Towards a theory of non-commutative optimization: geodesic 1st and 2nd order methods for moment maps and polytopes

Peter Bürgisser*, Cole Franks†, Ankit Garg‡, Rafael Oliveira§, Michael Walter¶ and Avi Wigderson‖
*Institut für Mathematik, Technische Universität Berlin
†Department of Mathematics, Rutgers University
‡Microsoft Research India
§Department of Computer Science, University of Toronto
¶Korteweg-de Vries Institute for Mathematics & Institute for Theoretical Physics, University of Amsterdam
‖Institute for Advanced Study, Princeton

***Abstract*—This paper initiates a systematic development of a theory of *non-commutative* optimization, a setting which greatly extends ordinary (Euclidean) convex optimization. It aims to unify and generalize a growing body of work from the past few years which developed and analyzed algorithms for natural *geodesically convex* optimization problems on Riemannian manifolds that arise from the symmetries of non-commutative groups. More specifically, these are algorithms to minimize the *moment map* (a non-commutative notion of the usual *gradient*), and to test membership in *moment polytopes* (a vast class of polytopes, typically of exponential vertex and facet complexity, which quite magically arise from this a-priori non-convex, non-linear setting).**

The importance of understanding this very general setting of geodesic optimization, as these works unveiled and powerfully demonstrate, is that it captures a diverse set of problems, many non-convex, in different areas of CS, math, and physics. Several of them were solved efficiently for the first time using non-commutative methods; the corresponding algorithms also lead to solutions of purely structural problems and to many new connections between disparate fields.

In the spirit of standard convex optimization, we develop two general methods in the geodesic setting, a first order and a second order method, which respectively receive first and second order information on the "derivatives" of the function to be optimized. These in particular subsume all past results. The main technical work, again unifying and extending much of the previous work, goes into identifying the key parameters of the underlying group actions which control convergence to the optimum in each of these methods. These non-commutative analogues of "smoothness" in the commutative case are far more complex, and require significant algebraic and analytic machinery (much existing and some newly developed here). Despite this complexity, the way in which these parameters control convergence in both methods is quite simple and elegant. We also bound these parameters in several general cases.

Our work points to intriguing open problems and suggests further research directions. We believe that extending this theory, namely understanding geodesic optimization better, is both mathematically and computationally fascinating; it provides a great meeting place for ideas and techniques from several very different research areas, and promises better algorithms for existing and yet unforeseen applications.

***Keywords*-computational complexity, convex optimization, geodesic convexity, invariant theory, moment polytopes, non-commutative optimization, null cone, representation theory, scaling algorithms**

I. Introduction

Consider a group G that acts by *linear* transformations on the complex Euclidean space $V = \mathbb{C}^m$. This partitions V into *orbits*: For a vector $v \in V$, the orbit $\mathcal{O}_v$ is simply all vectors of the form $g \cdot v$ to which the action of a group element $g \in G$ can map v.

The most basic algorithmic question in this setting is as follows. Given a vector $v \in V$, compute (or approximate) the smallest ℓ_2-norm of any vector in the orbit of v, that is, $\inf\{\|w\|_2 : w \in \mathcal{O}_v\}$. Remarkably, this simple question, for different groups G, captures natural important problems in computational complexity, algebra, analysis, and quantum information. Even when restricted only to *commutative* groups, it already captures all linear programming problems!

Funding: PB acknowledges support by DFG grant BU 1371 2-2 and by the ERC under the European's Horizon 2020 research and innovation programme (grant agreement no. 787840). CF acknowledges support by Simons Foundation award 332622. MW acknowledges support by the NWO through Veni grant no. 680-47-459. AW acknowledges support by NSF grant CCF-1412958.

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00055

Starting with [1], a series of recent works including [2–7] designed algorithms and analysis tools to handle this basic and other related optimization problems over *non-commutative* groups G. These provided efficient solutions for some applications, and *through algorithms*, the resolution of some purely structural mathematical open problems. We will mention some of these below.

A great deal of understanding gradually evolved in this sequence of works. These new algorithms are all essentially iterative methods, progressing from the input vector v to the desired optimum in small steps, as do convex optimization algorithms. This seems surprising, as the basic question above is patently *non-convex* for non-commutative groups (in the commutative case, a simple change of variables discussed below convexifies the problem). Indeed, neither the domain nor the function to be optimized are convex! However, in hindsight, a key to all of them are the notions of *geodesic convexity* (which generalizes the familiar Euclidean notion of convexity) and the *moment map* (which generalizes the familiar Euclidean gradient) in the curved space and new metrics induced by the group action. A rich duality theory of geometric invariant theory (greatly generalizing LP duality), together with tools from algebraic geometry, representation theory and differential equations are used in the convergence analysis of these algorithms.

The main objective of this paper is to unify and generalize these works, in a way which naturally extends the familiar first and second order methods of standard convex optimization. We design geodesic analogs of these methods, which, respectively, have oracle access to first and second order "derivatives" of the function being optimized. Our first order method (which is a non-commutative version of gradient descent) replaces and extends the use of "alternate minimization" in most past works, and thus can accommodate more general group actions. Our second order method greatly generalizes the one used for the particular group action corresponding to operator scaling in [6]. It may be thought of as a geodesic analog of the "trust region method" [8] or the "box-constrained Newton method" [9, 10] applied to a regularized function. For both methods, in this non-commutative setting, we recover the familiar convergence behavior of the classical commutative case: to achieve "proximity" ε to the optimum, our first order method converges in $O(1/\varepsilon)$ iterations and our second order method in $O(\mathrm{poly}\log(1/\varepsilon))$ iterations.

As in the commutative case, the fundamental challenge is to understand the "constants" hidden in the big-O notation of each method. These depend on "smoothness" properties of the function optimized, which in turn are determined by the action of the group G on the space V that defines the optimization problem. The main technical contributions of the theory we develop are to identify the key parameters which control this dependence, and to bound them for various actions to obtain concrete running time bounds. These parameters depend on a combination of algebraic and geometric properties of the group action, in particular the irreducible representations occurring in it. As mentioned, despite the technical complexity of defining (and bounding) these parameters, the way they control convergence of the algorithms is surprisingly elegant.

We also develop important technical tools which naturally extend ones common in the commutative theory, including regularizers, diameter bounds, numerical stability, and initial starting points, which are key to the design and analysis of the presented (and hopefully future) algorithms in the geodesic setting.

As in previous works, we also address other optimization problems beyond the basic "norm minimization" question above, in particular the minimization of the moment map (which turns out to be a dual problem), and the membership problem for *moment polytopes*; a very rich class of polytopes (typically with exponentially many vertices and facets) which arises magically from any such group action.

The paradigm of optimization described above resulted in efficient algorithms for problems from various diverse areas of CS and mathematics. We mention some of these applications in the full version.

A. Some unexpected applications and connections

We mention here some of the diverse applications of the paradigm of optimization over non-commutative groups:

1) **Algebraic identities:** Given an arithmetic formula (with inversion gates) in non-commuting variables, is it identically zero?
2) **Quantum information:** Given density matrices describing local quantum states of various parties, is there a global pure state consistent with the local states?

3) **Eigenvalues of sums of Hermitian matrices:** Given three real n-vectors, do there exist three Hermitian $n \times n$ matrices A, B, C with these *prescribed* spectra, such that $A + B = C$?
4) **Analytic inequalities:** Given m linear maps $A_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ and $p_1, \ldots, p_m \geqslant 0$, does there exist a finite constant C such that for all integrable functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}_+$ we have

$$\int_{x \in \mathbb{R}^n} \prod_{i=1}^m f_i(A_i x) dx \leqslant C \prod_{i=1}^m \|f_i\|_{1/p_i}?$$

These inequalities are the celebrated Brascamp-Lieb inequalities, which capture the Cauchy-Schwarz, Hölder, Loomis-Whitney, and many further inequalities.

At first glance, it is far from obvious that solving any of these problems has any relation to *either* optimization or groups. We will clarify this mystery below, showing not only how symmetries naturally exist in all of them, but also how these help both in formulating them as optimization problems over groups, suggesting natural algorithms (or at least heuristics) for them, and finally in providing tools for analyzing these algorithms. It perhaps should be stressed that symmetries exist in many examples in which the relevant groups are commutative (e.g., perfect matching in bipartite graphs, matrix scaling, and more generally in linear, geometric, and hyperbolic programming); however in these cases, standard convex optimization or combinatorial algorithms can be designed and analyzed *without* any reference to these existing symmetries. Making this connection explicit is an important part of our exposition.

Polynomial time algorithms were first given in [1] for Problem 1 (the works [11–13] later discovered completely different algebraic algorithms), in [7] for Problem 2, in [4, 14–16] for Problem 3 (the celebrated structural result in [14] and the algorithmic results of [15, 16] solved the decision problem, while [4] solved the search problem), and in [2] for Problem 4. However the algorithms in [2, 4, 7] remain exponential time in various input parameters, exemplifying only one aspect of many in which the current theory and understanding is lacking.

The unexpected connections revealed in this study are far richer than the mere relevance of optimization and symmetries to such problems. One type are connections between problems in disparate fields. For example, the analytic Problem 4 turns out to be a special case of the algebraic Problem 1. Moreover, Problem 1 has (well-studied) differently looking but equivalent formulations in quantum information theory and in invariant theory, which are automatically solved by the same algorithm. Another type of connections are of purely structural open problems solved through such geodesic algorithms, reasserting the importance of the computational lens in mathematics. One was the first dimension-independent bound on the Paulsen problem in operator theory, obtained ingeniously through such an algorithm in [5] (this work was followed by [17], who gave a strikingly simpler proof and stronger bounds). Another was a quantitative bound on the continuity of the best constant C in Problem 4 (in terms of the input data), important for non-linear variants of such inequalities. This bound was obtained through the algorithm in [2] and relies on its efficiency; previous methods used compactness arguments that provided no bounds.

