{t_i}\right) \leq \frac{\mathbb{E}[\tau - t_i | \mathcal{F}_{t_i}]}{2\beta^2} \leq \frac{1}{2},$$

which is also true for the essential supremum. Clearly $\mathbb{P}(\tau \geq 0) = 1$ which finishes the proof. □

Proof of Theorem 1.1. Our objective is to apply Theorem 1.10, defining $X_t = a_t$ and $\Gamma_t = P_t$ so that (X_t, Γ_t, τ) becomes a martingale embedding according to Proposition 1.20. In this case, we have that Γ_t is a projection matrix almost surely. Thus,

$$\operatorname{Tr}\left(\mathbb{E}[\Gamma_t^4]\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right) \leq d,$$

and

$$\operatorname{Tr}\left(\mathbb{E}[\Gamma_t^2]\right) \leq d\mathbb{P}(\tau > t).$$

Therefore, if G and S_n are defined as in Theorem 1.10, then

$$\begin{aligned}
\mathcal{W}_2^2(S_n, G) &\leq \int_0^{2\beta^2 \log_2(n)} \frac{d}{n} dt + \int_{2\beta^2 \log_2(n)}^{\infty} 4d\mathbb{P}(\tau > t) dt \\
&\leq \frac{2d\beta^2 \log_2(n)}{n} + 4d \int_{2\beta^2 \log_2(n)}^{\infty} \mathbb{P}\left(\tau > \left\lfloor \frac{t}{2\beta^2} \right\rfloor 2\beta^2\right) dt \\
&\overset{(1.10)}{\leq} \frac{2d\beta^2 \log_2(n)}{n} + 4d \int_{2\beta^2 \log_2(n)}^{\infty} \left(\frac{1}{2}\right)^{\left\lfloor \frac{t}{2\beta^2} \right\rfloor} dt \\
&\leq \frac{2d\beta^2 \log_2(n)}{n} + 8d\beta^2 \sum_{j=\lfloor \log_2(n) \rfloor}^{\infty} \frac{1}{2^j} \leq \frac{2d\beta^2 \log_2(n)}{n} + \frac{32d\beta^2}{n}.
\end{aligned}$$

Taking square roots, we finally have

$$\mathcal{W}_2(S_n, G) \leq \frac{\beta\sqrt{d}\sqrt{32 + 2\log_2(n)}}{\sqrt{n}},$$

as required. □

1.3.2 The case of log-concave vectors: proof of Theorem 1.2

In this section we fix μ to be an isotropic log concave measure. The processes $a_t = a_t^\mu, A_t = A_t^\mu$ are defined as in Section 1.2.3 along with the stopping time τ. To define the matrix process C_t, we first define a new stopping time

$$T := 1 \wedge \inf\{t | \, \|A_t\|_{op} \geq 2\}.$$

C_t is then defined in the following manner:

$$C_t = \begin{cases} \min(A_t^\dagger, \mathrm{I}_d) & \text{if } t \leq T \\ A_t^\dagger & \text{otherwise} \end{cases}$$

where, again, $A_t^\dagger$ denotes the pseudo-inverse of A_t and $\min(A_t^\dagger, \mathrm{I}_d)$ is the unique matrix which is diagonalizable with respect to the same basis as $A_t^\dagger$ and such that each of its eigenvalues corresponds to an an eigenvalue of $A_t^\dagger$ truncated at 1. Since $\mathrm{Tr}\left(A_t A_t^\dagger\right) \geq 1$ whenever $t \leq \tau$, then the conditions of Proposition 1.20 are clearly met for $t_0 = 1$ and a_τ has the law of μ.

In order to use Theorem 1.10, we will also need to demonstrate that τ has subexponential tails in the sense of (1.2). For this, we first relate τ to the stopping time T.

Lemma 1.24. $\tau < 1 + \frac{4}{T}$.

Proof. Let Σ_t be as in Proposition 1.18. As described in the proposition, μ_t is proportional to μ times a Gaussian of covariance Σ_t, on an appropriate affine subspace. In this case, an application of the Brascamp-Lieb inequality (see [133] for details) shows that $A_t = \mathrm{Cov}(\mu_t) \preceq \Sigma_t$. In particular, this means that for $t > T$, when restricted to the orthogonal complement of $\ker(A_t)$, the following inequality holds,

$$\frac{d}{dt}\Sigma_t = -\Sigma_t C_t^2 \Sigma_t \preceq -\mathrm{I}_d.$$

So, $\tau \leq T + \|\Sigma_T\|_{op}$.

It remains to estimate $\|\Sigma_T\|_{op}$. To this end, recall that for $0 < t \leq T$, we have $\|A_t\|_{op} \leq 2$, which implies

$$\frac{d}{dt}\Sigma_t = -\Sigma_t C_t^2 \Sigma_t \preceq -\frac{1}{4}\Sigma_t^2.$$

Now, consider the differential equation $f'(t) = -\frac{1}{4}f(t)^2$ with $f(T) = \|\Sigma_T\|_{op}$, which has solution $f(t) = \frac{4}{t-T+\frac{4}{\|\Sigma_T\|_{op}}}$. By Gronwall's inequality, $f(t)$ lower bounds $\|\Sigma_t\|_{op}$ for $0 < t \leq T$, and so, in particular, $f(t)$ must remain finite within that interval. Consequently, we have

$$\frac{4}{\|\Sigma_T\|_{op}} > T \implies \|\Sigma_T\|_{op} < \frac{4}{T}.$$

We conclude that

$$\tau \leq T + \|\Sigma_T\|_{op} < 1 + \frac{4}{T},$$

as desired. □

Lemma 1.25. *There exist universal constants $c, C > 0$ such that if $s > C \cdot \kappa_d^2 \ln(d)^2$ and $d \geq 8$ then*

$$\mathbb{P}(\tau > s) \leq e^{-cs},$$

where κ_d is the constant defined in (1.1).

Proof. First, by using the previous claim, we may see that for any $s \geq 5$,

$$\mathbb{P}\left(\tau > s\right) \leq \mathbb{P}\left(\frac{1}{T} \geq \frac{s-1}{4}\right) \leq \mathbb{P}\left(\frac{1}{T} \geq \frac{s}{5}\right) = \mathbb{P}\left(5s^{-1} \geq T\right) = \mathbb{P}\left(\max_{0 \leq t \leq 5s^{-1}} \|A_t\|_{op} \geq 2\right).$$

Recall from Proposition 1.19,

$$dA_t = \int_{\mathbb{R}^d} (x - a_t) \otimes (x - a_t) \langle C_t\left(x - a_t\right), dB_t \rangle \mu_t(dx) - A_t C_t^2 A_t dt.$$

Since we are trying to bound the operator norm of A_t, we might as well just consider the matrix $\tilde{A}_t = A_t + \int_0^t A_s C_s^2 A_s ds$. Note that, by definition of T, for any $t \leq T$,

$$\int_0^t A_s C_s^2 A_s ds \preceq \mathrm{I}_d.$$

Thus, for $t \in [0, T]$,

$$3\mathrm{I}_d \succeq A_t + \mathrm{I}_d \succeq \tilde{A}_t \succeq A_t. \tag{1.11}$$

Also, $\tilde{A}_t$ can be written as,

$$d\tilde{A}_t = \int_{\mathbb{R}^d} (x - a_t) \otimes (x - a_t) \langle C_t(x - a_t), dB_t \rangle \mu_t(dx), \ \tilde{A}_0 = \mathrm{I}_d. \tag{1.12}$$

The above shows

$$\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}\|A_t\|_{op}\ge 2\right)\le\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}||\tilde{A}_t||_{op}\ge 2\right).$$

We note than whenever $||\tilde{A}_t||_{op}\ge 2$ then also $\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)^{\frac{1}{4\ln(d)}}\ge 2$, so that

$$\begin{aligned}\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}||\tilde{A}_t||_{op}\ge 2\right)&\le\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)^{\frac{1}{4\ln(d)}}\ge 2\right)\\&\le\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}\ln\left(\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)\right)\ge 2\ln(d)\right)=\mathbb{P}\left(\max_{0\le t\le 5s^{-1}}(M_t+E_t)\ge 2\ln(d)\right),\end{aligned}\tag{1.13}$$

where M_t and E_t form the Doob-decomposition of $\ln\left(\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)\right)$. That is, M_t is a local martingale and E_t is a process of bounded variation. To calculate the differential of the Doob-decomposition, fix t, let $v_1,...,v_n$ be the unit eigenvectors of $\tilde{A}_t$ and let $\alpha_{i,j}=\langle v_i,\tilde{A}_t v_j\rangle$ with

$$d\alpha_{i,j}=\int_{\mathbb{R}^d}\langle x,v_i\rangle\langle x,v_j\rangle\langle C_t x,dB_t\rangle\mu_t(dx+a_t),$$

which follows from (1.12). Also define

$$\xi_{i,j}=\frac{1}{\sqrt{\alpha_{i,i}\alpha_{j,j}}}\int_{\mathbb{R}^d}\langle x,v_i\rangle\langle x,v_j\rangle C_t x\mu_t(dx+a_t).$$

So that

$$d\alpha_{i,j}=\sqrt{\alpha_{i,i}\alpha_{j,j}}\langle\xi_{i,j},dB_t\rangle,\quad\frac{d}{dt}[\alpha_{i,j}]_t=\alpha_{i,i}\alpha_{j,j}\left\|\xi_{i,j}\right\|^2.$$

Now, since v_i is an eigenvector corresponding to the eigenvalue $\alpha_{i,i}$, we have

$$\xi_{i,j}=\int_{\mathbb{R}^d}\langle\tilde{A}_t^{-1/2}x,v_i\rangle\langle\tilde{A}_t^{-1/2}x,v_j\rangle C_t x\mu_t(dx+a_t).$$

If we define the measure $\tilde{\mu}_t(dx)=\det(\tilde{A}_t)^{1/2}\mu_t(\tilde{A}_t^{1/2}dx+a_t)$, then $\tilde{\mu}_t$ has the law of a centered log-concave random vector with covariance $\tilde{A}_t^{-1/2}A_t\tilde{A}_t^{-1/2}\preceq \mathrm{I}_d$. By making the substitution $y=\tilde{A}_t^{-1/2}x$, the above expression becomes

$$\xi_{i,j}=\int_{\mathbb{R}^d}\langle y,v_i\rangle\langle y,v_j\rangle C_t\tilde{A}_t^{1/2}y\tilde{\mu}_t(dy).$$

By (1.11) and the definition of T, C_t, for any $t\le T$, $\tilde{A}_t^{1/2}\preceq 2\mathrm{I}_d$ and $C_t\preceq\mathrm{I}_d$. So, $\left\|C_t\tilde{A}_t^{1/2}\right\|_{op}\le 2$. Under similar conditions, it was shown in [97], Lemma 3.2, that there exists a universal

constant $C > 0$ for which

- for any $1 \leq i \leq d$, $\|\xi_{i,i}\|^2 \leq C$.
- for any $1 \leq i \leq d$, $\sum_{j=1}^{d} \|\xi_{i,j}\|^2 \leq C\kappa_d^2$.

Furthermore, in the proof of Proposition 3.1 in the same paper it was shown

$$d\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right) \leq 4\ln(d)\sum_{i=1}^{d} \alpha_{i,i}^{4\ln(d)}\langle \xi_{i,i}, dB_t\rangle + 16C\kappa_d^2\ln(d)^2\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)dt.$$

So, using Itô's formula with the function $\ln(x)$ we can calculate the differential of the Doob decomposition (1.13). Specifically, we use the fact that the second derivative of $\ln(x)$ is negative and get

$$dE_t \leq 16C\kappa_d^2\ln(d)^2\frac{\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)}{\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)} = 16C\kappa_d^2\ln(d)^2,\ E_0 = \ln(d),$$

and

$$\frac{d}{dt}[M]_t \leq 16C^2\ln(d)^2\left(\frac{\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)}{\mathrm{Tr}\left(\tilde{A}_t^{4\ln(d)}\right)}\right)^2 = 16C^2\ln(d)^2. \tag{1.14}$$

Hence, $E_t \leq t \cdot 16C\kappa_n^2\ln(d)^2 + \ln(d)$, which together with (1.13) gives

$$\mathbb{P}(\tau > s) \leq \mathbb{P}\left(\max_{0\leq t\leq 5s^{-1}} M_t \geq 2\ln(d) - \ln(d) - 80s^{-1}C\kappa_d^2\ln(d)^2\right)\ \forall s \geq 5.$$

Under the assumption $s > 80C\kappa_d^2\ln(d)^2$, and since $d \geq 8$, the above can simplify to

$$\mathbb{P}(\tau > s) \leq \mathbb{P}\left(\max_{0\leq t\leq 5s^{-1}} M_t \geq \frac{1}{2}\ln(d)\right). \tag{1.15}$$

To bound this last expression, we will apply the Dubins-Schwartz theorem to write

$$M_t = W_{[M]_t},$$

where W_t is some Brownian motion. Combining this with (1.15) gives

$$\mathbb{P}(\tau > s) \leq \mathbb{P}\left(\max_{0\leq t\leq 5s^{-1}} W_{[M]_t} \geq \frac{\ln(d)}{2}\right).$$

An application of Doob's maximal inequality ([208] Proposition I.1.8) shows that for any $t', K > 0$

$$\mathbb{P}\left(\max_{0\leq t\leq t'} W_t \geq K\right) \leq \exp\left(-\frac{K^2}{2t'}\right).$$

We now integrate (1.14) and use the above inequality to obtain

$$\mathbb{P}\left(\max_{0\le t\le 5s^{-1}} W_{[M]_t} \ge \frac{\ln(d)}{2}\right) \le e^{-cs},$$

where $c > 0$ is some universal constant. □

Proof of Theorem 1.2. By definition of T and C_t, we have that for any $t \le T$, $A_t C_t \preceq 2\mathrm{I}_d$ and for any $t > T$, $A_t C_t = A_t A_t^\dagger \preceq \mathrm{I}_d$. We now invoke Theorem 1.10, with $\Gamma_t = A_t C_t$, for which

$$\mathrm{Tr}\left(\mathbb{E}[\Gamma_t^4]\mathbb{E}\left[\Gamma_t^2\right]^\dagger\right) \le 4d,$$

and, by Lemma 1.25

$$\mathrm{Tr}\left(\mathbb{E}[\Gamma_t^2]\right) \le 4d\mathbb{P}\left(\tau > t\right) \le 4de^{-ct} \quad \forall t > C\cdot\kappa_d^2\ln(d)^2.$$

If G is the standard d-dimensional Gaussian, then the theorem yields

$$\begin{aligned}\mathcal{W}_2^2(S_n, G) &\le \int_0^{C\cdot\kappa_d^2\ln(d)^2\ln(n)} 4\frac{d}{n}dt + \int_{C\cdot\kappa_d^2\ln(d)^2\ln(n)}^{\infty} 16d\mathbb{P}\left(\tau > t\right)\\ &\le 4\frac{dC\cdot\kappa_d^2\ln(d)^2\ln(n)}{n} + 16d\int_{C\cdot\kappa_d^2\ln(d)^2\ln(n)}^{\infty} e^{-ct}dt\\ &\le C'\frac{d\cdot\kappa_d^2\ln(d)^2\ln(n)}{n}.\end{aligned}$$

Thus

$$\mathcal{W}_2(S_n, G) \le \frac{C\kappa_d\ln(d)\sqrt{d\ln(n)}}{\sqrt{n}},$$

□

1.4 Convergence rates in entropy

Throughout this section, we fix a centered measure μ on $\mathbb{R}^d$ with an invertible covariance matrix Σ and $G \sim \mathcal{N}(0, \Sigma)$. Let $\{X^{(i)}\}$ be independent copies of $X \sim \mu$ and $S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n X^{(i)}$.

Our goal is to study the quantity $\mathrm{Ent}\left(S_n||G\right)$. In light of Theorem 1.12, we aim to construct a martingale embedding $(X_t, \Gamma_t, 1)$ such that $X_1 \sim \mu$ and which satisfies appropriate bounds on the matrix Γ_t. Our construction uses the process a_t from Proposition 1.15 with the choice

$C_t := \frac{1}{1-t}\mathrm{I}_d$. Property 2 in Proposition 1.15 gives

$$a_t = \int_0^t \frac{A_s}{1-s} dB_s.$$

Thus, we denote

$$\Gamma_t := \frac{A_t}{1-t}.$$

Since $\int_0^1 \lambda_{min}(C_t^2) = \infty$, Proposition 1.20 shows that the triplet $(a_t, \Gamma_t, 1)$ is a martingale embedding of μ. As above, the sequence $\Gamma_t^{(i)}$ will denote independent copies of Γ_t and we define $\tilde{\Gamma}_t := \sqrt{\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2}$.

1.4.1 Properties of the embedding

Let v_t stand for the F"ollmer drift, defined by 7, in the Introduction, and denote

$$Y_t := B_t + \int_0^t v_s ds.$$

In [100] (Section 2.2) it was shown that the density of the measure $Y_1|\mathcal{F}_t$ has the same dynamics as the density of μ_t. Thus, almost surely $Y_1|\mathcal{F}_t \sim \mu_t$ and since a_t is the expectation of μ_t, we have the identity

$$a_t = \mathbb{E}\left[Y_1|\mathcal{F}_t\right], \tag{1.16}$$

and in particular we have $a_1 = Y_1$. Moreover, the same reasoning implies that $A_t = \mathrm{Cov}(Y_1|\mathcal{F}_t)$ and

$$\Gamma_t = \frac{\mathrm{Cov}(Y_1|\mathcal{F}_t)}{1-t}. \tag{1.17}$$

The following identity, which is immediate from 7, will be crucial in the sequel,

$$\mathrm{Ent}(Y_1||\gamma) = \frac{1}{2}\int_0^1 \mathbb{E}\left[\|v_t\|^2\right] dt. \tag{1.18}$$

Lemma 1.26. *It holds that* $\frac{d}{dt}\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] = -\mathbb{E}\left[\Gamma_t^2\right]$.

Proof. From (1.16), we have

$$\mathrm{Cov}(Y_1|\mathcal{F}_t) = \mathbb{E}\left[Y_1^{\otimes 2}|\mathcal{F}_t\right] - \mathbb{E}\left[Y_1|\mathcal{F}_t\right]^{\otimes 2} = \mathbb{E}\left[Y_1^{\otimes 2}|\mathcal{F}_t\right] - a_t^{\otimes 2}.$$

a_t is a martingale, hence

$$\frac{d}{dt}\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] = -\frac{d}{dt}\mathbb{E}\left[[a]_t\right] = -\mathbb{E}\left[\Gamma_t^2\right]. \tag{1.19}$$

□

Our next goal is to recover v_t from the martingale a_t.

Lemma 1.27. *The drift v_t satisfies that identity $v_t = \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s$. Furthermore,*

$$\mathbb{E}\left[\|v_t\|^2\right] = \int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_s - \mathrm{I}_d)^2\right]\right)}{(1-s)^2} ds. \tag{1.20}$$

Proof. Recall identity (10), from the Introduction,

$$v_t = \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s.$$

The claim follows from a direct application of Itô's isometry. □

A combination of equations (1.18) and (1.20) gives the useful identity,

$$\mathrm{Ent}\left(Y_1||\gamma\right) = \frac{1}{2}\int_0^1 \int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_s - \mathrm{I}_d)^2\right]\right)}{(1-s)^2} ds dt = \frac{1}{2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]\right)}{1-t} dt, \tag{1.21}$$

which was also shown in (11). We can further capitalize on the the previous lemma to obtain a representation for $\mathbb{E}\left[\mathrm{Tr}\left(\Gamma_t\right)\right]$, in terms of $\mathbb{E}\left[\|v_t\|^2\right]$.

Lemma 1.28. *It holds that*

$$\mathbb{E}\left[\mathrm{Tr}(\Gamma_t)\right] = d - (1-t)\left(d - \mathrm{Tr}\left(\Sigma\right) + \mathbb{E}\left[\|v_t\|^2\right]\right).$$

Proof. The identity can be obtained through integration by parts. By Lemma 1.27,

$$\begin{aligned}
\mathbb{E}[\|v_t\|^2] &\overset{(1.20)}{=} \int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_s - \mathrm{I}_d)^2\right]\right)}{(1-s)^2} ds \\
&= \int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_s^2\right]\right)}{(1-s)^2} ds - 2\int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_s\right]\right)}{(1-s)^2} ds + \int_0^t \frac{\mathrm{Tr}\left(\mathrm{I}_d\right)}{(1-s)^2} ds.
\end{aligned}$$

Since, by Lemma 1.26, $\frac{d}{dt}\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right)\right] = -\mathbb{E}\left[\Gamma_t^2\right]$ integration by parts shows

$$\int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_s^2\right]\right)}{(1-s)^2}ds = -\frac{\mathrm{Tr}\left(\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_s\right)\right]\right)}{(1-s)^2}\Bigg|_0^t + 2\int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_s\right)\right]\right)}{(1-s)^3}ds$$
$$= \mathrm{Tr}\left(\Sigma\right) - \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t\right]\right)}{1-t} + 2\int_0^t \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_s\right]\right)}{(1-s)^2}ds,$$

where we have used (1.17) and the fact $\mathrm{Cov}\left(Y_1|\mathcal{F}_0\right) = \mathrm{Cov}\left(Y_1\right) = \Sigma$. Plugging this into the previous equation shows

$$\mathbb{E}[\|v_t\|^2] = \mathrm{Tr}\left(\Sigma\right) - \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t\right]\right)}{1-t} + \frac{d}{1-t} - d.$$

or equivalently

$$\mathbb{E}\left[\mathrm{Tr}(\Gamma_t)\right] = d - (1-t)\left(d - \mathrm{Tr}\left(\Sigma\right) + \mathbb{E}\left[\|v_t\|^2\right]\right).$$

□

Next, as in Theorem 1.12, we define σ_t to be the minimal eigenvalue of $\mathbb{E}\left[\Gamma_t\right]$, so that

$$\mathbb{E}\left[\Gamma_t\right] \succeq \sigma_t \mathrm{I}_d.$$

Note that by Jensen's inequality we also have

$$\mathbb{E}\left[\Gamma_t^2\right] \succeq \sigma_t^2 \mathrm{I}_d. \tag{1.22}$$

Lemma 1.29. *Assume that* $\mathrm{Ent}(Y_1||\gamma) < \infty$. *Then* Γ_t *is almost surely invertible for all* $t \in [0,1)$ *and, moreover, there exists a constant* $m = m_\mu > 0$ *for which*

$$\sigma_t \geq m, \quad \forall t \in [0,1).$$

Proof. We will show that for every $0 \leq t < 1$, $\sigma_t > 0$ and that there exists $c > 0$ such that $\sigma_t > \frac{1}{8}$ whenever $t > 1 - c$. The claim will then follow by continuity of σ_t. The key to showing this is identity (1.21), due to which,

$$\mathrm{Ent}\left(Y_1||\gamma\right) = \frac{1}{2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]\right)}{1-t}dt.$$

Recall that, by Equation (1.17), $\Gamma_t = \frac{\mathrm{Cov}(Y_1|\mathcal{F}_t)}{1-t}$ and observe that, by Proposition 1.21, if $\mathrm{Cov}\left(Y_1|\mathcal{F}_s\right)$ is not invertible for some $0 \leq s < 1$ then $\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right)$ is also not invertible for any $t > s$. Under this event, we would have that $\int_s^1 \frac{\mathrm{Tr}\left((\Gamma_t - \mathrm{I}_d)^2\right)}{1-t}dt = \infty$ which, using the above

display, implies that the probability of this event must be zero. Therefore, Γ_t is almost surely invertible and $\sigma_t > 0$ for all $t \in [0,1)$.

Suppose now that for some $t' \in [0,1]$, $\sigma_{t'} \leq \frac{1}{8}$. By Jensen's inequality, we have

$$\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]\right) \geq \mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t - \mathrm{I}_d\right]^2\right) \geq (1-\sigma_t)^2 \geq 1 - 2\sigma_t.$$

Since, by Lemma 1.26, $\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right)\right]$ is non increasing, for any $t' \leq t \leq t' + \frac{1-t'}{2}$,

$$\sigma_t \leq \frac{\sigma_{t'}(1-t')}{1-t} \leq \frac{1-t'}{8(1-t'-\frac{1-t'}{2})} = \frac{1}{4}.$$

Now, assume by contradiction that there exists a sequence $t_i \in (0,1)$ such that $\sigma_{t_i} \leq \frac{1}{8}$ and $\lim_{i\to\infty} t_i = 1$. By passing to a subsequence we may assume that $t_{i+1} - t_i \geq \frac{1-t_i}{2}$ for all i. The assumption $\mathrm{Ent}(Y_1||\gamma) < \infty$ combined with Equation (1.21) and with the last two displays finally gives

$$\infty > \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]\right)}{1-t} dt \geq \int_0^1 \frac{1-2\sigma_t}{1-t} dt \geq \sum_{i=1}^{\infty} \int_{t_i}^{t_i + \frac{1-t_i}{2}} \frac{1}{2}\frac{1}{1-t} dt \geq \log 2 \sum_{i=1}^{\infty} \frac{1}{2},$$

which leads to a contradiction and completes the proof. □

1.4.2 Proof of Theorem 1.5

Thanks to the assumption $\mathrm{Ent}\left(Y_1||G\right) < \infty$, an application of Lemma 1.29 gives that Γ_t is invertible almost surely, so we may invoke the second bound in Theorem 1.12 to obtain

$$\mathrm{Ent}(S_n||G) \leq \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{(1-t)^2} \left(\int_t^1 \sigma_s^{-2} ds\right) dt.$$

The same lemma also shows that for some $m > 0$ one has

$$\int_t^1 \sigma_s^{-2} ds \leq \frac{1-t}{m^2}.$$

Therefore, we attain that

$$\mathrm{Ent}(S_n||G) \leq \frac{1}{m^2} \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{1-t} dt. \tag{1.23}$$

Next, observe that, by Itô's isometry,

$$\mathrm{Cov}(X) = \int_0^1 \mathbb{E}\left[\Gamma_t^2\right] dt.$$

Hence, as long as $\mathrm{Cov}(X)$ is finite, $\mathbb{E}\left[\Gamma_t^2\right]$ is also finite for all $t \in A$ where $[0,1] \setminus A$ is a set of measure 0. We will use this fact to show that

$$\lim_{n\to\infty} \mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right) = 0, \quad \forall t \in A. \tag{1.24}$$

Indeed, by the law of large numbers, $\tilde{\Gamma}_t$ almost surely converges to $\sqrt{\mathbb{E}\left[\Gamma_t^2\right]}$. Since $\left(\Gamma_t^{(i)}\right)^2$ are integrable, we get that the sequence $\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2$ is uniformly integrable. We now use the inequality

$$\tilde{\Gamma}_t \preceq \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2 + \mathrm{I}_d} \preceq \frac{1}{n}\sum_{i=1}^n \left(\Gamma_t^{(i)}\right)^2 + \mathrm{I}_d,$$

to deduce that $\tilde{\Gamma}_t$ is uniformly integrable as well. An application of Vitali's convergence theorem (see [118], for example) implies (1.24).

We now know that the integrand in the right hand side of (1.23) convergence to zero for almost every t. It remains to show that the expression converges as an integral, for which we intend to apply the dominated convergence theorem. It thus remains to show that the expression

$$\frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{1-t}$$

is bounded by an integrable function, uniformly in n, which would imply that

$$\lim_{n\to\infty} \mathrm{Ent}(S_n||G) = 0,$$

and the proof would be complete. To that end, recall that the square root function is concave on positive definite matrices (see e.g., [9]), thus

$$\tilde{\Gamma}_t \succeq \frac{1}{n}\sum_{i=1}^n \Gamma_t^{(i)}.$$

It follows that

$$\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right) \leq \mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right] - \mathbb{E}\left[\Gamma_t\right]^2\right) \leq \mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]\right).$$

So we have

$$\frac{1}{m^2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right]-\mathbb{E}\left[\tilde{\Gamma}_t\right]^2\right)}{1-t}dt \leq \frac{1}{m^2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_t-\mathrm{I}_d)^2\right]\right)}{1-t}dt$$
$$\overset{(1.21)}{=} \frac{2}{m^2}\mathrm{Ent}\left(Y_1||\gamma\right) < \infty.$$

This completes the proof.

1.4.3 Quantitative bounds for log concave random vectors

In this section, we make the additional assumption that the measure μ is log concave. Under this assumption, we show how one can obtain explicit convergence rates in the central limit theorem. Our aim is to use the bound in Theorem 1.12 for which we are required to obtain bounds on the process Γ_t. We begin by recording several useful facts concerning this process.

Lemma 1.30. *The process Γ_t has the following properties:*

1. *If μ is log concave, then for every $t \in [0,1]$, $\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d$, almost surely.*

2. *If μ is also 1-uniformly log concave, then for every $t \in [0,1]$, $\Gamma_t \preceq \mathrm{I}_d$ almost surely.*

Proof. Denote by ρ_t the density of $Y_1|\mathcal{F}_t$ with respect to the Lebesgue measure with $\rho := \rho_0$ being the density of μ. By Proposition 1.18 with $C_t = \frac{\mathrm{I}_d}{1-t}$, we can calculate the ratio between ρ_t and ρ. In particular, we have

$$\frac{d}{dt}\Sigma_t^{-1} = -\Sigma_t^{-1}\left(\frac{d}{dt}\Sigma_t\right)\Sigma_t^{-1} = \frac{1}{(1-t)^2}\mathrm{I}_d.$$

Solving this differential equation with the initial condition $\Sigma_0^{-1} = 0$, we find that $\Sigma_t^{-1} = \frac{t}{1-t}\mathrm{I}_d$.

Since the ratio between ρ_t and ρ is proportional to the density of a Gaussian with covariance Σ_t, we thus have

$$-\nabla^2 \log(\rho_t) = -\nabla^2 \log(\rho) + \frac{t}{1-t}\mathrm{I}_d.$$

Now, if μ is log concave then $Y_1|\mathcal{F}_t$ is almost surely $\frac{t}{1-t}$-uniformly log-concave. By the Brascamp-Lieb inequality (as in [133]) we get $\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right) \preceq \frac{1-t}{t}\mathrm{I}_d$ and, using (1.17),

$$\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d.$$

If μ is also 1-uniformly log-concave then $-\nabla^2 \log(\rho) \succeq \mathrm{I}_d$ and almost surely

$$-\nabla^2 \log(\rho_t) \succeq \frac{1}{1-t}\mathrm{I}_d.$$

By the same argument this implies

$$\Gamma_t \preceq \mathrm{I}_d.$$

□

The relative entropy to the Gaussian of a log concave measure with non-degenerate covariance structure is finite (it is even universally bounded, see [173]). Thus, by Lemma 1.29, it follows that Γ_t is invertible almost surely. This allows us to invoke the first bound of Theorem 1.12,

$$\mathrm{Ent}(S_n||G) \leq \frac{1}{n}\int_0^1 \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)^2\sigma_t^2}\left(\int_t^1 \sigma_s^{-2} ds\right) dt. \tag{1.25}$$

Attaining an upper bound on the right hand side amounts to a concentration estimate for the process Γ_t^2 and a lower bound on σ_t. These two tasks are the objective of the following two lemmas.

Lemma 1.31. *If μ is log concave and isotropic then for any $t \in [0,1)$,*

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) \leq \frac{1-t}{t^2}\left(\frac{d(1+t)}{t^2} + 2\mathbb{E}\left[\|v_t\|^2\right]\right),$$

and

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) \leq C\frac{d^4}{(1-t)^4}$$

for a universal constant $C > 0$.

Proof. The isotropicity of μ, used in conjunction with the formula given in Lemma 1.28, yields

$$\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right]\right) \geq \frac{1}{d}\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t\right]\right)^2 \geq d - 2(1-t)\mathbb{E}\left[\|v_t\|^2\right],$$

where the first inequality follows by convexity. Since μ is log concave, Lemma 1.30 ensures that, almost surely, $\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d$. Therefore,

$$\begin{aligned}
\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) &\leq \mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \frac{1}{t^2}\mathrm{I}_d\right)^2\right]\right) \\
&= \frac{1}{t^4}\mathrm{Tr}\left(\mathbb{E}\left[\left(\mathrm{I}_d - t^2\Gamma_t^2\right)^2\right]\right) \\
&\leq \frac{1}{t^4}\mathrm{Tr}\left(\mathbb{E}\left[\mathrm{I}_d - t^2\Gamma_t^2\right]\right) \\
&\leq \frac{1-t}{t^2}\left(\frac{d(1+t)}{t^2} + 2\mathbb{E}\left[\|v_t\|^2\right]\right).
\end{aligned}$$

Which proves the first bound. Towards the second bound, we use (1.17) to write

$$\Gamma_t^2 \preceq \frac{1}{(1-t)^2}\mathbb{E}\left[Y_1^{\otimes 2}|\mathcal{F}_t\right]^2.$$

So,

$$\mathbb{E}\left[\left\|\Gamma_t^2\right\|_{HS}^2\right] \leq \frac{1}{(1-t)^4}\mathbb{E}\left[\left\|\|Y_1\|^2 Y_1^{\otimes 2}\right\|_{HS}^2\right] \leq \frac{1}{(1-t)^4}\mathbb{E}\left[\|Y_1\|^8\right].$$

For an isotropic log concave measure, the expression $\mathbb{E}\left[\|Y_1\|^8\right]$ is bounded from above by Cd^4 for a universal constant $C > 0$ (see [203]). Thus,

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) = \mathbb{E}\left[\left\|\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right\|_{HS}^2\right] \leq 2\mathbb{E}\left[\left\|\Gamma_t^2\right\|_{HS}^2\right] \leq C\frac{d^4}{(1-t)^4}.$$

□

Lemma 1.32. *Suppose that μ is log concave and isotropic, then there exists a universal constant $1 > c > 0$ such that*

1. *For any, $t \in [0, \frac{c}{d^2}]$, $\sigma_t \geq \frac{1}{2}$.*

2. *For any, $t \in [\frac{c}{d^2}, 1]$, $\sigma_t \geq \frac{c}{td^2}$.*

Proof. By Lemma 1.26, we have

$$\frac{d}{dt}\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] = -\mathbb{E}\left[\Gamma_t^2\right] \overset{(1.17)}{=} -\frac{\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right)^2\right]}{(1-t)^2}.$$

Moreover, by convexity,

$$\mathbb{E}\left[\mathrm{Cov}\left(Y_1|\mathcal{F}_t\right)^2\right] \preceq \mathbb{E}\left[\mathbb{E}\left[Y_1^{\otimes 2}|\mathcal{F}_t\right]^2\right] \preceq \mathbb{E}\left[\|Y_1\|^4\right] \mathrm{I}_d.$$

It is known (see [203]) then when μ is log concave and isotropic there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[\|Y_1\|^4\right] \leq Cd^2.$$

Consequently, $\frac{d}{dt}\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] \succeq -\frac{Cd^2}{(1-t)^2}\mathrm{I}_d$, and since $\mathrm{Cov}(Y_1|\mathcal{F}_0) = \mathrm{I}_d$,

$$\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] \succeq \left(1 - Cd^2\int_0^t \frac{1}{(1-s)^2}ds\right)\mathrm{I}_d = \left(1 - \frac{Cd^2 t}{1-t}\right)\mathrm{I}_d.$$

By increasing the value of C, we may legitimately assume that $\frac{1}{Cd^2} \leq 1$, thus for any $t \in [0, \frac{1}{3Cd^2}]$ we get that

$$\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] \succeq \frac{1}{2}\mathrm{I}_d,$$

which implies $\sigma_t \geq \frac{1}{2}$ and completes the first part of the lemma. In order to prove the second

part, we first write

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right] = \frac{d}{dt}\frac{\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right]}{1-t} \overset{\text{(Lemma 1.26)}}{=} \frac{\mathbb{E}\left[\mathrm{Cov}(Y_1|\mathcal{F}_t)\right] - (1-t)\mathbb{E}\left[\Gamma_t^2\right]}{(1-t)^2} = \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t}. \tag{1.26}$$

Since, by Lemma 1.30, $\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d$, we have the bound

$$\frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t} \succeq \frac{1-\frac{1}{t}}{1-t}\mathbb{E}\left[\Gamma_t\right] = -\frac{1}{t}\mathbb{E}\left[\Gamma_t\right].$$

Now, consider the differential equation $f'(t) = \frac{-f(t)}{t}$, $f\left(\frac{1}{3Cd^2}\right) = \frac{1}{2}$. Its unique solution is $f(t) = \frac{1}{6Cd^2t}$. Thus, Gromwall's inequality shows that $\sigma_t \geq \frac{1}{6Cd^2t}$, which concludes the proof. □

Proof of Theorem 1.6. Our objective is to bound from above the right hand side of Equation (1.25). As a consequence of Lemma 1.32, we have that for any $t \in [0,1)$,

$$\int_t^1 \sigma_s^{-2} ds \leq Cd^4(1-t),$$

for some universal constant $C > 0$. It follows that the integral in (1.25) admits the bound

$$\int_0^1 \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)^2\sigma_t^2}\left(\int_t^1 \sigma_s^{-2} ds\right) dt \leq Cd^4 \int_0^1 \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)\sigma_t^2} dt.$$

Next, there exists a universal constant $C' > 0$ such that

$$Cd^4 \int_0^{cd^{-2}} \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)\sigma_t^2} dt \leq C' \int_0^{cd^{-2}} \frac{d^8}{(1-t)^5} dt \leq C'd^8,$$

where we have used the second bound of Lemma 1.31 and the first bound of Lemma 1.32. Also, by applying the second bound of Lemma 1.32 when $t \in [cd^{-2}, d^{-1}]$ we get

$$Cd^4 \int_{cd^{-2}}^{d^{-1}} \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)\sigma_t^2} dt \leq C' \int_{cd^{-2}}^{d^{-1}} \frac{d^{12}t^2}{(1-t)^5} dt \leq C'd^9.$$

Finally, when $t > d^{-1}$, we have

$$\begin{aligned} Cd^4 \int_{d^{-1}}^{1} \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)\sigma_t^2} dt &\leq C' d^8 \int_{d^{-1}}^{1} \frac{t^2 \mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{1-t} dt \\ &\leq 2C' d^9 \int_{d^{-1}}^{1} \left(\frac{1}{t^2} + \mathbb{E}\left[\|v_t\|^2\right]\right) dt \\ &\overset{(1.18)}{\leq} 4C' d^{10}(1 + \mathrm{Ent}(Y_1||G)), \end{aligned}$$

where the first inequality uses Lemma 1.32 and the second one uses Lemma 1.31. This establishes

$$\mathrm{Ent}(S_n||G) \leq \frac{Cd^{10}(1 + \mathrm{Ent}(Y_1||G))}{n}.$$

□

Finally, we derive an improved bound for the case of 1-uniformly log concave measures, based on the following estimates.

Lemma 1.33. *Suppose that μ is 1-uniformly log concave, then for every $t \in [0, 1)$*

1. $\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) \leq 2(1-t)\left(d - \mathrm{Tr}\left(\Sigma\right) + \mathbb{E}\left[\|v_t\|^2\right]\right).$
2. $\sigma_t \geq \sigma_0$.

Proof. By Lemma 1.30, we have that $\Gamma_t \preceq \mathrm{I}_d$ almost surely. Using this together with the identity given by Lemma 1.28, and proceeding in similar fashion to Lemma 1.31 we obtain

$$\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t^2\right]\right) \geq \frac{1}{d}\mathrm{Tr}\left(\mathbb{E}\left[\Gamma_t\right]\right)^2 \geq d - 2(1-t)\left(d - \mathrm{Tr}\left(\Sigma\right) + \mathbb{E}\left[\|v_t\|^2\right]\right),$$

and

$$\begin{aligned} \mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right]\right) &\leq \mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^2 - \mathrm{I}_d\right)^2\right]\right) \leq \mathrm{Tr}\left(\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^2\right]\right) \\ &\leq 2(1-t)\left(d - \mathrm{Tr}\left(\Sigma\right) + \mathbb{E}\left[\|v_t\|^2\right]\right). \end{aligned}$$

Also, recalling (1.26) and since $\Gamma_t \preceq \mathrm{I}_d$ we get

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right] = \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t} \geq 0,$$

which shows that σ_t is bounded from below by a non-decreasing function and so $\sigma_t \geq \sigma_0$ which is the minimal eigenvalue of Σ. □

Proof of Theorem 1.7. Plugging the bounds given in Lemma 1.33 into Equation (1.25) yields

$$\begin{aligned}\mathrm{Ent}(S_n||G) &\leq \frac{1}{n}\int_0^1 \frac{\mathbb{E}\left[\mathrm{Tr}\left(\left(\Gamma_t^2 - \mathbb{E}\left[\Gamma_t^2\right]\right)^2\right)\right]}{(1-t)^2\sigma_t^2}\left(\int_t^1 \sigma_s^{-2} ds\right) dt \\ &\leq \frac{2\left(d + \int_0^1 \mathbb{E}\left[\|v_t\|^2\right] dt\right)}{\sigma_0^4 n} \overset{(1.18)}{=} \frac{2\left(d + 2\mathrm{Ent}\left(X||\gamma\right)\right)}{\sigma_0^4 n},\end{aligned}$$

which completes the proof. □

2

A Central Limit Theorem in Stein's Distance for Generalized Wishart Matrices and Higher Order Tensors

2.1 Introduction

Let μ be an isotropic probability measure on $\mathbb{R}^n$. For $2 \leq p \in \mathbb{N}$, we consider the following tensor analogue of the Wishart matrix,

$$\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \left(X_i^{\odot p} - \mathbb{E}\left[X_i^{\odot p} \right] \right),$$

where $X_i \sim \mu$ are i.i.d. and $X_i^{\odot p}$ stands for the symmetric p'th tensor power of X_i. We denote the law of this random tensor by $W_{n,d}^p(\mu)$. Such distributions arise naturally as the sample moment tensor of the measure μ, in which case d serves as the sample size. For reasons soon to become apparent, we will sometimes refer to such tensors as *Wishart tensors*.

When $p = 2$, $W_{n,d}^2(\mu)$ is the sample covariance of μ. If $\mathbb{X}$ is an $n \times d$ matrix with columns independently distributed as μ, then $W_{n,d}^2(\mu)$ may also be realized as the upper triangular part

of the matrix,

$$\frac{\mathbb{X}\mathbb{X}^T - d\mathrm{Id}}{\sqrt{d}}. \tag{2.1}$$

Hence, $W_{n,d}^2(\mu)$ has the law of a Wishart matrix. These matrices have recently been studied in the context of random geometric graphs ([53, 56, 58, 102]).

For fixed p, n, according to the central limit theorem (CLT), as $d \to \infty$, $W_{n,d}^p(\mu)$ approaches a normal law. The aim of this chapter is to study the *high-dimensional regime* of the problem, where we allow the dimension n to scale with the sample size d. Specifically, we investigate possible conditions on n and d for the CLT to hold. Observe that this problem may be reformulated as a question about the rate of convergence in the high-dimensional CLT, for the special case of Wishart tensors.

Our starting point is the paper [58], which obtained an optimal bound when $p = 2$, for log-concave product measures. Remark that when μ is a product measure, the entries of the matrix $\mathbb{X}$ in (2.1) are all independent. The proof [58] was information-theoretic and made use of the chain rule for relative entropy to account for the low-dimensional structure of $W_{n,d}^2(\mu)$. For now, we denote $\widetilde{W}_{n,d}^2(\mu)$ to be the same law as $W_{n,d}^2(\mu)$, but with the diagonal elements removed (see below for a precise definition).

Theorem 2.1 ([58, Theorem 1]). *Let μ be a log-concave product measure on $\mathbb{R}^n$ and let γ denote the standard Gaussian in $\mathbb{R}^{\binom{n}{2}}$. Then,*

1. *If $n^3 \ll d$ then* $\mathrm{Ent}\left(\widetilde{W}_{n,d}^2(\mu)||\gamma\right) \xrightarrow{n\to\infty} 0$.
2. *If $n^3 \gg d$, then $W_{n,d}^2(\mu)$ remains bounded away from any Gaussian law.*

Here, Ent *stands for relative entropy (see Section 2.1.1 for the definition).*

Thus, for log-concave product measures there is a sharp condition for the CLT to hold. Our results, which we now summarize, generalize Point 1 of Theorem 2.1 in several directions and are aimed to answer questions which were raised in [58].

- We show that it is not necessary for μ to have a product structure. So, in particular, the matrix $\mathbb{X}$ in (2.1) may admit some dependence between its entries.
- If μ is a product measure, we relax the log-concavity assumption and show the same result holds for a much larger class of product measures.
- The above results extend to the case $p > 2$, and we propose the new threshold $n^{2p-1} \ll d$.
- We show that Theorem 2.1 is still true when we take the full symmetric tensor $W_{n,d}^2(\mu)$ and include the diagonal.

Naively, since $W_{n,d}^p(\mu)$ can be realized as a sum of i.i.d. random vectors, one should be able to employ standard techniques of Stein's method (such as exchangeable pairs [69], as proposed in [58]) in order to deduce some bounds. However, it turns out that the obtained bounds are sub-optimal. The reason for this sub-optimality is that, while $X^{\odot p}$ is a random vector in a high-dimensional space, its randomness comes from the lower-dimensional $\mathbb{R}^n$. So, at least on the intuitive level, one must exploit the low-dimensional structure of the random tensor in order to produce better bounds. Our method is is based on a novel application of Stein's method which is particularly adapted to this situation and may be of use in other, similar, settings.

2.1.1 Definitions and notations

Probability measures: A measure is said to be unconditional, if its density satisfies

$$\frac{d\mu}{dx}(\pm x_1, ..., \pm x_n) = \frac{d\mu}{dx}(|x_1|, ..., |x_n|),$$

where in the left side of the equality we consider all possible sign patterns. Note that, in particular, if $X = (X_{(1)}, ..., X_{(n)})$ is isotropic and unconditional, then , for any choice of distinct indices $j_1, ..., j_k$ and powers $n_2, ..., n_k$,

$$\mathbb{E}\left[X_{(j_1)} \cdot X_{(j_2)}^{n_2} \cdot X_{(j_3)}^{n_3} \cdot ... \cdot X_{(j_k)}^{n_k}\right] = 0. \tag{2.2}$$

Finally, if $\varphi : \mathbb{R}^n \to \mathbb{R}^N$ for some $N \geq 0$, we denote $\varphi_* \mu$, to be the push-forward of μ by φ.

Tensor spaces: Fix $\{e_j\}_{j=1}^n$ to be the standard basis in $\mathbb{R}^n$. We identify the tensor space $(\mathbb{R}^n)^{\otimes p}$ with $\mathbb{R}^{n^p}$ where the base is given by

$$\{e_{j_1} e_{j_2} ... e_{j_p} | 1 \leq j_1, j_2, ..., j_p \leq n\}.$$

Under this identification, we may consider the symmetric tensor space $\mathrm{Sym}^p(\mathbb{R}^n) \subset (\mathbb{R}^n)^{\otimes p}$ with basis

$$\{e_{j_1} e_{j_2} ... e_{j_p} | 1 \leq j_1 \leq j_2 ... \leq j_p \leq n\}.$$

We will also be interested in the subspace of principal tensors, $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n) \subset \mathrm{Sym}^p(\mathbb{R}^n)$, spanned by the basis elements

$$\{e_{j_1} e_{j_2} ... e_{j_p} | 1 \leq j_1 < j_2 ... < j_p \leq n\}.$$

Our main result will deal with the marginal of $W_{n,d}^p(\mu)$ on the subspace $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$. We denote this marginal law by $\widetilde{W}_{n,d}^p(\mu)$. Put differently, if $X_i = (X_{i,1}, ..., X_{i,n})$ are i.i.d. random vectors

with law μ. Then, $\widetilde{W}^p_{n,d}(\mu)$ is the law of a random vector in $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$ with entries

$$\left(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} X_{i,j_1}\cdot X_{i,j_2}\cdot\dots\cdot X_{i,j_p}\right)_{1\le j_1<\dots<j_p\le n}.$$

2.1.2 Main results

Our main contribution is a new approach, detailed in Section 2.3, to Stein's method, which allows to capitalize on the fact that a high-dimensional random vector may have some latent low-dimensional structure. Thus, it is particularly well suited to study the CLT for $W^p_{n,d}(\mu)$. Using this approach, we obtain the following threshold for the CLT: Suppose that μ is a "nice" measure. Then, if $n^{2p-1} \ll d$, $W^p_{n,d}(\mu)$ is approximately Gaussian, as d tends to infinity.

We now state several results which are obtained using our method. The first result shows that, under some assumptions, the matrix $\mathbb{X}$ in (2.1), can admit some dependencies, even when considering higher order tensors.

Theorem 2.2. *Let μ be an isotropic L-uniformly log-concave measure on $\mathbb{R}^n$ which is also unconditional. Denote $\Sigma^{-\frac{1}{2}} = \sqrt{\widetilde{\Sigma}_p(\mu)^{-1}}$, where $\widetilde{\Sigma}_p(\mu)$ is the covariance matrix of $\widetilde{W}^p_{n,d}(\mu)$. Then, there exists a constant C_p, depending only on p, such that*

$$\mathcal{W}_2^2\left(\Sigma_*^{-\frac{1}{2}}\widetilde{W}^p_{n,d}(\mu),\gamma\right) \le \frac{C_p}{L^4}\frac{n^{2p-1}}{d},$$

where $\Sigma_*^{-\frac{1}{2}}\widetilde{W}^p_{n,d}(\mu)$ is the push-forward by the linear operator $\Sigma^{-\frac{1}{2}}$.

An important remark, which applies to the coming results as well, is that the bounds are formulated with respect to the quadratic Wasserstein distance. However, as will become evident from the proof, the bounds actually hold with a stronger notion of distance: namely, Stein's discrepancy (see Section 2.2 for the definition). We have decided to state our results with the more familiar Wasserstein distance to ease the presentation. Our next result is a direct extension of Theorem 2.1, as it both applies to a larger class of product measures and to $p > 2$.

Theorem 2.3. *Let μ be an isotropic product measure on $\mathbb{R}^n$, with independent coordinates. Then, there exists a constant $C_p > 0$, depending only on p, such that*

1. *If μ is log-concave, then*

$$\mathcal{W}_2^2\left(\widetilde{W}^p_{n,d}(\mu),\gamma\right) \le C_p\frac{n^{2p-1}}{d}\log(n)^2.$$

2. *If each coordinate marginal of μ satisfies the L_1-Poincaré inequality (see Section 2.2.2)*

with constant $c > 0$, then

$$\mathcal{W}_2^2\left(\widetilde{W}_{n,d}^p(\mu), \gamma\right) \leq C_p \frac{1}{c^{2p+2}} \frac{n^{2p-1}}{d} \log(n)^4.$$

3. *If there exists a uni-variate polynomial Q of degree k, such that each coordinate marginal of μ has the same law as the push-forward measure $Q_*\gamma_1$, then*

$$\mathcal{W}_2^2\left(\widetilde{W}_{n,d}^p(\mu), \gamma\right) \leq C_{Q,p} \frac{n^{2p-1}}{d} \log(n)^{2(k-1)},$$

where $C_{Q,p} > 0$ may depend both on p and the polynomial Q.

Observe that, when μ is an isotropic product measure, then $\widetilde{W}_{n,d}^p(\mu)$ is also isotropic (when considered as a random vector in $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$), which explains why the matrix $\Sigma^{-\frac{1}{2}}$ does not appear in Theorem 2.3. Our last result is an extension to Theorem 2.3 which shows that, sometimes, we may consider subspaces of $(\mathbb{R}^n)^{\otimes p}$ which are strictly larger than $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$. We specialize to the case $p = 2$, and show that one may consider the full symmetric matrix $W_{n,d}^2(\mu)$.

Theorem 2.4. *Let μ be an isotropic log-concave measure on $\mathbb{R}^n$. Assume that μ is a product measure with independent coordinates and denote $\Sigma^{-\frac{1}{2}} = \sqrt{\Sigma_2(\mu)^{-1}}$, where $\Sigma_2(\mu)$ is the covariance matrix of $W_{n,d}^2(\mu)$. Then, there exists a universal constant $C > 0$ such that*

$$\mathcal{W}_2^2\left(\Sigma_*^{-\frac{1}{2}} W_{n,d}^2(\mu), \gamma\right) \leq C \frac{n^3}{d} \log(n)^2.$$

2.1.3 Related work

The study of normal approximations for high-dimensional Wishart tensors was initiated in [56] (see [143] as well, for an independent result), which dealt with the case of $\widetilde{W}_{n,d}^2(\gamma)$. The authors were interested in detecting latent geometry in random geometric graphs. The main result of [56] was a particular case of Theorem 2.1, which gave a sharp threshold for detection in the total variation distance. The Gaussian setting was studied further in [207], where a smooth transition between the regimes $n^3 \gg d$ and $n^3 \ll d$, was shown to hold. The proof of such results was facilitated by the fact that $\widetilde{W}_{n,d}^2(\gamma)$ has a tractable density with respect to the Lebesgue measure. This is not the case in general though.

In a follow-up ([58]), as discussed above, the results of [56] were expanded to the relative entropy distance and to Wishart tensors $\widetilde{W}_{n,d}^2(\mu)$, where μ is a log-concave product measure. Specifically, it was shown that one may consider relative entropy in the formulation of Theorem 2.1, and that

$$\mathrm{Ent}\left(\widetilde{W}_{n,d}^2(\mu) || \gamma\right) \leq C \frac{n^3 \log(d)^2}{d},$$

for a universal constant $C > 0$. The main idea of the proof was a clever use of the chain rule for relative entropy along with ideas adapted from the one-dimensional entropic central limit theorem proven in [17]. We do note the this result is not directly comparable to our results. As remarked, our results hold in Stein's discrepancy. In general, Stein's discrepancy and relative entropy are not comparable. However, one may bound the relative entropy by the discrepancy, in some cases. One such case, is when the measure has a finite Fisher information. $\widetilde{W}^2_{n,d}(\gamma)$ is an example of such a measure.

The question of handling dependencies between the entries of the matrix $\mathbb{X}$ in (2.1) was also tackled in [197]. The authors considered the case where the rows of $\mathbb{X}$ are i.i.d. copies of a Gaussian measure whose covariance is a symmetric Toeplitz matrix. The paper employed Stein's method in a clever way, which seems to be somewhat different from our approach.

For another direction of handling dependencies, note that if the rows of $\mathbb{X}$ are independent, but not isotropic, Gaussian vectors, then by applying an orthogonal transformation to the rows we can obtain a matrix with independent entries which have different variances. Such measures were studied in [102]. Specifically if $\alpha = \{\alpha_i\}_{i=1}^d \subset \mathbb{R}^+$, with $\sum \alpha_i^2 = 1$ and $X_i \sim \gamma$ are independent, then the paper introduced $W^2_{n,\alpha}(\gamma)$, as the law of,

$$\sum \alpha_i \left(X_i^{\odot 2} - \mathbb{E}\left[X_i^{\odot 2} \right] \right).$$

The following variant of Theorem 2.1 was given:

$$\mathrm{Ent}\left(\widetilde{W}^2_{n,\alpha}(\gamma) || \gamma_{\binom{n}{2}} \right) \leq C n^3 \sum \alpha_i^4. \tag{2.3}$$

When $\alpha_i \equiv \frac{1}{\sqrt{d}}$, this recovers the previous known result. We mention here that our method applies to non-homogeneous sums as well, with the same dependence on α. See Section 2.9 for a comparison with the above result, as well as the one in [197].

The authors of [197] also dealt with Wishart tensors, when the underlying measure is the standard Gaussian. It was shown that for some constant C_p, which depends only on p,

$$\mathcal{W}_1\left(\widetilde{W}^p_{n,d}(\gamma), \gamma_{\binom{n}{p}} \right) \leq C_p \sqrt{\frac{n^{2p-1}}{d}}.$$

Thus, our results should also be seen as a direct generalization of this bound.

Wishart tensors have recently gained interest in the machine learning community (see [8, 220] for recent results and applications). To mention a few examples: In [144] the distribution of the maximal entry of $\widetilde{W}^p_{n,d}(\mu)$ is investigated. Using tools of random matrix theory, the spectrum of Wishart tensors is analyzed in [6], while [171] studies the central limit theorem

for spectral linear statistics of $W^p_{n,d}(\mu)$. Results of a different flavor are given in [239], where exponential concentration is studied for a class of random tensors.

2.1.4 Organization

The rest of this chapter is organized in the following way: In Section 2.2 we introduce some preliminaries from Stein's method and concentration of measure, which will be used in our proofs. In Section 2.3 we describe our method and present the necessary ideas with which we will prove our results. In particular, we will state Theorem 2.5, which will act as our main technical tool. In Section 2.4 we introduce a construction in Stein's theory which will be used in Section 2.5 to prove Theorem 2.5. Sections 2.6, 2.7 and 2.8 are then devoted to the proofs of Theorems 2.2, 2.3 and 2.4 respectively. Finally, in Section 2.9 we discuss a generalization of our results to non-homogeneous sums of the tensor powers.

2.2 Preliminaries

In this section we will describe our method and explain how to derive the stated results. We begin with some preliminaries on Stein's method.

2.2.1 Stein kernels

For convenience, we recall some of the definitions which appeared in the introduction to the thesis. We say that a measurable matrix valued map $\tau : \mathbb{R}^n \to \mathcal{M}_n(\mathbb{R})$ is a Stein kernel for μ, if the following equality holds, for all locally-Lipschitz test functions $f : \mathbb{R}^n \to \mathbb{R}^n$,

$$\int \langle x, f(x)\rangle d\mu(x) = \int \langle \tau(x), Df(x)\rangle_{HS} d\mu(x).$$

The Stein discrepancy is defined by,

$$S(\mu) := \inf_{\tau} \sqrt{\int \|\tau(x) - \mathrm{Id}\|^2_{HS}\, d\mu(x)}$$

The first property that we will care about here is the linear decay of discrepancy along the CLT (shown in (20))

$$S^2(S_d) \leq \frac{S^2(X)}{d}. \tag{2.4}$$

The second property is the relation to the quadratic Wasserstein distance ((23))

$$\mathcal{W}_2^2(\mu, \gamma) \leq S^2(\mu). \tag{2.5}$$

2.2.2 Smooth measures and concentration inequalities

Our result will mostly apply for measures which satisfy some regularity conditions. We detail here the main properties which will be relevant.

A measure μ is said to satisfy the L_1-Poincaré inequality with constant c if, for any differentiable function f with 0-median,

$$\int |f| \, d\mu \leq \frac{1}{c} \int \|\nabla f\|_2 \, d\mu.$$

Remark that the L_1-Poincaré inequality is equivalent, up to constants, to the Cheeger's isoperimetric inequality. That is, if μ satisfies the L_1-Poincaré inequality with constant $c > 0$, then for some other constant $c' > 0$, depending only on c, and for every measurable set B,

$$\mu^+ (\partial B) \geq c' \mu(B) \left(1 - \mu(B)\right).$$

where $\mu^+ (\partial B)$ is the outer boundary measure of B. Moreover, up to universal constants, the L_1-Poincaré inequality implies an L_2-Poincaré inequality. We refer the reader to [61] for further discussion of those facts.

For a given measure, the above conditions imply the existence sub-exponential tails. In particular, if μ is a centered measure which satisfies the L_1-Poincaré inequality (or L^2) with constant c, then, for any $m \geq 2$:

$$\mathbb{E}\left[\|X\|_2^m\right] \leq C_m \left(\frac{1}{c}\right)^{\frac{m}{2}} \mathbb{E}\left[\|X\|_2^2\right]^{\frac{m}{2}}, \tag{2.6}$$

where C_m depends only on m (see [181] for the connection between Poincaré inequalities and exponential concentration). All log-concave measures satisfy a Poincaré inequality, which implies that they have sub-exponential tails. In fact, a stronger statement holds for log-concave measures, and one may omit the dependence on the Poincaré constant in (2.6) (see [170, Theorem 5.22]). Thus, if X is a log-concave random vector,

$$\mathbb{E}\left[\|X\|_2^m\right] \leq C'_m \mathbb{E}\left[\|X\|_2^2\right]^{\frac{m}{2}}.$$

for some constant $C'_m > 0$, depending only on m.

2.3 The method

With the above results, the following theorem is our main tool, with which we may prove CLTs for $W_{n,d}^p(\mu)$.

Theorem 2.5. *Let $X \sim \mu$ be an isotropic random vector in $\mathbb{R}^n$ and let $G \sim \mathcal{N}(0, \mathrm{Id})$ stand for the standard Gaussian. Assume that $X \overset{\text{law}}{=} \varphi(G)$ for some $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, which is locally-Lipschitz and let $A : \mathrm{Sym}^p(\mathbb{R}^n) \to V$ be a linear transformation with $V \subset \mathrm{Sym}^p(\mathbb{R}^n)$, such that $A_* W^p_{n,d}(\mu)$ is isotropic. Then, for any $2 \leq p \in \mathbb{N}$,*

$$S^2\left(A_* W^p_{n,d}(\mu)\right) \leq 2\,\|A\|_{op}^2\, p^4 \cdot \frac{n}{d}\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]}\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]} + \frac{2n^p}{d}.$$

Some remarks are in order concerning the theorem. We first discuss the role of the matrix A. Recall that, in order to use the sub-additive property (2.4) of the Stein discrepancy, the random vectors need to be isotropic. This can be achieved via a normalizing linear transformation. However, the term $\|A\|_{op}$ which appears in the theorem tells us that if the covariance matrix of $W^p_{n,d}(\mu)$ has very small eigenvalues, the normalizing transformation might have adverse effects on the rate of convergence. To avoid this, we will sometimes project the vectors into a subspace, such as $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$, where the covariance matrix is easier to control. Thus, A should be thought of as a product of a projection matrix with the inverse of a covariance matrix on the projected space. For our applications we will make sure that, $\|A\|_{op} = O(1)$.

Concerning the other terms in the stated bound, there are two terms which we will need to control, $\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]}$ and $\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]}$. Since we are mainly interested in measures with sub-exponential tails, the first term will be of order n^{2p-2} and we will focus on the second term. Thus, in some sense, our bounds are meaningful mainly for measures which can be transported from the standard Gaussian with low distortion. Still, the class of measures which can be realized in such a way is rather large and contains many interesting examples.

A map ψ is said to transport G to X if $\psi(G)$ has the same law as X. To apply the result we must realize X by choosing an appropriate transport map. It is a classical fact ([52]) that, whenever μ has a finite second moment and is absolutely continuous with respect to γ, there is a distinguished map which transports G to X. Namely, the Brenier map which minimizes the quadratic distance,

$$\varphi_\mu := \inf_{\psi : \psi(G) \overset{\text{law}}{=} X} \mathbb{E}\left[\|G - \psi(G)\|_2^2\right].$$

The Brenier map has been studied extensively (see [62, 63, 78, 153] for example). Here, we will concern ourselves with cases where one can bound the derivative of φ_μ. The celebrated Caffareli's log-concave perturbation theorem ([64]) states that if μ is L-uniformly log-concave, then φ_μ is $\frac{1}{L}$-Lipschitz. In particular, φ is differentiable almost everywhere with

$$\|D\varphi_\mu(x)\|_{op} \leq \frac{1}{L}.$$

In this case we get

$$\sqrt{\mathbb{E}\left[\|D\varphi_\mu(G)\|_{op}^8\right]} \leq \frac{1}{L^4}. \tag{2.7}$$

Theorem 2.2 will follow from this bound. The reason why the theorem specializes to unconditional measures is that, in light of the dependence on the matrix A in Theorem 2.5, we need to have some control over the covariance structure of $\widetilde{W}_{n,d}^p(\mu)$. It turns out, that for unconditional log-concave measures the covariance of $\widetilde{W}_{n,d}^p(\mu)$ is well behaved. The result might be extended to uniformly log-concave measures which are not necessarily unconditional as long as we allow the bound to depend on the minimal eigenvalue of the covariance matrix of $\widetilde{W}_{n,d}^p(\mu)$.

There are more examples of measures for which the Brenier map admits bounds on the Lipschitz constant. In [79] it is shown that for measures μ which are bounded perturbation of the Gaussian, including radially symmetric measures, φ_μ is Lipschitz. The theorem may thus be applied to those measures as well.

One may also consider cases where the transport map is only locally-Lipschitz in a well behaved way. For example, consider the case where $X = (X_{(1)}, ..., X_{(n)}) \sim \mu$ is a product measure. That is, for $i \neq j$, $X_{(i)}$ is independent from $X_{(j)}$. Suppose that for $i = 1, \ldots, n$, there exist functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ such that, if G^1 is a standard Gaussian in $\mathbb{R}$, then $\varphi_i(G^1) \overset{\text{law}}{=} X_{(i)}$ and that φ has polynomial growth. Meaning, that for some constants $\alpha, \beta \geq 0$,

$$\varphi_i'(x) \leq \alpha(1 + |x|^\beta).$$

Since μ is a product measure, it follows that the map $\varphi = (\varphi_1, \ldots, \varphi_n)$ transports G to X and that,

$$\|D\varphi(x)\|_{op} \leq \alpha(1 + \|x\|_\infty^\beta).$$

Thus, for product measures, we can translate bounds on the derivative of one-dimensional transport maps into multivariate bounds involving the L_∞ norm. Theorem 2.3 will be proved by using these ideas and known estimates on the one-dimensional Brenier map (also known as monotone rearrangement). Results like Theorem 2.4 can then be proven by bounding the covariance matrix of $W_{n,d}^2(\mu)$. Indeed, this is the main ingredient in the proof of the theorem.

One may hope that Theorem 2.5 could be applied to general log-concave measures. However, this would be a highly non-trivial task. Indeed, if we wish to use Theorem 2.5 in order to verify the threshold $n^{2p-1} \ll d$, up to logarithmic terms, we should require that for any isotropic log-concave measure μ, there exists a map ψ_μ such that $\psi_\mu(G) \sim \mu$ and $\mathbb{E}\left[\|D\psi_\mu(G)\|_{op}^8\right] \leq \log(n)^\beta$, for some fixed $\beta \geq 0$. Then, by applying the Gaussian L_2-Poincaré inequality to the

function $\|\cdot\|_2$, we would get,

$$\begin{aligned}
\mathrm{Var}\left(\|\psi_\mu(G)\|_2\right) &\leq \mathbb{E}\left[\left\|D\left(\|\psi_\mu(G)\|_2\right)\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\frac{\psi_\mu(G)}{\|\psi_\mu(G)\|_2} D\psi_\mu(G)\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{\psi_\mu(G)}{\|\psi_\mu(G)\|_2}\right\|_2^2 \cdot \|D\psi_\mu(G)\|_{op}^2\right] \\
&= \mathbb{E}\left[\|D\psi_\mu(G)\|_{op}^2\right] \leq \log(n)^{\frac{\beta}{4}},
\end{aligned}$$

where the first equality is the chain rule and the second inequality is a consequence of considering $D\psi_\mu$ as an $n \times n$ matrix. This bound would, up to logarithmic factors, verify the *thin-shell* conjecture (see [13]), and, through the results of [97], also the *KLS* conjecture ([147]). These two conjectures are both famous long-standing open problems in convex geometry. Thus, while we believe that similar results should hold for general log-concave, it seems such claims would require new ideas.

Another evidence for the possible difficulty of determining optimal convergence rates for general log-concave vectors can be seen from the case $p = 1$. In the standard setting of the CLT, the best known convergence rates ([85, 106]), in quadratic Wasserstein distance, depend on the Poincaré constant of the isotropic log-concave measure. Bounding the Poincaré constant is precisely the object of the KLS conjecture. So, proving a convergence rate which does not depend on the specific log-concave measure seems to be intimately connected with the conjecture. This suggests the question might be a genuinely challenging one. On the brighter side, we remark that the recent breakthrough of Chen ([72]) towards the resolution of the KLS conjecture, can prove useful in establishing such bounds.

2.3.1 High-level idea

We now present the idea behind the proof of Theorem 2.5 and detail the main steps. We first provide an informal explanation of why standard techniques fail to give optimal bounds. We may treat $W_{n,d}^p(\mu)$ as a sum of independent random vectors and invoke Theorem 7 from [69] (similar results will encounter the same difficulty). So, if $X \sim \mu$, optimistically, the theorem will give,

$$\mathcal{W}_1(W_{n,d}^p(\mu), \gamma) \leq \frac{\mathbb{E}\left[\|X^{\odot p}\|_2^3\right]}{\sqrt{d}},$$

where we take the Euclidean norm of $X^{\odot p}$ when considered as a vector in $\mathrm{Sym}^p(\mathbb{R}^n)$. Since $\dim\left(\mathrm{Sym}^p(\mathbb{R}^n)\right) \simeq n^p$, and we expect each coordinate of $X^{\odot p}$ to have magnitude, roughly

$O(1)$, Jensen's inequality gives:

$$\mathbb{E}\left[\left\|X^{\odot p}\right\|_2^3\right] \geq \mathbb{E}\left[\left\|X^{\odot p}\right\|_2^2\right]^{\frac{3}{2}} \gtrsim \dim\left(\mathrm{Sym}^p(\mathbb{R}^n)\right)^{\frac{3}{2}} \gtrsim n^{\frac{3p}{2}}.$$

This is worse than the bound $\sqrt{n^{2p-1}}$, achieved by Theorem 2.5.

The high-level plan of our proof is to use the fact that $X^{\odot p}$ has some low-dimensional structure. We will construct a map which transports the standard Gaussian G, from the lower dimensional space $\mathbb{R}^n$ into the law of $X^{\odot p}$ in the higher dimensional space $\mathrm{Sym}^p(\mathbb{R}^n)$. In some sense, the role of this transport map is to preserve the low-dimensional randomness coming from $\mathbb{R}^n$. The map can be constructed in two steps, first use a transport map φ, such that $\varphi(G) \overset{\text{law}}{=} X$, and then take its tensor power $\varphi(G)^{\odot p}$. We will use this map in order to construct a Stein kernel and show that tame tails of the map's derivative translate into small norms for the Stein kernel.

2.4 From transport maps to Stein kernels

We now explain how to construct a Stein kernel from a given transport map. For the rest of this section let ν be a measure on $\mathbb{R}^N$ and $Y \sim \nu$. Recall the definition of a Stein kernel; A matrix-valued map, $\tau : \mathbb{R}^N \to \mathcal{M}_N(\mathbb{R})$, is a Stein kernel for ν, if for every smooth $f : \mathbb{R}^N \to \mathbb{R}^N$,

$$\mathbb{E}\left[\langle Y, f(Y)\rangle\right] = \mathbb{E}\left[\langle \tau(Y), Df(Y)\rangle_{HS}\right].$$

Our construction is based on differential operators which arise naturally when preforming analysis in Gaussian spaces. We incorporate into this construction the idea of considering transport measures between spaces of different dimensions. For completeness, we give all of the necessary details, but see [138, 194] for a rigorous treatment.

2.4.1 Analysis in finite dimensional Gauss space

We let γ stand for the standard Gaussian measure in $\mathbb{R}^N$ and consider the Sobolev subspace of weakly differentiable functions,

$$W^{1,2}(\gamma) := \{f \in L^2(\gamma) | f \text{ is weakly differntiable, and } \mathbb{E}_\gamma \|Df\|_2^2 < \infty\}.$$

where $D : W^{1,2}(\gamma) \to L^2(\gamma, \mathbb{R}^N)$ is the natural (weak) derivative operator. We will mainly care about the fact that locally-Lipschitz functions are weakly differentiable the reader is referred to the second chapter of [254] for the necessary background on Sobolev spaces.

The divergence δ is defined to be the formal adjoint of D, so that for $g : \mathbb{R}^N \to \mathbb{R}^N$,

$$\mathbb{E}_\gamma\left[\langle Df, g\rangle\right] = \mathbb{E}_\gamma\left[f\delta g\right].$$

δ is given explicitly by the relation

$$\delta g(x) = \langle x, g(x)\rangle - \mathrm{div}(g(x)),$$

where $\mathrm{div}(g(x)) = \sum_{i=1}^{N} \frac{\partial g_i}{\partial x_i}(x)$.

The Ornstein-Uhlenbeck (OU) operator is now defined by $L := -\delta \circ D$. On functions, L operates as $Lf(x) = -xDf(x) + \Delta f(x)$. The operator L also serves as the infinitesimal generator of the OU semi-group ([196, Proposition 1.3.6]). That is,

$$L = \frac{d}{dt} P_t \Big|_{t=0},$$

where

$$P_t f(x) := \mathbb{E}_{N\sim\gamma}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}N)\right]. \tag{2.8}$$

The following fact, which may be proved by the Hermite decomposition of $L^2(\gamma)$, will be useful; There exists an operator, denoted L^{-1} such that $LL^{-1}f = f$. In particular, on the subspace of functions whose Gaussian expectation vanishes, L^{-1} is the inverse of L ([194, Proposition 2.8.11]).

We now introduce a general construction for Stein kernels. By a slight abuse of notation, even when working in different dimensions, we will refer to the above differential operators as the same, as well as extending them to act of vector and matrix valued functions. Note that, in particular, if f is a vector-valued function and g is matrix-valued of compatible dimensions, then,

$$\mathbb{E}_\gamma\left[\langle Df, g\rangle_{HS}\right] = \mathbb{E}_\gamma\left[\langle f, \delta g\rangle\right].$$

Lemma 2.6. *Let γ_m be the standard Gaussian measure on $\mathbb{R}^m$ and let $\varphi : \mathbb{R}^m \to \mathbb{R}^N$ be weakly differentiable. Set $\nu = \varphi_*\gamma_m$ and suppose that $\int_{\mathbb{R}^N} x d\nu = 0$. Then, if the following expectation is finite for ν-almost every $x \in \mathbb{R}^N$,*

$$\tau_\varphi(x) := \mathbb{E}_{y\sim\gamma_m}\left[(-DL^{-1})\varphi(y)(D\varphi(y))^T | \varphi(y) = x\right],$$

is a Stein kernel of ν.

