Continuous Optimization

CO 466

Levent Tunçel

LATEXed by Sibelius Peng

Preface

Disclaimer Much of the information on this set of notes is transcribed directly/indirectly from the lectures of CO 466 during Fall 2020 as well as other related resources. I do not make any warranties about the completeness, reliability and accuracy of this set of notes. Use at your own risk.

$\mathbb{Z}_+ = \{0, 1, 2, 3, \ldots\}$; $\mathbb{Z}_{++} = \{1, 2, 3, \ldots\}$

For any questions, send me an email via https://notes.sibeliusp.com/contact/.

You can find my notes for other courses on https://notes.sibeliusp.com/.

Sibelius Peng

Contents

1

Introduction: Formulations, fundamental background and definitions

Let $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p$ all continuous.

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & g(x) \leq 0 \\ & h(x) = 0 \end{array} \tag{P}$$

$$S := \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$$

is called the feasible solution set of (P), equivalently feasible region of (P).

Definition 1: global minimizer

$\bar{x} \in \mathbb{R}^n$ is a **global minimizer** of (P) if $x \in S$ and $f(x) \geq f(\bar{x})$ for all $x \in S$.

Sometimes we simply say $\bar{x}$ is a minimizer of (P).

Definition 2: local minimizer

$\bar{x} \in \mathbb{R}^n$ is a **local minimizer** of (P) if $\bar{x} \in S$ and there exists a neighborhood U of $\bar{x}$ such that

$$f(x) \geq f(\bar{x}) \quad \forall x \in S \cap U$$

$\bar{x} \in \mathbb{R}^n$ is a **strict local minimizer** of (P) if $\bar{x} \in S$ and there exists a neighborhood U of $\bar{x}$ such that

$$f(x) > f(\bar{x}) \quad \forall x \in (S \cap U) \setminus \{\bar{x}\}$$

$\bar{x} \in \mathbb{R}^n$ is an **isolated local minimizer** of (P), if $\bar{x} \in S$ and there exists a neighborhood U of $\bar{x}$ such that $\bar{x}$ is the only local minimizer of (P) in $(S \cap U)$.

continuous optimization problem

A **continuous optimization problem** is a problem of optimizing (minimizing or maximizing) a continuous function of finitely many real variables subject to finitely many equations and inequalities on continuous functions of these variables.

What kind of problems can be formulated as Continuous Optimization problems? Almost everything.

Example 3: Fermat's Last Theorem

"There do not exist positive integers x, y, z and an integer $n \geq 3$ such that $x^n + y^n = z^n$."

Consider

$$\begin{aligned} \inf \quad & f(x) := \left(x_1^{x_4} + x_2^{x_4} - x_3^{x_4}\right)^2 + \sum_{i=1}^{4} \left(\sin(\pi x_i)\right)^2 \\ \text{s.t.} \quad & g_1(x) := 1 - x_1 \leq 0 \\ & g_2(x) := 1 - x_2 \leq 0 \\ & g_3(x) := 1 - x_3 \leq 0 \\ & g_4(x) := 3 - x_4 \leq 0 \end{aligned} \tag{P}$$

The optimal objective value of (P) is zero and attained if and only if FLT is false.

We can show that (P) has a sequence of feasible solutions $\{x^{(k)}\}$ such that $f(x^{(k)}) \searrow 0$. Since $f(x) \geq 0$ for all $x \in \mathbb{R}^4$, the optimal value of (P) is zero.

FLT is true if and only if (P) does not attain its optimal value (of zero).

Even when the number of variables in a continuous optimization problem is very small (e.g., 4) the optimization problem may be notoriously hard. Even discrete structures can be formulated in our framework. $\sin(\pi x_1) = 0 \iff x_1 \in \mathbb{Z}$. In Example 3, we have functions that are "highly nonlinear".

Example 4: Combinatorial Optimization, 0,1 Integer Programming

Let m, n be positive integers, $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$ be given. Consider the 0,1 Integer Programming problem.

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \in \{0,1\}^n \end{aligned} \tag{IP}$$

The first condition can be written as

$$g(x) := Ax - b \leq 0.$$

The second condition can be written as

$$x_j(x_j - 1) = 0 \quad \forall j \in [n] \leftrightarrow h(x) = 0$$

Our continuous optimization problem is only mildly nonlinear.

Some conclusions from Example 3 and 4

Continuous Optimization problems acn be very hard even when the number of variables and constraints are both small, the nonlinearity in f, g, h is very mild.

To successfully solve Continuous Optimization Problems, we must study the problem class at hand, discover special properties and structures and then exploit these special properties & structures.

1.1 Conic Form

Definition 5: cone

A set $K \subseteq \mathbb{R}^n$ is a **cone** if $\forall x \in K, \forall k \in \mathbb{R}_+, \lambda x \in K$.

Definition 6: convex

A set $S \subseteq \mathbb{R}^n$ is **convex** if for every pair of points in S, the line segment joining them lies entirely in S.

That is S is convex if $\forall u, v \in S, \forall \lambda \in [0,1], [\lambda u + (1-\lambda)v] \in S$.

Definition 7: convex cone

A set $K \subseteq \mathbb{R}^n$ is a **convex cone** if it is convex and is a cone.

Let $g : \mathbb{R}^n \to \mathbb{R}^m, f : \mathbb{R}^n \to \mathbb{R}$ be continuous functions. Consider

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & g(x) \preceq_K 0 \end{array}$$

where $K \subseteq \mathbb{R}^m$ is a convex cone and for every $u, v \in \mathbb{R}^m$, $u \succeq_K v$ means $(u - v) \in K$.

This is at least as general as our original (P). Consider $\mathbb{R}^p \ni K := \mathbb{R}^m_+ \oplus \{0\} \ldots$

1.2 Derivatives

Definition 8: directional derivative

The **directional derivative** of $f : \mathbb{R}^n \to \mathbb{R}$ at $\bar{x} \in \mathbb{R}^n$ along the direction $d \in \mathbb{R}^n$ is

$$f'(\bar{x}; d) := \lim_{\alpha \searrow 0} \frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha}$$

(Gâteaux (directional) derivative)

Exercise:
What is the directional derivative of $f : \mathbb{R}^n \to \mathbb{R}, f(x) := \|x\|_\infty$, for every $\bar{x}, d \in \mathbb{R}^n$?

Definition 9: differentiable

$f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $\bar{x} \in \mathbb{R}^n$ if $\exists \mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$, linear, such that

$$\lim_{\substack{h \to 0 \\ (h \in \mathbb{R}^n)}} \frac{\left\| f(\bar{x} + h) - [f(\bar{x}) + \mathcal{A}(h)] \right\|}{\|h\|} = 0$$

Such $\mathcal{A}$ is called the derivative of f at $\bar{x}$ and is denoted by $Df(\bar{x})$ or $f'(\bar{x})$ (matrix representation of $Df(\bar{x})$). We will also use $\nabla f(\bar{x}) := [f'(x)]^T$.

Suppose $f : \mathbb{E}_1 \to \mathbb{E}_2$, we have

$$\begin{array}{ll} Df(\bar{x}) \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2), & Df : \mathbb{E}_1 \to \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2) \\ D^2 f(\bar{x}) \in \mathcal{L}(\mathbb{E}_1, \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)), & D^2 f : \mathbb{E}_1 \to \mathcal{L}(\mathbb{E}_1, \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)) \end{array}$$

If $f : \mathbb{R}^n \to \mathbb{R}$, then $D^k f(\bar{x})[h^{(1)}, h^{(2)}, \ldots, h^{(k)}]$: k^{th} differential (derivative) along the directions $h^{(1)}, h^{(2)}, \ldots, h^{(k)} \in \mathbb{R}^n$.

Theorem 10: Taylor's Theorem

Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ be a $\mathcal{C}^r$ function on U. Let $x, d \in \mathbb{R}^n$. If $x, (x+d)$, and the line segment joining x and $(x+d)$ lie in U, then there exists $z \in (x, x+d)$ such that

$$f(x+d) = f(x) + \sum_{k=1}^{r-1} \frac{1}{k!} D^k f(x) \underbrace{[d, d, \dots, d]}_{k\text{-times}} + \frac{1}{r!} D^r f(z) \underbrace{[d, d, \dots, d]}_{r\text{-times}}$$

Definition 11: contraction mapping

Let $U \subseteq \mathbb{R}^n$ be a closed set. $f : U \to U$ is called a **contraction mapping** if there exists $\lambda \in [0,1)$ such that

$$\|f(x) - f(y)\| \leq \lambda \|x - y\| \quad \forall x, y \in U$$

1.3 Fixed Point

Theorem 12: Banach Fixed Point Theorem (1922)

Let $U \subseteq \mathbb{R}^n$ be a closed set and let $f : U \to U$ be a contraction mapping, then

(i) (Existence and uniqueness of solution - fixed point)

the mapping f has a unique fixed point $\bar{x} \in U$.

(ii) (Algorithm and convergence)

For all $x^{(0)} \in U$, the sequence $\{x^{(k)}\}$ generated by $x^{(k+1)} := f\left(x^{(k)}\right), k \in \{0, 1, 2, \dots\}$ (fixed point iteration) converges to $\bar{x}$. In particular,

$$\|x^{(k)} - \bar{x}\| \leq \lambda^k \|x^{(0)} - \bar{x}\| \quad \forall k \in \{0, 1, 2, \dots\}$$

Proof:
Suppose $U \subseteq \mathbb{R}^n$ is a nonempty closed set, and $f : U \to U$ is a contraction mapping with $\lambda \in [0,1)$. Let $x^{(k+1)} := f\left(x^{(k)}\right)$ for all $k \in \mathbb{Z}_+$. Then for all $k \in \mathbb{Z}_+$,

$$\|x^{(k+1)} - x^{(k)}\| = \|f(x^{(k)}) - f(x^{(k-1)})\| \leq \lambda \|x^{(k)} - x^{(k-1)}\|$$

By induction on k, ... we obtain

$$\|x^{(k)} - x^{(k-1)}\| \leq \lambda^k \|x^{(1)} - x^{(0)}\| \quad \forall k \in \mathbb{Z}_+ \tag{$*$}$$

$\forall m \in \mathbb{Z}_{++}, \forall k \in \mathbb{Z}_{++}$,

$$\begin{aligned}
\left\|x^{(m+k)} - x^{(m)}\right\| &= \left\|x^{(m+k)} - x^{(m+k-1)} + x^{(m+k-1)} - x^{(m+k-2)} + \dots + x^{(m+1)} - x^{(m)}\right\| \\
&\leq \sum_{i=1}^{k} \left\|x^{(m+i)} - x^{(m+i-1)}\right\| && \triangle\text{-ineq} \\
&\leq (\lambda^{m+k-1} + \lambda^{m+k-2} + \dots + \lambda^m)\left\|x^{(1)} - x^{(0)}\right\| && \text{by } (*) \\
&= \lambda^m (1 + \lambda + \lambda^2 + \dots + \lambda^{k-1})\left\|x^{(1)} - x^{(0)}\right\| \\
&= \frac{\lambda^m (1 - \lambda^k)}{1 - \lambda}\left\|x^{(1)} - x^{(0)}\right\| \\
&\leq \frac{\lambda^m}{1 - \lambda}\left\|x^{(1)} - x^{(0)}\right\| \to 0 \text{ as } m \to +\infty \text{ (independent of } k\text{)}
\end{aligned}$$

Therefore $\{x^{(k)}\}$ os a Cauchy sequence and hence it converges. Let $\bar{x}$ be its limit. $\bar{x} \in U$ (U is closed).

$\forall k \in \mathbb{Z}_+$, we have

$$\|f(\bar{x}) - \bar{x}\| \leq \|f(\bar{x}) - x^{(k)}\| + \|x^{(k)} - \bar{x}\| \leq \lambda \underbrace{\|\bar{x} - x^{(k-1)}\|}_{\to 0} + \underbrace{\|x^{(k)} - \bar{x}\|}_{\to 0}$$

As $k \to +\infty$, RHS $\to 0$. Thus $f(\bar{x}) = \bar{x}$. This proves the existence. Now we prove the uniqueness.

Suppose $\exists \bar{x}, \bar{y} \in U$ such that $f(\bar{x}) = \bar{x}$ and $f(\bar{y}) = \bar{y}$. Then

$$\|\bar{x} - \bar{y}\| = \|f(\bar{x}) - f(\bar{y})\| \leq \lambda \|\bar{x} - \bar{y}\| \implies (1-\lambda)\|\bar{x} - \bar{y}\| = 0 \underset{\lambda \in [0,1)}{\implies} \bar{x} = \bar{y}$$

Now that we have established existence and uniqueness of $|barx$, for a proof of convergence rate claim, we proceed as in the beginning of the proof. However, we use $\bar{x}$.

$$\|x^{(1)} - \bar{x}\| = \|f(x^{(0)}) - f(\bar{x})\| \leq \lambda \|x^{(0)} - \bar{x}\| \implies \|x^{(2)} - \bar{x}\| \leq \lambda^2 \|x^{(0)} - \bar{x}\|$$

By induction on k, we have

$$\|x^{(k)} - \bar{x}\| \leq \lambda^k \|x^{(0)} - \bar{x}\| \quad \forall k \in \mathbb{Z}_k$$

as desired. □

Theorem 13: Brouwer's Fixed Point Theorem (1910)

Let $U \subset \mathbb{R}^n$ be a nonempty, compact and convex set; let $f : U \to U$ continuous such that $f(U) = U$. Then there exists $\bar{x} \in U$ such that $f(\bar{x}) = \bar{x}$.

See the application in https://n.sibp.ro/co/456.

Theorem 14: Kakutani's Fixed Point Theorem (1941)

Let $U \subset \mathbb{R}^n$ be a nonempty, compact convex set and $f : U \to 2^U$ be a set valued map on U. If $\text{Graph}(f) := \left\{ \binom{x}{v} \in U \oplus U : v \in f(x) \right\}$ is closed and $f(x)$ is nonempty and convex for every $x \in U$, then there exists $\bar{x} \in U$ such that $\bar{x} \in f(\bar{x})$.

