

CS730: NONLINEAR OPTIMIZATION II: SPRING 2011

STEPHEN WRIGHT*

Abstract. This document is a record of CS730, Nonlinear Optimization II, as taught in Spring 2011.

Topics. Taken from curriculum and posted on the class web site.

- Geometric viewpoint of constrained optimization
 - Convex sets, cones, projections
 - Tangent and normal to polyhedral sets
 - Theorems of the alternative, separation results
 - First-order conditions: polyhedral case
- Optimality conditions for nonlinear programming
 - Constraint qualifications
 - First-order conditions and saddle points
 - Second-order conditions and critical cones
 - Degeneracy
- Duality for nonlinear programming.
- Nonlinear programming algorithms
 - Fundamentals: merit functions and filters, Maratos effect.
 - Interior-point and augmented Lagrangian methods
 - Sequential quadratic programming
- Semidefinite programming (SDP): applications, barrier methods.
- Second-order cone programming: applications, barrier methods.

Emphasize algorithms more (this is the focus of NW).

Possible other topics: modeling of practical SDP and SOCP problems (e.g. robust optimization, QCQP), use of SeDuMi to solve them.

*Computer Sciences Department, University of Wisconsin-Madison

Lecture 1. (1/19/11; 60 min)

Web page at

<http://www.cs.wisc.edu/~swright/cs730.html>

Class mailing list, mailing list archive linked to web page.

We will use learn@uw to post grades on homeworks. Check frequently to see that your grade was recorded properly. Log in now to check that you are enrolled. Other class material will be posted on the course web page.

Instructor: Steve Wright, 4379 CS.

TA: Probably none.

Email instructor and TA together at cs730-1@cs.wisc.edu

Office hours posted on web page. Not quite final yet.

Lectures: Hold open three 75-minute slots per week, MWF 11-12:15. Usually 3 60-minute lectures. However schedule may be varied according to my travels. Schedule posted on the web page at least a week in advance.

Assessment: TBD. Probably 4-5 graded assignments, midterm, final. Driven by the fact that we don't have at TA, apparently. Another 4-5 non-graded assignments.

Go over curriculum, including additions above.

Go over texts and references.

We work in \mathbf{R}^n . Recall vector notation (subscripts for components), inner products, Euclidean norms, sequences, subsequences, limits, accumulation points.

Open and closed sets.

Compact sets: all sequences in S have a limit in S (Bolzano-Weierstrass), any cover of S (i.e. collection of open sets whose union includes S) has a finite subcover - this is the Heine-Borel theorem. Compact \equiv closed and bounded.

Convex Sets, Cones, and Functions: Define.

Cones: sets C such that $\lambda x \in C$ for all $\lambda > 0$ and $x \in C$. Are not necessarily closed or convex. Give example of non-closed and non-convex cone. "Salient" cone has no vectors $x \neq 0$ such that x and $-x$ are both in C . "Pointed" cone is one that contains 0. These definitions vary. We're particularly interested in closed convex cones.

Convex cone defined by $\alpha x + \beta y \in C$ for all $x, y \in C$ and all $\alpha > 0$ and $\beta > 0$.

Polar cone: $C^\circ := \{x \mid \langle x, v \rangle \leq 0 \text{ for all } v \in C\}$. Dual cone is negative of polar cone.

Lecture 2. (1/21/10; 60 min)

Normal cone to closed, convex set Ω at $x \in \Omega$:

$$N_{\Omega}(x) := \{v \mid \langle v, y - x \rangle \leq 0 \text{ for all } y \in \Omega\}.$$

Draw pictures.

Show that $N_{\Omega}(x)$ is outer semicontinuous as a function of x , i.e. given sequence $\{x_k\} \subset X$ with $x_k \rightarrow x$ and $v_k \in N_{\Omega}(x_k)$ with $v_k \rightarrow v$, we have $v \in N_{\Omega}(x)$.

Tangent cone. Define limiting directions. Given $x \in \Omega$, y is a limiting direction if there exist sequence $y_k \rightarrow y$, $\alpha_k \rightarrow 0$ such that $x + \alpha_k y_k \in \Omega$ for all k . Tangent cone $T_{\Omega}(x)$ is set of all limiting directions to Ω at x .

For closed convex Ω , normal and tangent are polar to each other. Prove after defining projection operator P below!

Projection P onto closed convex set:

$$P(y) = \arg \min_{z \in \Omega} \|z - y\|.$$

Recap and illustrate results proved in 726: (i) $(P(y) - z)^T(y - z) \geq 0$ for all $z \in \Omega$, with equality if and only if $z = P(y)$; (ii) $(y - P(y))^T(z - P(y)) \leq 0$ for all $z \in \Omega$.

(ii) implies that $y - P(y) \in N_{\Omega}(P(y))$. It also implies that the projection is uniquely defined. (Prove in class by showing that if there are two possible values $P(y)$ and $\bar{P}(y)$ for the projection of y , they must be the same.)

Also use (ii) to prove that projection operator is a contraction. (Proved in Bertsekas.) We have

$$(x_2 - P(x_2))^T(P(x_1) - P(x_2)) \leq 0, \quad (x_1 - P(x_1))^T(P(x_2) - P(x_1)) \leq 0.$$

Adding, we obtain

$$[x_2 - P(x_2) - x_1 + P(x_1)]^T[P(x_1) - P(x_2)] \leq 0,$$

which after rearrangement gives

$$\|P(x_1) - P(x_2)\|_2^2 \leq [x_1 - x_2]^T[P(x_1) - P(x_2)] \leq \|x_1 - x_2\| \|P(x_1) - P(x_2)\|,$$

proving the claim.

Proof that tangent and normal are polar. ($T_{\Omega}(x) \subset N_{\Omega}(x)^{\circ}$) Suppose $y \in T_{\Omega}(x)$ and let α_k and y_k be sequences with the stated properties. Let v be any vector in $N_{\Omega}(x)$. Then since $x + \alpha_k y_k \in \Omega$ we have $\langle v, \alpha_k y_k \rangle \leq 0$ and so $\langle v, y_k \rangle \leq 0$, so taking limits we have $\langle v, y \rangle \leq 0$. Hence y is in the polar of $N_{\Omega}(x)$.

($N_{\Omega}(x)^{\circ} \subset T_{\Omega}(x)$) Suppose y is in the polar of $N_{\Omega}(x)$. Consider the sequences $\alpha_k = 1/k$ and $y_k = k(P(x + y/k) - x)$. Clearly $x + \alpha_k y_k \in \Omega$ and $\alpha_k \rightarrow 0$. Since y_k is bounded (by contraction property of P) we have that there is a vector u and subsequence \mathcal{S} such that $\lim_{k \in \mathcal{S}} y_k = u$. Since

$$\langle (x + y/k) - P(x + y/k), x - P(x + y/k) \rangle \leq 0,$$

we have $\langle y/k - y_k/k, -y_k/k \rangle \leq 0$ and so $\langle y - y_k, y_k \rangle \geq 0$ for all k . Hence taking limits we have $\langle y - u, u \rangle \geq 0$. Now use outer semicontinuity of N_{Ω} . We have

$$k[(x + y/k) - P(x + y/k)] \in N_{\Omega}(P(x + y/k)),$$

so that

$$y - k[P(x + y/k) - x] \in N_{\Omega}(P(x + y/k)).$$

Taking limits of the subsequence \mathcal{S} we have $y - u \in N_{\Omega}(x)$. Since $y \in N_{\Omega}(x)^{\circ}$, we have $\langle y - u, y \rangle \leq 0$. Combining with the previous inequality we have

$$0 \geq \langle y - u, y \rangle - \langle y - u, u \rangle = \|y - u\|_2^2,$$

so that $u = y$. Get the result by taking the subsequence \mathcal{S} and renumbering.

Lecture 3. (1/24/11; 55 min)

Optimization concepts for the setting $\min_{x \in \Omega} f(x)$. Assume f is continuously differentiable.

Now prove our first first-order optimality conditions. First recall Taylor's theorem (NW, pp. 14-15). In particular we use this form, which requires f continuously differentiable:

$$f(y) = f(z) + \nabla f(z + t(y - z))^T (y - z)$$

for some $t \in (0, 1)$.

Define local solution, strict local solution, isolated local solution (NW, p. 306).

Show that local solution definition is equivalent to there being *no* sequence $\{z_k\} \subset \Omega$ with $z_k \rightarrow x^*$ and $f(z_k) < f(x^*)$. *Proof:* If x^* is not a local solution then for any $k > 0$ we have that there is a point $z_k \in \Omega$ with $\|z_k - x^*\| \leq 1/k$ such that $f(z_k) < f(x^*)$. Clearly $z_k \rightarrow x^*$ so the "sequence" definition is not satisfied. Conversely, if the "sequence" definition is not satisfied, the elements of the violating sequence enter any given neighborhood \mathcal{N} so will violate the original definition too. (Draw pictures.)

THEOREM 0.1. Consider $\min_{x \in \Omega} f(x)$, where f is continuously differentiable and Ω is closed and convex. If x^* is a local solution, then

$$-\nabla f(x^*) \in N_{\Omega}(x^*).$$

Proof. Suppose that $-\nabla f(x^*) \notin N_{\Omega}(x^*)$, so that by our earlier result, $-\nabla f(x^*) \notin T_{\Omega}(x^*)^{\circ}$. Thus there is a tangent direction $y \in T_{\Omega}(x^*)$ such that $\langle -\nabla f(x^*), y \rangle > 0$. By definition of tangent, we there are sequences $\{\alpha_k\}$ (with $\alpha_k > 0$ and $\alpha_k \rightarrow 0$) and $\{y_k\}$ with $y_k \rightarrow y$ such that $x^* + \alpha_k y_k \in \Omega$. By Taylor's theorem we have

$$f(x^* + \alpha_k y_k) = f(x^*) + \langle \nabla f(x^* + \gamma_k \alpha_k y_k), \alpha_k y_k \rangle,$$

for some $\gamma_k \in (0, 1)$ and all k . Thus by adding and subtracting terms we have

$$\begin{aligned} f(x^* + \alpha_k y_k) &= f(x^*) + \alpha_k \langle \nabla f(x^*), y \rangle + \alpha_k \langle \nabla f(x^*), y_k - y \rangle \\ &\quad + \alpha_k \langle \nabla f(x^* + \gamma_k \alpha_k y_k) - \nabla f(x^*), y_k \rangle \\ &= f(x^*) + \alpha_k \langle \nabla f(x^*), y \rangle + o(\alpha_k), \end{aligned}$$

so that $f(x^* + \alpha_k y_k) < f(x^*)$ for all k sufficiently large. Hence the sequence $z_k := x^* + \alpha_k y_k$ has the properties that show that x^* is not a local min. \square

These results link the geometric condition $-\nabla f(x^*) \in N_{\Omega}(x^*)$ to the KKT conditions described later, when constraint qualifications are satisfied.

Polyhedral set: Define as intersection of half-planes: $\{x \mid Dx \leq d\}$.

Draw some pictures of such sets and their normal cones. Make the point that normal cone at x depends only on constraints active at x .

Prove results from Robinson Lecture 6 about tangent and normal cones to a polyhedral convex set.

Let $C = \{x \mid Dx \leq d\}$ be a nonempty polyhedral convex set in \mathbb{R}^n . For a point $x_0 \in C$ define $\mathcal{A}(x_0) = \{i \mid D_i x_0 = d_i\}$ to be the set of active indices, and $\mathcal{I}(x_0) = \{i \mid D_i x_0 < d_i\}$ contains the inactive indices. Let $D_{\mathcal{A}}, d_{\mathcal{A}}$ contain the active rows of D and d , and let $D_{\mathcal{I}}, d_{\mathcal{I}}$ contain the inactive rows. Thus $D_{\mathcal{A}} x_0 = d_{\mathcal{A}}$ and $D_{\mathcal{I}} x_0 < d_{\mathcal{I}}$.

Lecture 4. (1/26/11; 55 min)

PROPOSITION 0.2. Let $K = \{z \mid D_{\mathcal{A}}z \leq 0\}$. Then $(C - x_0) \subset K$, and there is a neighborhood U of the origin such that $U \cap K = U \cap (C - x_0)$.

Proof. Let $c \in C$, then $D_{\mathcal{A}}c \leq d_{\mathcal{A}}$, so $D_{\mathcal{A}}(c - x_0) \leq 0$, so $c - x_0 \in K$, proving the first assertion. For the second assertion, note first that $U \cap (C - x_0) \subset U \cap K$ for every U . For the reverse inclusion, since $D_{\mathcal{I}}x_0 < d_{\mathcal{I}}$, we can choose U so that $D_{\mathcal{I}}(x_0 + z) < d_{\mathcal{I}}$ for every $z \in U$. Hence for $z \in U \cap K$ we have $D_{\mathcal{I}}(x_0 + z) < d_{\mathcal{I}}$ and $D_{\mathcal{A}}(x_0 + z) \leq d_{\mathcal{A}}$. Hence $x_0 + z \in C$ and so $z \in U \cap (C - x_0)$. \square

PROPOSITION 0.3.

$$N_C(x_0) = D_{\mathcal{A}}^T(R_+^{|\mathcal{A}|}) = \left\{ \sum_{i \in \mathcal{A}} \mu_i D_i^T \mid \mu_i \geq 0, \text{ for all } i \in \mathcal{A} \right\}.$$

Proof. (\supset) Let $u \in \mathbf{R}_+^{|\mathcal{A}|}$ and $y = D_{\mathcal{A}}^T u$. Choose any $c \in C$. Then from previous result we have $c - x_0 \in K$, so $D_{\mathcal{A}}(c - x_0) \leq 0$. Then

$$\langle y, c - x_0 \rangle = \langle D_{\mathcal{A}}^T u, c - x_0 \rangle = \langle u, D_{\mathcal{A}}(c - x_0) \rangle \leq 0.$$

Hence, $y \in N_C(x_0)$.

