

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:

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

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

Strong Duality: There are three possibilities:

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

Same applies for the primal-dual pair:

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

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

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

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

has this dual:

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

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

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

whose dual is

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

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

Lecture 2. (1/25/13; 60 min)

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

To prove define primal-dual pair:

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

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

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

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

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

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

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

for some y_2, y_3, y_4 .

Define the LP pair:

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

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

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

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

Convex Analysis Basics. We work mostly in $\mathbb{R}^n$ and sometimes in $S\mathbb{R}^{n \times n}$. 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 $S\mathbb{R}^{n \times n}$, 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 cone. "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 $S\mathbb{R}^{n \times n}$.

Lecture 3. (1/28/13; 60 min)

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

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

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

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: define and explain.

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

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

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

Projection P onto closed convex set:

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

Recall a result proved in 726 when I last taught it (in 2008).

LEMMA 0.1.

(i) $(P(y) - z)^T(y - z) \geq 0$ for all $z \in \Omega$, with equality if and only if $z = P(y)$.

(ii) $(y - P(y))^T(z - P(y)) \leq 0$ for all $z \in \Omega$.

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

Let z be an arbitrary vector in Ω . 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

$$2(P(y) - z)^T(y - z) = \|P(y) - z\|_2^2 + [\|z - y\|_2^2 - \|P(y) - y\|_2^2]. \quad (0.1)$$

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 (0.1) in particular is zero, so we have $z = P(y)$. $\square$

(ii) implies that $y - P(y) \in N_\Omega(P(y))$.

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

Lecture 4. (1/30/13; 65 min)

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

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

Adding, we obtain

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

which after rearrangement gives

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

proving the claim.

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

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

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

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

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

so that

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

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

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

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

Tangent and Normal Cones are Polar: Proof: Simpler and nicer version (proposed by a student in 2013). We prove instead that $N_\Omega(x) = T_\Omega(x)^\circ$. Start with $N_\Omega(x) \subset T_\Omega(x)^\circ$. Taking an arbitrary $v \in N_\Omega(x)$, we wish to show that $\langle v, y \rangle \leq 0$ for all $y \in T_\Omega(x)$. By definition of $T_\Omega(x)$, there exist sequences $\alpha_k \downarrow 0$ and $y_k \rightarrow y$ such that $x + \alpha_k y_k \in \Omega$. Thus $\langle v, (x + \alpha_k y_k) - x \rangle \leq 0$ and so $\alpha_k \langle v, y_k \rangle \leq 0$. By using $\alpha_k > 0$ and taking limits, we conclude that $\langle v, y \rangle \leq 0$, as required. Now we prove that $T_\Omega^\circ \subset N_\Omega(x)$. Taking any $u \in T_\Omega^\circ$, we have that $\langle u, y \rangle \leq 0$ for all $y \in T_\Omega(x)$. Given any $z \in \Omega$, we have that $z - x \in T_\Omega(x)$. (To see this, simply define $\alpha_k = 1/k$ and $y_k \equiv y$ in the definition of $T_\Omega(x)$.) Thus $\langle u, z - x \rangle \leq 0$ for all $z \in \Omega$. Thus $u \in N_\Omega(x)$, as required.

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 $i = 1, 2, \dots, n$, so $y \leq 0$.

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

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

Second-order cone:

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

Have $N_\Omega(0) = -\Omega = \{(u, v, w) \mid w \leq -\sqrt{u^2 + v^2}\}$. 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 $z \geq 0$.

Lecture 5. (2/1/13; 65 min)

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

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

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

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

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

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

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

$$-\nabla f(x^*) \in N_\Omega(x^*). \quad (0.2)$$

Proof. Suppose that $-\nabla f(x^*) \notin N_\Omega(x^*)$, so that by our earlier result, $-\nabla f(x^*) \notin T_\Omega(x^*)^\circ$. Thus there is a tangent direction $y \in T_\Omega(x^*)$ such that $\langle -\nabla f(x^*), y \rangle > 0$.

