Analysis And Stochastics

Pan Jing Bin

Supervisor: Assistant Professor Subhroshekhar Ghosh

Examiner: Associate Professor Sun Rongfeng

Department of Mathematics

National University of Singapore

Honours Year Project for Semester 2, AY2022/2023

Summary

The written report is a summary of the material that I have read while working as a research assistant under Assistant Professor Subhroshekhar Ghosh. The main focus of the report is the seminal paper [3], which introduced the notion of algebraically structured models. Chapter 1, 2, 3 and 4 consists of a detailed exposition of the mathematical tools and proof techniques that were developed in [3]. The intention of studying each proof in painstaking detail is so as to apply these very same techniques to obtain minimax rates of estimation in the setting of sparse signals, as a continuation of the work in [14]. Some of the other supplementary materials which I have read are briefly mentioned in Chapter 5.

My contributions consist mainly of filling in the details in the proofs given in [3].

In Theorem 2.3.1, I have modified the original proof in [3, Theorem 9, page 22] as there was a minor error. The formula for the χ^2-divergence should be $\int \frac{(f_\theta(y)-f_\phi(y))^2}{f_\phi(y)}\,dy$ instead of $\int \frac{(f_\theta(y)-f_\phi(y))^2}{f_\theta(y)}\,dy$.

Contents

1 Algebraically Structured Models

1.1 Introduction

A fundamental problem arising in many scientific fields is the recovery of data that is corrupted not only by noise, but also by latent transformations. In many cases, these latent transformations can be modelled by the action of an unknown element in a known group. Until recently, these problems have mainly been tackled from computational perspectives and their theoretical properties remain largely unexplored. However, with recent improvements in technology, understanding the statistical properties of such problems have become a much more pressing concern.

One such problem which have received widespread attention in recent years is Cryogenic electron microscopy (cryo-EM). In the simplest sense, the cryo-EM problem involves reconstructing a three-dimensional molecule from noisy two dimensional tomographic projections, which are taken after an unknown three-dimensional rotation has been applied to the molecule.

With the goal of establishing a theoretical framework for the systematic study of such problems, the notion of algebraically structured models was introduced in [3]. The paper have since generated a flurry of new research and led to a much better understanding of the problem. In this report, we will give a detailed exposition on the tools and techniques that were developed in the paper to derive the key results.

1.2 Algebraically Structured Models

In this section, we formally introduce the algebraically structured model. Let d be a positive integer and let $\mathcal{G}$ denote a known compact subgroup of the orthogonal group

$$O(d) := \left\{ \boldsymbol{A} \in \mathrm{M}_d(\mathbb{R}) \ : \ \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_d \right\}.$$

The group $\mathcal{G}$ naturally acts on the vector space $\mathbb{R}^d$ by left-multiplication. The objective is to recover an unknown parameter $\theta \in \mathbb{R}^d$, known in the literature as a ***signal***, from independent noisy observations $X_1, X_2, \cdots, X_n$ given by

$$X_i = G_i\theta + \sigma\xi_i. \tag{1}$$

Here, the G_i's are drawn from $\mathcal{G}$ independently and uniformly at random via the Haar measure and each $\xi_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ is i.i.d standard Gaussian noise that is independent of G_i.

Both the dimension d and the variance σ^2 are fixed and assumed to be known throughout the report. Motivated by practical applications (including cryo-EM) in which the noise level is high, we focus on the case in which the quantity $\|\theta\|/\sigma$, known as the ***signal-to-noise ratio (SNR)***, is small. In many such applications, technological improvements can directly decrease the value of σ. As such, understanding the relationship between the difficulty of the estimation problem and the signal-to-noise ratio is of fundamental importance and is the main focus of this report. In particular, we will focus on the ***sampling complexity*** of the problem, which is the number of samples needed to estimate the true signal θ at a prescribed accuracy.

We remark that another problem that is outside the scope of this report is the dependence of the difficulty of the estimation problem on the dimension d. The quantity d corresponds to the level of discretization of the object of interest and can be quite large ($\gtrsim 10^6$) in practice [5]. As such, the problem is also of significant interest and has been investigated in work in [20].

For a signal θ, let P_θ denote the distribution of a random variable X satisfying

$$X = G\theta + \sigma\xi,$$

where G is drawn from $\mathcal{G}$ uniformly and $\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ is independent of G. Also let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ denote the density function of P_θ. Explicitly,

$$\begin{aligned} f_\theta(x) &= \frac{1}{\sigma^d(2\pi)^{d/2}}\mathbb{E}\Big[\exp\Big(-\frac{1}{2\sigma^2}\|x - G\theta\|^2\Big)\Big] && (2) \\ &= \frac{1}{\sigma^d(2\pi)^{d/2}}\exp\left(-\frac{1}{2\sigma^2}\big(\|x\|^2 + \|\theta\|^2\big)\right)\mathbb{E}\left[\exp\left(\frac{1}{\sigma^2}x^T G\theta\right)\right]. && (3) \end{aligned}$$

If ϕ is another signal that lies in the same $\mathcal{G}$-orbit as θ, then the translational invariance of the Haar measure implies that the two distributions P_θ and P_ϕ are indistinguishable. Consequently, the parameter θ is only identifiable up to the action of $\mathcal{G}$ and so we seek to obtain an estimator $\tilde{\theta}$ whose distance to the orbit of θ defined by

$$\rho(\tilde{\theta}, \theta) := \min_{G \in \mathcal{G}} \|\tilde{\theta} - G\theta\|$$

is small in expectation.

Throughout this report, we impose the following assumptions on the parameter θ.

(i) There exists an absolute constant K such that $K^{-1} \le \|\theta\| \le K$;

(ii) $\|\theta\| \le \sigma$.

Let $\mathcal{S} \subseteq \mathbb{R}^d$ denote the set of vectors satisfying the above two assumptions. For the key results in the paper, the true signal θ will always be assumed to lie in $\mathcal{S}$.

The first assumption is imposed so that we can normalise $\|\theta\| = 1$. This allows σ to capture entirely the signal-to-noise ratio of the problem. On the other hand, the second assumption is adopted as we will only focus on the high noise regime. For the low noise regime, the current leading approach is the synchronization approach [2].

1.3 Notation

We use log to denote the natural logarithm and $\mathbb{N}_0$ to denote the set of nonnegative integers. For a positive integer m, let $\mathbb{N}_0^m$ denote a m-tuple of nonnegative integers. Given $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m) \in \mathbb{N}_0^m$, define $\alpha! = \alpha_1!\alpha_2! \cdots \alpha_m!$ and $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_m$. Let $[m]$ denote the set $\{1, 2, \cdots, m\}$. If k is another positive integer, then $[m]^k$ denotes the Cartesian product $\{1, 2, \cdots, m\}^k$.

For two probability measures P and Q on a measure space $(\Omega, \mathcal{F})$, we use $P \ll Q$ to denote that P is absolutely continuous with respect to Q. If that is the case, we also let $\frac{dP}{dQ}$ denote the corresponding Radon-Nikodym derivative.

Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the usual ℓ_2 inner product and ℓ_2 norm on $\mathbb{R}^d$ respectively. For a matrix $\boldsymbol{A} \in \mathrm{M}_d(\mathbb{R})$, we let $\|\boldsymbol{A}\|_{\mathrm{op}}$ denote its operator norm.

For each positive integer n, let $P_\theta^{\otimes n}$ denote the n-fold product measure of P_θ, usually used to denote the joint distribution of n i.i.d samples $X_1, \cdots, X_n$ drawn according to P_θ. For any estimator $\tilde{\theta}_n$ based on the n samples $X_1, \cdots, X_n$ and any measurable function $g : \mathbb{R}^d \to \mathbb{R}^k$, we also use $\mathbb{E}_\theta[g(\tilde{\theta}_n)]$ to denote the corresponding expectation.

In informal discussions, we will sometimes use $A \lesssim B$ and $A = O(B)$ to denote $A \leq CB$ for some constant C. We use $A \approx B$ to denote both $A \lesssim B$ and $B \lesssim A$. If this constant depend on parameters (say p and q), we will use $A \lesssim_{p,q} B$ and $A = O_{p,q}(B)$ to denote $A \leq C_{p,q}B$ for a constant $C_{p,q}$ depending only on p and q.

We adopt the convention that $0 \log 0 = 0$, which is commonly used in information theory [11].

2 Kullback-Leibler Divergence and Moment Tensors

In this chapter, we will begin the journey of developing the general theory of algebraically structured models. In the process, we will also derive some key properties about the model that will be used throughout the report.

2.1 Kullback-Leibler Divergence

The first piece of machinery that we need to acquire is a way to mathematically quantify the concept of "distance" between different probability distributions.

Definition 2.1.1. Let P and Q be two probability measures on a measurable space $(\Omega, \mathcal{F})$. Define the ***Kullback-Leibler(KL) divergence of P from Q*** by

$$D_{\mathrm{KL}}(P \parallel Q) := \begin{cases} \displaystyle\int_\Omega \log\left(\frac{dP}{dQ}\right)\, dP & \text{if } P \ll Q \\ +\infty & \text{otherwise.} \end{cases}$$

The KL divergence will serve as our main tool for gauging the difficulty of the estimation problem. Roughly speaking, the smaller the KL divergence between two probability distributions, the more difficult it is to distinguish between them. To make this notion precise, we adopt Le Cam's method of two hypotheses [17]. The outline is given below.

One of the main objectives of this report is to obtain lower bounds on the quantity

$$\inf_{\tilde{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[\rho(\tilde{\theta}_n, \theta)\right]$$

where the infimum is taken over all estimators $\tilde{\theta}_n$ on n samples $X_1, \cdots, X_n$. Here, Θ denotes the parameter space (taken to be a subspace of $\mathcal{S}$ in this report). By Markov's inequality, we have

$$\inf_{\tilde{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[\rho(\tilde{\theta}_n, \theta)\right] \geq C \cdot \inf_{\tilde{\theta}_n} \sup_{\theta\in\Theta} P_\theta^{\otimes n}\left(\rho(\tilde{\theta}_n, \theta) \geq C\right)$$

where C is a positive constant whose exact value will be determined later.

Here, we slightly abuse notation and also use $P_\theta^{\otimes n}$ to denote the probability of an event when the samples $X_1, \cdots, X_n$ are drawn according to the distribution P_θ.

For any pair of hypotheses $\phi, \psi \in \Theta$, clearly

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} P_\theta^{\otimes n}\big(\rho(\tilde{\theta}_n, \theta) \geq C\big) \geq \inf_{\tilde{\theta}_n} \max_{\theta \in \{\phi, \psi\}} P_\theta^{\otimes n}\big(\rho(\tilde{\theta}_n, \theta) \geq C\big).$$

Now suppose that ϕ and ψ are well-separated in the sense that $\rho(\phi, \psi) \geq 2C$. Then for any estimator $\tilde{\theta}_n$, we have

$$P_\phi^{\otimes n}\big(\rho(\tilde{\theta}_n, \phi) \geq C\big) \geq P_\phi^{\otimes n}\big(\Psi_n \neq \phi\big) \quad \text{and} \quad P_\psi^{\otimes n}\big(\rho(\tilde{\theta}_n, \psi) \geq C\big) \geq P_\psi^{\otimes n}\big(\Psi_n \neq \psi\big),$$

where $\Psi_n : \Omega \to \{\phi, \psi\}$ is the minimum-distance estimator

$$\Psi_n := \arg\min_{\theta \in \{\phi, \psi\}} \rho(\tilde{\theta}_n, \theta).$$

In summary, if we are able to find two signals $\phi, \psi \in \Theta$ satisfying $\rho(\phi, \psi) \geq 2C$, then

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\big] \geq C \cdot \inf_{\Phi_n} \max_{\theta \in \{\phi, \psi\}} P_\theta^{\otimes n}\big(\Phi_n \neq \theta\big)$$

where the infimum on the right-hand is taken over all binary classifiers $\Phi_n : \Omega \to \{\phi, \psi\}$.

The following lemma establishes a uniform lower bound for the error probability of binary classifiers under the new simplified setting of two hypotheses.

Lemma 2.1.2. Let P_1 and P_2 be two probability measures on a measure space $(\Omega, \mathcal{F})$. Then

$$\inf_{\Phi} \max_{j \in \{1,2\}} P_j(\Phi \neq j) \geq \frac{2 - \sqrt{2D_{\mathrm{KL}}(P_1 \parallel P_2)}}{4}$$

where the infimum is taken over all measurable functions $\Phi : \Omega \to \{1, 2\}$.

Proof. Let μ be any finite measure on $(\Omega, \mathcal{F})$ satisfying $P_1 \ll \mu$ and $P_2 \ll \mu$ (such a measure always exist, by considering $\mu = P_1 + P_2$) and let f_1 and f_2 denote the Radon-Nikodym derivatives of P_1 and P_2 with respect to μ. Then

$$\begin{aligned}
\inf_{\Phi} \max_{j \in \{1,2\}} P_j(\Phi \neq j) &\geq \frac{1}{2} \inf_{\Phi} \Big\{ P_1(\Phi \neq 1) + P_2(\Phi \neq 2) \Big\} \\
&= \frac{1}{2} \inf_{\Phi} \left\{ \int_{\{\Phi \neq 1\}} f_1(\omega)\ d\mu(\omega) + \int_{\{\Phi \neq 2\}} f_2(\omega)\ d\mu(\omega) \right\} \\
&\geq \frac{1}{2} \int_{\Omega} \min\big\{f_1(\omega), f_2(\omega)\big\}\ d\mu(\omega) \\
&= \frac{1 - V(P_1, P_2)}{2}
\end{aligned}$$

where the last line follows from Scheffé's lemma [24, Lemma 2.1] applied to the total variation distance $V(P_1, P_2)$ defined by

$$V(P_1, P_2) := \sup_{A \in \mathcal{F}} \big|P_1(A) - P_2(A)\big|.$$

The conclusion follows from the First Pinsker's inequality [24, Lemma 2.5]. □

With $C = \rho(\phi, \psi)/2$, we formalise the above discussion into the following proposition.

Proposition 2.1.3. For any $\phi, \psi \in \Theta$, we have

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \big[\rho(\tilde{\theta}_n, \theta)\big] \geq \rho(\phi, \psi) \cdot \left(\frac{2 - \sqrt{2nD_{\mathrm{KL}}(P_\phi \parallel P_\psi)}}{8} \right).$$

Informally speaking, the above result suggests that if $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \lesssim \sigma^{-m}$, then $n \gtrsim \sigma^m$ is required to keep the error probability and the expected loss low. This allows us to reduce the problem of obtaining bounds on the performance of estimators to obtaining bounds on the quantity $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ instead. We will henceforth focus our attention on understanding the behaviour of the KL divergence in the algebraically structured model. Our first order of business is to derive an explicit formula for the KL divergence. Recall that

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)} \; dx = \mathbb{E}\left[\log \frac{f_\theta(X)}{f_\phi(X)} \right]$$

where $X \sim P_\theta$.

Lemma 2.1.4. Let $\theta, \phi \in \mathbb{R}^d$. Then

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = \frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) + \mathbb{E}_\xi \left[\log \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\phi\big)\big]} \right].$$

Proof. Directly from (3), we have

$$\frac{f_\theta(x)}{f_\phi(x)} = \exp\left(\frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) \right) \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2} x^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2} x^T G\phi\big)\big]}.$$

Taking expectation with respect to X then gives

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = \frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) + \mathbb{E}_X \left[\log \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2} X^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2} X^T G\phi\big)\big]} \right].$$

Write $X = H\theta + \sigma\xi$ where $H \in \mathcal{G}$ denotes an independent and identically distributed copy of G. As X and $H(\theta + \sigma\xi)$ have identical distributions, we obtain

$$\begin{aligned} D_{\mathrm{KL}}(P_\theta \parallel P_\phi) &= \frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) + \mathbb{E}_\xi \left[\mathbb{E}_H \left[\log \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T H^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T H^T G\phi\big)\big]} \right] \right] \\ &= \frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) + \mathbb{E}_\xi \left[\mathbb{E}_H \left[\log \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\phi\big)\big]} \right] \right] \\ &= \frac{1}{2\sigma^2}\big(\|\phi\|^2 - \|\theta\|^2 \big) + \mathbb{E}_\xi \left[\log \frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta + \sigma\xi)^T G\phi\big)\big]} \right] \end{aligned}$$

as desired. □

Remark 2.1.5. From the explicit formula in Lemma 2.1.4, it is easy to see that replacing the quantities $\theta,\ \phi$ and σ by $\theta/\left\|\theta\right\|,\ \phi/\left\|\theta\right\|$ and $\sigma/\left\|\theta\right\|$ respectively does not change $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$. This allows us to normalise $\|\theta\| = 1$ and $\sigma \geq 1$ when studying properties associated to the KL divergence. This technique will be employed repeatedly throughout the report.

In general, the formula in Lemma 2.1.4 is too complicated to be of much use in explicit computations. As such, we will introduce another tool in the next section to help us to control the KL divergence.

2.2 Moment Tensors

Inspired by the classical method of moments in statistical estimation, we compare the moments between different probability distributions as a way of quantifying their relative similarity.

Definition 2.2.1. Let m be any positive integer. The ***mth moment tensor*** of a vector $\theta \in \mathbb{R}^d$ is defined to be the quantity $\mathbb{E}[(G\theta)^{\otimes m}] \in (\mathbb{R}^d)^{\otimes m}$. This is a order-$m$ symmetric tensor whose $(i_1, i_2, \cdots, i_m)$-entry is given by

$$\mathbb{E}\big[(G\theta)_{i_1}(G\theta)_{i_2}\cdots(G\theta)_{i_m}\big].$$

Suppose that $\phi \in \mathbb{R}^d$ is another vector. Define the ***mth moment difference tensor*** between θ and ϕ to be

$$\Delta_m(\theta,\phi) := \mathbb{E}\big[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\big] \in (\mathbb{R}^d)^{\otimes m}.$$

Remark 2.2.2. Both the moment tensors $\mathbb{E}[(G\theta)^{\otimes m}]$ and the KL divergence $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ are invariant under the action of $\mathcal{G}$. As we will soon see, this frequently gives us the freedom to choose the orbit representatives of θ and ϕ. We will leverage on this by choosing the representatives θ and ϕ such that $\|\theta - \phi\|$ is minimised (i.e $\rho(\theta,\phi) = \|\theta - \phi\|$).

While seemingly unrelated to the KL divergence at first glance, it turns out that the two quantities are closely related in rather intricate ways. As an first example, the following proposition allows us to reduce the task of bounding $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ to the case in which both θ and ϕ are mean zero signals.