Theorem 15: Borsuk-Ulam Theorem (1930-1933)

Let $f : \{x \in \mathbb{R}^{n+1} : \|x\|_2 = 1\} \to \mathbb{R}^n$ be continuous. Then there exists $\bar{x} \in \mathbb{R}^{n+1}$ such that $\|\bar{x}\|_2 = 1$ and $f(\bar{x}) = f(-\bar{x})$.

Example:

Let $n := 2$. Assuming temperature and barometric air pressure are continuous functions on the Earth's surface, and Earth's surface is homeomorphic to a sphere, there always exists an antipodal pair of points on Earth with the same temperature & the same air pressure.

1.4 Other

$\mathbb{S}^n := n \times n$ symmetric matrices with real entries.

Theorem 16: Spectral Decomposition Theorem

For every $A \in \mathbb{S}^n$, there exists $Q \in \mathbb{R}^{n\times n}$ orthogonal ($Q^TQ = I$) such that $A = QDQ^T$, where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix.

In the above theorem, the diagonal matrix D contains all eigenvalues of A, and the columns of Q are the corresponding eigenvectors of A.

Definition 17: positive definite

$A \in \mathbb{R}^{n\times n}$ is **positive semidefinite** if $h^TAh \geq 0$ for all $h \in \mathbb{R}^n$; such A is **positive definite** if $h^TAh > 0$ for all $h \in \mathbb{R}^n \setminus \{0\}$.

If $A \in \mathbb{R}^{n\times n}$ is skew-symmetric ($A = -A^T$), then $h^TAh = (h^TAh)^T = -h^TAh = 0$ for all $h \in \mathbb{R}^n$. Therefore, such A is positive semidefinite but not positive definite.

$$\begin{aligned} \mathbb{S}^n_+ &:= \text{positive semidefinite matrices in } \mathbb{S}^n, \\ \mathbb{S}^n_{++} &:= \text{positive definite matrices in } \mathbb{S}^n. \end{aligned}$$

In fact, $\mathbb{S}^n_{++} = \operatorname{int}(\mathbb{S}^n_+)$.

Theorem 18: Choleski Decomposition Theorem

Let $A \in \mathbb{S}^n$, then

(a) A is positive semidefinite if and only if there exists $L \in \mathbb{R}^{n\times n}$ lower triangular such that $A = LL^T$;

(b) A is positive definite if and only if there exists $L \in \mathbb{R}^{n\times n}$ lower triangular and nonsingular such that $A = LL^T$.

Note that Taylor's Theorem (Theorem 10) cannot be completely generalized to functions $f : \mathbb{R}^n \to \mathbb{R}^m$ with $m \geq 2$, even for $r = 1$. However, we have

Theorem 19

Let $U \subseteq \mathbb{R}^n$ be an open set and $f : U \to \mathbb{R}^m$ be a C^1 on U. Suppose for $\bar{x}, d \in \mathbb{R}^n$, $[\bar{x}, \bar{x} + d] \subset U$. Then

$$f(\bar{x} + d) - f(\bar{x}) = \int_0^1 Df(\bar{x} + \alpha d)d\ (\partial\alpha)$$

A consequence of this result is obtained when $Df(\cdot)$ is Lipschitz continuous on U (in a neighborhood of $[\bar{x}, \bar{x} + d]$ suffices). Let L denote the Lipschitz constant. Then

$$\|Df(x) - Df(y)\| \leq L\|x - y\| \quad \forall x, y \in U$$

Then we have

$$\begin{aligned} \left\|f(\bar{x} + d) - f(\bar{x}) - Df(\bar{x})d\right\|_2 &= \left\|\int_0^1 [Df(\bar{x} + \alpha d) - Df(\bar{x})]d\ (\partial\alpha)\right\|_2 \\ &\leq \int_0^1 \underset{\substack{\uparrow \\ \text{operator} \\ \text{2-norm}}}{\left\|Df(\bar{x} - \alpha d) - Df(\bar{x})\right\|_2} \cdot \underset{\substack{\uparrow \\ \text{2-norm} \\ \text{on } \mathbb{R}^n}}{\|d\|_2} (\partial\alpha) && \text{see aside} \\ &\leq \int_0^1 L\|d\|_2 \cdot \|d\|_2\alpha(\partial\alpha) \\ &= \frac{1}{2}L\|d\|_2^2 \end{aligned}$$

So, if $\|d\|_2 < \epsilon$, then this error in this first-order estimate of $f(\bar{x}+d)$ is bounded above by $\frac{1}{2}L\epsilon^2$.

> Aside:
>
> Let $h := \int_0^1 \left[Df(\bar{x}+\alpha d) - Df(\bar{x})\right]d\ (\partial\alpha)$, then
>
> $$\begin{aligned} \|h\|_2^2 = h^T h = h^T \int_0^1 \left[Df(\bar{x}+\alpha d) - Df(\bar{x})\right]d\ (\partial\alpha) \\ = \int_0^1 h^T\left[Df(\bar{x}+\alpha d) - Df(\bar{x})\right]d\ (\partial\alpha) \\ \leq \int_0^1 \|h\|_2 \left\|\left[Df(\bar{x}+\alpha d) - Df(\bar{x})\right]d\right\|_2 (\partial\alpha) \qquad \text{Cauchy-Schwarz} \end{aligned}$$
>
> $$\implies \|h\|_2 \leq \int_0^1 \left\|\left[Df(\bar{x}+\alpha d) - Df(\bar{x})\right]d\right\|_2 (\partial\alpha)$$

Note that we may replace f in Theorem 19 by $Df^r(\cdot)$ (assuming $f \in \mathcal{C}^{r+1}$) and apply the same reasoning. Indeed, Theorem 19 can be very useful in the design and analysis of continuous optimization algorithms.

Theorem 20: Inverse Function Theorem

Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^n$ be $\mathcal{C}^1$, $\bar{x} \in U$, $\det(\nabla f(\bar{x})) \neq 0$. Then there exists an open neighborhood V of $\bar{x}$ in U and an open neighborhood W of $f(\bar{x})$ such that

- $f(V) = W$,
- f has a local $\mathcal{C}^1$ inverse $f^{-1} : W \to V$,
- $\forall y \in W$, with $x = f^{-1}(y)$, we have $Df^{-1}(y) = \left[Df(x)\right]^{-1}$.

In the above, if f is $\mathcal{C}^r$, then there exists such an $f^{-1} \in \mathcal{C}^r$. Theorem 20 can be proved by utilizing Theorem 12 (in showing that the inverse is well-defined, i.e., one-to-one).

Theorem 21: Implicit Function Theorem

Let $h : \mathbb{R}^n \to \mathbb{R}^p$, $h \in \mathcal{C}^1$ in a neighborhood of $\bar{x} \in \mathbb{R}^n$ where $h(\bar{x}) = 0$. Suppose $h'(\bar{x})$ has full row rank ($\text{rank}(h'(\bar{x})) = p \leq n$). Define a partition $[B|N]$ of columns of $h'(\bar{x})$:

$$h' =: \left[\underset{\substack{\uparrow \\ \in \mathbb{R}^{p\times p}, \\ \text{nonsingular,} \\ \text{partition}}}{h'_B(\bar{x})}\ |h'_N(\bar{x})\right]$$

$\bar{x}$ and x with respect to the same $[B|N]$. Then there exist neighborhoods U_B of $\bar{x}_B$ and U_N of $\bar{x}_N$ and a $\mathcal{C}^1$ function $f : U_N \to U_B$ such that

- $f(\bar{x}_N) = \bar{x}_B$,
- $h\binom{x_B}{x_N} = 0 \iff x_B = f(x_N)$ for all $x_B \in U_B, x_N \in U_N$.

Moreover, $f'(x_N) = -[h'_B(\bar{x})]^{-1}h'_N(\bar{x})$.

Recall the very special case (e.g., equality constraints in an LP problem): $A \in \mathbb{R}^{p\times n}$, $\text{rank}(A) = p$ given

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

$$
\begin{aligned}
h(x) := Ax - b &\implies h'(x) = A \\
\bar{x}_B &= A_B^{-1}b - A_B^{-1}A_N\bar{x}_N \\
x_B &= A_B^{-1}b - A_B^{-1}A_N x_N \\
f(x_N) &:= A_B^{-1}b - A_B^{-1}A_N x_N
\end{aligned}
$$

In this setting $U_B := \mathbb{R}^p, U_N := \mathbb{R}^{n-p}$.

Lemma 22: Chain Rule

Let $U \subseteq \mathbb{R}^n, V \subseteq \mathbb{R}^m$ be both open sets. $f_1 : U \to \mathbb{R}^m, f_2 : V \to \mathbb{R}^p$ be differentiable on U and V respectively such that $f_1(U) \subseteq V$. Then $(f_2 \circ f_1)$ is differentiable on U and

$$D(f_2 \circ f_1)(\bar{x}) = Df_2(f_1(\bar{x})) \circ Df(\bar{x}) \qquad \forall \bar{x} \in U$$

Example: Line search, directional derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$. Also, given are a current point $\bar{x} \in \mathbb{R}^n$ and a “search direction” $d \in \mathbb{R}^n$. We define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(\alpha) := f(\bar{x} + \alpha d)$. Then $\phi'(\alpha) = \langle \nabla f(\bar{x} + \alpha d), d \rangle$. If f is $\mathcal{C}^2$, then $\phi''(\alpha) = d^T \nabla^2 f(\bar{x} + \alpha d) d$. Note $\phi'(0) = \langle \nabla f(\bar{x}), d \rangle, \phi''(0) = d^T \nabla^2 f(\bar{x}) d$.

Corollary 23

Suppose h and $\bar{x}$ are as in Theorem 21 (Implicit Function Theorem). Also assume $Z \in \mathbb{R}^{n \times p}$ $(q \le n - p)$ such that $h'(\bar{x})Z = 0$. Then there exists a neighborhood U of $0 \in \mathbb{R}^q$ and a $\mathcal{C}^1$ function $t : U \to \mathbb{R}^n$ such that

- $t(0) = 0$,
- $t'(0) = 0$,
- $h\big(\bar{x} + Zd_Z + t(d_Z)\big) = 0$ for all $d_Z \in U$.

So the function t gives us a way of moving away from $\bar{x}$ (a solution of $h(x) = 0$) in a way that keeps feasible with respect to $h(x) = 0$.

Proof:

Let $h, \bar{x}$ and Z be as in the assumptions. Using the partition $[B|N]$, define $z =: \begin{bmatrix} z_B \\ z_N \end{bmatrix}$ (recall $h'(\bar{x}) = [h'_B(\bar{x})|h'_N(\bar{x})]$). Let $U := \big\{ d_Z \in \mathbb{R}^q : (\bar{x}_N + Z_N d_Z) \in \underset{\substack{\uparrow \\ \text{neighborhood} \\ \text{of } \bar{x}_N \text{ from} \\ \text{Theorem 21}}}{U_N} \big\}$. Define t by

$$
\begin{aligned}
t_N(d_Z) &:= 0 \\
t_B(d_Z) &:= f(\bar{x}_N + Z_N d_Z) - \bar{x}_B - Z_B d_Z
\end{aligned}
$$

Thus,

$$h\big(\bar{x} + Zd_Z + t(d_Z)\big) = h \begin{bmatrix} \bar{x}_B + Z_B d_Z + f(\bar{x}_N + Z_N d_Z) - \bar{x}_B - z_B d_Z \\ \bar{x}_N + Z_N d_Z + 0 \end{bmatrix} = h \begin{bmatrix} f_N(\bar{x}_N + Z_N d_Z) \\ \bar{x}_N + Z_N d_Z \end{bmatrix} \underset{\substack{\uparrow \\ \text{By Theorem 21}}}{=} 0$$

Also,

$$t(0) = f(\bar{x}_N) - \bar{x}_B = 0, \qquad t'_N(0) = 0$$

$$t'_B(0) \underset{\substack{\uparrow \\ \text{Chain rule (Lemma 22)}}}{=} f'(\bar{x}_N)Z_N - Z_B = -[h'_B(\bar{x})]^{-1} h'_N(\bar{x}) Z_N - Z_B = [h'_B(\bar{x})]^{-1} \underbrace{[-h'_N(\bar{x})Z_N - h'_B(\bar{x})Z_N]}_{=-h'(\bar{x})Z=0} = 0$$

□

What does the size of the neighborhood depend on?

Note in LPs $t(d_Z) := 0$ for all $d_Z \in \mathbb{R}^q$.

Corollary 24

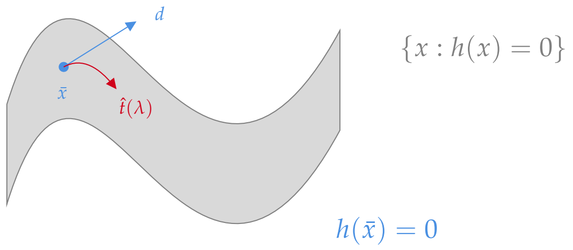
Assume h and $\bar{x}$ are as described in Theorem 21. Let $d \in \mathbb{R}^n$ such that $h'(\bar{x})d = 0$. Then there exists $\bar{\lambda} > 0$ and a $\mathcal{C}^1$ arc (directed curve) $\hat{t}$ with properties:

$$\begin{cases} \hat{t}(0) = \bar{x} \\ h\big(\hat{t}(\lambda)\big) = 0 \quad \forall \lambda \in [0, \bar{\lambda}) \\ \hat{t}'(0) = d \end{cases}$$

Proof:
In the statement of Corollary 23, plug in $Z := d$ and then using the resulting t,

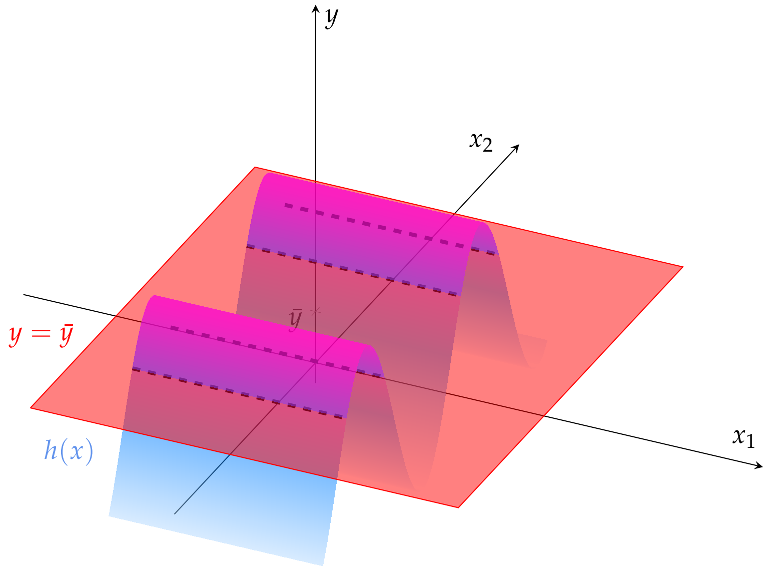
$$\hat{t}(\lambda) := \bar{x} + \underset{\substack{\uparrow \\ d_Z}}{\lambda}\ \underset{\substack{\uparrow \\ Z}}{d} + t(\lambda)$$

□



How applicable are the Theorems 20, 21 and their Corollaries?

