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
Avi Wigderson

Overview

- Study algorithmic questions about the action of **groups**, such as the group of invertible matrices, on **vector spaces**.
- **Primary goal:** provide a unified framework for optimizing the ℓ_2 **norm** and its **gradient** over group orbits.
- Generalize familiar **first and second order methods** to work in the **non-Euclidean geometry** of the group.

Outline

1. Definitions and examples of group actions
2. Problem statements and results
3. The algorithms
4. Open questions

Group actions

Group actions

- **Group** such as

$$G = \mathrm{GL}(n) = \{\text{invertible } n \times n \text{ matrices}\}.$$

  or $G =$ diagonal matrices, or $G = \mathrm{GL}(n) \times \mathrm{GL}(n)$
- **Action** on vector space $V (= \mathbb{C}^m)$:

$$\text{homomorphism } G \to \mathrm{GL}(V) \quad (m \times m \text{ invertible matrices}),$$

- g acting on v written

$$g \cdot v.$$

Conjugation

$G = \mathrm{GL}(n)$, $V = \mathrm{Mat}(n)$,

$$\boxed{g \cdot A = gAg^{-1}}$$

More examples

Operator scaling: $G = \mathrm{GL}(n) \times \mathrm{GL}(n)$, $V = \mathrm{Mat}(n)^k$ by

$$\boxed{(g, h) \cdot (A_1, \ldots, A_k) = gA_1h^T, \ldots, gA_kh^T.}$$

Tensor scaling: $G = \mathrm{GL}(n)^3$, $V = (\mathbb{C}^n)^{\otimes 3}$

$$\boxed{(g_1, g_2, g_3) \cdot |\phi\rangle = g_1 \otimes g_2 \otimes g_3 \, |\phi\rangle}$$

Norm optimization

Given a vector $\boldsymbol{v} \in V$, compute

$$\inf_{g \in G} \|g \cdot v\|.$$

Many surprising applications!

- **Combinatorics:** approximating the permanent [LSW98]
- **Functional analysis:** Brascamp-Lieb inequalities [CCT05]
- **Machine learning:** radial isotropic position [MH13]
- **Polynomial identity testing:** noncommutative identity testing [GGOW16]
- **Quantum information:** one body quantum marginal problem [BFGGOW18]
- **Computational invariant theory:** null cone problem

Example: perfect matchings and matrix scaling

Let

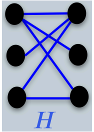
$$G = \{\text{pairs of diagonal matrices with det } 1\}$$

act on matrices A by $(X, Y) \cdot A = XAY$.

Ancient theorem

1. H has a perfect matching $\iff$
2. $\boxed{\inf_{(X,Y)\in G} \|XA_HY\|_F > 0}$ $\iff$
3. exist X, Y diagonal with $B = XA_HY$ doubly stochastic*:

$$\boxed{\operatorname{diag} BB^T = I, \operatorname{diag} B^TB = I.}$$



$$A_H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Why? $\nabla\|XAY\|_F = (\operatorname{diag} AA^T - I, \operatorname{diag} A^TA - I)$! "row and column sums"

Noncommutative analogue of ancient theorem

Analogue of row and column sums: **gradient** of (log) norm.

For historical reasons, called **moment map**, written

$$\boldsymbol{\mu}(v) := \nabla_X \log \|e^X \cdot v\|.$$

Matrix scaling: $\boldsymbol{\mu}(A) = \frac{1}{\|A\|_F^2}(\operatorname{diag} AA^T - I, \operatorname{diag} A^T A - I)$

Conjugation: $\boldsymbol{\mu}(A) = \frac{1}{\|A\|_F^2}(AA^T - A^T A)$

Ancient theorem

1. $\inf_{(X,Y)\in G} \|XA_H Y\|_F > 0 \iff$
2. A_H has (approx) doubly stochastic scalings $\iff$
3. H has perfect matching.