(\subset) If $y \in N_C(x_0)$ and $v \in K$, then for some small $\mu > 0$ we have by the result above that $\mu v \in C - x_0$. Hence $0 \geq \langle y, \mu v \rangle = \mu \langle y, v \rangle$. Hence, $\langle y, v \rangle \leq 0$ for all $v \in K$, and so $y \in K^\circ$. We now have that $K = [D_{\mathcal{A}}^T(\mathbf{R}_+^{|\mathcal{A}|})]^\circ$ by the following argument:

$$\begin{aligned} z \in [D_{\mathcal{A}}^T(\mathbf{R}_+^{|\mathcal{A}|})]^\circ &\Leftrightarrow \langle z, D_{\mathcal{A}}^T u \rangle \leq 0, \quad \forall u \geq 0 \\ &\Leftrightarrow \langle D_{\mathcal{A}} z, u \rangle \leq 0, \quad \forall u \geq 0 \\ &\Leftrightarrow D_{\mathcal{A}} z \leq 0 \Leftrightarrow z \in K. \end{aligned}$$

(This result is a special case of Farkas' Lemma.) The result follows from $y \in K^\circ = [D_{\mathcal{A}}^T(\mathbf{R}_+^{|\mathcal{A}|})]^\circ = D_{\mathcal{A}}^T(\mathbf{R}_+^{|\mathcal{A}|})$. \square

It follows that

$$T_C(x_0) = N_C(x_0)^\circ = K.$$

Make the link to the first-order conditions of Theorem 0.1. Obvious corollary of the results above: If C is polyhedral of the form $C = \{x \mid Dx \leq d\}$, with \mathcal{A} and \mathcal{I} defined as above, then

$$-\nabla f(x^*) \in N_C(x^*) = \left\{ \sum_{i \in \mathcal{A}} \mu_i D_i^T \mid \mu_i \geq 0, \text{ for all } i \in \mathcal{A} \right\}.$$

In other words, if D and d have m rows there exists $\mu \in \mathbf{R}^m$ with

$$-\nabla f(x^*) = D^T \mu, \quad 0 \leq \mu \perp Dx - d \leq 0. \quad (0.1)$$

These extend the KKT conditions for linear programming to the case of a nonconvex objective - but constraints still linear. We'll extend further to nonlinear constraints below.

Theorems of the Alternative. Give further insight into the relationships between linear equalities/inequalities and the cones that they generate. Provide keys to link between the geometry of sets and their algebraic descriptions.

Theorems of the alternative typically have two logical statements I and II, and an assertion that exactly one of I and II is true. Typically prove by showing that $I \Leftrightarrow \sim II$ or $II \Leftrightarrow \sim I$.

Can prove from first principles and using projection operator. We'll prove using tools of LP duality, in particular, strong duality. recap this.

Consider first standard form:

$$(P) \quad \min_x c^T x \text{ s.t. } Ax \geq b, x \geq 0.$$

$$(D) \quad \max_u b^T u \text{ s.t. } A^T u \leq c, u \geq 0.$$

Strong Duality: There are three possibilities:

- (a) P and D both have solutions, and their optimal objectives are equal.
- (b) One is unbounded and the other is infeasible.
- (c) Both are feasible.

Lecture 5. (1/28/11; 60 min)

Same applies for the primal-dual pair:

$$(P) \quad \min_x c^T x \text{ s.t. } Ax \geq b.$$

$$(D) \quad \max_u b^T u \text{ s.t. } A^T u = c, u \geq 0.$$

Same applies for more general statements of LP. e.g. this primal:

$$\begin{aligned} \min_{x,y} p^T x + q^T y \\ \text{s.t. } Bx + Cy \geq d, \\ Ex + Fy = g, \\ Hx + Jy \leq k, \\ x \geq 0, \end{aligned}$$

has this dual:

$$\begin{aligned} \max_{u,v,w} d^T u + g^T v + k^T w \\ \text{s.t. } B^T u + E^T v + H^T w \leq p, \\ C^T u + F^T v + J^T w = q, \\ u \geq 0, w \leq 0. \end{aligned}$$

Farkas' Lemma is key to optimality conditions for nonlinear programming. Given matrix $A \in \mathbf{R}^{m \times n}$ and vector $b \in \mathbf{R}^m$, it says that either (I) b is a nonnegative linear combination of the rows of A (that is $A^T \lambda = b$ for some $\lambda \geq 0$), or (II) there is a plane through the origin that strictly separates the cone generated by the rows of A from b , that is, there is an x such that $Ax \leq 0$ and $b^T x > 0$, but not both. (Illustrate in 2d and 3d.)

Prove by defining an LP

$$(P) \quad \min -b^T x \text{ s.t. } Ax \leq 0,$$

whose dual is

$$(D) \quad \max 0^T u \text{ s.t. } A^T u = b, u \geq 0.$$

Apply strong duality. (I) is true \Rightarrow (D) has a solution, with optimal objective zero \Rightarrow (P) also has a solution with optimal objective zero (strong duality case (a)) \Rightarrow for every vector x with $Ax \leq 0$, we have $b^T x \leq 0$ iff (II) is not true. Conversely, (I) is false \Rightarrow (D) is infeasible \Rightarrow (P) is unbounded (since it is clearly feasible) \Rightarrow (II) is true.

Gordan's Theorem: Given A have either that (I) $Ax > 0$ has a solution, or (II) $A^T y = 0, y \geq 0, y \neq 0$ has a solution, but not both.

To prove define primal-dual pair:

$$(P) \quad \min_{(x,\alpha)} -\alpha \text{ s.t. } Ax - \alpha e \geq 0.$$

$$(D) \quad \max_{\lambda} 0^T \lambda \text{ s.t. } A^T \lambda = 0, e^T \lambda = 1, \lambda \geq 0,$$

where $e = (1, 1, \dots, 1)^T$ as usual.

$II \Leftrightarrow$ there exists λ feasible for (D) (scale if necessary) \Leftrightarrow (D) has a solution with optimal objective 0 \Leftrightarrow (P) has a solution with optimal objective 0 \Leftrightarrow there is no x with $Ax > 0$.

Tucker's theorem of the alternative has these two complementary statements.

$$(I) \quad Bx \geq 0, \quad Bx \neq 0, \quad Cx \geq 0, \quad Dx = 0 \text{ for some } x;$$

$$(II) \quad B^T y_2 + C^T y_3 + D^T y_4 = 0, \quad y_2 > 0, \quad y_3 \geq 0,$$

for some y_2, y_3, y_4 .

Define the LP pair:

$$(P): \quad \min -e^T Bx \text{ s.t. } Bx \geq 0, \quad Cx \geq 0, \quad Dx = 0.$$

$$(D): \quad \max 0 \text{ s.t. } -B^T e - B^T z_2 - C^T z_3 - D^T z_4 = 0, \quad z_2 \geq 0, \quad z_3 \geq 0.$$

(I) \Rightarrow (P) unbounded \Rightarrow (D) infeasible \Rightarrow there are no $z_2 \geq 0, z_3 \geq 0, z_4$ with $B^T(z_2 + e) + C^T z_3 + D^T z_4 = 0 \Rightarrow$ there are no $y_2 > 0, y_3 \geq 0, y_4$ with $B^T y_2 + C^T y_3 + D^T y_4$ (otherwise we could easily construct z_2, z_3, z_4 with the required properties from y_2, y_3, y_4) $\Leftrightarrow \sim II$.

$\sim (II) \Rightarrow$ there is no $y_2 \geq e$ with $B^T y_2 + C^T y_3 + D^T y_4 = 0$ for some $y_3 \geq 0$ and $y_4 \Rightarrow$ (D) is infeasible (setting $z_2 = y_2 - e, z_3 = y_3, z_4 = y_4$) \Rightarrow (P) is either infeasible or unbounded - but it is clearly not infeasible so must be unbounded \Rightarrow (I).

Carathéodory's Theorem. (e.g. p.43 of Mangasarian). Given a set $\Gamma \subset \mathbb{R}^n$, if x is a convex combination of points in Γ , then it is a convex combination of $n + 1$ or fewer points in Γ . (Daw a picture in \mathbb{R}^2 .)

Proof: Suppose $x = \sum_{i=1}^m \alpha_i x^i$ for $x^i \in \Gamma$ for all i and $\alpha_i \geq 0$ for all i and $\sum_{i=1}^m \alpha_i = 1$. Suppose that m is the minimal number such that x is expressible in this way, but for C! assume that $m > n + 1$.

Since the set $\{x^1 - x^m, x^2 - x^m, \dots, x^{m-1} - x^m\}$ has at least n elements, we can find coefficients $\rho_i, i = 1, \dots, m - 1$, not all zero, such that $\sum_{i=1}^{m-1} \rho_i (x^i - x^m) = 0$. Defining $\rho_m = -\sum_{i=1}^{m-1} \rho_i$, we have $\sum_{i=1}^m \rho_i = 0$. Now consider $\beta_i = \alpha_i + \gamma \rho_i, i = 1, 2, \dots, m$ for some γ , and note that

$$\sum_{i=1}^m \beta_i x^i = x, \quad \sum_{i=1}^m \beta_i = 1.$$

Since at least one of ρ_i is negative, we can choose a positive value of γ such that $\beta_i = 0$ for some i . Explicitly:

$$\gamma = \min_{i: \rho_i < 0} \frac{\alpha_i}{-\rho_i}.$$

For this γ , the β_i define a set of coefficients *fewer than m of which are positive*, such that $x = \sum_{i=1}^m \beta_i x^i, \beta_i \geq 0$, and $\sum_{i=1}^m \beta_i = 1$.

Lecture 6. (1/31/11; 60 min)

Separation Theorem. First discuss the following fundamental result about compact sets: Let Λ be compact and let Λ_x be a collection of subsets of Λ , closed in Λ , indexed by a set X (such that $x \in X$). Then if for any finite collection of points in X , call them $\{x_1, x_2, \dots, x_m\}$, we have $\bigcap_{i=1}^m \Lambda_{x_i}$ nonempty, then $\bigcap_{x \in X} \Lambda_x$ is also nonempty. Prove by using the property that the complement of each Λ_x is open in Λ . If $\bigcap_{x \in X} \Lambda_x = \emptyset$, then the complements form an open cover of Λ . So there is a finite cover, that is, a finite set of points $\{x_1, x_2, \dots, x_m\}$, such that the complements of Λ_{x_i} cover Λ , so that the intersection of the Λ_{x_i} is empty. C!

Separation Theorem: Let X be any convex set in \mathbb{R}^n not containing the origin. Then there is a vector $\bar{t} \in \mathbb{R}^n$ with $\bar{t} \neq 0$ such that $\bar{t}^T x \leq 0$ for all $x \in X$.

Proof: We use Λ_x to denote the subset of the unit ball $\|v\|_2 = 1$ such that $v^T x \leq 0$. Note that for each $x \in X$, Λ_x is closed and in fact compact. Now let $\{x_1, x_2, \dots, x_m\}$ be any finite set of vectors in X . Since $0 \notin X$, there cannot be a vector $p \in \mathbb{R}^m$ such that

$$0 = \sum_{i=1}^m p_i x_i = Xp, \quad p \geq 0, \quad e^T p = 1,$$

So there cannot be a vector p such that

$$0 = \sum_{i=1}^m p_i x_i = Xp, \quad p \geq 0, \quad p \neq 0.$$

Hence, by Gordan's theorem, there is a vector t such that

$$X^T t > 0$$

that is, $X^T(-t) < 0$. In other words, $-t/\|t\|_2 \in \bigcap_{i=1,2,\dots,m} \Lambda_{x_i}$, so this intersection is nonempty. This is true regardless of what finite collection of vectors we choose in X . Hence by the fundamental compactness result above, we have $\bigcap_{x \in X} \Lambda_x$ is also nonempty, so there is \bar{t} such that $\|\bar{t}\|_2 = 1$ and $\bar{t}^T x \leq 0$ for all $x \in X$.

An example where only $t^T \bar{x} \leq 0$ is possible (not strict inequality) is the set Ω in \mathbb{R}^2 consisting of the closed right-half plane but excluding the half-line $\{(0, x_2) \mid x_2 \leq 0\}$.

Theorem. Let X be nonempty, convex, and closed, with $0 \notin X$. Then there is $\bar{t} \in \mathbb{R}^n$ and $\alpha > 0$ such that $\bar{t}^T x \leq -\alpha$ for all $x \in X$.

Proof. Using $P(\cdot)$ to define projection onto closed convex X , we have by assumption that $P(0) \neq 0$. By the elementary property $(y - P(y))^T(z - P(y)) \leq 0$ for all $z \in X$, we have by setting $y = 0$ that $(0 - P(0))^T(z - P(0)) \leq 0$ for all $z \in X$, which implies $P(0)^T z \geq \|P(0)\|_2^2 > 0$. We obtain the result by taking $\bar{t} = -P(0)$ and $\alpha = \|P(0)\|_2^2$.

Now consider strict separation between two disjoint convex closed sets X and Y , and define

$$X - Y := \{x - y \mid x \in X, y \in Y\}.$$

First consider the question: Is $X - Y$ convex? closed? Clearly convex. But may not be closed. e.g. consider a sequence $\{z_k\} \subset X - Y$, with $z_k \rightarrow z$. We have $z_k = x_k - y_k$ for some $\{x_k\} \subset X$ and $\{y_k\} \subset Y$, but there is no guarantee that these two sequences even converge, so can't say for sure that $z \in X - Y$. Example:

$$X = \{(x_1, x_2) \mid x_1 > 0, x_2 \geq 1/x_1\}, \quad Y = \{(y_1, y_2) \mid y_1 > 0, y_2 \leq -1/y_1\}.$$

Now define $z_k = (0, 2/k)$ for $k = 1, 2, \dots$. We have $z_k \in X - Y$ (by taking $x_k = (k, 1/k)$ and $y_k = (k, -1/k)$). But $z_k \rightarrow (0, 0) \notin X - Y$.

(* Class cancelled for blizzard on 2/2/11 *)

Lecture 7. (2/4/11; 60 min)

However if we add the condition that X is compact, we have that $X - Y$ is closed.

Proof: Define z_k, x_k, y_k as above with $z_k \rightarrow z$. Then $x_k = z_k + y_k$ is in a compact set so we can take a subsequence approaching some $x \in X$. Thus $y_k = -z_k + x_k$ approaches $z - x$, and this limit must be in Y by closedness. Thus $z \in X - Y$.

So can prove the following: *Theorem.* Let X and Y be two disjoint closed convex nonempty sets with X compact. Then there is $c \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}$ such that $c^T x - \alpha < 0$ for all $x \in X$ and $c^T y - \alpha > 0$ for all $y \in Y$.

Proof. Consider $Z = X - Y$ and note that Z is closed as above, and $0 \notin X - Y$. By previous theorem, can find \bar{t} and $\beta > 0$ such that $\bar{t}^T(x - y) < -\beta$ for all $x \in X$ and $y \in Y$. Fix some $\bar{y} \in Y$ and note that $\bar{t}^T x < \bar{t}^T \bar{y} - \beta$. Thus $\bar{t}^T x$ is bounded above and so has a supremal value γ , while $\bar{t}^T y$ is bounded below and so has an infimal value δ , and $\gamma + \beta \leq \delta$. We have

$$\bar{t}^T x \leq \gamma < \gamma + \beta/2 < \gamma + \beta \leq \bar{t}^T y$$

for all $x \in X$ and $y \in Y$. Hence, the result is proved with $c = \bar{t}$ and $\alpha = \gamma + \beta/2$.

