Lecture 15: Scaling Algorithms

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

rafael.oliveira.teaching@gmail.com

March 8, 2021

Overview

- Nullcone Problem, and Scaling Algorithms
- Matrix Scaling & Analysis
- Crash Course on Convex Optimization
- Conjugation Action - Teaser
- Conclusion

Orbit Closure Problems

- G acts linearly on V
- *Orbit Closure Intersection*: Given two points $u, w \in V$, do their orbit closures intersect?

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w \neq \emptyset?$$

Orbit Closure Problems

- G acts linearly on V
- *Orbit Closure Intersection*: Given two points $u, w \in V$, do their orbit closures intersect?

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w \neq \emptyset?$$

- If $w = 0$, we get the *null cone* problem:

$$0 \in \overline{\mathcal{O}}_u?$$

$\{g_t\}_{t \in \mathbb{N}} \subset G$ s.t.

$$\lim_{t \to \infty} \|g_t \circ u\| = 0$$

Orbit Closure Problems

- G acts linearly on V
- *Orbit Closure Intersection*: Given two points $u, w \in V$, do their orbit closures intersect?

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w \neq \emptyset?$$

- If $w = 0$, we get the *null cone* problem:

$$0 \in \overline{\mathcal{O}}_u?$$

- Null-cone problem has its name from the definition that the *nullcone* is the set of elements that have zero in their orbit closure.

$$\mathcal{N} = \{u \in V \mid 0 \in \overline{\mathcal{O}}_u\}$$

Orbit Closure Problems

- G acts linearly on V
- *Orbit Closure Intersection*: Given two points $u, w \in V$, do their orbit closures intersect?
$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w \neq \emptyset?$$
- If $w = 0$, we get the *null cone* problem:
$$0 \in \overline{\mathcal{O}}_u?$$
- Null-cone problem has its name from the definition that the *nullcone* is the set of elements that have zero in their orbit closure.
$$\mathcal{N} = \{u \in V \mid 0 \in \overline{\mathcal{O}}_u\}$$
- *Orbit Closure Containment*: Given two points $u, w \in V$, does the orbit of u contain the orbit closure of w?
$$\overline{\mathcal{O}}_w \subset \overline{\mathcal{O}}_w?$$

Orbit Closure Problems

- G acts linearly on V
- *Orbit Closure Intersection*: Given two points $u, w \in V$, do their orbit closures intersect?

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w \neq \emptyset?$$

- If $w = 0$, we get the *null cone* problem:

$$0 \in \overline{\mathcal{O}}_u?$$

- Null-cone problem has its name from the definition that the *nullcone* is the set of elements that have zero in their orbit closure.

$$\mathcal{N} = \{u \in V \mid 0 \in \overline{\mathcal{O}}_u\}$$

- *Orbit Closure Containment*: Given two points $u, w \in V$, does the orbit of u contain the orbit closure of w?

$$\overline{\mathcal{O}}_w \subset \overline{\mathcal{O}}_{\cancel{w}\, u}?$$

- Given $u \in V$, is its orbit closed?

$$\mathcal{O}_u = \overline{\mathcal{O}}_u?$$

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.
- **[Hilbert-Mumford]**: orbit closure intersection

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w = \emptyset \Leftrightarrow p(u) \neq p(w) \quad \textbf{for some} \quad \underbrace{p \in \mathbb{C}[V]^G}_{\text{invariant}}$$

do not intersect

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.
- **[Hilbert-Mumford]**: orbit closure intersection

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w = \emptyset \Leftrightarrow p(u) \neq p(w) \ \ \textbf{for some} \ \ p \in \mathbb{C}[V]^G$$

1. $G = \mathbb{SL}(2),\ V = \mathbb{C}^d$ linear transformations of curves

   Discriminants, Catalecticants (and more)

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.
- **[Hilbert-Mumford]**: orbit closure intersection

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w = \emptyset \Leftrightarrow p(u) \neq p(w) \ \ \textbf{for some} \ \ p \in \mathbb{C}[V]^G$$

1. $G = \mathbb{SL}(2),\ V = \mathbb{C}^d$ linear transformations of curves

   Discriminants, Catalecticants (and more)

2. $G = \mathbb{SL}(n),\ V = \mathrm{Mat}(n)$ left multiplication

   Determinant

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.
- **[Hilbert-Mumford]**: orbit closure intersection

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w = \emptyset \Leftrightarrow p(u) \neq p(w) \quad \textbf{for some} \quad p \in \mathbb{C}[V]^G$$

1. $G = \mathbb{SL}(2),\ V = \mathbb{C}^d$ — linear transformations of curves

   Discriminants, Catalecticants (and more)

2. $G = \mathbb{SL}(n),\ V = \mathrm{Mat}(n)$ — left multiplication

   Determinant

3. $G = \mathbb{GL}(n),\ V = \mathrm{Mat}(n)$ — conjugation

   Trace polynomials.