We have no doubt that more unexpected applications and connections will follow. The most extreme and speculative perhaps among such potential applications is to develop a deterministic polynomial-time algorithm for the polynomial identity testing (PIT) problem. Such an algorithm will imply major algebraic or Boolean lower bounds, namely either separating VP from VNP, or proving that NEXP has no small Boolean circuits [18]. We note that this goal was a central motivation of the initial work in this sequence [1], which provided such a deterministic algorithm for Problem 1 above, the non-commutative analog of PIT. The "real" PIT problem (in which variables commute) also has a natural group of symmetries acting on it, which does not quite fall into the frameworks developed so far (including the one of this paper). Yet, the hope of proving lower bounds via optimization methods is a fascinating (and possibly achievable) one. This agenda of hoping to shed light on the PIT problem by the study of invariant theoretic questions was formulated in the fifth paper of the Geometric Complexity Theory (GCT) series [19, 20].

II. Non-commutative optimization: a primer

We now give an introduction to non-commutative optimization and contrast its geometric structure and convexity properties with the familiar commutative setting. The basic setting is that of a continuous group G acting (linearly) on an m-dimensional complex vector space $V \cong \mathbb{C}^m$. For this section, and the rest of the introduction, think

of G as either the group of $n \times n$ complex invertible matrices, denoted $GL(n)$, or the group of *diagonal* such matrices, denoted $T(n)$. The latter corresponds to the commutative case and the former is a paradigmatic example of the non-commutative case. An *action* (also called a *representation*) of a group G on a complex vector space V is a group homomorphism $\pi : G \to GL(V)$, that is, an association of an invertible linear map $\pi(g)$ from $V \to V$ for every group element $g \in G$ satisfying $\pi(g_1 g_2) = \pi(g_1)\pi(g_2)$ for all $g_1, g_2 \in G$. Further suppose that V is also equipped with an inner product $\langle \cdot, \cdot \rangle$ and hence a norm $\|v\| := \langle v, v \rangle$.[1]

Given a vector $v \in V$ one can consider the optimization problem of taking the infimum of the norm in the *orbit* of the vector v under the action of G. More formally, define the *capacity* of v by[2]

$$\mathrm{cap}(v) := \inf_{g \in G} \|\pi(g)v\|.$$

This notion generalizes the matrix and operator capacities developed in [21, 22] (to see this, carry out the optimization over one of the two group variables) as well as the polynomial capacity of [23]. It turns out that this simple-looking optimization problem is already very general in the commutative case and, in the non-commutative case, captures *all* examples discussed in Section I-A.

Let us first consider the commutative case, $G = T(n)$ acting on V. In this simple case, *all* actions π have a very simple form. We give two equivalent descriptions, first of how any representation π splits into one-dimensional irreducible representations, and another describing π as a natural scaling action on n-variate polynomials with m monomials.

The irreducible representations are given by an orthonormal basis $v_1, \ldots, v_m$ of V such that the v_j are simultaneous eigenvectors of all the matrices $\pi(g)$. That is, for all $g = \mathrm{diag}(g_1, \ldots, g_n) \in T(n)$ and $j \in [m]$,

$$\pi(g)v_j = \lambda_j(g)v_j, \quad \text{where} \quad \lambda_j(g) = \textstyle\prod_{i=1}^n g_i^{\omega_j(i)} \tag{II.1}$$

[1]In general, the theory works whenever the group is connected, algebraic and reductive, and our results hold in this generality. However, for purposes of exposition we only discuss very simple groups in this introduction. We also suppress some technical details which are spelled out later, e.g., that the representations are regular and map unitary matrices to unitary matrices (both are essentially without loss of generality).

[2]For notational convenience, we suppress the dependence on the group G and representation π of $\mathrm{cap}(v)$ (likewise for the null cone and the moment polytopes defined below).

for fixed integer vectors $\omega_1, \ldots, \omega_m \in \mathbb{Z}^n$, which are called *weights* and encode the simultaneous eigenvalues, and completely determine the action. Below we also refer to the weights of representation π of $GL(n)$, defined as the weights of π restricted to $T(n)$.

A natural way to view all these actions is as follows. The natural action of $T(n)$ on $\mathbb{C}^n$ by matrix-vector multiplication, induces an action of $T(n)$ on n-variate polynomials $V = \mathbb{C}[x_1, x_2, \ldots, x_n]$: simply, any group element $g = \mathrm{diag}(g_1, \ldots, g_n)$ "scales" each x_i to $g_i x_i$. Note that any monomial $x^\omega = \prod_{i=1}^n x_i^{\omega(i)}$ (where ω is the integer vector of exponents) is an eigenvector of this action, with an eigenvalue $\lambda(g) = \prod_{i=1}^n g_i^{\omega(i)}$.

Now fix m integer vectors ω_j as above. Consider the linear space of n-variate Laurent polynomials (i.e., polynomials where the variables can have negative exponents, too) with the following m monomials: $v_j = x^{\omega_j} = \prod_{i=1}^n x_i^{\omega_j(i)}$. The action on any polynomial $v = \sum_{j=1}^m c_j v_j$ is precisely the one described above, scaling each coefficient by the eigenvalue of its monomial. The norm $\|v\|$ of a polynomial is the sum of the square moduli of its coefficients. Now let us calculate the capacity of this action. For any $v = \sum_{j=1}^m c_j v_j$,

$$\begin{aligned} \mathrm{cap}(v)^2 &= \inf_{g_1, \ldots, g_n \in \mathbb{C}^*} \sum_{j=1}^m |c_j|^2 \prod_{i=1}^n |g_i|^{2\omega_j(i)} \\ &= \inf_{x \in \mathbb{R}^n} \sum_{j=1}^m |c_j|^2 e^{x \cdot \omega_j}, \end{aligned} \tag{II.2}$$

where we used the change of variables $x_i = \log |g_i|^2$, which makes the problem convex (in fact, log-convex)! This class of optimization problems (of optimizing norm in the orbit of a commutative group) is known as *geometric programming* and is well-studied in the optimization literature (see, e.g., Chapter 4.5 in [24]). Hence for non-commutative groups, one can view $\mathrm{cap}(v)$ as *non-commutative geometric programming*. Is there a similar change of variables that makes the problem convex in the non-commutative case? It does not seem so. However, the non-commutative case also satisfies a notion of convexity, known as geodesic convexity, which we will study next.

1) Geodesic convexity: Geodesic convexity generalizes the notion of convexity in the Euclidean space to arbitrary Riemannian manifolds. We will not go into the notion of geodesic convexity in this generality but just mention what it amounts

to in our concrete setting of norm optimization for $G = \mathrm{GL}(n)$.

It turns out the appropriate way to define geodesic convexity in this case is as follows. Fix an action π of $\mathrm{GL}(n)$ and a vector v. Then $\log\|\pi(e^{tH}g)v\|$ is convex in the real parameter t for every Hermitian matrix H and $g \in \mathrm{GL}(n)$. This notion of convexity is quite similar to the notion of Euclidean convexity, where a function is convex iff it is convex along all lines. However, it is far from obvious how to import optimization techniques from the Euclidean setting to work in this non-commutative geodesic setting. An essential ingredient we describe next is the geodesic notion of a gradient, called the *moment map*.

2) Moment map: The moment map is by definition the gradient of the function $\log\|\pi(g)v\|$ (understood as a function of v), at the identity element of the group, $g = I$. It captures how the norm of the vector v changes when we act on it by infinitesimal perturbations of the identity.

Again, we start with the commutative case $G = \mathrm{T}(n)$ acting on the multivariate Laurent polynomials. For a ("direction") vector $h \in \mathbb{R}^n$ and a real ("perturbation") parameter t, let $e^{th} = \mathrm{diag}\left(e^{th_1}, \ldots, e^{th_n}\right)$. Then, for $G = \mathrm{T}(n)$, the moment map is the function $\mu\colon V\backslash\{0\} \to \mathbb{R}^n$, defined by the following property:

$$\mu(v) \cdot h = \partial_{t=0}\left[\log\left\|\pi(\mathrm{diag}(e^{th})v\right\|\right],$$

for all $h \in \mathbb{R}^n$. That is, the directional derivative in direction h is given by the dot product $\mu(v) \cdot h$. Here one can see that the moment map matches the notion of Euclidean gradient. For the action of $\mathrm{T}(n)$ in Eq. (II.1),

$$\mu(v) = \nabla_{x=0} \log\left(\sum_{j=1}^{m} |c_j|^2 e^{x\cdot\omega_j}\right) = \frac{\sum_{j=1}^m |c_j|^2 \omega_j}{\sum_{j=1}^m |c_j|^2}. \tag{II.3}$$

Note that the gradient $\mu(v)$ at any point v is a convex combination of the weights! Viewing v as a polynomial, the gradient thus belongs to the so-called *Newton polytope* of v, namely the convex hull of the exponent vectors of its monomials! Conversely, every point in that polytope is a gradient of some polynomial v with these monomials. We will soon return to this curious fact!

We now proceed to the non-commutative case, focusing on $G = \mathrm{GL}(n)$. Denote by $\mathrm{Herm}(n)$ the set of $n \times n$ complex Hermitian matrices. Here "directions" will be parametrized by $H \in \mathrm{Herm}(n)$. For the case of $G = \mathrm{GL}(n)$, the moment map is the function $\mu\colon V\backslash\{0\} \to \mathrm{Herm}(n)$ defined (in complete analogy to the commutative case above) by the following property that

$$\mathrm{tr}[\mu(v)H] = \partial_{t=0}\left[\log\left\|\pi(e^{tH})v\right\|\right]$$

for all $H \in \mathrm{Herm}(n)$. That is, the directional derivative in direction H is given by $\mathrm{tr}[\mu(v)H]$.