Proposition 2.2.3. Let $\theta, \phi \in \mathbb{R}^d$. If $\vartheta = \theta - \mathbb{E}[G\theta]$ and let $\varphi = \phi - \mathbb{E}[G\phi]$, then

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = D_{\mathrm{KL}}(P_\vartheta \parallel P_\varphi) + \frac{1}{2\sigma^2}\left\|\Delta_1(\theta,\phi)\right\|^2.$$

Proof. The proof, while almost entirely computational in nature, is nevertheless a good illustration of how certain standard tricks can be used to simplify expressions that involve an expectation. By viewing G' as an independent and identically distributed copy of $\mathcal{G}$, the invariance of the Haar measure gives

$$\mathbb{E}\big[G\theta\big]^T\mathbb{E}\big[G\theta\big] = \mathbb{E}\big[\theta^T(G')^TG\theta\big] = \mathbb{E}_{G'}\Big[\mathbb{E}_G\big[\theta^T(G')^TG\theta\big]\Big] = \mathbb{E}_G\big[\theta^TG\theta\big]. \tag{4}$$

Analogues of (4) can be obtained by replacing one or both copies of θ on the left-hand side with ϕ. With this observation, we obtain

$$\begin{aligned}\frac{1}{2\sigma^2}\big(\,\|\varphi\|^2 - \|\vartheta\|^2\,\big) &= \frac{1}{2\sigma^2}\big(\,\|\phi - \mathbb{E}[G\phi]\|^2 - \|\theta - \mathbb{E}[G\theta]\|^2\,\big)\\ &= \frac{1}{2\sigma^2}\big(\,\|\phi\|^2 - \|\theta\|^2\,\big) + \frac{1}{2\sigma^2}\big(\mathbb{E}[\theta^TG\theta] - \mathbb{E}[\phi^TG\phi]\big). \end{aligned}\tag{5}$$

On the other hand, for any fixed $G_0 \in \mathcal{G}$,

$$\begin{aligned}(\vartheta + \sigma\xi)^TG_0\vartheta &= (\theta - \mathbb{E}[G\theta] + \sigma\xi)^TG_0(\theta - \mathbb{E}[G\theta])\\ &= (\theta + \sigma\xi)^TG_0\theta - \mathbb{E}[\theta^TG\theta] - \sigma\xi^T\mathbb{E}[G\theta]\end{aligned}\tag{6}$$

where we have used the fact that $G_0\mathbb{E}[G\theta] = \mathbb{E}[G\theta]$. Similarly,

$$(\vartheta + \sigma\xi)^TG_0\varphi = (\theta + \sigma\xi)^TG_0\phi - \mathbb{E}[\theta^TG\phi] - \sigma\xi^T\mathbb{E}[G\phi]. \tag{7}$$

Substituting (5), (6) and (7) into Lemma 2.1.4, we have

$$\begin{aligned}&D_{\mathrm{KL}}(P_\vartheta \parallel P_\varphi)\\ &= \frac{1}{2\sigma^2}\big(\,\|\varphi\|^2 - \|\vartheta\|^2\,\big) + \mathbb{E}_\xi\left[\log\frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^TG\vartheta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^TG\varphi\big)\big]}\right]\\ &= \frac{1}{2\sigma^2}\big(\,\|\phi\|^2 - \|\theta\|^2\,\big) + \mathbb{E}_\xi\left[\log\frac{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta+\sigma\xi)^TG\theta\big)\big]}{\mathbb{E}_G\big[\exp\big(\frac{1}{\sigma^2}(\theta+\sigma\xi)^TG\phi\big)\big]}\right]\\ &- \mathbb{E}_\xi\left[\frac{\sigma\xi^T(\mathbb{E}_G[G\theta] - \mathbb{E}_G[G\phi])}{\sigma^2}\right] - \frac{\mathbb{E}[\theta^TG\theta] - \mathbb{E}[\theta^TG\phi]}{\sigma^2} + \frac{\mathbb{E}[\theta^TG\theta] - \mathbb{E}[\phi^TG\phi]}{2\sigma^2}\\ &= D_{\mathrm{KL}}(P_\theta \parallel P_\phi) - \frac{1}{2\sigma^2}\,\|\Delta_1(\theta,\phi)\|^2\,.\end{aligned}$$

□

Next, we introduce a way to measure the size of the moment difference tensors.

Definition 2.2.4. Let m be any positive integer. The Euclidean inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ can be canonically extended to an inner product $\langle\cdot,\cdot\rangle_{(\mathbb{R}^d)^{\otimes m}} : (\mathbb{R}^d)^{\otimes m} \times (\mathbb{R}^d)^{\otimes m} \to \mathbb{R}$ by setting

$$\big\langle u_1 \otimes \cdots \otimes u_m,\ v_1 \otimes \cdots \otimes v_m\big\rangle_{(\mathbb{R}^d)^{\otimes m}} := \prod_{i=1}^m \langle u_i, v_i\rangle_{\mathbb{R}^d}$$

for all $u_1, \cdots, u_m, v_1, \cdots, v_m \in \mathbb{R}^d$ and extending by linearity.

When the context is clear, we will use $\langle \cdot, \cdot \rangle$ as shorthand for $\langle \cdot, \cdot \rangle_{(\mathbb{R}^d)^{\otimes m}}$.

Another routine application of the technique in (4) can be used to derive the following formula.

Lemma 2.2.5. For any positive integer m and vectors $\theta, \phi \in \mathbb{R}^d$,

$$\|\Delta_m(\theta, \phi)\|^2 = \mathbb{E}\big[(\theta^T G\theta)^m - 2(\theta^T G\phi)^m + (\phi^T G\phi)^m\big].$$

Proof. Again by letting G' denote an independent and identically distributed copy of G, we have

$$\begin{aligned}\Big\langle \mathbb{E}\big[(G\theta)^{\otimes m}\big], \mathbb{E}\big[(G\phi)^{\otimes m}\big] \Big\rangle &= \mathbb{E}\Big[\big\langle (G\theta)^{\otimes m}, (G'\phi)^{\otimes m} \big\rangle\Big] \\ &= \mathbb{E}\big[(\theta^T G^T G'\phi)^m\big] \\ &= \mathbb{E}\big[(\theta^T G\phi)^m\big].\end{aligned}$$

The conclusion then follows by applying the above equality to

$$\|\Delta_m(\theta, \phi)\|^2 = \Big\langle \mathbb{E}\big[(G\theta)^{\otimes m}\big] - \mathbb{E}\big[(G\phi)^{\otimes m}\big], \mathbb{E}\big[(G\theta)^{\otimes m}\big] - \mathbb{E}\big[(G\phi)^{\otimes m}\big] \Big\rangle.$$

□

2.3 Upper Bound for the Kullback-Leibler Divergence

The remainder of this chapter will be dedicated to proving the first major result in the theory of algebraically structured models, which provides a connection between the KL divergence and the family of moment different tensors via matching lower and upper bounds. The theorem is stated as follows.

Theorem 2.3.1. Let $\theta, \phi \in \mathbb{R}^d$ be two vectors satisfying $3\rho(\theta, \phi) \leq \|\theta\| \leq \sigma$ and $\mathbb{E}[G\theta] = \mathbb{E}[G\phi] = 0$. There exist universal constants $\underline{C}$ and $\overline{C}$ such that for any positive integer k,

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m(\theta, \phi)\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \leq 3 \sum_{m=1}^{k-1} \frac{\|\Delta_m(\theta, \phi)\|^2}{\sigma^{2m} m!} + \overline{C} \frac{\|\theta\|^{2k-2} \rho(\theta, \phi)^2}{\sigma^{2k}}. \tag{8}$$

In particular, the above theorem implies that if the first $k - 1$ moment difference tensors vanish, then $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ is of order $\sigma^{-2k}\rho(\theta, \phi)^2$. We will explore this notion in greater depth in Chapter 4.

In this section, we will first tackle the easier direction, which is the upper bound.

Proof. (of the upper bound in Theorem 2.3.1) If $\theta = 0$, then since $3\rho(\theta, \phi) \leq \|\theta\|$, we must have $\phi = 0$ as well so the statement trivially holds. Otherwise, first note that each term in (8) remains unchanged if the quantities θ, ϕ and σ are replaced by $\theta / \|\theta\|$, $\phi / \|\theta\|$ and $\sigma / \|\theta\|$ respectively (Remark 2.1.5). The same is true when ϕ is replaced by another vector $G_0\phi$ in the same $\mathcal{G}$-orbit (Remark 2.2.2). As such, we henceforth assume $\|\theta\| = 1$, $\sigma \geq 1$ and $\rho(\theta, \phi) = \|\theta - \phi\|$.

Instead of establishing an upper bound on the KL divergence $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ directly, we instead work with the χ^2-divergence

$$\chi^2(P_\theta, P_\phi) := \int_{\mathbb{R}^d} \frac{(f_\theta(x) - f_\phi(x))^2}{f_\phi(x)}\, dx.$$

and then pass to the KL divergence via the upper bound $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \leq \chi^2(P_\theta, P_\phi)$ [24, Lemma 2.7]. Since $\mathbb{E}[G\phi] = 0$, Jensen's inequality implies that

$$f_\phi(x) \geq \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{\|x\|^2 + \|\phi\|^2}{2\sigma^2}} e^{\frac{1}{\sigma^2}\mathbb{E}[x^T G\phi]} = \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{\|x\|^2 + \|\phi\|^2}{2\sigma^2}}.$$

Hence

$$\frac{(f_\theta(x) - f_\phi(x))^2}{f_\phi(x)} \leq \frac{1}{\sigma^d (2\pi)^{d/2}} e^{\frac{-\|x\|^2 + \|\phi\|^2}{2\sigma^2}} \Big(e^{-\frac{\|\theta\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{1}{\sigma^2} x^T G\theta}\big] - e^{-\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{1}{\sigma^2} x^T G\phi}\big] \Big)^2.$$

To obtain our desired bound on the χ^2-divergence, we will integrate both sides with respect to x. Expanding out the square on the right-hand side yield three terms, which we will evaluate separately. The first term is

$$e^{\frac{\|\phi\|^2 - 2\|\theta\|^2}{2\sigma^2}} \int_{\mathbb{R}^d} \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{1}{\sigma^2} x^T (G + G')\theta}\big]\, dx$$

where G' denotes an independent and identically distributed copy of G. To simplify the expression, we seek to rewrite it as the integral of the density of a Gaussian by completing the square. We obtain

$$\begin{aligned}
& e^{\frac{\|\phi\|^2 - 2\|\theta\|^2}{2\sigma^2}} \int_{\mathbb{R}^d} \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{1}{\sigma^2} x^T (G + G')\theta}\big]\, dx \\
&= e^{\frac{\|\phi\|^2 - 2\|\theta\|^2}{2\sigma^2}} \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{1}{2\sigma^2}(x - (G+G')\theta)^T (x - (G+G')\theta)} dx \cdot e^{\frac{1}{2\sigma^2}((G+G')\theta)^T((G+G')\theta)} \right] \\
&= e^{\frac{\|\phi\|^2 - 2\|\theta\|^2}{2\sigma^2}} \mathbb{E}\Big[e^{\frac{1}{2\sigma^2}((G+G')\theta)^T((G+G')\theta)} \Big] \\
&= e^{\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{\theta^T G\theta}{\sigma^2}} \big]
\end{aligned}$$

Via similar computations, the second and third terms evaluate to

$$-2e^{\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{\theta^T G\phi}{\sigma^2}}\big] \qquad \text{and} \qquad e^{\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}\big[e^{\frac{\phi^T G\phi}{\sigma^2}}\big]$$

respectively. As $\|\theta-\phi\|\le 1/3$ and $\|\theta\|=1$, we have that $\|\phi\|^2\le 16/9$ and so $e^{\frac{\|\phi\|^2}{2\sigma^2}}\le 3$. A power series expansion, together with Lemma 2.2.5, yields

$$\begin{aligned}\chi^2(P_\theta,P_\phi) &\le 3\mathbb{E}\Big[e^{\frac{\theta^T G\theta}{\sigma^2}}-2e^{\frac{\theta^T G\phi}{\sigma^2}}+e^{\frac{\phi^T G\phi}{\sigma^2}}\Big]\\ &=3\sum_{m=1}^{\infty}\frac{1}{\sigma^{2m}m!}\mathbb{E}\big[(\theta^T G\theta)^m-2(\theta^T G\phi)^m+(\phi^T G\phi)^m\big]\\ &=3\sum_{m=1}^{\infty}\frac{1}{\sigma^{2m}m!}\left\|\Delta_m(\theta,\phi)\right\|^2.\end{aligned}$$

To complete the proof, it remains to bound the tail of the summation.

Lemma 2.3.2. For any positive integer m and vectors $\theta,\phi\in\mathbb{R}^d$ satisfying $\|\theta\|=1$ and $\rho(\theta,\phi)\le 1/3$, we have

$$\left\|\Delta_m(\theta,\phi)\right\|^2\le 12\cdot 2^m\rho(\theta,\phi)^2.$$

Proof. (of Lemma 2.3.2) We may again assume that $\rho(\theta,\phi)=\|\theta-\phi\|$. Let $\epsilon=\|\theta-\phi\|$ and $\gamma=\langle\theta,\phi-\theta\rangle$. By Jensen's inequality,

$$\begin{aligned}\left\|\Delta_m(\theta,\phi)\right\|^2 &\le \mathbb{E}\Big[\left\|(G\theta)^{\otimes m}-(G\phi)^{\otimes m}\right\|^2\Big]\\ &=\left\|\theta^{\otimes m}-\phi^{\otimes m}\right\|^2\\ &=1-2\langle\theta,\phi\rangle^m+\|\phi\|^{2m}\\ &=1-2(1+\gamma)^m+(1+2\gamma+\epsilon^2)^m.\end{aligned}$$

Note that $|\gamma|\le\epsilon\le 1/3$ by Cauchy-Schwartz. Since ϵ and γ are small, we perform binomial expansion to the second order. For any $x\in\mathbb{R}$ satisfying $|x|\le 1$, a direct expansion

$$(1+x)^m=1+mx+\sum_{k=2}^{m}\binom{m}{k}x^k$$

gives the bound $1+mx-2^m x^2\le(1+x)^m\le 1+mx+2^m x^2$.

Since $|2\gamma+\epsilon^2|\le 3\epsilon\le 1$, we apply the above to obtain

$$\begin{aligned}\left\|\Delta_m(\theta,\phi)\right\|^2 &\le 1-2(1+m\gamma-2^m\gamma^2)+1+m(2\gamma+\epsilon^2)+2^m(2\gamma+\epsilon^2)^2\\ &\le 2\cdot 2^m\epsilon^2+m\epsilon^2+9\cdot 2^m\epsilon^2\le 12\cdot 2^m\epsilon^2\end{aligned}$$

as desired. □

With the lemma, and using the fact that $\sigma\ge 1$, we conclude the proof with

$$3\sum_{m=k}^{\infty}\frac{\left\|\Delta_m(\theta,\phi)\right\|^2}{\sigma^{2m}m!}\le 36\sum_{m=k}^{\infty}\frac{2^m\rho(\theta,\phi)^2}{\sigma^{2m}m!}\le\frac{36\rho(\theta,\phi)^2}{\sigma^{2k}}\sum_{m=k}^{\infty}\frac{2^m}{m!}=\frac{36e^2\rho(\theta,\phi)^2}{\sigma^{2k}}.$$

□

2.4 Lower Bound for the Kullback-Leibler Divergence

We now turn to the task of establishing the lower bound in Theorem 2.3.1. While the proof of the upper bound consist mostly of elementary (but tedious) calculations, the proof for the lower bound relies on Fourier-analytic arguments and is considerably more advanced in nature. In preparation for the proof, we first introduce a new piece of technology.

Definition 2.4.1. The ***(probabilist's) Hermite polynomials*** are a family of polynomials $\{h_k\}_{k=0}^{\infty}$ defined by

$$h_k(x) := (-1)^k e^{\frac{x^2}{2}} \frac{\partial^k}{\partial x^k} e^{-\frac{x^2}{2}}, \qquad k \in \mathbb{Z}_{\geq 0}.$$

Fact 2.4.2. The Hermite polynomials satisfy the following basic properties:

(i) The polynomial $h_k(x)$ has degree k;

(ii) The family $\{h_k\}_{k=0}^{\infty}$ is an orthogonal basis for $L^2(\mathbb{R}, \gamma)$, where γ denotes the standard Gaussian measure on $\mathbb{R}$;

(iii) We have $\|h_k\|^2_{L^2(\mathbb{R},\gamma)} = k!$;

(iv) For any $\mu \in \mathbb{R}$, we have $\underset{Y\sim\mathcal{N}(\mu,1)}{\mathbb{E}}\big[h_k(Y)\big] = \mu^k$.

The Hermite polynomials serve as the analogue of the Fourier series for the Gaussian Hilbert space $L^2(\mathbb{R}, \gamma)$. Our first step will be to generalise the Hermite polynomials to the multidimensional setting to suit our context.

Definition 2.4.3. For each multi-index $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}_0^d$, define the multivariate polynomial h_α by

$$h_\alpha(x_1, \cdots, x_d) := \prod_{i=1}^{d} h_{\alpha_i}(x_i).$$

The family $\left\{h_\alpha : \alpha \in \mathbb{N}_0^d\right\}$ is called the ***multivariate Hermite polynomials***.

The multivariate Hermite polynomials form an orthogonal basis for the product space $L^2(\mathbb{R}^d, \gamma^{\otimes d})$ [15, Theorem 2.6]. By properties (ii), (iii) and (iv) of Fact 2.4.2, the family of rescaled Hermite polynomials $\{H_\alpha \ : \ \alpha \in \mathbb{N}_0^d\}$ defined by

$$H_\alpha(x_1, \cdots, x_d) := \sigma^{|\alpha|} h_\alpha(\sigma^{-1}x_1, \cdots, \sigma^{-1}x_d) \tag{9}$$

satisfy the following identities

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)}\big[H_\alpha(Z)\big] = \prod_{i=1}^d \mu_i^{\alpha_i} \tag{10}$$

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}_d)}\big[H_\alpha(Z)H_\beta(Z)\big] = \begin{cases}\sigma^{2|\alpha|}\alpha! & \text{if } \alpha=\beta\\ 0 & \text{otherwise.}\end{cases} \tag{11}$$

Next, for each positive integer m, we define the function $H_m : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes m}$ in the following way. Given $x \in \mathbb{R}^d$, set $H_m(x)$ to be the order-m symmetric tensor whose $(i_1,\cdots,i_m)$th-entry is given by $H_{\alpha^{(i_1\cdots i_m)}}(x)$, where $\alpha^{(i_1\cdots i_m)} \in \{0,\cdots,m\}^d$ is the multi-index associated to $(i_1,\cdots,i_m)$:

$$\alpha_\ell^{(i_1\cdots i_m)} := \big|\big\{j\in[m] \ : \ i_j = \ell\big\}\big|, \qquad 1\le \ell\le d.$$

Note that $|\alpha^{(i_1\cdots i_m)}| = m$ for each m-tuple $(i_1,\cdots,i_m)\in[d]^m$.

