# Capacity and Invariant Theory

## Polynomial Capacity: Theory, Applications, Generalizations

Jonathan Leake

Technische Universität Berlin

February 18th, 2021

# Outline

1. Last time
   - Matrix scaling
   - Operator scaling
2. Null-cone problem
   - Motivation
   - The moment map and moment polytope
   - Connection to capacity
   - Scaling-type algorithm
3. Further questions

# Outline

1. Last time
   - Matrix scaling
   - Operator scaling
2. Null-cone problem
   - Motivation
   - The moment map and moment polytope
   - Connection to capacity
   - Scaling-type algorithm
3. Further questions

# The matrix scaling problem

Let $M$ be an $m \times n$ matrix with $\mathbb{R}_+$ entries, and fix $\boldsymbol{r} \in \mathbb{R}_+^m$ and $\boldsymbol{c} \in \mathbb{R}_+^n$.

**Definition:** A **scaling** of $M$ is given by multiplying $M$ on the left and right by diagonal matrices with positive entries:

$$\text{scaling} = AMB \quad \implies \quad (AMB)_{ij} = a_{ii} m_{ij} b_{jj}.$$

**Question:** Given $M$, do there exist such $A, B$ such that the row sums and column sums of $AMB$ are $\boldsymbol{r}$ and $\boldsymbol{c}$ respectively?

**Sinkhorn's algorithm** is a very simple iterative algorithm for $M_t$. For $\boldsymbol{r} = \boldsymbol{c} = \mathbf{1}$ (doubly stochastic scaling), the algorithm is:

1. Scale the columns so that col-sums$(M_{t+1}) = \mathbf{1}$.
2. Scale the rows so that row-sums$(M_{t+2}) = \mathbf{1}$ (changes col sums).
3. Repeat iterations until $M_t$ is almost doubly stochastic.

Keep track of $M_t = \cdots A_6 A_4 A_2 M B_1 B_3 B_5 \cdots$, which gives $A$ and $B$.

**Almost doubly sotchastic** $\implies$ scalable to doubly stochastic.

# Why do we care about matrix scaling?

**Application:** Deterministic approximation to the permanent. **How?**

Given an $n \times n$ matrix $M$, set $\boldsymbol{r} = \boldsymbol{c} = \mathbf{1}$. Suppose we have obtained the matrices $A, B$ which scale $M$ to the correct row/column sums.

Since $AMB$ is doubly stochastic, we can use van der Waerden bound:

$$1 \geq \text{per}(AMB) \geq \frac{n!}{n^n} \geq e^{-n} \quad \text{(e.g., recall } \text{Cap}_{\mathbf{1}}(p) \geq p_{\mathbf{1}} \geq \frac{n!}{n^n} \text{Cap}_{\mathbf{1}}(p)\text{)}.$$

**Now:** $\text{per}(AMB) = \det(A) \, \text{per}(M) \det(B)$.

**Therefore:** $[\det(A) \det(B)]^{-1} \geq \text{per}(M) \geq e^{-n} [\det(A) \det(B)]^{-1}$.

This says that $\det(AB)^{-1}$ is an $e^n$-approximation to the permanent of $M$. (And a similar bound holds when $AMB$ close to doubly stochastic.)

**[Linial-Samorodnitsky-Wigderson '00]:** No capacity at the time, but the vdW bound was already proven by Egorychev and Falikman.

# The LSW algorithm

Given $M$, want to compute $A, B$ so that $AMB$ is almost doubly stochastic.

**Main algorithm steps:**

1. **Preprocessing:** Scale to get $M_1$ such that $\text{per}(M_1) \geq \frac{1}{n^n}$.
2. **Sinkhorn:** Apply iterative scaling until $\|\mathbf{1} - \boldsymbol{c}_t\|_2$ is small.
3. **Approximation:** $M_t$ is close to doubly stochastic $\implies \approx e^n$-approx.

**Output:** $A = A_2 A_4 A_6 \cdots$ and $B = B_1 B_3 B_5 \cdots$ and $\text{per}(M) \approx \det(AB)^{-1}$.

**Different "marginals":** Similar algorithm given in [LSW '00].

**General form of multiplicative iterative scaling algorithms:**

1. **Lower bound:** Only need "small" number of steps to get close to DS.
2. **Progress:** Apply Sinkhorn until "marginals" close to DS.
3. **Approximation:** Once close to DS, use vdW-type approximation.

This framework works in more general operator / tensor scaling setting.

# The operator scaling problem

Let $T$ be a linear operator from $m \times m$ matrices to $n \times n$ matrices which maps PSD matrices to PSD matrices.