Remark II.1. *The reason we are restricting to directions in $\mathbb{R}^n$ in the $\mathrm{T}(n)$ case and to directions in Herm_n in the $\mathrm{GL}(n)$ case is that imaginary and skew-Hermitian directions, respectively, do not change the norm.*

In the commutative case, Eq. (II.3) is a convex combination of the weights ω_j. Thus, the image of μ is the convex hull of the weights – a convex polytope. This brings us to moment polytopes.

3) Moment polytopes: One can ask whether the above fact is true for actions of $\mathrm{GL}(n)$ i.e., is the set $\{\mu(v) : v \in V\backslash\{0\}\}$ convex? This turns out to be blatantly false. Consider the action of $\mathrm{GL}(n)$ on $\mathbb{C}^n$ by matrix-vector multiplication. The moment map in this setting is $\mu(v) = vv^\dagger/\|v\|^2$, and its image is clearly not convex. However, there is still something deep and non-trivial that can be said. Given a Hermitian matrix $H \in \mathrm{Herm}(n)$, define its *spectrum* to be the vector of its eigenvalues arranged in non-increasing order. That is, $\mathrm{spec}(H) := (\lambda_1, \ldots, \lambda_n)$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ are the eigenvalues of H. Amazingly, the set of spectra of moment map images, that is,

$$\Delta := \left\{\mathrm{spec}\big(\mu(v)\big) : 0 \neq v \in V\right\},$$

is a convex polytope for every representation π [25–29]! These polytopes are called *moment polytopes*.

Let us mention two important examples of moment polytopes. The examples are for actions of products of $\mathrm{GL}(n)$'s but the above definitions generalize almost immediately.

Example II.2 (Star quiver with two arrows, or Horn's problem). $G = \mathrm{GL}(n) \times \mathrm{GL}(n) \times \mathrm{GL}(n)$ *acts on* $\mathrm{Mat}(n) \oplus \mathrm{Mat}(n)$*, as follows:* $\pi(g_1, g_2, g_3)(X, Y) := (g_1 X g_3^{-1}, g_2 Y g_3^{-1})$*. This is one of the simplest examples of a quiver representation [30]. The moment map in this case is*

$$\mu(X, Y) = \frac{(XX^\dagger, YY^\dagger, -(X^\dagger X + Y^\dagger Y))}{\|X\|_F^2 + \|Y\|_F^2}.$$

Using that $XX^\dagger$ and $X^\dagger X$ are PSD and isospectral, we obtain the following moment polytope, which characterizes the eigenvalues of sums of Hermitian matrices, i.e.,

Horn's problem [31]:

$$\Delta = \{ (\mathrm{spec}(A), \mathrm{spec}(B), \mathrm{spec}(-A-B)) : A \geqslant 0, B \geqslant 0, \mathrm{tr}\, A + \mathrm{tr}\, B = 1\}$$

It is known as the Horn polytope *and corresponds to Problem 3 in Section I-A.*

Example II.3 (Tensor action)**.** $G = \mathrm{GL}(n) \times \mathrm{GL}(n) \times \mathrm{GL}(n)$ *acts on* $V = \mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n$*, as follows:* $\pi(g_1, g_2, g_3)v := (g_1 \otimes g_2 \otimes g_3)v$*. We can think of vectors* $\psi \in V$ *as tripartite quantum states with local dimension* n*. Then the moment map for this group action captures precisely the notion of* quantum marginals*. That is,* $\mu(\psi) = (\rho_1, \rho_2, \rho_3)$*, where* $\rho_k = \mathrm{tr}_{k^c}(\psi\psi^\dagger)$ *denotes the reduced density matrix describing the state of the* k*-th particle. This corresponds to Problem 2 in Section I-A. The moment polytope in this case is also known as the* Kronecker polytope*, since it can be equivalently described in terms of the Kronecker coefficients of the symmetric group.*

There is a more refined notion of a moment polytope. One can look at the collection of spectra of moment maps of vectors in the *orbit closure* of a particular vector $v \in V$. Its closure,

$$\Delta(v) := \overline{\{\mathrm{spec}(\mu(w)) : w \in \mathcal{O}_v\}}$$

is a convex polytope as well, called the *moment polytope of* v [25, 32]!

4) Null cone: Fix a representation π of a group G on a vector space V (recall G is $T(n)$ or $\mathrm{GL}(n)$ for the introduction). The *null cone* for this group action is defined as the set of vectors v such that $\mathrm{cap}(v) = 0$:

$$\mathcal{N} := \{v \in V : \mathrm{cap}(v) = 0\}$$

In other words, v is in the null cone if and only if 0 lies in the orbit-closure of v. It is of importance in invariant theory due to the results of Hilbert and Mumford [33, 34] which state that the null cone is the algebraic variety defined by non-constant homogeneous invariant polynomials of the group action (see, e.g., the excellent textbooks [35, 36]).

Let us see what the null cone for the action of $T(n)$ in Eq. (II.1) is. Recall from Eq. (II.2), the formulation for $\mathrm{cap}(v)$. It is easy to see that $\mathrm{cap}(v) = 0$ iff there exists $x \in \mathbb{R}^n$ such that $x \cdot \omega_j < 0$ for all $j \in \mathrm{supp}(v)$, where $\mathrm{supp}(v) = \{j \in [m] : c_j \neq 0\}$ for $v = \sum_{j=1}^m c_j v_j$. Thus the property of v being in the null cone is captured by a simple linear program defined by $\mathrm{supp}(v)$ and the weights ω_j's. Hence the null cone membership problem for non-commutative group actions can be thought of as *non-commutative linear programming*.

We know by Farkas' lemma that there exists a $x \in \mathbb{R}^n$ such that $x \cdot \omega_j < 0$ for all $j \in \mathrm{supp}(v)$ iff 0 does not lie in $\mathrm{conv}\{\omega_j : j \in \mathrm{supp}(v)\}$. In other words, $\mathrm{cap}(v) = 0$ iff $0 \notin \Delta(v)$. Is this true in the non-commutative world? It is! This is the Kempf-Ness theorem [37] and it is a consequence of the geodesic convexity of the function $g \to \log\|\pi(g)v\|$. The Kempf-Ness theorem can be thought of as a *non-commutative duality theory* paralleling the linear programming duality given by Farkas' lemma (which corresponds to the commutative world). Let us now mention an example of an interesting null cone in the non-commutative case.

Example II.4 (Operator scaling, or left-right action)**.** $G = \mathrm{SL}(n) \times \mathrm{SL}(n)$ *(where* $\mathrm{SL}(n)$ *denotes the group of* $n \times n$ *matrices with determinant* 1*) acts on* $\mathrm{Mat}(n)^k$ *as follows:* $\pi(g, h)(X_1, \ldots, X_k) := (gX_1h^T, \ldots, gX_kh^T)$*. This family of actions is called the* left-right *action.*

The null cone for this action captures non-commutative singularity *(see, e.g., [1, 11–13]) and Problem 1 in Section I-A. The left-right action has been crucial in getting deterministic polynomial time algorithms for the non-commutative rational identity testing problem* [1, 11–13]*. The commutative analogue is the famous polynomial identity testing (PIT) problem, for which designing a deterministic polynomial time algorithm remains a major open question in derandomization and complexity theory.*

Remark II.5 (Generalized Kronecker quivers)**.** *Also sometimes referred to as the left-right action, the action* $\pi(g, h)(X_1, \ldots, X_k) := (gX_1h^{-1}, \ldots, gX_kh^{-1})$ *of matrices* $g, h \in \mathrm{GL}(n)$ *on* k*-tuples of matrices* $(X_1, \ldots, X_k)$ *can be obtained from action of Example II.4 via the isomorphism* $h \mapsto (h^{-1})^T$ *of* $\mathrm{GL}(n)$*. These actions are called* representations of the generalized Kronecker quivers*.*

III. Computational problems and state of the art

In this section, we describe the main computational questions that are of interest for the optimization problems discussed in the previous section and then discuss what is known about them in the commutative and non-commutative worlds.

Problem III.1 (Null cone membership)**.** *Given* (π, v)*, determine if* v *is in the null cone, i.e., if* $\mathrm{cap}(v) = 0$*. Equivalently, test if* $0 \notin \Delta(v)$*.*

The null cone membership problem for $GL(n)$ is interesting only when the action $\pi(g)$ is given by rational functions in the g_{ij} rather than polynomials. This is completely analogous to the commutative case (e.g., the convex hull of weights ω_j with positive entries never contains the origin). In the important case that π is homogeneous, the null cone membership problem is interesting precisely when the total degree is zero, so that scalar multiples of the identity matrix act trivially. Thus, in this case the null cone membership problem for $G = GL(n)$ is equivalent to the one for $G = SL(n)$. We will come back to this perspective in Section V.

Problem III.2 (Scaling). *Given* (π, v, ε) *such that* $0 \in \Delta(v)$, *output a group element* $g \in G$ *such that* $\|\operatorname{spec}(\mu(g)v)\|_2 = \|\mu(\pi(g)v)\|_F \leqslant \varepsilon$.

In particular, the following promise problem can be reduced to Problem III.2: Given (π, v, ε), decide whether $0 \notin \Delta(v)$ under the promise that either $0 \in \Delta(v)$ or 0 is ε-far from $\Delta(v)$. In fact, there always exists $\varepsilon > 0$, depending only on the group action, such that this promise is satisfied! Thus the null cone membership problem can always be reduced to the scaling problem (see Corollary IV.5 below).