The motivation behind the above definition will gradually become apparent once we write the quantities $\|\Delta_m(\theta,\phi)\|^2$ in terms of the family $(H_m)_{m=1}^\infty$, where θ and ϕ are arbitrary vectors in $\mathbb{R}^d$. For a positive integer k, consider the degree k polynomial

$$T_k(x) := \sum_{m=1}^k \frac{\langle\Delta_m(\theta,\phi),H_m(x)\rangle}{(\sqrt{3}\sigma)^{2m}m!}.$$

If $X\sim P_\theta$, then (10) implies that

$$\mathbb{E}[T_k(X)] = \mathbb{E}_G\left[\sum_{m=1}^k \frac{\big\langle\Delta_m(\theta,\phi),\mathbb{E}_\xi[H_m(X)]\big\rangle}{(\sqrt{3}\sigma)^{2m}m!}\right] = \sum_{m=1}^k \frac{\big\langle\Delta_m(\theta,\phi),\mathbb{E}[(G\theta)^{\otimes m}]\big\rangle}{(\sqrt{3}\sigma)^{2m}m!}.$$

Hence if $Y\sim P_\phi$, we get

$$\mathbb{E}[T_k(X)] - \mathbb{E}[T_k(Y)] = \sum_{m=1}^k \frac{\|\Delta_m(\theta,\phi)\|^2}{(\sqrt{3}\sigma)^{2m}m!} =: \delta.$$

To proceed, we will use the following lemma to relate the KL divergence between P_θ and P_ϕ to the quantity δ.

Lemma 2.4.4. Let P_1 and P_2 be any two probability distributions on a measure space $(\Omega,\mathcal{F})$. Suppose that there exists a measurable function $F:\Omega\to\mathbb{R}$ such that $\big(\mathbb{E}_{P_1}[F(X)] - \mathbb{E}_{P_2}[F(X)]\big)^2 = \mu^2$ and $\max\big\{\mathrm{var}_{P_1}(F(X)),\mathrm{var}_{P_2}(F(X))\big\}\le\sigma^2$. Then

$$D_{\mathrm{KL}}(P_1 \,\|\, P_2) \ge \frac{\mu^2}{4\sigma^2+\mu^2}. \tag{12}$$

Proof. By replacing F by $F + \lambda$ for a suitably chosen constant λ, we may assume that $\mathbb{E}_{P_1}[F(X)] = \mu/2$ and $\mathbb{E}_{P_2}[F(X)] = -\mu/2$. Let Q_1 and Q_2 denote the corresponding probability distributions of $F(X)$ when X is distributed according to P_1 and P_2 respectively. By the data processing inequality, it suffices to prove the claimed bound for $D_{\mathrm{KL}}(Q_1 \parallel Q_2)$. We further assume that Q_1 is absolutely continuous with respect to Q_2 (otherwise the bound is trivial).

As the quantities involved in (12) arise from taking the expectation of random variables, our approach to establish the inequality will be to pass to the convex function $f : [0, +\infty) \to \mathbb{R}$ defined by

$$f(x) := x \log x - \frac{(x-1)^2}{2(x+1)}$$

and then apply Jensen's inequality. This yields

$$\mathbb{E}_{Q_2}\left[f\left(\frac{dQ_1}{dQ_2}\right) \right] \geq f\left(\mathbb{E}_{Q_2}\left[\frac{dQ_1}{dQ_2}\right] \right) = f(1) = 0.$$

Let μ be a dominating measure (i.e. $Q_1 \ll \mu$ and $Q_2 \ll \mu$) and let q_1 and q_2 denote the densities of Q_1 and Q_2 with respect to μ. The previous calculation implies that

$$D_{\mathrm{KL}}(Q_1 \parallel Q_2) = \mathbb{E}_{Q_2}\left[\frac{dQ_1}{dQ_2} \log \frac{dQ_1}{dQ_2}\right] \geq \frac{1}{2} \int_{\mathbb{R}} \frac{(q_1(x) - q_2(x))^2}{(q_1(x) + q_2(x))} \, d\mu(x).$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} \mu^2 &= \left(\int_{\mathbb{R}} x\big(q_1(x) - q_2(x)\big) \, d\mu(x) \right)^2 \\ &\leq \int_{\mathbb{R}} x^2\big(q_1(x) + q_2(x)\big) \, d\mu(x) \int_{\mathbb{R}} \frac{(q_1(x) - q_2(x))^2}{q_1(x) + q_2(x)} \, d\mu(x) \\ &= (2\sigma^2 + \mu^2/2) \int_{\mathbb{R}} \frac{(q_1(x) - q_2(x))^2}{q_1(x) + q_2(x)} \, d\mu(x). \end{aligned}$$

Hence

$$D_{\mathrm{KL}}(Q_1 \parallel Q_2) \geq \frac{\mu^2}{4\sigma^4 + \mu^2}$$

as desired. □

With the above lemma, our strategy for establishing lower bounds for the KL divergence will be to lower bound the quantity δ and upper bound the variances of both $T_k(X)$ and $T_k(Y)$. To control δ, we will apply Lemma 2.3.2. On the other hand, to control the variances, we will use its Hermite decomposition as a gateway to bring in heavy machinery from harmonic analysis.

Lemma 2.4.5. Fix a positive integer k. Let $\zeta \in \mathbb{R}^d$ and suppose that $Y \sim P_\zeta$. Then for any symmetric tensors $S_1, \cdots, S_k$, where $S_m \in (\mathbb{R}^d)^m$, we have

$$\operatorname{var}\left(\sum_{m=1}^{k} \frac{\langle S_m, H_m(Y)\rangle}{(\sqrt{3}\sigma)^{2m} m!}\right) \leq e^{\frac{\|\zeta\|^2}{2\sigma^2}} \sum_{k=1}^{k} \frac{\|S_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}.$$

Proof. Let $F(x) = \sum_{m=1}^{k} \frac{\langle S_m, H_m(x)\rangle}{(\sqrt{3}\sigma)^{2m} m!}$. To upper bound the variance, it suffices to upper bound the second moment $\mathbb{E}[F(Y)^2]$. Before we are able to bring in results from the theory of Gaussian spaces, we first need to replace P_ζ with the centered multivariate normal distribution $Z \sim P_0 = \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$. To that end, we apply the Cauchy-Schwartz inequality to obtain

$$\begin{aligned}
\mathbb{E}[F(Y)^2] &= \int_{\mathbb{R}^d} f_\zeta(x) F(x)^2 \, dx \\
&\leq \left(\int_{\mathbb{R}^d} f_0(x) F(x)^4 \, dx\right)^{1/2} \left(\int_{\mathbb{R}^d} \frac{f_\zeta(x)^2}{f_0(x)} \, dx\right)^{1/2} \\
&= \mathbb{E}\big[F(Z)^4\big]^{1/2} \left(\int_{\mathbb{R}^d} \frac{f_\zeta(x)^2}{f_0(x)} \, dx\right)^{1/2}. \qquad (13)
\end{aligned}$$

We first address the second term. This is done by proceeding in a similar fashion as in the proof of the upper bound of Theorem 2.3.1. Observe that

$$\int_{\mathbb{R}^d} \frac{f_\zeta(x)^2}{f_0(x)} \, dx = \frac{1}{\sigma^d (2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\mathbb{E}\big[\exp(-\frac{1}{2\sigma^2}(\|x\|^2 - 2x^T G\zeta + \|\zeta\|^2))\big]^2}{\exp(-\frac{1}{2\sigma^2}\|x\|^2)} \, dx.$$

By applying Jensen's inequality and then completing the square afterwards, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{f_\zeta(x)^2}{f_0(x)} \, dx \\
&\leq \mathbb{E}\left[\frac{1}{\sigma^d (2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\exp\big(-\frac{1}{\sigma^2}(\|x\|^2 - 2x^T G\zeta + \|\zeta\|^2)\big)}{\exp(-\frac{1}{2\sigma^2}\|x\|^2)} \, dy.\right] \\
&= \mathbb{E}\left[\frac{1}{\sigma^d (2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\sigma^2}\|x\|^2 + \frac{2}{\sigma^2} x^T G - \frac{2}{\sigma^2}\|\zeta\|^2\right) \, dy \cdot \exp\left(\frac{1}{\sigma^2}\|\zeta\|^2\right)\right] \\
&= \exp\left(\frac{1}{\sigma^2}\|\zeta\|^2\right). \qquad (14)
\end{aligned}$$

We now come to the crux of the matter, which is to establish an upper bound on the first term $(\mathbb{E}[F(Z)^4])^{1/2}$. To accomplish that goal, we bring in some standard results about the ***Ornstein-Uhlenbeck semigroup***, which is a family of operators $U_\rho : L^2(\mathbb{R}^d, \gamma^{\otimes d}) \to L^2(\mathbb{R}^d, \gamma^{\otimes d})$ defined by

$$U_\rho(f)(z) := \mathop{\mathbb{E}}_{Z' \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)} \Big[f\big(\rho z + \sqrt{1-\rho^2} \cdot Z'\big) \Big], \qquad \rho \in [-1, 1].$$

For a detailed overview, see for example [18, Chapter 11]. Here, we highlight that our definition of U_ρ differs from the standard definition in the literature in the sense that expectation is taken with respect to a multivariate normal distribution with covariance matrix $\sigma^2 \boldsymbol{I}_d$ as opposed to $\boldsymbol{I}_d$ to compensate for the scaling in (9).

The set $\{H_\alpha \,:\, \alpha \in \mathbb{N}_0^d\}$ is an eigenbasis for the family (U_ρ), with $U_\rho(H_\alpha) = \rho^{|\alpha|} H_\alpha$ [18, Proposition 11.37]. By viewing $\langle S_m, H_m(x) \rangle$ as a polynomial in x, we get

$$\begin{aligned} U_\rho\big(\langle S_m, H_m(x) \rangle\big) &= U_\rho \left(\sum_{1 \le i_1, \cdots, i_m \le d} (S_m)_{i_1 \cdots i_m} (H_m)_{i_1 \cdots i_m} \right) \\ &= \sum_{1 \le i_1, \cdots, i_m \le d} (S_m)_{i_1 \cdots i_m} U_\rho\big(H_{\alpha^{(i_1 \cdots i_m)}}\big) \\ &= \rho^m \sum_{1 \le i_1, \cdots, i_m \le d} (S_m)_{i_1 \cdots i_m} H_{\alpha^{(i_1 \cdots i_m)}} = \rho^m \langle S_m, H_m(x) \rangle, \end{aligned}$$

where we have used the fact that $|\alpha^{(i_1 \cdots i_m)}| = m$. Thus if we define the degree k polynomial

$$\widetilde{F}(x) := \sum_{m=1}^{k} \frac{\langle S_m, H_m(x) \rangle}{(\sqrt{3})^m \sigma^{2m} m!},$$

then $U_{1/\sqrt{3}}(\widetilde{F}) = F$. The Gaussian hypercontractivity theorem [18, Theorem 11.23] implies that

$$\mathbb{E}\big[F(Z)^4\big]^{1/2} \le \mathbb{E}\big[\widetilde{F}(Z)^2\big].$$

It remains to compute $\mathbb{E}[\widetilde{F}(Z)^2]$. Thankfully, the orthogonality relations in (11) means that when we expand the square, most of the terms will vanish. Since both S_m and $H_m(x)$ are both symmetric tensors, for any tuple $(i_1, \cdots, i_m) \in [d]^m$, the quantities $(S_m)_{i_1 \cdots i_m}$ and $(H_m(x))_{i_1 \cdots i_m}$ depend only on the multi-set $\{i_1, \cdots, i_m\}$. Thus for each $\alpha \in \mathbb{N}_0^d$ such that $|\alpha| = m$, if we define

$$S_\alpha := (S_m)_{i_1 \cdots i_m}$$

where $(i_1, \cdots, i_m)$ is any m-tuple satisfying $\alpha^{(i_1 \cdots i_m)} = \alpha$, we have that

$$\langle S_m, H_m(x) \rangle = \sum_{|\alpha| = m} \frac{m!}{\alpha!} S_\alpha H_\alpha(x).$$

Applying the orthogonality relations in (11), we obtain

$$\begin{aligned}
\mathbb{E}\big[F(Z)^4\big]^{1/2} \le \mathbb{E}\big[\widetilde{F}(Z)^2\big] &= \sum_{m=1}^{k} \mathbb{E}\Bigg[\bigg(\frac{1}{(\sqrt{3})^m \sigma^{2m} m!} \sum_{|\alpha|=m} \frac{m!}{\alpha!} S_\alpha H_\alpha(Z)\bigg)^2\Bigg] \\
&= \sum_{m=1}^{k} \frac{1}{3^m \sigma^{4m}} \sum_{|\alpha|=m} \frac{S_\alpha^2}{\alpha!^2} \cdot \mathbb{E}\big[H_\alpha(Z)^2\big] \\
&= \sum_{m=1}^{k} \frac{1}{(\sqrt{3}\sigma)^{2m} m!} \sum_{|\alpha|=m} \frac{m!}{\alpha!} S_\alpha^2 \\
&= \sum_{m=1}^{k} \frac{\|S_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}. \qquad (15)
\end{aligned}$$

Plugging in (14) and (15) back into (13) gives the desired conclusion. □

We now conclude our discussion in this section by putting together everything that was introduced.

Proof. (of the lower bound in Theorem 2.3.1) As in the proof of the upper bound, we assume that $\rho(\theta, \phi) = \|\theta - \phi\|$, $\|\theta\| = 1$ and $\sigma \ge 1$. By Lemma 2.3.2, we have

$$\delta = \sum_{m=1}^{k} \frac{\|\Delta_m(\theta,\phi)\|^2}{(\sqrt{3}\sigma)^{2m} m!} \le 12\rho(\theta,\phi)^2 \sum_{m=0}^{\infty} \frac{2^m}{3^m m!} \le 12 \cdot \frac{e^{2/3}}{9} \le 4.$$

By Lemma 2.4.5, since $\|\theta\|^2 = 1$ and $\|\phi\|^2 \le 16/9$, the variances of $T_k(X)$ and $T_k(Y)$ are bounded above by $e\delta$. Applying Lemma 2.4.4 then gives

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \ge \frac{\delta^2}{4e\delta + \delta^2} \ge \frac{\delta^2}{4(e\delta + \delta)} > \frac{1}{4(e+1)} \sum_{m=1}^{k} \frac{\|\Delta_m(\theta,\phi)\|^2}{(\sqrt{3}\sigma)^{2m} m!}.$$

Finally, recall that the integer k was arbitrarily chosen. As the summands are nonnegative, letting $k \to \infty$ gives the desired result. □

By far the most important application of Theorem 2.3.1 is that it allows us to reduce the task of establishing bounds on the KL divergence to establishing bounds on the family of moment tensors. In some cases, it suffices to obtain a good bound on the low-order moment tensors ($m = 1, 2, 3$), which are well-understood objects. The full power of the theorem will be illustrated in chapter 4, where we will look at specific subgroups of the orthogonal group.

3 Maximum Likelihood Estimation in Algebraically Structured Models

In this chapter, we will continue our journey in developing the technology needed to study algebraically structured models. The tools introduced in this chapter are geared towards analysing the maximum likelihood estimator (MLE) $\tilde{\theta}_n$ defined by

$$\tilde{\theta}_n := \operatorname*{argmax}_{\phi \in \mathbb{R}^d} \sum_{i=1}^{n} \log \mathbb{E}_G \left[\exp\Big(-\frac{1}{2\sigma^2} \left\| X_i - G\phi \right\|^2 \Big) \right].$$

Here, $X_1, \cdots, X_n$ are i.i.d samples drawn according to P_θ, where $\theta \in \mathbb{R}^d$ is the true signal. We assume that θ is fixed throughout this chapter.

In the final section of this chapter, we will prove the following theorem, which gives a uniform upper bound on the rate of convergence of the MLE under certain conditions.

Theorem 3.0.1. Let $\mathcal{L}$ be any $\mathcal{G}$-invariant vector subspace of $\mathbb{R}$. Suppose that there exists a positive integer k and a positive constant C such that for all $\theta \in \mathcal{S} \cap \mathcal{L}$ and $\phi \in \mathcal{L}$,

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \geq C\sigma^{-2k} \rho(\theta, \phi)^2. \tag{16}$$

Then there exist a positive constant $C'_{d,k}$, depending on d and k, such that the MLE $\tilde{\theta}_n$ constrained to lie in $\mathcal{L}$ satisfies

$$\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\big] \leq C'_{d,k} \left(\frac{\sigma^k}{\sqrt{n}} + \sigma^{6k-5} \frac{\log n}{n} \right)$$

uniformly over all $\theta \in \mathcal{S} \cap \mathcal{L}$.

3.1 Information Geometry for Algebraically Structured Models

In the discussion preceding Proposition 2.1.3 of the previous chapter, we have already witnessed the dependence of the performance of an estimator on the KL divergence $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$. Thus, to understand the rate of convergence of the MLE, our first task will be to understand the relationship between the KL divergence and the orbit distance $\rho(\theta, \phi)$ for vectors ϕ in a (small) neighbourhood of the true value θ. By Remark 2.2.2, this problem is essentially equivalent to understanding how the quantity $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ varies with $\|\theta - \phi\|$. This motivates us to study the geometry of the map $D : \mathbb{R}^d \to \mathbb{R}$ defined by

$$D(\phi) := D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = \mathbb{E}\left[\log \frac{f_\theta(X)}{f_\phi(X)}\right]$$

where $X \sim P_\theta$, using tools from differential calculus.

Our strategy will be to expand D as a Taylor series in a neighbourhood of θ. Since $D(\phi)$ has a global minimum at $\phi = \theta$, both $D(\theta)$ and the gradient $\nabla D(\theta)$ vanishes. Consequently, the Hessian matrix $H_D(\theta)$ becomes the leading term in the Taylor expansion and determines the local behaviour of the KL divergence around θ.

We proceed in two stages. In the first stage, we fix an arbitrary point $x \in \mathbb{R}^d$ and study the behaviour of the log-likelihood ratio function $g_x : \mathbb{R}^d \to \mathbb{R}$ defined by

$$g_x(\phi) := \log \frac{f_\theta(x)}{f_\phi(x)} = \frac{1}{2\sigma^2}\big(\,\|\phi\|^2 - \|\theta\|^2\,\big) + \log \frac{\mathbb{E}\big[\exp(\frac{1}{\sigma^2}x^T G\theta)\big]}{\mathbb{E}\big[\exp(\frac{1}{\sigma^2}x^T G\phi)\big]}.$$

In the second stage, we will study the function D by viewing it as the expectation of the random variable $g_X(\phi)$. The intention is to carry over the results established in the first stage by differentiating under the integral sign.

As the Taylor expansion of multivariate functions typically involve complicated indexed sums, we temporary pause the discussion and introduce the notion of the derivative tensor to simplify notations.

Definition 3.1.1. Let m be a positive integer and let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. For any point $\zeta \in \mathbb{R}^d$, define the ***mth derivative tensor*** $T_\zeta^{(m)} f$ of f to be the order-m symmetric tensor whose $(i_1, \cdots, i_m)$-entry is given by

$$\big(T_\zeta^{(m)} f\big)_{i_1,\cdots,i_m} := \frac{\partial^m f}{\partial \zeta_{i_1} \cdots \partial \zeta_{i_m}}(\zeta).$$

The first and second derivative tensors of f can be canonically identified with the gradient and the Hessian of f respectively.

