Convex Analysis Basics. We work mostly in $\mathbf{R}^n$ and sometimes in $S\mathbf{R}^{n\times n}$. Recall vector notation (subscripts for components), inner products, Euclidean norms, sequences, subsequences, limits, accumulation points.

Open and closed sets. Set of positive semidefinite matrices is closed in $S\mathbf{R}^{n\times n}$, positive definite matrices is open.

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 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 cones. "Pointed" cone has no vectors $x \neq 0$ such that x and $-x$ are both in C (i.e. $\{0\}$ is the only subspace in C.) We're particularly interested in closed convex cones.

Positive definite and positive semidefinite matrices are cones in $S\mathbf{R}^{n\times n}$.

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

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

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

$$N_\Omega(x) := \{v \,|\, \langle v, y - x\rangle \le 0 \;\text{ for all } y \in \Omega\}. \tag{0.1}$$

Draw pictures.

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.

Normal cone is outer semicontinuous: If $\{x_k\}$ and $\{v_k\}$ are sequences in $\mathbf{R}^n$ such that $x_k \in \Omega$, $v_k \in N_\Omega(x_k)$, $x_k \to x$, and $v_k \to v$, then $v \in N_\Omega(x)$.

Lecture 2. (1/23/15; 60 min)

Tangent cone to a closed set Ω (not necessarily convex). Define limiting directions. Given $x \in \Omega$, y is a limiting direction if there exist sequence $y_k \to y$, $\alpha_k \downarrow 0$ such that $x + \alpha_k y_k \in S$ for all k. Tangent cone $T_\Omega(x)$ is set of all limiting directions to Ω at x. That is,

$$T_\Omega(x) := \{u \,|\, \text{there exist sequences } u_k \to u \text{ and } \alpha_k \downarrow 0 \text{ such that } x + \alpha_k u_k \in \Omega\}. \tag{0.2}$$

The tangent cone $T_\Omega(x)$ to a closed convex Ω is closed and convex (prove!). But for Ω not convex, $T_\Omega(x)$ may not be convex! Example: the set $\Omega = \{x \in \mathbf{R}^2 \;:\; x_2 = \pm x_1\}$. At $x = 0$, we have $T_\Omega(0) = \Omega$.

Having defined T_Ω as the set of limiting feasible directions, we can define the normal cone $N_\Omega(x)$ as its polar:

$$N_\Omega(x) = T_\Omega(x)^\circ = \{v \mid \langle v, u\rangle \le 0 \text{ for all } u \in T_\Omega(x)\}.$$

When Ω is convex, this definition of $N_\Omega(x)$ coincides with the definition (0.1) above. Otherwise, it may not. (Draw examples of nonconvex sets Ω where in fact $\langle z - x, v\rangle \ge 0$ for $v \in N_\Omega(x)$ and all $z \in \Omega$! The exact opposite of the definition (0.1).)

Examples of Normal Cones. Set $\Omega = \mathbb{R}^n_+$. $N_\Omega(0) = -\mathbb{R}^n_+$ because we need $y^T e_i \leq 0$ for $i = 1, 2, \ldots, n$, so $y \leq 0$.

Hyperplane: $\Omega := \{x \,|\, p^T x \leq \alpha\}$. Have $N_\Omega(x^*) = \{\beta p \,|\, \beta \geq 0\}$ at any $x^* \in \Omega$.

Disk: $\Omega := \{x \,|\, \|x\|_2 \leq 1\}$. $N_\Omega(x^*) = \{\beta x^* \,|\, \beta \geq 0\}$ for any x^* with $\|x^*\| = 1$.

Second-order cone:

$$\Omega := \{(x, y, z) \,|\, z \geq \sqrt{x^2 + y^2}\}.$$

Have $N_\Omega(0) = -\Omega = \{(u, v, w) \,|\, w \leq -\sqrt{u^2 + v^2}\}$. Partial verification: Need (u, v, w) such that $ux + vy + wz \leq 0$. Have

$$ux + vy + wz \leq wz + \sqrt{u^2 + v^2}\sqrt{x^2 + y^2} \leq wz + |w||z| = 0,$$

since $w \leq 0$ and $z \geq 0$.

All these examples are for Ω convex. We also did several nonconvex examples in class.

Projection P onto closed convex set X:

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

Lemma 0.1.
(i) $(P(y) - z)^T (y - z) \geq 0$ *for all* $z \in X$*, with equality if and only if* $z = P(y)$*.*
(ii) $(y - P(y))^T (z - P(y)) \leq 0$ *for all* $z \in X$*.*

Proof. We prove (i) and leave (ii) as an exercise.

Let z be an arbitrary vector in X. We have

$$\begin{aligned}\|P(y) - y\|_2^2 &= \|P(y) - z + z - y\|_2^2 \\ &= \|P(y) - z\|_2^2 + 2(P(y) - z)^T (z - y) + \|z - y\|_2^2\end{aligned}$$

which implies by rearrangement that

$$2(P(y) - z)^T (y - z) = \|P(y) - z\|_2^2 + \left[\|z - y\|_2^2 - \|P(y) - y\|_2^2\right]. \tag{0.3}$$

The term in [] is nonnegative, from the definition of P. The first term on the right-hand side is trivially nonnegative, so the nonnegativity claim is proved.

If $z = P(y)$, we obviously have $(P(y) - z)^T (y - z) = 0$. If the latter condition holds, then the first term on the right-hand side of (0.3) in particular is zero, so we have $z = P(y)$. □

(ii) implies that $y - P(y) \in N_X(P(y))$ for closed convex X.

(ii) also implies that the projection onto a closed convex X is uniquely defined. (Proved in class by showing that if there are two possible values $P(y)$ and $\bar{P}(y)$ for he projection of y, they must be the same.)

Lecture 3. (1/26/15; 70 min)

Using the result above, we can prove the following: *If C is a closed convex cone, then $C^{\circ\circ} = C$.* Proof: Show $C \subset C^{\circ\circ}$: If $v \in C$ then $\langle v, u\rangle \leq 0$ for all $u \in C^\circ$ so that $v \in C^{\circ\circ}$. Show $C^{\circ\circ} \subset C$: Let $v \in C^{\circ\circ}$ and let $P(v)$ be projection onto C. We have by earlier result that

$$\langle v - P(v), z - P(v)\rangle \leq 0 \quad \text{for all } z \in C$$

Setting $z = 0$ we have $\langle v - P(v), -P(v)\rangle \le 0$ and setting $z = 2P(v)$ we have $\langle v - P(v), P(v)\rangle \le 0$, so $\langle v - P(v), P(v)\rangle = 0$. Thus by substituting into the inequality above we have $\langle v - P(v), z\rangle \le 0$ for all $z \in C$. Thus $v - P(v) \in C^\circ$, and so because $v \in C^{\circ\circ}$, we have $\langle v - P(v), v\rangle \le 0$. Using again that $\langle v - P(v), P(v)\rangle = 0$, we deduce that $\langle v - P(v), v - P(v)\rangle \le 0$, which implies $v = P(v)$. Thus $v \in C$, as required.

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

$$(x_2 - P(x_2))^T(P(x_1) - P(x_2)) \le 0, \qquad (x_1 - P(x_1))^T(P(x_2) - P(x_1)) \le 0.$$

Adding, we obtain

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

which after rearrangement gives

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

proving the claim.

Polarity of Tangent and Normal Cones.

For tangent defined by (0.2) and normal defined by $N_\Omega(x) = T_\Omega(x)^\circ$, we clearly have a polarity relationship *by definition*, even if Ω is not convex.

When $T_\Omega(x)$ is also a *closed convex cone*, then by the result above, we have $T_\Omega(x)^{\circ\circ} = T_\Omega(x)$, so we have *in addition* that $N_\Omega(x)^\circ = T_\Omega(x)$.

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.

Consider first standard form:

$$(P) \qquad \min_x c^Tx \ \text{ s.t. } Ax \ge b, \ x \ge 0.$$

$$(D) \qquad \max_u b^Tu \ \text{ s.t. } A^Tu \le c, \ u \ge 0.$$

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.

Same applies for the primal-dual pair:

$$(P) \qquad \min_x c^Tx \ \text{ s.t. } Ax \ge b.$$

$$(D) \qquad \max_u b^Tu \ \text{ s.t. } A^Tu = c, \ u \ge 0.$$

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 $b \in \mathbf{R}^n$, 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

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

whose dual is

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

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

Lecture 4. (1/28/15; 60 min)

Review and clarify the earlier stuff on tangent and normal cones, and ask some questions.

- Definition (0.2) and $N_\Omega(x) := T_\Omega(x)^\circ$. These definitions do not require convexity of Ω.
- Is $T_\Omega(x)$ convex, even if Ω is nonconvex? No, there are counterexamples.
- Is $N_\Omega(x)$ convex, even if Ω is nonconvex? Yes.
- Prove that $T_\Omega(x)$ is convex, when Ω is convex.
- Prove that $T_\Omega(x)$ is closed (even when Ω is nonconvex).
- Prove that $N_\Omega(x)$ is closed. (It is the polar of a cone, so is closed by appeal to this general result.)
- When Ω is closed and convex, the definition $N_\Omega(x) := T_\Omega(x)^\circ$ is equivalent to the definition (0.1).

Review definitions of local, global, strict local, isolated solutions from p. 305-306 of NW.

Lecture 5. (1/30/15; 60 min)

Show that "local solution" is equivalent to there being *no* sequence $\{z_k\} \subset \Omega$ with $z_k \to x^*$ and $f(z_k) < f(z^*)$. *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 $z_k \to x^*$ 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.)

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 $t \in (0, 1)$.

THEOREM 0.2. *Consider* $\min_{x \in \Omega} f(x)$, *where* f *is continuously differentiable and* Ω *is closed. If* x^* *is a local solution, then*

$$-\nabla f(x^*) \in N_\Omega(x^*). \tag{0.4}$$

Proof. Suppose that $-\nabla f(x^*) \notin N_\Omega(x^*)$. Then there is some $v \in T_\Omega(x^*)$ such that $\langle -\nabla f(x^*), v \rangle > 0$. Since v is a limiting feasible direction, there exist sequences $v_k \to v$ and $\alpha_k \downarrow 0$ such that $x^* + \alpha_k v_k \in \Omega$. We have by Taylor's theorem that

$$f(x^* + \alpha_k v_k) = f(x^*) + \alpha_k \langle \nabla f(x^*), v_k \rangle + o(\alpha_k) < f(x^*),$$

since $\langle \nabla f(x^*), v_k \rangle \leq (1/2) \langle \nabla f(x^*), v \rangle < 0$ for all k sufficiently large, so the α_k term dominates the $o(\alpha_k)$ term. Hence x^* is not a local solution. □

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

More on Theorems of the Alternative.. 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:

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

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

where $e = (1, 1, \ldots, 1)^T$ 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 $Ax > 0$ $\Leftrightarrow$ $\sim$I.

Separation Theorem. Preview the two main results below: separation and strict separation.

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 *every* finite collection of points in X, call them

$\{x_1, x_2, \ldots, x_m\}$, we have $\cap_{i=1}^m \Lambda_{x_i}$ nonempty, then $\cap_{x \in X} \Lambda_x$ is also nonempty. Prove by using the property that the complement of each Λ_x is open in Λ. If $\cap_{x \in X} \Lambda_x = \emptyset$, then the complements form an open cover of Λ. So by the Heine-Borel Theorem, there is a finite cover, that is, a finite set of points $\{x_1, x_2, \ldots, x_m\}$, such that the complements of Λ_{x_i} cover Λ, so that the intersection of the Λ_{x_i} is empty. C!

Lecture 6. (2/2/15; 60 min)

Separation Theorem: Let X be any convex set in $\mathbf{R}^n$ not containing the origin. Then there is a vector $\bar{t} \in \mathbf{R}^n$ 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 $\{x_1, x_2, \ldots, x_m\}$ be any finite set of vectors in X. Since $0 \notin X$, there cannot be a vector $p \in \mathbf{R}^m$ such that

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

So there cannot be a vector p such that

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

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, $-t/\|t\|_2 \in \cap_{i=1,2,\ldots,m} \Lambda_{x_i}$, 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 $\cap_{x \in X} \Lambda_x$ 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 $\mathbf{R}^2$ consisting of the closed right-half plane but excluding the half-line $\{(0, x_2) \,|\, x_2 \leq 0\}$.

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

Proof. Using $P()$ 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 separation between two disjoint convex closed sets X and Y, and define

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

First consider the question: Is $X - Y$ convex? closed? Clearly convex by elementary argument. But may not be closed. e.g. consider a sequence $\{z_k\} \subset X - Y$, with $z_k \to z$. 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:

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

Now define $z_k = (0, 2/k)$ for $k = 1, 2, \ldots$. We have $z_k \in X - Y$ (by taking $x_k = (k, 1/k)$ and $y_k = (k, -1/k)$). But $z_k \to (0, 0) \notin X - Y$. Since 0 is in the closure of $X - Y$ but not a member of $X - Y$, we conclude that $X - Y$ is not closed. However we have this result.

THEOREM 0.3. *Let X and Y be two nonempty disjoint closed convex nonempty sets. Then there is $c \in \mathbb{R}^n$ with $c \neq 0$, and $\alpha \in \mathbb{R}$ such that $c^T x - \alpha \le 0$ for all $x \in X$ and $c^T y - \alpha \ge 0$ for all $y \in Y$.*

Proof. By applying separation result to $X - Y$, we have that there exists $c \neq 0$ such that $c^T x \le c^T y$ for all $x \in X$ and $y \in Y$. By choosing an arbitrary $\hat{x} \in X$, we have that $c^T y$ is bounded below by $c^T \hat{x}$ for all $y \in Y$. Hence the infimum of $c^T y$ over $y \in Y$ exists; we denote it by α. Clearly $c^T y \ge \alpha$ for all $y \in Y$. We claim that α also satisfies $c^T x \le \alpha$ for all $x \in X$. If this is not true, there exists some $\bar{x} \in X$ such that $c^T \bar{x} > \alpha$. Since $c^T y \ge c^T \bar{x}$, we have that $\inf_{y \in Y} c^T y \ge c^T \bar{x} > \alpha$, which contradicts the definition of α, so there is no such $\bar{x}$ we conclude that $c^T y - \alpha \ge 0$ for all $y \in Y$ and $c^T x - \alpha \le 0$ for all $x \in X$. □

Strict Separation between Sets. (Not covered in 2015.) 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 $z_k \to z$. 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 strict separation result.

THEOREM 0.4. *Let X and Y be two disjoint closed convex nonempty sets with X compact. Then there is $c \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ 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 $\gamma + \beta \le \delta$. We have

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

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

Tangent and Normal Cones to a Polyhedral Set.. Consider the set defined by linear equalities and inequalities:

$$C := \{x \,|\, Dx \le d, \; Gx = g\}.$$

Given a feasible point x_0, define the active set $\mathcal{A}$ as the set of indices of inequality constraints that are satisfied at equality:

$$\mathcal{A} := \{i \,|\, D_i x = d_i\}.$$

By using the "limiting feasible direction" definition of tangent cone, we can easily verify the following. (We leave it as a homework exercise.)

$$T_C(x_0) = \{z \,|\, Gz = 0, \; D_i z \le 0, \text{ for all } i \in \mathcal{A}\}.$$

We let $D_{\mathcal{A}}$ be the row submatrix of D corresponding to the index set $\mathcal{A}$, and likewise for $d_{\mathcal{A}}$.