Let $h : \mathbb{R}^n \to \mathbb{R}^p$, where $p \leq n$. Call $\bar{x} \in \mathbb{R}^n$ **regular** if $\operatorname{rank}(h'(\bar{x})) = p$; call $\bar{y} \in \mathbb{R}^p$ a **regular value** if $\forall x \in h^{-1}(\bar{y})$ are regular.



Union of these curves is the set $h^{-1}(\bar{y})$.

If h is affine, then $h(x) = Ax - b$ for some given $A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$. Let $\bar{y} \in \mathbb{R}^p$ be given. Then $h^{-1}(\bar{y}) = \{x \in \mathbb{R}^n : Ax = \bar{y} + b\}$.

Theorem 25: Sard's Theorem, Morse-Sard Theorem

Let $h : \mathbb{R}^n \to \mathbb{R}^p$, where $p \leq n$, $h \in \mathcal{C}^r$ with $r \geq n - p + 1$. Then the p-dimensional Lebesgue measure of $\{y \in \mathbb{R}^p : y \text{ is not a regular value}\}$ is zero.

Morse (1939) proved the $p = 1$ case, Sard (1942) proved the generalization above. Smale (1965) proved an infinite dimensional version.

2 Unconstrained Continuous Optimization

$f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p$.

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & g(x) \leq 0 \\ & h(x) = 0 \end{array} \tag{P}$$

$S := \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$. Here, we assume $S = \mathbb{R}^n$.

Theorem 26: First-order necessary conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^1$ and $S = \mathbb{R}^n$. Then, $\bar{x} \in \mathbb{R}^n$ is a local minimizer for (P) $\implies f'(\bar{x}) = 0$.

$\bar{x}$ is a **stationary point** of f.

Proof:

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1, S = \mathbb{R}^n$, and $\bar{x} \in \mathbb{R}^n$ is a local minimizer for (P). For the sake of seeking a contradiction, suppose $f'(\bar{x}) \neq 0$. Then, there exists $d \in \mathbb{R}^n$ such that $\langle f'(\bar{x}), d\rangle < 0$ (e.g., let $A \in \mathbb{S}^n_{++}$, and set $d := -Af'(\bar{x})$). Consider $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(\alpha) := f(\bar{x} + \alpha d)$. Then, $\phi'(0) = \langle f'(\bar{x}), d\rangle < 0$. Thus, for all sufficiently small, positive α, $f(\bar{x} + \alpha d) < f(\bar{x})$. Therefore, $\bar{x}$ is not a local minimizer for (P). □

Optimality conditions are widely used in algorithm design. E.g., for many software $\|\nabla f(x^{(k)})\| < \epsilon$ is a part of the stopping criteria.

Definition 27

$d \in \mathbb{R}^n$ is a **decent direction** for f at $\bar{x} \in \mathbb{R}^n$, if $\langle f'(\bar{x}), d\rangle < 0$.

$d \in \mathbb{R}^n$ is an **improving direction** for f at $\bar{x}$, if $f(\bar{x} + \alpha d) < f(\bar{x})$ $\forall \alpha > 0$ and sufficiently small.

Theorem 28: Second-order necessary conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^2$ and $S = \mathbb{R}^n$. If $\bar{x} \in \mathbb{R}^n$ a local minimizer for (P), then $f'(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \in \mathbb{S}^n_+$.

Proof:

Suppose $\bar{x}$ is a local minimizer for (P). Since f is $\mathcal{C}^2$ by Theorem 27, $f'(\bar{x}) = 0$. Suppose for the sake of contradiction that $\nabla^2 f(\bar{x}) \notin \mathbb{S}^n_+$. Since $f \in \mathcal{C}^2, \nabla^2 f(\bar{x}) \in \mathbb{S}^n$. Therefore, there exists $d \in \mathbb{R}^n$

such that $d^T\nabla^2 f(\bar{x})d < 0$. Define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(\alpha) := f(\bar{x}+\alpha d)$. Then $\phi'(0) = \langle \nabla f(\bar{x}), d\rangle = 0$, $\phi''(0) = d^T\nabla^2 f(\bar{x})d < 0$. Therefore, for all $\epsilon > 0$ and sufficiently small $f(\bar{x}+\epsilon d) < f(\bar{x})$ which contradicts the fact that $\bar{x}$ is a local minimizer for (P). □

Definition 29: direction of negative curvature

$d \in \mathbb{R}^n$ is called a **direction of negative curvature** for f at $\bar{x}$ if $d^T\nabla^2 f(\bar{x})d < 0$.

Theorem 30: Taylor's Theorem - implicit remainder version

Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ be $\mathcal{C}^r$ on U. Let $\bar{x}, d \in \mathbb{R}^n$, assume $[\bar{x}, \bar{x}+d] \subset U$. Then,

$$f(\bar{x}+d) = f(\bar{x}) + \sum_{k=1}^{r} \frac{1}{k!} D^k f(\bar{x})[\underbrace{d,\ldots,d}_{k\text{-times}}] + \mathcal{R}(\bar{x}, d),$$

where $\mathcal{R}(\bar{x}, \cdot) : \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{\mathcal{R}(\bar{x}, h)}{\|h\|^r} = 0.$$

Theorem 31: Second order sufficient conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$, $S = \mathbb{R}^n$. Let $\bar{x} \in \mathbb{R}^n$. If $f'(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \in \mathbb{S}^n_{++}$, then $\bar{x}$ is a strict local minimizer for (P).

Proof:
Let $\bar{x} \in \mathbb{R}^n$ such that $f'(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \in \mathbb{S}^n_{++}$,

$$\underset{\substack{\uparrow \\ \lambda_n(\nabla^2 f(\bar{x}))}}{\delta} := \min\{d^T\nabla^2 f(\bar{x})d : \|d\|_2 = 1\} > 0$$

By Theorem 30, for all $d \in \mathbb{R}^n$, $\|d\|_2 = 1$, and $\alpha > 0$ and small enough, we have

$$f(\bar{x}+\alpha d) = f(\bar{x}) + \underbrace{\alpha\langle \nabla f(\bar{x}), d\rangle}_{=0} + \frac{\alpha^2}{2} d^T\nabla^2 f(\bar{x})d + o(\alpha^2) \geq f(\bar{x}) + \frac{\delta}{2}\alpha^2 + o(\alpha^2)$$

Choose a neighborhood U of $\bar{x}$ such that $\frac{\delta}{2}\alpha^2 > |o(\alpha^2)|$. Then for all $x \in U \setminus \{\bar{x}\}$, $f(x) > f(\bar{x})$. Therefore, $\bar{x}$ is a strict local minimizer for (P). □

How applicable is this last theorem?

Proposition 32

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^2$ and consider $\tilde{f}(x) := f(x) + c^T x$, where $c \in \mathbb{R}^n$ is given. Then for almost all $c \in \mathbb{R}^n$, $\tilde{f}(\bar{x}) = 0 \implies \nabla^2 f(\bar{x})$ is nonsingular.

Proof:
Apply Sard's Theorem (Theorem 25) to $g(x) := f'(x)$, with $r := 1$ and $p := n$. □

What if f has some nice structure, can we say more?

Definition 33: convex function

$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is **convex** if $\text{epi}(f) := \left\{ \begin{pmatrix} \mu \\ x \end{pmatrix} \in \mathbb{R} \oplus \mathbb{R}^n : f(x) \leq \mu \right\}$ is convex.

Here $\text{epi}(f)$ denotes the epi graph of f.

Theorem 34

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $S := \mathbb{R}^n$. Then every local minimizer of (P) is a global minimizer of (P). If in addition, f is differentiable on $\mathbb{R}^n$, then every stationary point of f is a global minimizer of (P).

2.1 Affine Subspace Constraints

One of the most popular form of continuous optimization problems is

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & Ax = b \end{array}$$

where $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$ are given.

At a first glance (and strictly speaking), (P) does not belong to the class of unconstrained continuous optimization problems. We may assume $\text{rank}(A) = p$; otherwise

- we easily prove $Ax = b$ has no solution, which implies (P) is infeasible, or
- we easily find all redundant equations and $\bar{x} \in \mathbb{R}^n$ such that $A\bar{x} = b$.

So, $\text{rank}(A) = p$. Find a basis B of A and form the partitions

$$[A_B | A_N] := A, \quad \left[\frac{x_B}{x_N} \right] := x.$$

Then,

$$Ax = b \iff x_B = A_B^{-1} b - A_B^{-1} A_N x_N.$$

Therefore, for every $x \in S$,

$$f(x) = f \begin{pmatrix} A_B^{-1} b - A_B^{-1} A_N x_N \\ x_N \end{pmatrix}.$$

We define $\tilde{f} : \mathbb{R}^{n-p} \to \mathbb{R}$ by

$$\tilde{f}(x_N) := f \begin{pmatrix} A_B^{-1} b - A_B^{-1} A_N x_N \\ x_N \end{pmatrix}.$$

Thus (P) is equivalent to

$$\inf_{x \in \mathbb{R}^{n-p}} \tilde{f}(x) \tag{$\tilde{\text{P}}$}$$

and we can start any algorithm from any starting point $x^{(0)} \in \mathbb{R}^{n-p}$.

Another equivalent approach:

Let $\bar{x} \in S$ (i.e., $A\bar{x} = b$). Then, $S = \{\bar{x} + u : u \in \text{Null}(A)\}$.

Let columns of $Z \in \mathbb{R}^{n \times (n-p)}$ form a basis for $\text{Null}(A)$. Then (P) is also equivalent to

$$\inf_{v \in \mathbb{R}^{n-p}} \hat{f}(v),$$

where $\hat{f} : \mathbb{R}^{n-p} \to \mathbb{R}$ is defined as $\hat{f}(v) := f(\bar{x} + Zv)$.

In applications, with either of these two approaches, we must be very careful about exploiting sparsity as well as making sure we can efficiently and accurately evaluate all ingredients of the algorithms we choose to use on such problems.

2.2 Applications

Some other ways of dealing with constrained optimization problems using unconstrained optimization algorithms: Form the Lagrangian for (P):

$$\mathcal{L}(x, v) := f(x) + v^T(b - Ax),$$

where $v \in \mathbb{R}^p$ represents the Lagrange multipliers (dual variables corresponding to the constraints).

Use a penalty function (penalizing any violation of the constraints):

$$\rho(x, \eta) := f(x) + \eta\|Ax - b\|_\beta^\gamma,$$

where $\beta, \gamma \in \mathbb{R}$ suitably defined, $\eta \in \mathbb{R}_{++}$ a penalty parameter.

In compressed sensing and related applications, one seeks a solution of

$$\inf\{f(x) + \eta\|x\|_0 : Ax = b\},$$

where $\|x\|_0 :=$ number of nonzero entries of x. As an approximation, many researches and practitioners work with

$$\inf\{f(x) + \eta_1\|x\|_1 + \eta_2\|Ax = b\|_2^\gamma\},$$

where $\eta_1, \eta_2, \gamma \in \mathbb{R}$, usually fixed.

We can generalize such approaches to matrix variables. Very many interesting applications in Machine Learning, AI and modern Data Science. In many of these applications, we want to find a low-rank solution.

> Example:
>
> $\min\{\text{rank}(X) : A(X) = b\}$, where $A : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ linear, $b \in \mathbb{R}^p$, both A, b are given.

2.3 Prototype low-rank approximation problem

Given $A \in \mathbb{R}_+^{m\times n}$ (both m and n are huge). We want to find matrices $U \in \mathbb{R}_+^{m\times k}$, $V \in \mathbb{R}_+^{n\times k}$ such that $A = UV^T$ and k is as small as possible.

If we do not require U and V to be nonnegative, the problem is solved by Singular Value Decomposition (SVD) and optimal k is the rank of A.

$A = Q_1 D Q_2^T$ where $Q_1 \in \mathbb{R}^{m\times m}, Q_2 \in \mathbb{R}^{n\times n}$ are orthogonal and $D \in \mathbb{R}^{m\times n}$ diagonal. Let's assume $m \le n$, then

$$D = \begin{bmatrix} \sigma_1(A) & & & 0 & 0 & \dots & 0 \\ & \sigma_2(A) & & & 0 & \dots & 0 \\ & & \ddots & & & & \vdots \\ 0 & & & \sigma_m(A) & 0 & \dots & 0 \end{bmatrix}$$

where $\sigma_1(A) \ge \sigma_2(A) \ge \cdots \ge \sigma_m(A) \ge 0$ are the singular values of A.

Theorem 35

Every $A \subseteq \mathbb{R}^{m\times n}$ has a SVD.

