Unit 2: Nonlinear Programming Models

Outline

1. Calculus and Convexity
2. Models [Quadratic, Second-Order Cone, Semidefinite, Conic Programs]
3. Optimality Conditions
4. Algorithms (given simple constraints)

[Projected Subgradient, Proximal Point, Frank-Wolfe]

1. Calculus and Convexity

In the remainder of this course, we will focus on potentially "nonlinear" objectives and constraints.

We will allow our variables x to live in any Euclidean space E with inner product $\langle \cdot, \cdot \rangle$.

- This can be vectors $E = \mathbb{R}^n$ with $\langle x, y \rangle = x^T y$.
- This can be matrices $E = \mathbb{R}^{n \times m}$ with $\langle X, Y \rangle = \mathrm{tr}(X^T Y)$.

We will only consider finite-dimensional settings, but many ideas here do generalize, enabling optimization over, say, functions.

We are interested in optimization problems of the form

$$\begin{cases} \min f(x) \\ \text{s.t. } x \in Q \end{cases} \quad \text{or} \quad \begin{cases} \min f(x) \\ \text{s.t. } g_i(x) \le 0 \quad \forall i = 1 \dots m \end{cases}$$

for potentially nonlinear $f, g_i : E \to \mathbb{R} \cup \{\pm\infty\}$
and nonpolyhedral $Q \subseteq E$.

Denote the <u>domain</u> of $h : E \to \mathbb{R} \cup \{\pm\infty\}$ as

$$\operatorname{dom} h = \{x \mid h(x) \in \mathbb{R}\} \subseteq E.$$

Calculus Preliminaries (i.e. How to "linearize" nonlinear functions)

We say h is <u>(Frechet) differentiable</u> at $x \in \operatorname{dom} f$ if there exists $\nabla f(x) \in E$ such that

$$\lim_{\Delta x \to 0} \frac{f(x+\Delta x) - (f(x) + \langle \nabla f(x), \Delta x \rangle)}{\|\Delta x\|} = 0.$$

← this is the norm induced by $\langle \cdot, \cdot \rangle$, namely, $\|x\|^2 = \langle x, x \rangle$.

We say $\nabla f(\cdot)$ is the <u>gradient</u> of f (w.r.t. $\langle \cdot, \cdot \rangle$).

If $\nabla f(\cdot)$ exists everywhere in $\operatorname{dom} f$ and is continuous, we say $f \in C^1$.

Likewise, we say $f \in C^1$ is <u>twice (Frechet) differentiable</u> at $x \in \text{dom } f$ if there exists linear operator $\nabla^2 f(x): E \to E$ such that

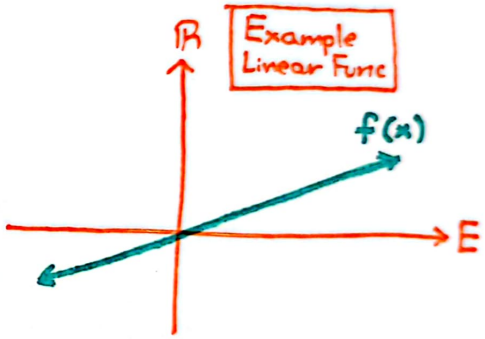
$$\lim_{\Delta x \to 0} \frac{\| \nabla f(x+\Delta x) - (\nabla f(x) + \nabla^2 f(x)\Delta x) \|}{\|\Delta x\|} = 0.$$

We say $\nabla^2 f(\cdot)$ is the <u>Hessian</u> of f (w.r.t. $\langle \cdot, \cdot \rangle$).

Convexity Preliminaries (a useful weakening of linearity)

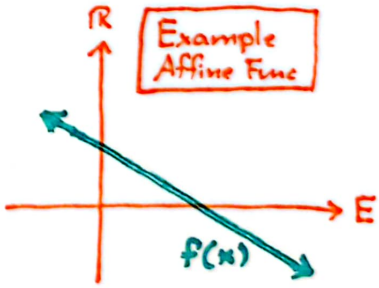
Recall a function is <u>linear</u> if

$$f(\lambda x) = \lambda f(x) \quad \forall_{\substack{x \in E \\ \lambda \in \mathbb{R}}}$$



and is <u>affine</u> if

$$f(\lambda x + (1-\lambda)y) = \lambda f(x) + (1-\lambda) f(y) \quad \forall_{\substack{x,y \in E \\ \lambda \in [0,1]}}$$



[Note Linear programs are exactly the problems like $\begin{cases} \min f(x) \\ \text{s.t. } g_i(x) \leq 0 \end{cases}$ with affine f, g_i.]

We say a function $f: E \to \mathbb{R} \cup \{\pm\infty\}$ is <u>convex</u> if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall_{\substack{x,y \in E \\ \lambda \in [0,1]}}.$$

We will find that convex optimization (when f, g_i are convex) is (fairly) tractable having a rich duality like linear optimization.

We will also find generic (nonconvex) problems are very hard to globally solve, typically lacking duality theories.

A few more definitions

We say a function is <u>strictly convex</u> if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \quad \forall_{\substack{x,y \in E \\ \lambda \in (0,1)}}.$$

We say a function is <u>concave</u> if

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) \quad \forall_{\substack{x,y \in E \\ \lambda \in [0,1]}}.$$

We say a set $Q \subseteq E$ is <u>convex</u> if

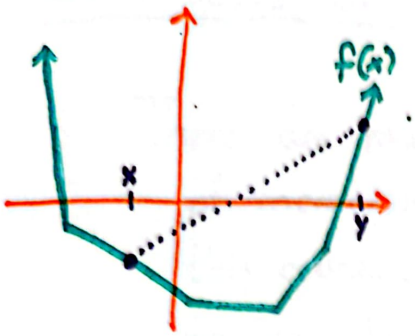
$$\lambda x + (1-\lambda)y \in Q \quad \forall x, y \in Q, \ \lambda \in [0,1].$$

Some good exercises to play with these definitions

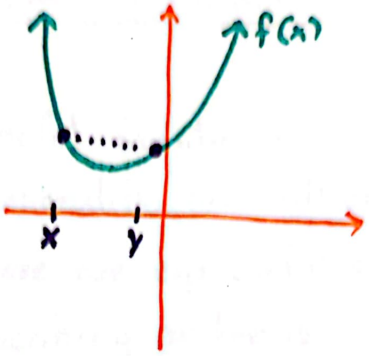
(a) Show f is convex iff its epigraph, $\operatorname{epi} f = \{(x,t) \mid f(x) \leq t\}$ is a convex set.

(b) Show Q is convex iff its indicator $\delta_Q(x) = \begin{cases} 0 & \text{if } x \in Q \\ \infty & \text{if } x \notin Q \end{cases}$ is a convex function.

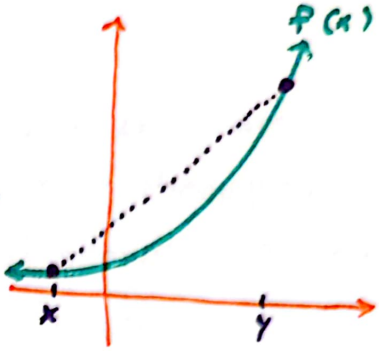
Examples of Convex Functions (E = ℝ so they are "drawable")



A maximum of several affine functions



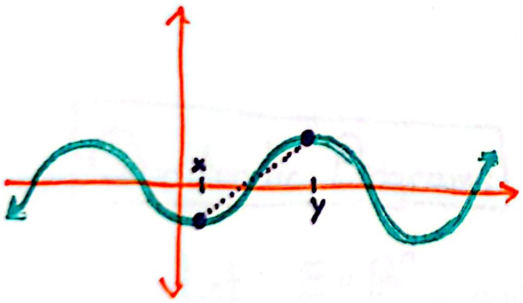
A quadratic with positive curvature

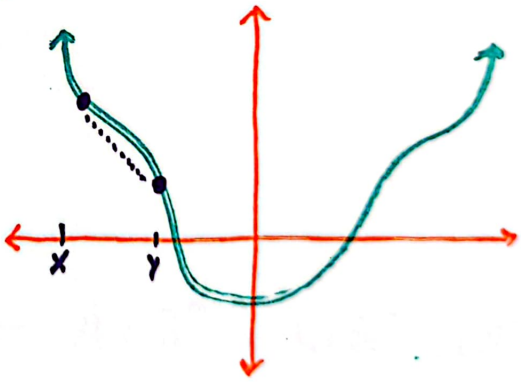


The exponential function

[Graphically, convexity is defined as all dotted line segments, as drawn above, lie above the function's graph.]