By definition of tangent, we there are sequences $\{\alpha_k\}$ (with $\alpha_k > 0$ and $\alpha_k \rightarrow 0$) and $\{y_k\}$ with $y_k \rightarrow y$ such that $x^* + \alpha_k y_k \in \Omega$. By Taylor's theorem we have

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

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

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

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

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

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

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

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

PROPOSITION 0.3. *Let $K = \{z \mid D_{\mathcal{A}} z \leq 0\}$. Then $(C - x_0) \subset K$, and there is a neighborhood U of the origin such that $U \cap K = U \cap (C - x_0)$.*

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

We prove the following result using Farkas' Lemma (proved in Lecture 1).

PROPOSITION 0.4.

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

Proof. We have from the previous result that

$$N_C(x_0) = \{v \mid \langle v, u \rangle \leq 0 \text{ for all } u \text{ such that } D_{\mathcal{A}} u \leq 0\}.$$

Thus for any $v \in N_C(x_0)$, there is no vector u such that $\langle u, v \rangle > 0$ and $D_{\mathcal{A}} u \leq 0$. By Farkas' Lemma, this statement is equivalent to the statement that $v = D_{\mathcal{A}}^T w$ for some $w \geq 0$. $\square$

It follows that

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

Make the link to the first-order conditions of Theorem 0.2. Obvious corollary of the results above: If C is polyhedral of the form $C = \{x \mid Dx \leq d\}$, with $\mathcal{A}$ and $\mathcal{I}$ defined as above, then

$$-\nabla f(x^*) \in N_C(x^*) = \{D_{\mathcal{A}}^T w^* \mid w^* \geq 0\}.$$

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

$$-\nabla f(x^*) = D^T \mu, \quad 0 \leq \mu \perp Dx - d \leq 0. \quad (0.3)$$

(We identify the components $i \in \mathcal{A}$ of μ with the components of w^* .) These extend the KKT conditions for linear programming to the case of a nonconvex objective - but constraints are still linear. We'll extend further to nonlinear constraints below.

Lecture 6. (2/4/13; 70 min)

Gradient Projection. Consider the problem

$$\min_{x \in \Omega} f(x), \quad (0.4)$$

where f is smooth and Ω is closed and convex.

THEOREM 0.5. *If*

$$P(x^* - \bar{\alpha} \nabla f(x^*)) = x^*, \quad \text{for some } \bar{\alpha} > 0, \quad (0.5)$$

then (0.2) holds. Conversely, if (0.2) holds, then

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

Proof. Suppose first that (0.5) holds. In Lemma 0.1(ii) we set

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

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 (0.2).

Now supposed that (0.2) 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

$$\|x^* - x_\alpha\|_2^2 \leq \alpha \nabla f(x^*)^T (x^* - x_\alpha) \leq 0, \quad (0.6)$$

by (0.2). Thus $x^* = x_\alpha$, as claimed. $\square$

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?)

A basic algorithm is *gradient projection*. In a sense it is the natural extension of steepest descent to the problem (0.4). The search path from a point x is the projection of the steepest descent path onto the feasible set Ω , that is,

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

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 difficult, even for simple functions f . 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:

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

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:

$$f(x_k(\beta^m \bar{\alpha})) \leq f(x_k) + c_1 \nabla f(x_k)^T (x_k(\beta^m \bar{\alpha}) - x_k). \quad (0.7)$$

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.

Algorithm 1 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 0.6. *For all $x \in \Omega$ and $z \in \mathbf{R}^n$, the function $g : [0, \infty) \rightarrow \mathbf{R}$ defined by*

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

is monotonically nonincreasing.

Our convergence result is then:

THEOREM 0.7.

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

$$f(x) - f(x(s)) \geq c_1 \nabla f(x)^T (x - x(s)), \quad \forall s \in [0, s_x]. \quad (0.8)$$

(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

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

By rearrangement this becomes

$$\nabla f(x)^T(x - x(s)) \geq \frac{\|x - x(s)\|^2}{s}, \quad \forall x \in \Omega, \quad s > 0. \quad (0.9)$$

If x is stationary (that is, satisfies conditions and (0.5)) we have $x(s) = x$ for all $s > 0$, so the conclusion (0.8) 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, (0.8) can be written as

$$(1 - c_1)\nabla f(x)^T(x - x(s)) \geq (\nabla f(x) - \nabla f(\zeta_s))^T(x - x(s)). \quad (0.10)$$

From (0.9) and Lemma 0.6, 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 (0.10) is satisfied 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 7. (2/6/13; 60 min)

Note that part (a) ensures that we can find an α_k of the form $\alpha_k = \beta^{m_k}$ that satisfies (0.7) from each point x_k . Suppose there is a subsequence K such that $x_k \rightarrow_{k \in K} \bar{x}$. Since the sequence of function values $\{f(x_k)\}$ is nonincreasing we have $f(x_k) \rightarrow f(\bar{x})$.

