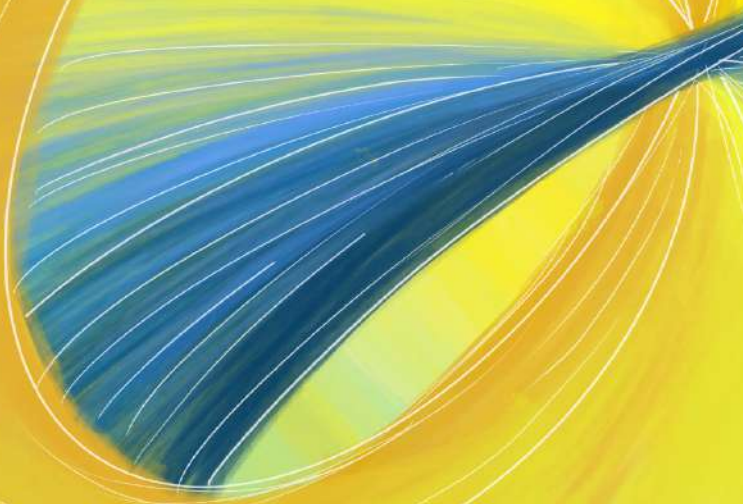
Scaling problems, algorithms and applications to Computer Science and Statistics



Rafael Oliveira
Akshay Ramachandran

impa

33º Colóquio Brasileiro de Matemática

Scaling problems, algorithms and applications to Computer Science and Statistics

Scaling problems, algorithms and applications to Computer Science and Statistics
Primeira impressão, julho de 2021

Publicado no Brasil / Published in Brazil.

ISBN 978-65-89124-18-4
MSC (2020) Primary: 14L24, Secondary: 62D99, 47N10, 90C26

Coordenação Geral Carolina Araujo

Produção Books in Bytes **Capa** Izabella Freitas & Jack Salvador

Realização da Editora do IMPA
IMPA
Estrada Dona Castorina, 110
Jardim Botânico
22460-320 Rio de Janeiro RJ

www.impa.br
editora@impa.br

Contents

1 Introduction

Several problems in mathematics, computer science, machine learning and statistics exhibit inherent symmetries which can be described by a group acting linearly on a vector space. Oftentimes, these symmetries are implicit or disguised in the "natural" description of the problems. Thus, many qualitative and quantitative properties inherent to these problems have laid dormant or unexplored until recent developments, which made crucial use of the group action structure, allowed for significant progress in such problems.
In this survey, we will give an overview of the phenomenon described above. Our emphasis will be on the geometric properties of such group actions and on the generalization of convexity that arises from natural optimization problems along group orbits, which we term by *geodesic convexity*.

1.1 Brief history

In the early days of invariant theory, the period known as classical invariant theory (late 1800s), the question of understanding geometric properties of plane curves which were invariant under changes of bases received a lot of attention. Notable mathematicians who worked on this question at the time include Aronhold, Clebsch, Gordan, Cayley, Sylvester and Hilbert. During this time, their focus was on

finding functions which associate a number to each curve that was independent of the choice of basis.

Mathematicians at the time came to realize that such a problem (invariance under change of basis) was about the action of a group on the ambient vector space, usually the action of the special linear group $\mathrm{SL}_n(\mathbb{C})$, and that the functions that they were studying were polynomial functions over the coefficients of the polynomials defining the curves being studied.

A simple example of the problem above, which is familiar to us all (but most likely not in this language), is the problem of deciding when a quadratic form in two variables, given by $ax^2+bxy+cy^2 \in \mathbb{C}[x,y]$, has a double root. As it turns out, the property of "having a double root" is independent of the choice of basis (that is, if we change basis $(x',y') = (x,y)A$, the quadratic will still have a double root) and it is characterized by the vanishing of the *discriminant* $\Delta := b^2-4ac$. Thus, the property of having a double root is completely captured by a polynomial function on the coefficients of the quadratic form (i.e. a,b,c).

The major research effort at the time was to determine the set of all polynomial invariants of "nice" group actions on certain vector spaces. Since the set of all invariant polynomials forms a $\mathbb{C}$-algebra, one of the main questions at the time, which was termed the first fundamental theorem of invariant theory, was to prove whether a group action had a finite set of generating invariants as a $\mathbb{C}$-algebra.

This research effort culminated in Hilbert's seminal works Hilbert (1890, 1893), where he proved such fundamental theorems as the Hilbert Basis Theorem, the Nullstellensatz, the Syzygy theorem, and the rationality of the Hilbert series. Hilbert's motivation to prove these theorems was to give a constructive proof that the ring of invariants was finitely generated, and to give a full description of the ring of invariants.

While the algebraic side of invariant theory has received much attention since the nineteenth century, it was only in the seminal works of Mumford and the striking developments by Kempf, Ness, Kostant and Kirwan, among others, that the geometric side of invariant theory really flourished. In geometric invariant theory,[1] given a group G acting on a vector space V, the goal is to understand the quotient space V/G given by the set of orbits of the group action on V.

In the development of geometric invariant theory by Mumford, a special optimization problem is central: the *null-cone problem*, which was already defined in the work of Hilbert (1893). We will study this problem in greater detail in Chap-

[1] The setting of geometric invariant theory is more general, and we have decided to remain with the setting of a group acting on a vector space for simplicity. For the more general treatment we refer the reader to Mumford, Fogarty, and Kirwan (1994) and Wallach (2017).

ter 2, but now through the lens of optimization over a Riemannian manifold.

1.2 Examples of scaling problems

In this section we describe some concrete examples of scaling problems which have seen important progress in recent years by the use of the optimization approach to geometric invariant theory. The beauty of these concrete examples, apart from being fundamental problems in their respective subareas of mathematics, is that we can state them even without the definitions from invariant theory, and we will do so in order to motivate the reader and to showcase how the inherent symmetries of a problem may be disguised in its statement.

1.2.1 Matrix Scaling

Given a non-negative $n \times n$ matrix $A \in \mathrm{Mat}_n(\mathbb{R})$, we say that A is *doubly-stochastic* if all row and column sums of A are equal to 1. An important problem, which appears in several disciplines ranging from economics, engineering, transportation theory and computer science, is the question of deciding when one can "transform" a non-negative matrix A (approximately) into a doubly-stochastic matrix B by multiplying the rows and columns of A by positive scalars. This problem motivates the following definition:

Definition 1 (Scaling of a matrix)**.** Given a non-negative matrix $A \in \mathrm{Mat}_n(\mathbb{R})$, we say that $\hat{A}$ is a *scaling* of A if it can be obtained by multiplying the rows and columns of A by positive scalars. In other words, $\hat{A}$ is a scaling of A if there exist positive diagonal matrices $R, C \in \mathrm{Mat}_n(\mathbb{R})$ such that $\hat{A} = RAC$.

As the reader can realize, the approximate version of the question is often needed, since the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ can be scaled arbitrarily close to a doubly-stochastic matrix (i.e. the identity) but it cannot be scaled exactly to a doubly-stochastic matrix (since the non-zero pattern of the matrix does not change by scaling). This motivates us to define a measure for how close a matrix is to being doubly-stochastic:

Definition 2 (Distance to doubly-stochastic)**.** Given a non-negative matrix $A \in \mathrm{Mat}_n(\mathbb{R})$, define its distance to doubly stochastic to be

$$\mathrm{ds}(A) = \sum_{i=1}^{n} (r_i - 1)^2 + \sum_{j=1}^{n} (c_j - 1)^2$$

where r_i and c_j denote the i^{th} row sum and the j^{th} column sum, respectively.

With the definition of distance as above, we can say that a non-negative matrix A is *approximately scalable* to doubly-stochastic if, and only if, for every $\varepsilon > 0$, there exists a scaling A_ε of A such that $\mathrm{ds}(A_\varepsilon) \leqslant \varepsilon$. We call such a scaling A_ε an ε-scaling of A.

Thus, given a non-negative matrix A, two natural questions arise: when is a matrix approximately scalable? If a matrix is scalable, can one efficiently find an ε-scaling, for a given parameter $\varepsilon > 0$?

We have arrived at the (computational version of the) matrix scaling problem:

Question 1.2.1 (Matrix Scaling)**.** Given a non-negative matrix $A \in \mathrm{Mat}_n(\mathbb{R})$ and an accuracy parameter $\varepsilon > 0$, is there a scaling B of A such that $\mathrm{ds}(B) \leqslant \varepsilon$? If there is such a scaling, output it.

As mentioned in the beginning of this section, the matrix scaling problem has historically appeared independently in several scientific areas, and to solve the matrix scaling problem the following natural iterative algorithm has often been used: if the matrix is not *row-stochastic* (that is, the row sums are 1), make it row-stochastic by properly normalizing the rows. This may change the column sums. If the matrix is not *column-stochastic*, make it column-stochastic by normalizing the columns.

Input: a non-negative matrix $A \in \mathrm{Mat}_n(\mathbb{R})$, $\varepsilon > 0$.
Output: a scaling B of A such that $\mathrm{ds}(B) \leqslant \varepsilon$, if one exists. NO, otherwise.

- Set $B \leftarrow A$
- For T steps, while $\mathrm{ds}(B) > \varepsilon$:
  1. if B is not row-stochastic, multiply i^{th} row of B by $r_i(B)^{-1}$ for all $i \in [n]$
  2. if B is not column-stochastic, multiply j^{th} column by $c_j(B)^{-1}$ for all $j \in [n]$
- If at any point above $\mathrm{ds}(B) \leqslant \varepsilon$, return B, otherwise, after the T steps, return NO.

Algorithm 1: RAS algorithm

The algorithm above is a special case of a general optimization paradigm

known as *alternating minimization*, where to minimize a function one tries to alternately minimize simpler functions in an alternate fashion, where the idea is that the simpler functions are much easier to optimize (sometimes the optimum for the simpler functions can even be written in closed form, as is our case).

In Section 1.3, we will see an analysis of the algorithm shown above, as well as a striking application of using matrix scaling to obtain a deterministic approximation to the permanent of non-negative matrices, and the connection between matrix scaling and bipartite matchings.

For more background on the matrix scaling problem, we refer the reader to the surveys Garg and Oliveira (2018) and Idel (2016).

1.2.2 Optimal Transport Distances in Finite Distributions

Given two discrete probability measures $r, c \in \mathbb{R}_+^d$ over a finite set $[d] := \{1, 2, \ldots, d\}$, we define $U(r, c)$ to be the *transportation polytope* of r and c, which is given by

$$U(r, c) := \{P \in \mathrm{Mat}_d(\mathbb{R}_+) \mid P1_d = r,\ P^\dagger 1_d = c\}$$

where 1_d is the all ones vector of dimension d. An element of $U(r, c)$ is called a *transportation matrix* or joint distribution, as we will now see.

One can view $U(r, c)$ as the set of all *joint probability distributions* of two discrete random variables X, Y each taking values in $[d] := \{1, 2, \ldots, d\}$ where X has probability distribution r and Y has probability distribution c. In this case, each matrix $P \in U(r, c)$ is such that $P_{i,j} = \Pr[X = i, Y = j]$.

Given a cost matrix $M \in \mathrm{Mat}_d(\mathbb{R})$, the cost of mapping measure r to c using a transportation matrix P can be quantified by the Frobenius inner product $\langle M, P \rangle := \mathrm{Tr}[M^\dagger P]$. Thus, we have arrived at the *optimal transport* problem between r and c given cost M:

$$d_M(r, c) := \min_{P \in U(r,c)} \langle M, P \rangle.$$

Optimal transport of measures is a problem of great practical importance, having originated in the works of Monge (in 1871) and developed further by Kantorovich[2] (in 1942) in their studies on optimal allocation and transportation of

[2]Interestingly, Kantorovich is regarded as the father of Linear Programming.

resources. While the formulation above can be solved via standard convex optimization methods, or more specialized methods for linear programs, the complexity of solving the optimal transport problem above turns out to be $O(d^3 \log d)$ in practice, which turns out to be prohibitive for many applications.

In Cuturi (2013), the author proposed to add entropic constraints on the optimal transport problem to find optimal joint distributions which have *small mutual information*, as these solutions have applications to machine learning. Thus, Cuturi proposed to find solutions in the convex set

$$U_\alpha(r,c) := \{P \in U(r,c) \mid d_{KL}(P \mid\mid rc^\dagger) \leqslant \alpha\}$$

where $\alpha \geqslant 0$. Moreover, in the same work, Cuturi showed how one can use the matrix scaling algorithm from the previous section to solving the modified optimal transport problem! This yields a much simpler algorithm with a much better runtime in practice for computing such distances, and as showed in Cuturi (ibid.), these new distances have much better practical applications than the unconstrained original distances.

For more background on optimal transport and its connections to matrix scaling and machine learning, we refer the reader to Cuturi (ibid.), where the connection presented above was first made, and where we drew this example from. For connections to image retrieval, see the seminal work of Rubner, Tomasi, and Guibas (2000). For a comprehensive treatment of optimal transport, see Villani (2008).

1.2.3 Paulsen Problem

The Paulsen problem is a central question in frame theory as discussed in Casazza and Kutyniok (2013).

Question 1.2.2. Let $U = \{u_1, ..., u_n\} \subseteq \mathbb{C}^d$ be a spanning set of vectors satisfying

$$\frac{1-\varepsilon}{d} I_d \preceq \sum_{j=1}^{n} u_j u_j^* \preceq \frac{1+\varepsilon}{d} I_d, \qquad \forall j \in [n] : \frac{1-\varepsilon}{n} \leqslant \|u_j\|_2^2 \leqslant \frac{1+\varepsilon}{n}. \tag{1.2.1}$$

What is the minimum distance $\sum_{j=1}^n \|v_j - u_j\|_2^2$ over all $V = \{v_1, ..., v_n\}$ satisfying Equation (1.2.1) exactly:

$$\sum_j v_j v_j^* = \frac{1}{d} I_d, \qquad \forall j \in [n] : \|v_j\|_2^2 = \frac{1}{n}.$$

Note that this is a different normalization, by a factor d, than normally given in the literature.

Vectors satisfying Equation (1.2.1) are known as ε-doubly balanced frames. The balance properties of doubly balanced frames, where $\varepsilon = 0$, are exploited to give strong results in coding theory and signal processing Casazza and Kutyniok (ibid.). Constructions of exactly doubly balanced frames are difficult and often rely on complicated algebraic structures. On the other hand, there are many simple algorithms to construct ε-doubly balanced frames. For example, a large enough set of random vectors will satisfy Equation (1.2.1) for some small ε with high probability. The Paulsen problem asks, for a given ε-doubly balanced frame, whether the conditions in Equation (1.2.1) can be corrected without moving too much. Since randomly generated frames are nearly doubly balanced, analyzing the distance bound in this case is of special importance.

Holmes and Paulsen (2004) studied frames from the perspective of coding theory, and showed that doubly balanced frames were optimally robust with respect to a single erasure. They also showed that Grassmannian frames, doubly balanced frames with large pairwise angles, were optimal for two erasures.

To address the difficulty of constructing these structured frames, the authors of Holmes and Paulsen (ibid.) suggested a simple numerical approach: first generate random frames, which approximately satisfy Equation (1.2.1), and then correct the conditions. Random frames are good candidates for both of these settings because they are approximately doubly balanced and have large pairwise angles with high probability. One goal of the Paulsen problem is then to validate this numerical algorithm as a simple method of constructing structured frames. The formalization below is from Cahill and Casazza (2013).

Conjecture 1.2.3 (Paulsen Problem)**.** Let $p(d, n, \varepsilon)$ be the smallest function such that for all ε-doubly balanced $U = \{u_1, ..., u_n\} \subseteq \mathbb{C}^d$, there exists a doubly balanced $V = \{v_1, ..., v_n\} \subseteq \mathbb{C}^d$ such that

$$\|V - U\|_F^2 = \sum_{j=1}^{n} \|v_j - u_j\|_2^2 \leqslant p(d, n, \varepsilon).$$

Then this distance function p can be taken independent of n.

The optimal function p has been unknown for almost twenty years, despite considerable attention in the frame theory literature. Prior to the work of Kwok, Lau, Lee, et al. (2017), the only known results on the function p were given

by Casazza, Fickus, and Mixon (2012) and Bodmann and Casazza (2010), and showed $p \leqslant poly(d, n, \varepsilon)$ when d, n are relatively prime and ε is small enough.

These results left open Conjecture 1.2.3, which was positively resolved in Kwok, Lau, Lee, et al. (2017).

Theorem 1.2.4 (Theorem 1.3.1 in Kwok, Lau, Lee, et al. (ibid.))**.** *The distance function can be bounded by* $p(d, n, \varepsilon) \lesssim d^{11/2}\varepsilon$. *In particular it can be taken independent of* n.

The new idea in this work was to use scaling algorithms like those studied recently in Garg, Gurvits, et al. (2016). To carry out this approach, Kwok, Lau, Lee, et al. (2017) defined a dynamical system which corrected approximately doubly balanced frames. This dynamical system could then be analyzed using tools from the operator scaling analysis of Garg, Gurvits, et al. (2016). The full proof of Kwok, Lau, Lee, et al. (2017) required a smoothed analysis approach coupled with an involved convergence analysis of the dynamical system.

Subsequently, in the aptly titled "Paulsen Problem made Simple", Hamilton and Moitra (2019) improved the distance bound to $p(d, n, \varepsilon) \lesssim d\varepsilon$, using a totally different and much shorter method. This almost matches the known lower bound, as there are simple examples showing $p \gtrsim \varepsilon$. Ramachandran (2021) revisits the dynamical system approach and closes this gap by using tools from geodesic convex optimization.

This dynamical system can also be analyzed to give a refined distance bound for the case of random frames, which answers the original motivation of the Paulsen problem.

Theorem 1.2.5 (Theorem 1.12 in Kwok, Lau, and Ramachandran (2019))**.** *For any* $n \geqslant poly(d)$ *large enough, if* $U = \{u_1, ..., u_n\} \subseteq \mathbb{R}^d$ *is generated such that each* u_j *is independent and uniformly distributed on* $\frac{1}{\sqrt{n}} S^{d-1}$*, then with high probability* U *is* ε*-doubly balanced for* $\varepsilon \leqslant \widetilde{O}(\sqrt{\frac{d}{n}})$*, and there exists doubly balanced* V *such that*

$$\|V - U\|_F^2 \lesssim \varepsilon^2.$$

This result validates the numerical approach suggested in Holmes and Paulsen (2004) to generate doubly balanced frames, and therefore gives a satisfactory answer to the original motivation for Question 1.2.2. It also gives the following corollary on Grassmannian frames.

Corollary 1.2.6. *With the same conditions as Theorem 1.2.5, U has large pairwise angles with high probability:*

$$\max_{j \neq j' \in [n]} \langle u_j, u_{j'} \rangle^2 \leqslant \frac{\widetilde{O}(1)}{dn^2}.$$

This result further validates the numerical algorithm in Holmes and Paulsen (ibid.) as a simple way to generate nearly optimal Grassmannian frames.

1.2.4 Operator Scaling

The operator scaling was first proposed and studied by Gurvits (2004) as a quantum generalization of the matrix scaling problem. In this setting, the objects of study are *completely positive operators*, which can be defined by a tuple of matrices[3] $A = (A_1, \ldots, A_m) \in \mathrm{Mat}_{\mathbb{C}}(d, n)^m$ in the following way: $T_A : \mathrm{Mat}_{\mathbb{C}}(n) \to \mathrm{Mat}_{\mathbb{C}}(d)$ is the map

$$T_A(X) := \sum_{j=1}^{m} A_j X A_j^{\dagger}.$$

An important property of such operators is that they define a map from the set of positive semidefinite matrices to themselves, and positive semidefinite matrices encode mixed quantum states.

In the case where $n = d$, that is, we have square matrices, Gurvits also defined a notion of doubly-stochasticity for such operators, which as the reader may notice generalizes the definitions in the matrix scaling setting and also captures the definition of balanced frames in the Paulsen problem!

Definition 3 (Doubly Stochastic Operators)**.** We say that a completely positive operator defined by $(A_1, \ldots, A_m) \in \mathrm{Mat}_{\mathbb{C}}(n)^m$ is *doubly stochastic* if the following conditions hold:

$$\sum_{j=1}^{m} A_j A_j^{\dagger} = \sum_{j=1}^{m} A_j^{\dagger} A_j = I_n.$$

Gurvits' generalization of a scaling for a completely positive operator is by simultaneous pre and post multiplication by invertible matrices.

Definition 4 (Scaling for Operators)**.** Given a tuple $A = (A_1, \ldots, A_m) \in \mathrm{Mat}_{\mathbb{C}}(n, d)$, we say that tuple $B = (B_1, \ldots, B_m)$ is a scaling of the tuple A if there exist $L \in \mathrm{GL}_n(\mathbb{C})$, $R \in \mathrm{GL}_d(\mathbb{C})$ such that $B_j = L A_j R^T$.

[3] these tuple of matrices are known as Kraus operators of the completely positive map.

As we saw in the case of the matrix scaling problem, the approximate scaling problem is the more interesting one[4] and the situation is no different here. With this in mind, analogously to the distance to doubly-stochasticity in the matrix setting, Gurvits defined the distance to doubly-stochasticity as follows:

Definition 5. Given $A = (A_1, \ldots, A_m) \in \mathrm{Mat}_{\mathbb{C}}(n)^m$, define its distance to doubly-stochasticity as

$$\mathrm{ds}(A) := \left\| \sum_{j=1}^{m} A_j A_j^\dagger - I_n \right\|_F^2 + \left\| \sum_{j=1}^{m} A_j^\dagger A_j - I_n \right\|_F^2$$

where $\|\cdot\|_F$ is the Frobenius norm.