First-Order Conditions: Constraint Qualifications and the Nonconvex Case. Recall the results proved already. For closed convex C have first order conditions $-\nabla f(x^*) \in N_C(x^*)$, expressed in KKT form as ((0.1)). Now refer to (NW, chapter 12) for extension to the case in which constraints are expressed *algebraically* by possibly nonconvex functions.

First extend Proposition 0.3. Given this polyhedral set:

$$C = \{x \mid Dx \leq d, Gx = g\},$$

where D has m rows and G has p rows, show that

$$N_C(x_0) = \left\{ \sum_{i:D_i x=d_i} \mu_i D_i^T + \sum_{j=1}^p \gamma_j G_j^T \mid \mu_i \geq 0, \text{ for all } i \text{ with } D_i x_0 = d_i \right\}.$$

(Express C as $C = \{x \mid Dx \leq d, Gx \leq g, -Gx \leq -g\}$, apply the Proposition, and tweak.)

Special case of this extended result: Suppose that $d = 0$ and $g = 0$ in the definition of C , and let $x_0 = 0$. Then $C = T_C(0)$ (easy to show!) and so $C^\circ = N_C(0) = T_C(x)^\circ$ is defined by

$$C^\circ = \left\{ \sum_{i=1}^m \mu_i D_i^T + \sum_{j=1}^p \gamma_j G_j^T \mid \mu_i \geq 0, \text{ for all } i = 1, 2, \dots, m \right\}$$

Note that Def 12.2 (p. 316) of $T_\Omega(x^*)$ is equivalent to definition of tangent presented earlier. We're no longer assuming convex feasible set Ω however. Draw some pictures of tangent cones for nonconvex Ω . Note that T_Ω is a cone (according to our original definition of tangent) but no longer necessarily a *convex* cone. Nevertheless its polar is still defined, and is a closed convex cone. (See homework.) In fact for this more general definition of tangent, we *define* the normal cone $N_\Omega(x^*)$ as the polar of $T_\Omega(x^*)$ — see p.340. This coincides with our definition of the normal cone for the case of convex Ω , but may be different in the nonconvex case.

Define *linearized feasible direction set* $\mathcal{F}(x)$ as in Def 12.3. Note that this definition uses the algebraic description of the feasible set.

Lecture 8. (2/7/11; 70 min)

Now mention Theorem 12.3 on p.325. This says that x^* is a local solution implies that $-\nabla f(x^*) \in T_{\Omega}(x^*)^{\circ}$. Note that this result holds even if $T_{\Omega}(x^*)$ is not closed and convex, which is what we assumed when we proved the corresponding result in Lecture 3. The proof is the same here as earlier.

Draw some examples where $T_{\Omega}(x^*)$ and $\mathcal{F}(x^*)$ are the same. Now give examples where they differ e.g. pp. 319-320 and Figure 12.10.

Constraint qualifications are conditions under which $\mathcal{F}(x)$ and $T_{\Omega}(x)$ are similar. That is, linearization of the functions describing the set yields a tangent cone that is the same as the tangent cone to the actual set.

“Linear Constraint CQ”: all active c_i are linear functions. Clearly here, $T_{\Omega}(x^*)$ and $\mathcal{F}(x^*)$ are the same.

LICQ. Define and draw pictures. Note that it is neither implied by nor implies the “Linear Constraint CQ”.

State Lemma 12.2 on p. 323:

LEMMA 0.4. *For a feasible point x^* , we have*

(i) $T_{\Omega}(x^*) \subset \mathcal{F}(x^*)$.

(ii) *if LICQ is satisfied at x^* , then $T_{\Omega}(x^*) = \mathcal{F}(x^*)$.*

Stress that (i) holds independently of CQ! Hence to prove that a CQ yields equality of $T_{\Omega}(x^*)$ and $\mathcal{F}(x^*)$, we need to prove only $\mathcal{F}(x^*) \subset T_{\Omega}(x^*)$.

Prove $\mathcal{F}(x^*) \subset T_{\Omega}(x^*)$ when *all constraints are linear*.

Now prove (ii) from the lemma above. First discuss Implicit function theorem: p. 631. Now outline (not in complete detail) the proof of (ii) on pp. 324-325.

Lecture 9. (2/9/11; 60 min)

Recall optimality condition $-\nabla f(x^*) \in N_{\Omega}(x^*) := T_{\Omega}(x^*)^{\circ}$. Theorem 12.3 on p. 325.

If LICQ holds, we have $T_{\Omega}(x^*) = \mathcal{F}(x^*)$, so first-order optimality condition is

$$-\nabla f(x^*) \in \mathcal{F}(x^*)^{\circ}.$$

Now apply Proposition 0.3, extended as described above, to $\mathcal{F}(x^*)$ at the point $d = 0$. Recall that $\mathcal{A}(x^*)$ contains indices from \mathcal{E} as well as the active indices from \mathcal{I} . We have

$$\begin{aligned} -\nabla f(x^*) &\in \mathcal{F}(x^*)^{\circ} \\ &= \{d \mid \nabla c_i(x^*)^T d = 0 \ (i \in \mathcal{E}), \ \nabla c_i(x^*)^T d \geq 0 \ (i \in \mathcal{A}(x^*) \cap \mathcal{I})\}^{\circ} \\ &= \sum_{i \in \mathcal{A}(x^*) \cap \mathcal{I}} \lambda_i \nabla c_i(x^*) + \sum_{i \in \mathcal{E}} \lambda_i \nabla c_i(x^*), \quad \lambda_i \geq 0 \text{ for } i \in \mathcal{A}(x^*) \cap \mathcal{I}. \end{aligned}$$

This leads to KKT conditions - See Theorem 12.1 on p.321. Summarize by stating this theorem formally.

Define the Lagrangian \mathcal{L} and note how the KKT conditions can be stated using this function.

Review several examples with two constraints, where LICQ is satisfied, degenerate and nondegenerate, and where LICQ is not satisfied.

Lecture 10. (2/11/11; 60 min)

Uniqueness of multipliers λ^* when LICQ holds. Prove by contradiction.

MFCQ. See Definition 2.6 (p. 339). Show that it is weaker than LICQ.

Draw some pictures of where MFCQ holds but LICQ does not. Draw some pictures of problems where MFCQ doesn't hold. How to prove that MFCQ yields $T_\Omega(x^*) = \mathcal{F}(x^*)$? The hard part is to show that $\mathcal{F}(x^*) \subset T_\Omega(x^*)$.

Prove the classic result that the set of optimal multipliers is bounded when MFCQ holds. Do it just for the case of inequality constraints. For contradiction, assume there is an unbounded sequence of optimal multipliers $\lambda_{\mathcal{A}}^k$ with property that $\|\lambda_{\mathcal{A}}^k\| \rightarrow \infty$, $\lambda_{\mathcal{A}}^k \geq 0$, and $-\nabla f(x^*) = \nabla c_{\mathcal{A}}(x^*)\lambda_{\mathcal{A}}^k$ for all k . Dividing by $\|\lambda_{\mathcal{A}}^k\|$ and taking limits we have

$$\nabla c_{\mathcal{A}}(x^*)\lambda_{\mathcal{A}}^k/\|\lambda_{\mathcal{A}}^k\| \rightarrow 0.$$

Note that $\lambda_{\mathcal{A}}^k/\|\lambda_{\mathcal{A}}^k\|$ is a unit vector in $\mathbf{R}^{|\mathcal{A}|}$. By taking a subsequence if necessary, we can identify $t \in \mathbf{R}^{|\mathcal{A}|}$ with $\|t\| = 1$, $t \geq 0$, $t \neq 0$ such that $\lambda_{\mathcal{A}}^k/\|\lambda_{\mathcal{A}}^k\| \rightarrow t$. Thus $\nabla c_{\mathcal{A}}(x^*)t = 0$. MFCQ says there is $w \in \mathbf{R}^n$ such that $\nabla c_{\mathcal{A}}(x^*)^T w > 0$. Thus

$$0 = w^T (\nabla c_{\mathcal{A}}(x^*)t) = (w^T \nabla c_{\mathcal{A}}(x^*)) t > 0,$$

yielding a contradiction.

Strict complementarity.

Define and motivate critical cone $\mathcal{C}(x^*, \lambda^*)$. It's a subset of $\mathcal{F}(x^*)$ by definition.

Lecture 11. (2/14/11; 60 min)

Second-Order Conditions. See p. 330-337.

Discuss both necessary and sufficient. (We don't have sufficient first-order conditions in general, just in the convex case.)

Show that for a direction $w \in \mathcal{F}(x^*) \setminus \mathcal{C}(x^*, \lambda^*)$, we have that f increases along w just by looking at first-order conditions.

Second-order conditions essentially play a tiebreaking role for directions $w \in \mathcal{F}(x^*)$ that are marginal - that is, they are feasible directions but not strongly feasible so the behavior of f along these directions can't be resolved by first-order information alone.

State and prove Theorem 12.5: 2oN conditions. Note that it needs a CQ (we use LICQ but others would work).

State and prove Theorem 12.6: 2oS conditions. *Note that no CQ is needed for this case, even though we make use of the KKT conditions.*

Lecture 12. (2/16/11; 60 min)

Do example 12.9 and discuss.

Also do my favorite two-circle example.

$$\min -x_1 \text{ s.t. } 1 - x_1^2 - x_2^2 \geq 0, \quad 4 - (x_1 + 1)^2 - x_2^2 \geq 0.$$

$x^* = (1, 0)$ is a solution - in fact a strict solution. Review KKT conditions (discussed earlier), KKT multiplier set, 2oN and 2oS conditions.

Section 12.8: Sensitivity of solution of NLP to constraint perturbation, and role of Lagrange multipliers. Showed how under suitable assumptions, the KKT conditions could be reduced (locally to (x^*, λ^*)) to a system of nonlinear equations. The implicit function can then be applied to obtain sensitivity information.

Specified Newton's method for nonlinear equations and derived the KKT system and its Jacobian. Said that a sufficient condition for Newton system of linear equations to have a solution is that LICQ and 2oS hold. Preview of SQP.

Different flavors of duality: Lagrange, Fenchel, Wolfe.

Lecture 13. (2/18/11; 60 min)

Wolfe Duality. See Sec 12.9, pp. 343–350. Follow the book, proving most theorems: 12.10, 12.11, 12.12, 12.13. Show that LP duality is a special case

Lecture 14. (2/21/11; 60 min)

Do duality for convex QP with positive definite Hessian (eliminate x).

Wolfe duality and Theorem 12.14. Do Wolfe duality for convex QP for case of positive *semidefinite* Q (i.e. not eliminating x).

Algorithms: Fundamentals. (Chapter 15.) Discuss SQP approach. Linearized constraints, quadratic objective. Second-order term in objective should incorporate curvature information about the constraints. “Ideal” choice is Lagrangian Hessian $\nabla_{xx}^2 \mathcal{L}(x, \lambda)$. To see why, look at equality-constrained nonlinear programming. Show that Newton’s method applied to the KKT conditions (which are a square system of equations of size $m + n$) gives same update formula as SQP, provided that $\nabla_{xx}^2 \mathcal{L}(x, \lambda)$ is the Hessian term in the objective.

Examine the form of the Jacobian of the nonlinear equations. When LICQ and 2oS hold, this matrix is nonsingular at (x^*, λ^*) . (Details left for homework.)

Lecture 15. (2/23/11; 60 min)

Prove the result about nonsingularity of the KKT Hessian by doing Homework 3, Q3.

Chapter 15. Sketch other different types of NLP algorithm.

Quadratic penalty. Embed in loop of increasing penalty values. Demonstrate on $\min_x x$ s.t. $x - 1 = 0$. Adv: smooth, Hessian at minimizer is positive definite if 2oS satisfied. Disadv: non exact, Hessian is ill conditioned.

Look closely at structure of gradient and Hessian, and related to KKT and 2oS conditions. Proved this lemma: Given H with $w^T H w > 0$ for all $w \neq 0$ with $A^T w = 0$, the matrix $H + \nu A A^T$ is positive definite for all ν sufficiently large. Proof by contradiction: Suppose not, then for all $i = 1, 2, \dots$ there is w_i with $\|w_i\| = 1$ such that $w_i^T (H + i A A^T) w_i \leq 0$. By compactness of unit ball, we have subsequence \mathcal{S} with $w_i \rightarrow w^*$ for some w^* with $\|w^*\| = 1$. Since

$$\|A^T w_i\|_2^2 \leq -\frac{1}{i} w_i^T H w_i,$$

we have by taking limits that $A^T w^* = 0$. Since

$$w_i^T H w_i \leq -i \|A^T w_i\|_2^2 \leq 0,$$

we have too that $(w^*)^T H w^* \leq 0$. This contradicts the assumption on H , so we are done.

Lecture 16. (2/25/11; 65 min)

Quadratic penalty on inequality constraints.

Lagrangian-based. Motivate augmented Lagrangian. Work through details for equality-constrained case. Show that multipliers are good estimates, Hessian of the augmented Lagrangian should be pos definite at a minimizer when 2oS conditions are satisfied for the nonlinear program. Sketch an aug Lagr strategy. The penalty coefficient needs to be large enough only to make the aug Lagr Hessian positive definite, NOT to force x closer to feasibility. (The Lagrange multiplier estimates do this.)

Log-barrier function. Show that optimality conditions for this function are related to KKT. Again, Hessian is pos definite when 2oS conditions are satisfied. Examples:

$$\min x \text{ s.t. } x \geq 0,$$

$$\min x_1 + 0.5(x_2 - .5)^2 \text{ s.t. } x \geq 0.$$

Find log barrier minimizer $x(\mu)$ for each, examine Hessian of log barrier for each. The second one has an ill conditioned Hessian.

Lecture 17. (2/28/11; 65 min)

Primal-dual interior-point. Newton on perturbed KKT. Jacobian is still nonsingular near (x^*, λ^*) . (Details left for homework.)

Show the variant with slack s , so that the variables are (x, s, λ) .

Show relationship to primal log barrier. Introduction of the dual variable explicitly in p-d i-p avoids ill conditioning of Hessian near solution.

Nonsmooth penalty function. Discuss simple example.

Desirable behavior: Algorithms should converge easily to “nice” KKT points (satisfying LICQ and 2oS, say). Tools to promote robust behavior: Descent in penalty functions, filters.

Lecture 18. (3/2/11; 60 min)

Go over solution of Homework 3, Q1.