Examples of Not Convex Functions





[Some good exercises to play with these definitions:

For convex functions $f_i : E \to \mathbb{R} \cup \{\pm\infty\}$, $i = 1 \dots n$

(a) Prove $\max_i \{ f_i(\cdot) \}$ is convex.

(b) Prove $\sum_{i=1}^{n} f_i(\cdot)$ is convex.

(c) If $f_i(x) \geq 0 \ \forall i, x$, prove $f_i^2(\cdot)$ is convex.

(d) If f_i is concave, prove f_i is affine.]

2. Models of Nonlinear Programs

There are many useful families of nonlinear programs, of increasing generality, we will encounter throughout this course. These are esp useful as standardized models to use when describing problems in software.

An overview of the models we will survey here:

(We will introduce these left to right, proving containments as we go.)

Quadratic Programs

Let $E = \mathbb{R}^n$, and consider $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and positive semidefinite $H \in \mathbb{R}^{n \times n}$.

↑ (that is, it is symmetric with all nonnegative eigenvalues, or equivalently $v^T H v \geq 0 \ \forall v$)

Then QPs minimize the following convex quadratic over a polyhedron

$$\begin{cases} \min & c^T x + \frac{1}{2} x^T H x \\ \text{s.t.} & Ax = b \\ & x \geq 0. \end{cases}$$

↑ Just as with LPs, we could equivalently consider constraints $Ax \leq b$

[A good exercise: Prove $x \mapsto c^T x + \frac{1}{2} x^T H x$ is convex if and only if the matrix H is positive semidefinite.]

Trivially the set of QPs contains all LPs by letting $H = 0$.

Example 1 Projection onto a polyhedron $\mathcal{P} = \{x \mid Ax = b, x \geq 0\}$.

Given a point $x_0 \in E$, find the nearest element of $\mathcal{P}$.

We can model this problem as

$$\begin{cases} \min & \frac{1}{2}\|x - x_0\|_2^2 = \frac{1}{2}(x-x_0)^T I (x-x_0) = \frac{1}{2}x^T I x - x_0^T x + \frac{1}{2}x_0^T I x_0 \\ \text{s.t.} & Ax = b \\ & x \geq 0. \end{cases}$$

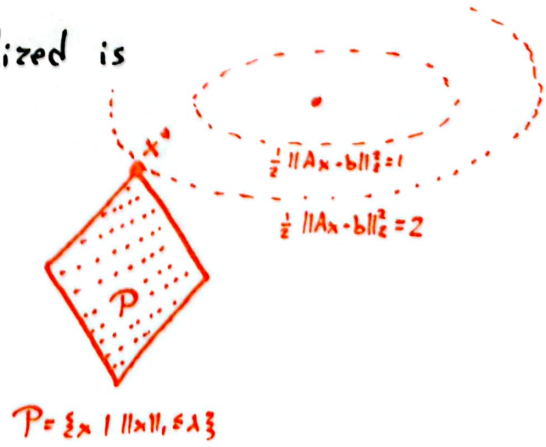
quadratic, linear, constant

Example 2 LASSO Regression, seeking sparse solutions (approx) to a system of equations $Ax = b$.

One way this is often formalized is

$$\begin{cases} \min & \frac{1}{2}\|Ax - b\|_2^2 \\ \text{s.t.} & \|x\|_1 \leq \lambda \end{cases}$$

for a given $\lambda > 0$.



Second-Order Cone Programs

Let $E = \mathbb{R}^n$, and consider $A_i \in \mathbb{R}^{m_i \times n}$, $b_i \in \mathbb{R}^{m_i}$, $c \in \mathbb{R}^n$, $d_i \in \mathbb{R}$, $f_i \in \mathbb{R}^n$.

Then SOCPs minimize the following

$$\begin{cases} \min & c^T x \\ \text{s.t.} & \| A_i x - b_i \|_2 \leq f_i^T x + d_i \quad \forall_{i=1\dots k} \end{cases}$$

These constraints can describe any polyhedron by selecting $A_i = 0$ and $b_i = 0$. Then it is just the halfspace constraint $0 \leq f_i^T x + d_i$.

Every QP can be rewritten as an SOCP:

$$\begin{cases} \min & c^T x + \frac{1}{2} x^T H x \\ \text{s.t.} & a_i^T x \leq b_i \ \forall_i \end{cases} = \begin{cases} \min & t \\ \text{s.t.} & c^T x + \frac{1}{2} x^T H x \leq t \\ & a_i^T x \leq b_i \quad \forall_i \end{cases}$$

$$= \begin{cases} \min & t \\ \text{s.t.} & [\text{redacted}] \quad a_i^T x \leq b_i \ \forall_i \\ & \frac{1}{2}(x,t)^T \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix} (x,t) + (c,-1)^T (x,t) \leq 0 \end{cases}$$

↙ SOCP constraint as argued above

↑ SOCP constraint by completing the square. (note this relies on H being p.s.d.)

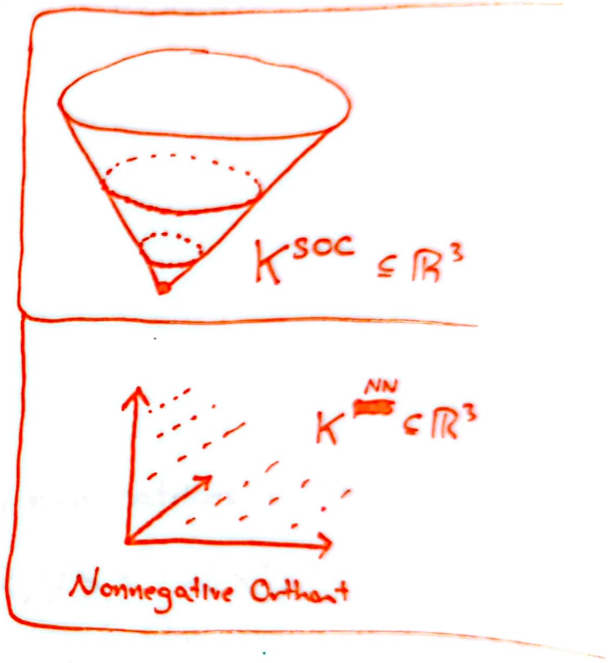
Note each SOC constraint can be written in the following "standard" form (mirroring LPs)

$$\|A_i x - b_i\|_2 \le f_i^T x + d_i \iff \begin{array}{l} y = A_i x - b_i \\ t = f_i^T x + d_i \\ (y,t) \in K^{soc} = \{(y,t) \mid \|y\|_2 \le t\} \end{array}$$

Hence just like LPs can be viewed as "$Ax=b, x \ge 0$".

$x \ge 0 \Leftrightarrow x \in K^{NN} = \{x \mid x \ge 0\}$

SOCP constraints can be viewed as "$A(x,t)=b, (x,t) \in K^{soc}$".



Example 1 Another LASSO regression problem formulation

$$\begin{cases} \min \|x\|_1 \\ \text{s.t. } \|Ax-b\|_2 \le \eta \end{cases}$$

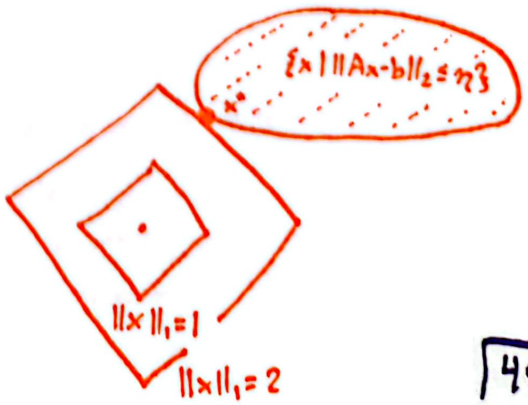
given some $\eta > 0$ (note this reverses the obj and constraints previous seen).