We consider two cases. First, suppose that

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

for some $\hat{\alpha} > 0$. From (0.9) and Lemma 0.6, we have for all $k \in K$ sufficiently large that

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

since $\bar{\alpha}$ is the initial choice of steplength and $\alpha_k \leq \bar{\alpha}$. Taking the limit as $k \rightarrow \infty$, $k \in K$, we have $\bar{x} - \bar{x}(\bar{\alpha}) = 0$, which implies that $\bar{x}$ is stationary (see (0.5)).

In the second case, suppose that $\liminf_{k \in K, k \rightarrow \infty} \alpha_k = 0$. Then by taking a further subsequence $\bar{K} \subset K$, we have $\lim_{k \in \bar{K}, k \rightarrow \infty} \alpha_k = 0$. For all $k \in \bar{K}$, the Armijo test (0.7) will fail at least once, so we have

$$f(x_k) - f(x_k(\beta^{-1}\alpha_k)) < c_1 \nabla f(x_k)^T (x_k - x_k(\beta^{-1}\alpha_k)). \quad (0.11)$$

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

$$\|x_k - x_k(\beta^{-1}\alpha_k)\| > 0. \quad (0.12)$$

By Taylor's theorem we have

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

for some ζ_k on the line segment between x_k and $x_k(\beta^{-1}\alpha_k)$. By combining this expression with (0.11), 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 (0.9) with Lemma 0.6, 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 (0.12) we obtain

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

Since $\alpha_k \rightarrow 0$ and $x_k \rightarrow \bar{x}$ as $k \rightarrow \infty$, $k \in \bar{K}$, it follows that $\zeta_k \rightarrow \bar{x}$. Hence by taking limits in the expression above we obtain that $\bar{x} = \bar{x}(1)$, which implies that $\bar{x}$ is stationary, as claimed. $\square$

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 satisfies 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 search along the straight line defined by

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

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.

- 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.

Lecture 8. (2/8/13; 60 min)

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.

Showed that gradient projection steps have the “identification property” when the local solution x^* is nondegenerate, when Ω is the positive orthant: $\Omega := \{x \mid x \geq 0\}$.

We need the concept of a *nondegenerate stationary point*. For the positive orthant, the first-order condition $-\nabla f(x^*) \in N_{\Omega}(x^*)$ can be restated as

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

A nondegenerate stationary (first-order) point requires the additional condition $x^* + \nabla f(x^*) > 0$ (strict positivity). Equivalently, nondegeneracy means that we can define the optimal nonzero set to be $\mathcal{A} := \{i \mid x_i^* > 0\}$, and have $[\nabla f(x^*)]_i > 0$ for all $i \notin \mathcal{A}$.

We discussed and stated this theorem (but didn't prove it in class). It says that gradient projection identifies the correct active set, for all x close enough to x^* .

THEOREM 0.8. *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 $\underline{\alpha}$ and $\bar{\alpha}$ such that $0 < \underline{\alpha} < \bar{\alpha}$. Then there is $\epsilon > 0$ such that for all x with $\|x - x^*\| \leq \epsilon$ and all $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, we have*

$$\begin{aligned} i \in \mathcal{A} &\Rightarrow x_i - \alpha[\nabla f(x)]_i > 0, \\ i \notin \mathcal{A} &\Rightarrow x_i - \alpha[\nabla f(x)]_i < 0. \end{aligned}$$

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

Proof. We define

$$\bar{\epsilon} := \min \left(\min_{i \in \mathcal{A}} x_i^*, \min_{i \notin \mathcal{A}} [\nabla f(x^*)]_i \right).$$