Now that we have all the definitions we need, we say that an operator $A = (A_1, \ldots, A_m)$ is *scalable* if for all $\varepsilon > 0$, there is an ε-scaling B of A, that is, B is such that $\mathrm{ds}(B) \leqslant \varepsilon$. And with these definitions we have the same algorithmic problems as in matrix scaling: given a matrix A, can we decide if it is scalable? If so, given a parameter $\varepsilon > 0$, can we find an ε-scaling B of A?

Gurvits also generalized the RAS algorithm to the operator scaling setting, and generalized structural results characterizing when a given operator is scalable. In addition, Gurvits was able to analyze the running time of the alternating minimization algorithm for operator scaling for a special class of operators. However, he left open the question of proving whether the alternating minimization algorithm runs in polynomial time, a task that was achieved only recently with the use of techniques from invariant theory by Garg, Gurvits, et al. (2016). The latter work marked a new beginning in the combination of techniques from invariant theory and optimization to solve problems from diverse areas of mathematics.

A further generalization to the operator scaling problem is the tensor scaling problem, which has further connections to physics, as tensor scaling is the algorithmic problem of *entanglement distillation* by SLOCC operations, which are natural actions on quantum systems when each of the subsystems is owned by a different party.

In Chapter 3 and Chapter 4 we will properly define and study the tensor scaling problem and see a striking application of the tensor scaling problem in statistics. For more on the connections between tensor scaling and physics, we refer the reader to the works Bürgisser, Franks, et al. (2018) and Bürgisser, Garg, et al. (n.d.) and references therein.

[4] This has several reasons, some of which will only become clear in the next chapter when we discuss the geometry of scaling problems.

1.2.5 Maximum Likelihood Estimation

A basic problem in statistics is the problem of fitting a set of data to a statistical model (i.e. a parametrized family of probability distributions). In general one tries to recover from the data the best probability distribution which fits the given data. To achieve this task, one wants to set the parameters of the probability distribution in order to maximize the likelihood of observing the input data that was given. A distribution with such property is called a *maximum likelihood estimate*, or MLE for short.

The usual way to compute an MLE for a given statistical model is to setup an optimization problem given the input data and use standard optimization techniques to find a local maximum, and hope that this maximum is a global maximum. Thus, a fundamental task is to understand the properties of such statistical models and optimization problems to provide provably efficient and optimum algorithms which compute the MLEs, or to prove that such task is computationally hard.

Another fundamental task in statistics is to understand the number of input samples that one needs (i.e. the size of the input data set) in order for an MLE to actually exist. Such a problem is known as the sample complexity problem for a statistical model, and in a very concrete sense it is a prerequisite for the MLE problem to be well-defined! The sample complexity problem also has its computational variant, as in addition to existence of an MLE, we are also interested in knowing whether the MLE actually approximates the true distribution where the data was sampled from, as well as whether there exists an efficient algorithm to compute such an MLE.

As was recently discovered by Améndola et al. (2020), scaling problems naturally appear in the maximum likelihood estimation problem of two fundamental settings in statistics: the log-linear models and Gaussian transformation families. The latter family of models in particular include the two main statistical models that we will be studying in Chapter 4: the Matrix Normal Model and the Tensor Normal Model. As was recently proved by Franks et al. (2021), a deep geometric understanding of the latter models coupled with tools and ideas from invariant theory and from techniques (old and new) from optimization yield to nearly optimal sample complexity bounds for the MLE problem for such models!

1.3 Approximation of the Permanent

An important application of matrix scaling in computer science is given in Linial, Samorodnitsky, and Wigderson (2000), where the goal is to obtain a deterministic

approximation to the permanent of non-negative matrices.

Definition 6. For matrix $A \in \mathbb{R}^{n\times n}$ the permanent is

$$\mathrm{Per}(A) := \sum_{\sigma\in S_n}\prod_{i=1}^{n} A_{i,\sigma(i)}.$$

This quantity has important implications in algebraic complexity and enumerative combinatorics, and it is known to be #P-hard to compute even the sign Aaronson (2011). When the matrix is non-negative, there is a PTAS given in Jerrum and Sinclair (1989).

The permanent has a combinatorial interpretation: if $A \in \{0,1\}^{n\times n}$, then A is the adjacency matrix of a bipartite graph on $n+n$ vertices, and the permanent of A counts the number of perfect matchings in this graph.

The goal of Linial, Samorodnitsky, and Wigderson (2000) was to give a strongly polynomial deterministic algorithm to multiplicatively approximate the permanent of a non-negative matrix. Of course, if the permanent is 0, any multiplicative approximation must also output 0. But $\mathrm{Per}(A) = 0$ for $A \in \mathbb{R}_+^{n\times n}$ iff the weighted bipartite graph associated with A has no perfect matchings, and this can be ascertained easily in polynomial time. The following is their main theorem.

Theorem 1.3.1 (Theorem 1.1 in Linial, Samorodnitsky, and Wigderson (ibid.))**.** *For non-negative matrix $A \in \mathbb{R}_+^{n\times n}$, there is a quantity $f(A)$ computable in strongly polynomial time such that*

$$\mathrm{Per}(A) \leqslant f(A) \leqslant e^{n+1}\,\mathrm{Per}(A).$$

When the matrix is doubly stochastic, the following powerful theorem gives the required approximation.

Theorem 1.3.2 (Van der Waerden)**.** *If $A \in \mathbb{R}_+^{n\times n}$ is doubly stochastic, i.e. $A1_n = 1_n$, $A^T 1_n = 1_n$, then*

$$e^{-n} \leqslant \mathrm{Per}(A) \leqslant 1.$$

The upper bound holds more generally for non-negative matrices:

$$\mathrm{Per}(A) \leqslant \prod_{i=1}^{n}\left(\sum_{j=1}^{n} A_{ij}\right),$$

as this quantity only has more terms than Definition 6. The difficult part of Theorem 1.3.2 is the lower bound, for which we now have several proofs (e.g. Gurvits (2006)).

So the approach to Theorem 1.3.1 is by transforming our input A to some doubly stochastic B while maintaining control of the change in permanent. We can use the following simple observation to execute this plan.

Fact 1.3.3. *For $A \in \mathbb{R}^{n\times n}$ and scalings $L, R \in \mathrm{diag}(n)$*

$$\mathrm{Per}(LAR) = \det(L)\,\mathrm{Per}(A)\,\det(R).$$

Proof. Each term in Definition 6 contains exactly one entry from each row and column, so this scaling will contribute $\prod_{i=1}^{n} L_{ii} \prod_{j=1}^{n} R_{jj}$ to every term. □

So we have reduced the problem of approximating the permanent (up to simply exponential factor) to finding a doubly stochastic row-column scaling. To do this in strongly polynomial time, Linial, Samorodnitsky, and Wigderson (2000) use the classical Sinkhorn scaling algorithm. This simple iterative algorithm can detect when the permanent is 0, and otherwise produces a nearly doubly-stochastic scaling in polynomial time. We first need a few definitions.

Definition 7. For matrix $A \in \mathbb{R}_{+}^{n\times n}$, we define the row and column sums to be $\boldsymbol{r}(A) := A\mathbf{1}_n, \boldsymbol{c}(A) := A^T\mathbf{1}_n$. We also define the error from doubly stochastic

$$\Delta(A) := \sum_{i=1}^{n}(r_i(A)-1)^2 + \sum_{j=1}^{n}(c_j(A)-1)^2.$$

Note that the following simple transformations produce row/column stochastic (but not necessarily doubly stochastic) matrices respectively:

$$A \leftarrow \mathrm{diag}(r(A))^{-1}A, \qquad A \leftarrow A\,\mathrm{diag}(c(A))^{-1}. \tag{1.3.1}$$

The Sinkhorn algorithm for matrix scaling alternates these steps until some termination condition. To decide on a termination condition, we need the following fact on permanent 0 matrices.

Lemma 1.3.4 (Lemma 5.2 in Linial, Samorodnitsky, and Wigderson (ibid.))**.** *For non-negative stochastic matrix $A \in R_{+}^{n\times n}$ with permanent 0, the error is lower bounded by*

$$\Delta(A) \geqslant \frac{1}{n}.$$

The above lemma in the contrapositive tells us that in order to distinguish between the zero and non-zero cases, it is enough to scale our input to be $\frac{1}{n}$-close to doubly stochastic.

On the other hand, we can show that the Sinkhorn algorithm makes progress while our input is far from doubly stochastic.

Lemma 1.3.5. *For non-negative stochastic matrix $A \in \mathbb{R}_+^{n\times n}$ with non-zero permanent, a single iteration of Sinkhorn algorithm produces A' with*

$$\log \mathrm{Per}(A') - \log \mathrm{Per}(A) \geqslant \frac{1}{6} \min\{\Delta(A), 1\}.$$

The proof uses the following robust version of AMGM, which can be shown by Taylor approximation.

Lemma 1.3.6 (Lemma 3.10 in Linial, Samorodnitsky, and Wigderson (2000))**.** *For $x \in \mathbb{R}_{++}^n$ with $\sum_{i=1}^n x_i = n$, if $\Delta := \sum_{i=1}^n (x_i - 1)^2$, then*

$$\log \prod_{i=1}^n x_i \leqslant \frac{\Delta}{2} + O(\Delta^{3/2}).$$

To approximate the permanent for a general input, we need a simple preprocessing step which guarantees that the input is stochastic and if its permanent is non-zero, then $\log \mathrm{Per}(A) \geqslant -n \log n$. From here, the proof of Theorem 1.3.1 then follows simply by iterating Sinkhorn's algorithm for $poly(n)$ steps. If $\Delta(A_t) < \frac{1}{n \log n}$ for any of these steps, then an approximate version of Theorem 1.3.2 shows that

$$1 \geqslant \mathrm{Per}(A_t) \geqslant e^{-(n+1)}.$$

Otherwise, we must be in the case where $\mathrm{Per}(A) = 0$, as otherwise

$$0 \geqslant \log \mathrm{Per}(A_T) \geqslant \log \mathrm{Per}(A) + \sum_{t=1}^T \frac{1}{6} \min\{\Delta(A_t), 1\} \geqslant -n \log n + \frac{T}{12n} > 0,$$

where the first step is due to the simple approximate for stochastic matrices, and the second is by Lemma 1.3.5. This is a contradiction for T chosen large enough.

1.4 References

In this brief introduction we just outlined some of the scaling problems and their striking applications. For a more complete reference on the history of scaling problems, its independent discoveries by several disparate communities of researchers in different research fields, and the recent unification of such problems in the perspective of invariant theory, we refer the reader to the surveys Garg and Oliveira (2018) and Idel (2016), to the recent work Bürgisser, Franks, et al. (2019) and references therein.

Scaling problems are only part of the optimization and geometric side of invariant theory, and an understanding of the algebraic side of invariant theory has also greatly contributed to progress in the solution to scaling problems. Recently, with the works of Améndola et al. (2020), Derksen and Makam (2020), and Derksen, Makam, and Walter (2020), the algebraic side of invariant theory has shown to be quite effective also in solving some of the structural problems in statistics. And as it is usual for the discipline of invariant theory, the algebraic, geometric and optimization perspectives often inform one another, and the connections generated from these perspectives greatly enhance our understanding of the field as well as their intended applications. In this survey, we will unfortunately not have enough space to an introduction to the algebraic developments in the past decades which have connections to our topics being discussed.

For developments on the algebraic front, we refer the reader to Améndola et al. (2020) and Derksen, Makam, and Walter (2020) for recent results, and the references therein, as well as the books by Derksen and Kemper (2015) and Sturmfels (2008).

1.5 Outline

In Chapter 2 we will provide a very brief introduction to the geometric side of invariant theory, describing the fundamental computational problems which will be of importance to us in the subsequent chapters. We will also see how optimization problems naturally appear in this setting, and we will revisit some of the problems discussed in the introduction through the perspective of such invariant-theoretic optimization problems. Finally, we conclude the chapter with an exploration of *geodesic convexity* and the beautiful non-commutative duality theory that comes from such optimization problems.

In Chapter 3, we will give a more rigorous analysis of matrix scaling. Specifically, we will cover the strongly convex analysis of Kwok, Lau, and Ramachan-

dran (2019), which will allow us to introduce many tools from convexity in this simpler setting. Then we will analyze the tensor scaling problem, and again consider the strongly convex case. In this more general setting, we will use tools from geodesic convexity.

In Chapter 4 we will apply the results on strongly convex tensor scaling to give near-optimal sample complexity bounds for a well-studied covariance estimation problem in statistics. These results will be based on Franks et al. (2021).

2 Geometric invariant theory

In this chapter we give a very brief overview of the general setting in which scaling problems arise and the convex-like properties inherent in such problems. For a more thorough and rigorous exposition of this chapter, we refer the reader to Bürgisser, Franks, et al. (2019).

As it turns out, since the origins of geometric invariant theory in the seminal work of Hilbert (1893), with the definition of the null cone of a group action, an optimization problem was implicit in the characterization of "singular orbits" (which are the orbits whose closure contain the zero vector). Hilbert used the null cone, which we will define in this chapter, to provide a constructive proof of the finiteness of generators for the invariant ring of $\mathrm{SL}_3(\mathbb{C})$ acting on the space of trivariate homogeneous polynomials of degree d. One important result in Hilbert's paper is the characterization of the null cone as the zero set of all homogeneous non-constant invariant polynomials. In the 1960s, Mumford proved that the set of points outside of the null cone can be given a structure of an algebraic variety, and thus the null cone appears prominently in geometric invariant theory. In the 1970s, Kempf and Ness (1979) proved that the null cone can also be characterized by a non-commutative duality arising from the group action on the vector space. In this chapter we will be exploring this connection, as well as the computational aspects of the optimization problems appearing in geometric invariant theory.

As we are only concerned with the geometric aspects, we will not describe the algebraic properties of geometric invariant theory, such as the ring of invariant polynomials the the problems of finite generation as an algebra, separating invariants, and other structural results. For two thorough introductions to the algebraic side of invariant theory, we refer the reader to the books Derksen and Kemper (2015) and Sturmfels (2008).

2.1 General setting

As mentioned in Chapter 1, scaling problems arise when we have a "continuous group" acting linearly on a vector space. The formal general setting where scaling problems arise is when we take G to be a connected symmetric matrix Lie group, that is, G is a subgroup of $\mathrm{GL}_n(\mathbb{C})$ which is a Zariski-closed, connected (under the standard topology) and such that $g^\dagger \in G$ whenever $g \in G$.

For the sake of concreteness, in this survey we will only study the cases where G is one of the groups $\mathrm{SL}_n(\mathbb{C})$, $\mathrm{diag}_n(\mathbb{C})$, or products of these groups, V is a finite-dimensional Hermitian space ($\mathbb{C}^m$ for some $m \in \mathbb{N}$), and the action of G is given by a representation $\pi : G \to \mathrm{GL}(V)$. For the most general setting and a more thorough discussion of scaling problems, we refer the interested reader to Bürgisser, Franks, et al. (2019). And for a reference on geometric invariant theory, we refer the reader to Wallach (2017).

2.2 Orbits and orbit closures

Given an element $u \in V$, we can define the **G-orbit** of u as the set of all elements in the vector space V that can be reached from u via an action of the group G, that is, the orbit is the following set:

$$G \cdot u := \{w \in V \mid w = \pi(g)u, \text{ for some } g \in G\}.$$

Since we are studying actions of continuous groups on finite-dimensional Hermitian spaces, it is natural (and as we will see much more important) to consider the closure of the orbits under the natural metric induced by the norm on V. This yields the **orbit closure**, which is the set defined by the union of the orbit and its limit points, denoted by $\overline{G \cdot u}$.

With these geometric definitions, two natural computational problems arise:

Problem 2.2.1 (Orbit closure intersection)**.** *Given two elements $u, w \in V$, do their orbit closures intersect? That is,*

$$\overline{G \cdot u} \cap \overline{G \cdot w} \neq^{?} \emptyset$$

As we will see in the next section, a special case of the orbit closure intersection problem, when one of the vectors is the zero vector, is very important in several areas of mathematics. Moreover, the set of all such elements of V whose orbit closures contain the zero element, the null cone of the group action, was defined in Hilbert (1893) where he showed the importance of the null cone for the algebraic setting of invariant theory.

Problem 2.2.2 (Orbit closure containment)**.** *Given two elements $u, w \in V$, is it the case that the orbit closure of u contains the orbit closure of w? That is:*

$$\overline{G \cdot w} \subseteq^{?} \overline{G \cdot u}$$

The orbit closure containment problem appears to be a much harder problem than the orbit closure intersection one, as it contains as a subproblems a geometric version of the famous VP vs VNP problem from algebraic complexity theory Mulmuley and Sohoni (2001) as well as matrix completion problems and the slice-rank problem Bläser et al. (2021).

2.3 Null cone & optimization

In his seminal paper, Hilbert (1893) defined the null cone as the set of points which contain zero in their orbit closures. He proved that the null cone is the zero set of all homogeneous, non-constant invariant polynomials, and used the null cone to construct a set of generators for the invariant ring of polynomials for the $\mathrm{SL}_3(\mathbb{C})$ action on degree d homogeneous polynomials in 3 variables. The null cone also appears prominently in the construction by Mumford of moduli spaces, since these are the "bad points" that one must remove in order to give the quotient space V/G the structure of an algebraic variety in a way that the quotient map becomes a morphism. We will not discuss these aspects of the null cone here, but rather emphasize the importance of the null cone, and of the null-cone problem, in computer science and other areas of mathematics.

Definition 8 (Null Cone)**.** The set of all elements $u \in V$ such that $0 \in \overline{G \cdot u}$ is called the *null cone*, denoted by $\mathcal{N}(G, V)$.

As we will see in this and later chapters, the null cone is an important subset of our vector space, and in particular the *nul-cone problem*, which we now define, is an important optimization problem which naturally appears in many areas of science and mathematics.

Problem 2.3.1 (Null-cone problem)**.** *Given an element $u \in V$, is $0 \in \overline{G \cdot u}$?*

Since we are working over an inner product space V, we can consider the null-cone problem as a norm minimization problem, since 0 is in the orbit closure if, and only if, the element of minimum norm in the orbit closure has zero norm! Motivated by this connection, we have the following definition:

Definition 9 (Capacity)**.** The *capacity* of a vector $u \in V$ is given by the value of the following optimization problem:

$$\operatorname{cap}(u) := \inf_{g \in G} \|\pi(g)u\|^2$$

where the norm is the norm induced by the inner product on V.

Problem 2.3.2 (Null-cone problem - optimization version)**.** *Given a vector $u \in V$, decide whether* $\operatorname{cap}(u) = 0$.

Given the definition above, and the importance of the norm function along a group orbit for geometric invariant theory, given any vector $u \in V$, the family of functions $f_u : G \to \mathbb{R}$ defined by

$$f_u(g) := \|\pi(g)u\|^2$$

is the *Kempf-Ness* function.

Since the norm is induced by the inner product, we see that the Kempf-Ness family of functions can actually be interpreted as a real-valued functions from the manifold of positive definite matrices. The importance of this connection, together with the properties of positive definite matrices will be quite important for our optimization problems. In Section 2.5 we will briefly study the necessary properties of the manifold of positive definite matrices that we need, and then come back to see the Kempf-Ness family of functions in this new setup.

2.4 Examples of Scaling Problems

Now that we have seen a brief picture of the general setting, let us see how the problems that we discussed in Chapter 1, as well as many additional known objects in mathematics, are instances of the null cone of particular group actions!

We encourage the reader to work out the details of these examples.

2.4.1 Left-right multiplication

Let $G = \mathrm{SL}_n(\mathbb{C}) \times \mathrm{SL}_n(\mathbb{C})$ act on $V = \mathrm{Mat}_n(\mathbb{C})$ by left-right multiplication, taking $(L, R) \circ A \mapsto LAR^T$. In this case, we have that the orbit of a matrix $A \in V$ of rank $r < n$ is the same as the orbit of the matrix $I_r \oplus 0_{n-r}$. One can see this since the group action encodes row reduction, column permutations, and column scaling (without changing the determinant), and therefore we can always obtain the canonical form just mentioned. The orbit of an invertible matrix A is the same as the orbit of $I_{n-1} \oplus \det(A)$.

When we look at the orbit closures, we note that the orbit closures of singular matrices will contain the zero matrix. Since the determinant is an invariant polynomial for this action, we see that in this case the null cone is the set of all singular matrices! Moreover, the scaling problem here in particular captures the problem of deciding whether a matrix is singular.

2.4.2 Matrix scaling

In the matrix scaling action, we have $G = \mathrm{diag}^*_{\mathbb{C}}(n) \times \mathrm{diag}^*_{\mathbb{C}}(n)$, where $\mathrm{diag}^*_{\mathbb{C}}(n)$ is the set of diagonal $n \times n$ matrices with determinant 1. Our group G acts on the space of matrices $\mathrm{Mat}(n, \mathbb{C})$ by left and right multiplication. In this setting, by using the exponential map $x \to e^x$ for the group action and also for the entries of the matrix being acted upon, one can see that the orbits of the matrix scaling problem correspond to weighted bipartite graphs.

The set of matrices that are in the null cone correspond to the (weighted) bipartite graphs which have no perfect matching, so the matrix scaling problem captures the perfect matching problem in bipartite graphs!

2.4.3 Conjugation action

In the conjugation action our group is $G = \mathrm{GL}_n(\mathbb{C})$ acting on $V = \mathrm{Mat}(n, \mathbb{C})$ by conjugation, that is, $g \circ A \mapsto gAg^{-1}$. The orbit of a matrix A corresponds to the orbit of its Jordan normal form. The orbit closure of a matrix will contain the diagonal matrices with the same eigenvalues (counting multiplicities). Thus, in this case the null cone corresponds to exactly the nilpotent matrices! Moreover, the scaling problem in this case corresponds to diagonalizing a matrix by conjugation.