We just need to do some minor work to reformulate this to have a linear objective function...

Claim: This problem is equivalent to

$$\begin{cases} \min \sum t_i \\ \text{s.t. } t_i \ge x_i \\ \quad\; t_i \ge -x_i \\ \quad\; \|Ax-b\|_2 \le \eta \end{cases}$$

Proof left as an exercise.



Example 2 Stochastic Linear Programs

Suppose we do not exactly know the constraints for our LP. Instead, we know $a_i \sim N(\bar{a}_i, \Sigma_i)$.

For given $p \geq 0.5$, we aim to solve

$$\begin{cases} \min\ c^T x \\ \text{s.t. } \text{Prob}(a_i^T x \leq b_i) \geq p \quad \forall i = 1 \dots m. \end{cases}$$

[In homework you will show this can be written as a Second-Order Cone Program.]

Semidefinite Programming

Let $E = \mathbb{S}^{n \times n}$ = the set of symmetric $n \times n$ matrices with the trace inner product $\langle X, Y \rangle = \text{tr}(X^T Y)$.

Consider any linear $\mathcal{A}: \mathbb{S}^{n \times n} \to \mathbb{R}^m$, $b \in \mathbb{R}^m$, $C \in \mathbb{S}^{n \times n}$.

Then a SDP is the following optimization problem

$$\begin{cases} \min & \langle C, X \rangle \\ \text{s.t} & \mathcal{A}X = b \\ & X \succeq 0 \end{cases}$$

$X \succeq 0$ meaning X is positive semidefinite.
(i.e. $X \in K^{PSD} = \{X \text{ is p.s.d.}\}$)
A 3D printed copy of K^{PSD} was shown in lecture (embedded in $\mathbb{R}^3$ by $\begin{bmatrix} x & y \\ y & z \end{bmatrix} \to (x, y, z)$).

Note this model allows for several p.s.d. constraints implicitly: If you want matrices $X_1, \dots, X_K$ to all be p.s.d, require $\begin{bmatrix} X_1 & & 0 \\ & X_2 & \\ & & \ddots \\ 0 & & X_K \end{bmatrix} \succeq 0$.

↑ block diagonal matrix

Every SOCP can be written as an SDP.

Proof. Observe that

$$\|x\|_2 \le t \iff \begin{bmatrix} tI & x \\ x^T & t \end{bmatrix} \succeq 0.$$ (Why? Schur complements.)

Then we can reformulate any SOC constraints as

$$\|A_i x - b_i\|_2 \le f_i^T x + d_i \overset{\forall i=1\dots K}{\iff} \begin{cases} y_i = A_i x - b_i \\ t_i = f_i^T x + d_i \quad \forall i \\ \|y_i\|_2 \le t_i \end{cases}$$

$$\iff \begin{cases} y_i = A_i x - b_i \\ t_i = f_i^T x + d_i \\ \begin{bmatrix} t_1 I & y_1 & & & \\ y_1^T & t_1 & & 0 & \\ & & t_2 I & y_2 & \\ & & y_2^T & t_2 & \ddots \\ & 0 & & & \end{bmatrix} \succeq 0. \end{cases}$$

Example 1 NP-Hard Combinatorial Opt Approximations

Suppose we have a complete graph on nodes $1 \dots n$.
Our task is to split them in to two groups with cost A_{ij} for placing i and j together and profit A_{ij} for seperating them.

This can be modeled as the combinatorial problem $p^* = \begin{cases} \min & \sum A_{ij} x_i x_j = x^T A x \\ \text{s.t.} & x \in \{\pm 1\}^n. \end{cases}$

[This problem is NP-Hard to globally optimize.]

Despite this problem's hard combinatorial nature, we can approximate its solution via an SDP:

$$p^* = \begin{cases} \min & \langle A, xx^T \rangle \\ \text{s.t.} & x \in \{\pm 1\}^n \end{cases}$$

← using the cyclic property of trace: $\langle A, xx^T \rangle = tr(A^T xx^T) = tr(x^T A^T x) = x^T A^T x = x^T A x$.

$$= \begin{cases} \min & \langle A, X \rangle \\ \text{s.t.} & X \in \text{convexhull}\{xx^T \mid x \in \{\pm 1\}^n\} \end{cases}$$

← using that linear optimization over a set of pts equals the LP over their hull.

$$\geq \begin{cases} \min & \langle A, X \rangle \\ \text{s.t.} & \text{diag } X = 1 \\ & X \succeq 0 \end{cases} = r^*$$

← relaxing the above constraints since all xx^T have 1's on diag and are p.s.d.

Famously r^* is within a constant factor of p^* and so this SDP is a provably good approximation. (Very related to Goemans-Williamson Alg)

Example 2 Optimizing Stepsizes for Gradient Descent
(Performance Estimation Problems)

Suppose you have just implemented gradient descent

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad , \quad x \in \mathbb{R}^n$$

for your favorite stepsize sequence $\alpha_k > 0$. You then want to know which f your method does worst on.

Lets compute the worst f after fixed T steps ← assumed $< n$

satisfying: (i) f is a convex, piecewise linear function,
(ii) f has bounded slope, $\|\nabla f(x)\|_2 \le L \quad \forall x$,
(iii) Bounded distance to optimal $\|x_0 - x^*\|_2 \le D$.

Then we can formulate this as the nonconvex problem:

(Variables are x_0, x^*, f_i, g_i with all x_j determined by these.)

$$\begin{cases} \max & f_T - f_* \\ \text{s.t.} & f_i + g_i^T(x_j - x_i) \le f_j \quad \forall i,j \in \{1,\dots,T,*\} \\ & \|g_i\|_2^2 \le L^2 \\ & \|x_0 - x^*\|_2^2 \le D^2 \\ & x_j = x_0 - \sum_{i=0}^{j-1} \alpha_i g_i \quad \forall j \end{cases}$$

← enforces convexity (i)
← enforces (ii)
← enforces (iii)

A clever change of variables turns this into a SDP:

Consider $P = \begin{bmatrix} | & | & | & & | \\ x_0 & x_* & g_0 & \cdots & g_T \\ | & | & | & & | \end{bmatrix}$. Then $X = P^T P$ contains every quadratic term above.

[Omitting writing all the gorey details of this formulation.]
Happy to do so in office hours sometime.

After solving for X, the worst case function in P is recovered by factoring X into its squareroots.

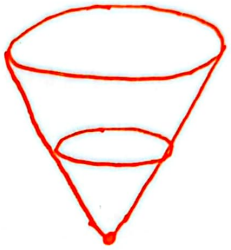
Conic Programs

For a generic E, we say a set $K \subseteq E$ is a <u>cone</u> if $\lambda x \in K \quad \forall x \in K, \lambda > 0$.

Our previous models have all had closely related closed, convex, cones:

LPs and QPs solved over $K^{NN} = \{x \geq 0\}$

SOCPs solved over $K^{SOC} = \{\|x\|_2 \leq t\}$



SDPs solved over $K^{SDP} = \{X \succeq 0\}$

(Hard to draw, see 3D printed version)

Conic Programs for any $c \in E$, $\mathcal{A}: E \to \mathbb{R}^m$ linear, $b \in \mathbb{R}^m$ solve

$$\begin{cases} \min & \langle c, x \rangle \\ \text{s.t.} & \mathcal{A}x = b \\ & x \in K \end{cases}$$

for some closed convex cone K.

More wild cones include the cone of copositive matrices

$$K = \{X \mid v^T X v \geq 0 \quad \forall v \geq 0\}.$$

Checking membership of this cone is NP-Hard, so we will not aim to solve generic conic programs efficiently.