Proof. Let $f : \mathbb{R}^N \to \mathbb{R}^N$ be a smooth function and set $Y \sim \nu, G \sim \gamma_m$. Our goal is to show

$$\mathbb{E}\left[\langle Df(Y), \tau_\varphi(Y)\rangle_{HS}\right] = \mathbb{E}\left[\langle f(Y), Y\rangle\right].$$

Before turning to the calculations let us make explicit the dimensions of the objects which will be involved. Df is an $N \times N$ matrix, while $D\varphi$ is an $N \times m$ matrix. Since $DL^{-1}\varphi$ is also an

$N \times m$ matrix, it holds that τ_φ is an $N \times N$ matrix, as required. Now,

$$\begin{aligned}
\mathbb{E}\left[\langle Df(Y), \tau_\varphi(Y)\rangle_{HS}\right] &= \mathbb{E}\left[\langle Df(Y), \mathbb{E}\left[(-DL^{-1})\varphi(G)(D\varphi(G))^T | \varphi(G) = Y\right]\rangle_{HS}\right] \\
&= \mathbb{E}\left[\langle Df(\varphi(G))D\varphi(G), (-DL^{-1})\varphi(G)\rangle_{HS}\right] \\
&= \mathbb{E}\left[\langle D(f\circ\varphi(G)), (-DL^{-1})\varphi(G)\rangle_{HS}\right] && \text{(Chain rule)} \\
&= \mathbb{E}\left[\langle f\circ\varphi(G), (-\delta DL^{-1})\varphi(G)\rangle\right] && (D \text{ is adjoint to } \delta) \\
&= \mathbb{E}\left[\langle f\circ\varphi(G), LL^{-1}\varphi(G)\rangle\right] && L = -\delta D \\
&= \mathbb{E}\left[\langle f\circ\varphi(G), \varphi(G)\rangle\right] && \mathbb{E}[\varphi(G)] = 0 \\
&= \mathbb{E}\left[\langle f(Y), Y\rangle\right]. && \varphi_*\gamma_m = \nu
\end{aligned}$$

In the first line, the inner product is taken in the space of $N \times N$ matrices and in the next two lines, in the space of $N \times m$ matrices. Also, note that in the penultimate equality the fact $\mathbb{E}\left[\varphi(G)\right]$ was important for the cancellation of LL^{-1}. □

The above formula suggests that one might control the the kernel τ_φ by controlling the gradient of the transport map, φ. This will be the main step in proving Theorem 2.5. The following formula from [194, Proposition 29.3] will be useful:

$$-DL^{-1}\varphi = \int_0^\infty e^{-t} P_t D\varphi dt. \tag{2.9}$$

We thus have the corollary:

Corollary 2.7. *With the same notations as in Lemma 2.6,*

$$\tau_\varphi(x) = \int_0^\infty e^{-t}\mathbb{E}_{y\sim\gamma_m}\left[P_t D\varphi(y)\left(D\varphi(y)\right)^T | \varphi(y) = x\right] dt.$$

We remark that the construction is based on ideas which have appeared implicitly in the literature, at least as far as [68] (see [194] for a more modern point of view). Our main novelty lies in interpreting the transport map, used in the construction, as an embedding from a low-dimensional space. Other constructions of Stein kernels use different ideas, such as Malliavin calculus ([196]), other notions of transport problems ([112]) or calculus of variation ([83]). However, as will become clear in the next section, our construction seems particularly well adapted to the current problem, since it is well behaved with respect to compositions. That is, if ψ, φ are two compatible maps and τ_φ is 'close' to the identity, then as long ψ is not too wild, the same can be said about $\tau_{\psi\circ\varphi}$. In our setting, one should think about φ as the transport map and $\psi(v) := v^{\otimes p}$. It is an interesting question whether other constructions for Stein kernels could be used in a similar way.

As a warm up we present a simple case in which we can show that the Stein kernel obtained from the construction is bounded almost surely.

Lemma 2.8. *Let the notations of Lemma 2.6 prevail and suppose that* $\|D\varphi(x)\|_{op} \leq 1$ *almost surely. Then*

$$\|\tau_\varphi(x)\|_{op} \leq 1,$$

almost surely.

Proof. From the representation (2.8) and by Jensen's inequality, P_t is a contraction. That is, for any function h,

$$\mathbb{E}_{y\sim\gamma_m}\left[P_t(h(y))h(y)\right] \leq \mathbb{E}_{y\sim\gamma_m}\left[h(y)^2\right]. \tag{2.10}$$

So, from Corollary 2.7 and since $\|D\varphi(x)\|_{op} \leq 1$, we get,

$$\begin{aligned}
\|\tau_\varphi(x)\|_{op} &\leq \int_0^\infty e^{-t}\mathbb{E}_{y\sim\gamma_m}\left[\|P_t D\varphi(y)\left(D\varphi(y)\right)\|_{op}\,|\varphi(y)=x\right]dt \\
&\leq \int_0^\infty e^{-t}\mathbb{E}_{y\sim\gamma_m}\left[P_t\left(\|D\varphi(y)\|_{op}\right)\|D\varphi(y)\|_{op}\,|\varphi(y)=x\right]dt \\
&\leq \int_0^\infty e^{-t}\mathbb{E}_{y\sim\gamma_m}\left[\|D\varphi(y)\|_{op}^2\,|\varphi(y)=x\right]dt \leq 1.
\end{aligned}$$

The first inequality follows from Jensen's inequality. The second inequality uses the fact that the matrix norm is sub-multiplicative combined with Jensen's inequality for P_t. The last inequality is the contractive property (2.10) and the a-priori bound on $\|D\varphi\|_{op}$. □

2.5 Proof of Theorem 2.5

Let $A : (\mathbb{R}^n)^{\otimes p} \to V$ be any linear transformation such that $A\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)$ is isotropic, and let τ be a Stein kernel for $X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$. In light of (17), we know that

$$\begin{aligned}
S^2\left(A\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)\right) &\leq \mathbb{E}\left[\left\|A\tau\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)A^T - \mathrm{Id}\right\|_{HS}^2\right] \\
&\leq 2\mathbb{E}\left[\left\|A\tau\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)A^T\right\|_{HS}^2\right] + 2\|\mathrm{Id}\|_{HS}^2 \\
&\leq 2\|A\|_{op}^2\,\mathbb{E}\left[\left\|\tau\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)\right\|_{HS}^2\right] + 2\dim(V).
\end{aligned}$$

Thus, by combining the above with (2.4), Theorem 2.5 is directly implied by the following lemma.

Lemma 2.9. *Let X be an isotropic random vector in $\mathbb{R}^n$ and let $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable almost everywhere, such that $\varphi(G) \overset{\text{law}}{=} X$, where G is the standard Gaussian in $\mathbb{R}^n$. Then, for any integer $p \geq 2$, there exists a Stein kernel τ of $X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$, such that*

$$\mathbb{E}\left[\left\|\tau\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)\right\|_{HS}^2\right] \leq p^4 n \sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]} \sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]}.$$

Proof. Consider the map $u \to \varphi(u)^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$, which transports G to $X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$. For a vector $v \in \mathbb{R}^n$ we will denote $\tilde{v}^{\otimes p} := v^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$. Corollary 2.7 shows that the function defined by,

$$\tau(\tilde{v}^{\otimes p}) := \int_0^\infty e^{-t} \mathbb{E}\left[P_t\left(D(\varphi(G)^{\otimes p})\right) \cdot D(\varphi(G)^{\otimes p})^T | \varphi(G)^{\otimes p} = v^{\otimes p}\right] dt,$$

and which vanishes on tensors which are not of the form $\tilde{v}^{\otimes p}$, is a Stein kernel for $X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]$. Note that for any two matrices A, B,

$$\|AB\|_{HS} \leq \sqrt{\operatorname{rank}(A)}\, \|A\|_{op} \|B\|_{op}.$$

Thus, by applying Jensen's inequality several times, both for the integrals and for P_t, we have the bound

$$\begin{aligned}
\mathbb{E}\left[\left\|\tau(X^{\otimes p})\right\|_{HS}^2\right] &\leq \int_0^\infty e^{-t} \mathbb{E}\left[\left\|P_t\left(D(\varphi(G)^{\otimes p})\right) \cdot D(\varphi(G)^{\otimes p})^T\right\|_{HS}^2\right] dt \\
&\leq \int_0^\infty e^{-t} \mathbb{E}\left[\operatorname{rank}(D(\varphi(G)^{\otimes p}))\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2 P_t\left(\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2\right)\right] dt \\
&\leq \int_0^\infty n e^{-t} \mathbb{E}\left[\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2 P_t\left(\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2\right)\right] dt.
\end{aligned}$$

To see the why the last inequality is true, observe that $v \to \varphi(v)^{\otimes p}$ is a map from $\mathbb{R}^n$ to $\mathbb{R}^{n^p}$, hence $D(\varphi(G)^{\otimes p})$ is an $n^p \times n$ matrix, which leads to $\operatorname{rank}(D(\varphi(G)^{\otimes p})) \leq n$. We now use the fact that P_t is a contraction, as in (2.10), so that for every $t > 0$,

$$\mathbb{E}\left[\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2 P_t\left(\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^2\right)\right] \leq \mathbb{E}\left[\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^4\right].$$

So,

$$\mathbb{E}\left[\left\|\tau(X^{\otimes p})\right\|_{HS}^2\right] \leq n \mathbb{E}\left[\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^4\right].$$

We may realize the map $v \to \varphi(v)^{\otimes p}$ as the p-fold Kronecker power (the reader is referred to [206] for the relevant details concerning the Kronecker product) of $\varphi(v)$. With $\otimes$ now standing

for the Kronecker product, the following Leibniz law holds for the Jacobian:

$$D\left(\varphi(x)^{\otimes p}\right) = \sum_{i=1}^{p} \varphi(x)^{\otimes i-1} \otimes D\varphi(x) \otimes \varphi(x)^{\otimes p-i}.$$

The Kronecker product is multiplicative with respect to singular values, and for any $A_1, ..., A_p$ matrices,

$$\|A_1 \otimes ... \otimes A_p\|_{op} = \prod_{i=1}^{p} \|A_i\|_{op}\,.$$

We then have,

$$\begin{aligned}
\mathbb{E}\left[\left\|\tau(X^{\otimes p})\right\|_{HS}^2\right] &\leq n\mathbb{E}\left[\left\|D(\varphi(G)^{\otimes p})\right\|_{op}^4\right] \\
&= n\mathbb{E}\left[\left\|\sum_{i=1}^{p} \varphi(G)^{\otimes i-1} \otimes D\varphi(G) \otimes \varphi(G)^{\otimes p-i}\right\|_{op}^4\right] \\
&\leq n\mathbb{E}\left[\left(\sum_{i=1}^{p} \left\|\varphi(G)^{\otimes i-1} \otimes D\varphi(G) \otimes \varphi(G)^{\otimes p-i}\right\|_{op}\right)^4\right],
\end{aligned}$$

where the operator norm here is considered on the space of $n^p \times n$ matrices. The multiplicative property of the Kronecker product shows that for every $i = 1, \dots p$,

$$\left\|\varphi(G)^{\otimes i-1} \otimes D\varphi(G) \otimes \varphi(G)^{\otimes p-i}\right\|_{op} = \|\varphi(G)\|_2^{(p-1)} \|D\varphi(G)\|_{op}\,,$$

where now the operator norm is considered on the space of $n \times n$ matrices, and one can think about the Euclidean norm as the operator on the space of $1 \times n$ matrices. Thus,

$$\begin{aligned}
\mathbb{E}\left[\left\|\tau(X^{\otimes p})\right\|_{HS}^2\right] &\leq np^4\mathbb{E}\left[\|\varphi(G)\|_2^{4(p-1)} \|D\varphi(G)\|_{op}^4\right] \\
&\leq np^4\sqrt{\mathbb{E}\left[\|\varphi(G)\|_2^{8(p-1)}\right]}\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]} \\
&= np^4\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]}\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]},
\end{aligned}$$

where the last inequality is Cauchy-Schwartz. □

2.6 Unconditional log-concave measures; Proof of Theorem 2.2

We now wish to apply Theorem 2.5 to unconditional measures which are uniformly log-concave. In this case, we begin by showing that the covariance of $\widetilde{W}_{n,d}^p(\mu)$ is well conditioned.

Lemma 2.10. *Let μ be an unconditional log-concave measure on $\mathbb{R}^n$ and let $\widetilde{\Sigma}_p(\mu)$ denote the covariance matrix $\widetilde{W}^p_{n,d}(\mu)$. Then, there exists a constant $c_p > 0$, depending only on p, such that if $\tilde{\lambda}_{\min}$ stands for the smallest eigenvalue of $\widetilde{\Sigma}_p(\mu)$, then*

$$c_p \leq \tilde{\lambda}_{\min}.$$

Proof. We write $X = (X_{(1)}, ..., X_{(n)})$ and observe that $\Sigma^p(\mu)$ is diagonal. Indeed, if $1 \leq j_1 < j_2 < ... < j_p \leq n$ and $1 \leq j'_1 < j'_2 < ... < j'_p \leq n$ are two different sequences of indices then the covariance between $X_{(j_1)} \cdot ... \cdot X_{(j_p)}$ and $X_{(j'_1)} \cdot ... \cdot X_{(j'_p)}$ can be written as

$$\mathbb{E}\left[X_{(i_1)} \cdot X^{n_2}_{(i_2)} \cdot ... \cdot X^{n_k}_{(i_k)}\right],$$

where $p+1 \leq k \leq 2p$ and for every $i = 2, ..., k$, $n_i \in \{1, 2\}$. By (2.2), those terms vanish. Thus, in order to prove the lemma, it will suffice to show that for every set of distinct indices $j_1, ..., j_p$,

$$c_p \leq \mathbb{E}\left[\left(X_{(j_1)} \cdot ... \cdot X_{(j_p)}\right)^2\right],$$

for some constant $c_p > 0$, which depends only on p. If we consider the random isotropic and log-concave vector $(X_{j_1}, ..., X_{j_p})$ in $\mathbb{R}^p$, the existence of such a constant is assured by the fact that the density of this vector is uniformly bounded from below on some ball around the origin (see [170, Theorem 5.14]). □

We now prove Theorem 2.2.

Proof of Theorem 2.2. Set $P : (\mathbb{R}^n)^{\otimes p} \to \widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$ to be the linear projection operator and $\widetilde{\Sigma}_p(\mu)$ to be as in Lemma 2.10. Denote $A = \sqrt{\widetilde{\Sigma}_p^{-1}(\mu)}P$. Then, $A\left(X^{\otimes p} - \mathbb{E}\left[X^{\otimes p}\right]\right)$ is isotropic and has the same law as $\sqrt{\widetilde{\Sigma}_p^{-1}(\mu)}_* \widetilde{W}^p_{n,d}(\mu)$. The lemma implies

$$\|A\|^2_{op} \leq \frac{1}{c_p}.$$

As X is log-concave and isotropic, from (2.6), we get

$$\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]} \leq C_p n^{2p-2}.$$

X is also L-uniformly log-concave. So, as in (2.7), if φ_μ is the Brenier map, sending the standard Gaussian G to X,

$$\sqrt{\mathbb{E}\left[\|D\varphi_\mu(G)\|^8_{op}\right]} \leq \frac{1}{L^4}.$$

Combining the above displays with Theorem 2.5, gives the desired result. □

2.7 Product measures; Proof of Theorem 2.3

As mentioned in Section 2.3, when μ is a product measure, transport bounds on the marginals of μ may be used to construct a transport map φ whose derivative satisfies an L_∞ bound of the form,

$$\|D\varphi(x)\|_{op} \leq \alpha(1 + \|x\|_\infty^\beta). \tag{2.11}$$

for some $\alpha, \beta \geq 0$. Such conditions can be verified for a wide variety of product measures. For example, it holds, a fortiori, when the marginals of μ are polynomials of the standard one-dimensional Gaussian with bounded degrees. Furthermore, we mention now two more cases where the one-dimensional Brenier map is known to have tame growth. Those estimates will serve as the basis for the proof of Theorem 2.3.

In [83] it is shown that if μ is an isotropic log-concave measure in $\mathbb{R}$, and φ_μ is its associated Brenier map, then for some universal constant $C > 0$,

$$\varphi_\mu'(x) \leq C(1 + |x|). \tag{2.12}$$

If, instead, μ satisfies an L_1-Poincaré inequality with constant $c_\ell > 0$, then for another universal constant $C > 0$

$$\varphi_\mu'(x) \leq C\frac{1}{c_\ell}(1 + x^2).$$

Thus, for log-concave product measures (2.11) holds with $\beta = 1$ and for products of measures which satisfy the L_1-Poincaré inequality it holds with $\beta = 2$. Using these bounds, Theorem 2.3 becomes a consequence of the following lemma.

Lemma 2.11. *Let X be a random vector in $\mathbb{R}^n$ and let G stand for the standard Gaussian. Suppose that for some $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, $\varphi(G) \overset{\text{law}}{=} X$, and that φ is differentiable almost everywhere with*

$$\|D\varphi(x)\|_{op} \leq \alpha(1 + \|x\|_\infty^\beta),$$

for some $\beta, \alpha > 0$. Then, there exists a constant C_β, depending only on β, such that

$$\mathbb{E}\left[\|D\varphi(x)\|_{op}^8\right] \leq C_\beta \alpha^8 \log(n)^{4\beta}.$$

Proof. For any $x, y \geq 0$, the following elementary inequality holds,

$$(x + y)^8 \leq 2^7\left(x^8 + y^8\right).$$

Thus, we begin the proof with,

$$\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right] \leq \alpha^8\mathbb{E}\left[(1 + \|G\|_\infty^\beta)^8\right] \leq 256\alpha^8\mathbb{E}\left[\|G\|_\infty^{8\beta}\right].$$

Observe that the density function of $\|G\|_\infty$ is given by $n\psi \cdot \Phi^{n-1}$, where ψ is the density of the standard Gaussian in $\mathbb{R}$ and Φ is its cumulative distribution function. Since the product of log-concave functions is also log-concave, we deduce that $\|G\|_\infty$ is a log-concave random variable. From (2.6), we thus get

$$\mathbb{E}\left[\|G\|_\infty^{8\beta}\right] \leq C'_\beta \mathbb{E}\left[\|G\|_\infty^2\right]^{4\beta},$$

where C'_β depends only on β. The proof is concluded by applying known estimates to $\mathbb{E}\left[\|G\|_\infty^2\right]$. □

Proof of Theorem 2.3. We first observe that, since μ is an isotropic product measure, $\widetilde{W}_{n,d}^p(\mu)$ is an isotropic random vector in $\widetilde{\mathrm{Sym}}^p(\mathbb{R}^n)$. Thus, the matrix A in Theorem 2.5, reduces to a projection matrix and $\|A\|_{op} = 1$.
Let $X \sim \mu$. For the first case, we assume that X is log-concave. Since it is also isotropic, from (2.6) there exists a constant C'_p depending only p, such that

$$\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]} \leq C'_p n^{2p-2}. \tag{2.13}$$

We now let φ_μ stand for the Brenier map between the standard Gaussian G and X. Since X has independent coordinates it follows from (2.12) that for some absolute constant $C > 0$.

$$\|D\varphi(x)\|_{op} \leq C(1 + \|x\|_\infty).$$

In this case, Lemma 2.11 gives:

$$\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]} \leq C' \log(n)^2, \tag{2.14}$$

where $C' > 0$ is some other absolute constant. Plugging these estimates into Theorem 2.5 and taking $C_p = 2C' \cdot C'_p \cdot p^4$ shows Point 1.

The proof of Point 2 is almost identical and we omit it. For Point 3, when each marginal of μ is a polynomial of the standard Gaussian, observe that the map $\tilde{Q} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\tilde{Q}(x_1, ..., x_n) = (Q(x_1), ..., Q(x_n)),$$

is by definition a transport map between G and X. Since Q is a degree k polynomial, there exists some constant C_Q, such that

$$Q'(x) \leq C_Q(1 + |x|^{k-1}).$$

So, from Lemma 2.11, there is some constant $C'_{Q,p}$, such that

$$\sqrt{\mathbb{E}\left[\left\|D\tilde{Q}(G)\right\|_{op}^{8}\right]} \leq C'_{Q,p}\log(n)^{2(k-1)}.$$

Moreover, using hypercontractivity (see [140, Theorem 5.10]), since X is given by a degree k polynomial of the standard Gaussian, we also have the following bound on the moments of X:

$$\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]} \leq (8p)^{2kp}\mathbb{E}\left[\|X\|_2^2\right]^{2p-2} = (8p)^{2kp}n^{2p-2}.$$

Using the above two displays in Theorem 2.5 finishes the proof. □

2.8 Extending Theorem 2.3; Proof of Theorem 2.4

We now fix $X \sim \mu$ to be an unconditional isotropic log-concave measure on $\mathbb{R}^n$ with independent coordinates. If $\Sigma_2(\mu)$ stands for the covariance matrix of $W^2_{n,d}(\mu)$, then, using the same arguments as in the proof of Theorem 2.3, it will be enough to show that $\Sigma_2(\mu)$ is bounded uniformly from below. Towards that, we first prove:

Lemma 2.12. *Let Y be an isotropic log-concave random variable in $\mathbb{R}$. Then*

$$\mathrm{Var}(Y^2) \geq \frac{1}{100}.$$

Proof. Denote by ρ the density of Y. We will use the following 3 facts, pertaining to isotropic log-concave densities in $\mathbb{R}$ (see Section 5.2 in [170]).

- ρ is uni-modal. That is, there exists a point $x_0 \in \mathbb{R}$, such that ρ is non-decreasing on $(-\infty, x_0)$ and non-increasing on (x_0, ∞).
- $\rho(0) \geq \frac{1}{8}$ and $\rho(x) \leq 1$, for every $x \in \mathbb{R}$.
- $\int\limits_{|x|\geq 2} \rho(x)dx \leq \frac{1}{e}$.

The first observation is that either $\rho\left(\frac{1}{9}\right) \geq \frac{1}{10}$, or $\rho\left(-\frac{1}{9}\right) \geq \frac{1}{10}$. Indeed, if not, then as ρ is uni-modal and $\rho(0) \geq \frac{1}{10}$,

$$\int\limits_{-2}^{2} \rho(x)dx \leq \int\limits_{-\frac{1}{9}}^{\frac{1}{9}} \rho(x)dx + \frac{4}{10} \leq \frac{2}{9} + \frac{4}{10} < 1 - \frac{1}{e},$$

which is a contradiction. We assume, without loss of generality, that $\rho\left(\frac{1}{9}\right) \geq \frac{1}{10}$. Similar considerations then show

$$\mathrm{Var}(Y^2) = \int_{\mathbb{R}} (x^2-1)^2 \rho(x) dx \geq \frac{1}{10} \int_0^{\frac{1}{9}} (x^2-1)^2 dx \geq \frac{1}{100}.$$

□

Using the lemma, we now prove Theorem 2.4.

Proof of Theorem 2.4. First, as in Lemma 2.10, the product structure of μ implies that $\Sigma_2(\mu)$, the covariance matrix of $W^2_{n,d}(\mu)$, is diagonal. Write $X = \left(X_{(1)}, ..., X_{(n)}\right)$. There are two types of elements on the diagonal:

- The first corresponds to elements of the form $\mathrm{Var}(X_{(i)}X_{(j)})$. For those elements, by independence, $\mathrm{Var}(X_{(i)}X_{(j)}) = 1$.
- The other type of elements are of the form $\mathrm{Var}\left(X^2_{(i)}\right)$. By Lemma 2.12, $\mathrm{Var}(X_i^2) \geq \frac{1}{100}$.

So, if $P : (\mathbb{R}^n)^{\otimes 2} \to \mathrm{Sym}^2(\mathbb{R}^n)$ is the projection operator, we have that

$$\left\| \Sigma^{-\frac{1}{2}} P \right\|^2_{op} \leq 100.$$

The estimates (2.13) and (2.14) are valid here as well. Thus, Theorem 2.5 implies the result. □

2.9 Non-homogeneous sums

In this section we consider a slight variation on the law of $W^p_{n,d}(\mu)$. Specifically, we let $\alpha := \{\alpha_i\}_{i=1}^d \subset \mathbb{R}^+$, be a sequence of positive numbers and $X_i \sim \mu$ be i.i.d. random vectors in $\mathbb{R}^n$. Define $W^p_{\alpha,d}(\mu)$ as the law of the non-homogeneous sum,

$$\frac{1}{\|\alpha\|_2} \sum_{i=1}^d \alpha_i \left(X_i^{\odot p} - \mathbb{E}\left[X_i^{\odot p}\right] \right), \tag{2.15}$$

where for $q > 0$, $\|\alpha\|_q^q := \sum_{i=1}^d \alpha_i^q$. The marginal law $\widetilde{W}^p_{\alpha,d}(\mu)$ is defined accordingly. The case $\alpha_i \equiv 1$ corresponds to $W^p_{n,d}(\mu)$. As it turns out, controlling the Stein discrepancy of $\widetilde{W}^p_{\alpha,d}(\mu)$ poses no new difficulties and Theorem 2.5 may be readily adapted to deal with these laws as well. The basic observation is that the calculation in (18) also applies to this case.

Indeed, in the general case, if Y_i are i.i.d. isotropic random vectors with Stein kernel given by τ_Y, then, $S_\alpha := \frac{1}{\|\alpha\|_2}\sum_{i=1}^d \alpha_i Y_i$ is isotropic as well, and it has a Stein kernel given by

$$\tau_{S_\alpha}(x) = \frac{1}{\|\alpha\|_2^2}\sum_{i=1}^d \alpha_i^2 \mathbb{E}\left[\tau_Y(Y_i)|S_\alpha = x\right].$$

By repeating the calculations which led to (20), we may see

$$S^2(S_\alpha) \leq \mathbb{E}\left[\|\tau_{S_\alpha}(S_\alpha) - \mathrm{Id}\|_{HS}^2\right] = \mathbb{E}\left[\left\|\frac{1}{\|\alpha\|_2^2}\sum_{i=1}^d \alpha_i^2 \mathbb{E}\left[\tau_Y(Y_i) - \mathrm{Id}|S_\alpha\right]\right\|_{HS}^2\right]$$
$$\leq \frac{1}{\|\alpha\|_2^4}\sum_{i=1}^d \alpha_i^4 \mathbb{E}\left[\|\tau_Y(Y_i) - \mathrm{Id}]\|_{HS}^2 = \frac{\|\alpha\|_4^4}{\|\alpha\|_2^4}\mathbb{E}\left[\|\tau_Y(Y) - \mathrm{Id}\|_{HS}^2\right],$$

which implies

$$S^2(S_\alpha) \leq \frac{\|\alpha\|_4^4}{\|\alpha\|_2^4} S^2(Y).$$

Combining this inequality with Lemma 2.9 we obtain the following variant of Theorem 2.5.

Theorem 2.13. *With the same notations as in Theorem 2.5,*

$$S^2\left(A_* W_{\alpha,d}^p(\mu)\right) \leq 2\frac{\|\alpha\|_4^4}{\|\alpha\|_2^4}\left(\|A\|_{op}^2 p^4 \cdot n\sqrt{\mathbb{E}\left[\|X\|_2^{8(p-1)}\right]}\sqrt{\mathbb{E}\left[\|D\varphi(G)\|_{op}^8\right]} + n^p\right).$$

Thus, all of our results apply to non-homogeneous sums as well. We state here only the analogue for uniformly log-concave measures as reference.

Theorem 2.14. *Let μ be an isotropic L-uniformly log-concave measure on $\mathbb{R}^n$ which is also unconditional. Denote $\Sigma^{-\frac{1}{2}} = \sqrt{\widetilde{\Sigma}_p(\mu)^{-1}}$, where $\widetilde{\Sigma}_p(\mu)$ is the covariance matrix of $\widetilde{W}_{\alpha,d}(\mu)$. Then, there exists a constant C_p, depending only on p, such that*

$$S^2\left(\Sigma_*^{-\frac{1}{2}}\widetilde{W}_{\alpha,d}^p(\mu)\right) \leq \frac{C_p}{L^4} n^{2p-1}\frac{\|\alpha\|_4^4}{\|\alpha\|_2^4}.$$

By specializing to $\mu = \gamma$ and $p = 2$, the theorem gives the same bound as in (2.3), which was obtained in [102].

As noted in the introduction, when $p = 2$, the symmetric matrix defined by (2.15) can be realized as normalized version of a Gram matrix $\mathbb{X}\mathbb{X}^T$, where $\mathbb{X}$ is an $n \times d$ matrix with independent columns.

By taking a different perspective on Theorem 2.14, we now show that, in some special cases, one may also allow dependencies between the columns of $\mathbb{X}$. Let Σ be a $d \times d$ positive definite

matrix and $\{X_i\}_{i=1}^n$ i.i.d random vectors in $\mathbb{R}^d$ with common law $\mathcal{N}(0,\Sigma)$. Suppose that for every $i=1,...,d$, $\Sigma_{i,i}=1$ and define $\mathbb{X}_\Sigma$ to be an $n\times d$ matrix whose i^{th} row equals X_i. So, the rows of $\mathbb{X}_\Sigma$ are independent while its columns might admit dependencies. Now, set

$$W_n(\Sigma) := \frac{1}{\sqrt{d}}\left(\mathbb{X}_\Sigma\mathbb{X}_\Sigma^T - d\cdot \mathrm{Id}\right).$$

Our result will apply by a change of variables. If U is a $d\times d$ orthogonal matrix which diagonalizes Σ the following identity holds:

$$\mathbb{X}_\Sigma\mathbb{X}_\Sigma^T = \left(\mathbb{X}_\Sigma U\right)\left(\mathbb{X}_\Sigma U\right)^T,$$

with the columns of $\mathbb{X}_\Sigma U$ being independent. Specifically, the rows of $\mathbb{X}_\Sigma$ are given by U^TX_i. Thus, if $\{\alpha_i\}_{i=1}^d$ are the eigenvalues of Σ, then for every i,j, $(\mathbb{X}_\Sigma U)_{i,j}\sim\mathcal{N}(0,\alpha_j)$. This implies that $W^2_{\alpha,d}(\gamma)$ is the law of the upper triangular part of $\frac{\sqrt{d}}{\|\alpha\|_2^2}W_n(\Sigma)$. So,

$$S^2\left(\frac{\sqrt{d}}{\|\alpha\|_2^2}W_n(\Sigma)\right) \le Cn^3\frac{\|\alpha\|_4^4}{\|\alpha\|_2^4} = Cn^3\frac{\mathrm{Tr}\left(\Sigma^4\right)}{\mathrm{Tr}\left(\Sigma^2\right)^2}.$$

As a particular case, we can assume that the rows of $\mathbb{X}_\Sigma$ form a stationary Gaussian process. Let $s:\mathbb{N}\to\mathbb{R}$ be a function with $s(0)=1$ and define a symmetric $d\times d$ matrix by $(\Sigma_s)_{i,j} = s(|i-j|)$. If Σ_s is positive definite, then the proof of Theorem 1.2 in [194] shows:

$$\mathcal{W}_1^2(W_n(\Sigma_s),G_s) \le Cn^3\frac{1}{d^2}\sum_{i,j,k,\ell=1}^d s(|i-j|)s(|j-k|)s(|k-\ell|)s(|\ell-i|),$$

where G_s is the law of a Wigner matrix, normalized to have the same covariance structure as $W_n(\Sigma_s)$. Since $s(0)=1$, it is clear that $\mathrm{Tr}(\Sigma_s^2)^2\ge d^2$ and we also have

$$\mathrm{Tr}(\Sigma_s^4) = \sum_{i,j,k\ell=1}^d s(|i-j|)s(|j-k|)s(|k-\ell|)s(|\ell-i|).$$

Thus, our result is directly comparable to the one in [194].

3

A Central Limit Theorem for Neural Networks in a Space of Functions

3.1 Introduction

In the past decade, artificial neural networks have experienced an unprecedented renaissance. However, the current theory has yet to catch-up with the practice and cannot explain their impressive performance. Particularly intriguing is the fact that *over-parameterized* models do not tend to over-fit, even when trained to zero error on the training set. Owing to this seemingly paradoxical fact, researchers have focused on understanding the infinite-width limit of neural networks. This line of research has led to many important discoveries such as the 'lazy-training' regime [75, 234] which is governed by the limiting 'neural tangent kernel' (see [139]), as well as the 'mean-field' limit approach (see [148, 175, 176] for some examples) to study the training dynamics and loss landscape.

The first to study the limiting distribution of a neural network at (a random) initialization was Neal [189], who proved a *Central Limit Theorem (CLT)* for two-layered wide neural networks. According to Neal's result, when initialized with random weights, as the width of the network goes to infinity, its law converges, in distribution, to a Gaussian process. Subsequent works have generalized this result to deeper networks and other architectures ([122, 134, 174, 198, 243, 246, 247]). This correspondence between Gaussian processes and neural networks has proved to be highly influential and has inspired many new models (see [246] for a thorough review of these

models).

Towards supplying a theoretical framework to study *real world* neural networks, one important challenge is to understand the extent to which existing asymptotic results, which essentially only apply to infinite networks, may also be applied to finite ones. While there have been several works in this direction (c.f. [4, 11, 12, 16, 123, 132, 245]), to the best of our knowledge, all known results consider finite-dimensional marginals of the random process and the question of finding a finite-width quantitative analog to Neal's CLT, which applies in a functional space, has remained open. The main goal of this chapter is to tackle this question.

In essence, we prove a quantitative CLT in the space of functions. On a first glance, this is a completely different setting than the classical CLT, even in high-dimensional regimes. The function space is infinite-dimensional, while all quantitative bounds of CLT deteriorate with the dimension ([42,47,85]). However, by exploiting the special structure of neural networks we are able to reduce the problem to finite-dimensional sets, where we capitalize on recent advances made in understanding the rate of convergence of the high-dimensional CLT. In particular, we give quantitative bounds, depending on the network's width and the dimension of the input, which show that, when initialized randomly, wide but finite networks can be well-approximated by a Gaussian process. The functional nature of our results essentially means that when considering the joint distribution attained on a finite set of inputs to the function, ours bounds do not deteriorate as the number of input points increases.

Roughly speaking, we prove the following results:

- We first consider two-layered networks with polynomial activation functions. By embedding the network into a high-dimensional tensor space we prove a quantitative CLT, with a polynomial rate of convergence in a strong transportation metric.

- We next consider general activations and show that under a (very mild) integrability assumption, one can reduce this case to the polynomial case. This is done at the cost of weakening the transportation distance. The rate of convergence depends on the smoothness of the activation and is typically sub-polynomial.

Organization: The rest of this chapter is devoted to describing and proving these results. In Section 3.2 we give the necessary background concerning random initializations of neural networks and we introduce a metric between random processes on the sphere. Our main results are stated in Section 3.3. In Section 3.4 we prove results which concern polynomial activations, while in Section 3.5 we consider general activations.

3.2 Background

Let $\sigma : \mathbb{R} \to \mathbb{R}$ and fix a dimension $n > 1$. A two-layered network with activation σ is a function $N : \mathbb{R}^n \to \mathbb{R}$, of the form

$$N(x) = \sum_{i=1}^{k} c_i \sigma \left(u_i \cdot x \right),$$

where $u_i \in \mathbb{R}^n, c_i \in \mathbb{R}$, for every $i = 1, \ldots, k$. We will refer to k as the width of the network. In most training procedures, it is typical to initialize the weight as *i.i.d.* random vectors. Specifically, let $\{w_i\}_{i=1}^k$ be *i.i.d.* as standard Gaussians in $\mathbb{R}^n$ and let $\{s_i\}_{i=1}^k$ be *i.i.d.* with $\mathbb{P}(s_1 = 1) = \mathbb{P}(s_1 = -1) = \frac{1}{2}$. We consider the random network,

$$\mathcal{P}_k \sigma(x) := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i \sigma \left(w_i \cdot x \right).$$

Let $\mathbb{S}^{n-1}$ stand for the unit sphere in $\mathbb{R}^n$. By restricting our attention to $x \in \mathbb{R}^n$, with $\|x\| = 1$, we may consider $\mathcal{P}_k \sigma$ as a random process, indexed by $\mathbb{S}^{n-1}$. In other words, $\mathcal{P}_k \sigma$ is a random vector in $L^2(\mathbb{S}^{n-1})$, equipped with its canonical rotation-invariant probability measure.

A Gaussian process is a random vector $\mathcal{G} \in L^2(\mathbb{S}^{n-1})$, such that for any finite set $\{x_j\}_{j=1}^m \subset \mathbb{S}^{n-1}$ the random vector $\{\mathcal{G}(x_j)\}_{j=1}^m \in \mathbb{R}^m$, has a multivariate Gaussian law. Since $\mathcal{P}_k \sigma$ is a sum of independent centered vectors, standard reasoning suggests that as $k \to \infty$, $\mathcal{P}_k \sigma$ should approach a Gaussian law in $L^2(\mathbb{S}^{n-1})$, i.e. a Gaussian process. Indeed, this is precisely Neal's CLT, which proves the existence of a Gaussian process $\mathcal{G}$, such that $\mathcal{P}_k \sigma \xrightarrow{k \to \infty} \mathcal{G}$, where the convergence is in distribution.

To make this result quantitative, we must first specify a metric. Our choice is inspired by the classical Wasserstein transportation in Euclidean spaces. The observant reader may notice that our definition, described below, does not correspond to the p-Wasserstein distance on $L^2(\mathbb{S}^{n-1})$, but rather the p-Wasserstein distance on $L^p(\mathbb{S}^{n-1})$. We chose this presentation for ease of exposition and its familiarity.

For $\mathcal{P}, \mathcal{P}'$, random processes on the sphere, and $p \geq 1$ we define the functional p-Wasserstein distance as,

$$\mathcal{WF}_p(\mathcal{P}, \mathcal{P}') := \inf_{(\mathcal{P}, \mathcal{P}')} \left(\int_{\mathbb{S}^{n-1}} \mathbb{E} \left[|\mathcal{P}(x) - \mathcal{P}'(x)|^p \right] dx \right)^{\frac{1}{p}},$$

where the infimum is taken over all couplings of $(\mathcal{P}, \mathcal{P}')$ and where dx is to be understood as the normalized uniform measure on $\mathbb{S}^{n-1}$. For $p = \infty$ we define $\mathcal{WF}_\infty$ as,

$$\mathcal{WF}_\infty(\mathcal{P}, \mathcal{P}') := \inf_{(\mathcal{P}, \mathcal{P}')} \mathbb{E} \left[\sup_{x \in \mathbb{S}^{n-1}} |\mathcal{P}(x) - \mathcal{P}'(x)| \right].$$

The notation $\mathcal{W}_p$ is reserved to the p-Wasserstein distance in finite-dimensional Euclidean spaces.

There is a straightforward way to connect between the quadratic functional Wasserstein distance on the sphere and the L^2 distance in Gaussian space.

Lemma 3.1. *Let $f, g : \mathbb{R} \to \mathbb{R}$, and γ, the standard Gaussian in $\mathbb{R}$. Then,*

$$\mathcal{WF}_2^2(\mathcal{P}_k f, \mathcal{P}_k g) \leq \int_{\mathbb{R}} (f(x) - g(x))^2 d\gamma(x).$$

Proof. There is a natural coupling such that

$$\begin{aligned}
\mathcal{WF}_2^2(\mathcal{P}_k f, \mathcal{P}_k g) &\leq \frac{1}{k} \int_{\mathbb{S}^{n-1}} \mathbb{E}\left[\left(\sum_{i=1}^k s_i (f(w_i \cdot x) - g(w_i \cdot x))\right)^2\right] dx \\
&= \frac{1}{k} \int_{\mathbb{S}^{n-1}} \sum_{i=1}^k \mathbb{E}\left[(f(w_i \cdot x) - g(w_i \cdot x))^2\right] dx \\
&= \int_{\mathbb{R}} (f(x) - g(x))^2 d\gamma(x).
\end{aligned}$$

The first equality is a result of independence, while the second equality follows from the fact that for any $x \in \mathbb{S}^{n-1}$, $w_i \cdot x \sim \gamma$. □

3.3 Results

We now turn to describe the quantitative CLT convergence rates obtained by our method. Our first result deals with polynomial activations.

Theorem 3.2. *Let $p(x) = \sum_{m=0}^{d} a_m x^m$ be a degree d polynomial. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that*

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \leq C_d \max_m \{|a_m|^2\} \left(\frac{n^{5d-\frac{1}{2}}}{k}\right)^{\frac{1}{3}},$$

where $C_d \leq d^{Cd}$, for some numerical constant $C > 0$.

According to the result, when the degree d is fixed, as long as $k \gg n^{5d-\frac{1}{2}}$, $\mathcal{P}_k p$ is close to a Gaussian process. One way to interpret the metric $\mathcal{WF}_\infty$, in the result, is as follows. For any finite set $\{x_j\}_{j=1}^m \subset (\mathbb{S}^{n-1})^m$, the random vector $\{\mathcal{P}_k p(x_j)\}_{j=1}^m \subset \mathbb{R}^m$ converges to a Gaussian random vector, uniformly in m. Let us also mention that while the result is stated for Gaussian

weights, the Gaussian plays no special role here (as will become evident from the proof), and the weights could be initialized by any symmetric random vector with sub-Gaussian tails.

One drawback of using polynomial activations is that the resulting network will always be a polynomial of bounded degree, which limits its expressive power. For this reason, in practice, neural networks are usually implemented using non-polynomial activations. By using a polynomial approximation scheme in Gaussian space, we are able to extend our result to this setting as well. We defer the necessary definitions and formulation of the result to Section 3.5, but mention here two specialized cases of common activations.

We first consider the Rectified Linear Unit (ReLU) function, denoted as $\psi(x) := \max(0, x)$. For this activation, we prove:

Theorem 3.3. *There exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{WF}_2^2(\mathcal{P}_k\psi, \mathcal{G}) \leq C\left(\frac{\log(n)\log(\log(k))}{\log(k)}\right)^2,$$

where $C > 0$ is a numerical constant.

The reader might get the impression that this is a weaker result than Theorem 3.2. Indeed, the rate of convergence here is much slower. In order to get $\mathcal{WF}_2(\mathcal{P}_k\psi, \mathcal{G}) \leq \varepsilon$, the theorem requires that $k \gtrsim n^{\frac{1}{\varepsilon}\log\log n}$. Also, $\mathcal{WF}_2$ is a weaker metric than $\mathcal{WF}_\infty$. Let us point out that it may not be reasonable to expect similar behavior for polynomial and non-polynomial activations. The celebrated universal approximation theorem of Cybenko ([88], see also [25, 166]) states that any function in $L^2(\mathbb{S}^{n-1})$ can be approximated, to any precision, by a sufficiently wide neural network with a non-polynomial activation. Thus, as $k \to \infty$, the limiting support of $\mathcal{P}_k\psi$ will encompass all of $L^2(\mathbb{S}^{n-1})$. This is in sharp contrast to a polynomial activation function, for which the support of $\mathcal{P}_k p$ is always contained in some finite-dimensional subspace of $L^2(\mathbb{S}^{n-1})$, uniformly in k.

Another explanation for the slow rate of convergence, is the fact that ψ is non-differentiable. For smooth functions, the rate can be improved, but will still be typically sub-polynomial. As an example, we consider the hyperbolic tangent activation, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Theorem 3.4. *There exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{W}_2^2(\mathcal{P}_k \tanh, \mathcal{G}) \leq C\exp\left(-\frac{1}{C}\sqrt{\frac{\log(k)}{\log(n)\log(\log(k))}}\right),$$

where $c > 0$ is an absolute constant.

Finally, let us remark about possible improvements to our obtained rates. We do not know whether the constant C_d in Theorem 3.5 is necessarily exponential, and we have made no effort to optimize it. We do conjecture that the dependence on the ratio $\frac{n^{5d-\frac{1}{2}}}{k}$ is not tight. To support this claim we prove an improved rate when the activation is monomial.

Theorem 3.5. *Let $p(x) = x^d$ for some $d \in \mathbb{N}$. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that*

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \leq C_d \frac{n^{2.5d-1.5}}{k},$$

where $C_d \leq d^{Cd}$, for some numerical constant $C > 0$.

Remark 3.6. It is plausible the dependence on d and k could be further improved. Let us note that when $d = 2$, the best rate one could hope for is proportional to $\frac{n^3}{k}$. This is a consequence of the bounds proven in [56, 143], which show that if $n^3 \gg k$, then when considered as a random bi-linear form (or a Wishart matrix) $\mathcal{P}_k p$ is far from any Gaussian law. In fact, our proof of Theorem 3.5 can actually be improved when $d = 2$ (or, in general, for even d), and we are able to obtain the sharp rate $\frac{n^3}{k}$. It is an interesting question to understand the correct rates when $d > 2$.

3.4 Polynomial processes

For this section, fix a polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree d, $p(x) = \sum_{m=0}^{d} a_m x^m$. The goal of this section is to show that when k is large enough, $\mathcal{P}_k p$ can be well approximated by a Gaussian process in the $\mathcal{WF}_\infty$ metric. Towards this, we will use the polynomial p to embed $\mathbb{R}^n$ into some high-dimensional tensor space.

3.4.1 The embedding

For $m \in \mathbb{N}$, we make the identification $(\mathbb{R}^n)^{\otimes m} = \mathbb{R}^{n^m}$ and focus on the subspace of symmetric tensors, which we denote $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$. If $\{e_i\}_{i=1}^n$ is the standard orthonormal basis of $\mathbb{R}^n$, then an orthonormal basis for $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$, is given by the set

$$\{e_I | I \in \mathrm{MI}_n(m)\}.$$

where $\mathrm{MI}_n(m)$ is the set of multi-indices,

$$\mathrm{MI}_n(m) = \{(I_1, \ldots I_n) \in \mathbb{N}^n | I_1 + \cdots + I_n = m\},$$

With this perspective, we have $e_I = \otimes_{i=1}^n \left(e_i^{\otimes I_i}\right)$, and we denote the inner product on $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$ by $\langle \cdot, \cdot \rangle_m$. We also use the following multi-index notation: if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $x^I = \prod_{i=1}^n x_i^{I_i}$.

Define the feature space $H := \oplus_{m=0}^d \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$. If $\pi_m : H \to \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes m}\right)$ is the

natural projection, then an inner product on H may be defined by

$$\langle v, u\rangle_H := \sum_{m=0}^{d} \langle \pi_m v, \pi_m u\rangle_m.$$

We further define the embedding $P : \mathbb{R}^n \to H$, $P(x) = \sum_{m=0}^{d} \sqrt{|a_m|} x^{\otimes m}$, which induces a bi-linear form on H as,

$$Q(u, v) := \sum_{m=0}^{d} \operatorname{sign}(a_m) \langle \pi_m u, \pi_m v\rangle_m.$$

Observe that Q is not necessarily positive definite, but still satisfies the following Cauchy-Schwartz type inequality,

$$Q(u, v) \le \|u\|_H \|v\|_H. \tag{3.1}$$

Furthermore, it is clear that for any $x, y \in \mathbb{R}^n$,

$$Q(P(x), P(y)) = \sum_{m=0}^{d} a_m (x \cdot y)^m = p\,(x \cdot y)\,,$$

and we have the identity,

$$\mathcal{P}_k p(x) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i p(w_i \cdot x) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i Q(P(x), P(w_i)) = Q\left(P(x), \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)\right). \tag{3.2}$$

Consider the random vector $X_k := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)$ taking values in H. By the central limit theorem, we should expect X_k to approach a Gaussian law. The next result shows that approximate Gaussianity of X_k implies that the process $\mathcal{P}_k p$ is approximately Gaussian as well.

Lemma 3.7. *Let G be a Gaussian random vector in H and define the random process on $\mathcal{G}$ in $\mathbb{S}^{n-1}$ by $\mathcal{G}(x) := Q(P(x), G)$. Then, $\mathcal{G}$ is a Gaussian process and,*

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p, \mathcal{G}) \le \left(\sum_{m=0}^{d} |a_m|\right) \mathcal{W}_2^2(X_k, G).$$

Proof. Let (X_k, G) be the optimal coupling so that $\mathcal{W}_2^2(X_k, G) = \mathbb{E}\left[\|X_k - G\|_H^2\right]$. We then

have

$$\mathcal{WF}_\infty(\mathcal{P}_k p, \mathcal{G}) \leq \mathbb{E}\left[\sup_{x\in\mathbb{S}^{n-1}} |\mathcal{P}_k p(x) - \mathcal{G}(x)|\right] = \mathbb{E}\left[\sup_{x\in\mathbb{S}^{n-1}} |Q(P(x), X_k - G)|\right]$$
$$\leq \sup_{x\in\mathbb{S}^{n-1}} \|P(x)\|_H \sqrt{\mathbb{E}\left[\|X_k - G\|_H^2\right]} = \sup_{x\in\mathbb{S}^{n-1}} \|P(x)\|_H \cdot \mathcal{W}_2(X_k, G),$$

where we have used (3.1) in the second inequality. Now, for any $x \in \mathbb{S}^{n-1}$,

$$\|P(x)\|_H = \sqrt{\sum_{m=0}^{d} |a_m| \langle x^{\otimes m}, x^{\otimes m}\rangle_m} = \sqrt{\sum_{m=0}^{d} |a_m|}.$$

□

So, we wish to show that the random vector $X_k := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)$ is approximately Gaussian inside H. For this, we will apply the following Wasserstein CLT bound, recently proven by Bonis, in [47].

Theorem 3.8. *[47, Theorem 1] Let Y_i be i.i.d isotropic random vectors in $\mathbb{R}^N$ and let G be the standard Gaussian. Then, if $S_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} Y_i$,*

$$\mathcal{W}_2^2(S_k, G) \leq \frac{\sqrt{N}}{k} \left\| \mathbb{E}\left[YY^T \|Y\|_2^2\right] \right\|_{HS}.$$

Since the theorem applies to isotropic random vectors, for which the covariance matrix is the identity, we first need to understand $\Sigma := \mathrm{Cov}\,(P(w))$. Let us emphasize the fact that Σ is a bi-linear operator on H. Thus it can be regarded as a $\dim(H) \times \dim(H)$ positive semi-definite matrix.

3.4.2 The covariance matrix

We first show that one may disregard small eigenvalues of Σ. Let (λ_j, v_j) stand for the eigenvalue/vector pairs of Σ. Fix $\delta > 0$ define $V_\delta = \mathrm{span}\,(v_j | \lambda_j \leq \delta)$ and let $\Pi_\delta, \Pi_\delta^\perp$ be the orthogonal projection unto $V_\delta, V_\delta^\perp$, respectively.

Lemma 3.9. *Let $G \sim \mathcal{N}(0, \Sigma)$ be a Gaussian in H, then*

$$\mathcal{W}_2^2(X_k, G) \leq \mathcal{W}_2^2(\Pi_\delta^\perp X_k, \Pi_\delta^\perp G) + 8n^d\delta.$$

Proof. For any coupling (X_k, G) we have

$$\begin{aligned}\mathcal{W}_2^2(X_k, G) &\leq \mathbb{E}\left[\|X_k - G\|^2\right] = \mathbb{E}\left[\|\Pi_\delta X_k - \Pi_\delta G\|^2\right] + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right] \\ &\leq 2\mathbb{E}\left[\|\Pi_\delta G\|^2\right] + 2\mathbb{E}\left[\|\Pi_\delta X_k\|^2\right] + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right] \\ &\leq 4\dim(H)\delta + \mathbb{E}\left[\left\|\Pi_\delta^\perp X_k - \Pi_\delta^\perp G\right\|^2\right].\end{aligned}$$

The proof concludes by taking the coupling for which $\Pi_\delta^\perp X_k, \Pi_\delta^\perp G$ is optimal, and by noting $\dim(H) \leq 2n^d$. □

Next, we bound from above the eigenvalues of Σ.

Lemma 3.10. *Let $\Sigma = \mathrm{Cov}(P(w))$, where $P(w)$ is defined as in (3.2). Then*

$$\|\Sigma\|_{op} \leq (4d)! \max_m\{|a_m|\} n^{\frac{d-1}{2}}.$$

Proof. Let $\sum_I v_I e_I = v \in H$ be a unit vector, we wish to bound $\langle v, \Sigma v\rangle = \mathrm{Var}\left(\langle P(w), v\rangle_H\right)$ from above. Let us denote the degree d polynomial $\sum_{i=0}^{m} \sum_{I\in \mathrm{MI}_n(m)} \sqrt{|a_m|} v_I x^I = q(x) = \langle P(x), v\rangle_H$. We will prove the claim by induction on d. The case $d = 1$, is rather straightforward to check. For the general case, we will use the Gaussian Poincaré inequality (see [194, Proposition 1.3.7], for example) to reduce the degree. According to the inequality,

$$\begin{aligned}\mathrm{Var}\left(\langle P(w), v\rangle_H\right) = \mathrm{Var}\left(q(w)\right) &\leq \mathbb{E}\left[\|\nabla q(w)\|_2^2\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\left|\frac{d}{dx_i} q(w)\right|^2\right] \\ &= \sum_{i=1}^{n} \mathrm{Var}\left(\frac{d}{dx_i} q(w)\right) + \mathbb{E}\left[\frac{d}{dx_i} q(w)\right]^2. \qquad (3.3)\end{aligned}$$

Fix $i = 1, \ldots, n$, if $I \in \mathrm{MI}_n(m)$ we denote by $\partial_i I \in \mathrm{MI}_n(m-1)$, to be a multi-index set such that

$$\partial_i I_j = \begin{cases} I_j & \text{if } i \neq j \\ \max(0, I_i - 1) & \text{if } i = j. \end{cases}$$

With this notation, we have,

$$\begin{aligned}\frac{d}{dx_i} q(w) &= \frac{d}{dx_i}\left(\sum_{m=0}^{d} \sum_{I\in \mathrm{MI}_n(m)} \sqrt{|a_m|} w^I v_I\right) \\ &= \sum_{m=0}^{d} \sum_{I\in \mathrm{MI}_n(m)} \sqrt{|a_m|} I_i w^{\partial_i I} v_I.\end{aligned}$$

Since $\frac{d}{dx_i}q$ is a polynomial of degree $d-1$, we thus get by induction,

$$\mathrm{Var}\left(\frac{d}{dx_i}q(w)\right) \leq (4d-4)!\max_m\{|a_m|\}n^{\frac{d-2}{2}}\sum_{m=0}^{d}\sum_{I\in \mathrm{MI}_n(m)} I_i^2 v_I^2.$$

Observe that $I_i \leq d$ and that for every $I \in \mathrm{MI}_n(m)$, there are at most d different indices $i \in [n]$, for which $I_i \neq 0$. Therefore,

$$\sum_{i=1}^{n}\mathrm{Var}\left(\frac{d}{dx_i}q(w)\right) \leq d^2(4d-4)!\max_m\{|a_m|\}n^{\frac{d-2}{2}}\left(d\sum_I v_I^2\right) \leq (4d-1)!\max_m\{|a_m|\}n^{\frac{d-2}{2}}. \tag{3.4}$$

Furthermore, if for some $j \in [n]$, $\partial_i I_j$ is odd, then $\mathbb{E}\left[w^{\partial_i I}\right] = 0$. Otherwise,

$$|\mathbb{E}\left[w^{\partial_i I}\right]| \leq |\mathbb{E}\left[w_1^{d-1}\right]| \leq \sqrt{d!}.$$

It is easy to verify that the size of the following set,

$$A_i = \{I \in \cup_{m=0}^{d}\mathrm{MI}_n(m) | I_i\mathbb{E}[w^{\partial_i I}] \neq 0\},$$

is at most $(2n)^{\frac{d-1}{2}}$. Thus, since there are at most $(2n)^{\frac{d-1}{2}}$ elements which do not vanish, Cauchy-Schwartz's inequality shows,

$$\begin{aligned}\mathbb{E}\left[\frac{d}{dx_i}q(w)\right]^2 &\leq d^2\max_m\{|a_m|\}\mathbb{E}\left[\sum_{m=0}^{d}\sum_{I\in\mathrm{MI}_n(m)} w^{\partial_i I}v_I\right]^2 \\ &\leq (4d-1)!\max_m\{|a_m|\}n^{\frac{d-1}{2}}\sum_{I\in A_i} v_I^2.\end{aligned}$$

Note that if $I \in A_i$, then necessarily I_i is odd. In this case, it follows that for $j \neq i$, $\mathbb{E}\left[w^{\partial_j I}\right] = 0$. Hence, $A_i \cap A_j = \emptyset$, and

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[\frac{d}{dx_i}q(w)\right]^2 &\leq (4d-1)!\max_m\{|a_m|\}n^{\frac{d-1}{2}}\sum_{i=1}^{n}\sum_{I\in A_i} v_I^2 \\ &\leq (4d-1)!\max_m\{|a_m|\}n^{\frac{d-1}{2}}\sum_I v_I^2 = (4d-1)!\max_m\{|a_m|\}n^{\frac{d-1}{2}}. \end{aligned} \tag{3.5}$$

We now plug (3.4) and (3.5) into (3.3) to obtain

$$\mathrm{Var}\left(\langle P(w), v\rangle_H\right) \leq (4d)!\max_m\{|a_m|\}n^{\frac{d-1}{2}}.$$

□

Remark that, up to the multiplicative dependence on d, this bound is generally sharp. As an example, when $d = 2\ell - 1$ is odd, one can consider the degree d polynomial,

$$q(x) = \frac{1}{n^{\ell/2}} \sum_{i_1,\dots i_\ell=1}^{n} x_{i_1} x_{i_2}^2 \dots x_{i_\ell}^2.$$

For this polynomial it may be verified that $\mathrm{Var}(q(w)) = \Omega(n^{\ell-1}) = \Omega\left(n^{\frac{d-1}{2}}\right)$.

3.4.3 A functional CLT for polynomial processes

Proof of Theorem 3.2. Let δ be some small number to be determined later and set $\tilde{X}_k = \Sigma^{-1/2} X_k$ and $\tilde{G}$, the standard Gaussian in H. By Lemma 3.9,

$$\begin{aligned}\mathcal{W}_2^2(X_k, G) &\le \mathcal{W}_2^2(\Pi_\delta^\perp X_k, \Pi_\delta^\perp G) + 8n^d\delta \\ &= \mathcal{W}_2^2(\Sigma^{1/2}\Pi_\delta^\perp \tilde{X}_k, \Sigma^{1/2}\Pi_\delta^\perp \tilde{G}) + 8n^d\delta \le \|\Sigma\|_{op}\mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G}) + 8n^d\delta.\end{aligned}$$

We focus on the term $\mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G})$ for which Theorem 3.8 may be invoked,

$$\begin{aligned}\mathcal{W}_2^2(\Pi_\delta^\perp \tilde{X}_k, \Pi_\delta^\perp \tilde{G}) &\le \frac{\sqrt{\dim(H)}}{k}\mathbb{E}\left[\left\|\Pi_\delta^\perp \Sigma^{-1/2} P(w)\right\|_H^4\right] \\ &\le \frac{\sqrt{\dim(H)}}{k}\mathbb{E}\left[\|P(w)\|_H^4\right]\left\|\Pi_\delta^\perp \Sigma^{-1}\right\|_{op}^2 \\ &\le \frac{\sqrt{\dim(H)}}{\delta^2 k}\mathbb{E}\left[\|P(w)\|_H^4\right].\end{aligned}$$

In the first inequality, we have used Jensen's inequality on the bound from Theorem 3.8. Let us estimate $\mathbb{E}\left[\|P(w)\|_H^4\right]$. By definition,

$$\begin{aligned}\mathbb{E}\left[\|P(w)\|_H^4\right] &= \mathbb{E}\left[\left(\sum_{m=0}^{d} |a_m| \|w^{\otimes m}\|_m^{2m}\right)^2\right] \le \left(\sum_{m=0}^{d} a_m^2\right)\left(\sum_{m=0}^{d}\mathbb{E}\left[\|w\|_2^{4m}\right]\right) \\ &\le \left(\sum_{m=0}^{d} a_m^2\right)\left(\sum_{m=0}^{d}(2m)!\left(4\mathbb{E}\left[\|w\|_2^2\right]\right)^{2m}\right) \le \left(\sum_{m=0}^{d} a_m^2\right)\left(\sum_{m=0}^{d}(2m)!\,(4n)^{2m}\right) \\ &\le \left(\sum_{m=0}^{d} a_m^2\right) 16^d (2d)! n^{2d} \le \max_m\{a_m^2\}(100d)! n^{2d}\end{aligned}$$

The first inequality is Cauchy-Schwartz and in the second inequality we have used the fact that $\|w\|_2$ has sub-exponential tails.

Since $\dim(H) \le 2n^d$, it follows that,

$$\mathcal{W}_2^2(X_k, G) \le \|\Sigma\|_{op}\frac{(100d)! n^{\frac{5d}{2}}}{\delta^2 k}\max_m\{a_m^2\} + 8n^d\delta.$$

We plug the estimate for $\|\Sigma\|_{op}$ from Lemma 3.10 to deduce:

$$\mathcal{W}_2^2(X_k, G) \leq \frac{(110d)! n^{3d-0.5}}{\delta^2 k} \max_m\{|a_m|^3\} + 8n^d\delta.$$

We now take $\delta = \left(\frac{(110d)! n^{2d-0.5} \max_m\{|a_m|^3\}}{k}\right)^{\frac{1}{3}}$ to obtain

$$\mathcal{W}_2^2(X_k, G) \leq 16 \max_m\{|a_m|\} \left((110d)!\right) \frac{n^{\frac{5d-0.5}{3}}}{k^{\frac{1}{3}}}$$

To finish the proof, define the Gaussian process $\mathcal{G}$ by $\mathcal{G}(x) = Q\left(P(x), G\right)$, and invoke Claim 3.7. □

3.4.4 An improved rate for tensor powers

Throughout this section we assume that $p(x) = x^d$ for some $d \in \mathbb{N}$. Under this assumption, we improve Theorem 3.2. This improvement is enabled by two factors:

- A specialized CLT for tensor powers, proved in Chapter 2.
- An improved control on the eigenvalues of Σ, which allows to bypass Lemma 3.9.

Let us first state the result about approximating tensor powers by Gaussians. Note that for a polynomial p as above, we have the embedding map $P(x) = x^{\otimes d}$. Since the image of P is always a symmetric d-tensor, we allow ourselves to restrict the embedding map P and overload notations, so that $P : \mathbb{R}^n \to \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$. In this case, for $w \sim \mathcal{N}(0, \mathrm{I}_d)$, we have $\Sigma := \mathrm{Cov}(P(w))$, and $X_k := \frac{1}{\sqrt{k}} \sum_{i=1}^{k} s_i P(w_i)$. An immediate consequence of Theorems 2.2 and 2.5 is the following result:

Theorem 3.11. *Let the above notations prevail. Then, there exists a Gaussian random vector G, in $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$, such that,*

$$\mathcal{W}_2^2\left(X_k, G\right) \leq C_d \|\Sigma\|_{op} \|\Sigma^{-1}\|_{op}^2 \frac{n^{2d-1}}{k},$$

where $C_d = d^{Cd}$, for some universal constant $C > 0$.

Remark that the results in Chapter 2 actually deal with the random vector $\sqrt{\Sigma^{-1}} X_k$. Since we care about the un-normalized vector X_k we incur a dependence on $\|\Sigma\|_{op}$. We now show how to bound from below the eigenvalues of Σ.

Lemma 3.12. *Let $\lambda_{\min}(\Sigma)$ stand for the minimal eigenvalue of Σ. Then*

$$\lambda_{\min}(\Sigma) \geq \frac{1}{d!}.$$

Proof. Let $v \in \mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$ be a unit vector. We can thus write $v = \sum_{|I|=d} v_I e_I$, with $\sum_{I\in \mathrm{MI}_n(d)} v_I^2 = 1$. Define the degree d homogeneous polynomial $q : \mathbb{R}^n \to \mathbb{R}$, by $q(x) = \sum_{I\in \mathrm{MI}_n(d)} v_I x^I$. In this case we have $\langle v, P(w)\rangle = q(w)$, and it will be enough to show,

$$\mathrm{Var}\left(\langle v, P(w)\rangle\right) = \mathrm{Var}\left(q(w)\right) \geq \frac{1}{d!}.$$

We will use the variance expansion for functions of Gaussian vectors, which can be found at [194, Proposition 1.5.1]. According to this expansion, for any smooth enough function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathrm{Var}(f(w)) = \sum_{m=1}^{\infty} \frac{\|\mathbb{E}\left[\nabla^m f(w)\right]\|_m^2}{m!}. \tag{3.6}$$

Here $\nabla^m f$ is the m^{th} total derivative of f, which we regard as an element in $(\mathbb{R}^n)^{\otimes m}$. In particular, we have,

$$\mathrm{Var}\left(q(w)\right) \geq \frac{\left\|\mathbb{E}\left[\nabla^d q(w)\right]\right\|_d^2}{d!}.$$

Now, if $I \neq J$ are two multi-subsets of $[n]$, with $I, J \in \mathrm{MI}_n(d)$, we have

$$\frac{d}{dx^I} x^J = 0 \text{ and } \frac{d}{dx^I} x^I = I!.$$

So, since $\nabla^d f = \{\frac{d}{dx^I} q\}_{I\in \mathrm{MI}_n(d)}$,

$$\|\mathbb{E}\left[\nabla^d f(w)\right]\|_d^2 = \sum_{I\in \mathrm{MI}_n(d)} I! v_I^2 \geq 1,$$

and

$$\mathrm{Var}\left(q(w)\right) \geq \frac{1}{d!},$$

as required. □

We are now in a position to prove Theorem 3.5.

Proof of Theorem 3.5. By combining Lemma 3.10 and Lemma 3.12, there exists some numerical constant $C' > 0$, such that

$$\|\Sigma\|_{op}\|\Sigma^{-1}\|_{op}^2 \leq d^{C'd} n^{\frac{d-1}{2}}.$$

Thus, Theorem 3.11 shows that there exists a Gaussian vector G in $\mathrm{Sym}\left((\mathbb{R}^n)^{\otimes d}\right)$, such that

$$\mathcal{W}_2^2\left(X_k, G\right) \leq d^{Cd} \frac{n^{2.5d-1.5}}{k},$$

for some other constant $C > 0$. Define the Gaussian process $\mathcal{G}(x) = \langle P(x), G\rangle$, then Lemma 3.1 shows,

$$\mathcal{WF}^2_\infty(\mathcal{P}_k p, \mathcal{G}) \leq d^{Cd}\frac{n^{2.5d-1.5}}{k},$$

which concludes the proof. When $d = 2$, it is not hard to see that $\|\Sigma\|_{op}$ can be bounded by an absolute constant (see Lemma 3.13). In this case,

$$\|\Sigma\|_{op}\|\Sigma^{-1}\|^2_{op} \leq C,$$

which is the reason behind Remark 3.6. □

3.4.5 Dimension-free covariance estimates for quadratic tensor powers

When considering the polynomial $p(x) = x^2$, we can strictly improve upon Lemma 3.10 and obtain dimension-free bounds. As noted in the proof of Theorem 3.5, this explains Remark 3.6.

Lemma 3.13. *Suppose that $d = 2$. Then,*

$$\|\Sigma\|_{op} \leq 1.$$

Proof. As in the proof of Lemma 3.12, let $v = \sum_{i,j=1}^{n} v_{i,j} e_i \otimes e_j$, with $\sum v^2_{i,j} = 1$. Define $q(x) = \sum_{i,j=1} v_{i,j} x_i x_j$. It will suffice to bound $\mathrm{Var}(q(w))$ from above. Since q is a quadratic polynomial, the variance decomposition (3.6) gives,

$$\mathrm{Var}(q(w)) = \|\mathbb{E}\left[\nabla q(w)\right]\|^2 + \frac{1}{2}\|\mathbb{E}\left[\nabla^2 q(w)\right]\|^2.$$

for $i \in [n]$, we have $\frac{d}{dx_i} q(w) = \sum_{j=1}^{n}(1 + \delta_{i,j}) v_{i,j} w_j$. So, $\mathbb{E}\left[\nabla q(w)\right] = 0$. On the other hand,

$$\|\mathbb{E}\left[\nabla q(w)\right]\|^2 = \sum_{i,j=1}^{n} \mathbb{E}\left[\frac{d^2}{dx_i dx_j} q(w)\right]^2 \leq 2\sum_{i=1}^{n} v^2_{i,j} = 2,$$

and the claim is proven. □

3.5 General activations

In this section we consider a general (non-polynomial) activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Our goal is to derive a quantitative CLT for the random process $\mathcal{P}_k\sigma$. Our strategy will be to approximate σ by some polynomial, for which Theorem 3.2 applies. We set γ to be the law of the standard

Gaussian in $\mathbb{R}$. Lemma 3.1 suggests that, in order to control the remainder in the approximation, it would be beneficial to find a polynomial p, such that p and σ are close in $L^2(\gamma)$.

In $L^2(\gamma)$ there is a distinguished set of polynomials, the so-called Hermite polynomials. Henceforth we denote h_m to be the m^{th} normalized Hermite polynomial,

$$h_m(x) = \frac{(-1)^m}{\sqrt{m!}} \left(\frac{d^m}{dx^m} e^{-\frac{x^2}{2}} \right) e^{\frac{x^2}{2}}.$$

The reader is referred to [140] for the necessary definitions and proofs pertaining to Hermite polynomials. We will mainly care about the fact that $\{h_m\}_{m=0}^{\infty}$ forms a complete orthonormal system in $L^2(\gamma)$. Thus, assuming that $\sigma \in L^2(\gamma)$, it may be written as,

$$\sigma = \sum_{m=0}^{\infty} \hat{\sigma}_m h_m \text{ , where } \hat{\sigma}_m := \int_{\mathbb{R}} \sigma(x) h_m(x) d\gamma(x).$$

Let us also define the remainder function of σ as,

$$R_\sigma(d) = \sum_{m=d+1}^{\infty} \hat{\sigma}_r^2.$$

If we define the degree d polynomial

$$p_d := \sum_{m=1}^{d} \hat{\sigma}_m h_m, \tag{3.7}$$

we then have,

$$\|\sigma - p_d\|_{L^2(\gamma)}^2 \leq R_\sigma(d). \tag{3.8}$$

With these notations, the main result of this section is:

Theorem 3.14. *Suppose that $\sigma \in L^2(\gamma)$. Then, there exists a Gaussian process $\mathcal{G}$ on $\mathbb{S}^{n-1}$, such that,*

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C' \frac{\max_m |\hat{\sigma}_m|^2}{k^{\frac{1}{6}}} + R_\sigma \left(\frac{\log(k)}{C' \log(n) \log(\log(k))} \right),$$

where $C' > 0$ is a numerical constant.

Before proving the theorem, we first focus on the coefficients of the polynomial p_d, defined in (3.7), with respect to the standard monomial basis. For this, we write h_m, explicitly (see [140, Chapter 3]) as,

$$h_m(x) = \sqrt{m!} \sum_{j=0}^{\frac{m}{2}} \frac{(-1)^j}{j!(m-2j)!2^j} x^{m-2j}.$$

Write now $p_d = \sum_{m=0}^{d} a_m x^m$ and let us estimate a_m.

Lemma 3.15. *It holds that*

$$|a_m| \le \max_i |\hat{\sigma}_i| \frac{2}{\sqrt{m!}} 2^d.$$

Proof. We have:

$$\begin{aligned}
|a_m| &\le \sum_{i=m}^{d} |\hat{\sigma}_i| \frac{\sqrt{i!}}{m!((i-m)/2)!2^{(i-m)/2}} \\
&\le \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{\sqrt{i!}}{m!((i-m)/2)!2^{(i-m)/2}} \\
&\le \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{\sqrt{i!}}{m!\sqrt{(i-m)!}} = \max_i |\hat{\sigma}_i| \sum_{i=m}^{d} \frac{1}{\sqrt{m!}} \sqrt{\binom{i}{m}} \\
&\le \max_i |\hat{\sigma}_i| \frac{1}{\sqrt{m!}} \sum_{i=m}^{d} \binom{i}{m} = \max_i |\hat{\sigma}_i| \frac{1}{\sqrt{m!}} \binom{d+1}{m+1} \le \max_i |\hat{\sigma}_i| \frac{2}{\sqrt{m!}} 2^d,
\end{aligned}$$

where the last equality is Pascal's identity. □

We may now prove Theorem 3.14.

Proof of Theorem 3.14. Fix d and let $\mathcal{G}$ be the Gaussian process promised by Theorem 3.2, for p_d. By the triangle inequality,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \le 2\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) + 2\mathcal{WF}_2^2(\mathcal{P}_k p_d, \mathcal{G}) \le 2\mathcal{WF}_\infty^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) + 2\mathcal{WF}_2^2(\mathcal{P}_k p_d, \mathcal{G}).$$

We now invoke Lemma 3.1 with (3.8) to obtain,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{P}_k p_d) \le \|p_d - \sigma\|_{L^2(\gamma)}^2 \le R_\sigma(d).$$

For the other term, Theorem 3.2 along with Lemma 3.15 imply,

$$\mathcal{WF}_\infty^2(\mathcal{P}_k p_d, \mathcal{G}) \le \max_i |\hat{\sigma}_i|^2 \cdot d^{Cd} \frac{n^{2d}}{k^{\frac{1}{3}}},$$

for some numerical constant $C > 0$. So,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \le 2 \max_i |\hat{\sigma}_i|^2 d^{Cd} \frac{n^{2d}}{k^{\frac{1}{3}}} + 2R_\sigma(d).$$

Finally, choose $d = \lceil \frac{\log(k)}{100C \log(n) \log(\log(k))} \rceil$. It can be verified that for any $\delta > 0, \alpha > 0$,

$$d^{Cd} \cdot n^{2d} \le \log(k)^{\frac{\log(k)}{100 \log(\log(k))}} \cdot e^{\frac{\log(k)}{10}} = O(k^{\frac{1}{6}}).$$

This implies the existence of an absolute constant $C' > 0$, for which,

$$\mathcal{WF}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C' \left(\frac{\max_i |\hat{\sigma}_i|^2}{k^{\frac{1}{6}}} + R_\sigma(d) \right).$$

The proof is complete. □

3.5.1 ReLU activation

In this section we specialize Theorem 3.14 to the ReLU activation $\psi(x) := \max(0, x)$. The calculation of $\hat{\psi}_m$ may be found in [91, 126]. We repeat it here for completeness.

Lemma 3.16. *Let $m \in \mathbb{N}$. Then,*

$$|\hat{\psi}_m| = \begin{cases} \frac{1}{\sqrt{2}} & m = 1 \\ 0 & m > 1 \text{ and odd} \\ \frac{(m-3)!!}{\sqrt{\pi}\sqrt{m!}} & \text{otherwise} \end{cases}. \tag{3.9}$$

In particular, $|\hat{\psi}_m| \leq \frac{1}{m^{\frac{3}{2}}}$, and

$$R_\psi(d) \leq \frac{1}{d^2}.$$

Proof. Note that once (3.9) is established the rest of the proof is trivial. Thus, let us focus on calculating $\hat{\psi}_m$. We will use the following formula for the derivative of Hermite polynomials,

$$h_m'(x) = \sqrt{m} h_{m-1}(x). \tag{3.10}$$

Using this, we have, with an application of integration by parts,

$$\begin{aligned} \hat{\psi}_m = \int_{\mathbb{R}} h_m(x)\psi(x)d\gamma(x) &= \int_{x>0} h_m(x) x d\gamma(x) = \frac{h_m(0)}{\sqrt{2\pi}} - \int_{x>0} h_m'(x) d\gamma(x) \\ &= \frac{h_m(0)}{\sqrt{2\pi}} - \sqrt{m} \int_{x>0} h_{m-1}(x) d\gamma(x) \\ &= \frac{h_m(0)}{\sqrt{2\pi}} + (-1)^m \sqrt{\frac{m}{2\pi(m-1)!}} \int_{x>0} \frac{d^{m-1}}{dx^{m-1}} e^{-\frac{x^2}{2}}(x) dx \\ &= \frac{h_m(0)}{\sqrt{2\pi}} + (-1)^m \sqrt{\frac{m}{2\pi(m-1)!}} \frac{d^{m-2}}{dx^{m-2}} e^{-\frac{x^2}{2}}(0) \\ &= \frac{h_m(0) + \sqrt{\frac{m}{(m-1)}} h_{m-2}(0)}{\sqrt{2\pi}}. \end{aligned}$$

For $h_m(0)$, the following explicit formula holds:

$$h_m(0) = \begin{cases} 0 & \text{for } m \text{ odd} \\ (-1)^{m/2} \frac{(m-1)!!}{\sqrt{m!}} & \text{for } m \text{ even} \end{cases}.$$

In this case, for m even,

$$\sqrt{\frac{m}{(m-1)}} h_{m-2}(0) = (-1)^{m/2-1} \sqrt{\frac{m}{m-1}} \frac{(m-3)!!}{\sqrt{(m-2)!}} = (-1)^{m/2-1} \frac{m(m-3)!!}{\sqrt{(m-1)!}},$$

and (3.9) follows. □

Theorem 3.3 follows immediately, by plugging the above Lemma into Theorem 3.14.