Because of nondegeneracy, we have $\bar{\epsilon} > 0$. By the smoothness properties of f , we can choose $\epsilon \leq \min(\bar{\epsilon}/4, \bar{\epsilon}\underline{\alpha}/2)$ sufficiently small that the following inequalities hold for

$$\|x - x^*\| \leq \epsilon:$$

$$\begin{aligned} i \in \mathcal{A} &\Rightarrow \bar{\alpha}|\nabla f(x)_i| \leq \bar{\epsilon}/4, \\ i \notin \mathcal{A} &\Rightarrow |\nabla f(x)_i| \geq \bar{\epsilon}/2. \end{aligned}$$

Thus we have for $i \in \mathcal{A}$, with $\|x - x^*\| \leq \epsilon \leq \bar{\epsilon}/4$ and $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, that

$$x_i - \alpha[\nabla f(x)]_i \geq (x_i^* - \bar{\epsilon}/4) - \bar{\alpha}|\nabla f(x)_i| \geq (\bar{\epsilon} - \bar{\epsilon}/4) - \bar{\epsilon}/4 = \bar{\epsilon}/2 > 0.$$

For $i \notin \mathcal{A}$, with $\|x - x^*\| \leq \epsilon \leq \bar{\epsilon}\underline{\alpha}/2$ and $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, we have

$$x_i - \alpha[\nabla f(x)]_i \leq \bar{\epsilon}\underline{\alpha}/2 - \alpha(\bar{\epsilon}/2) = (\bar{\epsilon}/2)(\underline{\alpha} - \alpha) \leq 0.$$

□

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*.

Lecture 9. (2/11/13; 60 min)

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 to write this subproblem is

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

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 $x \notin \Omega$.

Can thus generalize the gradient projection idea to problems of the form

$$\min_x f(x) + \psi(x),$$

where f is smooth and ψ may be nonsmooth. Replacing I_Ω by ψ in the template above, we obtain the subproblem

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

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

“Barzilai-Borwein” / Spectral Methods. Recall that the Newton step, motivated by Taylor's theorem, satisfies $\nabla^2 f(x_k)p_k = -\nabla f(x_k)$. Quasi-Newton idea is to replace $\nabla^2 f(x_k)$ by a simpler matrix B_k - easier to store, less expensive to calculate, maintain, factor - and obtain the step by solving

$$B_k p_k = -\nabla f(x_k).$$

Properties that we want B_k to satisfy include:

- nonsingular: Then p_k is uniquely determined.
- positive definite: Then p_k is a descent direction.
- symmetric: B_k mimics the Hessian in this sense.
- secant condition. If $s_k := x_k - x_{k-1}$ is step just taken, and $y_k := \nabla f(x_k) - \nabla f(x_{k-1})$ is difference of last two gradients, we have by Taylor's theorem that $\nabla^2 f(x_k) s_k \approx y_k$. The secant condition requires B_k to satisfy a similar property, that is, $B_k s_k \approx y_k$.

Quasi-Newton methods proposed B_k with different structures and different update rules. For the present, we focus on a very simple type of B_k : one with the form $B_k = \gamma_k I$, where γ_k is some scalar. When we solve $B_k p_k = -\nabla f(x_k)$ with a B_k of this form, we obtain the step $p_k = -\alpha_k \nabla f(x_k)$, where $\alpha_k = 1/\gamma_k$. Hence, this form of B_k results in a negative-gradient step with a particular choice of step size.

The Barzilai-Borwein (BB) method [1] is obtained by choosing γ_k to satisfy $\gamma_k s_k \approx y_k$ in the least squares sense. That is,

$$\gamma_k := \arg \min_{\gamma} \|\gamma s_k - y_k\|_2^2.$$

Simple computation leads to the formula

$$\gamma_k = \frac{y_k^T s_k}{s_k^T s_k},$$

leading to the following updating scheme:

$$x_{k+1} = x_k - \frac{s_k^T s_k}{y_k^T s_k} \nabla f(x_k), k = 1, 2, \dots \quad (0.13)$$

At the first iterate ($k = 0$) we could just do a regular line search along the negative gradient direction, as discussed in Section 3.5 of the text.

Note that this procedure is *non-monotone*: it does not necessarily yield a decrease in f at every step, in fact, f can sometimes increase spectacularly. (Illustrate with the 2d ill-conditioned quadratic.)