Maratos effect (Sec 15.5). Work through the example. Defeats the obvious merit and filter approaches.

Outline remedies for Maratos (p. 442). Second-order correction. (Use KKT to show that indeed (15.36) is the min-norm solution of the linearized constraint.

Lecture 19. (3/4/11; 50 min)

Watchdog strategy (p. 444-446).

Discussed exact penalty functions. A specific result: Proved (loosely) that for equality constrained problem the ℓ_1 penalty function is exact when $\nu \geq \|\lambda^*\|_\infty$. Did this by comparing KKT conditions with optimality conditions for the ℓ_1 penalty function. (In the process, introduced the concept of subgradient of a convex nonsmooth function, optimality conditions for convex nonsmooth functions, subdifferential calculus including chain rule. Hand-waved about the extension to nonconvex nonsmooth.) This topic is discussed in the text on pp. 435 et seq. and 507 et seq.

Lecture 20. (3/7/11; 60 min)

Inequality constraints introduce a combinatorial element. (Sec 15.2) Elimination is tricky (Sec 15.3). For linear equality constraints it's easy, however.

Primal-dual interior-point: Chapter 19. Restate using the form with slacks s (as in (19.1)) and restate KKT conditions for this form.

Give symmetric form too, and compressed form (19.15).

Describe why this is nonsingular close to the solution. (Consider the case of inequality constraints only. Use similar reasoning to the homework questions above.)

Discuss linear algebra issues. Which form is better? Depends on fill-in.

Lecture 21. (3/9/11; 60 min)

Outline the basic method: Sec 19.2. Steps are from perturbed KKT. Restrict step length to maintain $(s, z) > 0$, Use norm of modified KKT to decide when to stop iterating for each value of μ .

Prove Theorem 19.1 but first prove the preliminary result that if there is a seq of matrices $B_k \in \mathbf{R}^{n \times t}$ with $n \geq t$ converging to \hat{B} with full column rank, and sequences of vectors $h_k \in \mathbf{R}^n$ converging to \hat{h} , and $z_k \in \mathbf{R}^t$ such that $B_k z_k - h_k \rightarrow 0$, then $\{z_k\}$ has a unique limit $\hat{z} \in \mathbf{R}^t$ satisfying $\hat{B}\hat{z} = \hat{h}$. Proof: Claim that \hat{z} defined uniquely by $\hat{B}^T \hat{B}\hat{z} = \hat{B}^T \hat{h}$ is the unique limit. Set $e_k := B_k z_k - h_k$. Thus:

$$\begin{aligned}(B_k^T B_k - \hat{B}^T \hat{B})z_k + \hat{B}^T \hat{B}z_k &= B_k^T (h_k + e_k) \\ \hat{B}^T \hat{B}\hat{z} &= \hat{B}^T \hat{h}.\end{aligned}$$

Taking differences we obtain

$$\hat{B}^T \hat{B}(z_k - \hat{z}) = (B_k^T h_k - \hat{B}^T \hat{h}) + B_k^T e_k + (B_k^T B_k - \hat{B}^T \hat{B})z_k.$$

Thus for some $\epsilon_k \rightarrow 0$ and $\eta_k \rightarrow 0$ we have

$$\|z_k - \hat{z}\| \leq \epsilon_k + \eta_k \|z_k\| \leq \epsilon_k + \eta_k (\|z_k - \hat{z}\| + \|\hat{z}\|)$$

where the second inequality is from the triangle inequality. Thus

$$(1 - \eta_k)\|z_k - \hat{z}\| \leq \epsilon_k + \eta_k \|\hat{z}\|$$

from which it follows easily that $z_k \rightarrow \hat{z}$.

Proceed with proof of Theorem 19.1.

Question arose in class: Why not just do Newton on original (unperturbed) KKT conditions, with line search to ensure positivity of s and z ? Answer: This is the affine scaling method. It can be made to work in the case of linear programming, with a special line search (see my paper with Renato), but it is very slow. The advantage of perturbation is that the solution of the perturbed KKT conditions is strictly interior, hence we will eventually be able to take *full* Newton steps. It's a path following approach with the path defined by μ ; the idea is to trace the path to the solution.

(* MIDTERM ON 3/9/11 *)

(* NO CLASS 3/11/11 *)

(* SPRING BREAK *)

Lecture 22. (3/21/11; 55 min)

Discuss strategies for updating μ_k : p.572-573. Adjust on every step (decreasing the multiplier to zero to force superlinear convergence), adaptive, probing.

Discuss decreasing μ_k rapidly to get superlinear convergence.

Discuss merit functions (log barrier term and nonsmooth penalties) - p.575. Alternative: Filter, based on two “objectives:” the barrier part $f(x) - \mu \sum \log s_i$, and the equality constraint violation part $\|(c_I(x) - s, c_E(x))\|$.

Quasi-Newton approximate Hessians p.575-576. BFGS or LBFGS updates to the Lagrangian Hessian estimate. Skip if update does not retain positive definiteness. Can use SR1 instead.

Sketch line search algorithm p.577.

Lecture 23. (3/23/11; 60 min)

Note that second-order correction enhancements could be added to the algorithm on page 577.

Wachter-Biegler failure: Section 19.7. Explain. Figure 19.2 is slightly wrong: the feasible region does not begin until the parabola crosses the horizontal axis. Also the second-last term $X^{(1)}$ on page 587 should probably be $s_1^{(1)}$.

Trust region methods. Since much of the class hadn't done trust regions in 726, cover the basics of TR for unconstrained problems. Extend to SQP trust region for constrained problems. Difficulty arises because subproblem may not be feasible, for small Δ_k . Need to adjust the linear constraint to make it "less ambitious" than trying to resolve all the infeasibility in a single step.

Trust-region: Sec 19.5. Implemented in KNITRO-CG. Write down (19.31) for inequality constraints only, and show that when the trust-region constraints are omitted, and $z + \Delta z$ is used for Lagrange multipliers, then KKT conditions for (19.31) are simply (19.6), with scaling by S^{-1} . Did not go into much detail on this.

Discuss use of implicit function theorem (p.631) to show that the solution vector (x, s, y, z) at μ is within $O(\mu)$ of the optimal values, provided that KKT, LICQ, 2oS, strict complementarity all hold. Do it for inequality constraints only. In IFT set $z \leftarrow (x, s, z)$ and $t = \mu$, and let

$$h(z, t) \leftarrow \begin{bmatrix} \nabla f(x) - A_I(x)^T z \\ c_I(x) - z \\ SZe - \mu e \end{bmatrix}.$$

Note that

$$\nabla_z h \leftarrow \begin{bmatrix} \nabla_{xx}^2 \mathcal{L}(x, z) & -A_I(x)^T & 0 \\ A_i(x) & 0 & -I \\ 0 & S & Z \end{bmatrix}, \quad \nabla_t h(z, t) \leftarrow \begin{bmatrix} 0 \\ 0 \\ -e \end{bmatrix}.$$

Verify conditions of IFT in turn: (i) (z^*, s^*, z^*) clearly satisfies the equations for $\mu = 0$; (ii) smoothness follows from second-order continuous; (ii) Nonsingularity of $\nabla_z h(z^*, 0)$ follows from LICQ, 2oS, strict complementarity (Tricky! Have to split the triple into 5 subvectors!) Conclusion of IFT is then that there is a smooth path of solutions with

$$\nabla z(\mu) = -\nabla_t h(z(t), t) [\nabla_z h(z(t), t)]^{-1},$$

which is nice. It follows from Taylor's theorem that

$$z(\mu) = z(0) + \mu \int_0^\mu \nabla z(t) dt$$

so that using boundedness of $\|z(t)\|$ for all t sufficiently small, we have

$$\|z(\mu) - z(0)\| \leq \mu \int_0^\mu \|\nabla z(t)\| dt = O(\mu).$$

Talk about how we get superlinearity from this:

- Can decrease μ rapidly;
- Steplengths α_s^{\max} and α_z^{\max} will be close to 1 — if we let $\tau \rightarrow 1$;
- One Newton step will suffice for each μ .

Lecture 24. (3/25/11; 60 min)

Augmented Lagrangian. Outlined the overall method for equality constrained problem. Write as

$$\min_x \max_{\lambda} \mathcal{L}(x, \lambda).$$

Now use prox-point stabilization on the dual variable:

$$\min_x \max_{\lambda} f(x) - \lambda^T c(x) + \frac{1}{2\mu_k} \|\lambda - \lambda^k\|_2^2.$$

Perform max explicitly, get $\lambda = \lambda^k + \mu_k c(x^k)$. Substitute to get

$$\min_x f(x) + (\lambda^k)^T c(x) + \frac{\mu_k}{2} \|c(x)\|^2.$$

Define \mathcal{L}_A accordingly. Denote the minimizer of this subproblem by x^k .

Framework 17.3: The approximate solution for iteration k can be used as start point for iteration $k + 1$. A possible termination condition for the subproblem is $\|\nabla \mathcal{L}_A(x^{k+1}, \lambda^k; \mu_k)\| \leq \tau_k$.

Do Example 17.4. (Compare Fig 17.5 with Fig 17.1.)

Do Theorem 17.5: Under nice conditions on solution x^* , and if λ is exact optimum, there is threshold value $\bar{\mu}$ such that for all $\mu \geq \bar{\mu}$, x^* is strict local minimum for $\mathcal{L}_A(\cdot, \lambda^*; \mu)$.

Lecture 25. (3/28/11; 60 min)

Explain Theorem 17.6.

Add explicit constraints — membership of a convex set Ω :

$$\min f(x) \text{ s.t. } c(x) = 0, \quad x \in \Omega,$$

where the augr Lagr is defined as

$$\mathcal{L}_A(x, \lambda; \mu) = f(x) - \lambda^T c(x) + \frac{\mu}{2} \|c(x)\|_2$$

and the x -subproblem at iteration k is

$$\min_{x \in \Omega} \mathcal{L}_A(x, \lambda_k; \mu_k).$$

Measure suboptimality by

$$\|x - P_\Omega(x - \nabla_x \mathcal{L}_A(x, \lambda_k; \mu_k))\|.$$

Application to equalities plus bounds. Show that general inequality constrained NLPs can be formulated in this way by introducing slacks. Show how we can explicitly minimize with respect to the slacks, and obtain a subproblem in the original variables x alone.

Discuss briefly projected gradient methods for bound-constrained minimization.

Discuss Algorithm 17.4 — minor modification for explicit constraints.

Splitting: Consider

$$\min_u f(u) + h(u)$$

where it is easy to minimize with respect to f and h separately but not jointly. Here can use splitting in conjunction with augmented Lagrangian. Reformulate as a constrained problem:

$$\min_{(u,v)} f(u) + h(v) \text{ s.t. } u - v = 0,$$

for which the augmented Lagrangian is

$$\mathcal{L}_A(u, v, \lambda; \mu) = f(u) + h(v) + \lambda^T (u - v) + \frac{\mu}{2} \|u - v\|_2^2.$$

Solving the subproblem with respect to u and v jointly may still be hard. Can do alternating directions minimization:

$$u_{k,j+1} = \arg \min_u f(u) + h(v_{k,j}) + \lambda_k^T (u - v_{k,j}) + \frac{\mu_k}{2} \|u - v_{k,j}\|_2^2;$$

$$v_{k,j+1} = \arg \min_v f(u_{k,j+1}) + h(v) + \lambda_k^T (u_{k,j+1} - v) + \frac{\mu_k}{2} \|u_{k,j+1} - v\|_2^2.$$

Start with $u_{k,0} = u_k$ and $v_{k,0} = v_k$ and cycle around $j = 0, 1, 2, \dots$ until termination criteria satisfied. Then set u_{k+1}, v_{k+1} to the final inner iterates.

Maybe take just one inner iteration.

Lecture 26. (3/30/11; 60 min)

Sent email to class re Homework 5. Common error: LICQ does NOT mean that $\nabla_{c_A}(x^*)^T \Delta x = 0 \Rightarrow \Delta x = 0$. Recall directional derivative and how it applied in Q2.

For inequality constraints, re-derive the update formulas without the use of slacks: pp. 523-524.

SQP. Section 18.1. Equality constraint first and solving KKT conditions: Lagrange-Newton. Equivalence to (local solution of a) QP subproblem. Linear equations are locally the same.

Natural extension to inequality constraints - yields a QP subproblem.

Lecture 27. (4/1/11; 60 min)

Discussed the nonsmooth equations formulation of the KKT conditions for inequality constrained problem, with $\min(\lambda_i, c_i(x)) = 0$ for $i \in \mathcal{I}$. Newton's method does not work because of the nonsmoothness, but semismooth methods work. Derive the basic semismooth Newton equations, and show that they are simply the KKT conditions for the SQP subproblem.

p.533, Theorem 18.1: Robinson result on existence of a subproblem solution.

Sec 18.2: EQP and IQP methods. IQP more popular. But SLQP, which is an EQP method, recently has some success.

Discussed in some detail the concept of active constraint identification. Discussed the SLP subproblem, with trust regions added: $\|\cdot\|_\infty$ and $\|\cdot\|_2$. The latter is

$$\min_d \nabla f(x)^T d \text{ s.t. } A(x)d + c(x) \geq 0, \quad \|d\|_2 \leq \Omega,$$

hard to solve directly, but we can form the equivalent QP with quadratic regularization:

$$\min_d \nabla f(x)^T d + \frac{\mu}{2} d^T d \text{ s.t. } A(x)d + c(x) \geq 0.$$

Discuss the relationship of these two problems. If we use the ℓ_∞ norm instead, the subproblem may be infeasible if the trust region radius is too small. (The same is true if we impose the ℓ_2 trust region as an explicit constraint $\|d\|_2 \leq \Omega$.) This formulation has the advantage of being a linear program, for which efficient software is available.

(* NO LECTURES 4/4/11 AND 4/6/11 DUE TO TRIP TO UBC. *)

Lecture 28. (4/8/11; 60 min)

SQP: Partitioning the step into Y and Z subspaces. Define them for the equality constrained case, pp. 538-539. Write out KKT conditions for the equality constrained subproblem again. Derive least-squares multiplier formula (18.21)-(18.22) from this partitioned system, but neglecting the lower-order p_k term. Write out formula for p_Z with $Z^T \nabla_{xx}^2 \mathcal{L} Z$ coefficient matrix. Neglect p_Y term from RHS.

Derive reduced Hessian update formulae: p.539-540. Since $Z^T \nabla_{xx}^2 \mathcal{L} Z$ is positive definite when second-order conditions hold, it's natural to use formulae such as BFGS. Motivate by saying that the Z space may be low dimensional.

What happens in the quasi-Newton case? Mention 2-step superlinear convergence for the reduced-Hessian approximation scheme.