Proof of Theorem 3.3. From Lemma 3.16 we see that $\max_i |\hat{\psi}_i| \leq 1$, and so coupled with Theorem 3.14, we get

$$\mathcal{W}_2^2(\mathcal{P}_k \sigma, \mathcal{G}) \leq C \left(\frac{1}{k^{\frac{1}{6}}} + \left(\frac{\log(n) \log(\log(k))}{\log(k)} \right)^2 \right).$$

It is now enough to observe,

$$\frac{1}{k^{\frac{1}{6}}} = O\left(\left(\frac{\log(n) \log(\log(k))}{\log(k)} \right)^2 \right).$$

□

3.5.2 Hyperbolic tangent activation

Let us now consider the function $\tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}}$ as an activation. Since it is smooth, we should expect it to have better polynomial approximations than the ReLU. This will lead to a faster convergence rate along the CLT. An explicit expression for $\widehat{\tanh}_m$ may be difficult to find. However, one may combine the smoothness of $\tanh$ with a classical result of Hille ([135]) in order to bound the coefficients from above.

This calculation was done in [202], where it was shown that for the derivative $|\widehat{\tanh'}_m| \leq e^{-C\sqrt{m}}$, where $C > 0$, does not depend on m. We now extend this result to $\tanh$.

Lemma 3.17. *Let $m \geq 0$. It holds that*

$$|\widehat{\tanh}_m| \leq e^{-C\sqrt{m}},$$

for some absolute constant $C > 0$.

Proof. Since $|\widehat{\tanh'}_m| \leq e^{-C\sqrt{m}}$, Hille's result ([135, Theorem 1]) shows that we have the point-wise equality,

$$\tanh'(x) = \sum_{m=0}^{\infty} \widehat{\tanh'}_m h_m(x).$$

We now use (3.10), and integrate the series, term by term, so that

$$\tanh(x) = \sum_{m=1}^{\infty} \frac{\widehat{\tanh'}_{m-1}}{\sqrt{m}} h_m(x).$$

So, $\widehat{\tanh}_m = \frac{\widehat{\tanh'}_{m-1}}{\sqrt{m}}$, which proves the claim. □

From the lemma, we get that there is some absolute constant $C > 0$, such that $R_{\tanh}(d) \leq e^{-C\sqrt{d}}$. This allows us to prove Theorem 3.4.

Proof of Theorem 3.4. From Lemma 3.17 along with Theorem 3.14, we get

$$\mathcal{W}_2^2(\mathcal{P}_k\sigma, \mathcal{G}) \leq C\left(\frac{1}{k^{\frac{1}{6}}} + \exp\left(-\frac{1}{C}\sqrt{\frac{\log(k)}{\log(n)\log(\log(k))}}\right)\right).$$

As before, the claim follows since,

$$\frac{1}{k^{\frac{1}{6}}} = O\left(\exp\left(-\frac{1}{C}\sqrt{\frac{\log(k)}{\log(n)\log(\log(k))}}\right)\right).$$

□

PART II

STABILITY OF FUNCTIONAL INEQUALITIES

"All are equal, but some are more equal than others."
- Paraphrasing pig Napoleon

4

Stability of the Shannon-Stam Inequality

4.1 Introduction

Let μ be a probability measure on $\mathbb{R}^d$ and $X \sim \mu$. Denote by $\mathrm{h}(\mu)$, the differential entropy of μ which is defined to be

$$\mathrm{h}(\mu) := \mathrm{h}(X) = -\int_{\mathbb{R}^d} \ln\left(\frac{d\mu}{dx}\right) d\mu.$$

One of the fundamental results of information theory is the celebrated Shannon-Stam inequality which asserts that for independent vectors X, Y and $\lambda \in (0,1)$

$$\mathrm{h}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\right) \geq \lambda \mathrm{h}(X) + (1-\lambda)\mathrm{h}(Y). \tag{4.1}$$

We remark that Stam [223] actually proved the equivalent statement

$$e^{\frac{2\mathrm{h}(X+Y)}{d}} \geq e^{\frac{2\mathrm{h}(X)}{d}} + e^{\frac{2\mathrm{h}(Y)}{d}}, \tag{4.2}$$

first observed by Shannon in [219], and known today as the entropy power inequality. To state yet another equivalent form of the inequality, for any positive-definite matrix, Σ, we set γ_Σ as

the centered Gaussian measure on $\mathbb{R}^d$ with density

$$\frac{d\gamma_\Sigma(x)}{dx} = \frac{e^{-\frac{\langle x,\Sigma^{-1}x\rangle}{2}}}{\sqrt{\det(2\pi\Sigma)}}.$$

For the case where the covariance matrix is the identity, I_d, we will also write $\gamma := \gamma_{\mathrm{I}_d}$. If $Y \sim \nu$ we set the relative entropy of X with respect to Y as

$$\mathrm{Ent}(\mu||\nu) := \mathrm{Ent}(X||Y) = \int_{\mathbb{R}^d} \ln\left(\frac{d\mu}{d\nu}\right) d\mu.$$

For $G \sim \gamma$, the differential entropy is related to the relative entropy by

$$\mathrm{Ent}(X||G) = -\mathrm{h}(X) - \frac{1}{2}\mathbb{E}\left[\|X\|_2^2\right] + \frac{d}{2}\ln(2\pi).$$

Thus, when X and Y are independent and centered the statement

$$\mathrm{Ent}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\middle||G\right) \leq \lambda\mathrm{Ent}(X||G) + (1-\lambda)\mathrm{Ent}(Y||G), \tag{4.3}$$

is equivalent to (4.1). Shannon noted that in the case that X and Y are Gaussians with proportional covariance matrices, both sides of (4.2) are equal. Later, in [223] it was shown that this is actually a necessary condition for the equality case. We define the deficit in (4.3) as

$$\delta_{EPI,\lambda}(\mu,\nu) := \delta_{EPI,\lambda}(X,Y) = \Big(\lambda\mathrm{Ent}(X||G)+(1-\lambda)\mathrm{Ent}(Y||G)\Big) - \mathrm{Ent}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\middle||G\right),$$

and are led to the question: *what can be said about X and Y when $\delta_{EPI,\lambda}(X,Y)$ is small?* One might expect that, in light of the equality cases, a small deficit in (4.3) should imply that X and Y are both close, in some sense, to a Gaussian. A recent line of works has focused on an attempt to make this intuition precise (see e.g., [84, 231]), which is also our main goal in the present work. In particular, we give the first stability estimate in terms of relative entropy. A good starting point is the work of Courtade, Fathi and Pananjady ([84]) which considers stability in terms of the Wasserstein distance (also known as quadratic transportation). A crucial observation made in their work is that without further assumptions on the measures μ and ν, one should not expect meaningful stability results to hold. Indeed, for any $\lambda \in (0,1)$ they show that there exists a family of measures $\{\mu_\varepsilon\}_{\varepsilon>0}$ such that $\delta_{EPI,\lambda}(\mu_\varepsilon,\mu_\varepsilon) < \varepsilon$ and such that for any Gaussian measure γ_Σ, $\mathcal{W}_2(\mu_\varepsilon,\gamma_\Sigma) \geq \frac{1}{3}$. Moreover, one may take μ_ε to be a mixture of Gaussians. Thus, in order to derive quantitative bounds it is necessary to consider a more restricted class of measures. We focus on the class of log-concave measures which, as our method demonstrates, turns out to be natural in this context.

Main Contributions

Our first result applies to uniformly log-concave vectors. Recall that, if there exists $\xi > 0$ such that

$$-\nabla^2 \ln(f(x)) \succeq \xi \mathrm{I}_d \text{ for all } x,$$

then we say that the measure is ξ-uniformly log-concave.

Theorem 4.1. *Let X and Y be 1-uniformly log-concave centered vectors, and denote by σ_X^2, σ_Y^2 the respective minimal eigenvalues of their covariance matrices. Then there exist Gaussian vectors G_X and G_Y such that for any $\lambda \in (0,1)$,*

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2}\left(\sigma_X^4 \mathrm{Ent}\left(X||G_X\right) + \sigma_Y^4 \mathrm{Ent}\left(Y||G_Y\right) + \frac{\sigma_X^4}{2}\mathrm{Ent}\left(G_X||G_Y\right) + \frac{\sigma_Y^4}{2}\mathrm{Ent}\left(G_Y||G_X\right)\right).$$

To compare this with the main result of [84] we recall the transportation-entropy inequality due to Talagrand ([229]) which states that

$$\mathcal{W}_2^2(X,G) \leq 2\mathrm{Ent}(X||G).$$

As a conclusion we get

$$\delta_{EPI,\lambda}(X,Y) \geq C_{\sigma_X,\sigma_Y}\frac{\lambda(1-\lambda)}{2}\left(\mathcal{W}_2^2\left(X,G_X\right) + \mathcal{W}_2^2\left(Y,G_Y\right) + \mathcal{W}_2^2\left(G_X,G_Y\right)\right),$$

where C_{σ_X,σ_Y} depends only on σ_X and σ_Y. Up to this constant, this is precisely the main result of [84]. In fact, our method can reproduce their exact result, which we present as a warm up in the next section. We remark that as the underlying inequality is of information-theoretic nature, it is natural to expect that stability estimates are expressed in terms of relative entropy.

A random vector is isotropic if it is centered and its covariance matrix is the identity. By a re-scaling argument the above theorem can be restated for uniform log-concave isotropic random vectors.

Corollary 4.2. *Let X and Y be ξ-uniformly log-concave and isotropic random vectors, then there exist Gaussian vectors G_X and G_Y such that for any $\lambda \in (0,1)$*

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2}\xi^2\left(\mathrm{Ent}\left(X||G_X\right) + \mathrm{Ent}\left(Y||G_Y\right) + \frac{1}{2}\mathrm{Ent}\left(G_X||G_Y\right) + \frac{1}{2}\mathrm{Ent}\left(G_Y||G_X\right)\right).$$

In our estimate for general log-concave vectors, the dependence on the parameter ξ will be replaced by the spectral gap of the measures. We say that a random vector X satisfies a Poincaré inequality if there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\mathrm{Var}(\psi(X))\right] \leq C\mathbb{E}\left[\|\nabla\psi(X)\|_2^2\right], \text{ for all test functions } \psi.$$

We define $\mathrm{C_p}(X)$ to be the smallest number such that the above equation holds with $C = \mathrm{C_p}(X)$, and refer to this quantity as the Poincaré constant of X. The inverse quantity, $\mathrm{C_p}(X)^{-1}$ is referred to as the *spectral gap* of X.

Theorem 4.3. *Let X and Y be centered log-concave vectors with σ_X^2, σ_Y^2 denoting the minimal eigenvalues of their covariance matrices. Assume that $\mathrm{Cov}(X) + \mathrm{Cov}(Y) = 2\mathrm{I}_d$ and set $\max\left(\frac{\mathrm{C_p}(X)}{\sigma_X^2}, \frac{\mathrm{C_p}(Y)}{\sigma_Y^2}\right) = \mathrm{C_p}$. Then, if G denotes the standard Gaussian, for every $\lambda \in (0,1)$*

$$\delta_{EPI,\lambda}(X,Y) \geq K\lambda(1-\lambda)\left(\frac{\min(\sigma_Y^2,\sigma_X^2)}{\mathrm{C_p}}\right)^3 \left(\mathrm{Ent}\left(X||G\right) + \mathrm{Ent}\left(Y||G\right)\right),$$

where $K > 0$ is a numerical constant, which can be made explicit.

Remark 4.4. For ξ-uniformly log-concave vectors, we have the relation, $\mathrm{C_p}(X) \leq \frac{1}{\xi}$ (this is a consequence of the Brascamp-Lieb inequality [50], for instance). Thus, considering Corollary 4.2, one might have expected that the term $\mathrm{C_p^3}$ could have been replaced by $\mathrm{C_p^2}$ in Theorem 4.3. We do not know if either result is tight.

Remark 4.5. Bounding the Poincaré constant of an isotropic log-concave measure is the object of the long standing Kannan-Lováz-Simonovits (KLS) conjecture (see [147, 160] for more information). The conjecture asserts that there exists a constant $K > 0$, independent of the dimension, such that for any isotropic log-concave vector X, $\mathrm{C_p}(X) \leq K$. The best known bound is due to Chen which showed in [72] that if X is a a d-dimensional log-concave vector, $\mathrm{C_p}(X) = O\left(d^{o_d(1)}\right)$.

Concerning the assumptions of Theorem 4.3; note that as the EPI is invariant to linear transformation, there is no loss in generality in assuming $\mathrm{Cov}(X) + \mathrm{Cov}(Y) = 2\mathrm{I}_d$. Remark that $\mathrm{C_p}(X)$ is, approximately, proportional to the maximal eigenvalue of $\mathrm{Cov}(X)$. Thus, for ill-conditioned covariance matrices $\frac{\mathrm{C_p}(X)}{\sigma_X^2}, \frac{\mathrm{C_p}(Y)}{\sigma_Y^2}$ will not be on the same scale. It seems plausible to conjecture that the dependence on the minimal eigenvalue and Poicnaré constant could be replaced by a quantity which would take into consideration all eigenvalues.

Some other known stability results, both for log-concave vectors and for other classes of measures, may be found in [83, 84, 231]. The reader is referred to [84, Section 2.2] for a complete discussion. Let us mention one important special case, which is relevant to our results; the so-called entropy jump, first proved for the one dimensional case by Ball, Barthe and Naor ([20]) and then generalized by Ball and Nguyen to arbitrary dimensions in [21]. According to the latter result, if X is a log-concave and isotropic random vector, then

$$\delta_{EPI,\frac{1}{2}}(X,X) \geq \frac{1}{8\mathrm{C_p}(X)}\mathrm{Ent}(X||G),$$

where $\mathrm{C_p}(X)$ is the Poincaré constant of X and G is the standard Gaussian. This should be compared to both Corollary 4.2 and Theorem 4.3. That is, in the special case of two identical

measures and $\lambda = \frac{1}{2}$, their result gives a better dependence on the Poincaré constant than the one afforded by our results.

Ball and Nguyen ([21]) also give an interesting motivation for these type of inequalities: They show that if for some constant $\kappa > 0$,

$$\delta_{EPI,\frac{1}{2}}(X, X) \geq \kappa \mathrm{Ent}(X||G),$$

then the density f_X of X satisfies, $f_X(0) \leq e^{\frac{2d}{\kappa}}$. The isotropic constant of X is defined by $L_X := f_X(0)^{\frac{1}{d}}$, and is the main subject of the slicing conjecture, which hypothesizes that L_X is uniformly bounded by a constant, independent of the dimension, for every isotropic log-concave vector X. Ball and Nguyen observed that using the above fact in conjunction with an entropy jump estimate gives a bound on the isotropic constant in terms of the Poincaré constant, and in particular the slicing conjecture is implied by the KLS conjecture.

Using ideas originating from additive combinatorics, the setting of the entropy jump is further investigated in [156, 172].

Our final results give improved bounds under the assumption that X and Y are already close to being Gaussian, in terms of relative entropy, or if one them is a Gaussian. We record these results in the following theorems.

Theorem 4.6. *Suppose that X, Y be isotropic log-concave vectors such that $\mathrm{C_p}(X), \mathrm{C_p}(Y) \leq \mathrm{C_p}$ for some $\mathrm{C_p} < \infty$. Suppose further that $\mathrm{Ent}(X||G), \mathrm{Ent}(Y||G) \leq \frac{1}{4}$, then*

$$\delta_{EPI,\lambda}(X, Y) \geq \frac{\lambda(1-\lambda)}{36\mathrm{C_p}} \left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right)$$

The following gives an improved bound in the case that one of the random vectors is a Gaussian, and holds in full generality with respect to the other vector, without a log-concavity assumption.

Theorem 4.7. *Let X be a centered random vector with finite Poincaré constant, $\mathrm{C_p}(X) < \infty$. Then*

$$\delta_{EPI,\lambda}(X, G) \geq \left(\lambda - \frac{\lambda\left(\mathrm{C_p}(X) - 1\right) - \ln\left(\lambda\left(\mathrm{C_p}(X) - 1\right) + 1\right)}{\mathrm{C_p}(X) - \ln\left(\mathrm{C_p}(X)\right) - 1}\right) \mathrm{Ent}(X||G).$$

Remark 4.8. When $\mathrm{C_p}(X) \geq 1$, the following inequality holds

$$\left(\lambda - \frac{\lambda\left(\mathrm{C_p}(X) - 1\right) - \ln\left(\lambda\left(\mathrm{C_p}(X) - 1\right) + 1\right)}{\mathrm{C_p}(X) - \ln\left(\mathrm{C_p}(X)\right) - 1}\right) \geq \frac{\lambda(1-\lambda)}{\mathrm{C_p}(X)}.$$

Remark 4.9. Theorem 4.7 was already proved in [84] by using a slightly different approach. Denote by $\mathrm{I}(X||G)$, the relative Fisher information of the random vector X. In [113] the authors

proof the following improved log-Sobolev inequality.

$$\mathrm{I}(X||G) \geq 2\mathrm{Ent}(X||G)\frac{(1-\mathrm{C_p}(X))^2}{\mathrm{C_p}(X)(\mathrm{C_p}(X)-\ln(\mathrm{C_p}(X)-1))}.$$

The theorem follows by integrating the inequality along the Ornstein-Uhlenbeck semi-group.

4.2 Bounding the deficit via martingale embeddings

Our approach is based on ideas somewhat related to the ones which appear in Chapter 1: the very high-level plan of the proof is to embed the variables X, Y as the terminal points of some martingales and express the entropies of X, Y and $X+Y$ as functions of the associates quadratic co-variation processes. One of the main benefits in using such an embedding is that the co-variation process of $X+Y$ can be easily expressed in terms on the ones of X, Y, as demonstrated below. In Chapter 1 these ideas where used to produce upper bounds for the entropic central limit theorem, so it stands to reason that related methods may be useful here. It turns out, however, that in order to produce meaningful bounds for the Shannon-Stam inequality, one needs a more intricate analysis, since this inequality corresponds to a second-derivative phenomenon: whereas for the CLT one only needs to produce upper bounds on the relative entropy, here we need to be able to compare, in a non-asymptotic way, two relative entropies.

In particular, our martingale embedding is constructed through the Föllmer process, defined by (7) in the introduction. This construction has several useful features, one of which is that it allows us to express the relative entropy of a measure in $\mathbb{R}^d$ in terms of a variational problem on the Wiener space. In addition, upon attaining a slightly different point of view on this process, that we introduce here, the behavior of this variational expression turns out to be tractable with respect to convolutions.

In order to outline the argument, fix centered measures μ and ν on $\mathbb{R}^d$ with finite second moment. Let $X \sim \mu, Y \sim \nu$ be random vectors and $G \sim \gamma$ a standard Gaussian random vector.

An entropy-minimizing drift. Let B_t be a standard Brownian motion on $\mathbb{R}^d$ and denote by $\mathcal{F}_t$ its natural filtration. In the sequel, we denote v_t^X, to be the Föllmer drift associated to X (and the same for Y) and,

$$X_t := B_t + \int_0^t v_s^X ds.$$

Let us recall some of the properties which were proven in the introduction to this thesis.

It turns out that the process v_t^X is a martingale (which goes together with the fact that it

minimizes a quadratic form) which is given by the equation

$$v_t^X = \nabla_x \ln \left(P_{1-t}(f_X(X_t)) \right), \tag{4.4}$$

where f_X is the density of X with respect to the standard Gaussian and P_{1-t} denotes the heat semi-group. Moreover, from Girsanov's formula,

$$\mathrm{Ent}(X||G) = \frac{1}{2} \int_0^1 \mathbb{E}\left[\left\| v_t^X \right\|_2^2 \right] dt. \tag{4.5}$$

Another important fact concerns the marginal X_t, which follows a rather simple law,

$$X_t \stackrel{d}{=} tX_1 + \sqrt{t(1-t)}G. \tag{4.6}$$

Lehec's proof of the Shannon-Stam inequality. For the sake of intuition, we now repeat Lehec's argument to reproduce the Shannon-Stam inequality (4.3) using this process. Let $X_t := B_t^X + \int_0^t v_s^X ds$ and $Y_t := B_t^Y + \int_0^t v_s^Y ds$ be the Föllmer processes associated to X and Y, where B_t^X and B_t^Y are independent Brownian motions. For $\lambda \in (0,1)$, define the new processes

$$w_t = \sqrt{\lambda} v_t^X + \sqrt{1-\lambda} v_t^Y,$$

and

$$\tilde{B}_t = \sqrt{\lambda} B_t^X + \sqrt{1-\lambda} B_t^Y.$$

By the independence of B_t^X and B_t^Y, $\tilde{B}_t$ is a Brownian motion and

$$\tilde{B}_1 + \int_0^1 w_t dt = \sqrt{\lambda} X_1 + \sqrt{1-\lambda} Y_1.$$

Note that as the v_t^X is martingale, we have for every $t \in [0,1]$,

$$\mathbb{E}\left[v_t^X \right] = \mathbb{E}\left[X_1 \right] = 0.$$

From the bound on relative entropy in (1.21) coupled with (4.5) and the independence of the

processes, we have,

$$\begin{aligned}\mathrm{Ent}(\sqrt{\lambda}X_1+\sqrt{1-\lambda}Y_1||G) &\leq \frac{1}{2}\int_0^1 \mathbb{E}\left[\|w_t\|_2^2\right]dt \\ &= \frac{\lambda}{2}\int \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]dt + \frac{1-\lambda}{2}\int \mathbb{E}\left[\left\|v_t^Y\right\|_2^2\right]dt \\ &= \lambda \mathrm{Ent}(X_1||G) + (1-\lambda)\mathrm{Ent}(Y_1||G).\end{aligned}$$

This recovers the Shannon-Stam inequality in the form (4.3).

An alternative point of view: Replacing the drift by a varying diffusion coefficient. Lehec's proof gives rise to the following idea: Suppose the processes v_t^X and v_t^Y could be coupled in a way such that the variance of the resulting process $\sqrt{\lambda}v_t^X+\sqrt{1-\lambda}v_t^Y$ was smaller than that of w_t above. Such a coupling would improve on (4.3) and that is the starting point of this work.

As it turns out, however, it is easier to get tractable bounds by working with a slightly different interpretation of the above processes, in which the role of the drift is taken by an adapted diffusion coefficient of a related process.

The idea is as follows: Suppose that $M_t := \int_0^t F_s dB_s$ is a martingale, where F_t is some positive-definite matrix valued process adapted to $\mathcal{F}_t$. Consider the drift defined by

$$u_t := \int_0^t \frac{F_s - \mathrm{I}_d}{1-s} dB_s. \tag{4.7}$$

We then claim that $B_1 + \int_0^1 u_t dt = M_1$. To show this, we use the stochastic Fubini Theorem ([238]) to write

$$\int_0^1 F_t dB_t = \int_0^1 \mathrm{I}_d dB_t + \int_0^1 (F_t - \mathrm{I}_d)\, dB_t = B_1 + \int_0^1\int_t^1 \frac{F_t-\mathrm{I}_d}{1-t} ds dB_t = B_1 + \int_0^1 u_t dt.$$

Since we now expressed the random variable M_1 as the terminal point of a standard Brownian motion with an adapted drift, the minimality property of the Föllmer drift together with equation (4.5) immediately produce a bound on its entropy. Namely, by using Itô's isometry and Fubini's theorem we have the bound

$$\mathrm{Ent}(M_1||G) \overset{(4.5)}{\leq} \frac{1}{2}\int_0^1 \mathbb{E}\left[\|u_t\|_2^2\right] = \frac{1}{2}\mathrm{Tr}\int_0^1\int_0^t \frac{\mathbb{E}\left[(F_s-\mathrm{I}_d)^2\right]}{(1-s)^2} ds dt = \frac{1}{2}\mathrm{Tr}\int_0^1 \frac{\mathbb{E}\left[(F_t-\mathrm{I}_d)^2\right]}{1-t} dt. \tag{4.8}$$

This hints at the following possible scheme of proof: in order to give an upper bound for the

expression $\mathrm{Ent}(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1||G)$, it suffices to find martingales M_t^X and M_t^Y such that M_1^X, M_1^Y have the laws of X and Y, respectively, and such that the λ-average of the covariance processes is close to the identity.

The Föllmer process gives rise to a natural martingale: Consider $\mathbb{E}[X_1|\mathcal{F}_t]$, the associated Doob martingale. By the martingale representation theorem ([199, Theorem 4.3.3]) there exists a uniquely defined adapted matrix valued process Γ_t^X, for which

$$\mathbb{E}[X_1|\mathcal{F}_t] = \int_0^t \Gamma_s^X dB_s^X. \tag{4.9}$$

We will require the following identity, which appeared in (10),

$$v_t^X = \int_0^t \frac{\Gamma_s^X - \mathrm{I}_d}{1-s} dB_s^X. \tag{4.10}$$

The matrix Γ_t^X turns out to be positive definite almost surely, (in fact, it has an explicit simple representation, see Proposition 4.12 below), which yields, as in (11),

$$\mathrm{Ent}(X||G) = \frac{1}{2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_s^X - \mathrm{I}_d\right)^2\right]\right)}{1-t} dt. \tag{4.11}$$

Given the processes Γ_t^X and Γ_t^Y, we are now in position to express $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$ as the terminal point of a martingale, towards using (4.8), which would lead to a bound on $\delta_{EPI,\lambda}$. We define

$$\tilde{\Gamma}_t := \sqrt{\lambda\left(\Gamma_t^X\right)^2 + (1-\lambda)\left(\Gamma_t^Y\right)^2},$$

and a martingale $\tilde{B}_t$ which satisfies

$$\tilde{B}_0 = 0 \text{ and } d\tilde{B}_t = \tilde{\Gamma}_t^{-1}\left(\sqrt{\lambda}\Gamma_t^X dB_t^X + \sqrt{1-\lambda}\Gamma_t^Y dB_t^Y\right).$$

Since Γ_t^X and Γ_t^Y are invertible almost surely and independent, it holds that

$$[\tilde{B}]_t = t\mathrm{I}_d,$$

where $[\tilde{B}]_t$ denotes the quadratic co-variation of $\tilde{B}_t$. Thus, by Levy's characterization, $\tilde{B}_t$ is a standard Brownian motion and we have the following equality in law

$$\int_0^1 \tilde{\Gamma}_t d\tilde{B}_t = \sqrt{\lambda}\int_0^1 \Gamma_t^X dB_t^X + \sqrt{1-\lambda}\int_0^1 \Gamma_t^Y dB_t^Y \stackrel{d}{=} \sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1.$$

We can now invoke (4.8) to get

$$\mathrm{Ent}\left(\sqrt{\lambda}X_1+\sqrt{1-\lambda}Y_1||G\right)\leq\frac{1}{2}\int_0^1\frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\tilde{\Gamma}_t-\mathrm{I}_d\right)^2\right]\right)}{1-t}dt.$$

Combining this with the identity (4.11) finally gives a bound on the deficit in the Shannon-Stam inequality, in the form

$$\begin{aligned}\delta_{EPI,\lambda}(X,Y)&\geq\frac{1}{2}\int_0^1\frac{\mathrm{Tr}\left(\lambda\mathbb{E}\left[\left(\Gamma_t^X-\mathrm{I}_d\right)^2\right]+(1-\lambda)\mathbb{E}\left[\left(\Gamma_t^Y-\mathrm{I}_d\right)^2\right]-\mathbb{E}\left[\left(\tilde{\Gamma}_t-\mathrm{I}_d\right)^2\right]\right)}{1-t}dt\\&=\int_0^1\frac{\mathrm{Tr}\left(\mathbb{E}\left[\tilde{\Gamma}_t\right]-\lambda\mathbb{E}\left[\Gamma_t^X\right]-(1-\lambda)\mathbb{E}\left[\Gamma_t^Y\right]\right)}{1-t}dt.\end{aligned}\tag{4.12}$$

The following technical lemma will allow us to give a lower bound for the right hand side in terms of the variances of the processes Γ_t^X,Γ_t^Y. Its proof is postponed to the end of the section.

Lemma 4.10. *Let A and B be positive definite matrices and denote*

$$(A,B)_\lambda:=\lambda A+(1-\lambda)B \text{ and } (A^2,B^2)_\lambda:=\lambda A^2+(1-\lambda)B^2.$$

Then

$$\mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda}-(A,B)_\lambda\right)=\lambda(1-\lambda)\mathrm{Tr}\left((A-B)^2\left(\sqrt{(A^2,B^2)_\lambda}+(A,B)_\lambda\right)^{-1}\right).$$

Combining the lemma with the estimate obtained in (4.12) produces the following result, which will be our main tool in studying $\delta_{EPI,\lambda}$.

Lemma 4.11. *Let X and Y be centered random vectors on $\mathbb{R}^d$ with finite second moment, and let Γ_t^X,Γ_t^Y be defined as above. Then,*

$$\begin{aligned}&\delta_{EPI,\lambda}(X,Y)\geq\\&\lambda(1-\lambda)\int_0^1\frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X-\Gamma_t^Y\right)^2\left(\sqrt{\lambda\left(\Gamma_t^X\right)^2+(1-\lambda)\left(\Gamma_t^Y\right)^2}+\lambda\Gamma_t^X+(1-\lambda)\Gamma_t^Y\right)^{-1}\right]\right)}{1-t}dt.\end{aligned}\tag{4.13}$$

The expression on the right-hand side of (4.13) may seem unwieldy, however, in many cases it can be simplified. For example, if it can be shown that, almost surely, $\Gamma_t^X,\Gamma_t^Y\preceq c_t\mathrm{I}_d$ for some

deterministic $c_t > 0$, then we obtain the more tractable inequality

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2} \int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]\right)}{(1-t)c_t} dt. \tag{4.14}$$

As we will show, this is the case when the random vectors are log-concave.

Proof of Lemma 4.10. We have

$$\begin{aligned}
&\mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda} - (A,B)_\lambda\right) \\
&= \mathrm{Tr}\left(\left(\sqrt{(A^2,B^2)_\lambda} - (A,B)_\lambda\right)\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right).
\end{aligned}$$

As

$$\begin{aligned}
&\left(\sqrt{(A^2,B^2)_\lambda} - (A,B)_\lambda\right)\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right) \\
&= \lambda(1-\lambda)\left(A^2 + B^2 - AB - BA\right) + \sqrt{(A^2,B^2)_\lambda}(A,B)_\lambda - (A,B)_\lambda\sqrt{(A^2,B^2)_\lambda},
\end{aligned}$$

we have the equality

$$\begin{aligned}
\mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda} - (A,B)_\lambda\right) &= \lambda(1-\lambda)\mathrm{Tr}\left(\left(A^2 + B^2 - (AB+BA)\right)\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right) \\
&\quad + \mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda}(A,B)_\lambda\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right) \\
&\quad - \mathrm{Tr}\left((A,B)_\lambda\sqrt{(A^2,B^2)_\lambda}\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right)
\end{aligned}$$

Finally, as the trace is invariant under any permutation of three symmetric matrices we have that

$$\mathrm{Tr}\left(AB\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right) = \mathrm{Tr}\left(BA\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right),$$

and

$$\begin{aligned}
&\mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda}(A,B)_\lambda\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right) \\
&\qquad = \mathrm{Tr}\left((A,B)_\lambda\sqrt{(A^2,B^2)_\lambda}\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right).
\end{aligned}$$

Thus,

$$\mathrm{Tr}\left(\sqrt{(A^2,B^2)_\lambda} - (A,B)_\lambda\right) = \lambda(1-\lambda)\mathrm{Tr}\left(\left((A-B)^2\right)\left(\sqrt{(A^2,B^2)_\lambda} + (A,B)_\lambda\right)^{-1}\right),$$

as required. □

4.2.1 The Föllmer process associated to log-concave random vectors

In this section, we collect several results pertaining to the Föllmer process. Throughout the section, we fix a random vector X in $\mathbb{R}^n$ and associate to it the Föllmer process X_t, defined in the previous section, as well as the process Γ_t^X, defined in equation (4.9) above. The next result lists some of its basic properties, and we refer to [100, 106] for proofs.

Proposition 4.12. *For $t \in (0,1)$ define*

$$f_X^t(x) := f_X(x) \exp\left(\frac{\|x - X_t\|_2^2}{2(1-t)}\right) Z_{t,X}^{-1},$$

where f_X is the density of X with respect to the standard Gaussian and $Z_{t,X}$ is a normalizing constant defined so that $\int_{\mathbb{R}^d} f_X^t = 1$. Then

- *f_X^t is the density of the random measure $\mu_t := X_1|\mathcal{F}_t$ with respect to the standard Gaussian and $\Gamma_t^X = \frac{\mathrm{Cov}(\mu_t)}{1-t}$.*
- *Γ_t^X is almost surely a positive definite matrix, in particular, it is invertible.*
- *For all $t \in (0,1)$, we have*

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t^X\right] = \frac{\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\left(\Gamma_t^X\right)^2\right]}{1-t}. \tag{4.15}$$

- *The following identity holds*

$$\mathbb{E}\left[v_t^X \otimes v_t^X\right] = \frac{\mathrm{I}_d - \mathbb{E}\left[\Gamma_t^X\right]}{1-t} + \mathrm{Cov}(X) - \mathrm{I}_d, \tag{4.16}$$

for all $t \in [0,1]$. In particular, if $\mathrm{Cov}(X) \preceq \mathrm{I}_d$, then $\mathbb{E}\left[\Gamma_t^X\right] \preceq \mathrm{I}_d$.

In what follows, we restrict ourselves to the case that X is log-concave. Using this assumption we will establish several important properties for the matrix Γ_t. For simplicity, we will write $\Gamma_t := \Gamma_t^X$ and $v_t := v_t^X$. The next result shows that the matrix Γ_t is bounded almost surely, this is essentially a generalization of Lemma 1.30 from Chapter 1 and the prove is similar. We repeat it here for completeness.

Lemma 4.13. *Suppose that X is log-concave, then for every $t \in (0,1)$*

$$\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d.$$

Moreover, if for some $\xi > 0$, X is ξ-uniformly log-concave then

$$\Gamma_t \preceq \frac{1}{(1-t)\xi + t} \mathrm{I}_d.$$

Proof. By Proposition 4.12, μ_t, the law of $X_1|\mathcal{F}_t$ has a density ρ_t, with respect to the Lebesgue measure, proportional to

$$f_X(x) \exp\left(\frac{\|x\|_2^2}{2}\right) \exp\left(-\frac{\|x - X_t\|_2^2}{2(1-t)}\right) = f_X(x) \exp\left(\frac{\|x\|_2^2 (1-t) - \|x - X_t\|_2^2}{2(1-t)}\right).$$

Consequently, since $-\nabla^2 f_X \succeq 0$,

$$-\nabla^2 \ln(\rho_t) = -\nabla^2 f_X - \left(1 - \frac{1}{1-t}\right) \mathrm{I}_d \succeq \frac{t}{1-t} \mathrm{I}_d.$$

It follows that, almost surely, μ_t is $\frac{t}{1-t}$-uniformly log-concave. According to the Brascamp-Lieb inequality ([50]) α-uniform log-concavity implies a spectral gap of α, and in particular $\mathrm{Cov}(\mu_t) \preceq \frac{1-t}{t} \mathrm{I}_d$ and so, $\Gamma_t = \frac{\mathrm{Cov}(\mu_t)}{1-t} \preceq \frac{1}{t} \mathrm{I}_d$. If, in addition, X is ξ-uniformly log-concave, so that $-\nabla^2 f_X \succeq \xi \mathrm{I}_d$, then we may write

$$-\nabla^2 \ln(\rho_t) \succeq \left(\xi + \frac{t}{1-t}\right) \mathrm{I}_d = \frac{(1-t)\xi + t}{(1-t)} \mathrm{I}_d$$

and the arguments given above show $\mathrm{Cov}(\mu_t) \preceq \frac{(1-t)}{(1-t)\xi + t} \mathrm{I}_d$. Thus,

$$\Gamma_t \preceq \frac{1}{(1-t)\xi + t} \mathrm{I}_d.$$

□

Our next goal is to use the formulas given in the above lemma in order to bound from below the expectation of Γ_t. We begin with a simple corollary.

Corollary 4.14. *Suppose that X is 1-uniformly log-concave, then for every $t \in [0,1]$*

$$\mathbb{E}[\Gamma_t] \succeq \mathrm{Cov}(X).$$

Proof. By (4.15), we have

$$\frac{d}{dt} \mathbb{E}[\Gamma_t] = \frac{\mathbb{E}[\Gamma_t] - \mathbb{E}[\Gamma_t^2]}{1-t}.$$

By Lemma 4.13, $\Gamma_t \preceq \mathrm{I}_d$, which shows

$$\frac{d}{dt} \mathbb{E}[\Gamma_t] \succeq 0.$$

Thus, for every t,

$$\mathbb{E}\left[\Gamma_t\right] \succeq \mathbb{E}\left[\Gamma_0\right] = \mathrm{Cov}(X|\mathcal{F}_0) = \mathrm{Cov}(X).$$

□

To produce similar bounds for general log-concave random vectors, we require more intricate arguments. Recall that $\mathrm{C_p}(X)$ denotes the Poincaré constant of X.

Lemma 4.15. *If X is centered and has a finite a Poincaré constant $\mathrm{C_p}(X) < \infty$, then*

$$\mathbb{E}\left[v_t^{\otimes 2}\right] \preceq \left(t^2 \mathrm{C_p}(X) + t(1-t)\right) \frac{d}{dt} \mathbb{E}\left[v_t^{\otimes 2}\right].$$

Proof. Recall that, by equation (4.6), we know that X_t has the same law as $tX_1 + \sqrt{t(1-t)}G$, where G is a standard Gaussian independent of X_1. Since $\mathrm{C_p}(tX) = t^2 \mathrm{C_p}(X)$ and since the Poincaré constant is sub-additive with respect to convolution ([82]) we get

$$\mathrm{C_p}(X_t) \leq t^2 \mathrm{C_p}(X) + t(1-t).$$

The drift, v_t, is a function of X_t and $\mathbb{E}\left[v_t\right] = 0$. Equation (4.4) implies that $\nabla_x v_t(X_t)$ is a symmetric matrix, hence the Poincaré inequality yields

$$\mathbb{E}\left[v_t^{\otimes 2}\right] \preceq \left(t^2 \mathrm{C_p}(X) + t(1-t)\right) \mathbb{E}\left[\nabla_x v_t(X_t)^2\right].$$

As $v_t(X_t)$ is a martingale, by Itô's lemma we have

$$dv_t(X_t) = \nabla_x v_t(X_t) dB_t.$$

An application of Itô's isometry then shows

$$\mathbb{E}\left[\nabla_x v_t(X_t)^2\right] = \frac{d}{dt} \mathbb{E}\left[v_t(X_t)^{\otimes 2}\right],$$

where we have again used the fact that $\nabla_x v_t(X_t)$ is symmetric. □

Using the last lemma, we can deduce lower bounds on the matrix Γ_t^X in terms of the Poincaré constant.

Corollary 4.16. *Suppose that X is log-concave and that σ^2 is the minimal eigenvalue of $\mathrm{Cov}(X)$. Then,*

- *For every $t \in \left[0, \frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma^2}+1}\right]$, $\mathbb{E}\left[\Gamma_t\right] \succeq \frac{\min(1,\sigma^2)}{3} \mathrm{I}_d$.*
- *For every $t \in \left[\frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma^2}+1}, 1\right]$, $\mathbb{E}\left[\Gamma_t\right] \succeq \frac{\min(1,\sigma^2)}{3} \frac{1}{t\left(2\frac{\mathrm{C_p}(X)}{\sigma^2}+1\right)} \mathrm{I}_d$.*

Proof. Using Equation (4.10), Itô's isometry and the fact that Γ_t is symmetric, we deduce that

$$\frac{d}{dt}\mathbb{E}\left[v_t^{\otimes 2}\right] = \mathbb{E}\left[\left(\frac{\Gamma_t - \mathrm{I}_d}{1-t}\right)^2\right],$$

Combining this with equation (4.16) and using Lemma 4.15, we get

$$\mathrm{Cov}(X) - \mathrm{I}_d + \frac{\mathrm{I}_d - \mathbb{E}\left[\Gamma_t\right]}{1-t} \preceq \left(t^2\mathrm{C_p}(X) + t(1-t)\right)\frac{\mathbb{E}\left[\Gamma_t^2\right] - 2\mathbb{E}\left[\Gamma_t\right] + \mathrm{I}_d}{(1-t)^2}. \tag{4.17}$$

In the case where X is log-concave, by Lemma 4.13, $\Gamma_t \preceq \frac{1}{t}\mathrm{I}_d$ almost surely, therefore $\mathbb{E}\left[\Gamma_t^2\right] \preceq \frac{1}{t}\mathbb{E}\left[\Gamma_t\right]$. The above inequality then becomes

$$\begin{aligned}(1-t)^2\left(\sigma^2 - 1\right)\mathrm{I}_d + (1-t)(\mathrm{I}_d - \mathbb{E}\left[\Gamma_t\right]) \\ \preceq \left(t\mathrm{C_p}(X) + (1-t)\right)\mathbb{E}\left[\Gamma_t\right] + \left(t^2\mathrm{C_p}(X) + t(1-t)\right)\left(\mathrm{I}_d - 2\mathbb{E}\left[\Gamma_t\right]\right).\end{aligned}$$

Rearranging the inequality shows

$$\frac{\sigma^2 - 2t\sigma^2 - \mathrm{C_p}(X)t^2 \ + t^2\sigma^2}{2 - 4t - 2\mathrm{C_p}(X)t^2 + \mathrm{C_p}(X)t + 2t^2}\mathrm{I}_d \preceq \mathbb{E}\left[\Gamma_t\right].$$

As long as $t \leq \frac{1}{2\left(\frac{\mathrm{C_p}(X)}{\sigma^2}\right)+1}$, we have

$$\begin{aligned}\text{if } \sigma^2 \geq 1, \quad \frac{1}{3}\mathrm{I}_d \preceq \frac{\sigma^2\left(4\mathrm{C_p}(X) - \sigma^2\right)}{2\mathrm{C_p}(X)(\sigma^2+4) - \sigma^4}\mathrm{I}_d \preceq \mathbb{E}\left[\Gamma_t\right], \\ \text{if } \sigma^2 < 1, \quad \frac{\sigma^2}{3}\mathrm{I}_d \preceq \frac{\sigma^2\left(4\mathrm{C_p}(X) - \sigma^2\right)}{2\mathrm{C_p}(X)(\sigma^2+4) - \sigma^4}\mathrm{I}_d \preceq \mathbb{E}\left[\Gamma_t\right],\end{aligned}$$

which gives the first bound. By (4.9), we also have the bound

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right] = \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t} \succeq \frac{1-\frac{1}{t}}{1-t}\mathbb{E}\left[\Gamma_t\right] = -\frac{1}{t}\mathbb{E}\left[\Gamma_t\right].$$

The differential equation

$$g'(t) = -\frac{g(t)}{t}, g\left(\frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma^2}+1}\right) = \frac{\min(1,\sigma^2)}{3}$$

has a unique solution given by

$$g(t) = \frac{\min(1,\sigma^2)}{3}\frac{1}{t\left(2\frac{\mathrm{C_p}(X)}{\sigma^2}+1\right)}.$$

Using Gromwall's inequality, we conclude that for every $t \in \left[\frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma^2}+1}, 1\right]$,

$$\mathbb{E}\left[\Gamma_t\right] \succeq \frac{\min(1,\sigma^2)}{3} \frac{1}{t\left(2\frac{\mathrm{C_p}(X)}{\sigma^2}+1\right)} \mathrm{I}_d.$$

□

We conclude this section with a comparison lemma that will allow to control the values of $\mathbb{E}\left[\|v_t\|_2^2\right]$.

Lemma 4.17. *Let $t_0 \in [0,1]$ and suppose that X is centered with a finite Poincaré constant $\mathrm{C_p}(X) < \infty$. Then*

- *For $t_0 \le t \le 1$,*
$$\mathbb{E}\left[\|v_t\|_2^2\right] \ge \mathbb{E}\left[\|v_{t_0}\|_2^2\right] \frac{t_0\left(\mathrm{C_p}(X)-1\right)t + t}{t_0\left(\mathrm{C_p}(X)-1\right)t + t_0}.$$
- *For $0 \le t \le t_0$,*
$$\mathbb{E}\left[\|v_t\|_2^2\right] \le \mathbb{E}\left[\|v_{t_0}\|_2^2\right] \frac{t_0\left(\mathrm{C_p}(X)-1\right)t + t}{t_0\left(\mathrm{C_p}(X)-1\right)t + t_0}.$$

Proof. Consider the differential equation

$$g(t) = \left(\mathrm{C_p}(X)t^2 + t(1-t)\right) g'(t) \text{ with initial condition } g(t_0) = \mathbb{E}\left[\|v_{t_0}\|_2^2\right].$$

It has a unique solution given by

$$g(t) = \mathbb{E}\left[\|v_{t_0}\|_2^2\right] \frac{t_0\left(\mathrm{C_p}(X)-1\right)t + t}{t_0\left(\mathrm{C_p}(X)-1\right)t + t_0}.$$

The bounds follow by applying Gromwall's inequality combined with the result of Lemma 4.15.
□

4.3 Stability for uniformly log-concave random vectors

In this section, we assume that X and Y are both 1-uniformly log-concave. Let B_t^X, B_t^Y be independent standard Brownian motions and consider the associated processes Γ_t^X, Γ_t^Y defined as in Section 4.2.

The key fact that makes the uniform log-concave case easier is Lemma 4.13, which implies

that $\Gamma_t^X, \Gamma_t^Y \preceq \mathrm{I}_d$ almost surely. In this case, Lemma 4.11 simplifies to

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2} \int_0^1 \left(\frac{\mathrm{Tr}\left(\mathrm{Var}(\Gamma_t^X)\right)}{1-t} + \frac{\mathrm{Tr}\left(\mathrm{Var}(\Gamma_t^Y)\right)}{1-t} + \frac{\mathrm{Tr}\left(\left(\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\Gamma_t^Y\right]\right)^2\right)}{1-t} \right) dt, \tag{4.18}$$

where we have used the fact that

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]\right) = \mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X - \mathbb{E}\left[\Gamma_t^X\right]\right)^2\right] + \mathbb{E}\left[\left(\Gamma_t^Y - \mathbb{E}\left[\Gamma_t^Y\right]\right)^2\right] + \left(\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\Gamma_t^Y\right]\right)^2\right).$$

Consider the two Gaussian random vectors defined as

$$G_X = \int_0^1 \mathbb{E}\left[\Gamma_t^X\right] dB_t^X \text{ and } G_Y = \int_0^1 \mathbb{E}\left[\Gamma_t^Y\right] dB_t^Y,$$

and observe that

$$X = \int_0^1 \Gamma_t^X dB_t^X = \int_0^1 \left(\Gamma_t^X - \mathbb{E}\left[\Gamma_t^X\right]\right) dB_t^X + \int_0^1 \mathbb{E}\left[\Gamma_t^X\right] dB_t^X = \int_0^1 \left(\Gamma_t^X - \mathbb{E}\left[\Gamma_t^X\right]\right) dB_t^X + G_X.$$

This induces a coupling between X and G_X from which we obtain, using Itô's Isometry,

$$\mathcal{W}_2^2\left(X, G_X\right) \leq \mathbb{E}\left[\left(\int_0^1 \left(\Gamma_t^X - \mathbb{E}\left[\Gamma_t^X\right]\right) dB_t^X\right)^2\right] = \int_0^1 \mathrm{Tr}\left(\mathrm{Var}\left(\Gamma_t^X\right)\right) dt,$$

and an analogous estimate also holds for Y. We may now use $\mathbb{E}\left[\Gamma_t^X\right]$ and $\mathbb{E}\left[\Gamma_t^Y\right]$ as the diffusion coefficients for the same Brownian motion to establish

$$\mathcal{W}_2^2(G_X, G_Y) \leq \mathbb{E}\left[\left(\int_0^1 \left(\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\Gamma_t^Y\right]\right) dB_t\right)^2\right] = \int_0^1 \mathrm{Tr}\left(\left(\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\Gamma_t^Y\right]\right)^2\right) dt.$$

Plugging these estimates into (4.18) reproves the following bound, which is identical to Theorem 1 in [84].

Theorem 4.18. *Let X and Y be 1-uniformly log-concave centered vectors and let G_X, G_Y be defined as above. Then,*

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2}\left(\mathcal{W}_2^2\left(X||G_X\right) + \mathcal{W}_2^2\left(Y||G_Y\right) + \mathcal{W}_2^2\left(G_X, G_Y\right)\right).$$

To obtain a bound for the relative entropy towards the proof of Theorem 4.1, we will require a slightly more general version of inequality (4.8). This is the content of the next lemma,

whose proof is similar to the argument presented above. The main difference comes from applying Girsanov's theorem to a re-scaled Brownian motion, from which we obtain an expression analogous to (4.5). This is essentially Lemma 1.13, which we restate here for convenience

Lemma 4.19. *Let F_t and E_t be two F_t-adapted matrix-valued processes and let X_t, M_t be two processes defined by*

$$Z_t = \int_0^t F_s dB_s, \text{ and } M_t = \int_0^t E_s dB_s.$$

Suppose that for every $t \in [0,1]$, $E_t \succeq c\mathrm{I}_d$ for some deterministic $c > 0$, then

$$\mathrm{Ent}(Z_1||M_1) \leq \mathrm{Tr} \int_0^1 \frac{\mathbb{E}\left[(F_t - E_t)^2\right]}{c^2(1-t)} dt.$$

Proof of Theorem 4.1. By Corollary 4.14

$$\mathbb{E}\left[\Gamma_t^X\right] \succeq \sigma_X \mathrm{I}_d \text{ and } \mathbb{E}\left[\Gamma_t^Y\right] \succeq \sigma_Y \mathrm{I}_d \text{ for every } t \in [0,1].$$

We invoke Lemma 4.19 with $E_t = \mathbb{E}\left[\Gamma_t^X\right]$ and $F_t = \Gamma_t^X$ to obtain

$$\sigma_X^2 \mathrm{Ent}(X||G_X) \leq \int_0^1 \frac{\mathrm{Tr}\left(\mathrm{Var}\left(\Gamma_t^X\right)\right)}{1-t} dt.$$

Repeating the same argument for Y gives

$$\sigma_Y^2 \mathrm{Ent}(Y||G_Y) \leq \int_0^1 \frac{\mathrm{Tr}\left(\mathrm{Var}\left(\Gamma_t^Y\right)\right)}{1-t} dt.$$

By invoking Lemma 4.19 with $F_t = \mathbb{E}\left[\Gamma_t^X\right]$ and $E_t = \mathbb{E}\left[\Gamma_t^Y\right]$ and then one more time after switching between F_t and E_t, and summing the results, we get

$$\frac{\sigma_Y^2}{2}\mathrm{Ent}(G_X||G_Y) + \frac{\sigma_X^2}{2}\mathrm{Ent}(G_Y||G_X) \leq \int_0^1 \frac{\mathrm{Tr}\left(\left(\mathbb{E}\left[\Gamma_t^X\right] - \mathbb{E}\left[\Gamma_t^Y\right]\right)^2\right)}{1-t} dt.$$

Plugging the above inequalities into (4.18) concludes the proof. ∎

4.4 Stability for general log-concave random vectors

Fix X, Y, centered log-concave random vectors in $\mathbb{R}^d$, such that

$$\mathrm{Cov}(Y) + \mathrm{Cov}(X) = 2\mathrm{I}_d, \tag{4.19}$$

with σ_X^2, σ_Y^2 the corresponding minimal eigenvalues of $\mathrm{Cov}(X)$ and $\mathrm{Cov}(Y)$. Assume further that $\frac{\mathrm{C_P}(Y)}{\sigma_Y^2}, \frac{\mathrm{C_P}(X)}{\sigma_X^2} \leq \mathrm{C_P}$, for some $\mathrm{C_P} > 1$. Again, let B_t^X and B_t^Y be independent Brownian motions and consider the associated processes Γ_t^X, Γ_t^Y defined as in Section 4.2.

The general log-concave case, in comparison with the case where X and Y are uniformly log-concave, gives rise to two essential difficulties. Recall that the results in the previous section used the fact that an upper bound for the matrices Γ_t^X, Γ_t^Y, combined with equation (4.13) gives the simpler bound (4.18). Unfortunately, in the general log-concave case, there is no upper bound uniform in t, which creates the first problem. The second issue has to do with the lack of respective lower bounds for $\mathbb{E}[\Gamma_t^X]$ and $\mathbb{E}[\Gamma_t^Y]$: in view of Lemma 4.19, one needs such bounds in order to obtain estimates on the entropies.

The solution of the second issue lies in Corollary 4.16, which gives a lower bound for the processes in terms on the Poincaré constants. We denote $\xi = \frac{1}{(2\mathrm{C_P}+1)}\frac{\min(\sigma_Y^2,\sigma_X^2)}{3}$, so that the corollary gives

$$\mathbb{E}\left[\Gamma_t^Y\right], \mathbb{E}\left[\Gamma_t^X\right] \succeq \xi \mathrm{I}_d. \tag{4.20}$$

Thus, we are left with the issue arising from the lack of a uniform upper bound for the matrices Γ_t^X, Γ_t^Y. Note that Lemma 4.13 gives $\Gamma_t^X \preceq \frac{1}{t}\mathrm{I}_d$, a bound which is not uniform in t. To illustrate how one may overcome this issue, suppose that there exists an $\varepsilon > 0$, such that

$$\int_0^{\varepsilon} \frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]\right)}{(1-t)} dt < \frac{1}{2}\int_0^1 \frac{\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]\right)}{(1-t)} dt.$$

In such a case, Lemma 4.11 would imply

$$\delta_{EPI,\lambda}(X,Y) \gtrsim \frac{\lambda(1-\lambda)}{\varepsilon}\mathrm{Tr}\int_0^1 \frac{\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]}{1-t} dt.$$

Towards finding an ε such that the above holds, note that since v_t^X is a martingale, and using (4.5) we have for every $t_0 \in [0,1]$,

$$(1-t_0)\mathrm{Ent}\left(X||G\right) = \frac{1-t_0}{2}\int_0^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt \leq \frac{1}{2}\int_{t_0}^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt \leq \mathrm{Ent}\left(X||G\right). \tag{4.21}$$

Observe that

$$\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X-\Gamma_t^Y\right)^2\right]\right)=\mathrm{Tr}\left(\mathbb{E}\left[\left(\Gamma_t^X-\mathrm{I}_d\right)^2\right]+\mathbb{E}\left[\left(\Gamma_t^Y-\mathrm{I}_d\right)^2\right]-2\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^X\right]\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^Y\right]\right).$$

Using the relation in (4.10), Fubini's theorem shows

$$\begin{aligned}\int_{t_0}^1\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]dt&=\mathrm{Tr}\int_{t_0}^1\int_0^t\frac{\mathbb{E}\left[\left(\Gamma_s^X-\mathrm{I}_d\right)^2\right]}{(1-s)^2}dsdt\\&=\mathrm{Tr}\int_0^{t_0}\int_{t_0}^1\frac{\mathbb{E}\left[\left(\Gamma_s^X-\mathrm{I}_d\right)^2\right]}{(1-s)^2}dtds+\mathrm{Tr}\int_{t_0}^1\int_s^1\frac{\mathbb{E}\left[\left(\Gamma_s^X-\mathrm{I}_d\right)^2\right]}{(1-s)^2}dtds\\&=(1-t_0)\mathbb{E}\left[\left\|v_{t_0}^X\right\|_2^2\right]+\mathrm{Tr}\int_{t_0}^1\frac{\mathbb{E}\left[\left(\Gamma_s^X-\mathrm{I}_d\right)^2\right]}{1-s}ds.\end{aligned}$$

Combining the last two displays gives

$$\begin{aligned}\mathrm{Tr}\int_{t_0}^1\frac{\mathbb{E}\left[\left(\Gamma_t^X-\Gamma_t^Y\right)^2\right]}{1-t}dt=\int_{t_0}^1\left(\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]+\mathbb{E}\left[\left\|v_t^Y\right\|_2^2\right]\right)dt-(1-t_0)\left(\mathbb{E}\left[\left\|v_{t_0}^X\right\|_2^2\right]+\mathbb{E}\left[\left\|v_{t_0}^Y\right\|_2^2\right]\right)\\-2\mathrm{Tr}\int_{t_0}^1\frac{\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^X\right]\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^Y\right]}{1-t}dt.\end{aligned}\tag{4.22}$$

Using (4.16), we have the identities:

$$\frac{\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^X\right]}{1-t}=\mathbb{E}\left[v_t^X\otimes v_t^X\right]+\mathrm{I}_d-\mathrm{Cov}(X)$$

and

$$\frac{\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^Y\right]}{1-t}=\mathbb{E}\left[v_t^Y\otimes v_t^Y\right]+\mathrm{I}_d-\mathrm{Cov}(Y),$$

from which we deduce

$$\begin{aligned}2\frac{\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^X\right]\mathbb{E}\left[\mathrm{I}_d-\Gamma_t^Y\right]}{1-t}&=\left(\mathrm{I}_d-\mathbb{E}\left[\Gamma_t^Y\right]\right)\mathbb{E}\left[v_t^X\otimes v_t^X\right]+\left(\mathrm{I}_d-\mathbb{E}\left[\Gamma_t^X\right]\right)\mathbb{E}\left[v_t^Y\otimes v_t^Y\right]\\&\quad+\left(\mathrm{I}_d-\mathbb{E}\left[\Gamma_t^Y\right]\right)\left(\mathrm{I}_d-\mathrm{Cov}(X)\right)+\left(\mathrm{I}_d-\mathbb{E}\left[\Gamma_t^X\right]\right)\left(\mathrm{I}_d-\mathrm{Cov}(Y)\right).\end{aligned}$$

Let $\{w_i\}_{i=1}^d$ be an orthornormal basis of eigenvectors corresponding to the eigenvalues $\{\lambda_i\}_{i=1}^d$ of $\mathrm{I}_d-\mathbb{E}\left[\Gamma_t^X\right]$. The following observation, which follows from the above identities, is crucial: if $\lambda_i\leq 0$ then necessarily $\langle w_i,\mathrm{Cov}(X)w_i\rangle\geq 1$. In this case, by assumption (4.19),

$\langle w_i, \mathrm{Cov}(Y) w_i \rangle \leq 1$ and

$$\left\langle w_i, \frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right] \mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right]}{1-t} w_i \right\rangle \leq 0.$$

Our aim is to bound (4.22) from below; thus, in the calculation of the trace in the RHS, we may disregard all w_i corresponding to negative λ_i. Moreover, if $\lambda_i \geq 0$, we need only consider the cases where

$$\langle w_i, \left(\mathrm{I}_d - \mathbb{E}\left[\Gamma_t^Y\right]\right) w_i \rangle \geq 0,$$

as well. Since,

$$\begin{aligned} 2\left\langle w_i, \frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right] \mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right]}{1-t} w_i \right\rangle =& \langle w_i, \mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right] w_i \rangle \left(\mathbb{E}\left[\langle v_t^Y, w_i\rangle^2\right] + 1 - \langle w_i, \mathrm{Cov}(Y) w_i \rangle\right) \\ &+ \langle w_i, \mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right] w_i \rangle \left(\mathbb{E}\left[\langle v_t^X, w_i\rangle^2\right] + 1 - \langle w_i, \mathrm{Cov}(X) w_i \rangle\right), \end{aligned}$$

under the assumptions taken on w_i, we see that all the terms are positive. Using the estimate (4.20), the previous equation is bounded from above by

$$\begin{aligned} &(1-\xi)\big(\mathbb{E}\left[\langle v_t^Y, w_i\rangle^2\right] + 1 - \langle w_i, \mathrm{Cov}(Y) w_i \rangle + \mathbb{E}\left[\langle v_t^X, w_i\rangle^2\right] + 1 - \langle w_i, \mathrm{Cov}(X) w_i \rangle\big) \\ &\quad = (1-\xi)\big(\mathbb{E}\left[\langle v_t^Y, w_i\rangle^2\right] + \mathbb{E}\left[\langle v_t^X, w_i\rangle^2\right]\big), \end{aligned}$$

where we have used (4.19). Summing over all the relevant w_i we get

$$2\mathrm{Tr}\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right] \mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right]}{1-t} \leq (1-\xi)\left(\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] + \mathbb{E}\left[\left\|v_t^Y\right\|_2^2\right]\right).$$

Plugging this into (4.22) and using (4.21) we have thus shown

$$\begin{aligned} \mathrm{Tr}\int_{t_0}^1 \frac{\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]}{1-t} dt \geq& 2\xi(1-t_0)\left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right) \\ &- (1-t_0)\left(\mathbb{E}\left[\left\|v_{t_0}^X\right\|_2^2\right] + \mathbb{E}\left[\left\|v_{t_0}^Y\right\|_2^2\right]\right). \end{aligned} \tag{4.23}$$

This suggests that it may be useful to bound $\mathbb{E}\left[\left\|v_{t_0}^X\right\|_2^2\right]$ from above, for small values of t_0, which is the objective of the next lemma.

Lemma 4.20. *If X is centered and has a finite Poincaré constant $\mathrm{C_p}(X) < \infty$, then for every $s \leq \frac{1}{3(2\mathrm{C_p}(X)+1)}$ the following holds*

$$\mathbb{E}\left[\left\|v_{s^2}^X\right\|_2^2\right] < \frac{s}{4} \cdot \mathrm{Ent}(X||G).$$

Proof. Suppose to the contrary that $\mathbb{E}\left[\left\|v_{s^2}^X\right\|_2^2\right] \geq \frac{s}{4} \cdot \mathrm{Ent}(X||G)$. Invoking Lemma 4.17 with $t_0 = s^2$ gives

$$\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] \geq \mathrm{Ent}(X||G) \cdot \frac{t\left((\mathrm{C_p}(X)-1)s^2+1\right)}{4\left((\mathrm{C_p}(X)-1)st+s\right)},$$

whenever $t \geq s^2$. Thus,

$$\begin{aligned}\int_{s^2}^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt &\geq \mathrm{Ent}(X||G) \int_{s^2}^1 \frac{t\left((\mathrm{C_p}(X)-1)s^2+1\right)}{4\left((\mathrm{C_p}(X)-1)st+s\right)} dt \\ &= \mathrm{Ent}(X||G)\left((\mathrm{C_p}(X)-1)s^2+1\right) \frac{(\mathrm{C_p}(X)-1)t - \ln\left(t\left(\mathrm{C_p}(X)-1\right)+1\right)}{4(\mathrm{C_p}(X)-1)^2 s}\Bigg|_{s^2}^1. \end{aligned} \tag{4.24}$$

Note now that for $s \leq \frac{1}{3(2\mathrm{C_p}(X)+1)}$

$$\frac{d}{ds}\frac{t\left((\mathrm{C_p}(X)-1)s^2+1\right)}{4\left((\mathrm{C_p}(X)-1)st+s\right)} = \frac{\left(\mathrm{C_p}(X)-1\right)s^2 t - 1}{s^2\left((\mathrm{C_p}(X)-1)t+1\right)} < 0,$$

and in particular we may substitute $s = \frac{1}{3(2\mathrm{C_p}(X)+1)}$ in (4.24). In this case, a straightforward calculation yields

$$\int_{\xi_X^2}^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt > \mathrm{Ent}(X||G),$$

which contradicts the identity (4.5), and concludes the proof by contradiction. □

We would like to use the lemma with the choice $s = \xi^2$. In order to verify the condition on the lemma which amounts to $\xi^2 \leq \frac{1}{3(2\mathrm{C_p}(X)+1)}$, we first remark that if $\sigma_X^2 \leq 1$, then it is clear that $\xi \leq \frac{1}{3(2\mathrm{C_p}(X)+1)}$. Otherwise, $\sigma_X^2 \geq 1$ and

$$\xi \leq \frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma_X^2}+1}\frac{\sigma_Y^2}{3} \leq \frac{1}{2\frac{\mathrm{C_p}(X)}{\sigma_X^2}+1}\frac{2-\sigma_X^2}{3} \leq \frac{1}{3(2\mathrm{C_p}(X)+1)}.$$

As the same reasoning is also true for Y, we now choose $t_0 = \xi^2$, which allows to invoke the previous lemma in (4.23) and to establish:

$$\mathrm{Tr}\int_{\xi^2}^1 \frac{\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]}{1-t} dt \geq \xi\left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right). \tag{4.25}$$

We are finally ready to prove the main theorem.

Proof of Theorem 4.3. Denote $\xi = \frac{1}{(2\mathrm{C_p}+1)}\frac{\min(\sigma_Y^2,\sigma_X^2)}{3}$. Since X and Y are log-concave, by

Lemma 4.13, $\Gamma_t^X, \Gamma_t^Y \preceq \frac{1}{t}\mathrm{I}_d$ almost surely. Thus, Lemma 4.11 gives

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\xi^2\lambda(1-\lambda)}{2}\int_{\xi^2}^{1}\frac{\mathrm{Tr}\left(\mathbb{E}\left[(\Gamma_t^X-\Gamma_t^Y)^2\right]\right)}{1-t}dt.$$

By noting that $\mathrm{C_p} \geq 1$, the bound (4.25) gives

$$\begin{aligned}\delta_{EPI,\lambda}(X,Y) &\geq \frac{\xi^3\lambda(1-\lambda)}{2}\left(\mathrm{Ent}\left(X||G\right)+\mathrm{Ent}\left(Y||G\right)\right)\\ &\geq K\lambda(1-\lambda)\left(\frac{\min(\sigma_Y^2,\sigma_X^2)}{\mathrm{C_p}}\right)^3\left(\mathrm{Ent}\left(X||G\right)+\mathrm{Ent}\left(Y||G\right)\right),\end{aligned}$$

for some numerical constant $K > 0$. □

4.5 Further results

4.5.1 Stability for low entropy log concave measures

In this section we focus on the case where X and Y are log-concave and isotropic. Similar to the previous section, we set $\xi_X = \frac{1}{3(2\mathrm{C_p}(X)+1)}$, so that by Corollary 4.16,

$$\mathbb{E}\left[\Gamma_t^X\right] \succeq \xi_X \mathrm{I}_d.$$

Towards the proof of Theorem 4.6, we first need an analogue of Lemma 4.20, for which we sketch the proof here.

Lemma 4.21. *If X is centred and has a finite Poincaré constant $\mathrm{C_p}(X) < \infty$,*

$$\mathbb{E}\left[\|v_{\xi_X}\|_2^2\right] < \frac{1}{4}\mathrm{Ent}(X||G).$$

Proof. Assume by contradiction that $\mathbb{E}\left[\|v_{\xi_X}\|_2^2\right] \geq \frac{1}{4}\mathrm{Ent}(X||G)$. In this case, Lemma 4.17 implies, for every $t \geq \xi_X$,

$$\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] \geq \mathrm{Ent}(X||G)\cdot\frac{t\left((\mathrm{C_p}(X)-1)\xi_X+1\right)}{4\left((\mathrm{C_p}(X)-1)\xi_X t+\xi_X\right)}.$$

A calculation then shows that

$$\int_{\xi_X}^{1}\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]dt \geq \mathrm{Ent}(X||G),$$

which is a contradiction to (4.5). □

Proof of Theorem 4.6. Since v_t^X is a martingale, $\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]$ is an increasing function. By (4.5) we deduce the elementary inequality

$$\mathbb{E}\left[\left\|v_s^X\right\|_2^2\right] \leq \frac{1}{1-s}\int_0^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt = \frac{2\mathrm{Ent}(X||G)}{1-s},$$

which holds for every $s \in [0,1]$. For isotropic X, Equation (4.16) shows that, for all $t \in [0,1]$,

$$(1-t)\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] = \mathrm{Tr}\left(\mathrm{I}_d - \mathbb{E}\left[\Gamma_t^X\right]\right) \leq 2\mathrm{Ent}(X||G) \leq \frac{1}{2},$$

where the second inequality is by assumption. Note that Equation (4.16) also shows that $\mathbb{E}\left[\Gamma_t^X\right] \preceq \mathrm{I}_d$ which yields, for every $t \in [0,1]$

$$0 \preceq \mathrm{I}_d - \mathbb{E}\left[\Gamma_t^X\right] \preceq \frac{1}{2}\mathrm{I}_d.$$

Applying this to Y as well produces the bound

$$\begin{aligned} 2\mathrm{Tr}\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right]\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right]}{1-t} &\leq \frac{1}{2}\mathrm{Tr}\left(\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^Y\right]}{1-t}\right) + \frac{1}{2}\mathrm{Tr}\left(\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t^X\right]}{1-t}\right) \\ &= \frac{1}{2}\left(\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] + \mathbb{E}\left[\left\|v_t^Y\right\|_2^2\right]\right). \end{aligned}$$

Set $\xi = \min(\xi_X, \xi_Y)$. Repeating the same calculation as in (4.22) and using the above gives that

$$\begin{aligned} \mathrm{Tr}\int_\xi^1 \frac{\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]}{1-t} dt &\geq (1-\xi)\left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right) \\ &\quad - (1-\xi)\left(\mathbb{E}\left[\left\|v_\xi^X\right\|_2^2\right] + \mathbb{E}\left[\left\|v_\xi^Y\right\|_2^2\right]\right). \end{aligned}$$

Lemma 4.21 implies

$$\mathrm{Tr}\int_\xi^1 \frac{\mathbb{E}\left[\left(\Gamma_t^X - \Gamma_t^Y\right)^2\right]}{1-t} dt \geq \frac{3}{4}(1-\xi)\left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right) \geq \frac{1}{2}\left(\mathrm{Ent}(X||G) + \mathrm{Ent}(Y||G)\right).$$

Finally, by Lemma 4.13, $\Gamma_t^X, \Gamma_t^Y \preceq \frac{1}{t}\mathrm{I}_d$ almost surely for all $t \in [0,1]$. We now invoke Lemma

4.11 to obtain

$$\delta_{EPI,\lambda}(X,Y) \geq \frac{\lambda(1-\lambda)}{2\xi}\mathrm{Tr}\int_{\xi}^{1}\frac{\mathbb{E}\left[\left(\Gamma_t^X-\Gamma_t^Y\right)^2\right]}{1-t}dt$$
$$\geq \frac{\lambda(1-\lambda)}{4\xi}\left(\mathrm{Ent}(X||G)+\mathrm{Ent}(Y||G)\right).$$

□

4.5.2 Stability under convolution with a Gaussian

Proof of Theorem 4.7. Fix $\lambda \in (0,1)$, by (4.6) we have that

$$\sqrt{\lambda}\left(\sqrt{\lambda}X_1+\sqrt{1-\lambda}G\right) \stackrel{d}{=} B_\lambda + \int_0^\lambda v_t^X dt.$$

As the relative entropy is affine invariant, this implies

$$\mathrm{Ent}\left(\sqrt{\lambda}\left(\sqrt{\lambda}X_1+\sqrt{1-\lambda}G\right)\middle|\middle|\sqrt{\lambda}G\right) = \mathrm{Ent}\left(\sqrt{\lambda}X_1+\sqrt{1-\lambda}G\middle|\middle|G\right) = \frac{1}{2}\int_0^\lambda \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right]dt. \tag{4.26}$$

Lemma 4.17 yields,

$$\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] \geq \mathbb{E}\left[\left\|v_\lambda^X\right\|_2^2\right]\frac{\lambda\left(\mathrm{C_p}(X)-1\right)t+t}{\lambda\left(\mathrm{C_p}(X)-1\right)t+\lambda} \text{ for } t \geq \lambda,$$

and

$$\mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] \leq \mathbb{E}\left[\left\|v_\lambda^X\right\|_2^2\right]\frac{\lambda\left(\mathrm{C_p}(X)-1\right)t+t}{\lambda\left(\mathrm{C_p}(X)-1\right)t+\lambda} \text{ for } t \leq \lambda.$$

Denote

$$I_1 := \int_\lambda^1 \frac{\lambda\left(\mathrm{C_p}(X)-1\right)t+t}{\lambda\left(\mathrm{C_p}(X)-1\right)t+\lambda}dt \text{ and } I_2 := \int_0^\lambda \frac{\lambda\left(\mathrm{C_p}(X)-1\right)t+t}{\lambda\left(\mathrm{C_p}(X)-1\right)t+\lambda}dt.$$

A calculation shows

$$I_1 = \frac{\left(\lambda\left(\mathrm{C_p}(X)-1\right)+1\right)\left((1-\lambda)\left(\mathrm{C_p}(X)-1\right)-\ln\left(\mathrm{C_p}(X)\right)+\ln\left(\lambda\left(\mathrm{C_p}(X)-1\right)+1\right)\right)}{\lambda\left(\mathrm{C_p}(X)-1\right)^2},$$

as well as

$$I_2 = \frac{\left(\lambda\left(\mathrm{C_p}(X)-1\right)+1\right)\left(\lambda(\mathrm{C_p}(X)-1)-\ln\left(\lambda\left(\mathrm{C_p}(X)-1\right)+1\right)\right)}{\lambda\left(\mathrm{C_p}(X)-1\right)^2}.$$

Thus, the above bounds give

$$\mathrm{Ent}(X||G) = \frac{1}{2}\int_0^1 \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt \geq \frac{1}{2}\int_0^\lambda \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt + \frac{\mathbb{E}\left[\left\|v_\lambda^X\right\|_2^2\right]}{2} I_1,$$

and

$$0 \leq \frac{1}{2}\int_0^\lambda \mathbb{E}\left[\left\|v_t^X\right\|_2^2\right] dt \leq \frac{1}{2} I_2.$$

Now, since the expression $\frac{\alpha}{\alpha+\beta}$ is monotone increasing with respect to α and decreasing with respect to β whenever $\alpha, \beta > 0$, those two inequalities together with (4.26) imply that

$$\begin{aligned}\mathrm{Ent}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}G\middle|\middle|G\right) &\leq \frac{I_2}{I_1 + I_2}\mathrm{Ent}(X||G) \\ &= \frac{\lambda\left(\mathrm{C_p}(X) - 1\right) - \ln\left(\lambda\left(\mathrm{C_p}(X) - 1\right) + 1\right)}{\mathrm{C_p}(X) - \ln\left(\mathrm{C_p}(X)\right) - 1}\mathrm{Ent}(X||G).\end{aligned}$$

Rewriting the above in terms of the deficit in the Shannon-Stam inequality, we have established

$$\begin{aligned}\delta_{EPI,\lambda}(X, G) &= \lambda\mathrm{Ent}(X||G) - \mathrm{Ent}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}G\middle|\middle|G\right) \\ &\geq \left(\lambda - \frac{\lambda\left(\mathrm{C_p}(X) - 1\right) - \ln\left(\lambda\left(\mathrm{C_p}(X) - 1\right) + 1\right)}{\mathrm{C_p}(X) - \ln\left(\mathrm{C_p}(X)\right) - 1}\right)\mathrm{Ent}(X||G).\end{aligned}$$

□

5

Stability of Talagrand's Gaussian Transport-Entropy Inequality

5.1 Introduction

Talagrand's Gaussian transport-entropy inequality, first proved in [229], states that for any measure μ in $\mathbb{R}^d$, with a finite second moment matrix,

$$\mathcal{W}_2^2(\mu,\gamma) \leq 2\mathrm{Ent}(\mu||\gamma). \tag{5.1}$$

Recall that γ is the standard Gaussian measure on $\mathbb{R}^d$, $\mathrm{Ent}(\mu||\gamma)$ stands for relative entropy, and $\mathcal{W}_p(\mu,\gamma)$ is the L_p-Wasserstein distance (with L_2 cost function).