Orbit Problems and Invariant Polynomials

- **[Hilbert 1893]**: *Nullcone* is the zero set of non-constant, homogeneous *invariant polynomials*.
- **[Hilbert-Mumford]**: orbit closure intersection

$$\overline{\mathcal{O}}_u \cap \overline{\mathcal{O}}_w = \emptyset \Leftrightarrow p(u) \neq p(w) \quad \textbf{for some} \quad p \in \mathbb{C}[V]^G$$

1. $G = \mathbb{SL}(2),\ V = \mathbb{C}^d$ linear transformations of curves

   Discriminants, Catalecticants (and more)

2. $G = \mathbb{SL}(n),\ V = \mathrm{Mat}(n)$ left multiplication

   Determinant

3. $G = \mathbb{GL}(n),\ V = \mathrm{Mat}(n)$ conjugation

   Trace polynomials.

4. $G = \mathbb{ST}(n) \times \mathbb{ST}(n),\ V = \mathrm{Mat}(n)$ row/column scaling

   Matching/Permutation monomials.

Scaling Algorithms

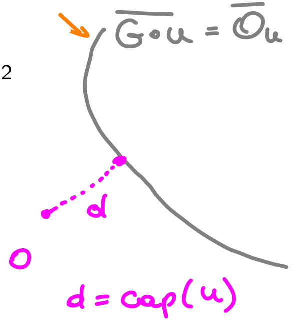
- Is there a geometric way to approach such problems?
- Would help us when invariants are hard to obtain

Scaling Algorithms

- Is there a geometric way to approach such problems?
- Would help us when invariants are hard to obtain
- When our vector space has an inner product, motivates the following optimization question:

$$\underline{\mathrm{cap}(u)} = \inf_{g \in G} \|g \circ u\|_2$$

Capacity of vector u



Scaling Algorithms

- Is there a geometric way to approach such problems?
- Would help us when invariants are hard to obtain
- When our vector space has an inner product, motivates the following optimization question:

$$\text{cap}(u) = \inf_{g \in G} \|g \circ u\|_2$$

- u in Nullcone iff $\text{cap}(u) = 0$

Matrix Scaling

$$\begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix}$$

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

Graphs without bipartite matching.

$$ST(n) = \left\{ \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{pmatrix} \;\middle|\; \prod_{i=1}^{n} a_i = 1 \right\}$$

$ST(n) \times ST(n) \cdot Mat(n)$

$(X, Y) \circ A \longmapsto XAY$

$$R_i(A) = C_j(A) = \frac{\|A\|_F^2}{n}$$

A is doubly-balanced :

$$R_i(A) = \sum_{j=1}^{n} |A_{ij}|^2$$ norm squared of i^{th} row

$$C_j(A) = \sum_{i=1}^{n} |A_{ij}|^2$$ norm squared of j^{th} column

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Graphs without bipartite matching.

- Distance to doubly-balanced:

$$db(A) = \sum_{i=1}^{n} \underbrace{\left(R_i(A) - \frac{\|A\|_F^2}{n} \right)^2}_{\text{how far ith row is from balanced condition}} +$$

$$+ \sum_{j=1}^{n} \underbrace{\left(C_j(A) - \frac{\|A\|_F^2}{n} \right)^2}_{\text{how far jth column is from balanced condition}}$$

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

Graphs without bipartite matching.

- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$ $\underline{\mathrm{db}}(R_\epsilon A C_\epsilon) \le \epsilon$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings R_ϵ, $C_\varepsilon$such that $\underline{\mathrm{ds}}(R_\epsilon A C_\epsilon) \leq \epsilon$.

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Graphs without bipartite matching.

- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings R_ϵ, $C_\varepsilon$such that $\mathrm{ds}(R_\epsilon A C_\epsilon) \leq \epsilon$.
- Null-cone problem for matrix scaling:
  - **Input:** $A \in \mathrm{Mat}(n)$
  - **Output:** Is A in the Nullcone of the matrix scaling action?
- How do these two relate?

they are the same problem!

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \text{Mat}(n)$ row/column scaling

Graphs without bipartite matching.

- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \text{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings $R_\epsilon, C_\varepsilon$ such that $\text{ds}(R_\epsilon A C_\epsilon) \leq \epsilon$.
- Null-cone problem for matrix scaling:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A in the Nullcone of the matrix scaling action?
- How do these two relate?
- Norm of a scaled element:

$$X = \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{pmatrix} \quad Y = \begin{pmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{pmatrix}$$

$$\|XAY\|_F^2 = \sum_{i,j=1}^{n} |A_{ij}|^2 x_i^2 y_j^2$$

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \text{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*

- Norm of a scaled element:

$$\|XAY\|_F^2 = \sum_{i,j=1}^{n} |A_{ij}|^2 x_i^2 y_j^2$$

$$\inf_{X,Y} \|XAY\|_F^2$$

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*

- Norm of a scaled element:
- Distance to doubly-balanced:

$$ds(A) = \sum_{i=1}^{n} \left(R_i(A) - \frac{\|A\|_F^2}{n} \right)^2 +$$

$$+ \sum_{j=1}^{n} \left(C_j(A) - \frac{\|A\|_F^2}{n} \right)^2$$

$$= \left\| \left(R_i(A) - \frac{\|A\|_F^2}{n} \right)_{i=1}^{n} \times \left(C_j(A) - \frac{\|A\|_F^2}{n} \right)_{j=1}^{n} \right\|_2^2$$

(norm of the vector above)

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*
- Norm of a scaled element:
- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings $R_\epsilon, C_\varepsilon$ such that $\mathrm{ds}(R_\varepsilon A C_\epsilon) \le \epsilon$.

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*
- Norm of a scaled element:
- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings R_ϵ, $C_\varepsilon$such that $\mathrm{ds}(R_\varepsilon A C_\epsilon) \leq \epsilon$.
- Normalized matrices

$$r_i(A)^{-1} = \frac{\|A\|_F}{\sqrt{n}\, R_i(A)^{1/2}}$$

A rows don't have same norm :

$$\underbrace{\left(\prod_{i=1}^{n} r_i\right)^{1/n} \begin{pmatrix} 1/r_1 & & \\ & \ddots & \\ & & 1/r_n \end{pmatrix}}_{\in ST(n)} A$$

all rows have same norm square

$$B = \underbrace{\left(\prod_{i=1}^{n} r_i\right)}_{c} \begin{pmatrix} 1/r_1 & & \\ & \ddots & \\ & & 1/r_m \end{pmatrix} A$$

$$R_i(B) = c^2 \sum_{j=1}^{n} \left(\frac{1}{r_i}\right)^2 \cdot |A_{ij}|^2$$

$$B_{ij} = c \cdot \frac{1}{r_i} \cdot A_{ij}$$

$$= c^2 \cdot \frac{1}{r_i^2} \cdot \underbrace{\sum_{j=1}^{n} |A_{ij}|^2}_{R_i(A)} = \frac{c^2 \cancel{R_i(A)}}{\underbrace{n \cancel{R_i(A)}}_{\|A\|_F^2}}$$

$$= \frac{c^2 \|A\|_F^2}{n}$$

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*

- Norm of a scaled element:
- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings $R_\epsilon, C_\varepsilon$such that $\mathrm{ds}(R_\varepsilon A C_\epsilon) \leq \epsilon$.
- Normalized matrices
- An ancient algorithm:

Alternating minimization

1. For T steps, repeat the following:
2. If $\mathrm{ds}(A) \leq \varepsilon$return A

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \mathrm{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*
- Norm of a scaled element:
- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings $R_\epsilon, C_\varepsilon$such that $\mathrm{ds}(R_\varepsilon A C_\epsilon) \leq \epsilon$.
- Normalized matrices
- An ancient algorithm:

Alternating minimization

1. For T steps, repeat the following:
2. If $\mathrm{ds}(A) \leq \varepsilon$return A
3. If $\mathrm{ds}(A) > \varepsilon$and A not row-normalized, then normalize rows of A

Matrix Scaling

- $G = \mathbb{ST}(n) \times \mathbb{ST}(n)$, $V = \text{Mat}(n)$ row/column scaling