By using the language of the derivative tensors, the sum of the all the mth order partial derivatives of f in its Taylor expansion can be written succinctly as

$$\sum_{1 \le i_1, \cdots, i_m \le d} \frac{\partial^m f}{\partial \zeta_{i_1} \cdots \partial \zeta_{i_m}}(\zeta) \cdot (\phi_{i_1} - \zeta_{i_1}) \cdots (\phi_{i_m} - \zeta_{i_m}) = \big\langle T_\zeta^{(m)} f, (\phi - \zeta)^{\otimes m} \big\rangle.$$

We now resume the discussion. In the first stage, we will focus on establishing upper bounds on the Hessian matrix of g_x. To that end, we compute the derivatives of g_x up to the third order. The computations are divided into three separate lemmas as each lemma warrants independent interest.

Lemma 3.1.2. Let $h_x(\zeta) := \mathbb{E}\big[\exp(\frac{1}{\sigma^2}x^T G\zeta)\big]$. Then for any positive integer m and any vectors $\zeta, u^{(1)}, \cdots, u^{(m)} \in \mathbb{R}^d$, we have that

$$\left| \left\langle \frac{T_\zeta^{(m)} h_x}{h_x(\zeta)}, u^{(1)} \otimes \cdots \otimes u^{(m)} \right\rangle \right| \le \sigma^{-2m} \|x\|^m \prod_{j=1}^m \|u^{(j)}\|.$$

Proof. By a direct computation, we obtain

$$\begin{aligned}
&\left|\left\langle \frac{T_\zeta^{(m)} h_x}{h_x(\zeta)}, u^{(1)} \otimes \cdots \otimes u^{(m)} \right\rangle\right| \\
&= \left| \frac{1}{h_x(\zeta)} \sum_{1 \le i_1, \cdots, i_m \le d} \frac{\partial^m h_x}{\partial \zeta_{i_1} \cdots \partial \zeta_{i_m}}(\zeta) \cdot u_{i_1}^{(1)} \cdots u_{i_m}^{(m)} \right| \\
&= \sigma^{-2m} \left| \frac{1}{h_x(\zeta)} \sum_{1 \le i_1, \cdots, i_m \le d} \mathbb{E}\Big[\exp\Big(\frac{1}{\sigma^2} x^T G \zeta\Big) \prod_{j=1}^m (G^T x)_{i_j} u_{i_j}^{(j)} \Big] \right| \\
&= \sigma^{-2m} \left| \frac{1}{h_x(\zeta)} \mathbb{E}\Big[\exp\Big(\frac{1}{\sigma^2} x^T G \zeta\Big) \Big\langle (G^T x)^{\otimes m}, u^{(1)} \otimes \cdots \otimes u^{(m)} \Big\rangle \Big] \right|.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality to the inner product gives

$$\begin{aligned}
\left|\left\langle \frac{T_\zeta^{(m)} h_x}{h_x(\zeta)}, u^{(1)} \otimes \cdots \otimes u^{(m)} \right\rangle\right| &\le \sigma^{-2m} \left| \frac{\mathbb{E}\big[\exp(\frac{1}{\sigma^2} x^T G \zeta) \, \|x\|^m \prod_{j=1}^m \|u^{(j)}\| \big]}{\mathbb{E}\big[\exp(\frac{1}{\sigma^2} x^T G \zeta) \big]} \right| \\
&= \sigma^{-2m} \|x\|^m \prod_{j=1}^m \|u^{(j)}\|.
\end{aligned}$$

as desired. □

Lemma 3.1.3. For any point $\zeta \in \mathbb{R}^d$ and any vectors $v_1, v_2, v_3 \in \mathbb{R}^d$,

$$|\langle T_\zeta^{(2)} g_x, v_1 \otimes v_2 \rangle| \le (\sigma^{-2} + 2\sigma^{-4} \|x\|^2) \|v_1\| \|v_2\| \tag{17}$$

$$|\langle T_\zeta^{(3)} g_x, v_1 \otimes v_2 \otimes v_3 \rangle| \le 6\sigma^{-6} \|x\|^3 \|v_1\| \|v_2\| \|v_3\|. \tag{18}$$

Proof. A direct calculation reveals that

$$T_\zeta^{(1)} g_x = \frac{1}{\sigma^2} \zeta - \frac{T_\zeta^{(1)} h_x(\zeta)}{h_x(\zeta)} \tag{19}$$

$$T_\zeta^{(2)} g_x = \frac{1}{\sigma^2} I_d - \frac{T_\zeta^{(2)} h_x}{h_x(\zeta)} + \left(\frac{T_\zeta^{(1)} h_x}{h_x(\zeta)} \right)^{\otimes 2} \tag{20}$$

$$T_\zeta^{(3)} g_x = -\frac{T_\zeta^{(3)} h_x}{h_x(\zeta)} + 3 \, \mathrm{sym} \left(\frac{T_\zeta^{(2)} h_x}{h_x(\zeta)} \otimes \frac{T_\zeta^{(1)} h_x}{h_x(\zeta)} \right) - 2 \left(\frac{T_\zeta^{(1)} h_x}{h_x(\zeta)} \right)^{\otimes 3} \tag{21}$$

where $h_x(\zeta) = \mathbb{E}\big[\exp(\frac{1}{\sigma^2} x^T G \zeta) \big]$ and sym is the symmetrization operator which acts on order-3 tensors by averaging over all permutations of the indices:

$$\mathrm{sym}(A)_{i_1 i_2 i_3} := \frac{1}{6} \sum_{\pi \in S_3} A_{i_{\pi(1)} i_{\pi(2)} i_{\pi(3)}} \qquad \text{for all } A \in (\mathbb{R}^d)^{\otimes 3}.$$

Substituting Lemma 3.1.2 into (20) and using the triangle inequality, we obtain the first inequality

$$|\langle T_\zeta^{(2)} g_x, v_1 \otimes v_2 \rangle| \leq \sigma^{-2} \|v_1\| \|v_2\| + 2\sigma^{-4} \|x\|^2 \|v_1\| \|v_2\| .$$

To obtain the second inequality, we use the self-adjoint property of the symmetrization operator to obtain

$$\begin{aligned}
&\left| \left\langle 3 \text{ sym} \left(\frac{T_\zeta^{(2)} h_x}{h_x(\zeta)} \otimes \frac{T_\zeta^{(1)} h_x}{h_x(\zeta)} \right), u^{(1)} \otimes u^{(2)} \otimes u^{(3)} \right\rangle \right| \\
&= \frac{1}{2} \left| \left\langle \frac{T_\zeta^{(2)}}{h_x(\zeta)} \otimes \frac{T_\zeta^{(1)} h_x}{h_x(\zeta)}, \sum_{\pi \in S_3} u^{(\pi(1))} \otimes u^{(\pi(2))} \otimes u^{(\pi(3))} \right\rangle \right| \\
&\leq \frac{1}{2} \sum_{\pi \in S_3} \left| \left\langle \frac{T_\zeta^{(2)} h_x}{h_x(\zeta)}, u^{(\pi(1))} \otimes u^{(\pi(2))} \right\rangle \right| \cdot \left| \left\langle \frac{T_\zeta^{(1)} h_x}{h_x(\zeta)}, u^{(\pi(3))} \right\rangle \right| \leq 3\sigma^{-6} \|x\|^3 \prod_{j=1}^{3} \|u^{(j)}\|.
\end{aligned}$$

Together with (21), we get

$$|\langle T_\zeta^{(3)} g_x, v_1 \otimes v_2 \otimes v_3 \rangle| \leq 6\sigma^{-6} \|x\|^3 \|v_1\| \|v_2\| \|v_3\| .$$

□

Lemma 3.1.4. For any $\phi, \eta \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|H_{g_x}(\phi)\|_{\text{op}} &\leq \sigma^{-2} + 2\sigma^{-4} \|x\|^2 \\
\|H_{g_x}(\phi) - H_{g_x}(\eta)\|_{\text{op}} &\leq 6\sigma^{-6} \|x\|^3 \|\phi - \eta\| .
\end{aligned}$$

Proof. The first inequality follows directly from (17) in Lemma 3.1.3 since

$$\|H_{g_x}(\phi)\|_{\text{op}} = \sup_{\substack{v,w \in \mathbb{R}^d \\ \|v\|=\|w\|=1}} |v^T H_{g_x}(\phi) w| = \sup_{\substack{v,w \in \mathbb{R}^d \\ \|v\|=\|w\|=1}} |\langle T_\phi^{(2)} g_x, v \otimes w \rangle|.$$

For the second inequality, first fix any $v, w \in \mathbb{R}^d$ and write

$$\langle T_\phi^{(2)} g_x - T_\eta^{(2)} g_x, v \otimes w \rangle = \sum_{1 \leq i,j \leq d} \left(\frac{\partial^2 g_x}{\partial \zeta_i \partial \zeta_j}(\phi) - \frac{\partial^2 g_x}{\partial \zeta_i \partial \zeta_j}(\eta) \right) v_i w_j.$$

Applying the fundamental theorem of calculus to the map $\lambda \mapsto \frac{\partial^2 g_x}{\partial \zeta_i \partial \zeta_j}\big(\eta + \lambda(\phi - \eta)\big)$,

$$\begin{aligned}
\big|\langle T_\phi^{(2)} g_x - T_\eta^{(2)} g_x, v \otimes w \rangle\big| &= \left| \sum_{1 \leq i,j \leq d} \sum_{k=1}^{d} \int_0^1 \frac{\partial^3 g_x}{\partial \zeta_i \partial \zeta_j \partial \zeta_k}\big(\eta + \lambda(\phi - \eta)\big) v_i w_j (\phi - \eta)_k \, d\lambda \right| \\
&= \left| \int_0^1 \big\langle T_{\eta + \lambda(\phi - \eta)}^{(3)} g_x, v \otimes w \otimes (\phi - \eta) \big\rangle \, d\lambda \right| \\
&\leq 6\sigma^{-6} \|x\|^3 \|v\| \|w\| \|\phi - \eta\| .
\end{aligned}$$

This proves the second statement. □

We now enter the second stage. Recall that the functions $D(\phi)$ and $g_x(\phi)$ are related by

$$D(\phi) = \mathbb{E}\big[g_X(\phi)\big]$$

where $X \sim P_\theta$. By differentiating under the integral sign, we obtain

$$T_\phi^{(m)} D = \mathbb{E}\big[T_\phi^{(m)} g_X\big] \tag{22}$$

for any $\phi \in \mathbb{R}^d$ and any positive integer m. As discussed earlier, we will expand D as a Taylor series at θ to the second order and use the upper bounds established in the first stage to control the error term.

Proposition 3.1.5. There exists a positive constant C_d, depending only on d, such that for any $\theta, \phi \in \mathbb{R}^d$ with $\sigma \geq \|\theta\|$,

$$\Big| D(\phi) - \frac{1}{2}(\phi - \theta)^T H_D(\theta)(\phi - \theta) \Big| \leq C_d \frac{\|\phi - \theta\|^3}{\sigma^3}.$$

Proof. By the mean-value form of Taylor's theorem, there exists a vector $\eta \in \mathbb{R}^d$ on the line segment between ϕ and θ such that

$$D(\phi) = \frac{1}{2}(\phi - \theta)^T H_D(\theta)(\phi - \theta) + \frac{1}{6}\big\langle T_\eta^{(3)} D, (\phi - \theta)^{\otimes 3}\big\rangle.$$

Our goal is to bound the last term. Let C_d be a constant, depending only on d, whose value may change from line to line. Firstly, by (22), we have that

$$\big|\langle T_\eta^{(3)} D, (\phi - \theta)^{\otimes 3}\rangle\big| = \Big|\mathbb{E}\Big[\big\langle T_\eta^{(3)} g_X, (\phi - \theta)^{\otimes 3}\big\rangle\Big]\Big| \leq \mathbb{E}\Big[\big|\langle T_\eta^{(3)} g_X, (\phi - \theta)^{\otimes 3}\rangle\big|\Big].$$

Together with (18) in Lemma 3.1.3,

$$\begin{aligned}
\big|\langle T_\eta^{(3)} D, (\phi - \theta)^{\otimes 3}\rangle\big| &\leq 6\sigma^{-6} \|\phi - \theta\|^3 \, \mathbb{E}\big[\, \|X\|^3 \,\big] \\
&= 6\sigma^{-6} \|\phi - \theta\|^3 \, \mathbb{E}_G\Big[\mathbb{E}_\xi\big[\, \|G\theta + \sigma\xi\|^3 \,\big]\Big] \\
&\leq C_d \sigma^{-6} \|\phi - \theta\|^3 \, \mathbb{E}_G\Big[\mathbb{E}_\xi\big[\, \|G\theta\|^3 + \sigma^3 \|\xi\|^3 \,\big]\Big] \\
&\leq C_d \sigma^{-3} \|\phi - \theta\|^3
\end{aligned}$$

where we have used the fact that $\sigma \geq \|\theta\|$ for the last inequality. □

In the remainder of this section, we delve deeper into the term $(\phi - \theta)^T H_D(\theta)(\phi - \theta)$, which controls the local behaviour of $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$. This is done by using the Hessian $H := H_D(\theta)$ to define a positive semidefinite bilinear form $\langle -, - \rangle_H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by setting

$$\langle v, w \rangle_H := v^T H w \qquad \text{for all } v, w \in \mathbb{R}^d.$$

Let $\|\cdot\|_H$ denote the seminorm induced by this bilinear form. We then define the corresponding dual norm $\|\cdot\|_H^* : \mathbb{R}^d \to [0, \infty]$ by

$$\|v\|_H^* := \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_H = 1}} v^T u.$$

For the sake of exposition, we assume $H \neq 0$. The following auxiliary result provides an explicit description of the dual norm $\|\cdot\|_H^*$ in terms of H.

Proposition 3.1.6. Let $H^\dagger$ denote the Moore-Penrose inverse of H. Then for any vector $v \in \mathbb{R}^d$,

$$\|v\|_H^* = \begin{cases} \sqrt{v^T H^\dagger v} & \text{if } v \text{ lies in the row space of } H \\ +\infty & \text{otherwise.} \end{cases}$$

Remark 3.1.7. As H is positive semidefinite, its Moore-Penrose inverse admits a very simple description. If $P \in O(d)$ is an orthogonal matrix such that

$$PHP^T = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_j, 0, \cdots, 0),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_j > 0$, then $H^\dagger$ is precisely the matrix satisfying

$$PH^\dagger P^T = \operatorname{diag}(1/\lambda_1, 1/\lambda_2, \cdots, 1/\lambda_j, 0, \cdots, 0).$$

In particular, if R and N are the row space and null space of H respectively, then $H^\dagger H|_R \equiv \mathrm{id}$ and $H^\dagger H|_N \equiv 0$.

Proof. We have the orthogonal decomposition of $v = v_1 + v_2$, where $v_1 \in R$ and $v_2 \in N$. If v does not lie in the row space of H, then $v_2 \neq 0$. For any $\lambda \in \mathbb{R}_{>0}$,

$$\|v\|_H^* \geq \frac{v^T(v_1 + \lambda v_2)}{\|v_1 + \lambda v_2\|_H} = \frac{\|v_1\|^2 + \lambda \|v_2\|^2}{\|v_1\|_H} \to +\infty$$

as $\lambda \to +\infty$, proving that $\|v\|_H^* = +\infty$.

On the other hand, suppose that v lies in the row space of H. Let $H^{1/2}$ and $(H^{1/2})^\dagger$ denote the nonnegative square roots of H and $H^\dagger$ respectively. From the discussion in Remark 3.1.7, we see that $v^T(H^{1/2})^\dagger H^{1/2} = v^T$ and so for any $u \in \mathbb{R}^d$ satisfying $\|u\|_H = 1$,

$$v^T u = v^T (H^{1/2})^\dagger H^{1/2} u \leq \|(H^{1/2})^\dagger v\| \|H^{1/2} u\| = \sqrt{v^T H^\dagger v}.$$

We conclude the proof by observing that equality is achieved when $u = \dfrac{H^\dagger v}{\|H^\dagger v\|_H}$. □

3.2 KL Divergence as a Subgaussian Process

The technology that we have acquired in the previous section are inherently local in nature, and are only effective when we restrict our analysis to a small open neighbourhood of θ. Such situations will arise when we analyse the MLE in Section 3.3; for any threshold $\epsilon \in \mathbb{R}_{>0}$, by choosing n sufficiently large, we expect $\rho(\tilde{\theta}_n, \theta) \leq \epsilon$ to hold with high probability. Nevertheless, to establish uniform rates of convergence, we must supplement our local bounds with corresponding global bounds to account for the event in which $\rho(\tilde{\theta}_n, \theta) > \epsilon$.

Thankfully, the density function

$$f_\theta(x) = \frac{1}{\sigma^d (2\pi)^{d/2}} \mathbb{E}\Big[\exp\Big(- \frac{1}{2\sigma^2} \|x - G\theta\|^2 \Big) \Big]$$

already enjoys Gaussian type decay with respect to $\|x\|$. As such, we expect the tail probability $P(\rho(\tilde{\theta}_n, \theta) > \epsilon)$ to decay sharply with n. To make this notion precise, we will tap on the theory of subgaussian processes.

Definition 3.2.1. A random variable Y is ***subgaussian with variance proxy s^2*** if

$$\mathbb{E}\Big[\exp\big(\lambda(Y - \mathbb{E}[Y])\big) \Big] \leq \exp\left(\frac{\lambda^2 s^2}{2} \right) \qquad \text{for all } \lambda \in \mathbb{R}$$

Definition 3.2.2. A random process $(Y_t)_{t \in T}$ on a metric space (T, d) is ***subgaussian with variance proxy s^2*** if $\mathbb{E}[Y_t] = 0$ for each t and

$$\mathbb{E}\Big[\exp\big(\lambda(Y_t - Y_v)\big) \Big] \leq \exp\left(\frac{\lambda^2 s^2 d(t,v)^2}{2} \right) \qquad \text{for all } t, v \in T \text{ and } \lambda \in \mathbb{R}_{\geq 0}.$$

The properties of subgaussian random variables and subgaussian processes have been extensively studied in the literature. For a detailed treatment, see for example [25].

Recall that $X_1, \cdots, X_n$ are i.i.d samples drawn according to P_θ. Thus the MLE $\tilde{\theta}_n$ can be viewed as the minimizer of

$$D_n(\phi) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta}{f_\phi}(X_i). \tag{23}$$

In this section, the bulk of the work will be dedicated to framing the KL divergence as a suitable subgaussian process. The end goal is the following theorem.

Theorem 3.2.3. Suppose that $\|\theta\| \leq \sigma$. With respect to the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^d$, the random process $\{\mathfrak{G}_n(\phi)\}_{\phi \in \mathbb{R}^d}$ defined by

$$\mathfrak{G}_n(\phi) := D(\phi) - D_n(\phi)$$

is a subgaussian process with variance proxy $20d/(n\sigma^2)$.