Lecture 7. (2/4/15; 60 min)

Farkas's Lemma gives us the following result for the normal cone.

Theorem 0.5.

$$N_C(x_0) = \{D_{\mathcal{A}}^T w + G^T y \,|\, w \geq 0\}.$$

Proof. From the definition of tangent cone, we have that

$$N_C(x_0) = \{v \,|\, v^T u \text{ for all } u \text{ with } D_{\mathcal{A}} u \leq 0 \text{ and } Gu = 0\}.$$

Thus for any $v \in N_C(x_0)$, there is no vector u such that

$$u^T v > 0, \quad D_{\mathcal{A}} u \leq 0, \quad Gu \leq 0, \quad -Gu \leq 0.$$

Hence by Farkas's Lemma, we must have vectors $w \geq 0$, $\hat{y} \geq 0$, $\bar{y} \geq 0$ such that

$$v = D_{\mathcal{A}}^T w + G^T \hat{y} - G^T \bar{y} = D_{\mathcal{A}}^T w + G^T (\hat{y} - \bar{y}).$$

In fact we can write v equivalently as $v = D_{\mathcal{A}}^T w + G^T y$ defining $y = \hat{y} - \bar{y}$, and there is no sign restriction on y. □

We proved about the first-order optimality condition $-\nabla f(x^*) \in N_C(x^*)$. It follows from this that the following condition is necessary for x^* to be optimal: For the set of indices $\mathcal{A}^*$ that are active at x^*, there are vectors $w_{\mathcal{A}^*} \geq 0$ and y such that

$$-\nabla f(x^*) = D_{\mathcal{A}^*}^T w_{\mathcal{A}^*} + G^T y, \quad w_{\mathcal{A}^*} \geq 0.$$

We can write these conditions equivalently without using $\mathcal{A}^*$ as follows:

$$-\nabla f(x^*) = D^T w + G^T y, \quad 0 \leq w \perp Dx^* - d \leq 0.$$

(Note that $w_i = 0$ for $i \notin \mathcal{A}^*$.) These are the **Karush-Kuhn-Tucker (KKT)** conditions for polyhedral sets C and general smooth (but not necessarily convex) objectives.

Conditional Gradient (Frank-Wolfe). Out of favor for a long time, but there has been an upsurge of activity in this area. Originally due to Frank and Wolfe (1956).

Consider the problem

$$\min_{x \in \Omega} f(x), \tag{0.5}$$

where f is smooth and Ω is **compact** and convex.

Get new search direction by minimizing a linear approx to f over Ω, then do a line search in the direction of the resulting point. Because Ω is compact, the minimizer of the linear approximation exists.

Show that the subproblem

$$\min_{z \in \Omega} \nabla f(x_k)^T (z - x_k)$$

has a nonpositive optimal objective (because x_k is a feasible point) and in fact the optimal objective is *negative* unless x_k is stationary.

Discuss sufficient decrease condition. Give Taylor series argument as to why it can be satisfied for $c_1 \in (0, 1)$, for steplength sufficiently small.

A conditional gradient scheme with backtracking is shown as Algorithm 1.

Algorithm 1 Conditional Gradient (backtracking)

```
Given $x_0$, $c_1 \in (0,1)$, $\beta \in (0,1)$, $\bar{\alpha} \in (0,1]$;
for $k = 0, 1, 2, \dots$ do
    $\bar{x}_k := \arg\min_{z \in \Omega} \nabla f(x_k)^T (z - x_k)$;
    $j \leftarrow 0$;
    while $f(x_k + \beta^j \bar{\alpha}(\bar{x}_k - x_k)) > f(x_k) + c_1 \beta^j \bar{\alpha} \nabla f(x_k)^T [\bar{x}_k - x_k]$ do
        $j \leftarrow j + 1$;
    end while
    $\alpha_k \leftarrow \beta^j \bar{\alpha}$;
    $x_{k+1} \leftarrow x_k + \alpha_k (\bar{x}_k - x_k)$;
    if $x_{k+1} = x_k$ then
        STOP;
    end if
end for
```

Lecture 8. (2/6/15; 60 min)

It is not hard to show, using a Taylor series argument identical to the one made in line-search methods for unconstrained optimization, that the line search will eventually succeed, that is, there exists some finite j such that the sufficient decrease condition is satisfied for $\alpha_k = \beta^j \bar{\alpha}$.

Recall: We say that a point x^* is *stationary* if condition (0.4) holds. We have the following convergence result for conditional gradient.

Theorem 0.6. *Suppose that f is continuously differentiable in an open neighborhood of the compact convex set Ω. Then any accumulation point of $\{x_k\}$ is a stationary point.*

Proof. Note that by the compactness and continuity assumptions, f is bounded below on Ω and in fact achieves its minimum value over this set.

Suppose for contradiction that there is an accumulation point $\hat{x}$ that is *not* stationary, that is, $-\nabla f(\hat{x}) \notin N_\Omega(\hat{x})$. Then by the convex-set characterization of N_Ω there exists $z \in \Omega$ such that $-\nabla f(\hat{x})^T (z - \hat{x}) > 0$, that is,

$$\bar{\gamma} > 0, \quad \text{where } \bar{\gamma} := -\nabla f(\hat{x})^T (z - \hat{x}) > 0.$$

Let K be the subsequence of iterates that approaches $\hat{x}$ that is, $\lim_{k \in K,\, k \to \infty} x_k = \hat{x}$. Note that for $k \in K$ sufficiently large, and by definition of $\bar{x}_k$, we have

$$\nabla f(x_k)^T (\bar{x}_k - x_k) \le \nabla f(x_k)^T (z - x_k) \le -\frac{1}{2} \bar{\gamma}. \tag{0.6}$$

Now consider two cases. In Case I, there exists $\hat{\alpha} > 0$ such that $\alpha_k \ge \hat{\alpha}$ for all $k \in K$ sufficiently large. We thus have from the acceptance condition that, for such k,

$$f(x_{k+1}) \le f(x_k) + c_1 \alpha_k \nabla f(x_k)^T (\bar{x}_k - x_k) \le f(x_k) - \frac{1}{2} \bar{\gamma} c_1 \hat{\alpha},$$

which implies, since $\{f(x_k)\}$ is a decreasing sequence, that $f(x_k) \downarrow -\infty$, which contradicts boundedness below. Hence Case I cannot happen.

Case II is the other alternative, in which by reducing the subsequence K if necessary, we have that $\lim_{k \in K,\, k \to \infty} \alpha_k = 0$. For each such k it must have been necessary to backtrack at least once, that is, the steplength $\beta^{-1} \alpha_k$ must have failed the sufficient decrease test, that is,

$$f(x_k + \beta^{-1} \alpha_k (\bar{x}_k - x_k)) > f(x_k) + c_1 \beta^{-1} \alpha_k \nabla f(x_k)^T (\bar{x}_k - x_k).$$

Because of Taylor's theorem, and the smoothness assumptions, we have

$$f(x_k + \beta^{-1}\alpha_k(\bar{x}_k - x_k)) = f(x_k) + \beta^{-1}\alpha_k \nabla f(x_k)^T(\bar{x}_k - x_k) + o(\alpha_k).$$

By combining these last two expressions, we obtain

$$(1 - c_1)\beta^{-1}\alpha_k \nabla f(x_k)^T(\bar{x}_k - x_k) > o(\alpha_k)$$

and by dividing both sides by α_k, using $(1 - c_1) > 0$ and $\beta^{-1} > 0$, we have

$$\lim_{k \in K,\, k \to \infty} \nabla f(x_k)^T(\bar{x}_k - x_k) \geq 0,$$

which contradicts (0.6). Hence Case II also cannot happen. We conclude that accumulation points are stationary. □

We can prove a simplified version of the convergence result in the case of f convex, if we use a different line search. We can even get a result concerning the speed of convergence. (It's sublinear.) We prove this result for the *exact* line search. (As we point out later, a similar result can be obtained when we relax the assumptions on the line search and on the accuracy of the subproblem.) The algorithm is shown as Algorithm 2.

Algorithm 2 Conditional Gradient (exact search)

```
Given $x_0$;
for $k = 0, 1, 2, \ldots$ do
    $\bar{x}_k := \arg\min_{z \in \Omega} \nabla f(x_k)^T(z - x_k)$;
    if $\nabla f(x_k)^T(\bar{x}_k - x_k) = 0$ then
        STOP;
    end if
    $\alpha_k := \arg\min_{\alpha \in (0,1]} f(x_k + \alpha(\bar{x}_k - x_k))$;
    $x_{k+1} \leftarrow x_k + \alpha_k(\bar{x}_k - x_k)$;
end for
```

We start with a few preliminaries. We are going to assume that f is convex and Lipschitz continuously differentiable, with Lipschitz constant L on an open neighborhood of the feasible set Ω. Thus for any $y, z \in \Omega$, we have

$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\|.$$

By taking the term that arises in one form of Taylor's theorem, we have in consequence that

$$\begin{aligned}
&\left| \int_0^1 [\nabla f(y + \gamma(z - y)) - \nabla f(y)]^T (y - z)\, d\gamma \right| \\
&\leq \int_0^1 \|\nabla f(y + \gamma(z - y)) - \nabla f(y)\| \|y - z\|\, d\gamma \\
&\leq L\|y - z\|^2 \int_0^1 \gamma\, d\gamma \leq \frac{1}{2} L\|y - z\|^2.
\end{aligned} \tag{0.7}$$

We also have from convexity of f that for any $y, z \in \Omega$, we have

$$\nabla f(y)^T(z - y) \leq f(z) - f(y). \tag{0.8}$$

Proof: We have by Taylor's theorem that for small positive α:

$$f(y+\alpha(z-y)) = f(y) + \alpha \nabla f(y)^T (z-y) + o(\alpha),$$

while from convexity we have

$$f(y+\alpha(z-y)) \le (1-\alpha) f(y) + \alpha f(z).$$

We get (0.8) by combining these two results, diving by α, and taking $\alpha \downarrow 0$.

We also define D to be the diameter of the compact feasible set Ω, that is, $D := \sup_{y,z \in \Omega} \|y-z\|$.

Lecture 9. (2/9/15; 60 min)

THEOREM 0.7. *Suppose that Ω is compact and convex with diameter D and that f is convex Lipschitz continuously differentiable on an open neighborhood of Ω with Lipschitz constant L. Let x^* be a minimizer of f over Ω. Then we have that $\{f(x_k)\}$ decreases monotonically to $f(x^*)$, with*

$$f(x_k) - f(x^*) \le \frac{2LD^2}{k+2}, \quad k = 1, 2, \dots .$$

Proof. By using (0.7) and the definition of D, we have

$$\begin{aligned} f(x_{k+1}) &= f(x_k + \alpha_k(\bar{x}_k - x_k)) \\ &= \min_{\alpha \in (0,1]} f(x_k + \alpha(\bar{x}_k - x_k)) \\ &= \min_{\alpha \in (0,1]} f(x_k) + \alpha \nabla f(x_k)^T (\bar{x}_k - x_k) \\ &\quad + \alpha \int_0^1 [\nabla f(x_k + \gamma\alpha(\bar{x}_k - x_k)) - \nabla f(x_k)]^T (\bar{x}_k - x_k)\, d\gamma \\ &\le \min_{\alpha \in (0,1]} f(x_k) + \alpha \nabla f(x_k)^T (\bar{x}_k - x_k) + \frac{1}{2}\alpha^2 L D^2. \end{aligned} \tag{0.9}$$

For the α term, we have by definition of $\bar{x}_k$ and feasibility of x^* that

$$\nabla f(x_k)^T (\bar{x}_k - x_k) \le \nabla f(x_k)^T (x^* - x_k) \le f(x^*) - f(x_k). \tag{0.10}$$

where the last inequality comes from (0.8). By subtracting $f(x^*)$ from both sides of (0.9), and using (0.10), we have

$$f(x_{k+1}) - f(x^*) \le \min_{\alpha \in (0,1]} (1-\alpha)[f(x_k) - f(x^*)] + \frac{1}{2}\alpha^2 L D^2. \tag{0.11}$$

We prove the claim by induction. First note that when we set $k = 0$ in (0.11), we have by considering the value $\alpha = 1$ that

$$f(x_1) - f(x^*) \le \min_{\alpha \in (0,1]} (1-\alpha)[f(x_0) - f(x^*)] + \frac{1}{2}\alpha^2 L D^2 \le \frac{1}{2} L D^2 < \frac{2}{3} L D^2,$$

so that the desired bound holds in the base case of $k = 1$. For the inductive step, we suppose that the claim holds for some $k \ge 0$ and show that it continues to hold for

$k+1$. By substituting the inductive hypothesis into (0.11), *and making the particular choice* $\alpha = 2/(k+2)$, we have

$$\begin{aligned}
f(x_{k+1}) - f(x^*) &\leq \left(1 - \frac{2}{k+2}\right) \frac{2LD^2}{k+2} + \frac{1}{2} \frac{4}{(k+2)^2} LD^2 \\
&= LD^2 \left[\frac{2k}{(k+2)^2} + \frac{2}{(k+2)^2} \right] \\
&= 2LD^2 \frac{(k+1)}{(k+2)^2} \\
&= 2LD^2 \frac{k+1}{k+2} \frac{1}{k+2} \\
&\leq 2LD^2 \frac{k+2}{k+3} \frac{1}{k+2} = \frac{2LD^2}{k+3},
\end{aligned}$$

as required □

Machine learning people love this kind of result, but it's important to bear in mind that it's *slow*! Only sublinear.

Variants:

- Can simply set $\alpha_k = 2/(k+2)$ — skip the line search — and the proof still holds. Better idea: use this value as the starting guess for a line search and take one or two steps. Provided a descent method is used for the line search, the proof still holds.
- Instead of solving the linearized subproblem exactly, find an inexact proof that still guarantees the key inequality

  $$\nabla f(x_k)^T (\bar{x}_k - x_k) \leq \eta_k [f(x^*) - f(x_k)],$$

  for some $\eta_k \in (0, 1]$. (The exact solution satisfies this inequality with $\eta_k = 1$.) Then have to modify the result of the analysis to let η_k ripple through, possibly use a different step length etc. (Exercise! Perhaps can show that if $\eta_k \equiv \eta > .5$, the step can be $2/(k+2\eta)$ and the bound can be $2LD^2/(k+2\eta)$. Didn't check it yet.)

Some examples where solving the lienarized subproblem in Cond Grad is much cheaper than solving the full problem:

- When $\Omega = [0, 1]^n$, solution is

  $$[\bar{x}_k]_i = \begin{cases} 0 & \text{if } [\nabla f(x_k)]_i \geq 0, \\ 1 & \text{if } [\nabla f(x_k)]_i < 0. \end{cases}$$

- When $\Omega = \{x \mid \|x\| \leq 1\}$, solution is $\bar{x}_k = -\nabla f(x_k)/\|\nabla f(x_k)\|$.

Gradient Projection. Consider the problem

$$\min_{x \in \Omega} f(x), \tag{0.12}$$

where f is smooth and Ω is closed and convex.