Convex Programs

For generic E, we say a function $f: E \to (-\infty, \infty]$ is <u>proper</u>

if $\emptyset \neq \text{dom } f = \{x \mid f(x) \neq \infty\}$.

(i.e. it is not infinity everywhere.)

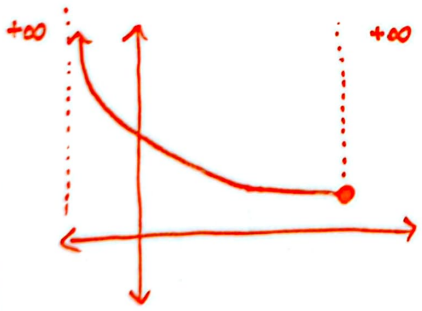
We say f is <u>closed</u> (or <u>lower semicontinuous</u>) if

$$\liminf_{x' \to x} f(x') = f(x) \quad \forall x \in E$$

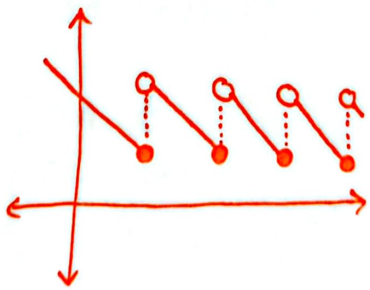
(i.e. the epigraph $\text{epi } f = \{(x,t) \mid f(x) \leq t\}$ is closed. Proving this is a good exercise.)

Examples

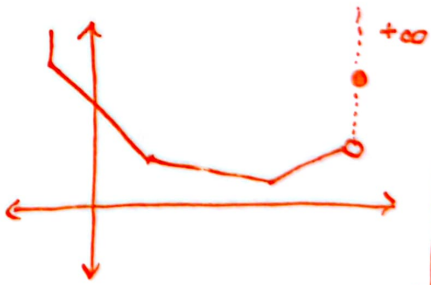
Closed and Convex



Closed, Not Convex



Not Closed, but Convex



Then convex programs solved problems of the form

$$\begin{cases} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0 \quad \forall i = 1 \dots m \end{cases}$$

for closed, convex, proper $f, g_i : E \to (-\infty, \infty]$.

Claim: For any nonempty closed convex S, $\text{dist}_S(x) = \min\{\|x'-x\| \mid x' \in S\}$ is closed, convex, proper.

Then it follows that every conic program can be written as a convex program as

$$\begin{cases} \min & \langle c, x \rangle \\ \text{s.t.} & \text{dist}_S(x) \leq 0 \end{cases}$$

where $S = \{x \mid Ax = b, x \in K\}$.

Conversely every convex program can be written as a conic program:

$$\begin{cases} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0 \end{cases} = \min_{x \in E} h(x) \quad \text{where } h(x) = \begin{cases} f(x) & \text{if } g_i(x) \leq 0 \ \forall i \\ +\infty & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \min & t \\ \text{s.t.} & (x,t) \in \text{epi } h \end{cases} \quad \text{where epi } h = \{(x,t) \mid h(x) \leq t\}$$

$$= \begin{cases} \min & t \\ \text{s.t.} & z = 1 \\ & (x,t,z) \in K^{\text{epi } h} \end{cases} \quad \text{where } K^{\text{epi } h} \text{ is the "recession cone"} = \{(x,t,z) \mid \tfrac{(x,t)}{z} \in \text{epi } h\}.$$

Hence Conic and Convex Programming are equivalent in terms of which problems they can model (and so generic convex programs are also NP-Hard to solve.)

3. Optimality Conditions

Equipt with these optimization models, we know what to characterize what optimal solutions look like.

We say $\bar{x} \in E$ is a <u>local minimizer</u> of f if

$$f(x) \geq f(\bar{x}) \quad \forall_{\text{feasible } x \text{ near } \bar{x}}.$$

We say $\bar{x} \in E$ is a <u>global minimizer</u> of f if

$$f(x) \geq f(\bar{x}) \quad \forall_{\text{feasible } x}.$$

We split our discussion in two parts, depending on the type of constraints:

3.1 Optimality with Set Constraints

$$\begin{cases} \min & f(x) \\ \text{s.t.} & x \in Q \end{cases}$$

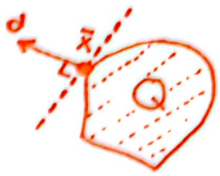
Denote directional derivatives of f at $\bar{x}$ in direction d by

$$f'(x;d) = \lim_{t \to 0^+} \frac{f(x+td) - f(x)}{d}$$

importantly we just consider $t > 0$ when taking this limit. As a result, $f(x) = |x|$ has $f'(0;1) = 1$, $f'(0;-1) = -1$.

Note if $f \in C^1$, $f'(\bar{x};d) = \langle \nabla f(\bar{x}), d \rangle$.

Define the normal cone of Q at $\bar{x} \in Q$ as



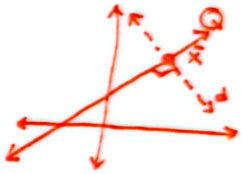
$$N_Q(\bar{x}) = \{ d \mid \langle d, x - \bar{x} \rangle \leq 0 \ \forall x \in Q \}.$$

[Checking this is a cone is a good sanity check (i.e. $d \in N_Q(\bar{x})$, $\lambda > 0 \Rightarrow \lambda d \in N_Q(\bar{x})$).]

Three Short Normal Cone Examples

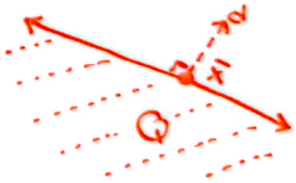
1. $Q = \{ x \mid Ax = b \}$, an affine subspace given by some linear $A: E \to F$, $b \in F$.
   (F: another Euclidean space)

   $\Rightarrow$ For any $\bar{x} \in Q$, $N_Q(\bar{x}) = \{ A^* y \mid y \in F \}$
   $= A^* F$.



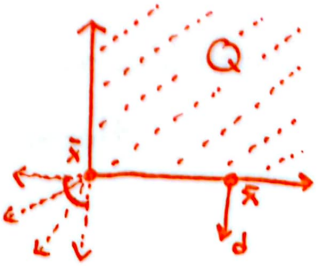
2. $Q = \{ x \mid \langle a, x \rangle \leq b \}$, a halfspace given by some $a \in E$, $b \in \mathbb{R}$.

   $\Rightarrow$ For any $\bar{x} \in Q$, $N_Q(\bar{x}) = \begin{cases} \{0\} & \text{if } \langle a, \bar{x} \rangle < b \\ \{ \lambda a \mid \lambda \geq 0 \} & \text{otherwise.} \end{cases}$



3. $Q = K^{NN} = \{ x \in \mathbb{R}^n \mid x \geq 0 \}$, the nonnegative orthant.

   $\Rightarrow$ For any $\bar{x} \in Q$, $N_Q(\bar{x}) = \{ d \mid d_i = 0 \text{ if } \bar{x}_i > 0$
   $d_i \leq 0 \text{ if } \bar{x}_i = 0 \}$.



Proposition (First-Order Necessary Condition for Optimality)

Suppose Q is a closed, convex set in E and $\bar{x} \in Q$ is a local minimizer of f over Q.

For every $x \in Q$, if $f'(\bar{x}, x-\bar{x})$ exists, it is nonnegative.

In particular, if f is differentiable at $\bar{x}$,

$$\langle \nabla f(\bar{x}), x-\bar{x} \rangle \geq 0 \quad \forall x \in Q$$

$$\Leftrightarrow -\nabla f(\bar{x}) \in N_Q(\bar{x}).$$

Proof. Suppose to the contrary some $x \in Q$ has $f'(\bar{x}, x-\bar{x}) < 0$.

Then for some $0 < t < \delta$ (for any choice of $\delta > 0$), we must have

$$\frac{f(\bar{x} + t(x-\bar{x})) - f(\bar{x})}{t} < 0.$$

$$\Rightarrow f(\bar{x} + t(x-\bar{x})) < f(\bar{x})$$

The convexity of Q ensures $\bar{x} + t(x-\bar{x}) \in Q$ for all $t \leq 1$.

