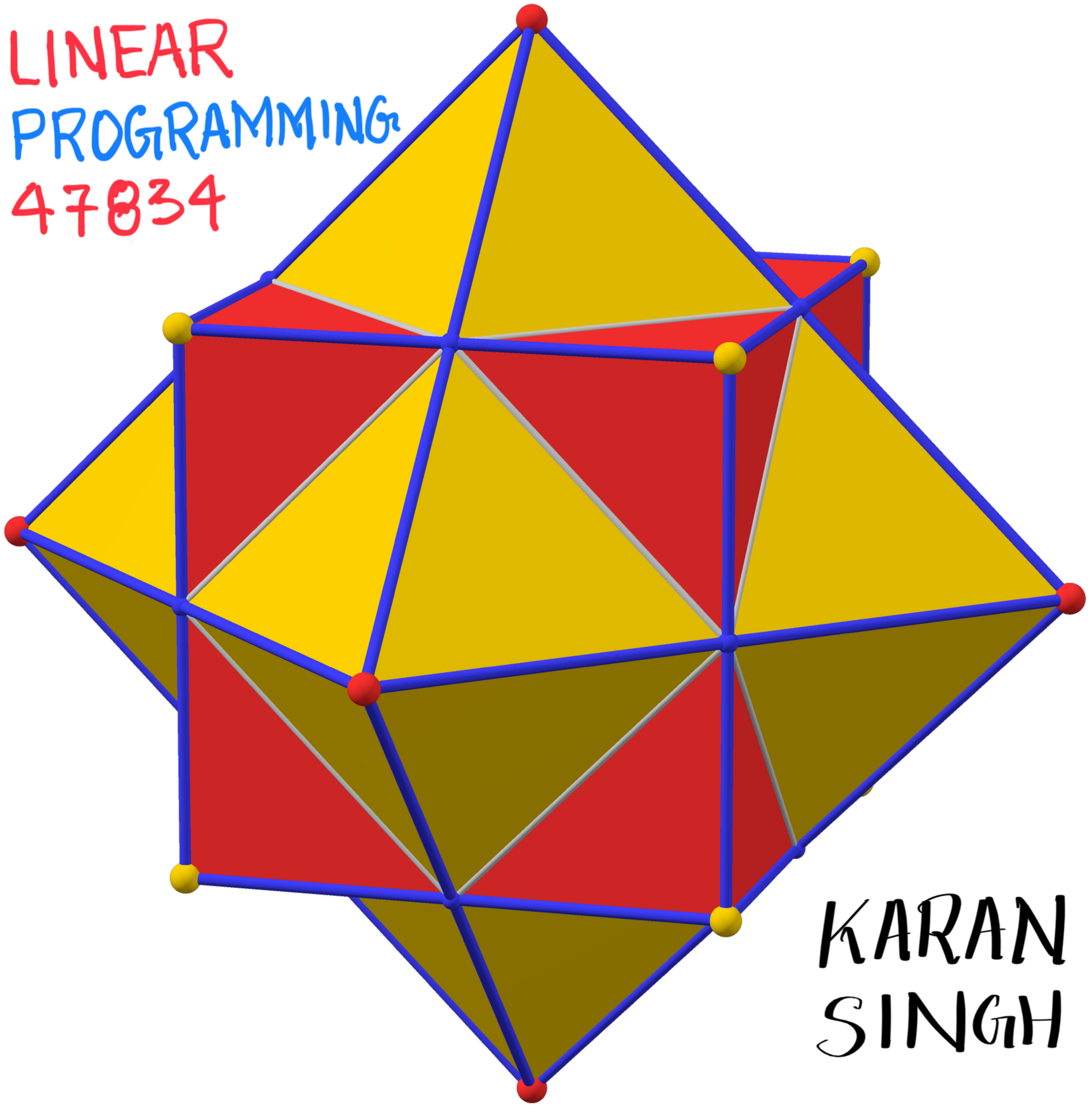
LINEAR PROGRAMMING 47834



KARAN SINGH

LECTURE 1: INTRO

LINEAR PROGRAMMING

47-834 M/W 4-6 pm (Typically 5:30 pm)
TEP 5219

KARAN SINGH

CORRECT: kuh-ruhn
APPROX: cur-RUN

TOPICS

THEORY:
* Standardization of LPs ++
* Geometry of LPs
* Algebra of LPs
* Minimax / LP Duality

ALGORITHMS:
* Simplex
* Center-of-mass
* Ellipsoid
* ? Interior-point

EXTENSIONS ~~APPLICATIONS~~:
* Regret Minimization
* Zero-sum games
* ? Optimal Transport
* ? Linear Integration

Headings
Suggested Exercises
Corrections

GRADING

45%	final exam
15% x3	Assignments
10%	Participation

POLICIES ETC.

- Don't buy textbooks.
- Prereq: Linear Algebra ($\det A$, A^{-1}, rank-nullity) Single-var calculus
- No late submissions.
- ONLY ASSIGNMENTS: Can discuss; write by self.

Front page of NYT twice

BREAKTHROUGH IN PROBLEM SOLVING
https://www.nytimes.com/1984/11/19/us/breakthrough-in-problem-solving.html
+ Karmakar (28 years old, recent UCB PhD) features twice in NYT in '84.
+ This was a poly-time interior point method. We'll study this.
+ "It has also set off a deluge of inquiries from brokerage houses, oil companies and airlines, industries with millions of dollars at stake in problems known as linear programming."
+ ''This is a path-breaking result,'' said Dr. Ronald L. Graham, director of mathematical sciences for Bell Labs in Murray Hill, N.J. ''Science has its moments of great progress, and this may well be one of them.''
+ K talks with American Airlines: How much fuel to carry? Where to fuel?
+ Exon's research head says "studies underway".
+ Dantzig is cautious; he was partial to the simplex method.

A Soviet Discovery Rocks World of Mathematics
https://www.nytimes.com/1979/11/07/archives/a-soviet-discovery-rocks-world-of-mathematics-russians-surprise.html
+ Khachiyan (late 20s?) features twice in NYT in '79.
+ "applicable in weather prediction, complicated industrial processes, petroleum refining, the scheduling of workers at large factories, secret codes and many other things."
+ This was the ellipsoid algorithm; also poly time. We'll study this.

Karan: Today, such press seems parallel to the coverage ML/deep learning gets.

George Dantzig's Story

+ From WWI/WWII era (distributed) logistics, productions problems. Questions around: what to do/when to do to arrive at some state/ achieve some objective. Semantics: programming ~ planning.
+ Dantzig (USAF) 1947 formulates/recognizes the general linear programming problem as a possible compromise between solvable and interesting problem classes. Also, proposes the simplex algorithm.
+ His claim (in his text): previous work did not have an objective function, i.e. only posed feasibility problems. An example is Motzkin's 1936 thesis which cites 42 pages, none considering an objective.
+ Some LP special cases (Koopmans, Leontif, Kantorvich) would win Nobel in Econ.
+ Meets von Neumann to discuss. Von Neumann is annoyed, "get to the point!". On seeing the problem, delivers an impromptu 1.5 hour lecture to Dantzig and describes both LP duality (including Farkas's Lemma) and an early interior point method. What triggered this?

Von Neumann's Story

+ See https://wwnorton.com/books/the-man-from-the-future. A prodigy, and reputed as a deep mathematician who interfaces with applied problems/worldly affairs, e.g., consults on Manhattan project.
+ Early contributions include a resolution to fundamental inconsistencies in mathematics (Russel's paradox: S is the set of all sets which are not members of themselves. Is S in S? Others resolved it simultaneously by better means.), and rigorous unification of wave equation and matrix mechanics in QM (earlier heuristic argument by Dirac using delta functions).
+ In 1944, a book with an Economist Morgenstern on The Theory of Games and Economic Behavior.
+ Minimax duality in 2-person zero-sum games is same as LP duality. Today, called von Neumann duality. But, it is von Neumann's?

Truer Origins of Duality

+ Monge proposes a question about the transportation problem in 1700's, used to model moving ores from mines to factories at minimum cost.
+ Kantorvich (1939) solves it, constructs the dual. Transportation is as general as LP. Kantorvich largely ignored in Russia. We will study this too.
+ Today, applications in PDEs, convex geometry, dynamical systems, probability. Cedric Villani (later member of French Parliament) wins a Fields Medal; see his book here https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.

Linear Programming – a class of optimization problems that are useful and simultaneously tractable.

What is LINEAR?

Def^n A function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq$ some vector space V, is LINEAR if

(a) $f(x+y) = f(x) + f(y) \quad \forall x, y \in \mathcal{X}$.

(b) $f(\alpha x) = \alpha f(x) \quad \forall x \in \mathcal{X}, \alpha \in \mathbb{R}$.

Proposition If $V = \mathbb{R}^n$, then for any $\mathcal{X} \subseteq V$ and LINEAR $f : \mathcal{X} \to \mathbb{R}$, $\exists c \in \mathbb{R}^n$ such that $f(x) = c^T x \quad \forall x \in \mathcal{X}$.
Furthermore, if $\mathcal{X} = \mathbb{R}^n$, then the choice of such $c \in \mathbb{R}^n$ given a fixed 'f' is unique.

Comments – $x \longrightarrow c^T x$ is of course linear. The interesting bit is that any linear function on $\mathcal{X} \subseteq \mathbb{R}^n$ can be written this way. Also this representation may not be unique if $\mathcal{X}$ is a subset of a proper subspace of $\mathbb{R}^n$, i.e., is lower-dimensional.

Ex Try to see why! This does not impugn existence.

Proof Sketch: We will only consider the case when $\mathcal{X} = \mathbb{R}^n$.

Ex Check the general case $V = \mathbb{R}^n$ on your own.

Any $x \in \mathbb{R}^n$ can be written as $x = \sum_{i=1}^{n} x_i e_i$, where x_i → scalar (i^{th} element), e_i → vector

where $e_i = [0\,0 \cdots \underset{i^{th}\text{ place}}{1}\, 0\, 0 \cdots 0]^T$.

Now, $f(x) = \sum_{i=1}^{n} f(e_i) x_i$.

Hence, $c = [f(e_1) \cdots f(e_n)]^T$ satisfies said claim. $\square$

Motivating Definition of LPs

ATTEMPT 1: $\max\ c^T x$ subject to $x \in \mathbb{R}^n$ (or such that / s.t.)

is either 0 when $c = 0$ or $+\infty$ when $c \neq 0$. Not interesting / useful.

ATTEMPT 2: Impose constraints ← WILL RETURN

A few definitions

(1) A set S is CLOSED if it contains all its limit points, i.e., for any $x_1, x_2, \dots \in S$ where $\lim x_n$ exists then $\lim x_n \in S$.

if for any $x_1, x_2 \ldots \in S$ such that $\lim_{n\to\infty} x_n$ exists, then $\lim_{n\to\infty} x_n \in S$.

(2) A set S is BOUNDED if $\exists C \in \mathbb{R}$ such that $\forall x \in S$, $\|x\| \leq C$.

Choice of norm (in finite-dimensional spaces) is not crucial / is immaterial to this definition.

(3) S is COMPACT if it is CLOSED and BOUNDED.

(4) A function $f: \mathcal{X} \to \mathcal{Y}$ is CONTINUOUS if for any $x_1, x_2, \ldots \in \mathcal{X}$ such that $\lim_{n\to\infty} x_n$ exists and is in $\mathcal{X}$, we have $\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right)$.

(5) Given a function $f: \mathcal{X} \to \mathbb{R}$, its maximum is ATTAINED on $S \subseteq \mathcal{X}$ if $\exists x^* \in S$ such that $\forall x \in S$, $f(x) \leq f(x^*)$.

$x^* \in \mathcal{X}$ is said to MAXIMIZE f on S.

This is a stronger requirement than existence of SUPREMUM.

A nice result; makes life easy.

<u>THEOREM:</u> A continuous function ATTAINS its maximum (and minimum) (Weierstrass) on any non-empty compact set.

RETURNING TO LP MOTIVATIONS

Take any compact set S.

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t} \quad & x \in S \end{aligned}$$

$\longrightarrow$ Good definition? No hope of computational tractability.

<u>FACT:</u> Maximizing general (non-concave) functions is hard.

$$\begin{aligned} \max \quad & f(x) \\ \text{s.t} \quad & x \in S \end{aligned} \quad \xleftrightarrow{\text{equivalent}} \quad \begin{aligned} \max_{(t,x)} \quad & t \\ \text{s.t.} \quad & x \in S \\ & f(x) \geq t. \end{aligned}$$

Optionally add $-10^{10^{10}} \leq t \leq 10^{10^{10}}$ to ensure compactness: NITPICK.

ATTEMPT 3: (George Dantzig 1947)

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & a_1^T x \leq b_1 \\ & a_2^T x \leq b_2 \\ & \vdots \\ & a_m^T x \leq b_m \end{aligned}$$

Can express many problems + efficiently solvable.

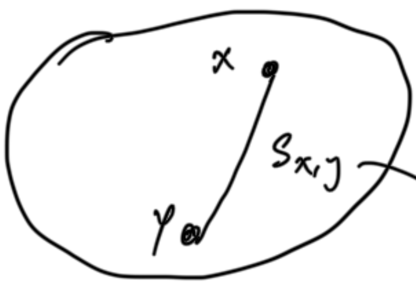
These are technically affine. (We will call them linear)

In addition to linearity, having a finite number 'm' of constraints is also important to guarantee tractability.

LPs as a special case of CONVEX PROGRAMS

(1) A <u>set</u> $S \subseteq$ vector space V is CONVEX if $\forall\, x, y \in S$ $\forall\, \lambda \in [0,1]$, $\lambda x + (1-\lambda) y \in S$.

$S_{x,y} = \{ \lambda x + (1-\lambda) y : \lambda \in [0,1] \}$ is a <u>line segment</u> between x & y.



Comments: (1) $\{x : a^T x \le b\}$ is a convex set.

(2) The intersection of two (or any number of) convex sets is convex. Ex

Hence, $\left\{ x : \begin{array}{c} a_1^T x \le b_1 \\ \vdots \\ a_m^T x \le b_m \end{array} \right\}$ is convex.

$A \cap B = \{ a : a \in A \text{ and } a \in B \}$

(3) If A & B are convex, then Ex

$A \times B = \{ (a,b) : a \in A, b \in B \}$ is convex.

CARTESIAN PRODUCT

(4) If A & B are convex, then Ex

$A + B = \{ a + b : a \in A, b \in B \}$ is convex.

MINKOWSKI SUM

(5) If S is a convex set in $\mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$,

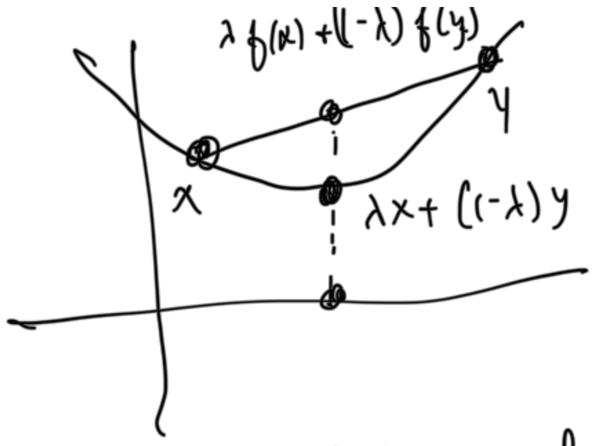
then $\{ Ax : x \in S \}$ is convex. Ex

(6) For any convex function f on a convex set S, Ex

$\operatorname{argmin}_{x \in S} f(x) = \left\{ x^* \in S : \begin{array}{c} \forall x \in S \\ f(x) \ge f(x^*) \end{array} \right\}$ is convex.

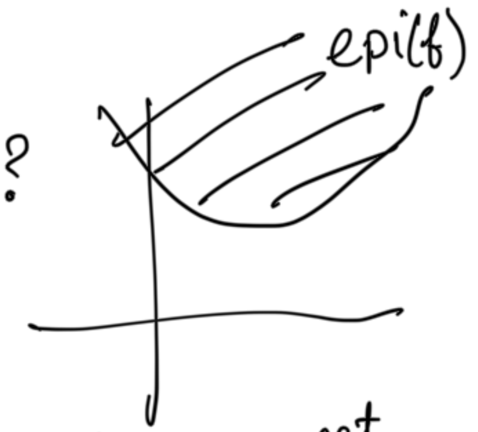
(2) A <u>function</u> $f : \mathcal{X} \text{(convex)} \to \mathbb{R}$ is said to CONVEX if $\forall\, x, y \in \mathcal{X}$, $\forall\, \lambda \in [0,1]$, $f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y)$.

In words, interpolating line lies above.



(3) How is this related to convex <u>sets</u>?

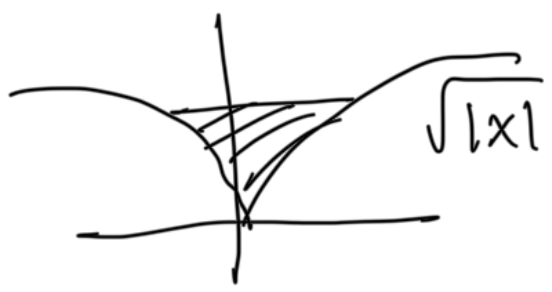
$$\text{epi}(f) = \{(t,x) : t \geq f(x)\}$$ Definition



Ex

<u>PROPOSITION.</u> f is a convex function if epi(f) is a convex set.

A tantalizing, but incorrect way to link convex <u>sets</u> and <u>functions</u> is: ask f is a convex function if and only if $\forall t \in \mathbb{R}$, $S_t(f) = \{x : f(x) \leq t\}$ is convex? FALSE.



$\sqrt{|x|}$ is quasi-convex, but not convex

Such functions are called quasi-convex.

Convex $\Longrightarrow$ Quasi-convex.

(4) Finally, $x \to c^T x$ is a convex function. Hence, LP are convex PROGRAMS.