2.4.4 Homogeneous bivariate polynomials

In this problem, we have the group $G = \mathrm{SL}_2(\mathbb{C})$ acting on binary forms of degree d, that is, $V = \mathbb{C}[x, y]_d$, by a change of variables. Thus, a polynomial $p(x, y)$ is taken to $A \circ p \mapsto p((x, y) \cdot A)$.

The orbits of this group action preserve the zero/nonzero pattern of the roots and their multiplicities, although they (and their closures) are more complex to describe. The null cone in this case is the set of polynomials with a root of multiplicity at least $\lfloor d/2 \rfloor + 1$. For more details on this action, see Weyman (1989) and references therein.

2.5 Geodesics in Positive Definite Manifold

Before describing the geometry that arises from a group action and geodesic convexity, we need to review some preliminaries from optimization and linear algebra, which we do here.

2.5.1 Linear Algebra Preliminaries

Definition 10 (Linear Operators)**.** For vector spaces U, V, $L(U, V)$ is the space of linear operators $T : U \to V$. If $U = V$ we will denote this as $L(U)$.

Any choice of bases $\{u_1, \ldots, u_{\dim(U)}\} \subseteq U, \{v_1, \ldots, v_{\dim(V)}\} \subseteq V$ induces a matrix representation for $L(U, V)$. Namely, for the operator $T \in L(U, V)$ if $Tu_j = \sum_{i=1}^{\dim(V)} M_{ij} v_i$ then the matrix representation is $\{M_{ij}\}$. Note that this representation is unique due to the linear independence of bases.

Definition 11 (Invertible Linear Operators)**.** $GL(V) \subseteq L(V)$ is the subset of invertible linear operators on V. It is a group by left (or right) composition. $SL(V) \subseteq GL(V)$ is the subgroup of unit determinant operators.

If $\dim(V) = d$ and $\{v_1, \ldots, v_d\} \subseteq V$ is a basis, then by the matrix representation given above $GL(V)$ is isomorphic to the group of invertible matrices $GL(n, \mathbb{F})$ with matrix multiplication. $SL(n, \mathbb{F}) \subseteq GL(n, \mathbb{F})$ is the subgroup of unit determinant matrices.

Definition 12 (Inner Product Spaces)**.** If $V, \langle \cdot, \cdot \rangle$ is an inner product, then $U(V) \subseteq GL(V)$ is the subgroup of unitary matrices which preserve the inner product:

$$\forall x, y \in V, U \in U(V) : \langle Ux, Uy \rangle = \langle x, y \rangle.$$

Definition 13 (Adjoint Operator)**.** A^* is the adjoint of matrix $A \in L(V)$, and is defined as the unique operator satisfying

$$\forall x, y \in V : \langle Ax, y\rangle = \langle x, A^* y\rangle.$$

Theorem 2.5.1 (Spectral Theorem)**.** *$A \in L(V)$ is normal if $AA^* = A^*A$. If $A^* = A$, then A is Hermitian and we denote this $H(V)$.*

Every normal A can be written

$$A = UDU^*,$$

where U is unitary and D is diagonal. If A is Hermitian then D can be taken to be real.

Definition 14 (Positive Semidefinite Matrices)**.** A is positive semidefinite if it is Hermitian and

$$\forall x \in V : \langle x, Ax\rangle \geqslant 0.$$

It is positive definite if the inequality is strict for all $x \neq 0$. We denote $\mathrm{PD}(V) \subseteq GL(V)$ to be the subset of positive definite matrices.

Theorem 2.5.2 (Polar Decomposition)**.** *Any $A \in GL(V)$ can be uniquely written $A = UP$ for unitary U and positive definite P. This decomposes the group $GL(V) \simeq U(V) \times \mathrm{PD}(V)$, and this decomposition is a diffeomorphism.*

If $A \in SL(V)$, then U, P will also have unit determinant. This gives the decomposition $SL(V) \simeq SU(V) \times \mathrm{SPD}(V)$, where $SU(V) \subseteq U(V), \mathrm{SPD}(V) \subseteq \mathrm{PD}(V)$ denote the subset of unit determinant elements.

Proof. The first statement is standard and we will not prove it here.

For the second statement, by the multiplicativity of det we have

$$1 = \det(A) = \det(UP) = \det(U)\det(P).$$

Since $U \in U(V)$, all of its eigenvalues are purely complex, and $P \in P(V)$ implies all of its eigenvalues are positive real. Therefore their determinants multiply to 1 iff they are both also 1. □

This statement can be vastly generalized by the Cartan decomposition to all semisimple Lie groups, but we will not need this fact for our applications, as the polar decomposition works for any matrix Lie groups. We are now ready to define geodesics in the manifold of positive definite matrices.

Definition 15 (Geodesics)**.** For any two elements $P, Q \in \mathrm{PD}(V)$, we can define the following geodesic from P to Q, or path $\gamma_{P,Q} : [0,1] \to \mathrm{PD}(V)$ as

$$\gamma_{P,Q}(t) := P^{1/2}(P^{-1/2}QP^{-1/2})^t P^{1/2} = P^{1/2}\exp(t\log(P^{-1/2}QP^{-1/2}))P^{1/2}.$$

This is in fact the shortest path according to a natural metric on $\mathrm{PD}(V)$ (Bhatia (2009)). It turns out that this is the only geometry on the PD manifold which satisfies natural equivariance properties, which we will now define.

First note that for any $P \in \mathrm{PD}(V)$, $X := \log P$ exists and is Hermitian since P is Hermitian with all eigenvalues strictly positive. Therefore there is a natural mapping $\gamma : H(V) \to P(V)$ at the identity:

$$\gamma_{I_V}(X) := e^X.$$

The curves from Definition 15 satisfy the following natural equivariance property:

$$\gamma_{P,Q}(t) \equiv P^{1/2}\gamma_{I_V, P^{-1/2}QP^{-1/2}}(t)P^{1/2}.$$

This equivariance property essentially tells us that all geodesics through a particular matrix look the same as the geodesics through the identity matrix.

When we discuss geodesically convex functions, our applications will center exclusively on $\mathrm{PD}(V)$, $\mathrm{SPD}(V)$ and direct products of these manifolds.

2.6 Convexity Preliminaries

In this section we define geodesic convexity and strong convexity, and state some properties which we will need about these functions for the later chapters.

2.6.1 Convexity in Euclidean Spaces

Before we define geodesically convex functions, let us quickly review the usual notions and some lemmas of convexity in the Euclidean setting.

Definition 16. Function $h : \mathbb{R} \to \mathbb{R}$ is convex if either of the following conditions hold

1. $\forall t \in \mathbb{R} : h''(t) := \partial_x^2 h(x)|_{x=t} \geqslant 0.$
2. $\forall s, t \in \mathbb{R} : h(t) - h(s) \geqslant h'(s)(t-s).$

As a strengthening, h is α-strongly convex at $s \in \mathbb{R}$ if

1. $h''(s) \geq \alpha$.
2. $\forall t \in \mathbb{R} : h(t) - h(s) \geq h'(s)(t-s) + \frac{\alpha}{2}(t-s)^2$.

Lemma 2.6.1. *For $h : \mathbb{R} \to \mathbb{R}$ that is α strongly convex, then for any $s \in \mathbb{R}$, the optimum can be lower bounded by*

$$h^* := \inf_{t \in \mathbb{R}} h(t) \geq h(s) - \frac{|h'(s)|^2}{2\alpha}.$$

Proof. By α-strong convexity we have for any $s \in \mathbb{R}$

$$\begin{aligned} \inf_{t \in \mathbb{R}} h(t) &\geq \inf_{t \in \mathbb{R}} h(s) + h'(s)(t-s) + \frac{\alpha}{2}(t-s)^2 \\ &= h(s) - \frac{(h'(s))^2}{2\alpha}, \end{aligned}$$

where in the last step we chose infimizer $t - s = \frac{-h'(s)}{\alpha}$. □

Remark 2.6.2. Note that we have only used strong convexity for t in the interval $[s, s - \frac{h'(s)}{\alpha}]$.

Lemma 2.6.3. *For convex $h : \mathbb{R} \to \mathbb{R}$ with optimizer t_*, assume h is α-strongly convex for all $|s - t| \leqslant R$. Then it is α-strongly convex on the level set*

$$L := \{s \in \mathbb{R} \mid f(s) - f(t_*) \leqslant \alpha R^2/2\}.$$

Further if $s \in \mathbb{R}$ such that $|h'(s)| < \alpha R$, then $s \in L$.

Proof. We show the first statement in contrapositive by giving a lower bound for all $s - t_* > R$. By a simple translation we can assume that $t_* = 0, h(t_*) = 0$. Then

$$h(s) = \int_{t=0}^{s} h'(t) = \int_{t=0}^{s} \left(h'(0) + \int_{r=0}^{t} h''(r) \right) \geq 0 + \int_{t=0}^{R} \alpha t = \frac{\alpha R^2}{2},$$

where the first two steps are by the fundamental theorem of calculus, and the third step was by optimality of $t_* = 0$ and α-strong convexity. The same lower bound holds for $h(-s)$, so by convexity if $h(s) \leqslant \frac{\alpha R^2}{2}$ then $|s| \leqslant R$ is in the strongly convex region.

To show the second statement, note that if $s \in \mathbb{R}$ such that $|h'(s)| \leqslant R$ and $|s| \leqslant R$, then we are done by Lemma 2.6.1 as

$$h(s) - h(t_*) \leqslant \frac{(h'(s))^2}{2\alpha} \leqslant \frac{\alpha^2 R^2}{2\alpha} = \frac{\alpha R^2}{2},$$

so $s \in L$. We show that the other case, $|s| > R$, contradicts the first derivative assumption.

$$\alpha R > h'(s) = h'(0) + \int_{t=0}^{s} h''(s) \geqslant 0 + \int_{t=0}^{R} \alpha + \int_{t=R}^{s} 0 = \alpha R,$$

where the first step was by fundamental theorem of calculus, and the second step was by optimality of $t_* = 0$ and α-strongly convexity for $t \in [0, R]$. This gives the required contradiction. □

All of these properties lift to convex functions on vector spaces.

Definition 17. For vector space V, function $f : V \to \mathbb{R}$ is convex if for every $x, y \in V$, the univariate restriction $t \to f(x + ty)$ is convex.

If V has inner product $\langle \cdot, \cdot \rangle$, then the convexity conditions can be written equivalently as

1. $\forall v \in V : \nabla^2 f(v) \succeq 0$.
2. $\forall u, v \in V : f(v) - f(u) \geqslant \langle \nabla f(u), v - u \rangle$.

f is α-strongly convex in norm $\| \cdot \|$ at point $v \in V$ if for every $x \in V$ the univariate restriction $t \to f(v + tx)$ is $\alpha \|x\|$-strongly convex.

Definition 18. For vector space V and function $f : V \to \mathbb{R}$, $x \in V$ is a critical point of f iff

$$\forall v \in V : \partial_{t=0} f(x + tv) = 0,$$

or equivalently that for every $v \in V$, the univariate restriction $t \to f(x + tv)$ has critical point $t = 0$.

2.6.2 Geodesic Convexity

Definition 19 (Geodesic Convexity)**.** Let f be a function $f : P \to \mathbb{R}$ where P is some connected submanifold of $\mathrm{PD}(V)$ for some vector space V that is closed under geodesics (Definition 15). Then f is geodesically convex if every univariate restriction $t \to f(\gamma_{P,Q}(t))$ is convex for every pair $P, Q \in P$.

In our case, since we are studying the Kempf-Ness family of functions $f_v : G \to \mathbb{R}$ given by $f_v(g) = \|g \circ v\|^2$, the inner product structure on V tells us that the Kempf-Ness function can be actually interpreted as a function $f_v : \mathrm{PD}(V) \to \mathbb{R}$ for every $v \in V$. One can see this as

$$f_v(g) = \|g \circ v\|^2 = \langle g \circ v, g \circ v \rangle = \langle v, g^* g \circ v \rangle = \langle v, pv \rangle,$$

where $p := g^* g$ is the polar component.

We will use these functions to give optimization formulations of scaling problems, and therefore the geodesic convexity property will be crucial for our bounds. Specifically in Proposition 3.1.1 and Lemma 3.2.4, we will use the added structure of this family of functions to prove geodesic convexity by considering the simpler geometry of P from the identity.

2.7 Optimization in Geometric Invariant Theory

2.7.1 Commutative case & convex optimization

Now that we have the basic definitions from linear algebra and convex optimization at hand, we can look at the general scaling problem in the case where we have commutative groups, and we will see that this problem corresponds to the setting of *geometric programming* in classical convex optimization.

Given a commutative group G and a representation (or group action) $\pi : G \to \mathrm{GL}(V)$, we have that the matrices $\pi(g)$ for all $g \in G$ are simultaneously diagonalizable. Thus, after an appropriate change of basis we can think of G as a subgroup of the invertible diagonal matrices $\mathrm{diag}_{\mathbb{C}}(n)$. Moreover, the group action can be described by the simultaneous eigenvalues of the group action, which after an appropriate change of basis become the standard basis vectors.

In particular, for an eigenvector $v \in V$, if we write $\pi(g) = \mathrm{diag}(t_1, \ldots, t_n)$, we can write the group action in the following way:

$$\pi(g)v = \prod_{j=1}^{n} t_i^{w_i} v$$

where each $w_j \in \mathbb{Z}$. The vectors $\omega := (w_1, \ldots, w_n)$ are called the *weights* of the representation π.

Hence, the Kempf-Ness function for the vector $u = \sum \alpha_v v$, where the vectors v form an orthonormal eigenbasis for the group action, becomes:

$$f_u(g) = \sum |\alpha_v|^2 \cdot \prod_{j=1}^{n} |t_i|^{2w_i}$$

In particular, our optimization problem (the capacity problem) becomes:

$$\begin{aligned} \inf \ & \sum |\alpha_v|^2 \cdot \prod_{j=1}^{n} |t_i|^{2w_i} \\ \text{s.t.} \ & t_i \in \mathbb{C}^* \end{aligned}$$

Since $t_i \in \mathbb{C}^*$, we have $|t_i| \in \mathbb{R}_{>0}$ and hence we can apply the variable substitution (via the exponential map) $|t_i| = e^{x_i}$ where $x_i \in \mathbb{R}$. This transformation makes our capacity problem become:

$$\begin{aligned} \inf \ & \sum |\alpha_v|^2 \cdot \exp\left(\sum_{i=1}^{n} 2x_i w_i\right) \\ \text{s.t.} \ & x_i \in \mathbb{R} \end{aligned}$$

Thus the problem of optimizing the function $f_u(g)$, after an appropriate change of coordinates, becomes an unconstrained geometric program! We can then use standard Euclidean optimization methods to solve the capacity problem in the case of commutative groups!

As it turns out, the commutative case above already has applications in statistics, as it captures the maximum likelihood estimation problem for log-linear models. We refer the reader to Améndola et al. (2020) for the explicit connections, as well as to Straszak and Vishnoi (2019) for complexity aspects on these optimization problems.

2.7.2 Non-commutative Case & geodesically convex optimization

After seeing the connection between capacity (i.e. optimization of a Kempf-Ness function) in the commutative setting and standard convex optimization, one is tempted to wonder whether in the non-commutative case such a global change of coordinates into a convex optimization problem is actually possible. While that does not seem to be the case, what we do know is that a more general convexity phenomenon happens when the group is non-commutative, albeit now in

the manifold of positive definite matrices. Here, we will briefly discuss how this phenomenon happens and state some of the properties we will need for the later chapters.

Given a non-commutative group G (again, think of G as being either $\mathrm{SL}(n, \mathbb{C})$ or products of such groups), and a representation (or group action) $\pi : G \to \mathrm{GL}(V)$, as we have seen in Section 2.6.2, the family of Kempf-Ness functions can be thought of as functions $f_u : \mathrm{PD}(V) \to \mathbb{R}$ for every $u \in V$. An important property of this family of functions is that they are *equivariant* with respect to the group action, in much the same way as the geodesics in $\mathrm{PD}(V)$ are equivariant.

Proposition 2.7.1 (Equivariance of Kempf-Ness functions)**.** *Given a group action G on a finite dimensional complex vector space V, we have that*

$$f_u(\pi(g)\pi(h)) = f_{\pi(h)u}(\pi(g))$$

for any $g, h \in G$ and $u \in V$.

One benefit of such equivariance property of the family of Kempf-Ness functions is that it is sufficient to study the properties of such functions in the neighborhood of the identity, since the functions "locally look the same." In particular, we can define the usual notions of gradient and Hessian for the family of Kempf-Ness functions around the identity, as it is done in Bürgisser, Franks, et al. (2019, Section 3). The gradient of the Kempf-Ness function is also known as the moment map.

In the case of a non-commutative group action, it turns our that the family of Kempf-Ness functions is geodesically convex, as defined in Section 2.6.2. Thus the norm minimization problem (or capacity) turns out to be a convex optimization problem, where the convexity is along the geodesics on the manifold of positive definite matrices. This gives us hopes to generalize the methods from convex optimization to this new setting, in order to solve the null cone problem. This has been done in the series of works Allen-Zhu et al. (2018) and Bürgisser, Franks, et al. (2019), and we refer the reader to the latter work for the development of this paradigm in the most general setting.

2.7.3 Non-commutative duality theory

The geodesic convexity of the Kempf-Ness family of functions can be used to establish a non-commutative duality theory, which greatly generalizes linear program duality to the non-commutative setting! This is the content of the *Kempf-Ness* theorem, proved in Kempf and Ness (1979). This theorem essentially states

that the capacity of an element $u \in V$ is zero (i.e., zero is in the orbit closure of u) iff the norm of the geodesic gradient is always nonzero along the orbit closure. In particular, the Kempf-Ness theorem states that the norm minimization problem (i.e. the capacity) is dual to the minimization of the geodesic gradient along the group orbit!

A quantitative version of the Kempf-Ness theorem has been developed in Bürgisser, Franks, et al. (2019).

2.8 References

All of the material presented here can be found, in much more generality, in Bürgisser, Franks, et al. (ibid.) and references therein. For the readers interested in learning more about the manifold of positive definite matrices, we recommend the book by Bhatia (2009)

3 Scaling problems and algorithms

3.1 Matrix Scaling

3.1.1 Sinkhorn Scaling as Convex Optimization

In this section, we can use convex optimization to reframe the scaling algorithm and analysis in Linial, Samorodnitsky, and Wigderson (2000). We first formulate a generalization of the problem to matrix tuples. This makes the problem slightly more natural, and allows to lift our results to the non-commutative setting (e.g. operator and tensor scaling).

Definition 20. For matrix tuple $A = \{A_1, ..., A_K\}$ where $A_k \in \mathrm{Mat}_{\mathbb{C}}(d, n)$, its size is defined

$$s(A) := \sum_{k=1}^{K} \|A_k\|_F^2.$$

The row and column sums for $i \in [d]$, $j \in [n]$ are defined as

$$r_i(A) := \sum_{k=1}^{K} \sum_{j=1}^{n} |A_k|_{ij}^2, \qquad c_j(A) := \sum_{k=1}^{K} \sum_{i=1}^{d} |A_k|_{ij}^2.$$

Definition 21. Tuple $A = \{A_1, ..., A_K\} \in \mathrm{Mat}_{\mathbb{C}}(d,n)^K$ is ε-doubly balanced if

$$r_i(A) \in \frac{s(A)}{d}(1 \pm \varepsilon), \quad c_j(A) \in \frac{s(A)}{n}(1 \pm \varepsilon), \tag{3.1.1}$$

for all $i \in [d]$, $j \in [n]$. A is doubly balanced if the above holds with $\varepsilon = 0$.

Definition 22 (Matrix Scaling Problem)**.** For matrix tuple $A \in \mathrm{Mat}_{\mathbb{C}}(d,n)^K$, we can define an action of diagonal matrices $X \in \mathrm{diag}(d), Y \in \mathrm{diag}(n)$ by left/right scaling:

$$e^X A e^Y := \{e^X A_1 e^Y, ..., e^X A_K e^Y\}.$$

The input to a matrix scaling problem is a matrix tuple A.

1. **Success**: Output scalings (X, Y) such that $e^X A e^Y$ is doubly balanced.
2. **Failure**: Proof that no scaling of A is doubly balanced.

When $d = n$ and $K = 1$, this is equivalent to the matrix scaling problem of Linial, Samorodnitsky, and Wigderson (2000) on $B_{ij} := |A_{ij}|^2$. Because Definition 21 is homogenous, we can assume the following normalization on scalings without loss of generality.

Definition 23. Matrix scalings can be restricted to the subspace

$$\mathfrak{t} := \{(X, Y) \in \mathrm{diag}(d) \oplus \mathrm{diag}(n) \mid Tr[X] = Tr[Y] = 0\}.$$

Note that scalings $\{(e^X, e^Y) \mid (X, Y) \in \mathfrak{t}\}$ are all determinant one, since $\det(e^X) = \exp(Tr[X])$. At times it will be convenient to view these elements as vectors, so by abuse of notation we will also use $\mathfrak{t}$ to refer to the following vector space:

$$\mathfrak{t} := \left\{ (X, Y) \in \mathbb{R}^d \oplus \mathbb{R}^n \mid \sum_{i=1}^{d} X_i = \sum_{j=1}^{n} Y_j = 0 \right\}.$$

It turns out that the work of Kempf and Ness (1979) gives a convex formulation for the matrix scaling problem. In fact, this phenomena is far more general, and we will revisit this in Section 3.2 for the tensor scaling problem.