A basic algorithm is *gradient projection.* In a sense it is the natural extension of steepest descent to the problem (0.12). The search path from a point x is the projection of the steepest descent path onto the feasible set Ω, that is,

$$P(x - \alpha \nabla f(x)), \quad \alpha > 0.$$

Draw pictures of various cases, emphasizing the path traced out by $P(x_k - \alpha \nabla f(x_k)$ for $\alpha > 0$.

***** NO CLASS ON 2/11/15 (DOE PANEL) *****

Lecture 10. (2/13/15; 70 min)

THEOREM 0.8. *Assume that Ω is closed convex. If*

$$P(x^* - \bar{\alpha}\nabla f(x^*)) = x^*, \ \ \textit{for some } \bar{\alpha} > 0, \tag{0.13}$$

then (0.4) holds. Conversely, if (0.4) holds, then

$$P(x^* - \alpha \nabla f(x^*)) = x^*, \ \ \textit{for all } \alpha > 0.$$

Proof. Suppose first that (0.13) holds. In Lemma 0.1(ii) we set

$$y = x^* - \bar{\alpha}\nabla f(x^*), \ \ P(y) = x^*,$$

and let z be any element of Ω. We then have

$$0 \geq (y - P(y))^T (z - P(y)) = (-\bar{\alpha}\nabla f(x^*))^T (z - x^*),$$

which implies that $\nabla f(x^*)^T (z - x^*) \geq 0$ for all $z \in \Omega$, proving (0.4).

Now supposed that (0.4) holds, and denote

$$x_\alpha = P(x^* - \alpha \nabla f(x^*)).$$

Setting $y = x^* - \alpha \nabla f(x^*)$, $P(y) = x_\alpha$, $z = x^*$ in Lemma 0.1, we have

$$(x^* - \alpha \nabla f(x^*) - x_\alpha)^T (x^* - x_\alpha) \leq 0,$$

which implies that

$$\|x^* - x_\alpha\|_2^2 \leq \alpha \nabla f(x^*)^T (x^* - x_\alpha) \leq 0, \tag{0.14}$$

by (0.4). Thus $x^* = x_\alpha$, as claimed. □

An immediate consequence of this theorem is that $P(x^* - \bar{\alpha}\nabla f(x^*)) = x^*$ for *some* $\bar{\alpha} > 0$, then $P(x^* - \alpha \nabla f(x^*)) = x^*$ for *all* $\alpha > 0$. (Why?)

Sufficent decrease for GP: We seek an α for which the function decreases, that is,

$$f(x) > f(P(x - \alpha \nabla f(x))).$$

For global convergence we need a stronger condition than this. Ideally, we would like to find an α that minimizes the line search function $\phi(\alpha) \equiv f(P(x - \alpha \nabla f(x)))$. But this is usually too much work, even for simple functions f, because of the complications introduced by the projection operation. In general it is not a smooth function of α. For example, when Ω is polyhedral and f is quadratic, ϕ is piecewise quadratic. When Ω is polyhedral and f is smooth, ϕ is piecewise smooth. But because of the "kinks" in the search path, we cannot apply line search procedures like the one developed in Chapter 3, which assume smoothness. Backtracking may be more suitable.

To describe this strategy we use the following notation, for convenience:

$$x(\alpha) \overset{\text{def}}{=} P(x - \alpha \nabla f(x)).$$

We consider an Armijo backtracking strategy along the projection arc, in which we choose an $\bar{\alpha} > 0$ and take the step α_k to be the first element in the sequence $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \dots$ for which the following condition is satisfied:

$$f(x_k(\beta^m \bar{\alpha})) \le f(x_k) + c_1 \nabla f(x_k)^T (x_k(\beta^m \bar{\alpha}) - x_k). \tag{0.15}$$

We can show that the resulting algorithm has stationary limit points. The analysis is quite complicated (see pp. 236–240 of Bertsekas), but we state it in part because the proof techniques are interesting and relevant to convergence results in other settings.

See Algorithm 3 for a backtracking gradient projection algorithm.

Algorithm 3 Gradient Projection

```
Given $x_0$, $c_1 \in (0,1)$, $\beta \in (0,1)$, $\bar{\alpha} > 0$;
for $k = 0, 1, 2, \dots$ do
  $j \leftarrow 0$;
  while $f(x_k(\beta^j \bar{\alpha})) > f(x_k) + c_1 \nabla f(x_k)^T [x_k(\beta^j \bar{\alpha}) - x_k]$ do
    $j \leftarrow j + 1$;
  end while
  $\alpha_k \leftarrow \beta^j \bar{\alpha}$;
  $x_{k+1} \leftarrow x_k(\alpha_k)$;
  if $x_{k+1} = x_k$ then
    STOP;
  end if
end for
```

We need a preliminary "geometric" lemma whose proof is omitted. It's proved as Lemma 2.3.1 of Bertsekas.

Lemma 0.9. *For all $x \in \Omega$ and $z \in \mathbb{R}^n$, the function $g : [0, \infty) \to \mathbb{R}$ defined by*

$$g(s) \overset{\text{def}}{=} \frac{\|P(x + sz) - x\|}{s}$$

is monotonically nonincreasing.

Our convergence result is then:

Theorem 0.10.

(a) For every $x \in \Omega$ there exists a scalar $s_x > 0$ such that

$$f(x) - f(x(s)) \ge c_1 \nabla f(x)^T (x - x(s)), \quad \forall s \in [0, s_x]. \tag{0.16}$$

(b) Let x_k be the sequence generated by Algorithm GP. Then every accumulation point of $\{x_k\}$ is stationary

Proof. We follow the proof of Bertsekas Proposition 2.3.3. For Part (a), we have by Lemma 0.1(ii) with $y = x - s\nabla f(x)$ (thus $P(y) = x(s)$) that

$$(x - x(s))^T (x - s\nabla f(x) - x(s)) \le 0, \quad \forall x \in \Omega, \ s > 0.$$

By rearrangement this becomes

$$\nabla f(x)^T (x - x(s)) \ge \frac{\|x - x(s)\|^2}{s}, \quad \forall x \in \Omega, \ s > 0. \tag{0.17}$$

If x is stationary (that is, satisfies conditions and (0.13)) we have $x(s) = x$ for all $s > 0$, so the conclusion (0.16) holds trivially. Otherwise, x is nonstationary, so $x \neq x(s)$ for all $s > 0$. By Taylor's theorem we have

$$f(x) - f(x(s)) = \nabla f(x)^T (x - x(s)) + (\nabla f(\zeta_s) - \nabla f(x))^T (x - x(s)),$$

for some ζ_s on the line segment between x and $x(s)$. Hence, (0.16) can be written as

$$(1 - c_1) \nabla f(x)^T (x - x(s)) \geq (\nabla f(x) - \nabla f(\zeta_s))^T (x - x(s)). \tag{0.18}$$

From (0.17) and Lemma 0.9, we have for all $s \in (0, \bar{\alpha}]$ that

$$\nabla f(x)^T (x - x(s)) \geq \frac{\|x - x(s)\|^2}{s} \geq \frac{1}{\bar{\alpha}} \|x - x(\bar{\alpha})\| \|x - x(s)\|.$$

Therefore (0.18) is satisified for all $s \in (0, 1]$ such that

$$(1 - c_1) \frac{1}{\bar{\alpha}} \|x - x(\bar{\alpha})\| \geq (\nabla f(x) - \nabla f(\zeta_s))^T \frac{x - x(s)}{\|x - x(s)\|}.$$

Obviously the left-hand side of this expression is strictly positive, while the right-hand side goes to zero continuously, as $s \downarrow 0$. Hence there is an s_x with the desired property, so the proof of part (a) is complete.

Lecture 11. (2/16/15; 70 min)

Note that part (a) ensures that we can find an α_k of the form $\alpha_k = \beta^{m_k}$ that satisfies (0.15) from each point x_k. Suppose there is a subsequence K such that $x_k \to_{k \in K} \bar{x}$. Since the sequence of function values $\{f(x_k)\}$ is nonincreasing we have $f(x_k) \to f(\bar{x})$.

We consider two cases. First, suppose that

$$\liminf_{k \in K} \alpha_k \geq \hat{\alpha}$$

for some $\hat{\alpha} > 0$. From (0.17) and Lemma 0.9, we have for all $k \in K$ sufficiently large that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq c_1 \nabla f(x_k)^T (x_k - x_{k+1}) \\ &\geq c_1 \frac{\|x_k - x_{k+1}\|^2}{\alpha_k} \\ &= \frac{c_1 \alpha_k \|x_k - x_{k+1}\|^2}{\alpha_k^2} \\ &\geq c_1 \hat{\alpha} \frac{1}{\bar{\alpha}^2} \|x_k - x_k(\bar{\alpha})\|^2, \end{aligned}$$

since $\bar{\alpha}$ is the initial choice of steplength and $\alpha_k \leq \bar{\alpha}$. Taking the limit as $k \to \infty$, $k \in K$, we have $\bar{x} - \bar{x}(\bar{\alpha}) = 0$, which implies that $\bar{x}$ is stationary (see (0.13)).

In the second case, suppose that $\liminf_{k \in K, k \to \infty} \alpha_k = 0$. Then by taking a further subsequence $\bar{K} \subset K$, we have $\lim_{k \in \bar{K}, k \to \infty} \alpha_k = 0$. For all $k \in \bar{K}$, the Armijo test (0.15) will fail at least once, so we have

$$f(x_k) - f(x_k(\beta^{-1} \alpha_k)) < c_1 \nabla f(x_k)^T (x_k - x_k(\beta^{-1} \alpha_k)). \tag{0.19}$$

Further, no x_k with $k \in \bar{K}$ can be stationary, since for stationary points we have $\alpha_k = 1$. Hence,

$$\|x_k - x_k(\beta^{-1}\alpha_k)\| > 0. \tag{0.20}$$

By Taylor's theorem we have

$$\begin{aligned} f(x_k) - f(x_k(\beta^{-1}\alpha_k)) = \nabla f(x_k)^T(x_k - x_k(\beta^{-1}\alpha_k)) \\ + (\nabla f(\zeta_k) - \nabla f(x_k))^T(x_k - x_k(\beta^{-1}\alpha_k)), \end{aligned}$$

for some ζ_k on the line segment between x_k and $x_k(\beta^{-1}\alpha_k))$. By combining this expression with (0.19), we have

$$(1-c_1)\nabla f(x_k)^T(x_k - x_k(\beta^{-1}\alpha_k)) < (\nabla f(x_k) - \nabla f(\zeta_k))^T(x_k - x_k(\beta^{-1}\alpha_k)),$$

From (0.17) with Lemma 0.9, we have

$$\begin{aligned} \nabla f(x_k)^T(x_k - x_k(\beta^{-1}\alpha_k)) &\geq \frac{\|x_k - x_k(\beta^{-1}\alpha_k)\|^2}{\beta^{-1}\alpha_k} \\ &\geq \frac{1}{\bar{\alpha}}\|x_k - x_k(\bar{\alpha})\|\|x_k - x_k(\beta^{-1}\alpha_k)\|. \end{aligned}$$

By combining the last two results, and using the Schwartz inequality, we have for large $k \in \bar{K}$ that

$$\begin{aligned} (1-c_1)\frac{1}{\bar{\alpha}}\|x_k - x_k(\bar{\alpha})\|\|x_k - x_k(\beta^{-1}\alpha_k)\| &< (\nabla f(x_k) - \nabla f(\zeta_k))^T(x_k - x_k(\beta^{-1}\alpha_k)) \\ &\leq \|\nabla f(x_k) - \nabla f(\zeta_k)\|\|x_k - x_k(\beta^{-1}\alpha_k)\|. \end{aligned}$$

Using this expression together with (0.20) we obtain

$$(1-c_1)\frac{1}{\bar{\alpha}}\|x_k - x_k(\bar{\alpha})\| < \|\nabla f(x_k) - \nabla f(\zeta_k)\|.$$

Since $\alpha_k \to 0$ and $x_k \to \bar{x}$ as $k \to \infty$, $k \in \bar{K}$, it follows that $\zeta_k \to \bar{x}$. Hence by taking limits in the expression above we obtain that $\bar{x} = \bar{x}(\bar{\alpha})$, which implies that $\bar{x}$ is stationary, as claimed. □

Proved that when f is convex in the problem $\min_{x\in\Omega} f(x)$, where Ω is convex and closed, that stationary points are actually minimizers.

Interpolated here a discussion of Barzilai-Borwein, motivated as a quasi-Newton method with a single parameter. Derived regular and inverse BB formulae, discussed nonmonotone property, showed the 2-d convex quadratic example of its behavior. This is needed for Homework 3.

Lecture 12. (2/18/15; 60 min)

Now discuss several variants and enhancements to gradient projection.

- We need a practical termination criterion. This is a major topic by itself. One simple possibility is to stop when $\|x_k - P(x_k - \bar{\alpha}\nabla f(x_k))\|$ is small (less than 10^{-6}, say). Note that this quantity is zero if and only if x_k satistifies first-order conditions, and is a continuous function of x_k.
- For many sets Ω of interest, the path defined by $P(x - t\bar{\alpha}\nabla f(x))$ for $t \in [0,1]$ is piecewise smooth. For example, when Ω is polyhedral, the path is piecewise linear. We could devise specialized line search techniques that exploit this structure.

- An alternative to the projected search path is to project just once, then searh along the straight line defined by

$$x_k + t[P(x_k - \bar{\alpha}\nabla f(x_k)) - x_k], \ \ t \in [0,1].$$

If we do a backtracing line search, we would try the values $t = 1, \beta, \beta^2, \dots$. However, since the search is along a straight line, we could easily implement a more sophisticated line search technique. *This is a kind of combination of gradient projection and conditional gradient.*

- Rather than trying the constant value $\bar{\alpha}$ as the initial guess of step size at every iteration, we could start with the steplength from the previous iteration, perhaps augmented slightly. Specifically, use α_{k-1}/β as the initial guess of α_k. This can save many backtracking steps, when $\bar{\alpha}$ is too large. (Many variants on this idea are possible.)
- We could use Barzilai-Borwein step lengths / nonmonotone methods, extending the ideas from unconstrained optimization in a straightforward way. More below on BB.

Higher-Order Gradient Projection. Newton step for unconstrained optimization has fast local convergence: $\nabla^2 f(x_k)p_k = -\nabla f(x_k)$. Why not search along $P(x_k + \alpha p_k)$? Draw the picture to show that this doesn't work. Use it to motivate two-metric methods and methods that take reduced Newton steps. Focus on the latter. Specialize to the case of bound-constrained optimization.

Talked about the "right" implementation of Newton's method for $\min_{x\in\Omega} f(x)$, which is to obtain steps as follows:

$$x^{k+1} := \min_{z\in\Omega} \nabla f(x^k)^T(z - x^k) + \frac{1}{2}(z - x^k)^T \nabla^2 f(x^k)(z - x^k). \tag{0.21}$$

This has nice properties, but is not used often, as the subproblem is hard to solve. (In some settings however the subproblem can be solved inexactly and efficiently.)

Active Face Identification.. Showed that gradient projection steps have the "identification property" when the local solution x^* is nondegenerate, when Ω is the positive orthant: $\Omega := \{x \,|\, x \ge 0\}$.