Requiring U, V to be nonnegative, makes the problem hard. Let p be an upper bound on k (taking p as $(mn + 1)$ suffices, but in practice, better guesses can help). Suppose our guess for the minimum nonnegative rank of A is p. Then let $U \in \mathbb{R}^{m\times p}$ and $V \in \mathbb{R}^{n\times p}$ denote the variable matrices and consider

$$\inf f(U,V) := \eta_1\|A - UV^T\| + \eta_2\|U_-\| + \eta_3\|V_-\|,$$

where $\eta_1, \eta_2, \eta_3 \in \mathbb{R}^+$ are parameters that we can fix, and U_- denotes the $\mathbb{R}^{m\times p}$ matrix with only negative entries of U.

2.4 Classical Algorithmic Approaches

I. Search direction + line-search strategies

Pick a search direction $d^{(k)}$, pick a step-size $\alpha_k > 0$. $x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}$. Repeat.

- $d^{(k)} := -\nabla f(x^{(k)})$ steepest-descent direction
- any $d^{(k)}$ with $\langle \nabla f(x^{(k)}), d^{(k)}\rangle < 0$
- $d^{(k)} := -\big[\underset{\substack{\uparrow \\ \text{Assuming } \nabla^2 f(x^{(k)}) \in \mathbb{S}^n_{++}}}{\nabla^2 f(x^{(k)})}\big]^{-1}\nabla f(x^{(k)})$ Newton direction

  For convex optimization problems and also near local minimizers of nonconvex problems, we want $\alpha_k \approx 1$ with this direction → superlinear or quadratic convergence

Exact Line-Search Find $\alpha > 0$ such that $\phi(\alpha) := f(x^{(k)} + \alpha d^{(k)})$ is minimized. Typically *not* practical.

Inexact Line-Search Armijo-Goldstein (1966-67) conditions (or Wolfe (1969) conditions):
Choose $\alpha > 0$ so that

$$f(x^{(k)} + \alpha d^{(k)}) \leq f(x^{(k)}) + c_1 \cdot \alpha\langle \nabla f(x^{(k)}), d^{(k)}\rangle$$

(sufficiently good rate for the decrease in the objective function) and **"curvature condition"**

$$\langle \nabla f(x^{(k)} + \alpha d^{(k)}), d^{(k)}\rangle \geq c_2\langle \nabla f(x^{(k)}), d^{(k)}\rangle$$

(step size should not be too small) where constants c_1, c_2 satisfy $0 < c_1 < c_2 < 1$.

Strong Wolfe Conditions

$$f(x^{(k)} + \alpha d^{(k)}) \leq f(x^{(k)}) + c_1 \cdot \alpha\langle \nabla f(x^{(k)}), d^{(k)}\rangle$$

and

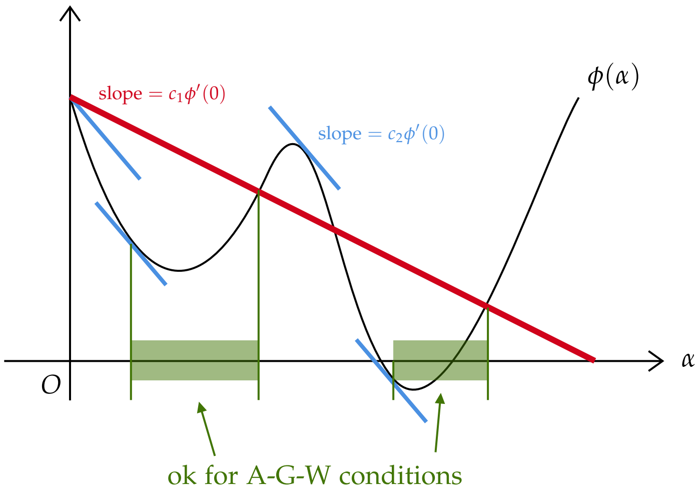
$$|\underset{\substack{\uparrow \\ \text{The second conditions disallows} \\ \text{this being too large and positive}}}{\langle \nabla f(x^{(k)} + \alpha d^{(k)}), d^{(k)}\rangle}| \leq c_2|\langle \nabla f(x^{(k)}), d^{(k)}\rangle|$$

Lemma 36

Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^1$, and $d \in \mathbb{R}^n$ be a descent direction at $\bar{x} \in \mathbb{R}^n$ for f. Suppose f is bounded from below on the ray$\{\bar{x} + \alpha d : \alpha \in \mathbb{R}_+\}$. Then $\forall 0 < c_1 < c_2 < 1$, there exists step lengths $\alpha > 0$ satisfying Armijo-Goldstein-Wolfe as well as Strong Wolfe conditions.

With $\phi(\alpha) := f(\bar{x} + \alpha d), 0 < c_1 < c_2 < 1$, choose $\alpha > 0$ such that

$$\text{Armijo-Goldstein-Wolfe}\begin{cases}\phi(\alpha) \leq \phi(0) + c_1\alpha\phi'(0) \\ \phi'(\alpha) \geq c_2\phi'(0)\end{cases} \qquad \text{Strong Wolfe}\begin{cases}\phi(\alpha) \leq \phi(0) + c_1\alpha\phi'(0) \\ |\phi'(\alpha)| \leq c_2|\phi'(0)|\end{cases}$$



Proof of Lemma 36:

Suppose the stated assumptions hold. We adopt the above mentioned notion with ϕ. Then $\phi(\alpha)$ is bounded from below on $\{\alpha \in \mathbb{R} : \alpha \geq 0\}$. Since $c_1 \in (0,1)$ and $\phi'(0) = \langle \nabla f(\bar{x}), d\rangle < 0$ (d is a descent direction for f), the ray $\{\phi(0) + (c_1 \cdot \phi'(0))\alpha : \alpha \geq 0\}$ is unbounded below and therefore, intersects the graph of ϕ at least once for $\alpha > 0$. Let $\bar{\alpha} > 0$ denote the smallest value of α for which the ray intersects the graph of ϕ. Then,

$$\phi(\bar{\alpha}) = \phi(0) + \bar{\alpha}c_1\phi'(0) \qquad (*)$$

Thus, the first condition of A-G-W holds on $(0, \bar{\alpha}]$.

By the Mean Value Theorem, there exists $\hat{\alpha} \in (0, \bar{\alpha})$ such that $\phi(\bar{\alpha}) - \phi(0) = \bar{\alpha}\phi'(\hat{\alpha})$. Therefore,

$$\phi(\bar{\alpha}) - \phi(0) \overset{(*)}{=} \bar{\alpha}\phi'(\hat{\alpha}) \overset{(*)}{=} \bar{\alpha}c_1\phi'(0) \underset{\substack{\uparrow \\ c_2 > c_1,\, \phi'(0) < 0}}{>} c_2\bar{\alpha}\phi'(0).$$

Thus, A-G-W conditions strictly hold at $\hat{\alpha}$. Since $\phi'(\hat{\alpha}) < 0$, Strong Wolfe conditions also hold at $\hat{\alpha}$ as well as in a sufficiently small neighborhood of $\hat{\alpha}$. □

In the textbook, Backtracking line search.

II. Trust-Region Strategies

Use the information gathered about f so far and construct an approximation ("model") m_k of the function f. Then solve

$$\begin{array}{ll} \min & m_k(d) \\ \text{s.t.} & d \in \text{Trust Region (around } x^{(k)}) \end{array}$$

$x^{(k)} \in \mathbb{R}^n$ is our current iterate. Let B_k denote $\nabla^2 f(x^{(k)})$ or an approximation of it. Choose $\delta_k > 0$, and solve

$$\begin{array}{ll} \min & m_k(d) := f(x^{(k)}) + \langle \nabla f(x^{(k)}), d\rangle + \frac{1}{2}d^T B_k d \\ \text{s.t.} & \|d\|_2 \leq \delta_k \end{array}$$

Let $\bar{d}$ denote an optimal solution of this trust-region subproblem. If $x^{(k)} + \bar{d}$ satisfies certain criteria, then set $x^{(k+1)} := x^{(k)} + \bar{d}$; otherwise either modify δ_k, or the step size, ...

Depending on how well we did with the latest δ_k choose a suitable value for δ_{k+1} and repeat. (size of the Trust-Region is being adjusted.)

2.5 Convergence Properties of Descent Algorithms

Let $f : \mathbb{R}^n \to \mathbb{R}$. For every $\beta \in \mathbb{R}$,

$$\{x \in \mathbb{R}^n : f(x) \leq \beta\}$$

is called a **sublevel set** of f (some literature use level set).

$$\{x \in \mathbb{R}^n : f(x) = \beta\}$$

is called a **level set** of f (some also call it a contour of f)

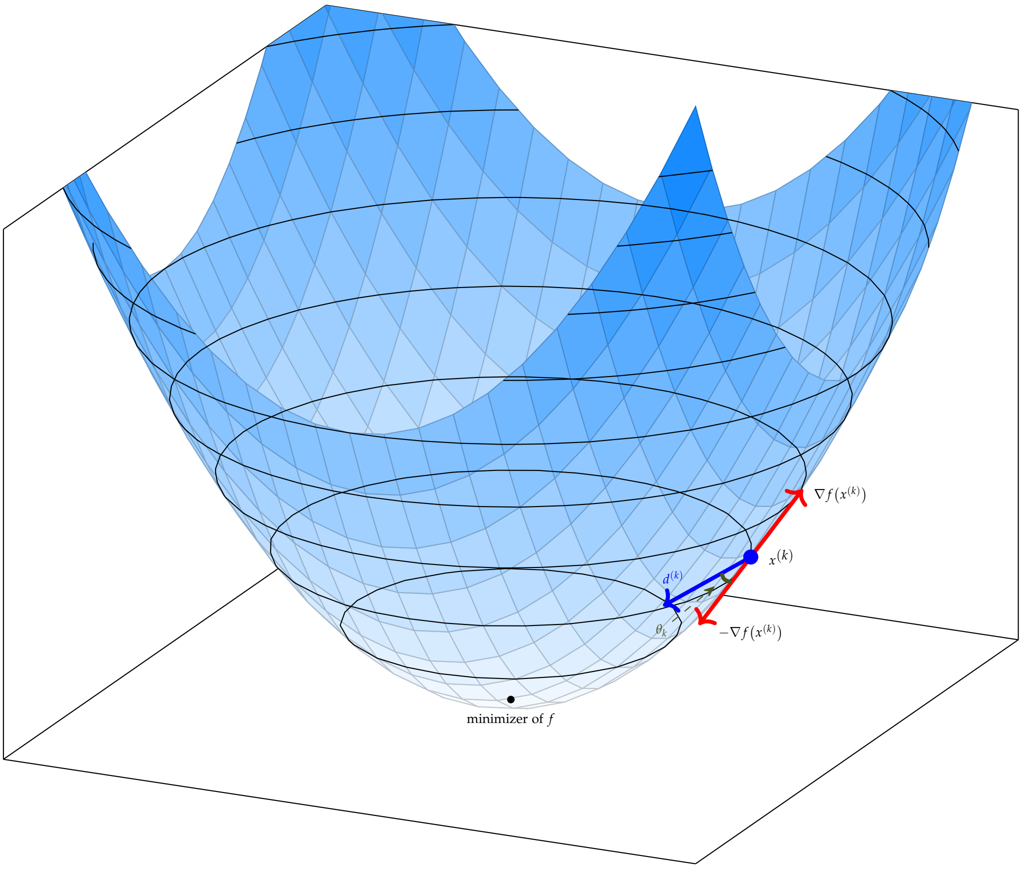
Consider a descent algorithm: Start with $x^{(0)} \in \mathbb{R}^n$, at each iteration k, choose $d^{(k)} \in \mathbb{R}^n$ such that $\langle \nabla f(x^{(k)}), d^{(k)} \rangle < 0$ and choose $\alpha_k > 0$, $x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}$.

Recall the geometric fact: $\forall u, v \in \mathbb{R}^n$,

$$\langle u, v \rangle = \|u\|_2 \cdot \|v\|_2 \cos\theta$$

where $\theta :=$ angle between u and v.

Define $\theta_k := \arccos\left(-\frac{\langle \nabla f(x^{(k)}, d^{(k)}) \rangle}{\|\nabla f(x^{(k)})\|_2 \cdot \|d^{(k)}\|_2} \right)$



Theorem 37: Zoutendijk (1970), Wolfe (1969)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a bounded from below, $x^{(0)} \in \mathbb{R}^n$, and f be $\mathcal{C}^1$ on

$$N := \text{nbhd}\{x \in \mathbb{R}^n : f(x) \leq f(x^{(0)})\}.$$

Assume ∇f is Lipschitz continuous on N with Lipschitz constant $L \in \mathbb{R}_{++}$. Then, every descent algorithm following Armijo-Goldstein-Wolfe conditions for stepsize selection satisfies:

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^{(k)})\|_2^2 < \infty.$$

Proof:
Suppose the assumptions in the statement hold. For every iteration k, due to the second A-G-W condition, we have $\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle \geq c_2 \langle \nabla f(x^{(k)}), d^{(k)} \rangle$, which implies

$$\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), d^{(k)} \rangle \geq (c_2 - 1)\langle \nabla f(x^{(k)}), d^{(k)} \rangle \tag{$\star$}$$

Due to the fact that we are working with a descent algorithm ($\langle \nabla f(x^{(k)}), d^{(k)} \rangle < 0$ for all k) and the first condition of A-G-W, $\{x^{(k)}\} \subset N$. Since ∇f is Lipschitz continuous on N with Lipschitz constant L,

$$\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle \underset{\substack{\uparrow \\ \text{Cauchy-Schwarz}}}{\leq} \|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\|_2 \|d^{(k)}\|_2 \underset{\substack{\uparrow \\ \nabla f \in \text{Lip}(L)}}{\leq} \alpha_k L \|d^{(k)}\|_2 \tag{$\spadesuit$}$$

By ($\star$) and ($\spadesuit$), we have

$$\alpha_k \geq \frac{(c_2 - 1)\langle \nabla f(x^{(k)}), d^{(k)} \rangle}{L\|d^{(k)}\|_2^2}$$

Substituting this lower bound on α_k into the first A-G-W condition, we obtain

$$f(x^{(k)} + \alpha_k d^{(k)}) \leq f(x^{(k)}) - \frac{c_1(c_2 - 1)\langle \nabla f(x^{(k)}), d^{(k)} \rangle}{L\|d^{(k)}\|_2^2}$$

$$\iff f(x^{(k+1)}) \leq f(x^{(k)}) - \left(\frac{c_1(1 - c_2)}{L}\right) \cos^2\theta_k \|\nabla f(x^{(k)})\|_2^2$$

Applying the above to pairs of consecutive iterates, we obtain:

$$f(x^{(k+1)}) \leq f(x^{(0)}) - \frac{c_1(1 - c_2)}{L} \sum_{\ell=0}^{k} \cos^2\theta_\ell \|\nabla f(x^{(\ell)})\|_2^2$$

Since f is bounded from below, $\left[f(x^{(0)}) - f(x^{(k)})\right]$ is bounded from above, and

$$\frac{c_1(1 - c_2)}{L} \sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^{(k)})\|_2^2 < +\infty$$

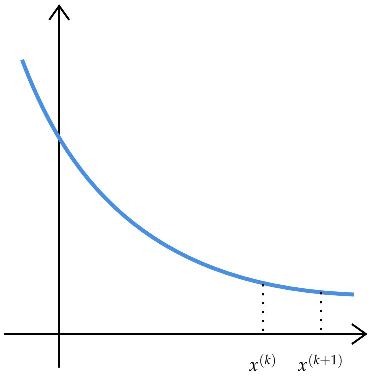
□

A consequence of Theorem 37:

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^{(k)})\|_2^2 < +\infty \implies \cos^2\theta_k \|\nabla f(x^{(k)})\|_2^2 \to 0 \text{ as } k \to +\infty$$

Therefore, if $\cos^2\theta_k \geq \delta > 0$ for all $k \in \mathbb{Z}_+$, then $\lim\limits_{k \to +\infty} \|\nabla f(x^{(k)})\|_2 \to 0$. In some places, including the textbook, this criterion is used to conclude that Steepest-Descent Algorithm is "globally convergent".