Definition 24. For tuple $\{A_1, ..., A_K\} \in \mathrm{Mat}(d,n)^K$, the Kempf-Ness function $f_A : \mathfrak{t} \to \mathbb{R}$ is defined

$$f_A(2X, 2Y) := s(e^X A e^Y) = \sum_{k=1}^{K} \|e^X A_k e^Y\|_F^2 = \sum_{k=1}^{K} \sum_{ij} e^{2X_i} |A_k|_{ij}^2 e^{2Y_j},$$

where size is given in Definition 20. The factor 2 is just to remove leading constants for future calculations.

Proposition 3.1.1. *For matrix input $A \in \mathrm{Mat}(d,n)^K$, f_A is convex on its domain* $\mathfrak{t}$*, and $(X,Y) \in \mathfrak{t}$ is a doubly balanced scaling of A iff (X,Y) is a critical point for f_A iff (X,Y) is a global minimum on* $\mathfrak{t}$.

We omit the proof, which is a straightforward calculation of derivatives, as we will prove a generalization of this statement for tensor scaling in Section 3.2.

Proposition 3.1.1 shows that the Kempf-Ness function f_A gives a convex formulation for the matrix scaling problem in Definition 22. Using this perspective, we can reframe Sinkhorn scaling as a convex optimization algorithm.

To apply tools from convex optimization, our first off-the-shelf approach would be gradient descent. For this to be well-defined, we need to choose an inner product on our domain $\mathfrak{t}$.

Definition 25 ($\mathfrak{t}$ Inner Product)**.** For elements $(X,Y),(X',Y') \in \mathfrak{t}$ (Definition 23), we define inner product

$$\langle (X,Y),(X',Y')\rangle_{\mathfrak{t}} := \frac{1}{d}\sum_{i=1}^{d} X_i X_i' + \frac{1}{n}\sum_{j=1}^{n} Y_j Y_j'.$$

The induced norm is $\|(X,Y)\|_{\mathfrak{t}} = \sqrt{\langle (X,Y),(X,Y)\rangle_{\mathfrak{t}}}$.

The justification of this normalization will become clearer in Section 3.1.2. For now we observe that this makes the gradient well-defined.

Proposition 3.1.2. *For input $A \in \mathrm{Mat}(d,n)^K$, the gradient of f_A at point $(2X,2Y) \in \mathfrak{t}$ is:*

$$\nabla f_A(2X,2Y) = \{d \cdot r_i(e^X A e^Y) - s(e^X A e^Y)\} \oplus \{n \cdot c_j(e^X A e^Y) - s(e^X A e^Y)\}$$

where r,c,s refer to the row/column sums and size of $e^X A e^Y$. We will often use shorthand ∇_A for $\nabla f_A(0,0)$, and $(\nabla_A^L, \nabla_A^R) \in \mathfrak{t}$ for the left and right parts, which involve the row/column sums respectively.

We omit the proof as it is a straightforward calculation, and we revisit this proposition more explicitly in the more general tensor setting.

We can now reframe Lemma 1.3.5 in this more general setting as a guarantee of the progress of Sinkhorn scaling. In this setting, we consider the following variant of Sinkhorn

$$A' := \left(\frac{dR}{\det(dR)^{1/d}}\right)^{-1/2} A, \quad A' := A\left(\frac{nC}{\det(nC)^{1/d}}\right)^{-1/2},$$

where $R := \operatorname{diag}\{r_i(A)\}_{i\in[d]}, C := \operatorname{diag}\{c_j(A)\}_{j\in[n]}$. This is so that the scalings remain within $\mathfrak{t}$.

Lemma 3.1.3. *Let $A \to A'$ represent one iteration of Sinkhorn scaling. Then the size decreases by*

$$\log s(A') \leqslant \log s(A) - \frac{1}{6}\min\left\{\frac{\|\nabla_A^{L,R}\|_{\mathfrak{t}}^2}{s(A)^2}, \gamma\right\}$$

where $\gamma = \frac{1}{d}, \frac{1}{n}$ for row and column normalization steps respectively. This can be written in terms of the Kempf-Ness function as

$$\log f_A(X,Y) - \log f_A(0,0) \leqslant -\frac{1}{6}\min\{\|\nabla \log f_A(0,0)\|_{\mathfrak{t}}^2, \gamma\},$$

where (X,Y) represents one iteration of Sinkhorn, and we have omitted the L,R superscript depending on whether it is a row or column normalization step.

Proof. We show the lemma for the case when $t = 0$ and we are normalizing the left marginal. The other cases follow by induction.

$$s(A') = \sum_{i=1}^{d} \det(dR)^{1/d} \sum_{j=1}^{n} \frac{|A_{ij}|^2}{dr_i(A)} = \left(\prod_{i=1}^{d} dr_i(A)\right)^{1/d},$$

where in the last step we used $\sum_{j=1}^{n} |A_{ij}|^2 = r_i(A)$. To bound this value, we can use Lemma 1.3.6 with $x_i := \frac{dr_i}{s(A)}$ to show

$$-\log \prod_{i=1}^{d} x_i = -\log \frac{\prod_{i=1}^{d} dr_i}{s(A)^d} \geqslant \frac{1}{6}\sum_{i=1}^{d}\left(\frac{dr_i}{s(A)} - 1\right)^2 \geqslant \frac{1}{6}\min\left\{\frac{d\|\nabla_A^L\|_{\mathfrak{t}}^2}{s(A)^2}, 1\right\}.$$

Plugging this calculation into the formula for size gives the result. The calculation for column-normalization is the same. Since $\nabla \log f = \frac{\nabla f}{f}$, and $f_A(0,0) = s(A)$, we can rewrite

$$\log s(A') - \log s(A) = \log f_A(X,Y) - \log f_A(0,0),$$

$$\frac{\|\nabla_A^L\|_{\mathfrak{t}}^2}{s(A)^2} = \|\nabla^L \log f_A(0,0)\|_{\mathfrak{t}}^2$$

which gives the statement in terms of the Kempf-Ness function. □

The algorithm of Linial, Samorodnitsky, and Wigderson (2000) used a simple preprocessing step to guarantee that the permanent was lower bounded, or equivalently that $\log f_A$ was finite. We can generalize this analysis in the case when we have a guarantee on the optimum of f_A.

Theorem 3.1.4. *Consider matrix tuple $A \in \mathrm{Mat}(d,n)^K$ such that $s(A) = f_A(0,0) = 1$ and*

$$f^* := \inf_{(X,Y)\in\mathfrak{t}} f_A(2X,2Y) \geqslant \exp(-\nu).$$

Then for any $\delta \leqslant \frac{1}{n}$, some iteration of Sinkhorn scaling satisfies $\|\nabla \log f_A(X_t,Y_t)\|_{\mathfrak{t}}^2 \leqslant \delta$ for some

$$T \lesssim \frac{\nu}{\delta}.$$

Proof. Let T be the first time $\|\nabla \log f_A(X_t,Y_t)\|_{\mathfrak{t}}^2 \leqslant \delta$. Then until this time we make significant progress:

$$\begin{aligned}\log f^* - \log f_A(0,0) &\leqslant \log f_A(X_T,Y_T) - 0 \\ &= \sum_{t<T} \log f_A(X_{t+1},Y_{t+1}) - \log f_A(X_t,Y_t) \leqslant -T\frac{\delta}{6},\end{aligned}$$

where the first step was by definition of f^* and the final step was by Lemma 3.1.3, since the gradient is large for every step before T. Therefore by the assumed lower bound $\log f^* \geqslant -\nu$, we can rearrange to show $T \leqslant \frac{6\nu}{\delta}$. □

The optimization perspective also gives a more principled proof of an approximate version of Theorem 1.3.2 that was needed in Linial, Samorodnitsky, and Wigderson (ibid.). The proof goes by showing the following version of convex duality for the Kempf-Ness function.

Theorem 3.1.5. *For matrix tuple $A \in \mathrm{Mat}(d,n)^K$ with $s(A) = 1$, the optimum of f_A can be lower bounded by*

$$\log f^* - \log f_A(0,0) \geqslant -n\,d\,\|\nabla \log f_A(0,0)\|_{\mathrm{t}}.$$

A grand generalization of this theorem is proved in Bürgisser, Franks, et al. (2019).

3.1.2 Strongly Convex Setting

In this section we will show strong convergence results for abstract convex functions. These will be applied to analyze Sinkhorn's algorithm for "strongly convex" matrix scaling in the following section.

Lemma 3.1.6. *If $f : V \to \mathbb{R}$ is α-strongly convex in norm $\|\cdot\|$ on vector space V, then for any $x \in V$ we have the lower bound*

$$f^* := \inf_{z \in V} f(z) \geqslant f(x) - \frac{\|\nabla f(x)\|^2}{2\alpha}.$$

Furthermore, if $f(x)\|\nabla \log f(x)\|^2 \leqslant \alpha$, we have multiplicative lower bound

$$\log f^* \geqslant \log f(x) - \frac{f(x)\|\nabla \log f(x)\|^2}{\alpha}.$$

Proof. Let z_* be the optimizer of f and $\widetilde{z} := \frac{z_* - z}{\|z_* - z\|}$ the normalized direction towards the optimizer. Then this follows simply from Lemma 2.6.1 by considering the univariate function $h(t) := f(z + t\widetilde{z})$:

$$f(z_*) - f(z) = \inf_{t \in \mathbb{R}} h(t) - h(0) \geqslant -\frac{|h'(0)|^2}{2\alpha} = -\frac{\langle \nabla f(z), \widetilde{z}\rangle^2}{2\alpha} \geqslant \frac{\|\nabla f(z)\|^2}{2\alpha},$$

where the final step was by Cauchy Schwarz.

To show the multiplicative lower bound, note $f(x)\nabla \log f(x) = \nabla f(x)$. We can plug this into the previous statement to show

$$\begin{aligned}
\log f^* &\geqslant \log\left(f(x) - f(x)^2 \frac{\|\nabla \log f(x)\|^2}{2\alpha}\right) \\
&= \log f(x) + \log\left(1 - f(x)\frac{\|\nabla \log f(x)\|^2}{2\alpha}\right) \\
&\geqslant \log f(x) - \frac{f(x)\|\nabla \log f(x)\|^2}{\alpha},
\end{aligned}$$

where the last step was by Taylor approximation for $f(x)\|\nabla \log f(x)\|^2 \leqslant \alpha$ small enough by assumption. □

Remark 3.1.7. The above in fact holds whenever the univariate restriction of f between z, z_* is α-strongly convex. We will use this weaker assumption when analyzing strongly convex matrix scaling. This will also allow us to generalize these results to the geodesically convex setting for tensor scaling (Section 3.2).

In our application in Chapter 4, we will require faster algorithmic convergence. The following is a standard analysis of convex optimization algorithms in the strongly convex setting.

Definition 26. $\mathcal{A}$ is an L-descent algorithm for F if

$$F(\mathcal{A}(z)) \leqslant F(z) - \frac{1}{2L}\|\nabla F(z)\|^2.$$

Note that if F is convex and L-smooth, then standard gradient descent with step size $\frac{1}{2L}$ is an L-descent algorithm for F. Also we have shown in Lemma 3.1.3 that Sinkhorn scaling is an exponential $O(1)$-descent algorithm for $\log f_A$ where f_A is the Kempf-Ness function.

Theorem 3.1.8. *Let $f : V \to \mathbb{R}_+$ be α-strongly convex and $\mathcal{A}$ be an L-descent algorithm for $\log f$ with initial point $z_0 \in V$ and iterates $z_{t+1} := \mathcal{A}(z_t)$. Further assume that initially $\alpha \geqslant f(z_0)\|\nabla \log f(z_0)\|^2$. Then for every $k \in \mathbb{N}$,*

$$\|\nabla \log f(z_T)\|_{\mathfrak{t}}^2 \leqslant 2^{-k}\|\nabla \log f(z_0)\|_{\mathfrak{t}}^2$$

for some $T \leqslant k\frac{4f(z_0)L}{\alpha}$.

Proof. Define $\nabla_t := \nabla \log f(z_t)$ for shorthand. By strong convexity and Lemma 2.6.1, we have

$$\log f^* - \log f(z_0) \geqslant -\frac{f(z_0)\|\nabla_0\|^2}{\alpha}.$$

Now let t_1 be the first time that $\|\nabla_t\|^2 \leqslant \frac{1}{2}\|\nabla_0\|^2$. Then by Definition 26,

$$\begin{aligned}\log f(z_{t_1}) - \log f(z_0) &= \sum_{t<t_1} \log f(z_{t+1}) - \log f(z_t) \\ &\leqslant -\sum_{t<t_1} \frac{\|\nabla_t\|^2}{2L} \leqslant -\frac{t_1\|\nabla_0\|^2}{4L},\end{aligned}$$

where the last step was by the assumption that $\|\nabla_t\|^2 \geqslant \frac{\|\nabla_0\|^2}{2}$ for all $t < t_1$. If $t_1 > \frac{4Lf(z_0)}{\alpha}$, then this contradicts the lower bound derived above.

Since $\mathcal{A}$ is a descent function, we must have $f(z_{t_1}) \leqslant f(z_0)$. Letting t_k be the first time $\|\nabla_t\|^2 \leqslant 2^{-k}\|\nabla_0\|^2$, we can show the statement by induction. □

Note in fact that we don't need strong convexity of f everywhere, but just enough to apply the lower bound in Lemma 2.6.1. If we are using a descent method, it is enough to have strong convexity in a level set containing the initial point. For this purpose, we extend Lemma 2.6.3 to the vector setting.

Lemma 3.1.9. *Let z_* be the optimizer of $f : V \to \mathbb{R}$, and assume f is α-strongly convex on $\{z \in V \mid \|z - z_*\| \leqslant R\}$. Then it is strongly convex on the level set*

$$L := \{z \in V \mid f(z) - f(z_*) \leqslant \alpha R^2/2\}.$$

Further $\|\nabla f(z)\| < \alpha R$ implies $z \in L$.

The proof is a straightforward application of Lemma 2.6.3 on univariate restrictions of f.

3.1.3 Putting it Together for Matrix Scaling

As we've shown above, Sinkhorn scaling is a L-descent method for $\log f_A$ where f_A is the Kempf-Ness function. So in order to get fast convergence results, we can apply the strong convexity analysis derived above.

Definition 27. Matrix tuple $A \in \mathrm{Mat}(d,n)^K$ is α-strongly convex if f_A is α-strongly convex in norm $\|\cdot\|_t$ at the origin.

Theorem 3.1.10 (Kwok, Lau, and Ramachandran (2019), Ramachandran (2021))**.** *For $A \in \mathrm{Mat}(d,n)$ with $s(A) = 1$, if A is ε-doubly balanced and $\alpha \gtrsim \varepsilon \log d$-strongly convex, then*

1. *The optimizer of f_A exists and satisfies*

$$\max_{i\in[d]} |(X_*)_i| + \max_{j\in[n]} |(Y_*)_j| \lesssim \frac{\varepsilon \log d}{\alpha}.$$

2. *The optimum value is lower bounded*

$$f^* = f_A(X_*, Y_*) \geqslant 1 - \frac{\varepsilon^2}{\alpha}.$$

3. *The doubly balanced scaling* $A_* := e^{X_*/2} A e^{Y_*/2}$ *is* $\Omega(\alpha)$*-strongly convex. Furthermore,* $e^X A e^Y$ *is* $\Omega(\alpha)$*-strongly convex for any* $\|(X - X_*, Y - Y_*)\|_{\mathfrak{t}}^2 \leq \frac{1}{n}$.

As a consequence, we have the desired fast convergence of Sinkhorn scaling.

Theorem 3.1.11. *Under the same conditions as Theorem 3.1.10, Sinkhorn scaling outputs* δ*-doubly balanced iterate* A_T *for some* $T \lesssim \frac{1}{\alpha}(n + \log(n\varepsilon/\delta))$.

Proof. We would like to apply Theorem 3.1.8 to show fast convergence. By part (3) of the above, we have that f_A is $\Omega(\alpha)$-strongly convex for a $\|\cdot\|_{\mathfrak{t}}$-norm ball of size $\frac{1}{\sqrt{n}}$ around (X_*, Y_*). Lemma 3.1.9 shows that any iterate satisfying $\|\nabla f_A(X_t, Y_t)\|_{\mathfrak{t}}^2 \leq \frac{\alpha^2}{n}$ will be in a strongly convex level set of f_A, after which point Sinkhorn scaling will have fast convergence by Theorem 3.1.8.

So let T be the first time the gradient condition holds. Since Sinkhorn scaling makes progress on $\log f$ instead of f, we need to translate this condition. Because Sinkhorn scaling is a descent method on f, we have

$$1 - \frac{\varepsilon^2}{\alpha} \leq f^* \leq f_A(X_t, Y_t) \leq f_A(0) = 1,$$

where the first inequality is by part (2) of the previous theorem, and the final equality is by definition $s(A) = 1$. Therefore, the gradient of f and $\log f$ are similar, and we can apply the analysis of Theorem 3.1.4.

$$\log f_A(X_T, Y_T) - \log f_A(0, 0) \lesssim -T\frac{\alpha^2}{n}$$

by the assumption that the gradient is large until step T. Combining with the lower bound and rearranging gives

$$T \lesssim \frac{n\varepsilon^2}{\alpha^3} \leq \frac{n}{\alpha}.$$

At this point, we can use Theorem 3.1.8 to show fast convergence, i.e. that $\|\nabla_t\|_{\mathfrak{t}}^2$ halves every $O(1/\alpha)$ iterations. In particular, if $\|\log \nabla f_{A_t}\|_{\mathfrak{t}}^2 \leq \frac{\delta^2}{n}$, then A_t is δ-doubly balanced, so the result follows. □

In the next section, we will vastly generalize these results to the setting of non-commutative scaling problems. We will be able to lift these results by using geodesic convexity of the Kempf-Ness function and tools from convex optimization.

3.2 Tensor Scaling

The tensor scaling problem is a generalization of matrix scaling where the inputs are higher order tensors, and scalings come from the non-commutative group of unit determinant matrices.

Definition 28. For $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$ where $\mathbb{C}^D = \mathbb{C}^{d_1} \otimes ... \otimes \mathbb{C}^{d_m}$, its size is defined

$$s(x) := \sum_{k=1}^{K} \|x_k\|_2^2,$$

where the norm is the standard Euclidean norm on $\mathbb{C}^D$.

For this tuple, we will also define an operator in $\mathrm{Mat}(D)$ by

$$\rho_x := \sum_{k=1}^{K} x_k x_k^*.$$

Note that it is positive semidefinite and $Tr[\rho_x] = s(x)$.

Definition 29 (Partial trace)**.** Let $\rho \in \mathrm{Mat}(D)$ be an operator on $\mathbb{C}^D = \mathbb{C}^{d_1} \otimes ... \otimes \mathbb{C}^{d_m}$, and $J \subseteq [m]$. Define the *partial trace* $\rho^{(J)}$ as the element of $\mathrm{Mat}(d_J)$, $\mathbb{C}^{d_J} := \otimes_{a \in J} \mathbb{C}^{d_a}$, that satisfies the following property.

$$\langle \rho^{(J)}, H \rangle = \langle \rho, H \otimes I_{\overline{J}} \rangle \tag{3.2.1}$$

for any $H \in \mathrm{Mat}(d_J)$, where $I_{\overline{J}}$ is the identity on $\otimes_{a \notin J} \mathbb{C}^{d_a}$, and we have used the standard inner product $\langle A, B \rangle := Tr[A^* B]$. This property uniquely determines $\rho^{(J)}$. We will omit brackets for small sets and write e.g. $\rho^{(a)}$ and $\rho^{(abc)}$ for $J = \{a\}$ and $J = \{a, b, c\}$, respectively.

Definition 30. Tuple $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$ is ε-balanced if

$$\forall a \in [m]: \quad s(x) \frac{1-\varepsilon}{d_a} I_a \preceq \rho_x^{(a)} \preceq s(x) \frac{1+\varepsilon}{d_a} I_a,$$

and x is balanced if $\varepsilon = 0$.

Definition 31 (Tensor Scaling Problem)**.** For tuple $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$ where $\mathbb{C}^D = \mathbb{C}^{d_1} \otimes ... \otimes \mathbb{C}^{d_m}$, let $G := SL(d_1) \times ... \times SL(d_m)$ act by tensor product

$$((g_1, ..., g_m) \in G) \cdot x := \{(g_1 \otimes ... \otimes g_m) x_1, ..., (g_1 \otimes ... \otimes g_m) x_K\}.$$

By abuse of notation, we will also use G to refer to the embedded subgroup $\{g_1 \otimes ... \otimes g_m \mid (g_1, ..., g_m) \in SL(d_1) \times ... \times SL(d_m)\} \subseteq GL(D)$.

The input to the tensor scaling problem is a tuple x.

1. **Success**: Output scalings $g = (g_1, ..., g_m)$ such that $g \cdot x$ is balanced.
2. **Failure**: Proof that no G-scaling of x is balanced.

Similar to Proposition 3.1.1 for matrix scaling, there is a geodesically convex formulation for tensor scaling.

Definition 32. For tuple $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$, the Kempf-Ness function $\widetilde{f_x} : G \to \mathbb{R}$ is defined as

$$\widetilde{f_x}(g \in G) := s(g \cdot x) = \sum_{k=1}^{K} \|(g_1 \otimes ... \otimes g_m) \cdot x_k\|_2^2.$$

The family of functions $\{\widetilde{f_{x \in (\mathbb{C}^D)^K}}\}$ satisfies the following equivariance property. We will use it repeatedly in order to simplify calculations.