Often *safeguard* the BB step by bounding above and below by positive constants, i.e.

$$\gamma_k = \text{mid} \left(\epsilon, \frac{y_k^T s_k}{s_k^T s_k}, \frac{1}{\epsilon} \right),$$

for some small constant $\epsilon > 0$.

For smooth convex functions, we are guaranteed that $s_k^T y_k \geq 0$ and in fact $s_k^T y_k \geq \epsilon s_k^T s_k$ for some $\epsilon > 0$ if a strict convexity assumption is made. In the latter case, the coefficient in (0.13) is well defined and lies in the range $[0, \epsilon^{-1}]$. For cases in which $y_k^T s_k < 0$, the step (0.13) does not move in the negative gradient direction. Should we (a) take it anyway? This is a non-monotone method after all, so what if we are moving opposite to the negative gradient; (b) do a line search in the negative gradient direction instead? Not obvious what to do here. Most analysis has been done on convex functions. The issue does not arise when we're close to a solution satisfying second-order sufficient conditions, as f is locally convex here.

(* No lecture on 2/13/13 *)

Lecture 10. (2/15/13; 55 min)

The *alternative BB* formula seeks to approximate $[\nabla^2 f(x_k)]^{-1}$ rather than the Hessian itself. Rewriting the secant condition as $B_k^{-1}y_k \approx s_k$, we choose τ_k that satisfies $\tau_k y_k \approx s_k$ in the least-squares sense, and set $B_k^{-1} = \tau_k I$. We obtain by the usual scalar minimization that

$$\tau_k = \frac{y_k^T s_k}{y_k^T y_k}$$

and

$$p_k = -\tau_k \nabla f(x_k) = -\frac{y_k^T s_k}{y_k^T y_k} \nabla f(x_k).$$

The *alternating BB* strategy is obtained by switching between the two strategies above on successive iterations. That is, BB is used on even iterations and alternative BB on odd iterations.

The *cyclic BB* strategy (see for example [2]) is obtained by choosing a cycle length M (a positive integer) and setting

$$\alpha_k = \begin{cases} \frac{s_k^T s_k}{y_k^T s_k} & \text{if } k \bmod M = 0 \\ \alpha_{M \lfloor k/M \rfloor} & \text{otherwise,} \end{cases}$$

and then setting $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$. That is, we calculate the steplength using the BB formula only every M th iteration, and fix the steplength between such iterations to its last calculated value.

Another possibility is *lagged steepest descent*, in which a line search is used to calculate an approximately minimizing α_k (satisfying Wolfe conditions, say) along the negative gradient direction. However, when we take the step, we use the step α_{k-1} calculated at the *previous* iteration, that is,

$$x_{k+1} = x_k - \alpha_{k-1} \nabla f(x_k).$$

A variant is *half-lagged steepest descent* which is essentially line-search steepest descent with a cycle length of 2, that is, on even iterations k we calculate α_k as an approximately minimizing step length, while on odd iterations k we set $\alpha_k = \alpha_{k-1}$.

It is interesting to analyze and experiment with the performance of the various BB strategies on *convex quadratic objective* functions. Consider

$$f(x) = \frac{1}{2} x^T Q x,$$

where Q is spd. The minimizer is obviously $x^* = 0$. There is no loss of generality by omitting the linear term $b^T x$; the iterates are invariant under the translation by the vector $Q^{-1}b$. In fact, we can assume that Q is *diagonal*, with $Q = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. This can be assumed WLOG because the iterates are invariant under a nonsingular transformation.

Noting that $\nabla f(x) = Qx$, and defining $p_k = -\nabla f(x_k) = -Qx_k$, we have in general that $x_{k+1} = x_k + \alpha_k p_k$. Moreover,

$$s_k = x_k - x_{k-1} = \alpha_{k-1} p_{k-1}, \quad y_k = Qx_k - Qx_{k-1} = \alpha_{k-1} Q p_{k-1}.$$

For all the steps described above, and several others, we can derive explicit formulae.

- steepest descent: $\alpha_k^{SD} = -p_k^T Q x_k / p_k^T Q p_k = p_k^T p_k / p_k^T Q p_k$.
- BB: $\alpha_k = p_{k-1}^T p_{k-1} / p_{k-1}^T Q p_{k-1} = \alpha_{k-1}^{SD}$.
- BB-alternative: $\alpha_k = p_{k-1}^T Q p_{k-1} / p_{k-1}^T Q^2 p_{k-1}$.