General convex program: $\min f(x)$ $\to$ convex function

s.to $x \in S$ $\to$ convex set

More generally, for

$$\max f(x)$$
$$\text{s.to } x \in S$$

S is called feasible region

x^* maximizing f on S is called an optimal solution.

References:

1. History—
    1. p 209 in Schrijver
    2. Dantzig's article
2. Basics of Convexity— chapters 2 & 3 in Boyd

LECTURE 2: STANDARDIZATION

EXAMPLES AND NOTATIONS INVOLVING LPS

* $x \leq y$ for vectors x & y in $\mathbb{R}^n$ if $x_i \leq y_i \ \forall i \in [n] = \{1 \dots n\}$.

  Note for vectors, it is NOT true that either $x \leq y$ or $y \geq x$.

  Also, very much basis dependent (here, standard basis).

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & a_1^T x \leq b_1 \\ & \vdots \\ & a_m^T x \leq b_m \end{array} \iff \begin{array}{ll} \max & c^T x \\ \text{s.t} & \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} x \leq \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \end{array}$$

* EXAMPLE: UN Diet Problem

  n food item; food j has cost c_j.

  m nutrients; minimum acceptable level of nutri i is b_i.

  a_{ij} is amount of nutri i in food j. $A = [a_{ij}] \in \mathbb{R}^{m \times n}$

$$\begin{array}{ll} \min & \sum_{j=1}^{n} c_j x_j \\ \text{s.t.} & \sum_{j=1}^{n} a_{ij} x_j \geq b_i \end{array} \iff \begin{array}{ll} \min & c^T x \\ \text{s.t.} & Ax \geq b \end{array}$$

* STANDARDIZATION OF LPS

How to transform

Note! Size doesn't blow up.

(1) $\max_{x \in S} c^T x = -\min_{x \in S} (-c)^T x$

(2) $a^T x \leq b \iff (-a)^T x \geq (-b)$

(3) $a^T x = b \iff a^T x \leq b \ \wedge \ a^T x \geq b$

(4) $a^T x \leq b \iff a^T x + s = b \ \wedge \ s \geq 0$

(5) unconstrained $x \iff$ replace with $x^+ - x^-$ where $x^+, x^- \geq 0$.

$$\begin{array}{ll} \max & c^T x \\ \text{s.t} & Ax \leq b \end{array}$$

General Form

$$\begin{array}{ll} \min & c^T x \\ \text{s.t} & Ax = b, \ x \geq 0 \end{array}$$

Standard form

EXAMPLE

* $\min_{x \in \mathbb{R}^n} \max_{i \in [m]} (a_i^T x + b_i) \Longleftrightarrow \min_{(t,x)} t$ s.t $t \geq a_i^T x + b_i \;\; \forall i \in [m]$

why is this tractable? Ex

Given convex functions $f_1 \cdots f_m$, $f(x) = \max_{i \in [m]} f_i(x)$ is convex.

* EXAMPLE

$\max \; c^T x$

s.t. $\|x\|_1 = \sum_{i=1}^{n} |x_i| \leq 1$

$\longrightarrow$ Manually solving this.

Max Value $= \max_{i \in [n]} |c_i|$

Let $S = \{ i : |c_i| \geq \max_{i \in [n]} |c_i| \}$

Argmax $= \text{sign}(c_i)\, e_i$ for $i \in S$

Let's express this as a LP.

Attempt 1: $\max \; c^T x$

s.t $\underbrace{\sigma^T x \leq 1 \;\; \forall \sigma \in \{\pm 1\}^n}_{\text{exponentially many constraints.}}$

Attempt 2: $\max \; c^T x$

s.t. $y_i \geq x_i$, $y_i \geq -x_i$, $\sum_{i=1}^{n} y_i \leq 1$

$2n+1$ constraints! Exponential improvement.

Extended formulations (area of study: extension complexity) aim to achieve drastic reductions in the number of constraints, by introducing a few (polynomially) more variables.

HISTORY:

* Early 1990s / late 1980s; numerous attempts at solving NP-complete problems using LPs; hence, at P=NP.

  Idea: Typical hard problems (like TSP) can be written as LPs with exponentially many constraints. Introduce a few more variables to achieve (similar to the above example) polynomially many constraints.

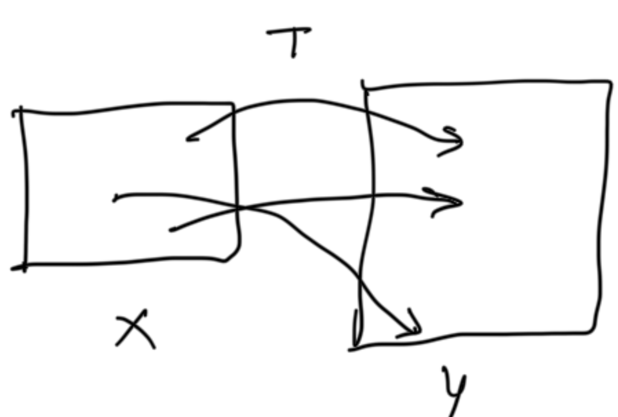
* Yannakakis '91: Any symmetric formulation of TSP as LP has exponentially many constraints.
* FMPTW '12: Any formulation of TSP as LP has exponentially many constraints.
* Killing the 1980/90's hope.

* EXAMPLE: OPTIMAL TRANSPORT (equivalent to general LPs)

Monge 1700's: Want to move iron ore from mines to factories. Mines produce $\mu(x)dx$ ore at 'x'. Factories at 'y' consume $\nu(y)\,dy$ ore.

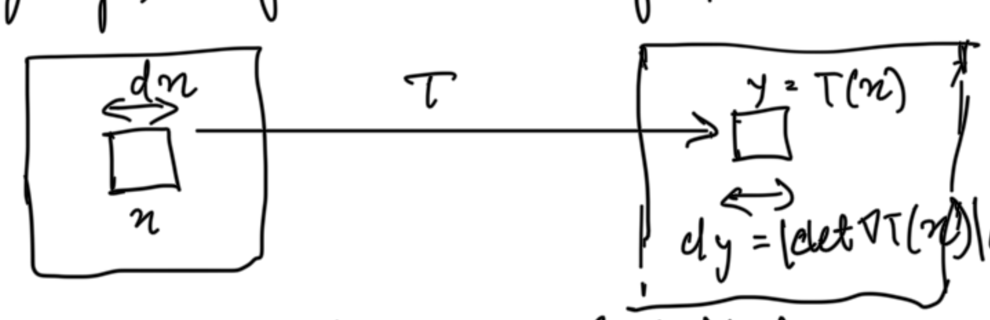
$c(x, y)$ — cost of transporting unit quantity from 'x' to 'y'.

$$\min_{\substack{T: X \to Y \\ \text{invertible}}} \int_{\mathcal{X}} c(x, T(x))\, \mu(x)\, dx$$



$$\text{s.t. } \nu(y) = \underbrace{|\det \nabla T^{-1}(y)|\, \mu(T^{-1}(y))}_{\text{push-forward of } \mu \text{ through } T.} \quad \forall y \in \mathcal{Y} \quad \text{(assume invertible)}$$

Roughly, by conserving probability mass



$$\mu(x)\,dx = \underbrace{\nu(T(x))}\, \underbrace{|\det \nabla T(x)|\, dx}$$
$$= \nu(y) \quad dy$$

Assume, above that

$$\int_X \mu(x)\,dx = \int_Y \nu(y)\,dy = 1$$

Interpretable as Probability measures.

Highly non-linear problem, * 2 Econ Nobel Prizes

TODAY applications in * concentration of measure
* dynamical systems

(now, French MP) * PDEs.

Cedric Villani, won a Fields Medal solving related problems.

Kantorovich 1939: Tractable solution by reformulating the problem. Notice that this is pre-LP / von-Neumann minimax duality. Also, game dual.

$$\min_{\Gamma: X \times Y \to \mathbb{R}} \int_X \int_Y \Gamma(x,y)\, u(x)\, v(y)\, dx\, dy$$

$$\text{s.t. } \forall y \in Y \quad \int_X \Gamma(x,y)\, u(x)\, dx = v(y)$$

$$\forall x \in X \quad \int_Y \Gamma(x,y)\, v(y)\, dy = u(x)$$

Notice this is linear in Γ. If u, v are probability mass functions, instead of density, same as LP below.

$$\min_{\Gamma} \sum_{x \in X} \sum_{y \in Y} \Gamma_{x,y}\, u_x\, v_y$$

$$\text{s.t.} \quad \forall y \in Y \quad \sum_{x \in X} \Gamma_{x,y}\, u_x = v_y$$

$$\forall x \in X \quad \sum_{y \in Y} \Gamma_{x,y}\, v_y = u_x$$

Intuitively, in Kantorovich's Transport Plan formulation, production & supply capacities arise simultaneously while maintaining marginal rates of u, y.

Also, related: "coupling" in probability theory.

* Convention in Convex Optimization

$\min_{x \in S} f(x)$ is equivalent to $\min_{x \in \mathbb{R}^n} f_S(\cancel{x})$, where $f_S(x) = \begin{cases} +\infty & \text{if } x \notin S \\ f(x) & \text{if } x \in S \end{cases}$

Similarly, $\max_{x \in S} f(x) \equiv \max_{x \in \mathbb{R}^n} f_S'(x)$,

where $f_S'(x) = \begin{cases} -\infty & \text{if } x \notin S \\ f(x) & \text{if } x \in S \end{cases}$

* Feasibility programs ask does there exist $x \in \mathbb{R}^n$ such that $x \in S$?

Linear Feasibility : $\exists ?\, x \in \mathbb{R}^n$ s.t. $Ax \leq b$
(generally represented)

* Optimization & feasibility are closely related in computational terms.

Using an optimization solver to check feasibility.

$$\theta(c) = \max \; c^T x \quad \text{s.t. } x \in S$$

Observe, $\theta(0) = 0$ iff $\exists x \in \mathbb{R}^n$ s.t. $x \in S$.

Using a feasibility solver for optimization

$$\mathcal{F}(S) = \begin{cases} \text{YES} & \text{if } \exists x : x \in S \\ \text{NO} & \text{otherwise.} \end{cases}$$

Let's say we have some apriori range for $-10^{100} \leq \max_{x \in S} f(x) \leq 10^{100}$.

Using $\mathcal{F}$, we will solve $\max_{x \in S} f(x)$ to ε-accuracy

ALGORITHM $A = [l, u]$

1. If $|l - u| \leq \varepsilon$, then output any value in $[l, u]$.
2. Else, $t = \frac{l + u}{2}$.

If $\mathcal{F}(S \cap \{x : f(x) \geq t\}) = \text{YES}$,

$[t, u]$

call A on $[l, u]$.

Else, call A on $[l, t]$.

Start with A on $[-10^{10}, 10^{10}]$: INITIALIZATION

Comments:

1. $\max_{x \in S} f(x) \in [l, u]$ holds at the start of any call to the algorithm A; because

$$\mathcal{F}(S \cap \{x : f(x) \geq t\}) = YES \Rightarrow \mathcal{F}(S \cap \{x : f(x) \geq t'\}) = YES$$

for all $t' \leq t$.

2. In each successive call, the length of the argument interval $[l, u]$ is halved.

After T calls, we have a $\frac{10^{10} \times 2}{2^T}$-sized interval containing $\max_{x \in S} f(x)$.

If $T \approx \log \frac{1}{\varepsilon}$, we know $\max_{x \in S} f(x)$ to ε-accuracy.

3. If the feasibility oracle also returns a feasible point 'x' on YES. Then, can recover a point $\tilde{x} \in S$ s.t.

$$f(\tilde{x}) \geq \max_{x \in S} f(x) - \varepsilon.$$

Ex

References:

1. Standardization— section 1.1 in Nemirovski
2. (Beyond this course) optimal transport— chapter 1 in Thorpe
3. (Beyond this course) extension complexity— Gerard's survey
4. Feasibility-optimization reduction— 4.2.5 in Boyd

LECTURE 3: ALGEBRA

LPs as a proof system

$$\begin{aligned} t = \max \ & c^T x \\ \text{s.t } \ & Ax \le b \end{aligned}$$ can be interpretted as $\forall x,\ Ax \le b \Rightarrow c^T x \le t.$

One way to prove statements of the latter form is by combining existing inequalities via non-negative multipliers (these don't flip the sign.)

$$\left.\begin{array}{l} (x_1 \le 2) \times 3 \\ (x_2 \le 4) \times 2 \end{array}\right\} \longrightarrow 3x_1 + 2x_2 \le 14$$

also valid

The non-trivial/interesting bit for LPs is that such a proof (in the restricted language of multiplying existing inequalities with non-negative statements, then adding) always exists for any valid inequality. We will see an algorithmic proof. This remarkable fact is not true about mathematics in general, i.e., courtesy Godel, there are 'true' but 'unprovable' statements in mathematics.

Fourier-Motzkin Elimination

Rough Idea: Eliminate variables by adding more constraints. $\approx$ opposite of extended formulations.

An algorithm to solve linear feasibility problems, i.e. does there exist $x \in \mathbb{R}^n$ such that $Ax \le b$?

Note, can use this for (high-accuracy) optimization via the optimization $\rightarrow$ feasibility reduction.

1-step of FM Elimination

(can always ensure this by Lec 2)

Input : 'm' inequalities of the form $a_i^T x \le b_i \ \forall i \in [m]$

1. Divide all inequalities into 3 sets

$Z = \{$ all inequalities that don't involve x_1; $a_{i1}=0\}$

$P = \{$ all " with $a_{i1}>0\}$

Each can be rewritten as

$$x_1 \leq \underbrace{\frac{b_i - \sum_{j\neq 1} a_{ij} x_j}{a_{i1}}}_{\text{Call this } p(x_2 \dots x_n)}$$

$N = \{$ all " with $a_{i1}<0\}$

Each can be rewritten as

$$x_1 \geq \underbrace{\frac{b_i - \sum_{j\neq 1} a_{ij} x_j}{a_{i1}}}_{\text{Call this } n(x_2 \dots x_n)}$$

2. Construct a new feasibility problem by

(a) Copying all of Z.

(b) $\forall p \in P \ \forall n \in N$, introduce $n(x_2 \dots x_n) \leq p(x_2 \dots x_n)$

These no longer contain x_1.

Rearrange (b) into $a^T x \leq b$ form.

<u>Claim:</u> New LP is feasible iff old LP is feasible.

<u>Proof.</u> $x_1 \dots x_n$ satisfies old LP. (IF)

$\Rightarrow x_2 \dots x_n$ satisfies Z; and

$p(x_2 \dots x_n) \geq x_1 \geq n(x_2 \dots x_n) \ \forall n \in N, p \in P$

$\Rightarrow x_2 \dots x_n$ satisfies new LP.

$x_2 \dots x_n$ satisfies new LP. (ONLY IF)

By construction, $x_2 \dots x_n$ satisfies Z.

Also, $\max_{n \in N} n(x_2 \dots x_n) \leq \min_{p \in P} p(x_2 \dots x_n)$.

(Recall max over ϕ is $-\infty$, min over ϕ is $+\infty$).

Choose $x_1 \in \left[\max_{n \in N} n(x_2 \dots x_n), \min_{p \in P} p(x_2 \dots x_n)\right]$ arbitrarily.

$x_1 \dots x_n$ satisfies old LP. $\square$

<u>Comments</u>. 1. n-steps of FM elimination solves any feasibility problem. At termination, we either have all tautological inequalities or a contradiction.

2. If old LP has m constraints, new LP has $\leq m^2$ constraints. Therefore, the transcript produced by the algorithm (over n-steps), and hence the running time is $\approx m^{2^n}$.

$$\begin{matrix} m \text{ ineq} \\ n \text{ var} \end{matrix} \longrightarrow \begin{matrix} m^2 \text{ ineq} \\ n-1 \text{ var} \end{matrix} \longrightarrow \begin{matrix} m^4 \text{ ineq} \\ n-2 \text{ var} \end{matrix} \longrightarrow \begin{matrix} m^8 \text{ ineq} \\ n-3 \text{ var} \end{matrix}$$

So, FM is largely a conceptual algorithm.

3. If A, b only contain rationals, then

$Ax \leq b$ is feasible $\Rightarrow \exists x$ rational feasible.

Why? Because FM only creates inequalities with rational coefficients, given rational A, b.

<u>Observation</u>: Any new inequality produced during FM is done by combining existing ones with non-negative coefficients.

(1) $a_1 x_1 + \sum_{i=2}^{n} a_i x_i \leq b \quad (a_1 > 0)$

(2) $a_1' x_1 + \sum_{i=2}^{n} a_i' x_i \leq b' \quad (a_1' < 0)$

$$\Longrightarrow \frac{b' - \sum_{i=2}^{n} a_i' x_i}{a_1'} \leq x_1 \leq \frac{b - \sum_{i=2}^{n} a_i x_i}{a_1}$$

$$\Downarrow$$

Same as $\frac{1}{a_1} \times (1) + \frac{1}{|a_1'|} \times (2)$

$$\longrightarrow \sum_{i=2}^{n} \left(\frac{a_i}{a_1} - \frac{a_i'}{a_1'} \right) x_i \leq \frac{b}{a_1} - \frac{b'}{a_1'}$$

<u>Farkas' Lemma</u>: $Ax \leq b$ is infeasible iff $\exists u \geq 0$, $u^T A = 0$, $u^T b < 0$.

<u>Interpretation</u>. If $Ax \leq b$, then for any $u \geq 0$, $u^T A x \leq u^T b$. So if $\exists u \geq 0, u^T A = 0, u^T b < 0$, that implies $Ax \leq b$ is infeasible. Because, otherwise $0 = u^T A x \leq u^T b < 0$; a contradiction.

'u' is therefore a certificate of infeasibility; its existence guarantees infeasibility of $Ax \leq b$. The interesting bit is that such a 'blatant' certificate always exists whenever $Ax \leq b$ is infeasible. In this sense, the linear inequality proof system is complete, not just sound.