What if



i.e., what if $\|\nabla f(x^{(k)})\| < 10^{-8}$ but (even assuming convexity of f etc.) the unique minimizer $\bar{x}$ is far away from $x^{(k)}$?

2.6 A General Conversation about Convergence

Example:

$f : \mathbb{R} \to \mathbb{R}$, $f(x) := \frac{1}{4}x^4 - 5x$. Then f is convex, global minimizer is unique and attained at $\bar{x} := \sqrt[3]{5}$ which is irrational (even though the data $\subseteq \mathbb{Z}^{\cdots}$)

Thus we cannot expect finite algorithms in the worst-case. We will generate a sequence $x^{(1)}, x^{(2)}, \ldots$. We will hope fore conclusions like:

- For every $x^{(0)}$, $x^{(k)} \to \bar{x}$ a global minimizer.
- For every $x^{(0)}$, $x^{(k)} \to \bar{x}$ a local minimizer.
- For every $x^{(0)}$, all limit points of $\{x^{(k)}\}$ are global (local) minimizers, or $f(x^{(k)}) \to -\infty$.
- For every $x^{(0)}$, all limit points of $\{x^{(k)}\}$ satisfy second-order necessary conditions.
- For every $x^{(0)}$, all limit points of $\{x^{(k)}\}$ satisfy first-order necessary conditions.
- For every $x^{(0)}$, $\lim\limits_{k\to+\infty} \|\nabla f(x^{(k)})\| = 0$.

Locally, replace "every $x^{(0)} \in \mathbb{R}$" by "every $x^{(0)} \in B(\bar{x}, \eta) := \{x \in \mathbb{R}^n : \|x - \bar{x}\|_2 < \eta\}$," and hope that the 2nd-order sufficient conditions hold.

How fast? $\epsilon_k := \|x^{(k)} - \bar{x}\|$.

Example:

$\epsilon_k := (0.1)^k \to 10^{-1}, 10^{-2}, \ldots$	linear convergence
$\epsilon_k := (0.9)^k \to 0.9, 0.081, 0.729, \ldots$	linear convergence
$\epsilon_k := (0.1)^{2^k} \to (10)^{-2}, (10)^{-4}, (10)^{-8}, \ldots$	quadratic convergence
$\epsilon_k := (0.9)^{2^k} \to 0.81, 0.65, 0.43, 0.185, 0.034, 10^{-3}, \ldots$	quadratic convergence
$\epsilon_k := (0.1)^{3^k} \to \ldots$	cubic convergence

Definition 38

If $\epsilon_k \searrow 0$ and $\epsilon_{k+1} \leq \beta(\epsilon_k)^p$ for some constants $p \geq 1$ and β ($\beta < 1$ if $p = 1$) and for all sufficiently large k, then we say $\underline{\epsilon_k \to 0 \text{ with Q-order (at least) } p}$. If $\epsilon_k \searrow 0$ and $\frac{\epsilon_{k+1}}{\epsilon_k} \to 0$ (as $k \to +\infty$) then the convergence is **Q-superlinear**.

Q-linear := Q-order 1, Q-quadratic := Q-order 2

Example:

$\epsilon :_k = \left(\frac{1}{l}\right)^k$, $k \in \mathbb{Z}_{++}$. Then $\epsilon_k \searrow 0$, Q-superlinearly, but it does not have Q-order $p > 1$.

Given a sequence $\{\epsilon_k\} \subset \mathbb{R}_+$, let $\eta_i := \sup\{\epsilon_k : k \geq i\}$. Then $\limsup\limits_{k\to+\infty}\{\epsilon_k\} := \lim i \to \infty\{\eta_i\}$.

Definition 39

If $\epsilon_k \searrow 0$ and $\limsup\limits_{k\to+\infty} \left\{\epsilon_k^{1/q^k}\right\} < 1$ for all $0 < q < p, p > 1$, then $\underline{\epsilon_k \to 0 \text{ with R-order (at least) } p}$.

This is the same as $\limsup\limits_{k\to+\infty} \left\{\frac{1}{q^k}\ln(\epsilon_k)\right\} < 0$

Proposition 40

(i) If $x^{(k)} \to \bar{x}$ with Q-order p (R-order p), so does $\{x^{(k+\ell)}\}$ for every fixed $\ell \in \mathbb{Z}_+$.

(ii) If $\epsilon_k \searrow 0$ with Q-order p and $0 < \eta_k \leq \epsilon_k$ for $\forall k \in \mathbb{Z}_{++}$, then $\eta_k \searrow 0$ with R-order p.

Fast Local Convergence of Newton's Method

This goes back at least to Kantorovich (Nobel Prize in Economics for his work on "the theory of optimal allocation of resources", 1975). In addition to his foundational work on the convergence theory of Newton's Method, Kantorovich also made significant contributions to functional analysis and operator theory.

Lemma 41

Let $A, B \in \mathbb{R}^{n\times n}$, A nonsingular, $\|A^{-1}\|_2 \leq \gamma$ and $\|A - B\|_2 \leq \frac{1}{3\gamma}$. Then B is nonsingular and $\|B^{-1}\|_2 \leq \frac{3\gamma}{2}$.

$\|A^{-1}\|_2 \leq \gamma \iff \text{dist}(A, \text{singular matrices}) \geq \frac{1}{\gamma}$

Proof:

Suppose $A, B \in \mathbb{R}^n$ satisfy the assumptions. Note that

$$B = A - (A - B) = A[I - A^{-1}(A - B)],$$

and

$$\|A^{-1}(A - B)\|_2 \leq \|A^{-1}\|_2\|A - B\|_2 \leq \gamma \cdot \frac{1}{3\gamma} = \frac{1}{3}.$$

If $C \in \mathbb{R}^{n\times n}$ nonsingular such that $\|C\|_2 \leq \frac{1}{3}$, then $(I - C)$ is invertible and

$$(I - C)^{-1} = I + C + C^2 + \cdots = \sum_{k=0}^{\infty} C^k.$$

This implies

$$\|(I-C)^{-1}\|_2 \le \sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^k = \frac{1}{2/3} = \frac{3}{2}.$$

Thus, with $C := A^{-1}(A-B)$, (then $B = A(I-C)$) B is invertible. $B^{-1} = (I-C)^{-1}A^{-1}$ and

$$\|B^{-1}\|_2 \le \|(I-C)^{-1}\|_2 \cdot \|A^{-1}\|_2 \le \frac{3}{2}\gamma.$$

□

Lemma 42

Let $g : \mathbb{R}^n \to \mathbb{R}^m$, $g \in \mathcal{C}^1$ and $\nabla g \in \text{Lip}(L)$ on some open and convex set $D \subseteq \mathbb{R}^n$. Then

$$\|g(y) - g(x) - \nabla g(x)(y-x)\|_2 \le \frac{L}{2}\|y-x\|_2^2,$$

$\forall x, y \in D$.

Proof:
We already proved this as a part of our discussion following Theorem 19. □

Newton's Method: $x^{(0)} \in \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2$.

$$\forall k \in \mathbb{Z}_+ : \begin{cases} d^{(k)} := -\left[\nabla^2 f(x^{(k)})\right]^{-1}\nabla f(x^{(k)}) \\ x^{(k+1)} := x^{(k)} + d^{(k)} \end{cases}$$

Theorem 43

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$, $x^{(0)} \in \mathbb{R}^n, L \ge 1$. Assume $\nabla f(\bar{x}) = 0$, $\nabla^2 f(\bar{x})$ is nonsingular, $\nabla^2 f \in \text{Lip}(L)$ in an open neighborhood of $\bar{x}$. Then there exists an open neighborhood N_1 of $\bar{x}$ such that for all $x^{(0)} \in N_1$, Newton's Method converges to $\bar{x}$ linearly and the method is **locally Q-quadratically convergence** (there exists an open neighborhood $N_2 \subseteq N_1$ of $\bar{x}$ such that

$$\forall x^{(0)} \in N_2, \|x^{(k+1)} - \bar{x}\|_2 \le \text{constant} \cdot \|x^{(k)} - \bar{x}\|_2^2,$$

$\forall k \in \mathbb{Z}_+$). Moreover, $\|\nabla f(x^{(k)})\|$ also converges to zero in N_1, locally Q-quadratically

$$(\forall x^{(0)} \in N_2, \|\nabla f(x^{(k+1)})\|_2 \le \text{constant} \cdot \|\nabla f(x^{(k)})\|_2^2, \quad \forall k \in \mathbb{Z}_+).$$

Proof:
Suppose the assumptions hold.

Let $\gamma := \left\|\left[\nabla^2 f(\bar{x})\right]^{-1}\right\|_2$. Choose $\eta > 0$ such that with $\mathcal{B} := B(\bar{x}, \eta) := \{x \in \mathbb{R}^n : \|x - \bar{x}\|_2 < \eta\}$, $\nabla^2 f \in \text{Lip}(L)$ on $\mathcal{B}$ and $\eta \le \frac{1}{3\gamma L}$. Then for all $x \in \mathcal{B}$,

$$\left\|\nabla^2 f(x) - \nabla^2 f(\bar{x})\right\|_2 \le L\|x - \bar{x}\|_2 < L\eta \le \frac{1}{3\gamma} \qquad (\diamondsuit)$$

Therefore, by lemma 41 (with $A := \nabla^2 f(\bar{x})$, $B := \nabla^2 f(x), x \in \mathcal{B}$), $\nabla^2 f(x)$ is nonsingular $\forall x \in \mathcal{B}$; thus, Newton's Method is well-defined for $\{x^{(k)}\} \subset \mathcal{B}$.

We will prove the theorem by induction on the iteration number k.

Let $x^{(0)} \in \mathcal{B}$ (in general, $x^{(k)} \in \mathcal{B}$), then

$$\begin{aligned}
\left\|x^{(k+1)} - \bar{x}\right\|_2 &= \left\|x^{(k)} - \left[\nabla^2 f(x^{(k)})\right]^{-1} \nabla f(x^{(k)}) - \bar{x}\right\|_2 \\
&= \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1} - \left[\underset{\overset{\uparrow}{\nabla f(\bar{x})}}{0} - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)})(\bar{x} - x^{(k)})\right]\right\|_2 \\
&\leq \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1}\right\|_2 \left\|\nabla f(\bar{x}) - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)})(\bar{x} - x^{(k)})\right\|_2 \\
&\leq \frac{3\gamma}{2} \cdot \frac{L}{2} \left\|x^{(k)} - \bar{x}\right\|_2^2 \qquad \text{by } (\diamondsuit) \text{ and Lemma 41, Lemma 42} \\
&= \frac{3\gamma L}{4} \left\|x^{(k)} - \bar{x}\right\|_2^2 \qquad \text{by induction on } k \text{, we get quadratic convergence}
\end{aligned}$$

Also, if $x^{(k)} \in \mathcal{B}$, then

$$\left\|x^{(k+1)} - \bar{x}\right\|_2 \leq \frac{3\gamma L}{4} \cdot \frac{1}{3\gamma L} \left\|x^{(k)} - \bar{x}\right\|_2 = \frac{1}{4} \left\|x^{(k)} - \bar{x}\right\|_2$$

which is linear convergence.

Next, let $d := x^{(k+1)} - x^{(k)}$. Then

$$\begin{aligned}
\|\nabla f(x^{(k+1)})\|_2 &= \left\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)}) d\right\|_2 \\
&\leq \frac{L}{2} \|d\|_2^2 && \text{Lemma 42} \\
&= \frac{L}{2} \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1} \nabla f(x^{(k)})\right\|_2^2 && (\clubsuit) \\
&\leq \frac{L}{2} \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1}\right\|_2^2 \left\|\nabla f(x^{(k)})\right\|_2^2 \\
&\leq \frac{9\gamma^2 L}{8} \left\|\nabla f(x^{(k)})\right\|_2^2 && \text{by } x^{(k)} \in \mathcal{B}, (\clubsuit) \text{ and Lemma 41}
\end{aligned}$$

Now, we have all the ingredients for an induction proof. We proved

> "$\forall x^{(0)} \in \mathcal{B}, \|x^{(1)} - \bar{x}\|_2 \leq \frac{1}{4}\|x^{(0)} - \bar{x}\|_2$; so, $x^{(1)} \in \mathcal{B}$, and $\|x^{(1)} - \bar{x}\|_2 \leq \frac{3\gamma L}{4}\|x^{(0)} - \bar{x}\|_2^2$, and $\|\nabla f(x^{(1)})\|_2 \leq \frac{9\gamma^2 L}{8}\|\nabla f(x^{(0)})\|_2^2$."