  Nullcone: *Graphs without bipartite matching.*

- Norm of a scaled element:
- Distance to doubly-balanced:
- Matrix scaling problem:
  - **Input:** $A \in \text{Mat}(n)$, $\varepsilon > 0$
  - **Output:** can A be scaled to ϵ-doubly balanced? If yes, return scalings $R_\epsilon, C_\varepsilon$such that $\text{ds}(R_\varepsilon A C_\epsilon) \leq \epsilon$.
- Normalized matrices
- An ancient algorithm:

  Alternating minimization

  1. For T steps, repeat the following:
  2. If $\text{ds}(A) \leq \varepsilon$return A
  3. If $\text{ds}(A) > \varepsilon$and A not row-normalized, then normalize rows of A
  4. If $\text{ds}(A) > \varepsilon$and A not column-normalized, normalize columns of A

$B^{(0)} = A$

$B^{(t+1)} = R^{(t)} B^{(t)}$ (if $B^{(t)}$ was not row normalized)

$B^{(t+1)} = B^{(t)} C^{(t)}$ (if $B^{(t)}$ not column normalized)

Matrix Scaling - Analysis

A can be scaled to ϵ-doubly balanced iff support(A) has matching.

(there is an nice invariant polynomial P s.t. $P(A) \neq 0$)

Potential function : invariant polynomial P !

$\prod A_{ii} \neq 0$ in our original matrix.
$(=1)$

throughout algorithm we know that $\prod A_{ii}$ won't change!

Matrix Scaling - Analysis

Potential function $\prod_{i=1}^{n} A_{ii}$ $(=1)$

$$\Phi(B^{(t)}) = \|B^{(t)}\|_F^2$$

1) throughout $\underline{\prod_{i=1}^{n} B_{ii}^{(t)}}$ remains 1

$\therefore \|B^{(t)}\|_F^2 \geq 1$ (AM-GM)

2) initially we have a bound

on $\|B^{(0)}\|_F^2 = \|A\|_F^2 = C$ (based on size of input)

Matrix Scaling - Analysis

$$r_i(A) = \frac{\sqrt{n R_i(A)}}{\|A\|}$$

when normalize

$$B^{(t+1)} = \underbrace{\prod_{i=1}^{n} r_i(B^{(t)})}_{\alpha} \begin{pmatrix} 1/r_1 & & \\ & \ddots & \\ & & 1/r_n \end{pmatrix} B^{(t)}$$

$$\|B^{(t+1)}\|_F^2 = \alpha^2 \cdot \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{n} |B_{ij}^{(t+1)}|^2}_{\frac{1}{r_i^2} \cdot \sum_{j=1}^{n} |B_{ij}^{(t)}|^2}$$

$$\frac{\|B^{(t)}\|^2}{n \cancel{R_i(B^{(t)})}} \cdot \cancel{R_i(B^{(t)})} = \alpha^2 \|B^{(t)}\|^2$$

$$\|B^{(t+1)}\|^2 = \alpha^2 \|B^{(t)}\|^2$$

$$\alpha = \prod_{i=1}^{n} r_i \qquad \boxed{r_i = \frac{n R_i(B^{(t)})}{\|B^{(t)}\|^2}}$$

$$ds(B^{(t)}) > \varepsilon$$

''

$$\sum_{i=1}^{n} \left(R_i - \frac{\|B^{(t)}\|^2}{n} \right)^2 > \varepsilon$$

''

$$\boxed{\sum_{i=1}^{n} \left(r_i - 1 \right)^2 > \frac{\varepsilon n^2}{\|B^{(t)}\|^4}}$$

$r_i > 0 \qquad \sum_{i=1}^{n} r_i = n$

$$\sum_{i=1}^{n} (r_i - 1)^2 = \delta \leftarrow \delta\text{ - far from equality to } \perp$$

$$\prod_{i=1}^{n} r_i \leq \prod_{i=1}^{n} \left(1 + (r_i - 1)\right) \leq e^{\cancelto{0}{\sum r_i - 1} - \frac{\sum (r_i - 1)^2}{2}}$$