We first discussed the concept of identification at some length in the general context of $\min_{x\in\Omega} f(x)$ and drew lots of pictures. The central question is whether the point $P(x - \alpha\nabla f(x))$ lies on the same face/vertex as x^*, where α is a fixed positive parameter and x is close to x^*. Generally this is true when $-\nabla f(x^k)$ is in the *relative interior* of the normal cone $N_\Omega(x^*)$. Such points are said to be *nondegenerate* or to satisfy *strict complementary.*

Lecture 13. (2/20/15; 60 min)

For the positive orthant, the first-order condition $-\nabla f(x^*) \in N_\Omega(x^*)$ can be restated as

$$0 \le x^* \perp \nabla f(x^*) \ge 0.$$

A nondegenerate stationary (first-order) point requires the additional condition $x^* + \bar{\alpha}\nabla f(x^*) > 0$ (strict positivity) for some $\bar{\alpha} > 0$. Equivalently, nondegneracy means that *exactly one* of x_i^* and $[\nabla f(x^*)]_i$ is zero for each i (and the other is strictly positive).

We stated this theorem and sketched a proof. Gradient projection identifies the correct active set, for all x close enough to x^*.

THEOREM 0.11. *Suppose that f is continuously differentiable, that Ω is the positive orthant, and that x^* is a nondegenerate stationary point for the problem $\min_{x\in\Omega} f(x)$. Define $\bar{\alpha} > 0$. Then there is $\epsilon > 0$ such that for all x with $\|x - x^*\| \le \epsilon$, we have*

$$\begin{aligned} x_i^* > 0 \;&\Rightarrow\; x_i - \bar{\alpha}[\nabla f(x)]_i > 0, \\ x_i^* = 0 \;&\Rightarrow\; x_i - \bar{\alpha}[\nabla f(x)]_i < 0. \end{aligned}$$

that is, $P(x - \alpha\nabla f(x))$ has the same nonzero indices as x^*.

Proof. It is easy to verify that for a nondegenerate stationary point x^*, the following are true:

$$\begin{aligned} x_i^* > 0 \;&\Rightarrow\; x_i^* - \bar{\alpha}[\nabla f(x^*)]_i > 0, \\ x_i^* = 0 \;&\Rightarrow\; x_i^* - \bar{\alpha}[\nabla f(x^*)]_i < 0. \end{aligned}$$

The result follows by noting that $x - \bar{\alpha}\nabla f(x)$ is a continuous function of x, so for $x - x^*$ sufficiently small, the components of $x - \bar{\alpha}\nabla f(x)$ will have the same sign as the components of $x^* - \bar{\alpha}\nabla f(x^*)$. □

The identification property generalizes to the case of general polyhedral Ω, but we need more geometric machinery, namely, the concepts of *facets* and *relative interior.*

Forward-Backward Methods. Note that

$$\begin{aligned} P(x - \alpha\nabla f(x)) &= \arg\min_{z\in\Omega} \frac{1}{2}\|z - (x - \alpha\nabla f(x))\|_2^2 \\ &= \arg\min_{z\in\Omega} \nabla f(x)^T(z - x) + \frac{1}{2\alpha}\|z - x\|_2^2. \end{aligned}$$

i.e. it is the minimizer of a simple quadratic model to f, with the linear term coming from Taylor's theorem and a simple quadratic term (with Hessian $(1/\alpha)I$). Decreasing α means more damping.

Another way is to write the original problem $\min_{x\in\Omega} f(x)$ as

$$\min_x f(x) + I_\Omega(x)$$

and to write this subproblem as

$$\min_z \nabla f(x)^T(z - x) + \frac{1}{2\alpha}\|z - x\|_2^2 + I_\Omega(z),$$

where $I_\Omega(\cdot)$ denotes the indicator function for Ω, defined so that $I_\Omega(x) = 0$ for $x \in \Omega$ and $I_\Omega(x) = \infty$ for $x \notin \Omega$. This formulation gives us a way to generalize the gradient projection idea to problems of the form

$$\min_x f(x) + \lambda\psi(x), \tag{0.22}$$

where f is smooth and ψ may be nonsmooth, and $\lambda \ge 0$ is a regularization parameter. Replacing I_Ω by ψ in the template above, we obtain the subproblem

$$\min_z \nabla f(x)^T(z - x) + \frac{1}{2\alpha}\|z - x\|_2^2 + \lambda\psi(z).$$

This is sometimes known as the *forward-backward* method. It is the basis for many recently proposed successful techniques in regularized optimization.

We discussed regularized optimization - motivating it as the search for *structured approximate minimizers* of f. Need to choose regularization function ψ to induce the desired kind of structure in the solution. Larger λ implies more structred solution with less emphasis on minimizing f.

In the case of $\psi(x) = \|x\|_1$, we get *sparse* solutions — x^* has few nonzero elements.

(* NO CLASSES 2/23/15 or 2/25/15 *)

Lecture 14. (2/27/15; 65 min)

First-Order Conditions for Algebraic Constraints. Example:

$$\min\,(x_1+1)^2 + (x_2-T)^2 \;\text{ s.t. }\; x_2 \le (1/2)x_1,\; x_2 \ge -(1/2)x_1,$$

where T is a parameter. Show that the KKT conditions are satisfied at $x^* = (0,0)$ when $T = 0$. For what values of T does the solution stay at $(0,0)$? Then seek solutions on which just the first constraint is active. What values of T lead to solutions in this regime?

Note that Def 12.2 (p.316) of $T_\Omega(x^*)$ is equivalent to the "limiting feasible directions" definition of tangent presented earlier in these notes: just set $d_k = (z_k - x)/t_k$ and note that $x + t_k d_k \in \Omega$ and $d_k \to d$ and $t_k \downarrow 0$.

Now mention Theorem 12.3 on p.325. It's equivalent to Theorem 0.2 proved in Lecture 5, where we stated it equivalently as "x^* is a local solution implies that $-\nabla f(x^*) \in T_\Omega(x^*)^\circ = N_\Omega(x^*)$" using our definition of normal cone as polar of the tangent (which applies even when Ω is nonconvex).

Define *linearized feasible direction set* $\mathcal{F}(x)$ as in Def 12.3 (p.316). Note that this definition uses the algebraic description of the feasible set. Draw some examples where $T_\Omega(x^*)$ and $\mathcal{F}(x^*)$ are the same. Now note some examples where they differ e.g. pp. 319-320 and Figure 12.10 (p.320).

Lecture 15. (3/2/15; 65 min)

Went over solution to Homework 4, Q5 in detail. It's relevant to the implicit-function-theorem argument to be discussed next..

State Lemma 12.2 on p. 323:

Lemma 0.12. *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 LICQ or any other similar condition.

We have $\mathcal{F}(x^*) \subset T_\Omega(x^*)$ when *all constraints are linear,* that is, when Ω is polyhedral. (Discussed this earlier, numerous times).

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.

Constraint Qualifications are conditions under which $\mathcal{F}(x)$ and $T_\Omega(x)$ are the same — or at least, their normal cones coincide. Because $\mathcal{F}(x)$ is a polyhedral set, we have a nice way to express membership of the normal cone: the Farkas Lemma. This allows us to derive the the KKT conditions, in the same for as for polyhedral Ω, above.

Lecture 16. (3/4/15; 65 min)

Define $\mathcal{F}(x^*)$ and $T_\Omega(x^*)$ again. Apply Theorem 0.5 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

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

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.

Reminder: At a point satisfying KKT conditions we say that *strict complementarity* holds if for all Lagrange multipliers λ_i^* corresponding to active inequality constraints, we may have $\lambda_i^* > 0$ (strictly positive). Note: For a given point x^*, there might be multiple sets of multipliers λ^* that satisfy KKT. Some might be strictly complementary, others not.

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

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

Note that LICQ is neither implied by nor implies the "Linear Constraint CQ".

Example: Problem in which Linear Constraint CQ is satisfied but not LICQ: Suppose have constraint set $\Omega = \{(x_1, x_2) \,|\, x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \geq 0\}$. "Linear Constraint CQ is satisfied at $x^* = 0$ but not LICQ¿ Suppose $x^* = 0$ have $\nabla f(x^*) = (2/3, 1/3)$. Then optimal multipliers satisfy

$$
\nabla f(x^*) = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \lambda_1 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \lambda_2 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \lambda_3,
$$

as well as $(\lambda_1, \lambda_2, \lambda_3) \geq 0$. Thus the optimal multiplier set has

$$
2/3 = \lambda_1 + \lambda_3,\ \ 1/3 = \lambda_2 + \lambda_3,\ \ (\lambda_1, \lambda_2, \lambda_3) \geq 0.
$$

By manipulating we obtain optimal multipler set

$$
(2/3 - t, 1/3 - t, t),\ \ t \in [0, 1/3].
$$

Clearly strict complementarity holds when $t \in (0, 1/3)$ (open interval).

Discuss

$$
\Omega := \{(x_1, x_2) \,|\, x_1 \geq 0,\ x_2 \geq 0,\ x_1 x_2 = 0\}. \tag{0.23}
$$

Does it satisfy LICQ? Consider various feasible points.

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

$$
\min\ x_1 + x_2 \ \text{ s.t. }\ x_2 \geq 0,\ x_2 \leq x_1^3.
$$

Solution at $(0, 0)$, but KKT conditions are not satisfied, because CQ fails to hold.

Show that we can still have KKT condition satisfied at a local solution even when LICQ does not hold. Example:

$$\min \frac{1}{2}(x_1+1)^2 + \frac{1}{2}(x_2+1/2)^2 \;\text{ s.t. }\; x_2 \le x_1,\; x_2 \ge -x_1,\; x_1 \ge 0.$$

Derive an explicit expression for the set of optimal multipliers λ_i^*, $i = 1, 2, 3$. Show that it's finite line segment.

Lecture 17. (3/6/15; 60 min)

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

The set of optimal multipiers is bounded when MFCQ holds. Prove 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 $\|\lambda_{\mathcal{A}}^k\| \to \infty$, $\lambda_{\mathcal{A}}^k \ge 0$, 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

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

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 $\|t\| = 1$, $t \ge 0$, $t \ne 0$ such that $\lambda_{\mathcal{A}}^k/\|\lambda_{\mathcal{A}}^k\| \to t$. 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

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

yielding a contradiction.

Discuss CQ properties at points in the set Ω defined by (0.23).

Did my favorite two-circle exampleL

$$\min -x_1 \;\text{ s.t. }\; 1 - x_1^2 - x_2^2 \ge 0,\;\; 4 - (x_1+1)^2 - x_2^2 \ge 0. \tag{0.24}$$

$x^* = (1, 0)$ is a solution - in fact a strict solution.) Showed that MFCQ is satisfied but not LICQ. Derived the polyhedral set of optimal λ^*.

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.

Suppose the constraints in the NLP are changed to

$$c_i(x) = \epsilon_i \;\; i \in \mathcal{E}; \quad c_i(x) \ge \epsilon_i, \;\; i \in \mathcal{I},$$

where the ϵ_i are all small. Suppose the perturbed problem has solution $x(\epsilon)$ near x^*. Suppose that LICQ holds. Suppose that the active set remains unchanged at $x(\epsilon)$. Then from the KKT condition

$$\nabla f(x^*) = \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*),$$

we have

$$\begin{aligned} f(x(\epsilon)) - f(x^*) &\approx (x(\epsilon) - x^*)^T \nabla f(x^*) \\ &= \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* (x(\epsilon) - x^*)^T \nabla c_i(x^*) \\ &= \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* c_i(x(\epsilon)) \\ &= \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \epsilon_i. \end{aligned}$$

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

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

Recall 1oN, 2oN, 2oS conditions for unconstrained optimization.

Define and motivate critical cone $\mathcal{C}(x^*, \lambda^*)$. It's a subset of $\mathcal{F}(x^*)$ by definition. Show how the definition simplifies in the case of λ^* strictly complementary to

$$\mathcal{C}(x^*, \lambda^*) = \{w \mid \nabla c_i(x^*)^T w = 0, \ \ i \in \mathcal{A}(x^*)\}.$$

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 2oN and 2oS.

(NO CLASS ON 3/9/15 **)**

Lecture 18. (3/11/15; 65 min)

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.

Do example 12.9 and discuss.

Lecture 19. (3/13/15; 60 min)

Review some interesting problem classes and formulations and see if / when CQs are satisfied.

(i) Optimization with a complementarity constraint:

$$\min f(x) \ \text{ s.t. } \ 0 \le g(x) \perp h(x) \ge 0,$$

where we assume for simplicity that g and h are scalar functions. Could formulate this as

$$\min f(x) \ \text{ s.t. } \ g(x) \ge 0, \ h(x) \ge 0, \ g(x)h(x) = 0,$$

but this fails to satisfy LICQ or MFCQ at *every feasible point.* (Homework!)

(ii) Reformulating equalities as inequalities. Given original problem min $f(x)$ s.t. $c(x) = 0$, reformulate as min $f(x)$ s.t. $c(x) \le 0$, $c(x) \ge 0$. This is equivalent mathematically but fails to satisfy LICQ or MFCQ even if the original problem does.

(iii) Introducing slack variables for inequalities. If LICQ or MFCQ satisfied by the original problem it is satisfied for the slack formulation too.

(iv) What about squared slack variables? Write down KKT conditions for original and reformulated problems, given a set of inequality constraints $c_i(x) \ge 0$, $i = 1, 2, \dots, m$. The problem is that a KKT point for the reformulated problem may not yield a KKT point for the original problem, because the multipliers λ_i^* may have the wrong sign.

For problems in which the linearized constraints are a good approximation to reality (i.e. when a CQ holds) can we design an algorithm that is based on constraint linearization?

Yes. Consider SLP. Usually unbounded subproblems. Can add ℓ_2 or ℓ_∞ trust regions.

Now consider SQP. What should the Hessian be? $\nabla^2 f(x)$ or $\nabla^2 \mathcal{L}(x, \lambda)$?

A preliminary tool: Specify Newton's method for nonlinear equations using notation $F : \mathbf{R}^N \to \mathbf{R}^N$ and Jacobian $J(z) \in \mathbf{R}^{N\times N}$.

Prove quadratic convergence of Newton's method for $F(z) = 0$ at a nondegenerate point z^* (where $J(z^*)$ is nonsingular). Need Lipschitz continuous $J(z)$. See book Theorem 11.2. Summarize as follows: Under Lipschitz continuity of J, we have by Taylor's theorem that

$$-J(z)d = F(z) = F(z) - F(z^*) = J(z)(z - z^*) + O(\|z - z^*\|^2),$$

and thus

$$J(z)[z + d - z^*] = O(\|z - z^*\|^2).$$

By nonsingularity of $J(z^*)$ there is a neighborhood of z^* in which $\|J(z)^{-1}\| \le \beta$ for some $\beta > 0$. Thus

$$\|z + d - z^*\| \le \|J(z)^{-1}\| O(\|z - z^*\|^2) \le \beta O(\|z - z^*\|^2).$$

Derived the KKT system and its Jacobian for the equality constrained case. Use notation

$$\min\ f(x)\ \text{ s.t }\ c_i(x) = 0,\ i = 1, 2, \ldots, m.$$

Lecture 20. (3/16/15; 60 min)

Said that a sufficient condition for Newton system of linear equations to have a solution (near (x^*, λ^*)) is that LICQ and 2oS hold. Prove this by showing nonsingularity of the Jacobian of the KKT system at (x^*, λ^*). Twice continuous diff of f and c_i.