Unfortunately, the exponential function is not very well-suited to arguments that rely on the transferring of inequalities; if Y and Z are two random variables, then $|Y| \leq |Z|$ does not necessarily imply that $\mathbb{E}[\exp(\lambda Y)] \leq \mathbb{E}[\exp(\lambda Z)]$ for all $\lambda \in \mathbb{R}$. To rectify this issue, we work with symmetrized random variables instead.

Lemma 3.2.4. Let Y and Z be two random variables such that $|Y| \leq |Z|$ almost surely and let ϵ be a Rademacher random variable independent of Y and Z. Then

$$\mathbb{E}\big[\exp(\lambda \epsilon Y)\big] \leq \mathbb{E}\big[\exp(\lambda \epsilon Z)\big] \qquad \text{for all } \lambda \in \mathbb{R}.$$

Proof. The function $x \mapsto \cosh(x)$ is increasing on $\mathbb{R}_{\geq 0}$ so conditioning on Y and Z gives

$$\mathbb{E}\big[\exp(\lambda \epsilon Y) \mid Y\big] = \cosh\big(|\lambda Y|\big) \leq \cosh(|\lambda Z|\big) = \mathbb{E}\big[\exp(\lambda \epsilon Z) \mid Z\big]$$

almost surely. The claim follows. □

Symmetrization works particularly well with subgaussian random variables and subgaussian processes as many crucial properties are preserved under symmetrization (albeit at the cost of increased variance proxy or additional universal constants).

Lemma 3.2.5. If Y is a subgaussian random variable with variance proxy s^2 and ϵ is a Rademacher random variable independent of Y, then ϵY is subgaussian with variance proxy $s^2 + \mathbb{E}[Y]^2$.

Proof. By a direct computation,

$$\begin{aligned}
&\mathbb{E}\big[\exp(\lambda \epsilon Y)\big] \\
&= \frac{1}{2}\mathbb{E}\big[\exp(\lambda Y)\big] + \frac{1}{2}\mathbb{E}\big[\exp(-\lambda Y)\big] \\
&= \frac{\exp(\lambda \mathbb{E}[Y])}{2}\mathbb{E}\big[\exp(\lambda(Y - \mathbb{E}[Y]))\big] + \frac{\exp(-\lambda \mathbb{E}[Y])}{2}\mathbb{E}\big[\exp(-\lambda(Y - \mathbb{E}[Y]))\big] \\
&\leq \cosh\big(\lambda \mathbb{E}[Y]\big) \exp\left(\frac{\lambda^2 s^2}{2}\right) \\
&\leq \exp\left(\frac{\lambda^2(s^2 + \mathbb{E}[Y]^2)}{2}\right)
\end{aligned}$$

where we used the elementary inequality $\cosh(x) \leq \exp(x^2/2)$ in the last step. □

Lemma 3.2.6. Let Y be a random variable and let ϵ be a Rademacher random variable independent of Y. Then

$$\mathbb{E}\Big[\exp\big(Y - \mathbb{E}[Y]\big)\Big] \leq \mathbb{E}\Big[\exp\big(2\epsilon|Y|\big)\Big].$$

Proof. Let Y' denote an independent and identically distributed copy of Y. An application of Jensen's inequality gives

$$\mathbb{E}\Big[\exp\big(Y - \mathbb{E}[Y]\big)\Big] = \mathbb{E}_Y\Big[\exp\big(\mathbb{E}_{Y'}[Y - Y']\big)\Big] \leq \mathbb{E}\Big[\exp\big(Y - Y'\big)\Big].$$

As $Y - Y'$ is symmetric, the distributions of $Y - Y'$ and $\epsilon(Y - Y')$ are identical. By Lemma 3.2.4,

$$\mathbb{E}\Big[\exp\big(Y - Y'\big)\Big] = \mathbb{E}\Big[\exp\big(\epsilon(Y - Y')\big)\Big] \leq \mathbb{E}\Big[\exp\big(\epsilon|Y| + \epsilon|Y'|\big)\Big].$$

Since ϵY and $\epsilon Y'$ are independent given ϵ, we obtain

$$\begin{aligned}\mathbb{E}\Big[\exp\big(Y - Y'\big)\Big] &\leq \mathbb{E}_\epsilon\Big[\mathbb{E}_Y\Big[\exp\big(\epsilon|Y|\big)\Big]\mathbb{E}_{Y'}\Big[\exp\big(\epsilon|Y'|\big)\Big]\Big] \\ &= \mathbb{E}_\epsilon\Big[\mathbb{E}_Y\Big[\exp\big(\epsilon|Y|\big)\Big]^2\Big] \\ &\leq \mathbb{E}\Big[\exp\big(2\epsilon|Y|\big)\Big]\end{aligned}$$

as desired. □

Remark 3.2.7. The symmetrization toolkit provides an algorithmic approach to obtain subgaussian bounds on a random variable Y. Namely, we first upper bound Y by a random variable Z that is known to be subgaussian. By replacing both Y and Z by their symmetrized versions and then applying Lemma 3.2.4, we can transfer the subgaussian property of Z over to Y. The proof of Theorem 3.2.3 provides an illustration of this technique. Another instance of this technique can be found in the proof of Proposition 4.5.1. See [3, Lemma B.10] for more details.

Proof. (of Theorem 3.2.3) Fix $\phi, \zeta \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$. Let $X \sim P_\theta$. A preliminary calculation yields

$$\begin{aligned}\mathfrak{G}_n(\phi) - \mathfrak{G}_n(\zeta) &= \mathbb{E}\big[\log f_\zeta(X) - \log f_\phi(X)\big] - \frac{1}{n}\sum_{i=1}^n \big(\log f_\zeta(X_i) - \log f_\phi(X_i)\big) \\ &= \mathbb{E}_X\left[\log\frac{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}X^T G\zeta)\big]}{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}X^T G\phi)\big]}\right] - \frac{1}{n}\sum_{i=1}^n \log\frac{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}X_i^T G\zeta)\big]}{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}X_i^T G\phi)\big]} \\ &= \mathbb{E}[F(X)] - \frac{1}{n}\sum_{i=1}^n F(X_i)\end{aligned}$$

where $F(x) := \log\dfrac{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}x^T G\zeta)\big]}{\mathbb{E}_G\big[\exp(\frac{1}{\sigma^2}x^T G\phi)\big]}$.

Next, by applying Lemma 3.2.6 to each term, we have

$$\mathbb{E}\Big[\exp\Big(\lambda\big(\mathfrak{G}_n(\phi)-\mathfrak{G}_n(\zeta)\big)\Big)\Big] \leq \prod_{i=1}^{n}\mathbb{E}\Big[\exp\Big(\epsilon_i\Big|\frac{2\lambda}{n}F(X_i)\Big|\Big)\Big] \tag{24}$$

where $\epsilon_1, \cdots, \epsilon_n$ are i.i.d Rademacher random variables independent of the X_i's. To proceed, we will bound each term in the product on the right-hand side. To that end, we will upper bound the $F(X_i)$'s by subgaussian random variables. For each fixed $x \in \mathbb{R}^d$, let $h_x : \mathbb{R}^d \to \mathbb{R}$ denote the map

$$h_x(\psi) := \mathbb{E}_G\Big[\exp\Big(\frac{1}{\sigma^2}x^T G\psi\Big)\Big].$$

Then

$$|F(X_i)| = |\log h_{X_i}(\zeta) - \log h_{X_i}(\phi)| \leq \sup_{\psi\in\mathbb{R}^d}\|\nabla \log h_{X_i}(\psi)\|\,\|\phi-\zeta\|\,.$$

By Lemma 3.1.2,

$$|F(X_i)| \leq \frac{1}{\sigma^2}\|X_i\|\,\|\phi-\zeta\| \leq \frac{1}{\sigma^2}\big(\,\|\theta\| + \sigma\,\|\xi_i\|\,\big)\,\|\phi-\zeta\|$$

where $\xi_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. As the map $\xi \mapsto \|\theta\| + \sigma\,\|\xi\|$ is σ-Lipschitz, the Tsirelson-Ibragimov-Sudakov inequality [10, Theorem 5.5] implies that $\|\theta\|+\sigma\,\|\xi_i\|$ is subgaussian with variance proxy σ^2. Next, Jensen's inequality gives

$$\mathbb{E}\big[\,\|\xi_i\|\,\big] \leq \sqrt{\mathbb{E}\big[\,\|\xi_i\|^2\,\big]} = \sqrt{d}$$

and so

$$\Big(\mathbb{E}\big[\,\|\theta\| + \sigma\,\|\xi_i\|\,\big]\Big)^2 \leq (\sigma+\sigma\sqrt{d})^2 \leq 4d\sigma^2.$$

Therefore, the symmetrized random variable $\epsilon_i(\|\theta\| + \sigma\,\|\xi_i\|)$ is subgaussian with variance proxy $5d\sigma^2$ by Lemma 3.2.5. By Lemma 3.2.4, we then have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\epsilon_i\Big|\frac{2\lambda}{n}F(X_i)\Big|\right)\right] &\leq \mathbb{E}\left[\exp\left(\frac{2\lambda\epsilon_i(\|\theta\| + \sigma\,\|\xi\|)\,\|\phi-\zeta\|}{n\sigma^2}\right)\right]\\ &\leq \exp\left(\frac{10d\lambda^2\,\|\phi-\zeta\|^2}{n^2\sigma^2}\right)\end{aligned}$$

The conclusion follows by substituting the above upper bound back into (24). □

The payoff of Theorem 3.2.3 is that we are now free to bring in various tools from the literature of subgaussian processes. We will take advantage of this to accomplish our original purpose by employing a standard technique known as the slicing method; see [25, Chapter 5.4] for an overview.

Proposition 3.2.8. Assume the conditions of Theorem 3.0.1. There exists a constant C_d, depending on d, such that the MLE $\tilde{\theta}_n$ constrained to lie in $\mathcal{L}$ satisfies

$$\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n,\theta)^2\big] \leq C_d \frac{\sigma^{4k-2}}{n}$$

uniformly over all $\theta \in \mathcal{S} \cap \mathcal{L}$

Proof. In what follows, let C_d be a constant, depending on d, whose value may change from line to line. To remove the dependence of the variance proxy on the parameters σ and n, we consider the rescaled process $\{\sigma\sqrt{n}\mathfrak{G}_n(\phi)\}_{\phi\in\mathbb{R}^d}$ instead, which is subgaussian with variance proxy $20d$. We first employ the tail bound variant of Dudley's integral inequality [26, Theorem 8.1.6]. Using a volume argument [26, Corollary 4.2.13], an upper bound for the entropy number of the open unit ball $B_1(\theta)$ in $\mathbb{R}^d$ is given by

$$\mathcal{N}\big(B_1(\theta), \|\cdot\|, \epsilon\big) \leq C_d \left(\frac{1}{\epsilon}\right)^d.$$

For any fixed $\delta \in \mathbb{R}_{>0}$, the above display gives the following upper bound for the entropy integral after a change of variables $\gamma = \epsilon/\delta$

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}\big(B_\delta(\theta), \|\cdot\|, \epsilon\big)}\, d\epsilon &= \delta \int_0^1 \sqrt{\log \mathcal{N}\big(B_1(\theta), \|\cdot\|, \gamma\big)}\, d\gamma \\ &\leq \delta \int_0^1 \sqrt{\log C_d - d \log \gamma}\, d\gamma \\ &\leq \delta C_d.\end{aligned}$$

Here, the last line follows from the fact that the integral converges. Applying Dudley's integral inequality gives

$$P\left(\sup_{\phi\in B_\delta(\theta)} \sqrt{n}\sigma\big(\mathfrak{G}_n(\phi) - \mathfrak{G}_n(\theta)\big) \geq C_d\delta + \sqrt{n}\sigma x\right) \leq C_d \exp\left(-\frac{n\sigma^2x^2}{C_d\delta^2}\right)$$

and since $\mathfrak{G}_n(\theta) = 0$, we have that

$$P\left(\sup_{\phi\in B_\delta(\theta)} \mathfrak{G}_n(\phi) \geq C_d\frac{\delta}{\sqrt{n}\sigma} + x\right) \leq C_d \exp\left(-\frac{n\sigma^2x^2}{C_d\delta^2}\right). \tag{25}$$

We now proceed to decompose the supremum into slices. Define the sequence $(\alpha_j)_{j=0}^\infty$ by setting $\alpha_0 = 0$ and $\alpha_j = M_d\sigma^{2k-1}2^{j-1}$ for $j \in \mathbb{Z}_{\geq 1}$, where M_d is another constant, depending on d, whose exact value will be specified later. For each $j \in \mathbb{Z}_{\geq 0}$, define

$$S_j := \Big\{\phi \in \mathbb{R}^d \ : \ \alpha_j \leq \sqrt{n}\rho(\phi,\theta) < \alpha_{j+1}\Big\}.$$

We obtain

$$\mathbb{E}_\theta\big[n\rho(\tilde{\theta}_n,\theta)^2\big] = \sum_{j=0}^{\infty}\mathbb{E}_\theta\Big[n\rho(\tilde{\theta}_n,\theta)^2\mathbb{1}_{\{\tilde{\theta}_n\in S_j\}}\Big] \le M_d^2\sigma^{4k-2} + \sum_{j=1}^{\infty}\alpha_{j+1}^2 P\big(\tilde{\theta}_n\in S_j\big). \tag{26}$$

For each $j\in\mathbb{Z}_{\ge 1}$, we will bound the quantity $P(\tilde{\theta}_n\in S_j)$ by showing that $\mathfrak{G}_n(\tilde{\theta}_n)$ is large when $\tilde{\theta}_n\in S_j$ and then apply (25). On one hand, by definition of the maximum likelihood estimator, we have $D_n(\tilde{\theta}_n)\le D_n(\theta)=0$. On the other hand, on the subspace $\mathcal{L}$, we have $D(\tilde{\theta}_n)\ge C\sigma^{-2k}\rho(\tilde{\theta}_n,\theta)^2$, where C is the constant given in Theorem 3.0.1. Hence if $\tilde{\theta}_n\in S_j$, then

$$\mathfrak{G}_n(\tilde{\theta}_n) = D(\tilde{\theta}_n) - D_n(\tilde{\theta}_n) \ge C\sigma^{-2k}\rho(\tilde{\theta}_n,\theta)^2 \ge C\sigma^{-2k}\frac{\alpha_j^2}{n}.$$

This in turn implies that

$$P\big(\tilde{\theta}_n\in S_j\big) \le P\bigg(\sup_{\phi\in S_j}\mathfrak{G}_n(\phi)\ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\bigg) \le P\bigg(\sup_{\phi\in B_{\frac{\alpha_{j+1}}{\sqrt{n}}}(\theta)}\mathfrak{G}_n(\phi)\ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\bigg).$$

As long as M_d is chosen to be sufficiently large (this is why we require $j\ge 1$), we have $\sigma^{-2k}\alpha_j^2\ge C\alpha_{j+1}/2\sigma$. Applying (25) with $\delta=\alpha_{j+1}/\sqrt{n}$ and $x=C\sigma^{-2k}\alpha_j^2/2n$ gives

$$P\bigg(\sup_{\phi\in B_{\frac{\alpha_{j+1}}{\sqrt{n}}}(\theta)}\mathfrak{G}_n(\phi)\ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\bigg) \le C_d\exp\left(-\frac{\alpha_j^4}{C_d\alpha_{j+1}^2\sigma^{4k-2}}\right) \le C_d\exp\left(-\frac{2^{2j}}{C_d}\right).$$

Plugging the above upper bound back into (26) results in a convergent sum

$$\mathbb{E}\big[\rho(\tilde{\theta}_n,\theta)^2\big] \le \sigma^{4k-2}\frac{M_d^2}{n} + \sigma^{4k-2}\frac{C_d}{n}\sum_{j=1}^{\infty}2^{2j}\exp\left(-\frac{2^{2j}}{C_d}\right) \le C_d\frac{\sigma^{4k-2}}{n}$$

as desired. □

Another instance of the slicing method appears in the proof of Theorem 4.5.3. See [3, Theorem 2] for the full proof.

3.3 Convergence of the Maximum Likelihood Estimator

We now have all the tools necessary to prove the main result of this chapter.

Proof. (of Theorem 3.0.1) As the KL divergence is invariant under the action of $\mathcal{G}$, we assume $\rho(\tilde{\theta}_n,\theta)=\|\tilde{\theta}_n-\theta\|$. Furthermore, since $K^{-1}\le\|\theta\|\le K$ for all $\theta\in\mathcal{S}$, we may also assume $\|\theta\| = 1$ and $\sigma \geq 1$ by replacing θ and σ by $\theta / \|\theta\|$ and $\sigma / \|\theta\|$ respectively (this is the reason why the constant $C'_{d,k}$ depend on k).

Let δ_d and M_d denote positive constants, depending on d, whose value may change from line to line. Define the event $\mathcal{E} := \{\rho(\tilde{\theta}_n, \theta) \leq \epsilon\}$, where ϵ is another positive constant whose exact value will be determined later. We decompose

$$\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\big] = \mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\mathbb{1}_{\mathcal{E}}\big] + \mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\mathbb{1}_{\mathcal{E}^c}\big] \tag{27}$$

and apply different tools to bound each term.