Lecture 29. (4/11/11; 60 min)

More on the equality constrained problem: Rate of convergence of SQP with exact Hessian information. Q-superlinear in (x, λ) , hence R-superlinear in x . Follows from analysis of general Newton method for nonlinear equations, using 2oS and LICQ to get nonsingularity of the matrix.

If we estimate λ by the least-squares procedure, we can get Q-quadratic in x alone, since $\|\lambda^k - \lambda^*\| = O(\|x^k - x^*\|)$.

Inequality constrained case: With accurate identification of active constraints, it reduces to equality constrained case, locally, so superlinear convergence follows from the analysis for the equality constrained case.

Recall reduced-Hessian quasi-Newton methods discussed last time. Problematic in the inequality constrained case - we have to use an explicit active set estimate, and need a way to expand and shrink the reduced Hessian approx when the active set changes (its size changes).

Lecture 30. (4/13/11; 60 min)

Quasi-Newton approx to full Lagrangian Hessian is more appealing as we don't need to estimate active sets. Damped updating (p. 537) because the usual update doesn't maintain positive definiteness of the full Hessian, which property may not be present anyway in a constrained problem.

Discuss $\ell_1 PF(\mu)$ for the equality constrained case. In this context, discuss the directional derivative stuff from p. 628 and compute directional derivatives of $\|c(x)\|_1$ in a general direction d .

Prove Theorem 18.2 (p. 541) about how the SQP step affects this function. Discuss how to increase penalty parameter μ if it seems too small. The book on p. 542-543 describes a detailed strategy for adjusting μ - perhaps assign this as part of homework.

Discuss second-order corrections: p. 544. Derive the formulas.

Lecture 31. (4/15/11; 60 min)

Discuss line-search framework, Sec 18.4., p. 545.

Conic Optimization. Given cone $K \in \mathbb{R}^n$, we say $x \succeq_K y$ if $x - y \in K$. We say $x \succ y$ if $x - y \in \text{int}K$.

We can also have cones in the space of symmetric matrices. Here $X \succeq 0$ means that X is positive semidefinite. (Ex: Prove that this set is a cone.)

General form of conic optimization:

$$\min c^T x \text{ s.t. } Fx + g \succeq_K 0, \quad Ax = b,$$

where g the range of F and K are in some space (e.g. a Euclidean space or the space $\mathcal{SR}^{n \times n}$ of $n \times n$ symmetric matrices). Examples:

- When K is the nonnegative orthant, this is a linear program.
- Define K_i to be the second-order cone:

$$K_i = \{(y, t) \in \mathbb{R}^{k_i+1} \mid \|y\|_2 \leq t\}.$$

Then the second-order cone programming problem is as follows:

$$\min c^T x \text{ s.t. } (A_i x + b_i, c_i^T x + d_i) \succeq_{K_i} 0, \quad i = 1, 2, \dots, m, \quad Fx = g.$$

- Semidefinite programming:

$$\max c^T x \text{ s.t. } x \sum_{i=1}^m x_i F_i + G \succeq 0, \quad Ax = b,$$

where F_i and G are all symmetric matrices.

All these cases admit a barrier function that leads to algorithms with polynomial complexity.

A dual pair of SDP can be obtained by defining the following inner product between symmetric matrices: $A \bullet B = \text{tr}(A^T B) = \sum_{i,j=1}^n A_{ij} B_{ij}$. Then

$$\text{(SDP-P)} \quad \min_X C \bullet X \text{ s.t. } A_i \bullet X = b_i, \quad i = 1, 2, \dots, m, \quad X \succeq 0, \quad (0.2)$$

where A_i and C and X are all in $\mathcal{SR}^{n \times n}$, and

$$\text{(SDP-D)} \quad \max b^T y \text{ s.t. } \sum_{i=1}^m y_i A_i \preceq C. \quad (0.3)$$

An alternative form of (SDP-D) uses a “slack matrix” S :

$$\text{(SDP-D)} \quad \max_{y,S} b^T y \text{ s.t. } \sum_{i=1}^m y_i A_i + S = C, \quad S \succeq 0.$$

Showed how LP is a special case of SDP. Given LP:

$$\min c^T x \text{ s.t. } Ax = b, \quad x \geq 0,$$

define $C := \text{diag}c$, $A_i := \text{diag}A_i$, E_{ij} is the matrix in $\mathcal{SR}^{n \times n}$ with all zeros except for 1 in the (i, j) and (j, i) position, and define the SDP as:

$$\min_X C \bullet X \text{ s.t. } A_i \bullet X = b_i, \quad X \succeq 0, \quad E_{ij} \bullet X = 0.$$

The optimal X will be diagonal and the diagonal elements correspond to the solution x of the LP.

Lecture 32. (4/18/11; 65 min)

Also did an example in which there are TWO positive semidefinite matrix variables X and Z - showed that it can be set up in standard form in terms of a matrix

$$\begin{bmatrix} X & 0 \\ 0 & Z \end{bmatrix},$$

Using matrices E_{ij} as above to set the desired off-diagonals to zero.

If the problem has a symmetric matrix variable X that is *not* required to be positive semidefinite, can express it in standard form by splitting: $X = X^+ - X^-$ with $X^+ \succeq 0$ and $X^- \succeq 0$. Justification: Write

$$X = Z\Lambda Z^T = \sum_{i=1}^n \lambda_i z_i z_i^T,$$

with λ_i all real and Z orthogonal. The corresponding X^+ and X^- are then

$$X^+ = \sum_{\lambda_i \geq 0} \lambda_i z_i z_i^T, \quad X^- = - \sum_{\lambda_i < 0} \lambda_i z_i z_i^T.$$

Often convenient to have both matrix and vector variables in the formulation. Codes allow this e.g. SDPT3.

Example (Boyd and Vandenberghe): Given matrix function $A(x) := A_0 + \sum_{i=1}^m x_i A_i$, with all $A_i \in \mathbf{SR}^{n \times n}$, problem is $\min_x \|A(x)\|_2$. Express first as

$$\min_{t,x} t \quad \text{s.t.} \quad t^2 I - A(x)^T A(x) \geq 0,$$

which can be written as

$$\min_{t,x} t \quad \text{s.t.} \quad \begin{bmatrix} tI & A(x) \\ A(x)^T & tI \end{bmatrix} \succeq 0,$$

which clearly has (SDP-D) form.

Example: Relaxation of nonconvex quadratic programming. (Nonconvex QP is NP-hard, so there is no hope of an exact SDP formulation.)

$$\min \frac{1}{2} x^T Q x + c^T x + \alpha \quad \text{s.t.} \quad Hx = d, \quad x \geq 0,$$

where Q is symmetric but not necessarily positive definite. Set up an SDP with variable \tilde{X} of the form:

$$\tilde{X} = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix},$$

where we would like to have $X_{ij} = x_i x_j$ for all i, j . If this is indeed true then \tilde{X} will be spsd and rank-one. Define SDP by setting

$$C := \begin{bmatrix} \alpha & 0.5c^T \\ 0.5c & 0.5Q \end{bmatrix}, \quad A_i = \begin{bmatrix} -d_i & 0.5H_i \\ 0.5H_i^T & 0 \end{bmatrix},$$

and also

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and E_i is all zeros except for 1 in the $(1, i + 1)$ and $(i + 1, 1)$ positions. The SDP is then

$$\min_{\tilde{X}} C \bullet \tilde{X} \quad \text{s.t.} \quad A_i \bullet \tilde{X} = b_i \quad (i = 1, 2, \dots, m), \quad A_0 \bullet \tilde{X} = 1, \quad E_i \bullet \tilde{X} \geq 0.$$

(Can express the last constraint in standard SDP form by introducing a slack, putting it into a matrix and requiring the matrix to be spsd. If we are willing to let real variables into the formulation, we can just impose $s_i \geq 0$ directly on each slack.)

We cannot express our desire for $\text{rank}(\tilde{X}) = 1$ in the formulation, since this is a nonconvex constraint. By leaving this constraint out, we get the (convex) SDP relaxation above. Thus the optimal \tilde{X} gives a lower bound on the optimal objective of the QP.

If the optimal \tilde{X} is in fact rank-1, the lower bound is attained by the QP and we can extract the optimal x for the QP from \tilde{X} .

Lecture 33. (4/20/11; 65 min)

Extend to quadratic constraints: A constraint

$$\frac{1}{2}x^T G_i x + H_i x - d_i = 0$$

can be relaxed as $A_i \bullet X = 0$ by setting

$$A_i = \begin{bmatrix} -d_i & 0.5H_i \\ 0.5H_i^T & G_i \end{bmatrix}.$$

Continuing the example above, suppose that we have a general QP over BINARY variables $x_i \in \{0, 1\}$. Enforce binary-ness with this ingenious quadratic constraint:

$$x_i(1 - x_i) = 0.$$

Thus the SDP relaxation will include $A_i \bullet X = 0$, where

$$A_i = \begin{bmatrix} 0 & 0.5e_i^T \\ 0.5e_i & -E_i \end{bmatrix}.$$

where e_i is the i th unit vector and E_i has 1 in the (i, i) position and zeros elsewhere. Could also include the constraints $0 \leq x_i \leq 1$.

Other wacky formulation tricks for SDP relaxation of QP:

- Square the constraint $H_i x = d_i$ to get $(H_i x)^2 = d_i^2$ and use the technique above to get an SDP relaxation of the squared constraint. Adds only m constraints to the SDP formulation.
- Add the quadratic constraints $x_j(H_i x - d_i) = 0$ for $j = 1, 2, \dots, n$ and $i = 1, 2, \dots, n$. Adds mn constraints to the SDP form - potentially a lot. Could be selective here.
- multiply constraints together: $(H_i x - d_i)(H_k x - d_k) = 0$.

CODE: Good codes available, but limited in the size of problems they can handle.

I use **SeDuMi**. Google it. Also **SDPT3**.

Example (Boyd and Vandenberghe): Estimate a vector n from measurements $y = Ax + w$, where $w \sim N(0, I)$ is measurement noise. Choose rows a_i^T of the matrix A from a “menu” of possible test vectors $v^{(1)}, \dots, v^{(M)}$, so as to design an experiment that is “maximally informative” about x .

Important role played by the error covariance $(A^T A)^{-1}$.

Suppose that λ_i is the fraction of rows of A that are chosen to be $v^{(i)}$. Then $A^T A = q \sum_{i=1}^M \lambda_i v^{(i)} (v^{(i)})^T$. (Ignore the fact that λ_i should be an integer multiple of $1/q$.) Can maximize the smallest eigenvalue of $A^T A$ via this SDP:

$$\max_{t, \lambda} t \quad \text{subject to} \quad \sum_{i=1}^M \lambda_i v^{(i)} (v^{(i)})^T - tI \geq 0, \quad \sum_{i=1}^M \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, 2, \dots, M.$$

(Exercise: Write the SDP to minimize the trace of $(A^T A)^{-1}$. Actually I don't know how to do this.)

We can formulate the constraint $\sum_{i=1}^M \lambda_i = 1$ as a matrix semidefiniteness constraint as in (??). Setting

$$F_0 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad F_i = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad i = 1, 2, \dots, M$$

we find that this constraint is equivalent to

$$F_0 + \sum_{i=1}^M \lambda_i F_i \succeq 0.$$

Similarly, the nonnegativity constraints $\lambda_i \geq 0$, $i = 1, 2, \dots, M$ can be expressed by setting F_0 to be the $M \times M$ matrix of zeros, and each F_i , $i = 1, 2, \dots, M$ to be the $M \times M$ matrix whose elements are all zero except for a 1 in the (i, i) position.

Weak Duality. Duality theory for SDP is more complex than linear programming.

Weak duality is comparatively easy.

THEOREM 0.5 (Weak Duality). *If X is feasible for (0.2) and (y, S) is feasible for (0.3), then*

$$C \bullet X - b^T y = X \bullet S \geq 0.$$

Lecture 34. (4/22/11; 60 min)

Proof. We have by substitution from the constraints in (0.2) and (0.3) that

$$C \bullet X - b^T y = \left(\sum_{i=1}^m y_i A_i + S \right) \bullet X - b^T y = \sum_{i=1}^m y_i (A_i \bullet X) + S \bullet X - b^T y = S \bullet X = X \bullet S.$$

Since X is positive semidefinite, it has a square root $X^{1/2}$. Since $\text{trace}(PQ) = \text{trace}(QP)$ for any symmetric Q, P , we have

$$X \bullet S = \text{trace}(XS) = \text{trace}(X^{1/2} X^{1/2} S) = \text{trace}(X^{1/2} S X^{1/2}).$$

Since the trace of a symmetric matrix is the sum of its eigenvalues, we have

$$X \bullet S = \sum_{j=1}^n \lambda_j (X^{1/2} S X^{1/2}) \geq 0,$$

where λ_j denotes the j th eigenvalue. The final inequality follows from the fact that $X^{1/2} S X^{1/2}$ is positive semidefinite. \square

0.1. Strong Duality: Discussion and “Counterexamples”. Strong duality is not as easy. Some examples illustrate the difficulties that can arise.

Example 1. (No duality gap, but optimal value not attained by dual.) Let $n = 2$, $m = 2$. Define problem

$$\max -y_1, \quad \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} y_1 + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} y_2 \preceq \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The constraints are equivalent to

$$\begin{bmatrix} y_1 & 1 \\ 1 & y_2 \end{bmatrix} \succeq 0,$$

which is true if and only if $(y_1, y_2) > 0$, $y_1 y_2 \geq 1$. (Derive this by forming the quadratic for the eigenvalues.) Hence, optimal value is 0, but not attained. (We can come arbitrarily close with $(\epsilon, 1/\epsilon)$ for $\epsilon > 0$. The primal is

$$\min \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \bullet X, \quad \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \bullet X = -1, \quad \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \bullet X = 0, \quad X \succeq 0.$$

The only feasible (hence optimal) solution is

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Example 2. (Both primal and dual attain optimal values, but there remains a duality gap.) The problem data is:

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad b_1 = 0, \quad b_2 = 2.$$

Any X satisfying the equality constraints has the form

$$X = \begin{bmatrix} 0 & x_{12} & x_{13} \\ x_{12} & x_{22} & x_{23} \\ x_{13} & x_{23} & 1 - x_{12} \end{bmatrix}.$$

Applying $X \succeq 0$ then yields

$$X = \begin{bmatrix} 0 & 0 & 0 \\ 0 & x_{22} & x_{23} \\ 0 & x_{23} & 1 \end{bmatrix},$$

with $x_{22} \geq 0$ and $x_{22} - x_{23}^2 \geq 0$. Hence the primal optimal value is 1 and it is attained by *all* feasible X . For the dual, we require

$$S = \begin{bmatrix} -y_1 & -y_2 & 0 \\ -y_2 & 0 & 0 \\ 0 & 0 & 1 - 2y_2 \end{bmatrix} \succeq 0.$$

so that $y_2 = 0$ and $y_1 \leq 0$. Hence $y = (0, 0)$ is optimal, with value 0.