They are sometimes called *spectral methods* because α_k^{-1} always lies in the range of eigenvalues of Q , that is,

$$\alpha_k \in \left[\frac{1}{\lambda_n}, \frac{1}{\lambda_1} \right].$$

Convergence results for BB and other nonmonotone algorithms either follows the monotone line search theorems (when we impose additional criteria) or is quite complex. The original paper of Barzilai and Borwein (1988) has an ingenious analysis of the quadratic case, and see [2] for some analysis of cyclic algorithms. The latter paper also contains some useful references.

The basic idea is to use backtracking, starting with the BB steplength, in conjunction with a modified sufficient decrease (first Wolfe) condition. As noted in the text, a backtracking strategy can be used in place of the second Wolfe condition, ensuring that the step is *not too short because it is within a constant factor of a step that is too long*. The modification in sufficient decrease is that we require a sufficient decrease over the *largest function value seen during the past M iterations*, where M is a small integer, say 5 or 10. Formally, we require

$$f(x_k + \alpha p_k) \leq \max_{j=0,1,\dots,\min(M,k)} f(x_{k-j}) + c_1 \alpha_k \nabla f(x_k)^T p_k.$$

Using this condition together with several standard assumptions, we can prove that accumulation points are stationary.

First-Order Conditions: Constraint Qualifications and the Nonconvex Case. Recall the results proved already. For closed convex C have first order conditions $-\nabla f(x^*) \in N_C(x^*)$, expressed in KKT form as ((0.3)) when C is polyhedral. We now refer to (NW, chapter 12) for extension to the case in which constraints are expressed *algebraically* by possibly nonconvex functions.

First extend Proposition 0.4 to the case in which there are equalities as well as inequalities. Given this polyhedral set:

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

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

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

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

Thus for the problem $\min_{x \in C} f(x)$, and first-order condition $-\nabla f(x^*) \in N_C(x^*)$, we can derive the KKT conditions:

$$-\nabla f(x^*) = D^T \mu + G^T \nu, \quad 0 \leq \mu \perp Dx^* - d \leq 0, \quad Gx^* = g.$$

Example:

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

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 $(0, 0)$?

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

Note that Def 12.2 (p.316) of $T_\Omega(x^*)$ is equivalent to the “limiting feasible directions” definition of tangent presented earlier.

In dealing with nonconvex sets, we *define* the normal cone $N_\Omega(x^*)$ as the polar of $T_\Omega(x^*)$ — see p.340. As we noted earlier, this coincides with our definition of the normal cone for the case of convex Ω , but may be different in the nonconvex case. Give an example of a nonconvex set where the definition $N_\Omega(x^*) = T_\Omega(x^*)^\circ$ gives a different result from the (inappropriate) definition for the case of convex Ω .

Now mention Theorem 12.3 on p.325. This says that x^* is a local solution implies that $-\nabla f(x^*) \in T_\Omega(x^*)^\circ$.

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

State Lemma 12.2 on p. 323:

LEMMA 0.9. *For a feasible point x^* , we have*

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

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

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

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

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

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

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

Lecture 12. (2/20/13; 60 min)

Note that the tangent cone for nonconvex set may not be convex. Note that the first-order condition $-\nabla f(x^*) \in N_\Omega(x^*)$ does not require $T_\Omega(x^*)$ to be convex. Note that in the definition

$$N_\Omega(x^*) = T_\Omega(x^*)^\circ,$$

we can replace $T_\Omega(x^*)$ by $\text{co}T_\Omega(x^*)$ without affecting the definition of N_Ω . The normal cone is convex even when T_Ω is not.

Discuss these points in relation to the set defined by

$$\{(x_1, x_2) \mid x_1 \geq 0, x_2 \geq 0, x_1 x_2 = 0\}. \quad (0.14)$$

When CQ holds, we have

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

Now apply Proposition 0.4, extended as described above, to $\mathcal{F}(x^*)$ at the point $d = 0$. Recall that $\mathcal{A}(x^*)$ contains indices from $\mathcal{E}$ as well as the active indices from $\mathcal{I}$. We have

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

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.