<u>Proof</u>: Based on the correctness of FM, for any infeasible system $Ax \leq b$, FM must terminate in a contradiction of the form $0 \leq b_0$, where $b_0 < 0$. By the last observation FM implicitly produces a vector $u \geq 0$ such that $u^T A = 0$ and $u^T b = b_0 < 0$. $\square$

This is a central result in the theory of LPs, and only a step away from LP duality itself. We will complete this later.

3 VIEWS OF LPs

Will assume $P = \{x : \overset{m \times n}{A} x = b,\ x \geq 0\}$, $\max_{x \in P} c^T x$, where

(a) $Ax = b$ has atleast one solution, or $b \in COLSP(A)$.
Else, P is infeasible/empty.

(b) Rows of A are linearly independent. $(m \leq n)$.

OPTIMIZATION

<u>Defn.</u> $x \in P$ is a VERTEX if $\exists c$, $c^T x > c^T y \ \forall y \in P - \{x\}$.

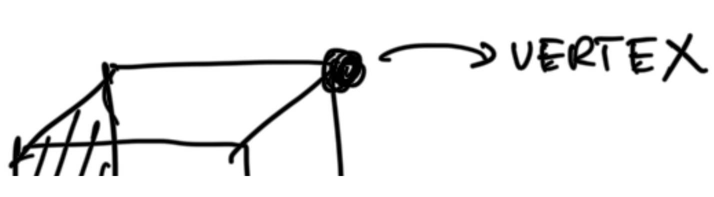
<u>Defn</u>: $F \subseteq P$ is a k-dimensional face of P if

(1) $\exists\, x_0$ and k-dimensional subspace V such $P \subseteq x_0 + V$,

(2) $\exists\, c, z$ such $c^T x = z \ \forall x \in P$, $\forall y \in P - F$, $c^T y < z$.

F is a <u>proper</u> k-dim face of P is it is a k-dim face, but not a $(k-1)$-dimensional face.

* VERTEX is a 0-dimensional proper face.
EDGE is a 1 " " " .



GEOMETRIC

<u>Def^n.</u> $x \in P$ is an EXTREME POINT if $\nexists\ u \neq v,\ \lambda \in (0,1)$ such that $\lambda u + (1-\lambda) v = x$.

ALGEBRAIC

<u>Def^n.</u> $x \in P$ is a BASIC FEASIBLE SOL^N if $\exists\ B \subseteq [n], |B| = m$ such that $A_B \in \mathbb{R}^{m \times m}$ is invertible and $x_{\overline{B}} = 0$.

Notation: For any $S \subseteq [n]$, (a) $\overline{S} = [n] - S$: complement of S

(b) A_S is a $m \times |S|$-sized matrix composed only of columns whose indices are in S.

(c) x_S is $|S|$-sized vector composed only of coordinates whose indices are in S.

For a vector x, $\mathrm{Supp}(x) = \{ j : x_j \neq 0 \}$.

* Notice that every B can correspond to atmost one BFS. Given B, $A_B^{-1} b \in \mathbb{R}^m$ extended to $\mathbb{R}^n$ by padding with 0's on $\overline{B}$ is the only possible candidate for BFS, but its possible that $A_B^{-1} b \geq 0$ fails.

These 3 views are equivalent.

<u>THEOREM.</u> $x \in P$ is a vertex $\iff$ it is an extreme point $\iff$ it is a BFS.

<u>PROOF.</u> $V \Rightarrow E$

x is a vertex. $\exists\ c$ such $c^T x > c^T y\ \forall y \in P - \{x\}$. Say $\exists\ u \neq v, \lambda \in (0,1)$ such $x = \lambda u + (1-\lambda) v$. But then $c^T x = \lambda (c^T u) + (1-\lambda)(c^T v)$.
This is a contradiction because $u \neq x$ and $v \neq x$, hence $c^T u < c^T x$ and $c^T v < c^T x$.

$E \Rightarrow BFS$.

x is an extreme point. Recall $\mathrm{supp}(x) = \{ j : x_j > 0 \}$.

<u>CASE A:</u> $A_{\mathrm{supp}(x)}$ has linearly independent columns.
Implies that $|\mathrm{supp}(x)| \leq m$. Since $x_{\overline{\mathrm{supp}(x)}} = 0$, it is tempting to think $B = \mathrm{supp}(x)$ concludes this case. This almost

works except $|B| = m$, while $|supp(x)|$ can be smaller.
Here's a fix: since rows of A are linearly independent, it's possible to construct B by starting with $supp(x)$, and

then expanding this set incrementally till it includes m indices by choosing column of $A_{\overline{B}}$ that are linearly independent of that of A_B. At the end, we have $B \subseteq [n]$, $|B| = m$, A_B is invertible. Finally, since $B \supseteq supp(x)$, $x_{\overline{B}} = 0$. Hence, x is a BFS.

<u>CASE B:</u> $A_{supp(x)}$ has linearly dependent columns.
$\therefore \exists w$, $A_{supp(x)} w = 0$, $w \neq 0$. By padding w with 0's we can construct $\tilde{w} \in \mathbb{R}^n$, $\tilde{w}_{supp(x)} = w$, $\tilde{w}_{\overline{supp(x)}} = 0$, $A\tilde{w} = 0$.
Define $y_+ = x + \varepsilon \tilde{w}$, $y_- = x - \varepsilon \tilde{w}$. Note $x = \frac{y_+ + y_-}{2}$; yet $y_+ \neq y_-$ for any $\varepsilon > 0$ since $\tilde{w} \neq 0$. Also $Ay_+ = Ax + A\tilde{w} = b$ and $Ay_- = b$. So, if we can ensure $y_+, y_- \geq 0$, then $y_+, y_- \in P$ implying we have reached a contradiction. Choose $0 < \varepsilon \leq \frac{\min_{i \in supp(x)} x_i}{\max_{i \in [n]} |w_i|}$; observe $supp(\tilde{w}) \subseteq supp(x)$ to conclude $y_+, y_- \geq 0$.

BFS $\Rightarrow$ V
x is a BFS. $\exists B$, $|B| = m$ such A_B is invertible and $x_{\overline{B}} = 0$.
Note that $A_B x_B = b$. Construct $c \in \mathbb{R}^n$ such $c_j = \begin{cases} -1 & \text{if } j \in \overline{B} \\ 0 & \text{if } j \in B \end{cases}$.

Now, $c^T x = 0$. Notice for any $y \in P$, since $y \geq 0$, $c^T y \leq 0$. We'll prove if $c^T y = 0$ then $y = x$ to establish that x is a vertex. Consider $y \in P$ such that $c^T y = 0$. $y_{\overline{B}} = 0$ and A_B is invertible. y is a BFS. But then for B can correspond to at most one BFS. $\therefore y = x$. $\square$

References:

1. Fourier-Motzkin Elimination—
    1. I like 3.1 and 3.2 in Gerard's book; includes proof of Farkas' Lemma
    2. Also section 6.7 in Matousek
    3. Alternative: Section 2.8 in Bertsimas
2. BFS-Vertex-Extreme Equivalence—
    1. Section 2.2 and 2.3 in Bertsimas
    2. Chapter 4 in Matousek

LECTURE 4 : GEOMETRY

Since every $B \subseteq [n]$ of size m corresponds to at most one BFS. The number of BFSs is at most $\binom{n}{m}$. Recall that we are looking at LPs of the form.

$$\max \; c^T x \quad \text{s.t.} \quad P = \left\{ \begin{array}{l} Ax = b \\ x \geq 0 \end{array} \right\}$$

① $Ax = b$ has atleast one solution, i.e. $b \in \mathrm{COLSP}(A)$.

② Rows of A are linearly independent.

The next result we proves states that any LP of this form chooses between one of these three fates:

(1) The LP is infeasible, i.e. $\max_{x \in P} c^T x = -\infty$.

(2) The LP has unbounded optima, i.e. $\max_{x \in P} c^T x = +\infty$.

(3) A (finite) optima exists, and a BFS x achieves it.

Implicitly, we have the following finite-time algorithm:

SOLVING LPs WITH BOUNDED OPTIMAL VALUE BY ENUMERATION

1. For each $B \subseteq [n]$ of size m, solve $x_B = A_B^{-1} b$. Check if $x_B \geq 0$. On 'YES', set $x_{\bar{B}} = 0$ and add x to the set of BFSs.
2. If no BFSs are found, output INFEASIBLE, else output the highest objective value achieved by any BFS.

This takes $\approx \binom{m}{n} \times \mathrm{poly}(n)$ time. The simplex algorithm reuses this idea, but searches for BFSs greedily.

FUNDAMENTAL THEOREM OF SIMPLEX. In any feasible LP with bounded optima (note this is weaker than saying P is bounded) $\exists$ a BFS x that achieves the optimal value.

Let's start with a somewhat seemingly unrelated observation.

LEMMA Every feasible LP in the standard form has an extreme point.

PROOF. Since $P = \{x : Ax = b,\, x \geq 0\}$ is feasible, choose x to be a feasible point with the smallest number of non-zero entries. We claim 'x' is an extreme point. If not, $\exists\, u \neq v \in P$, $\lambda \in (0,1)$ such that $\lambda u + (1-\lambda) v = x$. Since $u, v \geq 0$, we have $\mathrm{supp}(u), \mathrm{supp}(v) \subseteq \mathrm{supp}(x)$, i.e. there can be no coordinate cancellations. We claim $\exists\, j \in [n]$ such $(u-v)_j > 0$, since $u \neq v$ and if $u-v$ is all negative, we relabel $(u, v, \lambda) \longleftrightarrow (v, u, 1-\lambda)$.

Now, consider $y = x - \varepsilon(u-v)$. Choose $\varepsilon = \min_{j : (u-v)_j > 0} \frac{x_j}{(u-v)_j}$.

Now, $Ay = Ax - \varepsilon(Au - Av) = b$, and $y \geq 0$ since $\text{supp}(u-v) \subseteq \text{supp}(x)$. Yet y has at least one fewer non-zero coordinate than x. Hence, x must be an extreme point. $\square$

Because of BFS-V-E equivalence, will prove the fundamental theorem of simplex by proving the existence of an optimal extreme point.

PROOF OF FUND. THM. OF SIMPLEX. Since $\max c^T x$ over $x \in P = \{x : Ax = b, x \geq 0\}$ is feasible and has bounded optima, $\exists v^* \in \mathbb{R}$ such $\max_{x \in P} c^T x = v^*$. Consider $Q = \{x : Ax = b, c^T x = v^*, x \geq 0\}$.

We know $\exists$ an extreme point x in Q. Note that $c^T x = v^*$ and $x \in P$ since $Q \subseteq P$. So, it only remains to show such an x is extreme in P too. Suppose not. Then, $\exists\, u \neq v \in P$, $\lambda \in (0,1)$ such that $x = \lambda u + (1-\lambda) v$. Since x is extreme in Q, at least one of u, v must not be in Q. Hence, $\min\{c^T u, c^T v\} < v^*$. Also, $\max\{c^T u, c^T v\} \leq v^*$. But $c^T x = \lambda c^T u + (1-\lambda) c^T v$ implying $c^T x < v^*$. Hence, x is extreme in P. $\square$

APPLICATIONS OF THE FUNDAMENTAL THEOREM

* AFFINE HULL$(S) = \{x = \sum_{i=1}^{k} \lambda_i x_i : \exists k \geq 0,\ x_1 \ldots x_k \in S,\ \lambda_1 \ldots \lambda_k \in \mathbb{R},\ \sum_{i=1}^{k} \lambda_i = 1\}$.
* CONIC HULL$(S) = \{x = \sum_{i=1}^{k} \lambda_i x_i : \exists k \geq 0,\ x_1 \ldots x_k \in S,\ \lambda_1, \ldots \lambda_k \geq 0\}$.
* CONVEX HULL$(S) = \{x = \sum_{i=1}^{k} \lambda_i x_i : \exists k \geq 0,\ x_1 \ldots x_k \in S,\ \lambda_1 \ldots \lambda_k \geq 0,\ \sum_{i=1}^{k} \lambda_i = 1\}$.

The next theorems are on the sufficiency of minimal representations: sure you can represent any $x \in$ CONIC HULL(S) as non-negative combinations of finitely many points in S, but how many are needed in the worst-case?

CARATHEODORY'S THEOREM FOR CONES If $x \in$ CONIC HULL(S), where $S \subseteq \mathbb{R}^n$, then $\exists\, x_1 \ldots x_n \in S$, such that $x \in$ CONIC HULL$(\{x_1 \cdots x_n\})$.

PROOF. Since $x \in$ CONIC HULL(S), $\exists\, x_1 \ldots x_k \in S$, $\lambda_1 \ldots \lambda_k \geq 0$ for some $k \geq 0$. Let $X = [x_1 \ldots x_k]$ be the matrix whose columns are x_k's. Then $P = \{x : x = X\lambda, \lambda \geq 0\}$ is feasible. Consider $\max_{x \in P} 0$. Then $\exists$

a BFS λ in P with atmost $n =$ number of equality constraint many non-zero elements, proving the theorem. $\square$

CARATHEODORY'S THEOREM FOR CONVEX HULLS

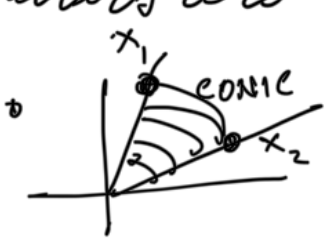
If $x \in \text{CONVEXHULL}(S)$, where $S \subseteq \mathbb{R}^n$, then $\exists x_1 \cdots x_{n+1} \in S$, such that $x \in \text{CONVEX-HULL}(\{x_1 \cdots x_{n+1}\})$.

PROOF. This time we have that $P = \{x : x = X\lambda, \lambda \geq 0, 1^T\lambda = 1\}$ is feasible. Hence, $\exists$ a BFS λ in P with atmost $n+1 =$ number of equality constraint many non-zeros. $\square$

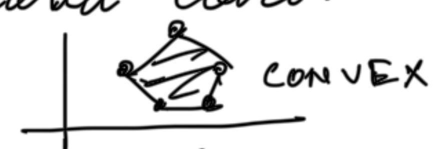
HINTS OF DUALITY

* $\{x : Ax \leq b\}$ is a polyhedron.
* $\{x : Ax \leq 0\}$ is a polyhedral cone.

These descriptions create a set by exclusion: each inequality rejects some subset of points in $\mathbb{R}^n$; survivors/members are those than satisfy simultaneously all these checks.



* $\text{CONICHULL}(\{x_1 \cdots x_k\})$ is a finitely generated cone.
* $\text{CONVEX HULL}(\{x_1 \cdots x_k\})$ is a polytope.



These description create sets by inclusion: as long as a small subset (even 2 points) linearly combine to produce a point, it is in the set. A deep result in polyhedral theory is that these ways of constructing sets are equally powerful.

SOME PRE REQS.

* Recall Farkas' Lemma: $Ax \leq b$ is infeasible iff $\exists \lambda \geq 0$, $\lambda^T A = 0$, $\lambda^T b < 0$.

FARKAS' LEMMA FOR STANDARD FORM

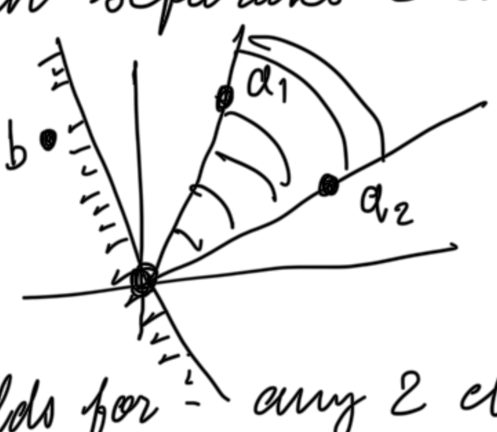
$Ax = b$, $x \geq 0$ is feasible iff $\forall \lambda \quad \lambda^T A \leq 0 \Rightarrow \lambda^T b \leq 0$.

Interpretation: $Ax = b$, $x \geq 0$ is feasible iff $b \in \text{CONICHULL}(\{a_1 \cdots a_n\})$.

The theorem says either this happens, or there is a hyperplane passing through the origin which separates b and $\{a_1 \cdots a_n\}$.

Either

or



The existence of such hyperplane holds for any 2 closed disjoint convex sets, at least one of which is compact. But for our purposes, Farkas' Lemma suffices.

PROOF. ($\Longrightarrow$) If $Ax=b$, $x\geq 0$ is feasible, then $\forall\lambda$ $\lambda^T A x = \lambda^T b$. If $\lambda^T A \leq 0$, then $\lambda^T b = (\lambda^T A)x \leq 0$ since $\lambda^T A x$ is a dot product between a non-negative 'x' & a non-positive $(\lambda^T A)^T$.

($\Longleftarrow$) We will prove the contra-positive; i.e. $A \Rightarrow B \equiv \neg B \Rightarrow \neg A$. If $Ax \leq b$, $-Ax \leq -b$, $-x \leq 0$ is infeasible, $\exists\, \lambda_1, \lambda_2, \lambda_3 \geq 0$ such that $(\lambda_1-\lambda_2)^T A - \lambda_3 = 0$ and $(\lambda_1-\lambda_2)^T b < 0$. We rewrite this as $(\lambda_2-\lambda_1)^T A = -\lambda_3 \leq 0$, and $(\lambda_2-\lambda_1)^T b > 0$ to complete the proof. $\square$

* (A,R) is a double description pair iff $\forall x$

$$Ax \leq 0 \iff \exists \lambda \geq 0 \quad x = R\lambda.$$

LEMMA. (A,R) is a DDP iff (R^T, A^T) is a DDP.

PROOF. By symmetry, it is enough to prove one side. Say (A,R) is a DDP. Then, we have for any 'x' that

$$R^T x \leq 0$$

$\iff \forall \lambda \geq 0 \quad \lambda^T R x = (R\lambda)^T x \leq 0$

$\iff \forall y \quad Ay \leq 0 \Longrightarrow y^T x \leq 0$ using (A,R) is DDP.