In the full version we develop a duality theory showing that an efficient agorithm to minimize the norm on an orbit closure of a vector v (i.e., approximate the capacity of v) under the promise that $0 \in \Delta(v)$ results in an efficient algorithm for the scaling problem and hence for the null cone membership problem. This motivates our next computational problem.

Problem III.3 (Norm minimization). *Given* (π, v, ε) *such that* $\operatorname{cap}(v) > 0$, *output a group element* $g \in G$ *such that* $\log\|\pi(g)v\| - \log \operatorname{cap}(v) \leqslant \varepsilon$.

We also consider the moment polytope membership problem for an arbitrary point $p \in \mathbb{Q}^n$.

Problem III.4 (Moment polytope membership). *Given* (π, v, p), *determine if* $p \in \Delta(v)$.

The moment polytope membership problem is more general than the null cone membership problem, but there is a reduction from the former to the latter via the "shifting trick" in the next subsection. This forms the basis of our algorithms for the moment polytope membership problem. As in the case of the null cone, we consider a scaling version of the moment polytope membership problem.

Problem III.5 (p-Scaling). *Given* (π, v, p, ε) *such that* $p \in \Delta(v)$, *output an element* $g \in G$ *such that* $\|\operatorname{spec}(\mu(\pi(g)v)) - p\|_2 \leqslant \varepsilon$.

The above problem has been referred to as *nonuniform scaling* [7] or, for operators, matrices and tensors, as *scaling with specified or prescribed marginals* [4]. The following problem can be reduced to Problem III.5: Given (π, v, p, ε), decide whether $p \in \Delta(v)$ under the promise that either $p \in \Delta(v)$ or p is ε-far from $\Delta(v)$. By combining the shifting trick with our duality theory, we show in the full version that there is a value ε of bit size polynomial in the input size such that the moment polytope membership problem can be reduced to p-scaling.

There are multiple interesting input models for these problems. One could explicitly describe the weights $\omega_1, \ldots, \omega_m$ for an action of $T(n)$ (Eq. (II.1)) and then describe v as $\sum_{j=1}^m c_j v_j$ by describing the c_j's. The analogous description in the non-commutative world would be to describe the irreducible representations occuring in V. Alternately, one could give black box access to the function $\|\pi(g)v\|$, or to the moment map $\mu(\pi(g)v)$, etc. Sometimes π can be a non-uniform input as well, such as a fixed family of representations like the simultaneous left-right action Example II.4 as done in [1]. The inputs p and ε will be given in their binary descriptions but we will see that some of the algorithms run in time polynomial in their unary descriptions.

Remark III.6 (Running time in terms of ε). *By standard considerations about the bit complexity of the facets of the moment polytope, it can be shown that polynomial time algorithms for the scaling problems (Problems III.2 and III.5) result in polynomial time algorithms for the exact versions (Problems III.1 and III.4, respectively). Polynomial time requires, in particular,* $\operatorname{poly}(\log(1/\varepsilon))$ *dependence on* ε*; a* $\operatorname{poly}(1/\varepsilon)$ *dependence is only known to suffice in special cases.*

A. Commutative groups and geometric programming

In the commutative case, the preceding problems are reformulations of well-studied optimization problems and much is known about them computationally. To see this, consider the action of $T(n)$ as in Eq. (II.1), and a vector $v = \sum_{j=1}^m c_j v_j$. It follows from Section II-4 that v is in the null cone iff $0 \notin \Delta(v) = \operatorname{conv}\{\omega_j : c_j \neq 0\}$. Recall from Eq. (II.2), the formulation for $\operatorname{cap}(v)$. Since this formulation is convex, it follows that, given $\omega_1, \ldots, \omega_m \in \mathbb{Z}^n$ (recall this is the description

of π) and $c_1, \ldots, c_m \in \mathbb{Q}[i]$ (each entry described in binary), there is a polynomial-time algorithm for the null cone membership problem via linear programming [38, 39]. The same is true for the moment polytope membership problem. The capacity optimization problem is an instance of *(unconstrained) geometric programming* and one can design polynomial time algorithms in the same input model. It is hard to find an exact reference for this, but this can be done, for example, using the ellipsoid algorithm as done for the same problem in slightly different settings in the papers [40–42]. There has been work in the oracle setting as well, in which one has oracle access to the function $\|\pi(g)v\|$. The advantage of the oracle setting is that one can handle exponentially large representations of $T(n)$ when it is not possible to describe all the weights explicitly. A very general result of this form is proved in [42]. While not explicitly mentioned in [42], their techniques can also be used to design polynomial time algorithms for *commutative* null cone and moment polytope membership in the oracle setting. Thus, in the commutative case, Problems III.1, III.3 and III.4 are well-understood.

B. Non-commutative actions

Comparatively very little is known in the non-commutative case. The two non-trivial group actions for which there are known polynomial-time algorithms for null cone membership (Problem III.1) are *simultaneous conjugation* [43, 44] and the *left-right* action [1, 11–13]. Approximate algorithms for null cone membership have been designed for the *tensor action* of products of $SL(n)$'s [3]. However the running time is exponential in the binary description of ε (i.e., polynomial in $1/\varepsilon$). This is the reason the algorithm does not lead to a polynomial time algorithm for the exact null cone membership problem for the tensor action.

The first work on moment polytope membership (Problem III.4) in the noncommutative case focused on *Brascamp-Lieb polytopes* [2] (which are affine slices of moment polytopes) and solved the moment polytope membership problem in time depending polynomially on the *unary* complexity of the target point. In [7], efficient algorithms were designed for the p-scaling problem (Problem III.5) for tensor actions, extending the earlier work of [4] for the simultaneous left-right action. The running times of both algorithms are $\mathrm{poly}(1/\varepsilon)$; for this reason both algorithms result in moment polytope algorithms depending exponentially on the binary bitsize of p as in [2].

Regarding the approximate computation of the capacity (Problem III.3), efficient algorithms were previously known only for the simultaneous left-right action. [1] gave an algorithm to approximate the capacity in time polynomial in all of the input description except ε, on which it had dependence $\mathrm{poly}(1/\varepsilon)$. The paper [6] gave an algorithm that depended polynomially on the input description; it has running time dependence $\mathrm{poly}(\log(1/\varepsilon))$ on the error parameter ε.

In terms of algorithmic techniques, all prior works fall into two categories. One is that of *alternating minimization* (which can be thought of as a large-step coordinate gradient descent, i.e., roughly speaking as a first order method). However, alternating minimization is limited in applicability to 'multilinear' actions of products of $T(n)$'s or $GL(n)$'s, where the action is linear in each component so that it is easy to optimize over one component when fixing all the others. This is true for all the actions described above and hence explains the applicability of alternating minimization (in fact, in all the above examples, one can even get a closed-form expression for the group element that has to be applied in each alternating step). The second category are geodesic analogues of *box-constrained Newton's methods* (second order). Recently, [6] designed an algorithm tailored towards the specific case of the simultaneous left-right action (Example II.4), but no second order algorithms were known for other group actions. However, many group actions of interest – from classical problems in invariant theory about symmetric forms to the important variant of Problem 2 in Section I-A for fermions – are not multilinear nor can otherwise be captured by the left-right action, and no efficient algorithms were known. All this motivates the development of new techniques.

In this paper, we show how these limitations can be overcome. Specifically, we provide both first and second order algorithms (geodesic variants of gradient descent and box-constrained Newton's method) that apply in great generality and identify the main structural parameters that control the running time of these algorithm. We now describe our contributions in more detail.

IV. Algorithmic and structural results

We describe here our algorithmic and structural contributions to the theory of non-commutative

optimization. In Section IV-A, we describe the main parameters that govern the running time of our algorithms. In Section IV-B, we describe the first order algorithm for $\mathrm{cap}(v)$ and the structural results we prove for its analysis. In Section IV-B1, we describe a first order algorithm for the problem of membership in moment polytopes and the relevant structural results. In Section IV-C, we describe the second order algorithm for $\mathrm{cap}(v)$ and the techniques and ideas used in its analysis.

A. Essential parameters and structural results

In this section, we define the essential parameters related to the group action which, in addition to dictating the running times of our first and second order methods, control the relationships between the null cone, the norm of the moment map, and the capacity, i.e., between Problems III.1 to III.3.

We saw in Section II that for all actions of $T(n)$ on a vector space V, one can find a basis of V consisting of simultaneous eigenvectors of the matrices $\pi(g)$, $g \in T(n)$. While this is in general impossible for non-commutative groups, one can still decompose V into building blocks known as irreducible subspaces (or subrepresentations), as will be discussed in further detail in the full version.

For $GL(n)$, these are uniquely characterized by nonincreasing sequences $\lambda \in \mathbb{Z}^n$; such sequences λ are in bijection with irreducible representations $\pi_\lambda \colon GL(n) \to GL(V_\lambda)$. We say that λ *occurs in* π if one of its irreducible subspaces is of type λ. If all the λ occuring in π have nonnegative entries, then the entries of the matrix $\pi(g)$ are polynomials in the entries of g. Such representations π are called *polynomial*, and if all λ occuring in π have sum exactly (resp. at most) d, then π is said to be a *homogeneous polynomial representation of degree (resp. at most)* d. We elaborate further on the representation theory of $GL(n)$ in Section V and in the full version.

Now we can define the complexity measure which captures the smoothness of the optimization problems of interest. In the full version we discuss how to think of the following measure as a *norm* of the Lie algebra representation Π, hence the name *weight norm*.

Definition IV.1 (Complexity measure I: weight norm)**.** *We define the* weight norm $N(\pi)$ *of an action* π *of* $GL(n)$ *by* $N(\pi) := \max\{\|\lambda\|_2 : \lambda \text{ occurs in } \pi\}$, *where* $\|\cdot\|_2$ *denotes the Euclidean norm.*

Another use of the weight norm is to provide a bounding ball for the moment polytope. As shown in the full version, the moment polytope is contained in a Euclidean ball of radius $N(\pi)$. The weight norm is in turn controlled by the degree of a polynomial representation. More specifically, if π is a polynomial representation of $GL(n)$ of degree at most d, then $N(\pi) \leqslant d$.