Apply Newton to KKT system for the equality constrained problem.

(Chapter 15.)

Show that this is exactly SQP. Write down the QP approximation, and show that the subproblem KKT conditions are identical to Lagrange-Newton.

Say that convergence of (x^k, λ^k) to (x^*, λ^*) is quadratic. This does *not* imply that convergence of $\{x^k\}$ alone is quadratic.

Discuss variant in which the λ are updated not from the solution of the Lagrange-Newton subproblem, but rather obtained from the least-squares estimate

$$\lambda^{k+1} := \arg\min_{\lambda} \|\nabla f(x^{k+1}) - A(x^{k+1})\lambda\|_2^2,$$

where $A(x) := [\nabla c_1(x) ... \nabla c_m(x)]$. When LICQ holds, show that $\|\lambda^k - \lambda^*\| = O(\|x^k - x^*\|)$.

Extend to inequality constraints.

Discussed the version of SQP that uses only $\nabla^2 f(x^k)$ in the second-order term. For linear constraints, of course, this is identical to the Lagrangian Hessian.

Also discussed QCQP - subproblem are generally too hard to solve, and are usually not convex.

Didn't do this: Extend SQP to quasi-Newton approximate Hessians.

Lecture 21. (3/18/15; 70 min)

Quadratic penalty for equality constraints.

$$P_\mu(x) = f(x) + \frac{\mu}{2}\sum_{i=1}^m c_i^2(x).$$

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 KKT conditions and compare with $\nabla P_\mu(x) = 0$, which is satisfied by the minimizer $x(\mu)$ of P_μ. Note relationship $-\lambda_i^k \approx \mu c_i(x^k)$. See p.506 of book. Expanding a bit on that argument, and using this setting for λ^k, we have that the SQP step satisfies

$$\begin{bmatrix} \nabla^2\mathcal{L}^k & A(x^k)^T \\ A(x^k) & 0 \end{bmatrix} \begin{bmatrix} d_x \\ d_\lambda \end{bmatrix} = \begin{bmatrix} -\nabla_x\mathcal{L}^k \\ -c(x^k) \end{bmatrix},$$

while after substitution and rearrangement, the quadratic penalty step p_x satisfies

$$\begin{bmatrix} \nabla^2\mathcal{L}^k & A(x^k)^T \\ A(x^k) & (-1/\mu_k)I \end{bmatrix} \begin{bmatrix} p_x \\ p_\lambda \end{bmatrix} = \begin{bmatrix} -\nabla_x\mathcal{L}^k \\ 0 \end{bmatrix},$$

for some value of p_λ. By taking differences and rearranging, we obtain

$$\begin{bmatrix} \nabla^2\mathcal{L}^k & A(x^k)^T \\ A(x^k) & 0 \end{bmatrix} \begin{bmatrix} d_x - p_x \\ d_\lambda - p_\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ -c(x^k) \end{bmatrix} + \begin{bmatrix} 0 \\ (-1/\mu_k)p_\lambda \end{bmatrix} = -\frac{1}{\mu_k}\begin{bmatrix} 0 \\ \lambda^k + p_\lambda \end{bmatrix}.$$

Since the coefficient matrix is eventually nonsingular (we showed this last time), we can say that

$$\left\| \begin{bmatrix} d_x - p_x \\ d_\lambda - p_\lambda \end{bmatrix} \right\| = O(\mu_k^{-1})\|\lambda^k + p_\lambda\|,$$

from which we can show by some standard manipulation that for μ_k sufficiently large, we have

$$\left\| \begin{bmatrix} d_x - p_x \\ d_\lambda - p_\lambda \end{bmatrix} \right\| = O(\mu_k^{-1})\|\lambda^k + d_\lambda\| \approx O(\mu_k^{-1})\|\lambda^*\|.$$

Advantages: P_μ is as smooth as f and c_i are, and under LICQ and 2oS, $x(\mu)$ near x^* is a strict local minimizer of P_μ. Disadvantages: systematic offset in $c_i(x(\mu))$, leading to error of $O(1/\mu)$ in $x(\mu)$. Need $\mu \sim 1/\epsilon$ to drive error down to size ϵ. Ill conditioning in $\nabla^2 P_\mu$ (see below) means that first-order methods perform poorly. Newton is better, but still may get numerical issues in solving the Newton linear system.

Look closely at Hessian $\nabla^2 P_\mu(x)$ for x near x^*. If LICQ and 2OS hold, have m eigenvalues of size $O(\mu)$ and the other $n - m$ of size $O(1)$.

Prove nonsingularity of Hessian $\nabla^2 P_\mu$ via 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 $i = 1, 2, \ldots$ there is w_i with $\|w_i\| = 1$ such that $w_i^T (H + iAA^T) w_i \le 0$. By compactness of unit ball, we have subsequence $\mathcal{S}$ with $w_i \to w^*$ for some w^* with $\|w^*\| = 1$. Since

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

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

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

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

Lecture 22. (3/20/15; 60 min)

Quadratic penalty on inequality constraints. Constraints $c_i(x) \ge 0$ are penalized via $\min(c_i(x), 0)^2$. Leads to discontinuity in second derivatives. Alternatively, reformulate as

$$\min f(x) \text{ s.t. } c(x) - s = 0, \; s \ge 0$$

leading to penalty formulation:

$$\min_{s \ge 0, x} P_\mu(x, s) := f(x) + \frac{\mu}{2} \|c(x) - s\|_2^2.$$

Could apply gradient projection (with second-order enhancements) to this. Show that if we eliminate s explicitly, we simply get the naive version above, with discontinuous second derivatives.

Lagrangian-based. Motivate augmented Lagrangian via quadratic penalty. Show that introduction of Lagrange multiplier estimates into the penalty could remove the “systematic offset” effect.

Augmented Lagrangian. Work through details for equality-constrained case. See text, Section 17.3.

Show that multipliers are good estimates, Hessian of the augmented Lagrangian should be positive 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.)

Discuss Theorem 17.6.

Lecture 23. (3/23/15; 60 min)

AL for inequality constraints. Reformulate using bounds, as in Lancelot, and as we did for quadratic penalty.

Note (p.524 of the book) that there is a more complicated but more compact form of AL for inequality constraints, that is obtained by minimizing $\mathcal{L}_A(x, s, \lambda^k, \mu_k)$ over $s \ge 0$ (for which a closed-form solution can be found) then eliminating s by substituting its optimal value into $\mathcal{L}_A$ to obtain a more compact function of (x, λ^k, μ_k) alone.

Discuss ADMM (“alternating-direction method of multipliers”) applied to

$$\min_{x_1, x_2} f_1(x_1) + f_2(x_2) \text{ s.t.} A_1 x_1 + A_2 x_2 = b.$$

Aug Lagr is

$$\mathcal{L}_A(x_1, x_2, \lambda, \mu) = f_1(x_1) + f_2(x_2) - \lambda^T(A_1x_1 + A_2x_2 - b) + \frac{\mu}{2}\|A_1x_1 + A_2x_2 - b\|_2^2.$$

Main subproblem in implementing AL is

$$\min_{(x_1,x_2)} \mathcal{L}_A(x_1, x_2, \lambda^k, \mu_k),$$

but there is coupling between x_1 and x_2 via the penalty term $\|A_1x_1 + A_2x_2 - b\|_2^2$. In ADMM, we minimize wrt x_1 and x_2 separately and sequentially:

$$\begin{aligned} x_1^{k+1} &:= \arg\min_{x_1} \mathcal{L}_A(x_1, x_2^k, \lambda^k, \mu_k), \\ x_2^{k+1} &:= \arg\min_{x_2} \mathcal{L}_A(x_1^{k+1}, x_2, \lambda^k, \mu_k), \\ \lambda^{k+1} &:= \lambda^k - (1/\mu_k)(A_1x_1^{k+1} + A_2x_2^{k+1} - b). \end{aligned}$$

This approach makes sense when it is much easier to minimize $f_1(x_1)+$ (convex quadratic in x_1) and $f_2(x_2)+$ (convex quadratic in x_2) than to minimize $f_1(x_1) + f_2(x_2)+$ (convex quadratic in (x_1, x_2)).

Example:

$$\min_{x\in\Omega} f(x) + h(x),$$

where Ω is closed and convex. Formulate as:

$$\min_{x\in\Omega, z\in\Omega} f(x) + h(z) \ \text{ s.t. } \ x = z.$$

Aug Lagr is

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

Example:

$$\min_{x\in\Omega_1\cap\Omega_2} f(x)$$

can be reformulated as

$$\min_{x_1\in\Omega_1, x_2\in\Omega_2} f(x_1) \ \text{ subject to } x_1 = x_2,$$

for which the ADMM steps are

$$\begin{aligned} x_1^{k+1} &:= \min_{x_1\in\Omega_1} f(x_1) - (\lambda^k)(x_1 - x_2^k) + \frac{\mu_k}{2}\|x_1 - x_2^k\|_2^2, \\ x_2^{k+1} &:= \min_{x_2\in\Omega_2} -(\lambda^k)(x_1^k - x_2) + \frac{\mu_k}{2}\|x_1^{k+1} - x_2\|_2^2, \\ \lambda^{k+1} &:= \lambda^k - (1/\mu_k)(x_1^k - x_2^k). \end{aligned}$$

Nice application of this is to "doubly nonnegative" matrix optimization, where

$$\Omega_1 = \{X \in S\mathbb{R}^{n\times n} \ : \ X \succeq 0\}, \quad \Omega_2 = \{X \in S\mathbb{R}^{n\times n} \ : \ X \geq 0\}.$$

Apply ADMM to the following "consensus optimization" problem:

$$\min_x \sum_{i=1}^m f_i(x),$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$. Replicate x, but maintain a "master copy:"

$$\min_{x_1,x_2,\dots,x_m,x} \sum_{i=1}^m f_i(x_i) \;\text{ s.t. }\; x = x_i,\; i = 1, 2, \dots, p,$$

We have

$$\mathcal{L}_A(x, x_1, x_2, \dots, x_m, \lambda_1, \lambda_2, \dots, \lambda_m, \mu) := \sum_{i=1}^m f_i(x_i) - \sum_{i=1}^m \lambda_i^T (x - x_i) + \frac{\mu}{2} \sum_{i=1}^m \|x - x_i\|_2^2.$$

Minimization wrt x_i, $i = 1, 2, \dots, m$ can be performed concurrently:

$$\min_{x_i} f_i(x_i) - \lambda_i^T (x - x_i) + \frac{\mu}{2} \|x - x_i\|^2, \;\; i = 1, 2, \dots, p.$$

Minimization wrt x can be done in closed form: The updated x must satisfy

$$-\sum_{i=1}^m \lambda_i^{k+1} + \mu_k \sum_{i=1}^p (x - x_i^{k+1}) = 0,$$

leading to

$$x^{k+1} = \frac{1}{p} \sum_{i=1}^p (x_i^{k+1} + \lambda_i^{k+1}/\mu_k).$$

The update to λ_i is

$$\lambda_i^{k+1} \leftarrow \lambda_i^k - \mu_k (x^{k+1} - x_i^{k+1}), \;\; i = 1, 2, \dots, p.$$

(MIDTERM EXAM GIVEN AS A TAKE-HOME OVER 3/23/15-3/24/15.)**

Lecture 24. (3/25/15; 60 min)

Discuss merit functions. (Section 15.4) Can use them to (a) judge quality of a step generated by other means; (b) actually minimize then directly.

(b) only really makes sense if the merit function is exact i.e. its local min correspond to KKT points of the nonlinear program.

Quadratic penalty (smooth) is not exact for finite μ — there is an "offset" of size $O(1/\mu)$. But can argue that it is close to exact for large μ, which is how it becomes the basis for an algorithm.

Fletcher's augmented Lagrangian — call it $\mathcal{L}_F(x, \mu)$. $\lambda(x)$ is a function of x. Write details for equality constrained case. Take gradient and Hessian and outline why x^* satisfying LICQ, 2OS is indeed a local minimizer of $\mathcal{L}_F$ for μ sufficiently large. Disadvantage: Need ∇c_i to evaluate $\lambda(x)$ and thus $\mathcal{L}_F$. Difficult to compute $\nabla \mathcal{L}_F$ and $\nabla^2 \mathcal{L}_F$ since they require $\nabla \lambda(x)$ and $\nabla^2 \lambda(x)$. Not clear how to extend to inequality constraints.

Also discussed the "norm of KKT:"

$$\left\| \begin{bmatrix} \nabla \mathcal{L}(x,\lambda) \\ \min(c_i,\lambda_i) \end{bmatrix} \right\|.$$

This depends on λ too, but is OK for a "primal-dual" algorithms that generate estimates of λ as well as x at each iteration. Extend to inequality constraints by looking for functions that are zero when KKT conditions are satisfied and positive otherwise. For constraints $c_i(x) \geq 0$, $i = 1, 2, \ldots, m$,

Discussed ℓ_1 penalty too. Advantage: exact. Disadvantage: nonsmooth. Do this example:

$$\min\, x \;\text{ s.t. }\; x - 1 = 0.$$

Construct ℓ_1 penalty function $\phi_\mu(x) = x + \mu|x-1|$ and plot explicitly for $\mu = 1/2$ and $\mu = 2$.

Is it possible to find an exact *smooth* penalty function of the form $\phi(x) = f(x) + h(c(x))$, where $h(c)$ is zero when x is feasible and positive otherwise? Answer: NO! If such a function did exist we could apply it to the example above to get $\phi(x) = x + h(x-1)$. Because h is required to be smooth, along with $h(x-1) = 0$ if $x = 1$ and positive if $x \neq 1$, we have $h'(x-1) = 0$ when $x = 1$. Thus $\phi'(1) = 1$, so that $x^* = 1$ could not be a minimizer of the penalty function. So it cannot be exact, and the existence of a smooth exact penaly function of this form is contradicted.

Lecture 25. (3/27/15; 60 min)

Do a 2d example of exact penalty function, graphically. (See example in book of $\min\, x_1 + x_2$ s.t. $x_1^2 + x_2^2 - 2 = 0$.)

Do directional derivatives: formula (A.51). Work through details of finding directional derivatives for $|\cdot|$ and $\max(-c, 0)$, and thus of $\|c(x)\|_1$ and $\|\max(c(x), 0)\|_1$ (using chain rule).

Do Theorem 17.4 (p.509) for the case of $\hat{x}$ feasible. Show that KKT are satisfied. (Thus, referring back to Theorem 12.3, 1oN conditions are satisfied if a CQ holds as well.)

(SPRING BREAK **)**

Lecture 26. (4/6/15; 60 min)

Define subgradients of a convex nonsmooth function.

$$\partial\psi(x) = \{d \,|\, \psi(z) \geq \psi(x) + d^T(z-x)\} \quad \text{for all } z.$$

Draw pictures. Derive $\partial\psi(x)$ for $\psi(\cdot) = |\cdot|$ and $\psi(\cdot) = \|\cdot\|_1$ and $\psi(\cdot) = \|\max(\cdot, 0)\|_1$.