$\iff \exists \lambda \geq 0,\ A\lambda = x$ by Farkas' Standard-form Lemma. $\square$

MINKOWSKI WEYL THEOREM FOR CONES: Any polyhedral cone is a finitely generated cone, and vice versa.

PROOF. We'll prove that $\forall R, \exists A$ such that $\text{CONICHULL}(R) \ni x$ iff $Ax \leq 0$. Take any x. Consider $\{\lambda R - x \leq 0,\ x - R\lambda \leq 0,\ -\lambda \leq 0\}$. Run Fourier-Motzkin to eliminate all λ's. Since we start with homogenous inequalities. We arrive at $\{Ax \leq 0\}$ for some A such that the new system is feasible in x iff the former system is feasible in (x,λ) for $\lambda \geq 0$. This establishes that any finitely generated cone is a polyhedral cone. Finally, using the DDP lemma, we also get that $\forall R\ \exists A$ $\text{CONIC HULL}(A^T) \ni x$ iff $R^T x \leq 0$, completing the proof. $\square$

MINKOWSKI WEYL THEOREM FOR POLYHEDRA. Any polyhedra is expressible as the sum of a polytope and a finitely generated cone, and vice-versa.

* $P = \{x : Ax \leq b\}$. $C_P = \text{CONICHULL}\left(\left\{\begin{pmatrix} x \\ 1 \end{pmatrix} : x \in P\right\}\right) = \left\{\begin{matrix} Ay - bt \leq 0 \\ t \geq 0 \end{matrix}\right\}$. $x \in P$ iff $\begin{pmatrix} x \\ 1 \end{pmatrix} \in C_P$. from HW.

PROOF: ($\Rightarrow$) x satisfies $Ax \le b$ iff $\binom{x}{1} \in C_P$. But $C_P = \text{CONE}\left(\left\{\binom{p_1}{1} \cdots \binom{p_k}{1}, \binom{q_1}{0}, \cdots \binom{q_l}{0}\right\}\right)$ by MW for cones for some p's, q's. $\binom{x}{1} \in C_P$ iff $x = \sum_{i=1}^{k} \lambda_i p_i + \sum_{i=1}^{l} u_i q_i$, and $\sum_{i=1}^{k} \lambda_i = 1$. Therefore $P = \text{CONVEX}(p_1 \cdots p_k) + \text{CONE}(q_1 \cdots q_l)$.

($\Leftarrow$) Note $x \in \text{CONVEX}(p_1 \cdots p_k) + \text{CONE}(q_1 \cdots q_l)$ iff $x \in C_P = \text{CONE}\left(\binom{p_1}{1} \cdots \binom{p_k}{1} \binom{q_1}{0} \cdots \binom{q_l}{0}\right)$. Then, by MW for cones, $\exists A, b$ such that $C_P = \{(y,t) : Ay - bt \le 0\}$. Take $P = \{x : \binom{x}{1} \in C_P\} = \{x : Ax \le b\}$. $\square$

* Note that $\{0\}$ is the bounded cone. Hence, we reach the following.

COROLLARY: Any bounded polyhedra is a polytope, and vice-versa.

In fact, we can characterize such polytopes.

COROLLARY: Any bounded polyhedra is a convex hull of its vertices.

PROOF: Any bounded polyhedron $P = \text{CONVEX HULL}(\{x_1 \cdots x_p\})$ for some vectors x_p's. Iteratively delete any $x_i \in \text{CONVEX HULL}(\{x_1 \cdots x_p\} - \{x_i\})$ so that $P = \text{CONVEX HULL}(Q = \{x_1 \cdots x_q\})$ where $x_i \notin \text{CONVEX HULL}(Q - \{x_i\})$ $\forall i \in [q]$. Let V be the set of all vertices of P. We claim $Q \subseteq V$. Else, generically say $x_1 \notin V$. Then $\exists u \neq v \in P$, $\lambda \in (0,1)$ such $x_1 = \lambda u + (1-\lambda) v$. Also, note $u = \sum_{i=1}^{q} \alpha_i x_i$, $v = \sum_{i=1}^{q} \beta_i x_i$, where $\alpha, \beta \geq 0$, $\mathbb{1}^T \alpha = \mathbb{1}^T \beta = 1$ and $\alpha_1, \beta_1 < 1$. But then

$$x_1 = \frac{1}{1 - \lambda\alpha_1 - (1-\lambda)\beta_1} \sum_{i=2}^{q} (\lambda\alpha_i + (1-\lambda)\beta_i) x_i;$$ a contradiction.

Hence, $P = \text{CONVEX HULL}(Q) \subseteq \text{CONVEX HULL}(V)$. Since $V \subseteq P$, $\text{CONVEX HULL}(V) \subseteq P$. $\square$

COMMENTS ON GENERAL FORM LPs

- * No good reason why LPs $Ax \le b$ should have sparse solutions at all.
- * Modified BFS definition: x is a BFS for $Ax \le b$ ($x \in \mathbb{R}^n$) if there are atleast 'n' active/tight constraints at x, with linearly independent a_i's.
- * But consider: $\max C^T x$ s.t. $C^T x = d$. It's feasible, has bounded optima.

Yet $\max C^T x$ s.t. $C^T x \le d$, $-C^T x \le -d$ has no extreme points/vertices/BFS.

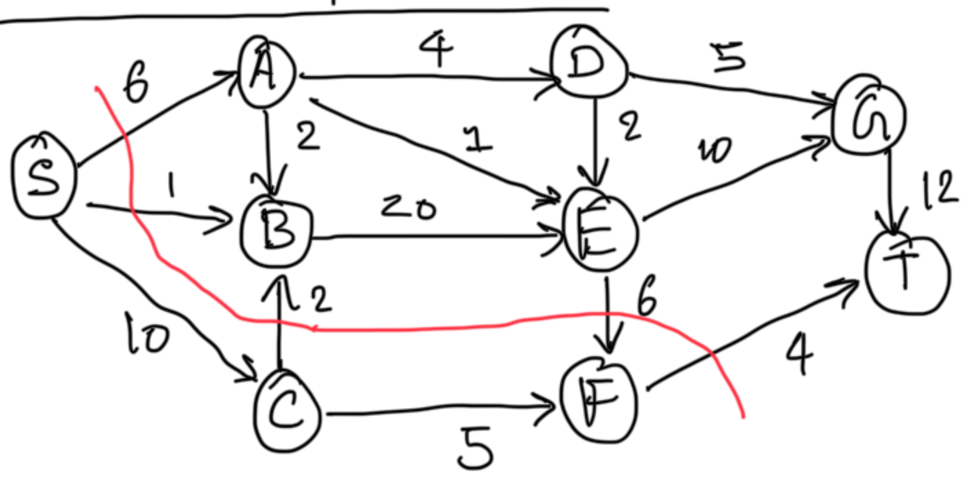
* A set S is pointed if it does not contain a line (extending infinitely in both directions), i.e. if $\nexists\, d \in \mathbb{R}^n$ such $\forall \lambda \in \mathbb{R}$, $x \in S$, $x + \lambda d \in S$.

* Fundamental Theorem of Simplex: For feasible LP in the general form with a pointed feasible set and bounded optima, $\exists$ a BFS which attains the optimal value.

References:

1. Optimality of BFSs—
    1. Section 4.2 in Matousek
2. Minkowski Weyl Theorems—
    1. Section 3.5 in Gerard's book
    2. Section 3.5 in Fukuda
3. Results on LPs in general form
    1. Section 2.2 and 2.3 in Bertsimas

LECTURE 5: DUALITY



Max Flow $= 5+4+4 = 13$.
How to know/certify this is optimal? RED LINE.
Magic: such certificate always exists.

Heuristically: Computing Duals

$$\max_x c^T x \quad \text{s.t. } Ax \le b \quad \text{PRIMAL} \;=\; \max_x \min_{y \ge 0} c^T x + y^T(b - Ax)$$
$$= y^T b + x^T(c - A^T y)$$
$$\overset{?}{=} \min_{y \ge 0} \max_x y^T b + x^T(c - A^T y)$$
$$= \min b^T y \quad \text{s.t. } c = A^T y,\; y \ge 0 \quad \text{DUAL}$$

$$\min c^T x \quad \text{s.t. } Ax = b,\; x \ge 0 \;=\; \min_{x \ge 0} \max_\lambda c^T x + \lambda^T(b - Ax) \overset{?}{=} \max_\lambda \min_{x \ge 0} + b^T\lambda + x^T(c - A^T\lambda)$$
$$= \max b^T \lambda \quad \text{s.t. } c \ge A^T \lambda.$$

We'll make the questionable steps concrete by the end of the lecture.

DUAL

* Notice that dual of the dual of a LP is the LP itself.

Interpretation: Recall the diet problem. $\min c^T x$ s.t. $Ax \ge b$, $x \ge 0$ $\longrightarrow$ $\max b^T y$ s.t. $A^T y \le c$, $y \ge 0$.
A seller wants to sell nutrient tablets at prices y while making sure that food is costlier than its constituents.

WEAK DUALITY THM. Any feasible y in $\max b^T y$ s.t. $A^T y \le c$ (DUAL) provides a lower bound on value of $\min c^T x$ s.t. $Ax = b$, $x \ge 0$ (PRIMAL).

PROOF. Take any feasible x, y; if primal is infeasible, any real $< +\infty$.
Then, $c^T x \ge (A^T y)^T x = y^T A x = y^T b$ since $x \ge 0$, $c - A^T y \ge 0$. $\square$

Implication: If dual is unbounded, then primal is infeasible & vice-versa.

Weak duality is not an accident. Whenever we exchanged the orders of min/max operators in the heuristic derivation, there's a consistent assignment of $\ge / \le$ that held consistently. Take $f: X \times Y \to \mathbb{R}$. Clearly, $\forall x \in X\; \forall y \in Y$, $\max_{y \in Y} f(x,y) \ge f(x,y)$.
Then, $\forall y \in Y$, $\min_{x \in X} \max_{y \in Y} f(x,y) \ge \min_{x \in X} f(x,y)$. $\therefore$ $\min_{x \in X} \max_{y \in Y} f(x,y) \ge \max_{y \in Y} \min_{x \in X} f(x,y)$

Any PRIMAL / DUAL pair suffers from 1 of 4 fates:

(1) Both P & D are infeasible.

(2) P is infeasible, D has unbounded optima.

(3) D is infeasible, P has unbounded optima

(4) P is feasible and has bounded optima. Then D is feasible and has bounded optima. Further, the optimal values of P & D match.

STRONG DUALITY THM. If $\min_{\substack{Ax=b\\x\geq 0}} c^T x$ is feasible & has bounded optima, then $\max_{c \geq A^T y} b^T y$ is feasible and

$$\min_{\substack{Ax=b\\x\geq 0}} c^T x = \max_{c\geq A^T y} b^T y.$$

PROOF. Let $v^* = \min_{\substack{Ax=b\\x\geq 0}} c^T x$. We will prove $\exists y$ such that $A^T y \leq c$, $b^T y \geq v^*$.

This is enough since for any feasible y, $b^T y \leq v^*$ by weak duality. Let's assume $\nexists y$ such $A^T y \leq c$, $-b^T y \leq -v^*$. Then, by Farkas' lemma,

$\exists \begin{pmatrix}\lambda\\ \lambda_0\end{pmatrix} \geq 0$ such $\lambda^T A^T - \lambda_0 b^T = 0$ and $\lambda^T c - \lambda_0 v^* < 0$.
(or $A\lambda = \lambda_0 b$) (or $\lambda^T c < \lambda_0 v^*$)

Case A: If $\lambda_0 > 0$, then $\tilde{x} = \lambda/\lambda_0$ satisfies $A\tilde{x} = \frac{A\lambda}{\lambda_0} = b$, $\tilde{x} \geq 0$ and $c^T \tilde{x} = \frac{c^T\lambda}{\lambda_0} < v^*$. A contradiction.

Case B: If $\lambda_0 = 0$, then take any feasible x^* with $c^T x^* = v^*$. Consider $\tilde{x} = x + \lambda$. Then, $\tilde{x} \geq 0$, $A\tilde{x} = Ax + A\lambda = b + 0 = b$ and $c^T\tilde{x} = c^T x^* + c^T\lambda < v^*$. A contradiction, again. □

Although we are using Farkas's lemma here, morally, strong duality is an 'obvious' consequence of the completeness of the Fourier Motzkin algorithm in deriving valid linear inequalities. Concretely $\max_{x} \; c^T x = v^*$ with $Ax \leq b$ is equivalent to $\forall x \; Ax \leq b \Longrightarrow c^T x \leq v^*$. If the last implication is true, FM can prove it by combining rows of $Ax \leq b$ with non-negative multipliers $y \geq 0$. If so, $y^T A = c$ & $y^T b \leq v^*$. Also, for any such y, by weak duality we have $y^T b \geq v^*$.

THM. ON COMP. SLACKNESS If x^* maximizes $\begin{matrix} c^Tx \\ \text{s.t. } Ax \le b \end{matrix}$, y^* minimizes $\begin{matrix} b^Ty \\ \text{s.t. } A^Ty = c \\ y \ge 0 \end{matrix}$, then

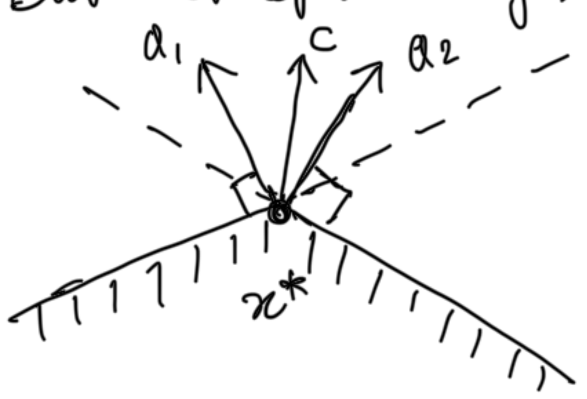
$$\forall i \in [m], \quad y_i^*(a_i^T x^* - b_i) = 0.$$

In words, y_i^* can be positive only when the corresponding inequality $a_i^T x^* \le b_i$ is tight for x^*.

PROOF. By strong duality, $0 = c^Tx^* - b^Ty^* = y^{*T}(Ax^* - b) = 0$.
Yet, $y^* \ge 0$, and $Ax^* - b \le 0$. Each term in the dot product is non-positive, yet the sum is zero $\Rightarrow$ each term is zero. □

Geometric Interpretation: $\begin{matrix} A^Ty = c \\ y \ge 0 \end{matrix}$ is same as $c \in \mathrm{CONE}(\underbrace{a_1, \dots, a_m}_{\text{rows of } A})$.

But at optimality, we can say more.



Intuitively, or by KKT conditions, at x^*,

$$c \in \mathrm{CONE}(a_i\text{'s of active constraints at } x^*).$$

This is exactly what comp. slackness implies.

$$c = A^Ty^* = (A^T)_{supp(y^*)} \, y^*_{supp(y^*)}$$

$$\in \mathrm{CONE} \text{ of active constraints at } x^*.$$

In fact, Strong Duality is same as $\begin{pmatrix} c \\ v^* \end{pmatrix} \in \mathrm{CONE}\left(\begin{pmatrix} a_i \\ b_i \end{pmatrix} \text{ for active } a_i^Tx^* \le b \text{ at } x^*\right)$.

APPLICATION ONE: ROBUST LPs

(c, A, b) are inputs to a LP. But do we know these to absolute certainity? Often not. Say we know $c \in U_c$, $a_i \in U_{a_i}$, $b_i \in U_{b_i}$. Can we optimize for the worst-case value?

$$\begin{matrix} \min\limits_{x} \max\limits_{c \in U_c} c^Tx \\ \forall i \in [m] \;\; a_i^Tx \le b_i \;\; \forall a_i \in U_{a_i}, b_i \in U_{b_i} \end{matrix} \iff \begin{matrix} \min\limits_{x,z} \; z \\ c^Tx \le z \;\; \forall c \in U_{c_i} \\ \forall i \in [m] \;\; a_i^Tx \le b_i \;\; \forall a_i \in U_{a_i}, b_i \in U_{b_i}. \end{matrix}$$

Thus we can assume w.l.o.g that there's uncertainity in c.

Similarly, we can get rid of uncertainity in b_i, by choosing $b_i = \min\limits_{b_i \in U_{b_i}} b_i$ regardless of x, since $a^Tx \le b_1 \Rightarrow a^Tx \le b_2 \; \forall b_2 \ge b_1$

So, there's just U_{a_i}. Let's stick to polyhedral uncertainity.

$U_{a_i} = \{a_i : D_i a_i \le d_i\}$. A priori, this is a LP with uncountably infinite constraints. Can we solve it? Assume for simplicity that U_{a_i} is bounded & feasible.

$$\min_x \; c^T x \quad \forall i \in [m] \; a_i^T x \le b_i \; \forall a_i \text{ such } D_i a_i \le d_i \iff \min_x \; c^T x \quad \forall i \in [m] \begin{bmatrix} \max_{a_i} & a_i^T x \\ D_i a_i \le d_i \end{bmatrix} \le b_i .$$

By duality, $\max_{a_i} a_i^T x$ s.t. $D_i a_i \le d_i$ $=$ $\min_{p_i} d_i^T p_i$ s.t. $D_i^T p_i = x$, $p_i \ge 0$. But now, we have

$$\min_x \; c^T x \quad \forall i \in [m] \begin{bmatrix} \min_{p_i} & d_i^T p_i \\ & D_i^T p_i = x \\ & p_i \ge 0 \end{bmatrix} \le b_i \iff \begin{aligned} &\min_{x, \{p_i\}} \; c^T x \\ &\forall i \in [m] \quad d_i^T p_i \le b_i \\ &\qquad\qquad D_i^T p_i = x \\ &\qquad\qquad p_i \ge 0 . \end{aligned}$$

Ex Prove the last equivalence.