Fact 3.2.1 (Equivariance)**.** *For tuple $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$, the Kempf-Ness function in Definition 32 satisfies the following relation for $g \in G$:*

$$\widetilde{f_x}(g) = s(g \cdot x) = \widetilde{f_{g \cdot x}}(e),$$

where $e = (I_1, ..., I_m)$ is the identity element of G.

These functions also inherit unitary invariance from the Euclidean norm.

Definition 33. For $G := SL(d_1) \times ... \times SL(d_m)$, let $K := G \cap SU(D) = SU(d_1) \times ... \times SU(d_m)$ be a maximal compact subgroup of unit determinant unitary matrices in G, and $P := G \cap \mathrm{PD}(D) = \mathrm{SPD}(d_1) \times ... \mathrm{SPD}(d_m)$ be the set of unit determinant positive definite matrices in G. Again U is a subgroup of $U(D)$ and P is a subgroup of $P(D)$ by the same embedding used for G.

By Theorem 2.5.2 on the polar decomposition, we have that $G = KP$.

Definition 34. For G, K, P given above, we define vector space

$$\mathfrak{g} := \{(Z_1, ..., Z_m) \in \mathrm{Mat}(d_1) \times ... \times \mathrm{Mat}(d_m) \mid \forall a \in [m] : Z_a^* = Z_a, Tr[Z_a] = 0\}.$$

By abuse of notation, we will also use $\mathfrak{g}$ to refer to the following embedding:

$$(Z_1, ..., Z_m) \to \sum_{a \in [m]} I_{\overline{a}} \otimes Z_a \in \mathrm{Mat}(D).$$

Note that $\sqrt{-1}\mathfrak{g}$ is the set of anti-Hermitian traceless matrices and is the Lie algebra of K. Also the Spectral Theorem shows that $P = e^{\mathfrak{g}}$, where this is the standard exponential map on Hermitian matrices.

Lemma 3.2.2. *The Kempf-Ness function is invariant under K. Therefore, by the decomposition $G = KP$, it descends to a function on P:*

$$f_x(p \in P) := s(p^{1/2} \cdot x) = \langle \rho_x, p \rangle.$$

Proof. Since $K \subseteq U(D)$, it does not change the norm. Therefore we can let f_x be a function on K orbits and choose the positive definite element as a representative. Explicitly, let $g = kp = ke^Z$ for $k \in K, Z \in \mathfrak{g}$.

$$\widetilde{f}_x(g) = \sum_{k=1}^{K} \|ke^Z \cdot x_k\|_2^2 = \sum_{k=1}^{K} \langle x_k x_k^*, e^Z k^* k e^Z \rangle = \langle \rho_x, e^{2Z} \rangle.$$

Since $(e^{2Z})^{1/2} = e^Z$ is the polar part of g, f_x is well-defined. □

We distinguish between the two functions $f, \widetilde{f}$ for the following reason: the domain of f is contained in the domain of $\widetilde{f}$, but the two functions are normalized differently, which leads to the following discrepancy on their common domain:

$$\widetilde{f}(e^Z) = f(e^{2Z}).$$

Therefore, we will tend to use f exclusively for positive definite elements to avoid confusion.

Note importantly that x is ε-balanced iff $K \cdot x$ is, so we do not lose any information by restricting to P. We next show that the Kempf-Ness function gives an optimization formulation for the tensor scaling problem (Definition 31).

Lemma 3.2.3. *For tuple $x := \{x_1, ..., x_K\} \in (\mathbb{C}^D)^K$, if $g \in G$ is the global minimizer of $\widetilde{f}_x$, then $g \cdot x$ is a balanced tensor.*

Proof. Let $g = k \cdot e^{Z_*}$ be the polar decomposition for $k \in K, Z_* \in \mathfrak{g}$. By Lemma 3.2.2, g is the global optimizer of $\widetilde{f_x}$ iff e^{2Z_*} is the global optimizer of f_x. To show the lemma, we examine the optimality conditions of $y := g \cdot x$ under small perturbations. For any $Z \in \mathfrak{g}$

$$\begin{aligned} \partial_{\eta=0} f_y(e^{\eta Z}) &= \partial_{\eta=0} \Big\langle \rho_y, \bigotimes_{a\in[m]} e^{\eta Z_a} \Big\rangle = \sum_{a\in[m]} \langle \rho_y, I_{\overline{a}} \otimes Z_a \rangle \\ &= \sum_{a\in[m]} \langle \rho_y^{(a)}, Z_a \rangle = \sum_{a\in[m]} \Big\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_a, Z_a \Big\rangle, \end{aligned} \tag{3.2.2}$$

where the second step was by the product rule, the third step was by Definition 29 on marginals, and the final step was by the constraint that $Tr[Z_a] = 0$.

By local optimality, we must have $\partial_{\eta=0} f_y(e^{\eta Z}) \geq 0$ for every $Z \in \mathfrak{g}$. Since the Frobenius inner product is non-degenerate, this happens iff $\rho_y^{(a)} = \frac{s(y)}{d_a} I_a$ for all $a \in [m]$, which is equivalent to y being a balanced tensor. Therefore the lemma follows by Fact 3.2.1 and the definition $y = g \cdot x$. □

The above calculation showed that local optimality conditions imply tensor balance. To show the converse, we will exploit the structure of P and the appropriate notion of convexity of f.

Lemma 3.2.4. *For $x \in (\mathbb{C}^D)^K$, g is the global minimum of $\widetilde{f_x}$ iff g^*g is a local minimum of f_x iff $g \cdot x$ is a balanced tensor.*

Proof. The first equivalence follows by Lemma 3.2.2. To show that local optimality implies $g \cdot x$ is balanced, we will use geodesic convexity of f_x.

Letting $y := g \cdot x$ be the balanced scaling, Equation (3.2.2) shows that for any $Z \in \mathfrak{g}$

$$\partial_{\eta=0} f_y(e^{\eta Z}) = \sum_{a\in[m]} \Big\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_a, Z_a \Big\rangle = 0.$$

We will show that the the family of functions $\{f_x \mid x \in (\mathbb{C}^D)^K\}$ are all geodesically convex on P. By Definition 19, it is enough to show that the univariate restriction $\eta \to f_x(e^{\eta Z})$ is convex at the origin for every $Z \in \mathfrak{g}$.

$$\partial^2_{\eta=0} f_x(e^{\eta Z}) = \partial^2_{\eta=0} \Big\langle \rho_x, \bigotimes_{a\in[m]} e^{\eta Z_a} \Big\rangle = \Big\langle \rho_x, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a \Big)^2 \Big\rangle, \tag{3.2.3}$$

and this inner product is non-negative as both terms are positive semi-definite.

Therefore, we can use convexity properties to show $s(y)$ is the global minimum. Specifically, consider the univariate restriction $h(\eta) := f_y(e^{\eta Z})$ for $Z \in \mathfrak{g}$.

$$f_y(e^Z) - s(y) = h(1) - h(0) \geqslant h'(0)(1-0) = 0,$$

where we used Definition 16 for the lower bound, and the local optmality of y for the final equality. As $Z \in \mathfrak{g}$ was arbitrary, $s(y)$ is the minimum value of f_y, and this is also the minimum value of f_x by Fact 3.2.1. □

This formulation will allow us to use tools from convex optimization to give strong bounds for the tensor scaling problem.

3.2.1 Geodesic Gradient

To use ideas from convex optimization, we would like to define the appropriate notion of gradient for the Kempf-Ness function. For an inner product space $(V, \langle \cdot, \cdot \rangle)$ the gradient ∇h of function $h : V \to \mathbb{R}$ represents the infinitesimal rate of change with respect to linear perturbations:

$$\langle \nabla h(u), x \in V \rangle = \partial_{\eta=0} h(u + \eta x).$$

The Kempf-Ness function in Definition 32 is not defined on a vector space, so the Euclidean gradient is not well-defined. But we can use the reduction in Lemma 3.2.2 and the geometry of P to define the analogous notion of geodesic gradients. Since $\mathfrak{g}$ is a vector space and $P = e^{\mathfrak{g}}$, the following gives a natural condition to define the geodesic gradient at the identity $I_D \in G$:

$$\langle \nabla f_x(I_D), Z \rangle = \partial_{\eta=0} f_x(e^{\eta Z}).$$

To lift this properly to a geodesic gradient at all points, we can use Fact 3.2.1.

Definition 35. [$\mathfrak{g}$ Inner Product] For elements $Z, Z' \in \mathfrak{g}$, we define inner product

$$\langle Z, Z' \rangle_{\mathfrak{g}} := \sum_{a \in [m]} \frac{1}{d_a} \langle Z_a, Z'_a \rangle,$$

where the right hand side uses $\langle X, Y \rangle := Tr[X^* Y]$.

The induces norm $\|Z\|_{\mathfrak{g}}^2 = \langle Z, Z\rangle_{\mathfrak{g}}$ which is in fact equivalent to the standard Frobenius norm in $\mathrm{Mat}(D)$ by the embedding

$$\left\langle \sum_{a\in[m]} I_{\overline{a}} \otimes Z_a, \sum_{b\in[m]} I_{\overline{b}} \otimes Z'_b \right\rangle = \sum_{a\in[m]} \|I_{\overline{a}}\|_F^2 \langle Z_a, Z'_a\rangle + \sum_{a\neq b\in[m]} \langle I_{\overline{a}} \otimes Z_a, I_{\overline{b}} \otimes Z'_b$$
$$= \sum_{a\in[m]} \frac{D}{d_a}\langle Z_a, Z'_a\rangle = D\langle Z, Z'\rangle_g$$

where the $a \neq b$ terms vanish as $Tr[Z_a] = Tr[Z'_b] = 0$.

Proposition 3.2.5. *For $x \in (\mathbb{C}^D)^K$, the geodesic gradient of f_x at point $pinP$ is the following element of $\mathfrak{g}$:*

$$\nabla f_x(p) = \left\{ d_a \cdot \rho^{(a)}_{p^{1/2}\cdot x} - s(p^{1/2}\cdot x) I_a \right\}_{a\in[m]}.$$

We will often use shorthand $\nabla_x := \nabla f_x(I_D)$, and $\{\nabla_x^{(a)}\}_{a\in[m]}$ for the marginals.

Proof. By Fact 3.2.1 and Definition 15, if $y = p^{1/2}\cdot x$ then

$$\partial_{\eta=0} f_x(p^{1/2} e^{\eta Z} p^{1/2}) = \partial_{\eta=0} f_y(e^{\eta Z})$$

for any $Z \in \mathfrak{g}$. So it is enough to define the geodesic gradient at the identity. By the dual definition above, for $Z \in \mathfrak{g}$ we calculate

$$\partial_{\eta=0} f_y(e^{\eta Z}) = \sum_{a\in[m]} \left\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_a, Z_a \right\rangle = \sum_{a\in[m]} \frac{1}{d_a}\langle d_a \rho_y^{(a)} - s(y) I_a, Z_a\rangle,$$

where the first step was by Equation (3.2.2). The final expression is exactly $\langle \nabla_y, Z\rangle_{\mathfrak{g}}$ by Definition 35. □

3.2.2 Strong Convexity

Convergence results for convex optimization can generally be strengthened under the assumption that the function is strongly convex. Here we define the correct notion of strong geodesic convexity which will allow us to give strong convergence results in the non-commutative optimization setting. Since we have already chosen the $\mathfrak{g}$-inner product, our notion of strong convexity is clear.

Definition 36. $x \in (\mathbb{C}^D)^K$ is α-strongly convex iff f_x is α-geodesically strong convex at the identity:

$$\forall Z \in \mathfrak{g}: \quad \partial^2_{\eta=0} f_x(e^{\eta Z}) \geqslant \alpha \|Z\|^2_{\mathfrak{g}} = \alpha \sum_{a\in[m]} \frac{\|Z_a\|_F^2}{d_a}.$$

In order to show strong convexity of a particular input x, we can further examine the second-order derivatives by expanding the terms of Equation (3.2.3).

$$\begin{aligned} \partial^2_{\eta=0} f_x(e^{\eta Z}) &= \left\langle \rho_x, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a \Big)^2 \right\rangle \\ &= \sum_{a\in[m]} \langle \rho_x^{(a)}, Z_a^2 \rangle + \sum_{a\neq b\in[m]} \langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle. \end{aligned} \tag{3.2.4}$$

The cases of interest to us will be when x is a nearly balanced tensor. Our plan will then be to show that the diagonal terms $\langle \rho_x^{(a)}, Z_a^2 \rangle$ are large. To show strong convexity, we will need a bound on the off-diagonal terms for which we define the following spectral condition.

Definition 37. For tensor $x \in (\mathbb{C}^D)^K$ and $a \neq b \in [m]$, x satisfies the λ-spectral condition in the (ab) part if

$$\sup_{Z=(Z_a,Z_b,0)\in\mathfrak{g}} \frac{\langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle}{\|Z_a\|_F \|Z_b\|_F} \leqslant \frac{\lambda}{\sqrt{d_a d_b}},$$

and x satisfies the λ-spectral condition if it does so for every part $a \neq b \in [m]$.

This condition originated in the operator scaling analysis of Kwok, Lau, and Ramachandran (2019). For this simpler case, when $m = 2$, the condition is actually a spectral upper bound for a particular linear map $\Phi : \mathrm{Mat}(d_b) \to \mathrm{Mat}(d_a)$. This interpretation is quite valuable for the proving robustness of strong convexity, e.g. Theorem 3.2.11.

We can use this condition to show strong convexity.

Proposition 3.2.6. *If tensor $x \in (\mathbb{C}^D)^K$ is ε-balanced and satisfies the λ-spectral condition, then x is α-strongly convex for $\alpha \geqslant s(x)(1-\varepsilon) - (m-1)\lambda$.*

Proof. To show that the diagonal terms in Equation (3.2.4) are large, we use the fact that x is ε-balanced so

$$\langle \rho_x^{(a)}, Z_a^2 \rangle \geqslant s(x) \frac{1-\varepsilon}{d_a} \langle I_a, Z_a^2 \rangle = s(x) \frac{1-\varepsilon}{d_a} \|Z_a\|_F^2,$$

where the first step was by Definition 30 and the fact that $Z_a^2 \succeq 0$.

The off-diagonal terms in Equation (3.2.4) are bounded by Definition 37, so for any $Z \in \mathfrak{g}$ we can bound the second order derivative

$$\begin{aligned}
\partial_{\eta=0}^2 f_x(e^{\eta Z}) &= \sum_{a \in [m]} \langle \rho_x^{(a)}, Z_a^2 \rangle + \sum_{a \neq b \in [m]} \langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle \\
&\geqslant \sum_{a \in [m]} \frac{s(x)(1-\varepsilon)}{d_a} \|Z_a\|_F^2 - \sum_{a \neq b \in [m]} \frac{\lambda}{\sqrt{d_a d_b}} \|Z_a\|_F \|Z_b\|_F \\
&= \left(s(x)(1-\varepsilon) + \lambda \right) \|Z\|_{\mathfrak{g}}^2 - \lambda \left(\sum_{a \in [m]} \frac{\|Z_a\|_F}{\sqrt{d_a}} \right)^2 \\
&\geqslant \left((s(x)(1-\varepsilon) + \lambda \right) \|Z\|_{\mathfrak{g}}^2 - m\lambda \|Z\|_{\mathfrak{g}}^2,
\end{aligned}$$

where the third and fourth step used Definition 35 on $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$, and the last step was by Cauchy-Schwarz. As $Z \in \mathfrak{g}$ was arbitrary, the statement follows by combining terms. □

3.2.3 Strong Convergence Bound

In this section, we will generalize Theorem 3.1.10 to the geodesic setting. Specifically, we will need to show that f_x is strongly convex on all univariate restrictions for some neighborhood of I_D. This gives a strong distance bound for the solution of the tensor scaling problem. We will need the following version of operator norm for our bounds.

Definition 38. For $Z \in \mathfrak{g}$, the operator norm is defined

$$\|Z\|_{op} := \sum_{a \in [m]} \|Z_a\|_{op},$$

where $\|Z_a\|_{op}$ is the standard Euclidean operator norm on $\mathrm{Mat}(d_a)$. Note that this is exactly the Euclidean operator norm of Z with respect to the embedding $Z \to \sum_{a \in [m]} I_{\overline{a}} \otimes Z_a \in \mathrm{Mat}(D)$.

Lemma 3.2.7. *For vector space $\mathfrak{g}$, the two norms are equivalent up to the following factors:*

$$\|Z\|_{\mathfrak{g}}^2 \leqslant \|Z\|_{op}^2 \leqslant \left(\sum_{a\in[m]} d_a\right) \|Z\|_{\mathfrak{g}}^2.$$

Proof. By Definition 35 of $\|\cdot\|_{\mathfrak{g}}$ and Definition 38 of $\|\cdot\|_{op}$

$$\|Z\|_{\mathfrak{g}}^2 = \sum_{a\in[m]} \frac{\|Z_a\|_F^2}{d_a} \leqslant \sum_{a\in[m]} \frac{d_a\|Z_a\|_{op}^2}{d_a} \leqslant \left(\sum_{a\in[m]} \|Z_a\|_{op}\right)^2 = \|Z\|_{op}^2,$$

where the second step was by standard equivalence of operator and Frobenius norm. To show the reverse equivalence, we calculate

$$\|Z\|_{op}^2 = \left(\sum_{a\in[m]} \|Z_a\|_{op}\right)^2 \leqslant \left(\sum_{a\in[m]} \|Z_a\|_F\right)^2 \leqslant \left(\sum_{a\in[m]} d_a\right) \|Z\|_{\mathfrak{g}}^2,$$

where we used $\|Z_a\|_{op} \leqslant \|Z_a\|_F$ and the final step was by Cauchy-Schwarz. □

Lemma 3.2.8. *If $x \in (\mathbb{C}^D)^K$ is α-strongly convex, then for any $Z \in \mathfrak{g}$ the univariate restriction $h(\eta) := f_x(e^{\eta Z})$ is $\alpha \exp(-\|\eta Z\|_{op})$-strongly convex at η.*

Proof. Consider univariate restriction $\eta \to f_x(e^{\eta Z})$ for some $\|Z\|_{\mathfrak{g}} = 1$. Then Equation (3.2.3) shows that the second derivative is

$$\left\langle \rho_{e^{\eta Z/2}x}, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a\Big)^2 \right\rangle = \left\langle e^{\eta Z/2}\rho_x e^{\eta Z/2}, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a\Big)^2 \right\rangle$$
$$= \left\langle \rho_x, e^{\eta Z}\Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a\Big)^2 \right\rangle \geqslant e^{-\|\eta Z\|_{op}} \left\langle \rho_x, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a\Big)^2 \right\rangle,$$

where the second step was from Equation (3.2.3), the third was by Fact 3.2.1, the fourth step was by the fact that e^Z commutes with the term in paranthesis, and therefore the fifth step was by a spectral lower bound. This last term is $\geqslant \alpha e^{-\|\eta Z\|_{op}}$ by strong convexity of x. □

Theorem 3.2.9. *If $x \in (\mathbb{C}^D)^K$ is α-strongly convex for*

$$\alpha \geqslant e^2 \sqrt{\sum_{a\in[m]} d_a} \|\nabla_x\|_{\mathfrak{g}},$$

then there exists a balanced scaling $e^{Z_*} \cdot x$ such that

$$\|Z_*\|_{\mathfrak{g}} \leqslant \frac{e\|\nabla_x\|_{\mathfrak{g}}}{\alpha}.$$

Furthermore, the size of this scaling can be lower bounded

$$s(x_*) \geqslant s(x) - \frac{e\|\nabla_x\|_{\mathfrak{g}}^2}{2\alpha}.$$

Proof. Let $R := e^2\|\nabla_x\|_{\mathfrak{g}}/\alpha$ be the desired radius bound, and consider geodesically convex set $B := \{e^Z \mid Z \in \mathfrak{g}, \|Z\|_{\mathfrak{g}} \leqslant R\}$. Since B is closed and bounded, the infimum of continuous function f_x over B exists and is attained at some point e^{Z_*}. If there are many such infimum, choose the one that minimizes $\|Z_*\|_{\mathfrak{g}}$. If we can show that e^{Z_*} is in fact strictly in the interior of B, then it is a local optimum, which by Lemma 3.2.3 shows that it is a global optimum and gives a balanced scaling.

The condition $Z_* \in B$ implies the following operator norm bound.

$$\|Z_*\|_{op}^2 \leqslant \left(\sum_{a\in[m]} d_a\right) \|Z_*\|_{\mathfrak{g}}^2 \leqslant \left(\sum_{a\in[m]} d_a\right) \frac{e^4\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha^2} \leqslant 1,$$

where the first step is by Lemma 3.2.7, and the final inequality is by the assumption on α.

Letting $T := \|Z_*\|_{\mathfrak{g}}$ and $Z := Z_*/T$, Lemma 3.2.8 shows that the univariate restriction $h(\eta) := f_x(e^{\eta Z})$ is $\frac{\alpha}{e}$-strongly convex for $\eta \in [0, T]$. By definition of Z_*, we must have $h'(T) \leqslant 0$ as otherwise we could choose some $e^{\eta Z_*}$ for $\eta < 1$ that is also an optimizer. So we can bound T as follows:

$$0 \geqslant h'(T) = h'(0) + \int_{\eta=0}^{T} h''(\eta) = \langle \nabla_x, Z\rangle + \int_{\eta=0}^{T} \partial_\eta^2 f_x(e^{\eta Z}) \geqslant -\|\nabla_x\|_{\mathfrak{g}} + T\frac{\alpha}{e},$$

where the second step was by the fundamental theorem of calculus, the third step was by Proposition 3.2.5 of geodesic gradient, and the final step was by Cauchy-Schwarz on $h'(0) = \langle \nabla_x, Z\rangle$ and the strong convexity of h. Rearranging terms,

we see that $T = \|Z_*\|_{\mathfrak{g}} \leqslant \frac{e\|\nabla_x\|_{\mathfrak{g}}}{\alpha} < R$, and so e^{Z_*} is strictly in the interior of B and therefore is the global optimum which satisfies the required radius bound.