We now describe our second measure of complexity which will govern the running time bound for our second order algorithm. This parameter, which will be discussed further in the full version, also features in Theorem IV.3 concerning quantitative non-commutative duality.

Definition IV.2 (Complexity measure II: weight margin)**.** *The* weight margin $\gamma(\pi)$ *of an action* π *of* $GL(n)$ *is the minimum Euclidean distance between the origin and the convex hull of any subset of the weights of* π *that does not contain the origin.*

Our running time bound will depend inversely on the weight margin. Two interesting examples with large (inverse polynomial) weight margin are the left-right action (Example II.4) and simultaneous conjugation. The existing second order algorithm for the left-right action relied on the large weight margin of the action [6]. It is interesting that the simultaneous conjugation action ($GL(n)$ acts on $(\mathrm{Mat}(n))^{\oplus d}$, $\pi(g)(X_1, \ldots, X_d) = (gX_1g^{-1}, \ldots, gX_dg^{-1})$), the sole other interesting example of an action of a non-commutative group for which there are efficient algorithms for the null cone membership problem [43–45] (which have nothing to do with the weight margin), also happens to have large weight margin! The only generally applicable lower bound on the weight margin is $n^{-1}N(\pi)^{-n}$, and indeed this exponential behavior is seen for the somewhat intractable 3-tensor action (Example II.3), which has weight margin at most $2^{-n/3}$ and weight norm $\sqrt{3}$ (implicit in [46]). For the convenience of the reader, we arrange in a tabular form the above information about the weight margin for various paradigmatic group actions in Table I (using a definition of the weight margin and weight norm, given in the full version, that naturally generalizes the one given above for $GL(n)$):

As the moment map is the gradient of the geodesically convex function $\log\|v\|$, it stands to

[3]This commutative example is modelled as follows: $G = ST(n) \times ST(n)$ acts on $\mathrm{Mat}(n)$ by $\pi(A, B)M = AMB$, where $ST(n)$ is the group of diagonal $n \times n$ matrices with unit determinant.

Group action	Weight margin $\gamma(\pi)$
Matrix scaling[3]	$\geqslant n^{-3/2}$ [47]
Simultaneous conjugation	$\geqslant n^{-3/2}$ (full version)
Simultaneous left-right action	$\geqslant \Omega(n^{-3/2})$ [22]
3-tensor action	$\leqslant 2^{-n/3}$; implicit in [46]
GL(n)-action of degree d	$\geqslant d^{-n}n^{-1}$ (full version)

Table I: Weight margin for various representations.

reason that as $\mu(v)$ tends to zero, $\|v\|$ tends to the capacity $\mathrm{cap}(v)$. However, in order to use this relationship to obtain efficient algorithms, we need this to hold in a precise quantitative sense. To this end, in the full version we show the following fundamental relation between the capacity and the norm of the moment map.

Theorem IV.3 (Noncommutative duality). *For $v \in V\backslash\{0\}$ we have*

$$1 - \frac{\|\mu(v)\|_F}{\gamma(\pi)} \leqslant \frac{\mathrm{cap}(v)^2}{\|v\|^2} \leqslant 1 - \frac{\|\mu(v)\|_F^2}{4N(\pi)^2}.$$

Equipped with this inequality, it is easy to relate Problems III.2 and III.3.

Corollary IV.4. *An output g for the norm minimization problem on input (π, v, ε) is a valid output for the scaling problem on input $(\pi, v, N(\pi)\sqrt{8\varepsilon})$. If $\varepsilon/\gamma(\pi) < \frac{1}{2}$ then an output g for the scaling problem on input (π, v, ε) is a valid output for the norm minimization problem on input $(\pi, v, \frac{2\log 2\varepsilon}{\gamma(\pi)})$.*

Because $0 \in \Delta(v)$ if and only if $\mathrm{cap}(v) > 0$, Theorem IV.3 and Corollary IV.4 immediately yield the accuracy to which we must solve the scaling problem or norm minimization problem to solve the null cone membership problem:

Corollary IV.5. *It holds that $0 \in \Delta(v)$ if and only if $\Delta(v)$ contains a point of norm smaller than $\gamma(\pi)$. In particular, solving the scaling problem with input $(\pi, v, \gamma(\pi)/2)$ or the norm minimization problem with $(\pi, v, \frac{1}{8}(\gamma(\pi)/2N(\pi))^2)$ suffices to solve the null cone membership problem for (π, v).*

In the full version we prove an analogous statement relating p-scaling to the moment polytope membership problem.

B. First order methods: structural results and algorithms

As discussed above, in order to approximately compute the capacity in the commutative case, one can just run a Euclidean gradient descent on the convex formulation in Eq. (II.2). We will see that gradient descent method naturally generalizes to the non-commutative setting. It is worth mentioning that there are several excellent sources of the analysis of gradient descent algorithms for geodesically convex functions (in the general setting of Riemannian manifolds and not just the group setting that we are interested in); see e.g., [48–53] and references therein. In this paper, our contribution is mostly in understanding the geometric properties (such as smoothness) of the optimization problems that we are concerned with, which allow us to carry out the classical analysis of Euclidean gradient descent in our setting.

The natural analogue of gradient descent for the optimization problem $\mathrm{cap}(v)$ is the following: start with $g_1 = I$ and repeat, for $T - 1$ iterations and a suitable step size η:

$$g_{t+1} = e^{-\eta\mu(\pi(g_t)v)}g_t$$

Finally, return the group element g among $g_1, \ldots, g_T$, which minimizes $\|\mu(\pi(g)v)\|_F$. This algorithm is described in Algorithm IV.1. A natural geometric parameter which governs the complexity (number of iterations T, step size η) of gradient descent is the *smoothness* of the function to be optimized. The smoothness parameter for actions of $T(n)$ in Eq. (II.1) turns out to be $O(\max_{j\in[m]}\|\omega_j\|_2^2)$ (see, e.g., [42]). We prove in the full version that the function $\log\|\pi(g)v\|$ is geodesically smooth, with a smoothness parameter exactly analogous to the commutative case. The smoothness parameter turns out to be the weight norm $N(\pi)$ defined in Definition IV.1.

Input:
- Oracle access to the moment map restricted to a group orbit, i.e., to the map $g \mapsto \mu(\pi(g)v)$,
- a number of iterations T.

Output: A group element $g \in G$.

Algorithm:
1) Set $g_0 = I$. Set a step size $\eta = \frac{1}{2N(\pi)^2}$.
2) For $t = 0, \ldots, T - 1$: Set
$$g_{t+1} := e^{-\eta\,\mu(\pi(g_t)v)}g_t.$$
3) **return** $\arg\min_{g\in\{g_0,\ldots,g_{T-1}\}}\|\mu(\pi(g)v)\|_F^2$

Algorithm IV.1: Algorithm for the scaling problem.

We now state the running time for our geodesic gradient descent algorithm for Problem III.2, which is proved in the full version.

Theorem IV.6 (First order algorithm for scaling)**.** *Fix a representation* $\pi : \mathrm{GL}(n) \to \mathrm{GL}(V)$ *and a unit vector* $v \in V$ *such that* $\mathrm{cap}(v) > 0$ *(i.e., v is not in the null cone). Then Algorithm IV.1 with a number of iterations at most*

$$T = O\left(\frac{N(\pi)^2}{\varepsilon^2}\left|\log \mathrm{cap}(v)\right|\right)$$

outputs a group element $g \in G$ *satisfying* $\|\mu(\pi(g)v)\|_F \leqslant \varepsilon$.

Theorem V.2 in Section V states concrete running time bounds in terms of the bit complexity of the input.

The analysis of Theorem IV.6 relies on the smoothness of the function $F_v(g) := \log\|\pi(g)v\|$, which implies that

$$F_v(e^H g) \leqslant F_v(g) + \mathrm{tr}\left[\mu\big(\pi(g)v\big)H\right] + N(\pi)^2\|H\|_F^2,$$

for all $g \in \mathrm{GL}(n)$ and for all Hermitian $H \in \mathrm{Herm}(n)$.

1) First order method for moment polytope membership: Next, we describe our first order algorithm for the p-scaling problem. Theorem IV.6 solves the problem of minimizing the moment map (equivalent to capacity computation), hence can be used to determine if $0 \in \Delta(v)$. Can we reduce the general moment polytope membership problem, $p \in \Delta(v)$, to this case? This is straightforward in the commutative case, $G = T(n)$. It follows from the reasoning in Section II-4 that, for $p \in \mathbb{R}^n$, we have $p \notin \Delta(v)$ iff

$$\mathrm{cap}_p(v)^2 := \inf_{x\in\mathbb{R}^n} \textstyle\sum_{j=1}^m |c_j|^2 e^{x\cdot(\omega_j - p)} = 0. \tag{IV.1}$$

Thus, all we need to do is shift the relevant vectors by p. Is there an analog of this trick in the non-commutative world? There is! It is called, unsurprisingly, the *shifting trick* [32]. Let us describe it here. A nice property about Eq. (IV.1) is that (recall Eq. (II.3)) $\nabla_{x=0} \log\left(\sum_{j=1}^m |c_j|^2 e^{x\cdot(\omega_j - p)}\right) = \mu(v) - p$. How do we shift the moment map in the case of $\mathrm{GL}(n)$? It relies on the following two elementary properties of the moment map:

1) The moment map of the tensor product π of two representations π_1, π_2 of $\mathrm{GL}(n)$, which is defined as $\pi(g)(v\otimes w) := (\pi_1(g)v)\otimes(\pi_2(g)w)$, satisfies $\mu(v\otimes w) = \mu(v) + \mu(w)$.
2) There is a vector v_λ (known as a *highest weight vector*) in the vector space V_λ of the irreducible action π_λ such that $\mu(v_\lambda) = \mathrm{diag}(\lambda)$.