Finally, we have an LP & can solve robust LPs with polyhedral uncertainty efficiently.

CAUTIONARY TALE: BILEVEL LPs

Generally, LPs embedded inside other LPs is a recipe for computational intractability. We get lucky above.

General Bilevel LP

$$\begin{aligned} \max_x \;& c^T x + d^T y \\ \text{s.t. } & Ax + By^* \le f \\ & y^* = \arg\max_y \; d'^T y \\ & \quad \text{s.t. } Cx + Dy \le g . \end{aligned}$$

Knapsack is NP-hard.

$$\begin{aligned} \max \;& \sum_{i=1}^n a_i x_i \\ \text{s.t. } & \sum_{i=1}^n a_i x_i \le B \\ & x_i \in \{0,1\} . \end{aligned}$$

We will embed Knapsack in a bilevel LP to demonstrate that solving general bilevel LPs is NP-hard.

Integral to Switching Constraints

$$\begin{aligned} \max \;& \sum_{i=1}^n a_i x_i - 10^{100} \sum_{i=1}^n y_i \\ \text{s.t. } & \sum_{i=1}^n a_i x_i + 10^{100} \sum_{i=1}^n y_i \le B \\ & y_i = \min\{x_i, 1 - x_i\} \\ & 0 \le x_i \le 1 \end{aligned}$$

Knapsack as Bilinear LP

$$\begin{aligned} \max \;& \sum_{i=1}^n a_i x_i - 10^{100} \sum_{i=1}^n y_i \\ \text{s.t. } & \sum_{i=1}^n a_i x_i + 10^{100} \sum_{i=1}^n y_i \le B \\ & 0 \le x_i \le 1 \\ & y_{1:n} = \arg\max \sum_{i=1}^n y_i \\ & \quad \text{s.t. } y_i \le x_i \\ & \qquad\;\; y_i \le 1 - x_i . \end{aligned}$$

Ex Prove that if $\max_i a_i \le 10^{100}$, then optima of switching formulation and knapsack coincide.

APPLICATION TWO: TWO PLAYER ZERO SUM GAMES

	R	P	S
R	0,0	−1,1	1,−1
P	1,−1	0,0	−1,1
S	−1,1	1,−1	0,0

payoff of row player, column player

Note these sum to zero.

Let A be a matrix of payoffs for the column player.

If row player goes first, $\min_i \max_j A_{ij}$.

If column player goes first, $\max_j \min_i A_{ij}$.

These are clearly not equal, i.e. $+1 \neq -1$.

__Thm.__ If players are permitted to choose randomized/mixed strategies, then order of play is irrelevant.

[VON NEUMANN]

$$\min_{x\in\Delta} \max_{y\in\Delta} x^T A y = \max_{y\in\Delta} \min_{x\in\Delta} x^T A y.$$

__Proof Sketch__

$$\min_{x\in\Delta} \max_{y\in\Delta} x^T A y \geq \max_{y\in\Delta} \min_{x\in\Delta} x^T A y$$

$$= \min_{x\in\Delta} \max_{j} (A^T x)_j \qquad = \max_{y\in\Delta} \min_{i} (Ay)_i$$

$$P = \min_{z, x\in\Delta} z \quad \text{s.t.} \quad z\mathbb{1} \geq A^T x$$

Feasible: take any $x \in \Delta$ & $z \geq \max_{i,j} |a_{ij}|$.

$$D = \max_{w, y\in\Delta} w \quad \text{s.t.} \quad w\mathbb{1} \leq Ay.$$

Feasible: take any $y \in \Delta$ & $w \leq -\max_{i,j} |a_{ij}|$.

Therefore, by strong duality, enough to prove P & D are duals.

EX Verify this via explicit computation.

More subtle point: P & D are automatically duals since this is exactly how we derive duals. $\square$

References:

1. Computing duals
   1. Mechanically— Section 6.2 in Matousek; also see this
   2. Via minimax inequality— Sections 5.2.1 and 5.4 in Boyd
2. Proofs of Strong Duality—
   1. Lecture 5 from Amir Ali's course notes are the tidiest; also discusses robust LPs
   2. Section 3.3 in Gerard's book provides a direct proof via FM
3. (Beyond this course) More on robust programs by Nemirovski
4. Zero sum games— Section 5.2.5 in Boyd
5. Hardness of Bilevel LPs— this paper

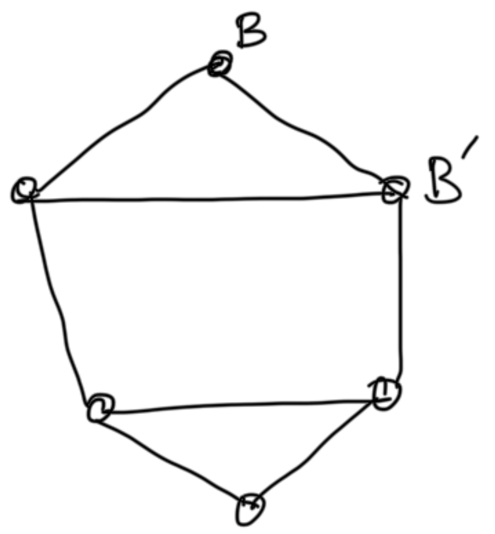
LECTURE 6: SIMPLEX

Let us recall the BFS enumeration algorithm from GEOMETRY. The simplex algorithm searches for an optimal BFS in a local manner. Questions that arise:

$$\min c^T x \quad Ax = b \quad x \geq 0$$

(1) What's this notion of locality?
(2) Why does local search lead to optimality?
(3) How to even obtain an initial BFS? Recall that checking for feasibility is almost as hard as optimization itself.

__Answer 1.__ Recall that each $B \subseteq [n]$ of size 'm' can correspond to at most one BFS in $\{Ax = b, x \geq 0\}$. We can draw a graph over B's that yield a (feasible) BFS.

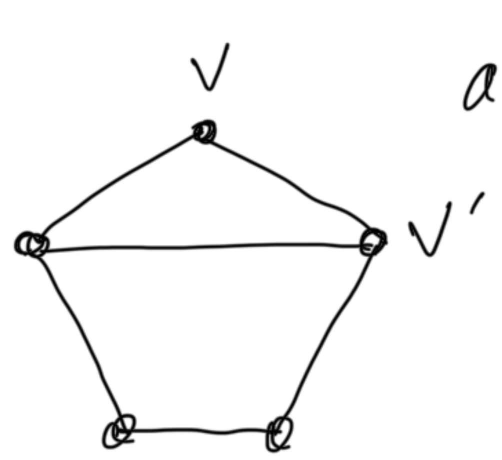

ALGEBRAIC GRAPH

$B - B'$ iff $|B \cap B'| = m-1$, i.e., if B & B' share all but one coordinate.

Roughly, simplex searches for better (or not worse) neighbours in this graph.

Note multiple B's might correspond to same BFS. This creates complications later. To stop this:

NON-DEGENERACY: All BFSs have m-non-zero coordinates, or equivalently, each feasible B produces a unique BFS.


GEOMETRIC GRAPH

$V - V'$ are connected iff there are connected by an edge, i.e., a 1-dimensional face.

Under non-degeneracy, ALGEBRAIC GRAPH = GEOMETRIC GRAPH.

Generally, Geometric Graph is an edge contraction of the Algebraic Graph, formed by collapsing B's that lead to the same BFS. By its nature, simplex performs local search over the algebraic graph. But local search over the geometric graph is easier to analyze.

Answer 3. In short, by cheating. Construct an auxilliary LP for which a BFS is easy to guess.

$$\text{(ORIGINAL LP)}\quad \begin{array}{c} \min c^T x \\ Ax = b \\ x \geq 0 \end{array} \qquad\qquad \text{(AUXILLIARY LP)}\quad \begin{array}{c} \min \mathbb{1}^T s \\ Ax + s = b \\ x \geq 0 \\ s \geq 0 \end{array}$$

Without loss of generality, $b \geq 0$.
Else, flip sign.

OBSERVATION: Original LP is feasible $\Longleftrightarrow$ Auxilliary LP's OPTIMUM is 0.

Idea: Run simplex on Aux LP starting with $x=0, s=b$ (which is a BFS). If opt >0 or unbounded, original LP is infeasible. Else, we end up with a BFS for original LP to run simplex on (maybe after using basis expansion from LECTURE 3).

Answer 2: Okay, this is a bit involved / annoying.

ASSUMING NON-DEGENERACY, SIMPLEX:

1. Start with a BFS.
2. Check if there's a neighboring BFS with strictly better value. If so, then move to it & repeat. Else, declare current BFS is optimal.

Consider any neighbor of B with BFS x, B' with BFS y. Let $B' - B = \{i\}$, $d_B = y_B - x_B$. Then, $Ay = A_B(x_B + d_B) + a_i y_i = b$, or $A_B d_B + a_i y_i = 0$, where a_i is i^{th} column of A. Also,

$$c^T y - c^T x = c_i y_i + c_B^T d_B = (c_i - c_B^T A_B^{-1} a_i) y_i.$$

OBSERVATION: $c^T x \leq \min\limits_{y \in \text{NEIGHBOR of } x} c^T y$ for B generating BFS x.

$$\Longleftrightarrow (c_i - c_B^T A_B^{-1} a_i) y_i \geq 0 \qquad \forall i \in \overline{B}.$$

where y is a BFS for B' such $B' - B = \{i\}$.

$\Leftarrow c_i - c_B^T A_B^{-1} a_i \geq 0$ $\forall i \in [n]$ since $y_i \geq 0$.

Defn: i^{th} reduced cost at B

$\Longrightarrow c_i - c_B^T A_B^{-1} a_i \geq 0$ $\forall i \in [n]$ since $y_i > 0$.

for non-degenerate LPs

Here, we are using that $\forall i \in B$, $c_i = c_B^T A_B^{-1} a_i = c_B^T e_i$, using definition of matrix inverse.

THEOREM: Define $\bar{c}^T = c^T - c_B^T A_B^{-1} A$ to be the reduced cost at B. Then, (1) $\bar{c} \geq 0 \Longrightarrow$ BFS x with B is optimal; (2) BFS x at B is optimal and LP is non-degenerate $\Rightarrow \bar{c} \geq 0$.

PROOF: Let's start with (2). If x is optimal, it must be be at least as good as its neighbors. Then for non-degenerate LPs, $\bar{c} \geq 0$, by the previous observation. For (1), we'll certify optimality by constructing a dual feasible solution. $\bar{c} \geq 0 \Longleftrightarrow c^T \geq c_B^T A_B^{-1} A = y^T A$, where $y^T = c_B^T A_B^{-1}$, or $A^T y \leq c$. Thus y is a dual feasible solution. Yet, $b^T y = c_B^T A_B^{-1} b = c_B^T x_B = c^T x$. Hence, by weak duality, 'x' is optimal. $\square$

COROLLARY: For non-denegrate LPs, simplex with strictly better neighbor rule terminates in a finite number of steps and reaches an optimum.

This corollary is immediate since non-zero improvement at each step implies no vertex/BFS is visited twice, and recall that there are at most $\binom{n}{m}$ of them. The last theorem guarantees optimality at stopping.

SIMPLEX FOR POSSIBLY DEGENERATE LP.

1. Start at some BFS x with base B.
2. Check if $\bar{c} \geq 0$. If so, declare optimality. Else,

choose a neighbouring vertex B' such $B' - B = \{i\}$ such $\bar{c}_i < 0$, using a PIVOTING RULE. Move to it; repeat.

Now, that we have given up the invariant of strict improvement every step, there's the possibility that simplex cycles (visits the same base B twice) and never terminates. Recall that asking for strict improvement (and stopping when it is not possible) impugns on the correctness/optimality; that's worse. Many natural pivoting rules for simplex cycle.

<u>BLAND'S RULE.</u> When at a base B, choose the smallest index i for which $\bar{c}_i < 0$; move to B' such $B' - B = \{i\}$.

<u>THEOREM.</u> Simplex with Bland's rule does not cycle, and hence, terminates at an optimum in finite steps.

We will not prove this in interest of time. Maybe in a future iteration of this course. Generally, such lexicographic (not invariant to naming/order of indexing) rule provide a consistent way of tie-breaking.

COMMENTS OF RUNNING TIME OF SIMPLEX

* A reasonable algorithm for LPs arising in practice. Mixed evidence on if cycling is a real concern.
* Many many pivoting rules require an exponential number of steps in the worst-case.
  KLEE-MINTY cube & variants are a common source of such hard examples.
* Multiple decades-long push to find a pivoting rule that results in polynomial complexity.

Q. Is this even true for an "omniscient" pivoting rule?

For a polytope P, let G_P be its BFS graph.

HIRSCH CONJECTURES. (1957) $\mathrm{diam}(G_P) \leq n-d$ where P is a d-dimensional polyhedron with n constraints.

True when $d \leq 3$

True when $n-d \leq 6$.

COUNTEREXAMPLE (2010 SANTOS) $\exists$ a counter-example to the conjecture in $d=43$.

THEOREM. (Kalai-Kleitman) $\mathrm{diam}(G_P) \leq 2n^{\log_2 d+1}$.

OPEN QUESTION $\mathrm{diam}(G_P) \leq \mathrm{poly}(n,d)$?

* Concession: Maybe worst-case complexity is really bad, but perhaps average-case is good. Caveat: notion of "average" in average-case is tricky. For example: an early result of this kind proved that if $(a_i, b_i) \sim D$ are sampled independently from some distribution D satisfying equipodal symmetry, i.e. if $\Pr_D((a,b)) = \Pr_D((-a,-b))$, then with n such constraints in d dimensions, simplex on $\max c^T x$ s.t. $Ax \leq b$ takes polynomially many steps with high probability. This is difficult to judge the significance of since for $m > 2n$, such LPs are infeasible with high probability for most (non-degenerate) distributions.

* SMOOTHED ANALYSIS: In ~2002, Speilman & Teng showed that given any (A, b), with

$$\tilde{A} = A + \text{random noise of size } \sigma$$
$$\tilde{b} = b + \text{random noise of size } \sigma,$$

the simplex algorithm takes $\text{poly}(n, d, \frac{1}{\sigma})$ steps on $\max\limits_{\tilde{A}x \leq \tilde{b}} c^T x$. Notice that although this is a statement about random instances, the randomness is very "localized". Such mode of analysis between worst & average-cases is called SMOOTHED ANALYSIS, and has proved useful in studying efficiency of algorithms in beyond worst-case settings more generally.

References:

1. Finding an initial BFS— page 70 in Matousek
2. Simplex algorithm
    1. Sections 3.1 and 3.5 in Bertsimas
    2. Section 11.1 in Schrijver; also proves termination of Bland's rule
3. (Beyond this course) Survey on Hirsch Conjecture
4. (Beyond this course) Smoothed analysis— Daniel Dadush's talk

LECTURE 7: CENTER OF MASS

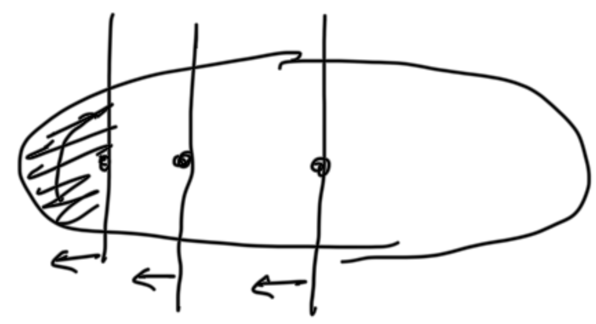
For any compact $K \subseteq \mathbb{R}^n$, $vol(K) = \int_K dx$, $COM(K) = \int_K \frac{x\,dx}{vol(K)}$.

Note $COM(K) = \mathbb{E}_{x \sim Unif \text{ on } K}[x]$.

Take any $\min_{x \in K} c^T x$; this can represent any convex prog.

$x \in K \longrightarrow$ compact, full-dimensional, convex.

ALGORITHM.



1. Set $K_1 = K$.
2. For $t = 1 \dots T$

$$\text{compute } x_t = \frac{1}{vol(K_t)} \int_{K_t} x\,dx.$$

$$\text{Take } K_{t+1} \leftarrow K_t \cap \{x : c^T x \le c^T x_t\}.$$

3. Output $\bar{x} = \arg\min_{x \in \{x_1 \dots x_{T+1}\}} c^T x$.

CLAIM. If $\max_{x,y \in K} c^T(x-y) \le F$, then $c^T \bar{x} \le \min_{x \in K} c^T x + F\left(1 - \frac{1}{e}\right)^{T/n}$.

In particular, if $T \geq n \log \frac{F}{\varepsilon}$, then we must be ε-optimal.

GRUNBAUM'S LEMMA. For any convex, compact K, with COM x_0,

$$\forall c \in \mathbb{R}^n, \quad \frac{vol(K \cap \{x : c^T(x - x_0) \le 0\})}{vol(K)} \geq \frac{1}{e}.$$

In words, any half-space through the center of mass of a convex body rejects at-least $1/e$ fraction of the volume. With this interpretation, the COM algorithm has the same flavor as binary search.

PROOF OF CLAIM: Let $x^* = \arg\min_{x \in K} c^T x$. Then, take $X^*_\varepsilon = \{(1-\varepsilon)x^* + \varepsilon x : x \in K\}$ $= (1-\varepsilon)x^* + \varepsilon K$.

Now, $vol(X^*_\varepsilon) = \varepsilon^n vol(K)$.

Also, $X^*_\varepsilon \subseteq K = K_1$, by convexity of K.

$\max_{x \in X^*_\varepsilon} c^T x \geq \min_{x \in K} c^T x + \varepsilon F$, by construction of X^*_ε.