$$\boxed{1 - x \leq e^{x - x^2/2}}$$

$$\leq e^{-\frac{\sum (r_i - 1)^2}{2}} \leq e^{-\delta/2}$$

$\boxed{\leq \varepsilon}$

if $\boxed{db(B^{(t)}) > \varepsilon}$ then

$$\|B^{(t+1)}\| < e^{-\varepsilon/6} \cdot \|B^{(t)}\|$$

$$\|B^{(0)}\| \leq c \qquad \boxed{\|B^{(t)}\| \geq 1}$$

invariants
non
zero

Cannot happen indefinitely has to stop in

$$T = \frac{6}{\varepsilon} \cdot \log c \text{ steps.}$$

Outline what we did

non zero invariant $\Downarrow$ lower bd on norm

1) if there is invariant $P : \rho(A) \neq 0$ then $\|B^{(t)}\| \geq 1$

norm always decreases

2) if $ds(B^{(t)}) > \varepsilon$ then

$$\|B^{(t+1)}\| < e^{-\varepsilon/6} \cdot \|B^{(t)}\|$$

(norm decreases)

upper bd on initial norm

3) $\|B^{(0)}\| = c$ some c

Matrix Scaling - Thoughts

$\{B^{(t)}\}_{t \geq 0}$ descent sequence for $f(A) := \|A\|_F^2$

$f(B^{(t+1)}) \leq f(B^{(t)})$

$ds(A)$ large

Matrix Scaling - Thoughts

$$ds(A) = \sum_{i=1}^{n}\left(R_i(A) - \frac{\|A\|^2}{n}\right)^2 +$$

$$\sum_{j=1}^{n}\left(C_j(A) - \frac{\|A\|_F^2}{n}\right)^2$$

"norm of the gradient over

$f_A(X,Y) = \|XAY\|_F^2$ "

$$X = \begin{pmatrix} e^{x_1} & & & \\ & e^{x_2} & & \\ & & \ddots & \\ & & & e^{x_n} \end{pmatrix} \qquad Y = \begin{pmatrix} e^{y_1} & & \\ & \ddots & \\ & & e^{y_n} \end{pmatrix}$$

$$\det(X) = \det(Y) = 1 \iff \underline{\sum x_i = \sum y_i = 0}$$

$$\mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$$

$$f_A(x_1, \dots, x_n, y_1, \dots, y_n) = \|XAY\|_1^2 = \sum_{i,j} |A_{ij}|^2 \cdot e^{2(x_i + y_j)}$$

$$\left\langle \underline{\nabla f_A}, \underline{(r_1, \dots, r_n, c_1, \dots, c_n)} \right\rangle = \frac{d}{dt}\left(f_A\left(r_1 t, \dots, r_n t, c_1 t, \dots, c_n t\right)\right)\Big|_{t=0}$$

$$= \sum_{i,j} |A_{ij}|^2 \, 2 \cdot (r_i + c_j) \underbrace{\cancel{e^{2(r_i + c_j)t}}}_{1} = \langle (R, C), (r, c) \rangle$$

$$R = \left(R_1(A), \dots, R_n(A)\right)$$
$$C = \left(C_1(A), \dots, C_n(A)\right)$$

$$\underbrace{\nabla f_A}_{H \subset \mathbb{R}^{2n}} = \left(\underbrace{R_1(A), \dots, R_n(A)}_{\perp I}, \underbrace{C_1(A), \dots, C_n(A)}_{\perp I}\right)$$

$$\nabla f_A = \left(R_1(A) - \frac{\|A\|_F^2}{n}, \dots, R_n(A) - \frac{\|A\|_F^2}{n},\right.$$
$$\left. C_1(A) - \frac{\|A\|_F^2}{n}, \dots, C_n(A) - \frac{\|A\|_F^2}{n}\right)$$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \underline{\nabla f(\mathbf{a})}, \underline{\mathbf{b}} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

direction of derivative

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \nabla f(\mathbf{a}), \mathbf{b} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

- Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the matrix $\nabla^2 f(\mathbf{a}) \in \mathbb{R}^{n \times n}$ such that:

$$\langle \mathbf{c}, \nabla^2 f(\mathbf{a}) \cdot \mathbf{b} \rangle = \partial_s \partial_t f(\mathbf{a} + \mathbf{b} \cdot t + \mathbf{c} \cdot s)|_{s,t=0}$$