**Definition:** A **scaling** of $T$ is given by PD matrices $A, B$:

$$\text{scaling} = A^{1/2} T(B^{1/2} X B^{1/2}) A^{1/2}, \quad \text{another PSD-preserving operator.}$$

**Question:** Given $T$, do there exist $A, B$ to scale to "doubly stochastic"?

**Doubly stochastic operator:** $T(I_m) = I_n$ and $T^*(I_n) = I_m$ ($\implies m = n$).

**Translated to matrices:** $M \cdot \mathbf{1} = \mathbf{1}$ and $M^* \cdot \mathbf{1} = \mathbf{1}$ (doubly stochastic).

**Gurvits-Sinkhorn algorithm:** Alternate scaling $T$ and $T^*$:

$$\cdots A_3^{1/2} A_1^{1/2}\, T\left(\cdots B_4^{1/2} B_2^{1/2}\, X\, B_2^{1/2} B_4^{1/2} \cdots\right) A_1^{1/2} A_3^{1/2} \cdots$$

The $A_i$ matrices scale to $T(I_n) = I_n$, the $B_j$ matrices scale to $T^*(I_n) = I_n$.

# Why do we care about operator scaling?

**Recall:** $T$ is almost scalable to DS iff rank-nondecreasing.

**CP operator:** $T(X) = \sum_{k=1}^{\ell} M_k^* X M_k \implies T^*(X) = \sum_{k=1}^{\ell} M_k X M_k^*$.

**Why do we care about rank non-decreasing?** Equivalent properties (see [Garg-Gurvits-Oliveira-Wigderson '15], Theorem 1.4):

1. $\mathrm{rank}(T(X)) \geq \mathrm{rank}(X)$ for all $X \succ 0$.
2. For some $B_1, \ldots, B_\ell$, the matrix $\sum_{k=1}^{\ell} B_k \otimes M_k$ is non-singular.
3. For some $d$, the polynomial $\det\left(\sum_{k=1}^{\ell} X_k \otimes M_k\right)$ is not identically 0 where $X_k$ is a $d \times d$ matrix of variables.
4. The "polynomial" $\mathrm{Det}\left(\sum_{k=1}^{\ell} M_k x_k\right)$ is not identically 0, where $x_1, \ldots, x_\ell$ are *non-commuting* variables (non-commutative "Det").
5. The tuple $(M_1, \ldots, M_\ell)$ is not in **null-cone** of left-right action of $\mathrm{SL}_n^2$.

**#4: (non-commutative) polynomial identity testing, (NC)PIT:**
When is the determinant of a matrix of linear forms identically zero?

**[Kabanets-Impagliazzo]:** Poly-time PIT $\implies$ complexity *lower* bounds.

# The general form of the algorithm

**Recall** the form, for some "measure of progress" $\mu$:

1. **Preprocess:** Scale to $T_1$ such that $\mu(T_1) \geq e^{-\text{poly}(n)}$.
2. **Iterations:** Iterate poly($n$) times, improving $\mu(T_t)$ multiplicatively by $1 + \frac{1}{O(\text{poly}(n))}$ each time based on "closeness of marginals".
3. **Approximation:** Once "marginals" are close to doubly stochastic, we can approximate / know $T$ is almost scalable.

**Matrix case:** $\mu =$ permanent. Could have also used $\mu = \text{Cap}_{\mathbf{1}}$, since $p$ is doubly stochastic iff $\text{Cap}_{\mathbf{1}}(p) = 1$ and $\text{Cap}_{\mathbf{1}}(p) \leq 1$ otherwise.

**Operator case:** $\mu =$ matrix capacity, $\text{Cap}(T) := \inf_{X \succ 0} \dfrac{\det(T(X))}{\det(X)}$.

**[Gurvits '04]:** The following are equivalent:

1. $\text{Cap}(T) > 0$.
2. $T$ is rank non-decreasing.
3. For all $\epsilon > 0$, we have $T_t(I_n) = I_n$ and $\|T_t^*(I_n) - I_n\|_F \leq \epsilon$ for $t \gg 0$.
4. For some $t$, we have $T_t(I_n) = I_n$ and $\|T_t^*(I_n) - I_n\|_F \leq \frac{1}{n+1}$.

# Outline

# The null-cone problem

Let $\pi : G \to \mathrm{GL}(V)$ be a representation of a group $G$ (i.e., $\pi$ is a group homomorphism and $V$ is a vector space).