In words, X^*_ε is a small set of points, all with good objective value. We'll prove that although initially completely inside K_1, some of it must fall outside K_t for large enough t. Whenever this first happens, x_t must be better than some $x \in X^*_\varepsilon$. To see this:

$$\mathrm{vol}(K_{t+1}) \le \left(1-\frac{1}{e}\right)\mathrm{vol}(K_t) \qquad \text{GRUNBAUM}$$

$$\le \left(1-\frac{1}{e}\right)^t \mathrm{vol}(K_1). \qquad \text{By repetition.}$$

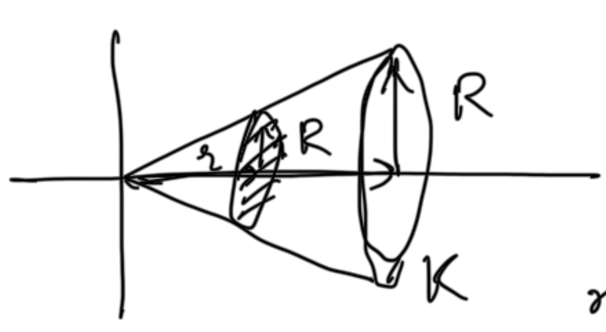
Now, set $\varepsilon > \left(1-\frac{1}{e}\right)^{T/N}$. Then, $X^*_\varepsilon \subseteq K_1$, yet $\mathrm{vol}(K_T) < \mathrm{vol}(X^*_\varepsilon)$.

Hence, $\exists\, t \in [T], x^*_\varepsilon \in X^*_\varepsilon$ such $x^*_\varepsilon \in K_t$, $x^*_\varepsilon \notin K_{t+1}$.

By construction, $c^T x_t < c^T x^*_\varepsilon \le \min_{x \in K} c^T x + \varepsilon F$. □

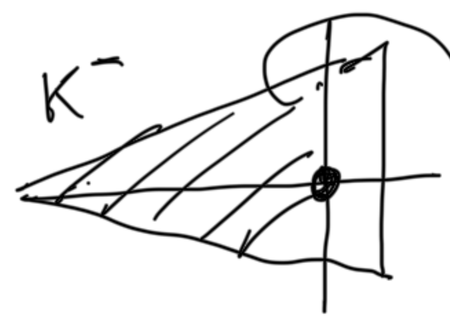
In the rest of this note, we will prove Grunbaum's lemma.

<u>OBSERVATION:</u> Say $(n-1)$-dimension volume of a sphere (radius r) is $C_{n-1} r^{n-1}$.



$$\mathrm{vol}(K) = \int_0^R C_{n-1} r^{n-1}\,dr = \frac{C_{n-1}}{n} R^n.$$

$$\underset{r\text{-coord}}{\mathrm{COM}(K)} = \frac{n}{C_{n-1}R^n}\int_0^R C_{n-1} r^n\,dr = \frac{n}{n+1} R.$$



$$\mathrm{vol}\left(K \cap \left\{x : x_1 \le \underset{r\text{-coord}}{\mathrm{COM}(K)}\right\}\right) = \int_0^{\frac{n}{n+1}R} C_{n-1} r^{n-1}\,dr = \frac{C_{n-1}}{n}\left(\frac{n}{n+1}\right)^n R.$$

$$\frac{\mathrm{vol}(K^-)}{\mathrm{vol}(K)} = \left(\frac{n}{n+1}\right)^n \ge \frac{1}{e}.$$

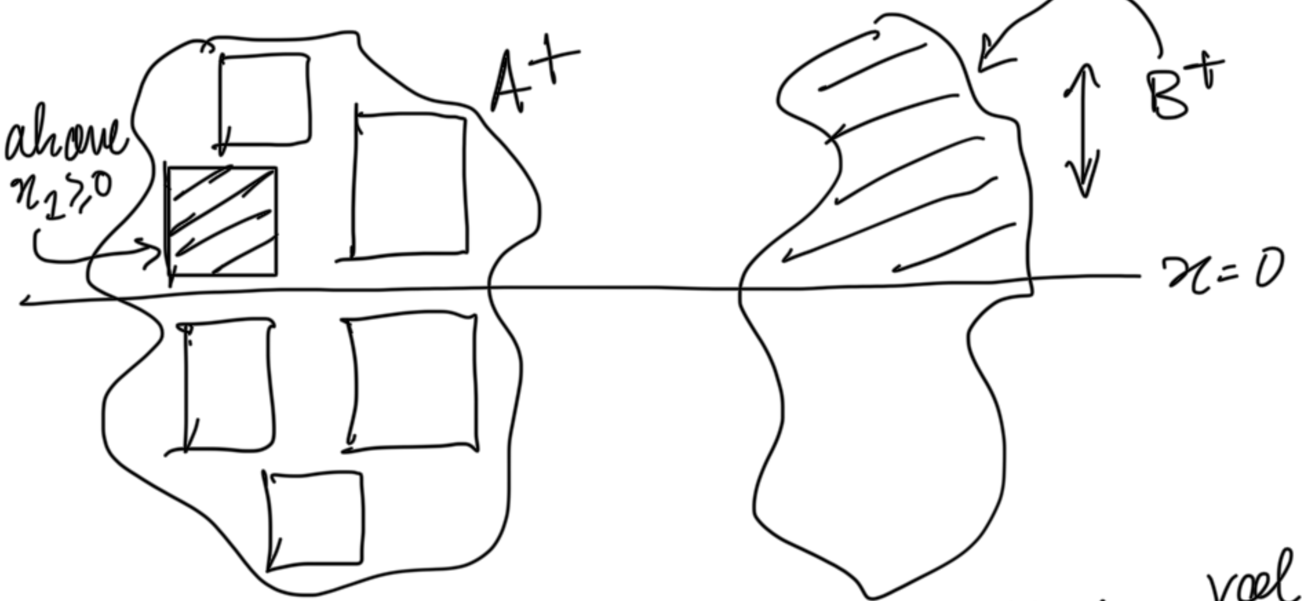
In some sense, cone is the worst case for Grunbaum.

Although, this is an example, we will use it as a proof strategy. We will reduce general convex bodies to (right) cones.

BRUNN MINKOWSKI INEQUALITY For non-empty compact sets $A, B \subseteq \mathbb{R}^n$,

$$\mathrm{vol}(A+B)^{1/n} \geq \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}.$$

PROOF. In the HW, you have proven this for the case when A & B are axis-aligned (hyper) rectangles. We will extend this to when A & B are unions of disjoint cuboids. By limiting argument, this extends to compact bodies. Our induction hypothesis is that the stated inequality is true when A & B contain 'n' disjoint cuboids in total. Volume is translation invariant. Hence, shift the $x_1 = 0$ plane so that at least one cuboid lies entirely above in A.



$A_+ = A \cap \{x : x_1 \geq 0\}$

$A_- = A - A_+$

$B_+ = B \cap \{x : x_1 \geq 0\}$

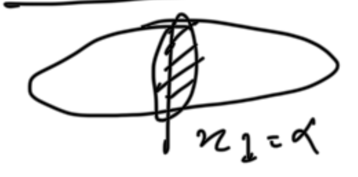
$B_- = B - B_+$

Translate B along x_2 so that $\dfrac{\mathrm{vol}(A^+)}{\mathrm{vol}(A)} = \dfrac{\mathrm{vol}(B^+)}{\mathrm{vol}(B)}$; such translation always exists due to Int. Value Theorem. Notice that $(A^+ + B^+) \cap (A^- + B^-) = \emptyset$ since $x_1 = 0$ separates them, yet $(A^+ + B^+) \cup (A^- + B^-) \subseteq A + B$. Hence, we have

$$\begin{aligned}
\mathrm{vol}(A+B) &\geq \mathrm{vol}(A^+ + B^+) + \mathrm{vol}(A^- + B^-) \\
&\geq \left(\mathrm{vol}(A^+)^{1/n} + \mathrm{vol}(B^+)^{1/n}\right)^n + \left(\mathrm{vol}(A^-)^{1/n} + \mathrm{vol}(B^-)^{1/n}\right)^n \\
&= \left(\mathrm{vol}(A^+) + \mathrm{vol}(A^-)\right)\left(1 + \left(\frac{\mathrm{vol}(B)}{\mathrm{vol}(A)}\right)^{1/n}\right)^n \\
&= \left(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}\right)^n.
\end{aligned}$$

$\square$

COROLLARY: $\underbrace{\mathrm{vol}(K \cap \{X_1 = \alpha\})}_{\text{this is } (n-1)\text{-dimensional volume.}}^{\frac{1}{n-1}}$ is concave in α for any compact, convex set $K \subseteq \mathbb{R}^n$.

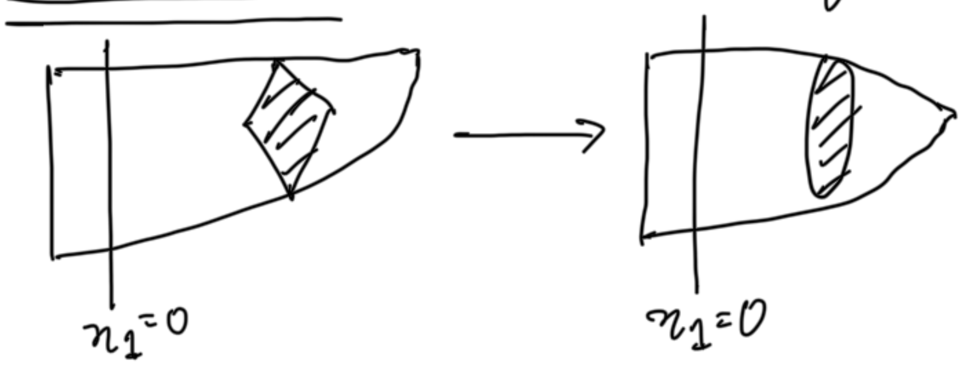


PROOF: Let $K_\alpha = K \cap \{x_1 = \alpha\}$. Note that $\lambda K_\alpha + (1-\lambda) K_\beta \subseteq K_{\lambda\alpha + (1-\lambda)\beta}$ $\forall \alpha, \beta \in \mathbb{R}$, $\lambda \in [0,1]$ since K is convex. Then,

$$\mathrm{vol}\left(K_{\lambda\alpha+(1-\lambda)\beta}\right)^{\frac{1}{n-1}} \geq \lambda \left(\mathrm{vol}(K_\alpha)\right)^{\frac{1}{n-1}} + (1-\lambda)\left(\mathrm{vol}(K_\beta)\right)^{\frac{1}{n-1}}. \quad \square$$

Now, we are ready to complete Grunbaum's lemma.

PROOF OF GRUNBAUM'S LEMMA. Without loss of generality, we can orient our convex body so that $x_1 = 0$ is the cutting hyperplane.



Replace every slice of K along x_1 axis with a $(n-1)$-dimensional sphere of equal $(n-1)$-dimensional volume. This step preserves volumes of both sections on either side of $x_1=0$; also x_1 coordinate of center-of-mass stays the same. So, it suffices to establish the claim for this new body.

$K_+ = K \cap \{x : x_1 \geq 0\}$. First note, this new body is convex, since we didn't modify $\mathrm{vol}(K \cap \{x_1 = \alpha\})$

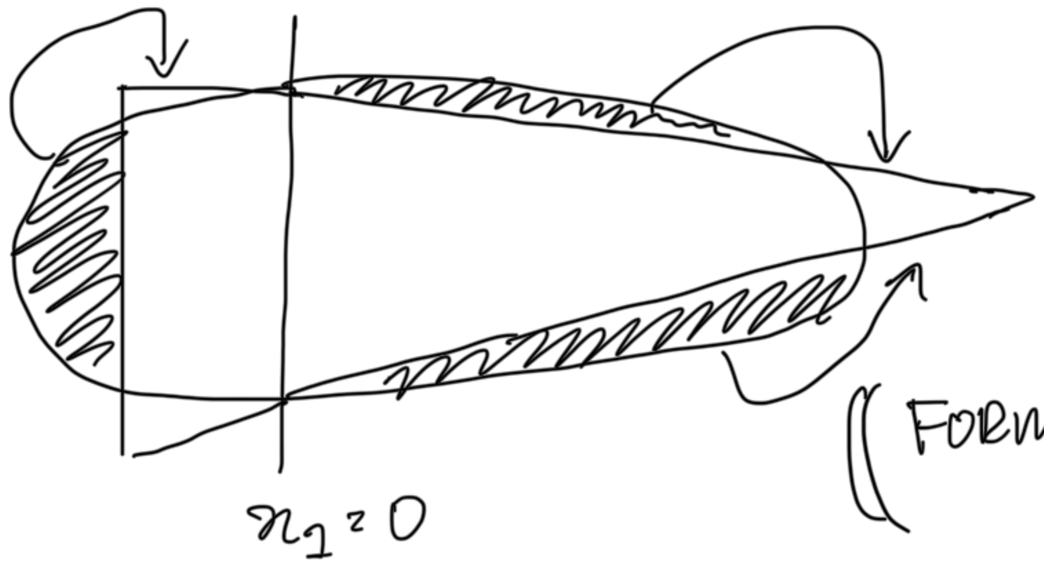
$K_- = K \cap \{x : x_1 \leq 0\}$.

and $\mathrm{vol}(K \cap \{x_1 = \alpha\})^{\frac{1}{n-1}}$ was concave in α for the old (and hence, the new) body.

Replace K^+ with a cone with the same spherical base as K^+, so that the cone and K^+ are equi-volume. Extend this cone in the negative x_1-region till this extension has volume equal to $\mathrm{vol}(K^-)$; again always possible by intermediate value theorem. These operations are volume preserving, but what happens to the center of mass? The claim is that it can only move rightwards. In other words, this transformation increases the x_1 coordinate of center of mass from 0 to something non-negative. This is once again a consequence of concavity of $\mathrm{vol}(K \cap \{x_1 = \alpha\})^{\frac{1}{n-1}}$ in α. Post this transformation, we have a perfect cone with non-negative x_1 coordinate of COM. Hence,

$$\frac{vol(K_+)}{vol(K)} = \frac{vol(K\cap\{x_1 \geq 0\})}{vol(K)} \geq \frac{vol(K\cap\{x_1 \geq x_1^{com}\})}{vol(K)} \geq \frac{1}{e},$$

where $x_1^{com} \geq 0$ is x_1-coordinate of COM. □



Mass moves rightward. Hence, this transformation shifts COM rightward.

((FORMAL proof by constructing a transport map.))

THE TROUBLE WITH CENTER OF MASS ALGORITHM

Computing COM of a general convex body given a polyhedral description is #P-hard (i.e. quite hard). But, nevertheless it is possible to get an ε-far point to COM in polynomial (in $\frac{1}{\varepsilon}$, m, n) time. This absolves the COM algorithm somewhat. But, nevertheless even the ε-approximation algorithm is almost as tedious as solving a LP. We will however see a better algorithm, in fact the first poly-time algorithm for solving LPs, inspired by the COM algorithm next lecture.

References:

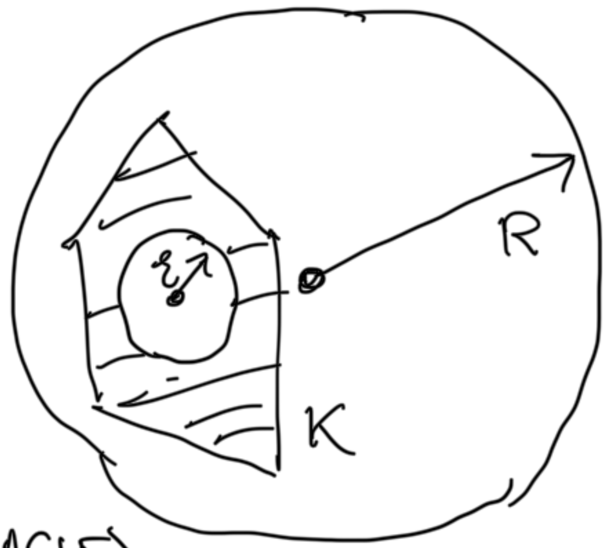
1. Center-of-mass Algorithm
    1. Section 2.1 in Bubeck
    2. Sections 3.4 and 1.7 in Lee-Vempala
2. Proofs of the Brunn-Minkowski and Grunbaum's inequality
    1. Chapter 2 in Vempala; also proves Grunbaum
    2. Lecture 13 in Kelner; also proves Grunbaum
    3. Section 9.1 in Tkocz
    4. Lecture 5 in Ball— a good intro to convex geometry; proves Prekopa-Leindler, a generalization of BM
3. (Beyond this course) Computing ~COM in poly time Bertsimas-Vempala

LECTURE 8: ELLIPSOID

The ellipsoid method can be thought of as a variant of the center-of-mass method, but one implementable in polynomial time. It was the first provably poly-time algorithm for LPs.

ASSUMPTION: K is convex, compact and $rB_2 \subseteq K \subseteq RB_2$ where $B_2 = \{x : \|x\|_2 \leq 1\}$.

ALGORITHM



1. Initialize $\mathcal{E}_1 = RB_2$.
2. For $t = 1 \ldots T$
   1. let x_1 be the center of $\mathcal{E}_1$.
   2. Is $x_t \in K$? (MEMBERSHIP ORACLE)
   3. If yes, construct an ellipse $\mathcal{E}_{t+1}$ containing $\mathcal{E}_t \cap \{x : c^T(x - x_t) \geq 0\}$.
   4. If no, ask for a half space w_t such that $\forall x \in K$, $w_t^T(x - x_t) \geq 0$. Thus, all of K is contained in $w^T(x - x_t) \geq 0$. Construct an ellipse $\mathcal{E}_{t+1}$ containing $\mathcal{E}_t \cap \{x : w_t^T(x - x_t) \geq 0\}$. (SEPARATION ORACLE)

IMPLEMENTATION

If $K = \{x : Ax \leq b\}$, then $x_t \overset{?}{\in} K$ can be answered in linear time by checking all constraints one-by-one $a_i^T x_t \leq b_i$. If $x_t \notin K$, then $\exists i \in [m]$ such $a_i^T x_t > b_i$. But $\forall x \in K$, $a_i^T x \leq b_i$, implying $a_i^T(x - x_t) \leq 0$ $\forall x \in K$. This gives us the required separating hyperplane.