Example 3. (Primal is infeasible, but dual is feasible and attains its optimum.) The data is

$$C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad b_1 = 0, \quad b_2 = 2.$$

From the equality constraints, X must have the form

$$X = \begin{bmatrix} 0 & 1 \\ 1 & x_{22} \end{bmatrix},$$

but this matrix cannot be positive semidefinite, so the primal is infeasible. For the dual, we require

$$S = \begin{bmatrix} -y_1 & -y_2 \\ -y_2 & 0 \end{bmatrix} \succeq 0,$$

which implies that $y_1 \leq 0$ and $y_2 = 0$. Hence $y = (0, 0)$ is optimal, with optimal value 0.

We now define feasible and strictly feasible sets in much the same way as for LP.

$$F(P) \stackrel{\text{def}}{=} \{X \in \mathbf{SR}^{n \times n} \mid A_i \bullet X = b_i, \quad i = 1, 2, \dots, m, \quad X \succeq 0\};$$

$$F^0(P) \stackrel{\text{def}}{=} \{X \in F(P) \mid X \succ 0\};$$

$$F(D) \stackrel{\text{def}}{=} \{(y, S) \in \mathbf{R}^m \times \mathbf{SR}^{n \times n} \mid \sum_i A_i y_i + S = C, \quad S \succeq 0\};$$

$$F^0(D) \stackrel{\text{def}}{=} \{(y, S) \in F(D) \mid S \succ 0\}.$$

We state the main duality results here. Their proofs are delayed until after we examine properties of the central path.

THEOREM 0.6. *Suppose that (0.2) and (0.3) are feasible and that $F^0(D)$ is nonempty. Then (0.2) has a nonempty compact set of solutions, and the optimal values of (0.2) and (0.3) are identical.*

Similarly, if (0.2) and (0.3) are feasible and that $F^0(P)$ is nonempty, then (0.3) has a nonempty compact set of solutions, and the optimal values of (0.2) and (0.3) are identical.

COROLLARY 0.7. *Suppose that both $F^0(P)$ and $F^0(D)$ are nonempty. Then both (0.2) and (0.3) have nonempty compact sets of solutions, and the optimal values are identical.*

0.2. Logarithmic Barrier Functions and the SDP Central Path. We define a barrier function on $S\mathbf{R}^{n \times n}$ as follows:

$$f(X) = -\ln \det X \quad \text{if } X \succ 0; \quad f(X) = +\infty \quad \text{otherwise.}$$

Since

$$\det X = \prod_{i=1}^n \lambda_i(X),$$

where $\lambda_i(X)$ are the eigenvalues of X , ordered with largest first, we can write

$$f(X) = -\sum_{i=1}^n \ln \lambda_i(X), \quad \text{if } X \succ 0.$$

Lecture 35. (4/25/11; 60 min)

As a barrier function should, $f(X)$ approaches $+\infty$ as X approaches the boundary of the set $S\mathbf{R}^{n \times n}$. This is because $\lambda_n(X)$ and possibly some other eigenvalues approach zero, so the contributions of these terms to the sum goes to ∞ , while the other terms are bounded.

We now define a *barrier subproblem* for (0.2), in which the positive semidefiniteness constraint is replaced by a barrier function term, parametrized by a coefficient ν :

$$BP(\nu) : \quad \min_{X \in S\mathbf{R}^{n \times n}} C \bullet X + \nu f(X), \quad A_i \bullet X = b_i, \quad i = 1, 2, \dots, m. \quad (X \succ 0)$$

We hope that the solution $X(\nu)$ of this problem approximates the solution of (0.2), and that $X(\nu) \rightarrow X^*$ as $\nu \downarrow 0$.

(Maybe discuss example of the trivial problem $\min x$ s.t. $x \geq 0$ solved by a barrier function.)

The log barrier subproblem for the dual (0.3) is

$$BD(\nu) : \quad \max_{y \in \mathbf{R}^m, S \in S\mathbf{R}^{n \times n}} b^T y - \nu f(S) \quad \sum_{i=1}^m y_i A_i + S = C. \quad (S \succ 0)$$

Now examine derivatives of f . First note that for any matrix E , we have

$$\det(I + \alpha E) = 1 + \alpha \text{trace} E + O(\alpha^2).$$

Demonstrate this for $n = 2$. It follows for larger n by induction.

For first derivative, have

$$\begin{aligned} f(X + \alpha H) &= -\ln \det[X(I + \alpha X^{-1}H)] \\ &= -\ln \det X - \ln(1 + \alpha \text{trace} X^{-1}H + O(\alpha^2)) \\ &= f(X) - \alpha X^{-1} \bullet H + O(\alpha^2), \end{aligned}$$

which implies that $f'(X) = -X^{-1}$.

Note that $f'(X)$ is a linear operator, so that “ $f'(X)H$ ” is not to be understood as a matrix multiplication, but rather the action of the linear operator on a given matrix H . The “action” in this case is the bullet-product of $-X^{-1}$ with the given matrix.

We also have

$$\begin{aligned} f'(X + \alpha H) &= -[X(I + \alpha X^{-1}H)]^{-1} \\ &= -[I - \alpha X^{-1}H + O(\alpha^2)]X^{-1} \\ &= f'(X) + \alpha X^{-1}HX^{-1} + O(\alpha^2). \end{aligned}$$

Hence, if we define the matrix operator \odot by

$$(P \odot Q)U = \frac{1}{2}(PUQ^T + QUP^T),$$

we can see from the expression above that $f''(X)H = X^{-1}HX^{-1}$, so we can represent $f''(X)$ by $X^{-1} \odot X^{-1}$. Here $f''(X)$ is to be understood as a bilinear operator, where

$$f''(X)UV = (X^{-1}UX^{-1}) \bullet V.$$

Note that $f(X)$ satisfies a strict convexity property in that $f''(X)HH > 0$ for all $H \neq 0$. We have

$$\begin{aligned}
f''(X)HH &= (X^{-1}HX^{-1}) \bullet H \\
&= \text{trace}(X^{-1}HX^{-1}H) \\
&= \text{trace}(X^{-1/2}X^{-1/2}HX^{-1}H) \\
&= \text{trace}(X^{-1/2}HX^{-1}HX^{-1/2}) \\
&= \text{trace}((X^{-1/2}HX^{-1/2})(X^{-1/2}HX^{-1/2})) \\
&= \|X^{-1/2}HX^{-1/2}\|_F^2 > 0 \text{ for all } H \neq 0
\end{aligned}$$

If $BP(\nu)$ has a solution, then we must have $X \in F^0(P)$ (in particular $X \succ 0$), and the first order conditions yield

$$0 = C + \nu f'(X) - \sum_{i=1}^m y_i A_i = C - \nu X^{-1} - \sum_{i=1}^m y_i A_i,$$

for some $y_i, i = 1, 2, \dots, m$. If we define $S = \nu X^{-1} \succ 0$, we have that $(y, S) \in F^0(D)$, and so (X, y, S) solve the following systems:

$$\begin{aligned}
\sum_i y_i A_i + S &= C, \quad (S \succ 0), \\
A_i \bullet X &= b_i, \quad i = 1, 2, \dots, m, \quad (X \succ 0), \\
XS &= \nu I.
\end{aligned}$$

To simplify this expression we introduce the following notation:

$$\mathcal{A}X = [A_i \bullet X]_{i=1}^m, \quad \mathcal{A}^*y = \sum_{i=1}^m y_i A_i.$$

We then have

$$\begin{aligned}
CPE(\nu) : \quad \mathcal{A}^*y + S &= C, \quad (S \succ 0), \\
\mathcal{A}X &= b, \quad (X \succ 0), \\
XS &= \nu I.
\end{aligned}$$

“ $CPE(\nu)$ ” stands for “central path equations.”

Lecture 36. (4/27/11; 60 min)

Note the similarity of $CPE(\nu)$ to the equations defining the central path for linear programming, which are

$$\begin{aligned} A^T y + s &= c, \quad (s > 0), \\ Ax &= b, \quad (x > 0), \\ XSe &= \nu e, \end{aligned}$$

where $X = \text{diag}(x_1, x_2, \dots, x_n)$ and $S = \text{diag}(s_1, s_2, \dots, s_n)$.

For the results of this section we need a *linear independence assumption*, which is simply that the matrices A_i , $i = 1, 2, \dots, m$ are linearly independent.

THEOREM 0.8. *Suppose that $F^0(P)$ and $F^0(D)$ are nonempty and that the linear independence condition holds. Then for any positive ν , there is a unique solution $(X(\nu), y(\nu), S(\nu))$ to $CPE(\nu)$. Further, $X(\nu)$ is the unique solution to $BP(\nu)$ and $(y(\nu), S(\nu))$ is the unique solution to $BD(\nu)$. Finally, if the assumption of strict feasibility fails, then $CPE(\nu)$, $BP(\nu)$, and $BD(\nu)$ have no solutions.*

Proof. (Todd [1]) First we establish existence. Choose any $\hat{X} \in F^0(P)$ and $(\hat{y}, \hat{S}) \in F^0(D)$. Since $\hat{S} \succ 0$, we have its smallest eigenvalue is positive: $\sigma \stackrel{\text{def}}{=} \lambda_{\min}(\hat{S}) > 0$. We have that \hat{X} is feasible for $BP(\nu)$, and for feasible X , $C \bullet X$ differs from $\hat{S} \bullet X$ by a constant, because

$$\hat{S} \bullet X = (C - \mathcal{A}^* \hat{y}) \bullet X = C \bullet X - \hat{y}^T \mathcal{A} \bullet X = C \bullet X - \hat{y}^T b.$$

Hence $BP(\nu)$ has the same set of solutions as the following subproblem, which we call $BP'(\nu)$:

$$BP'(\nu) : \quad \min \hat{S} \bullet X + \nu f(X), \quad \mathcal{A}X = b, \quad \hat{S} \bullet X + \nu f(X) \leq \hat{S} \bullet \hat{X} + \nu f(\hat{X}). \quad (\hat{X} \succ 0)$$

We aim to show that this problem is one of minimizing a continuous function over a compact set, hence its solution exists.

Let $\lambda(X)$ be the vector of eigenvalues $\lambda_j(X)$, $j = 1, 2, \dots, n$. For X feasible for $BP'(\nu)$, we have that $\lambda(X) > 0$ and that

$$f(X) = -\ln \det X = -\sum_{i=1}^n \ln \lambda_i.$$

Note that for any symmetric $P \succeq 0$ and $Q \succeq 0$ we have that $P \bullet Q \succeq 0$ (as we showed during the weak duality proof). Hence,

$$\hat{S} \bullet X = (\hat{S} - \sigma I) \bullet X + \sigma I \bullet X \geq \sigma I \bullet X = \sigma \text{trace}(X) = \sigma \sum_{i=1}^n \lambda_i,$$

where the first inequality follows from $X \succeq 0$ and $\hat{S} - \sigma I \succeq 0$. Therefore,

$$\sigma \sum_{i=1}^n \lambda_i - \nu \sum_{i=1}^n \ln \lambda_i \leq \hat{S} \bullet \hat{X} + \nu f(\hat{X}) \stackrel{\text{def}}{=} \alpha,$$

and hence

$$\sum_{i=1}^n (\sigma \lambda_i - \nu \ln \lambda_i) \leq \alpha.$$

Now the function $t(\tau) = \sigma\tau - \nu \ln \tau$ has a unique minimizer at $\tau^* = \nu/\sigma$, and it goes to $+\infty$ as $\tau \downarrow 0$ or $\tau \uparrow \infty$. Suppose that the minimum value of $t(\tau)$ is β , and choose $\underline{\tau}$ and $\bar{\tau}$ such that

$$\sigma\tau - \nu \ln \tau \leq \alpha - (n-1)\beta \Rightarrow \tau \in [\underline{\tau}, \bar{\tau}].$$

(Draw a picture of $t(\tau)$, $\underline{\tau}$, $\bar{\tau}$, α , β .) If we were to have $\lambda_j \notin [\underline{\tau}, \bar{\tau}]$ for some $j = 1, 2, \dots, n$, then

$$\sigma\lambda_j - \nu \ln \lambda_j > \alpha - (n-1)\beta$$

and so

$$\begin{aligned} \alpha &\geq \sum_{i=1}^n (\sigma\lambda_i - \nu \ln \lambda_i) \\ &= (\sigma\lambda_j - \nu \ln \lambda_j) + \sum_{i \neq j} (\sigma\lambda_i - \nu \ln \lambda_i) > (\alpha - (n-1)\beta) + (n-1)\beta = \alpha, \end{aligned}$$

a contradiction. Hence we must have $\lambda_j \in [\underline{\tau}, \bar{\tau}]$ for all $j = 1, 2, \dots, n$. Therefore we have $\|X\|_F = \|\lambda(X)\|_2 \leq \sqrt{n\bar{\tau}}$, so the feasible set for $BP'(\nu)$ is bounded.

Moreover since $\lambda_j \geq \underline{\tau} > 0$ for all j , the objective function $\hat{S} \bullet X + \nu f(X)$ is continuous. Hence the feasible set for $BP'(\nu)$ is also closed. Hence, it is compact.

Since the objective function for $BP'(\nu)$ is continuous and since the feasible set is compact, the minimum is attained. Hence, by the first-order necessary conditions, the conditions $CPE(\nu)$ hold at the solution. Since $BP'(\nu)$ is a convex minimization problem, the necessary conditions are also sufficient. In fact, since the objective is *strictly* convex (see above for the proof that $f''(X)$ is “positive definite”) then the minimizer X is unique. But then since $XS = \nu I$ we have that $S = \nu X^{-1}$ is also unique, and the linear independence property together with $\mathcal{A}^*y = C - S$ shows that the y is also unique. Hence, the (y, S) satisfying $CPE(\nu)$ also yield the solution of the dual problem $BD(\nu)$.

Finally, note that if the primal (0.2) is not strictly feasible, there is no solution of $BP(\nu)$ yielding a finite value of the objective. Hence there can be no solution of $BD(\nu)$ that satisfies the necessary conditions (otherwise these conditions $CPE(\nu)$ would yield a solution to $BP(\nu)$). Similar reasoning applies to the dual.