**Definition:** An **orbit** of $v \in V$ is the set $\mathcal{O}_v := \{\pi(g)v : g \in G\} \subset V$.

**Definition:** The **null-cone** of $V$ or $\pi$ is the set $\{v : 0 \in \overline{\mathcal{O}_v}\}$.

**[Hilbert], [Mumford '65]:** $v$ is in the null-cone iff for every non-constant homogeneous $G$-invariant polynomial $p$ on $V$ we have $p(v) = 0$.

**E.g.:** $v$ in null-cone $\implies \pi(g_i)v \to 0 \implies p(v) = p(\pi(g_i)v) = p(0) = 0$.

**[Kempf-Ness '79]:** $v$ is *not* in the null-cone iff $\mu(w) = 0$ for some $w \in \overline{\mathcal{O}_v}$, where $\mu$ is the **moment map** of $\pi$.

**Moment map:** Something like the "gradient" of the action of $\pi$ at $g = \mathrm{id}$:

$$\text{``}\mu(w) = \nabla|_{X=0} \log \|\pi(e^X)w\|\text{''}.$$

**Convex programming:** $f = \|w\|$ attains minimum at $w_0$ iff $\nabla f(w_0) = 0$.

# Why do we care about the null-cone problem?

**Last slide:** Given $\pi : G \to \mathrm{GL(V)}$, the null-cone is the set $\{v : 0 \in \overline{\mathcal{O}_v}\}$.

**Operator scaling:** Let $G = \mathrm{SL}_n^2(\mathbb{C})$ acting on $V = (\mathbb{C}^{n\times n})^\ell$ given by

$$\pi(g,h)\cdot(M_1,\ldots,M_\ell) := (gM_1h^{-1},\ldots,gM_\ell h^{-1}).$$

**Recall (NC-PIT):** $(M_1,\ldots,M_\ell)$ in null-cone iff $\mathrm{Det}\left(\sum_{k=1}^{\ell} M_k x_k\right) \equiv 0$.

**Non-convex optimization:** $v$ in the null-cone iff $\inf_{g\in G}\|\pi(g)v\| = 0$.

Other less obvious applications (see [BFGOWW '19]):

- **Horn's problem:** Given vectors $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma} \in \mathbb{R}^n$, are there Hermitian matrices $A, B, C$ with these spectra such that $A + B + C = 0$?
- **Brascamp-Lieb:** Given linear maps $A_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ and $p_1,\ldots,p_m > 0$, is there a finite constant $C$ such that

$$\int_{\mathbb{R}^n}\prod_i f_i(A_i\boldsymbol{x})d\boldsymbol{x} \le C\cdot\prod_i \|f_i\|_{1/p_i}$$

  for all $f_i$? Cauchy-Schwarz, Hölder, Loomis-Whitney, ...

# The moment map and moment polytope

**Throughout:** Think $G = \mathrm{GL}_n(\mathbb{C})$ or $G = \mathbb{T}^n$ with $\pi : G \to \mathrm{GL}(V)$.

**Definition:** The moment map $\mu(v)$ for $v \in V$ is defined via

$$\langle H, \mu(v) \rangle := \partial_t|_{t=0} \log \|\pi(e^{tH})v\|,$$

and $\mu(v)$ is Hermitian for $\mathrm{GL}_n(\mathbb{C})$ or a real (diagonal) vector for $\mathbb{T}^n$.

**Idea:** $\mu(v)$ is the "gradient" of $\log \|\pi(e^X)v\|$ at $X = 0$.

**Moment polytope:** $\Delta(v) := \overline{\{\mathrm{eig}(\mu(w)) : w \in \mathcal{O}_v\}}$ is a convex polytope.

**Kempf-Ness:** $v$ *not* in null-cone iff $\mu(w) = 0$ for a $w \in \overline{\mathcal{O}_v}$ **iff** $\mathbf{0} \in \Delta(v)$.

**Recap:** The following solve the same problem.

1. Null-cone membership problem.
2. Polytope membership problem (for $\boldsymbol{x} = \mathbf{0}$)
3. Norm/gradient minimization problem
4. Scaling problem: find $g \in G$ which minimizes $\|\pi(g)v\|$.
5. Capacity minimization problem?

## The commutative case: $G = \mathbb{T}^n = (\mathbb{C}^\times)^n$

**Rep. theory:** Commutative $G \implies$ basis of simultaneous eigenvectors.

**Further:** Orthonormal basis $v_1, \ldots, v_n$ such that $\pi(\boldsymbol{g})v_k = \lambda_k(\boldsymbol{g})v_k$ and:

$$\lambda_k(\boldsymbol{g}) = \boldsymbol{g}^{\boldsymbol{\omega}_k} := \prod_{i=1}^{n} g_i^{\omega_{k,i}},$$

where $\boldsymbol{\omega}_k$ are fixed **integer** vectors independent of $\boldsymbol{g}$.