By induction on k, we establish the desired inequalities on $x^{(k)}$. For the gradient, from $(\clubsuit)$,

$$\begin{aligned}
\|\nabla f(x^{(k+1)})\|_2 &\leq \frac{L}{2} \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1} \nabla f(x^{(k)})\right\|_2^2 \\
&= \frac{L}{2} \left\|x^{(k+1)} - x^{(k)}\right\|_2 \cdot \left\|\left[\nabla^2 f(x^{(k)})\right]^{-1} \nabla f(x^{(k)})\right\|_2 \\
&\leq \frac{L}{2} \cdot \frac{2}{3\gamma L} \cdot \frac{3\gamma}{2} \left\|\nabla f(x^{(k)})\right\|_2 \qquad x^{(k)}, x^{(k+1)} \in \mathcal{B}, \text{ Lemma 41} \\
&= \frac{1}{2} \left\|\nabla f(x^{(k)})\right\|_2
\end{aligned}$$

Therefore, for every $x^{(0)} \in \mathcal{B}$,

$$\begin{aligned}
\left\|x^{(k)} - \bar{x}\right\| &\to 0, && \text{Q-linearly, and locally Q-quadratically} \\
\left\|\nabla f(x^{(k)})\right\| &\to 0, && \text{Q-linearly, and locally Q-quadratically}
\end{aligned}$$

□

This proof also applies to the problem of solving systems of nonlinear equations.

$g : \mathbb{R}^n \to \mathbb{R}^n, g \in \mathcal{C}^1$ on open and convex set $D \subseteq \mathbb{R}^n$. $\exists \bar{x} \in D$ such that $g(\bar{x}) = 0, \nabla g(\bar{x})$ is nonsingular, $\nabla g \in \text{Lip}(L)$ on D. Let $x^{(0)} \in N \subseteq D$,

$$x^{(k+1)} := x^{(k)} - [\nabla g(x^{(k)})]^{-1} g(x^{(k)}) \qquad \forall k \in \mathbb{Z}_+.$$

Potential Problems with Newton's Method

(I) Fails if $\nabla^2 f(x^{(k)})$ is singular (or very ill-conditioned)

(II) $x^{(k+1)}$ is not the minimizer of the local quadratic model $\tilde{f}$ for f if $\nabla^2 f(x^{(k)})$ is not positive definite.

(III) Not globally convergent in general.

(IV) May not even provide descent in general.

Possible Remedies:

- To address (I) & (II) modify $\nabla^2 f(x^{(k)})$, if necessary, to a "nearby" symmetric positive definite matrix B_k. (do this in an efficient & numerically stable way.)
- Together with the above remedy, use A-G-W or Strong Wolfe based line searches to address (III) & (IV).

Still there are some disadvantages:

(i) We must evaluate Hessians at every iteration,

(ii) We must solve $n \times n$ linear systems of equations in every iteration.

For some problems evaluating the Hessian is very little extra work compared to $f, \nabla f$. Also, in some cases Automatic Differentiation via a small number of $\nabla f(\cdot)$ evaluations suffice. (Chapter 8 of the textbook)

2.7 Quasi-Newton Methods

Consider $B_k \in \mathbb{S}^n_{++}$, then $-B_k^{-1}\nabla f(x^{(k)})$ is a descent direction for f at $x^{(k)}$. Consider a quadratic model of f (near $x^{(k)}$):

$$\tilde{f}(d) := f(x^{(k)}) + \langle \nabla f(x^{(k)}), d\rangle + \frac{1}{2} d^T B_k d.$$

Since $B_k \in \mathbb{S}^n_{++}$, $\tilde{f}$ has a unique global minimizer at $\bar{d} = -B_k^{-1}\nabla f(x^{(k)})$. Now we can do a line search and find $x^{(k+1)}$, then we have $f(x^{(k+1)})$ and $\nabla f(x^{(k+1)})$. How do we find B_{k+1}?

Wish List for B_{k+1}

$B_{k+1} \in \mathbb{S}^n_{++}$. B_{k+1} should incorporate newly discovered information about $\nabla^2 f$.

$$\begin{aligned} s^{(k)} &:= x^{(k+1)} - x^{(k)} && \text{(primal step at iteration } k\text{)} \\ y^{(k)} &:= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) && \text{(dual step at iteration } k\text{)} \end{aligned}$$

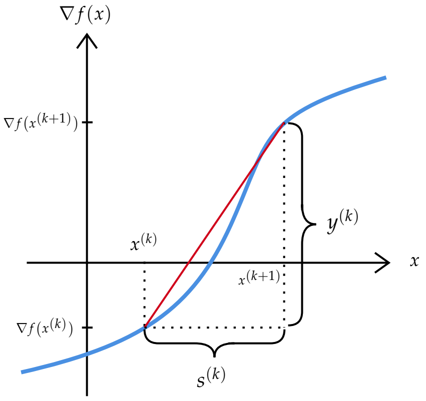
Magical solution: BFGS

$$B_{k+1} := \underbrace{B - \frac{1}{s^T B s} B s s^T B + \frac{1}{y^T s} y y^T}_{\text{we dropped the iteration number } k}$$

> **Note:**
> By Theorem 19, we have
>
> $$y^{(k)} = \left[\int_0^1 \nabla^2 f\left(x^{(k)} - \alpha s^{(k)}\right) \partial\alpha\right] s^{(k)}$$
>
> i.e., $y^{(k)}$ tells us the behavior of the "average" Hessian (along the line segment $[x^{(k)}, x^{(k+1)}]$) on the subspace span$\{s^{(k)}\}$. So, we want $B_{k+1} \in \mathbb{R}^{n\times n}$ such that $y^{(k)} = B_{k+1}s^{(k)}$ which is **secant equation**. By enforcing this equation on B_{k+1}, we can incorporate new "secant" information about $\nabla^2 f$.

Example:

Consider $n := 1$.



If $B_{k+1} \succ 0$ satisfies the secant equation, then $\langle y^{(k)}, s^{(k)} \rangle = \langle B_{k+1} s^{(k)}, s^{(k)} \rangle > 0$ since $s^{(k)} \neq 0, B_{k+1} \succ 0$. Notice that $\langle y^{(k)}, s^{(k)} \rangle$ is positively proportional to:

$$\langle y^{(k)}, d^{(k)} \rangle = \langle \nabla f(x^{(k+1)}), d^{(k)} \rangle - \langle \nabla f(x^{(k)}), d^{(k)} \rangle = \phi'(\alpha_k) - \phi'(0) \underset{\substack{\uparrow \\ \text{If we use A-G-W} \\ \text{or Strong Wolfe} \\ \text{based line-search}}}{>} 0$$

The condition $\langle y^{(k)}, s^{(k)} \rangle > 0$ is called the **curvature condition**.

How do we ensure B_{k+1} is close to B_k?

Solve the optimization problem

$$\begin{array}{rl} \min & \|B - H\|_F \\ \text{s.t.} & Bs = y, \quad B \in \mathbb{R}^{n \times n} \end{array} \tag{P$_1$}$$

for a fixed $H \in \mathbb{R}^{n \times n}$, e.g., $H := B_k$, and fixed $y, s \in \mathbb{R}^n$.

Here,

$$\underset{\substack{\uparrow \\ \text{Frobenius norm}}}{\|A\|_F} := \left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2} = \left[\mathrm{Tr}(A^T A) \right]^{1/2} = \left[\mathrm{vec}(A)^T \mathrm{vec}(A) \right]^{1/2}$$

(P$_1$) always has a unique solution $\bar{B}$. $Z = \bar{B} - H$. Note $\bar{B}s = y \iff \bar{B}s - Hs = y - Hs =: r$.

With this change of variable and definitions, (P$_1$) is equivalent to

$$\begin{array}{rl} \min & \|Z\|_F \\ \text{s.t.} & Zs = r \end{array} \tag{P$_2$}$$

Suppose $s \neq 0$ (i.e., we moved!). Let $Q \in \mathbb{R}^{n \times n}$ be orthogonal such that $Qs = \beta e_1$, $\beta \neq 0$. $\tilde{Z} := ZQ^T$. Then (P$_2$) is equivalent to

$$\begin{array}{rl} \min & \|\tilde{Z}\|_F \\ \text{s.t.} & \tilde{Z} e_1 = \frac{1}{\beta} r \end{array} \tag{P$_3$}$$

which implies

$$\tilde{Z} = \begin{bmatrix} \frac{1}{\beta} r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

Using our definitions, we compute

$$Z = \tilde{Z} Q = \frac{1}{\beta} r e_1^T Q = \frac{1}{\beta^2} r s^T Q^T Q = \frac{1}{\beta^2} r s^T$$

$$Zs = r \implies \frac{1}{\beta^2} r (s^T s) = r \underset{\substack{\uparrow \\ \text{unless } r = 0 \text{ in} \\ \text{which } Z = 0}}{\implies} s^T = \beta^2$$

Therefore, the unique optimal solution of (P$_1$) is

$$Z = \frac{r s^T}{s^T s} = \frac{1}{s^T s} (y - Hs) s^T$$

Theorem 44: Broyden (1965)

Let $s, y \in \mathbb{R}^n, s \neq 0, H \in \mathbb{R}^{n\times n}$ be given. Then the unique optimal solution of (P_1) is

$$B := H + \frac{1}{s^T s}(y - Hs)s^T.$$

Good Broyden

Setting $v := H^T y$ and $B := H + \dfrac{1}{v^T s}(y - Hs)\, v^T$ leads to "Broyden's Second Method"

Bad Broyden

Let us modify problem (P_1) by requiring $B \in \mathbb{S}^n$ (and $H \in \mathbb{S}^n$ in the data). Consider

$$\begin{array}{rl} \min & \|B - H\|_F \\ \text{s.t.} & Bs = y \\ & B = B^T \\ & B \in \mathbb{R}^{n\times n} \end{array}$$

Theorem 45: Powell (1970)

The unique optimal solution ($s \neq 0$) of the above problem is given by

$$B := H + \frac{1}{s^T s}\Big[(y - Hs)s^T + s(y - Hs)^T - (y - Hs)ss s^T\Big].$$

In the formula above, B may *not* be *positive definite* even if the curvature condition is satisfied ($y^T s > 0$) and H is symmetric positive definite.

We want B to be symmetric, positive definite, provided $H \succ 0$ and $y^T > 0$. We consider solving

$$\begin{array}{rl} \min & \left\|W^{1/2}(B - H)W^{1/2}\right\|_F \\ \text{s.t.} & Bs = y \\ & B \in \mathbb{S}^n \end{array} \tag{P_W}$$

where

$$W := \left[\int_0^1 \nabla^2 f\big(x^{(k)} + t\alpha_k d^{(k)}\big)\partial t\right]^{-1},$$

but any $W \in \mathbb{S}^n_{++}$ satisfying $Wy^{(k)} = s^{(k)}$ works.

Theorem 45

For every $H \in \mathbb{S}^n_{++}, y, s \in \mathbb{R}^n$ such that $y^T s > 0$ and $W \in \mathbb{S}^n_{++}$ such that $Wy^{(k)} = s^{(k)}$, the unique solution of (P_W) is

$$B := \left(I - \frac{ys^T}{y^T s}\right) H \left(I - \frac{sy^T}{y^T s}\right) + \frac{yy^T}{y^T s}.$$

This is called the Davidon–Fletcher–Powell update.

Moreover, $B \in \mathbb{S}^n_{++}$.

Note that

$$B^{-1} = H^{-1} - \frac{H^{-1}yy^T H^{-1}}{y^T H^{-1} y} + \frac{ss^T}{y^T s} \longleftarrow \text{DFP for the Hessian inverse}$$

Sherman-Morrison-Woodbury
Formula applied to the above

Next, consider

$$\begin{array}{rl} \min & \|W^{-1/2}(B - H^{-1})W^{-1/2}\|_F \\ \text{s.t.} & By = s \\ & B \in \mathbb{S}^n \end{array} \tag{$\mathrm{P}_W^{\mathrm{BFGS}}$}$$

Theorem 46

For every $H \in \mathbb{S}^n_{++}$, $y, s \in \mathbb{R}^n$ such that $y^Ts > 0$ and $W \in \mathbb{S}^n_{++}$ such that $Wy^{(k)} = s^{(k)}$, the unique solution of $(\mathrm{P}^{\mathrm{BFGS}}_W)$ is

$$B := \left(I - \frac{sy^T}{y^Ts}\right) H \underset{\uparrow}{\left(I - \frac{ys^T}{y^Ts}\right)} + \frac{ss^T}{y^Ts}.$$

Moreover, $B \in \mathbb{S}^n_{++}$.

Broyden–Fletcher–Goldfarb–Shanno formula for approximation Hessian inverse

To approximate the Hessian we invert the above formula and obtain (in terms of H as an approximation to the Hessian):

$$H - \frac{Hss^TH}{s^THs} + \frac{yy^T}{y^Ts}$$

In practice, this is most successful with updating its Choleski decomposition.

Summary of DFP, BFGS

$$\mathcal{P} := \{B \in \mathbb{S}^n : Bs = y, B \succ 0\}$$
$$\mathcal{D} := \{B \in S^n : By = s, B \succ 0\}$$

With

$$W := \left[\int_0^1 \nabla^2 f\big(x^{(k)} + t\alpha_k d^{(k)}\big)\partial t\right]^{-1},$$

DFP: Solve

$$\begin{array}{rl} \min & \left\|W^{1/2}(B-H)W^{1/2}\right\|_F \\ \text{s.t.} & Bs = y \\ & B \in \mathbb{S}^n \end{array}$$

BFGS: Solve

$$\begin{array}{rl} \min & \|W^{-1/2}(B-H^{-1})W^{-1/2}\|_F \\ \text{s.t.} & By = s \\ & B \in \mathbb{S}^n \end{array}$$

then the inverse of the solution is the BFGS estimate of the Hessian $\nabla^2 f$.