Thus we have points arbitrarily close to $\bar{x}$ with strictly better objective value, contradicting $\bar{x}$'s local optimality. □

The converse does not hold in general. However, if we additionally assume that f is convex, we get such a result.

Proposition (First-Order Sufficient Condition for Optimality)

Suppose Q and f are both closed and convex.

If all $x \in Q$ have $f'(\bar{x}; x-\bar{x}) \geq 0$ for some fixed $\bar{x} \in Q$,

then $\bar{x}$ is a global minimizer of f over Q.

In particular, if f is differentiable at $\bar{x}$,

$$-\nabla f(\bar{x}) \in N_Q(\bar{x}) \Rightarrow \bar{x} \text{ is globally optimal.}$$

Proof. First we claim the convexity of f ensures the function

$t \mapsto \dfrac{f(\bar{x}+t(x-\bar{x})) - f(\bar{x})}{t}$ is nondecreasing for $t>0$.

Proof. Consider $0<t_1<t_2$, then we have

$$\frac{f(\bar{x}+t_2(x-\bar{x})) - f(\bar{x})}{t_2} - \frac{f(\bar{x}+t_1(x-\bar{x})) - f(\bar{x})}{t_1}$$

$$= \frac{t_1 f(\bar{x}+t_2(x-\bar{x})) - t_2\{f(\bar{x}+t_1(x-\bar{x})) + (t_2-t_1)f(\bar{x})}{t_2 t_1}$$

$$\geq \frac{t_1 f(\bar{x}+t_2(x-\bar{x})) - t_2\left(\frac{t_1}{t_2} f(\bar{x}+t_2(x-\bar{x})) + \left(1-\frac{t_1}{t_2}\right) f(\bar{x})\right) + (t_2-t_1)f(\bar{x})}{t_2 t_1}$$

$= 0$.

Using the convexity of f on $\bar{x}+t_1(x-\bar{x})$ being rewritten as :

$$\frac{t_1}{t_2}(x+t_2(x-\bar{x})) + \left(1-\frac{t_1}{t_2}\right)\bar{x}$$

By assumption, for all small $t>0$, this function is nonnegative.

Hence for $t=1$ it is also nonnegative.

$\Rightarrow f(\bar{x}+1\cdot(x-\bar{x})) - f(\bar{x}) \geq 0$

$\Rightarrow f(x) \geq f(\bar{x}) \quad \forall x \in Q$.

$\Rightarrow \bar{x}$ is globally optimal. $\square$

3.2 Some Existence Theorems and Theorems of Alternatives (Farka's Lemma)

(this results will be needed as helpers to give optimality conditions for functional constraints.)

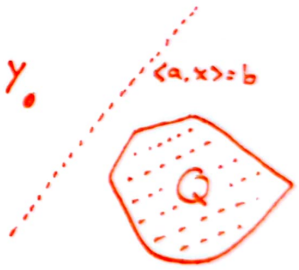
Theorem (Weierstrass)

For any closed, nonempty $Q \subseteq E$, $f: Q \to \mathbb{R}$ with bounded levelsets $\{x \mid f(x) \leq \lambda\}$ for all $\lambda \in \mathbb{R}$, then there exists a global minimizer.

Theorem (Basic Seperation)

For any closed, convex $Q \subseteq E$, $y \notin Q$ there exists $a \in E$, $b \in \mathbb{R}$ such that

$$\langle a, y \rangle > b \geq \langle a, x \rangle \quad \forall x \in Q.$$



Proof. Consider $f(x) = \frac{1}{2}\|x - y\|^2$.

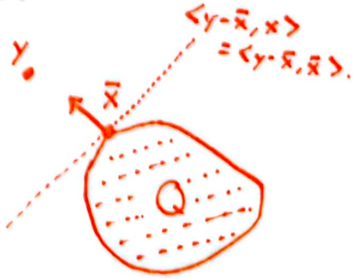
Note f has bounded level sets.

$\Rightarrow$ There exists $\bar{x}$ globally minimizing f over Q.

$\Rightarrow -\nabla f(\bar{x}) = y - \bar{x} \in N_Q(\bar{x})$.

$\Rightarrow \langle y - \bar{x}, x - \bar{x} \rangle \leq 0 \quad \forall x \in Q$.

Pick $a = y - \bar{x}$, $b = \langle y - \bar{x}, \bar{x} \rangle$. $\square$



Theorem (Small Gradients Exist)

If $f: E \to \mathbb{R}$ is differentiable and bound below, then there exist $x_i \in E$ with $\nabla f(x_i) \to 0$.

Proof. For $\varepsilon > 0$, consider minimizing $f(\cdot) + \varepsilon \|\cdot\|$.

Note this has bounded level sets.

$\Rightarrow$ There exists a global minimizer x_ε.

For $d = -\nabla f(x_\varepsilon)$ and small $t > 0$, we have

$$\frac{f(x_\varepsilon + td) - f(x_\varepsilon)}{t} = \frac{f(x_\varepsilon + td) + \varepsilon\|x + td\| - (f(x_\varepsilon) + \varepsilon\|x_\varepsilon\|) - \varepsilon\|x + td\| + \varepsilon\|x_\varepsilon\|}{t}$$

$$\overset{\text{using that } x_\varepsilon \text{ minimizes}}{\geq} \varepsilon \frac{\|x_\varepsilon\| - \|x_\varepsilon + td\|}{t}$$

$$\overset{\text{using the triangle inequality}}{\geq} -\varepsilon \|d\|.$$

Taking the limit as $t \to 0^+$ gives the result

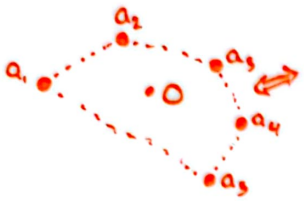
$$-\varepsilon \underbrace{\|d\|}_{\|-\nabla f(x_\varepsilon)\|} \leq \lim_{t \to 0^+} \frac{f(x_\varepsilon + td) - f(x_\varepsilon)}{t} = \langle \nabla f(x_\varepsilon), \underbrace{d}_{-\nabla f(x_\varepsilon)} \rangle = -\|\nabla f(x_\varepsilon)\|^2.$$

$\Rightarrow \|\nabla f(x_\varepsilon)\| \leq \varepsilon.$ $\square$

Finally, we present two specialized existence theorems, known as "Theorems of Alternatives" stating exactly one of two things exist.

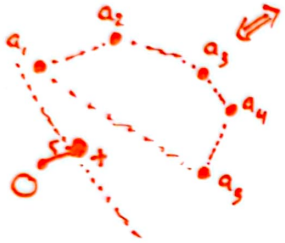
Theorem (Gordan's Theorem of Alternatives)

For any collection $a_1, \dots, a_m \in E$, exactly one of the following is true



(i) $\exists_{\lambda \in \mathbb{R}^m}, \sum_{i=1}^m \lambda_i a_i = 0, \sum_{i=1}^m \lambda_i = 1, \lambda \geq 0$

(think 0 is in the convex hull of $\{a_i\}$)



(ii) $\exists x \in E, \ \langle a_i, x \rangle < 0 \quad \forall_{i=1 \dots m}$.

(think 0 can be seperated from the convex hull)

Proof. We prove this by showing the following are equivalent

(1) $f(x) = \log\left(\sum_{i=1}^m \exp(\langle a_i, x \rangle)\right)$ is bounded below,
(2) System (i) is solvable,
(3) System (ii) is not solvable.

In homework, you will be asked to prove (2) $\Rightarrow$ (3) $\Rightarrow$ (1).

Lets show here that (1) $\Rightarrow$ (2).

Since f is bounded below, we can find x_k with $\nabla f(x_k) \to 0$.