**Null-cone objective:** $\|\pi(\boldsymbol{g})v\|_2^2 = \left\| \sum_{k=1}^{n} c_k v_k \boldsymbol{g}^{\boldsymbol{\omega}_k} \right\| = \sum_{k=1}^{n} |c_k|^2 \cdot |\boldsymbol{g}|^{2\omega_k}$.

**Optimization:** $\inf_{\boldsymbol{g}\in\mathbb{T}} \|\pi(\boldsymbol{g})v\|_2^2 = \inf_{\boldsymbol{g}\in\mathbb{T}} \sum_{k=1}^{n} |c_k|^2 \cdot |\boldsymbol{g}|^{2\omega_k} = \inf_{\boldsymbol{x}>0} \sum_{k=1}^{n} |c_k|^2 \boldsymbol{x}^{2\omega_k}$.

This is essentially capacity. Abusing notation: $\mathrm{Cap}_{\mathbf{0}} \left(\sum_{k=1}^{n} |c_k|^2 \boldsymbol{x}^{2\omega_k}\right)$.

**So:** Null-cone optimization becomes "polynomial capacity". **What about moment map formulation (finding zero of gradient)?**

# The commutative case: $G = \mathbb{T}^n = (\mathbb{C}^\times)^n$

**Last slide:** $\|\pi(\boldsymbol{g})v\|_2^2 = \sum_{k=1}^n |c_k|^2 \cdot |\boldsymbol{g}|^{2\omega_k} \implies$ "capacity" problem.

**Moment map:** $\langle \boldsymbol{y}, \mu(v)\rangle = \partial_t|_{t=0} \log \|\pi(e^{t\boldsymbol{y}})v\|$. We have:

$$\langle \boldsymbol{e}_j, \mu(v)\rangle = \partial_t|_{t=0} \log \sum_{k=1}^n |c_k|^2 \cdot e^{2t\langle \boldsymbol{e}_j, \omega_k\rangle} = \frac{\sum_{k=1}^n |c_k|^2 \cdot 2\omega_{k,j}}{\sum_{k=1}^n |c_k|^2}.$$

**Therefore:** $\mu(v) = \frac{\sum_{k=1}^n |c_k|^2 \cdot 2\omega_k}{\sum_{k=1}^n |c_k|^2} \implies$ convex combination of $2\boldsymbol{\omega}_k$.

**Further:** Moment polytope $\Delta(v) = \overline{\{\mu(w) : w \in \mathcal{O}_v\}}$ is precisely the "Newton" polytope of the "polynomial" $\|\pi(\boldsymbol{g})v\|_2^2$. ($c_k$ vary, but not $\boldsymbol{\omega}_k$)

**Kempf-Ness:** $\mathrm{Cap}_{\mathbf{0}}\left(\sum_{k=1}^n |c_k|^2 \boldsymbol{x}^{2\omega_k}\right) > 0$ iff $\mathbf{0} \in \mathrm{Newt}\left(\sum_{k=1}^n |c_k|^2 \boldsymbol{x}^{2\omega_k}\right)$.

Already proven before via direct computation, entropy, etc. Also known in this case as **Farkas' lemma**.

# Invariant-theoretic capacity

**Last slide:** $\inf_g \|\pi(g)v\|$ is a capacity problem in the commutative case.

In more general cases, let's just **make this the definition**:

$$\mathrm{Cap}_{\mathbf{0}}(v) := \inf_{g \in G} \|\pi(g)v\|.$$

"Non-commutative" capacity, "invariant-theoretic" capacity, etc.

Also called **non-commutative geometric programming** since the commutative case captures unconstrained **geometric programming** (see [Bürgisser-Li-Nieuwboer-Walter '20]).

**Kempf-Ness:** $\mathrm{Cap}_{\mathbf{0}}(v) > 0$ iff $\mathbf{0}$ is in the moment polytope $\implies$ Generalization of the same statement for polynomial capacity.

**Recall:** $\inf_{\boldsymbol{y} \in \mathbb{R}^n} \log \sum_{k=1}^n |c_k|^2 e^{\langle \boldsymbol{y}, 2\omega_k \rangle}$ is a convex program. **Can we do the same thing to non-commutative capacity?**

Appears to be "no"... (but general capacity is still **geodesically convex**). **Is there a scaling-type algorithm?**

# Scaling-type algorithm

**Recall** the form, for some "measure of progress" $\mu$:

1. **Preprocess:** Scale to $T_1$ such that $\mu(T_1) \geq e^{-\text{poly}(n)}$.
2. **Iterations:** Iterate poly($n$) times, improving $\mu(T_t)$ each time based on "closeness of marginals".
3. **Approximation:** Once "marginals" are close to desired, we know $T$ is almost scalable.