Since this fundamental inequality tensorizes, it holds in any dimension. Using this quality, the inequality was shown to imply a sharp form of the dimension-free concentration of measure phenomenon in Gaussian space. The reader is referred to [129, 159, 240] for further information on the topic. By setting the measure μ to be a translation of γ, we can see that the inequality is tight and that, in particular, the constant 2 in (5.1) cannot be improved. One, in fact, may show that these examples account for the only equality cases of (5.1). We are thus led to consider the question of stability of the inequality. Consider the deficit

$$\delta_{\mathrm{Tal}}(\mu) := 2\mathrm{Ent}(\mu||\gamma) - \mathcal{W}_2^2(\mu,\gamma).$$

Suppose that $\delta_{\mathrm{Tal}}(\mu)$ is small. In this case, must μ be necessarily close to a translate of γ?
A first step towards answering this question, which serves as a starting point for the current work, was given in [113] (see also [155]), where it was shown that there exists a numerical constant $c > 0$, such that if μ is centered,

$$\delta_{\mathrm{Tal}}(\mu) \geq c \min\left(\frac{\mathcal{W}_{1,1}^2(\mu,\gamma)}{d}, \frac{\mathcal{W}_{1,1}(\mu,\gamma)}{\sqrt{d}}\right). \tag{5.2}$$

Here, $\mathcal{W}_{1,1}$ stands for the L_1-Wasserstein distance with L_1-cost function. The inequality was later improved in [81], and $\frac{\mathcal{W}_{1,1}(\mu,\gamma)}{\sqrt{d}}$ was replaced by the larger quantity $\mathcal{W}_1(\mu,\gamma)$. One could hope to improve this result in several ways; First, one may consider stronger notions of distance than $\mathcal{W}_{1,1}$, like relative entropy. Indeed by Jensen's inequality and (5.1),

$$\frac{\mathcal{W}_{1,1}^2(\mu,\gamma)}{d} \leq \mathcal{W}_2^2(\mu,\gamma) \leq 2\mathrm{Ent}(\mu||\gamma). \tag{5.3}$$

Second, note that for product measures, $\delta_{\mathrm{Tal}}(\mu)$ grows linearly in d, while the RHS of (5.2) may grow like $\sqrt{d}$ (this remains true for the improved result, found in [81]). The dimension-free nature of (5.1) suggests that the dependence on the dimension in (5.2) should, hopefully, be removed. The goal of the present work is to identify cases in which (5.2) may be improved. Specifically, we will be interested in giving dimension-free stability bounds with respect to the relative entropy distance. We will also show that, without further assumptions on the measure μ, (5.2) cannot be significantly improved.
This work adds to a recent line of works which explored dimension-free stability estimates for functional inequalities in the Gaussian space, such as the log-Sobolev inequality [39, 101, 113, 116, 162], the Shannon-Stam inequality [83, 103] and the Gaussian isoperimetric inequality [23, 77, 185].

Results

In our first main result, we restrict our attention to the subclass of probability measures which satisfy a Poincaré inequality. As in Chapter 4, a measure μ is said to satisfy a Poincaré inequality with constant $\mathrm{C_p}(\mu)$, if for every smooth function $g : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d} g^2 d\mu - \left(\int_{\mathbb{R}^d} g d\mu\right)^2 \leq \mathrm{C_p}(\mu) \int_{\mathbb{R}^d} \|\nabla g\|_2^2 \, d\mu,$$

where we implicitly assume that $\mathrm{C_p}(\mu)$ is the smallest constant for which this inequality holds. If μ satisfies such an inequality, then, in some sense, μ must be regular. Indeed, μ must have finite moments of all orders. For such measures we prove:

Theorem 5.1. *Let μ be a centered measure on $\mathbb{R}^d$ with finite Poincaré constant $\mathrm{C_p}(\mu) < \infty$.*

Then

$$\delta_{\mathrm{Tal}}(\mu) \geq \min\left(\frac{1}{4}, \frac{(\mathrm{C_p}(\mu)+1)\left(2-2\mathrm{C_p}(\mu)+(\mathrm{C_p}(\mu)+1)\ln\left(\mathrm{C_p}(\mu)\right)\right)}{(\mathrm{C_p}(\mu)-1)^3}\right)\mathrm{Ent}(\mu||\gamma).$$

Note that as the deficit is invariant to translations, there is no loss in generality in assuming that μ is centered. Furthermore, the Poincaré constant tensorizes, in the sense that for any two measures ν and μ, $\mathrm{C_p}(\nu\otimes\mu) = \max\left(\mathrm{C_p}(\nu), \mathrm{C_p}(\mu)\right)$. So, if μ is a product measure $\mathrm{C_p}(\mu)$ does not depend on the dimension and we regard it as a dimensionless quantity. For a more applicable form of the result we may use the inequality

$$\min\left(\frac{1}{4}, \frac{(x+1)(2-2x+(x+1)\ln(x))}{(x-1)^3}\right) \geq \frac{\ln(x+1)}{4x},$$

valid for $x>0$, to get

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{\ln(\mathrm{C_p}(\mu)+1)}{4\mathrm{C_p}(\mu)}\mathrm{Ent}(\mu||\gamma).$$

Theorem 5.1 should be compared with Theorem 1 in [113] and Theorem 4.7 which give similar stability estimates, involving the Poincaré constant, for the log-Sobolev and Shannon-Stam inequalities.

Regarding the conditions of the theorem; as will be shown in Section 5.2 below, there exists a measure μ for which $\delta_{\mathrm{Tal}}(\mu)$ may be arbitrarily close to 0, while $\mathcal{W}_2(\mu,\gamma)$ remains bounded away from 0. Thus, in order to establish meaningful stability results, in relative entropy, it is necessary to make some assumptions on the measure μ.

In case the measure μ does not satisfy a Poincaré inequality, we provide estimates in terms of its covariance matrix. It turns out, that if $\mathrm{Cov}(\mu)$ is strictly smaller than the identity, at least in some directions, we may still produce a dimension-free bound for $\delta_{\mathrm{Tal}}(\mu)$.

Theorem 5.2. *Let μ be a centered measure on $\mathbb{R}^d$ and let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $\mathrm{Cov}(\mu)$, counted with multiplicity. Then*

$$\delta_{\mathrm{Tal}}(\mu) \geq \sum_{i=1}^{d} \frac{2(1-\lambda_i)+(\lambda_i+1)\log(\lambda_i)}{\lambda_i-1}\mathbb{1}_{\{\lambda_i<1\}}.$$

Remark that for $0<x<1$, the function $g(x) := \frac{2(1-x)+(x+1)\log(x)}{x-1}$ is positive and that it is a decreasing function of x. Also, it can be verified that g' is actually concave on this domain, from which we may see $g(x)\geq\frac{1}{6}(x-1)^2$. Thus, if $\mathrm{Cov}(\mu)\preceq \mathrm{I}_d$, then the Theorem implies the weaker result

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{1}{6}\left\|\mathrm{I}_d-\mathrm{Cov}(\mu)\right\|_{HS}^2,$$

where $\|\cdot\|_{HS}$, stands for the Hilbert-Schmidt norm. In line with the above discussion, we may

regard $\|\mathrm{Cov}(\mu) - \mathrm{I}_d\|_{HS}^2$ as a certain distance between μ and the standard Gaussian. Theorem 3 in [101] gives a similar estimate for the log-Sobolev inequality. Indeed, our methods are based on related ideas.

If $\mathrm{Cov}(\mu) = \mathrm{I}_d$, Theorem 5.2 does not give any new insight beyond (5.1). The next result applies, among others, to this case.

Theorem 5.3. *Let μ be a centered measure on $\mathbb{R}^d$, such that $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) \leq d$. Then*

$$\delta_{\mathrm{Tal}}(\mu) \geq \min\left(\frac{\mathrm{Ent}(\mu||\gamma)^2}{6d}, \frac{\mathrm{Ent}(\mu||\gamma)}{4}\right).$$

As opposed to the previous two results, Theorem 5.3 is not dimension-free and is directly comparable to (5.2). Under the assumption $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) \leq d$, by using (5.3) we may view the theorem as a strengthening of (5.2). We should also comment that by Pinsker's inequality ([86]), relative entropy induces a stronger topology than the $\mathcal{W}_1$ metric. On the other hand, (5.2) holds in greater generality than Theorem 5.3 as it makes no assumptions on the measure μ. It is then natural to ask whether one can relax the conditions of the theorem. We give a negative answer to this question.

Theorem 5.4. *Fix $d \in \mathbb{N}$ and let $\xi > d$. There exist a sequence of centered measures μ_k on $\mathbb{R}^d$ such that:*

- $\lim_{k\to\infty} \mathrm{Tr}\left(\mathrm{Cov}(\mu_k)\right) = \xi.$
- $\lim_{k\to\infty} \delta_{\mathrm{Tal}}\left(\mu_k\right) = 0.$
- $\liminf_{k\to\infty} \mathcal{W}_2^2(\mu_k, \gamma) \geq \xi - d > 0.$

Thus, even for one dimensional measures, in order to obtain general stability estimates in relative entropy or even in the quadratic Wasserstein distance, the assumption $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) \leq d$ is necessary.

The counterexample to stability, guaranteed by Theorem 5.4, may be realized as a Gaussian mixture. In fact, as demonstrated by recent works ([65, 83, 101]), Gaussian mixtures may serve as counterexamples to stability of several other Gaussian functional inequalities. This led the authors of [101] to note that if a measure μ saturates the log-Sobolev inequality, then it must be close, in L_2-Wasserstein distance, to some Gaussian mixture. We show that this is also true, in relative entropy, for Talagrand's inequality.

Theorem 5.5. *Let μ be a centered measure on $\mathbb{R}^d$. Then there exists another measure ν with $\mathrm{Cov}(\nu) \preceq \mathrm{Cov}(\mu)$, such that if $\delta_{\mathrm{Tal}}(\mu) \geq d$,*

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{\mathrm{Ent}\left(\mu||\nu * \gamma\right)}{6},$$

and if $\delta_{\mathrm{Tal}}(\mu) < d$,

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{1}{3\sqrt{3}} \frac{\mathrm{Ent}(\mu||\nu * \gamma)^{\frac{3}{2}}}{\sqrt{d}}.$$

Note that, in light of Theorem 5.4, the above theorem is not true without the convolution, and we cannot, in general, replace $\nu * \gamma$ by γ.

For our last result, define the Fisher information of μ, relative to γ, as

$$\mathrm{I}(\mu||\gamma) := \int_{\mathbb{R}^d} \left\| \nabla \ln \left(\frac{d\mu}{d\gamma} \right) \right\|_2^2 d\mu.$$

Gross' log-Sobolev inequality ([130]) states that

$$\mathrm{I}(\mu||\gamma) \geq 2\mathrm{Ent}(\mu||\gamma).$$

For this we define the deficit as

$$\delta_{\mathrm{LS}}(\mu) = \mathrm{I}(\mu||\gamma) - 2\mathrm{Ent}(\mu||\gamma).$$

As will be described in Section 5.3 below, our approach draws a new connection between Talagrand's and the log-Sobolev inequalities. One benefit of this approach is that all of our results apply verbatim to the log-Sobolev inequality. Some of the results improve upon existing estimates in the literature. We summarize those in the following corollary.

Corollary 5.6. *Let* μ *be a centered measure on* $\mathbb{R}^d$. *Then there exists a measure* ν *such that* $\mathrm{Cov}(\nu) \preceq \mathrm{Cov}(\mu)$ *and*

$$\delta_{\mathrm{LS}}(\mu) \geq \min \left(\frac{1}{3\sqrt{3}} \frac{\mathrm{Ent}\left(\mu||\nu * \gamma\right)^{\frac{3}{2}}}{\sqrt{d}}, \frac{\mathrm{Ent}\left(\mu||\nu * \gamma\right)}{6} \right).$$

Moreover, if $\mathrm{Tr}\left(\mathrm{Cov}\left(\mu\right)\right) \leq d$ *then*

$$\delta_{\mathrm{LS}}(\mu) \geq \min \left(\frac{\mathrm{Ent}(\mu||\gamma)^2}{6d}, \frac{\mathrm{Ent}(\mu||\gamma)}{4} \right),$$

The second point of the corollary is an improvement of Corollary 1.2 in [39] which shows, under the same hypothesis,

$$\delta_{\mathrm{LS}}(\mu) \geq c \frac{\mathcal{W}_2^4(\mu, \gamma)}{d},$$

for some universal constant $c > 0$. The improved bound can actually be deduced from Theorem 1.1 in the same paper, but it does not seem to appear in the literature explicitly

The first point of Corollary 5.6 strengthens Theorem 7 in [101] which states, that for some measure ν:

$$\delta_{\mathrm{LS}}(\mu) \geq \frac{1}{15} \frac{\mathcal{W}_2^3(\mu, \nu * \gamma)}{\sqrt{d}}. \tag{5.4}$$

Our proof closely resembles theirs, but our analysis yields bounds in the stronger relative entropy distance. The authors of [101] raise the natural question, whether the dependence on the dimension in (5.4) can be completely discarded. The same question is also relevant to $\delta_{\mathrm{Tal}}(\mu)$. We do not know the answer to either of the questions, which seem related.

Organization

The remainder of the chapter is organized as follows: In Section 5.2 we give a counter-example to stability of Talagrand's inequality, proving Theorem 5.4. Section 5.3 is devoted to explaining our method and proving some of its basic properties which will then be used in Section 5.4 to prove the stability estimates. Finally, in Section 5.5 we give an application of our results to Gaussian concentration inequalities.

5.2 A counterexample to stability

In this section we show that one cannot expect any general stability result to hold if $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) > d$. We present a one-dimensional example, which may be easily generalized to higher dimensions. The following notations will be used in this section:

- For $\sigma^2 > 0$, γ_{σ^2} denotes the law of the centered 1-dimensional Gaussian with variance σ^2.
- Fix $\xi > 1$ and $k \in \mathbb{N}$, we set

$$\mu_k := \left(1 - \frac{1}{k}\right)\gamma_1 + \frac{1}{k}\gamma_{k(\xi-1)}.$$

Recall now the Kantorovich dual formulation (see [124,240], for example) of the L_2-Wasserstien distance. For ν and μ measures on $\mathbb{R}$, we have

$$\mathcal{W}_2^2(\mu,\nu) = \sup_g \left\{ \int_{\mathbb{R}} g(x)d\mu(x) - \int_{\mathbb{R}} (Qg)(x)d\nu(x) \right\}, \tag{5.5}$$

where the supremum runs over all measurable functions, and Qg denotes the sup-convolution of g, namely

$$Qg(x) = \sup_{y\in\mathbb{R}}\{g(y) - (x-y)^2\}.$$

Proof of Theorem 5.4. We first note that $\mathrm{Var}(\mu_k) \xrightarrow{k\to\infty} \xi > 1$. Towards understanding $\delta_{\mathrm{Tal}}(\mu_k)$ we use the fact that relative entropy is convex with respect to mixtures of measures ([86]), so

$$\mathrm{Ent}(\mu_k||\gamma) \leq \frac{1}{k}\mathrm{Ent}\left(\gamma_{k(\xi-1)}||\gamma\right) = \frac{1}{2k}\left(k(\xi-1) - 1 - \ln\left(k(\xi-1)\right)\right) \leq \frac{\xi-1}{2}. \tag{5.6}$$

To control the Wasserstein distance, define the functions

$$g_k(x) = \begin{cases} 0 & \text{if } |x| < \frac{\sqrt{k}}{\ln(k)} \\ \left(1 - \frac{1}{\ln(k)}\right) x^2 & \text{otherwise} \end{cases}.$$

The main idea is that as k increases, Qg_k vanishes in an ever expanding region, while growing slowly outside of the region. Formally, for $0 \leq x \leq \frac{\sqrt{k}}{\ln(k)} - \frac{\sqrt{k(\ln(k)-1)}}{\ln(k)^{\frac{3}{2}}}$, it holds that

$$g_k\left(\frac{\sqrt{k}}{\ln(k)}\right) - \left(x - \frac{\sqrt{k}}{\ln(k)}\right)^2 = \left(1 - \frac{1}{\ln(k)}\right)\left(\frac{\sqrt{k}}{\ln(k)}\right)^2 - \left(x - \frac{\sqrt{k}}{\ln(k)}\right)^2 \leq 0.$$

and in particular, if $\frac{\sqrt{k}}{\ln(k)} < y$,

$$g_k(y) - (x-y)^2 < 0,$$

which shows $Qg_k(x) = 0$. There exists a constant $c > 0$ such that

$$\frac{\sqrt{k}}{\ln(k)} - \frac{\sqrt{k\left(\ln(k) - 1\right)}}{\ln(k)^{\frac{3}{2}}} \geq ck^{\frac{1}{4}},$$

which, combined with the previous observation shows that for $|x| \leq ck^{\frac{1}{4}}$, $Qg_k(x) = 0$. If $|x| > ck^{\frac{1}{4}}$ it is standard to show $Qg_k(x) \leq \ln(k)x^2$. So,

$$\int_{\mathbb{R}} Qg_k(x)d\gamma_1(x) \leq \ln(k) \int_{|x| \geq ck^{\frac{1}{4}}} x^2 d\gamma_1(x) = \ln(k)\left(\frac{c\sqrt{2}}{\sqrt{\pi}} k^{\frac{1}{4}} e^{-\frac{c^2\sqrt{k}}{2}} + \int_{|x| \geq ck^{\frac{1}{4}}} d\gamma_1\right) \xrightarrow{k\to\infty} 0,$$

where the equality is integration by parts. Also, it is clear that

$$\int_{\mathbb{R}} g_k(x)d\gamma_1(x) \xrightarrow{k\to\infty} 0.$$

Now, if φ denotes the density of the standard Gaussian, then by a change of variables we have

$$\begin{aligned} \frac{1}{k}\int_{\mathbb{R}} g_k(x)d\gamma_{k(\xi-1)}(x) &= \left(1 - \frac{1}{\ln(k)}\right)\frac{1}{k}\int_{|x| \geq \frac{\sqrt{k}}{\ln(k)}} \frac{x^2}{\sqrt{k\left(\xi-1\right)}}\varphi\left(\frac{x}{\sqrt{k\left(\xi-1\right)}}\right)dx \\ &= \left(1 - \frac{1}{\ln(k)}\right)(\xi - 1)\int_{|y| \geq \frac{1}{\ln(k)\sqrt{\xi-1}}} y^2\varphi(y)dy \xrightarrow{k\to\infty} \xi - 1. \end{aligned}$$

Combining the above displays with (5.5) we get,

$$\begin{aligned}\mathcal{W}_2^2(\mu_k,\gamma_1) &\geq \int_{\mathbb{R}} g_k(x)d\mu_k(x) - \int_{\mathbb{R}} Qg_k(x)d\gamma_1(x) \\ &= \left(1-\frac{1}{k}\right)\int_{\mathbb{R}} g_k(x)d\gamma_1(x) + \frac{1}{k}\int_{\mathbb{R}} g_k(x)d\gamma_{k(\xi-1)}(x) - \int_{\mathbb{R}} Qg_k(x)d\gamma_1(x) \xrightarrow{k\to\infty} \xi - 1.\end{aligned}$$

Finally, from (5.6) we obtain

$$\delta_{\text{Tal}}(\mu_k) = 2\text{Ent}(\mu_k||\gamma_1) - \mathcal{W}_2^2(\mu_k,\gamma_1) \xrightarrow{k\to\infty} 0.$$

□

We remark that, in light of Theorem 5.5, it would seem more natural to have as a counterexample a mixture of Gaussians with unit variance, as was done in [101] for the log-Sobolev inequality. However, (5.2) tells us that the situation in Talagrand's inequality is a bit more delicate, since the inequality is stable with respect to the $\mathcal{W}_1$ metric. Thus, as in the given example, a counterexample to stability (in the $\mathcal{W}_2$ metric or relative entropy) should satisfy $\lim_{k\to\infty} \mathcal{W}_1(\mu_k,\gamma) = 0$, while $\liminf_{k\to\infty} \mathcal{W}_2(\mu_k,\gamma) > 0$. keeping this in mind, it seems more straightforward to allow the second moments of the summands in the mixture to vary while keeping their means fixed at the origin. This is also very similar to the counterexample, obtained in [83], for the entropy power inequality.

5.3 The Föllmer process

Our method is based on the Föllmer process which was defined in (7). Recall that if μ is a measure on $\mathbb{R}^d$ with expectation 0, a finite second moment matrix and a density f, relative to γ. Then, we associated to it the Föllmer drift, v_t, adapted to $\mathcal{F}_t$. Define

$$X_t := B_t + \int_0^t v_s(X_s)ds,$$

which satisfies $X_1 \sim \mu$, and as in (4),

$$X_t \overset{\text{law}}{=} tX_1 + \sqrt{t(1-t)}G,$$

for G a standard Gaussian, independent from X_1. Further recall the defining property of v_t,

$$\mathrm{Ent}(\mu||\gamma) = \frac{1}{2}\int_0^1 \mathbb{E}\left[\|v_t\|_2^2\right] dt. \tag{5.7}$$

As in the previous chapters, we also introduce the martingale counterpart,

$$\mathbb{E}\left[X_1|\mathcal{F}_t\right] = \int_0^t \Gamma_s dB_s,$$

for which it was shown in (10)

$$v_t = \int_0^t \frac{\Gamma_s - \mathrm{I}_d}{1-s} dB_s. \tag{5.8}$$

Finally, recall the representations, (11),(12),(13), which implied both the log-Sobolev and Talagrand's inequality,

$$\mathrm{I}(\mu||\gamma) \geq 2\mathrm{Ent}(\mu||\gamma) \geq \mathcal{W}_2^2\left(\mu,\gamma\right),$$

through

$$\mathrm{Tr}\int_0^1 \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{(1-t)^2} dt \geq \mathrm{Tr}\int_0^1 \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{1-t} dt \geq \mathrm{Tr}\int_0^1 \mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right] dt.$$

The above representations are especially useful, since they yield formulas for the deficits,

$$\delta_{\mathrm{LS}}\left(\mu\right) = \mathrm{Tr}\int_0^1 t \cdot \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{(1-t)^2} dt, \tag{5.9}$$

$$\delta_{\mathrm{Tal}}\left(\mu\right) \geq \mathrm{Tr}\int_0^1 t \cdot \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{1-t} dt. \tag{5.10}$$

The above formulas are the key to Corollary 5.6.

Proof of Corollary 5.6. Note that by (5.9) and (5.10), any estimate on $\delta_{\mathrm{Tal}}(\mu)$ which is achieved by bounding

$$\mathrm{Tr}\int_0^1 t \cdot \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{1-t} dt$$

from below will also imply a bound for $\delta_{\mathrm{LS}}(\mu)$. Since Theorem 5.3 and Theorem 5.5 are proved using this method, the corollary follows. □

5.3.1 Properties of the Föllmer process

Our objective is now clear: In order to produce any stability estimates it will be enough to show, roughly speaking, that the process Γ_t is far from I_d, not too close to time 0. In order to establish such claims we will use several other properties of the processes Γ_t, v_t, which we now state and prove. First, as in Lemma 1.28, it is possible use (5.8) along with integration by parts to obtain the identity:

$$\mathbb{E}\left[v_t \otimes v_t\right] = \frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t\right]}{1-t} + \left(\mathrm{Cov}(\mu) - \mathrm{I}_d\right). \tag{5.11}$$

Combining the fact that v_t is a martingale with (5.8) we also see

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[\|v_t\|_2^2\right] &= \mathrm{Tr}\frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{(1-t)^2} \\ &\geq \frac{1}{d}\left(\mathrm{Tr}\left(\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t\right]}{1-t}\right)\right)^2 = \frac{\left(\mathbb{E}\left[\|v_t\|_2^2\right] - \mathrm{Tr}\left(\mathrm{Cov}(\mu) - \mathrm{I}_d\right)\right)^2}{d},\end{aligned} \tag{5.12}$$

where we have used Cauchy-Schwartz for the inequality. Using this we prove the following two lemmas:

Lemma 5.7. *It holds that*

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right] = \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t}.$$

Proof. Since Γ_t is a symmetric matrix equation (5.8) implies

$$\frac{d}{dt}\mathbb{E}\left[v_t \otimes v_t\right] = \frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{(1-t)^2}.$$

Combined with (5.11), this gives

$$\frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{(1-t)^2} = \frac{d}{dt}\frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t\right]}{1-t} = \frac{\mathbb{E}\left[\mathrm{I}_d - \Gamma_t\right] - (1-t)\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right]}{(1-t)^2}.$$

Rearranging the terms yields the result. □

Lemma 5.8. *Suppose that* $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) \leq d$ *and let* v_t *be as defined above. Then:*

- *For* $0 \leq t \leq \frac{1}{2}$, $\mathbb{E}\left[\|v_t\|_2^2\right] \leq \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right] \frac{2d}{\mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right](1-2t)+2d}$.
- *For* $\frac{1}{2} \leq t \leq 1$, $\mathbb{E}\left[\|v_t\|_2^2\right] \geq \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right] \frac{2d}{\mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right](1-2t)+2d}$.

Proof. Since $\mathrm{Tr}\left(\mathrm{Cov}(\mu)\right) \leq d$, (5.12) gives

$$\frac{d}{dt}\mathbb{E}\left[\|v_t\|_2^2\right] \geq \frac{\left(\mathbb{E}\left[\|v_t\|_2^2\right]\right)^2}{d}.$$

The unique solution to the differential equation

$$g'(t) = \frac{g(t)^2}{d}, \text{ with inital condition } g\left(\frac{1}{2}\right) = \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right],$$

is given by

$$g(t) = \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right] \frac{2d}{\mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right](1-2t) + 2d}.$$

The result follows by Gronwall's inequality □

To get a different type of inequality, but of similar flavor, recall (4),

$$X_t \overset{\text{law}}{=} tX_1 + \sqrt{t(1-t)}G,$$

where G is a standard Gaussian, independent from X_1. Now, suppose that μ satisfies a Poincaré inequality with optimal constant $\mathrm{C_p}(\mu)$. In this case X_t satisfies a Poincaré inequality with a constant smaller than $t^2\mathrm{C_p}(\mu) + t(1-t)$. This follows from the fact that the Poincaré constant is sub-additive with respect to convolutions ([49]) and that if $X \sim \nu$ and $aX \sim \nu_a$ for some $a \in \mathbb{R}$, then $\mathrm{C_p}(\nu_a) = a^2\mathrm{C_p}(\nu)$. Applying the Poincaré inequality to $v_t(X_t)$, we get

$$\mathbb{E}\left[\|v_t\|_2^2\right] \leq \left(t^2\mathrm{C_p}(\mu) + t(1-t)\right)\left[\|\nabla v_t\|_2^2\right] = \left(t^2\mathrm{C_p}(\mu) + t(1-t)\right)\frac{d}{dt}\left[\|v_t\|_2^2\right], \tag{5.13}$$

where the equality is due to the fact that v_t is a martingale. Repeating the proof of Lemma 5.8 for the differential equation

$$g(t) = \left(t^2\mathrm{C_p}(\mu) + t(1-t)\right)g'(t), \text{ with inital condition } g\left(\frac{1}{2}\right) = \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right],$$

proves:

Lemma 5.9. *Assume that μ has a finite Poincaré constant $\mathrm{C_p}(\mu) < \infty$. Then, for v_t defined as above:*

- *For $0 \leq t \leq \frac{1}{2}$,*
$$\mathbb{E}\left[\|v_t\|_2^2\right] \leq \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right]\frac{(\mathrm{C_p}(\mu)+1)\,t}{(\mathrm{C_p}(\mu)-1)\,t+1}.$$
- *For $\frac{1}{2} \leq t \leq 1$,*
$$\mathbb{E}\left[\|v_t\|_2^2\right] \geq \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right]\frac{(\mathrm{C_p}(\mu)+1)\,t}{(\mathrm{C_p}(\mu)-1)\,t+1}.$$

5.4 Stability for Talagrand's transportation-entropy inequality

We begin this section by showing two ways the Föllmer process may be used to establish quantitative stability estimates. As before, μ is a fixed measure on $\mathbb{R}^d$ with finite second moment matrix. Γ_t and v_t are defined as in the previous section. Fix $t_0 \in [0,1]$, by (5.10), we see

$$\delta_{\mathrm{Tal}}(\mu) \geq t_0 \mathrm{Tr} \int_{t_0}^{1} \frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{1-t} dt.$$

Now, using (5.8), we obtain, by Fubini's theorem,

$$\int_{t_0}^{1} \left(\mathbb{E}\left[\|v_s\|_2^2\right] - \mathbb{E}\left[\|v_{t_0}\|_2^2\right]\right) ds = \mathrm{Tr} \int_{t_0}^{1} \int_{t_0}^{s} \frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{(1-t)^2} dt ds = \mathrm{Tr} \int_{t_0}^{1} \frac{\mathbb{E}\left[(\mathrm{I}_d - \Gamma_t)^2\right]}{1-t} dt,$$

and

$$\delta_{\mathrm{Tal}}(\mu) \geq t_0 \left(\int_{t_0}^{1} \mathbb{E}\left[\|v_t\|_2^2\right] dt - (1-t_0)\mathbb{E}\left[\|v_{t_0}\|_2^2\right] \right) \geq t_0(1-t_0)\left(2\mathrm{Ent}(\mu||\gamma) - \mathbb{E}\left[\|v_{t_0}\|_2^2\right]\right), \tag{5.14}$$

where we have used (11) and the fact that v_t is a martingale. Another useful bound will follow by applying (5.8) to rewrite (5.10) as

$$\delta_{\mathrm{Tal}}(\mu) \geq \mathrm{Tr} \int_{0}^{1} t(1-t) \cdot \frac{\mathbb{E}\left[(\Gamma_t - \mathrm{I}_d)^2\right]}{(1-t)^2} dt = \int_{0}^{1} t(1-t) \frac{d}{dt} \mathbb{E}\left[\|v_t\|_2^2\right] dt.$$

Integration by parts then gives

$$\delta_{\mathrm{Tal}}(\mu) \geq \int_{0}^{1} (2t-1)\mathbb{E}\left[\|v_t\|_2^2\right] dt. \tag{5.15}$$

At an informal level, the above formula becomes useful if one is able to show that $\mathbb{E}\left[\|v_t\|_2^2\right]$ is large for $t \geq \frac{1}{2}$ and small otherwise.

5.4.1 Measures with a finite Poincaré constant

We now assume that the measure μ has a finite Poincaré constant $\mathrm{C_p}(\mu) < \infty$.

Proof of Theorem 5.1. First, suppose that $\mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right] \leq \mathrm{Ent}(\mu||\gamma)$. In this case (5.14) shows

$$\delta_{\mathrm{Tal}} \geq \frac{1}{4}\mathrm{Ent}(\mu||\gamma).$$

Otherwise, $\mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right] > \mathrm{Ent}(\mu||\gamma)$, and plugging Lemma 5.9 into (5.15) shows

$$\begin{aligned}\delta_{\mathrm{Tal}}(\mu) &\geq \mathrm{Ent}(\mu||\gamma)\int_0^1 (2t-1)\frac{(\mathrm{C_p}(\mu)+1)t}{(\mathrm{C_p}(\mu)-1)t+1}dt \\ &= \mathrm{Ent}(\mu||\gamma)\frac{(\mathrm{C_p}(\mu)+1)\left(2-2\mathrm{C_p}(\mu)+(\mathrm{C_p}(\mu)+1)\ln\left(\mathrm{C_p}(\mu)\right)\right)}{(\mathrm{C_p}(\mu)-1)^3},\end{aligned}$$

where the equality relies on the fact

$$\begin{aligned}&\frac{d}{dt}\frac{(\mathrm{C_p}(\mu)+1)\left((\mathrm{C_p}(\mu)-1)t(\mathrm{C_p}(\mu)(t-1)-1-t)+(\mathrm{C_p}(\mu)+1)\ln\left((\mathrm{C_p}(\mu)-1)t+1\right)\right)}{\left(\mathrm{C_p}(\mu)-1\right)^3} \\ &\quad = (2t-1)\frac{(\mathrm{C_p}(\mu)+1)t}{(\mathrm{C_p}(\mu)-1)t+1}.\end{aligned}$$

The proof is complete. □

5.4.2 Measures with small covariance

Here we work under the assumption $\mathrm{Tr}(\mathrm{Cov}(\mu)) \leq d$ and prove Theorem 5.3.

Proof of Theorem 5.3. Denote $c_\mu = \mathbb{E}\left[\left\|v_{1/2}\right\|_2^2\right]$. We begin by considering the case $c_\mu \leq \mathrm{Ent}(\mu||\gamma)$. In this case, (5.14) shows

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{1}{4}\mathrm{Ent}(\mu||\gamma).$$

In the other case, $c_\mu > \mathrm{Ent}(\mu||\gamma)$ and Lemma 5.8, along with (5.15), gives

$$\begin{aligned}\delta_{\mathrm{Tal}}(\mu) &\geq 2d\int_0^1 \frac{c_\mu(2t-1)}{c_\mu(1-2t)+2d}dt \\ &= 2d\left(\frac{-d\ln\left(c_\mu+2d-2c_\mu t\right)-c_\mu t}{c_\mu}\right)\Big|_0^1 \\ &= \frac{2d\left(d\ln(2d+c_\mu)-d\ln(2d-c_\mu)-c_\mu\right)}{c_\mu} \\ &= 2d\left(\frac{2d\coth^{-1}\left(\frac{2d}{c_\mu}\right)}{c_\mu}-1\right).\end{aligned}$$

Note that (5.11) implies $c_\mu \leq 2d$, so the above is well defined. Also, for any $x \geq 1$, we have the inequality $\coth^{-1}(x) \cdot x - 1 \geq \frac{1}{3x^2}$, applying it to the previous bound then gives

$$\delta_{\mathrm{Tal}}(\mu) \geq \frac{c_\mu^2}{6d} > \frac{\mathrm{Ent}\left(\mu||\gamma\right)^2}{6d}.$$

□

We can get a dimension free bound by considering directions $v \in \mathbb{R}^d$ in which $\mathrm{Cov}(\mu)$ is strictly smaller than the identity. For this we use Lemma 5.7 to establish:

$$\frac{d}{dt}\mathbb{E}\left[\Gamma_t\right] = \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t^2\right]}{1-t} \preceq \frac{\mathbb{E}\left[\Gamma_t\right] - \mathbb{E}\left[\Gamma_t\right]^2}{1-t}.$$

Fix $v \in \mathbb{R}^d$, a unit vector, and define $f(t) = \langle v, \mathbb{E}\left[\Gamma_t\right] v\rangle$. As $\mathbb{E}\left[\Gamma_t\right]$ is symmetric, by Cauchy-Schwartz

$$\langle v, \mathbb{E}\left[\Gamma_t\right] v\rangle^2 \leq \left\langle v, \mathbb{E}\left[\Gamma_t\right]^2 v\right\rangle.$$

This implies

$$\frac{d}{dt} f(t) \leq \frac{f(t)(1-f(t))}{1-t}.$$

If $\langle v, \mathbb{E}\left[\Gamma_0\right] v\rangle = \lambda$, from Gronwall's inequality we get

$$\langle v, \mathbb{E}\left[\Gamma_t\right] v\rangle \leq \frac{\lambda}{(\lambda-1)t+1}. \tag{5.16}$$

Using this, we prove Theorem 5.2.

Proof of Theorem 5.2. For $\lambda_i < 1$, let w_i be the unit eigenvector of $\mathrm{Cov}(\mu)$, corresponding to λ_i. From (5.16) we deduce, for every $t \in [0,1]$,

$$0 \leq \langle w_i, \mathbb{E}\left[\Gamma_t\right] w_i\rangle \leq 1.$$

We now observe that as v_t is a martingale, and since μ is centered, it must hold that $v_0 = 0$, almost surely. Combining this with (5.11) shows $\mathbb{E}\left[\Gamma_0\right] = \mathrm{Cov}(\mu)$ and in particular

$$\langle w_i, \mathbb{E}\left[\Gamma_0\right] w_i\rangle = \lambda_i.$$

Using (5.16) and the fact that $\mathbb{E}\left[\Gamma_t\right]$ is symmetric, we obtain:

$$t\frac{\left\langle w_i, \mathbb{E}\left[\left(\mathrm{I}_d - \Gamma_t\right)^2\right] w_i\right\rangle}{1-t} \geq t\frac{\left(\langle w_i, \mathbb{E}\left[\mathrm{I}_d - \Gamma_t\right] w_i\rangle\right)^2}{1-t} \geq \frac{t\left(1 - \frac{\lambda_i}{(\lambda_i-1)t+1}\right)^2}{1-t} = t(1-t)\left(\frac{\lambda_i - 1}{(\lambda_i-1)t+1}\right)^2.$$

So, by (5.10),

$$\begin{aligned}\delta_{\mathrm{Tal}}(\mu) &\geq \mathrm{Tr}\int_0^1 t\cdot\frac{\mathbb{E}\left[(\mathrm{I}_d-\Gamma_t)^2\right]}{1-t}dt \geq \sum_{i=1}^d \mathbb{1}_{\{\lambda_i<1\}}\int_0^1 t\cdot\frac{\left\langle v_i,\mathbb{E}\left[(\mathrm{I}_d-\Gamma_t)^2\right]v_i\right\rangle}{1-t}dt\\ &\geq \sum_{i=1}^d \mathbb{1}_{\{\lambda_i<1\}}\int_0^1 t(1-t)\left(\frac{\lambda_i-1}{(\lambda_i-1)t+1}\right)^2 dt\\ &= \sum_{i=1}^d \frac{2(1-\lambda_i)+(\lambda_i+1)\log(\lambda_i)}{\lambda_i-1}\mathbb{1}_{\{\lambda_i<1\}}.\end{aligned}$$

□

5.4.3 Stability with respect to Gaussian mixtures

In this section we prove Theorem 5.5. Our proof is based on [101], but we use our framework to give an improved analysis. To control the relative entropy we will use a specialized case of the bound given in Lemma 1.13. We supply here a sketch of the proof for the convenience of the reader.

Lemma 5.10. *Let H_t be an $\mathcal{F}_t$-adapted matrix-valued processes and let N_t be defined by*

$$N_t = \int_0^t H_s dB_s.$$

Suppose that $\tilde{H}_t$ is such that for some $t_0 \in [0,1]$:

1. *$\tilde{H}_t = H_t$ almost surely, for $t < t_0$.*
2. *For $t \geq t_0$, $\tilde{H}_t$ is deterministic and $\tilde{H}_t \succ \mathrm{I}_d$.*

Then, if M_t is defined by

$$M_t = \int_0^t \tilde{H}_s dB_s,$$

we have

$$\mathrm{Ent}\left(N_1||M_1\right) \leq \mathrm{Tr}\int_{t_0}^1 \frac{\mathbb{E}\left[\left(H_t-\tilde{H}_t\right)^2\right]}{1-t}dt.$$

Proof. Define the process

$$Y_t = \int_0^t \tilde{H}_s dB_s + \int_0^t\int_0^s \frac{H_r-\tilde{H}_r}{1-r}dB_r ds.$$

Denote $u_t = \int_0^t \frac{H_s - \tilde{H}_s}{1-s} dB_s$, so that, $dY_t = \tilde{H}_t dB_t + u_t dt$, and, by assumption $u_t = 0$, whenever $t < t_0$. It follows that $Y_t = M_t$ for $t < t_0$ and that, using Fubini's theorem, $Y_1 = N_1$. Indeed,

$$Y_1 = \int_0^1 \tilde{H}_t dB_t + \int_0^1 \int_0^t \frac{H_s - \tilde{H}_s}{1-s} dB_s dt = \int_0^1 \tilde{H}_t dB_t + \int_0^1 \left(H_t - \tilde{H}_t \right) dB_t = N_1. \tag{5.17}$$

We denote now by P, the measure under which B is a Brownian motion. If

$$\mathcal{E} := \exp\left(- \int_0^1 \tilde{H}_t^{-1} u_t dB_t - \frac{1}{2} \int_0^1 \left\| \tilde{H}_t^{-1} u_t \right\|^2 dt \right),$$

and we define the tilted measure $Q = \mathcal{E} P$, then by Girsanov's theorem, $\tilde{B}_t = B_t + \int_0^t \tilde{H}_s^{-1} u_s ds$ is a Brownian motion under Q and we have the representation

$$Y_t = \int_0^t \tilde{H}_s d\tilde{B}_s.$$

If $t < t_0$, then as $u_t = 0$, we have $\tilde{B}_t = B_t$ and Y_{t_0} has the same law under Q and under P, which is the law of M_{t_0}. Moreover, for $t \geq t_0$, $\tilde{H}_t$ is deterministic. Therefore, it is also true that the law of

$$Y_{t_0} + \int_{t_0}^1 \tilde{H}_t d\tilde{B}_t,$$

under Q and the law of

$$Y_{t_0} + \int_{t_0}^1 \tilde{H}_t dB_t,$$

under P coincide. We thus conclude that, under Q, Y_1 has the same law as M_1 under P. In particular, if ρ is the density of Y_1 with respect to M_1, this implies

$$1 = \mathbb{E}_P \left[\rho(M_1) \right] = \mathbb{E}_Q \left[\rho(Y_1) \right] = \mathbb{E}_P \left[\rho(Y_1) \mathcal{E} \right].$$

By Jensen's inequality, under P,

$$0 = \ln\left(\mathbb{E}\left[\rho(Y_1) \mathcal{E} \right] \right) \geq \mathbb{E}\left[\ln\left(\rho(Y_1) \mathcal{E} \right) \right] = \mathbb{E}\left[\ln(\rho(Y_1)) \right] + \mathbb{E}\left[\ln(\mathcal{E}) \right].$$

But,

$$
\begin{aligned}
-\mathbb{E}\left[\ln(\mathcal{E})\right] &= \frac{1}{2}\int_0^1 \mathbb{E}\left[\left\|\tilde{H}_t^{-1}u_t\right\|^2\right]dt \leq \int_{t_0}^1 \mathbb{E}\left[\|u_t\|^2\right]dt \\
&= \mathrm{Tr}\int_{t_0}^1\int_{t_0}^s \frac{\mathbb{E}\left[\left(H_s-\tilde{H}_s\right)^2\right]}{(1-s)^2}dsdt = \mathrm{Tr}\int_{t_0}^1\int_s^1 \frac{\mathbb{E}\left[\left(H_s-\tilde{H}_s\right)^2\right]}{(1-s)^2}dtds \\
&= \mathrm{Tr}\int_{t_0}^1 \frac{\mathbb{E}\left[\left(H_s-\tilde{H}_s\right)^2\right]}{1-s}ds,
\end{aligned}
$$

and, from (5.17)

$$
\mathbb{E}_P\left[\ln(\rho(Y_1))\right] = \mathrm{Ent}(N_1||M_1),
$$

which concludes the proof. □

Remark 5.11. In order to apply Girsanov's theorem in the proof above, one must also require some integrability condition from the drift u_t. It will suffice to assume

$$
\int_0^1 \mathbb{E}\left[\left\|\tilde{H}_t^{-1}u_t\right\|^2\right]dt < \infty.
$$

Indeed, if $\int_0^1 \left\|\tilde{H}_t^{-1}u_t\right\|^2 dt$ is uniformly bounded, then Novikov's criterion applies. The general case may then be obtained by an approximation argument (see [165, Proposition 1] for more details). In our application below this condition will be satisfied.

We are now in a position to prove that stability with respect to Gaussian mixtures holds in relative entropy.

Proof of Theorem 5.5. Fix $t_0 \in [0,1]$, by (5.10) we get

$$
\delta_{\mathrm{Tal}}(\mu) \geq t_0 \mathrm{Tr}\int_{t_0}^1 \frac{\mathbb{E}\left[(\mathrm{I}_d-\Gamma_t)^2\right]}{1-t}dt. \tag{5.18}
$$

Define the matrix-valued process

$$
\tilde{\Gamma}_t = \begin{cases} \Gamma_t & 0 \leq t < t_0 \\ \frac{1-t_0}{t_0(t-2)+1}\mathrm{I}_d & t_0 \leq t \leq 1 \end{cases},
$$

and the martingale

$$M_t = \int_0^t \tilde{\Gamma}_s dB_s.$$

One may verify that

$$\int_{t_0}^1 \left(\frac{1-t_0}{t_0(t-2)+1} \right)^2 dt = 1,$$

which implies, $M_1 - M_{t_0} = \int_{t_0}^1 \tilde{\Gamma}_t(M_t) dB_t \sim \gamma$. Also, from (9),

$$M_t = \int_0^{t_0} \tilde{\Gamma}_t dB_t = \int_0^{t_0} \Gamma_t dB_t = \mathbb{E}\left[X_1 | \mathcal{F}_{t_0}\right].$$

If ν_{t_0} is the law of $\mathbb{E}\left[X_1|\mathcal{F}_{t_0}\right]$, then since $\{B_s\}_{s>t_0}$ is independent from $\mathbb{E}\left[X_1|\mathcal{F}_{t_0}\right]$, we have that $\nu_{t_0} * \gamma$ is the law of M_1. We now invoke Lemma 5.10 with the process $\mathbb{E}\left[X_1|\mathcal{F}_t\right]$ as N_t. Since $\tilde{\Gamma}_t$ meets the conditions of the lemma, we get

$$\begin{aligned}
\mathrm{Ent}(X_1||M_1) = \mathrm{Ent}(\mu||\nu_{t_0} * \gamma) &\leq \mathrm{Tr} \int_{t_0}^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \tilde{\Gamma}_t\right)^2\right]}{1-t} dt \\
&\leq 2\mathrm{Tr} \int_{t_0}^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]}{1-t} dt + 2\mathrm{Tr} \int_{t_0}^1 \frac{\mathbb{E}\left[\left(\tilde{\Gamma}_t - \mathrm{I}_d\right)^2\right]}{1-t} dt.
\end{aligned}$$

Observe that by showing that the above integrals are finite we will also verify the integrability condition from Remark 5.11. Applying (5.18),

$$2\mathrm{Tr} \int_{t_0}^1 \frac{\mathbb{E}\left[\left(\Gamma_t - \mathrm{I}_d\right)^2\right]}{1-t} dt \leq 2\frac{\delta_{\mathrm{Tal}}(\mu)}{t_0}.$$

To bound the second term we calculate

$$\begin{aligned}
2\mathrm{Tr} \int_{t_0}^1 \frac{\mathbb{E}\left[\left(\tilde{\Gamma}_t - \mathrm{I}_d\right)^2\right]}{1-t} dt &= 2d \int_{t_0}^1 \frac{\left(\frac{1-t_0}{t_0(t-2)+1} - 1\right)^2}{1-t} dt \\
&= 2d \left(-\ln(1 + t_0(t-2)) - \frac{1-t_0}{2(t-t_0)+1} \right) \Big|_{t_0}^1 \\
&= 2d \left(\ln(1-t_0) + \frac{t_0}{1-t_0} \right).
\end{aligned}$$

Combining the last displays, we get

$$\mathrm{Ent}(\mu||\nu_{t_0} * \gamma) \leq 2\left(\frac{\delta_{\mathrm{Tal}}(\mu)}{t_0} + d\left(\ln(1-t_0) + \frac{t_0}{1-t_0}\right)\right).$$

Suppose that $\delta_{\mathrm{Tal}}(\mu) \geq d$, then choosing $t_0 = \frac{1}{2}$ gives

$$\frac{\mathrm{Ent}(\mu||\nu_{t_0} * \gamma)}{6} \leq \delta_{\mathrm{Tal}}(\mu).$$

Otherwise, $\delta_{\mathrm{Tal}}(\mu) < d$ and we choose $t_0 = \left(\frac{\delta_{\mathrm{Tal}}(\mu)}{d}\right)^{\frac{1}{3}} \leq \frac{1}{2}$. A second order approximation, shows that for $s \in [0, \frac{1}{2}]$,

$$\ln(1-s) + \frac{s}{1-s} \leq 2s^2.$$

Hence, for the above choice of t_0,

$$\mathrm{Ent}(\mu||\nu_{t_0} * \gamma) \leq 2\frac{\delta_{\mathrm{Tal}}(\mu)}{t_0} + 4d{t_0}^2 = 3\delta_{\mathrm{Tal}}(\mu)^{\frac{2}{3}} d^{\frac{1}{3}}.$$

This implies

$$\frac{1}{3\sqrt{3}} \frac{\mathrm{Ent}(\mu||\nu_{t_0} * \gamma)^{\frac{3}{2}}}{\sqrt{d}} \leq \delta_{\mathrm{Tal}}(\mu),$$

which is the desired claim. Finally, by the law of total variance, it is immediate that

$$\mathrm{Cov}\left(\nu_{t_0}\right) \preceq \mathrm{Cov}\left(\mu\right).$$

□

5.5 An application to Gaussian concentration

We now show that our stability bounds imply an improved Gaussian concentration inequality for concave functions.

Corollary 5.12. *Let f be a concave function and $G \sim \gamma$ in $\mathbb{R}^d$. Suppose that f is even, then for any $t \geq 0$,*

$$\mathbb{P}\left(f(G) \geq t\right) \leq e^{-\frac{4t^2}{7\mathbb{E}\left[\|\nabla f(G)\|_2^2\right]}}.$$

Before proving the result we mention that our proof follows the one presented in [212]. We use Theorem 5.1 to improve the constant obtained there. One should also compare the corollary to the main result of [205] which used Ehrhard's inequality in order to show that $\mathbb{E}\left[\|\nabla f(G)\|_2^2\right]$ may be replaced by the smaller quantity $\mathrm{Var}(f(G))$, at the cost of a worse constant in the exponent.

The assumption that f is even is used here for simplicity and could be relaxed.

Proof of Corollary 5.12. For $\lambda > 0$, denote the measure $\nu_\lambda = \frac{e^{\lambda f}}{\mathbb{E}_\gamma\left[e^{\lambda f}\right]} d\gamma$ and let (X, Y) be a random vector in $\mathbb{R}^{2d}$ which is a realization of the optimal coupling between ν_λ and γ. That is, $X \sim \nu_\lambda, Y \sim \gamma$ and

$$\mathcal{W}_2(\nu_\lambda, \gamma) = \sqrt{\mathbb{E}\left[\|X - Y\|_2^2\right]}.$$

As f is concave, we have by using Cauchy-Schwartz:

$$\begin{aligned}\mathbb{E}_{\nu_\lambda}\left[\lambda f\right] - \mathbb{E}_\gamma\left[\lambda f\right] &\leq \mathbb{E}\left[\langle \nabla \lambda f(Y), X - Y\rangle\right] \leq \sqrt{\lambda^2 \mathbb{E}\left[\|\nabla f(Y)\|_2^2\right]}\sqrt{\mathbb{E}\left[\|X - Y\|_2^2\right]} \\ &= \sqrt{\lambda^2 \mathbb{E}_\gamma\left[\|\nabla f\|_2^2\right]}\mathcal{W}_2(\nu_\lambda, \gamma). \end{aligned} \tag{5.19}$$

Since f is concave, ν_λ has a log-concave density with respect to the standard Gaussian. For such measures, Brascamp-Lieb's inequality ([50]) dictates that $\mathrm{C_p}(\nu_\lambda) \leq 1$. Note that

$$\frac{(x+1)(2 - 2x + (x+1)\ln(x)}{(x-1)^3} \geq \frac{1}{3}, \text{ whenever } x \in [0, 1].$$

In this case, since f is even and ν_λ is centered, Theorem 5.1 gives us,

$$\delta_{\mathrm{Tal}}(\nu_\lambda) \geq \frac{1}{4}\mathrm{Ent}\left(\nu_\lambda || \gamma\right),$$

which is equivalent to

$$\mathcal{W}_2^2\left(\nu_\lambda, \gamma\right) \leq \frac{7}{4}\mathrm{Ent}(\nu_\lambda || \gamma).$$

Combining this with (5.19) and the assumption, $\mathbb{E}_\gamma\left[\lambda f\right] = 0$, yields

$$\mathbb{E}_{\nu_\lambda}\left[\lambda f\right] \leq \sqrt{\lambda^2 \frac{7}{4}\mathbb{E}_\gamma\left[\|\nabla f\|_2^2\right]\mathrm{Ent}(\nu_\lambda || \gamma)}.$$

For any $x, y \geq 0$ we have the inequality, $\sqrt{xy} \leq \frac{x}{4} + y$. Observe as well that

$$\mathrm{Ent}(\nu_\lambda || \gamma) = \mathbb{E}_{\nu_\lambda}\left[\lambda f\right] - \ln\left(\mathbb{E}_\gamma\left[e^{\lambda f}\right]\right).$$

Thus, $\ln\left(\mathbb{E}_\gamma\left[e^{\lambda f}\right]\right) \leq \lambda^2 \frac{7}{16}\mathbb{E}_\gamma\left[\|\nabla f\|_2^2\right]$. By Markov's inequality, for any $\lambda, t > 0$

$$\mathbb{P}\left(f(G) \geq t\right) = \mathbb{P}\left(e^{\lambda f(G)} \geq e^{\lambda t}\right) \leq \mathbb{E}_\gamma\left[e^{\lambda f}\right] e^{-\lambda t} \leq \exp\left(\lambda^2 \frac{7}{16}\mathbb{E}_\gamma\left[\|\nabla f\|_2^2\right] - \lambda t\right).$$

We now optimize over λ to obtain,

$$\mathbb{P}\left(f(G) \geq t\right) \leq e^{-\frac{4t^2}{7\mathbb{E}_\gamma\left[\|\nabla f\|_2^2\right]}}.$$

□

6

Stability of Invariant Measures, with Applications to Stability of Moment Measures and Stein Kernels

6.1 Introduction

Let X_t, Y_t be stochastic processes in $\mathbb{R}^d$ which satisfy the SDEs,

$$dX_t = a(X_t)dt + \sqrt{2\tau(X_t)}dB_t, \;\; dY_t = b(Y_t)dt + \sqrt{2\sigma(Y_t)}dB_t. \tag{6.1}$$

Here B_t is a standard Brownian motion, a, b are vector-valued functions, and σ, τ take values in the cone of symmetric $d \times d$ positive definite matrices, which we shall denote by $\mathcal{S}_d^{++}$. Given X_0 and Y_0, we shall write the marginal laws of the processes as $X_t \sim \mu_t$ and $Y_t \sim \nu_t$.

Suppose that, in some sense to be made precise later, a is close to b, and τ is close σ. One can ask whether the measure μ_t must then be close to ν_t. Our goal here is to study the quantitative regime of this problem. The method used here is an adaptation of a technique developed by Crippa and De Lellis for transport equations. That method was introduced in the SDE setting in [67, 167]. Our implementation here will be a bit different, to allow for estimates in weighted Sobolev space that behave better for large times, and will allow us to compare the invariant

measures of the two processes, under suitable ergodic assumptions.

We will be especially interested in the case where, X_t and Y_t admit unique invariant measures, which we shall denote, respectively, as μ and ν. In this setting, we will think about ν as the reference measure and quantify the discrepancy in the coefficients as:

$$\beta := \|a - b\|_{L^1(\nu)} + \|\sqrt{\sigma} - \sqrt{\tau}\|_{L^2(\nu)}.$$

Remark that, throughout the paper, unless otherwise specified, when a matrix norm is considered, we treat it as the Hilbert-Schmidt norm. Our main result is an estimate of the form (see precise formulation below)

$$\mathrm{dist}(\mu, \nu) \leq h(\beta),$$

where $\mathrm{dist}(\cdot, \cdot)$ stands for an appropriate notion of distance, which will here be a transport distance, and $\lim_{\beta \to 0} h(\beta) = 0$.

While the assumption of unique invariant measures is certainly non-trivial to verify for general coefficients, we also show how to apply our stability estimates in some specific cases of interest. In particular, we will show that if two uniformly log-concave measures satisfy certain similar integration by parts formulas, in the sense arising in Stein's method, then the measures must be close. This problem was the original motivation of our study.

6.1.1 Background on stability for transport equations

Part of the present work is a variant in the SDE setting of a now well-established quantitative theory for transport equations with non-smooth coefficients, pioneered by Crippa and De Lellis [87].

As demonstrated by the DiPerna-Lions theory ([94] and later estimates in [87]), there is a significant difference in the stability of solutions to differential equations when the coefficients are Lipschitz continuous versus when they only belong to some Sobolev space (and are not necessarily globally Lipschitz). Our focus will be on the latter, and arguably more challenging, case. As we shall later discuss, the techniques can be carried over to the setting of stochastic differential equations we are interested in here, as worked out in [67, 167].

The strategy for quantitative estimates, introduced in [87], in the Lagrangian setting relies on controlling the behavior of $\ln(1+|X_t - Y_t|/\delta)$ for two flows (X_t) and (Y_t), with a parameter δ that will be very small, of the order of the difference between the vector fields driving the flows. A crucial idea is the use of the Lusin-Lipschitz property of Sobolev vector fields [169], which allows to get Lipschitz-like estimates on large regions, in a controlled way. We will discuss this idea in more details in Section 6.2.2. The ideas of [87] were adapted to the Eulerian setting in [215, 216], using a transport distance with logarithmic cost.

Existence and uniqueness of solutions to SDEs and Fokker-Planck equations with non-smooth coefficients by adapting DiPerna-Lions theory was first addressed in [117, 158], inspiring many further developments, such as [67, 109, 232, 244]. We will not discuss much the issue of well-posedness here, and focus on more quantitative aspects of the problem. We refer for example to [167, Section 3.1] and [232] for a comprehensive discussion of the issues, in particular with respect to the different ways of defining a notion of solution. As pointed out in [67, 167], the kind of quantitative methods used here could also prove well-posedness by approximation with processes with smoother coefficients.

6.2 Results

6.2.1 Main result

We consider two diffusion processes of the form (6.1), and assume that they admit unique invariant measures μ and ν. We fix some real number $p \geq 2$ with q its Hölder conjugate, so that $\frac{1}{q} + \frac{1}{p} = 1$. We then make the following assumptions:

H1. (Regularity of coefficients) There exists a function $g : \mathbb{R}^d \to \mathbb{R}$, such that for almost every $x, y \in \mathbb{R}^d$:

$$\|a(x) - a(y)\|, \|\sqrt{\tau(x)} - \sqrt{\tau(y)}\| \leq (g(x) + g(y))\|x - y\|, \tag{6.2}$$

and

$$\|g\|_{L^{2q}(\mu)} < \infty.$$

To simplify some notations later on, we also assume that $g \geq 1$ pointwise, which does not strengthen the assumption since we can always replace g by $\max(g, 1)$.

H2. (Integrability of relative density) Both μ and ν have finite second moments, and it holds that

$$\left\| \frac{d\nu}{d\mu} \right\|_{L^p(\mu)} < \infty.$$

H3. (Exponential convergence to equilibrium) There exist constants $\kappa, C_H > 0$ such that for any initial data μ_0 and any $t \geq 0$ we have

$$\mathcal{W}_2(\mu_t, \mu) \leq C_H e^{-\kappa \cdot t} \mathcal{W}_2(\mu_0, \mu).$$

We denote for the second moments

$$m_2^2(\mu) = \int_{\mathbb{R}^d} |x|^2 d\mu, \; m_2^2(\nu) = \int_{\mathbb{R}^d} |x|^2 d\nu.$$

Concerning Assumption (H1), it is the Lusin-Lipschitz property we mentioned above, and which we shall discuss in some depth in Section 6.2.2 below. One should think about it as being a generalization of Lipschitz continuity, as it essentially means σ and a may be well approximated by Lipschitz functions on arbitrarily large sets. In particular, this property holds for Sobolev functions when the reference measure is log-concave. When $C_H = 1$, the third assumption corresponds to contractivity, which for reversible diffusion processes is equivalent to a lower bound on the Bakry-Emery curvature [241]. Allowing for a constant $C_H > 1$ allows to cover other examples, such as hypocoercive dynamics, notably because the assumption is then invariant by change of equivalent metric, up to the value of C_H. See for example [28, 182]

Now, for $R > 0$, define the truncated quadratic Wasserstein distance by:

$$\widetilde{\mathcal{W}}_{2,R}^2(\nu,\mu) := \inf_{\pi} \int \min(|x-y|^2, R) d\pi, \tag{6.3}$$

where the infimum is taken over all couplings of μ and ν. Here, we say that π is a coupling of μ and ν if π is a measure on $\mathbb{R}^{2d}$ whose marginals on the first and last d coordinates equal μ and ν, respectively. This is a distance on the space of probability measures, weaker than the classical $\mathcal{W}_2$. With the above notations our main result reads:

Theorem 6.1. *Assume* (H1),(H2) *and* (H3) *hold, and denote*

$$\beta := \|a-b\|_{L^1(\nu)} + \|\sqrt{\sigma} - \sqrt{\tau}\|_{L^2(\nu)}.$$

Then, for any $R > 1$,

$$\widetilde{\mathcal{W}}_{2,R}^2(\nu,\mu) \leq 100 C_H^2 R \cdot \|g\|_{L^{2q}(\mu)}^2 \, \|d\nu/d\mu\|_{L^p(\mu)} \frac{\ln\left(\ln\left(1+\frac{R}{\beta}\right)\right) + \ln\left(m_2(\mu)+m_2(\nu)\right) + \kappa \cdot R}{\kappa \cdot \ln\left(1+\frac{R}{\beta}\right)},$$

where $\frac{1}{q} + \frac{1}{p} = 1$.

Remark 6.2. Essentially, the theorem says that $\widetilde{\mathcal{W}}_{2,R}(\nu,\mu)$ decreases at a rate which is proportional to $\sqrt{\ln\left(1+\frac{R}{\beta}\right)^{-1}}$. We could improve the rate to $\ln\left(1+\frac{R}{\beta}\right)^{-1}$ by considering a truncated $\mathcal{W}_1$ distance, as shall be discussed in Remark 6.8. However, for our application to Stein kernels it is more natural to work with $\mathcal{W}_2$.

Let us discuss now the role of the term $\|d\nu/d\mu\|_{L^p(\mu)}$ in Theorem 6.1. In order to use $\|a-b\|_{L^1(\nu)} + \|\sqrt{\sigma}-\sqrt{\tau}\|_{L^2(\nu)}$ as a measure of discrepancy, it seems necessary that the supports of μ and ν intersect, otherwise we could just change σ and a on a μ-negligible set and have $\beta = 0$ in the conclusion of the theorem, which obviously fails in this particular situation. Thus, since the bound in Theorem 6.1 only makes sense when $\|d\nu/d\mu\|_{L^p(\mu)}$ is finite, one may view this term as an a-priori guarantee on the common support of μ and ν.

The logarithmic rate obtained here may seem quite weak. In the setting of transport equations, the logarithmic bounds obtained by the method considered here are sometimes actually sharp [216]. We do not know much about optimality in the stochastic setting, since it may be that the presence of noise would help, while the method used here cannot do better in the stochastic setting than in the deterministic setting.

As alluded to in the introduction, if the coefficients a and $\sqrt{\tau}$ are actually L-Lipschitz, in which case $g \equiv \frac{L}{2}$ in (H1), then one may greatly improve the rate in Theorem 6.1.

Theorem 6.3. *Assume a and $\sqrt{\tau}$ are L-Lipschitz and that (H3) holds, and denote*

$$\beta := \|a - b\|_{L^2(\nu)} + \|\sqrt{\tau} - \sqrt{\sigma}\|_{L^2(\nu)}.$$

Then,

$$\mathcal{W}_2(\nu, \mu) \leq 15 C_H^{\frac{4L^2+1}{2\kappa}} \beta \left(\frac{L}{\kappa} + 1 \right).$$

This type of estimate is part of the folklore, and a version of it appears for example in [45].

6.2.2 About the Lusin-Lipschitz property for Sobolev functions

We will now discuss in some more depth Assumption (H1). As mentioned previously, it is motivated by the Lusin-Lipschitz property of Sobolev functions with respect to the Lebesgue measure [169]: if a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ satisfies $\int |\nabla f|^p dx < \infty$ then for a.e. $x, y \in \mathbb{R}^d$ we have

$$|f(x) - f(y)| \leq (M|\nabla f|(x) + M|\nabla f|(y))|x - y|, \tag{6.4}$$

where M is the Hardy-Littlewood maximal operator, defined on a non-negative function g as,

$$Mg(x) := \sup_{r>0} |B_r|^{-1} \int_{B_r(x)} g(y) dy,$$

C a dimension-free constant, and B_r is the Euclidean ball of radius r. This operator satisfies the dimension-free continuity property

$$||Mf||_{L^p(dx)} \leq C_p ||f||_{L^p(dx)},$$

when $p > 1$. The dimension-free bound on C_p is due to E. Stein [226]

In particular, if $\nabla f \in L^p(dx)$, then f is λ-Lipschitz on the regions where $M|\nabla f|$ is smaller than $\lambda/2$, which are large when λ is, by the Markov inequality. The important distinction between using an estimate on $M|\nabla f|$ instead of ∇f is that, even if both $\nabla f(x)$ and $\nabla f(y)$ are controlled, we do not automatically get an estimate on $f(x) - f(y)$, since the straight line from x to y may well go through a region where $|\nabla f|$ is arbitrarily large. The use of (6.4) nicely bypasses this issue.

An important issue for the applications we shall discuss here is that working in functional spaces weighted with the Lebesgue measure is not always the most natural when dealing with stochastic processes. It is often preferable to work in $L^p(\mu)$ with a probability measure adapted to the problem considered, which here shall be the invariant measure of the reference stochastic process. However, in general the maximal operator has no reason to be continuous over $L^p(\mu)$, unless μ has density with respect to the Lebesgue measure that is uniformly bounded from above and below on its support. As soon as μ is not compactly supported, this cannot be the case. Therefore, we shall make strong use of a work of Ambrosio, Brué and Trevisan [7], which proves a Lusin-type property for Sobolev functions with respect to a log-concave measure. The proof uses an operator different from the Hardy-Littlewood maximal operator, more adapted to the setting. We shall not discuss here the specifics of that operator, since we only need the Lusin property, and not the maximal operator itself. The exact statement of their result, in the restricted setting of log-concave measures on $\mathbb{R}^d$, is as follows:

Proposition 6.4. *Let μ be a log-concave measure, and $p \geq 2$. Then for any function $f \in W^{1,p}(\mu)$, there exists a function g such that $|f(x) - f(y)| \leq (g(x) + g(y))|x - y|$ for a.e. x and y, and with $||g||_{L^p(\mu)} \leq C_p ||\nabla f||_{L^p(\mu)}$ with C_p some universal constant, that only depends on p.*

This statement is proved in [7, Theorem 4.1], and also holds for maps taking values in some Hilbert space. It is written there only for $p = 2$, but the reason for that restriction is that they work in the more general setting of possibly nonsmooth RCD spaces, rather than just $\mathbb{R}^d$ endowed with a measure. The only point where they require the restriction to $p = 2$ is when using the Riesz inequality [7, Remark 3.9], which in the smooth setting is known for general values of p, as proved in [19].

6.2.3 Related works

As mentioned previously, the adaptation of the Crippa-De Lellis method to derive quantitative estimates for stochastic differential equations was already considered in [67] and [167].

The results of [167] give stability estimates with bounds that depend on $||\nabla \sigma||_{L^p(dx)}$. Considering estimates weighted with the Lebesgue measure allows to use the Hardy-Littlewood maximal function directly. As mentioned above, the main focus here is to get estimates that are weighted with respect to a probability measure adapted to the problem, which may behave very differently, for example when the coefficients of the two SDE are uniformly close, but not compactly supported.

To use estimates in weighted space, [67] considers functions such that

$$\int (M|f|)^2 d\mu_t + \int (M|\nabla f|)^2 d\mu_t < \infty,$$

with μ_t the flow of the SDE. The authors also consider other function spaces of the same nature, sharper in dimension one, or that handle weaker integrability conditions on $M|\nabla f|$ than L^p (but

stronger than L^1). Since the space depends on the law of the flow at all times, it may be difficult to determine estimates on such norms. For the application considered in Section 6.2.4, we do not know whether the approach of [67] could apply.

We shall also focus on establishing very explicit quantitative estimates in transport distance, highlighting in particular the dependence on the dimension.

Another approach was developed in [45] to directly obtain relative entropy estimates between the distributions at finite times. The upside of that approach is that the quantitative estimates are quite stronger, depending polynomially on some distance between the coefficients. The two downsides are that they depend on stronger Sobolev norms, requiring that the derivatives of the two diffusion coefficients are close in some sense, as well as a-priori Fisher information-like bounds on the relative densities, rather than L^p bounds. Fisher information-like estimates were then derived in [46] by directly comparing generators via a Poisson equation, also using stronger Sobolev norms.

Finally, when the two diffusion coefficients match, one can derive relative entropy bounds via Girsanov's theorem. Unfortunately, this strategy cannot work when the two diffusion coefficients differ.

6.2.4 An application to Stein kernels

We now explain how our result might be applied in the context of Stein's method for bounding distances between probability measures. Given a measure ν on $\mathbb{R}^d$ and $X \sim \nu$, we slightly relax the definition of Stein kernels, given in the introduction to this thesis and define a matrix valued map $\tau : \mathbb{R}^d \to M^d(\mathbb{R})$, such that:

$$\mathbb{E}\left[\langle \nabla f(X), X\rangle\right] = \mathbb{E}\left[\langle \nabla^2 f(X), \tau(X)\rangle_{HS}\right]. \tag{6.5}$$

The main difference is that here we consider test functions $f : \mathbb{R}^d \to \mathbb{R}$. As before, the map $\tau_\gamma \equiv \mathrm{I_d}$, which is constantly identity, is a Stein kernel for γ. Stein's lemma suggests that if τ_ν is close to the identity then ν should be close γ. This is in fact true, and there are many examples of precise quantitative statements implementing this idea, for various distances between measures, such as transport distances, the total variation distance, or the Kolmogorov distance in dimension 1. The one most relevant to the present work is inequality (20), which states that for any τ which is a Stein kernel for a measure ν,

$$\mathcal{W}_2^2(\nu, \gamma) \leq \|\tau - \mathrm{I_d}\|_{L^2(\nu)}. \tag{6.6}$$

The proof of this inequality strongly relies on Gaussian algebraic identities, such as the Mehler formula for the Ornstein-Uhlenbeck semigroup. We are interested in similar estimates when neither of the measures are Gaussian. The main motivation comes from the fact that many ways of implementing Stein's method, including the one developed in [161], are hard to use for target

measures that do not satisfy certain exact algebraic properties (typically, explicit knowledge of the eigenvectors of an associated Markov semigroup). We shall prove a weaker inequality holds for certain non-Gaussian reference measures and for one particular construction of Stein kernels. To understand this construction we require the following definition.

Definition 6.5 (Moment map)**.** Let μ be a measure on $\mathbb{R}^d$. A moment map of μ is a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $e^{-\varphi}$ is a centered probability density whose push-forward by $\nabla\varphi$ is μ.

As was shown in [80, 213], if μ is centered and has a finite first moment and a density, then its moment map exists and is unique as long as we enforce essential continuity at the boundary of its support. The moment map φ can be realized as the optimal transport map between some source log-concave measure and the target measure μ, where we enforce that gradient of the source measure's potential must equal the transport map itself. The correspondence between the convex function φ and the measure μ is actually a bijection, up to a translation of φ, and the measure associated with a given convex function is known as its moment measure.

If μ has a density ρ, then φ solves the Monge-Ampère-type PDE

$$e^{-\varphi} = \rho(\nabla\varphi)\det(\nabla^2\varphi).$$

This PDE, sometimes called the toric Kähler-Einstein PDE, first appeared in the geometry literature [34, 95, 164, 242], where it plays a role in the construction of Kähler-Einstein metrics on certain complex manifolds. Variants with different nonlinearities have recently been considered, for example in [150].

The connection between moment maps and Stein kernels was made in [112]. Specifically, it was proven that if φ is the moment map of μ, then (up to regularity issues) the matrix valued map,

$$\tau_\mu := \nabla^2\,\varphi(\nabla\varphi^{-1}), \tag{6.7}$$

is a Stein kernel for μ. Since φ is a convex function, τ_μ turns out to be supported on positive semi-definite matrices. For this specific construction of a Stein kernel we will prove the following analogue of (6.6).

Theorem 6.6. *Let μ be a log-concave measure on $\mathbb{R}^d$ such that*

$$\alpha \mathrm{I_d} \leq -\nabla^2 \ln\left(d\mu/dx\right) \leq \frac{1}{\alpha}\mathrm{I_d},$$

for some $\alpha \in (0, 1]$ and let τ_μ be its Stein kernel defined in (6.7). If ν is any other probability measure and σ is a Stein kernel for ν which is almost surely positive definite and bounded from

below, then for $\beta = \|\sqrt{\tau} - \sqrt{\sigma}\|_{L^2(\mu)}$ *and* $M = \max\left(m_2^2(\mu), m_2^2(\nu)\right)$,

$$\mathcal{W}_2^2(\mu,\nu) \leq C\alpha^{-6}d^3M^2\ln(M)^2\|d\nu/d\mu\|_\infty \frac{\ln\left(\ln\left(1+\frac{M^2}{\beta}\right)\right) + M^2\ln(M)^2}{\ln\left(1+\frac{M^2}{\beta}\right)}.$$

Moreover, if μ *is radially symmetric, has full support, and* $d > c$*, for some universal constant* $c > 0$,

$$\mathcal{W}_2^2(\mu,\nu) \leq C\alpha^{-20}d^{7/2}M^2\ln(M)^2\|d\nu/d\mu\|_{L^2(\mu)} \frac{\ln\left(\ln\left(1+\frac{M^2}{\beta}\right)\right) + M^2\ln(M)^2}{\ln\left(1+\frac{M^2}{\beta}\right)}.$$

Finally, if $\sqrt{\tau_\mu}$ *is L-Lipschitz, then*

$$\mathcal{W}_2^2(\mu,\nu) \leq 100\alpha^{-(4L^2+1)}\left(2L+1\right)\beta.$$

It should be emphasized that, except in dimension one, Stein kernels are *not* unique. Different constructions than the one studied here have been provided for example in [68,85,178,195]. Unlike the functional inequalities of [161] for the Gaussian measure, our results will only work for the Stein kernels constructed from moment maps (at least for one of the two measures). In particular, in order to define a stochastic flow from a Stein kernel, we must require the kernel to take positive values, which to our knowledge is not guaranteed for other constructions.