Kempf-Ness/Hilbert Mumford

1. $\inf_{g\in \boldsymbol{G}} \|g \cdot \mathbf{v}\| > 0. \iff$
2. $\inf_{g\in \boldsymbol{G}} \|\boldsymbol{\mu}(g \cdot \mathbf{v})\| = 0$.
3. $\iff \exists$ homogeneous invariant polynomial nonzero on v.

Problem statements and results

Back to the problems

Set $F(g) = \log \|g \cdot v\|$, and set $\mathrm{OPT} := \inf_{g \in G} F(g)$.

Norm optimization

Given v, produce g^* with $\boxed{F(g^*) \leq \mathrm{OPT} + \varepsilon}$ or determine that $\mathrm{OPT} = -\infty$.

$\mathrm{poly}(\log(1/\varepsilon))$ algorithm for special case; [AGLOW '17], algebraic algorithms for decision version [DM '19, IQS '17].

While we want to approximately optimize F, often the easier task of solving $\nabla F = \mu \approx 0$ *is still quite useful.*

Scaling

Given v and $\varepsilon > 0$, produce g with $\boxed{\|\mu(g \cdot v)\| < \varepsilon}$ or conclude that $\mathrm{OPT} = -\infty$.

$\mathrm{poly}(1/\varepsilon)$ time for operators [GGOW16], tensors [BFGGOW18]

The commutative case: Polynomial optimization

Suppose $\boldsymbol{p}$ is a Laurent polynomial $\boldsymbol{p}$ with nonnegative coefficients.

Ancient theorem

$$\inf_{x_i > 0} p(x) > 0 \iff 0 \in \mathrm{conv}(\Omega),$$

$\Omega \subset \mathbb{Z}^n$, set of exponents in polynomial.

Easy to optimize, but what about with **oracle access** to p, ∇p?

Weight margin Γ; Weight norm N

Γ : The closest the convex hull of a subset of Ω can come to the origin without containing it.

N : Maximum ℓ_2 norm of element of Ω.

[SV17:] can optimize in $\mathrm{poly}(1/\Gamma, N, \log(1/\varepsilon))$. with oracle access.

Contributions

Before our work, ad hoc range of algebraic/optimization algorithms.

New work implies all others, + new efficient algorithms

First order algorithm [BFGOWW 19]

Given oracle access to μ, outputs g with $\boxed{\|\mu(g \cdot v)\| \leq \varepsilon}$ in time $\mathrm{poly}(N, \mathrm{OPT}, 1/\varepsilon)$ or concludes that $\mathrm{OPT} = -\infty$.

Second order algorithm [BFGOWW 19]

Given oracle access to μ, Hessian, outputs g with $\boxed{\log \|g \cdot v\| \leq \mathrm{OPT} + \varepsilon}$ in time $\mathrm{poly}(1/\Gamma, N, \mathrm{OPT}, \log(1/\varepsilon))$ or concludes that $\mathrm{OPT} = -\infty$.

$|\mathrm{OPT}| \leq \mathrm{poly}$ for reasonable input models.

Size of $1/\Gamma$ explains previous **hard/easy cases:**

$\leq n^{3/2}$ for operator scaling, conjugation, $\geq 2^{n/3}$ for tensor scaling.

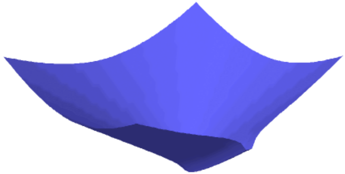
Algorithms

Geodesic convexity

Set $F(g) = \log \|g \cdot v\|$.

$F(e^X)$ **not convex** in Hermitian X!

but $F(e^{tX})$ is convex in t, i.e. $F(e^X)$ convex along **lines**!