Note that subgradient is closely related to the normal cone to the epigraph of the function, i.e. defining

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

we have

$$(d, -1) \in N_{\text{epi}(f)}(x, \psi(x)) \;\Leftrightarrow\; d \in \partial\psi(x).$$

Proof: $(d,-1) \in N_{\text{epi}(f)}(x,\psi(x))$ means that for all $(z, f(z))$ we have $d^T(z-x) - (f(z) - f(x)) \leq 0$ which implies $f(z) \geq f(x) + d^T(z-x)$, so that $d \in \partial\psi(x)$. Conversely, for $d \in \partial\psi(x)$, we have $d^T(z-x)-(f(z)-f(x)) \le 0$ so that $d^T(z-x)-(t-f(x)) \le 0$ for all $t \ge f(z)$, thus $(d,-1)$ satisfies the definition of $N_{\mathrm{epi}(f)}(x,\psi(x))$ since (z,t) is an arbitrary element of epi(f).

Chain rule for subdifferential calculus.

Optimality conditions for convex nonsmooth functions: $0 \in \partial\psi(x)$.

Proved that for equality constrained problem the ℓ_1 penalty function is exact when $\nu \ge \|\lambda^*\|_\infty$. Did this by comparing KKT conditions with optimality conditions for the ℓ_1 penalty function.

Talked about algorithms for nonsmooth functions. Can't just pick an arbitrary subderivative and search in this direction, as it may not be a descent direction. Illustrated this fact with a 2D example with contours. The negative of a *minimum-norm subgradient vector* is a descent direction, however. To show this we use Theorem 23.4 of Rockafellar's *Convex Analysis*, which says that

$$D(\psi(x);p) = \sup_{s\in\partial\psi(x)} s^T p.$$

(The rest of the proof is an exercise.)

Lecture 27. (4/8/15; 60 min)

Hand-wave about the extension to nonconvex nonsmooth: generalized gradients. Clarke's generalized derivative[1] starts with the following generalized directional derivative:

$$f^\circ(x;v) = \limsup_{y\to x} \inf_{t\downarrow 0} \frac{f(y+tv)-f(y)}{t}$$

and defines the generalized gradient to be

$$\partial f(x) := \{\xi \mid f^\circ(x;v) \ge \langle \xi, v\rangle,\ \forall v \in \mathbf{R}^n\}.$$

Talked about Sℓ_1LP and Sℓ_1QP approach. (p. 549–550) Formulation of the subproblems as LPs and QPs. Use trust regions. A lot in common with SLP and SQP, but in this case, we are guaranteed to have a feasible point.

Lecture 28. (4/10/15; 60 min)

p.437–440: Alternative: Filter, based on two "objectives:" the barrier part $f(x)-\mu\sum\log s_i$, and the equality constraint violation part $\|(c_I(x)-s, c_E(x))\|$.

Maratos effect (Sec 15.5). Work through the example. Defeats the obvious merit and filter approaches. (Fletcher's augmented Lagrangian is immune to it, however.) Main reason is that steps based on constraint linearizations fail to account well enough for curvature in the constraints.

Outline remedies for Maratos (p. 443-446).

- Watchdog: take some provisional steps and seek a decrease in merit function over multiple iterations; if the decrease is not seen, backtrack to the latest "good" iterate and decrease steplength or otherwise generate a more conservative step.
- Second-order correction. (See also p.543-544.) Use KKT to show that indeed (15.36) is the min-norm solution of the linearized constraint.

[1] See F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, 1990.

Lecture 29. (4/13/15; 60 min)

Details of second-order correction, for the case of equality constraints $c(x) = 0$ for $c : \mathbb{R}^n \to \mathbb{R}^m$. The normal step p_k that satisfies linearized gradient equations has

$$c(x_k) + A(x_k)p_k = 0, \quad p_k = O(\|x_k - x^*\|).$$

In second-order correction we seek an approximate solution to $c(x_k + p_k + \hat{p}_k) = 0$ for which $\|\hat{p}_k\|$ is minimized. We can't solve this exactly because of nonlinearity of c, so instead note that

$$\begin{aligned} c(x_k + p_k + \hat{p}_k) &= c(x_k + p_k) + A(x_k + p_k)\hat{p}_k + O(\|\hat{p}_k\|^2) \\ &= c(x_k + p_k) + A(x_k)\hat{p}_k + O(\|p_k\|\|\hat{p}_k\|) + O(\|\hat{p}_k\|^2), \end{aligned}$$

so if we choose $\hat{p}_k$ to satisfy $c(x_k + p_k) + A(x_k)\hat{p}_k$ we will have $c(x_k + p_k + \hat{p}_k) = O(\|p_k\|\|\hat{p}_k\|) + O(\|\hat{p}_k\|^2)$. Under LICQ assumption, and assuming that $\hat{p}_k$ is chosen to have minimal norm, we have

$$\|\hat{p}_k\| = O(\|c(x_k + p_k)\|) = O(\|p_k\|^2) = O(\|x_k - x^*\|^2),$$

so it follows that $c(x_k + p_k + \hat{p}_k) = O(\|x_k - x^*\|^3)$.

Log-barrier function (Sec 19.6, p. 583) for inequality constrained problem

$$\min f(x) \text{ s.t. } c_i(x) \geq 0, \quad i = 1, 2, \dots, m.$$

Show that optimality conditions for this function are related to KKT. Again, Hessian is positive definite when 2oS conditions are satisfied.

Show that $c_i(x(\mu)) \approx \mu/\lambda_i^*$ where $x(\mu)$ denotes minimizer of

$$P_\mu(x) := f(x) - \mu \sum_{i=1}^m \log c_i(x).$$

Example:

$$\min x \text{ s.t. } x \geq 1,$$

Find log barrier minimizer $x(\mu)$, examine Hessian of log barrier.

See also Example 19.1 from p.585.

Newton-like methods can be used to minimize P_μ but we need to restrict the Newton steps to ensure that $c_i(x) > 0$ at every candidate iterate, so need to curtail steps to ensure that this happens.

Lecture 30. (4/15/15; 60 min)

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

This one has an ill conditioned Hessian.

Returning to the example above: $\min x$ s.t. $x \geq 1$, we have

$$P_\mu(x) = x - \mu \log(x - 1), \quad \nabla P_\mu(x) = 1 - \frac{\mu}{x - 1}, \quad \nabla^2 P_\mu(x) = \frac{\mu}{(x - 1)^2}.$$

Thus the Newton step is

$$x^+ = x - \nabla^2 P_\mu(x)^{-1} \nabla P_\mu(x) = x - \frac{(x-1)^2}{\mu}\left[1 - \frac{\mu}{x-1}\right] = 2x - 1 - \frac{(x-1)^2}{\mu}.$$

Even if we start quite close to the minimizer $x(\mu) = 1+\mu$, Newton's method might give a totally inappropriate step. For example $x = 1+2\mu$ gives $x^+ = 1$ and $x = 1+3\mu$ gives something that violates the constraint. Domain of convergence of Newton's method is small, the quadratic approximation on which it's based does not capture the true behavior of the function except in a small neighborhood of $x(\mu)$.

Show generically that the log-barrier function has an ill-conditioned Hessian. Under KKT, 2oS, LICQ, Hessian has $|\mathcal{A}^*|$ eigenvalues of $O(1/\mu)$ and $n - |\mathcal{A}^*|$ eigenvalues of $O(1)$. Proved this.

Primal-dual interior-point. See Chapter 19. Newton on perturbed KKT. Do it for inequality constraints only. Introduce Lagrange multipliers $\lambda_i = \mu/c_i(x)$ explicitly into the formation. Write the approximate KKT conditions as a system of $m+n$ equations in $m+n$ unknowns. Stress that the positivity conditions $\lambda_i > 0$ and $c_i(x) > 0$ are important. If we neglect these we may converge to a point (x^*, λ^*) that bears no resemblance to a solution. So we have a constrained system of nonlinear equations to which we can apply a version of Newton's method, modified to maintain positivity.

how the variant with slack s, so that the variables are (x, s, λ). Can add equality constraint easily. Restate KKT conditions for this form. Jacobian is still nonsingular near (x^*, s^*, λ^*). Note that it's easy to curtail steps in this formulation. Outline the basic method: Sec 19.2. Steps are from perturbed KKT. Restrict step length to maintain $(s, \lambda) > 0$, Use norm of modified KKT to decide when to stop iterating for each value of μ.

Why use path-following with parameter μ rather than just applying Newton's method for nonlinear equations to the system with $\mu = 0$? Avoid spurious solutions, Newton has larger convergence domain for larger μ.

Lecture 31. (4/17/15; 60 min)

Compare with KKT conditions for slack form of the barrier function:

$$\min_{x,s} f(x) - \mu \sum_{i=1}^{m} \log s_i \;\text{ s.t. }\; c(x) - s = 0,$$

which are

$$\nabla f(x) - \mu \sum_{i=1}^{n} \lambda_i \nabla c_i(x), \quad c(x) - s = 0, \;\; -\mu S^{-1} e + \lambda = 0.$$

If we scale the last equation by S we recover the central path equation for the primal-dual formulation.

Restrict steplength to keep s and λ strictly positive. Explain why, for small μ, the steplengths are not going to be much different from 1, hence rapid convergence of Newton-based methods will still be seen.

Motivation for using a decreasing sequence of μ: The problems with large μ are easier to solve. (Newton has a bigger convergence domain.) Slow decrease of μ keeps starting points in the Newton convergence domain.

For Theorem 19.1, prove the preliminary result that if there is a seq of matrices $B_k \in \mathbb{R}^{n\times t}$ with $n \geq t$ converging to $\hat{B}$ with full column rank, and sequences of vectors $h_k \in \mathbb{R}^n$ converging to $\hat{h}$, and $z_k \in \mathbb{R}^t$ such that $B_k z_k - h_k \to 0$, then $\{z_k\}$ has a unique limit $\hat{z} \in \mathbb{R}^t$ satistfying $\hat{B}\hat{z} = \hat{h}$. Proof: $B_k z_k - h_k = e_k$ with $e_k \to 0$. By full rank we have that $z_k = (B_k^T B_k)^{-1} B_k^T (h_k + e_k) \to (\hat{B}^T\hat{B})^{-1}\hat{B}^T\hat{h} = \hat{z}$.

Proceed with proof of Theorem 19.1.

Give symmetric form and compressed form (19.15) of the linear system at each PDIP iteration.

Lecture 32. (4/20/15; 60 min)

Discuss linear algebra issues. Which form is better? Depends on fill-in, to some extent. Ill conditioning of some forms - benign and less benign.

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

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.

Briefly discuss line search algorithm p. 577 — a little more specific than Algorithm 19.1. 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. And I think that the linear set of feasible points is not necessarily tangent to the parabola. Also the second-last term $X^{(1)}$ on page 587 should be $s_1^{(1)}$.

Lecture 33. (4/22/15; 60 min)

Interior-Point Methods for Linear Programming. Material from "Primal-Dual Interior-Point Methods" (1997).

Primal and Dual LP, KKT conditions, central path equations defined by $\tau > 0$.

Discussed key steps: Newton step on central path equations and line search to stay in neighborhood.

Discussed PD framework (Chapter 1, page 8). Noted relationships to the same framework for nonlinear programming.

Defined central-path neighborhoods and discussed their properties. Illustrated with the "xs" diagram.

Defined Algorithm LPF.

Lecture 34. (4/24/15; 60 min)

Full analysis of LPF from Chapter 5, up to polynomial complexity result. Proved all the technical results.

In 2013: Defined Algorithm SPF. Derived complexity result: $O(\sqrt{n}\log\epsilon)$ iterations, or effort of $O(n^{3.5}\log\epsilon)$.

Lecture 35. (4/27/15; 60 min)

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:

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

where g the range of F and K are in some space (e.g. a Euclidean space or the space $S\mathbf{R}^{n\times n}$ 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:

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

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

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

- Semidefinite programming:

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

  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: $A \bullet B = \mathrm{tr}(A^T B) = \sum_{i,j=1}^n A_{ij} B_{ij}$. Then

$$\text{(SDP-P)} \qquad \min_X C \bullet X \text{ s.t. } A_i \bullet X = b_i, \; i = 1, 2, \dots, m, \; X \succeq 0, \tag{0.25}$$

where A_i and C and X are all in $S\mathbf{R}^{n\times n}$, and

$$\text{(SDP-D)} \qquad \max b^T y \text{ s.t. } \sum_{i=1}^{m} y_i A_i \preceq C. \tag{0.26}$$

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

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

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

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

define $C := \mathrm{diag} c$, $A_i := \mathrm{diag} A_{i\cdot}$, E_{ij} is the matrix in $S\mathbf{R}^{n\times n}$ with all zeros except for 1 in the (i,j) and (j,i) position, and define the SDP as:

$$\min_{X \in S\mathbf{R}^{n\times n}} C \bullet X \text{ s.t. } A_i \bullet X = b_i, \;\; X \succeq 0, \;\; E_{ij} \bullet X = 0, \;\; \text{for all } i \ne j.$$

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

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. **CODE: SeDuMi** and **SDPT3**. Now have **CVX** (see `cvxr.com`) — Matlab add-on that includes nice modeling framework and calls SeDuMi by default.

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

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

which can be written as

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

which clearly has (SDP-D) form.

Example (Boyd and Vandenberghe): Estimate a vector x 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 $v^{(1)}, \ldots, v^{(M)}$, so as to design an experiment that is "maximially 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:

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

(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 (0.26). Setting

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

we find that this constraint is equivalent to

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

Similarly, the nonnegativity constraints $\lambda_i \geq 0$, $i = 1, 2, \ldots, M$ can be expressed by setting F_0 to be the $M \times M$ matrix of zeros, and each F_i, $i = 1, 2, \ldots, M$ to be the $M \times M$ matrix whose elements are all zero except for a 1 in the (i, i) position. CVX does not require these kinds of reformulations; it does them automatically.

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

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

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

$$C := \begin{bmatrix} \alpha & 0.5c^T \\ 0.5c & 0.5Q \end{bmatrix}, \; A_i = \begin{bmatrix} -d_i & 0.5H_{i\cdot} \\ 0.5H_{i\cdot}^T & 0 \end{bmatrix},$$

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

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

(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.)

Lecture 36. (4/29/15; 60 min)

Revisit nonconvex QP formulation, highlighting $E_i \bullet \tilde{X} \geq 0$, $i = 1, 2, \ldots, n$ to formulate the $x \geq 0$ constraint.

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 $\tilde{X}$.

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 ith unit vector and E_i has 1 in the (i, i) position and zeros elsewhere. Could also include the constraints $0 \le x_i \le 1$.

Other wacky formulation tricks for SDP relaxation of QP:

- Square the constraint $H_{i\cdot}x = d_i$ to get $(H_{i\cdot}x)^2 = d_i^2$ 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 $x_j(H_{i\cdot}x - d_i) = 0$ 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: $(H_{i\cdot}x - d_i)(H_{k\cdot}x - d_k) = 0$.

Duality theory for SDP is more complex than linear programming.

Weak duality is comparatively easy.