The size lower bound now follows by standard strong convexity.

$$s(x_*) - s(x) = h(T) - h(0) \geqslant -\frac{e|h'(0)|^2}{\alpha} \geqslant -\frac{e\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha},$$

where the second step used Lemma 2.6.1, and the final step was by Cauchy-Schwarz on $h'(0) = \langle \nabla_x, Z \rangle_{\mathfrak{g}}$. □

3.2.4 Convergence of Algorithms

In this section, we will generalize the results of Theorem 3.1.11 to tensor scaling. We first give the appropriate generalization of Sinkhorn scaling.

Definition 39 (Flip-Flop Algorithm)**.** For input $x \in (\mathbb{C}^D)^K$, one iteration of the Flip-Flop algorithm chooses $a := \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_{\mathfrak{g}}$, and then normalizes this marginal

$$x \leftarrow I_{\overline{a}} \otimes (d_a \rho_x^{(a)})^{-1/2} x.$$

To solve the tensor scaling problem, we can apply a natural variant of this algorithm which stays within $G = SL(d_1) \times ... \times SL(d_m)$.

$$x \leftarrow I_{\overline{a}} \otimes \left(\frac{d_a \rho_x^{(a)}}{\det(d_a \rho_x^{(a)})^{1/d_a}} \right)^{-1/2} x.$$

Proposition 3.2.10. *For input $x \in (\mathbb{C}^D)^K$, let (a) be the largest marginal and x' be the output of one iteration of Flip-Flop (normalizing the (a) marginal). Then*

$$\log s(x') - \log s(x) \leqslant -\frac{1}{6} \min \left\{ \frac{\|\nabla_x^{(a)}\|_{\mathfrak{g}}^2}{s(x)^2}, \frac{1}{d_a} \right\}.$$

This can be rewritten in terms of the Kempf-Ness function as

$$\log f_x(e^Z) \leqslant \log f_x(e) - \frac{1}{6} \min \left\{ \frac{\|\nabla \log f_x(e)\|_{\mathfrak{g}}^2}{m}, \frac{1}{d_{\max}} \right\}$$

where e^Z is one iteration of the Flip-Flop algorithm and e is the identity element of G, and $\nabla \log f_x := \frac{\nabla_x}{s(x)}$ by abuse of notation.

Proof. This statement is shown in exactly the same way as Lemma 3.1.3, except that we consider eigenvalues of $\rho_x^{(a)}$ instead of row and column sums. □

In Theorem 3.2.9, we have proven a strong bound on the optimizer of f_x when the input $x \in (\mathbb{C}^D)^K$ is sufficiently strongly convex. In order to show fast convergence of algorithms we require the tensor scaling solution $x_* := e^{Z_*} \cdot x$ to be geodesically strongly convex, whereas Theorem 3.2.9 only proves strong convexity of the univariate restriction $\eta \to f_x(e^{\eta Z_*})$. In order to generalize our results from Theorem 3.1.11, we will need a stronger robustness result.

Theorem 3.2.11. *If $x \in (\mathbb{C}^D)^K$ is α-strongly convex, then for any perturbation $\delta \in \mathfrak{g}$ such that $\|\delta\|_{op} \leqslant \frac{1}{20}$, scaling $x' := e^{\delta} \cdot x$ is $\alpha - O(m\|\delta\|_{op})$-strongly convex.*

The full proof is given in Franks et al. (2021) and proceeds by showing that each block of the Hessian does not change too much under small perturbations.

This result allows us to conclude that if the initial point in Theorem 3.2.9 is α-strongly convex for large enough α, then the tensor scaling solution $x_* := e^{Z_*} \cdot x$ is also $\Omega(\alpha)$-strongly convex. Therefore we can generalize the arguments of Section 3.1.3 to the geodesic setting and show fast convergence of Flip-Flop. In fact, many of the lemmas go through verbatim by replacing Euclidean convexity with geodesic convexity. The specific condition we need for fast convergence is defined below.

Definition 40. Let $f : P \to \mathbb{R}$ be a geodesically convex function and assume for simplicity that the optimizer is at the identity. Then $Z \in \mathfrak{g}$ or $e^Z \in G$ is called α-strongly convergent if the univariate restriction $\eta \to f(e^{\eta Z})$ is $\alpha\|Z\|_{\mathfrak{g}}^2$-strongly convex for $\eta \in [0, 1]$. Equivalently, $\eta \to f(e^{\eta Z/\|Z\|_{\mathfrak{g}}})$ is α-strongly convex for $\eta \in [0, \|Z\|_{\mathfrak{g}}]$.

$B \subseteq \mathfrak{g}$ is α-strongly convergent if every $Z \in B$ is α-strongly convergent.

The above assumption is sufficient to derive the fast convergence properties of strongly convex functions.

Lemma 3.2.12. *Let $f : P \to \mathbb{R}$ be geodesically convex with optimizer at the identity. If $Z \in \mathfrak{g}$ is α-strongly convergent, then*

1. *(Function):* $f^* \geqslant f(e^Z) - \frac{\|\nabla f(e^Z)\|_{\mathfrak{g}}^2}{2\alpha}$.
2. *(Distance):* $\|Z\|_{\mathfrak{g}} \leqslant \frac{\|\nabla f(e^Z)\|_{\mathfrak{g}}}{\alpha}$.

Proof. Both statements follow simply from the univariate versions. Specifically consider $h(\eta) := f(e^{\eta Z})$ which is $\alpha\|Z\|_{\mathfrak{g}}^2$-strongly convex for $\eta \in [0,1]$. Then Lemma 2.6.1 gives

$$f^* = h(0) \geqslant h(1) - \frac{|h'(1)|^2}{2\alpha\|Z\|_{\mathfrak{g}}^2} \geqslant f(e^Z) - \frac{\|\nabla f(e^Z)\|_{\mathfrak{g}}^2}{2\alpha},$$

where in the last inequality we used Cauchy-Schwarz on $h'(1) = \langle \nabla f(e^Z), Z\rangle_{\mathfrak{g}}$.

To show the second statement, we use the fact that the optimizer is at the identity so $h'(0) = 0$.

$$-\|\nabla f(e^Z)\|_{\mathfrak{g}}\|Z\|_{\mathfrak{g}} \geqslant h'(1) = h'(0) + \int_{\eta=0}^{1} h''(\eta) \geqslant \alpha\|Z\|_{\mathfrak{g}}^2,$$

where the first inequality is again by Cauchy-Schwarz, and the final inequality is by strong convexity. The bound follows by rearranging. □

Using this notion, we can generalize Theorem 3.1.8 to the geodesic setting.

Lemma 3.2.13. *Let $f : P \to \mathbb{R}_+$ be a geodesically convex function with optimizer at the identity, and consider $\mathcal{A}$ an L-descent algorithm for $\log f$. If all the iterates Z_t are α-strongly convergent, and the initial point e^{Z_0} satisfies*

$$f(e^{Z_0})\|\nabla_0\|_{\mathfrak{g}}^2 \leqslant \alpha$$

then $T \lesssim \frac{L}{\alpha}\log\frac{1}{\delta}$ iterations suffice to produce

$$\|\nabla \log f(e^{Z_T})\|_{\mathfrak{g}}^2 \leqslant \delta^2\|\nabla \log f(e^{Z_0})\|_{\mathfrak{g}}^2.$$

Proof. For shorthand, let $\nabla_t := \nabla \log f(e^{Z_t})$, and let T be the first time $\|\nabla_T\|_{\mathfrak{g}}^2 \leqslant \frac{1}{2}\|\nabla_0\|_{\mathfrak{g}}^2$. Then by Lemma 3.2.12 (1) we have lower bound

$$f^* \geqslant f(e^{Z_0}) - \frac{\|\nabla f(e^{Z_0})\|_{\mathfrak{g}}^2}{2\alpha}.$$

To show a lower bound for $\log f$, we use $\nabla f = f \cdot \nabla \log f$ to show

$$\log f^* - \log f(e^{Z_0}) \geqslant \log\left(1 - f(e^{Z_0})\frac{\|\nabla_0\|_{\mathfrak{g}}^2}{2\alpha}\right) \geqslant -f(e^{Z_0})\frac{\|\nabla_0\|_{\mathfrak{g}}^2}{\alpha},$$

where in the last step we used the assumption on the initial gradient and Taylor approximation. Now by the descent property of $\mathcal{A}$, we get

$$\log f(e^{Z_T}) - \log f(e^{Z_0}) \leqslant \sum_{t<T} \log f(e^{Z_{t+1}}) - \log f(e^{Z_t}) \leqslant -T\frac{\|\nabla_0\|_{\mathfrak{g}}^2}{4L},$$

where in the last step we used the assumption that $\|\nabla_t\|_{\mathfrak{g}}^2 > \frac{1}{2}\|\nabla_0\|_{\mathfrak{g}}^2$ for all $t < T$. Combining with the lower bound and rearranging, we get

$$-f(e^{Z_0})\frac{\|\nabla_0\|_{\mathfrak{g}}^2}{\alpha} \leqslant -T\frac{\|\nabla_0\|_{\mathfrak{g}}^2}{4L} \implies T \leqslant f(e^{Z_0})\frac{4L}{\alpha}.$$

Letting T_k be the first time $\|\nabla_{T_k}\|_{\mathfrak{g}}^2 \leqslant 2^{-k}\|\nabla_0\|_{\mathfrak{g}}^2$ and continuing by induction, we get the result. □

Going back to Theorem 3.2.9, we can show by Theorem 3.2.11 that $x_* = e^{Z_*} \cdot x$ is strongly convex. To show that the iterates of Flip-Flop become strongly convergent, we use the level set strategy of Theorem 3.1.8.

Lemma 3.2.14. *Let $f : P \to \mathbb{R}$ be a geodesically convex function with optimizer at the identity. Assume further that the geodesic ball*

$$B := \{Z \in \mathfrak{g} \mid \|Z\|_{\mathfrak{g}} \leqslant R\}$$

is an α-strongly convergent zone. Then

1. *The level set $L := \{Z \in \mathfrak{g} \mid f(e^Z) - f^* \leqslant \alpha R^2/2\}$ is α-strongly convergent.*
2. *If $Z \in \mathfrak{g}$ is such that $\|\nabla f(e^Z)\|_{\mathfrak{g}} \leqslant \alpha R$, then $Z \in L$.*

Proof. Both statements follow from Lemma 2.6.3 applied to $h(\eta) := f(e^{\eta Z})$ for $\|Z\|_{\mathfrak{g}} = 1$ which is α-strongly convex for $\eta \in [0, R]$. □

With these tools, we can generalize Theorem 3.1.11 to the tensor setting.

Theorem 3.2.15. *Let $x \in (\mathbb{C}^D)^K$ have size $s(x) = 1$ and assume x is α-strongly convex for*

$$\alpha^2 \gtrsim \sqrt{\sum_{a\in[m]} d_a}\|\nabla_x\|_{\mathfrak{g}}.$$

By Theorem 3.2.9, there exists a balanced scaling $x_* := e^{Z_*} \cdot x$.

Then for any $\delta > 0$, $T \lesssim\lesssim \frac{m}{\alpha}\log\frac{1}{\delta}$ iterations of Flip-Flop suffice to produce $x_T := e^{Z_T} \cdot x_*$ such that

1. *(Distance):* $\|Z_T\|_{\mathfrak{g}}^2 \leqslant \left(\sum_{a\in[m]} d_a\right)^{-1} \delta^2$.

2. *(Function):* $\log f_x(e^{Z_T}) - \log f^* \lesssim \left(\sum_{a\in[m]} d_a\right)^{-1} \delta^2\alpha$.

Proof. We first apply Theorem 3.2.9 to show that e^{Z_*} is the optimizer of f_x and satisfies

$$\|Z_*\|_{\mathfrak{g}} \leqslant \frac{e\|\nabla_x\|_{\mathfrak{g}}}{\alpha}.$$

By Lemma 3.2.7, this gives bound

$$\|Z_*\|_{op}^2 \leqslant \left(\sum_{a\in[m]} d_a\right)\|Z_*\|_{\mathfrak{g}}^2 \leqslant \left(\sum_{a\in[m]} d_a\right)\frac{e\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha}.$$

By the assumption that α^2 is large enough, Theorem 3.2.11 implies that $x_* := e^{Z_*}\cdot x$ is a balanced tensor that is also $\frac{\alpha}{2}$-strongly convex.

In order to apply Lemma 3.2.13, we want to show that eventually all iterates of the Flip-Flop algorithm are $\Omega(\alpha)$-strongly convergent. By the above discussion, f_{x_*} is $\frac{\alpha}{2}$-geodesically convex at the origin, and therefore by Lemma 3.2.8 all points $\|Z\|_{op} \leqslant 1$ are $\frac{\alpha}{2e}$-strongly convergent. By Lemma 3.2.7, this means that the geodesic ball

$$B := \left\{ e^Z \in G \mid \|Z\|_{\mathfrak{g}}^2 \leqslant \left(\sum_{a\in[m]} d_a\right)^{-1} \right\}$$

is $\frac{\alpha}{2e}$-strongly convergent. Lemma 3.2.14 then shows

$$\|\nabla f_{x_*}(e^Z)\|_{\mathfrak{g}}^2 \leqslant \left(\sum_{a\in[m]} d_a\right)^{-1} \frac{\alpha^2}{4e^2}$$

is a sufficient condition for linear convergence of Flip-Flop. So let T be the first time this occurs. Note that the proof of Theorem 3.2.9 in fact shows that the initial point $Z_0 := -Z_*$ is such that $x = e^{Z_0}\cdot x_*$ and is $\frac{\alpha}{2}$-strongly convergent. Therefore Lemma 3.2.12 shows

$$s(x_*) \geqslant s(x) - \frac{\|\nabla f_{x_*}(e^{Z_0})\|_{\mathfrak{g}}^2}{\alpha}.$$

We can rewrite this in terms of $\log f_{x_*}$ as

$$\log f_{x_*}(I_D) - \log f_{x_*}(e^{Z_0}) \geqslant \log\left(1 - \frac{\|\nabla f_{x_*}(e^{Z_0})\|_{\mathfrak{g}}^2}{s(x)\alpha}\right) \geqslant -\frac{2\|\nabla f_{x_*}(e^{Z_0})\|_{\mathfrak{g}}^2}{\alpha},$$

where we used $s(x) = 1$ and $\|\nabla_x\|_{\mathfrak{g}} \ll \alpha$ for the Taylor approximation. Since Flip-Flop is a descent method, this gives a nearly tight bounds for all iterates

$$1 - \frac{\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha} \leqslant s(x_*) \leqslant s(x_t) \leqslant s(x) = 1.$$

By the fact that $\nabla f = f \cdot \nabla \log f$, this allows us to show that Flip-Flop makes progress with respect to ∇f. So for any $t < T$ we have

$$\log f_{x_*}(e^{Z_t}) - \log f_{x_*}(e^{Z_{t+1}}) \gtrsim \min\left\{\frac{\|\nabla f_{x_*}(e^{Z_t})\|_{\mathfrak{g}}^2}{m}, \frac{1}{d_{\max}}\right\} \gtrsim \frac{\alpha^2}{m\sum_{a\in[m]} d_a},$$

where in the final step we used the bound on the objective function and gradient. Combining this with the lower bound, we get

$$\frac{2\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha} \geqslant \log s(x) - \log s(x_*) \geqslant \log s(x_0) - \log s(x_T) \gtrsim \frac{T\alpha^2}{m\sum_{a\in[m]} d_a}.$$

$$\implies T \lesssim \left(\sum_{a\in[m]} d_a\right)\frac{m\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha^3}.$$

By the assumption on α^2, this is $O(m\alpha)$, which is usually negligible. In fact, even under the weaker assumption $\alpha \geqslant \sqrt{\sum_{a\in[m]} d_a}\|\nabla_x\|_{\mathfrak{g}}$, we can conclude that $T \leqslant \frac{O(m)}{\alpha}$, which is negligible for small δ.

After this point, every iterate $Z_{t>T}$ is $\Omega(\alpha)$-strongly convergent, so we can conclude that there is $t - T \lesssim \frac{m}{\alpha}\log\frac{1}{\delta}$ such that

$$\|\nabla f_{x_*}(e^{Z_t})\|_{\mathfrak{g}}^2 \leqslant \|\nabla \log f_{x_*}(e^{Z_t})\|_{\mathfrak{g}}^2 \leqslant \delta^2\|\nabla \log f_{x_*}(e^{Z_T})\|_{\mathfrak{g}}^2 \lesssim \delta^2\frac{\alpha^2}{\sum_{a\in[m]} d_a},$$

where the first step was by the bounds on $s(x_t)$ shown above, the second step was by Lemma 3.2.13, and the final step was by the gradient bound on Z_T derived above. The bounds on $\|Z_t\|_{\mathfrak{g}}$ and $f_{x_*}(e^{Z_t})$ then follow from Lemma 3.2.12. □

Remark 3.2.16. Note that the requirement on α is larger by a quadratic factor, because we need to show $x_* := e^{Z_*} \cdot x$ is still geodesically strongly convex. But even with this weaker assumption, the number of iterations before we reach $\Omega(\alpha)$-strongly convergent is only $\frac{O(m)}{\alpha}$. In our application in the next chapter, we will have $\alpha \approx 1$ strong convexity, so this will not make much difference. This quadratic loss is crucial in e.g. the application to the Paulsen problem.

4 Applications to statistics

This chapter is based on the work of Franks et al. (2021), which presents improved results in full details. So the reader can consult the more thorough treatment of Franks et al. (ibid.), we will attempt to maintain notational consistency with this work as far as possible

4.1 Statistical Background

4.1.1 Statistical Inference

Many problems in statistics relate to identifying an unknown distribution based on samples from that distribution. A statistical model is a set of assumptions which constrains the possible family of distributions $\mathcal{F}$ that we are dealing with. Given a model, the task of statistical inference is to extract some concrete information about the fixed unknown distribution $\mathcal{D} \in \mathcal{F}$. The quality of this estimate can be measured according to various metrics depending on the application requirements, and the theoretical goal is to give an upper bound on the number of samples required to give a good estimator.

Example 4.1.1. Bernoulli Estimation
Input: $X_1, ..., X_n \sim Ber(p)$ from a Bernoulli distribution with bias p.

Output: The sample mean $\widehat{p} := \frac{1}{n}\sum_{i=1} X_i$ is a natural high-quality estimator for the bias.

Example 4.1.2. Gaussian Estimation
Input: $X_1, ..., X_n \sim N(0, \Theta^{-1})$ from an unknown centered Gaussian distribution with positive definite precision matrix $\Theta \in \mathrm{Mat}(d)$. This is the inverse of the covariance matrix.
Output: The inverse sample covariance $\widehat{\Theta} := \left(\frac{1}{n}\sum_{i=1} X_i X_i^*\right)^{-1}$ is a natural high-quality estimator for the precision matrix.

4.1.2 Maximum Likelihood Estimation

In this section, we can derive the natural estimator used in Example 4.1.2 as the solution to an optimization problem.

Suppose we are given sample $X \in \mathbb{R}^d$ from some unknown centered Gaussian distribution, and we guess that the true distribution $N(0, \Theta^{-1})$. In this case, the probability density function (pdf) would be given by

$$f_\Theta(x \in \mathbb{R}^d) = \sqrt{\frac{\det(\Theta)}{(2\pi)^d}} \exp\left(-\frac{1}{2}x^*\Theta x\right).$$

So if $X^*\Theta X$ is very large, then the pdf states that this sample was very unlikely, and in some sense Θ is a bad guess. This intuition is formalized below.

Definition 41. Given samples $X_1, ..., X_n \in \mathbb{R}^d$ from some unknown distribution in $\mathcal{F} := \{\mathcal{D}_\omega\}_{\omega\in\Omega}$, the likelihood function of guess $\theta \in \Omega$ is

$$L(\theta) := f_\theta(X_1, ..., X_n),$$

where f_θ is the pdf of $\mathcal{D}_\theta$. We also often consider the log-likelihood function $\ell(\theta) := \log L(\theta)$ because independent terms become additive.

The maximum likelihood estimator (MLE) is the choice that maximizes the likelihood function

$$\widehat{\theta} := \arg\max_{\omega\in\Omega} L(\omega).$$

It turns out that the natural estimator given in Example 4.1.2 can be derived using this MLE perspective.

Proposition 4.1.3. *Given samples $X_1, ..., X_n \in \mathbb{R}^d$ from an unknown centered Gaussian distribution, the log-likelihood function for $\Theta \in \mathrm{PD}(d)$ is given by*

$$\ell(\Theta) = \frac{n}{2}\log\det(\Theta) - \frac{nd}{2}\log(2\pi) - \frac{1}{2}\left\langle \sum_{i=1}^{n} X_i X_i^*, \Theta \right\rangle,$$

and the inverse sample covariance $\widehat{\Theta} := \left(\frac{1}{n}\sum_{i=1} X_i X_i^*\right)^{-1}$ is the MLE.