$\mathcal{P}$ and $\mathcal{D}$ are convex sets. $U \in \mathcal{P} \iff U^{-1} \in \mathcal{D}$.

Therefore, $\forall U \in \mathcal{P}, \forall V \in \mathcal{D}$, $\forall \lambda \in [0,1]$, $[\lambda U + (1-\lambda)V^{-1}] \in \mathcal{P}$ and $[\lambda U^{-1} + (1-\lambda)V] \in \mathcal{D}$.

Broyden's convex class: $\left\{\lambda B^{\mathrm{DFP}} + (1-\lambda)B^{\mathrm{BFGS}} : \lambda \in [0,1]\right\}$

2.7.1 Convergence Results

(I) Global

(a) Powell (1972): If f is strictly convex (f is convex and $f(\lambda u + (1-\lambda)v) < \lambda f(u) + (1-\lambda)v$, for all $\lambda \in (0,1)$ and $u \neq v$), $\{x \in \mathbb{R}^n : f(x) \leq f(x^{(0)})\}$ is compact, $f \in \mathcal{C}^2$, and exact line search is used then quasi-Newton method based on DFP converges.

(b) Dixon (1972): If exact line search is used then DPF, BFGS (and many others) all give identical sequence of iterates $\{x^{(k)}\}$ for the same $(x^{(0)}, B_0)$.

(c) Powell (1976): Same assumptions on f as in (a), but line-search satisfying A-G-W conditions imply global convergence of BFGS.

(d) Byrd, Nocedal and Yuan (1987): Result of (c) holds for all Broyden's convex class, except for DFP $\big(\text{i.e., } \lambda \in [0,1)\big)$.

It seems that DFP is worse than BFGS in practice (with inexact line search).

(II) Local Convergence

Assume $f \in \mathcal{C}^2, x^{(k)} \to \bar{x}, \nabla f(x^{(k)}) \to 0, \nabla^2 f(\bar{x}) \in \mathbb{S}^n_{++}$.

(a) Powell (1971): with exact line search, DFP, BFGS both attain Q-superlinear convergence.

(b) Broyden, Dennis, Moré (1973): If we use $\alpha_k := 1$ for all $k \in \mathbb{Z}_+$ and for suitably small $\epsilon > 0, \delta > 0$, we have $\|x^{(0)} - \bar{x}\| \leq \epsilon$ and $\|B_0 - \nabla^2 f(\bar{x})\| \leq \delta$, then $x^{(k)} \to \bar{x}$ Q-superlinearly.

(c) Powell (1976): Assumptions as in I (c), BFGS with $\alpha_k := 1$ chosen whenever possible (i.e., whenever $\alpha_k := 1$ satisfies A-G-W conditions), attains Q-superlinear convergence (note: no assumptions on B_0)

(d) Byrd Nocedal and Yuan (1987): II (c) applies to every update in Broyden's COnvex Class, except DFP

2.7.2 Implementation of Quasi-Newton Methods

The most popular and the most successful (generally speaking) quasi-Newton algorithms belong to the class of **Limited Memory BFGS (L-BFGS)**: only keep the most recent r updates

$$(s^{(k-r)}, y^{(k-r)}), (s^{(k-r+1)}, y^{(k-r+1)}), \ldots, (s^{(k)}, y^{(k)}).$$

Typically $r \in \{10, 11, \ldots, 20\}$.

Implementing L-BFGS is relatively straightforward by utilizing the formula from Theorem 46.

Suppose for the current estimate of the Hessian, H, we have a Choleski decomposition: $H = LL^T$, where L is lower triangular. We would like a Choleski decomposition of B^{BFGS}.

Lemma 47

Let $H \in \mathbb{S}^n_{++}, y, s \in \mathbb{R}^n$ such that $y^T s > 0$. Also let $L \in \mathbb{R}^{n \times n}$, lower triangular satisfy $LL^T = H$. Then,

$$B^{\text{BFGS}} = \left(L + \frac{(y - \beta H s)s^T L}{\beta s^T H s}\right)\left(L^T + \frac{L^T s (y - \beta H s)^T}{\beta s^T H s}\right),$$

where $\beta := \sqrt{\dfrac{y^T s}{s^T H s}}$.

Proof:

Computation. □

So B is written as $(L + uv^T)(L^T + vu^T)$ which is not a Choleski decomposition. However, one can recover a Choleski factorization $\bar{L}\bar{L}^T$ of B as follows:

Remark:

For every orthogonal matrix $Q \in \mathbb{R}^{n \times n}$, $B = (L + uv^T)Q^T Q(L^T + vu^T)$.

We will use a sequence of orthogonal matrices on $L^T + vu^T$. First, forces on vu^T.

Using Given's Rotations

James Wallace Givens Jr. (1958)

For $\forall a, b \in \mathbb{R}$, $\exists \theta \in [0, 2\pi)$ such that

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}$$

$$\underset{\substack{\uparrow \\ \text{orthogonal}}}{\begin{bmatrix} 1 & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & \mathbf{0} & & & & c & s \\ & & & & & -s & c \end{bmatrix}} v = \begin{bmatrix} * \\ * \\ \vdots \\ \vdots \\ \vdots \\ * \\ 0 \end{bmatrix}$$

Then,

$$\begin{bmatrix} 1 & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & & c & s \\ & \mathbf{0} & & & -s & c \\ & & & & & & 1 \end{bmatrix} \begin{bmatrix} * \\ * \\ \vdots \\ \vdots \\ \vdots \\ * \\ 0 \end{bmatrix} = \begin{bmatrix} * \\ * \\ \vdots \\ \vdots \\ * \\ 0 \\ 0 \end{bmatrix}$$

$$\vdots$$

We find $Q_1 \in \mathbb{R}^{n \times n}$ orthogonal such that

$$Q_1(L^T + vu^T) = \begin{bmatrix} & & & \\ * & & & \\ 0 & * & & \\ \vdots & 0 & \ddots & \\ 0 & \cdots & 0 & * \end{bmatrix} + \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} [u^T] = \begin{bmatrix} * & * & \cdots & \cdots & * \\ * & & & & \\ 0 & * & & & \\ \vdots & 0 & \ddots & & \\ 0 & \cdots & 0 & * & \end{bmatrix}$$

Upper Hessenberg

Next we apply $(n-1)$ special orthogonal matrices (Givens' Rotations), to zero-out the nonzeros below the diagonal.

$$\begin{bmatrix} c & s & & & & \\ -s & c & & & \mathbf{0} & \\ & & 1 & & & \\ & & & \ddots & & \\ & \mathbf{0} & & & \ddots & \\ & & & & & 1 \\ & & & & & & 1 \end{bmatrix}, \text{ then } \begin{bmatrix} 1 & & & & & \\ & c & s & & \mathbf{0} & \\ & -s & c & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & \mathbf{0} & & & & 1 \\ & & & & & & 1 \end{bmatrix} \cdots$$

$\rightarrow$ Orthogonal matrix $Q_2 \in \mathbb{R}^{n \times n}$ such that $(Q_2 Q_1)(L^T + vu^T) =: \underset{\substack{\uparrow \\ \text{Choleski factor of } B}}{\bar{L}^T}$

Total work: $O(n^2)$ arithmetic operations.

2.8 Conjugate Gradient Methods

Let $C \subseteq \mathbb{R}^n$ be a convex set. Note that every $\mathcal{C}^2$ function $f : C \to \mathbb{R}$ with $\nabla^2 f(x) \succ 0$ for all $x \in C$, is strictly convex on C.

On strictly convex quadratic functions $f : \mathbb{R}^n \to \mathbb{R}$ ($f(x) := \gamma + c^T x + \frac{1}{2} x^T Hx$, with $\gamma \in \mathbb{R}, c \in \mathbb{R}^n, H \in \mathbb{S}^n_{++}$ given), BFGS and many other Quasi-Newton Methods require at most n iterations (with exact line search).

Special case: $f(x) := \gamma + c^T x + \frac{1}{2} x^T Dx$, D is diagonal and positive definite. In this case, the problem $\inf_{x \in \mathbb{R}^n} f(x)$ is separable. Coordinate Descent solves this problem in n iterations.

Now, consider an arbitrary $H \in \mathbb{S}^n_{++}$, with $f(x) := \gamma + c^T x + \frac{1}{2} x^T Hx$. Let $Q \in \mathbb{R}^{n \times n}$ be orthogonal such that $H = QDQ^T$, where $D \in \mathbb{R}^{n \times n}$ is diagonal and positive definite (Theorem 16, Spectral Decomposition Theorem).

Then upon defining $v := Q^T x$, we have

$$f(x) = \gamma + c^T x + \frac{1}{2} x^T QDQ^T x = \gamma + c^T Qv + \frac{1}{2} v^T Dv.$$

Thus, Coordinate Descent is the same as a search along the columns of Q in the x-space (if we are told ahead of time what the eigenvectors are).

This also shows how Coordinate Descent might suffer, if we do not have the "right basis".

Definition 48: H-conjugate

Let $H \in \mathbb{S}^n_{++}$. Then, $u, v \in \mathbb{R}^n$ are called H-**conjugate** if $u^T Hv = 0$.

Observation 49

If we have n, H-conjugate non-zero vectors, search along them sequentially will minimize $f(x) := c^T x + x^T Hx$, where $H \in \mathbb{S}^n_{++}$.

Lemma 50

Let $H \in \mathbb{S}^n_{++}$, suppose $d^{(1)}, \ldots, d^{(k)} \in \mathbb{R}^n \setminus \{0\}$ are pairwise H-conjugate. Then $\{d^{(1)}, \ldots, d^{(k)}\}$ is linearly independent.

Proof:

Let H and $d^{(1)}, \ldots, d^{(k)}$ be as in the statement of the lemma. Then, $H^{1/2} d^{(1)}, \ldots, H^{1/2} d^{(k)} \in \mathbb{R}^n \setminus \{0\}$ since $d^{(1)}, \ldots, d^{(k)} \in \mathbb{R}^n \setminus \{0\}$ and $H^{1/2}$ is nonsingular. Moreover, $H^{1/2} d^{(1)}, \ldots, H^{1/2} d^{(k)}$ are pairwise orthogonal (since they are H-conjugates), therefore, they are linearly independent. Thus under a change of basis with $H^{-1/2}$, we see that $\{d^{(1)}, \ldots, d^{(k)}\}$ is linearly independent. □

Theorem 51

Let $c \in \mathbb{R}^n$, $H \in \mathbb{S}^n_{++}$ be given. Define $f : \mathbb{R}^n \to \mathbb{R}$ by $f(x) := c^T x + \frac{1}{2} x^T Hx$. Further assume $d^{(0)}, d^{(1)}, \ldots, d^{(n-1)} \in \mathbb{R}^n \setminus \{0\}$ are pairwise H-conjugate, Let

$$D := \begin{bmatrix} d^{(0)} & d^{(1)} & \ldots & d^{(n-1)} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Then D is nonsingular and with $\hat{f}(y) := f(x^{(0)} + Dy)$ for any $x^{(0)} \in \mathbb{R}^n$, $\hat{f}$ is separable.

Proof:

Suppose $c, H, f, D, x^{(0)}, \hat{f}$ are as described in the statement of the theorem. Then, D is nonsingular, by Lemma 50. Moreover,

$$\begin{aligned}\hat{f}(y) &= c^T x^{(0)} + c^T Dy + \frac{1}{2}(x^{(0)} + Dy)^T H(x^{(0)} + Dy) \\ &= \left(c^T x^{(0)} + \frac{1}{2}x^{(0)^T} Hx^{(0)}\right) + (D^T c + D^T Hx^{(0)})^T y + \frac{1}{2}y^T(D^T HD)y\end{aligned}$$

$$(ij)^{\text{th}} \text{ entry of } (D^T HD) = \langle d^{(i)}, Hd^{(j)}\rangle = \begin{cases} 0 & \text{if } i \neq j \\ > 0 & \text{if } i = j \end{cases}$$

Therefore, $\hat{f}$ is separable. □

Corollary 52

Let $f, d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}$ be as above. IF we start with an arbitrary $x^{(0)} \in \mathbb{R}^n$ and successively search along the directions $d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}$ using exact line searches to get $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, then $x^{(j)}$ minimizes f on the affine space

$$\left\{x^{(0)} + \sum_{i=0}^{j-1} \mu_i d^{(i)} : \mu_i \in \mathbb{R}\right\} \qquad \forall j \in \{1, \ldots, n\}$$

and $x^{(n)}$ is the global minimizer of f.

Proof:

Follows from the last theorem. □

Conjugate Gradient Algorithm

Let f be as above, assume $x^{(0)} \in \mathbb{R}^n$ is given. $d^{(0)} := -\nabla f(x^{(0)})$.

Iteration k: (we have $x^{(k)}$ and $d^{(k)}$)

If $\nabla f(x^{(k)}) = 0$, set $x^{(k+1)} := x^{(k)}$

else $x^{(k+1)} := x^{(k)} + \underset{\substack{\uparrow \\ \text{obtained by exact line search}}}{\alpha_k} d^{(k)}$, $d^{(k+1)} := -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}$, where $\beta_k := \dfrac{\langle \nabla f(x^{(k+1)}), Hd^{(k)}\rangle}{\langle d^{(k)}, Hd^{(k)}\rangle}$.

Theorem 53

In the above algorithm, $d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}$ are pairwise H-conjugate and $x^{(n)}$ is the global minimizer of f.

Proof:

If $\nabla f(x^{(0)}) = 0$, then there is nothing left to prove. So we may assume $d^{(0)} \neq 0$. Proof is by induction on the iterate number k. Assume that $d^{(0)}, d^{(1)}, \ldots, d^{(k)}$ are all nonzero and pairwise H-conjugate. We will prove that

- either "$\nabla f(x^{(k+1)}) = 0$" → then, we are done!
- or "$d^{(0)}, d^{(1)}, \ldots, d^{(k+1)}$ are all nonzero and pairwise H-conjugate" → this will finish the proof.