Now suppose $p \in \mathbb{Q}^n$ and let $\ell > 0$ be the least integer such that $\lambda := \ell p \in \mathbb{Z}^n$. Let $\lambda^* := (-\lambda_n, \ldots, -\lambda_1)$. Then one can see that the tensor product action of $\mathrm{GL}(n)$ on the space $\mathrm{Sym}^\ell(V)\otimes V_{\lambda^*}$ satisfies $\frac{1}{\ell}\mu\left(v^{\otimes\ell}\otimes v_{\lambda^*}\right) = \mu(v) + \mathrm{diag}(\lambda^*)/\ell = \mu(v) - \Lambda$, where Λ is the diagonal matrix with entries $\Lambda_{i,i} = p_{n-i+1}$, which has spectrum p. We have managed to shift the moment map! So we are led to the following optimization problem,

$$\mathrm{cap}_p(v)^\ell := \inf_{g\in G} \|(\pi(g)v)^{\otimes\ell}\otimes(\pi_{\lambda^*}(g)v_{\lambda^*})\|.$$

In the noncommutative case, the relation between this p-*capacity* and the moment polytope is slightly more subtle. While $\mathrm{cap}_p(v) > 0$ always guarantees that $p \in \Delta(v)$, these two conditions are in general *not* equivalent (unless $p = 0$, when $\mathrm{cap}_p(v)$ reduces to $\mathrm{cap}(v)$). However, what holds is that $p \in \Delta(v)$ if and only if $\mathrm{cap}_p(\pi(g)v) > 0$ for *generic* $g \in G$. We can thus solve the p-scaling problem by first applying a random group element and then applying an optimization algorithm to approximate $\mathrm{cap}_p(v)$.

We now outline our optimization algorithm for $\mathrm{cap}_p(v)$. The optimization problem defining $\mathrm{cap}_p(v)$ is defined in terms of actions on a space of exponential dimension. However, it turns out that the gradients can be explicitly computed and the geodesic gradient descent can be described explicitly as follows: start with $g_1 = I$ and repeat, for $T - 1$ iterations and suitable step size η:

$$g_{t+1} = e^{-\eta\left(\mu(\pi(g_t)v) - Q_t\Lambda Q_t^\dagger\right)} g_t,$$

where $g_t = Q_t R_t$ is the QR decomposition of g_t. Finally return group element g among $g_1, \ldots, g_T$, which minimizes $\|\mu(\pi(g_t)v) - Q_t\Lambda Q_t^\dagger\|_F$. This algorithm is stated precisely in the full version.

Theorem IV.7 (First order algorithm for p-scaling)**.** *Fix a representation* $\pi : \mathrm{GL}(n) \to \mathrm{GL}(V)$*, a unit vector* $v \in V$*, and a target point* $p \in \mathbb{Q}^n$ *such that* $\mathrm{cap}_p(v) > 0$*. Let* $N^2 := N(\pi)^2 + \|p\|_2$*. Then our first order algorithm for* p*-scaling with a number of iterations at most*

$$T = O\left(\frac{N^2}{\varepsilon^2}\left|\log \mathrm{cap}_p(v)\right|\right)$$

outputs a group element $g \in G$ *satisfying* $\|\mathrm{spec}\left(\mu(\pi(g)v)\right) - p\|_2 \leqslant \varepsilon$.

A precise calculation of the smoothness of the function $g \mapsto \log\|\pi(g)v\| + \frac{1}{\ell}\log\|\pi_{\lambda^*}(g)v_{\lambda^*}\|$ (which underlies the p-capacity) features crucially in our analysis.

As described above, Theorem IV.7 preceded by a randomization step can be used to solve the p-scaling problem. Theorem V.5 in Section V describes the performance of such a randomized algorithm for $G = \mathrm{GL}(n)$.

C. Second order methods: structural results and algorithms

Here we discuss our second order algorithm for Problem III.3, the approximate norm minimization problem. As mentioned in Section III, the paper [6] (following the algorithms developed in [9, 10] for the commutative Euclidean case) developed a second order polynomial-time algorithm for approximating the capacity for the simultaneous left-right action (Example II.4) with running time polynomial in the bit description of the approximation parameter ε. In the full version, we generalize this algorithm to arbitrary groups and actions. It repeatedly optimizes quadratic Taylor expansions of the objective in a small neighbourhood. Such algorithms also go by the name "trust-region methods" in the Euclidean optimization literature [8]. The running time of our algorithm will depend inversely on the weight margin defined in Definition IV.2.

Theorem IV.8 (Second-order algorithm for norm minimization)**.** *Fix a representation* $\pi : \mathrm{GL}(n) \to \mathrm{GL}(V)$ *and a unit vector* $v \in V$ *such that* $\mathrm{cap}(v) > 0$*. Put* $C := |\log \mathrm{cap}(v)|$, $\gamma := \gamma(\pi)$ *and* $N := N(\pi)$*. Then our second order g-convex optimization algorithm for a suitably regularized objective function outputs* $g \in G$ *satisfying* $\log \|\pi(g)v\| \leqslant \log \mathrm{cap}(v) + \varepsilon$ *with a number of iterations at most*

$$T = O\left(\frac{N\sqrt{n}}{\gamma}\left(C + \log\frac{n}{\varepsilon}\right)\log\frac{C}{\varepsilon}\right).$$

Theorem V.3 in Section V specializes Theorem IV.8 to the group $\mathrm{SL}(n)$ by obtaining running time bounds in terms of the bit complexity of the input.

The two main structural parameters which govern the runtime of our second order g-convex optimization algorithm in general are the *robustness* (controlled by the weight norm) and a *diameter bound* (controlled by the weight margin). The robustness of a function bounds third derivatives in terms of second derivatives, similarly to the well-known notion of self concordance (however, in contrast to the latter, the robustness is not scale-invariant). As a consequence of the robustness, we show in the full version that the function $F_v(g) = \log\|\pi(g)v\|$ is sandwiched between two quadratic expansions in a small neighbourhood:

$$\begin{aligned} F(g) + \partial_{t=0}F(e^{tH}g) + \frac{1}{2e}\partial^2_{t=0}F(e^{tH}g) \\ \leqslant F(e^H g) \leqslant F(g) + \partial_{t=0}F(e^{tH}g) + \frac{e}{2}\partial^2_{t=0}F(e^{tH}g) \end{aligned}$$

for every $g \in \mathrm{GL}(n)$ and $H \in \mathrm{Herm}(n)$ such that $\|H\|_F \leqslant 1/(4N(\pi))$.

Another ingredient in the analysis of our second order g-convex optimization algorithm is to prove the existence of "well-conditioned" approximate minimizers, i.e. $g_\star \in G$ with small condition number satisfying $\log\|\pi(g_\star)v\| \leqslant \log \mathrm{cap}(v) + \varepsilon$. The bound on the condition numbers of approximate minimizers helps us ensure that the algorithm's trajectory always lies in a compact region with the use of appropriate regularizers. As in [6], we obtain this "diameter bound" by designing a suitable gradient flow and bounding the (continuous) time it takes for it to converge! A crucial ingredient of this analysis is our Theorem IV.3 relating capacity and norm of the moment map.

Our gradient flow approach, which can be traced back to works in symplectic geometry [54], is the only one we know for proving diameter bounds in the non-commutative case. In contrast, in the commutative case several different methods are available (see, e.g., [41, 42]). It is an important open problem to develop alternative methods for diameter bounds in the non-commutative case, which will also lead to improved running time bounds for algorithms like our second order g-convex optimization algorithm.

V. Explicit time complexity bounds for $\mathrm{SL}(n)$ and $\mathrm{GL}(n)$

Moving beyond the number of oracle calls, we now describe the running time of our algorithms in terms of the bitsize needed to describe the vector v and the action π. For concreteness, we restrict to *homogeneous, polynomial* actions of $\mathrm{GL}(n)$, i.e., those for which there is a degree d such that entries of the map π are homogeneous polynomials of degree d. This important class includes the setting studied by Hilbert in his seminal paper [33]. The results in this section extend readily to products of $\mathrm{GL}(n)$'s, a setting which captures all of the interesting examples discussed so far (tensor scaling, left-right action, simultaneous conjugation action, etc).

Up to isomorphism, irreducible polynomial representations of $\mathrm{GL}(n)$ can be specified by *partitions* of

length n, or nonincreasing vectors in $\mathbb{Z}^n_{\geqslant 0}$; the partition corresponding to an irreducible representation is called its *highest weight*. If the representation is of degree d, then the corresponding partition λ is a partion of (sums to) d.

We must be careful that we specify representations in such a way that the action can be computed. To this end, if λ is a partition, we take $\pi_\lambda : \mathrm{GL}(n) \to \mathrm{GL}(m)$ to be the unique irreducible representation of highest weight λ such that the standard basis of $\mathbb{C}^m$ is a *Gelfand-Tsetlin basis* for π_λ. The Gelfand-Tsetlin basis, as described in [55], is a well-studied basis for irreducible representations in which the entries of the map π_λ are polynomials with bounded, rational coefficients. Moreover, the group action and moment map can be computed in polynomial time when working in this basis.