Proof. By independence, the log-likelihood of Θ for samples $X_1, ..., X_n$ is just the sum of log-likelihoods for each individual sample. So we compute

$$\log f_\Theta(x) = \frac{1}{2}\log\det(\Theta) - \frac{d}{2}\log(2\pi) - \frac{1}{2}x^*\Theta x$$

$$\implies \ell(\Theta) = \frac{n}{2}\log\det(\Theta) - \frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} X_i^*\Theta X_i.$$

To find the MLE, we want to find the maximizer of $L(\theta)$. We will perform some simple transformations to simplify this optimization problem and more clearly show the similarity to our scaling perspective. We can drop the $\frac{nd}{2}\log(2\pi)$ term, since it does not depend on Θ, and renormalize to find the MLE as

$$\arg\min_{\Theta\in\mathrm{PD}(d)} F_X(\Theta) := -\frac{2}{n}\ell(\Theta) - \log 2\pi = \left\langle \frac{1}{n}\sum_{i=1}^{n} X_i X_i^*, \Theta \right\rangle - \log\det(\Theta),$$

where we have used the natural Frobenius (entrywise) inner product on $\mathrm{Mat}(d)$. We can find the optimizer by solving for critical points.

$$0 = \nabla_\Theta F_X(\Theta) = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^* - \Theta^{-1} \implies \widehat{\Theta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^*\right)^{-1}.$$

To show that this is in fact the MLE, we can compute the second order derivative and show it is the global minimizer. We leave this computation as an exercise. □

This problem and the optimization formulation for estimation enjoy a certain linear invariance.

Proposition 4.1.4. *For samples $X_1, ..., X_n \in \mathbb{R}^d$ and $A \in GL(d)$, let $Y_i := AX_i$. Then*

$$F_Y(\Theta) = F_X(A\Theta A^*) + \log\det(AA^*).$$

Therefore, $\widehat{\Theta}_Y$ is the MLE for Y iff $\widehat{\Theta}_X = A\widehat{\Theta}_Y A^*$ is the MLE for X.

Proof. This is a simple change of variable calculation.

$$\begin{aligned} F_Y(\Theta) &= \frac{1}{n}\sum_{i=1}^n \langle Y_i Y_i^*, \Theta\rangle - \log\det(\Theta) \\ &= \frac{1}{n}\sum_{i=1}^n \langle X_i X_i^*, A\Theta A^*\rangle - \log\det(A\Theta A^*) + \log\det(AA^*) \\ &= F_X(A\Theta A^*) + \log\det(AA^*). \end{aligned}$$

Since this $\log\det(AA^*)$ does not depend on Θ, we can drop this term without changing the optimizer, so the second statement follows. □

4.1.3 Quality of Gaussian Covariance Estimator

There are many ways to measure how good an estimator of covariance is. One natural measure of error is the following.

Definition 42. For $A, B \in \mathrm{PD}(d)$, we define relative Frobenius and operator error

$$d_F(A,B) = \|I_d - B^{-1/2}AB^{-1/2}\|_F, \qquad d_{op}(A,B) = \|I_d - B^{-1/2}AB^{-1/2}\|_{op}.$$

These measures are not symmetric, but note $d(A,B) = d(B^{-1/2}AB^{-1/2}, I_d)$. Intuitively, this gives a multiplicative form of error between A, B. For example

$$\sup_{v\in\mathbb{R}^d} \frac{\langle vv^*, I_d - B^{-1/2}AB^{-1/2}\rangle}{\|v\|_2^2} = \sup_{u\in\mathbb{R}^d} \frac{\langle u, Bu\rangle - \langle u, Au\rangle}{\langle u, Bu\rangle},$$

where the last line was a change of variable $v = B^{1/2}u$. Therefore $d_{op}(A,B) \leqslant \varepsilon$ implies a multiplicative approximation of the quadratic form

$$\forall u \in \mathbb{R}^d : \langle u, Au\rangle \in (1 \pm \varepsilon)\langle u, Bu\rangle.$$

This kind of approximation is common in the literature on Laplacian solvers and graph sparsification (e.g. Spielman and Teng (2014), Spielman and Srivastava (2011)).

This definition is also a natural measure of distance from the geodesic perspective we will consider. Recall from Definition 15 that the unique geodesic curve from $B \to A$ is defined

$$\gamma(t) := B^{1/2} \exp(tX) B^{1/2}$$

where $X := \log(B^{-1/2} A B^{1/2})$ so that $\gamma(0) = B, \gamma(1) = A$. If X is small, then up to constant factors, we can rewrite the error measures as

$$d(A, B) = \|I_d - \exp(X)\| \lesssim \|X\|,$$

where the last step was by Taylor approximation for small enough $\|X\|_F, \|X\|_{op}$ respectively. Since our results will rely on geodesic convex optimization, we will achieve strong bounds on geodesic distance, which then implies strong bounds on d_F, d_{op} by the above calculation.

4.1.4 Analysis of the MLE

In this section, we will give explicit sample complexity bounds for the MLE to be a high-quality estimator for the covariance on an unknown centered Gaussian distribution.

Theorem 4.1.5. *Let $X_1, ..., X_n \in \mathbb{R}^d$ be samples from Gaussian distribution $N(0, \Theta^{-1})$, and let $\widehat{\Theta}$ be the MLE for the precision matrix. If $\varepsilon \leq \frac{1}{10}$ and $n \geq \frac{d}{\varepsilon^2}$, then the MLE has the following error bounds with probability at least $1 - 2\exp(-\varepsilon^2 n/2)$:*

$$d_{op}(\widehat{\Theta}, \Theta) \lesssim \varepsilon, \qquad d_F(\widehat{\Theta}, \Theta)^2 \lesssim d\varepsilon^2.$$

We will use the linear invariance of the MLE and distance measure to reduce to the case when $\Theta = I_d$. The result will then follow from the following standard result on Gaussian matrix concentration.

Theorem 4.1.6 (Corollary 5.35 of Vershynin (2012))**.** *Let $G \in \mathrm{Mat}(d, n)$ be a random matrix with standard Gaussian entries, or equivalently standard Gaussian columns $g_1, ..., g_n \sim N(0, I_d)$. Then for any $t > 0$,*

$$\sqrt{n} - \sqrt{d} - t \leq \sigma_{\min}(G) \leq \sigma_{\max}(G) \leq \sqrt{n} + \sqrt{d} + t$$

with probability at least $1 - 2e^{-t^2/2}$:

Proof of Theorem 4.1.5. The distribution of $X \sim N(0, \Theta^{-1})$ is equivalent to the distribution of $\Theta^{-1/2}Y$ where $Y \sim N(0, I_d)$. By the discussion in Proposition 4.1.4, the MLE of X and Y are related as follows:

$$\widehat{\Theta}_X = \Theta^{-1/2}\widehat{\Theta}_Y\Theta^{-1/2}.$$

The error measures also satisfy a similar invariance

$$d(\widehat{\Theta}_X, \Theta) = d(\Theta^{-1/2}\widehat{\Theta}_X\Theta^{-1/2}, I_d) = d(\widehat{\Theta}_Y, I_d).$$

So to prove the theorem, it is enough to show the error bound in the case when $\Theta = I_d$. In this case, the sample covariance has spectrum concentrated close to one. Specifically, if $t = \varepsilon\sqrt{n}$ in Theorem 4.1.6, then with probability at least $1 - 2\exp(-\varepsilon^2 n/2)$, we have the bound

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^*\right) = \left(\frac{\sigma_{\min}(Y)}{\sqrt{n}}\right)^2 \geqslant \left(1 - \frac{\sqrt{d} + \varepsilon\sqrt{n}}{\sqrt{n}}\right)^2 \geqslant 1 - 5\varepsilon,$$

where in the last step we used the assumption that $n \geqslant d/\varepsilon^2$ and the Taylor approximation for $(1 - 2\varepsilon)^2$ for small ε. By the same calculation, we have an upper bound

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^*\right) \leqslant 1 + 5\varepsilon.$$

Therefore, when this event occurs we can bound

$$d_{op}(\widehat{\Theta}_Y, I_d) = \left\| \left(\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^*\right)^{-1} - I_d \right\|_{op} = \max\{|\lambda_{\max}^{-1} - 1|, |\lambda_{\min}^{-1} - 1|\} \leqslant 10\varepsilon,$$

where again in the last step we used the Taylor approximation for $\frac{1}{1\pm x}$ and the fact that $5\varepsilon \leqslant \frac{1}{2}$. Similarly, we can calculate

$$d_F(\widehat{\Theta}_Y, I_d)^2 = \left\| \left(\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^*\right)^{-1} - I_d \right\|_F^2 \leqslant d(10\varepsilon)^2.$$

□

This is in fact best possible error bound up to constant factors. In fact, the sample covariance is non-invertible for $n < d$ samples so in this case we cannot have any constant error estimator. The sample complexity requirement can be rewritten $nd \gtrsim d^2$, where the right hand side represents the degrees of freedom of the unknown precision matrix, and the left hand side is the information content of n samples of d-dimensional vectors. This intuition will generalize to sample complexity results on the matrix and tensor normal model.

4.2 Matrix and Tensor Normal Model

4.2.1 Setup

In the previous section, we saw tight results for Gaussian covariance estimation. In this section we will consider the case when our random data is in the form of a matrix or tensor. Explicitly $X \in \mathbb{R}^D$ where $\mathbb{R}^D := \mathbb{R}^{d_a} \otimes ... \otimes \mathbb{R}^{d_m}$ for some $m \geqslant 2$. The discussion after Theorem 4.1.5 shows that in this setting, $n \gtrsim D = \prod_{a \in [m]} d_a$ samples are required in order to get any reasonable estimator.

In order to bypass this sample lower bound, we can add constraints to the model. The tensor normal model is one such natural assumption where the precision matrix also respects the tensor structure.

Definition 43 (Matrix and Tensor Normal Model)**.** The tensor normal model with $m \geqslant 2$ and dimensions $d_1, ..., d_m$ is the family of centered Gaussian distributions $N(0, \Theta^{-1})$ where

$$\Theta = \Theta_1 \otimes ...\Theta_m$$

for some $\{\Theta_a \in \mathrm{PD}(d_a)\}_{a \in [m]}$. When $m = 2$, this is known as the matrix normal model.

Definition 44 (Covariance Estimation for Matrix and Tensor Normal Model)**. Input**: Samples $X_1, ..., X_n \sim N(0, \Theta^{-1})$ where $\Theta = \Theta_1 \otimes ...\Theta_m$.
Output: $\widehat{\Theta} := \widehat{\Theta}_1, ..., \widehat{\Theta}_m$ such that

$$\forall a \in [m] : d(\widehat{\Theta}_a, \Theta_a)$$

is small for $d = d_{op}, d_F$. A weaker requirement is for $d(\widehat{\Theta}, \Theta)$ to be small.

In the Gaussian model in Example 4.1.2, the inverse sample covariance was a natural estimator which had optimal error. In the tensor setting, this is not even a feasible solution as the sample covariance will almost surely not factorize into a

tensor product. Another natural guess would be the set of marginals of the tensor product.

$$\forall a \in [m] : \widehat{\Theta}_a := \left(Tr_{\overline{a}} \left[\frac{1}{n} \sum_{i=1}^{n} X_i X_i^* \right] \right)^{-1}.$$

By properties of Gaussian concentration, this estimator has very good error properties when the true covariance is I_D. But it does not enjoy the same linear invariance properties discussed in Proposition 4.1.4, and so in general we cannot derive strong error bounds with high probability.

We illustrate this for the case $\mathbb{R}^D = \mathbb{R}^d \otimes \mathbb{R}^n$ and $\Theta = \Theta_1 \otimes \Theta_2$. In this case, X is distributed as $\Theta^{-1/2} Y$ where $Y \sim N(0, I_D)$. Letting $G_i := \mathrm{Mat}(Y_i) \in \mathrm{Mat}(d, n)$ be a random matrix with standard Gaussian entries, the marginal is distributed as

$$Tr_2 \left[\sum_{i=1}^{n} X_i X_i^* \right] = Tr_2 \left[\sum_{i=1}^{n} \Theta^{-1/2} Y_i Y_i^* \Theta^{-1/2} \right] = \sum_{i=1}^{n} \Theta_1^{-1/2} G \Theta_2^{-1} G^* \Theta_1^{-1/2}.$$

Standard Gaussian concentration results show that the variation of this marginal is on the order of $\kappa(\Theta_2) := \|\Theta_2\|_{op} \|\Theta_2^{-1}\|_{op}$. So for general Θ, the error could be arbitrarily bad, and this will not be a good estimator in general.

On the other hand, the MLE is still well-defined for this problem, though it is more difficult to compute than Proposition 4.1.3.

Proposition 4.2.1. *For samples $X_1, ..., X_n \in \mathbb{R}^D$ from the tensor normal model, the MLE is given by the minimizer of the following function,*

$$F_X(\Theta_1, ..., \Theta_m) := \frac{1}{nD} \sum_{i=1}^{n} \langle X_i X_i^*, \otimes_{a \in [m]} \Theta_a \rangle - \frac{1}{D} \log \det(\otimes_{a \in [m]} \Theta_a) \tag{4.2.1}$$

$$= \left\langle \frac{1}{nD} \sum_{i=1}^{n} X_i X_i^*, \otimes_{a \in [m]} \Theta_a \right\rangle - \sum_{a \in [m]} \frac{1}{d_a} \log \det(\Theta_a). \tag{4.2.2}$$

over all $\{\Theta_a \in \mathrm{PD}(d_a)\}_{a \in [m]}$.

The above should look very familiar. In fact this is almost exactly the Kempf-Ness function for tensor scaling given in Definition 31. Therefore, by applying the strongly convex analysis of Theorem 3.2.9, we can derive strong error bounds for the MLE.

Theorem 4.2.2. *Let $X_1, ..., X_n \in \mathbb{R}^D$ be samples from the tensor normal model $\mathbb{R}^D := \mathbb{R}^{d_1} \otimes ... \mathbb{R}^{d_k}$ with $m \geqslant 2$ and distribution $N(0, \Theta^{-1})$ with $\Theta := \Theta_1 \otimes ... \otimes \Theta_m$. If $\varepsilon^2 \lesssim \left(m^2 \sum_{a\in[m]} d_a\right)^{-1}$, and $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, then the MLE $\widehat{\Theta} := \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_k$ satisfies*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim Dm\varepsilon^2$$

with probability at least $1 - k^2 \exp(-\Omega(d_{\min}))$.

The tensor normal model has $\sum_{a\in[m]} d_a^2$ degrees of freedom, so intuitively this result is only a small polynomial factor away from optimal. These same techniques also show similar results for each tensor factor. Also, in the Matrix Normal Model setting, we can apply the sharper analysis of Kwok, Lau, and Ramachandran (2019) to derive strong error bounds in the operator norm for each factor. These and other results are given in full detail in Franks et al. (2021).

As a further contribution, we also show that the MLE in this sample regime is efficiently computable. In fact, we give a rigorous analysis for a well-studied algorithm used in practice.

Theorem 4.2.3. *With the same conditions as Theorem 4.2.2, for any $\frac{1}{2} \geqslant \delta > 0$ the Flip-Flop algorithm in Definition 39 produces an estimator Θ_T such that*

$$d_F(\widehat{\Theta}_T, \widehat{\Theta})^2 \lesssim D\delta^2$$

in $T \lesssim \frac{m}{\alpha} \log \frac{md_{\max}}{\delta}$ iterations.

In the rest of this section, we will explicitly show the connection between the MLE and tensor scaling. Then we will use properties of Gaussian concentration to show that the input to the tensor normal model satisfies the conditions of Theorem 3.2.9 with high probability, which allows us to prove strong error bounds.

4.2.2 Reduction

To make the relation to scaling clearer, we first reduce to the set of determinant one matrices.

Lemma 4.2.4. *For samples $X_1, ..., X_n \in \mathbb{R}^D$ for the tensor normal model, let $\{\widehat{\Theta}_1, ..., \widehat{\Theta}_m\}$ be the minimizers of*

$$\inf f_X(\Theta) := \frac{1}{nD} \sum_{i=1}^{n} \langle X_i X_i^*, \Theta_1 \otimes ... \otimes \Theta_m \rangle,$$

over all choices $\{\Theta_1 \in \mathrm{SPD}(d_1), ..., \Theta_m \in \mathrm{SPD}(d_m)\}$ *of unit determinant positive definite matrices, and let* f^* *be the minimum value. Then the MLE is*

$$\widehat{\Theta} := e^{\lambda_*} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m,$$

where $\lambda_* := -\log f^*$ *and* $F_X(\widehat{\Theta}) = 1 + \log f^*$.

Proof. For any $\Theta \in \mathrm{PD}(d_1) \otimes ... \otimes \mathrm{PD}(d_m)$, we can decompose $\Theta = e^{\lambda} \cdot \Theta_1 \otimes ... \otimes \Theta_m$ where $\det(\Theta_a) = 1$. If we fix $\{\Theta_a\}_{a \in [m]}$, then this becomes

$$F_X(\Theta) = \frac{e^{\lambda}}{nD} \langle \sum_{i=1}^{n} X_i X_i^*, \otimes_{a \in [m]} \Theta_a \rangle - \frac{1}{D} \log \det(\Theta) = e^{\lambda} \nu - \lambda,$$

where we denote $\nu := \frac{1}{nD} \langle \sum_{i=1}^{n} X_i X_i^*, \otimes_{a \in [m]} \Theta_a \rangle$, and the other terms vanish by the determinant condition. The global minimum of this univariate function is calculated simply.

$$0 = \partial_\lambda (e^{\lambda} \nu - \lambda) = e^{\lambda} \nu - 1 \implies \lambda_* = -\log \nu, \qquad \partial_\lambda^2 (e^{\lambda} \nu - \lambda) = e^{\lambda} \nu \geq 0.$$

$$\implies \inf_{\lambda} F_X(e^{\lambda} \Theta) = e^{\lambda_*} \nu - \lambda_* = 1 + \log \nu.$$

Therefore in order to find the MLE, we can equivalently optimize

$$f_X(\Theta_1, ..., \Theta_m) := \frac{1}{nD} \langle \sum_{i=1}^{n} X_i X_i^*, \otimes_{a \in [m]} \Theta_a \rangle$$

for $\{\Theta_a \in \mathrm{SPD}(d_a)\}_{a \in [m]}$, and then choose the appropriate value of λ_*. □

This is exactly the Kempf-Ness function from the tensor scaling problem, and we reduce from F_X to f_x because we can derive stronger convexity properties in this SPD setting.

In order to apply Theorem 3.2.9, we would like to show that random inputs are nearly balanced and strongly convex. Just as in Theorem 4.1.5, we can reduce to the case where $\Theta = I_D$. Therefore, in the next two sections, we will show the conditions of Theorem 3.2.9 using properties of Gaussian concentration for $N(0, I_D)$ inputs.

4.2.3 Bounding the Gradient

We showed earlier that the marginals of ρ_X are have arbitrarily bad error for general Θ. On the other hand, when $\Theta = I_D$ we can use Gaussian concentration to show that these estimators concentrate well. This is equivalent to showing strong bounds on the gradient ∇_X.

Proposition 4.2.5. *For $X_1, ..., X_n \sim N(0, I_D)$, if $nD \geq \frac{d_{\max}^2}{\varepsilon^2}$, then the following bounds hold simultaneously with probability at least $1 - 2k \exp(-\frac{\varepsilon^2 nD}{d_{\max}})$:*

1. $\frac{1}{nD}|s(X) - nD| \lesssim \varepsilon$.
2. *X is an $O(\varepsilon)$-balanced tensor.*

Proof. For $x = \frac{1}{\sqrt{nD}}\{X_1, ..., X_n\}$ and any $a \in [m]$, by Proposition 3.2.5 we have

$$\nabla_x^{(a)} = d_a \rho_x^{(a)} - s(x) I_a.$$

In order to use standard concentration of Gaussian matrices, we can define $G_1, ..., G_n \in \mathrm{Mat}(d_a, D/d_a)$ as

$$G_i[j_a, (\boldsymbol{j})] := (X_i)_{j_1,...,j_a,...,j_m}$$

where $\boldsymbol{j} = \{j_b\}_{b \neq a}$ runs over all possible indices $j_b \in [d_b]$.

Then we can rewrite the gradient as

$$\nabla_x^{(a)} = \frac{d_a}{nD} \sum_{i=1}^{n} G_i G_i^* - \frac{1}{nD} Tr \left[\sum_{i=1}^{n} G_i G_i^* \right] I_a.$$

The spectrum can now be bounded using Gaussian concentration. Explicitly, choosing $t = \varepsilon \sqrt{\frac{nD}{d_{\max}}}$ in Theorem 4.1.6, we can bound the singular values of $G := [G_1, ..., G_n] \in \mathrm{Mat}(d_a, \frac{nD}{d_a})$.

$$\lambda_{\min}\left(\frac{d_a}{nD} GG^*\right) = \frac{d_a}{nD} \sigma_{\min}(G)^2 \geq \frac{d_a}{nD} \left(\sqrt{\frac{nD}{d_a}} - \sqrt{d_a} - \varepsilon \sqrt{\frac{nD}{d_{\max}}} \right)^2 \geq 1 - 5\varepsilon,$$

where in the last step we used $\varepsilon^2 nD \geq d_{\max}^2 \geq d_{\max} d_a$ and the assumption that $\varepsilon \leq \frac{1}{20}$. By the same calculation we get an upper bound

$$\lambda_{\max}\left(\frac{d_a}{nD} GG^*\right) \leq 1 + 5\varepsilon.$$

Putting these two together, we have

$$\|\nabla_x^{(a)}\|_{op} \leqslant \left|\lambda_{\max}\left(\frac{d_a}{nD}GG^*\right) - \lambda_{\min}\left(\frac{d_a}{nD}GG^*\right)\right| \leqslant 10\varepsilon.$$

To bound $s(X) = \sum_{i=1}^n \|X_i\|_2^2$, we can take the trace of the inequality derived for any marginal. We can in fact derive even stronger bounds on $s(X)$ using standard concentration of χ-square variables, but this will not be necessary for our application. □

4.2.4 Spectral Gap for Random Input

We will show strong convexity using the following theorem.