While this estimate is somewhat weak, it seems to be one of the few instances where we can estimate a distance from a discrepancy for a class of target measures, without explicit algebraic requirements for an associated Markov generator. Recently, there has been progress on implementing Stein's method for wide classes of target measures via Malliavin calculus [111, 127].

Note that if one of the two measures is Gaussian, since the natural Stein kernel for the standard Gaussian is constant, and hence Lipschitz, one could use the stronger Theorem 6.3 to get a stability estimate, which would still be weaker than that of [161], but with the sharp exponent.

One may wonder why we do not prove this type of estimate directly using Stein's method. The key difference lies in that we do not need a second-order regularity bound on solutions of Stein's equation, which we do not even know how to prove here. To be more precise, the natural way to try to use Stein's method for this problem would be to apply the generator approach using the generator of the process $dX_t = -X_t dt + \sqrt{2\tau(X_t)}dB_t$, where τ is the Stein kernel for μ. Applying Stein's method to bound say the $\mathcal{W}_1$ distance would require us to bound $||\nabla f||_\infty$ and $||\nabla^2 f||_\infty$ for solutions to the Stein equation

$$-x\cdot\nabla f + \mathrm{Tr}(\tau\nabla^2 f) = g - \int g d\mu$$

for arbitrary 1-Lipschitz data g. While a slightly stronger version of Assumption (H3) could be used to prove bounds on $||\nabla f||_\infty$, the techniques used here would not help to bound $||\nabla^2 f||_\infty$. So using Stein's method would require some ingredients we do not have. Indeed, in general proving second-order bounds is usually the most difficult step in implementing Stein's method via diffusion processes, and in the literature has mostly been done for measures satisfying certain algebraic properties, such as having an explicit orthogonal basis of polynomials that are eigenvectors for an associated diffusion process (for example Gaussians or gamma distributions).

6.3 Proofs of stability bounds

A rough outline of the proofs is as follows: as a first step we will use Itô's formula to show that (H1) implies bounds on the measures μ_t and ν_t, for fixed t. Indeed, (H1) will allow us to replace quantities like $\|\tau(X_t) - \tau(Y_t)\|$, which will arise through the use of Itô's formula by something more similar to $\|X_t - Y_t\|$. We will then use (H2) to transfer those estimate to the measure ν as well.
After establishing that μ_t and ν_t are close, (H3) will be used to establish the same for μ and ν.

We first prove this in the easier case of globally Lipschitz coefficients encompassed by Theorem 6.3.

6.3.1 Lipschitz coefficients - proof of Theorem 6.3

Proof of Theorem 6.3. By Itô's formula, we have

$$d\|X_t-Y_t\|^2 = 2\langle X_t-Y_t, a(X_t)-b(Y_t)\rangle dt+\sqrt{8}(X_t-Y_t)(\sqrt{\sigma}(X_t)-\sqrt{\tau}(Y_t))dB_t+\|\sqrt{\sigma}(X_t)-\sqrt{\tau}(Y_t)\|^2 dt.$$

So,

$$\frac{d}{dt}\mathbb{E}\left[\|X_t - Y_t\|^2\right] \leq \mathbb{E}\left[\|X_t - Y_t\|^2\right] + \mathbb{E}\left[\|a(X_t) - b(Y_t)\|^2\right] + \mathbb{E}\left[\|\sqrt{\sigma}(X_t) - \sqrt{\tau}(Y_t)\|^2\right].$$

We have

$$\begin{aligned}\mathbb{E}\left[\|a(X_t) - b(Y_t)\|^2\right] &\leq 2\mathbb{E}\left[\|a(X_t) - a(Y_t)\|^2\right] + 2\mathbb{E}\left[\|a(Y_t) - b(Y_t)\|^2\right] \\ &\leq 2L^2\mathbb{E}\left[\|X_t - Y_t\|^2\right] + 2\|a - b\|^2_{L^2(\nu_s)},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\|\sqrt{\sigma}(X_t)-\sqrt{\tau}(Y_t)\|^2\right] &\le 2\mathbb{E}\left[\|\sqrt{\sigma}(X_t)-\sqrt{\sigma}(Y_t)\|^2\right]+2\mathbb{E}\left[\|\sqrt{\sigma}(Y_t)-\sqrt{\tau}(Y_t)\|^2\right]\\ &\le 2L^2\mathbb{E}\left[\|X_t-Y_t\|^2\right]+2\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu_s)}.\end{aligned}$$

Combine the above displays to obtain,

$$\frac{d}{dt}\mathbb{E}\left[\|X_t-Y_t\|^2\right]\le(1+4L^2)\mathbb{E}\left[\|X_t-Y_t\|^2\right]+2\|a-b\|^2_{L^2(\nu_s)}+2\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu_s)}.$$

We choose $\mu_0=\nu_0=\nu$ so that $\nu_s=\nu$ for all $s\ge 0$, and denote $r=2\|a-b\|^2_{L^2(\nu)}+2\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu)}$. To bound $\mathcal{W}_2^2(\nu,\mu)$, we consider the differential equation

$$f'(t)=(1+4L^2)f(t)+r,\ \text{with initial condition } f(0)=0.$$

Its unique solution is given by $f(t)=r\frac{e^{(4L^2+1)t}-1}{4L^2+1}$. Thus, by Gronwall's inequality

$$\mathcal{W}_2^2(\nu,\mu_t)=\mathcal{W}_2^2(\nu_t,\mu_t)\le\mathbb{E}\left[\|X_t-Y_t\|^2\right]\le r\frac{e^{(4L^2+1)t}-1}{4L^2+1}.$$

By Assumption (H3) we also know that

$$\mathcal{W}_2(\mu_t,\mu)\le C_H e^{-\kappa\cdot t}\mathcal{W}_2(\nu,\mu).$$

Thus,

$$\mathcal{W}_2(\nu,\mu)\le\mathcal{W}_2(\nu,\mu_t)+\mathcal{W}_2(\mu_t,\mu)\le\sqrt{r\frac{e^{(4L^2+1)t}-1}{4L^2+1}}+C_He^{-\kappa\cdot t}\mathcal{W}_2(\nu,\mu),$$

or equivalently when t is large enough

$$\mathcal{W}_2(\nu,\mu)\le\sqrt{\frac{r}{4L^2+1}}\frac{\sqrt{e^{(4L^2+1)t}-1}}{1-e^{-\kappa\cdot t}C_H}\le\sqrt{\frac{r}{4L^2+1}}\frac{e^{(2L^2+1)t}}{1-e^{-\kappa\cdot t}C_H}.$$

We now take $t=\frac{\ln\left(1+\frac{2\kappa}{4L^2+1}\right)+\ln(C_H)}{\kappa}$ to get

$$\begin{aligned}W_2(\nu,\mu)&\le\sqrt{\frac{r}{4L^2+1}}\left(\frac{4L^2+1}{2\kappa}+1\right)\left(1+\frac{2\kappa}{4L^2+1}\right)^{\frac{4L^2+1}{2\kappa}}C_H^{\frac{4L^2+1}{2\kappa}}\\ &\le 10C_H^{\frac{4L^2+1}{2\kappa}}\sqrt{r}\left(\frac{L}{\kappa}+1\right).\end{aligned}$$

To finish the proof it is enough to observe that $r\le 2\beta^2$. □

6.3.2 Proof of Theorem 6.1

To prove Theorem 6.1, we will first show that, under suitable assumptions, for a given $t > 0$, the measure μ_t cannot be too different than ν_t. Following [87] we define the logarithmic transport distance, which serves as a natural measure of distance between μ_t and ν_t:

$$\mathcal{D}_\delta(\mu, \nu) := \inf_\pi \int \ln\left(1 + \frac{|x-y|^2}{\delta^2}\right) d\pi,$$

where $\delta > 0$ and the infimum is taken over all couplings of μ and ν, i.e. $\mathcal{D}_\delta$ is a transport cost (but not a distance, and the cost is concave, not convex).

We have the following connection between $\mathcal{D}_\delta$ and $\widetilde{\mathcal{W}}^2_{2,R}$, which is essentially the same as [215, Lemma 5]. The proof of this lemma may be found in Section 6.5.

Lemma 6.7. *For any $R, \delta, \varepsilon > 0$, we have*

$$\widetilde{\mathcal{W}}^2_{2,R}(\mu, \nu) \leq \delta^2 \exp\left(\frac{\mathcal{D}_\delta(\mu,\nu)}{\varepsilon}\right) + R\varepsilon + R\frac{\mathcal{D}_\delta(\mu,\nu)}{\ln\left(1+\frac{R^2}{\delta^2}\right)}.$$

Remark 6.8. We can define $\widetilde{\mathcal{W}}_{1,R}$ in the same way, and a similar proof would also show that

$$\widetilde{\mathcal{W}}_{1,R}(\mu, \nu) \leq \delta \exp\left(\frac{\mathcal{D}_\delta(\mu,\nu)}{\varepsilon}\right) + R\varepsilon + R\frac{\mathcal{D}_\delta(\mu,\nu)}{\ln\left(1+\frac{R}{\delta^2}\right)},$$

which motivates Remark 6.2.

Observe that if $\delta < R$, then by choosing $\varepsilon = \frac{\mathcal{D}_\delta(\mu,\nu)}{\ln\left(1+\frac{R}{\delta}\right)}$ in the above lemma, we obtain

$$\widetilde{\mathcal{W}}^2_{2,R}(\mu, \nu) \leq 2R\left(\delta + \frac{\mathcal{D}_\delta(\mu,\nu)}{\ln\left(1+\frac{R}{\delta}\right)}\right). \tag{6.8}$$

Moreover, if both μ and ν have tame tails then it can be shown that for R large enough,

$$\mathcal{W}^2_2(\mu, \nu) \simeq \widetilde{\mathcal{W}}^2_{2,R}(\nu, \mu).$$

This is made rigorous in Lemma 6.18, in Section 6.5. For the logarithmic transport distance, we will prove:

Lemma 6.9. *Suppose that (H1) and (H2) hold and that $X_0 = Y_0$ almost surely with $Y_0 \sim \nu$. Then, for any $t, \delta > 0$,*

$$\mathcal{D}_\delta(\mu_t, \nu_t) \leq 2t\left(10\,\|d\nu/d\mu\|_{L^p(\mu)}\,\|g\|^2_{L^{2q}(\mu)} + \frac{1}{\delta}\|a-b\|_{L^1(\nu)} + \frac{2}{\delta^2}\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu)}\right).$$

where q is such that $\frac{1}{q} + \frac{1}{p} = 1$.

Proof of Theorem 6.1. To ease the notation we will denote

$$\alpha = 20\,\|d\nu/d\mu\|_{L^p(\mu)}\,\|g\|^2_{L^{2q}(\mu)}, \beta = 2\|a-b\|_{L^1(\nu)} + 2\|\sqrt{\sigma}-\sqrt{\tau}\|_{L^2(\nu)}.$$

We choose $\delta = \beta$ in Lemma 6.9 and obtain:

$$\mathcal{D}_\delta(\mu_t, \nu_t) \leq (\alpha+1)\,t.$$

Now, combine the above estimate with (6.8) to get

$$\widetilde{\mathcal{W}}^2_{2,R}(\mu_t,\nu_t) \leq 2R\left(\beta + \frac{(\alpha+1)}{\ln\left(1+\frac{R}{\beta}\right)}t\right) \leq 2R\frac{(\alpha+1)}{\ln\left(1+\frac{R}{\beta}\right)}(t+R). \tag{6.9}$$

To see the second inequality note that $\beta \leq R\ln(1+\frac{R}{\beta})^{-1}$. With Assumption (H3), we have

$$\widetilde{\mathcal{W}}_{2,R}(\mu_t,\mu) \leq \mathcal{W}_2(\mu_t,\mu) \leq C_H e^{-t\kappa}\mathcal{W}_2(\nu,\mu) \leq C_H e^{-t\kappa}\left(\sqrt{\int |x|^2 d\mu + \int |x|^2 d\nu}\right).$$

Observe as well that since ν is an invariant measure,

$$\widetilde{\mathcal{W}}_{2,R}(\nu_t,\nu) = 0.$$

We thus get,

$$\widetilde{\mathcal{W}}_{2,R}(\nu_t,\nu) + \widetilde{\mathcal{W}}_{2,R}(\mu_t,\mu) \leq C_H e^{-t\kappa}\sqrt{m_2^2(\mu)+m_2^2(\nu)}.$$

Take

$$t_0 := \frac{1}{\kappa}\ln\left(\sqrt{m_2^2(\mu)+m_2^2(\nu)\ln\left(1+\frac{R}{\beta}\right)}\right),$$

for which,

$$\widetilde{\mathcal{W}}_{2,R}(\nu_{t_0},\nu) + \widetilde{\mathcal{W}}_{2,R}(\mu_{t_0},\mu) \leq \frac{C_H}{\sqrt{\ln\left(1+\frac{R}{\beta}\right)}},$$

and, by using (6.9),

$$\widetilde{\mathcal{W}}_{2,R}(\mu_{t_0},\nu_{t_0}) \leq \sqrt{\frac{2R(\alpha+1)}{\ln\left(1+\frac{R}{\beta}\right)}(t_0+R)}.$$

To conclude the proof, we use the triangle inequality,

$$\begin{aligned}\widetilde{\mathcal{W}}_{2,R}(\mu,\nu) &\leq \widetilde{\mathcal{W}}_{2,R}(\nu_{t_0},\nu) + \widetilde{\mathcal{W}}_{2,R}(\mu_{t_0},\mu) + \widetilde{\mathcal{W}}_2(\mu_{t_0},\nu_{t_0}) \\ &\leq \frac{C_H}{\sqrt{\ln\left(1+\frac{R}{\beta}\right)}} + \sqrt{\frac{2R(\alpha+1)}{\ln\left(1+\frac{R}{\beta}\right)}(t_0+R)} \\ &\leq \frac{1}{\sqrt{\ln\left(1+\frac{R}{\beta}\right)}}\left(C_H + \sqrt{\frac{2R(\alpha+1)}{\kappa}\ln\left((m_2^2(\mu)+m_2^2(\nu))\ln\left(1+\frac{R}{\beta}\right)\right)+\kappa R}\right).\end{aligned}$$

□

Proof of Lemma 6.9

In this section our goal is to bound the logarithmic distance between X_t and Y_t and thus prove Lemma 6.9. Towards this, we let $Z_t = X_t - Y_t$. A straightforward application of Itô's formula gives the following result, whose proof may be found in [167, Section 4.1].

Lemma 6.10. *Suppose that $(X_0, Y_0) \sim \pi$ are coupled in the optimal way for way D_δ. That is, $\mathbb{E}_\pi\left[\ln\left(1+\frac{\|X_0-Y_0\|}{\delta^2}\right)\right] = \mathcal{D}_\delta(\mu_0,\nu_0)$. Then,*

$$\begin{aligned}\mathcal{D}_\delta(\mu_t,\nu_t) \leq \mathcal{D}_\delta(\mu_0,\nu_0) &+ 2\int_0^t \mathbb{E}\left[\frac{\langle Z_s, a(X_s)-b(Y_s)\rangle}{|Z_s|^2+\delta^2}\right]ds \\ &+ 2\int_0^t \mathbb{E}\left[\frac{||\sqrt{\sigma}(X_s)-\sqrt{\tau}(Y_s)||^2}{|Z_s|^2+\delta^2}\right]ds.\end{aligned}$$

With the above inequality we may then prove.

Lemma 6.11. *Let $t \geq 0$. Then,*

$$\mathcal{D}_\delta(\mu_t,\nu_t) \leq \mathcal{D}_\delta(\mu_0,\nu_0) + 2\int_0^t \left(5\left(\|g\|^2_{L^2(\mu_s)} + \|g\|^2_{L^2(\nu_s)}\right) + \frac{1}{\delta}\|a-b\|_{L^1(\nu_s)} + \frac{2}{\delta^2}\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu_s)}\right)ds.$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\frac{\langle Z_s, a(X_s)-b(Y_s)\rangle}{|Z_s|^2+\delta^2}\right] &\leq \mathbb{E}\left[\frac{|a(X_s)-b(Y_s)|}{\sqrt{|Z_s|^2+\delta^2}}\right] \\ &\leq \mathbb{E}\left[\frac{|a(X_s)-a(Y_s)|}{\sqrt{|Z_s|^2+\delta^2}}\right] + \mathbb{E}\left[\frac{|a(Y_s)-b(Y_s)|}{\sqrt{|Z_s|^2+\delta^2}}\right].\end{aligned}$$

Using Assumption (H1), we get

$$\mathbb{E}\left[\frac{|a(X_s)-a(Y_s)|}{\sqrt{|Z_s|^2+\delta^2}}\right] \leq \mathbb{E}\left[g(X_s)+g(Y_s)\right] = \|g\|_{L^1(\mu_s)} + \|g\|_{L^1(\nu_s)}.$$

We also have,

$$\mathbb{E}\left[\frac{|a(Y_s)-b(Y_s)|}{\sqrt{|Z_s|^2+\delta^2}}\right] \leq \frac{1}{\delta}\|a-b\|_{L^1(\nu_s)}.$$

So,

$$\mathbb{E}\left[\frac{\langle Z_s, a(X_s)-b(Y_s)\rangle}{|Z_s|^2+\delta^2}\right] \leq \|g\|_{L^1(\mu_s)} + \|g\|_{L^1(\nu_s)} + \frac{1}{\delta}\|a-b\|_{L^1(\nu_s)}.$$

Similar calculations yield,

$$\mathbb{E}\left[\frac{||\sqrt{\sigma}(X_s)-\sqrt{\tau}(Y_s)||^2}{|Z_s|^2+\delta^2}\right] \leq 4\|g\|^2_{L^2(\mu_s)} + 4\|g\|^2_{L^2(\nu_s)} + \frac{2}{\delta^2}\left\|\sqrt{\sigma}-\sqrt{\tau}\right\|^2_{L^2(\nu_s)}.$$

As it is fine to assume $\|g\|^2_{L^2(\rho)} \geq \|g\|_{L^1(\rho)} \geq 1$ for any probability measure ρ we consider, since we assumed for convenience that $g \geq 1$, we now plug the above displays into Lemma 6.10. □

Lemma 6.9 is now a consequence of the previous lemma.

Proof of Lemma 6.9. We start from Lemma 6.11. Since $\mu_0 = \nu_0 = \nu$, and ν is the invariant measure of the evolution equation for (Y_t), we have

$$\mathcal{D}_\delta(\mu_t,\nu_t) \leq 2\int_0^t \left(5\left(\|g\|^2_{L^2(\mu_s)} + \|g\|^2_{L^2(\nu)}\right) + \frac{1}{\delta}\|a-b\|_{L^1(\nu)} + \frac{2}{\delta^2}\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu)}\right) ds. \tag{6.10}$$

Let $q = \left(1-\frac{1}{p}\right)^{-1}$. By Hölder's inequality,

$$\|g\|^2_{L^2(\nu)} \leq \|g\|^2_{L^{2q}(\mu)} \left\|\frac{d\nu}{d\mu}\right\|_{L^p(\mu)}.$$

Also,

$$\|g\|^2_{L^2(\mu_s)} \leq \|g\|^2_{L^{2q}(\mu)} \left\|\frac{d\mu_s}{d\mu}\right\|_{L^p(\mu)} \leq \|g\|^2_{L^{2q}(\mu)} \left\|\frac{d\nu}{d\mu}\right\|_{L^p(\mu)},$$

where in the second inequality we have used that $\left\|\frac{d\mu_s}{d\mu}\right\|_{L^p(\mu)}$ is monotonic decreasing in s. We

plug the above displays into (6.10), to obtain

$$\mathcal{D}_\delta(\mu_t,\nu_t) \le 2t\left(10\,\|d\nu/d\mu\|_{L^p(\mu)}\,\|g\|^2_{L^{2q}(\mu)} + \frac{1}{\delta}\|a-b\|_{L^1(\nu)} + \frac{2}{\delta^2}\|\sqrt{\sigma}-\sqrt{\tau}\|^2_{L^2(\nu)}\right).$$

which concludes the proof. □

6.4 Proofs of the applications to Stein kernels

In this section we fix a measure μ on $\mathbb{R}^d$, with Stein kernel τ_μ, constructed as in (6.7). For now, we make the assumption that τ_μ is positive definite and uniformly bounded from below. In the sequel, when we say that a measure is isotropic we mean that it is centered and that its covariance matrix is the identity. To apply our result, we must first construct an Itô diffusion process with μ as its unique invariant measure. Define the process X_t to satisfy the following SDE:

$$dX_t = -X_t dt + \sqrt{2\tau_\mu(X_t)}dB_t. \tag{6.11}$$

Lemma 6.12. *μ is the unique invariant measure of the process X_t.*

Proof. Let L be the infinitesimal generator of of (X_t). That is, for a twice differentiable test function,

$$Lf(x) = \langle -x, \nabla f(x)\rangle + \langle \tau_\mu(x), \nabla^2 f(x)\rangle_{HS}.$$

From the definition of the Stein kernel (6.5), we have

$$\mathbb{E}_\mu\left[Lf(x)\right] = 0,$$

for any such test function. We conclude that μ is the invariant measure of the process. Uniqueness follows, since τ_μ is uniformly bounded from below ([38]). □

Before proving Theorem 6.6 we collect several facts concerning this process.

6.4.1 Lusin-Lipschitz Property for moment maps

We would now like to claim that the kernel τ_μ exhibits Lipschitz-like properties as in Assumption (H1). For this to hold we restrict our attention to a more regular class of measures. Henceforth, we assume that $\mu = e^{-V(x)}dx$ is an isotropic log-concave measure whose support equals $\mathbb{R}^d$ and that there exists a constant $\alpha > 0$, such that

$$\alpha \mathrm{I_d} \le \nabla^2 V \le \frac{1}{\alpha}\mathrm{I_d}. \tag{6.12}$$

In some sense, this assumption can be viewed as restricting ourselves to measures that are not too far from a Gaussian distribution. Under this assumption the main result of this section is that Stein kernels satisfy the Lusin-Lipschitz property that we need in order to apply Theorem 6.1. That is:

Lemma 6.13. *Let μ be an isotropic log-concave measure on $\mathbb{R}^d$ satisfying (6.12) and let τ_μ be its Stein kernel constructed from the moment map. Then, there exists a function $g : \mathbb{R}^d \to \mathbb{R}$ such that for almost every $x, y \in \mathbb{R}^d$:*

$$\left\| \sqrt{\tau_\mu(x)} - \sqrt{\tau_\mu(y)} \right\| \leq (g(x) + g(y))\|x - y\|,$$

and,

$$\|g\|_{L^2(\mu)} \leq C d^{3/2} \alpha^{-1},$$

where $C > 0$ is a universal constant. Moreover, there exists a constant c such that if μ is radially symmetric and has full support then we also have for $d > c$

$$\|g\|_{L^4(\mu)} < C d^{7/4} \alpha^{-8}.$$

In the sequel we will use the following notation, for $v \in \mathbb{R}^d$, $\partial_v \varphi$ is the directional derivative of φ along v. Repeated derivations will be denoted as $\partial^2_{uv}\varphi, \partial^3_{uvw}\varphi$, etc. If e_i, for $i = 1, \dots, d$, is a standard unit vector, we will abbreviate $\partial_i \varphi = \partial_{e_i}\varphi$. Finally, $\nabla\varphi^{-1}$ is the inverse map of $\nabla\varphi$.

Recall that $\tau_\mu = \nabla^2\varphi(\nabla\varphi^{-1})$, where $\nabla\varphi$ pushes the measure $e^{-\varphi}dx$ to μ. Thus, keeping in mind Proposition 6.4, our first objective is to show $\partial^3_{ijk}\varphi \in W^{1,2}(\mu)$, for every $i, j, k = 1, \dots, d$. This will be a consequence of the following result:

Proposition 6.14 (Third-order regularity bounds on moment maps)**.** *Assume that μ is isotropic and that $\nabla^2 V \geq \alpha \mathrm{I}_d$. Then, for $i, j = 1, ..., d$ and $j \neq i$,*

1. $\int |\nabla \partial^2_{ii}\varphi|^2 e^{-\varphi} dx \leq C\alpha^{-1}$.
2. $\int |\nabla \partial^2_{ij}\varphi|^2 e^{-\varphi} dx \leq C(d + \alpha^{-1})$.

Here C is a dimension-free constant, independent of μ.

Note that under the isotropy condition, necessarily $\alpha \leq 1$. These bounds build up on the following estimates :

Proposition 6.15. *Assume that $\mu = e^{-V(x)}dx$ is log-concave and isotropic and φ is its moment map.*

1. *For any direction $e \in \mathbb{S}^{d-1}$ we have*

$$\int (\partial^2_{ee}\varphi)^p e^{-\varphi} dx \leq 8^p p^{2p}$$

2. $\int \langle (\nabla^2\varphi)^{-1}\nabla\partial^2_{ee}\varphi, \nabla\partial^2_{ee}\varphi\rangle e^{-\varphi}dx \leq 32\sqrt{\int \langle x,e\rangle^4 d\mu} \leq C$, *with* C *a dimension-free constant, that does not depend on* μ.

3. *If* μ *has a convex support and* $\nabla^2 V \geq \alpha \mathrm{I}_d$ *with* $\alpha > 0$, *then* $\nabla^2\varphi \leq \alpha^{-1}\mathrm{I}_d$.

4. *If* μ *has full support and* $\nabla^2 V \leq \beta \mathrm{I}_d$ *with* $\beta > 0$ *then* $\nabla^2\varphi \geq \beta^{-1}\mathrm{I}_d$.

The first part was proved in [151] (see [112, Proposition 3.2] for the precise statement). The second part is an immediate consequence of [155, eq (55)]. The third part was proved in [151]. The last part is part of the proof of [155, Theorem 3.4]

Proof of Proposition 6.14. The first part is an immediate consequence of items 2 and 3 of Proposition 6.15. For the second part, with several successive integrations by parts, we have,

$$\begin{aligned}
\int (\partial^3_{ijk}\varphi)^2 e^{-\varphi}dx &= -\int (\partial^4_{iijk}\varphi)(\partial^2_{jk}\varphi)e^{-\varphi}dx + \int (\partial^3_{ijk}\varphi)(\partial^2_{jk}\varphi)(\partial_i\varphi)e^{-\varphi}dx \\
&= \int (\partial^3_{iik}\varphi)(\partial^3_{jjk}\varphi)e^{-\varphi}dx - \int (\partial^3_{iik}\varphi)(\partial^2_{jk}\varphi)(\partial_j\varphi)e^{-\varphi}dx \\
&\quad + \int (\partial^3_{ijk}\varphi)(\partial^2_{jk}\varphi)(\partial_i\varphi)e^{-\varphi}dx \\
&\leq \frac{1}{2}\int (\partial^3_{ijk}\varphi)^2 e^{-\varphi}dx + \int (\partial^3_{iik}\varphi)^2 e^{-\varphi}dx + \frac{1}{2}\int (\partial^3_{jjk}\varphi)^2 e^{-\varphi}dx \\
&\quad + \frac{1}{2}\int (\partial^2_{jk}\varphi)^4 e^{-\varphi}dx + \frac{1}{4}\int ((\partial_i\varphi)^4 + (\partial_j\varphi)^4 e^{-\varphi}dx. \qquad (6.13)
\end{aligned}$$

Moreover, since $\nabla^2\varphi$ is positive-definite, we have $|\partial^2_{jk}\varphi| \leq (\partial^2_{jj}\varphi + \partial^2_{kk}\varphi)/2$, and therefore

$$\begin{aligned}
\sum_k \int (\partial^2_{jk}\varphi)^4 e^{-\varphi}dx &\leq \frac{1}{8}\sum_k \int (\partial^2_{jj}\varphi + \partial^2_{kk}\varphi)^4 e^{-\varphi}dx \\
&\leq Cd. \qquad (6.14)
\end{aligned}$$

Summing (6.13) over k implies the result, via the moment bounds for isotropic log-concave distributions and the 2nd order bounds on φ. □

We will also need the following result about radially symmetric functions.

Proposition 6.16. *Suppose that* (6.12) *holds and that* $\mu = e^{-V(x)}dx$ *is radially symmetric and has full support. Then, there exists an absolute constant* $c > 0$, *such that, for any* $i, j, k = 1, \ldots, d$, *if* $d > c$:

$$\int |\partial^3_{ijk}\varphi|^4 e^{-\varphi}dx \leq C\alpha^{-30}d^4.$$

for some absolute constant $C > 0$.

Proof. Note that φ satisfies the Monge-Ampère equation

$$e^{-\varphi} = e^{-V(\nabla\varphi)} \det\left(\nabla^2\varphi\right),$$

and that it can be verified that if V is a radial function then so is φ. Let $i = 1, \ldots, d$, by taking the logarithm and differentiating the above equation we get:

$$\partial_i\varphi = \langle \nabla V(\nabla\varphi), \nabla\partial_i\varphi\rangle - \mathrm{Tr}\left(\nabla^2\partial_i\varphi \left(\nabla^2\varphi\right)^{-1}\right).$$

By Proposition 6.15, $\alpha \mathrm{I_d} \leq \nabla^2\varphi \leq \frac{1}{\alpha}\mathrm{I_d}$. Hence,

$$\mathrm{Tr}\left(\left(\nabla^2\varphi\right)^{-1}\nabla^2\partial_i\varphi\right) \leq |\partial_i\varphi| + |\langle \nabla V(\nabla\varphi), \nabla\partial_i\varphi\rangle| \leq |\partial_i\varphi| + \alpha^{-1}\sqrt{d}\|\nabla V(\nabla\varphi)\|, \quad (6.15)$$

where the second inequality used Cauchy-Schwartz along with $\|\nabla\partial_i\varphi\| \leq \sqrt{d}\|\nabla^2\varphi\|_{op}$. The proof will now be conducted in three steps:

1. We will bound $\mathrm{Tr}\left(\left(\nabla^2\varphi\right)^{-1}\nabla^2\partial_i\varphi\right)$ in terms of $\mathrm{Tr}\left(\nabla^2\partial_i\varphi\right) = \sum_{j=1}^{d}\partial^3_{jji}\varphi$.

2. Using (6.15), we'll show that $\int\left(\sum_{j=1}^{d}\partial^3_{jji}\varphi\right)^4 e^{-\varphi}dx$ cannot be large.

3. Finally, we will use the previous step to bound $\int\left(\partial^3_{kji}\varphi\right)^4 e^{-\varphi}dx$.

Step 1: We now wish to understand $\mathrm{Tr}\left(\left(\nabla^2\varphi\right)^{-1}\nabla^2\partial_i\varphi\right)$. Write $\varphi(x) = f(\|x\|^2)$, so that,

$$\nabla^2\varphi(x) = 2f'(\|x\|^2)\mathrm{I_d} + 4f''(\|x\|^2)xx^T. \quad (6.16)$$

The bounds on $\nabla^2\varphi$ imply the following inequalities, which we shall freely use below:

$$\alpha \leq 2f'(\|x\|), 2f'(\|x\|) + 4f''(\|x\|^2)\|x\|^2 \leq \alpha^{-1},$$

and

$$\left|4f''(\|x\|^2)\|x\|^2\right| \leq \alpha^{-1}.$$

By the Sherman-Morrison formula,

$$\left(\nabla^2\varphi\right)^{-1}(x) = \frac{1}{2f'(\|x\|^2)}\left(\mathrm{I_d} - \frac{4f''(\|x\|^2)}{2f'(\|x\|^2) + 4f''(\|x\|^2)\|x\|^2}xx^T\right).$$

So,

$$\mathrm{Tr}\left(\left(\nabla^2\varphi\right)^{-1}\nabla^2\partial_i\varphi\right) = \frac{1}{2f'(\|x\|^2)}\left(\sum_{j=1}^{d}\partial^3_{jji}\varphi - \frac{4f''(\|x\|^2)}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2}\partial^3_{xxe_i}\varphi\right). \tag{6.17}$$

A calculation shows

$$\partial^3_{jji}\varphi(x) = 4x_i(2x_j^2 f'''(\|x\|^2) + f''(\|x\|^2) + 2\delta_{ij}f''(\|x\|^2)),$$

and

$$\sum_{j=1}^{d}\partial^3_{jji}\varphi(x) = 4x_i(2\|x\|^2 f'''(\|x\|^2) + (d+2)f''(\|x\|^2)).$$

Also,

$$\partial^3_{xxe_i}\varphi = 4x_i\left(2\|x\|^4 f'''(\|x\|^2) + 3\|x\|^2 f''(\|x\|^2)\right).$$

Thus, if $D(x) := 1 - \frac{4f''(\|x\|^2)\|x\|^2}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2} = \frac{2f'(\|x\|^2)}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2}$,

$$\begin{aligned}
&\sum_{j=1}^{d}\partial^3_{jji}\varphi - \frac{4f''(\|x\|^2)}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2}\partial^3_{xxe_i}\varphi \\
&\quad= \left(1 - \frac{4f''(\|x\|^2)\|x\|^2}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2}\right)8x_i\|x\|^2 f'''(\|x\|^2) \\
&\qquad+ \left(d+2-3\frac{4f''(\|x\|^2)\|x\|^2}{2f'(\|x\|^2)+4f''(\|x\|^2)\|x\|^2}\right)4x_i f''(\|x\|^2) \\
&\quad= D(x)8x_i\|x\|^2 f'''(\|x\|^2) + (d-1+3D(x))\,4x_i f''(\|x\|^2) \\
&\quad= D(x)\sum_{j=1}^{d}\partial^3_{jji}\varphi + (d-1+3D(x)-D(x)(d+2))\,4x_i f''(\|x\|^2) \\
&\quad\geq D(x)\sum_{j=1}^{d}\partial^3_{jji}\varphi - Cd\alpha^{-3}\frac{|x_i|}{\|x\|^2} \\
&\quad= D(x)\mathrm{Tr}\left(\nabla^2\partial_i\varphi\right) - Cd\alpha^{-3}\frac{|x_i|}{\|x\|^2}.
\end{aligned}$$

In the inequality, we have used (6.16) along with the bounds $4|f''(\|x\|^2)| \leq \frac{\alpha^{-1}}{\|x\|^2}$, and $\alpha^2 \leq D(x) \leq \alpha^{-2}$. (6.17) then implies:

$$\mathrm{Tr}\left(\left(\nabla^2\varphi\right)^{-1}\nabla^2\partial_i\varphi\right) \geq \frac{D(x)}{2f'(\|x\|^2)}\mathrm{Tr}\left(\nabla^2\partial_i\varphi\right) - \frac{Cd\alpha^{-3}}{2f'(\|x\|^2)}\frac{1}{\|x\|}.$$

Step 2: We now integrate with respect to the moment measure, so the estimate from the previous step, along with the bounds $\alpha^2 \leq D(x)$, $\alpha \leq 2f''(\|x\|) \leq \alpha^{-1}$, and (6.15) give:

$$\int \mathrm{Tr}\left(\nabla^2 \partial_i \varphi\right)^4 e^{-\varphi} dx \leq C\alpha^{-12}\Big(\int |\partial_i \varphi|^4 e^{-\varphi} dx + \alpha^{-4} d^2 \int \|\nabla V(\nabla \varphi)\|^4 e^{-\varphi} dx + d^4 \alpha^{-16} \int \frac{1}{\|x\|^4} e^{-\varphi} dx \Big). \quad (6.18)$$

Let us look at each term on the right hand side. By a change of variable

$$\int |\partial_i \varphi|^4 e^{-\varphi} dx = \int \|x_i\|^4 d\mu \leq C,$$

since higher moments of coordinates of isotropic log-concave measures are controlled.

For the second term, since $\nabla \varphi$ is a transport map, we get that

$$\int \|\nabla V(\nabla \varphi)\|^4 e^{-\varphi} dx = \int \|\nabla V\|^4 d\mu.$$

Recalling that $\alpha \mathrm{I_d} \leq \nabla^2 V \leq \frac{1}{\alpha} \mathrm{I_d}$, we apply the Poincaré inequality for μ, and since $|\nabla |\nabla V|^2| = 2|\nabla^2 V \nabla V| \leq 2\alpha^{-1} |\nabla V|$,

$$\int \|\nabla V(\nabla \varphi)\|^4 e^{-\varphi} dx \leq \left(\int |\nabla V|^2 d\mu \right)^2 + 4\alpha^{-3} \left(\int |\nabla V|^2 d\mu \right).$$

By integration by parts, $\int |\nabla V|^2 d\mu = \int \Delta V d\mu \leq d\alpha^{-1}$, and hence

$$\int \|\nabla V(\nabla \varphi)\|^4 e^{-\varphi} dx \leq 5 d^2 \alpha^{-4}.$$

For the third integral, we may use the fact that when $d \geq c$, a reverse Hölder inequality holds for negative moments, and may be applied to radially symmetric log-concave measures (see [204, Theorem 1.4]). According to the inequality,

$$\int \frac{1}{\|x\|^4} e^{-\varphi} dx \leq C \left(\int \|x\|^2 e^{-\varphi} dx \right)^{-2} \leq C\alpha^{-2}.$$

We now plug the previous three displays into (6.18) to conclude,

$$\int \left(\sum_{j=1}^d \partial^3_{jji} \varphi \right)^4 e^{-\varphi} dx = \int \mathrm{Tr}\left(\nabla^2 \partial_i \varphi\right)^4 e^{-\varphi} dx \leq C\alpha^{-30} d^4. \quad (6.19)$$

Step 3: Now, let $i, j, k = 1, \dots, d$ be distinct. We have,

$$\begin{aligned}
\int \left(\partial^3_{ijk}\varphi\right)^4 e^{-\varphi} dx &= 8^4 \int \left(x_i x_j x_k f'''(\|x\|^2)\right)^4 e^{-\varphi} dx \\
&\leq \frac{8^4}{2^4}\left(\int x_i^4 \left(x_j^2 f'''(\|x\|^2)\right)^4 e^{-\varphi} dx + \int x_i^4 \left(x_k^2 f'''(\|x\|^2)\right)^4 e^{-\varphi} dx\right). \\
&\leq \frac{8^4}{2^3} \int x_i^4 \left(\|x\|^2 f'''(\|x\|^2)\right)^4 e^{-\varphi} dx \\
&\leq 8^4 \int x_i^4 \left(\|x\|^2 f'''(\|x\|^2) + (d+2) f''(\|x\|^2)\right)^4 e^{-\varphi} dx \\
&\qquad + 8^5 \int x_i^4 \left((d+2) f''(\|x\|^2)\right)^4 e^{-\varphi} dx \\
&= 8^4 \left(\int \left(\sum_{j=1}^d \partial^3_{jji}\varphi(x)\right)^4 e^{-\varphi} dx + \int \left((d+2) x_i f''(\|x\|^2)\right)^4 e^{-\varphi} dx\right) \\
&\leq \left(C\alpha^{-30} d^4 + (d+2)^4 \int \left(x_i f''(\|x\|^2)\right)^4 e^{-\varphi} dx\right),
\end{aligned}$$

where we have used (6.19) in the last inequality. For the remaining integral term, denote $x_{\sim i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$. Then

$$\begin{aligned}
\int \left(x_i f''(\|x\|^2)\right)^4 e^{-\varphi} dx &= \int \frac{1}{\|x_{\sim i}\|^4} \left(x_i \|x_{\sim i}\| \cdot f''(\|x\|^2)\right)^4 e^{-\varphi} dx \\
&\leq \int \frac{1}{\|x_{\sim i}\|^4} \sum_{j \neq i} x_i^4 x_j^4 f''(\|x\|^2)^4 e^{-\varphi} dx \\
&= \int \frac{1}{\|x_{\sim i}\|^4} \sum_{j \neq i} \left(\partial^2_{ij}\varphi\right)^4 e^{-\varphi} dx \\
&\leq d\alpha^{-4} \int \frac{1}{\|x_{\sim i}\|^4} e^{-\varphi} dx \leq Cd\alpha^{-6}.
\end{aligned}$$

In the last inequality we again used the reverse Hölder inequality for negative moments of radially symmetric log-concave measures. Plugging this estimate into the previous display finishes the proof, when $|\{i, j, k\}| = 3$. The other cases can be proven similarly. □

We are now in a position to prove Lemma 6.13.

Proof of Lemma 6.13. Since $\nabla\varphi^{-1}$ transports μ to $e^{-\varphi}dx$, from Proposition 6.14 we conclude that $\tau_\mu \in W^{1,2}(\mu)$ and that $\|D\tau_\mu\|_{L^2(\mu)} \leq Cd^{3/2}\alpha^{-1/2}$, where D stands for the total derivative operator. Here, we have used the identity $\|D^3\varphi\|_{L^2(e^{-\varphi}dx)} = \|D\tau_\mu\|_{L^2(\mu)}$, which follows from (6.7). Thus, by Proposition 6.4, there exists a function $\tilde{g}$ for which,

$$\|\tau_\mu(x) - \tau_\mu(y)\| \leq (\tilde{g}(x) + \tilde{g}(y))\|x - y\|,$$

and $\|\tilde{g}\|_{L^2(\mu)} \leq C\frac{d^{3/2}}{\alpha^{1/2}}$. Proposition 6.15 along with (6.12) shows

$$\sqrt{\alpha}\mathrm{I}_d \leq \sqrt{\tau_\mu}.$$

Hence,

$$\left\|\sqrt{\tau_\mu(x)} - \sqrt{\tau_\mu(y)}\right\| \leq \frac{1}{\sqrt{\alpha}}\|\tau_\mu(x) - \tau_\mu(y)\|.$$

Take now $g := \frac{1}{\sqrt{\alpha}}\tilde{g}$ to conclude the proof. If μ is radially symmetric, then Proposition 6.16 shows $\|D\tau_\mu\|_{L^4(\mu)} \leq Cd^{\frac{7}{4}}\alpha^{-\frac{15}{2}}$ and the proof continues in a similar way. □

6.4.2 Exponential convergence to equilibrium

We now show that the process (6.11) satisfies the exponential convergence to equilibrium property we require, as long as (6.12) is satisfied.

Lemma 6.17. *Assume that $\mu = e^{-V}dx$ with $\alpha^{-1}\mathrm{I}_d \geq \nabla^2 V \geq \alpha\mathrm{I}_d$. Then the diffusion process (6.11) satisfies Assumption H3 with $\kappa = 1/2$ and $C_H = \alpha^{-2}$.*

Proof. As demonstrated in [154], the diffusion process $\nabla\varphi^{-1}(X_t)$, where (X_t) solves (6.11), satisfies the Bakry-Emery curvature dimension condition $\mathrm{CD}(1/2, \infty)$ when viewed as the canonical diffusion process on the weighted manifold $(\mathbb{R}^d, (\nabla^2\varphi)^{-1}, e^{-\varphi})$. Therefore it is a contraction in Wasserstein distance, with respect to the Riemannian metric d_φ with tensor $(\nabla^2\varphi)^{-1}$ [241]. That is

$$W_{2,d_\varphi}(\mu_t \circ \nabla\varphi, e^{-\varphi}) \leq e^{-t/2}\mathcal{W}_{2,d_\varphi}(\mu_0 \circ \nabla\varphi, e^{-\varphi}).$$

From the bounds on $\nabla^2\varphi$ given by Proposition 6.15, we have

$$\alpha^{-1}|x-y|^2 \geq d_\varphi(x,y)^2 \geq \alpha|x-y|^2,$$

and the result follows, using again the two-sided Lipschitz bounds on $\nabla\varphi$. □

6.4.3 Stability for Stein kernels

Proof of Theorem 6.6. We consider the two processes

$$dX_t = -X_t + \sqrt{2\tau(X_t)}dB_t,$$
$$dY_t = -Y_t + \sqrt{2\sigma(Y_t)}dB_t.$$

By Lemma 6.12, μ and ν are the respective invariant measures of X_t and Y_t. By Lemma 6.17, Assumption H3 is satisfied with $\kappa = \frac{1}{2}, C_H = \alpha^{-2}$. By Lemma 6.13, Assumption H1 is satisfied with $\|g\|^2_{L^2(\mu)} \leq Cd^3\alpha^{-2}$.

Set $p = \infty, q = 1$ and $R > 0$. Plugging the above estimates to Theorem 6.1, we get

$$\widetilde{\mathcal{W}}^2_{2,R}(\nu,\mu) \leq C\alpha^{-6}d^3R\,\|d\nu/d\mu\|_\infty \frac{\ln\left(\ln\left(1+\frac{R}{\beta}\right)\right)+\ln(M)+R}{\ln\left(1+\frac{R}{\beta}\right)}.$$

ν and μ are log-concave and in-particular have sub-exponential tails. We apply Lemma 6.18, from the appendix, to obtain a constant $C' > 0$ such that

$$\mathcal{W}_2(\nu,\mu) \leq 2\widetilde{\mathcal{W}}_{2,C'M^2\ln(M)^2}(\nu,\mu),$$

which proves the first part of the theorem. For the second part, if μ is radially symmetric, then we take $p = q = 2$, and by Lemma 6.13, H1 is now satisfied with $\|g\|^2_{L^4(\mu)} \leq Cd^{\frac{7}{2}}\alpha^{-16}$ and the proof continues in the same way. The last part of the theorem is an immediate consequence of Theorem 6.3. □

6.5 Transport inequalities for the truncated Wasserstein distance

Proof of Lemma 6.7. We let π denote the optimal coupling for $\mathcal{D}_\delta$ and $(X,Y) \sim \pi$. Define the sets

$$D = \left\{(x,y) \in \mathbb{R}^{2d} : \|x-y\|^2 \leq R\right\},$$
$$\widetilde{D} = \left\{(x,y) \in D : \ln\left(1+\frac{\|x-y\|^2}{\delta^2}\right) \leq \frac{\mathcal{D}_\delta(\mu,\nu)}{\varepsilon}\right\}.$$

We now write

$$\mathbb{E}\left[\min\left(\|X-Y\|^2, R\right)\right] = \mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{\widetilde{D}}\right] + \mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{D\setminus\widetilde{D}}\right] + R\cdot\mathbb{E}\left[\mathbb{1}_{\mathbb{R}^{2d}}\setminus D\right],$$

and bound each term separately. Observe that for any $\alpha \in \mathbb{R}$,

$$\ln\left(1+\frac{x^2}{\delta^2}\right) \leq \alpha \iff |x| \leq \delta\sqrt{e^\alpha - 1}.$$

Thus,

$$\mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{\widetilde{D}}\right] \leq \delta^2\exp\left(\frac{\mathcal{D}_\delta(\mu,\nu)}{\varepsilon}\right).$$

Next, by Markov's inequality

$$\mathbb{P}\left(\ln\left(1+\frac{\|X-Y\|^2}{\delta^2}\right)\geq\frac{\mathcal{D}_\delta(\mu,\nu)}{\varepsilon}\right)\leq\frac{\varepsilon}{\mathcal{D}_\delta(\mu,\nu)}\mathbb{E}\left[\ln\left(1+\frac{\|X-Y\|^2}{\delta^2}\right)\right]=\varepsilon.$$

So,

$$\mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{D\setminus\widetilde{D}}\right]\leq R\mathbb{E}\left[\mathbb{1}_{\mathbb{R}^{2d}\setminus\widetilde{D}}\right]\leq R\varepsilon.$$

Finally, a second application of Markov's inequality gives

$$\begin{aligned}R\mathbb{E}\left[\mathbb{1}_{\mathbb{R}^{2d}}\setminus D\right]&\leq R\cdot\mathbb{P}\left(\|X-Y\|^2\geq R\right)\\&=R\cdot\mathbb{P}\left(\ln\left(1+\frac{\|X-Y\|^2}{\delta^2}\right)\geq\ln\left(1+\frac{R^2}{\delta^2}\right)\right)\leq R\frac{\mathcal{D}_\delta(\mu,\nu)}{\ln\left(1+\frac{R^2}{\delta^2}\right)}.\end{aligned}$$

□

Lemma 6.18. *Let $X\sim\mu, Y\sim\nu$ be two centered random vectors in $\mathbb{R}^d$. Assume that both X and Y are sub-exponential with parameter M, in the sense that for every $k\geq 2$,*

$$\frac{\mathbb{E}\left[\|X\|^k\right]^{\frac{1}{k}}}{k},\frac{\mathbb{E}\left[\|Y\|^k\right]^{\frac{1}{k}}}{k}\leq M. \tag{6.20}$$

Then

$$\mathcal{W}_2^2(\mu,\nu)\leq 2\widetilde{\mathcal{W}}^2_{2,CM^2\ln(M)^2}(\mu,\nu),$$

for a universal constant $C>0$.

Proof. Fix $R>0$ and let π denote the optimal coupling for $\widetilde{\mathcal{W}}_{2,R}$ and $(X,Y)\sim\pi$. Then,

$$\begin{aligned}\mathbb{E}\left[\|X-Y\|^2\right]&=\mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{\|X-Y\|^2\leq R}\right]+\mathbb{E}\left[\|X-Y\|^2\mathbb{1}_{\|X-Y\|^2>R}\right]\\&\leq\widetilde{W}^2_{2,R}(\mu,\nu)+\sqrt{\mathbb{E}\left[\|X-Y\|^4\right]\mathbb{P}\left(\|X-Y\|^2>R\right)}.\end{aligned}$$

$X-Y$ also has a sub-exponential law with parameter $C'M$, where $C'>0$ is a constant. Thus, by (6.20), for some other constant C,

$$\sqrt{\mathbb{E}\left[\|X-Y\|^4\right]}\leq CM\mathbb{E}\left[\|X-Y\|^2\right],$$

and $\sqrt{\mathbb{P}\left(\|X-Y\|^2>R\right)}\leq e^{-\frac{\sqrt{R}}{CM}}$. Take $R=(CM\ln(2CM))^2$, to get.

$$\mathbb{E}\left[\|X-Y\|^2\right]\leq\widetilde{\mathcal{W}}^2_{2,R}(\mu,\nu)+\frac{1}{2}\mathbb{E}\left[\|X-Y\|^2\right],$$

which implies,

$$\mathcal{W}_2^2(\mu,\nu)\leq\mathbb{E}\left[\|X-Y\|^2\right]\leq 2\widetilde{\mathcal{W}}^2_{2,R}(\mu,\nu).$$

□

PART III

APPLICATIONS IN DATA SCIENCE

"Thou art god, I am god. All that groks is god."
- Michael Valentine Smith

7

Methods in Non-Convex Optimization - Gradient Flow Trapping

7.1 Introduction

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function (i.e., the map $x \mapsto \nabla f(x)$ is 1-Lipschitz, and f is possibly non-convex). We aim to find an ε-approximate stationary point, i.e., a point $x \in \mathbb{R}^d$ such that $\|\nabla f(x)\|_2 \leq \varepsilon$. It is an elementary exercise to verify that for smooth and bounded functions, gradient descent finds such a point in $O(1/\varepsilon^2)$ steps, see e.g., [192]. Moreover, it was recently shown in [66] that this result is *optimal*, in the sense that any procedure with only black-box access to f (e.g., to its value and gradient) must, *in the worst case*, make $\Omega(1/\varepsilon^2)$ queries before finding an ε-approximate stationary point. This situation is akin to the non-smooth convex case, where the same result (optimality of gradient descent at complexity $1/\varepsilon^2$) holds true for finding an ε-approximate optimal point (i.e., such that $f(x) - \min_{y \in \mathbb{R}^d} f(y) \leq \varepsilon$), [190, 192].

There is an important footnote to both of these results (convex and non-convex), namely that optimality only holds in *arbitrarily high dimension* (specifically the hard instance in both cases require $d = \Omega(1/\varepsilon^2)$). It is well-known that in the convex case this large dimension requirement is actually necessary, for the cutting plane type strategies (e.g., center of gravity) can find ε-approximate optimal points on compact domains in $O(d \log(1/\varepsilon))$ queries. It is natural to

ask: **Is there some analogue to cutting planes for non-convex optimization?**[1] In dimension 1 it is easy to see that one can indeed do a binary search to find an approximate stationary point of a smooth non-convex function on an interval. The first non-trivial case is thus dimension 2, which is the focus of this chapter (although we also obtain new results in high dimensions, and in particular our approach does achieve $O(\mathrm{poly}(d)\log(1/\varepsilon))$ *parallel depth*, see below for details).

This problem, of finding an approximate stationary point of a smooth function on a compact domain of $\mathbb{R}^2$, was studied in 1993 by Stephen A. Vavasis in [237]. From an algorithmic perspective, his main observation is that in finite dimensional spaces one can speed up gradient descent by using a *warm start*. Specifically, observe that gradient descent only needs $O(\Delta/\varepsilon^2)$ queries when starting from a Δ-approximate optimal point. Leveraging smoothness (see e.g., Lemma 7.5 below), observe that the best point on a $\sqrt{\Delta}$-net of the domain will be Δ-approximate optimal. Thus starting gradient descent from the best point on $\sqrt{\Delta}$-net one obtains the complexity $O_d\left(\frac{\Delta}{\varepsilon^2} + \frac{1}{\Delta^{d/2}}\right)$ in $\mathbb{R}^d$. Optimizing over Δ, one obtains a $O_d\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2d}{d+2}}\right)$ complexity. In particular for $d = 2$ this yields a $O(1/\varepsilon)$ query strategy. In addition to this algorithmic advance, Vavasis also proved a lower bound of $\Omega(1/\sqrt{\varepsilon})$ for deterministic algorithms. In this chapter we close the gap up to a logarithmic term. Our main contribution is a new strategy loosely inspired by cutting planes, which we call *gradient flow trapping* (GFT), with complexity $O\left(\sqrt{\frac{\log(1/\varepsilon)}{\varepsilon}}\right)$. We also extend Vavasis lower bound to randomized algorithms, by connecting the problem with unpredictable walks in probability theory [31].

Although we focus on $d = 2$ for the description and analysis of GFT in this chapter, one can in fact easily generalize to higher dimensions. Before stating our results there, we first make precise the notion of approximate stationary points, and we also introduce the *parallel query* model.

7.1.1 Approximate stationary point

We focus on the constraint set $[0,1]^d$, although this is not necessary and we make this choice mainly for ease of exposition. Let us fix a differentiable function $f : [0,1]^d \to \mathbb{R}$ such that $\forall x, y \in [0,1]^d$, $\|\nabla f(x) - \nabla f(y)\|_2 \leq \|x - y\|_2$. Our goal is to find a point $x \in [0,1]^d$ such that for any $\varepsilon' > \varepsilon$, there exists a neighborhood $N \subset [0,1]^d$ of x such that for any $y \in N$,

$$f(x) \leq f(y) + \varepsilon' \cdot \|x - y\|_2 .$$

[1] We note that a different perspective on this question from the one developed in this chapter was investigated in [136], where the author asks whether one can adapt *actual cutting planes* to non-convex settings. In particular [136] shows that one can improve upon gradient descent and obtain a complexity $O(\mathrm{poly}(d)/\varepsilon^{4/3})$ with a cutting plane method, under a higher order smoothness assumption (namely third order instead of first order here).

We say that such an x is an ε-stationary point (its existence is guaranteed by the extreme value theorem). In particular if $x \in (0,1)^d$ this means that $\|\nabla f(x)\|_2 \leq \varepsilon$. More generally, for $x = (x^1, \ldots, x^d) \in [0,1]^d$ (possibly on the boundary), let us define the *projected gradient* at x, $g(x) = (g_1(x), \ldots, g_d(x))$ by:

$$g_i(x) = \begin{cases} \max\left(0, \frac{df}{dx^i}(x)\right) & \text{if } x^i = 0\,, \\ \frac{df}{dx^i}(x) & \text{if } x^i \in (0,1)\,, \\ \min\left(0, \frac{df}{dx^i}(x)\right) & \text{if } x^i = 1\,. \end{cases}$$

It is standard to show (see also [237]) that x is an ε-stationary point of f if and only if $\|g(x)\|_2 \leq \varepsilon$.

7.1.2 Parallel query model

In the classical black-box model, the algorithm can sequentially query an oracle at points $x \in [0,1]^d$ and obtain the value[2] of the function $f(x)$. An extension of this model, first considered in [191], is as follows: instead of submitting queries one by one sequentially, the algorithm can submit any number of queries in parallel. One can then count the *depth*, defined as the number of rounds of interaction with the oracle, and the *total work*, defined as the total number of queries.

It seems that the parallel complexity of finding stationary points has not been studied before. As far as we know, the only low-depth algorithm (say depth polylogarithmic in $1/\varepsilon$) is the naive grid search: simply query all the points on an ε-net of $[0,1]^d$ (it is guaranteed that one point in such a net is an ε-stationary point). This strategy has depth 1, and total work $O(1/\varepsilon^d)$. As we explain next, the high-dimensional version of GFT has depth $O(\mathrm{poly}(d)\log(1/\varepsilon))$, and its total work improves at least quadratically upon grid search.

7.1.3 Complexity bounds for GFT

In this chapter we give a complete proof of the following near-optimal result in dimension 2:

Theorem 7.1. *Let $d = 2$. The gradient flow trapping algorithm (see Section 7.5) finds a 4ε-stationary point with less than $10^5\sqrt{\frac{\log(1/\varepsilon)}{\varepsilon}}$ queries to the value of f.*

It turns out that there is nothing inherently two-dimensional about GFT. At a very high level, one can think of GFT as making hyperplane cuts, just like standard cutting planes methods in convex optimization. While in the convex case those hyperplane cuts are simply obtained by gradients, here we obtain them by querying a $\tilde{O}(\sqrt{\varepsilon})$-net on a carefully selected small set of

[2] Technically we consider here the zeroth order oracle model. It is clear that one can obtain a first order oracle model from it, at the expense of a multiplicative dimension blow-up in the complexity. In the context of this chapter an extra factor d is small, and thus we do not dwell on the distinction between zeroth order and first order.

hyperplanes. Note also that the meaning of a "cut" is much more delicate than for traditional cutting planes methods (here we use those cuts to "trap" gradient flows). All of these ideas are more easily expressed in dimension 2, but generalizing them to higher dimensions presents no new difficulties (besides heavier notation). In Section 7.5.4 we prove the following result:

Theorem 7.2. *The high-dimensional version of GFT finds an ε-stationary point in depth $O(d^2 \log(d/\varepsilon))$ and in total work $d^{O(d)} \cdot \left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)^{\frac{d-1}{2}}$.*

In particular we see that the three-dimensional version of GFT has complexity $O\left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)$. This improves upon the previous state of the art complexity $O(1/\varepsilon^{1.2})$ [237]. However, on the contrary to the two-dimensional case, we believe that here GFT is suboptimal. As we discuss in Section 7.6.3, in dimension 3 we conjecture the lower bound $\Omega(1/\varepsilon^{0.6})$.

In dimensions $d \geq 4$, the total work given by Theorem 7.2 is worse than the total work $O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2d}{d+2}}\right)$ of Vavasis' algorithm. On the other hand, the depth of Vavasis' algorithm is of the same order as its total work, in stark contrast with GFT which maintains a logarithmic depth even in higher dimensions. Among algorithms with polylogarithmic depth, the total work given in Theorem 7.2 is more than a quadratic improvement (in fixed dimension) over the previous state of the art (namely naive grid search).

We also propose a simplified version of GFT, which we call *Cut and Flow* (CF), that always improve upon Vavasis' algorithm (in fact in dimension d it attains the same rate as Vavasis in dimension $d-1$). In particular CF attains the same rate as GFT for $d = 3$, and improves upon on it for any $d > 3$. It is however a serial algorithm and does not enjoy the parallel properties of GFT.

Theorem 7.3. *Fix $d \in \mathbb{N}$. The cut and flow algorithm (see Section 7.4) finds an ε-stationary point with less than $5d^3 \log\left(\frac{d}{\varepsilon}\right)\left(\frac{1}{\varepsilon}\right)^{\frac{2d-2}{d+1}}$ queries to the values of f and ∇f.*

Organization: The rest of the chapter (besides Section 7.6 and Section 7.7) is dedicated to motivating, describing and analyzing our gradient flow trapping strategy in dimension 2 (from now on we fix $d = 2$, unless specified otherwise). In Section 7.2 we make a basic "local to global" observation about gradient flow which forms the basis of our "trapping" strategy. Section 7.3 is an informal section on how one could potentially use this local to global phenomenon to design an algorithm, and we outline some of the difficulties one has to overcome. As a warm-up, to demonstrate the use of our ideas, we introduce the "cut and flow" algorithm in Section 7.4 and prove Theorem 7.3. In Section 7.5 we formally describe our new strategy and analyze its complexity. In Section 7.6 we extend Vavasis' $\Omega(1/\sqrt{\varepsilon})$ lower bound to randomized algorithms. Finally we conclude the chapter in Section 7.7 by introducing several open problems related to higher dimensions.

7.2 A local to global phenomenon for gradient flow

We begin with some definitions. For an axis-aligned hyperrectangle $R = [a_1, b_1] \times \cdots \times [a_d, b_d]$ in $\mathbb{R}^d$, we denote its volume and diameter by

$$\mathrm{diam}(R) := \sqrt{\sum_{i=1}^{d} (b_i - a_i)^2} \text{ and } \mathrm{vol}(R) := \prod_{i=1}^{d} (b_i - a_i)$$

We further define the aspect ratio of R as $\frac{\max_i (b_i - a_i)}{\min_i (b_i - a_i)}$. The $2d$ faces of R are the subsets of the form:

$$[a_1, b_1] \times \cdots \times \{a_i\} \times \cdots \times [a_d, b_d] \text{ and } [a_1, b_1] \times \cdots \times \{b_i\} \times \cdots \times [a_d, b_d],$$

for $i = 1, \ldots, d$. The boundary of R, which we denote ∂R is the union of all faces.

If $E \subset [0,1]^d$ is a $(d-1)$-dimensional hyperrectangle and $\delta > 0$, we say that $N \subset E$ is a δ-net of E, if for any $x \in E$, there exists some $y \in N$ such that $\|x - y\|_2 \leq \delta$. We will always assume implicitly that if $N \subset E$ is a δ-net, then the vertices of E are elements of N.

We denote $f^*_\delta(E)$ for the largest value one can obtain by minimizing f on a δ-net of E. Formally,

$$f^*_\delta(E) = \sup_N \inf_{x \in N} f(x),$$

where the supremum is taken over all δ-nets of E. We say that a pair (E, x) of segment/point in $[0,1]^d$ (where E is *not* a subset of a face of $[0,1]^d$) satisfies the property P_c for some $c \geq 0$ if there exists $\delta > 0$ such that

$$f(x) < f^*_\delta(E) - \frac{\delta^2}{8} + c \cdot \mathrm{dist}(x, E)\,,$$

where

$$\mathrm{dist}(x, E) := \inf_{y \in E} \|x - y\|_2.$$

When E is a subset of $\partial [0,1]^d$ we *always* say that (E, x) satisfies P_c (for any $c \geq 0$ and any $x \in [0,1]^d$).

For an axis-aligned hyperrectangle R and $x \in R$, we say that (R, x) satisfies P_c if, for any of the $2d$ faces E of R, one has that (E, x) satisfies P_c. We refer to x as the *pivot* for R.

Our main observation is as follows:

Lemma 7.4. *Let R be a hyperrectangle such that (R, x) satisfies P_c for some $x \in R$ and $c \geq 0$. Then R must contain a c-stationary point (in fact the gradient flow emanating from x must visit a c-stationary point before exiting R).*

This lemma will be our basic tool to develop cutting plane-like strategies for non-convex optimization. From "local" information (values on a net of the boundary of R) one deduces a "global" property (existence of approximate stationary point in R).

Proof. Let us assume by contradiction that R does not contain a c-stationary point, and consider the unit-speed gradient flow $(x(t))_{t\geq 0}$ constrained to stay in $[0,1]^d$. That is, $x(t)$ is the piecewise differentiable function defined by $x(0) = x$ and $\frac{d}{dt}x(t) = -\frac{g(x(t))}{\|g(x(t))\|_2}$, where g is the projected gradient defined in the previous section. Since there is no stationary point in R, it must be that the gradient flow exits R. Let us denote $T = \inf\{t \geq 0 : x(t) \notin R\}$, and E a face of R such that $x(T) \in E$. Remark that E cannot be part of a face of $[0,1]^d$. Furthermore, for any $0 \leq t \leq T$, one has

$$f(x(t)) - f(x(0)) = \int_0^t g(x(s)) \cdot \frac{d}{ds}x(s)ds \leq -c \cdot t \leq -c \cdot \|x(t) - x(0)\|_2 .$$

where the first inequality uses that R does not contain a c-stationary point. In particular, this implies $f(x(T)) - f(x) \leq -c \cdot \mathrm{dist}(x, E)$, so that,

$$\min_{y\in E} f(y) \leq f(x) - c \cdot \mathrm{dist}(x, E) .$$

Lemma 7.5 below shows that for any $\delta > 0$ one has $f^*_\delta(E) \leq \min_{y\in E} f(y) + \frac{\delta^2}{8}$, and thus together with the above display it shows that (E, x) does *not* satisfy P_c, which is a contradiction. □

Lemma 7.5. *For any $(d-1)$-dimensional hyperrectangle $E \subset [0,1]^2$ and $\delta > 0$ one has:*

$$f^*_\delta(E) \leq \min_{y\in E} f(y) + \frac{\delta^2}{8} .$$

Proof. Let $x \in E$ be such that $f(x) = \min_{z\in E} f(z)$. If x is a vertex of E, then we are done since we require the endpoints of E to be in the δ-nets. Otherwise x is in the relative interior of E, and thus one has $\nabla f(x) \cdot (y - x) = 0$ for any $y \in E$. In particular by smoothness one has:

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x + t(y - x)) \cdot (y - x)dt \\ &\leq f(x) + \int_0^1 t \cdot \|y - x\|_2^2 dt = f(x) + \frac{1}{2}\|y - x\|_2^2 . \end{aligned}$$

Moreover for any δ-net of E there exists y such that $\|y-x\|_2 \leq \frac{\delta}{2}$, and thus $f(y) \leq f(x)+\delta^2/8$, which concludes the proof. □

Our algorithmic approach to finding stationary points will be to somehow shrink the domain of consideration over time. At first it can be slightly unclear how the newly created boundaries

interact with the definition of stationary points. To dispell any mystery, it might be useful to keep in mind the following observation, which states that if (R, x) satisfies P_c, then x cannot be on a boundary of R which was not part of the original boundary of $[0,1]^d$.

Lemma 7.6. *Let R be a rectangle such that (R, x) satisfies P_c for some $x \in R$ and $c \geq 0$. Then $x \notin \partial R \setminus \partial[0,1]^d$.*

Proof. Let E be a face of R which is not a subset of $\partial[0,1]^d$. Then by definition of P_c, and by invoking Lemma 7.5, one has:

$$f(x) < f^*_\delta(E) - \frac{\delta^2}{8} + c \cdot \mathrm{dist}(x, E) \leq \min_{y \in E} f(y) + c \cdot \mathrm{dist}(x, E)\,.$$

In particular if $x \in E$ then $\mathrm{dist}(x, E) = 0$, and thus $f(x) < \min_{y \in E} f(y)$ which is a contradiction. □

7.3 From Lemma 7.4 to an algorithm

Lemma 7.4 naturally leads to the following algorithmic idea (for sake of simplicity in this discussion we replace squares by circles): given some current candidate point x in some well-conditioned domain (e.g., such that the domain contains and is contained in balls centered at x and of comparable sizes), query a $\sqrt{\varepsilon}$-net on the circle $C = \{y : \|y - x\|_2 = 1\}$, and denote y for the best point found on this net. If one finds a significant enough improvement, say $f(y) < f(x) - \frac{3}{4}\varepsilon$, then this is great news, as it means that one obtained a **per query** improvement of $\Theta(\varepsilon^{-3/2})$ (to be compared with gradient descent which only yields an improvement of $\Theta(\varepsilon^{-2})$). On the other hand if no such improvement is found, then the gradient flow from x must visit an ε-stationary point inside C.[3] In other words one can now hope to restrict the domain of consideration to a region inside C, which is a constant fraction smaller than the original domain. Figure 7.1 illustrates the two possibilities.

Optimistically this strategy would give a $\tilde{O}(B/\varepsilon^{3/2})$ rate for B-bounded smooth functions (since at any given scale one could make at most $O(B/\varepsilon^{3/2})$ improvement steps). In particular together with the warm start this would tentatively yield a $\tilde{O}(1/\varepsilon^{3/4})$ rate, thus already improving the state-of-the-art $O(1/\varepsilon)$ by Vavasis.

[3] In "essence" (C, x) satisfies P_ε, this is only slightly informal since we defined P_ε for rectangles and C is a circle. In particular we chose the improvement $\frac{3}{4}\varepsilon$ instead of the larger $\frac{7}{8}\varepsilon$ (which is enough to obtain P_c) to account for an extra term due to polygonal approximation of the circle. We encourage the reader to ignore this irrelevant technicality.

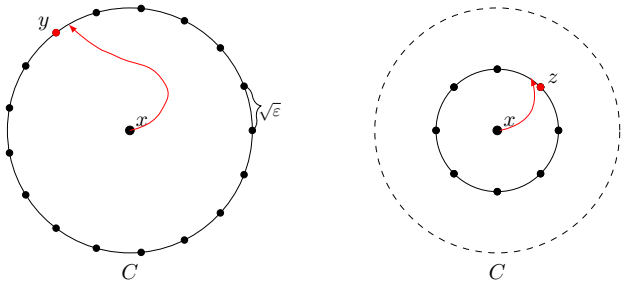

Figure 7.1: The red curve illustrates the gradient flow emanating from x. On the left, the flow does not visit an ε-stationary point and y has a significantly smaller function value than x. Otherwise, as in the right, we shrink the domain.

There is however a difficulty in the induction part of the argument. Indeed, what we know after a shrinking step is that the current point x satisfies $f(x) \leq f(y) + \varepsilon$ for any $y \in C$. Now we would like to query a net on $\{y : \|y - x\|_2 = 1/2\}$. Say that after such querying we find that we can't shrink, namely we found some point z with $f(z) < f(x) - \frac{\varepsilon}{2} + \frac{\delta^2}{8}$, and in particular $f(z) < f(y) + \frac{1}{2}\varepsilon + \frac{\delta^2}{8}$ for any $y \in C$. Could the gradient flow from z escape the original circle C without visiting an ε-stationary point? Unfortunately the answer is yes. Indeed (because of the discretization error $\delta^2/8$) one cannot rule out that there would be a point $y \in C$ with $f(y) < f(z) - \frac{\varepsilon}{2}$, and since C is only at distance $1/2$ from z, such a point could be attained from z with a gradient flow without ε-stationary points. Of course one could say that instead of satisfying P_ε we now only satisfy $P_{\varepsilon+\delta^2/4}$, and try to control the increase of the approximation guarantee, but such an approach would not improve upon the $1/\varepsilon^2$ of gradient descent (simply because we could additively worsen the approximation guarantee too many times).

The core part of the above argument will remain in our full algorithm (querying a $\sqrt{\varepsilon}$-net to shrink the domain). However it is made more difficult by the discretization error as we just saw. We also note that this discretization issue does not appear in discrete spaces, which is one reason why discrete spaces are much easier than continuous spaces for local optimization problems.

Technically we observe that the whole issue of discretization comes from the fact that when we update the center, we move closer to the boundary, which we "pay" in the term $\mathrm{dist}(x, E)$ in P_c, and we cannot "afford" it because of the discretization error term that we suffer when we update. Thus this issue would disappear if in our induction hypothesis we had P_0 for the boundary. Our strategy will work in two steps: first we give a querying strategy for a domain with P_0 that ensures that one can **always** shrink with P_ε guaranteed for the boundary, and secondly we give a method to essentially turn a P_ε boundary into P_0.

7.4 Cut and flow

We now fix $d \in \mathbb{N}$ and consider $[0,1]^d$. We say that a pair (H, x) is a *domain* if $H \subset [0,1]^d$ is an axis-aligned hyperrectangle and $x \in H$. In this section, we further require that if $H = [a_1, b_1] \times \cdots \times [a_d, b_d]$, then for every $1 \leq i, j \leq d$, $\frac{b_i - a_i}{b_j - a_j} \in \{\frac{1}{2}, 1, 2\}$. In other words, all edges of H either have the same length or differ by a factor of 2. The Cut and Flow (CF) algorithm is performed with two alternating steps, *bisection* and *descent* (See Figure 7.2 for an illustration of the two steps, when $d = 2$).

1. At the *bisection* step, we have a domain (H, x) satisfying P_0. Let $k \in [d]$ be any coordinate such that $b_k - a_k$ is maximal and set the midpoint, $m_k = \frac{a_k + b_k}{2}$. We now bisect H into two equal parts,

$$H_1 = [a_1, b_1] \times \cdots \times [a_k, m_k] \times \cdots \times [a_d, b_d],$$
$$H_2 = [a_1, b_1] \times \cdots \times [m_k, b_k] \times \cdots \times [a_d, b_d],$$

   so that $H_1 \cup H_2 = H$ and $E = H_1 \cap H_2$ is a $(d-1)$-dimensional hyperrectangle. Set $N \subset E$ to be a δ-net and,

$$x_N = \arg\min_{y \in N} f(y).$$

   Here δ is some small parameter to determined later. To choose a new pivot $\bar{x}$ for the domain we compare $f(x_N)$ and $f(x)$. If $f(x) \leq f(x_N)$, set $\bar{x} = x$ otherwise $\bar{x} = x_N$. We end the step with the two pairs $(H_1, \bar{x})$, $(H_2, \bar{x})$.

2. The *descent* step takes the two pairs produced by the *bisection* step and returns a new domain $(\tilde{H}, \tilde{x})$ satisfying P_0 such that $\tilde{H} \in \{H_1, H_2\}$. This is done be performing gradient descent iterations:

$$\bar{x}_i = \bar{x}_{i-1} - \nabla f(\bar{x}_{i-1}), \tag{7.1}$$

   where $\bar{x}_0 = \bar{x}$. Set $T = \frac{\delta^2}{\varepsilon^2}$, and $\tilde{x} = \bar{x}_T$. Then, $\tilde{H} = H_1$ if $\tilde{x} \in H_1$ and $\tilde{H} = H_2$ otherwise.

The CF algorithm starts with the domain (H_0, x_0) where $H_0 = [0,1]^d$ and x_0 is arbitrary. Given (H_t, x_t) the algorithm runs a *bisection* step, followed by a *descent* step and sets $(H_{t+1}, x_{t+1}) = (\tilde{H}, \tilde{x})$, as described above. The algorithm stops when the diameter of H_t is smaller than ε.