A list of partitions $\lambda^1, \ldots, \lambda^s$ specifies the representation $\pi \cong \oplus_{i=1}^s \pi_{\lambda^i}$; up to isomorphism, every finite dimensional representation π of $\mathrm{GL}(n)$ can be specified this way. If π is such a representation, the input size $\langle \pi \rangle$ of π is defined to be $\langle \lambda^1 \rangle + \cdots + \langle \lambda^s \rangle$ where $\langle \lambda^i \rangle$ is the binary size of λ^i. We may assume that π is specified in unary without loss of generality, because the dimensions of the representations we study grow exponentially quickly in the entries of the λ^i.

For a vector $v \in \mathbb{C}^m$ with coordinates in $\mathbb{Q} + i\mathbb{Q}$, $\langle v \rangle$ refers to the total binary size of its entries. In [56, 57] it is shown that, for π, v specified as above and $g \in \mathrm{Mat}(n, \mathbb{Q} + i\mathbb{Q})$ specified in binary, $\pi(g)v$ and $\mu(v)$ can be computed in polynomial time. If ε is a rational number, $\langle \varepsilon \rangle$ refers to its size in binary.

We now define instances for the problems discussed in Section III for the case $\mathrm{SL}(n)$ and $\mathrm{GL}(n)$. Further, we assume any target spectrum p for the moment polytope membership problem has nonnegative, rational entries adding to d because if π is polynomial and homogeneous of degree d then every element of $\Delta(v)$ has this property. For the problem of norm minimization (equivalently, the null cone membership problem), we consider the restriction of π to the smaller group $\mathrm{SL}(n)$. This is without loss of generality because, unless $d = 0$, the capacity for homogeneous actions of $\mathrm{GL}(n)$ is always zero.

1) A tuple $(\pi = \oplus_{i=1}^s \pi_{\lambda^i}, v)$ is called an *instance of the null cone membership problem for* $\mathrm{SL}(n)$ if
   - $\pi : \mathrm{GL}(n) \to \mathrm{GL}(m)$ is a homogeneous, polynomial representation of $\mathrm{GL}(n)$ of degree d.
   - $v \in V = \mathbb{C}^m$ is a Gaussian integer vector.
2) A tuple (π, v, ε) is called an *instance of the scaling problem for* $\mathrm{SL}(n)$ if (π, v) is an instance of the null cone membership problem for $\mathrm{SL}(n)$ and $\varepsilon > 0$ is a rational number.
3) A tuple (π, v, p) is an *instance of moment polytope membership for* $\mathrm{GL}(n)$ if (π, v) is an instance of the null–cone membership problem for $\mathrm{SL}(n)$ and $p \in \mathbb{Q}^n_{\geqslant 0}$ is a vector with nonnegative, rational entries adding to d.
4) A tuple (π, v, p, ε) is an *instance of the* p*-scaling problem for* $\mathrm{GL}(n)$ if (π, v, p) is an instance of moment polytope membership over $\mathrm{GL}(n)$ and $\varepsilon > 0$ is rational number.

Remark V.1 (Degree versus dimension)**.** *We may assume that for our input representations $\oplus_{i=1}^s \pi_{\lambda^i}$ we have $\lambda^i_n = 0$ for some $i \in [s]$; this is without loss of generality because simultaneously translating each λ^i by an integer multiple of the all–ones vector simply shifts the entire moment polytope in $\mathbb{R}^n$ by the same vector. If some $\lambda^i_n = 0$, then the bound $d \leqslant mn$ follows from classical formulae for the dimensions of irreducible representations, which ensures that our bounds in the coming theorems are polynomial in $\langle v \rangle, \langle \pi \rangle$.*

By deriving capacity lower bounds for vectors of bounded bit complexity, we prove in the full version that Theorem IV.6 implies the following time bounds for Problem III.2.

Theorem V.2 (First order algorithm for scaling in terms of input size)**.** *Let (π, v, ε) be an instance of the scaling problem over $\mathrm{SL}(n)$ such that $0 \in \Delta(v)$ and every entry of v is at most M in absolute value. Algorithm IV.1 with a number of iterations at most*

$$T = O\left(\frac{d^2}{\varepsilon^2} mn^3 d \log(Mmnd)\right)$$

returns a group element $g \in \mathrm{SL}(n)$ such that $\|\mu(\pi(g)v)\|_F \leqslant \varepsilon$. By Remark V.1, there is a $\mathrm{poly}(\langle v \rangle, \langle \pi \rangle, \varepsilon^{-1})$ time algorithm to solve the scaling problem (Problem III.2) for $\mathrm{SL}(n)$.

We also show a concrete version of Theorem IV.8 on norm minimization.

Theorem V.3 (Second order algorithm for norm minimization in terms of input size)**.** *Let (π, v, ε) be an instance of the scaling problem for $\mathrm{SL}(n)$ such that $0 \in \Delta(v)$ and every entry of v is at most M in absolute value. Let γ denote the weight margin $\gamma(\pi)$. Then our second order g-convex optimization algorithm, applied to a suitably regularized objective function, with*

a number of iterations at most

$$T = O\left(\frac{d\sqrt{n}}{\gamma} m n^3 d \log^2\left(\frac{Mmnd}{\varepsilon}\right)\right)$$

returns a group element $g \in G$ *such that* $\log\|\pi(g)v\| \leqslant \log \operatorname{cap}(v) + \varepsilon$.

By Remark V.1, there is an algorithm that solves the norm minimization problem for $SL(n)$ *in time* $\mathrm{poly}(\langle v\rangle, \langle\pi\rangle, \gamma^{-1}, \log\frac{1}{\varepsilon})$*, which is shown in the full version to be at most* $\mathrm{poly}(\langle v\rangle^n, \langle\pi\rangle^n, \log\frac{1}{\varepsilon})$*. Note that* $\langle\pi\rangle^n, \langle v\rangle^n$ *are polynomial in the input size if the group is fixed.*

Corollary IV.5 implies that both the first and second order algorithm result in a null cone membership algorithm with polynomial dependence on γ^{-1}; the tradeoffs are discussed further in the full version. Using the bound $\gamma \geqslant n^{-1}d^{-n}$ from Table I also gives a bound which is polynomial if n is fixed.

Corollary V.4 (Algorithm for null cone membership problem in terms of input size)**.** *There is an algorithm to solve the null cone membership problem for* $SL(n)$ *in time* $\mathrm{poly}(\langle v\rangle, \langle\pi\rangle, \gamma^{-1})$*, which is at most* $\mathrm{poly}(\langle v\rangle^n, \langle\pi\rangle^n)$.

In the important setting when n is constant, the above corollary asserts that our second order algorithm solves the null-cone problem for $GL(n)$ in deterministic polynomial time. Prior to this result, the only known polynomial time algorithms for this class of null-cone problems were given by the use of quantifier elimination (which is completely impractical) and, more recently, by Mulmuley in [20, Theorem 8.5] through a purely algebraic approach. Mulmuley constructs a circuit which encodes a generating set of invariants for the ring of invariants of the corresponding action, and then invokes previous results on Polynomial Identity Testing to obtain an algorithm for the null-cone problem.

Finally, we apply Theorem IV.7 to obtain a randomized algorithm for the p-scaling problem for $GL(n)$. Note that we consider the full group $GL(n)$ rather than $SL(n)$ as in the null cone membership problem.

Theorem V.5 (First-order randomized algorithm for p-scaling in terms of input size)**.** *Let* (π, v, p, ε) *be an instance of the moment polytope problem for* $GL(n)$ *such that* $p \in \Delta(v)$ *and every entry of* v *is at most* M *in absolute value. Then our first order algorithm for* p*-scaling with a randomized starting point and a number of iterations at most*

$$T = O\left(\frac{d^2}{\varepsilon^2} m n^5 d \log(Mmnd)\right).$$

outputs $g \in GL(n)$ *such that* $\|\operatorname{spec}(\mu(\pi(gg_0)v) - p\|_2 \leqslant \varepsilon$ *with probability at least* $1/2$*. By Remark V.1, there is a randomized algorithm for the* p*-scaling problem for* $GL(n)$ *that runs in time* $\mathrm{poly}(\langle v\rangle, \langle\pi\rangle, \langle p\rangle, \varepsilon^{-1})$.

VI. Conclusion

This paper initiates a systematic development of a theory of *non-commutative* optimization, a setting which greatly extends ordinary (Euclidean) convex optimization. This very general setting captures a diverse set of problems, many non-convex, in different areas of CS, math, and physics. Several of them were solved efficiently for the first time using non-commutative methods; the corresponding algorithms also lead to solutions of purely structural problems and to many new connections between disparate fields. Our work points to intriguing open problems and suggests further research directions. We believe that extending this theory, namely understanding geodesic optimization better, is both mathematically and computationally fascinating; it provides a great meeting place for ideas and techniques from several very different research areas, and promises better algorithms for existing and yet unforeseen applications. We mention a few concrete challenges:

1) Is the null cone membership problem for general group actions in P? A natural intermediate goal is to prove that they are in NP $\cap$ coNP. The duality theory explained in this paper makes such a result likely. The same question may be asked about the moment polytope membership problem for general group actions [57].
2) Can we find more general classes of problems or group actions where our algorithms converge in polynomial time? In view of the complexity parameters we have identified, it is of particular interest to understand in which cases the *weight margin* is only inverse polynomially rather than exponentially small.
3) Interestingly, when restricted to the commutative case discussed in Section III, our algorithms' guarantees do not match those of cut methods (in the spirit of the ellipsoid algorithm). Can we extend non-commutative/geodesic optimization to include cut methods (in the spirit of the el-

lipsoid algorithm), as well as interior point methods? The foundations we lay in extending first and second order methods to the non-commutative case makes one optimistic that similar extensions are possible of other methods in standard convex optimization.