The first term entails the local behaviour of $\tilde{\theta}_n$ and so we will tackle it using the tools developed in Section 3.1. Observe that if δ_d is sufficiently small, then for all $\epsilon \leq \delta_d \sigma^{3-2k}$, Proposition 3.1.5 and (16) implies that

$$\Big| D(\tilde{\theta}_n) - \frac{1}{2}\|\tilde{\theta}_n - \theta\|_H^2 \Big| \leq M_d \frac{\|\tilde{\theta}_n - \theta\|^3}{\sigma^3} \leq \delta_d M_d \sigma^{-2k} \|\tilde{\theta}_n - \theta\|^2 \leq \frac{1}{2} D(\tilde{\theta}_n).$$

Hence

$$\frac{1}{3}\|\tilde{\theta}_n - \theta\|_H^2 \leq D(\tilde{\theta}_n) \leq \|\tilde{\theta}_n - \theta\|_H^2.$$

Together with (16), we get

$$\|\tilde{\theta}_n - \theta\|_H^2 \geq C\sigma^{-2k}\|\tilde{\theta}_n - \theta\|^2. \tag{28}$$

The function D_n defined in (23) satisfy $D_n(\theta) = 0$ and is minimised by $\tilde{\theta}_n$. As such, we must have $D_n(\tilde{\theta}_n) \leq 0$. Expanding $D - D_n$ as a Taylor series about θ, we have

$$\begin{aligned}\frac{1}{3}\|\tilde{\theta}_n - \theta\|_H^2 &\leq D(\tilde{\theta}_n) - D_n(\tilde{\theta}_n) \\ &= -\nabla D_n(\theta)^T(\tilde{\theta}_n - \theta) + \frac{1}{2}(\tilde{\theta}_n - \theta)^T\big(H_D(\eta) - H_{D_n}(\eta)\big)(\tilde{\theta}_n - \theta)\end{aligned}$$

where $\eta \in \mathbb{R}^d$ is a vector on the line segment between θ and $\tilde{\theta}_n$. We employ the dual norm $\|\cdot\|_H^*$ to bound the first term and the operator norm $\|\cdot\|_{\mathrm{op}}$ to bound the second term. This gives

$$\frac{1}{3}\|\tilde{\theta}_n - \theta\|_H^2 \leq \|\nabla D_n(\theta)\|_H^* \, \|\tilde{\theta}_n - \theta\|_H + \frac{1}{2}\|\tilde{\theta}_n - \theta\|^2 \sup_{\phi \in \mathcal{B}_\epsilon} \|H_D(\phi) - H_{D_n}(\phi)\|_{\mathrm{op}}$$

where $\mathcal{B}_\epsilon := \big\{\phi \in \mathbb{R}^d \ : \ \rho(\phi, \theta) \leq \epsilon\big\}$. Using (28) and dividing by $\|\tilde{\theta}_n - \theta\|_H$ throughout on both sides,

$$\sigma^{-k}\|\tilde{\theta}_n - \theta\| \leq M_d\Big(\|\nabla D_n(\theta)\|_H^* + \sigma^k\|\tilde{\theta}_n - \theta\| \sup_{\phi \in \mathcal{B}_\epsilon} \|H_D(\phi) - H_{D_n}(\phi)\|_{\mathrm{op}}\Big).$$

Multiplying both sides by σ^k and applying Young's inequality to the second term on the right-hand side yields

$$\|\tilde{\theta}_n - \theta\| \leq M_d\Big(\sigma^k \left\|\nabla D_n(\theta)\right\|_H^* + \sigma^{2k+3} \sup_{\phi \in \mathcal{B}_\epsilon} \left\|H_D(\phi) - H_{D_n}(\phi)\right\|_{\mathrm{op}}^2 + \sigma^{2k-3}\|\tilde{\theta}_n - \theta\|^2\Big).$$

The purpose of applying Young's inequality to split σ^{2k} into σ^{2k+3} and σ^{2k-3} will become apparent when we merge the above expression with the term $\mathbb{E}_\theta[\rho(\tilde{\theta}_n, \theta)\mathbb{1}_{\mathcal{E}^c}]$ in (27). By choosing $\epsilon = \delta_d \sigma^{3-2k}$ and then taking expectation on both sides,

$$\begin{aligned}\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\mathbb{1}_{\mathcal{E}}\big] \leq{}& M_d\sigma^k \mathbb{E}_\theta\big[\left\|\nabla D_n(\theta)\right\|_H^*\big] + M_d\sigma^{2k+3}\mathbb{E}_\theta\left[\sup_{\phi \in \mathcal{B}_\epsilon} \left\|H_D(\phi) - H_{D_n}(\phi)\right\|_{\mathrm{op}}^2\right] \\ &+ M_d\sigma^{2k-3}\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)^2\big].\end{aligned} \tag{29}$$

Now for the second term in (27), since $\rho(\tilde{\theta}_n, \theta) \geq \delta_d \sigma^{3-2k}$ on $\mathcal{E}^c$, we obtain

$$\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\mathbb{1}_{\mathcal{E}^c}\big] \leq M_d\sigma^{2k-3}\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)^2\big]$$

which can be combined with the third term in (29). The end result is

$$\begin{aligned}\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\big] \leq{}& M_d\sigma^k \mathbb{E}_\theta\big[\left\|\nabla D_n(\theta)\right\|_H^*\big] + M_d\sigma^{2k+3}\mathbb{E}_\theta\left[\sup_{\phi \in \mathcal{B}_\epsilon} \left\|H_D(\phi) - H_{D_n}(\phi)\right\|_{\mathrm{op}}^2\right] \\ &+ M_d\sigma^{2k-3}\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)^2\big].\end{aligned} \tag{30}$$

We will establish an upper bound for each of the three terms in the above equation separately.

For the first term, we seek to apply Proposition 3.1.6. Bartlett's identities (Appendix A) state that for each $1 \leq i \leq n$,

$$\begin{aligned}\mathbb{E}_\theta\big[\nabla \log f_\phi(X_i)|_{\phi=\theta}\big] &= 0 \\ \mathbb{E}_\theta\big[(\nabla \log f_\phi(X_i)|_{\phi=\theta})(\nabla \log f_\phi(X_i)|_{\phi=\theta})^T\big] &= H_D(\theta) = H.\end{aligned}$$

Since $\nabla D_n(\theta) = -\dfrac{1}{n}\displaystyle\sum_{i=1}^n \nabla \log f_\phi(X_i)\Big|_{\phi=\theta}$ and $X_1, \cdots, X_n$ are independent, we get

$$\mathbb{E}_\theta\big[\nabla D_n(\theta)\nabla D_n(\theta)^T\big] = \frac{1}{n}H.$$

In particular, the nullspace of H is contained in the nullspace of $\nabla D_n(\theta)\nabla D_n(\theta)^T$ almost surely since any vector v lying in the nullspace of H satisfy

$$0 \leq \mathbb{E}_\theta\big[v^T\nabla D_n(\theta)\nabla D_n(\theta)^T v\big] = \frac{1}{n}v^T H v = 0.$$

Since the row space of a symmetric matrix is the orthogonal complement of its nullspace, this means that the row space of $\nabla D_n(\theta)\nabla D_n(\theta)^T$ (and hence the row space of $\nabla D_n(\theta)$) is contained in the row space of H almost surely. As a result,

$$\begin{aligned}
\mathbb{E}_\theta\big[\, \|\nabla D_n(\theta)\|_H^* \,\big] &= \mathbb{E}_\theta\left[\sqrt{\nabla D_n(\theta)^T H^\dagger \nabla D_n(\theta)}\right] \\
&\leq \sqrt{\mathbb{E}_\theta\big[\mathrm{Tr}\big(\nabla D_n(\theta)^T H^\dagger \nabla D_n(\theta)\big)\big]} \\
&= \sqrt{\frac{1}{n}\mathrm{Tr}(H^\dagger H)} \leq \sqrt{\frac{d}{n}} \qquad (31)
\end{aligned}$$

where the last inequality follows from Remark 3.1.7.

For the second term, matrix concentration bounds (Appendix B.1) can be applied to show that

$$\mathbb{E}_\theta\left[\sup_{\phi\in\mathcal{B}_\epsilon} \|H_D(\phi) - H_{D_n}(\phi)\|_{\mathrm{op}}^2\right] \leq M_d \frac{\log n}{n\sigma^4}. \qquad (32)$$

Finally, by Proposition 3.2.8,

$$\mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)^2\big] \leq M_d \frac{\sigma^{4k-2}}{n}. \qquad (33)$$

Putting (31, (32) and 33) back into (30), we have

$$\mathbb{E}\big[\rho(\tilde{\theta}_n, \theta)\big] \leq M_d\left(\frac{\sigma^k}{\sqrt{n}} + \frac{\sigma^{2k-1}\log n}{n} + \frac{\sigma^{6k-5}}{n}\right).$$

The conclusion follows after combining the second and third terms. □

Although the guarantee of a uniform rate of convergence in Theorem 3.0.1 is a very strong result, the condition (16) is a serious drawback and the subspace $\mathcal{L}$ must be carefully chosen in order for the theorem to work. In Section 4.5 of the next chapter, we will see an example of how the MLE can be modified to ensure that the condition (16) is met.

4 The Multi-reference Alignment Model

4.1 Introduction to the Multi-reference Alignment Model

Recall that Theorem 2.3.1 provides a way to bound the KL divergence $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ both from above and from below via the moment difference tensors $\{\Delta_m(\theta, \phi)\}_{m=1}^{\infty}$. In this chapter, we will exploit the close connection between these two objects and use the moment tensors to understand the properties of the KL divergence.

Unfortunately, the moment tensors themselves involve an integral over the entire subgroup. As such, explicit computations of the moment tensors for a general subgroup $\mathcal{G}$ quickly becomes intractable in higher dimensions due to the large number of parameters. This motivates us to look for specific subgroups of the orthogonal group in which explicit computation is still feasible.

Among all such instances, the ***Multi-reference Alignment (MRA) model*** is by far the most extensively studied and well-understood. In the MRA model, the subgroup is taken to be the cyclic group $\mathcal{R} := \big\{R_\ell \ : \ \ell \in [d]\big\}$, where the linear operator $R_\ell : \mathbb{R}^d \to \mathbb{R}^d$ is defined to be the shift of coordinates

$$(R_\ell \theta)_j := \theta_{j+\ell}, \qquad \theta \in \mathbb{R}^d \text{ and } j \in [d].$$

In this chapter, all indices are taken modulo d. This greatly simplifies notations when we introduce the discrete Fourier transform in Section 4.2 and the phase shift model in Section 4.4. The observations $X_1, \cdots, X_n$ are now drawn according to

$$X = R\theta + \sigma\xi$$

where R is drawn from $\mathcal{R}$ uniformly and $\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ is the usual standard Gaussian noise that is independent of R. The shift-invariant distance is now given by

$$\rho(\theta, \phi) := \min_{\ell \in [d]} \|\theta - R_\ell \phi\|.$$

While seemingly an oversimplification at first glance, the MRA model in fact holds key insights into the mathematical nature of the original problem. In addition to being a toy model for cryo-EM, this simplified model is also of independent interest in several applications such as radar classification and the theoretical study of the MRA model predates the algebraically structured model ([2]).

4.2 The Discrete Fourier Transform

The most striking characteristic of the MRA model is that the Haar measure on the group $\mathcal{R}$ is reduced to the counting measure. As such, the moment tensors $\mathbb{E}[(R\phi)^{\otimes m}]$

now admit much simpler descriptions

$$\mathbb{E}\big[(R\phi)^{\otimes m}\big] = \frac{1}{d}\sum_{\ell=1}^{d}(R_\ell \phi)^{\otimes m}.$$

In this section, we will delve deeper into the properties of the moment tensors by passing to the Fourier domain.

Definition 4.2.1. The ***discrete Fourier transform (DFT)*** of a vector $\phi \in \mathbb{R}^d$ is the vector $\hat{\phi} \in \mathbb{C}^d$ defined by

$$\hat{\phi}_k := \frac{1}{\sqrt{d}}\sum_{j=1}^{d} e^{\frac{2\pi i jk}{d}}\phi_j, \qquad k \in [d].$$

The $\mathbb{R}$-linearity of the discrete Fourier transform means that for each positive integer m, it can be naturally extended to a $\mathbb{R}$-linear operator on the space $(\mathbb{R}^d)^{\otimes m}$ of order-m tensors. This is done by setting, for each $T \in (\mathbb{R}^d)^{\otimes m}$,

$$\hat{T}_{k_1\cdots k_m} := \frac{1}{\sqrt{d^m}}\sum_{1\le j_1,\cdots,j_m\le d} e^{\frac{2\pi i(j_1k_1+j_2k_2+\cdots+j_mk_m)}{d}} T_{j_1\cdots j_m}, \qquad (k_1,\cdots,k_m)\in[d]^m$$

By Plancherel's theorem, the discrete Fourier transform is a linear isometry (with respect to the Euclidean norm) from $\mathbb{R}^d$ onto the $\mathbb{R}$-subspace

$$\mathcal{C} := \Big\{\phi \in \mathbb{C}^d \ :\ \phi_\ell = \overline{\phi}_{-\ell} \text{ for all } \ell \in [d]\Big\}.$$

Hence we may represent any vector $\phi \in \mathbb{R}^d$ by the d-tuple $\hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_d) \in \mathcal{C}$. We refer to $(\hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_d)$ as the ***Fourier coordinates*** of ϕ. Similarly, we may represent any tensor $\tau^{(1)} \otimes \tau^{(2)} \otimes \cdots \otimes \tau^{(m)} \in (\mathbb{R}^d)^{\otimes m}$ by

$$\widehat{\tau^{(1)}} \otimes \widehat{\tau^{(2)}} \otimes \cdots \otimes \widehat{\tau^{(m)}} \in \mathcal{C}^{\otimes m}.$$

As it turns out, the moment tensors are very simple monomials in Fourier coordinates.

Theorem 4.2.2. For each m-tuple $(k_1,\cdots,k_m) \in [d]^m$, we have

$$\mathbb{E}\big[(\widehat{R\phi})^{\otimes m}\big]_{k_1\cdots k_m} = \begin{cases} \hat{\phi}_{k_1}\hat{\phi}_{k_2}\cdots\hat{\phi}_{k_m} & \text{if } k_1+k_2+\cdots+k_m \equiv 0 \ (\text{mod } d) \\ 0 & \text{otherwise.}\end{cases}$$

Proof. First observe that for each $1 \le \ell \le d$, we have that $\widehat{R_\ell\phi}_k = e^{-\frac{2\pi i \ell k}{d}}\hat{\phi}_k$. Hence

$$\mathbb{E}\big[(\widehat{R\phi})^{\otimes m}\big]_{k_1\cdots k_m} = \frac{1}{d}\sum_{\ell=1}^{d}(\widehat{R_\ell\phi})_{k_1}\cdots(\widehat{R_\ell\phi})_{k_m} = \frac{1}{d}\sum_{\ell=1}^{d} e^{-\frac{2\pi i\ell(k_1+k_2+\cdots+k_m)}{d}}\hat{\phi}_{k_1}\cdots\hat{\phi}_{k_m}.$$

The right-hand side is a geometric series which can be easily computed to yield

$$\sum_{\ell=1}^{d} e^{-\frac{2\pi i\ell(k_1+k_2+\cdots+k_m)}{d}} = \begin{cases} d & \text{if } k_1+k_2+\cdots+k_m \equiv 0 \pmod{d} \\ 0 & \text{otherwise.} \end{cases}$$

The conclusion follows. □

An immediate consequence is that the Fourier domain is the natural setting for the analysis of the MRA model, in an manner analogous to how systems exhibiting radial symmetry are easier to analyse in spherical coordinates as compared to cartesian coordinates. Although deceptively simple in nature, the above theorem is a powerful passageway that connects the MRA model to the fields of harmonic analysis and signal processing. The second moment tensor consist of terms of the form $|\hat{\phi}_i|^2$ and is known as the power spectrum of the signal while the third moment tensor consist of terms of the form $\hat{\phi}_i\hat{\phi}_j\hat{\phi}_{-i-j}$ and is known as the bispectrum. Both quantities are powerful invariants (under the action of the group $\mathcal{R}$) and play an indispensable role in the study of the MRA model.

This is a good place to pause and introduce some new definitions. If $\phi \in \mathbb{R}^d$ is any vector, then the condition

$$\hat{\phi}_\ell = \overline{\hat{\phi}}_{-\ell}, \qquad \ell \in [d]$$

means that ϕ is completely determined by $\hat{\phi}_0, \hat{\phi}_1, \cdots, \hat{\phi}_{\lfloor d/2 \rfloor}$. Hence it suffices to analyse the coefficients of ϕ only at these indices.

Definition 4.2.3. For any vector $\phi \in \mathbb{R}^d$, define the ***positive support*** of ϕ to be

$$\operatorname{psupp}(\hat{\phi}) := \Big\{1 \leq j \leq \lfloor d/2 \rfloor \ : \ \hat{\phi}_j \neq 0\Big\}.$$

Next, we impose the following additional assumption in the MRA model:

There exists an absolute constant K_0 such that $|\hat{\phi}_j| \geq K_0$ for all $j \in \operatorname{psupp}(\hat{\phi})$.

Let $\mathcal{T}$ denote the set of vectors in $\mathcal{S}$ satisfying the above assumption. The above assumption rules out the existence of a sequence of signals $(\phi_k)_{k=1}^{\infty}$ having Fourier coordinates that are nonzero but which approaches 0 as $k \to \infty$. Such situations are not reflective of the signals that occur in real-world applications but yet poses some technical difficulties. For each $s \in \lfloor d/2 \rfloor$, let $\mathcal{T}_s$ denote the set of vectors $\phi \in \mathcal{T}$ satisfying the additional assumption that $\operatorname{psupp}(\hat{\phi}) \subseteq [s]$. The motivation for studying the Fourier support of the signal may not be evident at first, but it will gradually emerge as we see how it is used in the derivation of the results in the next section.

4.3 Moment Matching

The tight lower and upper bounds of the KL divergence $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ in terms of an infinite series in Theorem 2.3.1 suggests that the asymptotic behaviour of the KL divergence (for $\sigma \to \infty$) is controlled by the smallest positive integer m for which $\|\Delta_m(\theta,\phi)\|$ is nonvanishing. This gives rise to the following notion: if we are able to construct two signals θ and ϕ such that $\|\Delta_m(\theta,\phi)\| = 0$ for all $1 \le m \le k-1$, then $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \lesssim \sigma^{-2k}$. Consequently, the number of samples n needed in order to estimate the true signal at a prescribed accuracy should satisfy $n \gtrsim \sigma^{2k}$. The converse is also true in the following sense: if we have a subspace S with the property that for any two distinct vectors $\theta, \phi \in S$, there exists $1 \le m \le k-1$ such that $\|\Delta_m(\theta,\phi)\| > 0$ (i.e. it is possible to uniquely recover a vector from its first $k-1$ moment tensors), then we expect a corresponding sampling complexity upper bound of $n \lesssim \sigma^{2k-2}$.

To realise this concept, we will use Theorem 4.2.2 to either construct the desired pair of signals or show that no such pair exists. This forms the basis for the technique of moment matching, which turns out to have powerful applications. As a first example of this technique, we will establish the following lower bound on the sampling complexity by constructing two signals with matching low order tensors.

Theorem 4.3.1. Let $1 \le s \le \lfloor d/2 \rfloor$. In the MRA model, there exists a universal constant M such that

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \mathcal{T}_s} \mathbb{E}_\theta\left[\rho(\tilde{\theta}_n, \theta)\right] \ge \frac{M}{d} \min\left\{\frac{\sigma^d}{\sqrt{n}},\ 1\right\},$$

where the infimum is taken over all estimators $\tilde{\theta}_n$ on n samples from P_θ.

Proof. Define $\phi \in \mathbb{R}^d$ by

$$\hat{\phi}_j = \begin{cases} 1/\sqrt{2} & \text{if } j \in \{\pm 1\} \\ 0 & \text{otherwise} \end{cases}$$

and define $\tau \in \mathbb{R}^d$ by

$$\hat{\tau}_j = \begin{cases} e^{i\delta}/\sqrt{2} & \text{if } j = 1 \\ e^{-i\delta}/\sqrt{2} & \text{if } j = -1 \\ 0 & \text{otherwise} \end{cases}$$

where $\delta = \dfrac{m_1}{d} \min\left\{\dfrac{\sigma^d}{\sqrt{n}},\ 1\right\}$ for some small universal constant m_1 to be chosen later.