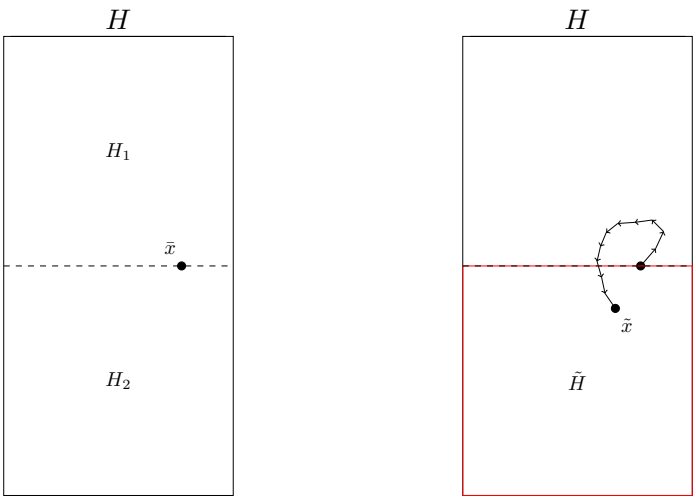

Figure 7.2: The left image shows the bisection of H into two equal parts H_1 and H_2. The right image shows the trajectory of gradient descent, starting from $\bar{x}$ and terminating at $\tilde{x}$, inside $\tilde{H}$.

Let us first prove that at the end of the *descent* step, the obtained domain satisfies P_0.

Lemma 7.7. *Suppose that (H, x) satisfies P_0, then either the* descent *step finds an ε-stationary point or $(\tilde{H}, \tilde{x})$ satisfies P_0 as well.*

Proof. Let us first estimate the value of $f(\tilde{x})$. Observe that, by smoothness of f, if we consider the gradient descent iterates (7.1), we have

$$f(\bar{x}_{i-1}) - f(\bar{x}_i) \geq \|\nabla f(\bar{x}_{i-1})\|_2^2 - \frac{1}{2}\|\nabla f(\bar{x}_{i-1})\|_2^2 = \frac{1}{2}\|\nabla f(\bar{x}_{i-1})\|_2^2 \geq \frac{\varepsilon^2}{2},$$

where the last inequality holds as long as $\bar{x}_{i-1}$ is not an ε-stationary point (see also Section 3.2 in [55]). It follows that,

$$f(\tilde{x}) = f(\bar{x}_T) \leq f(\bar{x}) - \frac{T}{2}\varepsilon^2 \leq f(x) - \frac{T}{2}\varepsilon^2.$$

Now, let $E' \subset \partial\tilde{H}$ be a face such that $E' \neq H_1 \cap H_2$. Then, $E' \subset \partial H$ and by assumption, (E', x) satisfied P_0. Since $f(\tilde{x}) \leq f(x)$, it is clear that $(E', \tilde{x})$ satisfies P_0 as well.

We are left with showing that, if $E = H_1 \cap H_2$, then $(E, \tilde{x})$ satisfies P_0. Indeed, from the construction, and using $T = \frac{\delta^2}{\varepsilon^2}$, we have,

$$f(\tilde{x}) \leq f(\bar{x}) - \frac{T}{2}\varepsilon^2 \leq f_\delta^*(E) - \frac{T}{2}\varepsilon^2 \leq f_\delta^*(E) - \frac{\delta^2}{8}.$$

□

Let us now prove Theorem 7.3.

Proof of Theorem 7.3. Observe that $\mathrm{diam}(H_0) = \sqrt{d}$ and that after performing d consecutive *bisection* steps, necessarily, every face of H_t was bisected into two equal parts. Hence, $\mathrm{diam}(H_{t+d}) \leq \frac{1}{2}\mathrm{diam}(H_t)$, and,

$$\mathrm{diam}(H_t) \leq \left(\frac{1}{2}\right)^{\lfloor \frac{t}{d} \rfloor} \sqrt{d}.$$

Choose $T = \lceil d \log_2 \left(\frac{\sqrt{d}}{\varepsilon}\right) \rceil$, so that $\mathrm{diam}(H_T) \leq \varepsilon$. We claim that x_T is an ε-stationary point. Indeed, by iterating Lemma 7.7 we know that the pair (H_T, x_T) satisfies P_0. By Lemma 7.4, there exists $x_* \in H_T$ which is a stationary point and $\|x_* - x_T\|_2 \leq \varepsilon$.

All that remains is to calculate the number of queries made by the algorithm. At the *bisection* step we query a δ-net N, over a $(d-1)$-dimensional hyperrectangle, contained in the unit cube. Elementary computations show that we can take,

$$|N| \leq \frac{(2d)^{d-1}}{\delta^{d-1}}.$$

Combined with the number of queries made by the *descent step*, we see that the total number of queries made by the algorithm is,

$$\left\lceil d \log_2 \left(\frac{\sqrt{d}}{\varepsilon}\right) \right\rceil \left(\frac{(2d)^{d-1}}{\delta^{d-1}} + \frac{\delta^2}{\varepsilon^2}\right).$$

We now optimize and choose $\delta = \varepsilon^{\frac{2}{d+1}} 2d$. Substituting into the above equations shows that the number of queries is smaller than

$$5d^3 \log_2 \left(\frac{d}{\varepsilon}\right) \varepsilon^{-\frac{2d-2}{d+1}}.$$

□

7.5 Gradient flow trapping

In this section we focus on the case $d = 2$. We say that a pair (R, x) is a *domain* if R is an axis-aligned rectangle with aspect ratio bounded by 3, and $x \in R$ (note that the definition of a domain is slightly different than in the previous section). The gradient flow trapping (GFT) algorithm is decomposed into two subroutines:

1. The first algorithm, which we call the *parallel trap*, takes as input a domain (R, x) satisfying P_0. It returns a domain $(\tilde{R}, \tilde{x})$ satisfying P_ε and such that $\mathrm{vol}(\tilde{R}) \leq 0.95\ \mathrm{vol}(R)$. The cost of this step is at most $2\sqrt{\frac{\mathrm{diam}(R)}{\varepsilon}}$ queries.

2. The second algorithm, which we call *edge fixing*, takes as input a domain (R, x) satisfying $P_{\varepsilon'}$ (for some $\varepsilon' \in [\varepsilon, 2\varepsilon]$) and such that for $k \in \{0, 1, 2, 3\}$ edges E of R one also has P_0 for (E, x). It returns a domain $(\tilde{R}, \tilde{x})$ such that either (i) it satisfies $P_{\varepsilon'}$ and for $k+1$ edges it also satisfies P_0, or (ii) it satisfies $P_{\left(1+\frac{1}{500\log(1/\varepsilon)}\right)\varepsilon'}$ and furthermore $\mathrm{vol}(\tilde{R}) \leq 0.95\ \mathrm{vol}(R)$. The cost of this step is at most $90\sqrt{\frac{\mathrm{diam}(R)\log(1/\varepsilon)}{\varepsilon}}$ queries.

Equipped with these subroutines, GFT proceeds as follows. Initialize $(R_0, x_0) = ([0,1]^2, (0.5, 0.5))$, $\varepsilon_0 = \varepsilon$, and $k_0 = 4$. For $t \geq 0$:

- if $k_t = 4$, call *parallel trap* on (R_t, x_t), and update $k_{t+1} = 0$, $(R_{t+1}, x_{t+1}) = (\tilde{R}_t, \tilde{x}_t)$, and $\varepsilon_{t+1} = \varepsilon$.
- Otherwise call *edge fixing*, and update $(R_{t+1}, x_{t+1}) = (\tilde{R}_t, \tilde{x}_t)$. If $R_{t+1} = R_t$ then set $k_{t+1} = k_t + 1$ and $\varepsilon_{t+1} = \varepsilon_t$, and otherwise set $k_{t+1} = 0$ and $\varepsilon_{t+1} = \left(1 + \frac{1}{500\log(1/\varepsilon)}\right)\varepsilon_t$.

We terminate once the diameter of R_t is smaller than 2ε.

Next we give the complexity analysis of GFT assuming the claimed properties of the subroutines *parallel trap* and *edge fixing* in 1. and 2. above. We then proceed to describe in details the subroutines, and prove that they satisfy the claimed properties.

7.5.1 Complexity analysis of GFT

The following three lemmas give a proof of Theorem 7.1.

Lemma 7.8. *GFT stops after at most* $200\log(1/\varepsilon)$ *steps.*

Proof. First note that at least one out of five steps of GFT reduces the volume of the domain by 0.95 (since one can do at most four steps in a row of edge fixing without volume decrease). Thus on average the volume decrease per step is at least 0.99, i.e., $\mathrm{vol}(R_T) \leq 0.99^T$. In particular since R_T has aspect ratio smaller than 3, it is easy to verify $\mathrm{diam}(R_T) \leq 2\sqrt{\mathrm{vol}(R_T)} \leq 2 \times 0.99^{T/2}$. Thus for any $T \geq \log_{100/99}(1/\varepsilon^2)$, one must have $\mathrm{diam}(R_T) \leq 2\varepsilon$. Thus we see that GFT performs at most $\log_{100/99}(1/\varepsilon^2) \leq 200\log(1/\varepsilon)$ steps. □

Lemma 7.9. *When GFT stops, its pivot is a* 4ε*-stationary point.*

Proof. First note that $\varepsilon_T \leq \left(1 + \frac{1}{500\log(1/\varepsilon)}\right)^T \varepsilon$, thus after $T \leq 200\log(1/\varepsilon)$ steps we know that (R_T, x_T) satisfies at least $P_{2\varepsilon}$. In particular by Lemma 7.4, R_T must contain a 2ε-stationary point, and since the diameter is less than 2ε, it must be (by smoothness) that x_T is a 4ε-stationary point. □

Lemma 7.10. *GFT makes at most* $10^5\sqrt{\frac{\log(1/\varepsilon)}{\varepsilon}}$ *queries before it stops.*

Proof. As we saw in the proof of Lemma 7.8, one has $\operatorname{diam}(R_t) \leq 2 \times 0.99^{t/2}$. Furthermore the t^{th} step requires at most $90\sqrt{\frac{\operatorname{diam}(R_t)\log(1/\varepsilon)}{\varepsilon}}$ queries. Thus the total number of queries is bounded by:

$$90\sqrt{\frac{2\log(1/\varepsilon)}{\varepsilon}}\sum_{t=0}^{\infty} 0.99^{t/4} \leq 10^5\sqrt{\frac{\log(1/\varepsilon)}{\varepsilon}}\,.$$

□

7.5.2 A parallel trap

Let (R, x) be a domain. We define two segments E and F in R as follows. Assume that R is a translation of $[0, s] \times [0, r]$. For sake of notation assume that in fact $R = [0, s] \times [0, r]$ with $s \in [r, 3r]$ and $x^1 \geq r/2$, where $x = (x^1, x^2)$ (in practice one always ensures this situation with a simple change of variables). Now we define $E = \{r/6\} \times [0, r]$ and $F = \{r/3\} \times [0, r]$ (See Figure 7.3).

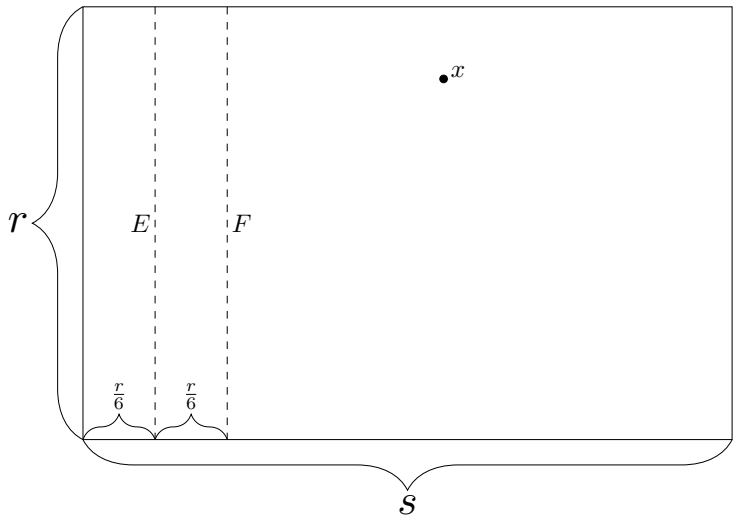

Figure 7.3: The *parallel trap*

The parallel trap algorithm queries a $\sqrt{r\varepsilon}$-net on both E and F (which cost at most $2\frac{r}{\sqrt{r\varepsilon}} = 2\sqrt{\frac{r}{\varepsilon}}$). Denote $\bar{x}$ to be the best point (in terms of f value) found on the union of those nets. That is, denoting $N \subset F \cup E$ for the queried $\sqrt{r\varepsilon}$-net, then

$$\bar{x} = \arg\min_{y \in N} f(y).$$

One has the following possibilities (see Figure 7.4 for an illustration):

- If $f(x) \leq f(\bar{x})$ then we set $\tilde{x} = x$ and $\tilde{R} = [r/3, s] \times [0, r]$.
- Otherwise we set $\tilde{x} = \bar{x}$. If $\bar{x} \in F$ we set $\tilde{R} = [r/6, s] \times [0, r]$, and if $\bar{x} \in E$ we set $\tilde{R} = [0, r/3] \times [0, r]$.

The above construction is justified by the following lemma (a trivial consequence of the definitions), and it proves in particular the properties of *parallel trap* described in 1. at the beginning of Section 7.5.

Lemma 7.11. *The rectangle $\tilde{R}$ has aspect ratio smaller than 3, and it satisfies* $\mathrm{vol}(\tilde{R}) \leq 0.95\,\mathrm{vol}(R)$. *Furthermore if (R, x) satisfies P_0, then $(\tilde{R}, \tilde{x})$ satisfies P_ε.*

Proof. The first sentence is trivial to verify. For the second sentence, first note that for any edge E of R one has P_0 for $(E, \tilde{x})$ since by assumption one has P_0 for (E, x) and furthermore $f(\tilde{x}) \leq f(x)$. Next observe that $\tilde{R}$ has at most one new edge $\tilde{E}$ with respect to R, and this edge is at distance at least $r/6$ from $\tilde{x}$, thus in particular one has $\varepsilon \cdot \mathrm{dist}(\tilde{x}, \tilde{E}) - \delta^2/8 > 0$ for $\delta = \sqrt{r\varepsilon}$. Furthermore by definition $f(\tilde{x}) \leq f^*_\delta(\tilde{E})$, and thus $f(\tilde{x}) < f^*_\delta(\tilde{E}) - \frac{\delta^2}{8} + \varepsilon \cdot \mathrm{dist}(\tilde{x}, \tilde{E})$, or in other words $(\tilde{E}, \tilde{x})$ satisfies P_ε. ◻

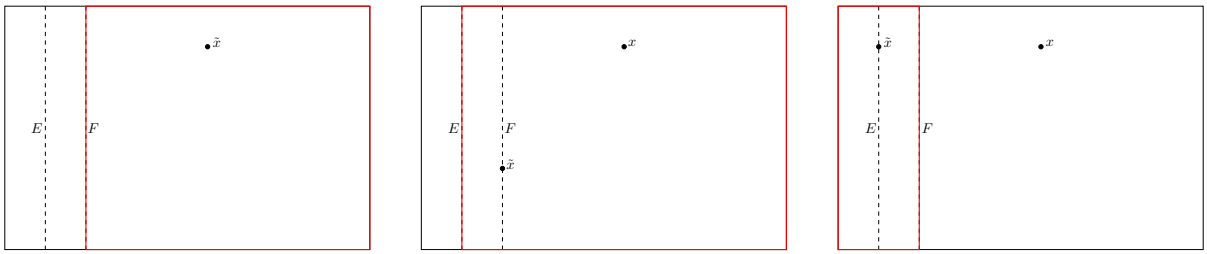

Figure 7.4: The three possible cases for $(\tilde{R}, \tilde{x})$. $\tilde{R}$ is marked in red.

7.5.3 Edge fixing

Let (R, x) be a domain satisfying $P_{\varepsilon'}$ for some $\varepsilon' \in [\varepsilon, 2\varepsilon]$, and with some edges possibly also satisfying P_0. Denote $\mathcal{E}$ for the closest edge to x that does not satisfy P_0, and let $r = \mathrm{dist}(x, \mathcal{E})$. We will consider three[4] candidate smaller rectangles, R_1, R_2 and R_3, as well as three candidate pivots (in addition to x) $x_1 \in \partial R_1$, $x_2 \in \partial R_2$ and $x_3 \in \partial R_3$. The rectangles are defined by $R_i = R \cap \{y : \|x_{i-1} - y\|_\infty \leq \frac{r}{3}\}$, where we set $x_0 = x$. The possible output $(\tilde{R}, \tilde{x})$ of *edge fixing* will be either (R_i, x_{i-1}) for some $i \in \{1, 2, 3\}$, or (R, x_3) (see Figure 7.5 for a demonstration of how to construct the rectangles).

To guarantee the properties described in 2. at the beginning of Section 7.5 we will prove the following: if the output is (R_i, x_{i-1}) for some i then all edges will satisfy $P_{\left(1+\frac{1}{500\log(1/\varepsilon)}\right)\varepsilon'}$ (Lemma 7.14 below) and the domain has shrunk (Lemma 7.12 below), and if the output is (R, x_3) then one more edge satisfies P_0 compared to (R, x) while all edges still satisfy at least $P_{\varepsilon'}$ (Lemma 7.13 below).

Lemma 7.12. *For any $i \in \{1, 2, 3\}$ one has* $\mathrm{vol}(R_i) \leq \frac{2}{3}\mathrm{vol}(R)$. *Furthermore if the aspect ratio of R is smaller than 3, then so is the aspect ratio of R_i.*

[4] We need three candidates to ensure that the domain will shrink.

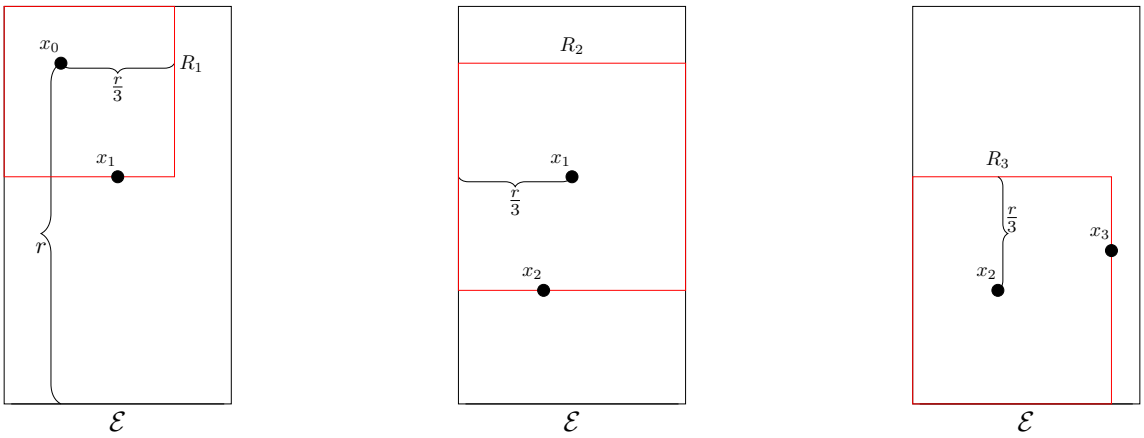

Figure 7.5: *Edge fixing*: the rectangles R_1, R_2 and R_3 are marked in red, from left to right.

Proof. Let us denote $\ell_1(R)$ for the length of R in the axis of $\mathcal{E}$ (the edge whose distance to x defines r), and $\ell_2(R)$ for the length in the orthogonal direction (and similarly define $\ell_1(R_i)$ and $\ell_2(R_i)$).

Since $R_i \subset R$ one has $\ell_1(R_i) \leq \ell_1(R)$. Furthermore $\ell_2(R_i) \leq \frac{2}{3}r$ and $\ell_2(R) \geq r$, so that $\ell_2(R_i) \leq \frac{2}{3}\ell_2(R)$. This implies that $\mathrm{vol}(R_i) \leq \frac{2}{3}\mathrm{vol}(R)$.

For the second statement observe that $\ell_1(R) \geq \frac{\ell_2(R)}{3} \geq \frac{r}{3}$ (the first inequality is by assumption on the aspect ratio of R, the second inequality is by definition of r). Given this estimate, the construction of R_i implies that $\frac{1}{3}r \leq \ell_2(R_i), \ell_1(R_i) \leq \frac{2}{3}r$, which concludes the fact that R_i has aspect ratio smaller than 2. □

Queries and choice of output. The edge fixing algorithm queries a $\sqrt{\frac{\varepsilon' r}{500\log(1/\varepsilon)}}$-net on ∂R_i for all $i \in \{1,2,3\}$ (thus a total of $4\sqrt{\frac{500r\log(1/\varepsilon)}{\varepsilon'}} \leq 90\sqrt{\frac{r\log(1/\varepsilon)}{\varepsilon}}$ queries), and we define x_i to be the best point found on each respective net.

If for all $i \in \{1,2,3\}$ one has

$$f(x_i) \leq f(x_{i-1}) - \frac{\varepsilon' r}{3}\,, \tag{7.2}$$

then we set $(\tilde{R}, \tilde{x}) = (R, x_3)$. Otherwise denote $i^* \in \{1,2,3\}$ for the smallest number which violates (7.2), and set $(\tilde{R}, \tilde{x}) = (R_{i^*}, x_{i^*-1})$.

Lemma 7.13. *If $(\tilde{R}, \tilde{x}) = (R, x_3)$ then $(\mathcal{E}, x_3)$ satisfies P_0. Furthermore for any edge E of R, if (E, x) satisfies P_0 (respectively $P_{\varepsilon'}$) then so does (E, x_3).*

Proof. Since $(\tilde{R}, \tilde{x}) = (R, x_3)$ it means that $f(x_3) \leq f(x_0) - \varepsilon' r$. In particular since $(\mathcal{E}, x_0)$ satisfies $P_{\varepsilon'}$ one has $f(x_0) < f^*_\delta(\mathcal{E}) - \frac{\delta^2}{8} + \varepsilon' r$, and thus now one has $f(x_3) < f^*_\delta(\mathcal{E}) - \frac{\delta^2}{8}$ which means that $(\mathcal{E}, x_3)$ satisfies P_0.

Let us now turn to some other edge E of R. Certainly if (E, x_0) satisfies P_0 then so does (E, x_3) since $f(x_3) \leq f(x_0)$. But, in fact, even $P_{\varepsilon'}$ is preserved since by the triangle inequality

(and $\|x_3 - x_0\|_2 \leq r$) one has

$$f(x_3) - \varepsilon' \cdot \mathrm{dist}(x_3, E) \leq f(x_3) + \varepsilon' r - \varepsilon' \cdot \mathrm{dist}(x_0, E) \leq f(x_0) - \varepsilon' \cdot \mathrm{dist}(x_0, E).$$

□

Lemma 7.14. *If $(\tilde{R}, \tilde{x}) = (R_i, x_{i-1})$ for some $i \in \{1, 2, 3\}$, then $(\tilde{R}, \tilde{x})$ satisfy $P_{\left(1+\frac{1}{500\log(1/\varepsilon)}\right)\varepsilon'}$.*

Proof. By construction, if $(\tilde{R}, \tilde{x}) = (R_i, x_{i-1})$, then for any edge E of R_i one has $f(x_{i-1}) < f^*_\delta(E) + \frac{\varepsilon' r}{3}$. Furthermore one has $\frac{\varepsilon' r}{3} = -\frac{\varepsilon' r}{8\times 500\log(1/\varepsilon)} + \left(1 + \frac{3}{8\times 500\log(1/\varepsilon)}\right)\frac{\varepsilon' r}{3}$, and thus one has $P_{\left(1+\frac{3}{8\times 500\log(1/\varepsilon)}\right)\varepsilon'}$ for (E, x_{i-1}) whenever $\mathrm{dist}(x_{i-1}, E) = \frac{r}{3}$. Indeed, since $\delta = \sqrt{\frac{\varepsilon' r}{500\log(1/\varepsilon)}}$,

$$f(x_{i-1}) < f^*_\delta(E) - \frac{\delta^2}{8} + \left(1 + \frac{3}{8 \times 500\log(1/\varepsilon)}\right)\varepsilon' \cdot \mathrm{dist}\left(x_{i-1}, E\right).$$

If $\mathrm{dist}(x_{i-1}, E) < \frac{r}{3}$ then by the triangle inequality, $\mathrm{dist}(x_0, E) < r$, and moreover E is also an edge with respect to R. Thus from the definition of r, (E, x_0) satisfies P_0. Also by our choice of x_{i-1}, we know that $f(x_{i-1}) \leq f(x_0)$. Hence (E, x_{i-1}) satisfies P_0 as well. □

7.5.4 Generalization to higher dimensions

As explained in the introduction, there is no reason to restrict GFT to $[0, 1]^2$ and, in fact, the algorithm may be readily adapted to higher-dimensional spaces, such as $[0, 1]^d$, for some $d > 2$. We now detail the necessary changes and derive the complexity announced in Theorem 7.2.

First, if F is an affine hyperplane, and $x \in [0, 1]^d$, we define P_c for (F, x) in the obvious way (i.e., same definition except that we consider a δ-net of F). Similarly for (R, x), when R is an axis-aligned hyperrectangle.

Gradient flow trapping in higher dimensions replaces every line by a hyperplane, and every rectangle by a hyperrectangle. In particular at each step GFT maintains a domain (R, x), where R is a hyperrectangle with aspect ratio bounded by 3, and $x \in R$. The two subroutines are adapted as follows:

1. *Parallel trap* works exactly in the same way, except that the two lines E and F are replaced by two corresponding affine hyperplanes. In particular the query cost of this step is now $O\left(\left(\frac{\mathrm{diam}(R)}{\varepsilon}\right)^{\frac{d-1}{2}}\right)$, and the volume shrinks by at least 0.95.

2. In *edge fixing*, we now have three hyperrectangles R_i, and we need to query nets on their $2d$ faces. Thus the total cost of this step is $O\left(d\left(\frac{\mathrm{diam(R)}\log(1/\varepsilon)}{\varepsilon}\right)^{\frac{d-1}{2}}\right)$. Moreover, suppose that domain does not shrink at the end of this step and the output is a domain $(R, \tilde{x})$ for some other $\tilde{x} \in R$. In this case we know that R has some face F, such that

(F, x) did not satisfy P_0, but $(F, \tilde{x})$ does satisfy P_0. It follows that we can run *edge fixing*, at most $2d$ times before the domain shrinks.

We can now analyze the complexity of the high-dimensional version of GFT:

Proof of Theorem 7.2. First observe that, if R is a hyperrectangle in $[0,1]^d$ with aspect ratio bounded by 3, then we have the following inequality,

$$\operatorname{diam}(R) \leq 3\sqrt{d} \cdot \operatorname{vol}(R)^{\frac{1}{d}}.$$

By repeating the same calculations done in Lemma 7.8 and the observation about *parallel trap* and *edge fixing* made above, we see that the domain shrinks at least once in every $2d+1$ steps, so that at step T,

$$\operatorname{vol}(R_T) \leq 0.95^{\frac{T}{2d+1}},$$

and

$$\operatorname{diam}(R_T) \leq 3\sqrt{d} \cdot 0.95^{\frac{T}{(2d+1)d}}.$$

Since the algorithm stops when $\operatorname{diam}(R_T) \leq 2\varepsilon$, we get

$$T = O\left(d^2 \log\left(\frac{d}{\varepsilon}\right)\right).$$

The total work done by the algorithm is evident now by considering the number of queries at each step. □

7.6 Lower bound for randomized algorithms

In this section, we show that any randomized algorithm must make at least $\tilde{\Omega}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ queries in order to find an ε-stationary point. This extends the lower bound in [237], which applied only to deterministic algorithms. In particular, it shows that, up to logarithmic factors, adding randomness cannot improve the algorithm described in the previous section.

For an algorithm $\mathcal{A}$, a function $f : [0,1]^2 \to \mathbb{R}$ and $\varepsilon > 0$ we denote by $\mathcal{Q}(\mathcal{A}, f, \varepsilon)$ the number of queries made by $\mathcal{A}$, in order to find an ε-stationary point of f. Our goal is to bound from below

$$\mathcal{Q}_{\text{rand}}(\varepsilon) := \inf_{\mathcal{A} \text{ random}} \sup_{f} \mathbb{E}_{\mathcal{A}}\left[\mathcal{Q}(\mathcal{A}, f, \varepsilon)\right],$$

where the infimum is taken over all random algorithms and the supremum is taken over all smooth functions, f. The expectation is with respect to the randomness of $\mathcal{A}$. By Yao's minimax

principle we have the equality

$$\mathcal{Q}_{\mathrm{rand}}(\varepsilon) = \sup_{\mathcal{D}} \inf_{\mathcal{A} \text{ determinstic}} \mathbb{E}_{f\sim\mathcal{D}} \left[\mathcal{Q}\left(\mathcal{A}, f, \varepsilon\right)\right].$$

Here, $\mathcal{A}$ is a deterministic algorithm and $\mathcal{D}$ is a distribution over smooth functions. The rest of this section is devoted to proving the following theorem:

Theorem 7.15. *Let $h : \mathbb{N} \to \mathbb{R}$ be a decreasing function such that*

$$\sum_{k=1}^{\infty} \frac{h(k)}{k} < \infty,$$

and set

$$S_h(n) := \sum_{k=1}^{n} \frac{1}{k \cdot h(k)}. \tag{7.3}$$

Then,

$$\mathcal{Q}_{\mathrm{rand}}(\varepsilon) = \Omega\left(\frac{1}{\sqrt{\varepsilon} \cdot S_h\left(\left\lceil \frac{1}{\sqrt{\varepsilon}} \right\rceil\right)}\right).$$

Remark that one may take $h(k) := \frac{1}{\log(k)^2+1}$ in the theorem. In this case $S_h(k) = O(\log^3(k))$, and $\mathcal{Q}_{\mathrm{rand}}(\varepsilon) = \Omega\left(\frac{1}{\log^3(1/\varepsilon)\sqrt{\varepsilon}}\right)$, which is the announced lower bound.

One of the main tools utilized in our proof is the construction introduced in [237]. We now present the relevant details.

7.6.1 A reduction to monotone path functions

Let $G_n = (V_n, E_n)$ stand for the $n+1 \times n+1$ grid graph. That is,

$$V_n = \{0, \ldots, n\} \times \{0, \ldots, n\} \text{ and } E_n = \{(v, u) \in V_n \times V_n : \|v - u\|_1 = 1\}.$$

We say that a sequence of vertices, $(v_0, ..., v_n)$ is a *monotone path* in G_n if $v_0 = (0, 0)$ and for every $0 < i \leq n$, $v_i - v_{i-1}$ either equals $(0, 1)$ or $(1, 0)$. In other words, the path starts at the origin and continues each step by either going right or up. If $(v_0, ..., v_n)$ is a monotone path, we associate to it a *monotone path function* $P : V_n \to \mathbb{R}$ by

$$P(v) = \begin{cases} -\|v\|_1 & \text{if } v \in \{v_0, ..., v_n\} \\ \|v\|_1 & \text{otherwise} \end{cases}.$$

By a slight abuse of notation, we will sometimes refer to the path function and the path itself as the same entity. If $i = 0, ..., n$ we write P_i for $P^{-1}(-i)$ and $P[i]$ for the prefix $(P_0, P_1, ..., P_i)$.

If $v \in V_n$ is such that $P(v) > 0$, we say that v does not lie on the path.
We denote the set of all monotone path functions on G_n by F_n. It is clear that if $P \in \mathrm{F}_n$ then P_n is the only local minimum of P and hence the global minimum.

Informally, the main construction in [237] shows that for every $P \in \mathrm{F}_n$ there is a corresponding smooth function $\hat{P} : [0,1]^2 \to \mathbb{R}$, which 'traces' the path in P and preserves its structure. In particular, finding an ε-stationary point of $\hat{P}$ is not easier than finding the minimum of P.

To formally state the result we fix $\varepsilon > 0$ and assume for simplicity that $\frac{1}{\sqrt{\varepsilon}}$ is an integer. We henceforth denote $n(\varepsilon) := \frac{1}{\sqrt{\varepsilon}}$ and identify $V_{n(\varepsilon)}$ with $[0,1]^2$ in the following way: if $(i,j) = v \in V_{n(\varepsilon)}$ we write $\mathrm{square}(v)$ for the square:

$$\mathrm{square}(v) = \left[\frac{i}{n(\varepsilon)+1}, \frac{i+1}{n(\varepsilon)+1}\right] \times \left[\frac{j}{n(\varepsilon)+1}, \frac{j+1}{n(\varepsilon)+1}\right].$$

If $\varphi : [0,1]^2 \to \mathbb{R}$, then $\mathrm{supp}(\varphi)$ denotes the closure of the set $\{x \in [0,1]^2 : \varphi(x) \neq 0\}$.

Lemma 7.16 (Section 3, [237])**.** *Let $P \in \mathrm{F}_{n(\varepsilon)}$. Then there exists a function $\hat{P} : [0,1]^2 \to \mathbb{R}$ with the following properties:*

1. *$\hat{P}$ is smooth.*
2. *$\hat{P} = f_P + \ell$, where ℓ is a linear function, which does not depend on P, and*

$$\mathrm{supp}(f_P) \subset \bigcup_{i=0}^{n} \mathrm{square}\,(P_i)\,.$$

3. *If $x \in [0,1]^2$ is an ε-stationary point of $\hat{P}$ then $x \in \mathrm{square}\,(P_n)$.*
4. *if $P' \in \mathrm{F}_{n(\varepsilon)}$ is another function and for some $i = 0, ..., n$, $(P'_{i-1}, P'_i, P'_{i+1}) = (P_{i-1}, P_i, P_{i+1})$. Then*

$$\hat{P'}|_{\mathrm{square}(P_i)} = \hat{P}|_{\mathrm{square}(P_i)}$$

We now make precise of the fact that finding the minimum of P is as hard as finding an ε-stationary point of $\hat{P}$. For this we define $\mathcal{G}(\mathcal{A}, P)$, the number of queries made by algorithm $\mathcal{A}$, in order to find the minimal value of the function P.

Lemma 7.17. *For any algorithm $\mathcal{A}$, which finds an ε-stationary point of smooth functions on $[0,1]^2$, there exists an algorithm $\tilde{\mathcal{A}}$ such that*

$$\mathcal{Q}(\mathcal{A}, \hat{P}, \varepsilon) \geq \frac{1}{5}\mathcal{G}(\tilde{\mathcal{A}}, P),$$

for any $P \in \mathrm{F}_{n(\varepsilon)}$.

Proof. Given an algorithm $\mathcal{A}$ we explain how to construct $\tilde{\mathcal{A}}$. Fix $P \in \mathrm{F}_{n(\varepsilon)}$. If $\mathcal{A}$ queries a point $x \in \mathrm{square}\,(v) \subset [0,1]^2$. Then $\tilde{\mathcal{A}}$ queries v and all of its neighbors. When $\mathcal{A}$ terminates it has found an ε-stationary point. By Lemma 7.16, this point must lie in $\mathrm{square}\,(P_n)$. By querying P_n and its neighbors, $\tilde{\mathcal{A}}$ will determine that P_n is a local minimum and hence the minimum of P.

Since each vertex has at most 4 neighbors, it will now suffice to show that $\tilde{\mathcal{A}}$ can remain consistent with $\mathcal{A}$. We thus need to show that after querying the neighbors of v, $\tilde{\mathcal{A}}$ may deduce the value of $\hat{P}(x)$.

As we are only interested in the number of queries made by $\tilde{\mathcal{A}}$, it is fine to assume that $\tilde{\mathcal{A}}$ has access to the construction used in Lemma 7.16. Now, suppose that $P(v) > 0$ and v does not lie on the path. In this case, by Lemma 7.16, $\hat{P}(x) = \ell(x)$, which does not depend on P itself and $\ell(x)$ is known. Otherwise $v = P_i$ for some $i = 0, ..., n$. So, after querying the neighbors of v, $\tilde{\mathcal{A}}$ also knows P_{i-1} and P_{i+1}. The lemma then tells us that $\hat{P}|_{\mathrm{square}(v)}$ is uniquely determined and, in particular, the value of $\hat{P}(x)$ is known. □

7.6.2 A lower bound for monotone path functions

Denote $\mathcal{D}_{\mathrm{p}}(n)$ to be the set of all distributions supported on F_n. By Lemma 7.17,

$$\mathcal{Q}_{\mathrm{rand}}(\varepsilon) \geq \sup_{\mathcal{D} \in \mathcal{D}_{\mathrm{p}}(n(\varepsilon))} \inf_{\mathcal{A}\ \text{determinstic}} \mathbb{E}_{P \sim \mathcal{D}} \left[\mathcal{Q}\left(\mathcal{A}, \hat{P}, \varepsilon\right) \right] \geq \frac{1}{5} \sup_{\mathcal{D} \in \mathcal{D}_{\mathrm{p}}(n(\varepsilon))} \inf_{\mathcal{A}\ \text{determinstic}} \mathbb{E}_{P \sim \mathcal{D}} \left[\mathcal{G}\left(\tilde{\mathcal{A}}, P\right) \right].$$

In [227], the authors present a family of random paths $(X_\delta)_{\delta > 0} \subset \mathcal{D}_{\mathrm{p}}(n)$. Using these random paths it is shown that for every $\delta > 0$,

$$\mathcal{G}_{\mathrm{rand}}(n) := \sup_{\mathcal{D} \in \mathcal{D}_{\mathrm{p}}(n)} \inf_{\mathcal{A}\ \text{determinstic}} \mathbb{E}_{P \sim \mathcal{D}} \left[\mathcal{G}\left(\mathcal{A}, P\right) \right] = \Omega(n^{1-\delta}).$$

This immediately implies,

$$\mathcal{Q}_{\mathrm{rand}}(\varepsilon) = \Omega\left(\left(\frac{1}{\sqrt{\varepsilon}} \right)^{1-\delta} \right).$$

Their proof uses results from combinatorial number theory in order to construct a random path which, roughly speaking, has unpredictable increments. This distribution is then used in conjunction with a method developed by Aaronson ([1]) in order to produce a lower bound.

We now present a simplified proof of the result, which also slightly improves the bound. We simply observe that known results concerning unpredictable random walks, can be combined with Aaronson's method. Theorem 7.15 will then be a consequence of the following theorem:

Theorem 7.18. *Let the notations of Theorem 7.15 prevail. Then*

$$\mathcal{G}_{\mathrm{rand}}(n) = \Omega\left(\frac{n}{S_h(n)}\right).$$

The theorem of Aaronson, reformulated using our notations (see also [227, Lemma 2]), is given below.

Theorem 7.19 (Theorem 5, [1])**.** *Let $w : \mathrm{F}_n \times \mathrm{F}_n \to \mathbb{R}^+$ be a weight function with the following properties:*

- $w(P, P') = w(P', P)$.
- $w(P, P') = 0$*, whenever* $P_n = P'_n$.

Define

$$T(w, P) := \sum_{Q \in \mathrm{F}_n} w(P, Q),$$

and for $v \in V_n$

$$T(w, P, v) := \sum_{Q \in \mathrm{F}_n : Q(v) \neq P(v)} w(P, Q).$$

Then

$$\mathcal{G}_{\mathrm{rand}}(n) = \Omega\left(\min_{\substack{P,P',v \\ PP(v) \neq P'(v), w(P,P') > 0}} \max\left(\frac{T(w,P)}{T(w,P,v)}, \frac{T(w,P')}{T(w,P',v)}\right)\right).$$

For $P \in \mathrm{F}_n$, one should think about w as inducing a probability measure according to $w(P, \cdot)$. If Q is sampled according to this measure, then the quantity $\frac{T(w,P,v)}{T(w,P)}$ is the probability that $P(v) \neq Q(v)$. That is, either $v \in P$ or $v \in Q$, but not both. The theorem then says that if this probability is small, for at least one path in each pair (P, P') such that $P_n \neq P'_n$, then any randomized algorithm must make as many queries as the reciprocal of the probability.

We now formalize this notion; For a random path $X \in \mathcal{D}_{\mathrm{p}}(n)$, define the following weight function:

$$w_X(P, P') = \begin{cases} 0 & \text{if } P_n = P'_n \\ \mathbb{P}(X = P) \cdot \sum_{i=0}^{n-1} \mathbb{P}(X = P' | X[i] = P[i]) & \text{otherwise} \end{cases}.$$

Here $w_X(P, P')$ is proportional to the probability that $X = P'$, conditional on agreeing with P on the first i steps, where i is uniformly chosen between 0 and $n - 1$. Note that, for any i,

$$\begin{aligned} &\mathbb{P}(X = P) \cdot \mathbb{P}(X = P' | X[i] = P[i]) \\ &\qquad = \mathbb{P}(X[i] = P[i]) \cdot \mathbb{P}(X = P | X[i] = P[i]) \cdot \mathbb{P}(X = P' | X[i] = P[i]) \\ &\qquad = \mathbb{P}(X = P') \cdot \mathbb{P}(X = P | X[i] = P'[i]). \end{aligned}$$

Hence, $w_X(P,P') = w_X(P',P)$. We will use the following theorem from [131], which generalizes the main result of [31].

Theorem 7.20 (Theorem 1.4, [131])**.** *Let h be as in Theorem 7.15, Then there exists a random path $X^h \in \mathcal{D}_{\mathrm{p}}(n)$ and a constant $c_h > 0$, such that for all $m \geq k$, and for every $(v_0, v_1 ..., v_{m-k})$, sequence of vertices,*

$$\sup_{\|u\|_1 = m} \mathbb{P}\left(X^h_m = u | X^h_0 = v_0, ..., X^h_{m-k} = v_{m-k}\right) \leq \frac{c_h}{kf(k)}. \tag{7.4}$$

For X^h as in the theorem abbreviate $w_h := w_{X^h}$ and recall $S_h(n) := \sum_{k=1}^{n} \frac{1}{k \cdot h(k)}$. We now prove the main quantitative estimates which apply to w_h.

Lemma 7.21. *For any $P \in \mathrm{F}_n$,*

$$\sum_{Q \in \mathrm{F}_n} w_h(P,Q) \geq \mathbb{P}(X^h = P) \cdot (n - c_h S_h(n)) .$$

Proof. We write

$$\begin{aligned}
\sum_{Q \in \mathrm{F}_n} w_h(P,Q) &= \mathbb{P}(X^h = P) \sum_{Q : Q_n \neq P_n} \sum_{i=0}^{n-1} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right) \\
&= \mathbb{P}(X^h = P) \sum_{i=0}^{n-1} \left(1 - \sum_{Q : Q_n = P_n} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right)\right) \\
&= \mathbb{P}(X^h = P) \left(n - \sum_{i=0}^{n-1} \sum_{Q : Q_n = P_n} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right)\right).
\end{aligned}$$

Using (7.4), we get

$$\sum_{Q : Q_n = P_n} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right) \leq \mathbb{P}(X^h_n = P_n | X^h[i] = P[i]) \leq \frac{c_h}{(n-i) \cdot h(n-i)},$$

and

$$\sum_{i=0}^{n-1} \sum_{Q : Q_n = P_n} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right) \leq \sum_{k=1}^{n} \frac{c_h}{k \cdot h(k)} = c_h S_h(n).$$

□

Lemma 7.22. *Let $P \in \mathrm{F}_n$ and $v \in V_n$ such that $\|v\|_1 = \ell$ and $P_\ell \neq v$. Then,*

$$\sum_{\substack{Q \in \mathrm{F}_n \\ Q_\ell = v}} w_h(P,Q) \leq 2\mathbb{P}(X^h = P) c_h S_h(n).$$

Proof.

$$\sum_{\substack{Q\in \mathrm{F}_n\\ Q_\ell = v}} w_h(P,Q) = \mathbb{P}(X^h = P) \sum_{i=0}^{\ell-1} \sum_{\substack{Q: Q_n \neq P_n\\ Q_\ell = v}} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right)$$
$$\leq \mathbb{P}(X^h = P) \sum_{i=0}^{\ell-1} \sum_{Q: Q_\ell = v} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right).$$

Observe that if $Q_\ell = v$, then $Q_{\ell+1}$ must equal $v + (0,1)$ or $v + (1,0)$. In particular, for $i < \ell$, (7.4) shows

$$\begin{aligned}
&\sum_{Q: Q_\ell = v} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right)\\
&\quad \leq \mathbb{P}\left(X^h_{\ell+1} = v + (0,1) \text{ or } X^h_{\ell+1} = v + (1,0) | X^h[i] = P[i]\right)\\
&\quad \leq \mathbb{P}\left(X^h_{\ell+1} = v + (0,1) | X^h[i] = P[i]\right) + \mathbb{P}\left(X^h_{\ell+1} = v + (1,0) | X^h[i] = P[i]\right)\\
&\quad \leq \frac{2c_h}{(\ell + 1 - i) \cdot h(\ell + 1 - i)}.
\end{aligned}$$

So,

$$\sum_{i=0}^{\ell-1} \sum_{Q: Q_\ell = v} \mathbb{P}\left(X^h = Q | X^h[i] = P[i]\right) \leq \sum_{i=0}^{\ell-1} \frac{2c_h}{(\ell + 1 - i) \cdot h(\ell + 1 - i)}$$
$$\leq 2c_h S_h(n).$$

□

We are now in a position to prove Theorem 7.18.

Proof of Theorem 7.18. Let $P \in \mathrm{F}_n$ and let $v \in V_n$, with

$$\|v\|_1 = \ell \text{ and } P_\ell \neq v.$$

Note that $P(v) = \ell$. So, if $Q \in \mathrm{F}_n$ is such that $Q(v) \neq P(v)$, then necessarily $Q_\ell = v$. We now set $P' \in \mathrm{F}_n$, with $P(v) \neq P'(v)$. In this case, the previous two lemmas show

$$\begin{aligned}
&\max\left(\frac{T(w_h, P)}{T(w_h, P, v)}, \frac{T(w_h, P')}{T(w_h, P', v)}\right)\\
&\quad \geq \frac{T(w_h, P)}{T(w_h, P, v)} = \frac{\sum_{Q\in \mathrm{F}_n} w_h(P,Q)}{\sum_{\substack{Q\in \mathrm{F}_n\\ Q(v)\neq P(v)}} w_h(P,Q)} = \frac{\sum_{Q\in \mathrm{F}_n} w_h(P,Q)}{\sum_{\substack{Q\in \mathrm{F}_n\\ Q_\ell = v}} w_h(P,Q)} \geq \frac{n - c_h S_h(n)}{2c_h S_h(n)}.
\end{aligned}$$

Since we are trying to establish a lower bound, we might as well assume that $S_h(n) = o(n)$. So, for n large enough

$$\frac{n - c_h S_h(n)}{2c_h S_h(n)} \geq \frac{n}{4c_h S_h(n)}.$$

Plugging this estimate into Theorem 7.19 yields the desired result □

7.6.3 Heuristic extension to higher dimensions

In this section we propose a heuristic approach to extend the lower bound to higher dimensions. In the 2 dimensional case, the proof method of Section 7.6 consisted of two steps: first reduces the problem to the discrete setting of monotone paths in $[n]^2$, and then analyze the query complexity of finding the minimal point for such path functions. Thus, to extend the result we should consider path functions on the d-dimensional grid, as well as a way to build smooth functions on $[0,1]^d$ from those paths.

The lower bound for finding minimal points of path functions in high-dimensional grids was obtained in [253], where it was shown that, in the worst case, any randomized algorithm must make $\Omega\left(n^{\frac{d}{2}}\right)$ queries in order to find the end point of a path defined over $[n]^d$. Thus, if we can find a discretization scheme, analogous to Lemma 7.16, in higher dimensions, we could obtain a lower bound for finding ε-stationary points. What are the constraints on such a discretization?

First note that necessarily the construction of [253] must be based on paths of lengths $\Omega\left(n^{\frac{d}{2}}\right)$, for otherwise one could simply trace the path to find its endpoint. In particular, since each cube has edge length $\frac{1}{n}$, an analogous construction to Lemma 7.16 will reach value smaller than $-\varepsilon \cdot n^{\frac{d}{2}-1}$ at the stationary point (i.e., the endpoint of the path). On the other hand, in at least one of the neighboring cubes (which are at distance less than $1/n$ from the stationary point), the background linear function should prevail, meaning that the function should reach a positive value. Since around the stationary point the function is quadratic, we get the constraint:

$$-\varepsilon \cdot n^{\frac{d}{2}-1} + \left(\frac{1}{n}\right)^2 > 0 \Leftrightarrow n < \left(\frac{1}{\varepsilon}\right)^{\frac{2}{d+2}}.$$

In particular the lower bound $\Omega\left(n^{\frac{d}{2}}\right)$ now suggests that for finding stationary point one has the complexity lower bound $\left(\frac{1}{\varepsilon}\right)^{\frac{d}{d+2}}$.

7.7 Discussion

We introduced a near-optimal algorithm for finding ε-stationary points in dimension 2. Finding a near-optimal algorithm in dimensions $d \geq 3$ remains open. Specific challenges include:

1. Finding a strategy in dimension 3 which improves upon GFT's $\tilde{O}(1/\varepsilon)$ complexity.

2. The heuristic extension of the lower bound in Section 7.6.3 suggests $\Omega\left(\frac{1}{\varepsilon^{\frac{d}{d+2}}}\right)$ as a complexity lower bound for any dimension d (note in particular that the exponent tends to 1 as d tends to infinity). On the other hand, [66] proved that for $d = \Omega(1/\varepsilon^2)$, one has the complexity lower bound $\Omega(1/\varepsilon^2)$. How do we reconcile these two results? Specifically we raise the following question: Is there an algorithm with complexity C_d/ε for some constant C_d which depends only on d? (Note that C_d as small as $O(\sqrt{d})$ would remain consistent with [66].) Alternatively we might ask whether the [66] lower bound holds for much smaller dimensions, e.g. when $d = \Theta(\log(1/\varepsilon))$, are we in the $1/\varepsilon$ regime as suggested by the heuristic, or are we already in the high-dimensional $1/\varepsilon^2$ of [66]?

3. Especially intriguing is the limit of low-depth algorithms, say as defined by having depth smaller than $\mathrm{poly}(d\log(1/\varepsilon))$. Currently this class of algorithms suffers from the curse of dimensionality, as GFT's total work degrades significantly when the dimension increases (recall from Theorem 7.2 that it is $\tilde{O}\left(\frac{1}{\varepsilon^{\frac{d-1}{2}}}\right)$). Is this necessary? A much weaker question is to simply show a separation between low-depth and high-depth algorithms. Namely can one show a lower bound $\Omega(1/\varepsilon^c)$ with $c > 2$ for low-depth algorithms? We note that lower bounds on depth have been investigated in the convex setting, see [191], [59].

4. A technically challenging problem is to adapt the construction in [Section 3, [237]] to non-monotone paths in higher dimensions. In particular, to formalize the heuristic argument from Section 7.6.3, such construction should presumably avoid creating saddle points.

Many more questions remain open on how to exploit the low-dimensional geometry of smooth gradient fields, and the above four questions are only a subset of the fundamental questions that we would like to answer. Other interesting questions include closing the logarithmic gap in dimension 2, or understanding better the role of randomness for this problem (note that GFT is deterministic, but other type of strategies include randomness, such as Hinder's non-convex cutting plane [136]).

8

Memorization with Two-Layers Neural Networks

8.1 Introduction

We study two-layers neural networks in $\mathbb{R}^d$ with k neurons and non-linearity $\psi : \mathbb{R} \to \mathbb{R}$. These are functions of the form:

$$x \mapsto \sum_{\ell=1}^{k} a_\ell \psi(w_\ell \cdot x + b_\ell)\,, \tag{8.1}$$

with $a_\ell, b_\ell \in \mathbb{R}$ and $w_\ell \in \mathbb{R}^d$ for any $\ell \in [k]$. We are mostly concerned with the Rectified Linear Unit non-linearity, namely $\mathrm{ReLU}(t) = \max(0, t)$, in which case wlog one can restrict the recombination weights (a_ℓ) to be in $\{-1, 1\}$ (this holds more generally for positively homogeneous non-linearities). We denote by $\mathcal{F}_k(\psi)$ the set of functions of the form (8.1). Under mild conditions on ψ (namely that it is not a polynomial), such neural networks are *universal*, in the sense that for k large enough they can approximate any continuous function [88, 166].

In this chapter we are interested in approximating a target function on a *finite* data set. This is also called the *memorization* problem. Specifically, fix a data set $(x_i, y_i)_{i\in[n]} \in (\mathbb{R}^d \times \mathbb{R})^n$ and an approximation error $\varepsilon > 0$. We denote $\mathbf{y} = (y_1, \ldots, y_n)$, and for a function $f : \mathbb{R}^d \to \mathbb{R}$ we write $\mathbf{f} = (f(x_1), \ldots, f(x_n))$. The main question concerning the memorization capabilities of

$\mathcal{F}_k(\psi)$ is as follows: How large should be k so that there exists $f \in \mathcal{F}_k(\psi)$ such that $\|\mathbf{f}-\mathbf{y}\|^2 \leq \varepsilon\|\mathbf{y}\|^2$ (where $\|\cdot\|$ denotes the Euclidean norm)? A simple consequence of universality of neural networks is that $k \geq n$ is sufficient (see Proposition 8.4). In fact (as was already observed in [29] for threshold ψ and binary labels, see Proposition 8.5) much more compact representations can be achieved by leveraging the high-dimensionality of the data. Namely we prove that for $\psi = \mathrm{ReLU}$ and a data set in general position (i.e., any hyperplane contains at most d points), one only needs $k \geq 4 \cdot \lceil \frac{n}{d} \rceil$ to memorize the data perfectly, see Proposition 8.6. The size $k \approx n/d$ is clearly optimal, by a simple parameter counting argument. We call the construction given in Proposition 8.6 a *Baum network*, and as we shall see it is of a certain combinatorial flavor. In addition we also prove that such memorization can in fact essentially be achieved in a kernel regime (with a bit more assumptions on the data): we prove in Theorem 8.8 that for $k = \Omega\left(\frac{n}{d}\log(1/\varepsilon)\right)$ one can obtain approximate memorization with the Neural Tangent Kernel [139], and we call the corresponding construction the *NTK network*. Specifically, the kernel we consider is,

$$\mathbb{E}\left[\nabla_w \psi(w \cdot x) \cdot \nabla_w \psi(w \cdot y)\right] = \mathbb{E}\left[(x \cdot y)\psi'(w \cdot x)\psi'(w \cdot y)\right],$$

where ∇_w is the gradient with respect to the w variable and the expectation is taken over a random initialization of w.

Measuring regularity via total weight. One is often interested in fitting the data using functions which satisfy certain regularity properties. The main notion of regularity in which we are interested is the *total weight*, defined as follows: For a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form (8.1), we define

$$\mathbf{W}(f) := \sum_{\ell=1}^{k} |a_\ell| \sqrt{\|w_\ell\|^2 + b_\ell^2}.$$

This definition is widely used in the literature, see Section 8.2 for a discussion and references. Notably, it was shown in [26] that this measure of complexity is better associated with the network's generalization ability compared to the size of the network. We will be interested in constructions which have both a small number of neurons and a small total weight.

Our main contribution: The complex network. As we will see below, both the Baum network and the NTK networks have sub-optimal total weight. The main technical contribution in this chapter is a third type of construction, which we call the *harmonic network*, that under the same assumptions on the data as for the NTK network, has both near-optimal memorization size and near-optimal total weight:

Theorem 8.1. *(Informal). Suppose that $n \leq \mathrm{poly}(d)$. Let $x_1, .., x_N \in \mathbb{S}^{d-1}$ such that*

$$|x_i \cdot x_j| = \tilde{O}\left(\frac{1}{\sqrt{d}}\right).$$

For every $\varepsilon > 0$ and every choice of labels $(y_i)_{i=1}^n$ such that $|y_i| = O(1)$ for all i, there exist $k = \tilde{O}\left(\frac{n}{d\varepsilon}\right)$ and $f \in \mathcal{F}_k(\psi)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \min\left(\left(y_i - f(x_i)\right)^2, 1\right) \leq \varepsilon$$

and such that $\mathbf{W}(f) = \tilde{O}\left(\sqrt{n}\right)$.

We show below in Proposition 8.3 that for random data one necessarily has $\mathbf{W}(f) = \tilde{\Omega}\left(\sqrt{n}\right)$, thus proving that the harmonic network has near-optimal total weight. Moreover we also argue in the corresponding sections that the Baum and NTK networks have total weight at least $n\sqrt{n}$ on random data, thus being far from optimal.

An iterative construction. Both the NTK network and the harmonic network will be built by iteratively adding up small numbers of neurons. This procedure, akin to boosting, is justified by the following lemma. It shows that to build a large memorizing network it suffices to be able to build a small network f whose scalar product with the data $\mathbf{f} \cdot \mathbf{y}$ is comparable to its variance $\|\mathbf{f}\|^2$:

Lemma 8.2. *Fix $(x_i)_{i=1}^n$. Suppose that there are $m \in \mathbb{N}$ and $\alpha, \beta > 0$ such that the following holds: For any choice of $(y_i)_{i=1}^n$, there exists $f \in \mathcal{F}_m(\psi)$ with $\mathbf{y} \cdot \mathbf{f} \geq \alpha\|\mathbf{y}\|^2$ and $\|\mathbf{f}\|^2 \leq \beta\|\mathbf{y}\|^2$. Then for all $\varepsilon > 0$, there exists $g \in \mathcal{F}_{mk}(\psi)$ such that*

$$\|\mathbf{g} - \mathbf{y}\|^2 \leq \varepsilon\|\mathbf{y}\|^2$$

with

$$k \leq \frac{\beta}{\alpha^2}\log(1/\varepsilon).$$

Moreover, if the above holds with $\mathbf{W}(f) \leq \omega$, then $\mathbf{W}(g) \leq \frac{\omega}{\alpha}\log(1/\varepsilon)$.

Proof. Denote $\eta = \frac{\alpha}{\beta}$ and $\mathbf{r}_1 = \mathbf{y}$. Then, there exists $f_1 \in \mathcal{F}_m(\psi)$, such that

$$\|\eta\mathbf{f}_1 - \mathbf{r}_1\|^2 = \|\mathbf{r}_1\|^2 - 2\eta\mathbf{y}\cdot\mathbf{f}_1 + \eta^2\|\mathbf{f}_1\|^2 \leq \|\mathbf{r}_1\|^2\left(1 - 2\frac{\alpha^2}{\beta} + \frac{\alpha^2}{\beta}\right)$$
$$\leq \|\mathbf{r}_1\|^2\left(1 - \frac{\alpha^2}{\beta}\right) = \|\mathbf{y}\|^2\left(1 - \frac{\alpha^2}{\beta}\right)$$

The result is obtained by iterating the above inequality with $\mathbf{r}_i = \mathbf{y} - \eta\sum_{j=1}^{i-1}\mathbf{f}_j$ taken as the

residuals. By induction, if we set $g = \eta \sum_{j=1}^{k} f_j$, we get

$$\|\mathbf{g} - \mathbf{y}\| = \|\eta \mathbf{f}_k - \mathbf{r}_k\| \leq \|\mathbf{r}_k\|^2 \left(1 - \frac{\alpha^2}{\beta}\right) = \|\mathbf{y}\|^2 \left(1 - \frac{\alpha^2}{\beta}\right)^k .$$

□

In both the NTK and harmonic constructions, the function f will have the largest possible correlation with the data set attainable for a network of constant size. However, the harmonic network will have the extra advantage that the function f will be composed of a single neuron whose weight is the smallest one attainable. Thus, the harmonic network will enjoy both the smallest possible number of neurons and smallest possible total weight (up to logarithmic factors). Note however that the dependency on ε is worse for the harmonic network, which is technically due to a constant order term in the variance which we do not know how to remove.

We conclude the introduction by showing that a total weight of $\Omega(\sqrt{n})$ is necessary for approximate memorization. Just like for the upper bound, it turns out that it is sufficient to consider how well can one correlate a single neuron. Namely the proof boils down to showing that a single neuron cannot correlate well with random data sets.

Proposition 8.3. *There exists a data set $(x_i, y_i)_{i\in[n]} \in (\mathbb{S}^{d-1} \times \{-1, 1\})^n$ such that for every function f of the form (8.1) with ψ L-Lipschitz and which satisfies $\|\mathbf{f} - \mathbf{y}\|^2 \leq \frac{1}{2}\|\mathbf{y}\|^2$, it holds that $\mathbf{W}(f) \geq \frac{\sqrt{n}}{8L}$.*

Proof. We have

$$\frac{1}{2}\|\mathbf{y}\|^2 \geq \|\mathbf{f} - \mathbf{y}\|^2 \geq \|\mathbf{y}\|^2 - 2\mathbf{f} \cdot \mathbf{y} \Rightarrow \mathbf{f} \cdot \mathbf{y} \geq \frac{1}{4}\|\mathbf{y}\|^2 ,$$

that is

$$\sum_{\ell=1}^{k} \sum_{i=1}^{n} y_i a_\ell \psi(w_\ell \cdot x_i - b_\ell) \geq \frac{n}{4} ,$$

which implies:

$$\max_{w,b} \sum_{i=1}^{n} y_i \frac{\psi(w \cdot x_i - b)}{\sqrt{\|w\|^2 + b^2}} \geq \frac{n}{4\mathbf{W}(f)} .$$

Now let us assume that y_i are ± 1 uniformly at random (i.e., Rademacher random variables), and thus by Talagrand's contraction lemma for the Rademacher complexity (see [Lemma 26.9, [218]]) we have:

$$\begin{aligned}
\mathbb{E} \max_{w,b} \sum_{i=1}^{n} y_i \frac{\psi(w \cdot x_i - b)}{\sqrt{\|w\|^2 + b^2}} &\leq L \cdot \mathbb{E} \max_{w,b} \sum_{i=1}^{n} y_i \frac{w \cdot x_i - b}{\sqrt{\|w\|^2 + b^2}} \\
&\leq L \cdot \mathbb{E} \sqrt{\left\| \sum_{i=1}^{n} y_i x_i \right\|^2 + n} \leq 2L\sqrt{n} ,
\end{aligned}$$

and thus $\mathbf{W}(f) \geq \frac{\sqrt{n}}{8L}$. □

8.2 Related works

Exact memorization. The observation that n neurons are sufficient for memorization with essentially arbitrary non-linearity was already made in [18] (using Carathéodory's theorem), and before that a slightly weaker bound with $n+1$ neurons was already observed in [30] (or more recently $2n+d$ in [252]). The contribution of Proposition 8.4 is to show that this statement of exactly n neurons follows in fact from elementary linear algebra.

As already mentioned above, [29] proved that for threshold non-linearity and binary labels one can obtain a much better bound of n/d neurons for memorization, as long as the data is in general position. This was generalized to the ReLU non-linearity (but still binary labels) in [250] (we note that this paper also considers some questions around memorization capabilities of deeper networks). Our modest contribution here is to generalize this to arbitrary real labels, see Proposition 8.6.

Gradient-based memorization. A different line of works on memorization studies whether it can be achieved via gradient-based optimization on various neural network architectures. The literature here is very large, but early results with minimal assumptions include [168,221] which were notably generalized in [5,96]. Crucially these works leverage very large overparametrization, i.e., the number of neurons is a large polynomial in the number of data points. For a critique of this large overparametrization regime see [76,125,248], and for a different approach based on a certain scaling limit of stochastic gradient descent for sufficiently overparametrized networks see [74,175]. More recently the amount of overparametrization needed was improved to a small polynomial dependency in n and d in [149,201,222]. In the *random features* regime, [54] have also considered an iterative construction procedure for memorization. This is somewhat different than our approach, in which the iterative procedure updates the w_j's, and a much smaller number of neurons is needed as a result. Finally, very recently Amit Daniely [89,90] showed that gradient descent already works in the optimal regime of $k = \tilde{O}(n/d)$, at least for random data (and random labels). This result is closely related to our analysis of the NTK network in Section 8.4. Minor distinctions are that we allow for arbitrary labels, and we take a "boosting approach" were neurons are added one by one (although we do not believe that this is an essential difference).

Total weight complexity. It is well-known since [26] that the total weight of a two-layers neural network is a finer measure of complexity than the number of neurons to control its generalization (see [193] and [15] for more recent discussions on this, as well as [27] for other notions of norms for deeper networks). Of course the bound $\mathbf{W} = \tilde{O}(\sqrt{n})$ proved here leads

to vacuous generalization performance, as is necessary since the Harmonic network can memorize completely random data (for which no generalization is possible). It would be interesting to see if the weight of the Harmonic network can be smaller for more structured data, particularly given the context raised by the work [252] (where it was observed that SGD on deep networks will memorize arbitrary data, hence the question of where does the seeming generalization capabilities of those networks come from). We note the recent work [141] which proves for example that polylogarithmic size network is possible for memorization under a certain margin condition. Finally we also note that the effect in function space of bounding $\mathbf{W}$ has been recently studied in [200, 214].

Complex weights. It is quite natural to consider neural networks with complex weights. Indeed, as was already observed by Barron [25], the Fourier transform $f(x) = \int \hat{f}(\omega) \exp(i\omega \cdot x) d\omega$ exactly gives a representation of f as a two-layers neural network with the non-linearity $\psi(t) = \exp(it)$. More recently, it was noted in [10] that randomly perturbing a neuron with *complex weights* is potentially more beneficial than doing a mere real perturbation. We make a similar observation in Section 8.5 for the construction of the Harmonic network, where we show that complex perturbations allow to deal particularly easily with higher order terms in some key Taylor expansion. Moreover we also note that [10] considers non-linearity built from Hermite polynomials, which shall be a key step for us too in the construction of the Harmonic network (the use of Hermite polynomials in the context of learning theory goes back to [146]).

While orthogonal to our considerations here, we also note the work of Fefferman [115], where he used the analytical continuation of a (real) neural network to prove a certain uniqueness property (essentially that two networks with the same output must have the same weights up to some obvious symmetries and obvious counter-examples).

8.3 Elementary results on memorization

In this section we give a few examples of elementary conditions on k, ψ and the data set so that one can find $f \in \mathcal{F}_k(\psi)$ with $\mathbf{f} = \mathbf{y}$ (i.e., exact memorization). We prove three results: (i) $k \geq n$ suffices for any non-polynomial ψ, (ii) $k \geq \frac{n}{d} + 3$ with $\psi(t) = \mathbb{1}\{t \geq 0\}$ suffices for binary labels with data in general position (this is exactly [29]'s result), and (iii) $k \geq 4 \cdot \lceil \frac{n}{d} \rceil$ with $\psi = \mathrm{ReLU}$ suffices for data in general position and arbitrary labels.

We start with the basic linear algebraic observation that having a number of neurons larger than the size of the data set is always sufficient for perfect memorization:

Proposition 8.4. *Assuming that ψ is not a polynomial, there exists $f \in \mathcal{F}_n(\psi)$ such that $\mathbf{f} = \mathbf{y}$.*

Proof. Note that the set of functions of the form (8.1) (with arbitrary k) corresponds to the vector space V spanned by the functions $\psi_{w,b} : x \mapsto \psi(w \cdot x + b)$. Consider the linear operator Ψ :

$V \to \mathbb{R}^n$ that corresponds to the evaluation on the data points (x_i) (i.e., $\Psi(f) = (f(x_i))_{i\in[n]}$). Since ψ is not a polynomial, the image of Ψ is $\mathrm{Im}(\Psi) = \mathbb{R}^n$. Moreover $\mathrm{Im}(\Psi)$ is spanned by the set of vectors $\Psi(\psi_{w,b})$ for $w \in \mathbb{R}^d, b \in \mathbb{R}$. Now, since $\dim(\mathrm{Im}(\Psi)) = n$, one can extract a subset of n such vectors with the same span, that is there exists $w_1, b_1, \ldots, w_n, b_n$ such that

$$\mathrm{span}(\Psi(\psi_{w_1,b_1}), \ldots, \Psi(\psi_{w_n,b_n})) = \mathbb{R}^n \,,$$

which concludes the proof. □

In [29] it is observed that one can dramatically reduce the number of neurons for high-dimensional data:

Proposition 8.5. *Fix $\psi(t) = \mathbb{1}\{t \geq 0\}$. Let $(x_i)_{i\in[n]}$ be in general position in $\mathbb{R}^d$ (i.e., any hyperplane contains at most d points), and assume binary labels, i.e., $y_i \in \{0, 1\}$. Then there exists $f \in \mathcal{F}_{\frac{n}{d}+3}(\psi)$ such that $\mathbf{f} = \mathbf{y}$.*

Proof. [29] builds a network iteratively as follows. Pick d points with label 1, say $x_1, \ldots, x_d$, and let $H = \{x : u \cdot x = b\}$ be a hyperplane containing those points and no other points in the data, i.e., $x_i \notin H$ for any $i > d$. With two neurons (i.e., $f \in \mathcal{F}_2(\psi)$) one can build the indicator of a small neighborhood of H, namely $f(x) = \psi(u \cdot x - (b - \tau)) - \psi(u \cdot x - (b + \tau))$ with τ small enough, so that $f(x_i) = 1$ for $i \leq d$ and $f(x_i) = 0$ for $i > d$. Assuming that the label 1 is the minority (which is without loss of generality up to one additional neuron), one thus needs at most $2\lceil \frac{n}{2d} \rceil$ neurons to perfectly memorize the data. □

We now extend Proposition 8.5 to the ReLU non-linearity and arbitrary real labels. To do so we introduce the *derivative neuron* of ψ defined by:

$$f_{\delta,u,v,b} : x \mapsto \frac{\psi((u + \delta v) \cdot x - b) - \psi(u \cdot x - b)}{\delta} \,, \tag{8.2}$$

with $\delta \in \mathbb{R}$ and $u, v \in \mathbb{R}^d$. As δ tends to 0, this function is equal to

$$f_{u,v,b}(x) = \psi'(u \cdot x - b)v \cdot x \tag{8.3}$$

for any x such that ψ is differentiable at $u \cdot x - b$. In fact, for the ReLU one has for any x such that $u \cdot x \neq b$ that $f_{\delta,u,v,b}(x) = f_{u,v,b}(x)$ for δ small enough (this is because the ReLU is piecewise linear). We will always take δ small enough and u such that $f_{\delta,u,v,b}(x_i) = f_{u,v,b}(x_i)$ for any $i \in [n]$, for example by taking

$$\delta = \frac{1}{2} \min_{i\in[n]} \frac{|u \cdot x_i - b|}{|v \cdot x_i|} \,. \tag{8.4}$$

Thus, as far as memorization is concerned, we can assume that $f_{u,v,b} \in \mathcal{F}_2(\mathrm{ReLU})$. With this observation it is now trivial to prove the following extension of Baum's result:

Proposition 8.6. *Let $(x_i)_{i\in[n]}$ be in general position in $\mathbb{R}^d$ (i.e., any hyperplane contains at most d points). Then there exists $f \in \mathcal{F}_{4\cdot\lceil \frac{n}{d} \rceil}(\mathrm{ReLU})$ such that $\mathbf{f} = \mathbf{y}$.*

Proof. Pick an arbitrary set of d points, say $(x_i)_{i\le d}$, and let $H = \{x : u\cdot x = b\}$ be a hyperplane containing those points and no other points in the data, i.e., $x_i \notin H$ for any $i > d$. With four neurons one can build the function $f = f_{u,v,b-\tau} - f_{u,v,b+\tau}$ with τ small enough so that $f(x_i) = x_i \cdot v$ for $i \le d$ and $f(x_i) = 0$ for $i > d$. It only remains to pick v such that $v \cdot x_i = y_i$ for any $i \le d$, which we can do since the matrix given by $(x_i)_{i\le d}$ is full rank (by the general position assumption). □

Let us now sketch the calculation of this network's total weight in the case that the x_i's are independent uniform points on $\mathbb{S}^{d-1}$ and y_i are ± 1-Bernoulli distributed. We will show that the total weight is at least $n^2/\sqrt{d}$, thus more than n times the optimal attainable weight given in Proposition 8.3.

Consider the matrix X whose rows are the vectors $(x_i)_{i\le d}$. The vector v taken in the neuron corresponding to those points solves the equation $Xv = y$ and since the distribution of X is absolutely continuous, we have that X is invertible almost surely and therefore $v = X^{-1}y$, implying that $|v| \ge \|X\|_{OP}^{-1}\sqrt{d}$. It is well-known (and easy to show) that with overwhelming probability, $\|X\|_{\mathrm{OP}} = O(1)$, and thus $\|v\| = \Omega(\sqrt{d})$.

Observe that by normalizing the parameter δ accordingly, we can assume that $\|u\| = 1$. By definition we have $u\cdot x_i = b$ for all $i = 1, \dots, d$. A calculation shows that with probability $\Omega(1)$ we have $b = \Theta(1/\sqrt{d})$.

Next, we claim that $|v \cdot u| \le (1-\rho)\|v\|$ for some $\rho = \Omega(1)$. Indeed, suppose otherwise. Denote $c = \frac{1}{d}\sum_{i\in[d]} x_i$. It is easy to check that with high probability, $\|c\| = O\left(\frac{1}{\sqrt{d}}\right)$. Note that $v \cdot c = \frac{1}{d}\sum_{i\in[d]} y_i = O(1/\sqrt{d})$. This implies that

$$b(|v\cdot u| - O(1)) \le |v\cdot(bu-c)| \le \sqrt{\|v\|^2 - (v\cdot u)^2}\|bu - c\| \le \sqrt{2\rho}\frac{\|v\|}{\sqrt{d}},$$

where we used the fact that $(bu-c) \perp (v\cdot u)u$. Thus we have

$$\Omega(1-2\rho) = b(1-2\rho)\|v\| = O(\sqrt{\rho}).$$

leading to a contradiction. To summarize, we have $\|v\| = \Omega(\sqrt{d})$, $\|u\| = 1$, $|u\cdot v| \le (1-\rho)\|v\|$, $\rho = \Omega(1)$, and $b = O(1/\sqrt{d})$. Since spherical marginals are approximately Gaussian, if x is uniform in $\mathbb{S}^{d-1}$ we have that the joint distribution of $(x\cdot u, x\cdot v)$ conditional on v and u is approximately $\mathcal{N}\left(0, \frac{1}{d}\begin{pmatrix} 1 & (1-\rho)\beta \\ (1-\rho)\beta & \beta \end{pmatrix}\right)$ with $\rho = \Omega(1)$ and $\beta = \Theta(d)$. Therefore, with probability $\Omega(1/n)$ we have $|x\cdot v| = \Omega(1)$ and $|x\cdot u - b| = O(1/(n\sqrt{d}))$.

We conclude that

$$\mathbb{P}\left(\exists i \geq d+1 \text{ s.t. } \frac{|x_i \cdot u - b|}{|x_i \cdot v|} = O\left(\frac{1}{n\sqrt{d}}\right) \middle| x_1, ..., x_d\right) = \Omega(1).$$

Therefore, we get $\delta = O(1/n\sqrt{d})$ which implies that the weight of the neuron is of order at least $\frac{\|u\|}{\delta} = \Omega(n\sqrt{d})$. This happens with probability $\Omega(1)$ for every one of the first $n/(2d)$ neurons, implying that the total weight is of order $n^2/\sqrt{d}$.