4) Can geodesic optimization lead to new or different efficient algorithms in combinatorial optimization? We know that it captures known algorithms like bipartite matching (and more generally matroid intersection). How about perfect matching in general graphs – is the Edmonds polytope a moment polytope of a natural group action?
5) Can geodesic optimization lead to new or different efficient algorithms in algebraic complexity and derandomization? We know that it captures PIT (polynomial identity testing) in non-commutative variables. Is the classical PIT problem a null cone membership problem for some group action? Can we identify the required generalization and extend the current methods to solve it? Which algebraic varieties are *not* null cones of group actions?

References

[1] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson, "A deterministic polynomial time algorithm for non-commutative rational identity testing," in *Proceedings of the Symposium on Foundations of Computer Science (FOCS 2016)*. IEEE, 2016, pp. 109–117.

[2] ——, "Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling," in *Proceedings of the Symposium on the Theory of Computing (STOC 2017)*. ACM, 2017, pp. 397–409.

[3] P. Bürgisser, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, "Alternating minimization, scaling algorithms, and the null-cone problem from invariant theory," in *Proceedings of Innovations in Theoretical Computer Science (ITCS 2018)*, 2017.

[4] C. Franks, "Operator scaling with specified marginals," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 190–203.

[5] T. C. Kwok, L. C. Lau, Y. T. Lee, and A. Ramachandran, "The Paulsen problem, continuous operator scaling, and smoothed analysis," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 182–189.

[6] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson, "Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing," in *Proceedings of the Symposium on the Theory of Computing (STOC 2018)*, 2018, pp. 172–181.

[7] P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, "Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS 2018)*. IEEE, 2018, pp. 883–897.

[8] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*, ser. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2000, vol. 1.

[9] M. B. Cohen, A. Madry, D. Tsipras, and A. Vladu, "Matrix scaling and balancing via box constrained Newton's method and interior point methods," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017, pp. 902–913.

[10] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson, "Much faster algorithms for matrix scaling," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017, pp. 890–901.

[11] G. Ivanyos, Y. Qiao, and K. Subrahmanyam, "Non-commutative Edmonds' problem and matrix semi-invariants," *Computational Complexity*, vol. 26, no. 3, pp. 717–763, 2017.

[12] H. Derksen and V. Makam, "Polynomial degree bounds for matrix semi-invariants," *Advances in Mathematics*, vol. 310, pp. 44–63, 2017.

[13] G. Ivanyos, Y. Qiao, and K. Subrahmanyam, "Constructive non-commutative rank computation is in deterministic polynomial time," in *Proceedings of Innovations in Theoretical Computer Science (ITCS 2017)*, 2017.

[14] A. Knutson and T. Tao, "The honeycomb model of $\mathrm{GL}_n(\mathbb{C})$ tensor products I: Proof of the saturation conjecture," *Journal of the American Mathematical Society*, vol. 12, no. 4, pp. 1055–1090, 1999.

[15] K. D. Mulmuley, H. Narayanan, and M. Sohoni, "Geometric complexity theory III: on deciding nonvanishing of a Littlewood–Richardson coefficient," *Journal of Algebraic Combinatorics*, vol. 36, no. 1, pp. 103–110, 2012.

[16] P. Bürgisser and C. Ikenmeyer, "A max-flow algorithm for positivity of Littlewood-

Richardson coefficients," *Discrete Mathematics & Theoretical Computer Science*, vol. DMTCS Proceedings vol. AK, 21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2009), 2009.

[17] L. Hamilton and A. Moitra, "The Paulsen problem made simple," in *Proceedings of Innovations in Theoretical Computer Science (ITCS 2019)*, 2018.

[18] V. Kabanets and R. Impagliazzo, "Derandomizing polynomial identity tests means proving circuit lower bounds," *Computational Complexity*, vol. 13, no. 1-2, pp. 1–46, 2004.

[19] K. D. Mulmuley, "Geometric complexity theory V: Equivalence between blackbox derandomization of polynomial identity testing and derandomization of Noether's normalization lemma," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 2012, pp. 629–638.

[20] K. Mulmuley, "Geometric complexity theory V: Efficient algorithms for Noether normalization," *Journal of the American Mathematical Society*, vol. 30, no. 1, pp. 225–309, 2017.

[21] L. Gurvits and P. N. Yianilos, "The deflation-inflation method for certain semidefinite programming and maximum determinant completion problems," *Technical Report, NECI*, 1998.

[22] L. Gurvits, "Classical complexity and quantum entanglement," *Journal of Computer and System Sciences*, vol. 69, no. 3, pp. 448–484, 2004.

[23] ——, "Hyperbolic polynomials approach to van der Waerden/Schrijver-Valiant like conjectures: sharper bounds, simpler proofs and algorithmic applications," in *Proceedings of the thirty-eighth annual ACM Symposium on Theory of Computing*. ACM, 2006, pp. 417–426.

[24] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[25] L. Ness and D. Mumford, "A stratification of the null cone via the moment map," *American Journal of Mathematics*, vol. 106, no. 6, pp. 1281–1329, 1984.

[26] B. Kostant, "On convexity, the Weyl group and the Iwasawa decomposition," *Ann. scient. E.N.S*, vol. 6, pp. 413–455, 1973.

[27] M. F. Atiyah, "Convexity and commuting Hamiltonians," *Bulletin of the London Mathematical Society*, vol. 14, no. 1, pp. 1–15, 1982.

[28] V. Guillemin and S. Sternberg, "Convexity properties of the moment mapping," *Inventiones mathematicae*, vol. 67, pp. 491–513, 1982.

[29] F. Kirwan, "Convexity properties of the moment mapping, III," *Inventiones mathematicae*, vol. 77, no. 3, pp. 547–552, 1984.

[30] H. Derksen and J. Weyman, *An introduction to quiver representations*. American Mathematical Society, 2017, vol. 184.

[31] W. Fulton, "Eigenvalues, invariant factors, highest weights, and Schubert calculus," *Bulletin of the American Mathematical Society*, vol. 37, no. 3, pp. 209–249, 2000.

[32] M. Brion, "Sur l'image de l'application moment," in *Séminaire d'algebre Paul Dubreil et Marie-Paule Malliavin*, ser. Lecture Notes in Mathematics. Springer, 1987, vol. 1296, pp. 177–192.

[33] D. Hilbert, "Über die vollen Invariantensysteme," *Math. Ann.*, vol. 42, pp. 313–370, 1893.

[34] D. Mumford, *Geometric invariant theory*. Springer-Verlag, 1965.

[35] H. Derksen and G. Kemper, *Computational invariant theory*. Springer, 2015.

[36] B. Sturmfels, *Algorithms in invariant theory*. Springer, 2008.

[37] G. Kempf and L. Ness, "The length of vectors in representation spaces," in *Algebraic geometry*. Springer, 1979, pp. 233–243.

[38] L. G. Khachiyan, "A polynomial algorithm in linear programming," in *Doklady Academii Nauk SSSR*, vol. 244, 1979, pp. 1093–1096.

[39] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proceedings of Symposium on the Theory of Computing (STOC 1984)*. ACM, 1984, pp. 302–311.

[40] L. Gurvits, "Combinatorial and algorithmic aspects of hyperbolic polynomials," 2004.

[41] M. Singh and N. K. Vishnoi, "Entropy, optimization and counting," in *Proceedings of the Symposium on the Theory of Computing (STOC 2014)*. ACM, 2014, pp. 50–59.

[42] D. Straszak and N. K. Vishnoi, "Computing maximum entropy distributions everywhere," in *Proceedings of Machine Learning Research, 32nd Annual Conference on Learning Theory*, 2019.

[43] R. Raz and A. Shpilka, "Deterministic polynomial identity testing in non commutative models," *Computational Complexity*, vol. 14, pp. 1–19, 2005.

[44] M. A. Forbes and A. Shpilka, "Explicit Noether normalization for simultaneous conjugation via polynomial identity testing," *Lecture Notes in Computer Science*, pp. 527–542, 2013.

[45] H. Derksen and V. Makam, "Algorithms for orbit closure separation for invariants and semi-invariants of matrices," 2018.

[46] V. M. Kravtsov, "Combinatorial properties of noninteger vertices of a polytope in a three-index axial assignment problem," *Cybernetics and Systems Analysis*, vol. 43, no. 1, pp. 25–33, 2007.

[47] N. Linial, A. Samorodnitsky, and A. Wigderson, "A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents," in *Proceedings of the Symposium on the Theory of Computing (STOC 1998)*, 1998, pp. 644–652.

[48] C. Udriste, *Convex functions and optimization methods on Riemannian manifolds*. Springer, 1994, vol. 297.

[49] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[50] H. Zhang and S. Sra, "First-order methods for geodesically convex optimization," in *Conference on Learning Theory*, 2016, pp. 1617–1638.

[51] H. Zhang, S. J. Reddi, and S. Sra, "Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds," in *Advances in Neural Information Processing Systems*, 2016, pp. 4592–4600.

[52] H. Sato, H. Kasai, and B. Mishra, "Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport," *SIAM Journal on Optimization*, vol. 29, pp. 1444–1472, 2019.

[53] H. Zhang and S. Sra, "Towards Riemannian accelerated gradient methods," 2018.

[54] F. C. Kirwan, *Cohomology of quotients in symplectic and algebraic geometry*. Princeton University Press, 1984, vol. 31.

[55] A. I. Molev, "Gelfand-Tsetlin bases for classical Lie algebras," in *Handbook of Algebra*. Elsevier, 2006, vol. 4, pp. 109–170.

[56] P. Bürgisser, "The computational complexity to evaluate representations of general linear groups," *SIAM Journal on Computing*, vol. 30, no. 3, pp. 1010–1022, 2000.

[57] P. Bürgisser, M. Christandl, K. D. Mulmuley, and M. Walter, "Membership in moment polytopes is in NP and coNP," *SIAM Journal on Computing*, vol. 46, no. 3, pp. 972–991, 2017.