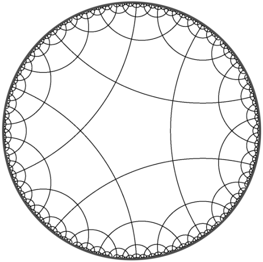


$F(e^X)$ for a subspace of 2×2 matrices

Geodesics:

analogues of lines in a non-Euclidean space. In G they are of the form

$$\boxed{e^{tX}g \text{ for } X \text{ hermitian}}$$



hyperbolic plane

Then F **geodesically convex**: convex along geodesics.

Geodesic gradient descent for scaling

Follow steepest **geodesic** at each step: steepest is

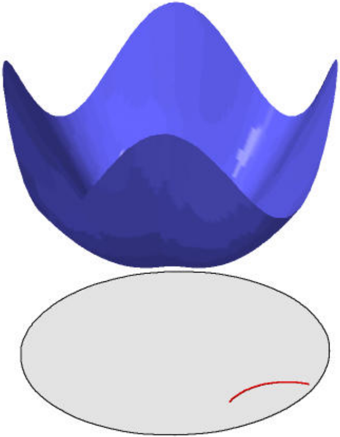
$$\boxed{\nabla_X F(e^X g) = \boldsymbol{\mu}(g \cdot v)}.$$

moment map = geodesic gradient!

Algorithm

Initially $g = I$, step size η.
For $i = 1 \ldots T$,

- Set $H = \boldsymbol{\mu}(g \cdot v)$
- $g \leftarrow e^{-\eta H} g$.



Analysis

We want to show that at some iteration, the geodesic gradient

$$\boldsymbol{\mu}(g \cdot v)$$

is small.

F is N-smooth

Second derivative bounded along geodesics:

$$\partial_t^2 F(e^{tX} g) \leq \mathbf{N}$$

for unit norm X

Standard analysis carries over!

Theorem

Take $\eta = 1/\mathbf{N}$, and $T \geq \frac{2\mathbf{N}}{\varepsilon^2}|\mathbf{OPT}|$, then at some step $\|\boldsymbol{\mu}(g \cdot v)\| \leq \varepsilon$.

Second order:

Trust region method: consider

$$Q(X) \text{ second order approx for } \mathbf{F}(e^X g).$$

Algorithm

Set $g = I$. For $i = 1...T$,

- Choose Hermitian H to minimize $Q(H)$ subject to $\|H\|_F \leq \eta$.
- Set $g \leftarrow e^H g$.

Second order analysis

Say F satisfies diameter bound D if

$$\inf_{\|X\|_F \leq D} F(e^X) \leq \mathrm{OPT} + \varepsilon.$$

Standard; [AGLOW17, CMTV17]

F can be regularized such that the algorithm takes $\mathrm{poly}(\log(1/\varepsilon), D, \mathrm{OPT})$ time.

Diameter bounds

Diameter bounded for large weight margin! $D \leq \mathrm{poly}(1/\boldsymbol{\Gamma})$.

Moment polytopes

Analogue of (r, c)-scaling; ask that μ take prescribed values.

μ takes value in **Hermitian matrices**, but...

Surprising and beautiful theorem [Bri87, NM84]

Eigenvalues of $\mu(g \cdot v)$ range over a convex polytope $\Delta(v)$!

$\Delta(v)$ can have exponentially many facets and vertices; examples include **polymatroids, matching polytopes, permutahedra**.

Weak moment polytope membership

Given v, Decide if $p \in \Delta$ or p at least ε-far from $\Delta(v)$.

Our work gives a $\text{poly}(1/\varepsilon)$ time algorithm for weak membership.
To put decision problem in P, need $\text{poly}(\log(1/\varepsilon))$!

Open problems

Very easy optimization algorithms seem to carry over: **alternating minimization, geodesic gradient descent, trust regions.**

What about the more powerful algorithms?

- Geodesic ellipsoid method? There is one [R18], but oracle calls take forever.
- Geodesic interior point methods?

Solve norm minimization in $\mathrm{poly}(\log(1/\varepsilon))$ time?

Thanks!