Proposition 4.2.6. *For $X_1, ..., X_n \sim N(0, I_D)$ from the tensor normal model, if $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, then X satisfies the λ-spectral condition in the ab part with*

$$\lambda \lesssim \frac{d_a + d_b}{\sqrt{nD}} \lesssim \varepsilon,$$

for all $a \neq b \in [m]$ simultaneously with probability at least $1 - k^2 \exp(-\Omega(d_{\min}))$.

The proof follows from a powerful theorem by Pisier.

Theorem 4.2.7 (Pisier (2012, 2014))**.** *Let $G_1, ...G_N \in \mathrm{Mat}(d, d')$ be independent random matrices with independent standard Gaussian entries. Then with probability at least $1 - \exp(-\Omega(d + d'))$,*

$$\sup \frac{\sum_{i=1}^N \langle Y, G_i Y' G_i^* \rangle}{\|Y\|_F \|Y'\|_F} \lesssim \sqrt{N}(d + d'),$$

where the supremum is over all traceless matrices $Y \in \mathrm{Mat}(d), Y' \in \mathrm{Mat}(d')$ such that $Tr[Y] = Tr[Y'] = 0$.

Proof of Proposition 4.2.6. For $x = \frac{1}{\sqrt{nD}}\{X_1, ..., X_n\}$ and any $a \neq b \in [m]$, x satisfies the λ-spectral condition in Definition 37 if

$$\sup_{(Z_a, Z_b, 0) \in \mathfrak{g}} \frac{\langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle}{\|Z_a\|_F \|Z_b\|_F} \leqslant \frac{\lambda}{\sqrt{d_a d_b}}.$$

In order to use Theorem 4.2.7, we rewrite the samples as random matrices. So for each $X_{i\in[n]}$ define $G_i^{\boldsymbol{j}} \in \mathrm{Mat}(d_a, d_b)$ as

$$G_i^{\boldsymbol{j}}[j_a, j_b] := (X_i)_{j_1,\ldots,j_m}$$

where $\boldsymbol{j}$ runs over all choice of indices or *indexes* $\{j_c \in [d_c]\}_{c\neq a,b}$. Then we can rewrite the spectral condition as

$$\langle \rho_x^{(ab)}, Z_a \otimes Z_b\rangle = \frac{1}{nD}\sum_{i=1}^{n}\sum_{j\in[N]}\langle Z_a, G_{ij} Z_b G_{ij}^*\rangle.$$

By assumption $Tr[Z_a] = Tr[Z_b] = 0$, so we can use Theorem 4.2.7 to show

$$\sup_{(Z_a,Z_b,0)\in\mathfrak{g}} \frac{\frac{1}{nD}\sum_{i=1}^{n}\sum_{j\in[N]}\langle Z_a, G_{ij} Z_b G_{ij}^*\rangle}{\|Z_a\|_F\|Z_b\|_F} \lesssim \sqrt{\frac{nD}{d_a d_b}}\frac{d_a+d_b}{nD} = \frac{d_a+d_b}{\sqrt{nDd_ad_b}},$$

i.e. x satisfies the spectral property for the ab part with $\lambda \lesssim \frac{d_a+d_b}{\sqrt{nD}} \lesssim \varepsilon$, by the assumption that $\varepsilon^2 nD \geqslant d_{\max}^2$. This event occurs with probability at least $1 - \exp(-\Omega(d_a + d_b))$. Therefore by the union bound x satisfies the λ-spectral condition for all parts with the required probability. □

4.2.5 Proof of Main Results

We can now show that our MLE optimization problem is strongly convex.

Lemma 4.2.8. *For $X_1, ..., X_n \sim N(0, I_D)$ from the tensor normal model, if $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, then $\frac{1}{\sqrt{nD}}X$ is $1 - O(m\varepsilon)$-strongly convex with probability at least $1 - 2m^2\exp(-\Omega(d_{\min}))$.*

Proof. By Proposition 4.2.5, $x := \frac{1}{\sqrt{nD}}X$ is $O(\varepsilon)$-balanced, and by Proposition 4.2.6 x satisfies the $O(\varepsilon)$-spectral gap condition. Therefore by Proposition 3.2.6, x is α-strongly convex for

$$\alpha \geqslant s(x)(1 - O(\varepsilon)) - (m-1)O(\varepsilon) \geqslant 1 - O(m\varepsilon).$$

□

Now to prove our main sample complexity theorems, we can apply the analysis of strongly convex tensor scaling.

Proof of Theorem 4.2.2. We first reduce to the case of $\Theta = I_D$. The distribution of X is equivalent to $\Theta^{-1/2}Y_1, ..., \Theta^{-1/2}Y_n$ where $Y \sim N(0, I_D)$. If $\widehat{\Theta}_X, \widehat{\Theta}_Y$ are the MLE for X, Y respectively, then by Proposition 4.1.4 we have the relation

$$\widehat{\Theta}_X = \Theta^{1/2}\widehat{\Theta}_Y\Theta^{1/2}.$$

Further, the relative error also has the same invariance

$$d(\widehat{\Theta}_X, \Theta) = d(\Theta^{-1/2}\widehat{\Theta}_X\Theta^{-1/2}, I_D) = d(\widehat{\Theta}_Y, I_D).$$

So to prove the error bounds in the theorem, it is enough to analyze the case when $Y \sim N(0, I_D)$.

For this case, consider the tensor $x := \frac{1}{\sqrt{nD}}X$. By the condition $\varepsilon^2 nD \geqslant d_{\max}^2$, we can apply Proposition 4.2.5 to show

$$s(x) \geqslant 1 - O(\varepsilon), \quad \|\nabla_x^{(a)}\|_{op} \leqslant O(\varepsilon),$$

simultaneously for all $a \in [m]$ with probability at least $1-2k\exp(-\Omega(\varepsilon^2 nD/d_{\max}))$. Similarly, by Proposition 4.2.6, x satisfies the spectral condition with parameter

$$\frac{O(d_{\max})}{\sqrt{nD}} \leqslant O(\varepsilon)$$

with probability at least $1 - k^2\exp(-\Omega(d_{\min})$. Therefore by the union bound and Proposition 3.2.6, x is α-strongly convex for

$$\alpha \geqslant s(x)(1 - O(\varepsilon)) - (m-1)O(\varepsilon) \geqslant 1 - O(m\varepsilon),$$

with probability at least $1 - k^2\exp(-\Omega(d_{\min}))$. This satisfies the conditions of Theorem 3.2.9 as

$$\left(\sum_{a\in[m]} d_a\right)\|\nabla_x\|_{\mathfrak{g}}^2 \leqslant \left(m\sum_{a\in[m]} d_a\right)\|\nabla_x\|_{op}^2 \lesssim 1 - O(m\varepsilon) \leqslant \alpha^2,$$

where the first step was by Lemma 3.2.7, the second step was by the assumption that ε is small enough, and the final step was by the calculation above for $\alpha \geqslant 1 - O(m\varepsilon)$. Therefore by Theorem 3.2.9, there is a balanced scaling $x_* := e^{Z_*}\cdot x$ such that

$$\|Z_*\|_{\mathfrak{g}}^2 \leqslant \frac{e^2\|\nabla_x\|_{\mathfrak{g}}^2}{\alpha^2} \leqslant \frac{e^2 m\|\nabla_x\|_{op}^2}{(1 - O(m\varepsilon))^2} \leqslant O(m\varepsilon^2),$$

where we again used Lemma 3.2.7 and the fact that x is ε-balanced.

To turn this into a bound on relative error, we use Lemma 4.2.4, which shows that the MLE is $e^{\lambda_*} \cdot e^{Z_*}$ for $\lambda_* = -\log f^*$. We can bound this by

$$|\lambda_*| = |\log s(x_*)| \leq \frac{O(m\varepsilon^2)}{\alpha} + O(\varepsilon) \leq O(\varepsilon),$$

where we used the function lower bound in in Theorem 3.2.9 to bound $\log s(x_*)$ and the error bound derived above to bound $\log s(x)$. Finally we can bound the error of the MLE as

$$d_F(e^{\lambda_*} e^{Z_*}, I_D)^2 \lesssim \left\| \lambda_* I_D + \sum_{a \in [m]} I_{\overline{a}} \otimes (Z_*)_a \right\|_F^2 = D(\lambda_*^2 + \|Z_*\|_{\mathfrak{g}}^2) \lesssim Dm\varepsilon^2,$$

where the first step used the Taylor approximation $e^X \approx I + X$ by the remark after Definition 42, and the second step used the fact that $\|Z\|_{\mathfrak{g}}$ is the standard Frobenius norm on the embedding $Z \to \sum_{a \in [m]} I_{\overline{a}} \otimes Z_a$. □

We can similarly use the strong convexity analysis of Theorem 3.2.15 to show convergence of the Flip-Flop algorithm.

Proof of Theorem 4.2.3. Recall that we have parametrized the MLE optimization by

$$\Theta = e^{\lambda} \cdot \Theta_1 \otimes \ldots \otimes \Theta_m$$

where $\Theta_a \in \mathrm{SPD}(d_a)$. We will use Theorem 3.2.15 to show that the SPD portion of the Flip-Flop algorithm converges quickly to the MLE, and then use Lemma 4.2.4 to compute the normalizing factor.

Once again we can reduce to the case of $\Theta = I_D$. We omit the details since the argument is exactly the same. The proof of Theorem 4.2.2 for samples $X_1, \ldots, X_n \sim N(0, I_D)$ shows that $x := \frac{1}{\sqrt{nD}} X$ satisfies the conditions of Theorem 3.2.15. Therefore let $x_* := e^{Z_*} \cdot x$ be the balanced scaling for $Z_* \in \mathfrak{g}$. Then by Lemma 4.2.4, the MLE is

$$\widehat{\Theta} = e^{\lambda_*} \cdot e^{Z_*}$$

for $\lambda_* = -\log s(x_*)$.

We denote the iterations of Flip-Flop from this perspective so $x_t := e^{Z_t} \cdot x_*$. Then Theorem 3.2.15 shows that $T \lesssim \frac{m}{\alpha} \log \frac{m d_{\max}}{\delta}$ iterations suffice to produce

$$\|Z_T\|_{\mathfrak{g}}^2 \lesssim \delta^2, \quad \log s(x_T) - \log s(x_*) \lesssim \delta^2.$$

Therefore we can bound the relative error as

$$\begin{aligned} d_F(\Theta_T, \widehat{\Theta})^2 &\lesssim \|(\lambda_T - \lambda_*) I_D + Z_T\|_F^2 \\ &\leqslant D(\log s(x_T) - \log s(x_*))^2 + D\|Z_T\|_{\mathfrak{g}}^2 \lesssim D\delta^2, \end{aligned}$$

where we used the normalization $\lambda_T := -\log s(x_T)$ from Lemma 4.2.4 in the second step, and the bounds above for the final step. □

Bibliography

S. Aaronson (2011). "A Linear-Optical Proof that the Permanent is #P-Hard." *Proceedings of the Royal Society A*. MR: 2853286 (cit. on p. 12).

Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson (2018). "Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing." In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 172–181. MR: 3826244 (cit. on p. 29).

C. Améndola, K. Kohn, P. Reichenbach, and A. Seigal (2020). "Invariant theory and scaling algorithms for maximum likelihood estimation." arXiv: 2003.13662 (cit. on pp. 11, 15, 28).

R. Bhatia (2009). *Positive definite matrices*. Princeton university press. MR: 2284176 (cit. on pp. 24, 30).

M. Bläser, C. Ikenmeyer, V. Lysikov, A. Pandey, and F.-O. Schreyer (2021). "On the Orbit Closure Containment Problem and Slice Rank of Tensors." In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, pp. 2565–2584 (cit. on p. 19).

B. G. Bodmann and P. G. Casazza (2010). "The road to equal-norm Parseval frames." *Journal of Functional Analysis* 258.2, pp. 397–420. MR: 2557942 (cit. on p. 8).

P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson (2018). "Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes." In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 883–897. MR: 3899650 (cit. on p. 10).

P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson (2019). "Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes." In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 845–861. MR: 3899650 (cit. on pp. 15, 17, 18, 29, 30, 36).

P. Bürgisser, A. Garg, R. Oliveira, M. Walter, and A. Wigderson (n.d.). "Alternating minimization, scaling algorithms, and the null-cone problem from invariant theory. 2017." *Proceedings of Innovations in Theoretical Computer Science (ITCS 2018)* (). MR: 3761760 (cit. on p. 10).

J. Cahill and P. G. Casazza (2013). "The Paulsen Problem in Operator Theory." *Operators and Matrices*. MR: 3076462 (cit. on p. 7).

P. G. Casazza, M. Fickus, and D. G. Mixon (2012). "Auto-tuning unit norm frames." *Applied and Computational Harmonic Analysis* 32.1, pp. 1–15. MR: 2854158 (cit. on p. 8).

P. G. Casazza and G. Kutyniok, eds. (2013). *Finite Frames: Theory and Applications*. Birkhauser Basel. MR: 2964005 (cit. on pp. 6, 7).

M. Cuturi (2013). "Sinkhorn distances: Lightspeed computation of optimal transport." *Advances in neural information processing systems* 26, pp. 2292–2300 (cit. on p. 6).

H. Derksen and G. Kemper (2015). *Computational invariant theory*. Springer. MR: 3445218 (cit. on pp. 15, 18).

H. Derksen and V. Makam (2020). "Maximum likelihood estimation for matrix normal models via quiver representations." arXiv: 2007.10206 (cit. on p. 15).

H. Derksen, V. Makam, and M. Walter (2020). "Maximum likelihood estimation for tensor normal models via castling transforms." arXiv: 2011.03849 (cit. on p. 15).

C. Franks, R. Oliveira, A. Ramachandran, and M. Walter (2021). "Logarithmic sample complexity for dense matrix and tensor normal models" (cit. on pp. 11, 16, 51, 57, 65).

A. Garg, L. Gurvits, R. de Oliveira, and A. Wigderson (2016). "A deterministic polynomial time algorithm for non-commutative rational identity testing." In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 109–117. arXiv: 1511.03730. MR: 3630971 (cit. on pp. 8, 10).

A. Garg and R. Oliveira (2018). "Recent progress on scaling algorithms and applications." *Bulletin of EATCS* 2.125. MR: 3888184 (cit. on pp. 5, 15).

L. Gurvits (2004). "Classical complexity and quantum entanglement." *Journal of Computer and System Sciences* 69.3, pp. 448–484. MR: 2087945 (cit. on p. 9).

— (2006). “Hyperbolic polynomials approach to van der Waerden/Schrijver-Valiant like conjectures: Sharper bounds, simpler proofs and algorithmic applications.” In: *Symposium on Theory of Computing (STOC)*. MR: 2277167 (cit. on p. 13).

L. Hamilton and A. Moitra (2019). “The Paulsen Problem Made Simple.” In: *Innovations in Theoretical Computer Science (ITCS)*. MR: 3899835 (cit. on p. 8).

D. Hilbert (1890). “Ueber die Theorie der algebraischen Formen.” *Mathematische annalen* 36.4, pp. 473–534. MR: 1510634 (cit. on p. 2).

— (1893). “Über die vollen Invariantensysteme.” *Mathematische annalen* 42.3, pp. 313–373. MR: 1510781 (cit. on pp. 2, 17, 19).

R. Holmes and V. Paulsen (2004). “Optimal frames for erasures.” *Linear Algebra and its Applications*. MR: 2021601 (cit. on pp. 7–9).

M. Idel (2016). “A review of matrix scaling and Sinkhorn’s normal form for matrices and positive maps.” arXiv: 1609.06349. MR: 3314325 (cit. on pp. 5, 15).

M. Jerrum and A. Sinclair (1989). “Approximating the Permanent.” *SIAM Journal of Computing*. MR: 1025467 (cit. on p. 12).

G. Kempf and L. Ness (1979). “The length of vectors in representation spaces.” In: *Algebraic Geometry*. Ed. by K. Lønsted. Vol. 732. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 233–243 (cit. on pp. 17, 29, 32).

T. C. Kwok, L. C. Lau, Y. T. Lee, and A. Ramachandran (2017). “The Paulsen Problem, Continuous Operator Scaling, and Smoothed Analysis.” In: *Symposium on Theory of Computing (STOC)*. ACM. MR: 3826245 (cit. on pp. 7, 8).

T. C. Kwok, L. C. Lau, and A. Ramachandran (2019). “Spectral Analysis of Matrix Scaling and Operator Scaling.” In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 1184–1204 (cit. on pp. 8, 15, 38, 46, 65).

N. Linial, A. Samorodnitsky, and A. Wigderson (2000). “A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents.” *Combinatorica* 20.4, pp. 545–568. MR: 1804826 (cit. on pp. 11–14, 31, 32, 35).

K. D. Mulmuley and M. Sohoni (2001). “Geometric complexity theory I: An approach to the P vs. NP and related problems.” *SIAM Journal on Computing* 31.2, pp. 496–526. MR: 1861288 (cit. on p. 19).

D. Mumford, J. Fogarty, and F. Kirwan (1994). *Geometric invariant theory*. Vol. 34. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer Science & Business Media. MR: 1304906 (cit. on p. 2).

G. Pisier (2012). "Grothendieck's theorem, past and present." *Bulletin of the American Mathematical Society* 49.2, pp. 237–323. MR: 2888168 (cit. on p. 68).

G. Pisier (2014). "Quantum expanders and geometry of operator spaces." *Journal of the European Mathematical Society* 16.6, pp. 1183–1219. MR: 3226740 (cit. on p. 68).

A. Ramachandran (2021). "Geodesic Convex Analysis of Group Scaling for the Paulsen Problem and Tensor normal model." PhD thesis. University of Waterloo (cit. on pp. 8, 38).

Y. Rubner, C. Tomasi, and L. J. Guibas (2000). "The earth mover's distance as a metric for image retrieval." *International journal of computer vision* 40.2, pp. 99–121 (cit. on p. 6).

D. Spielman and N. Srivastava (2011). "Graph Sparsification by Effective Resistances." *SIAM Journal on Computing*. MR: 2863199 (cit. on p. 60).

D. Spielman and S.-H. Teng (2014). "Neraly Linear Time Algorithms for Preconditioning and Solving Symmetric, Diagonally Dominant Linear Systems." *SIAM Journal of Matrix Analysis and Applications*. MR: 3228466 (cit. on p. 60).

D. Straszak and N. K. Vishnoi (2019). "Maximum entropy distributions: Bit complexity and stability." In: *Conference on Learning Theory*. PMLR, pp. 2861–2891 (cit. on p. 28).

B. Sturmfels (2008). *Algorithms in invariant theory*. Springer Science & Business Media. MR: 2667486 (cit. on pp. 15, 18).

R. Vershynin (2012). "Introduction to the non-asymptotic analysis of random matrices." In: *Compressed sensing*. Cambridge Univ. Press, Cambridge, pp. 210–268. arXiv: 1011.3027. MR: 2963170 (cit. on p. 61).

C. Villani (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media. MR: 2459454 (cit. on p. 6).

N. R. Wallach (2017). *Geometric invariant theory*. Springer. MR: 3700428 (cit. on pp. 2, 18).

J. Weyman (1989). "The equations of strata for binary forms." *Journal of Algebra* 122.1, pp. 244–249. MR: 0994946 (cit. on p. 22).

Títulos Publicados — 33º Colóquio Brasileiro de Matemática

Geometria Lipschitz das singularidades – *Lev Birbrair e Edvalter Sena*

Combinatória – *Fábio Botler, Maurício Collares, Taísa Martins, Walner Mendonça, Rob Morris e Guilherme Mota*

Códigos geométricos, uma introdução via corpos de funções algébricas – *Gilberto Brito de Almeida Filho e Saeed Tafazolian*

Topologia e geometria de 3-variedades, uma agradável introdução – *André Salles de Carvalho e Rafał Marian Siejakowski*

Ciência de dados: algoritmos e aplicações – *Luerbio Faria, Fabiano de Souza Oliveira, Paulo Eustáquio Duarte Pinto e Jayme Luiz Szwarcfiter*

Discovering Poncelet invariants in the plane – *Ronaldo A. Garcia e Dan S. Reznik*

Introdução à geometria e topologia dos sistemas dinâmicos em superfícies e além – *Víctor León e Bruno Scárdua*

Equações diferenciais e modelos epidemiológicos – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

Differential Equation Models in Epidemiology – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

A friendly invitation to Fourier analysis on polytopes – *Sinai Robins*

PI-álgebras: uma introdução à PI-teoria – *Rafael Bezerra dos Santos e Ana Cristina Vieira*

First steps into Model Order Reduction – *Alessandro Alla*

The Einstein Constraint Equations – *Rodrigo Avalos e Jorge H. Lira*

Dynamics of Circle Mappings – *Edson de Faria e Pablo Guarino*

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology – *Antonio Galves, Florencia Leonardi e Guilherme Ost*

Transfer operators in Hyperbolic Dynamics - an introduction – *Mark F. Demers, Niloofar Kiamari e Carlangelo Liverani*

A course in Hodge Theory: Periods of Algebraic Cycles – *Hossein Movasati e Roberto Villaflor Loyola*

A dynamical system approach for Lane-Emden type problems – *Liliane Maia, Gabrielle Nornberg e Filomena Pacella*

Visualizing Thurston's geometries – *Tiago Novello, Vinícius da Silva e Luiz Velho*

Scaling problems, algorithms and applications to Computer Science and Statistics – *Rafael Oliveira e Akshay Ramachandran*

An introduction to Characteristic Classes – *Jean-Paul Brasselet*

impa

Instituto de
Matemática
Pura e Aplicada

ISBN 978-65-89124-18-4

9 786589 124184