8.4 The NTK network

The constructions in Section 8.3 are based on a very careful set of weights that depend on the entire dataset. Here we show that essentially the same results can be obtained in the *neural tangent kernel* regime. That is, we take pair of neurons as given in (8.2) (which corresponds in fact to (8.3) since we will take δ to be small, we will also restrict to $b = 0$), and crucially we will also have that the "main weight" u will be chosen at random from a standard Gaussian, and only the "small perturbation" v will be chosen as a function of the dataset. The guarantee we obtain is slightly weaker than in Proposition 8.6: we have a $\log(1/\varepsilon)$ overhead in the number of neurons, and moreover we also need to assume that the data is "well-spread". Specifically we consider the following notion of "generic data":

Definition 8.7. We say that $(x_i)_{i\in[n]}$ are (γ, ω)-generic (with $\gamma \in (\frac{1}{2n}, 1)$ and $\omega > 0$) if:

- $\|x_i\| \geq 1$ for all $i \in [n]$,
- $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \preceq \frac{\omega}{d} \cdot \mathrm{I}_d$,
- and $|x_i \cdot x_j| \leq \gamma \cdot \|x_i\| \cdot \|x_j\|$ for all $i \neq j$.

In the following we fix such a (γ, ω)-generic data set. Note that i.i.d. points on the sphere are $\left(O\left(\sqrt{\frac{\log(n)}{d}}\right), O(1)\right)$-generic. We now formulate our main theorem concerning the NTK network.

Theorem 8.8. *There exists $f \in \mathcal{F}_k(\mathrm{ReLU})$, produced in the NTK regime (see Theorem 8.9 below for more details) with $\mathbb{E}[\|\mathbf{f} - \mathbf{y}\|^2] \leq \varepsilon\|\mathbf{y}\|^2$ (the expectation is over the random initialization of the "main weights") provided that*

$$k \cdot d \geq 20\omega \cdot n \log(1/\varepsilon) \cdot \frac{\log(2n)}{\log(1/\gamma)}\,. \tag{8.5}$$

In light of Lemma 8.2, it will be enough to produce a width-2 network, $f \in \mathcal{F}_2(\mathrm{ReLU})$, whose correlation with the data set is large.

Theorem 8.9. *There exists $f \in \mathcal{F}_2(\mathrm{ReLU})$ with*

$$\mathbf{y} \cdot \mathbf{f} \geq \frac{1}{10} \cdot \sqrt{\frac{\log(1/\gamma)}{\log(2n)}} \cdot \|\mathbf{y}\|^2 \,, \tag{8.6}$$

and

$$\|\mathbf{f}\|^2 \leq \frac{\omega \cdot n}{d} \|\mathbf{y}\|^2 \,. \tag{8.7}$$

In fact, one can take the construction (8.2) with:

$$u \sim \mathcal{N}(0, \mathrm{I}_d), \;\; v = \sum_{i : u \cdot x_i \geq 0} y_i x_i, \;\; \delta = \frac{1}{2} \frac{\min_{i \in [n]} |u \cdot x_i|}{|v \cdot x_i|}. \tag{8.8}$$

which produces $f \in \mathcal{F}_2(\mathrm{ReLU})$ such that (8.6) holds in expectation and (8.7) holds almost surely.

To deduce Theorem 8.8 from Theorem 8.9, apply Lemma 8.2 with $\alpha = \frac{1}{10} \cdot \sqrt{\frac{\log(1/\gamma)}{\log(2n)}}$ and $\beta = \frac{\omega \cdot n}{d}$.

For $u \in \mathbb{R}^d$, set

$$f_u(x) = \psi'(u \cdot x) v \cdot x, \tag{8.9}$$

where v is defined as in (8.8). Observe that as long as $u \cdot x_i \neq 0, \forall i \in [n]$, a small enough choice of δ ensures the existence of $f \in \mathcal{F}_2(\mathrm{ReLU})$ such that $\mathbf{f} = \mathbf{f}_u$.

To prove Theorem 8.9, it therefore remains to show that $\mathbf{f}_u$ satisfies (8.6) and (8.7) with positive probability as $u \sim \mathcal{N}(0, \mathrm{I}_d)$. This will be carried out in two steps: First we show that the correlation $\mathbf{y} \cdot \mathbf{f}$ for a derivative neuron has a particularly nice form as a function of u, see Lemma 8.10. Then, in Lemma 8.11 we derive a lower bound for the expectation of the correlation under $u \sim \mathcal{N}(0, \mathrm{I}_d)$. Taken together these lemmas complete the proof of Theorem 8.9.

Lemma 8.10. *Fix $u \in \mathbb{R}^d$, the function f_u defined in (8.9) satisfies*

$$\sum_{i=1}^n y_i f_u(x_i) = \left\| \sum_{i : u \cdot x_i \geq 0} y_i x_i \right\|^2 , \tag{8.10}$$

and furthermore

$$\sum_{i=1}^n f_u(x_i)^2 \leq \frac{\omega \cdot n}{d} \cdot \sum_{i=1}^n y_i f(x_i) \,. \tag{8.11}$$

Proof. We may write

$$\sum_{i=1}^n f_u(x) y_i = \sum_{i=1}^n \psi'(u \cdot x_i) y_i x_i \cdot v \,.$$

To maximize this quantity we take $v = \sum_{i=1}^n \psi'(u \cdot x_i) y_i x_i$ so that the correlation is exactly equal to:

$$\|v\|^2 = \left\| \sum_{i=1}^n \psi'(u \cdot x_i) y_i x_i \right\|^2 , \tag{8.12}$$

which concludes the proof of (8.10) (note also that $\psi'(t) = \mathbb{1}\{t \geq 0\}$ for the ReLU). Moreover for (8.11) it suffices to also notice that (recall that for ReLU, $|\psi'(t)| \leq 1$)

$$\sum_{i=1}^n f_u(x_i)^2 = \sum_{i=1}^n (\psi'(x_i \cdot u))^2 (x_i \cdot v)^2 \leq \lambda_{\max} \left(\sum_{i=1}^n x_i x_i^\top \right) \cdot \|v\|^2 . \tag{8.13}$$

□

Lemma 8.11. *One has:*

$$\mathbb{E}_{u \sim \mathcal{N}(0, \mathrm{I}_n)} \left\| \sum_{i:\ u \cdot x_i \geq 0} y_i x_i \right\|^2 \geq \frac{1}{10} \cdot \sqrt{\frac{\log(1/\gamma)}{\log(2n)}} \cdot \sum_{i=1}^n y_i^2 \|x_i\|^2 .$$

Proof. First note that

$$\mathbb{E} \left\| \sum_{i:\ u \cdot x_i \geq 0} y_i x_i \right\|^2 = \mathbf{y}^\top H \mathbf{y} ,$$

where

$$H_{i,j} = \mathbb{E}[x_i \cdot x_j \mathbb{1}\{u \cdot x_i \geq 0\} \mathbb{1}\{u \cdot x_j \geq 0\}] = \frac{2}{\pi} x_i \cdot x_j \left(\frac{1}{4} + \arcsin \left(\frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|} \right) \right) .$$

Let us denote V the matrix with entries $V_{i,j} = \frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|}$ and D the diagonal matrix with entries $\|x_i\|$. Note that $V \succeq 0$ and thus we have (recall also that $\arcsin(t) = \sum_{i=0}^\infty \frac{(2i)!}{(2^i i!)^2} \cdot \frac{t^{2i+1}}{2i+1}$):

$$D^{-1} H D^{-1} \succeq \frac{2}{\pi} \sum_{i=0}^\infty \frac{(2i)!}{(2^i i!)^2} \cdot \frac{V^{\circ 2(i+1)}}{2i+1} .$$

Now observe that for any i, by the Schur product theorem one has $V^{\circ i} \succeq 0$. Moreover $V^{\circ i}$ is equal to 1 on the diagonal, and off-diagonal it is smaller than γ^i, and thus for $i \geq \frac{\log(2n)}{\log(1/\gamma)}$ one has $V^{\circ i} \succeq \frac{1}{2} \mathrm{I}_n$. In particular we obtain:

$$D^{-1} H D^{-1} \succeq \left(\frac{1}{\pi} \sum_{i \geq \frac{\log(2n)}{2 \log(1/\gamma)}}^\infty \frac{(2i)!}{(2^i i!)^2} \cdot \frac{1}{2i+1} \right) \mathrm{I}_n .$$

It is easy to verify that $\frac{(2i)!}{(2^i i!)^2} \geq \frac{1}{8 \cdot i^{3/2}}$, and moreover $\sum_{i \geq N} \frac{1}{i^{3/2}} \geq \frac{2}{\sqrt{N}}$, so that for $\gamma \in (\frac{1}{2n}, 1)$,

$$\frac{1}{\pi} \sum_{i \geq \frac{\log(2n)}{2\log(1/\gamma)}}^{\infty} \frac{(2i)!}{(2^i i!)^2} \cdot \frac{1}{2i+1} \geq \frac{1}{10} \cdot \sqrt{\frac{\log(1/\gamma)}{\log(2n)}},$$

which concludes the proof. □

We conclude the section by sketching the calculation of the total weight of this network. Recall that the neurons are of the form (8.9). According to (8.12) and Lemma 8.11, we have that for typical neurons, $\|v\| = \Omega(\sqrt{n})$. Moreover, with high probability we have $\|u\| = \Theta(\sqrt{d})$, and thus the weight of a single neuron is at least $\frac{\|u\|}{\delta} = \frac{\sqrt{d}}{\delta}$. Adding up the neurons, this shows that the total weight is of order $\frac{\sqrt{d}}{\delta}$ (since $k = \tilde{\Theta}(n/d)$ and the coefficient in front of the neurons is of order $\tilde{\Theta}(\frac{d}{n})$).

Now suppose that δ is taken according to (8.4). The main observation (we omit the details of proof) is that u and v have a mutual distribution of roughly independent Gaussian random vectors (without loss of generality we can assume that $\sum y_i = 0$ which implies $\mathbb{E} u \cdot v = 0$). In this case we have $\delta = \tilde{O}\left(\frac{\sqrt{d}}{n\sqrt{n}}\right)$. This implies a total weight of order at least $n\sqrt{n}$.

8.5 The complex network

We now wish to improve upon the NTK construction, by creating a network with similar memorization properties and which has almost no excess total weight. We will work under the assumptions that

$$\|x_i\| = 1 \text{ for every } i \in [n], \text{ and, } |x_i \cdot x_j| \leq \gamma \text{ for } i \neq j. \tag{8.14}$$

In light of Lemma 8.2, it is enough to find a single neuron whose scalar product with the data set is large. Thus, the rest of this section is devoted to proving the following theorem.

Theorem 8.12. *Assume that (8.14) holds, that m is large enough so that $n\gamma^{m-2} \leq \frac{1}{2}$ and that for all $i \in [n]$, $y_i^2 \leq n\gamma^2$ with $\|\mathbf{y}\|^2 \leq n$. Then, there exist $w \in \mathbb{R}^d$ and $b, \sigma \in \mathbb{R}$, with*

$$\|w\|^2, |b|^2 \leq C_m d \log(n)^m, |\sigma| = 1,$$

such that for

$$f(x) = \sigma \cdot \mathrm{ReLU}\big(w \cdot x + b\big),$$

we have

$$\mathbf{y} \cdot \mathbf{f} \geq \frac{c_m}{\log(n)^{m^2/2}} \frac{1}{\sqrt{n\gamma^2}} \|\mathbf{y}\|^2,$$

and

$$\|\mathbf{f}\|^2 \leq \frac{n}{c_m}\log(n)^m,$$

where $c_m, C_m > 0$ are constants which depends only on m.

By invoking an iterative procedure as in Lemma 8.2, we obtain our main estimate. As it turns out, our construction will give a good fit for almost all points. If $A \subset [n]$ and $v \in \mathbb{R}^n$ we denote below by v_A the projection of v unto the indices contained in A. With this notation our result is:

Theorem 8.13. *Assume that (8.14) holds, that m is large enough so that $n\gamma^{m-2} \leq \frac{1}{2}$ and that $\|\mathbf{y}\|^2 = n$. There exists $f \in \mathcal{F}_k(\mathrm{ReLU})$ and $A \subset [n]$, with*

$$k = \left\lceil C_m \gamma^2 \frac{\log(1/\varepsilon)}{\varepsilon} n \log(n)^{(m^2+m)} \right\rceil,$$

such that

$$\mathbb{E}[\|\mathbf{f}_A - \mathbf{y}_A\|^2] \leq \varepsilon \|\mathbf{y}\|^2, \quad |A| \geq n - \frac{1}{\gamma^2}, \tag{8.15}$$

and

$$\mathbf{W}(f) = \tilde{O}\left(\frac{\log(1/\varepsilon)}{\varepsilon}\sqrt{n\gamma^2 d}\right), \tag{8.16}$$

where C_m is a constant which depends only on m.

Observe that if $(x_i)_{i\in[n]}$ are uniformly distributed in the $\mathbb{S}^{d-1}$ then $\gamma = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ and we get that $\mathbf{W}(f) = \tilde{O}\left(\frac{\log(1/\varepsilon)}{\varepsilon}\sqrt{n}\right)$, which is optimal up to the logarithmic factors and the dependence on ε.

The proof of Theorem 8.13 follows an iterative procedure similar to the one carried out in Lemma 8.2. The only caveat is the condition $y_i^2 \leq n\gamma^2$ which appears in Theorem 8.12. Due to this condition we need to consider a slightly smaller set of indices at each iteration, ignoring ones where the residue becomes too big.

Proof of Theorem 8.13. We build the network iteratively. Set $f_0 \equiv 0$, $A_0 = [n]$ and $r_{0,i} = y_i$. Now, for $\ell \in \mathbb{N}$, suppose that there exists $f_\ell \in \mathcal{F}_\ell(\mathrm{ReLU})$ with

$$\|(\mathbf{f}_\ell)_{A_\ell} - \mathbf{y}_{A_\ell}\| \leq \left(1 - \frac{c_m^3}{\log(n)^{m^2+m}}\frac{\varepsilon}{n\gamma^2}\right)\|\mathbf{y}\|^2.$$

Set $r_{\ell,i} = y_i - f_\ell(x_i)$ and $A_\ell = \{i \in A_{\ell-1} | r_{\ell,i}^2 \leq n\gamma^2\}$. We now invoke Theorem 8.12 with the residuals $\{r_{\ell,i} | i \in A_\ell\}$ to obtain a neuron $f \in \mathcal{F}_1(\mathrm{ReLU})$, which satisfies

$$(\mathbf{r}_\ell)_{A_\ell} \cdot \mathbf{f} \geq \frac{c_m}{\log(n)^{m^2/2}}\frac{1}{\sqrt{n\gamma^2}}\|\mathbf{r}_\ell\|^2,$$

and

$$\|\mathbf{f}_{A_\ell}\|^2 \leq \frac{n}{c_m}\log(n)^m.$$

Since we may assume $\|(\mathbf{r}_\ell)_{A_\ell}\|^2 \geq n\varepsilon$ (otherwise we are done), the second condition can be rewritten as

$$\|\mathbf{f}_{A_\ell}\|^2 \leq \frac{\log(n)^m}{c_m\varepsilon}\|(\mathbf{r}_\ell)_{A_\ell}\|^2.$$

In this case the calculation done in Lemma 8.2 with $\alpha = \frac{c_m}{\log(n)^{m^2/2}}\frac{1}{\sqrt{n\gamma^2}}$ and $\beta = \frac{\log(n)^m}{c_m\varepsilon}$ shows that for $\eta := \frac{c_m^2\varepsilon}{\log(n)^{m^2/2+m}}$, one has

$$\|\eta\mathbf{f}_{A_\ell} - (\mathbf{r}_\ell)_{A_\ell}\|^2 \leq \left(1 - \frac{c_m^3}{\log(n)^{m^2+m}}\frac{\varepsilon}{n\gamma^2}\right)\|(\mathbf{r}_\ell)_{A_\ell}\|^2.$$

In other words, if we define $f_{\ell+1} \in \mathcal{F}_{\ell+1}(\mathrm{ReLU})$ by $f_{\ell+1} = f_\ell + \eta f$,

$$\|(\mathbf{f}_{\ell+1})_{A_\ell} - \mathbf{y}_{A_\ell}\|^2 \leq \left(1 - \frac{c_m^3}{\log(n)^{m^2+m}}\frac{\varepsilon}{n\gamma^2}\right)^{\ell+1}\|\mathbf{y}\|^2.$$

The estimate (8.15) is now obtained with the appropriate choice of k. Let us also remark that for any ℓ,

$$\|(\mathbf{r}_{\ell+1})_{A_\ell}\|^2 \leq \|(\mathbf{r}_\ell)_{A_\ell}\|^2 \leq \|(\mathbf{r}_\ell)_{A_{\ell-1}}\|^2 - n\gamma^2|A_{\ell-1}\setminus A_\ell|.$$

By induction

$$\|(\mathbf{r}_{\ell+1})_{A_\ell}\|^2 \leq \|\mathbf{y}\|^2 - n\gamma^2\left(n - |A_\ell|\right)$$

This shows that $|A_\ell| \geq n - \frac{1}{\gamma^2}$. The bound on $\mathbf{W}(f_k)$ a direct consequence of Lemma 8.2. □

8.5.1 Correlation of a perturbed neuron with random sign

Towards understanding our construction, let us first revisit the task of correlating a *single* neuron with the data, namely we want to maximize over w the ratio between $|\sum_{i=1}^n y_i\psi(w\cdot x_i)|$ and $\sqrt{\sum_{i=1}^n \psi(w\cdot x_i)^2}$. Note that depending on whether the sign of the correlation is positive or negative, one would eventually take either neuron $x \mapsto \psi(w\cdot x)$ or $x \mapsto -\psi(w\cdot x)$. Let us first revisit the NTK calculation from the previous section, emphasizing that one can take a random sign for the recombination weight a.

The key NTK-like observation is that a single neuron perturbed around the parameter w_0 and with random sign can be interpreted as a linear model over a feature mapping that depends on w. More precisely (note that the random sign cancels the 0^{th} order term in the Taylor expansion):

$$\mathbb{E}_{a\sim\{-\delta,\delta\}}\, a^{-1}\psi\big((w+av)\cdot x\big) = \Phi_w(x)\cdot v + O(\delta)\,,\ \text{where } \Phi_w(x) = \psi'(w\cdot x)x\,. \tag{8.17}$$

In particular the correlation to the data of such a single random neuron is equal in expectation to $\sum_i y_i \Phi_w(x_i) \cdot v + O(\delta)$, and thus it is natural to take the perturbation vector v to be equal to $v_0 = \eta \sum_i y_i \Phi_w(x_i)$ (where η will be optimized to balance with the variance term), and we now find that:

$$\mathbb{E}_{a\sim\{-\delta,\delta\}} \sum_{i=1}^{n} y_i a^{-1} \psi((w+av_0)\cdot x_i) = \left\| \eta \sum_i y_i \Phi_w(x_i) \right\|^2 + O(\delta) = \eta y^\top H(w) y + O(\delta)\,, \quad (8.18)$$

where $H(w)$ is the Gram matrix of the feature embedding, namely

$$H(w)_{i,j} = \Phi_w(x_i) \cdot \Phi_w(x_j).$$

Note that for $\psi = ReLU$, one has in fact that the term $O(\delta)$ in (8.17) disappears for δ small is enough, and thus the correlation to the data is simply $\eta y^\top H(w) y$ in that case.

As we did with the NTK network, we now also take the base parameter w at random from a standard Gaussian. As we just saw, understanding the expected correlation then reduces to lower bound (spectrally) the Gram matrix H defined by $H_{i,j} = \mathbb{E}_{w\sim\mathcal{N}(0,\mathrm{I}_d)}[\psi'(w\cdot x_i)\psi'(w\cdot x_j) x_i \cdot x_j]$. This was exactly the content of Lemma 8.11 for $\psi = \mathrm{ReLU}$.

8.5.2 Eliminating the higher derivatives with a complex trick

The main issue of the strategy described above is that it requires to take δ small, which in turn may significantly increase the total weights of the resulting network. Our next idea is based on the following observation: Taking a random sign in (8.17) eliminates all the even order term in the Taylor expansion since $\mathbb{E}_{a\sim\{-1,1\}}[a^{-1}a^m] = 0$ for any even m (while it is $= 1$ for any odd m). However, taking a *complex* a, would rid us of *all* terms except the first order term. Namely, one has $\mathbb{E}_{a\in\mathbb{C}:|a|=1}[a^{-1}a^m] = 0$ for any $m \neq 1$. This suggests that it might make sense to consider neurons of the form

$$x \mapsto \mathrm{Re}\left(a^{-1}\psi\big((w+av)\cdot x\big)\right),$$

where a is a complex number of unit norm.

The challenge is now to give sense to $\psi(z)$ for a complex z, so that the rest of the argument remains unchanged. This gives rise to two caveats:

- There is no holomorphic extension of the ReLU function.
- The holomorphic extension of the activation function, even if exists, is a function of two (real) variables. The expression $\psi\big((w+av)\cdot x\big)$ when $a \notin \mathbb{R}$ is not a valid neuron to be

used in our construction since we're only allowed to use the original activation function as our non-linearity.

To overcome these caveats, the construction will be carried out in two steps, where in the first step we use *polynomial* activation functions, and in the second step, we replace these by the original activation function. It turns out that the calculation in Lemma 8.11 is particularly simple when the derivative of the activation function is a Hermite polynomial (see Section 8.6 for definitions), which is in particular obviously well-defined on $\mathbb{C}$ and in fact holomorphic. In the sequel, we fix $m \in \mathbb{N}$ so that

$$n\gamma^{m-2} \leq \frac{1}{2}. \tag{8.19}$$

Define

$$\varphi(z) = \frac{1}{\sqrt{m}} H_m(z), \quad z \in \mathbb{C}$$

where H_m is the m-th Hermite polynomial. Note that we also have $\varphi' = H_{m-1}$.

The first step of our proof will be to obtain a result analogous to Theorem 8.12 where ψ is replaced by φ.

Lemma 8.14. *Assume that (8.14) holds, that m is large enough so that $n\gamma^{m-2} \leq \frac{1}{2}$ and that for all $i \in [n]$, one has $y_i^2 \leq n\gamma^2$. There exist $\tilde{w}, \tilde{w}' \in \mathbb{R}^d$ and $z \in \mathbb{C}, |z| = 1$, such that for*

$$g(x) = Re\left(z \cdot \varphi\left(\left(\tilde{w} + \mathbf{i}\tilde{w}'\right) \cdot x\right)\right), \tag{8.20}$$

we have,

$$\mathbf{y} \cdot \mathbf{g} \geq \frac{1}{2C_m\sqrt{n\gamma^2}} \|\mathbf{y}\|^2.$$

Moreover, its weights admit the bounds

$$\|\tilde{w}\|^2, \|\tilde{w}'\|^2 \leq d(4C_m \log(n))^m \tag{8.21}$$

and for all $i \in [n]$,

$$|\tilde{w} \cdot x_i|, |\tilde{w}' \cdot x_i| \leq (4C_m \log(n))^{\frac{m}{2}}.$$

Given the above lemma, the second step towards Theorem 8.12 is to replace the polynomial attained by the above lemma by a ReLU. This will be achieved by:

- Observing that any polynomial in two variables $p(x, y)$ can be written as a linear combination of polynomials which only depend on one direction, hence polynomials of the form $q(ax + by)$.
- Using the fact that any nice enough function of one variable can be written as a mixture of ReLUs, due to the fact that the second derivative of the ReLU is a Dirac function (this was observed before, see e.g., [Lemma A.4, [142]]).

- The above implies that one can write the function $(x, y) \mapsto \varphi(x + iy)$ as the expectation of ReLUs such that the variance at points close to the origin is not too large.

These steps will be carried out in Section 8.5.4 below.

8.5.3 Constructing the complex neuron

Our approach to Lemma 8.14 will be to construct an appropriate distribution on neurons of type (8.20), and then show that the desirable properties are attained with positive probability. In what follows, let $w \sim \mathcal{N}(0, \mathrm{I}_d)$. Define

$$v(w) := \frac{1}{\sqrt{n\gamma^2}} \sum_{i=1}^{n} y_i \varphi'(w \cdot x_i) x_i.$$

Next, let a be uniformly distributed in the complex unit circle, and finally define

$$g(x) = \mathrm{Re}\left(a^{-1} \varphi((w + av(w)) \cdot x)\right). \tag{8.22}$$

We will prove the following two bounds.

Lemma 8.15. *Under the assumptions* (8.14) *and* (8.19)*, one has*

$$\mathbb{E}\left[\mathbf{y} \cdot \mathbf{g}\right] \geq \frac{1}{2\sqrt{n\gamma^2}} \|\mathbf{y}\|^2.$$

Lemma 8.16. *Suppose that the assumptions* (8.14) *and* (8.19) *hold. Assume also that for every* i *we have* $y_i \leq n\gamma^2$*. Then one has, for a constant* $C_m > 0$ *which depends only on* m*,*

$$\mathbb{E}[\|\mathbf{g}\|^2] \leq C_m n.$$

Moreover, for every $i \in [n]$ *and* $s > s_0$*, for some constant* s_0*,*

$$\mathbb{P}\left(|\mathrm{Re}((w + v(w)) \cdot x_i| > s\right), \mathbb{P}\left(|\mathrm{Im}((w + v(w))) \cdot x_i| > s\right) \leq \exp\left(\frac{1}{C_m} s^{-2/m}\right). \tag{8.23}$$

Recall the definition of the Gram matrix H,

$$H_{i,j} = \mathbb{E}_{w \sim \mathcal{N}(0, \mathrm{I}_d)}\left[\varphi'(w \cdot x_i)\varphi'(w \cdot x_j) x_i \cdot x_j\right].$$

As suggested in (8.18), we will need to bound H from below. We will need the following lemma.

Lemma 8.17. *Under the assumptions* (8.14) *and* (8.19)*, one has* $H \succeq \frac{1}{2}\mathrm{I}_n$*.*

Proof. If X and Y are standard, jointly-normal random variables with $\mathbb{E}\left[XY\right] = \rho$, by Lemma 8.22 one has $\mathbb{E}[H_{m-1}(X)H_{m-1}(Y)] = \rho^{m-1}$ and thus here $H_{i,j} = (x_i \cdot x_j)^m$. In particular if

$n \cdot \gamma^m \leq 1/2$ we obtain that for all $i \in [n]$ one has $1 = H_{i,i} \geq 2\sum_{j \neq i} |H_{i,j}|$. By diagonal dominance we conclude that $H \succeq \frac{1}{2}\mathrm{I}_n$. □

Proof of Lemma 8.15. For any $\beta \in \mathbb{N}, \beta \neq 1$, we have that $\mathbb{E}\left[a^{-1+\beta}\right] = 0$. Thus, since φ is an entire function, by taking its Taylor expansion around the point w, we obtain the identity

$$\mathbb{E}_a\left[a^{-1}\varphi((w + av(w)) \cdot x)\right] = \sum_{\beta=0}^{\infty} \frac{1}{\beta!}\mathbb{E}_a\left[a^{-1+\beta}\varphi^{(\beta)}(w \cdot x_i)(v(w) \cdot x)^{\beta}\right] = \varphi'(w \cdot x)v(w) \cdot x.$$

So we can estimate

$$\begin{aligned}\mathbb{E}_{w,a}\left[\sum_{i=1}^{n} y_i \mathrm{Re}\left(a^{-1}\varphi((w + av(w)) \cdot x_i)\right)\right] &= \sum_{i=1}^{n} y_i \mathbb{E}_w\left[\varphi'(w \cdot x_i)v(w) \cdot x_i\right] \\ &= \frac{1}{\sqrt{n\gamma^2}} \sum_{i,j} y_i y_j \mathbb{E}_w\left[\varphi'(w \cdot x_i)\varphi'(w \cdot x_j)x_i \cdot x_j\right] \\ &= \frac{1}{\sqrt{n\gamma^2}} \mathbf{y}^\top H \mathbf{y} \geq \frac{1}{2\sqrt{n\gamma^2}}\|\mathbf{y}\|^2,\end{aligned}$$

where the last inequality follows from Lemma 8.17. □

Proof of Lemma 8.16. In what follows, the expression C_m will denote a constant depending only on m, whose value may change between different appearances. Our objective is to obtain an upper bound on

$$\|\mathbf{g}\|^2 = \sum_{i=1}^{n} |\mathrm{Re}\left(a^{-1}\varphi((w + av(w)) \cdot x_i)\right)|^2.$$

Since φ is a polynomial of degree m we have

$$\|\mathbf{g}\|^2 \leq C_m \sum_{i=1}^{n} \left(1 + |w \cdot x_i|^{2m} + |v(w) \cdot x_i|^{2m}\right).$$

Moreover $w \cdot x_i$ is a standard Gaussian and thus $\mathbb{E}[|w \cdot x_i|^{2m}] \leq (2m)^m$. It therefore remains to control, for $x \in \{x_1, \ldots, x_n\}$, the expression

$$|v(w) \cdot x|^{2m} = \frac{1}{(n\gamma^2)^m}\left|\sum_{i=1}^{n} y_i H_{m-1}(w \cdot x_i)x_i \cdot x\right|^{2m}.$$

From hypercontractivity and the fact that the Hermite polynomials are eigenfunctions of the Ornstein-Uhlenbeck operator we have (see [140, Theorem 5.8])

$$\mathbb{E}\left[|v(w) \cdot x|^{2m}\right] \leq (2m)^{2m^2}\mathbb{E}\left[|v(w) \cdot x|^2\right]^m.$$

Thus, it will be enough to show that $\mathbb{E}_w[|v(w)\cdot x_j|^2] = O(1)$. We calculate

$$\begin{aligned}\mathbb{E}_w[|v(w)\cdot x_j|^2] &= \frac{1}{n\gamma^2}\mathbb{E}\left|\sum_{i=1}^n y_i H_{m-1}(w\cdot x_i)x_i\cdot x_j\right|^2 \\ &= \frac{1}{n\gamma^2}\left(\mathbb{E}\sum_{i=1}^n y_i^2\mathbb{E}[(H_{m-1}(w\cdot x_i))^2]|x_i\cdot x_j|^2\right. \\ &\quad \left.+\sum_{i\neq i'} y_i y_{i'}\mathbb{E}[H_{m-1}(w\cdot x_i)H_{m-1}(w\cdot x_i')](x_i\cdot x_j)(x_{i'}\cdot x_j)\right) \\ &\leq \frac{1}{n\gamma^2}\left(\sum_{i=1}^n y_i^2|x_i\cdot x_j|^2 + \frac{\gamma^{m-1}}{n\gamma^2}\sum_{i\neq i'}|y_i y_{i'}(x_{i'}\cdot x_j)(x_i\cdot x_j)|\right),\end{aligned}$$

where we used that $\mathbb{E}[(H_{m-1}(w\cdot x_i))^2] = 1$ and

$$|\mathbb{E}[H_{m-1}(w\cdot x_i)H_{m-1}(w\cdot x_{i'})]| = |x_i\cdot x_{i'}|^{m-1} \leq \gamma^{m-1},$$

valid whenever $i\neq i'$. By using that $\|\mathbf{y}\|^2 = O(n)$, we get

$$\frac{1}{n\gamma^2}\sum_{i=1}^n y_i^2|x_i\cdot x_j|^2 \leq \frac{y_j^2}{n\gamma^2} + \frac{\|\mathbf{y}\|^2}{n} = O(1).$$

To deal with the last term, observe that since $i\neq i'$ then $|(x_{i'}\cdot x_j)(x_i\cdot x_j)|\leq\gamma$, thus

$$\frac{\gamma^{m-1}}{n\gamma^2}\sum_{i\neq i'}|y_i y_{i'}(x_{i'}\cdot x_j)(x_i\cdot x_j)| \leq \frac{\gamma^{m-2}}{n}\left(\sum_{i=1}^n|y_i|\right)^2 \leq \gamma^{m-2}\|\mathbf{y}\|^2 = O(1),$$

where in the last inequality we've used $\gamma^{m-2}\leq\frac{1}{n}$. So, $\mathbb{E}_w[|v(w)\cdot x_i|^2] = O(1)$ as required.

Finally, to see (8.23) observe that both $\mathrm{Re}(w+v(w))$ and $\mathrm{Im}(w+v(w))$ are given by degree m polynomials of w, a standard Gaussian random vector. In [140, Theorem 6.7] it is shown that there exists a constant a_m depending only on m, such that if P is a polynomial of degree m and X is a standard normal random variable, then for every $t>2$,

$$\mathbb{P}\left(|p(X)| > t\sqrt{\mathbb{E}\left[p(X)^2\right]}\right) \leq \exp\left(-a_m t^{2/m}\right)$$

Thus, since

$$\mathbb{E}\left[|\mathrm{Re}(w+v(w))\cdot x_i|^2\right], \mathbb{E}\left[|\mathrm{Im}(w+v(w))\cdot x_i|^2\right] \leq \mathbb{E}\left[1+|w\cdot x_i|^{2m}+|v(w)\cdot x_i|^{2m}\right] \leq C_m,$$

the bound (8.23) follows. □

We are finally ready to prove the existence of the complex neuron.

Proof of Lemma 8.14. Consider the random variable

$$F = \mathbf{g} \cdot \mathbf{y} = \sum_{i=1}^{n} y_i g(x_i)$$

and set $W = \mathrm{Re}(w + v(w))$ and $W' = \mathrm{Im}(w + v(w))$. Lemma 8.15 gives

$$\mathbb{E}\left[F\right] \geq \frac{1}{2\sqrt{n\gamma^2}} \|\mathbf{y}\|^2.$$

Using Lemma 8.16 and Cauchy-Schwartz we may see that

$$\mathbb{E}\left[F^2\right] \leq \sum_{i=1}^{n} y_i^2 \mathbb{E}_{w,a}\left[\sum_{i=1}^{n} g(x_i)^2\right] \leq C_m n \|\mathbf{y}\|^2.$$

Define $G = \mathbb{1}_{\left\{\exists i : |W \cdot x_i|, |W' \cdot x_i| \geq (4C_m \log(n))^{\frac{m}{2}}\right\}}$. A second application of Cauchy-Schwartz gives

$$\mathbb{E}\Big[FG\Big] \leq \sqrt{C_m n \|\mathbf{y}\|^2 \mathbb{E}\left[G\right]}.$$

Now, the estimate (8.23) and a union bound yields

$$\mathbb{E}\left[G\right] \leq n \exp\left(-4\log(n)\right) \leq \frac{1}{n^3}.$$

Therefore,

$$\mathbb{E}\Big[FG\Big] \leq \frac{1}{n} C_m \|\mathbf{y}\|.$$

Combining this with the lower bound of $\mathbb{E}[F]$, we finally have

$$\mathbb{E}\Big[F(1-G)\Big] \geq \frac{1}{2\sqrt{n\gamma^2}} \|\mathbf{y}\|^2 - \frac{1}{n} C_m \|\mathbf{y}\| \geq \frac{1}{4\sqrt{n\gamma^2}} \|\mathbf{y}\|^2,$$

where the last inequality is valid as long as n is large enough. The claim now follows via taking a realization that exceeds the expectation. Since we might as well assume that the sample contains an orthonormal basis, (8.21) follows as well. □

8.5.4 Approximating a complex neuron with ReLU activation

Our goal in this section is to prove the following lemma, showing that the complex polynomial can be essentially replaced by a ReLU. We write $\psi(t) = \mathrm{ReLU}(t)$ and recall that $\phi(t) = \frac{1}{\sqrt{m}} H_m(t)$.

Lemma 8.18. *For any $w, w' \in \mathbb{R}^d, z \in \mathbb{C}$ with $|z| = 1$ and $M > 0$, there exist a pair of random variables S, B and a random vector $W \in \mathbb{R}^d$ such that for any $x \in \mathbb{S}^{d-1}$ with $m\left(|w \cdot x| + |w' \cdot x|\right) \leq M$,*

$$\mathbb{E}\left[S\psi(W \cdot x - B)\right] = \frac{c_{z,m}}{M^m}\mathrm{Re}\left(z \cdot \varphi(w \cdot x + \mathbf{i}w' \cdot x)\right),$$

where $c_{z,m}$ depends only on m and z and there exists another constant c_m, such that

$$\frac{1}{c_m} \geq c_{z,m} \geq c_m. \tag{8.24}$$

Moreover,

$$|S| = 1, |B| \leq M \quad \textit{almost surely,}$$

and

$$W = w + j \cdot w' \textit{ for some } j \in \{0, 1, \ldots, m\}.$$

Let us first see how to complete the proof of Theorem 8.12 using the combination of the above with Lemma 8.14.

Proof of Theorem 8.12. Invoke Lemma 8.14 to obtain a function

$$g(x) = Re\left(z \cdot \varphi\left(x \cdot \tilde{w} + \mathbf{i}x \cdot \tilde{w}'\right)\right)$$

such that

$$\mathbf{y} \cdot \mathbf{g} \geq \frac{1}{2C_m\sqrt{n\gamma^2}}\|\mathbf{y}\|^2,$$

and such that for every $i \in [n]$,

$$|\tilde{w} \cdot x_i|, |\tilde{w}' \cdot x_i| \leq C_m \log(n)^{\frac{m}{2}}.$$

Set $M = 2C_m m \log(n)^{\frac{m}{2}}$, so that $m(|\tilde{w} \cdot x_i| + |\tilde{w}' \cdot x_i|) \leq M$. By Lemma 8.18, we may find σ, w, b, such that

$$|b|^2 \leq M^2, \|w\|^2 \leq m^2(\|\tilde{w}\| + \|\tilde{w}'\|)^2 \leq 4C_m m^2 d \log(n)^m, \quad |\sigma| = 1,$$

for which we define $f(x) = \sigma\psi(w \cdot x - b)$. The lemma then implies,

$$\mathbf{y} \cdot \mathbf{f} \geq \frac{c_m}{M^m}\mathbf{y} \cdot \mathbf{g} \geq \frac{c'_m}{M^m\sqrt{n\gamma^2}}\|\mathbf{y}\|^2,$$

and

$$\|\mathbf{f}\|^2 = \sum_{i=1}^{n} \left(\psi(w \cdot x_i - b)\right)^2 \leq 2 \sum_{i=1}^{n} \left(|w \cdot x_i|^2 + b^2\right)$$
$$\leq 2M^2 n + 2 \sum_{i=1}^{n} |w \cdot x_i|^2.$$

By Lemma 8.18, $w = \tilde{w} + j \cdot \tilde{w}'$ for some $j = 0, ..., m$. Hence, $|w \cdot x_i|^2 \leq 2m(|\tilde{w} \cdot x_i|^2 + |\tilde{w}' \cdot x_i|^2)$ and

$$\|\mathbf{f}\|^2 \leq 2M^2 n + 4m \sum_{i=1}^{n} (|\tilde{w} \cdot x_i|^2 + |\tilde{w} \cdot x_i|^2) \leq 10mM^2 n.$$

The proof is concluded by substituting M. □

It remains to prove Lemma 8.18. This is done in the next subsections.

On homogeneous polynomials

Since our aim is to approximate a polynomial by ReLU, we first find an appropriate polynomial basis to work with.

Lemma 8.19. *Any polynomial of the form* $(x, y) \to Re(z \cdot (x + \mathbf{i}y)^m)$ *has the form,*

$$\sum_{j=0}^{m} a_j (x + j \cdot y)^m.$$

Proof. Define

$$\mathcal{H}_m = \{p(x, y) | p \text{ is a degree } m \text{ homogeneous polynomial}\},$$

and

$$A_m = \{(x + j \cdot y)^m | j = 0, ..., m\}.$$

It will suffice to show that A_m forms a basis for $\mathcal{H}_m$. The result will follow since $Re(z \cdot (x + \mathbf{i}y)^m)$ is clearly homogeneous. For $0 \leq j \leq m$, set $p_j = (x + j \cdot y)^m$, so that $p_j \in A_m$ and

$$p_j = \sum_{k=0}^{m} \binom{m}{k} j^k y^k x^{m-k}.$$

Note that the set $\{\binom{m}{k} y^k x^{m-k} | k = 0, ..., m\}$ forms a basis for $\mathcal{H}_m$ and in that basis p_j has coordinates $(1, j, \ldots, j^m)$. Taking the Vandermonde determinant of the matrix whose columns are $\{p_j : j = 0, ..., m\}$, we see that it must also be a basis for $\mathcal{H}_m$. □

Corollary 8.20. *Let $w, w' \in \mathbb{R}^d$ and $z \in \mathbb{C}$, then we have the following representation:*

$$\mathrm{Re}\left(z \cdot \varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) = \sum_{j=0}^{m} p_{z,j}((w + jw') \cdot x),$$

where each $p_{z,j}$ is a polynomial of degree m, which depends continuously on z.

Proof. The representation is immediate from the previous lemma. To address the point of continuity, we write

$$\begin{aligned}
\mathrm{Re}\left(z \cdot \varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) &= \mathrm{Re}(z)\mathrm{Re}\left(\varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) - \mathrm{Im}(z)\mathrm{Im}\left(\varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) \\
&= \mathrm{Re}(z)\mathrm{Re}\left(\varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) + \mathrm{Im}(z)\mathrm{Re}\left(\mathbf{i} \cdot \varphi(w \cdot x + \mathbf{i} w' \cdot x)\right) \\
&= \sum_{j=0}^{m} \left(\mathrm{Re}(z) p_{1,j}((w + jw') \cdot x) + \mathrm{Im}(z) p_{\mathbf{i},j}((w + jw') \cdot x)\right).
\end{aligned}$$

So, $p_{z,j}$ is a linear combination of $p_{1,j}$ and $p_{\mathbf{i},j}$, with coefficients that vary continuously in z. □

ReLUs as universal approximators

Next, we show how ReLU functions might be used to universally approximate compactly supported functions.

Proposition 8.21. *Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable and compactly supported on $[-M, M]$. Then, there exists a pair of random variables S, B, such that, for every $x \in [-M, M]$,*

$$\mathbb{E}\left[S\psi(x - B)\right] = \frac{f(x)}{\int |f''|},$$

and such that, almost surely $|B| \leq M$ and $|S| = 1$.

Proof. Observe that, when considered as a distribution, $\psi(x)'' = \delta_0$. Therefore, there exists a linear function L such that

$$f(x) + L(x) = \int_{-M}^{M} \psi(x - y) f''(y) dy.$$

$f''(x)$ is the second derivative of a compactly supported function which implies that $f(x) + L(x)$ is compactly supported as well. Hence, $L(x) \equiv 0$. Let B be the random variable whose density is $\frac{|f''|}{\int_{-M}^{M} |f''|}$ and set $S = \mathrm{sign}(f''(B)))$. We now have

$$\mathbb{E}\left[S\psi(x - B)\right] = \frac{\int_{-M}^{M} \psi(x - y) f''(y) dy}{\int_{-M}^{M} |f''|} = \frac{f(x)}{\int |f''|}.$$

□

Completing the proof of Lemma 8.18

Set χ_M to be a bump function for the interval $[-M, M]$. That is,

- $\chi_M : \mathbb{R} \to \mathbb{R}$ is smooth.
- $0 \leq \chi_M \leq 1$.
- $\chi_M(x) = 1$ for $x \in [-M, M]$.
- $\chi_M(x) = 0$ for $|x| > 2M$.

By Corollary 8.20, for any $w, w' \in \mathbb{R}^d, z \in \mathbb{C}$ we have the representation

$$\mathrm{Re}\left(z \cdot \varphi(w \cdot x + \mathrm{i}w' \cdot x)\right) \chi_M(|w\cdot x|+m|w'\cdot x|) = \sum_{j=0}^{m} p_{z,j}((w+jw')\cdot x)\chi_M(|w\cdot x|+m|w'\cdot x|). \tag{8.25}$$

Proof of Lemma 8.18. Define $X = \left\{x \in \mathbb{S}^{d-1}; m\left(|w \cdot x| + |w' \cdot x|\right) \leq M\right\}$. Observe that for all $x \in X$,

$$\mathrm{Re}\left(\varphi(w \cdot x + \mathbf{i}w' \cdot x)\right) = \mathrm{Re}\left(\varphi(w \cdot x + \mathbf{i}w' \cdot x)\right) \chi_M\left(m\left(|w \cdot x| + |w' \cdot x|\right)\right).$$

Moreover, if $j = 0, \ldots, m$, then $\chi_M((w + jw') \cdot x) = 1$, as well. By invoking Proposition 8.21 we deduce that for every $j = 0, ..., m$, there exists a pair of random variables S_j, B_j and a constant $c_{z,j} > 0$ depending only on j, m and z, such that

$$\mathbb{E}\left[S_j \psi\left((w + jw') \cdot x - B_j\right)\right] = \frac{c_{z,j}}{M^m} p_{z,j}((w + jw') \cdot x)\chi_M((w + jw') \cdot x), \quad \forall x \in X,$$

Here we have used the fact that if p_j is one of the degree m polynomials in the decomposition (8.25), then there exist some constants $C'_{z,j}, C_{z,j} > 0$, for which

$$C'_{z,j} M^m \leq \int_{-M}^{M} |p''_{z,j}| \leq \int_{-2M}^{2M} |p''_{z,j}| \leq C_{z,j} M^m.$$

We now set J to be a random index from the set $\{0, \ldots, m\}$ with

$$\mathbb{P}(J = j) = \frac{c_{z,j}^{-1}}{\sum_{j'} c_{z,j'}^{-1}}.$$

If we set $c_{z,m} = \frac{1}{\sum_{j'} c_{z,j'}^{-1}}$, and $S := S_J, B = B_J, W = w + Jw'$ it follows from (8.25) that

$$\begin{aligned}\mathbb{E}\left[S\psi(W \cdot x - B)\right] &= \frac{c_{z,m}}{M^m} \sum_{j=0}^{m} p_{z,j}((w + jw') \cdot x)\chi_M((w + jw') \cdot x) \\ &= \frac{c_{z,m}}{M^m} \mathrm{Re}\left(z \cdot \varphi(w \cdot x + \mathbf{i}w' \cdot x)\right) \chi_M\left(m\left(|w \cdot x| + |w' \cdot x|\right)\right).\end{aligned}$$

Finally since, by Corollary 8.20, $c_{z,m}$ depends continuously on z, a compactness argument implies (8.24). □

8.6 Hermite polynomials

Define the m'th Hermite polynomial by:

$$H_m(x) = \frac{(-1)^m}{\sqrt{m!}} \left(\frac{d^m}{dx^m} e^{-\frac{x^2}{2}}\right) e^{\frac{x^2}{2}}.$$

For ease of notion we also define $H_{-1} \equiv 0$. The Hermite polynomials may also be regarded as the power series associated to the function $F(t, x) = \exp(tx - \frac{t^2}{2})$. Indeed,

$$\begin{aligned}F(t, x) &= \exp\left(\frac{x^2}{2} - \frac{(x - t)^2}{2}\right) \\ &= e^{\frac{x^2}{2}} \sum_{\ell=0}^{\infty} \frac{t^m}{m!} \left(\frac{d^m}{dt^m} e^{-\frac{(x-t)^2}{2}}\right)\Big|_{t=0} \\ &= \sum_{m=0}^{\infty} \frac{t^m}{\sqrt{m!}} H_m(x). \qquad (8.26)\end{aligned}$$

Observe that $\frac{d}{dx}F(t, x) = tF(t, x)$, so that, since $H_0 \equiv 1$,

$$\sum_{m=1}^{\infty} \frac{t^m}{\sqrt{(m-1)!}} H_{m-1}(x) = \sum_{m=1}^{\infty} \frac{t^m}{\sqrt{m!}} H'_m(x),$$

and we deduce

$$H'_m = \sqrt{m} H_{m-1}. \qquad (8.27)$$

Also $\frac{d}{dt}F(t, x) = (x - t)F(t, x)$ and a similar argument shows that

$$\sqrt{\frac{m}{m-1}} H_m(x) = \frac{x}{\sqrt{m-1}} H_{m-1}(x) - H_{m-2}(x). \qquad (8.28)$$

Furthermore, we show that the family $\{H_m\}$ satisfies the following orthogonality relation, which we shall freely use.

Lemma 8.22. *Let $X, Y \sim \mathcal{N}(0,1)$ be jointly Gaussian with $\mathbb{E}[XY] = \rho$. Then*

$$\mathbb{E}\left[H_m(X) H_{m'}(Y)\right] = \delta_{m,m'} \rho^m .$$

Proof. Fix $s, t \in \mathbb{R}$. We have the following identity

$$\mathbb{E}\left[F(s,X)F(t,Y)\right] = \mathbb{E}\left[\exp(sX + tY)\right] \exp\left(-\frac{s^2+t^2}{2}\right) = e^{st\cdot\rho},$$

where in the second equality we have used the formula for the moment generating functions of bi-variate Gaussians. In particular, we have

$$\frac{d^{m+m'}}{ds^m dt^{m'}} \mathbb{E}\left[F(s,X)F(t,Y)\right]\Big|_{t=0,s=0} = \frac{d^{m+m'}}{ds^m dt^{m'}} e^{st\cdot\rho}\Big|_{t=0,s=0}.$$

By (8.26), the left hand side equals $\mathbb{E}\left[H_\ell(X) H_{\ell'}(Y)\right]$, while the right hand side is $\delta_{m,m'}\rho^m$. The proof is complete. □

8.7 More general non-linearities

We now consider an arbitrary L-Lipschitz non-linearity ψ that is differentiable except at a finite number of points and such that $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[(\psi'(X))^2] < +\infty$. In particular, with $H_1, H_2, \ldots$ being the Hermite polynomials (normalized such that it forms an orthonormal basis) we have that there exists a sequence of real numbers (a_ℓ) such that

$$\psi' = \sum_{\ell \geq 0} a_\ell H_\ell \,.$$

Our generalization of Theorem 8.8 now reads as follows:

Theorem 8.23. *Under the above assumptions on ψ, there exists $f \in \mathcal{F}_k(\psi)$ with $\|\mathbf{f} - \mathbf{y}\|^2 \leq \varepsilon\|\mathbf{y}\|^2$ provided that*

$$k \cdot d \geq \frac{16\omega \cdot L}{\sum_{\ell \geq \frac{\log(2n)}{2\log(1/\gamma)}} a_\ell^2} \cdot n \log(1/\varepsilon)\,.$$

In fact there is an efficient procedure that produces a random $f \in \mathcal{F}_k(\psi)$ with $\mathbb{E}[\|\mathbf{f} - \mathbf{y}\|^2] \leq \varepsilon\|\mathbf{y}\|^2$ when (8.5) holds.

Proof. First we follow the proof of Lemma 8.10, with the only change being: (i) in (8.9) there is an additive $O(\delta)$ term (also now the condition on u is that $u \cdot x_i$ is not in the finite set of points where ψ is not differentiable), and (ii) in (8.13) we use that $|\psi'| \leq L$. We obtain that for $u \in \mathbb{R}^d$

there exists $f \in \mathcal{F}_2(\psi)$ such that

$$\sum_{i=1}^{n} y_i f(x_i) \geq \frac{1}{2} \left\| \sum_{i=1}^{n} \psi'(u \cdot x_i) y_i x_i \right\|^2 , \tag{8.29}$$

where the $1/2$ compared to (8.10) is due to modification (i) above, and furthermore

$$\sum_{i=1}^{n} f(x_i)^2 \leq \frac{2\omega \cdot n \cdot L}{d} \cdot \sum_{i=1}^{n} y_i f(x_i) , \tag{8.30}$$

where the added term L is due to modification (ii) above and the added 2 is due to (i).

Next we follow the proof of Lemma 8.11, noting that the matrix H is now defined by (recall Lemma 8.22) $H_{i,j} = \sum_{\ell \geq 0} a_\ell^2 (x_i \cdot x_j)^{\ell+1}$, to obtain:

$$\mathbb{E}_{u \sim \mathcal{N}(0, \mathrm{I}_n)} \left\| \sum_{i=1}^{n} \psi'(u \cdot x_i) y_i x_i \right\|^2 \geq \frac{1}{2} \sum_{\ell \geq \frac{\log(2n)}{2\log(1/\gamma)}} a_\ell^2 \cdot \sum_{i=1}^{n} y_i^2 . \tag{8.31}$$

In particular we obtain from (8.29) and (8.31) that (8.6) holds true with the term $\frac{1}{10} \cdot \sqrt{\frac{\log(1/\gamma)}{\log(2n)}}$ replaced by $\frac{1}{4} \sum_{\ell \geq \frac{\log(2n)}{2\log(1/\gamma)}} a_\ell^2$, and from (8.30) that (8.7) holds true with ω replaced by $2\omega \cdot L$. We can thus conclude as we concluded Theorem 8.8 from Theorem 8.9. □

9

Community Detection and Percolation of Information in a Geometric Setting

9.1 Introduction

Community detection in large networks is a central task in data science. It is often the case that one gets to observe a large network, the links of which depends on some unknown, underlying community structure. A natural task in this case is to detect and recover this community structure to the best possible accuracy.

Perhaps the most well-studied model in this topic is the *stochastic block model* (SBM) where a random graph whose vertex set is composed of several communities, $\{c_1, ..., c_k\}$ is generated in a way that every pair of nodes v, u which belong to communities $c(u), c(v)$, will be connected to each other with probability $p = p(c(v), c(u))$, hence with probability that only depends on the respective communities, and otherwise independently. The task is to recover the communities $c(\cdot)$ based on the graph (and assuming that the function $p(\cdot, \cdot)$ is known). The (unknown) association of nodes with communities is usually assumed to be random and independent between different nodes. See [3] for an extensive review of this model.

A natural extension of the stochastic block model is the geometric random graph, where the discrete set of communities is replaced by a metric space. More formally, given a metric space (X, d), a function $f : V \to X$ from a vertex set V to the metric space and a function $\varphi : \mathbb{R}_+ \to [0, 1]$, a graph is formed by connecting each pair of vertices u, v independently, with

probability

$$p(u,v) := \varphi\left(d(f(u),f(v))\right).$$

This model can sometimes mimic the behavior of real-world networks more accurately than the stochastic block model. For example, a user in a social network may be represented as a point in some linear space in a way that the coordinates correspond to attributes of her personality and her geographic location. The likelihood of two persons being associated with each other in the network will then depend on the proximity of several of these attributes. A flat community structure may therefore be two simplistic to reflect these underlying attributes.

Therefore, a natural extension of the theory of stochastic block models would be to understand under what conditions the geometric representation can be recovered by looking at the graph. Our focus is on the case that the metric is defined over a symmetric space, such as the Euclidean sphere in d-dimensions. By symmetry, we mean that the probability of two vertices to be connected, given their locations, is invariant under a natural group acting on the space. We are interested in the sparse regime where the expected degrees of the vertices do not converge to infinity with the size of the graph. This is the (arguably) natural and most challenging regime for the stochastic block model.

9.1.1 Inference in geometric random graphs

For the sake of simplicity, in what follows, we will assume that the metric space is the Euclidean sphere, and our main theorems will be formulated in this setting; It will be straightforward, yet technical, to generalize our results to more general symmetric space (see [92] for further discussion on this point).

In order to construct our model, we need some notation. Let σ be the uniform probability measure on $\mathbb{S}^{d-1}$ and let $\varphi : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [0,1]$, be of the form $\varphi(x,y) = f(\langle x,y\rangle)$ for some $f : [-1,1] \to [0,1]$. Define the integral operator $A_\varphi : L^2\left(\mathbb{S}^{d-1}\right) \to L^2\left(\mathbb{S}^{d-1}\right)$ by

$$A_\varphi(g)(x) = \int\limits_{\mathbb{S}^{d-1}} \varphi(x,y)g(y)d\sigma(y).$$

It is standard to show that A_φ is a self-adjoint compact operator (see [137], for example) and so has a discrete spectrum, except at 0. By definition, φ is invariant to rotations and so A_φ commutes with the Laplacian. It follows that the eigenfunctions of A_φ are precisely the *spherical harmonics* which we denote by $\{\psi_i\}_{i=0}^{\infty}$. Thus, if $\lambda_i(A_\varphi)$ denotes the eigenvalue of φ corresponding to ψ_i we have the following identity,

$$\varphi = \sum_{i=0}^{\infty} \lambda_i \psi_i \otimes \psi_i. \tag{9.1}$$

In particular, $\psi_0 = 1$ and for $i = 1, ..., d$, ψ_i are linear functionals such that, for $x, y \in \mathbb{S}^{d-1}$,

$$d \cdot \langle x, y \rangle = \sum_{l=1}^{d} \psi_l(x)\psi_l(y). \tag{9.2}$$

Note that in our notation the eigenvalues are indexed by the spherical harmonics, and are therefore not necessarily in decreasing order. By rotational invariance it must hold that

$$\lambda(\varphi) := \lambda_1 = ... = \lambda_d. \tag{9.3}$$

Define $\|\varphi\|_\infty = \sup_{x,y} \varphi(x, y)$. We make the following, arguably natural, assumptions on the function φ:

A1. There exist $\delta > 0$ such that $\min_{i \neq 1,...,d} |\lambda(\varphi) - \lambda_i| > \delta$.

A2. Reordering the eigenvalues in decreasing order $\lambda_{l_0} \geq \lambda_{l_1} \geq \lambda_{l_2} \geq \dots$ there exists $C > 0$ such that for every $i \geq 0, \lambda_{l_i} \leq \frac{C}{(i+1)^2}$.

Let $\{X_i\}_{i=1}^n \sim \sigma$ be a sequence of independently-sampled vectors, uniformly distributed on $\mathbb{S}^{d-1}$. Let $G\left(n, \frac{1}{n}\varphi(X_i, X_j)\right)$ be the inhomogeneous Erdös-Rényi model where edges are formed independently with probability $\frac{1}{n}\varphi(X_i, X_j)$ and let A be the adjacency matrix of a random graph drawn from $G\left(n, \frac{1}{n}\varphi(X_i, X_j)\right)$.

Definition 9.1. We say that the model is $\varepsilon -$ *reconstructible* if, for all n large enough, there is an algorithm which returns an $n \times n$ matrix A such that

$$\frac{1}{n^2} \sum_{i,j} |A_{i,j} - X_i \cdot X_j|^2 \leq \varepsilon.$$

Remark that, due the symmetry of the model, it is clear that the locations can only be reconstructed up to an orthogonal transformation, which is equivalent to reconstruction of the Gram matrix.

Theorem 9.2. *For every $\varepsilon > 0$ there exists a constant $C = C(\varepsilon, d)$, such that the model is ε-reconstructible whenever*

$$\min_{i \neq 0,...,d} |\lambda_i - \lambda(\varphi)|^2 \geq C\|\varphi\|_\infty. \tag{9.4}$$

Remark 9.3. Observe that, since the left hand side of condition (9.4) is 2-homogeneous, whereas its right hand side is 1-homogeneous, we have that as long as the left hand side is nonzero, by multiplication of the function φ by a large enough constant, the condition can be made to hold true.

Theorem 9.2 should be compared to the known bounds for recovery in the stochastic block model (see [3]). In particular, it is illuminating to compare the SBM with two communities to linear kernels in our model. In this case, both models are parameterized by two numbers a, b. In the SBM these are the inter- and intra- communities probabilities and in our model, the coefficients of the linear function. In the SBM, recovery of the communities depends on the ratio $\frac{(a-b)^2}{a+b}$. The example below gives the same result for linear kernels, with a dimensional affect, which typically makes reconstruction easier.

Example 1. Consider the linear kernel, $\varphi(x, y) = a + b\langle x, y\rangle$, with $|b| \leq a$. A calculation shows that

$$\lambda_0 = a$$
$$\lambda(\varphi) = \frac{b}{d}.$$

Applying our theorem, we show that the model is reconstructible whenever

$$\left|a - \frac{b}{d}\right|^2 \geq C \cdot (a + b).$$

Methods and related works. Our reconstruction theorem is based on a spectral method, via the following steps:

1. We observe that by symmetry of our kernel, linear functions are among its eigenfunctions. We show that the kernel matrix (hence the matrix obtained by evaluating the kernel at pairs of the points (X_i)) will have a respective eigenvalues and eigenvectors which approximate the ones of the continuous kernel.

2. Observing that the kernel matrix is the expectation of the adjacency matrix, we rely on a matrix concentration inequality due to Le-Levina-Vershynin [157] to show that the eigenvalues of the former are close to the ones of the latter.

3. We use the Davis-Kahan theorem to show that the corresponding eigenvectors are also close to each other.

The idea in Steps 2 and 3 is not new, and rather standard (see [157] and references therein). Thus, the main technical contribution in proving our upper bound is in Step 1, where we prove a bound for the convergence of eigenvectors of kernel matrices. So far, similar results have only been obtained in the special case that the Kernel is positive-definite, see for instance [51].
The paper [235] considers kernels satisfying some Sobolev-type hypotheses similar to our assumptions on φ (but gives results on the spectrum rather than the eigenvectors). Reconstruction of the eigenspace has been considered in [230] for positive definite kernels in the dense regime,

in [228] for random dot products graphs and in [14] in the dense and relatively sparse regimes again for kernels satisfying some Sobolev-type hypotheses.

Let us also mention other works which augmented the SBM with geometric information. The paper [93] considers the problem of discrete community recovery when presented with informative node covariates, inspired by the spiked covariance model. The authors derived a sharp threshold for recovery by introducing a spectral-like algorithm. However, the model is rather different than the one we propose in which the community structure is continuous.

A model which is slightly closer to the one we consider appears in [121]. In this model, communities are still discrete but the edge connectivity depends continuously on the latent positions of nodes on some d dimensional sphere. In such a model, recovery of the communities may be reduced to more combinatorial arguments. Indeed, the number of common neighbors can serve as an indicator for checking whether two nodes come from the same community. A similar idea was previously explored in [56], where a triangle count was used to establish a threshold for detecting latent geometry in random geometric graphs.

9.1.2 Percolation of geometric information in trees

The above theorem gives an upper bound for the threshold for reconstruction. The question of finding respective lower bounds, in the stochastic block model, is usually reduced to a related but somewhat simpler model of percolation of information on *trees*. The idea is that in the sparse regime, the neighborhood of each node in the graph is usually a tree, and it can be shown that recovering the community of a specific node based on observation of the entire graph, is more difficult than the recovery of its location based on knowledge of the community association of the leaves of a tree rooted at this node. For a formal derivation of this reduction (in the case of the SBM), we refer to [3].

This gives rise to the following model, first described by Mossel and Peres [186] (see also [184]): Consider a q-ary tree $T = (V, E)$ of depth k, rooted at $r \in V$. Suppose that each node in V is associated with a label $\ell : V \to \{1, .., k\}$ in the following way: The root r is assigned with some label and then, iteratively, each node is assigned with its direct ancestor's label with probability p and with a uniformly picked label with probability $1-p$ (independent between the nodes at each level). The goal is then to detect the assignment of the root based on observation of the leaves.

Let us now suggest an extension of this model to the geometric setting. We fix a Markov kernel $\varphi(x, y) = f(\langle x, y\rangle)$ such that $\int_{\mathbb{S}^{n-1}} \varphi(x, y)d\sigma(y) = 1$ for all $x \in \mathbb{S}^{d-1}$. We define $g : T \to \mathbb{S}^{d-1}$ in the following way. For the root r, $g(r)$ is picked according to the uniform measure. Iteratively, given that $g(v)$ is already set for all nodes v at the ℓ-th level, we pick the values $g(u)$ for nodes u at the $(\ell + 1)$th level independently, so that if u is a direct descendant of v, the label $g(u)$ is distributed according to the law $\varphi(g(v), \cdot)d\sigma$.

Denote by $T_k \subset V$ the set of nodes at depth k, and define by μ_k the conditional distribution of $g(r)$ given $(g(v))_{v \in T_k}$. We say that the model has positive information flow if

$$\lim_{k \to \infty} \mathbb{E}\left[\mathrm{TV}(\mu_k, \sigma)\right] > 0.$$

Remark that by symmetry, we have

$$\mathbb{E}\left[\mathrm{TV}(\mu_k, \sigma)\right] = \mathbb{E}\left[\mathrm{TV}(\mu_k, \sigma) | g(r) = e_1\right]$$

where r is the root and e_1 is the north pole.

Our second objective in this work is to make the first steps towards understanding under which conditions the model has positive information flow, and in particular, our focus is on providing nontrivial sufficient conditions on q, φ for the above limit to be equal to zero.

Let us first outline a natural sufficient condition for the information flow to be positive which, as we later show, turns out to be sharp in the case of Gaussian kernels. Consider the following simple observable,

$$Z_k := \frac{1}{|T_k|} \sum_{v \in T_k} g(v).$$

By Bayes' rule, we clearly have that the model has positive information flow if (but not only if)

$$\liminf_{k \to \infty} \frac{\mathbb{E}[\langle Z_k, e_1 \rangle | g(r) = e_1]}{\sqrt{\mathrm{Var}\left[\langle Z_k, e_1 \rangle | \, g(r) = e_1\right]}} > 0. \tag{9.5}$$

This gives rise to the parameter

$$\lambda(\varphi) := \int_{\mathbb{S}^{d-1}} \langle x, e_1 \rangle \varphi(e_1, x) d\sigma(x),$$

which is the eigenvalue corresponding to linear harmonics. By linearity of expectation, we have

$$\mathbb{E}[\langle Z_k, e_1 \rangle | g(r) = e_1] = \lambda(\varphi)^k.$$

For two nodes $u, v \in T$ define by $c(u, v)$ the deepest common ancestor of u, v and by $\ell(u, v)$ its

level. A calculation gives

$$\begin{aligned}
\mathrm{Var}\left[\langle Z_k, e_1\rangle |\, g(r) = e_1\right] &= \frac{1}{q^{2k}} \sum_{u,v \in T_k} \mathbb{E}\left[g(v)_1 g(u)_1 | g(r) = e_1\right] - \lambda(\varphi)^{2k} \\
&= \frac{1}{q^{2k}} \sum_{u,v \in T_k} \mathbb{E}\left[\mathbb{E}[g(v)_1 | g(c(u,v))]\mathbb{E}[g(u)_1 | g(c(u,v))] | g(r) = e_1\right] - \lambda(\varphi)^{2k} \\
&= \frac{1}{q^{2k}} \sum_{u,v \in T_k} \mathbb{E}\left[g(c(u,v))_1^2 | g(r) = e_1\right] \lambda(\varphi)^{2(k-\ell(u,v))} - \lambda(\varphi)^{2k} \\
&\leq \frac{1}{q^{2k}} \sum_{u,v \in T_k} \lambda(\varphi)^{2(k-\ell(u,v))} - \lambda(\varphi)^{2k} \\
&= \frac{\lambda(\varphi)^{2k}}{q^{2k}} \sum_{u,v \in T_k} \lambda(\varphi)^{-2\ell(u,v)} - \lambda(\varphi)^{2k} \\
&= \frac{\lambda(\varphi)^{2k}}{q^{2k}} \sum_{\ell=0}^{k} q^{\ell} q^{2(k-\ell)} \lambda(\varphi)^{-2\ell} - \lambda(\varphi)^{2k} = \lambda(\varphi)^{2k} \sum_{\ell=1}^{k} \left(q\lambda(\varphi)^2\right)^{-\ell}.
\end{aligned}$$

This gives a sufficient condition for (9.5) to hold true, concluding:

Claim 9.4. *The condition $q\lambda(\varphi)^2 > 1$ is sufficient for the model to have positive percolation of information.*

We will refer to this as the Kesten-Stigum (KS) bound.

We now turn to describe our lower bounds. For the Gaussian kernel, we give a lower bound which misses by a factor of 2 from giving a matching bound to the KS bound. To describe the Gaussian kernel, fix $\beta > 0$, let X be a normal random variable with law $\mathcal{N}(0, \beta)$ and suppose that $\varphi : \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{R}$ is such that

$$\varphi(x, \cdot) \text{ is the density of } (x + X) \mathrm{mod} 2\pi, \tag{9.6}$$

where we identify $\mathbb{S}^1$ with the interval $[0, 2\pi)$. We have the following result.

Theorem 9.5. *For the Gaussian kernel defined above, there is zero information flow whenever $q\lambda(\varphi) < 1$.*

Related results in the Gaussian setting. In the discrete setting,an analogous result was obtained in [183]. In fact, our method of proof is closely related. We use the inherent symmetries of the spherical kernels to decouple the values of $g(r)$ and $g(v)$, where v is some vertex which is sufficiently distant from r. This is same idea of Proposition 10 in [183] which uses a decomposition of the transition matrix to deduce a similar conclusion.

Another related result appears in [187]. In the paper the authors consider a broadcast model on a binary tree with a Gaussian Markov random field and obtain a result in the same spirit as

the one above. However, since the random field they consider is real valued, as opposed to our process which is constrained to the circle, the method of proof is quite different.

In the general case, we were unable to give a corresponding bound, nevertheless, using the same ideas, we are able to give some nontrivial sufficient condition for zero flow of information for some $q > 1$, formulated in terms of the eigenvalues of the kernel. To our knowledge, this is the first known result in this direction. In order to formulate our result, we need some definitions.

We begin with a slightly generalized notion of a q-ary tree.

Definition 9.6. Let $q > 1$, we say that T is a tree of growth at most q if for every $k \in \mathbb{N}$,

$$|T_k| \leq \lceil q^k \rceil .$$

Now, recall that $\varphi(x, y) = f(\langle x, y \rangle)$. Our bound is proven under the following assumptions on the kernel.

- f is monotone.
- f is continuous.
- $\lambda(\varphi) > 0$ and for every $i \geq 1$, $|\lambda_i| \leq \lambda(\varphi)$.

We obtain the following result.

Theorem 9.7. *Let φ satisfy the assumptions above and let T be a tree of growth at most q. There exists a universal constant $c > 0$, such that if*

$$q \leq \left(1 - c \frac{\ln(\lambda(\varphi))(1 - \lambda(\varphi))^2}{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{f(1)} \right)} \right)^{-1}$$

then the model has zero percolation of information.

9.2 The upper bound: Proof of Theorem 9.2

Recall that

$$\varphi = \sum_{k=0}^{\infty} \lambda_k \psi_k \otimes \psi_k,$$

with the eigenvalues λ_k indexed by the spherical harmonics. Define the random matrices M_n, Ψ_n by

$$(M_n)_{i,j} = \frac{1}{n} \varphi(X_i, X_j), \qquad (\Psi_n)_{i,k} = \frac{1}{\sqrt{n}} \psi_k(X_i).$$

Note that M_n is an $n \times n$ matrix, while Ψ_n, has infinitely many columns. Furthermore, denote by Λ the diagonal matrix $\mathrm{diag}\{\lambda_i\}_{i=0}^{\infty}$. Then

$$(M_n)_{i,j} = \frac{1}{n}\sum_{k=0}^{\infty} \lambda_k \psi_k(X_i)\psi_k(X_j) = (\Psi_n \Lambda \Psi_n^T)_{i,j}.$$

For $r \in \mathbb{N}$ we also denote

$$\varphi^r := \sum_{k=0}^{r} \lambda_k \psi_k \otimes \psi_k,$$

the finite rank approximation of φ, $\Lambda^r = \mathrm{diag}\{\lambda_k\}_{k=0}^r$, and Ψ_n^r the sub-matrix of Ψ_n composed of its first r columns. Finally, denote

$$M_n^r = \Psi_n^r \Lambda^r (\Psi_n^r)^T.$$

As before, let A be an adjacency matrix drawn from $G\left(n, \frac{1}{n}\varphi(X_i, X_j)\right)$ so that $\mathbb{E}A = M_n$. Our goal is to recover Ψ_n^{d+1} from the observed A. The first step is to recover Ψ_n^{d+1} from M_n. We begin by showing that the columns of Ψ_n^r are, up to a small additive error, eigenvectors of M_n^r. To this end, denote

$$E_{n,r} := (\Psi_n^r)^T \Psi_n^r - \mathrm{Id}_r,$$

$C(n,r) = \|E_{n,r}\|_{op}^2$, and $K = \max_i \lambda_i$.

Lemma 9.8. *Let u_i be the i'th column of Ψ_n and let $\eta > 0$. Then*

$$\|M_n^r u_i - \lambda_i u_i\|_2^2 \leq K^2(\sqrt{C(n,r)} + 1)C(n,r).$$

Moreover, whenever $n \geq l(r)\log(2r/\eta)$, we have with probability larger than $1-\eta$,

$$C(n,r) \leq \frac{4l(r)\log(2r/\eta)}{n}.$$

where $l(r)$ only depends on r and on the dimension.

Proof. Let $e_i \in \mathbb{R}^r$ be the i'th standard unit vector so that $u_i = \Psi_n^r e_i$. So,

$$(\Psi_n^r)^T u_i = (\Psi_n^r)^T \Psi_n^r e_i = (\mathrm{Id}_r + (\Psi_n^r)^T \Psi_n^r - \mathrm{Id}_r)e_i = e_i + E_{n,r}e_i.$$

We then have

$$\begin{aligned} M_n^r u_i &= \Psi_n^r \Lambda^r (\Psi_n^r)^T u_i = \Psi_n^r \Lambda^r e_i + \Psi_n^r \Lambda^r E_{n,r} e_i \\ &= \lambda_i \Psi_n^r e_i + \Psi_n^r \Lambda^r E_{n,r} e_i = \lambda_i u_i + \Psi_n^r \Lambda^r E_{n,r} e_i. \end{aligned}$$

To bound the error, we estimate $\|M_n^r u_i - \lambda_i u_i\|_2^2 = \|\Psi_n^r \Lambda^r E_{n,r} e_i\|_2^2$ as

$$\begin{aligned}\langle \Lambda^r E_{n,r} e_i, (\Psi_n^r)^T \Psi_n^r \Lambda^r E_{n,r} e_i \rangle &= \langle \Lambda^r E_{n,r} e_i, E_{n,r} \Lambda^r E_{n,r} e_i \rangle + \|\Lambda^r E_{n,r} e_i\|_2^2 \\ &\leq \left(\sqrt{C(n,r)} + 1\right) \|\Lambda^r E_{n,r} e_i\|_2^2 \leq K^2 \left(\sqrt{C(n,r)} + 1\right) C(n,r).\end{aligned}$$

It remains to bound $C(n,r)$. Let $X_i^r = (\psi_0(X_i), \ldots, \psi_{r-1}(X_i))$ stand for the i'th row of Ψ_n^r. Then, $E_{n,r} = \frac{1}{n}\sum_{i=1}^n \left((X_i^r)^T X_i^r - \mathrm{Id}_r\right)$, is a sum of independent, centered random matrices. We have

$$\begin{aligned}\sigma_{n,r}^2 &:= \left\| \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \left((X_i^r)^T X_i^r - \mathrm{Id}_r\right) \right)^2 \right\|_{op} \\ &= \frac{1}{n} \left\| \mathbb{E}\left((X_1^r)^T X_1^r - \mathrm{Id}_r\right)^2 \right\|_{op}\end{aligned}$$

Furthermore, the norm of the matrices can be bounded by

$$\begin{aligned}\left\| \frac{1}{n}\left((X_1^r)^T X_1^r - \mathrm{Id}_r\right) \right\|_{op} &= \frac{1}{n} \max(1, \|X_1^r\|_2^2 - 1) \\ &\leq \frac{1}{n} \max\left(1, \left\| \sum_{i=0}^r \psi_i^2 \right\|_\infty - 1\right).\end{aligned}$$

Note that the right hand side of the two last displays are of the form $\frac{1}{n} l(r)$ where $l(r)$ depends only on r and d (not not on n). Applying matrix Bernstein ([233, Theorem 6.1]) then gives

$$\mathbb{P}\left(\|E_{n,r}\|_{op} \geq t \right) \leq 2r \exp\left(-\frac{n}{2l(r)} \frac{t^2}{1 + t/3} \right),$$

where $l(r)$ depends only on r and d. Choose now $t_0 = \frac{4l(r)\log(2r/\eta)}{n}$. As long as $n \geq l(r)\log(2r/\eta)$, $t_0 \leq 4$, and the above bound may be refined to

$$\mathbb{P}\left(\|E_{n,r}\|_{op} \geq t_0 \right) \leq 2r \exp\left(-\frac{n}{l(r)} \frac{t_0^2}{7} \right).$$

With the above conditions, it may now be verified that $2r \exp\left(-\frac{n}{l(r)}\frac{t_0^2}{7}\right) \leq \eta$, and the proof is complete. □

We now show that as r increases, the eigenvectors of M_n^r converge to those of M_n. Order the eigenvalues in decreasing order $\lambda_{l_0} \geq \lambda_{l_1} \geq \lambda_{l_2} \geq \ldots$ and let $\Lambda_{>r} = \sum_{i=r}^\infty \lambda_{l_i}^2$. Note that it follows from assumption A2 that $\Lambda_{>r} = O(r^{-3})$. We will denote by $\lambda_i(M_n), \lambda_i(M_n^r)$ the respective eigenvalues of M_n and M_n^r, ordered in a decreasing way, and by $v_i(M_n), v_i(M_n^r)$

their corresponding unit eigenvectors. Suppose that s is such that

$$\lambda(\varphi) = \lambda_{l_{s+1}} = \cdots = \lambda_{l_{s+d}}. \tag{9.7}$$

Moreover, define

$$V_n := \operatorname{span}(v_{l_{s+1}}(M_n), ..., v_{l_{s+d}}(M_n)), \quad V_n^r := \operatorname{span}\left(v_{l_{s+1}}(M_n^r), ..., v_{l_{s+d}}(M_n^r)\right).$$

The next lemma shows that V_n is close to V_n^r whenever both n and r are large enough.