IMPORTANT NOTE: We've demonstrated that for LPs with polynomially many constraints, MEMBERSHIP & SEPARATION ORACLES can be implemented efficiently. However, this is not the only case when this is possible. For certain structured LPs with exponentially / infinitely many constraints, ellipsoid is still a poly-time algorithm as long as SEPARATION / MEMBERSHIP QUERIES are efficiently answerable.

ANALYSIS

VOLUME REDUCTION LEMMA. For any ellipsoid $\mathcal{E}_0$ with center $x_0 \in \mathbb{R}^n$, and vector w_0, we can efficiently construct an ellipsoid $\mathcal{E}_1$ containing $\mathcal{E}_0 \cap \{x : w_0^T(x-x_0) \geq 0\}$ with $\mathrm{vol}(\mathcal{E}_1) \leq \mathrm{vol}(\mathcal{E}_0)\, e^{-1/2(n+1)}$.

CLAIM. Let $\bar{x}$ be a feasible point in $\{x_1 \ldots x_T\}$ achieving the minimum objective value. Then

$$c^T\bar{x} - \min_{x \in K} c^T x \leq \frac{FR}{r} e^{-T/2(n+1)n}, \text{ where}$$

$$F = \max_{x,y \in K} c^T(x-y).$$

Hence, as long as $T \geq 2n(n+1) \log \frac{FR}{r\varepsilon}$, we must be ε-optimal. Notice that this is slower than COM by a factor of n, because volume reduction $\simeq 1 - \frac{1}{2(n+1)}$ in each step, instead of a constant. This is the price ellipsoid method pays for efficient implementability.

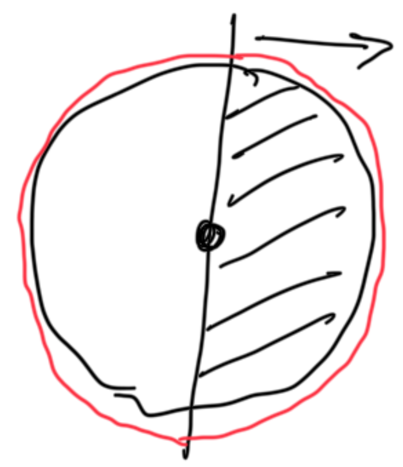
PROOF OF CLAIM. We will follow the same recipe as that for COM. Let $x^* \in \arg\min_{x \in K} c^T x$, and $X_\varepsilon^* = (1-\varepsilon)x^* + \varepsilon K \subseteq K$. Now, we have $\mathrm{vol}(X_\varepsilon^*) \geq \mathrm{vol}(\varepsilon r B_2) = C_n(\varepsilon r)^n$ for some C_n such that $\mathrm{vol}(B_2) = C_n$. Also, we have

$$\max_{x \in X_\varepsilon^*} c^T x \leq c^T x^* + \max_{x \in X_\varepsilon^*} c^T(x - x^*) \leq c^T x^* + \varepsilon F.$$

By volume reduction lemma, with repeated applications, we get $\mathrm{vol}(\mathcal{E}_{T+1}) \leq \mathrm{vol}(\mathcal{E}_1)\, e^{-\frac{T}{2(n+1)}} = C_n R^n e^{-\frac{T}{2(n+1)}}$.

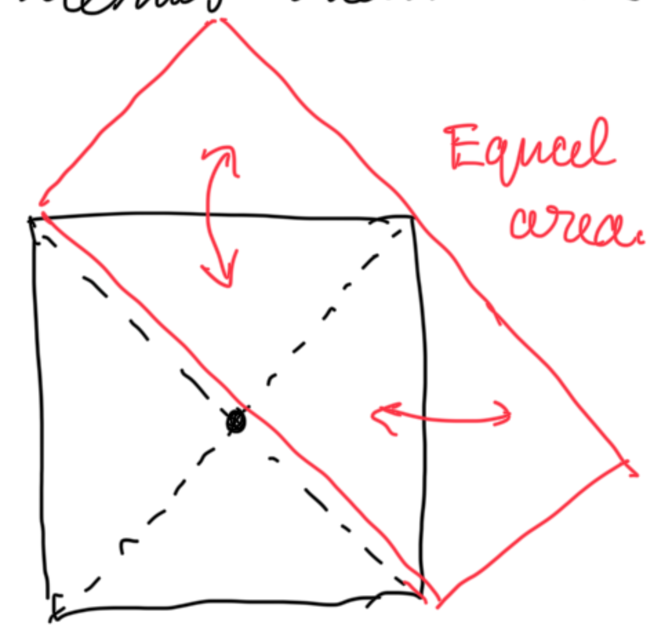
Choose $\varepsilon > \frac{R}{r} e^{-\frac{T}{2n(n+1)}}$. Then since $vol(\mathcal{E}_{T+1}) < vol(X^*_\varepsilon)$, $\exists\, t, x^*_\varepsilon \in X^*_\varepsilon$ such that $x^*_\varepsilon \in \mathcal{E}_t$, $x^*_\varepsilon \notin \mathcal{E}_{t+1}$. Further, note that this can only happen on the YES branch, because all existing feasible points are retained on the NO branch. Hence, $c^T x_t < c^T x^*_\varepsilon \le c^T x^* + \varepsilon F$. $\square$

Now, we will finish up the proof of the volume reduction lemma. Note that this lemma was crucial for the (sub)-optimality result. Also, such niceties don't work for some simpler shapes.



Smallest circle containing this hemicircle is the original circle itself.

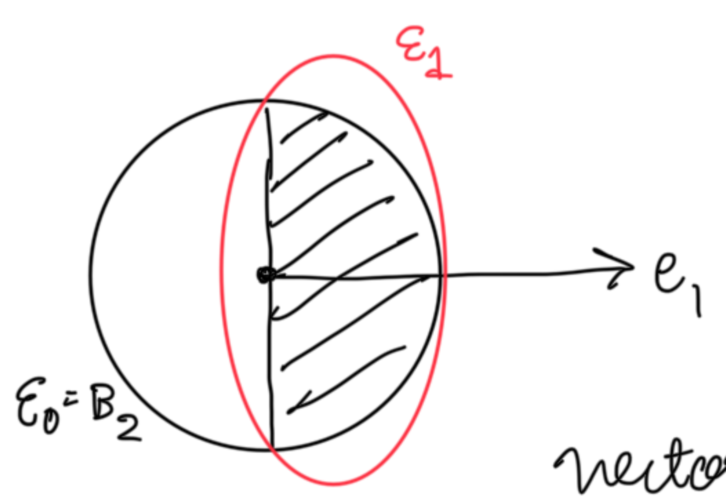
NO VOLUME REDUCTION for circle method



Equal area.

NO VOLUME REDUCTION. for rectangle method

PROOF OF VOLUME REDUCTION LEMMA. Consider a simpler case when we start with the unit ball $B_2 = \{x : \|x\|_2 \le 1\}$, $x_1 \ge 0$ is halfspace.



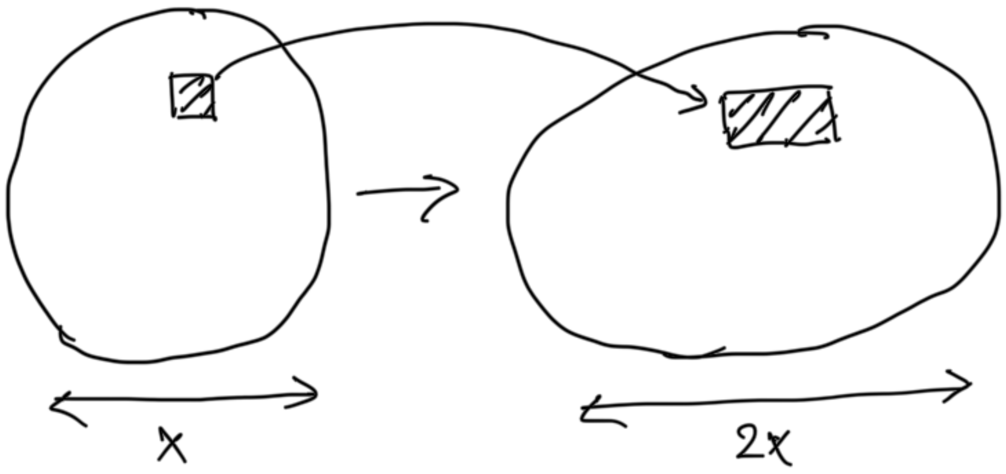
Any ellipse can be written as

$$\|x - x_0\|^2_{H_0^{-1}} = (x - x_0)^T H_0^{-1} (x - x_0) \le 1$$

where x_0 is the center, the eigen vectors of H_0 are its principal axes with lengths $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}$ where λ_i's are the eigenvectors of H_0.

Think of it as a generalization of $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \le 1$.

Clearly, an ellipse with principal axes of lengths $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}$ has volume $= C_n \sqrt{\lambda_1 \dots \lambda_n} = C_n \sqrt{\det(H_0)}$. Stretching a body along a single axis by $2\times$, increases volume by $2\times$.



Each volume element doubles in volume while stretching x-axis by $2x$.

Now, back to $\mathcal{E}_1$. By symmetry, we take $x_1 = t e_1$, as the center of $\mathcal{E}_1$. This ellipse touches e_1 and $B_2 \cap \{e_1 = 0\}$. So, our ansatz is $H = a e_1 e_1^T + b(I - e_1 e_1^T)$. This is an eigen-value decomposition. $H^{-1} = \frac{1}{a} e_1 e_1^T + \frac{1}{b}(I - e_1 e_1^T)$. So—

$$\frac{(1-t)^2}{a} = 1 \Rightarrow a = (1-t)^2 \qquad \frac{t^2}{a} + \frac{1}{b} = 1 \Rightarrow b = \frac{1}{1 - t^2/a}$$

$$= \frac{(1-t)^2}{1-2t}.$$

$$\text{volume}(\mathcal{E}_1) = C_n \sqrt{a b^{n-1}} = C_n \frac{(1-t)^n}{(1-2t)^{\frac{n-1}{2}}}.$$

Maximizing this for t, $\frac{(1-t)^{n-1}}{(1-2t)^{\frac{n-1}{2}}} n = \frac{(1-t)^n}{(1-2t)^{\frac{n-3}{2}}} \frac{n-1}{2} 2$

$$\Rightarrow \frac{n}{n-1} = \frac{1-t}{1-2t} \Rightarrow \frac{1}{n-1} = \frac{t}{1-2t} \Rightarrow t = \frac{1}{n+1}.$$

$$a = (1-t)^2 = \left(\frac{n}{n+1}\right)^2, \quad b = \frac{\left(\frac{n}{n+1}\right)^2}{\frac{n-1}{n+1}} = \frac{n^2}{n^2-1}.$$

$$\text{volume}(\mathcal{E}_1) = C_n \sqrt{a b^{n-1}} = C_n \left(\frac{n}{n+1}\right)\left(1 + \frac{1}{n^2-1}\right)^{\frac{n-1}{2}}$$

$$\leq C_n e^{-\frac{1}{n+1}} e^{\frac{n-1}{2(n^2-1)}} = e^{-\frac{1}{2(n+1)}} \text{volume}(\mathcal{E}_0 = B_2).$$

using $1 + x \leq e^x \ \forall x \in \mathbb{R}$.

Note that $\frac{\text{volume}(\mathcal{E}_1)}{\text{volume}(\mathcal{E}_0)}$ is invariant under rotations, (since all volumes are) but also under stretching of any axes as we have seen.

Thus $\frac{\text{volume}(\mathcal{E}_1)}{\text{volume}(\mathcal{E}_0)}$ is invariant under any invertible coordinate transformation, since any invertible linear map $A = U \Sigma V^T$ for orthogonal U, V

and diagonal Σ with positive entries by singular value decomposition. Thus our volume reduction result holds starting with any ellipse and half-space.

Ex Verify that $\mathcal{E}_1 \supseteq \mathcal{E}_0 \cap \{x_1 \geq 0\}$.

Finally, although unnecessary for our proof, note we have also constructed the smallest ellipse subject to the containment requirement; our upper bounds on its volume might have been a bit loose though.

For computationally explicit implementation, we extend this construction to the general case, i.e.,

$\mathcal{E}_0 = \{x : \|x - x_0\|^2_{H_0^{-1}} \leq 1\}$, we'll construct $\mathcal{E}_1 \supseteq \mathcal{E}_0 \cap \{x : w^T(x - x_0) \geq 0\}$

$$X \underset{y = H_0^{-1/2}(x - x_0)}{\overset{x = H_0^{1/2} y + x_0}{\longleftrightarrow}} Y$$

In y-space, $\mathcal{E}_0 = \{y : \|y\|_2^2 \leq 1\}$.

$$w^T(x - x_0) \geq 0 \iff (H_0^{1/2} w)^T y \geq 0 \iff \frac{(H_0^{1/2} w)^T y}{\|w\|_{H_0}} \geq 0$$

To make it a unit vector.

$$\mathcal{E}_1 = \left\{ y : \left\| y - \frac{1}{n+1} \frac{H_0^{1/2} w}{\|w\|_{H_0}} \right\|^2_{\left(\frac{n+1}{n}\right)^2 \frac{H_0^{1/2} w w^T H_0^{1/2}}{\|w\|^2_{H_0}} + \frac{n^2-1}{n^2}\left(I - \frac{H_0^{1/2} w w^T H_0}{\|w\|^2_{H_0}}\right)} \right\}$$

$$= \{x : \|x - x_1\|^2_{H_1^{-1}} \leq 1\}, \quad \text{where}$$

$$x_1 = x_0 + \frac{1}{n+1} \frac{H_0 w}{\|w\|_{H_0}}, \quad H_1 = \left(\left(\frac{n+1}{n}\right)^2 \frac{w w^T}{\|w\|^2_{H_0}} + \frac{n^2-1}{n^2}\left(H_0^{-1} - \frac{w w^T}{\|w\|^2_{H_0}}\right) \right)^{-1}.$$

$\square$

References:

1. Ellipsoid algorithm
    1. Section 2.2 in Bubeck
    2. Sections 3.2 and 3.3 in Lee-Vempala
2. (Beyond this course) Applying ellipsoid to large LPs— Chapter 3+ in GLS

LECTURE 9: REGRET

STORY SO FAR...

	ALGEBRAIC	GEOMETRIC	LOGICAL
HIGH ACCURACY $\log \frac{1}{\varepsilon}$	Simplex Vertex Enumeration ? Interior-point	Ellipsoid Oracle-based COM	Fourier-Motzkin
MODERATE ACCURACY $\text{poly} \frac{1}{\varepsilon}$	? Gradient Descent	Sampling-based COM	Multiplicative Weights

We will fill this quadrants now.

EXPERTS SETTING

$t = 1, \dots, T$

RAIN ↖ ↗ SUN

'N' experts make recommendations $\{\pm 1\}$ to the learner.
Learner chooses an expert to follow, say $i_t \in [N]$.
Adversary choose losses $\{0, 1\}$ for each recommendation.
(More loss is bad.)

Repeat

ASSUMPTION FOR NOW. $\exists$ an expert who is perfect, on all days incurs $\emptyset$ loss; learner doesn't know which one.

Q. What strategy should the learner follow to minimize her cumulative number of mistakes?

NAIVE STRATEGY: Follow friend i till they make a mistake. If/when they do, start following friend $i+1$.

Upon each of the learner's mistakes, one expert is eliminated. Hence, # mistakes for learner (or cumulative loss) $\leq N-1$ in the worst-case. But we can do much better.

SURVIVING MAJORITY: Each day take a majority vote among all surviving experts. At the conclusion of the day, eliminate those who made mistakes.

Every time the learner makes a mistake, $1/2$ of the expert pool is eliminated. Can only happen so many times, Hence, # mistakes $\leq \log_2 N$. This is an exponential improvement! Fantastic! But we would like to get rid of the realizability/perfect expert assumption. Natural generalization is to initially assign each expert some credibility that goes down, but doesn't become zero like before, when the experts make mistakes. This works to an extent, but comes up short against the following barrier.

Ex Try to produce an upper bound on number of mistakes by halving the credibility of a wrong expert in each round & taking weighted majority.

CLAIM: For any strategy* for the learner, there exists a worst-case assignment of losses guaranteeing the learner makes AT LEAST twice the number of mistakes for the best expert.

PROOF. Take 2 experts - A predicts +1 every day, B predicts -1. The adversary assigns +1 loss to whichever expert you as the learner pick, and 0 to the other. Thereby, on each day you make a mistake, i.e. after T days, T cumulative mistakes. However on each day exactly one expert makes a mistake. Hence, because minimum $\leq$ average, $\exists$ an expert that at the end of T days has made at most $T/2$ mistakes. $\square$

Let m^* = minimum number of mistakes for any expert. Thus, $2m^*$ seems like a natural barrier. However, the above lower bound construction crucially depends on the learner's strategy being deterministic. For a randomized strategy (where the adversary can't inspect the learner's random coins but everything else), one can plausibly randomized ... do better, and indeed we can.

MULTIPLICATIVE WEIGHTS / (REALLY) HEDGE ALGORITHM

Set $w_i^1 = 1 \quad \forall i \in [N]$

For $t = 1 \ldots T$

Play $i_t \sim P_t$ where $P_t^i = \dfrac{w_i^t}{\sum_{i \in [N]} w_i^t}$.

Adversary chooses loss vector $l_t \in [-1, +1]^N$, that can depend on past losses, weights – past & current, all actions $i_1 \ldots i_{t-1}$, but not i_t.

((Equivalently, it can depend on i_t as long as the learner's payoff $\triangleq \mathbb{E}_{i \sim P_t} l_t^i = P_t^T l_t$.))

Update $w_i^{t+1} = w_i^t e^{-\eta l_t^i}$.