THEOREM 0.13 (Weak Duality). *If X is feasible for (0.25) and (y, S) is feasible for (0.26), then*

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

Proof. We have by substitution from the constraints in (0.25) and (0.26) 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 trace(PQ) = 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

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

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

Lecture 37. (5/1/15; 60 min)

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

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

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:

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

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

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

From the equality constraints, X must have the form

$$X = \left[\begin{array}{cc} 0 & 1 \\ 1 & x_{22} \end{array} \right],$$

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

$$S = \left[\begin{array}{cc} -y_1 & -y_2 \\ -y_2 & 0 \end{array} \right] \succeq 0,$$

which implies that $y_1 \le 0$ 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.

$$\begin{aligned} F(P) &\overset{\text{def}}{=} \{X \in S\mathbb{R}^{n\times n} \,|\, A_i \bullet X = b_i, \; i = 1, 2, \ldots, m, \;\; X \succeq 0\}; \\ F^0(P) &\overset{\text{def}}{=} \{X \in F(P) \,|\, X \succ 0\}; \\ F(D) &\overset{\text{def}}{=} \{(y, S) \in \mathbb{R}^m \times S\mathbb{R}^{n\times n} \,|\, \sum_i A_i y_i + S = C, \; S \succeq 0\}; \\ F^0(D) &\overset{\text{def}}{=} \{(y, S) \in F(D) \,|\, S \succ 0\}. \end{aligned}$$

Logarithmic Barrier Functions and the SDP Central Path

We define a barrier function on $S\mathbb{R}^{n\times n}$ as follows:

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

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

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

As a barrier function should, $f(X)$ approaches $+\infty$ as X approaches the boundary of the set $S\mathbb{R}^{n\times n}$. 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 (0.25), in which the positive semidefiniteness constraint is replaced by a barrier function term, parametrized by a coefficient ν:

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

We hope that the solution $X(\nu)$ of this problem approximates the solution of (0.25), and that $X(\nu) \to X^*$ as $\nu \downarrow 0$.

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

The log barrier subproblem for the dual (0.26) is

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

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

Lecture 38. (5/4/15; 60 min)

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, $i = 1, 2, \ldots, m$. 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:

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

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

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

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

Work through optimality conditions for the dual barrier subproblem $BD(\nu)$ also, and show that they are the same: $CPE(\nu)$.

Strict convexity of $BP(\nu)$ suggests that the solution $X(\nu)$ is unique for each $\nu > 0$. (This is made rigorous in the theorem below; we also need that $X(\nu)$ lies in some compact set.) Similarly, from the fact that $BD(\nu)$ is strictly concave in S, we can deduce (subject to similar caveats) that the S component of the solution $(y(\nu), S(\nu))$ of $BD(\nu)$ is unique. For uniqeness of the $y(\nu)$ component, we need *linear independence* of the matrices $\{A_i\}_{i=1,2,\ldots,m}$.

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

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

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

Following LP, it seems plausible that we can design an algorithm whose search directions are the Newton directions for $CPE(\nu)$, with restrictions on the step length to ensure that X and S stay (strictly) positive definite. Theorem 0.15 below establishes the 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 $\mathbb{R}^{2n+m}$.) Unfortunately, it is *not* true for $CPE(\nu)$! The domain of these equations is $S\mathbb{R}^{n\times n} \times \mathbb{R}^m \times S\mathbb{R}^{n\times n}$ and their range is $S\mathbb{R}^{n\times n} \times \mathbb{R}^m \times \mathbb{R}^{n\times n}$ — 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

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

for which both range and domain are $S\mathbb{R}^{n\times n} \times \mathbb{R}^m \times S\mathbb{R}^{n\times n}$. Another possible symmetrization of the last equation is

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

Another possible symmetrization of the last equation was proposed by Alizadeh, Haeberley, Overton:

$$XS + SX = 2\nu I. \tag{0.27}$$

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

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

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

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

then the MZ directions are simply the AHO directions in the transformed space. The first two equations in CPE(ν) 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

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

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

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

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 $\hat{X} = \hat{S}$.

The three cases above correspond to the following choices of $\mathcal{E}$ and $\mathcal{F}$ in the system (0.28) below:

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

For these symmetrizations, let us consider under what conditions the linearized system (the Newton equations) has a solution, and how this solution is defined in general.

Lemma 0.14. *(Todd [?]) Let $\mathcal{E}$ and $\mathcal{F}$ be two operators that map $S\mathbb{R}^{n\times n}$ 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, $i = 1, 2, \ldots, m$ are linearly independent. Then for any $P, R \in S\mathbb{R}^{n\times n}$ and any $q \in \mathbb{R}^m$, the solution to the system*

$$\begin{array}{ccccccc} & & \mathcal{A}^*v & + & W & = & P, \\ \mathcal{A}U & & & & & = & q, \\ \mathcal{E}U & & & + & \mathcal{F}W & = & R \end{array} \tag{0.28}$$

is given uniquely by

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

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 for 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. □

Return to the AHO symetrization (0.27). After linearization we have

$$\frac{1}{2}(\Delta XS + S\Delta X + X\Delta S + \Delta SX) = \nu I - \tfrac{1}{2}(XS + SX). \tag{0.29}$$

This corresponds to choosing

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

in Lemma 0.14. We obtain the step by adding the feasibility conditions to (0.29) from (0.28), 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 0.14 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.

Let's discuss details of AHO approach. The key tool used here is solution of the Lyapunov equation with coefficient S (symmetric positive definite) and different right-hand sides H (symmetric), that is find U such that

$$US + SU = H.$$

The cost of finding U is $O(n^3)$.

Write the AHO equations as follows:

$$\sum_{i=1}^{m} \Delta y_i A_i + \Delta S = P \tag{0.30a}$$

$$A_i \bullet \Delta X = q_i, \;\; i = 1, 2, \ldots, m, \tag{0.30b}$$

$$\Delta X S + S \Delta X + X \Delta S + \Delta S X = R, \tag{0.30c}$$

for given right-hand sides P, q, R. By pre- and post-multiplying (0.30a) by X and adding, we obtain

$$\sum_{i=1}^{m} \Delta y_i (A_i X + X A_i) + (\Delta S) X + X (\Delta S) = PX + XP.$$

By combining this formula with (0.30c) we obtain

$$\sum_{i=1}^{m} \Delta y_i (A_i X + X A_i) - (\Delta X) S - S (\Delta X) = PX + XP - R. \tag{0.31}$$

For each $i = 1, 2, \ldots, m$, we solve a Lyapunov equation to find G_i such that

$$G_i S + S G_i = (A_i X + X A_i), \;\; i = 1, 2, \ldots, m. \tag{0.32}$$

By substituting into (0.31), we obtain

$$\sum_{i=1}^{m} \Delta y_i (G_i S + S G_i) - (\Delta X) S - S (\Delta X) = PX + XP - R.$$

which simplifies to

$$(\sum_{i=1}^{m} \Delta y_i G_i - \Delta X) S + S (\sum_{i=1}^{m} \Delta y_i G_i - \Delta X) = PX + XP - R. \tag{0.33}$$

Thus, by solving the following Lyapunov equation for U:

$$US + SU = (PX + XP) - R, \tag{0.34}$$

we have the identification

$$U = \sum_{i=1}^{m} \Delta y_i G_i - \Delta X. \tag{0.35}$$

We now combine this equation with (0.30b), taking products with A_j, $j = 1, 2, \ldots, m$, we obtain

$$A_j \bullet U = \sum_{i=1}^m \Delta y_i A_j \bullet G_i - A_j \bullet \Delta X = \sum_{i=1}^m \Delta y_i A_j \bullet G_i - q_j.$$

This is an $m \times m$ system of equations $M\Delta y = \tilde{q}$, where

$$M_{ji} = A_j \bullet G_i, \quad \tilde{q}_j = q_j + A_j \bullet U.$$

This system can be solved for Δy, then ΔX can be recovered by substituting into (0.35) and ΔS can be recovered from (0.30a).

To summarize, the procedure for solving (0.30) (with operation counts) is as follows:

- Solve m Lyapunov equations for G_i: $G_i S + S G_i = A_i X + X A_i$, $i = 1, 2, \ldots, m$. (Operations: $O(mn^3)$);
- Solve Lyapunov equation for U: $US + SU = (PX + XP) - R$. (Operations: $O(n^3)$);
- Form $M \in \mathbb{R}^{m \times m}$ and $\tilde{q} \in \mathbb{R}^m$ from

  $$M_{ij} = A_j \bullet G_i, \quad \tilde{q}_j = q_j + A_j \bullet U.$$

  (Operations: $O(m^2 n^2)$)
- Solve $M\Delta y = \tilde{q}$. (Operations: $O(m^3)$)
- Set $\Delta X = \sum i = 1^m \Delta y_i G_i - U$. (Operations: $O(mn^2)$)
- Set $\Delta S = P - \sum_{i=1}^m \Delta y_i A_i$. (Operations: $O(mn^2)$)

The dominant terms are the computation of M, which is an $O(m^2 n^2)$ operation, and the computation of the G_i, which is $O(mn^3)$. (Which of these is dominant depends on the relative sizes of m and n.)

Theorem 0.15. *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(ν). Further, $X(\nu)$ is the unique solution to BP(ν) and $(y(\nu), S(\nu))$ is the unique solution to BD(ν). Finally, if the assumption of strict feasibility fails, then CPE(ν), BP(ν), and BD(ν) have no solutions.*

(Proof skipped in class, 2013.)

Proof. (Todd [**?**]) 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: $\sigma \overset{\text{def}}{=} \lambda_{\min}(\hat{S}) > 0$. We have that $\hat{X}$ is feasible for BP(ν), 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(ν) has the same set of solutions as the following subproblem, which we call $BP'(\nu)$:

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

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 $\lambda_j(X)$, $j = 1, 2, \ldots, n$. 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,

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

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

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

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

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

(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 $j = 1, 2, \ldots, n$, 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 $j = 1, 2, \ldots, n$. 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 (0.25) 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. □

THEOREM 0.16. *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 (0.25) and (y^*, S^*) solves (0.26). In fact this is true, but hard to prove.

THEOREM 0.17. *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 now show that the solution set for the primal is nonempty and bounded, using similar techniques as in the proof of Theorem 0.15. 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:

$$\min_{X \in S\mathbb{R}^{n\times n}} \hat{S} \bullet X \quad \text{s.t.} \quad A_i \bullet X = b_i, \;\; i = 1, 2, \ldots, m, \quad \hat{S} \bullet X \le \hat{S} \bullet \hat{X}, \quad X \succeq 0. \quad (0.36)$$

Letting σ denote the smallest eigenvalue of $\hat{S}$ (as before) it is easy to show that the added constraint $\hat{S} \bullet X \le \hat{S} \bullet \hat{X}$ implies that the eigenvalues of X must be bounded by $\hat{S} \bullet \hat{X}/\sigma$. Hence, the feasible set for (0.36) is bounded. It is also nonempty because it contains $\hat{X}$. Finally it is easy to see that the objective of (0.36) is continuous over the feasible set. Hence, the problem attains its minimum, so the solution set of (0.36) (and hence (0.25)) is nonempty and bounded.

Zero duality gap follows from the theorem above about unique central path points. We have

$$C \bullet X(\nu) \ge C \bullet X^* \ge b^T y* \ge b^T y(\nu),$$

where X^* solves SDP-P and (y^*, S^*) solves SDP-D. Since we can take ν arbitrarily small, it follows that $C \bullet X^* = b^T y^*$. □

Second-order cone programming. Recall the general form of conic optimization:

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

Define K_i to be the second-order cone:

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

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

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

that is,

$$\min c^T x \;\; \text{s.t.} \;\; \|A_i x + b_i\|_2 \le c_i^T x + d_i, \;\; i = 1, 2, \ldots, m, \;\; Fx = g.$$

This is a useful formulation, but for purposes of development it is easier to consider the following standard form:

$$\min \sum_{i=1}^{m} c_i^T x_{[i]} \;\; \text{s.t.} \;\; \sum_{i=1}^{m} A_i x_{[i]} = b, \;\; x_{[i]} \in K_i, \; i = 1, 2, \dots, m,$$

where K_i is defined as above.

Recall that we calculated $N_{K_i}(0)$ in an earlier homework exercise. Optimality conditions for the formulation above can be derived from our optimality conditions for $\min_{x \in \Omega} f(x)$, where f is convex and Ω is a closed convex set, which are

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

We have

$$\begin{aligned} -c_i &\in A_i^T \lambda + v_i, \;\; \text{for some } v_i \in N_{K_i}(x_{[i]}) \\ \sum_{i=1}^{m} A_i x_{[i]} &= b, \\ x_{[i]} &\in K_i, \;\; i - 1, 2, \dots, m. \end{aligned}$$

The barrier function for a second-order cone K (with $(y,t) \in K$) would be

$$f(y,t) = -\log(t^2 - \|y\|_2^2).$$

This has derivatives:

$$\nabla f(y,t) = -\frac{2}{t^2 - \|y\|^2} \begin{bmatrix} -y \\ t \end{bmatrix},$$

$$\nabla^2 f(y,t) = -\frac{2}{t^2 - \|y\|^2} \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \frac{4}{(t^2 - \|y\|^2)^2} \begin{bmatrix} -y \\ t \end{bmatrix} \begin{bmatrix} -y \\ t \end{bmatrix}^T.$$

Exercise: show that this is positive definite.

Primal barrier function would be

$$P_\nu(x) := \sum_{i=1}^{m} c_i^T x_{[i]} - \nu \sum_{i=1}^{m} f_i(x_{[i]})$$

and our primal barrier subproblem would be

$$\min_x \; P_\nu(x) \;\; \text{s.t.} \;\; \sum_{i=1}^{m} A_i x_{[i]} = b.$$

Do some examples of SDP: QCQP, robust LP, sum-of-ℓ_2, least squares with sum-of-ℓ_2 regularizer.

Lecture 37. (5/6/13; 60 min)

Chapter 5 from "Primal-Dual Interior-Point Methods" (1997).

Proved uniqueness of solution to central path equations for all $\tau > 0$, under full row rank assumption on A.

Defined neighborhoods and analyzed their properties. Illustrated with the "xs" diagram.

Discussed key steps: Newton step and line search to stay in neighborhood.

Defined Algorithm SPF. Derived complexity result: $O(\sqrt{n}\log\epsilon)$ iterations, or effort of $O(n^{3.5}\log\epsilon)$.

Stochastic Gradient.

Setting

$$\min f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x),$$

with f convex and m large, and $x \in \mathbb{R}^n$. Example: SVM classification, where each term f_i depends on a single item of data. Evaluation of

$$\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x)$$

would require a complete scan of the data set. x is the vector of weights. Logistic regression is similar.

Could handle by sampling: choose $\mathcal{S} \subset \{1,2,\dots,m\}$ randomly with $|\mathcal{S}| \ll m$. Then apply standard methods to

$$\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f_i(x).$$

Basic stochastic gradient: At iteration k choose $i_k \in \{1,2,\dots,m\}$ randomly with equal probability. Set

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k),$$

for some steplength α_k.

Note that x_k depends on random indices $i_1, i_2, \dots, i_{k-1}$.