Calculating the gradient of f gives:

$$\nabla f(x_k) = \sum_{i=1}^m \lambda_i^k a_i \quad \text{where } \lambda_i^k = \frac{\exp(\langle a_i, x_k \rangle)}{\sum_j \exp(\langle a_j, x_k \rangle)}$$

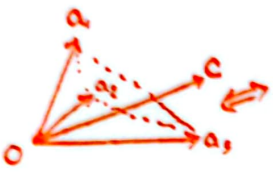
(note $\sum_{i=1}^m \lambda_i^k = 1$ and $\lambda_i^k \geq 0$).

Since λ^k is a bounded sequence, it must have a limit point λ.

This λ has $\sum_{i=1}^m \lambda_i = 1, \lambda \geq 0, \sum_{i=1}^m a_i \lambda_i = \lim_{k \to \infty} \nabla f(x_k) = 0$. $\square$

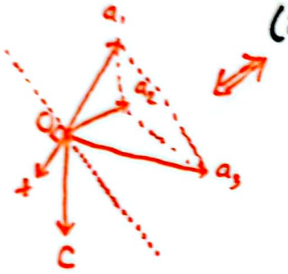
Lemma (Farkas Lemma)

For any collection $a_1 \dots a_m \in E$, $c \in E$, exactly one of the following is true



(i) $\exists \mu \in \mathbb{R}^m$, $\sum_{i=1}^{m} \mu_i a_i = c$, $\mu \geq 0$,

(think c lies in the cone produced by $\{a_i\}$)



(ii) $\exists x \in E$, $\langle a_i, x \rangle \leq 0$, $\langle c, x \rangle > 0$.

(think C can be seperated from this cone)

Proof. Lets prove this with our LP strong duality theory.

Consider $p^* = \begin{cases} \max & \langle c, x \rangle \\ \text{s.t.} & \langle a_i, x \rangle \leq 0. \end{cases}$

First we claim $p^* = 0$ or $+\infty$.

[Proof. Since $x=0$ is feasible, $p^* \geq 0$.
If $p^* \neq 0$, then some x has $\langle c, x \rangle > 0$ with $\langle a_i, x \rangle \leq 0$.
Note this is true for λx for any $\lambda > 0$.
Taking $\lambda \to \infty$ yields arbitrary large obj value. □]

If $p^* = 0$, then (2) has no solution.
Strong duality implies the dual must be feasible.
$\Rightarrow \exists \mu \in \mathbb{R}^m$ with $\begin{cases} \sum a_i \mu_i = c \\ \mu \geq 0 \end{cases}$
$\Rightarrow$ (1) has a solution.

If $p^* = \infty$, then (2) has a solution.
Weak duality implies the dual must be infeasible.
$\Rightarrow$ No solution to (1) exists. □

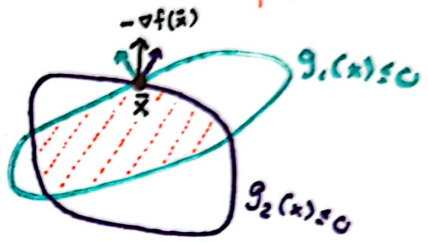
3.3 Optimality with Functional Constraints

$$\begin{cases} \min f(x) \\ \text{s.t. } g_i(x) \le 0 \quad \forall i = 1 \dots m \end{cases}$$

For our discussion here, we will assume $f, g_i \in C^1$ and consider some $\bar{x}$ with $g_i(\bar{x}) \le 0$.

Let $I(\bar{x}) = \{ i \mid g_i(\bar{x}) = 0 \}$ be called the <u>active set</u> at $\bar{x}$.

Basic Picture Example



Recall for set constraints optimality, needed $-\nabla f(\bar{x})$ normal to the feasible region, dotted above.

Classic Lagrange Approach

For $\lambda \ge 0$, $\lambda \in \mathbb{R}^m$, define the <u>Lagrangian</u> as

$$L(x; \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x).$$

We say $\lambda \ge 0$ is a <u>Lagrange Multiplier Vector</u> at $\bar{x}$ if the following hold

(i) $\bar{x}$ is a critical point on $L(\cdot\,; \lambda)$

(i.e. $\nabla f(\bar{x}) + \sum \lambda_i \nabla g_i(\bar{x}) = 0$)

(ii) $\lambda_i g_i(\bar{x}) = 0 \quad \forall i = 1, \dots, m$, known as complementary slackness.

(i.e. $\lambda_i = 0$ if $i \notin I(\bar{x})$)

The following pair of theorems will show existence of such vectors, first "weakly" in a general setting (Fritz John) and then "strongly" for a well behaved setting (the famous KKT conditions).

Theorem (Fritz John)

If $\bar{x}$ is a local minimizer of $\begin{cases} \min f(x) \\ \text{s.t. } g_i(x) \le 0 \quad \forall i=1,\dots,m \end{cases}$ with f, g_i all differentiable at $\bar{x}$,

then there exists $(\lambda_0, \lambda) \in \mathbb{R} \times \mathbb{R}^m$ nonnegative and not all zero such that

$$\lambda_0 \nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0, \quad \lambda_i g_i(\bar{x}) = 0.$$

Proof deferred to the end of this section

If $\lambda_0 > 0$, we could divide through by λ_0 and find $\frac{\lambda}{\lambda_0}$ is a Lagrange Multiplier Vector:

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_0} \nabla g_i(\bar{x}) = 0, \quad \frac{\lambda_i}{\lambda_0} g_i(\bar{x}) = 0 \ \forall i.$$

If $\lambda_0 = 0$, the above condition is rather weak since it becomes independent of the objective f.

[In homework you will explore examples of both types]

How can we guarantee we are not in the second case above?

We want to ensure no $\lambda \ge 0$, not all zero has $\sum_{i} \lambda_i \nabla g_i(\bar{x}) = 0$, $\lambda_i g_i(\bar{x}) = 0 \ \forall i$.

Equivalently, we want no $\lambda \ge 0$, $\sum_{i \in I(\bar{x})} \lambda_i = 1$, $\sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0$.

Gordan's Theorem ensures no solution to this if and only if

$$\exists d \in E \text{ s.t. } \langle \nabla g_i(\bar{x}), d \rangle < 0 \quad \forall i \in I(\bar{x}).$$

This is called the "Mangasarian-Fromovitz Constraint Qualification".

This MFCQ condition at $\bar{x}$ can be understood as a direction existing where all constraints become strictly held (i.e. $g_i'(\bar{x}; d) < 0$).

Lemma For convex, differentiable g_i, MFCQ holds at $\bar{x}$ if and only if there exists x_s with $g_i(x_s) < 0 \ \forall i$.

↑ this is known as a Slater point.

Proof. If MFCQ holds, for small enough t, $\bar{x} + td$ is strictly feasible.

If x_s exists, take $d = x_s - \bar{x}$. Then $\frac{g_i(\bar{x}+td) - g_i(\bar{x})}{t} < 0$ at $t = 1$.

(since $g_i(\bar{x} + 1 \cdot d) = g_i(x_s) < 0$ and $g_i(\bar{x}) = 0$).

Since this is nondecreasing in $t > 0$ for convex functions, taking the limit as $t \searrow 0^+$ ensures $\langle \nabla g_i(\bar{x}), d \rangle < 0$. □

Theorem (Karush-Kuhn-Tucker (KKT) Conditions)

If $\bar{x}$ is a local minimizer and f, g_i are differentiable with MFCQ at $\bar{x}$, then there exists a Lagrange Multiplier Vector.

(i.e. $\lambda \geq 0$ with $\nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0$

$\lambda_i g_i(\bar{x}) = 0 \quad \forall i = 1 \ldots m$.)

Proof. Fritz-John using Gordan's theorem to rule out $\lambda_0 = 0$. □

An aside Let's apply the KKT conditions to an LP to see how we recover our LP dual formulation.

Consider $\begin{cases} \min & f(y) = \langle -b, y \rangle \\ \text{s.t.} & g_i(y) = \langle A_i, y \rangle \le c_i \end{cases}$ (note this is the dual of a standard form LP.)

Assume a Slater pt exists.