THEOREM.
$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T} l_t^{i_t}\right]}_{\substack{\text{LEARNER'S} \\ \text{cumulative loss}}} - \underbrace{\min_{i \in [N]} \sum_{t=1}^{T} l_t^i}_{\substack{\text{Best expert's} \\ \text{loss in hindsight}}} = \sum_{t=1}^{T} P_t^T l_t - \min_{P \in \Delta_{N-1}} \sum_{t=1}^{T} P^T l_t \le \sqrt{T \log N}$$
for $\eta = \sqrt{\dfrac{\log N}{T}}$.

Also called REGRET

Let us explore the implication before diving into a proof. Dividing by T, we get

$$\text{LEARNER'S AVERAGE LOSS} \le \text{AVERAGE LOSS OF THE BEST EXPERT IN HINDSIGHT} + \underbrace{\sqrt{\frac{\log N}{T}}}_{\text{EXCESS AVERAGE LOSS}}.$$

Comments:

1. This guarantee holds for arbitrarily, or even adversarially, chosen loss assignment / vectors. No distributional assumptions were made unlike stats / ML / stochastics.
2. There's no $2\times$ multiplier associated with the best expert. This breaks our lower bound.
3. Excess average loss $\to 0$ as $T \to \infty$. In particular, if $T \geq \log N / \varepsilon^2$, excess average loss $\le \varepsilon$.

4. It's a relative error guarantee; no one can ensure low absolute error even for random loss functions. A relative (additive) error metric is something experts in other fields outside ML/statistical learning find hard to swallow (although situation is very rapidly changing), but it has proved to be one of the most far-reaching design choices in ML Theory.
5. The cost for having many inaccurate experts, as long as there is one good one, is small because of the $\log N$ dependence. Exponential N still yields reasonable bounds.
6. The nature of the bound in #3 is not an accident. It closely resembles uniform convergence results from statistical learning, precisely because online learning is an algorithmic theory as opposed to an analytic theory that generalizes the former.

PROOF OF THEOREM. Our basic proof strategy is to construct a potential function that decreases when a learner makes mistakes. Taking inspiration from MAJORITY / HALVING, we take $\Phi_1 = \sum_{i\in[n]} w_i^1 = N$. Now,

$$\Phi_{t+1} = \sum_{i\in[n]} w_i^{t+1} = \sum_{i\in[n]} w_i^t e^{-\eta l_t^i} = \Phi_t \sum_{i\in[n]} \frac{w_i^t}{\sum_{i\in[n]} w_i^t} e^{-\eta l_t^i}$$

$$\le \Phi_t \sum_{i\in[n]} p_t^i \left(1-\eta l_t^i + \eta^2 (l_t^i)^2\right)$$

$$\le \Phi_t \left(1-\eta p_t^T l_t + \eta^2\right) \le \Phi_t e^{-\eta p_t^T l_t + \eta^2}$$

using $e^x \le 1+x+x^2$ $\forall x \le 1$, and $1+t \le e^x$ $\forall x\in\mathbb{R}$ in successive steps of the derivation.

We are almost done. Let $i^* \in \arg\min_{i \in [n]} \sum_{t=1}^{T} l_t^i$ be the best expert in hindsight. Then using the above:

$$e^{-\eta \sum_{t=1}^{T} l_t^{i^*}} \le \Phi_{T+1} \le \Phi_1 e^{-\eta \sum_{t=1}^{T} p_t^T l_t + \eta^2 T}$$

Taking log on both sides, we get:

$$-\eta \sum_{t=1}^{T} l_t^{i^*} \le \log N - \eta \sum_{t=1}^{T} p_t^T l_t + \eta^2 T.$$

$$\Rightarrow \sum_{t=1}^{T} p_t^T l_t - \sum_{t=1}^{T} l_t^{i^*} \le \frac{\log N}{\eta} + \eta T \le 2\sqrt{T \log N}. \qquad \square$$

APPLICATIONS OF MULT WEIGHTS

EXAMPLE ONE: SOLVING LPs

We'll solve the feasibility problem: $\exists ?\, x \in K \quad Ax \le b$, where K is a "simple" convex set. Why restrict to simple sets? Because we'll use a subroutine/oracle that answers $\exists ?\, x \in K \quad c^T x \le d$. Note that this only has one inequality constraint, instead of m.

Example 1: $K = B_2 = \{\|x\|_2 \le 1\}$. $\exists ?\, x \in K,\ c^T x \le d$ is YES iff (a) $d \ge 0$ (or $0^T c \le d$), OR (b) $\mathrm{dist}(0, K) = \frac{|d|}{\|c\|_2} \le 1$.

Example 2: $K = \{x \ge 0\}$. $\exists ?\, x \in K,\ c^T x \le d$ is YES iff (a) $d \ge 0$ (or $0^T c \le d$), OR (b) $\exists i,\ c_i < 0$.

Ex Prove the correctness of the procedures above for $K = B_2$ and $K = \{x \ge 0\}$. Also, describe a procedure to efficiently solve $\exists ?\, x \in \{\|x\|_\infty \le 1\}\ c^T x \le d$.

We describe the algorithm next. Think of it as a game where the learner tries to prove the LP is infeasible assisted by the constraints as experts, by asking gotcha questions. The ORACLE assuages the learner's concerns.

v

ALGORITHM.

1. Each constraint is an expert, with $w_i^1 = 1 \ \forall i \in [m]$.
2. For $t = 1 \dots T$

   Learner chooses $p_t^i = \dfrac{w_i^t}{\sum_{i \in [m]} w_i^t}$.

   Asks the oracle $\exists ? x \in K, \quad p_t^T A x \le p_t^T b$.

   If NO, Output that the original LP is infeasible.

   If YES, ORACLE returns $x_t \in K$ such $p_t^T A x_t \le p_t^T b$.

   Each constraint i receives loss $\frac{1}{\rho}(b_i - a_i^T x_t)$. $\in [-1, +1]$

   $w_i^{t+1} = w_i^t \, e^{-\eta (b_i - a_i^T x_t)/\rho}$ update.

Here, $\rho = \max_{i \in [m], x \in K} |b_i - a_i^T x| =$ WIDTH of $Ax \le b$ against K.

CLAIM. For any $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, this algorithm either outputs:

(A) that $\exists ? x \in K$, $Ax \le b$ is infeasible, correctly, or

(B) a point $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \in K$ such $Ax \le b + \rho \sqrt{\frac{\log m}{T}} \mathbb{1}$.

In words, either the algorithm correctly decodes that the LP is infeasible or outputs an ε-feasible solution, when run for long enough, i.e., when $T = \dfrac{\rho^2 \log m}{\varepsilon^2}$.

PROOF. Clearly, if $Ax \le b$, $x \in K$ is feasible, then $\forall p \ge 0$, we have that $p^T A x \le p^T b$, $x \in K$ is feasible. Thus, we only need to prove part B assuming ORACLE says YES on all rounds. If so, by the regret guarantee:

$$\sum_{t=1}^{T} p_t^T (b - Ax_t) \le \min_{i \in [m]} \sum_{t=1}^{T} (b_i - a_i^T x_t) + \rho \sqrt{T \log m}.$$

But $\forall t$, $p_t^T b \ge p_t^T A x_t$. Hence

$$0 \le \frac{1}{T} \sum_{t=1}^{T} p_t (b - Ax_t) \le \min_{i \in [m]} \left(b_i - a_i^T \left(\sum_{t=1}^{T} x_t / T \right) \right) + \rho \sqrt{\frac{\log m}{T}}.$$

Rearranging, $\forall i \in [m]$, $a_i^T \bar{x} \le b_i + \rho \sqrt{\dfrac{\log m}{T}}$. $\square$

EXAMPLE 2: CONSTRUCTIVE MINIMAX THEOREM

Recall that $\min_{x \in \Delta} \max_{y \in \Delta} x^T A y = \max_{y \in \Delta} \min_{x \in \Delta} x^T A y$.

We arrived it using strong duality, in fact it is equivalent to strong duality. We will give a constructive algorithmically efficient proof of this statement. In fact, the previous algorithm can be seen as an efficient algorithmic Farkas' Lemma.

Assume $\max_{i,j} |a_{ij}| \leq 1$. Else, we scale.

ALGORITHM

1. Row player thinks of each row as an expert.
2. $t = 1 \ldots T$

   Row player plays $x_t \in \Delta$ as per Mult Weight.

   Column player plays $y_t \in \arg\max_{y \in \Delta} x_t^T A y$.

   Row i's loss is $e_i^T A y_t$.

CLAIM. If $T \geq \frac{\log m}{\varepsilon^2}$, then $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ satisfies

$$\max_{y \in \Delta} \bar{x}^T A y \leq \max_{y \in \Delta} \min_{x \in \Delta} x^T A y + \varepsilon.$$

Note that this is the non-trivial direction. By weak duality or definition of min/max, $\min_{x \in \Delta} \max_{y \in \Delta} x^T A y \geq \max_{y \in \Delta} \min_{x \in \Delta} x^T A y$

By compactness + continuity, we get $\exists x'$, $\max_{y \in \Delta} x'^T A y \leq \max_{y \in \Delta} \min_{x \in \Delta} x^T A y$

PROOF. $\max_{y \in \Delta} \bar{x}^T A y = \max_{y \in \Delta} \frac{1}{T} \sum_{t=1}^{T} x_t^T A y \leq \frac{1}{T} \sum_{t=1}^{T} \max_{y \in \Delta} x_t^T A y$

$$= \frac{1}{T} \sum_{t=1}^{T} x_t^T A y_t \leq \min_{x \in \Delta} \frac{1}{T} \sum_{t=1}^{T} x^T A y_t + \sqrt{\frac{\log m}{T}}$$

using regret guarantee

$$= \min_{x \in \Delta} x^T A \left(\frac{1}{T} \sum_{t=1}^{T} y_t \right) + \sqrt{\frac{\log m}{T}}$$

$$= \max_{y \in \Delta} \min_{x \in \Delta} x^T A y + \sqrt{\frac{\log m}{T}}. \qquad \square$$

So, low regret $\Longrightarrow$ minimax theorem. But can we go back? David Blackwell in 1956 proved a close equivalence between existence of low-regret strategies & a certain generalization of the minimax theorem.

References:

1. The best reference for regret minimization & applications to LPs/minimax duality is Elad's book— specifically chapters 1 & 8.
2. See this survey from Sanjeev, Elad and Satyen for applications of the multiplicative weights algorithm.
3. See this fantastic paper by Yoav Freund and Robert Schapire, who pioneered the Godel prize-winning boosting approach to machine learning using the regret-minimax link.
4. This NYTimes article quoting Rakesh Vohra chronicling the (independent) rediscovery of multiplicative weights in many academic fields; I think of this as convergent evolution. *In 1957, for example, a statistician named James Hanna called his theorem Bayesian Regret. He had been preceded by David Blackwell, also a statistician, who called his theorem Controlled Random Walks. Other, later papers had titles like "On Pseudo Games," "How to Play an Unknown Game," "Universal Coding" and "Universal Portfolios," Dr. Vohra said, adding, "It's not obvious how you do a literature search for this result."*

Assignment #1

RELEASED 11:59 pm Sep 6 Wed

DUE 11:59 pm Sep 13 Wed

ELECTRONICALLY (LaTex / scan / hand-written)

47834 LINEAR PROGRAMMING

Q1. PART A - 2 points.

Prove that $f : \mathcal{X} \to \mathbb{R}$, $f(x) = \dfrac{c^T x + p}{d^T x + q}$ is quasi-convex

on $\mathcal{X} = \{x : d^T x + q \geq 0\}$.

PART B - 8 points

Consider the following optimization problem.

$$\left.\begin{array}{ll} \max\limits_{x \in \mathcal{X}} & \dfrac{c^T x + p}{d^T x + q} \\ \text{s.t} & Ax \leq b \end{array}\right\} (O)$$

Propose a linear programming reformulation of this optimization problem. Also, describe how would one reconstruct a solution to (O) given an optimal solution of your proposed LP.

Comment: Upto 5 points, if you do not have a formulation, but propose an algorithm that solves (O) by solving multiple LPs.

Q2. PART A - 2 points

How far in Euclidean distance is a point $x' \in \mathbb{R}^n$ from the hyperplane $\mathcal{H} = \{x : a^T x = b\}$?

PART B - 8 points

Consider the set $P = \{x : Ax \leq b\}$; assume it's compact and non-empty. Provide a linear program to compute the center & the radius of the largest sphere contained (entirely) inside P.

Q3. PART A – 6 points.

In 2-dimensions, consider

$$A = \{ x \in \mathbb{R}^2 : \max\{|x_1|, |x_2|\} \leq 1 \}$$

$$B = \{ x \in \mathbb{R}^2 : |x_1| + |x_2| \leq 1 \}$$

$$C(\varepsilon) = \{ x \in \mathbb{R}^2 ; \; x_1^2 + x_2^2 \leq \varepsilon^2 \}$$

Sketch $A+B$ and $A+C(1)$; '+' is the Minkowski Sum.

Compute $\lim_{\varepsilon \to 0^+} \dfrac{\text{Area}(A + C(\varepsilon)) - \text{Area}(A)}{\varepsilon}$.

PART B – 4 points.

Now, consider 2 axis-aligned hyper-cuboids (i.e., 2 axis aligned 'n' dimensional rectangles) A and B, possibly of unequal sizes.

Prove $\text{vol}(A+B)^{1/n} \geq \text{vol}(A)^{1/n} + \text{vol}(B)^{1/n}$.

Assignment #2

RELEASED Sep 20 11:59 pm

DUE Sep 27 11:59 pm

47834 LINEAR PROGRAMMING

Q1. (10 points)

Minkowski-Weyl says that any bounded polyhedron can be written in two ways: either via inequalities defining it, or as convex hull of some set of points. But how do we know if ultimately we are talking about the same set expressed differently.

Concretely consider:

$$A = \text{CONVEX HULL}(\{x_1 \cdots x_m\}) \qquad B = \text{CONVEX HULL}(\{y_1 \cdots y_m\})$$

$$C = \{x : Ax \le b\} \qquad D = \{x : Cx \le d\}.$$

Assume you can solve any LP with 'n' variables and 'm' constraints in $\text{poly}(m,n) = (m+n)^{10}$ time. Here $A, C \in \mathbb{R}^{m \times n}$, $x_i, y_i \in \mathbb{R}^n$. Give polynomial time (e.g. $(m+n)^{100}$ time) algorithms to answer as many of these as possible:

(1) Is $A \subseteq B$?

(2) Is $A \subseteq C$?

(3) Is $C \subseteq D$?

(4) Is $C \subseteq A$?

Hint: Three of these are solvable in poly-time.

Q2. (10 points)

Recall that
1. conic hull requires $\lambda_i \ge 0 \; \forall i$.
2. affine hull requires $\sum \lambda_i = 1$.
3. convex hull requires both $\lambda_i \ge 0 \; \forall i$, $\sum \lambda_i = 1$

This suggests that for any set S,

$$\text{CONVEX HULL}(S) = \text{CONIC HULL}(S) \cap \text{AFFINE HULL}(S). \quad (*)$$

PART A: Prove that this is false. For example, construct a set S for which (*) is false.

PART B: What minimal conditions must $\underset{\text{HULL}}{\text{AFFINE}}(S)$ satisfy so that (*) is true $\forall S$? Prove that (*) indeed holds under your proposed condition

Q3. (10 points)

Let $2^{[n]}$ be the set of all subsets of $[n] = \{1, 2 \dots n\}$. Consider a function $f : 2^{[n]} \to \mathbb{R}_+$ such that

(1) $f(\emptyset) = 0$.

(2) $f(S) \leq f(T) \quad \forall S \subseteq T \subseteq [n]$.

(3) $f(S) + f(T) \geq f(S \cap T) + f(S \cup T) \quad \forall S, T \subseteq [n]$

Now, consider the following LP with exponentially many constraints, and some vector $c \in \mathbb{R}^n$.

$$\max \quad c^T x$$
$$\text{s.t.} \quad \sum_{j \in S} x_j \leq f(S) \quad \forall S \subseteq [n].$$
$$x \geq 0$$

Give a polynomial (in n) time algorithm to solve this LP; the algorithm must evaluate the function f polynomial (in n) times.

Prove your algorithm is correct by constructing a dual feasible solution that obtains the same objective value as the output of your algorithm.

Note: 3 points for listing the dual LP.

Assignment #3

RELEASED Oct 02 11:59 pm

DUE Oct 09 11:59 pm

47834

LINEAR PROGRAMMING

Q1. (10 points)

Let x be a BFS with basis B for $\min_{Ax=b,\, x\geq 0} c^T x$.

(A) Prove that if the reduced cost of every variable not in B is positive, then x is the UNIQUE minimum.

(B) If x is the UNIQUE minimum & the LP is non-degenerate, then the reduced cost of any variable not in B is positive.

Q2. (10 points)

Assume $\min_{x}\; {}_{Ax=b,\, x\geq 0}\; c^T x$ is non-degenerate.

Consider $f(\lambda) = \min_{x} (c+\lambda d)^T x$, for $\lambda \geq 0$.
$\qquad Ax=b, x\geq 0$

Say x^* with basis B is optimal at $\lambda = 0$.

(A) Prove the set of λ's for which x^* is an optimum is $[0, \lambda_1]$ for some $\lambda_1 \geq 0$. In fact, give as efficient of an algorithm as possible to compute the largest such λ_1.

(B) Prove that $\exists\, 0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_k \leq \lambda_{k+1} = +\infty$ where $k \geq 0$ and bases $B_0, \dots B_k$ such that some x_i with base B_i is optimal iff $\lambda \in [\lambda_i, \lambda_{i+1}]$.

Q3. (10 points)

(A) Prove $f(b) = \min_{x}\; {}_{Ax=b,\, x\geq 0}\; c^T x$ is convex.

(B) Prove $g(c) = \min_{x}\; {}_{Ax=b,\, x\geq 0}\; c^T x$ is concave.

(C) Rewrite the set $C(x^*) = \{c : c^T x^* \geq \max_{y \in P} c^T y\}$ where $P = \{x : Ax \leq b\}$ as a polyhedron with polynomially many constraints.