For all $1 \le m \le d-1$, the mth moment difference tensor $\Delta_m(\tau,\phi)$ vanishes. To see this, observe that the entry $\mathbb{E}[(\widehat{R\zeta})^{\otimes m}]_{k_1 \cdots k_m}$, for $\zeta \in \{\tau, \phi\}$, is nonzero only if $k_j \in \{\pm 1\}$ for all j and the congruence equation

$$k_1 + k_2 + \cdots + k_m \equiv 0 \pmod{d}$$

is satisfied. Since $m < d$, the only way for both conditions to hold is if $+1$ and -1 occur an equal number of times in the m-tuple $(k_1, \cdots, k_m)$, in which case

$$\mathbb{E}\big[(\widehat{R\tau})^{\otimes m}\big]_{k_1\cdots k_m} = |\hat{\tau}_1|^m = \frac{1}{2^m} = |\hat{\phi}_1|^m = \mathbb{E}\big[(\widehat{R\phi})^{\otimes m}\big]_{k_1\cdots k_m}.$$

Note that $\mathbb{E}\big[\widehat{R\tau}\big] = \mathbb{E}\big[\widehat{R\phi}\big] = 0$.

Next, we will show that the distance $\rho(\tau, \phi)^2$ between the orbits of τ and ϕ is not too small. By shrinking m_1 if necessary, we may assume $\delta < \pi/d$ (this is the reason for upper bounding $\sigma^d/\sqrt{n}$ by a constant). Plancherel's Theorem gives

$$\rho(\tau, \phi)^2 = \min_{\ell\in[d]} \|\tau - R_\ell\phi\|^2 = \min_{\ell\in[d]} \|\hat{\tau} - \widehat{R_\ell\phi}\|^2 = 2\min_{\ell\in[d]} |e^{i\delta} - e^{\frac{2\pi i\ell k}{d}}|^2 = 2|e^{i\delta} - 1|^2$$

and so there exist a universal constant m_2 such that $m_2\delta^2 \geq \rho(\tau, \phi)^2 \geq m_2^{-1}\delta^2$. Applying Theorem 2.3.1 to the mean zero signals τ and ϕ, we obtain

$$D_{\mathrm{KL}}(P_\tau \parallel P_\phi) \leq \overline{C}\sigma^{-2d}\rho(\tau, \phi)^2 \leq \frac{1}{2n}$$

where the second inequality follows by shrinking m_1 even further if necessary. The conclusion then follows from Proposition 2.1.3. □

Remark 4.3.2. When the two signals ϕ and τ are plotted in the time (or spatial) domain, it becomes clear that both are discretizations of the same underlying continuous signal. Yet, in the MRA model, the two signals lie in different orbits and are difficult to distinguish. This is one of the primary motivations for the phase shift model, which will be introduced in Section 4.4.

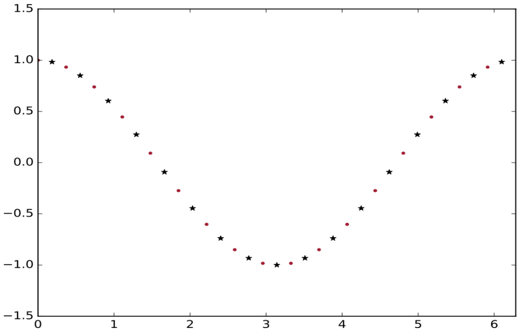

Fig 1. Figure taken from [3] Two (normalised) signals $\phi, \tau \in \mathbb{R}^{17}$ plotted coordinate-wise. The signal ϕ (red dots) and the signal τ (black stars) do not lie in the same orbit of $\mathcal{R}$; however, they do lie in the same orbit of $U(1)$ under the phase shift model.

At first glace, Theorem 4.3.1 paints a rather bleak picture; in real-world applications in which d can be very large, a sample complexity of order σ^d is too high to be considered practical. However, on closer inspection, one notices that the proof hinges on the fact that the Fourier support of the two signals is purposefully limited to the two indices $\{\pm 1\}$ in order to obtain very specific cancellations. Thankfully, such signals are not representative of what is encountered in practice.

The importance of the Fourier support in the MRA model have inspired researchers to investigate the sampling complexity under different assumptions of the Fourier support of the signal. One of the most important classes of signals is the ***generic signals***, which consists only of signals having full Fourier support. For generic signals, the following result in the companion paper [19] shows that sampling complexity can be significantly improved to $O_d(\sigma^6)$.

Theorem 4.3.3. Let $d \geq 3$, let $\sigma \geq 1$ and let $\epsilon > 0$ be sufficiently small. There exists a universal constant M such that for any estimator $\tilde{\theta}_n$ based on n samples from P_θ in the MRA model, there exists with probability at least $1/4$ a signal $\theta \in \mathbb{R}^d$ with $\|\theta\| = 1$ and $|\hat{\theta}_k| = 1/\sqrt{d}$ for all $k \in [d]$ (i.e. θ is a generic signal) such that $\rho(\tilde{\theta}_n, \theta) \geq \epsilon$ when $n \leq M\sigma^6\epsilon^{-2}$.

Sketch of proof. For the sake of exposition, we will only sketch the main idea. The approach will again be to construct two generic signals τ and ϕ whose first and second moment tensors agree. We work systematically in Fourier coordinates. We define $\hat{\tau} := (1/\sqrt{d},\ 1/\sqrt{d},\ \cdots\ ,\ 1/\sqrt{d})$ and define $\hat{\phi}$ by setting

$$\hat{\phi}_k := \begin{cases} e^{i\delta}/\sqrt{d} & \text{if } k = 1 \\ e^{-i\delta}/\sqrt{d} & \text{if } k = -1 \\ 1/\sqrt{d} & \text{otherwise,} \end{cases}$$

where $\delta = c_d\epsilon$ for some constant c_d (depending on d) chosen such that $|\hat{\tau}_1 - \hat{\phi}_1| \approx 2\epsilon$ and

$$\min_{\ell \in [d]} |1 - e^{\frac{2\pi i \ell}{d}}| \geq 2|1 - e^{i\delta}|$$

holds for all ϵ sufficiently small. Note that both ϕ and τ satisfy $\|\phi\| = \|\tau\| = 1$ and $|\hat{\phi}_k| = |\hat{\tau}_k| = 1/\sqrt{d}$ for each $k \in [d]$. An application of Plancherel's theorem gives

$$\rho(\tau, \phi) = \min_{\ell \in [d]} \|\tau - R_\ell\phi\| = \|\tau - \phi\| \approx \epsilon.$$

From Theorem 4.2.2, it is clear that the two signals which we have constructed satisfy

$$\mathbb{E}\big[R\tau\big] = \mathbb{E}\big[R\phi\big] \qquad \text{and} \qquad \mathbb{E}\big[(R\tau)^{\otimes 2}\big] = \mathbb{E}\big[(R\phi)^{\otimes 2}\big].$$

Since the observations $X_1, \cdots, X_n$ are independent and identically distributed, their joint probability distributions are given by the products $P_\tau^{\otimes n}$ and $P_\phi^{\otimes n}$ in accordance to whether $X_i \sim P_\tau$ or $X_i \sim P_\phi$. By Proposition 2.2.3 and Theorem 2.3.1,

$$D_{\mathrm{KL}}(P_\tau^{\otimes n} \parallel P_\phi^{\otimes n}) = nD_{\mathrm{KL}}(P_\tau \parallel P_\phi) \lesssim n\epsilon^2\sigma^{-6}.$$

Lemma 2.1.2 then implies that any estimator $\tilde{\theta}_n$ will fail to distinguish between τ and ϕ with probability at least

$$\frac{2-\sqrt{2D_{\mathrm{KL}}(P_\tau \parallel P_\phi)}}{4} \geq \frac{1}{4}$$

when $n \lesssim \sigma^6\epsilon^{-2}$. The conclusion follows since $\rho(\tau,\phi) \approx \epsilon$. □

The above result illustrates that a sample complexity of $O_d(\sigma^6)$ for the estimation problem on generic signals is asymptotically optimal. Unlike in the rather esoteric case of Theorem 4.3.1, generic signals do occur in practice (in fact, the set of non-generic signals form a subset of measure 0 in $\mathbb{R}^d$). As such, this fact has profound consequences on the MRA model. In particular, since the coefficients of the bispectrum $\hat{\phi}_i\hat{\phi}_j\hat{\phi}_{-i-j}$ has variance of order σ^6, the theorem implies that many practical estimation algorithms that employs the bispectrum to recover the signal, such as [6] and [22], are now asymptotically optimal. Bispectrum approaches using tensor decomposition have also been shown to achieve the optimal sample complexity of $O_d(\sigma^6)$ in the companion paper [19].

As another application of the moment matching technique, we will showcase the importance of high frequencies in the MRA model. An approach to the cryo-EM problem that has been proposed in [4] is to first estimate the low frequencies of the signal by blocking frequencies above a certain cutoff then use this initial estimate to estimate the higher frequencies. The underlying assumption of the strategy is that estimating a low-pass version of the signal is no harder than estimating the original signal. Surprisingly, this is not the case in general.

Example 4.3.4. Let $d \in \mathbb{Z}_{\geq 14}$ be an integer satisfying $d \equiv 2 \pmod 4$. Let $\mathcal{K}$ denote the set of vectors $\theta \in \mathbb{R}^d$ satisfying $\hat{\theta}_1 = \hat{\theta}_{-1} = 0$ and $\hat{\theta}_j \neq 0$ for all $j \notin \{1,-1\}$. We will prove below that θ can be recovered with $O_d(\sigma^6)$ samples. However, if we apply a low-pass filter to θ by setting $\hat{\theta}_j = 0$ for all $|j| > 4$, then the only nonzero entry of the bispectrum is $\hat{\theta}_2\hat{\theta}_2\hat{\theta}_{-4}$. In particular, the bispectrum carries no information about the phase of $\hat{\theta}_3$. This allows us to construct two signals τ and ϕ such that the first three moment tensors of their low-pass versions agree. By applying a similar argument as in Theorem 4.3.1, we conclude that any estimator must incur an error of at least 1/4 unless $n \gtrsim \sigma^8$.

To show that θ can be estimated by drawing $O_d(\sigma^6)$ samples from P_θ, it suffices to show that the orbit of θ can be uniquely recovered from its power spectrum and bispectrum.

Given a complex number z, let $\arg(z)$ denote its phase. We will show that the phases $\arg(\hat{\theta}_j)$ can be recovered from the bispectrum. In what follows, all computations are performed modulo 2π. Recall that a cyclic shift of R_ℓ to θ corresponds to a multiplication of $e^{-\frac{2\pi i \ell j}{d}}$ to each $\hat{\theta}_j$. Hence there exists a unique cyclic shift R_ℓ satisfying

$$\arg(\widehat{R_\ell \theta}_2) \in [0, 4\pi/d) \qquad \text{and} \qquad \arg(\widehat{R_\ell \theta}_3) \in [0, \pi).$$

By replacing θ with another representative in its $\mathcal{G}$-orbit if necessary, we may assume that $\arg(\hat{\theta}_2) \in [0, 4\pi/d)$ and $\arg(\hat{\theta}_3) \in [0, \pi)$. The following telescoping series evaluates to

$$\begin{aligned} 2\sum_{k=1}^{(d-6)/4} \arg\big(\hat{\theta}_2 \hat{\theta}_{2k} \hat{\theta}_{-2-2k}\big) &= 2\sum_{k=1}^{(d-6)/4} \arg(\hat{\theta}_2) + 2\sum_{k=1}^{(d-6)/4} \Big(\arg(\hat{\theta}_{2k}) - \arg(\hat{\theta}_{2+2k})\Big) \\ &= \frac{d-2}{2}\arg(\hat{\theta}_2) - 2\arg\big(\hat{\theta}_{\frac{d-2}{2}}\big) \\ &= \frac{d}{2}\arg(\hat{\theta}_2) - \arg\big(\hat{\theta}_{\frac{d-2}{2}} \hat{\theta}_{\frac{d-2}{2}} \hat{\theta}_2\big). \end{aligned}$$

Thus the quantity $\frac{d}{2}\arg(\hat{\theta}_2)$ can be recovered from the bispectrum modulo 2π. Together with the assumption that $\arg(\hat{\theta}_2) \in [0, 4\pi/d)$, we see that the choice of $\arg(\hat{\theta}_2)$ is unique. We can then inductively recover all even-indexed phases by considering the quantities $\arg(\hat{\theta}_2 \hat{\theta}_{2k} \hat{\theta}_{-2-2k})$ for $k \in \mathbb{Z}_{\geq 1}$. Next, the equality

$$\arg(\hat{\theta}_6) + \arg\big(\hat{\theta}_3 \hat{\theta}_3 \hat{\theta}_{-6}\big) = 2\arg(\hat{\theta}_3)$$

implies that the quantity $2\arg(\hat{\theta}_3)$ can also be recovered modulo 2π. The assumption $\arg(\hat{\theta}_3) \in [0, \pi)$ means that this choice is unique as well. We can then similarly proceed to recover all odd-indexed phases by considering the quantities $\arg(\hat{\theta}_2 \hat{\theta}_{1+2k} \hat{\theta}_{-3-2k})$ for $k \in \mathbb{Z}_{\geq 0}$.

4.4 The Phase Shift Model

While many deep mathematical insights can be drawn purely by studying the MRA model, it is ultimately still viewed as a toy model for the much more complicated cryo-EM problem. Therefore, another area of active research [7] is to look for other subgroups of the orthogonal group in which analysis is still tractable. In the previous section, we have seen that it is usually more convenient to work directly in the Fourier domain. This motivates the following natural generalisation of the MRA model to the circle group

$$U(1) := \big\{ z \in \mathbb{C} \ : \ |z| = 1 \big\}.$$

For the sake of convenience, the dimension d is assumed to be odd for the next two sections. Recall that for any $\ell \in [d]$ and $\theta \in \mathbb{R}^d$, the cyclic shift R_ℓ corresponds to a phase shift of the form

$$\big(\widehat{R_\ell \theta}\big)_j = e^{-\frac{2\pi i j \ell}{d}} \hat{\theta}_j.$$

Hence for any $z \in U(1)$, we specify a linear map $G_z : \mathbb{R}^d \to \mathbb{R}^d$ by working directly in the Fourier domain instead.

$$(\widehat{G_z\theta})_k := z^k \hat{\theta}_k, \qquad -\frac{d-1}{2} \leq k \leq \frac{d-1}{2}.$$

To identify $U(1)$ with a subgroup of the orthogonal group $O(d)$, we first observe that for each $z \in U(1)$, the map G_z is a $\mathbb{R}$-linear isometry from $\mathcal{C}$ to itself. Since the discrete Fourier transform DFT is also an isometry (by Plancherel's Theorem), the following composition

$$\begin{tikzcd} \mathbb{R}^d \arrow[r, "\mathrm{DFT}"] & \mathcal{C} \arrow[r, "G_z"] & \mathcal{C} \arrow[r, "\mathrm{DFT}^{-1}"] & \mathbb{R}^d \end{tikzcd}$$

gives a well-defined element of $O(d)$. The above setup is known as the ***phase shift model***. The MRA model can be viewed as a discretization of the phase shift model, in which the complex number z is restricted to the dth roots of unity $e^{\frac{2\pi ik}{d}}$ for $k \in \mathbb{Z}$.

An immediate implication of the Fourier-analytic definition of the group action of G_z on $\mathbb{R}^d$ is that the following analogue of Theorem 4.2.2 (which can be derived in an almost identical manner) holds for the phase shift model.

Theorem 4.4.1. For each m-tuple $(k_1, \cdots, k_m) \in [d]^m$, we have

$$\mathbb{E}\big[(\widehat{G_z\phi})^{\otimes m}\big]_{k_1 \cdots k_m} = \begin{cases} \hat{\phi}_{k_1}\hat{\phi}_{k_2}\cdots\hat{\phi}_{k_m} & \text{if } k_1 + k_2 + \cdots + k_m = 0 \\ 0 & \text{otherwise.} \end{cases}$$

This allows the technique of moment matching discussed in the previous section to be carried over to the new setting. Again by constructing two signals ϕ and τ with Fourier supports restricted to $\{\pm 1\}$ and leveraging on precise cancellations of their Fourier coefficients, the following analogue of Theorem 4.3.1 was established in [3] for the phase shift model.

Theorem 4.4.2. Let $0 \leq s \leq (d-1)/2$. In the phase shift model, there exists a universal constant M such that

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \mathcal{T}_s} \mathbb{E}_\theta\big[\rho(\tilde{\theta}_n, \theta)\big] \geq M \min\left\{\frac{\sigma^{\max\{2s-1, s+1\}}}{\sqrt{n}},\ 1\right\}$$

where the infimum is taken over all estimators $\tilde{\theta}_n$ on n samples from P_θ.

However, the condition $k_1 + k_2 + \cdots + k_m = 0$ is much more restrictive than the analogous condition in the MRA model, which only holds modulo d. This means that in general, the moment tensors in the phase shift model carry much less information about the underlying signal.

This reveals an interesting relationship between the complexity of the subgroup $\mathcal{G}$ in the algebraically structured model and the difficulty of the estimation problem. Intuitively, we expect the estimation problem on larger, more complex subgroups to be more difficult since the true signal θ can be corrupted by $\mathcal{G}$ in more complicated ways. However, the orbit of θ under the action of a larger complex subgroup tends to be larger as well, which potentially makes it easier to estimate. For example, the two signals ϕ and τ which we have constructed in the proof of Theorem 4.3.1 belongs to the same orbit under the phase shift model. The investigation of the fundamental trade-off between the complexity of the subgroup and the difficulty of the orbit retrieval problem is also an active subject of current research [1].

4.5 Constrained Maximum Likelihood Estimator

In the final section of this chapter, we will sketch the main idea of how the tools developed in Chapter 3 can be applied to analyse a constrained MLE $\check{\theta}_n$ for the phase shift model. A fully rigorous treatment involves a painstaking verification of a large number of technical details that contributes little conceptually.

The condition (16) means that Theorem 3.0.1 cannot be directly applied to the MRA model. Instead, we adopt a twofold strategy: given a sample $X_1, \cdots, X_{2n}$ of size $2n$ drawn from P_θ, we first split it into two subsamples $\mathcal{X}_1 = \{X_1, \cdots, X_n\}$ and $\mathcal{X}_2 = \{X_{n+1}, \cdots, X_{2n}\}$ of equal size. The first subsample $\mathcal{X}_1$ is used to estimate the Fourier support of θ. Define

$$M_j := \frac{1}{n}\sum_{i=1}^{n} |\widehat{(X_i)}_j|^2 - \sigma^2, \qquad 1 \leq j \leq \frac{d-1}{2}.$$

The set $\tilde{S}$ defined by

$$\tilde{S} := \left\{ 1 \leq j \leq \frac{d-1}{2} \ : \ M_j \geq \frac{1}{2}K_0^2 \right\}$$

will serve as our estimate for $\mathrm{psupp}(\hat{\theta})$. It can be shown, using similar techniques as outlined in Remark 3.2.7, that the probability of error decays exponentially with n.