Then $\bar{y}$ is a local minimizer implies

there exists $\bar{x} \ge 0$ such that

$$\begin{cases} -b + \sum_{i=1}^{m} \bar{x}_i A_i = 0 & \text{(i.e. } \bar{x} \text{ is primal feasible } A\bar{x}=b\text{.)} \\ \bar{x}^T(c - A^T\bar{y}) = 0. & \text{(i.e. complementary slackness)} \end{cases}$$

Hence our Lagrange Multipliers are exactly our dual variables from linear programming.

Proof of Fritz John Theorem

Suppose $\bar{x}$ is a local min of $\begin{cases} \min f(x) \\ \text{s.t. } g_i(x) \le 0 \ \forall i. \end{cases}$

Let $h(x) = \max\{ f(x) - f(\bar{x}),\ g_i(x) \mid i \in I(\bar{x}) \}$.

Note all x near $\bar{x}$ with $g_i(x) \le 0 \ \forall i$ have

$$h(x) \ge f(x) - f(\bar{x}) \ge 0.$$

and all x near $\bar{x}$ with some $g_i(x) > 0$ have

$$h(x) \ge g_i(x) > 0.$$

$\Rightarrow \bar{x}$ is a local minimizer of h (without any constraints).

Claim: $h'(\bar{x};d) = \max\{\langle \nabla f(\bar{x}), d\rangle, \langle \nabla g_i(\bar{x}), d\rangle \mid i \in I(\bar{x})\}$

Then our first-order optimality condition implies

$$h'(\bar{x};d) \geq 0 \quad \forall d \in E$$

$$\Leftrightarrow \text{No } d \in E \text{ has } \begin{cases} \langle \nabla f(\bar{x}), d\rangle < 0 \\ \langle \nabla g_i(\bar{x}), d\rangle < 0 \quad \forall i \in I(\bar{x}) \end{cases}$$

by Gordan's Theorem

$$\Leftrightarrow \exists \lambda_0, \lambda \in \mathbb{R}^m \text{ s.t. } \begin{cases} \lambda_0, \lambda \geq 0 \\ \lambda_0 + \sum \lambda_i = 1 \\ \lambda_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0. \end{cases}$$ □

4. Algorithms (given simple constraints)

The next unit of this course will develop a deep, strong duality theory extending the above KKT ideas and enabling several sophisticated families of algorithms.

Here we will first introduce three algorithms that work well when constraints are sufficiently simple:

- Frank-Wolfe (Conditional Gradient Method)
- Projected Gradient Descent.
- Proximal Point Method.

In order of least assumptions on f, Ω to most.

The Frank-Wolfe Method

Consider a nonlinear optimization problem $\begin{cases} \min f(x) \\ \text{s.t. } x \in Q \end{cases}$
(with $f \in C^1$ and $Q \subseteq E$ convex)

and assume Q is simple enough we can solve linear optimization problems over it:

$$c \mapsto \begin{cases} \min \langle c, x \rangle \\ \text{s.t. } x \in Q. \end{cases}$$

For example, if $Q = \text{convexhull}(p_1, \dots, p_m)$, then this can be computed as $\min_{i=1\dots m} c^T p_i$.

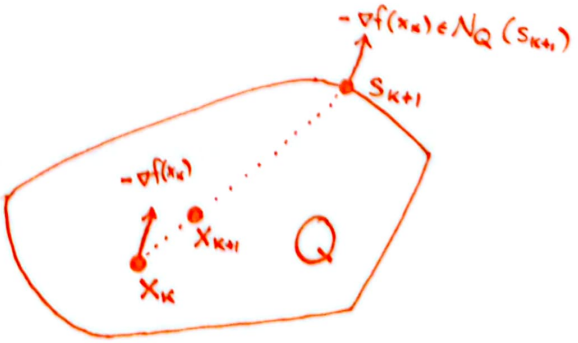
if $Q = \{x \mid \|x\| \leq 1\}$, then $x = \frac{-c}{\|c\|}$ minimizes with $-\|c\|$.
↑ using the norm induced by our inner product.

Frank-Wolfe aims to solve this by linearizing f at the current solution and moving a bit towards the linear min:

$$\text{(FW)} \quad \begin{cases} x_{k+1} = x_k + \alpha_k (s_{k+1} - x_k) \\ \text{where } s_{k+1} = \operatorname{argmin} \{\langle \nabla f(x_k), s \rangle \mid s \in Q\}, \end{cases}$$

with step length $\alpha_k \in (0, 1]$.

Example Picture of One FW Iteration:



Theorem (Convergence of Frank-Wolfe)

Consider any closed, convex, bounded Q (with all $x \in Q$ having $\|x\| \le D$) and any $f \in C^1$ with L-Lipschitz gradient.

Then setting $\alpha_k = 1/\sqrt{k+1}$, Frank-Wolfe has

$$0 \le \min_{k=0\dots T} \langle \nabla f(x_k), x_k - s_{k+1} \rangle \le \frac{f(x_0) - p^* + 2LD^2 \log(T)}{\sqrt{T+1}}.$$

If additionally f is convex, setting $\alpha_k = 1/k+1$ has

$$0 \le \min_{k=0\dots T} f(x_k) - p^* \le \frac{2LD^2 \log(T)}{T+1}.$$

These log factors are not fundamental and are removed by more careful analysis.

Note $\langle \nabla f(x_k), x_k - s_{k+1} \rangle =$ "the objective decrease found in our linearized subproblem" ≥ 0.

So the generic (nonconvex) result is guaranteeing some x_k is approximately optimal to its linearization of the problem.

The following two lemmas give more insight into the quantity: (when $y = s_{k+1}$, $x = x_k$)

Lemma 1 For any $f \in C^1$ with L-Lipschitz gradient,

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad \forall x, y \in E.$$

(Proof is applying the Fundamental Theorem of Calculus.)

Lemma 2 For any $f \in C^1$ that is convex,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in E.$$

(Proof uses the nondecreasingness of $t \mapsto \frac{f(x + t(y-x)) - f(x)}{t}$ at $t = 1$.)

Proof. Let $G_k = \langle \nabla f(x_k), x_k - s_{k+1} \rangle$.

Then we have the following key recurrence relation

$$f(x_{k+1}) - p^* = f(x_k + \alpha_k (s_{k+1} - x_k)) - p^*$$

$$\leq f(x_k) - p^* - \alpha_k G_k + \frac{L}{2} \|\alpha_k (s_{k+1} - x_k)\|^2 \quad \leftarrow \text{by Lemma 1}$$

$$\leq f(x_k) - p^* - \alpha_k G_k + \alpha_k^2 \, 2LD^2 \quad \leftarrow \text{by the triangle inequality and } \|x_k\| \leq D, \ \|s_{k+1}\| \leq D$$

Inductively applying this with $k = 0 \ldots T$ gives our first claim

$$0 \leq f(x_{T+1}) - p^* \leq f(x_0) - p^* - \sum_{k=0}^{T} \alpha_k G_k + 2LD^2 \sum_{k=0}^{T} \alpha_k^2$$

$$\Rightarrow \sum_{k=0}^{T} \alpha_k G_k \leq f(x_0) - p^* + 2LD^2 \sum_{k=0}^{T} \alpha_k^2$$

$$\Rightarrow \sqrt{T+1} \min_{k \leq T} G_k \leq f(x_0) - p^* + 2LD^2 \log(T). \quad \leftarrow \text{computing that } \sum \alpha_k \geq \sqrt{T+1}, \ \sum \alpha_k^2 \leq \log(T) \text{ for } \alpha_k = 1/\sqrt{k+1}.$$

Using Lemma 2 to bound $G_k \geq f(x_k) - p^*$, our key recurrence becomes

$$f(x_{k+1}) - p^* \leq (1 - \alpha_k)(f(x_k) - p^*) + \alpha_k^2 \cdot 2LD^2.$$

Plugging in $\alpha_k = 1/k+1$ and inductively applying this yields

$$f(x_{T+1}) - p^* \leq \underbrace{\prod_{k=0}^{T} \left(1 - \frac{1}{k+1}\right)}_{=0} (f(x_0) - p^*) + \sum_{k=0}^{T} \underbrace{\prod_{j=k+1}^{T} \left(1 - \frac{1}{j+1}\right)}_{= \frac{1}{k+1} \cdot \frac{1}{T+1}} \alpha_k^2 \, 2LD^2$$

$$= 2LD^2 \underbrace{\frac{\log(T)}{T+1}}_{\leq \frac{\log(T)}{T+1}}. \qquad \square$$

Projected Gradient Descent

Continue to consider nonlinear problems $\begin{cases} \min f(x) \\ \text{s.t. } x \in Q \end{cases}$
(with $f \in C^1$ and $Q \subseteq E$ convex)

and assume Q is simple enough we can solve the following (quadratic) orthogonal projection optimization problems over it:

$$\bar{x} \longmapsto \begin{cases} \min \frac{1}{2}\|x - \bar{x}\|^2 \\ \text{s.t. } x \in Q. \end{cases}$$

For example, if $Q = K^{NN} = \{x \geq 0\}$, then this problem is minimized at $x_i = \max\{0, \bar{x}_i\}$ (i.e. make all negative entries zero).