□

This theorem establishes that existence of unique solutions to $BP(\nu)$ and $CPE(\nu)$ for each $\nu > 0$. For algorithmic purposes, we would like to know that there is a “central path” of solutions that algorithms can follow to a solution. It would be enough for these purposes to show that the equations defining $CPE(\nu)$ are differentiable, with a derivative that is “square” and “nonsingular.” More precisely, the derivative is an operator whose range space is the same as its domain, and is invertible. (This is true for LP, where the central path equations have both range and domain \mathbf{R}^{2n+m} .) Unfortunately, it is *not* true for $CPE(\nu)$! The domain of these equations is $S\mathbf{R}^{n \times n} \times \mathbf{R}^m \times S\mathbf{R}^{n \times n}$ and their range is $S\mathbf{R}^{n \times n} \times \mathbf{R}^m \times \mathbf{R}^{n \times n}$ — note the last part is not symmetric because XS is not symmetric — this system has “more equations than unknowns.”

However, we can reformulate $CPE(\nu)$ in such a way that its range space is the same as its domain. We simply multiply the last equation by X^{-1} (which is legal,

since any X of interest has to be nonsingular). We then obtain

$$CPE(\nu) : \begin{aligned} \mathcal{A}^*y + S &= C, \quad (S \succ 0), \\ \mathcal{A}X &= b, \quad (X \succ 0), \\ S - \nu X^{-1} &= 0, \end{aligned}$$

for which both range and domain are $S\mathbf{R}^{n \times n} \times \mathbf{R}^m \times S\mathbf{R}^{n \times n}$.

Other possible symmetrizations of the last equation are

$$X - \nu S^{-1} = 0$$

and

$$(XS + SX) - 2\nu I = 0.$$

Lecture 37. (4/29/11; 60 min)

Hence the solutions of $\Phi_P(X, y, S; \nu) = 0$ with $X \succ 0$ and $S \succ 0$ give the central path, where

$$\Phi_P(X, y, S; \nu) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{A}^*y + S - C \\ \mathcal{A}X - b \\ -\nu X^{-1} + S \end{bmatrix}.$$

Its derivative with respect to (X, y, S) can be written (loosely) in matrix form as follows:

$$\Phi'_P(X, y, S; \nu) = \begin{bmatrix} 0 & \mathcal{A}^* & \mathcal{I} \\ \mathcal{A} & 0 & 0 \\ \nu(X^{-1} \odot X^{-1}) & 0 & \mathcal{I} \end{bmatrix},$$

where \mathcal{I} denotes the identity operator on $S\mathbf{R}^{n \times n}$. (Note that the “blocks” of this matrix are actually operators rather than matrices.)

We show nonsingularity of $\Phi'_P(X, y, S; \nu)$ at the solution $(X(\nu), y(\nu), S(\nu))$ by proving a more general result. This result is technical for present purposes, but it turns out to also have significance for algorithms.

LEMMA 0.9. (Todd [1]) *Let \mathcal{E} and \mathcal{F} be two operators that map $S\mathbf{R}^{n \times n}$ to itself, and that \mathcal{E} and \mathcal{F} are both nonsingular with $\mathcal{E}^{-1}\mathcal{F}$ positive definite (but not necessarily self-adjoint (i.e. “symmetric”). Assume that the A_i , $i = 1, 2, \dots, m$ are linearly independent. Then for any $P, R \in S\mathbf{R}^{n \times n}$ and any $q \in \mathbf{R}^m$, the solution to the system*

$$\begin{aligned} \mathcal{A}^*v + W &= P, \\ \mathcal{A}U &= q, \\ \mathcal{E}U + \mathcal{F}W &= R \end{aligned} \tag{0.4}$$

is given uniquely by

$$\begin{aligned} v &= (\mathcal{A}\mathcal{E}^{-1}\mathcal{F}\mathcal{A}^*)^{-1}(q - \mathcal{A}\mathcal{E}^{-1}(R - \mathcal{F}P)), \\ W &= P - \mathcal{A}^*v, \\ U &= \mathcal{E}^{-1}(R - \mathcal{F}W). \end{aligned}$$

Proof. This proof is via simple “block elimination” process on the operators. Note first that the expressions for W and U follow immediately from the first and third equations. Now substituting form W in the formula for U and inserting in the second equation, we obtain

$$(\mathcal{A}\mathcal{E}^{-1}\mathcal{F}\mathcal{A}^*)v = q - \mathcal{A}\mathcal{E}^{-1}(R - \mathcal{F}P).$$

Since $\mathcal{E}^{-1}\mathcal{F}$ is positive definite and since the A_i are linearly independent, the $m \times m$ coefficient matrix on the left is positive definite (easy to show) but not necessarily symmetric. Hence, it is nonsingular, so we can invert it to obtain a unique solution v . Since v is uniquely specified by this process, so are W and U when they are recovered by substitution. \square

Work through the details of solving the Newton equations at each iteration of a primal-dual path-following method, for the system $\Phi_P(X, y, S; \nu) = 0$ defined above. The analysis was quite similar to the following:

Newton/SQP steps for $\text{BP}(\nu)$ take the form of Newton steps for the system (0.8). Note first that this system is a suitable system for Newton's method because its domain and range space are both $\mathcal{S}\mathbf{R}^{n \times n} \times \mathbf{R}^m$. The Newton equations are as follows:

$$\begin{aligned} -\mathcal{A}^*(\Delta y) + \nu f''(X)(\Delta X) &= \mathcal{A}^*y - C - \nu f'(X), \\ \mathcal{A}(\Delta X) &= -\mathcal{A}X + b, \end{aligned} \quad (0.5)$$

which by using $f'(X) = -X^{-1}$ and $f''(X) = X^{-1} \odot X^{-1}$ becomes

$$\begin{aligned} -\mathcal{A}^*(\Delta y) + \nu(X^{-1} \odot X^{-1})(\Delta X) &= \mathcal{A}^*y - C + \nu X^{-1}, \\ \mathcal{A}(\Delta X) &= -\mathcal{A}X + b, \end{aligned}$$

By considering only *feasible* X , we have that the right-hand side of the second equation is zero. By substituting the definition of \odot , and replacing \mathcal{A} and \mathcal{A}^* by their definitions, we obtain

$$\begin{aligned} -\sum(\Delta y_i)A_i + \nu X^{-1}(\Delta X)X^{-1} &= \sum y_i A_i - C + \nu X^{-1}, \\ A_j \bullet \Delta X &= 0, \quad j = 1, 2, \dots, m. \end{aligned}$$

Multiplying the first equation from left and right by X , and by the scalar ν^{-1} , we obtain

$$-\nu^{-1} \sum(\Delta y_i)X A_i X + (\Delta X) = \nu^{-1} \sum y_i X A_i X - \nu^{-1} X C X + X.$$

Now substituting for ΔX in the second equation, we have

$$A_j \bullet \left[\nu^{-1} \sum(\Delta y_i)X A_i X \right] = -A_j \bullet \left[\nu^{-1} \sum y_i X A_i X - \nu^{-1} X C X + X \right], \quad j = 1, 2, \dots, m. \blacksquare$$

This is a system of m equations in m unknowns (the elements of Δy). Some elementary manipulations yield

$$\nu^{-1} \sum_{i=1}^m A_j \bullet (X A_i X)(\Delta y_i) = -A_j \bullet \left[\nu^{-1} X \left[\sum y_i A_i - C \right] X + X \right], \quad j = 1, 2, \dots, m.$$

Hence, the coefficient matrix of this system is the $m \times m$ matrix M whose (i, j) element is $A_j \bullet (X A_i X)$. Having solved this system for Δy , we can recover ΔX from one of the equations above.

Summarize the complexity of this process as a function of n and m , assuming dense data. (Note that even if data is sparse, the matrices fill in during the process of assembly, so the complexity estimates apply to many practical problems.) Assembly of coefficient matrix for the $m \times m$ system for Δy is the most expensive part. Need m matrix multiplications $A_i X$ at a cost of $O(n^3)$ each: total cost $O(mn^3)$. Need to do about $m^2/2$ bullet products, each costing $O(n^2)$: total cost $O(m^2n^2)$.

Different symmetrizations proposed for the final equation in the primal-dual formulation:

One of the first symmetrizations proposed, and still one of the most computationally successful, is that of Alizadeh-Haeberly-Overton, which is to replace $XS - \nu I$ by

$$\frac{1}{2}(XS + SX) = \nu I.$$

After linearization we have

$$\frac{1}{2}(\Delta XS + S\Delta X + X\Delta S + \Delta SX) = \nu I - \frac{1}{2}(XS + SX). \quad (0.6)$$

This corresponds to choosing

$$\mathcal{E} = S \odot I, \quad \mathcal{F} = X \odot I$$

in Lemma 0.9. We obtain the step by adding the feasibility conditions to (0.6) from (0.4), that is,

$$\mathcal{A}^* \Delta y + \Delta S = -(\mathcal{A}^* y + S - C), \quad \mathcal{A} \Delta X = -\mathcal{A} X + b.$$

However we cannot use the reduction in the proof of Lemma 0.9 to get the solution, as \mathcal{E} does not have an explicit inverse. We have to solve a Lyapunov equation to invert \mathcal{E} : if $V = \mathcal{E}U = \frac{1}{2}(SU + US)$, compute $\mathcal{E}^{-1}V$ by solving this system for U .

Other primal-dual directions can be derived by using the general symmetrization framework of Monteiro-Zhang, in which we replace $XS = \nu I$ by

$$\frac{1}{2}(PXS P^{-1} + P^{-T}SXP^T) - \nu I = 0$$

for a nonsingular matrix P . This approach can be motivated by noting that if we apply the following transformations:

$$\hat{X} = PXP^T, \quad \hat{S} = P^{-T}SP^{-1},$$

then the MZ directions are simply the AHO directions in the transformed space. The first two equations in $\text{CPE}(\nu)$ change as follows:

$$\sum y_i A_i + S = C \Leftrightarrow \sum y_i P^{-T} A_i P^{-1} + \hat{S} = P^{-T} C P^{-1},$$

so defining $\hat{A}_i = P^{-T} A_i P^{-1}$ and $\hat{C} = P^{-T} C P^{-1}$ we get an equation of the original form. For the other equation $A_i \bullet X = b_i$, we have

$$\begin{aligned} A_i \bullet X &= A_i \bullet P^{-1} \hat{X} P^{-T} = \text{tr}(A_i^T P^{-1} \hat{X} P^{-T}) \\ &= \text{tr}(P^{-T} A_i^T P^{-1} \hat{X}) = \text{tr}((P^{-T} A_i P^{-1})^T \hat{X}) = (P^{-T} A_i P^{-1}) \bullet \hat{X} = \hat{A}_i \bullet \hat{X}, \end{aligned}$$

so again we get an equation of the original form in the transformed space. For the final equation we have

$$\begin{aligned} 0 &= \frac{1}{2}(PXS P^{-1} + P^{-T}SXP^T) - \nu I \\ &= \frac{1}{2}(PXP^T P^{-T}SP^{-1} + P^{-T}SP^{-1}PXP^T) - \nu I = \frac{1}{2}(\hat{X}\hat{S} + \hat{S}\hat{X}) - \nu I. \end{aligned}$$

Three interesting (and practical) choices for P arise from the motivation of making \hat{X} and \hat{S} commute, i.e. $\hat{X}\hat{S} = \hat{S}\hat{X}$. They are:

- $P = S^{1/2}$ (HRVW/KSH/M) yields $\hat{S} = I$;
- $P = X^{-1/2}$ (dual HRVW/KSH/M) yields $\hat{X} = I$;
- $P = W^{-1/2}$, where $W = X^{1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}$. Here S is the unique positive definite matrix that ensures that $WSW = X$ so that $\hat{X} = \hat{S}$.

These three cases correspond to the following choices of \mathcal{E} and \mathcal{F} in the system (0.4):

- $\mathcal{E} = \mathcal{I}, \mathcal{F} = X \odot S^{-1}$;
- $\mathcal{E} = S \odot X^{-1}, \mathcal{F} = \mathcal{I}$;
- $\mathcal{E} = \mathcal{I}, \mathcal{F} = W \odot W$.

Lecture 38. (5/2/11; 60 min)

In $\Phi'_P(X, y, S; \nu)$ above, we have $\mathcal{E} = \nu(X^{-1} \odot X^{-1})$ and \mathcal{F} is the identity. (It is easy to verify that the assumptions hold.) Hence the theorem above shows that $\Phi'_P(X, y, S; \nu)$ is nonsingular along the central path (and indeed for all $(X, y, S) \in F^0(P) \times F^0(D)$).

THEOREM 0.10. *Assume that both $F^0(P)$ and $F^0(D)$ are nonempty and that the linear independence condition holds. Then the set of solutions to $CPE(\nu)$ forms a nonempty differentiable path called the central path. Moreover we have $X(\nu) \in F^0(P)$ and $(y(\nu), S(\nu)) \in F^0(D)$ and the duality gap is*

$$C \bullet X(\nu) - b^T y(\nu) = X(\nu) \bullet S(\nu) = n\nu.$$

An immediate consequence of this result is that there is no duality gap when the stated conditions hold. If we could show that $(X(\nu), y(\nu), S(\nu))$ has a limit (X^*, y^*, S^*) as $\nu \downarrow 0$, we could conclude also that X^* solves (0.2) and (y^*, S^*) solves (0.3). In fact this is true, but hard to prove.

Now we return to the **strong duality** results. We focus on the following result, which does not require us to show that the central path has a limit.

THEOREM 0.11. *If $F^0(P)$ and $F^0(D)$ are nonempty and the linear independence condition holds, then both the primal and dual SDP have bounded nonempty solution sets and the duality gap is zero.*

Proof. We have already noted the zero duality gap. We now show that the solution set for the primal is nonempty and bounded, using similar techniques as in the proof of Theorem 0.8. Let $\hat{X} \in F^0(P)$ and $(\hat{y}, \hat{S}) \in F^0(D)$. Similarly to the earlier proof, we can replace the primal problem by the following without changing its solution set:

$$\min_{X \in \mathbf{R}^{n \times n}} \hat{S} \bullet X \quad \text{s.t.} \quad A_i \bullet X = b_i, \quad i = 1, 2, \dots, m, \quad \hat{S} \bullet X \leq \hat{S} \bullet \hat{X}, \quad X \succeq 0. \quad (0.7)$$

Letting σ denote the smallest eigenvalue of \hat{S} (as before) it is easy to show that the added constraint $\hat{S} \bullet X \leq \hat{S} \bullet \hat{X}$ implies that the eigenvalues of X must be bounded by $\hat{S} \bullet \hat{X} / \sigma$. Hence, the feasible set for (0.7) is bounded. It is also nonempty because it contains \hat{X} . Finally it is easy to see that the objective of (0.7) is continuous over the feasible set. Hence, the problem attains its minimum, so the solution set of (0.7) (and hence (0.2)) is nonempty and bounded. \square

0.3. Algorithms for SDP. We discuss path-following algorithms for SDP, i.e. algorithms that attempt to follow the primal-dual central path defined by $CPE(\nu)$, or the primal or dual central paths defined by the minimizers of $BP(\nu)$ or $BD(\nu)$.