Lemma 9.9. *For all n, r, let $P_{n,r}$ be the orthogonal projection onto V_n^r. Then, for all $\eta > 0$ there exist constants n_0, r_0 such that for all $n > n_0$ and $r > r_0$, we have with probability at least $1 - \eta$ that, for all $w \in V_n$,*

$$\|w - P_{n,r}w\|_2 \leq \frac{4C}{\eta\delta^2 r^3}.$$

where δ and C are the constants from Assumption A1 and Assumption A2.

Proof. We have

$$\begin{aligned}
\mathbb{E}\,\|M_n - M_n^r\|_F^2 &= \sum_{i,j} \mathbb{E}(M_n - M_n^r)_{i,j}^2 \\
&= \sum_{i,j} \frac{1}{n^2} \mathbb{E}\left(\sum_{k=r}^{\infty} \lambda_k \psi_k(X_i)\psi_k(X_j)\right)^2 \\
&= \mathbb{E}_{x,y\sim\sigma}\left(\sum_{k=r}^{\infty} \lambda_k \psi_k(x)\psi_k(y)\right)^2 \\
&= \sum_{k=r}^{\infty} \lambda_k^2 = \Lambda_{>r}.
\end{aligned}$$

Applying Markov's inequality gives that with probability $1 - \eta$

$$\|M_n - M_n^r\|_F^2 \leq \frac{\Lambda_{>r}}{\eta} \leq \frac{C}{\eta r^3}. \tag{9.8}$$

Theorem 1 in [235] shows that there exists n large enough such that with probability larger than $1 - \eta$, one has

$$|\lambda_i(M_n) - \lambda_{l_{s+i}}| \leq \delta/4,$$

with δ being the constant from Assumption A1. It follows that

$$\lambda_{l_{s+1}}(M_n), ..., \lambda_{l_{s+d}}(M_n) \in \left[\lambda(\varphi) - \frac{\delta}{4}, \lambda(\varphi) + \frac{\delta}{4}\right], \tag{9.9}$$

while by (9.8) and Weyl's Perturbation Theorem (e.g., [36, Corollary III.2.6]), for r large

enough with probability $1-\eta$,

$$\lambda_i(M_n^r) \notin \left[\lambda(\varphi) - \frac{3\delta}{4}, \lambda(\varphi) + \frac{3\delta}{4}\right], \text{ for } i \neq l_{s+1}, \dots l_{s+d}. \tag{9.10}$$

Combining (9.8), (9.9) and (9.10) it follows from the classical Davis-Kahan theorem (see e.g. [36, Section VII.3]) that with probability at-least $1-2\eta$, for every $w \in V_n$,

$$\|w - P_{n,r}w\|_2^2 \leq \frac{4C}{\eta\delta^2 r^3}.$$

□

Denote

$$G_n := \frac{1}{d}\sum_{k=1}^{d} v_{l_{s+k}}(M_n)v_{l_{s+k}}(M_n)^T, \quad (G_n')_{i,j} = \frac{1}{n}\langle X_i, X_j\rangle.$$

A combination of the last two lemmas produces the following:

Theorem 9.10. *One has*

$$\|G_n - G_n'\|_F \to 0$$

in probability, as $n \to \infty$.

Proof. Denote

$$G_n^r := \frac{1}{d}\sum_{k=1}^{d} v_{l_{s+k}}(M_n^r)(v_{l_{s+k}}(M_n^r))^T.$$

Then

$$\|G_n - G_n'\|_F \leq \|G_n^r - G_n'\|_F + \|G_n - G_n^r\|_F$$

We will show that the two terms on the right hand side converge to zero. Let $r(n)$ be a function converging to infinity slowly enough so that $C(n,r) \to 0$, for the constant $C(n,r)$ defined in Lemma 9.8. Taking $\eta = \eta(n)$ to converge to zero slowly enough and applying Lemma 9.8, gives for all $2 \leq i \leq d+1$,

$$\|(M_n^r - \lambda(\varphi))u_i\|_2^2 \leq \varepsilon_n$$

with u_i the i'th column of Ψ_n^r and where $\varepsilon_n \to 0$ as $n \to \infty$. Now, if we write

$$u_i = \sum_{j=0}^{\infty} \alpha_{i,j} v_j(M_n^r),$$

the last inequality becomes

$$\sum_j |\lambda_j(M_n^r) - \lambda(\varphi)|^2 \alpha_{i,j}^2 = \sum_j |(M_n^r - \lambda(\varphi))v_j(M_n^r)|^2 \alpha_{i,j}^2 \leq \varepsilon_n, \ \ \forall 2 \leq i \leq d+1.$$

Using Equation (9.10), we have

$$\sum_{j\notin\{l_{s+1},..,l_{s+d}\}} \alpha_{i,j}^2 \leq \frac{4\varepsilon_n}{\delta} \to 0, \tag{9.11}$$

and thus

$$\left\| u_i - \sum_{k=1}^{d} \alpha_{i,l_{s+k}} v_{l_{s+k}}(M_n^r) \right\|_2^2 \to 0, \quad \forall 2 \leq i \leq d+1.$$

Define a $d \times d$-matrix B by $B_{i,j} = \alpha_{i,l_{s+j}}$. Then we can rewrite the above as

$$\left\| (u_1, \ldots, u_d) - (v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r)) \cdot B \right\|_F^2 \to 0.$$

Now, since for two $n \times d$ matrices R, S we have $\|RR^T - SS^T\|_F \leq (\|R\|_{op} + \|S\|_{op})\|R - S\|_F$. It follows that

$$\|G_n' - (v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r))BB^T(v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r))^T\|_F \to 0. \tag{9.12}$$

Observe that

$$B_{i,j} = \langle u_i, u_j \rangle - \sum_{k\notin\{l_{s+1},..,l_{s+d}\}} \alpha_{i,k}\alpha_{j,k},$$

implying that

$$|(BB^T)_{i,j} - E_{i,j}| \leq \sqrt{\sum_{k\notin\{l_{s+1},..,l_{s+d}\}} \alpha_{i,k}^2 \sum_{k\notin\{l_{s+1},..,l_{s+d}\}} \alpha_{j,k}^2} \overset{(9.11)}{\longrightarrow} 0.$$

where $E = E_{n,r}$. Consequently we have

$$\|BB^T - \mathrm{Id}_d\|_{op} \leq C(n,r) \to 0,$$

which implies that

$$\|v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r))(BB^T - \mathrm{Id}_d)(v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r))^T\|_F^2 \to 0.$$

Combining with (9.12) finally yields

$$\|G_n' - G_n^r\|_F \to 0.$$

in probability, as $n \to \infty$.

If P is the orthogonal projection onto $V_n^r = \mathrm{span}(v_{l_{s+1}}(M_n^r), \ldots, v_{l_{s+d}}(M_n^r))$, and Q is the orthogonal projection onto $\mathrm{span}(v_{l_{s+1}}(M_n), \ldots, v_{l_{s+d}}(M_n))$, then Lemma 9.9 shows that for all $\eta > 0$, with probability at least $1 - \eta$, as $n \to \infty$ (and $r = n^{\frac{1}{2d}} \to \infty$), we have for every unit

vector v

$$|(P - \mathrm{Id})Qv| \leq \varepsilon_n \tag{9.13}$$

with some $\varepsilon_n \to 0$. By symmetry, we also have for every unit vector v that

$$|(Q - \mathrm{Id})Pv| \leq \varepsilon_n$$

(this uses that fact that both P and Q are projections into subspaces of the same dimension). The last two inequalities easily yield that $\|P - Q\|_{op} \leq \varepsilon_n$. Since this is true for every $\eta > 0$, it follows that

$$\|G_n - G_n^r\|_F \to 0,$$

in probability, as $n \to \infty$. □

Now, after establishing that M_n is close to Ψ_n^{d+1}, the second step is to recover M_n (and therefore Ψ_n^{d+1}), from the observed A. For the proof we will need the following instance of the Davis-Kahan theorem.

Theorem 9.11 ([249, Theorem 2]). *Let X, Y be symmetric matrices with eigenvalues $\lambda_0 \geq \cdots \geq \lambda_p$ resp. $\hat{\lambda}_0 \geq \cdots \geq \hat{\lambda}_p$ with corresponding orthonormal eigenvectors $v_0, \ldots, v_p$ resp. $\hat{v}_0, \ldots, \hat{v}_p$. Let $V = (v_{s+1}, \ldots, v_{s+d})$ and $\hat{V} = (\hat{v}_{s+1}, \ldots, \hat{v}_{s+d})$. Then there exists an orthogonal $d \times d$ matrix R such that*

$$\|\hat{V}R - V\|_F \leq \frac{2^{3/2} \min(d^{1/2}\|Y - X\|_{op}, \|Y - X\|_F)}{\min(\lambda_s - \lambda_{s+1}, \lambda_{s+d} - \lambda_{s+d+1})}.$$

Our main tool to pass from the expectation of the adjacency matrix to the matrix itself is the following result regarding concentration of random matrices, which follows from [157, Theorem 5.1].

Theorem 9.12. *Let A be the adjacency matrix of a random graph drawn from $G\left(n, \frac{1}{n}\varphi(X_i, X_j)\right)$. Consider any subset of at most $10n/\|\varphi\|_\infty$ vertices, and reduce the weights of the edges incident to those vertices in an arbitrary way but so that all degrees of the new (weighted) network become bounded by $2\|\varphi\|_\infty$. Then with probability at least $1 - n^{-1}$ the adjacency matrix A' of the new weighted graph satisfies*

$$\|A' - M_n\| \leq C\sqrt{\|\varphi\|_\infty}.$$

We can now prove the main reconstruction theorem.

Proof of Theorem 9.2. Let A be the adjacency matrix of a random graph drawn from the model $G\left(n, \frac{1}{n}\varphi(X_i, X_j)\right)$. We first claim that with probability tending to 1, there exists a re-weighted

adjacency matrix A' as defined in Theorem 9.12. Indeed by Chernoff inequality for have for all $i \in [n]$,

$$\mathbb{P}(d_i > 1.5\|\varphi\|_\infty) \leq e^{-cn}$$

and therefore, by Markov's inequality, the expectation of the number of vertices whose degree exceeds $2\|\varphi\|_\infty$ goes to zero with n.

Denote by $\lambda'_0 \geq \lambda'_1 \geq \dots$ its eigenvalues and by $v'_0, v'_1, \dots$ the corresponding orthonormal eigenvectors of A'. Let $Y = (v'_{l_{s+1}}, \dots, v'_{l_{s+d}})$. By Theorem 9.11 there exists an $R \in \mathcal{O}(d)$ such that

$$\left\| \left(v_{l_{s+1}}(M_n), ..., v_{l_{s+d}}(M_n)\right) - YR \right\|_F \leq \frac{2^{3/2} d^{1/2} \|M_n - A'\|_{op}}{\min_{i \neq 1,\dots,d} |\lambda_i - \lambda(\varphi)|}.$$

Hence by Theorem 9.12 we have

$$\left\| \left(v_{l_{s+1}}(M_n), ..., v_{l_{s+d}}(M_n)\right) - YR \right\|_F \leq C\sqrt{d} \cdot \frac{\sqrt{\|\varphi\|_\infty}}{\min_{i \neq 1,\dots,d} |\lambda_i - \lambda(\varphi)|},$$

with probability $1 - n^{-1}$. It follows that

$$\left\| G_n - YY^T \right\|_F \leq C\sqrt{d} \cdot \frac{\sqrt{\|\varphi\|_\infty}}{\min_{i \neq 1,\dots,d} |\lambda_i - \lambda(\varphi)|}$$

Combining this with Theorem 9.10 yields

$$\left\| G'_n - YY^T \right\|_F \leq C\sqrt{d} \cdot \frac{\sqrt{\|\varphi\|_\infty}}{\min_{i \neq 1,\dots,d} |\lambda_i - \lambda(\varphi)|},$$

So,

$$\frac{1}{n^2} \sum_{i,j} \left| X_i \cdot X_j - (nYY^T)_{i,j} \right|^2 \leq Cd \frac{\|\varphi\|_\infty}{\min_{i \neq 1,\dots,d} |\lambda_i - \lambda(\varphi)|^2}$$

which gives the desired reconstruction bound.

□

9.3 Lower bounds

Our approach to proving lower bounds will be to exploit some symmetries which are inherent to well behaved kernel functions. We thus make the following definition:

Definition 9.13 (DPS property)**.** Let μ be a probability measure on $\mathbb{S}^{d-1}$, and let $w \in \mathbb{S}^{d-1}$. We say that μ has the Diminishing Post-translation Symmetry (DPS) around w property with

constant ε, if there exists a decomposition,

$$\mu = (1-\varepsilon)\mu_w + \varepsilon\mu_w^s.$$

Here μ_w^s is a probability measure, invariant to reflections with respect to $w^\perp$. In other words, if $R = \mathrm{I}_d - 2ww^T$, $R_*\mu_w^s = \mu_w^s$. For such a measure we denote $\mu \in \mathrm{DPS}_w(\varepsilon)$.

If instead μ is a measure on $(\mathbb{S}^{d-1})^{|T_k|}$, we say that $\mu \in \mathrm{DPS}_w^k(\varepsilon)$ if a similar decomposition exists but now the reflections should be applied to each coordinate separately.

We now explain the connection between the DPS property and percolation of information in trees. For this, let us recall the random function $g : T \to \mathbb{S}^{d-1}$, introduced in Section 9.1.2, which assigns to the root, r, a uniformly random value and for any other $u \in T$, the label $g(u)$ is distributed according to $\varphi(g(\mathrm{parent}(u)), \cdot)d\sigma$.

Lemma 9.14. *Suppose that there exist a sequence $(p_k)_k$ with $\lim_{k\to\infty} p_k = 1$ such that for every $w, x_0 \in \mathbb{S}^{d-1}$ and every $k > 0$,*

$$\mathrm{Law}((g(v))_{v\in T_k} | g(r) = x_0) \in \mathrm{DPS}_w^k(p_k).$$

Then there is zero percolation of information along T.

Proof. Denote $X = g(r)$ and $Y = g(v)_{v\in T_k}$ and let $\rho_{X|Y}$ be the density of $X|Y$ with respect to σ. Our aim is to show that $\mathbb{E}_Y\left[\mathrm{TV}(X|Y,\sigma)\right] = o(1)$. We first claim that it is enough to show that for all $x, x' \in \mathbb{S}^{d-1}$ and all $\delta > 0$ one has,

$$\mathbb{P}\left(\frac{\rho_{X|Y}(x)}{\rho_{X|Y}(x')} - 1 \le \delta \,\middle|\, X = x\right) = 1 - o(1). \tag{9.14}$$

Indeed, let $H = \left\{\frac{\rho_{X|Y}(x)}{\rho_{X|Y}(x')} - 1 \le \delta\right\}$. Note that, by Bayes' rule,

$$\frac{\rho_{X|Y}(x)}{\rho_{X|Y}(x')} = \frac{\rho_{Y|X=x}(Y)}{\rho_{Y|X=x'}(Y)}.$$

Let $x \in \mathbb{S}^{d-1}$ be some fixed point and let x' be uniformly distributed is $\mathbb{S}^{d-1}$ and independent from Y. We will use $\mathbb{E}_{x'}$ to denote integration with respect to x', which has density ρ_X. Similarly, $\mathbb{E}_Y$ stands for integration with respect to the density ρ_Y of Y and $\mathbb{E}_{x',Y}$ is with respect to the product $\rho_X \cdot \rho_Y$. The symbol $\mathbb{P}$ will always mean probability over x' and Y.

As a consequence of the assumption (9.14), there exists a function, as long as k, is large enough we have,

$$\mathbb{P}\left(H | X = x\right) = 1 - \delta.$$

as long as k is large enough. For any measurable set A, we denote the projection:

$$\Pi_Y A = \{y \in Y : (x', y) \in A \text{ for some } x' \in X\}.$$

Define now the set

$$H' := \{(x', y) | (x', y) \in H \text{ and } \mathbb{E}_{x'}[\mathbf{1}_H(\cdot, y)] \geq 1 - 4\delta\}.$$

In particular, for every $y \in \Pi_Y H'$,

$$\mathbb{E}_{x'}[\mathbf{1}_H(\cdot, y)] \geq 1 - 4\delta,$$

and by Fubini's theorem, $\mathbb{P}(H'|X = x) \geq 1 - 4\delta$.

Consider the random variables $\alpha := \alpha(Y) = \frac{\rho_{Y|X=x}(Y)}{\rho_Y(Y)}$ and $\beta := \beta(x', Y) = \frac{\rho_{Y|X=x'}(Y)}{\rho_Y(Y)}$. By definition of H , we have

$$(1 - \delta)\mathbf{1}_H \alpha \leq \mathbf{1}_H \beta. \tag{9.15}$$

Moreover, for almost every Y,

$$\mathbb{E}_{x'}[\beta] = \frac{1}{\rho_Y(Y)} \int \rho_{Y|X=x'}(Y)\rho_X(x') = \frac{1}{\rho_Y(Y)}\rho_Y(Y) = 1.$$

So,

$$(1 - \delta)(1 - 4\delta)\mathbb{E}_Y[\alpha \mathbf{1}_{\Pi_Y H'}] \leq (1 - \delta)\mathbb{E}_Y[\alpha \mathbb{E}_{x'}[\mathbf{1}_{H'}]] \leq \mathbb{E}_{x',Y}[\beta \mathbf{1}_{H'}] \leq \mathbb{E}_{x',Y}[\beta] = 1.$$

Observe that $\mathbb{E}_Y[\alpha \mathbf{1}_{\Pi_Y H'}] = \mathbb{P}(\Pi_Y H'|X = x) = 1 - o(1)$, where the second equality is a consequence of Fubini's theorem. Hence, let us write $h_1(\delta)$, so that

$$1 - h_1(\delta) = (1 - \delta)(1 - 4\delta)\mathbb{E}_Y[\alpha \mathbf{1}_{\Pi_Y H'}] \leq (1 - \delta)\mathbb{E}_{x',Y}[\alpha \mathbf{1}_{H'}].$$

Markov's inequality then implies,

$$\mathbb{P}\left(\beta \mathbf{1}_{H'} \geq (1 - \delta)\alpha \mathbf{1}_{H'} + \sqrt{h_1(\delta)}\right) \leq \frac{\mathbb{E}_{x',Y}[\beta \mathbf{1}_{H'}] - (1 - \delta)\mathbb{E}_{x',Y}[\alpha \mathbf{1}_{H'}]}{\sqrt{h_1(\delta)}} \leq \sqrt{h_1(\delta)}. \tag{9.16}$$

Now, we integrate over x' to obtain,

$$(1 - \delta)(1 - 4\delta)\alpha \mathbf{1}_{\Pi_Y H'} \leq (1 - \delta)\alpha \mathbb{E}_{x'}[\mathbf{1}_{H'}] \leq \mathbb{E}_{x'}[\beta \mathbf{1}_{H'}] \leq 1,$$

which implies

$$\alpha \mathbf{1}_{H'} \leq \frac{1}{(1 - \delta)(1 - 4\delta)}. \tag{9.17}$$

Keeping in mind the previous displays, we may choose $h_2(\delta)$, which satisfies $\lim_{\delta\to 0} h_2(\delta) = 0$, $\alpha \mathbf{1}_{H'} \leq 1 + h_2(\delta)$ and $\mathbb{E}_{x',Y}\left[\alpha \mathbf{1}_{H'}\right] \geq 1 - h_2(\delta)$.

So, an application of the the reverse Markov's inequality for bounded and positive random variables shows,

$$\mathbb{P}\left(\alpha \mathbf{1}_{H'} \geq 1 - \sqrt{h_2(\delta)}\right) \geq \frac{\mathbb{E}_{x',Y}\left[\alpha \mathbf{1}_{H'}\right] - (1 - \sqrt{h_2(\delta)})}{1 + h_2(\delta) - (1 - \sqrt{h_2(\delta)})} \geq \frac{1 - h_2(\delta) - (1 - \sqrt{h_2(\delta)})}{1 + h_2(\delta) - (1 - \sqrt{h_2(\delta)})}$$
$$= \frac{\sqrt{h_2(\delta)} - h_2(\delta)}{\sqrt{h_2(\delta)} + h_2(\delta)}. \tag{9.18}$$

Note that the as $\delta \to 0$ the RHS goes to 1. Thus, by combining the above displays, there exists a function h, which satisfies $\lim_{\delta\to 0} h(\delta) = 0$ and some $H'' \subset H'$, with $\mathbb{P}(H') \geq 1 - h(\delta)$, such that, by (9.17) and (9.18),

$$\mathbf{1}_{H''}|\alpha - 1| \leq h(\delta),$$

which implies, together with (9.15) and (9.16)

$$\mathbf{1}_{H''}|\alpha - \beta| \leq h(\delta).$$

This then gives

$$\mathbf{1}_{H''}|1 - \beta| \leq 2h(\delta).$$

We can thus conclude,

$$\begin{aligned}
\mathbb{E}_Y \mathrm{TV}\left(X|Y, X\right) &= \mathbb{E}_{Y,x'}\left[|\beta - 1|\right] = 2\mathbb{E}_{Y,x'}\left[(1-\beta)\mathbf{1}_{\beta\leq 1}\right] \\
&\leq 2\mathbb{E}_{Y,x'}\left[(1-\beta)\mathbf{1}_{\beta\leq 1}\mathbf{1}_{H''}\right] + 1 - \mathbb{P}\left(H''\right) \\
&= 2\mathbb{E}_{Y,x'}\left[|1-\beta|\mathbf{1}_{H''}\right] + h(\delta) \\
&\leq 5h(\delta).
\end{aligned}$$

Take now $\delta \to 0$ to get $\mathbb{E}_Y \mathrm{TV}\left(X|Y, X\right) \to 0$.

Thus, we may assume towards a contradiction that there exist $x, x' \in \mathbb{S}^{d-1}$ and a set $F \subset (\mathbb{S}^{d-1})^{|T_k|}$, such that

$$\mathbb{P}\left(Y \in F | X = x\right) \geq \delta, \tag{9.19}$$

and under $\{Y \in F\}$,

$$\frac{\rho_{X|Y}(x)}{\rho_{X|Y}(x')} \geq 1 + \delta, \tag{9.20}$$

for some constant $\delta > 0$.

Let $w \in \mathbb{S}^{d-1}$ be such that the reflection $R := \mathrm{I}_d - 2ww^T$ satisfies $Rx = x'$. Under our assumption, there exists an event A_k, which satisfies

$$\mathbb{P}(A_k|X = x) = 1 - o(1)$$

and such that $Y|(X = x, A_k)$ is R-invariant. By (9.19), we also have

$$\mathbb{P}(A_k|X = x, Y \in F) = 1 - o(1),$$

and thus there exists $y \in F$ such that

$$\mathbb{P}(A_k|X = x, Y = y) = 1 - o(1). \tag{9.21}$$

By continuity, we can make sense of conditioning on the zero probability event $E := \big\{X = x,\ Y \in \{y, Ry\}\big\}$. Note that we have by symmetry and since $y \in F$,

$$Y = y \ \ \Rightarrow \ \ 1 + \delta \leq \frac{\rho_{X|Y}(x)}{\rho_{X|Y}(x')} = \frac{\mathbb{P}(Y = y|E)}{\mathbb{P}(Y = Ry|E)}. \tag{9.22}$$

On the other hand, we have by definition of A_k,

$$\mathbb{P}(Y = Ry|E, A_k) = \mathbb{P}(Y = y|E, A_k),$$

which implies that

$$\begin{aligned}\mathbb{P}(Y = Ry|E) &\geq \mathbb{P}\big(\{Y = Ry\} \cap A_k|E\big)\\ &= \mathbb{P}\big(\{Y = y\} \cap A_k|E\big)\\ &\geq \mathbb{P}(Y = y|E)(1 - o(1))\end{aligned}$$

which contradicts (9.22). The proof is complete. □

9.3.1 The Gaussian case

Our aim is to show that certain classes of kernel functions satisfy the DPS condition. We begin by considering the case where the kernel φ is Gaussian, as in (9.6). In this case, the function g may be defined as follows. Let T be a q-ary tree. To each edge $e \in E(T)$ we associate a Brownian motion $(B_e(t))_{t\in(0,1)}$ of rate β such that for every node $v \in V$ we have

$$g(v) = \sum_{e\in P(v)} B_e(1) \bmod 2\pi$$

where $P(v)$ denotes the shortest path from the root to v.

For every node $v \in T_k$ let us now consider the Brownian motion $(B_v(t))_{t=0}^k$ defined by concatenating the Brownian motions B_e along the edges $e \in P(v)$. Define by E_v the event that the image of $B_v \mathrm{mod} \pi$ contains the entire interval $[0, \pi)$, and define $E_k = \bigcap_{v \in T_k} E_v$. Our lower bound relies on the following observation.

Claim 9.15. *Fix $v \in T_k$ and set $p_k := \mathbb{P}(E_k)$. Then, for every $\theta, x_0 \in \mathbb{S}^1$, $\mathrm{Law}(g(v)|g(r) = x_0) \in \mathrm{DPS}_\theta^k(p_k)$.*

Proof. Fix $\theta \in [0, \pi)$. Given the event E_v, we have almost surely that there exists a time $t_v \leq k$ such that $B_v(t_v) \in \{\theta - \pi, \theta\}$. By symmetry and by the Markov property of Brownian motion, we have that the distribution of $g(v)$ conditioned on the event $\{B_v(t_v) = \theta, \ \forall v\}$ is symmetric around θ. Thus, by considering $(B_v(t))_{v \in T_k}$, under the event $\{t_v \leq k | v \in T_k\}$ we get that for any x_0, $\mathrm{Law}((g(v))_{v \in T_k} | g(r) = x_0)$ is symmetric around θ. So,

$$\mathrm{Law}((g(v))_{v \in T_k} | g(r) = x_0) \in \mathrm{DPS}_\theta^k(p_k).$$

□

We will also need the following bound, shown for example in [107].

Lemma 9.16. *Let $B(t)$ be a Brownian motion of rate β on the unit circle. Then,*

$$\mathbb{P}(\mathrm{Image}(B(s) \mathrm{mod} \pi)_{0 \leq s \leq t} = [0, \pi)) \geq 1 - Ce^{-t\beta/2}.$$

We are now in a position to prove Theorem 9.5.

Proof of Theorem 9.5. Lemma 9.16 immediately implies that for all $v \in T_k$,

$$\mathbb{P}(E_v) \geq 1 - Ce^{-k\beta/2}.$$

On the other hand, a calculation gives $\lambda(\varphi) = \mathbb{E}[\cos(B_1)] = e^{-\beta/2}$. Thus, applying a union bound implies that for some constant $C > 0$, $\mathbb{P}(E_k) \geq 1 - Cq^k \lambda(\varphi)^k$. Hence, by Claim 9.15,

$$\mathrm{Law}((g(v))_{v \in T_k} | g(r) = x_0) \in \mathrm{DPS}_w(1 - Cq^k \lambda(\varphi)^k).$$

The result is now a direct consequence of Lemma 9.14. □

In the next section we generalize the above ideas and obtain a corresponding bound which holds for distributions other than the Gaussian one.

9.3.2 The general case

On symmetric functions

We begin this section with simple criterion to determine whether a measure belongs to some DPS class. In the sequel, for $w \in \mathbb{S}^{d-1}$, we denote,

$$H_w^+ = \{x \in \mathbb{S}^{d-1} | \langle x, w\rangle \geq 0\} \text{ and } H_w^- = \{x \in \mathbb{S}^{d-1} | \langle x, w\rangle < 0\}.$$

Lemma 9.17. *Let $f : [-1, 1] \to \mathbb{R}^+$ satisfy the assumptions of Theorem 9.7 and let $y \in \mathbb{S}^{d-1}$. If $\mu = f(\langle \cdot, y\rangle)d\sigma$, then for any $w \in \mathbb{S}^{d-1}$, $\mu \in \mathrm{DPS}_w(2 \cdot p_w)$, where $p_w = \min(\mu(H_w^+), \mu(H_w^-))$.*

Proof. Without loss of generality let us assume that $y \in H_w^+$. Monotonicity of f implies $\mu(H_w^+) \geq \mu(H_w^-)$. Now, if $R = \mathrm{I}_d - 2ww^T$ is the reflection matrix with respect to $w^\perp$, then we have for any $x \in H_w^-$,

$$f(\langle x, y\rangle) \leq f(\langle Rx, y\rangle).$$

This follows since $\langle x, y\rangle \leq \langle Rx, y\rangle$.

Let us now define the measure $\tilde{\mu}_w^s$ such that

$$\frac{d\tilde{\mu}_w^s}{d\sigma}(x) = \begin{cases} f(\langle x, y\rangle) & \text{if } x \in H_w^- \\ f(\langle Rx, y\rangle) & \text{if } x \in H_w^+. \end{cases}$$

$\tilde{\mu}_w^s$ is clearly R-invariant and the above observation shows that $\tilde{\mu}_w^s(\mathbb{S}^{d-1}) \leq 1$. We can thus define $\tilde{\mu}_w = \mu - \tilde{\mu}_w^s$.
To obtain a decomposition, define $\mu_w^s = \frac{\tilde{\mu}_w^s}{\tilde{\mu}_w^s(\mathbb{S}^{d-1})}$ and $\mu_w = \frac{\tilde{\mu}_w}{\tilde{\mu}_w(\mathbb{S}^{d-1})}$, for which

$$\mu = (1 - \tilde{\mu}_w^s(\mathbb{S}^{d-1}))\mu_w + \tilde{\mu}_w^s(\mathbb{S}^{d-1})\mu_w^s.$$

The proof is concluded by noting $\tilde{\mu}_w^s(\mathbb{S}^{d-1}) = 2\mu(H_w^-)$.

□

Our main object of interest will be the measure $\mu_\varphi(y)$ which, for a fixed y, is defined by

$$\mu_\varphi(y) := \varphi(x, y)d\sigma(x) = f(\langle x, y\rangle)d\sigma(x). \tag{9.23}$$

Let us now denote,

$$\beta_d(t) := \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})}(1 - t^2)^{(d-3)/2},$$

which is the marginal of the uniform distribution on the sphere. We now show that the spectral gap of φ may determine the DPS properties of $\mu_\varphi(y)$.

Lemma 9.18. *Let $w \in \mathbb{S}^{d-1}$ and suppose that f is monotone. If $|\langle w, y\rangle| \leq \frac{1-\lambda(\varphi)}{16\sqrt{d}}$, then*

$$\mu_\varphi(y) \in \mathrm{DPS}_w\left(\frac{1-\lambda(\varphi)}{35}\right).$$

Proof. Assume W.L.O.G. that $\langle y, w\rangle > 0$. By Lemma 9.17 it will be enough to bound $\int_{H_w^-} \mu_\varphi(y)$ from below. Let $X \sim \mu_\varphi(y)$ and define $Z = \langle X, y\rangle$. We have $\mathbb{E}[Z] = \lambda(\varphi)$ and by Markov's inequality,

$$\mathbb{P}\left(Z \leq \frac{1+\lambda(\varphi)}{2}\right) = \mathbb{P}\left(Z+1 \leq \frac{1+\lambda(\varphi)}{2}+1\right) \geq 1 - \frac{2(\lambda(\varphi)+1)}{3+\lambda(\varphi)} \geq \frac{1-\lambda(\varphi)}{4}.$$

For $t \in [-1,1]$, set $S_t = \{x \in \mathbb{S}^{d-1} | \langle x, y\rangle = t\}$ and let $\mathcal{H}^{d-2}$ stand for $d-2$-dimensional Hausdorff measure. To bound $\int_{H_w^-} \mu_\varphi(y)$ we would first like to estimate $\frac{\mathcal{H}^{d-2}(S_t \cap H_w^-)}{\mathcal{H}^{d-2}(S_t)}$.

We know that $0 \leq \langle w, y\rangle \leq \frac{1-\lambda(\varphi)}{16\sqrt{d}}$. Denote $t_y := \langle w, y\rangle$ and fix $t \leq t_0 := \frac{1+\lambda(\varphi)}{2}$. With no loss of generality, let us write $w = e_1$ and $y = t_y e_1 + \sqrt{1-t_y^2}e_2$. Define now $z = -\sqrt{1-t_y^2}e_1 + t_y e_2$. We claim that

$$\left\{v \in S_t \middle| \frac{\langle v, z\rangle}{\sqrt{1-t^2}} \geq \frac{1}{2\sqrt{d}}\right\} \subseteq S_t \cap H_w^-. \tag{9.24}$$

If $v \in S_t$, its projection onto the plane $\mathrm{span}(y, w)$, can be written as $t \cdot y + \sqrt{1-t^2}c \cdot z$, for some $c \in [-1,1]$. So,

$$\langle v, w\rangle = t \cdot t_y - \sqrt{1-t^2}\sqrt{1-t_y^2}c.$$

Now, whenever

$$c > \frac{t \cdot t_y}{\sqrt{1-t^2}\sqrt{1-t_y^2}},$$

we get $\langle v, w\rangle < 0$. Also,

$$\frac{t \cdot t_y}{\sqrt{1-t^2}\sqrt{1-t_y^2}} \leq \frac{1}{\sqrt{3}}\frac{t_y}{\sqrt{1-t_y^2}} \leq \frac{1}{2\sqrt{d}},$$

where we have used $t \leq \frac{1}{2}$ for the first inequality and $t_y \leq \frac{1}{2\sqrt{d}}$ in the second inequality. By combining the above displays with $\frac{\langle v,z\rangle}{\sqrt{1-t^2}} = c$, (9.24) is established.

Thus, by taking the marginal of S_t in the direction of $-z$, we see

$$\frac{\mathcal{H}^{d-2}(S_t \cap H_w^-)}{\mathcal{H}^{d-2}(S_t)} \geq \int_{-1}^{-\frac{1}{2\sqrt{d}}} \beta_{d-1}(s)ds \geq \int_{-\frac{1}{\sqrt{d}}}^{-\frac{1}{2\sqrt{d}}} \beta_{d-1}(s)ds \geq \frac{1}{2\sqrt{d}}\beta_{d-1}\left(\frac{1}{\sqrt{d}}\right)$$
$$\geq \frac{1}{2\sqrt{d}}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2})}\left(1-\frac{1}{d}\right)^{(d-4)/2} \geq \frac{1}{10\sqrt{e}},$$

where we used $\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2})} \geq \frac{\sqrt{d}}{5}$, valid for any $d \geq 3$. We use the above estimates with Fubini's theorem to obtain:

$$\mathbb{P}\left(X \in H_w^-\right) = \int_{-1}^{1} f(t)\mathcal{H}^{d-2}(S_t \cap H_w^-)dt \geq \int_{-1}^{t_0} f(t)\mathcal{H}^{d-2}(S_t \cap H_w^-)dt$$
$$\geq \frac{1}{10\sqrt{e}}\int_{-1}^{t_0} f(t)\mathcal{H}^{d-2}(S_t)dt = \frac{1}{10\sqrt{e}}\mathbb{P}\left(Z \leq \frac{1+\lambda(\varphi)}{2}\right) \geq \frac{1-\lambda(\varphi)}{70}.$$

□

Mixing

Recall the random function $g : T \to \mathbb{S}^{d-1}$, introduced in Section 9.1.2, which assignes to the root, r, a uniformly random value and for any other $u \in T$, the label $g(u)$ is distributed according to $\varphi(g(\text{parent}(u)), \cdot)d\sigma =: \mu_\varphi(\text{parent}(u))$.
Suppose that $v \in T_k$ and let $\{v_i\}_{i=0}^k$ denote the simple path from r to v in T. Fix $x_0 \in \mathbb{S}^{d-1}$, for $i = 0, \ldots, k$, we now regard,

$$X_i := g(v_i)|g(r) = x_0,$$

as a random walk on $\mathbb{S}^{d-1}$. Observe that given X_{i-1}, $X_i \sim \mu_\varphi(X_{i-1})$. The following lemma shows that this random walk is rapidly-mixing.

Lemma 9.19. *For $w \in \mathbb{S}^{d-1}$, let*

$$S(w) = \left\{u \in \mathbb{S}^{d-1} : |\langle u, w\rangle| \leq \frac{1-\lambda(\varphi)}{16\sqrt{d}}\right\},$$

and set $k_0 = \frac{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32f(1)}\right)}{\ln(\lambda(\varphi))}$. Then, if f satisfies the assumptions of Theorem 9.7,

$$\mathbb{P}(X_{k_0} \in S(w)) \geq \frac{1-\lambda(\varphi)}{32}.$$

Proof. Note that if $U \sim \text{Uniform}(\mathbb{S}^{d-1})$, then

$$\mathbb{P}(U \in S(w)) = \int\limits_{|t| \le \frac{1-\lambda(\varphi)}{16\sqrt{d}}} \beta_d(t)dt \ge 2\frac{\Gamma(\frac{d}{2})}{\Gamma\left(\frac{d-1}{2}\right)}\beta_d\left(\frac{1}{16\sqrt{d}}\right)\frac{1-\lambda(\varphi)}{16\sqrt{d}} \ge \frac{1-\lambda(\varphi)}{16}. \quad (9.25)$$

It will then suffice to show that $\mathbb{P}(X_{k_0} \in S(w))$ can be well approximated by $\mathbb{P}(U \in S(w))$. Since X_{k_0} has density $A_\varphi^{k_0-1}f(\langle x, x_0\rangle)$, the following holds true,

$$\begin{aligned}(\mathbb{P}(U \in S(w)) - \mathbb{P}(X_{k_0} \in S(w)))^2 &= \left(\int\limits_{S(w)} \left(A_\varphi^{k_0-1}f(\langle x, x_0\rangle) - 1\right) d\sigma(x)\right)^2 \\ &\le \int\limits_{S(w)} \left(A_\varphi^{k_0-1}f(\langle x, x_0\rangle) - 1\right)^2 d\sigma(x).\end{aligned}$$

We now decompose the density as $f(\langle x, x_0\rangle) = \sum\limits_{i=0}^{\infty} \lambda_i\psi_i(x)\psi_i(x_0)$. So that

$$A_\varphi^{k_0-1}f(\langle x, x_0\rangle) = \sum_{i=0}^{\infty} \lambda_i^{k_0}\psi_i(x)\psi_i(x_0).$$

We know that $\psi_0 \equiv 1$ with eigenvalue $\lambda_0 = 1$, and we've assumed that $|\lambda_i| \le \lambda_1 = \lambda(\varphi)$ for every $i \ge 1$. Thus,

$$\begin{aligned}\int\limits_{S(w)} \left(A_\varphi^{k_0-1}f(\langle x, x_0\rangle) - 1\right)^2 d\sigma(x) &= \int\limits_{S(w)} \left(\sum_{i=1}^{\infty} \lambda_i^{k_0}\psi_i(x)\psi_i(x_0)\right)^2 d\sigma(x) \\ &\le (\lambda(\varphi))^{2k_0-2} \int\limits_{S_\delta(w)} \sum_{i=1}^{\infty}(\lambda_i\psi_i(x))^2 \sum_{i=1}^{\infty}(\lambda_i\psi_i(x_0))^2 d\sigma(x) \\ &\le \lambda(\varphi)^{2k_0-2}f(1)^2.\end{aligned}$$

where in the last inequality we have used $f(1) = \sum \lambda_i\psi_i(y)\psi_i(y)$, which is valid for any $y \in \mathbb{S}^{d-1}$. Thus, since $k_0 = \frac{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32f(1)}\right)}{\ln(\lambda(\varphi))}$, by (9.25), we get,

$$\mathbb{P}(X_{k_0} \in S(w)) \ge \mathbb{P}(U \in S(w)) - \frac{1-\lambda(\varphi)}{32} \ge \frac{1-\lambda(\varphi)}{32}.$$

□

Since the random walk X_k mixes well, we now use Lemma 9.18 to show that after enough steps, X_k will be approximately invariant to a given reflection.

Lemma 9.20. *Let $w, x_0 \in \mathbb{S}^{d-1}$. Then,*

$$\mathrm{Law}((g(v))_{v\in T_k}|g(r)=x_0) \in \mathrm{DPS}_w^k\left(1-q^kp^k\right),$$

where $p = \left(1 - \frac{\ln(\lambda(\varphi))}{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32f(1)}\right)}\frac{(1-\lambda(\varphi))^2}{600}\right)$.

Proof. Let $R = \mathrm{I}_d - 2ww^T$ denote the linear reflection with respect to $w^\perp$. Then, the claim is equivalent to the decomposition,

$$X_k = P\tilde{X}_k + (1-P)X_k^R,$$

where X_k^R is invariant to reflections by R and $P \sim \mathrm{Bernoulli}(s_k)$ is independent from $\{\tilde{X}_k, X_k^R\}$ with $s_k \le \left(1 - \frac{\ln(\lambda(\varphi))}{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32f(1)}\right)}\frac{(1-\lambda(\varphi))^2}{600}\right)^k$.

Consider the case that for some $i = 0, ..., k$, $|\langle X_i, w\rangle| \le \frac{1-\lambda(\varphi)}{16\sqrt{d}}$. In this case, from Lemma 9.18, given X_i, we have the decomposition,

$$\mu_\varphi(X_i) = \left(1 - \frac{(1-\lambda(\varphi))}{35}\right)\mu_{\varphi,w} + \frac{(1-\lambda(\varphi))}{35}\mu^s_{\varphi,w}(X_i),$$

where $\mu^s_{\varphi,w}(X_i)$ is R-invariant.

We now generate the random walk in the following way. For $i = 0, \ldots, k$, let

$$P_i \sim \mathrm{Bernoulli}\left(\frac{(1-\lambda(\varphi))}{35}\right), \tag{9.26}$$

be an *i.i.d* sequence. Given X_i, if $|\langle X_i, w\rangle| > \frac{1-\lambda(\varphi)}{16\sqrt{d}}$ then $X_{i+1} \sim \mu_\varphi(X_i)$.

Otherwise, $|\langle X_i, w\rangle| \le \frac{1-\lambda(\varphi)}{16\sqrt{d}}$. To decide on the position of X_{i+1} we consider P_i. If $P_i = 0$ then $X_{i+1} \sim \mu_{\varphi,w}(X_i)$. Otherwise $P_i = 1$ and we generate $X_{i+1} \sim \mu^s_{\varphi,w}(X_i)$. We denote the latter event by A_i and $A = \cup_{i=0}^{k-1} A_i$.

It is clear that, conditional on A, $RX_k \overset{law}{=} X_k$. Thus, to finish the proof, if $\bar{A}$ is the complement of A, we will need to show

$$\mathbb{P}(\bar{A}) \le \left(1 - \frac{\ln(\lambda(\varphi))}{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32f(1)}\right)}\frac{(1-\lambda(\varphi))^2}{600}\right)^k.$$

Towards this, let $S(w)$ and k_0 be as in Lemma 9.19. Coupled with (9.26), the lemma tells us

that

$$\mathbb{P}\left(A_{k_0}\right) \geq \frac{(1-\lambda(\varphi))^2}{600}.$$

Now, by restarting the random walk from X_{k_0} if needed, we may show,

$$\mathbb{P}\left(\bar{A}\right) \leq \sum_{m \leq \frac{k}{k_0}} \mathbb{P}(\bar{A}_{m \cdot k_0}) \leq \left(1 - \frac{(1-\lambda(\varphi))^2}{600}\right)^{\frac{k}{k_0}} \leq \left(1 - \frac{(1-\lambda(\varphi))^2}{600 k_0}\right)^{k}.$$

Hence $\mathrm{Law}(X_k) \in \mathrm{DPS}(1-p)$ and the claim follows by taking a union bound over all paths. □

Proving Theorem 9.7

Proof. By Lemma 9.20, for every $w, x_0 \in \mathbb{S}^{d-1}$.

$$\mathrm{law}(g(v)_{v \in T_k} | g(r) = x_0) \in \mathrm{DPS}_w^k(1 - q^k p^k),$$

where $p = \left(1 - \frac{\ln(\lambda(\varphi))}{\ln\left(\frac{\lambda(\varphi)(1-\lambda(\varphi))}{32 f(1)}\right)} \frac{(1-\lambda(\varphi))^2}{600}\right)$. By assumption

$$q \leq p^{-1},$$

and Lemma 9.14 gives the result. □

Bibliography

[1] Scott Aaronson. Lower bounds for local search by quantum arguments. *SIAM Journal on Computing*, 35(4):804–824, 2006.

[2] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.*, 18:Paper No. 177, 86, 2017.

[3] Emmanuel Abbe. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.

[4] Kyle Aitken and Guy Gur-Ari. On the asymptotics of wide networks with polynomial activations. *arXiv preprint arXiv:2006.06687*, 2020.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

[6] Andris Ambainis, Aram W. Harrow, and Matthew B. Hastings. Random tensor theory: extending random matrix theory to mixtures of random product states. *Comm. Math. Phys.*, 310(1):25–74, 2012.

[7] Luigi Ambrosio, Elia Bruè, and Dario Trevisan. Lusin-type approximation of Sobolev by Lipschitz functions, in Gaussian and $\mathrm{RCD}(K,\infty)$ spaces. *Adv. Math.*, 339:426–452, 2018.

[8] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15:2773–2832, 2014.

[9] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra Appl.*, 26:203–241, 1979.

[10] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International conference on machine learning*, pages 1908–1916, 2014.

[11] Anders Andreassen and Ethan Dyer. Asymptotics of wide convolutional neural networks. *arXiv preprint arXiv:2008.08675*, 2020.

[12] Joseph M Antognini. Finite size corrections for neural network gaussian processes. *arXiv preprint arXiv:1908.10030*, 2019.

[13] Milla Anttila, Keith Ball, and Irini Perissinaki. The central limit problem for convex bodies. *Trans. Amer. Math. Soc.*, 355(12):4723–4735, 2003.

[14] Ernesto Araya Valdivia and Yohann De Castro. Latent distance estimation for random geometric graphs. In *Advances in Neural Information Processing Systems 32*, pages 8724–8734. Curran Associates, Inc., 2019.

[15] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2019.

[16] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8141–8150, 2019.

[17] Shiri Artstein, Keith M. Ball, Franck Barthe, and Assaf Naor. On the rate of convergence in the entropic central limit theorem. *Probab. Theory Related Fields*, 129(3):381–390, 2004.

[18] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[19] Dominique Bakry. Étude des transformations de Riesz dans les variétés riemanniennes à courbure de Ricci minorée. In *Séminaire de Probabilités, XXI*, volume 1247 of *Lecture Notes in Math.*, pages 137–172. Springer, Berlin, 1987.

[20] Keith Ball, Franck Barthe, and Assaf Naor. Entropy jumps in the presence of a spectral gap. *Duke Math. J.*, 119(1):41–63, 2003.

[21] Keith Ball and Van Hoang Nguyen. Entropy jumps for isotropic log-concave random vectors and spectral gap. *Studia Math.*, 213(1):81–96, 2012.

[22] Vlad Bally, Lucia Caramellino, et al. Asymptotic development for the CLT in total variation distance. *Bernoulli*, 22(4):2442–2485, 2016.

[23] Marco Barchiesi, Alessio Brancolini, and Vesa Julin. Sharp dimension free quantitative estimates for the Gaussian isoperimetric inequality. *Ann. Probab.*, 45(2):668–697, 2017.

[24] Andrew R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, 1986.

[25] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[26] Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[27] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[28] Fabrice Baudoin. Bakry-Émery meet Villani. *J. Funct. Anal.*, 273(7):2275–2291, 2017.

[29] Eric B Baum. On the capabilities of multilayer perceptrons. *Journal of complexity*, 4(3):193–215, 1988.

[30] Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in neural information processing systems*, pages 123–130, 2006.

[31] Itai Benjamini, Robin Pemantle, and Yuval Peres. Unpredictable paths and percolation. *The Annals of Probability*, 26(3):1198–1211, 1998.

[32] V. Bentkus. A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.*, 49(2):400–410, 2004.

[33] Harald Bergström. On the central limit theorem in the space $R_k, k > 1$. *Skand. Aktuarietidskr.*, 28:106–127, 1945.

[34] Robert J. Berman and Bo Berndtsson. Real Monge-Ampère equations and Kähler-Ricci solitons on toric log Fano varieties. *Ann. Fac. Sci. Toulouse Math. (6)*, 22(4):649–711, 2013.

[35] Andrew C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Amer. Math. Soc.*, 49:122–136, 1941.

[36] Rajendra Bhatia. *Matrix Analysis*. Springer-Verlag, New York, 1997.

[37] R. N. Bhattacharya. Refinements of the multidimensional central limit theorem and applications. *Ann. Probability*, 5(1):1–27, 1977.

[38] Carlo Bianca and Christian Dogbe. On the existence and uniqueness of invariant measure for multidimensional diffusion processes. *Nonlinear Stud.*, 24(3):[437–468?], 2017. Paging given as 1–32 in print version.

[39] S. G. Bobkov, N. Gozlan, C. Roberto, and P.-M. Samson. Bounds on the deficit in the logarithmic Sobolev inequality. *J. Funct. Anal.*, 267(11):4110–4138, 2014.

[40] S. G. Bobkov and A. Koldobsky. On the central limit property of convex bodies. In *Geometric aspects of functional analysis*, volume 1807 of *Lecture Notes in Math.*, pages 44–52. Springer, Berlin, 2003.

[41] Sergey G. Bobkov. Entropic approach to E. Rio's central limit theorem for W_2 transport distance. *Statist. Probab. Lett.*, 83(7):1644–1648, 2013.

[42] Sergey G. Bobkov. Berry-Esseen bounds and Edgeworth expansions in the central limit theorem for transport distances. *Probab. Theory Related Fields*, 170(1-2):229–262, 2018.

[43] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem. *Ann. Probab.*, 41(4):2479–2512, 2013.

[44] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Berry-Esseen bounds in the entropic central limit theorem. *Probab. Theory Related Fields*, 159(3-4):435–478, 2014.

[45] V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov. Distances between transition probabilities of diffusions and applications to nonlinear Fokker-Planck-Kolmogorov equations. *J. Funct. Anal.*, 271(5):1262–1300, 2016.

[46] V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov. The Poisson equation and estimates for distances between stationary distributions of diffusions. *J. Math. Sci. (N.Y.)*, 232(3, Problems in mathematical analysis. No. 92 (Russian)):254–282, 2018.

[47] Thomas Bonis. Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probab. Theory Related Fields*, 178(3-4):827–860, 2020.

[48] Christer Borell. The Ehrhard inequality. *C. R. Math. Acad. Sci. Paris*, 337(10):663–666, 2003.

[49] A. A. Borovkov and S. A. Utev. An inequality and a characterization of the normal distribution connected with it. *Teor. Veroyatnost. i Primenen.*, 28(2):209–218, 1983.

[50] Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

[51] Mikio L. Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *J. Mach. Learn. Res.*, 7:2303–2328, December 2006.

[52] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

[53] Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *arXiv preprint arXiv:1910.14167*, 2019.

[54] Guy Bresler and Dheeraj Nagaraj. A corrective view of neural networks: Representation, memorization and learning. *arXiv preprint arXiv:2002.00274*, 2020.

[55] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[56] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms*, 49(3):503–532, 2016.

[57] Sebastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and size of the weights in memorization with two-layers neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[58] Sébastien Bubeck and Shirshendu Ganguly. Entropic CLT and phase transition in high-dimensional Wishart matrices. *Int. Math. Res. Not. IMRN*, 1(2):588–606, 2018.

[59] Sebastien Bubeck, Qijia Jiang, Yin-Tat Lee, Yuanzhi Li, and Aaron Sidford. Complexity of highly parallel non-smooth convex optimization. In *Advances in Neural Information Processing Systems 32*, pages 13900–13909. 2019.

[60] Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In *Conference on Learning Theory*, pages 940–960. PMLR, 2020.

[61] Peter Buser. A note on the isoperimetric constant. *Ann. Sci. École Norm. Sup. (4)*, 15(2):213–230, 1982.

[62] L. A. Caffarelli. A localization property of viscosity solutions to the Monge-Ampère equation and their strict convexity. *Ann. of Math. (2)*, 131(1):129–134, 1990.

[63] Luis A. Caffarelli. The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, 5(1):99–104, 1992.

[64] Luis A. Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, 214(3):547–563, 2000.

[65] Eric A. Carlen, Maria C. Carvalho, Jonathan Le Roux, Michael Loss, and Cédric Villani. Entropy and chaos in the Kac model. *Kinet. Relat. Models*, 3(1):85–122, 2010.

[66] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 2019.

[67] Nicolas Champagnat and Pierre-Emmanuel Jabin. Strong solutions to stochastic differential equations with rough coefficients. *Ann. Probab.*, 46(3):1498–1541, 2018.

[68] Sourav Chatterjee. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields*, 143(1-2):1–40, 2009.

[69] Sourav Chatterjee and Elizabeth Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008.

[70] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011.

[71] Louis HY Chen and Xiao Fang. Multivariate normal approximation by stein's method: The concentration inequality approach. *arXiv preprint arXiv:1111.4073*, 2011.

[72] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

[73] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

[74] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3036–3046. 2018.

[75] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[76] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 32*, pages 2937–2947. 2019.

[77] A. Cianchi, N. Fusco, F. Maggi, and A. Pratelli. On the isoperimetric deficit in Gauss space. *Amer. J. Math.*, 133(1):131–186, 2011.

[78] Maria Colombo and Max Fathi. Bounds on optimal transport maps onto log-concave measures. *arXiv preprint arXiv:1910.09035*, 2019.

[79] Maria Colombo, Alessio Figalli, and Yash Jhaveri. Lipschitz changes of variables between perturbations of log-concave measures. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 17(4):1491–1519, 2017.

[80] D. Cordero-Erausquin and B. Klartag. Moment measures. *J. Funct. Anal.*, 268(12):3834–3866, 2015.

[81] Dario Cordero-Erausquin. Transport inequalities for log-concave measures, quantitative forms, and applications. *Canad. J. Math.*, 69(3):481–501, 2017.

[82] Thomas A Courtade. Bounds on the poincaré constant for convolution measures. *to appear in Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 2018.

[83] Thomas A Courtade. A quantitative entropic clt for radially symmetric random vectors. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1610–1614. IEEE, 2018.

[84] Thomas A. Courtade, Max Fathi, and Ashwin Pananjady. Quantitative stability of the entropy power inequality. *IEEE Trans. Inform. Theory*, 64(8):5691–5703, 2018.

[85] Thomas A. Courtade, Max Fathi, and Ashwin Pananjady. Existence of Stein kernels under a spectral gap, and discrepancy bounds. *Ann. Inst. Henri Poincaré Probab. Stat.*, 55(2):777–790, 2019.

[86] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[87] Gianluca Crippa and Camillo De Lellis. Estimates and regularity results for the DiPerna-Lions flow. *J. Reine Angew. Math.*, 616:15–46, 2008.

[88] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[89] Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. *arXiv preprint arXiv:1911.09873*, 2019.

[90] Amit Daniely. Memorizing gaussians with no over-parameterizaion via gradient decent on neural networks. *arXiv preprint arXiv:2003.12895*, 2020.

[91] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[92] Yohann De Castro, Claire Lacour, and Thanh Mai Pham Ngoc. Adaptive estimation of nonparametric geometric graphs. *arXiv preprint arXiv:1708.02107*, 2017.

[93] Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8581–8593. Curran Associates, Inc., 2018.

[94] R. J. DiPerna and P.-L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Invent. Math.*, 98(3):511–547, 1989.

[95] Simon K. Donaldson. Kähler geometry on toric manifolds, and some other manifolds with large symmetry. In *Handbook of geometric analysis. No. 1*, volume 7 of *Adv. Lect. Math. (ALM)*, pages 29–75. Int. Press, Somerville, MA, 2008.

[96] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019.

[97] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geom. Funct. Anal.*, 23(2):532–569, 2013.

[98] Ronen Eldan. Skorokhod embeddings via stochastic flows on the space of Gaussian measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(3):1259–1280, 2016.

[99] Ronen Eldan. Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.*, 28(6):1548–1596, 2018.

[100] Ronen Eldan and James R. Lee. Regularization under diffusion and anticoncentration of the information content. *Duke Math. J.*, 167(5):969–993, 2018.

[101] Ronen Eldan, Joseph Lehec, and Yair Shenfeld. Stability of the logarithmic Sobolev inequality via the Föllmer process. *arXiv preprint arXiv:1903.04522*, 2019.

[102] Ronen Eldan and Dan Mikulincer. Information and dimensionality of anisotropic random geometric graphs. *arXiv preprint arXiv:1609.02490*, 2016.

[103] Ronen Eldan and Dan Mikulincer. Stability of the Shannon-Stam inequality via the Föllmer process. *Probab. Theory Related Fields*, 177(3-4):891–922, 2020.

[104] Ronen Eldan, Dan Mikulincer, and Hester Pieters. Community detection and percolation of information in a geometric setting. *arXiv preprint arXiv:2006.15574*, 2020.

[105] Ronen Eldan, Dan Mikulincer, and Tselil Schramm. Non-asymptotic approximations of neural networks by gaussian processes. *arXiv preprint arXiv:2102.08668*, 2021.

[106] Ronen Eldan, Dan Mikulincer, and Alex Zhai. The CLT in high dimensions: quantitative bounds via martingale embedding. *Ann. Probab.*, 48(5):2494–2524, 2020.

[107] Philip Ernst and Larry Shepp. On the time for Brownian motion to visit every point on a circle. *J. Statist. Plann. Inference*, 171:130–134, 2016.

[108] Carl-Gustav Esseen. On the Liapounoff limit of error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28A(9):19, 1942.

[109] Shizan Fang, Dejun Luo, and Anton Thalmaier. Stochastic differential equations with coefficients in Sobolev spaces. *J. Funct. Anal.*, 259(5):1129–1168, 2010.

[110] Xiao Fang and Yuta Koike. New error bounds in multivariate normal approximations via exchangeable pairs with applications to wishart matrices and fourth moment theorems. *arXiv preprint arXiv:2004.02101*, 2020.

[111] Xiao Fang, Qi-Man Shao, and Lihu Xu. Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probab. Theory Related Fields*, 174(3-4):945–979, 2019.

[112] Max Fathi. Stein kernels and moment maps. *Ann. Probab.*, 47(4):2172–2185, 2019.

[113] Max Fathi, Emanuel Indrei, and Michel Ledoux. Quantitative logarithmic Sobolev inequalities and stability estimates. *Discrete Contin. Dyn. Syst.*, 36(12):6835–6853, 2016.

[114] Max Fathi and Dan Mikulincer. Stability estimates for invariant measures of diffusion processes, with applications to stability of moment measures and stein kernels. *arXiv preprint arXiv:2010.14178*, 2020.

[115] Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994.

[116] F. Feo, E. Indrei, M. R. Posteraro, and C. Roberto. Some remarks on the stability of the log-Sobolev inequality for the Gaussian measure. *Potential Anal.*, 47(1):37–52, 2017.

[117] Alessio Figalli. Existence and uniqueness of martingale solutions for SDEs with rough or degenerate coefficients. *J. Funct. Anal.*, 254(1):109–153, 2008.

[118] Gerald B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.

[119] H. Föllmer. Time reversal on Wiener space. In *Stochastic processes—mathematics and physics (Bielefeld, 1984)*, volume 1158 of *Lecture Notes in Math.*, pages 119–129. Springer, Berlin, 1986.

[120] Hans Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pages 156–163. Springer, Berlin, 1985.

[121] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. *arXiv preprint arXiv:1709.05510*, 2017.

[122] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2018.

[123] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stephane d'Ascoli, Giulio Biroli, Clement Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal Of Statistical Mechanics-Theory And Experiment*, 2020(ARTICLE):023401, 2020.

[124] Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. *Ann. Fac. Sci. Toulouse Math. (6)*, 26(3):569–601, 2017.

[125] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019.

[126] Surbhi Goel, Sushrut Karmalkar, and Adam Klivans. Time/accuracy tradeoffs for learning a relu with respect to Gussian marginals. In *Advances in Neural Information Processing Systems*, pages 8584–8593, 2019.

[127] Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, 2019.

[128] F. Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2):724–739, 1991.

[129] N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Related Fields*, 16(4):635–736, 2010.

[130] Leonard Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[131] Olle Häggström and Elchanan Mossel. Nearest-neighbor walks with low predictability profile and percolation in $2+\varepsilon$ dimensions. *The Annals of Probability*, 26(3):1212–1231, 1998.

[132] Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019.

[133] Gilles Hargé. A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces. *Probab. Theory Related Fields*, 130(3):415–440, 2004.

[134] Tamir Hazan and Tommi Jaakkola. Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*, 2015.

[135] Einar Hille. Contributions to the theory of Hermitian series II. the representation problem. *Transactions of the American Mathematical Society*, 47(1):80–94, 1940.

[136] Oliver Hinder. Cutting plane methods can be extended into nonconvex optimization. In *Conference On Learning Theory, COLT 2018*, pages 1451–1454, 2018.

[137] Francis Hirsch and Gilles Lacombe. *Elements of functional analysis*, volume 192. Springer Science & Business Media, 2012.

[138] Zhi-yuan Huang and Jia-an Yan. *Introduction to infinite dimensional stochastic analysis*, volume 502 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht; Science Press Beijing, Beijing, chinese edition, 2000.

[139] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[140] Svante Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997.

[141] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2020.

[142] Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2020.

[143] Tiefeng Jiang and Danning Li. Approximation of rectangular beta-Laguerre ensembles and large deviations. *J. Theoret. Probab.*, 28(3):804–847, 2015.

[144] Tiefeng Jiang and Junshan Xie. Limiting behavior of largest entry of random tensor constructed by high-dimensional data. *Journal of Theoretical Probability*, pages 1–21, 2019.

[145] Oliver Johnson and Andrew Barron. Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields*, 129(3):391–409, 2004.

[146] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[147] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, 13(3-4):541–559, 1995.

[148] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019.

[149] Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *57th Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019.

[150] Huynh Khanh and Filippo Santambrogio. q-moment measures and applications: A new approach via optimal transport. *arXiv preprint arXiv:2008.09362*, 2020.

[151] Bo'az Klartag. Logarithmically-concave moment measures I. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 231–260. Springer, Cham, 2014.

[152] Bo'az Klartag. Eldan's stochastic localization and tubular neighborhoods of complex-analytic sets. *J. Geom. Anal.*, 28(3):2008–2027, 2018.

[153] A. V. Kolesnikov. Global Hölder estimates for optimal transportation. *Mat. Zametki*, 88(5):708–728, 2010.

[154] Alexander V. Kolesnikov. Hessian metrics, $CD(K, N)$-spaces, and optimal transportation of log-concave measures. *Discrete Contin. Dyn. Syst.*, 34(4):1511–1532, 2014.

[155] Alexander V. Kolesnikov and Egor D. Kosov. Moment measures and stability for Gaussian inequalities. *Theory Stoch. Process.*, 22(2):47–61, 2017.

[156] Ioannis Kontoyiannis and Mokshay Madiman. Sumset and inverse sumset inequalities for differential entropy and mutual information. *IEEE Transactions on Information Theory*, 60(8):4503–4514, 2014.

[157] Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration of random graphs and application to community detection. *Proc. Int. Cong. of Math.*, 3:2913 – 2928, 2018.

[158] C. Le Bris and P.-L. Lions. Existence and uniqueness of solutions to Fokker-Planck type equations with irregular coefficients. *Comm. Partial Differential Equations*, 33(7-9):1272–1317, 2008.

[159] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[160] Michel Ledoux. Spectral gap, logarithmic sobolev constant, and geometric bounds. *Surveys in differential geometry*, 9(1):219–240, 2004.

[161] Michel Ledoux, Ivan Nourdin, and Giovanni Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geom. Funct. Anal.*, 25(1):256–306, 2015.

[162] Michel Ledoux, Ivan Nourdin, and Giovanni Peccati. A Stein deficit for the logarithmic Sobolev inequality. *Sci. China Math.*, 60(7):1163–1180, 2017.

[163] Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007. IEEE, 2017.

[164] Eveline Legendre. Toric Kähler-Einstein metrics and convex compact polytopes. *J. Geom. Anal.*, 26(1):399–427, 2016.

[165] Joseph Lehec. Representation formula for the entropy and functional inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(3):885–899, 2013.

[166] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[167] Huaiqian Li and Dejun Luo. Quantitative stability estimates for Fokker-Planck equations. *J. Math. Pures Appl. (9)*, 122:125–163, 2019.

[168] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[169] Fon Che Liu. A Luzin type property of Sobolev functions. *Indiana Univ. Math. J.*, 26(4):645–651, 1977.

[170] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms*, 30(3):307–358, 2007.

[171] A. Lytova. Central limit theorem for linear eigenvalue statistics for a tensor product version of sample covariance matrices. *J. Theoret. Probab.*, 31(2):1024–1057, 2018.

[172] Mokshay Madiman and Ioannis Kontoyiannis. Entropy bounds on abelian groups and the ruzsa divergence. *IEEE Transactions on Information Theory*, 64(1):77–92, 2018.

[173] Arnaud Marsiglietti and Victoria Kostina. A lower bound on the differential entropy of log-concave random vectors with applications. *Entropy*, 20(3):Paper No. 185, 24, 2018.

[174] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[175] S Mei, A Montanari, and PM Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33):E7665–E7671, 2018.

[176] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

[177] Max Mether. The history of the central limit theorem. *Sovelletun Matematiikan Erikoistyt, vol. 2, no. 1, pp. 8*, 2013.

[178] Guillaume Mijoule, Gesine Reinert, and Yvik Swan. Stein operators, kernels and discrepancies for multivariate continuous distributions. *arXiv preprint arXiv:1806.03478*, 2018.

[179] Dan Mikulincer. Stability of talagrand's gaussian transport-entropy inequality via the f\" ollmer process. *arXiv preprint arXiv:1906.05904*, 2019.

[180] Dan Mikulincer. A CLT in Stein's distance for generalized Wishart matrices and higher-order tensors. *International Mathematics Research Notices*, 01 2021. rnaa336.

[181] Emanuel Milman. On the role of convexity in isoperimetry, spectral gap and concentration. *Invent. Math.*, 177(1):1–43, 2009.

[182] Pierre Monmarché. Generalized Γ calculus and application to interacting particles on a graph. *Potential Anal.*, 50(3):439–466, 2019.

[183] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

[184] Elchanan Mossel. *Survey: information flow on trees*. 2004.

[185] Elchanan Mossel and Joe Neeman. Robust dimension free isoperimetry in Gaussian space. *Ann. Probab.*, 43(3):971–991, 2015.

[186] Elchanan Mossel and Yuval Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 08 2003.

[187] Elchanan Mossel, Sébastien Roch, and Allan Sly. Robust estimation of latent tree graphical models: inferring hidden states with inexact parameters. *IEEE Trans. Inform. Theory*, 59(7):4357–4373, 2013.

[188] S. V. Nagaev. An estimate of the remainder term in the multidimensional central limit theorem. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)*, pages 419–438. Lecture Notes in Math., Vol. 550, 1976.

[189] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer, 1996.

[190] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[191] Arkadi Nemirovski. On parallel complexity of nonsmooth convex optimization. *Journal of Complexity*, 10(4):451 – 463, 1994.

[192] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

[193] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

[194] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2012. From Stein's method to universality.

[195] Ivan Nourdin, Giovanni Peccati, and Anthony Réveillac. Multivariate normal approximation using Stein's method and Malliavin calculus. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(1):45–58, 2010.

[196] Ivan Nourdin, Giovanni Peccati, and Yvik Swan. Entropy and the fourth moment phenomenon. *J. Funct. Anal.*, 266(5):3170–3207, 2014.

[197] Ivan Nourdin and Guangqu Zheng. Asymptotic behavior of large gaussian correlated wishart matrices. *arXiv preprint arXiv:1804.06220*, 2018.

[198] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2018.

[199] Bernt Ø ksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[200] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

[201] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *CoRR*, abs/1902.04674, 2019.

[202] Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. Effect of activation functions on the training of overparametrized neural nets. In *International Conference on Learning Representations*, 2019.

[203] G. Paouris. Concentration of mass on convex bodies. *Geom. Funct. Anal.*, 16(5):1021–1049, 2006.

[204] Grigoris Paouris. Small ball probability estimates for log-concave measures. *Trans. Amer. Math. Soc.*, 364(1):287–308, 2012.

[205] Grigoris Paouris and Petros Valettas. A Gaussian small deviation inequality for convex functions. *Ann. Probab.*, 46(3):1441–1454, 2018.

[206] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012.

[207] Miklós Z. Rácz and Jacob Richey. A smooth transition from Wishart to GOE. *J. Theoret. Probab.*, 32(2):898–906, 2019.

[208] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[209] Emmanuel Rio. Upper bounds for minimal distances in the central limit theorem. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(3):802–817, 2009.

[210] Emmanuel Rio. Asymptotic constants for minimal distance in the central limit theorem. *Electron. Commun. Probab.*, 16:96–103, 2011.

[211] Nathan Ross. Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293, 2011.

[212] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[213] Filippo Santambrogio. Dealing with moment measures via entropy and optimal transport. *J. Funct. Anal.*, 271(2):418–436, 2016.

[214] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2667–2690. PMLR, 2019.

[215] Christian Seis. A quantitative theory for the continuity equation. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 34(7):1837–1850, 2017.

[216] Christian Seis. Optimal stability estimates for continuity equations. *Proc. Roy. Soc. Edinburgh Sect. A*, 148(6):1279–1296, 2018.

[217] V. V. Senatov. Some uniform estimates of the convergence rate in the multidimensional central limit theorem. *Teor. Veroyatnost. i Primenen.*, 25(4):757–770, 1980.

[218] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[219] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[220] Xin Shi, Robert Qiu, Xing He, Lei Chu, Zenan Ling, and Haosen Yang. Anomaly detection and location in distribution network: A data-driven approach. *arXiv preprint arXiv:1801.01669*, 2018.

[221] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[222] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *CoRR*, abs/1906.03593, 2019.

[223] Aart J Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.

[224] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, 1972.

[225] Charles Stein. *Approximate computation of expectations*, volume 7 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[226] E. M. Stein and J.-O. Strömberg. Behavior of maximal functions in $\mathbf{R}^n$ for large n. *Ark. Mat.*, 21(2):259–269, 1983.

[227] Xiaoming Sun and Andrew Chi-Chih Yao. On the quantum query complexity of local search in two and three dimensions. *Algorithmica*, 55(3):576–600, 2009.

[228] D. L. Sussman, M. Tang, and C. E. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57, Jan 2014.

[229] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, 6(3):587–600, 1996.

[230] Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

[231] Giuseppe Toscani. A strengthened entropy power inequality for log-concave densities. *IEEE Transactions on Information Theory*, 61(12):6550–6559, 2015.

[232] Dario Trevisan. Well-posedness of multidimensional diffusion processes with weakly differentiable coefficients. *Electron. J. Probab.*, 21:Paper No. 22, 41, 2016.

[233] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[234] Belinda Tzen and Maxim Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled Mckean-Vlasov dynamics. *arXiv preprint arXiv:2002.01987*, 2020.

[235] Ernesto Araya Valdivia. Relative concentration bounds for the kernel matrix spectrum. *arXiv preprint arXiv:1812.02108*, 2018.

[236] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 685–694. ACM, New York, 2011.

[237] Stephen A. Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.

[238] Mark Veraar. The stochastic fubini theorem revisited. *Stochastics An International Journal of Probability and Stochastic Processes*, 84(4):543–551, 2012.

[239] Roman Vershynin. Concentration inequalities for random tensors. *arXiv preprint arXiv:1905.00802*, 2019.

[240] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[241] Max-K. von Renesse and Karl-Theodor Sturm. Transport inequalities, gradient estimates, entropy, and Ricci curvature. *Comm. Pure Appl. Math.*, 58(7):923–940, 2005.

[242] Xu-Jia Wang and Xiaohua Zhu. Kähler-Ricci solitons on toric manifolds with positive first Chern class. *Adv. Math.*, 188(1):87–103, 2004.

[243] Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pages 295–301, 1997.

[244] Longjie Xie and Xicheng Zhang. Sobolev differentiable flows of SDEs with local Sobolev and super-linear growth coefficients. *Ann. Probab.*, 44(6):3661–3687, 2016.

[245] Sho Yaida. Non-gaussian processes and neural networks at finite widths. *arXiv preprint arXiv:1910.00019*, 2019.

[246] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[247] Greg Yang. Tensor programs I: Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. *arXiv preprint arXiv:1910.12478*, 2019.

[248] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

[249] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014.

[250] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems*, pages 15532–15543, 2019.

[251] Alex Zhai. A high-dimensional CLT in $\mathcal{W}_2$ distance with near optimal convergence rate. *Probab. Theory Related Fields*, 170(3-4):821–845, 2018.

[252] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.

[253] Shengyu Zhang. New upper and lower bounds for randomized and quantum local search. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 634–643. ACM, New York, 2006.

[254] William P. Ziemer. *Weakly differentiable functions*, volume 120 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1989. Sobolev spaces and functions of bounded variation.