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.

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

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

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

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

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

The set of optimal multipliers 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 $\|\lambda_{\mathcal{A}}^k\| \rightarrow \infty$, $\lambda_{\mathcal{A}}^k \geq 0$, and $-\nabla f(x^*) = \nabla c_{\mathcal{A}}(x^*) \lambda_{\mathcal{A}}^k$ for all k . Dividing by $\|\lambda_{\mathcal{A}}^k\|$ and taking limits we have

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

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

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

yielding a contradiction.

Lecture 13. (2/22/13; 60 min)

Discuss CQ properties at points in the set Ω defined by (0.14).

How to prove that MFCQ yields $T_{\Omega}(x^*) = \mathcal{F}(x^*)$? The hard part is to show that $\mathcal{F}(x^*) \subset T_{\Omega}(x^*)$.

Slater condition: With $\Omega = \{x \mid c_i(x) = 0 \ (i \in \mathcal{E}), \ c_i(x) \geq 0, \ (i \in \mathcal{I})\}$, requires that

- the equality constraints $c_i(x)$, $i \in \mathcal{E}$ are linear;
- the inequalities $c_i(x) \geq 0$, $i \in \mathcal{I}$ are concave (that is, $-c_i(\cdot)$ are convex);
- there is a point $\bar{x}$ such $c_i(\bar{x}) = 0$ for all $i \in \mathcal{E}$ and $c_i(\bar{x}) > 0$ for all $i \in \mathcal{I}$.

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

(i) Optimization with a complementarity constraint:

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

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

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

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

- (ii) Reformulating equalities as inequalities. Given original problem $\min f(x)$ e.t. $c(x) = 0$, reformulate as $\min f(x)$ s.t. $c(x) \leq 0$, $c(x) \geq 0$. 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. Proved this in class for LICQ.
- (iv) What about squared slack variables? Write down KKT conditions for original and reformulated problems, given a set of inequality constraints $c_i(x) \geq 0$, $i = 1, 2, \dots, m$. 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.

Lecture 14. (2/25/13; 60 min)

Discuss **Strict complementarity**.

For

$$\min \frac{1}{2}(x_1 + 1)^2 + \frac{1}{2}(x_2 + .25)^2 \text{ s.t. } x_2 \geq -.5x_1, x_2 \leq .5 * x_1, x_1 \geq 0,$$

calculate the set Λ^* of optimal Lagrange multipliers. It contains both strictly complementary multipliers (in its relative interior) and non-strictly complementary multipliers (the two extreme points).

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

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

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

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

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

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

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

Lecture 15. (2/27/13; 35 min)

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.

(Assigned my favorite two-circle example.

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

$x^* = (1, 0)$ is a solution - in fact a strict solution.)

(* no lecture Friday 3/1/13 due to travel *)

Lecture 16. (3/4/13; 60 min)

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

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

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

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

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 .

Lecture 17. (3/6/13; 60 min)

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 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 $\|J(z)^{-1}\| \leq \beta$ for some $\beta > 0$. Thus

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

Apply 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.

Discussed also SLP (and the need for a trust region of ℓ_2 or ℓ_{inf} type) and the version of SQP that uses only $\nabla^2 f(x^k)$ in the second-order term. Also discussed QCQP - subproblem are generally too hard to solve, and are usually not convex.

Extend SQP to (a) inequality constraints, (b) approximate Hessian, including quasi-Newton approx.

Lecture 18. (3/8/13; 60 min)

Quadratic penalty for equality constrained.

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

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 $\lambda_i^* \approx \mu c_i(x^*)$.

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 $i = 1, 2, \dots$ there is w_i with $\|w_i\| = 1$ such that $w_i^T (H + i A A^T) w_i \leq 0$. By compactness of unit ball, we have subsequence $\mathcal{S}$ with $w_i \rightarrow w^*$ for some w^* with $\|w^*\| = 1$. Since

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

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

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

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

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

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

leading to penalty formulation:

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

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.

Lecture 19. (3/11/13; 60 min)

Augmented Lagrangian. Work through details for equality-constrained case. 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.)

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

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

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

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

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

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

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

Nonsmooth penalty function. Discuss simple example.

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

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

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