Convex Optimization Crash Course

- Function $f : \mathbb{R} \to \mathbb{R}$ convex iff $\dfrac{d^2}{dt^2} f(t) \geq 0$ for all $t \in \mathbb{R}$
- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff the univariate function $g_{\mathbf{a}}(t) = f(\mathbf{a}t + \mathbf{b})$ is convex for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$
- Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the vector $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ such that:

$$\langle \nabla f(\mathbf{a}), \mathbf{b} \rangle = \partial_t f(\mathbf{a} + \mathbf{b} \cdot t)|_{t=0}$$

- Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{a}$ is the matrix $\nabla^2 f(\mathbf{a}) \in \mathbb{R}^{n \times n}$ such that:

$$\langle \mathbf{c}, \nabla^2 f(\mathbf{a}) \cdot \mathbf{b} \rangle = \partial_s \partial_t f(\mathbf{a} + \mathbf{b} \cdot t + \mathbf{c} \cdot s)|_{s,t=0}$$

- Function $f : \mathbb{R}^n \to \mathbb{R}$ convex iff $\nabla^2 f(\mathbf{a}) \succeq 0$ for all $\mathbf{a} \in \mathbb{R}^n$

Gradient descent (with line search)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.

Gradient descent (with line search)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$

Gradient descent (with line search)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $g_k : \mathbb{R} \to \mathbb{R}$ be the function $g_k(t) = f(\underline{\mathbf{x}^{(k)} + t\nabla^{(k)}})$

going in direction of gradient

Gradient descent (with line search)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $g_k : \mathbb{R} \to \mathbb{R}$ be the function $g_k(t) = f(\mathbf{x}^{(k)} + t\nabla^{(k)})$
  - Let $\mathbf{x}^{(k+1)} = \text{argmin}_t g_k$

Gradient descent (with fixed step size)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.

Gradient descent (with fixed step size)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$ $\quad \eta > 0$

Gradient descent (with fixed step size)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$

Gradient descent (with fixed step size)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \eta \cdot \cancel{\|}\nabla^{(k)}\cancel{\|_2}$

Gradient descent (with fixed step size)

- **Input:** convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$, initial point $\mathbf{a} \in \mathbb{R}^n$
- **Output:** Find point $\mathbf{y} \in \mathbb{R}^n$ such that $\|\nabla f(y)\|_2 \leq \varepsilon$.
- Start with your initial point $\mathbf{x}^{(0)} = \mathbf{a}$
- Let $\nabla^{(k)} := \nabla f(\mathbf{x}^{(k)})$.
  While $\|\nabla^{(k)}\|_2 > \varepsilon$
  - Let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \eta \cdot \|\nabla^{(k)}\|_2$

If your function is L-smooth, that is, has gradient L-Lipschitz, then can set $\eta = 2/L$.

Conjugation Action

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

  Nullcone: *Nilpotent Matrices*

Conjugation Action

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

  Nullcone: *Nilpotent Matrices*

- Norm of a scaled element:

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*

- Norm of a scaled element:
- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?

Conjugation Action

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

  Nullcone: *Nilpotent Matrices*
- Norm of a scaled element:
- Conjugation action scaling problem:
  - **Input:** $A \in \mathrm{Mat}(n)$
  - **Output:** Is A nilpotent?
- Cannot do Alternating Minimization here (nothing to alternate on!)

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*
- Norm of a scaled element:
- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Cannot do Alternating Minimization here (nothing to alternate on!)
- How do we generalize the notion of gradient here?

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*
- Norm of a scaled element:
- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Cannot do Alternating Minimization here (nothing to alternate on!)
- How do we generalize the notion of gradient here?
- Why would gradient descent work here?

Conjugation Action

- $G = \mathbb{GL}(n)$, $V = \text{Mat}(n)$ conjugation

  Nullcone: *Nilpotent Matrices*
- Norm of a scaled element:
- Conjugation action scaling problem:
  - **Input:** $A \in \text{Mat}(n)$
  - **Output:** Is A nilpotent?
- Cannot do Alternating Minimization here (nothing to alternate on!)
- How do we generalize the notion of gradient here?
- Why would gradient descent work here?
- Wouldn't gradient equal zero only give us a *local optimum*? Why would that work in general?

Geodesic Convexity!

Conclusion

- Today we learned the basics about scaling algorithms
- How optimization naturally comes up in geometric invariant theoretic questions
- Connections to other areas of mathematics
- Alternating minimization algorithms
- Next lecture: is this a general phenomenon?

  Yes – geodesic convexity!
- Will see how to compute gradients of conjugation action and show that it is geodesically convex next lecture!