Proposition 4.5.1. There exists a positive constant M, depending on K_0, such that

$$P\big(\tilde{S} \neq \mathrm{psupp}(\hat{\theta})\big) \leq 2d \exp\big(- Mn\sigma^{-4}\big).$$

Next, we use the second sample to construct the constrained MLE $\check{\theta}_n$. To that end, for any subset $S \subseteq [(d-1)/2]$, define the projection $P_S : \mathcal{C} \to \mathcal{C}$ by

$$\widehat{P_S(\phi)}_j = \begin{cases} \hat{\phi}_j & \text{if } j \in S \cup -S \\ \hat{\phi}_0 & \text{if } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The bulk of the work will be to show that condition (16) holds in this subspace.

Theorem 4.5.2. Fix $2 \leq s \leq (d-1)/2$ and $\theta \in \mathcal{T}_s$. Let $S = \mathrm{psupp}(\hat{\theta})$. There exists a constant $C_d > 0$ depending on d such that for all $\phi \in \mathrm{Im}(S)$, we have

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \geq C_d \sigma^{-4s+2} \rho(\theta, \phi)^2.$$

Finally, by invoking Theorem 3.0.1, and using probabilistic arguments (in the same spirit as Proposition 3.2.8) to handle the event in which $\tilde{S} \neq \mathrm{psupp}(\hat{\theta})$, we obtain the following result.

Theorem 4.5.3. For any $2 \leq s \leq (d-1)/2$, the modified MLE $\check{\theta}_n$ for the phase shift model satisfies

$$\sup_{\theta \in \mathcal{T}_s} \mathbb{E}_\theta[\rho(\check{\theta}_n, \theta)] \leq C' \frac{\sigma^{2s-1}}{\sqrt{n}} + C'' \sigma^{12s-11} \frac{\log n}{n}$$

for constants C' and C'' depending only on d and K_0.

5 Future Work

Since the original paper [3] in 2017, research in algebraically structured models have branched out in many different directions, bringing together tools from such diverse fields as algebraic geometry [1], [13], differential geometry [8], information theory [21], among others. In the next section, we will give a brief overview of one such tool from matrix analysis. Finally, we will conclude this report by describing the motivation behind the problem of sparse multi-reference alignment.

5.1 Löwner's Matrix Theory

An important class of matrices that arise naturally in many fields is the class of Hermitian matrices. Of particular interest are the positive definite and positive semidefinite Hermitian matrices, whose properties have already played an influential role (albeit indirectly) in Section 3.1 when we analysed the Hessian matrix. In this section, we will give a brief overview of Löwner's matrix theory, which have recently found important applications in information theory [16].

For any positive integer n, let $\mathcal{H}_n$ denote the set of $n \times n$ Hermitian matrices. In this section, all matrices are assumed to have entries in $\mathbb{C}$.

Definition 5.1.1. The ***Löwner Order*** on $\mathcal{H}_n$ is defined by

$$\begin{aligned} \boldsymbol{B} \geq \boldsymbol{A} &\text{ if and only if } \boldsymbol{B} - \boldsymbol{A} \text{ is positive semidefinite} \\ \boldsymbol{B} > \boldsymbol{A} &\text{ if and only if } \boldsymbol{B} - \boldsymbol{A} \text{ is positive definite.} \end{aligned}$$

Every real-valued function f on an interval I defines a corresponding matrix function in the following way. If $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix such that each eigenvalue λ_j is contained in I, we define $f(\boldsymbol{D}) = \operatorname{diag}(f(\lambda_1), \cdots, f(\lambda_n))$. This definition can be generalised to an arbitrary Hermitian matrix $\boldsymbol{A}$ by defining

$$f(\boldsymbol{A}) := \boldsymbol{U}\boldsymbol{f}(\boldsymbol{D})\boldsymbol{U}^*$$

where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ for $\boldsymbol{D}$ diagonal and $\boldsymbol{U}$ unitary.

Definition 5.1.2. A function $f : I \to \mathbb{R}$ is ***operator monotone*** if for all positive integers n and all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{H}_n$ whose eigenvalues are contained in I,

$$\boldsymbol{A} \leq \boldsymbol{B} \implies f(\boldsymbol{A}) \leq f(\boldsymbol{B}) \tag{34}$$

Definition 5.1.3. A function $f : I \to \mathbb{R}$ is ***operator convex*** if for all positive integers n, for all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{H}_n$ whose eigenvalues are contained in I and for all $\lambda \in [0, 1]$,

$$f\big((1-\lambda)\boldsymbol{A} + \lambda\boldsymbol{B}\big) \leq (1-\lambda)f(\boldsymbol{A}) + \lambda f(\boldsymbol{B}). \tag{35}$$

Although the definitions of operator monotonicity and operator convexity seem simple at first glace, the requirement that conditions (34) and (35) hold for all positive integers n turns out to be very restrictive. As a first illustration, operator monotone functions and operator convex functions are much more closely related as compared to their analogues in single-variable calculus. The following theorem is but one of the many manifestations of this phenomenon.

Theorem 5.1.4. Let f be a twice continuously differentiable operator convex function on an interval I satisfying $f(0) = 0$. Then the function $g(x) := f(x)/x$ is operator monotone.

Even more remarkably, operator monotone functions also have deep connections to the field of complex analysis. This correspondence is at the heart of Löwner's matrix theory.

Theorem 5.1.5 (Löwner). Let $f : (-1, 1) \to \mathbb{R}$ be an operator monotone function. Then there exist a probability measure μ on $[-1, 1]$ such that

$$f(t) = f(0) + f'(0) \int_{-1}^{1} \frac{t}{1 - \lambda t} \, d\mu(\lambda).$$

Results such as Theorem 5.1.4 imply that analogous statements also hold for operator convex functions.

Theorem 5.1.6 (Löwner). Let $f : (-1, 1) \to \mathbb{R}$ be an operator convex function. There exists a probability measure μ on $[-1, 1]$ such that

$$f(t) = f(0) + f'(0)t + \frac{1}{2} f''(0) \int_{-1}^{1} \frac{t^2}{1 - \lambda t} \, d\mu(\lambda).$$

In particular, the above integral representations show that operator monotone and operator convex functions admit an analytic continuation that is defined everywhere on the complex plane except possibly on $(-\infty, -1] \cup [1, \infty)$. For a detailed discussion on the other implications, see [9, Chapter 5].

Here is an application of Löwner's matrix theory to algebraically structured models. By bringing in tools from complex analysis, Epstein was able to give a relatively short proof of Lieb's theorem in [12].

Theorem 5.1.7 (Lieb). Let $\boldsymbol{H}$ be a fixed $n \times n$ Hermitian matrix. The map

$$\boldsymbol{A} \mapsto \operatorname{Tr}\Big(\exp\big(\boldsymbol{H} + \log \boldsymbol{A}\big)\Big)$$

is concave on the set of positive definite matrices.

Lieb's theorem in turn becomes a crucial ingredient in the proof of Theorem 4.6.1 in [23], which is used in Lemma B.1 of Appendix B.

5.2 Sparse Multi-reference Alignment

The sampling complexity of $O_d(\sigma^6)$ in Theorem 4.3.3, while a significant improvement from $O_d(\sigma^d)$ in Theorem 4.3.1, is nevertheless still considered to be too high for many practical applications. Yet, by Theorem 2.3.1, a sampling complexity of $O_d(\sigma^4)$ or better in the MRA model is only possible if the signal θ can be uniquely recovered from its power spectrum.

The task of recovering a signal from its power spectrum is a classical problem in signal processing and is called the phase retrieval problem. While it is well-known that unique recovery of the signal is not possible in general, there are many theoretical results that guarantees unique recovery under certain conditions. One of the most important success stories in this regard is in the field of compressed sensing, where the underlying signal is assumed to be sparse. This has motivated investigations into the sampling complexity of sparse signals in [14], where a sampling complexity of $O_d(\sigma^4)$ (and better) can indeed be established under certain sparsity conditions.

Appendices

A Bartlett's Identities

In this section, we will derive the Bartlett's identities that are used in the proof of Theorem 3.0.1.

Let $\mathcal{F} := \{f(x;\phi) : \phi \in \Theta\}$ be a family of densities on $\mathbb{R}^d$, indexed by an open subset Θ of $\mathbb{R}^n$. Then for every $\phi \in \Theta$, we have

$$\int_{\mathbb{R}^d} f(x;\phi)\ dx = 1.$$

Suppose that the family $\mathcal{F}$ satisfy the regularity conditions required to permit differentiation under the integral sign. Then for any $1 \leq i, j \leq n$, we have

$$\int_{\mathbb{R}^d} \frac{\partial}{\partial \phi_i} f(x;\phi)\ dx = 0 \qquad \text{and} \qquad \int_{\mathbb{R}^d} \frac{\partial^2}{\partial \phi_i \partial \phi_j} f(x;\phi)\ dx = 0. \tag{36}$$

The first equality in (36) implies that if X is a random variable with density $f(x;\theta)$, then we have

$$\mathbb{E}\left[\frac{\partial \log f}{\partial \phi_i}(X;\theta)\right] = \int_{\mathbb{R}^d} f(x;\theta) \frac{\frac{\partial f}{\partial \phi_i}(x;\theta)}{f(x;\theta)}\ dx = 0.$$

By a similar argument, the second equality in (36) implies that

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 \log f}{\partial \phi_i \partial \phi_j}(X;\theta)\right] &= \int_{\mathbb{R}^d} f(x;\theta) \frac{\frac{\partial^2 f}{\partial \phi_i \partial \phi_j}(x;\theta)}{f(x;\theta)}\ dx - \mathbb{E}\left[\frac{\frac{\partial f}{\partial \phi_i}(X;\theta) \cdot \frac{\partial f}{\partial \phi_j}(X;\theta)}{f(X;\theta)^2}\right] \\ &= -\mathbb{E}\left[\frac{\partial \log f}{\partial \phi_i}(X;\theta) \cdot \frac{\partial \log f}{\partial \phi_j}(X;\theta)\right].\end{aligned}$$

To summarise, we have derived the following two identities

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \log f}{\partial \phi_i}(X;\theta)\right] &= 0, \\ \mathbb{E}\left[\frac{\partial^2 \log f}{\partial \phi_i \partial \phi_j}(X;\theta)\right] &= -\mathbb{E}\left[\frac{\partial \log f}{\partial \phi_i}(X;\theta) \cdot \frac{\partial \log f}{\partial \phi_j}(X;\theta)\right].\end{aligned}$$

B Matrix Concentration Bound

Lemma B.1. Keep all notations introduced in Chapter 3. Suppose that $\|\theta\| \leq \sigma$. Let $\epsilon \in \mathbb{R}_{>0}$ and define the set $\mathcal{B}_\epsilon := \left\{\phi \in \mathbb{R}^d \;:\; \rho(\phi, \theta) \leq \epsilon\right\}$. Then there exists a constant C_d, depending on d, such that

$$\mathbb{E}\left[\sup_{\phi \in \mathcal{B}_\epsilon} \left\|H_D(\phi) - H_{D_n}(\phi)\right\|_{\mathrm{op}}^2\right] \leq C_d \frac{\log n}{n\sigma^4}.$$

Proof. Let C_d be a constant, depending on d, whose value may change from line to line. For each $1 \leq i \leq n$, let J_i denote the Hessian of the map g_{X_i} defined by

$$g_{X_i}(\phi) := \log \frac{f_\theta}{f_\phi}(X_i).$$

This allows us to write H_{D_n} as a sum of independent random matrices

$$H_{D_n}(\phi) = \frac{1}{n} \sum_{i=1}^n J_i(\phi).$$

By a similar symmetrization argument as in Lemma 3.2.6, we have that

$$\mathbb{E}\left[\sup_{\phi \in \mathcal{B}_\epsilon} \left\|H_D(\phi) - H_{D_n}(\phi)\right\|_{\mathrm{op}}^2\right] \leq \frac{4}{n^2} \mathbb{E}\left[\sup_{\phi \in \mathcal{B}_\epsilon} \left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\mathrm{op}}^2\right] \tag{37}$$

where $\epsilon_1, \cdots, \epsilon_n$ are i.i.d Rademacher random variables independent of the J_i's. By Lemma 3.1.4, for each i, we obtain

$$\begin{aligned}\|J_i(\phi) - J_i(\eta)\|_{\mathrm{op}} &\leq 6\sigma^{-6} \|X_i\|^3 \|\phi - \eta\| \leq C_d \frac{\|\theta\|^3 + \sigma^3 \|\xi_i\|^3}{\sigma^6} \|\phi - \eta\| \\ &\leq C_d \frac{1 + \|\xi_i\|^3}{\sigma^3} \|\phi - \eta\|.\end{aligned}$$

We apply the chaining method [25, Chapter 5]. Fix an arbitrary $\gamma \in (0, \epsilon)$ and let $\mathcal{Z}$ be a γ-net for $\mathcal{B}_\epsilon$. In other words, we require

$$\sup_{\phi \in \mathcal{B}_\epsilon} \min_{\eta \in \mathcal{Z}} \|\eta - \phi\| \leq \gamma.$$

By [26, Corollary 4.2.13], we can always choose the net $\mathcal{Z}$ such that $|\mathcal{Z}| \leq C_d (1/\gamma)^d$. A chaining argument gives

$$\sup_{\phi \in \mathcal{B}_\epsilon} \left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\mathrm{op}} \leq C_d \frac{\gamma}{\sigma^3} \sum_{i=1}^n \left(1 + \|\xi_i\|^3\right) + C_d \max_{\phi \in \mathcal{Z}} \left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\mathrm{op}}.$$

Young's inequality then implies that

$$\sup_{\phi\in\mathcal{B}_\epsilon}\left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\text{op}}^2 \leq C_d\frac{\gamma^2}{\sigma^6}\left(\sum_{i=1}^n 1+\|\xi_i\|^3\right)^2 + C_d\max_{\phi\in\mathcal{Z}}\left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\text{op}}^2. \tag{38}$$

The expectation of the first term is controlled using the fact that

$$\mathbb{E}\left[\left(\sum_{i=1}^n 1+\|\xi_i\|^3\right)^2\right] = \sum_{i,j=1}^n \mathbb{E}\big[(1+\|\xi_i\|^3)(1+\|\xi_j\|^3)\big] \leq C_d n^2. \tag{39}$$

For the second term, a matrix concentration bound [23, Theorem 4.6.1] gives

$$P\left(\max_{\phi\in\mathcal{Z}}\left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\text{op}}^2 \geq t \;\middle|\; X_1,\cdots,X_n\right) \leq 2d|\mathcal{Z}|\exp\left(-\frac{t}{2\max_{\phi\in\mathcal{Z}}\|\sum_{i=1}^n J_i(\phi)^2\|_{\text{op}}}\right)$$

for all $t\geq 0$. Integrating this tail bound, we get

$$\begin{aligned}\mathbb{E}\left[\max_{\phi\in\mathcal{Z}}\left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\text{op}}^2\right] &\leq C_d\log(d|\mathcal{Z}|)\mathbb{E}\left[\max_{\phi\in\mathcal{Z}}\left\|\sum_{i=1}^n J_i(\phi)^2\right\|_{\text{op}}\right]\\ &\leq C_d\log(|\mathcal{Z}|)n\mathbb{E}\left[\max_{\phi\in\mathcal{Z}}\left\|J_1(\phi)^2\right\|_{\text{op}}\right].\end{aligned}$$

By Lemma 3.1.4 and Young's inequality, we obtain an upper bound

$$\left\|J_1(\phi)^2\right\|_{\text{op}} = \left\|J_1(\phi)\right\|_{\text{op}}^2 \leq 2\sigma^{-4}+8\sigma^{-8}\|X_1\|^4 \leq C_d\frac{1+\|\xi_1\|^4}{\sigma^4}$$

that is uniform in ϕ. Therefore,

$$\mathbb{E}\left[\max_{\phi\in\mathcal{Z}}\left\|\sum_{i=1}^n \epsilon_i J_i(\phi)\right\|_{\text{op}}^2\right] \leq C_d\frac{\log(1/\gamma)}{\sigma^4}n. \tag{40}$$

Putting (38), (39) and (40) back into (37) and choosing $\gamma = n^{-1/2}$, we get

$$\mathbb{E}\left[\sup_{\phi\in\mathcal{B}_\epsilon}\left\|H(\phi)-H_n(\phi)\right\|_{\text{op}}^2\right] \leq C_d\left(\frac{\gamma^2}{\sigma^6}+\frac{\log(1/\gamma)}{n\sigma^4}\right) \leq C_d\frac{\log n}{n\sigma^4}$$

as desired. □

References

[1] Afonso S Bandeira, Ben Blum-Smith, Joe Kileel, Amelia Perry, Jonathan Weed, and Alexander S Wein. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

[2] Afonso S Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 459–470, 2014.

[3] Afonso S Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. Optimal rates of estimation for multi-reference alignment. *Mathematical Statistics and Learning*, 2(1):25–75, 2020.

[4] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak. Rapid solution of the cryo-em reconstruction problem by frequency marching. *SIAM Journal on Imaging Sciences*, 10(3):1170–1195, 2017.

[5] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE signal processing magazine*, 37(2):58–76, 2020.

[6] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on signal processing*, 66(4):1037–1050, 2017.

[7] Tamir Bendory, Dan Edidin, William Leeb, and Nir Sharon. Dihedral multi-reference alignment. *IEEE Transactions on Information Theory*, 68(5):3489–3499, 2022.

[8] Tamir Bendory, Ido Hadi, and Nir Sharon. Compactification of the rigid motions group in image processing. *SIAM Journal on Imaging Sciences*, 15(3):1041–1078, 2022.

[9] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[11] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

[12] H Epstein. Remarks on two theorems of e. lieb. *Communications in Mathematical Physics*, 31:317–325, 1973.

[13] Zhou Fan, Yi Sun, Tianhao Wang, and Yihong Wu. Likelihood landscape and maximum likelihood estimation for the discrete orbit recovery model. *Communications on Pure and Applied Mathematics*, 2020.

[14] Subhroshekhar Ghosh and Philippe Rigollet. Sparse multi-reference alignment: Phase retrieval, uniform uncertainty principles and the beltway problem. *Foundations of Computational Mathematics*, pages 1–48, 2022.

[15] Svante Janson et al. *Gaussian hilbert spaces*. Number 129. Cambridge university press, 1997.

[16] Eduard Jorswieck, Holger Boche, et al. Majorization and matrix-monotone functions in wireless communications. *Foundations and Trends® in Communications and Information Theory*, 3(6):553–701, 2007.

[17] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[18] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[19] Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019.

[20] Elad Romanov, Tamir Bendory, and Or Ordentlich. Multi-reference alignment in high dimensions: sample complexity and phase transition. *SIAM Journal on Mathematics of Data Science*, 3(2):494–523, 2021.

[21] Elad Romanov, Tamir Bendory, and Or Ordentlich. Multi-reference alignment in high dimensions: sample complexity and phase transition. *SIAM Journal on Mathematics of Data Science*, 3(2):494–523, 2021.

[22] Brian M Sadler and Georgios B Giannakis. Shift-and rotation-invariant object reconstruction using the bispectrum. *JOSA A*, 9(1):57–69, 1992.

[23] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[24] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

[25] Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[26] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.