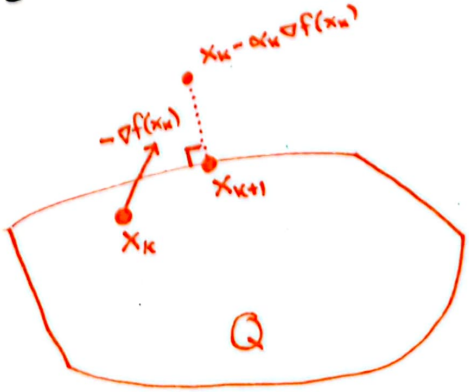
For a generic polyhedron Q, this subproblem is a QP.

Projected Gradient Descent (as the name implies) alternates between descending in obj value via $-\nabla f(x_k)$ and projecting on to Q to maintain feasibility:

(GD) $$x_{k+1} = \text{proj}_Q\{x_k - \alpha_k \nabla f(x_k)\}$$

with $\alpha_k > 0$ and $\text{proj}_Q(\bar{x}) = \text{argmin}\{\frac{1}{2}\|x - \bar{x}\|^2 \mid x \in Q\}$.

Example Picture of One GD Iteration:



Theorem (Convergence of Projected Gradient Descent)

Consider any closed, convex Q and $f \in C^1$ with L-Lipschitz gradient.

Then setting $\alpha_k = 1/L$, Projected Gradient Descent has

$$\min_{k=0\ldots T} \|x_{k+1} - x_k\| \leq \sqrt{\frac{2(f(x_0) - p^*)}{L(T+1)}}.$$

If additionally f is convex, it has

$$f(x_T) - p^* \leq \frac{2LD^2}{T+1}$$

where $D \geq \max_{k \leq T} \|x_k - x^*\|$.

Note $x_{k+1} - x_k$ can be understood as a sum of gradients and normal vectors below, equal to zero when the first-order optimality condition is attained... (exactly)

Lemma 3 Projected Gradient Descent has $n_k := (x_k - \alpha_k \nabla f(x_k)) - x_{k+1} \in N_Q(x_{k+1})$.

(Hence $\|x_{k+1} - x_k\| = \|\alpha_k \nabla f(x_k) + n_k\|$, so when $x_{k+1} - x_k$ is small, $\nabla f(x_k)$ is nearly in the normal of Q at x_{k+1}.)

Proof. The first order optimality condition for the projection subproblem ensures

$$\underbrace{-\nabla\left(\tfrac{1}{2}\|\cdot - (x_k - \alpha_k \nabla f(x_k))\|^2\right)^{(x_{k+1})}}_{= (x_k - \alpha_k \nabla f(x_k)) - x_{k+1} = n_k} \in N_Q(x_{k+1})$$

□

Proof. First we derive a recurrence showing the objective gap decreasing in each iteration of (PGD):

$$f(x_{k+1}) - p^* \le f(x_k) - p^* + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

(by Lemma 1 above)

$$\le f(x_k) - p^* + \frac{1}{\alpha_k}\langle \alpha_k \nabla f(x_k) + n_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

(by definition of normal vectors at x_{k+1})

$$= f(x_k) - p^* \underbrace{- \left(\frac{1}{\alpha_k} - \frac{L}{2}\right)}_{= -\frac{L}{2} \text{ (when } \alpha_k = \frac{1}{L})} \|x_{k+1} - x_k\|^2 .$$

(by Lemma 3 above)

Inductively applying this with $k = 0 \dots T$ gives our generic guarantee

as $0 \le f(x_{T+1}) - p^* \le f(x_0) - p^* - \sum_{k=0}^{T} \frac{L}{2}\|x_{k+1} - x_k\|^2$ and so

$$\min_{k \le T} \|x_{k+1} - x_k\|^2 \le \frac{1}{T+1}\sum_{k=0}^{T} \|x_{k+1} - x_k\|^2 \le \frac{2(f(x_0) - p^*)}{L(T+1)} .$$

Now suppose f is convex, we claim $\|x_{k+1} - x_k\| \ge \frac{f(x_{k+1}) - f(x^*)}{2L\|x_{k+1} - x^*\|}$.

[Proof. Convexity ensures $f(x^*) \ge f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle$.
Normality ensures $\langle n_k, x^* - x_{k+1} \rangle \le 0$
$\Rightarrow \langle \nabla f(x_{k+1}) + n_k, x^* - x_{k+1} \rangle \le f(x^*) - f(x_{k+1})$
By Lemma 3 and then bounding each inner product, one can conclude
$2L\|x_{k+1} - x_k\|\|x^* - x_{k+1}\| \ge f(x_{k+1}) - f(x_{k+1})$. □]

Then our recurrence in the convex case becomes:

$$f(x_{k+1}) - p^* \le f(x_k) - p^* - \frac{(f(x_{k+1}) - p^*)^2}{2LD^2} .$$

One can verify this is solved by $f(x_T) - p^* \le \frac{2LD^2}{T+1}$. □

The Proximal Point Method

Continue to consider nonlinear problems $\begin{cases} \min f(x) \\ \text{s.t. } x \in Q \end{cases}$
(with $f \in C^1$ and $Q \subseteq E$ convex)

and assume f and Q are simple enough we can solve the following quadratically penalized problem:

$$\bar{x} \longmapsto \begin{cases} \min & f(x) + \frac{\rho}{2}\|x - \bar{x}\|^2 \\ \text{s.t.} & x \in Q. \end{cases}$$

for $\rho > 0$.

This is a very hard subproblem in most cases. As a result, this method is more conceptually useful than practically.

Then the Proximal Point Method iterates

$$\text{(PPM)} \quad x_{k+1} \in \operatorname{argmin}\left\{ f(x) + \frac{\rho}{2}\|x - x_k\|^2 \;\middle|\; x \in Q \right\}.$$

Note the optimality condition for this subproblem ensures

$$n_{k+1} := -\nabla\left(f + \tfrac{\rho}{2}\|\cdot - x_k\|^2\right)(x_{k+1}) \in N_Q(x_{k+1})$$

$$\Rightarrow \quad x_{k+1} = x_k - \tfrac{1}{\rho}\left(\nabla f(x_{k+1}) + n_{k+1}\right)$$

and when this iteration is nearly stationary ($x_{k+1} \approx x_k$), we have $-\nabla f(x_{k+1})$ nearly in $N_Q(x_{k+1})$. This is similar to Lemma 3 for Projected Gradient Descent

The convergence theory here follows the same form as (GD)'s.

By definition, we have the recurrence

$$f(x_k) + \frac{\rho}{2}\|x_k - x_k\|^2 \geq f(x_{k+1}) + \frac{\rho}{2}\|x_{k+1} - x_k\|^2$$

$$\Rightarrow f(x_{k+1}) - p^* \leq f(x_k) - p^* - \frac{\rho}{2}\|x_{k+1} - x_k\|^2.$$

This then gives $O(1/\sqrt{T})$ general and $O(1/T)$ convex rates of convergence.