**Now:** $\mu = \text{Cap}_{\mathbf{0}}$. Can we generalize this to the null-cone problem?

1. **"Preprocess":** Set $g_0 = \text{id}$.
2. **Iterations:** Geodesic gradient descent, Taylor approx, "trust-region" methods... **I.e.:** Natural analogs to convex Euclidean techniques.
3. **Approximation:** How close do we need to get before stopping?

**Approximation step is key to determine computational complexity.**

**Need:** Relationship between value of capacity and norm of moment map.

This will heavily depend on the action $\pi$, the group $G$, etc.

# Complexity of the action

**Theorem [BFGOWW '19]:** For $\|v\| = 1$, we have

$$1 - \frac{\|\mu(v)\|}{\gamma(\pi)} \leq [\mathrm{Cap}_{\mathbf{0}}(v)]^2 \leq 1 - \frac{\|\mu(v)\|^2}{4N(\pi)^2}.$$

**Corollary:** $\mathbf{0} \in \Delta(v)$ iff $\Delta(v)$ contains a point smaller than $\gamma(\pi)$.
(This $\gamma(\pi)$ *is* how close we must get before stopping.)

**Proof of corollary:** ($\implies$) Obvious. ($\impliedby$) Kempf-Ness.

**Definition:** The **weight norm** $N(\pi)$ for $G = \mathrm{GL}_n(\mathbb{C})$ is:

$$N(\pi) := \max_{U \subseteq V, \text{ irreducible}} \|\boldsymbol{\lambda}_U\|, \quad \text{where } \boldsymbol{\lambda}_U \text{ is highest weight vector of } U.$$

**Commutative case:** $\boldsymbol{\lambda}_U$ are the simultaneous eigenvalue weights $\boldsymbol{\omega}_k$.

**Definition:** The **weight margin** $\gamma(\pi)$ is the minimum distance between $\mathbf{0}$ and any subset of the $\boldsymbol{\lambda}_U$'s whose convex hull does not contain $\mathbf{0}$.

**Fun fact:** For real stable polynomials and $\mathbf{1}$, the matroidal support condition implies the "weight margin" cannot be very small.

# Weight margin examples

**Last slide:** The **weight margin** $\gamma(\pi)$ is the minimum distance between $\mathbf{0}$ and any subset of the $\boldsymbol{\lambda}_U$'s whose convex hull does not contain $\mathbf{0}$.

**Matrix scaling:** Action of $(\mathrm{S}\mathbb{T}^n)^2$ via left-right action on matrices. $\gamma(\pi) \geq \frac{1}{\mathrm{poly}(n)}$ via [Linial-Samorodnitsky-Wigderson '00].

**Operator scaling:** Action of $(\mathrm{SL}_n(\mathbb{C}))^2$ on $(M_1, \ldots, M_\ell)$ via simultaneous left-right action. $\gamma(\pi) \geq \frac{1}{\mathrm{poly}(n)}$ via [Gurvits '04], [GGOW '15].

**Tensor scaling for 3-tensors:** Action of $(\mathrm{GL}_n(\mathbb{C}))^3$ on 3-tensors. $\gamma(\pi) \leq 2^{-\mathrm{poly}(n)}$ via [Franks-Reichenbach '21] (the other day).

**Last result:** Negative result for this method. **Open:** Other methods?

**Real stable polynomials formulation:** Given a real stabnle polynomial with $\mathbf{1}$ not in its Newton polytope, how far away can Newton polytope be?

# Outline

1. Last time
   - Matrix scaling
   - Operator scaling
2. Null-cone problem
   - Motivation
   - The moment map and moment polytope
   - Connection to capacity
   - Scaling-type algorithm
3. Further questions

# Further questions

**How do we handle other points in the moment polytope besides 0?**

**Commutative case:** Just change the denominator exponent in capacity.

**Non-commutative case [BFGOWW '19]:** Need to "shift" all weight vectors: tensor $\pi$ with another representation.

**Entropic capacity:** Is there any relation between entropic capacity and non-commutative capacity? (There is in the commutative case.)

**Further:** Connection to statistic via maximum likelihood, see [Améndola-Kohn-Reichenbach-Seigal '20]. **Connection between all three?**

**Open:** Does any connection give better algo for 3-tensors?