In a primal method, the basic idea, as described in our introduction to log barrier methods, is that we find an approximate solution of $BP(\nu)$ by taking Newton-SQP steps, terminating when some approximate minimization condition is satisfied, then reduce ν and repeat. We discuss the details here, including the details of step calculation, and sketch the theory.

The primal barrier method works with the subproblem $BP(\nu)$ defined (as earlier) by

$$BP(\nu) : \quad \min_{X \in \mathbf{R}^{n \times n}} C \bullet X + \nu f(X), \quad A_i \bullet X = b_i, \quad i = 1, 2, \dots, m. \quad (X \succ 0)$$

As shown in Theorem 0.8, the solution to $\text{BP}(\nu)$ exists for any $\nu > 0$ and satisfies $\text{CPE}(\nu)$, which we can write as follows (after eliminating S , and noting that $f'(X) = -X^{-1}$):

$$\begin{aligned} C + \nu f'(X) - \mathcal{A}^*y &= C - \nu X^{-1} - \mathcal{A}^*y = 0, \\ \mathcal{A}X - b &= 0, \quad (X \succ 0). \end{aligned} \quad (0.8)$$

Given that we know how to compute a Newton/SQP step for $\text{BP}(\nu)$, we now decide how to stop the iterations at an approximate solution. Since we deal only with feasible X , the only issue is ensuring that the first optimality condition is “nearly” satisfied, that is, $C - \nu X^{-1} - \mathcal{A}^*y = 0$. To measure the “size” of the deviation, we use a norm that is weighted by $f''(X)$, the Hessian of the objective (and of the Lagrangian for this problem). This norm turns out to be the most appropriate from the point of view of interior-point theory. We define the “ X -norm” of a symmetric matrix H as follows:

$$\begin{aligned} \|H\|_X &\stackrel{\text{def}}{=} ([f''(X)H] \bullet H)^{1/2} \\ &= ([X^{-1} \odot X^{-1}]H \bullet H)^{1/2} = ([X^{-1}HX^{-1}] \bullet H)^{1/2} = \text{trace}(X^{-1}HX^{-1}H)^{1/2} \\ &= \text{trace}(X^{-1/2}HX^{-1/2}X^{-1/2}HX^{-1/2})^{1/2} = \|X^{-1/2}HX^{-1/2}\|_F. \end{aligned}$$

We define a “dual X -norm” by making use of the inverse of the operator $f''(X)$. This inverse is simple to derive. If J and K satisfy $f''(X)J = K$, we have $X^{-1}JX^{-1} = K$, so that $J = XKX$. Therefore we have $K = [f''(X)]^{-1}J$ by defining $[f''(X)]^{-1}$ to be the operator $X \odot X$. The dual X -norm is defined as

$$\begin{aligned} \|H\|_X^* &\stackrel{\text{def}}{=} ([f''(X)]^{-1}H \bullet H)^{1/2} \\ &= ([X \odot X]H \bullet H)^{1/2} = ([XHX] \bullet H)^{1/2} = \text{trace}(XHXH)^{1/2} \\ &= \text{trace}(X^{1/2}HX^{1/2}X^{1/2}HX^{1/2})^{1/2} = \|X^{1/2}HX^{1/2}\|_F. \end{aligned}$$

Returning to the expression $C - \nu X^{-1} - \mathcal{A}^*y$; this “lives” in dual space, the same space inhabited by S . (It so happens here that both this space and the primal space are $S\mathbf{R}^{n \times n}$.) Hence, we measure its size by the dual X -norm. Our condition for accepting the approximate minimum is

$$\|C - \nu X^{-1} - \mathcal{A}^*y\|_X^* \leq \nu\tau, \quad (0.9)$$

where $\tau \in (0, 1)$ is some constant. We discuss the consequences of this choice in a moment.

A framework for a primal path-following method is as follows:

- Choose a strictly feasible X for (0.2) and values $y \in \mathbf{R}^m$ and $\nu > 0$.
- Perform Newton/SQP steps, possibly with damping or other globalization strategies, until (0.9) is satisfied.
- Stop if ν is sufficiently small (less than some tolerance).
- Otherwise, replace ν by $\theta\nu$ for some $\theta \in (0, 1)$, and continue.

Defining $S = C - \mathcal{A}^*y$, the bound (0.9) becomes $\|\nu^{-1}S - X^{-1}\|_X^* \leq \tau$, that is, $\|\nu^{-1}S - X^{-1}\|_{X^{-1}} \leq \tau$. Hence we have

$$\|\nu^{-1}S - X^{-1}\|_{X^{-1}} = \|\nu^{-1}X^{1/2}SX^{1/2} - I\|_F \leq \tau.$$

Since for symmetric A we have $\|A\|_F^2 = A \bullet A = \text{trace}(A^T A) = \sum \lambda_i(A^T A) = \sum \lambda_i^2(A)$, we have that all the eigenvalues of $\nu^{-1}X^{1/2}SX^{1/2} - I$ are less than 1; that

is, $\nu^{-1}X^{1/2}SX^{1/2}$ is positive definite, which implies that S is positive definite. Hence (y, S) is strictly feasible for the dual SDP, so we can derive a strictly feasible solution for dual SDP from each iterate of the barrier method for $\text{BP}(\nu)$. Another nice feature is that iterates satisfying (0.9) satisfy a guaranteed small duality gap. We have

$$X \bullet S = X \bullet (\nu X^1 + [S - \nu X^{-1}]) \leq \nu n + \|X\|_X \|S - \nu X^{-1}\|_X^* \leq \nu(n + \tau\sqrt{n}).$$

Hence, a small ν and satisfaction of the condition (0.9) ensures that we are close to a primal-dual solution of the SDP.

Convergence of a *short-step* variant of this approach (i.e. one that does not need damping) has been proved by Nesterov and Nemirovskii.

THEOREM 0.12. *Suppose that the initial X , y , and ν are chosen such that the condition (0.9) is satisfied. Then if we apply the procedure above with $\theta = 1 - .1/\sqrt{n}$, the criterion (0.9) will be satisfied after a single undamped Newton step, and ν will be reduced to a fraction ϵ of its original value after $O(\sqrt{n} \ln(1/\epsilon))$ iterations.*

Lecture 39. (5/4/11; 60 min)

We omit the proof, except to note that the complexity argument is the same as that used for LP methods. Denoting the value of ν after k iterations as ν_k , we have

$$\nu_k = \left(1 - \frac{.1}{\sqrt{n}}\right)^k \nu_0$$

Hence a sufficient condition for $\nu_k \leq \epsilon \nu_0$ is that

$$\left(1 - \frac{.1}{\sqrt{n}}\right)^k \leq \epsilon.$$

Taking the \ln of both sides and using $\ln(1 + \beta) \leq \beta$ for $\beta > -1$, we have

$$k \ln \left(1 - \frac{.1}{\sqrt{n}}\right) \leq \ln \epsilon,$$

which is true if

$$-k \frac{.1}{\sqrt{n}} \leq \ln \epsilon$$

which is true if

$$k \geq -10\sqrt{n}(\ln \epsilon) = 10\sqrt{n} \ln(1/\epsilon).$$

Proof that $K \bullet H \leq \|K\|_X \|H\|_X^* = \|X^{-1/2} K X^{-1/2}\|_F \|X^{1/2} H X^{1/2}\|_F$. First show that

$$K \bullet H = (X^{-1/2} K X^{-1/2}) \bullet (X^{1/2} H X^{1/2})$$

by using the definition of \bullet and its relationship to traces and sums of eigenvalues. Then use Cauchy-Schwarz inequality to show

$$U \bullet V = \sum_{i,j} U_{ij} V_{ij} \leq \left(\sum_{i,j} U_{ij}^2\right)^{1/2} \left(\sum_{i,j} V_{ij}^2\right)^{1/2} = \|U\|_F \|V\|_F.$$

Dual Path-Following. Note that the primal path-following approach method generally yields dense matrices X even if all data matrices C and A_i are sparse. Hence, it can be expensive to implement when n is large. We can use a dual path-following approach instead to get around this problem. This approach works with feasible matrices S , which because of the constraint

$$S = C - \sum_{i=1}^m y_i A_i$$

have a sparsity pattern that is the unification of that of the C and A_i 's. The development is analogous to the primal path-following approach. We work with $\text{BD}(\nu)$ instead of $\text{BP}(\nu)$, eliminating X from the $\text{CPE}(\nu)$ equations instead of S . We end up with a system involving ΔS and Δy , and after elimination we obtain the following $m \times m$ reduced system involving Δy :

$$M \Delta y = b - \nu A S^{-1},$$

where $M_{ij} = \nu A_i \bullet (S^{-1} A_j S^{-1})$. The proximity criterion for terminating the Newton iterates on $\text{BD}(\nu)$ is that $\|\Delta S\|_S \leq \tau$. A similar convergence/complexity result holds.

Primal-Dual Path-Following. Primal-dual methods work with variants of the $\text{CPE}(\nu)$ equations directly and generate iterates (X, y, S) . As discussed earlier, we cannot work with the original form of $\text{CPE}(\nu)$ because this system is “not square.” The technique used above, i.e. multiplying the last equation by X^{-1} to give $S - \nu X^{-1}$ leads to poor conditioning when X becomes nearly singular—the usual complaints about primal log barrier apply. There are other techniques for symmetrizing the condition $XS = \nu I$ that lead to nicer equations (even when X and S are nearly singular), so that a modified Newton’s method can be applied to the $\text{CPE}(\nu)$ equations to yield a primal-dual path-following method. These symmetrization techniques can be expressed in terms of the operators \mathcal{E} and \mathcal{F} introduced in Lemma 0.9 above.

Lecture ∞ . (possibly include later.)

Degeneracy. What does it mean to be a *degenerate* nonlinear program? Not well defined but usually refers to one that does *not* satisfy these conditions at x^* : LICQ, strict complementarity, second-order sufficient. These are minimal conditions under which local convergence should be provable. They should be able to identify active constraints locally, converge locally (often superlinearly), etc. Many algorithms get local convergence (possibly superlinear) under weaker conditions than this.

More Geometry. *Subspace* is a set $S \subset \mathbb{R}^n$ with the properties that $0 \in S$, $u \in S$ and $v \in S$ imply that $u+v \in S$ (closed under addition), and $u \in S$ implies that $\alpha u \in S$ for all α (closed under scalar multiplication).

Spanning set of a subspace: set of vectors $\{x_1, x_2, \dots, x_m\}$ in S such that any vector $v \in S$ can be expressed as a linear combination of the vectors in this set. Define *linear independence*. A *basis* is a linearly independent spanning set. Number of vectors m in a basis is called the *dimension* of S .

Affine set is a subset of \mathbb{R}^n with the form $x+S$, for some $x \in \mathbb{R}^n$ and some subspace S . Basically a shifted subspace.

Affine Hull of a set $C \subset \mathbb{R}^n$ is “the intersection of all affine sets containing C .” Can get it by picking a vector $x \in C$ then computing a basis of $C - x$, and defining the subspace S to be the span of this basis.

Given $C \in \mathbb{R}^n$, the *relative interior* is the interior of C relative to $\text{aff}(C)$. That is, $x \in \text{ri}(C)$ if we can choose a neighborhood U of x such that $U \cap \text{aff}(C) \subset C$. Draw some pictures, e.g. a 2d disc in 3d space.

Distance operator: $d_\Omega(x) = \inf_{z \in \Omega} \|z - x\|$. Clearly by definition, when Ω is closed too, we have $d_\Omega(x) = \|x - P(x)\|$.

Properties of $d_\Omega(x)$:

- Continuous? Yes (obvious that a small change in x produces only a small change in $d_\Omega(x)$).
- Smooth? No. Draw a picture to show that it's not differentiable.
- Convex? Yes, when Ω is convex. Prove this in the case of Ω also closed, and use the result below, that the minimizer is attained. Let $s_x \in \Omega$ be such that $d_\Omega(x) = \|x - s_x\|$ and s_y be such that $d_\Omega(y) = \|y - s_y\|$. Note that for any $\lambda \in [0, 1]$, $s_\lambda = \lambda s_x + (1 - \lambda)s_y \in \Omega$ by convexity. Now have

$$\begin{aligned} d_\Omega(\lambda x + (1 - \lambda)y) &\leq \|\lambda x + (1 - \lambda)y - s_\lambda\| \\ &\leq \lambda \|x - s_x\| + (1 - \lambda) \|y - s_y\| \\ &= \lambda d_\Omega(x) + (1 - \lambda) d_\Omega(y). \end{aligned}$$

- Draw a picture to show that $d_\Omega(\cdot)$ is not necessarily convex when Ω is non-convex.

Consider $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$. Define continuous function. Lower semicontinuous function definition: For all $x \in X$ we have $f(x) \leq \liminf f(x_k)$ for all sequences $\{x_k\} \subset X$ with $x_k \rightarrow x$.

Extended value function: $f : X \rightarrow [-\infty, \infty]$ (includes infinite values). Effective domain of a function is subset of f where f has values less than ∞ .

Epi graph of a function: $\{(x, w) \mid x \in X, f(x) \leq w\}$. Pictures.

Proper function if $f(x) < \infty$ for some $x \in X$ and $f(x) > -\infty$ for all $x \in X$. Epi graph does not contain a vertical line. Example of non-proper: $f(x) = -1/x$ for $x > 0$, $f(x) = -\infty$ for $x = 0$ is not proper on domain $X = [0, \infty)$.

f is *closed* if $\text{epi}(f)$ is a closed set.

Example of non-closed function: $f(x) = 0$ for $x < 0$ and $f(x) = 1$ for $x \geq 0$.

Curiosity: Linear transformation of a closed convex set is convex but not necessarily closed. Example:

$$C = \{(x_1, x_2) \mid x_1 > 0, x_2 > 0, x_1 x_2 \geq 1\}.$$

However closed polyhedral sets have closed projections.

Convex Hull: Define.

discussion of nonconvex Ω e.g. Rockafellar and Wets Chapter 6.

REFERENCES

- [1] M. J. TODD, *Semidefinite optimization*, Acta Numerica, 10 (2001), pp. 515–560.