Assume that f is smooth, and strongly convex with modulus μ, that is, for all x and z in region of interest we have

$$f(z) \ge f(x) + \nabla f(x)^T(z-x) + \frac{1}{2}\mu\|z-x\|^2.$$

Assume also that there is not too much variation in $\nabla f_i(x)$ over $i=1,2,\dots,m$, that is, for some constant M, we have

$$\frac{1}{m}\sum_{i=1}^{m}\|\nabla f_i(x)\|_2^2 \le M$$

for x in the region of interest.

Convergence: Define $a_k = \frac{1}{2}E(\|x_k - x^*\|^2)$.

$$\begin{aligned}
&\frac{1}{2}\|x_{k+1}-x^*\|_2^2 \\
&\qquad = \frac{1}{2}\|x_k - \alpha_k\nabla f_{i_k}(x_k) - x^*\|^2 \\
&\qquad = \frac{1}{2}\|x_k - x^*\|_2^2 - \alpha_k(x_k-x^*)^T\nabla f_{i_k}(x_k) + \frac{1}{2}\alpha_k^2\|\nabla f_{i_k}(x_k)\|^2.
\end{aligned}$$

Taking expectations, get

$$a_{k+1} \le a_k - \alpha_k E[(x_k - x^*)^T \nabla f_{i_k}(x_k)] + \frac{1}{2}\alpha_k^2 M^2.$$

For middle term, have

$$\begin{aligned} E[(x_k - x^*)^T \nabla f_{i_k}(x_k)] &= E_{i_1,i_2,\ldots,i_{k-1}} E_{i_k}[(x_k - x^*)^T \nabla f_{i_k}(x_k) | i_1, i_2, \ldots, i_{k-1}] \\ &= E_{i_1,i_2,\ldots,i_{k-1}} (x_k - x^*)^T \nabla f(x_k). \end{aligned}$$

By strong convexity, have

$$(x_k - x^*)^T \nabla f(x_k) \ge f(x_k) - f(x^*) + \frac{1}{2}\mu \|x_k - x^*\|^2 \ge \mu \|x_k - x^*\|^2.$$

Hence by taking expectations, we get $E[(x_k - x^*)^T \nabla f(x_k)] \ge 2\mu a_k$. Then, substituting above, we obtain

$$a_{k+1} \le (1 - 2\mu\alpha_k) a_k + \frac{1}{2}\alpha_k^2 M^2$$

When

$$\alpha_k \equiv \frac{1}{k\mu},$$

we claim that

$$a_k \le \frac{Q}{2k}, \qquad \text{for } Q := \max\left(\|x_1 - x^*\|^2, \frac{M^2}{\mu^2} \right).$$

This is proved by the following neat inductive argument:

$$a_{k+1} \le (1 - 2/k)\frac{Q}{2k} + \frac{1}{2}\frac{1}{k^2\mu^2} M^2 \le \left(1 - \frac{2}{k} + \frac{1}{k}\right)\frac{Q}{2k} \le \left(1 - \frac{1}{k+1}\right)\frac{Q}{2k} = \frac{Q}{2(k+1)}.$$

This is slower than linear, which would be

$$a_k \le \beta^k a_0,$$

for some $\beta \in (0,1)$. Consider how many iterations K in each case are required to get $a_K \le \epsilon$ for a given threshold ϵ.

Lecture 27. (4/8/13; 50 min)

We can also get rates of approximately $1/k$ for the strongly convex case, *without* performing iterate averaging and without requiring an accurate estimate of μ. The tricks are to (a) define the desired threshold for a_k in advance and (b) use a constant step size

Recall the bound from a few slides back, and set $\alpha_k \equiv \alpha$:

$$a_{k+1} \le (1 - 2\mu\alpha) a_k + \frac{1}{2}\alpha^2 M^2.$$

Define the "limiting value" α_∞ by

$$a_\infty = (1 - 2\mu\alpha) a_\infty + \frac{1}{2}\alpha^2 M^2.$$

Take the difference of the two expressions above:

$$(a_{k+1} - a_\infty) \le (1 - 2\mu\alpha)(a_k - a_\infty)$$

from which it follows that $\{a_k\}$ decreases monotonically to a_∞, and

$$(a_k - a_\infty) \le (1 - 2\mu\alpha)^k (a_0 - a_\infty).$$

Rearrange the expression for a_∞ to obtain

$$a_\infty = \frac{\alpha M^2}{4\mu}.$$

Thus have

$$\begin{aligned} a_k &\le (1 - 2\mu\alpha)^k (a_0 - a_\infty) + a_\infty \\ &\le (1 - 2\mu\alpha)^k a_0 + \frac{\alpha M^2}{4\mu}. \end{aligned}$$

Given threshold $\epsilon > 0$, we aim to find α and K such that $a_k \le \epsilon$ for all $k \ge K$. We ensure that both terms on the right-hand side of the expression above are less than $\epsilon/2$. The right values are:

$$\alpha := \frac{2\epsilon\mu}{M^2}, \qquad K := \frac{M^2}{4\epsilon\mu^2} \log\left(\frac{a_0}{2\epsilon}\right).$$

Clearly the choice of α guarantees that the second term is less than $\epsilon/2$.

For the first term, we obtain k from an elementary argument:

$$\begin{aligned} & (1 - 2\mu\alpha)^k a_0 \le \epsilon/2 \\ \Leftrightarrow\ & k \log(1 - 2\mu\alpha) \le -\log(2a_0/\epsilon) \\ \Leftarrow\ & k(-2\mu\alpha) \le -\log(2a_0/\epsilon) \qquad \text{since } \log(1+x) \le x \\ \Leftrightarrow\ & k \ge \frac{1}{2\mu\alpha} \log(2a_0/\epsilon), \end{aligned}$$

from which the result follows, by substituting for α in the right-hand side.

If μ is underestimated by a factor of β, we undervalue α by the same factor, and K increases by $1/\beta$. (Easy modification of the analysis above.)

Underestimating μ gives a mild performance penalty.

PRO: Avoid averaging, $1/k$ sublinear convergence, insensitive to underestimates of μ. CON: Need to estimate probably unknown quantities: besides μ, we need M (to get α) and a_0 (to get K).

Lecture 28. (4/10/13; 60 min)

The choice $\alpha_k = 1/(k\mu)$ requires strong convexity, with knowledge of the modulus μ. An underestimate of μ can greatly degrade the performance of the method (see example in Nemirovski et al. 2009).

Now describe a *Robust Stochastic Approximation* approach, which has a rate $1/\sqrt{k}$ (in function value convergence), and works for weakly convex nonsmooth functions and is not sensitive to choice of parameters in the step length. (This is the approach that generalizes to *mirror descent*.)

At iteration k:

- set $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$ as before;
- set

$$\bar{x}_k = \frac{\sum_{i=1}^k \alpha_i x_i}{\sum_{i=1}^k \alpha_i}.$$

For any $\theta > 0$ (not critical), choose step lengths to be

$$\alpha_k = \frac{\theta}{M\sqrt{k}}.$$

Then $f(\bar{x}_k)$ converges to $f(x^*)$ in expectation with rate approximately $(\log k)/k^{1/2}$. The choice of θ is not critical.

The analysis is again elementary. As above (using i instead of k), have:

$$\alpha_i E[(x_i - x^*)^T \nabla f(x_i)] \le a_i - a_{i+1} + \frac{1}{2}\alpha_i^2 M^2.$$

By convexity of f (not requiring strong convexity), we have

$$f(x^*) \ge f(x_i) + \nabla f(x_i)^T (x^* - x_i),$$

thus

$$\alpha_i E[f(x_i) - f(x^*)] \le a_i - a_{i+1} + \frac{1}{2}\alpha_i^2 M^2,$$

so by summing iterates $i = 1, 2, \dots, k$, telescoping, and using $a_{k+1} > 0$:

$$\sum_{i=1}^k \alpha_i E[f(x_i) - f(x^*)] \le a_1 + \frac{1}{2} M^2 \sum_{i=1}^k \alpha_i^2.$$

Thus dividing by $\sum_{i=1} \alpha_i$:

$$E\left[\frac{\sum_{i=1}^k \alpha_i f(x_i)}{\sum_{i=1}^k \alpha_i} - f(x^*)\right] \le \frac{a_1 + \frac{1}{2} M^2 \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

By convexity, we have

$$f(\bar{x}_k) \le \frac{\sum_{i=1}^k \alpha_i f(x_i)}{\sum_{i=1}^k \alpha_i},$$

so obtain the fundamental bound:

$$E[f(\bar{x}_k) - f(x^*)] \le \frac{a_1 + \frac{1}{2} M^2 \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

By substituting $\alpha_i = \frac{\theta}{M\sqrt{i}}$, we obtain

$$\begin{aligned} E[f(\bar{x}_k) - f(x^*)] &\le \frac{a_1 + \frac{1}{2}\theta^2 \sum_{i=1}^k \frac{1}{i}}{\frac{\theta}{M} \sum_{i=1}^k \frac{1}{\sqrt{i}}} \\ &\le \frac{a_1 + \theta^2 \log(k+1)}{\frac{\theta}{M}\sqrt{k}} \\ &= M\left[\frac{a_1}{\theta} + \theta \log(k+1)\right] k^{-1/2}. \end{aligned}$$

Mirror Descent. The step from x_k to x_{k+1} can be viewed as the solution of a subproblem:

$$x_{k+1} = \arg\min_z \nabla f_{i_k}(x_k)(z - x_k) + \frac{1}{2\alpha_k}\|z - x_k\|_2^2,$$

a linear estimate of f plus a prox-term. This provides a route to handling constrained problems, regularized problems, alternative prox-functions.

For the constrained problem $\min_{x\in\Omega} f(x)$, simply add the restriction $z \in \Omega$ to the subproblem above. In some cases (e.g. when Ω is a box), the subproblem is still easy to solve.

We may use other prox-functions in place of $(1/2)\|z-x\|_2^2$ above. Such alternatives may be particularly well suited to particular constraint sets Ω.

Mirror Descent is the term used for such generalizations of the SA approaches above. Given constraint set Ω, choose a norm $\|\cdot\|$ (not necessarily Euclidean). Define the *distance-generating function* ω to be a strongly convex function on Ω with modulus 1 with respect to $\|\cdot\|$, that is,

$$(\omega'(x) - \omega'(z))^T(x - z) \geq \|x - z\|^2, \quad \text{for all } x, z \in \Omega,$$

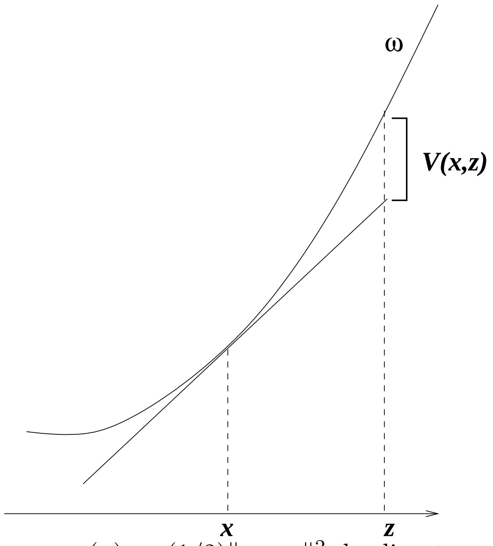
where $\omega'(\cdot)$ denotes an element of the subdifferential. Now define the *prox-function* $V(x, z)$ as follows:

$$V(x, z) = \omega(z) - \omega(x) - \omega'(x)^T(z - x).$$

This is also known as the *Bregman distance.* We can use it in the subproblem in place of $\frac{1}{2}\|\cdot\|^2$:

$$x_{k+1} = \arg\min_{z\in\Omega} \nabla f_{i_k}(x_k)^T(z - x_k) + \frac{1}{\alpha_k}V(z, x_k).$$

Bregman distance is the deviation from linearity:



For *any* Ω, we can use $\omega(x) := (1/2)\|x - \bar{x}\|_2^2$, leading to prox-function $V(x, z) = (1/2)\|x - z\|_2^2$.

For the simplex $\Omega = \{x \in \mathbf{R}^n \,:\, x \geq 0,\ \sum_{i=1}^n x_i = 1\}$, we can use instead the 1-norm $\|\cdot\|_1$, choose ω to be the entropy function

$$\omega(x) = \sum_{i=1}^n (x_i \log x_i - x_i),$$

leading to Bregman distance

$$V(x,z) = \sum_{i=1}^n z_i \log(z_i/x_i) - \sum_{i=1}^n z_i + \sum_{i=1}^n x_i$$

which is also known as the Kullback-Leibler divergence, a popular way to measure the distance between two probability distributions. (Note that for Ω given above, the last two terms cancel.)

Lecture 29. (4/12/13; 60 min)

To prove the claimed strong convexity property w.r.t. $\|\cdot\|_1$, note that from Taylor's theorem, we have

$$(\omega'(z) - \omega'(x))^T (z-x) = (z-x)^T \omega''(x)(z-x) + O(\|z-x\|^3),$$

so it suffices to prove that $(z-x)^T \omega''(x)(z-x) \geq \|z-x\|_1^2$. We have

$$\begin{aligned}(z-x)^T \omega''(x)(z-x) = \sum_{i=1}^n \frac{(z_i - x_i)^2}{x_i} &= \left(\sum_{i=1}^n \frac{(z_i-x_i)^2}{x_i}\right)(\sum_{i=1}^n x_i) \\ &\geq \left(\sum_{i=1}^n \frac{|z_i - x_i|}{x_i^{1/2}} x_i^{1/2}\right) = \|x-z\|_1^2.\end{aligned}$$

Convergence results for SA can be generalized to mirror descent.

Stochastic Dual Averaging. (Nesterov, Y., Math Programming B 120 (2009), pp. 221-259.) Define sequence $\hat{\beta}_k$ by

$$\hat{\beta}_0 = \hat{\beta}_1 = 1, \ \ \hat{\beta}_{i+1} = \hat{\beta}_i + \frac{1}{\hat{\beta}_i}.$$

We have that $\hat{\beta}_k \approx \sqrt{2k}$, specifically

$$\sqrt{2k-1} \leq \hat{\beta}_k \leq \sqrt{2k-1} + \frac{1}{1+\sqrt{3}}.$$

Algorithm is: Choose $\gamma > 0$ and set $s_0 = 0$, and choose starting point x_0. At iteration k:

- choose $i_k \in \{1, 2, \dots, m\}$ uniformly;
- set $s_{k+1} = s_k + \nabla f_{i_k}(x_k)$;
- set $x_{k+1} = x_0 - \frac{1}{\gamma \hat{\beta}_{k+1}} s_{k+1}$;

If we do primal averaging too, that is,

$$\bar{x}_k := \frac{1}{k} \sum_{i=1}^k x_i$$

we get the following convergence result (Nesterov, Theorem 7):

$$E(f(\bar{x}_k) - f^*) \leq \frac{\hat{\beta}_k}{k}\left(\gamma d(x^*) + \frac{L^2}{2\gamma}\right).$$

Here $d(x^*) = (1/2)\|x^* - x_0\|_2^2$ and L is an upper bound on $\nabla f(x)$ in the region of interest.

As in mirror descent we can choose a different prox function d, restrict the iterates to a feasible set, add a regularization terms, etc, redefining the new point x_{k+1} appropriately.