Thus, we may assume $\nabla f(x^{(k+1)}) \neq 0$. Then, by Corollary 52, $x^{(k+1)}$ minimizes f on the set $\left\{x^{(0)} + \sum_{i=0}^{k} \mu_i d^{(i)} : \mu \in \mathbb{R}^{k+1}\right\}$. Then $\langle \nabla f(x^{(k+1)}), d^{(j)}\rangle = 0 \; \forall j \in \{0, 1, \ldots, k\}$.

Since $\langle \nabla f(x^{(k+1)}), d^{(k)} \rangle = 0, d^{(k+1)} = -\nabla f(x^{(k+1)}) + \beta_k d^{(k)} \neq 0$.

Next, we prove $\langle d^{(k+1)}, Hd^{(j)} \rangle = 0 \; \forall j \in \{0, 1, \dots, k\}$. By definition of β_k,

$$\langle d^{(k+1)}, Hd^{(k)} \rangle = \langle -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}, Hd^{(k)} \rangle = 0.$$

Consider $d^{(j)}$, $j \in \{0, 1, \dots, k-1\}$. $x^{(j+1)} = x^{(j)} + \alpha_j d^{(j)}$, and $\alpha_j > 0$ since

$$\langle \nabla f(x^{(j)}), d^{(j)} \rangle = \langle \nabla f(x^{(j)}), -\nabla f(x^{(j)}) + \beta_{j-1} d^{\overset{\substack{d_{-1} := 0 \\ \beta_{-1} := 0 \\ \downarrow}}{(j-1)}} \rangle \underset{\substack{\uparrow \\ \text{We used } \langle \nabla f(x^{(j)}), d^{(j-1)} \rangle = 0 \text{ by Corollary 52.}}}{=} - \left\| \nabla f(x^{(j)}) \right\|_2^2 < 0$$

So, $Hd^{(j)} = \dfrac{1}{\alpha_j} H[x^{(j+1)} - x^{(j)}] = \dfrac{1}{\alpha_j}[\nabla f(x^{(j+1)}) - \nabla f(x^{(j+1)}) - \nabla f(x^{(j)})]$.

Since $\nabla f(x^{(j+1)}) \in \operatorname{span}\{d^{(j)}, d^{(j+1)}\}$ and $\nabla f(x^{(j)}) \in \operatorname{span}\{d^{(j-1)}, d^{(j)}\}$, we have $Hd^{(j)} \in \operatorname{span}\{d^{(j-1)}, d^{(j)}, d^{(j+1)}\}$ $(*)$. Then

$$\langle d^{(k+1)}, Hd^{(j)} \rangle = \langle -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}, Hd^{(j)} \rangle \underset{\substack{\uparrow \\ \text{by induction hypothesis}}}{=} \langle -\nabla f(x^{(k+1)}), Hd^{(j)} \rangle \underset{\substack{\uparrow \\ (*) \text{ and, } \nabla f(x^{(k+1)}) \\ \text{is orthogonal to} \\ \text{all previous } d^{(j)}\text{'s.}}}{=} 0$$

This finishes the inductive step. □

Note the relationships with Gram-Schmidt orthogonalization/conjugation and the appearance of Krylov subspaces.

What if f is not quadratic?

"Nonlinear Conjugate Gradient"

We can apply the algorithm to an arbitrary $\mathcal{C}^1$ function f using, $y^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$:

$$\beta_k := \begin{cases} \dfrac{\left\langle \nabla f(x^{(k+1)}), y^{(k)} \right\rangle}{\left\langle d^{(k)}, y^{(k)} \right\rangle} & \begin{array}{c}\text{Sorensen-Wolfe (SW)} \\ \text{Hestenes-Stiefel}\end{array} \\[2ex] \dfrac{\left\langle \nabla f(x^{(k+1)}), \nabla f(x^{(k+1)}) \right\rangle}{\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle} & \text{Fletcher-Reeves} \\[2ex] \dfrac{\left\langle \nabla f(x^{(k+1)}), y^{(k)} \right\rangle}{\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle} & \text{Polak-Reibiére} \end{cases}$$

Still, we have to do exact (or almost exact) line search. Quadratic or cubic splines are used in applications. All of the above choices for β_k become the same on quadratic functions. Performance depends on the spectral structure of $\nabla^2 f(x^{(k)})$, including distribution of its eigenvalues. Hager & Zhang (2005) use

$$\beta_{k+1} := \left\langle y^{(k)} - 2\frac{\|y^{(k)}\|_2^2}{\langle d^{(k)}, y^{(k)} \rangle} d^{(k)}, \frac{\nabla f(x^{(k+1)})}{\langle d^{(k)}, y^{(k)} \rangle} \right\rangle = SW + 2\frac{\|y^{(k)}\|_2^2}{\langle d^{(k)}, y^{(k)} \rangle} \langle d^{(k)}, \nabla f(x^{(k+1)}) \rangle$$

2.8.1 Preconditioned Conjugate Gradient

Let L be lower triangular such that $LL^T \approx \nabla^2 f(x^{(k)})$ (e.g., "approximate" possible "incomplete" Choleski decomposition). Then apply Conjugate Gradient Algorithm to $\tilde{f}(x) := f(\underbrace{L^{-T}\tilde{x}}_{:=x}) \implies \nabla \tilde{f}(\tilde{x}) = L^{-1} \nabla f(L^{-T}\tilde{x})$.

Conjugate Gradient Algorithms are related to "Memoryless BFGS". CGAs can be even slower than

Steepest-Descent. Even on strongly convex functions they are not “optimal algorithms” with respect to the worst-case behaviour (Nemirovskii & Yudin (1980)).

3

Constrained Optimization

$f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p$, all assumed to be $\mathcal{C}^1$.

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & g(x) \leq 0 \\ & h(x) = 0 \end{array} \tag{P}$$

$S := \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\}$. For $\bar{x} \in S$, $J(\bar{x}) := \{i : \underset{\substack{\uparrow \\ \text{active constraints at } \bar{x} \\ \text{tight constraints at } \bar{x}}}{g_i(\bar{x}) = 0}\}$

$J := J(\bar{x})$, then g_J is the corresponding "subfunction".

$d \in \mathbb{R}^n$ is a **feasible direction** for (P) at $\bar{x}$, if $\exists \bar{\alpha} > 0$ such that $(\bar{x} + \alpha d) \in S \ \forall \alpha \in [0, \bar{\alpha})$.

Lemma 54

If $d \in \mathbb{R}^n$ is a feasible direction for (P) at $\bar{x}$, then $\langle \nabla g_i(\bar{x}), d \rangle \leq 0 \ \forall i \in J$ and $h'(\bar{x})d = 0$.

Recall: Corollary 24.

Lemma 55

Let $\bar{x} \in S$ such that $h'(\bar{x})$ has rank p, and $d \in \mathbb{R}^n$ satisfies $g_J'(\bar{x})d < 0$ and $h'(\bar{x})d = 0$. Then $\exists \bar{\alpha} > 0$ and a $\mathcal{C}^1$ arc $\hat{t} : [0, \bar{\alpha}) \to \mathbb{R}^n$ such that

$$\begin{cases} \hat{t}(0) = \bar{x} \\ \hat{t}'(0) = d \\ \hat{t}(\alpha) \in S \qquad \forall \alpha \in [0, \bar{\alpha}) \end{cases}$$

Proof Sketch:
Apply Corollary 24, to determine $\bar{\alpha} > 0$ (and to prove its existence) note that $\forall i \in [m] \setminus J$, $g_i(\bar{x}) < 0$ (by definition of $J = J(\bar{x})$). □

Corollary 56

If $\bar{x} \in S$ is a local minimizer of (P) and $h'(\bar{x})$ has rank p, then $\nexists d \in \mathbb{R}^n$ satisfying

$$\begin{cases} \langle \nabla g_i(\bar{x}), d \rangle < 0 \quad \forall i \in J(\bar{x}) \\ h'(\bar{x})d = 0 \\ \langle \nabla f(\bar{x}), d \rangle < 0 \end{cases}$$

If such a direction $d \in \mathbb{R}^n$ existed, then by Lemma 55 we would have feasible solutions along the $\mathcal{C}^1$ arc $\hat{t}(\alpha)$ for $\alpha \in [0, \bar{\alpha})$ that are better than $\bar{x}$, contradicting the fact that $\bar{x}$ is a local minimizer for (P).

Lemma 57: a theorem of the alternative-Farkas-type

Let $A \in \mathbb{R}^{n \times q}, B \in \mathbb{R}^{n \times r}$. Then exactly one of the following two systems has a solution:

(I) $A^T d < 0, B^T d = 0$

(II) $Au + Bv = 0, u \geq 0, u \neq 0.$

Proof:

Suppose (I) has a solution $\bar{d} \in \mathbb{R}^n$ and (II) has a solution $(\bar{u}, \bar{v}) \in \mathbb{R}^q \times \mathbb{R}^r$. Then,

$$0 = A\bar{u} + B\bar{v} \implies 0 = \underbrace{\bar{d}^T A}_{<0^T} \underset{\substack{\uparrow \\ \geq 0 \\ \neq 0}}{\bar{u}} + \underbrace{\bar{d}^T B}_{=0^T} \bar{v} < 0,$$

a contradiction.

Suppose (II) does not have a solution. Consider the LP

$$\begin{array}{rl} \max & \mathbb{1}^T u \\ \text{s.t.} & Au + Bv = 0 \\ & u \geq 0 \end{array} \tag{LP}$$

which is the dual of

$$\begin{array}{rl} \min & 0^T d \\ \text{s.t.} & A^T d \geq \mathbb{1} \\ & B^T d = 0 \end{array} \tag{LD}$$

(LD) is equivalent to: $\min\{0^T d : A^T d \leq -\mathbb{1}, B^T d = 0\}$. Since (II) has no solution and $\bar{u} := 0, \bar{v} := 0$ give a feasible solution of (LP) with objective value zero, optimal objective value of (LP) is zero. By Strong Duality Theorem of linear programming, (LD) has an optimal solution $\bar{d}$. Therefore, system (I) has a solution. □

We used:

Theorem 58: Strong Duality Theorem of Linear Programming

Let (LP) be a linear programming problem, and let (LD) be its dual. If (LP) has an optimal solution then so does its dual (LD); moreover, in this case, the optimal objective values of (LP) and (LD) are the same.

Theorem 59: Karush (1939), Fritz John (1948)

Suppose $\bar{x} \in S$ is a local minimizer for (P). Then $\exists \bar{\lambda} \in \mathbb{R}_+, \bar{u} \in \mathbb{R}^m_+, \bar{v} \in \mathbb{R}^p$, $\begin{pmatrix} \bar{\lambda} \\ \bar{u} \\ \bar{v} \end{pmatrix} \neq 0$ such that

$$\begin{aligned} \bar{\lambda} \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{p} \bar{v}_i \nabla f_i(\bar{x}) &= 0, \\ \sum_{i=1}^{m} \bar{u}_i \cdot g_i(\bar{x}) = 0. & \end{aligned} \tag{★}$$

Since $\bar{u} \geq 0$ and $g(\bar{x}) \leq 0$, (★) is equivalent to $\forall i \in [m]$, either $g_i(\bar{x}) = 0$ or $\bar{u}_i = 0$ (possible both). Complementary Slackness conditions, or Complementarity Conditions.

Proof:

Suppose $\bar{x} \in S$ is a local minimizer for (P). If $h'(\bar{x})$ does not have rank p, then $\exists \bar{v} \in \mathbb{R}^p \setminus \{0\}$ such that $\bar{v}^T h'(\bar{x}) = 0^T$. So, we may set $\bar{\lambda} := 0$ and $\bar{u} := 0$, and we are done.

Otherwise (rank $\left(h'(\bar{x})\right) = p$), by Corollary 56, the system

$$\begin{cases} \langle \nabla f(\bar{x}), d \rangle < 0 \\ \langle \nabla g_i(\bar{x}), d \rangle < 0 \quad \forall i \in J(\bar{x}) \\ \langle \nabla h_i(\bar{x}), d \rangle = 0 \end{cases}$$

has no solution. Thus, by Lemma 57, $\exists \bar{\lambda} \in \mathbb{R}_+, \bar{u}_J \geq 0, \bar{v} \in \mathbb{R}^p$ such that $\begin{pmatrix} \bar{\lambda} \\ \bar{u}_J \end{pmatrix} \neq 0$ and

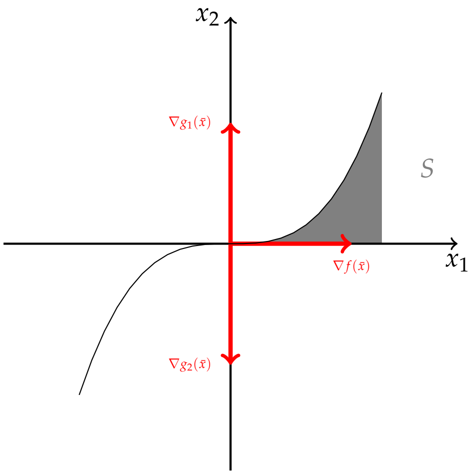
$$\bar{\lambda} \nabla f(\bar{x}) + \sum_{i \in J(\bar{x})} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{p} \bar{v}_i \nabla h_i(\bar{x}) = 0.$$

□

Note that being able to set $\bar{\lambda} = 0$ "makes the statement of the theorem work, without a Constraint Qualification" but it also takes away from its potential power.

Example 60:

$$\begin{array}{rl} \min & f(x) := x_1 \\ \text{s.t.} & g_1(x) := -x_1^3 + x_2 \leq 0 \\ & g_2(x) := -x_2 \leq 0 \end{array}$$



Here $\bar{x} := \begin{bmatrix}0\\0\end{bmatrix}, \nabla f(\bar{x}) = \begin{bmatrix}1\\0\end{bmatrix}, \nabla g_1(\bar{x}) = \begin{bmatrix}0\\1\end{bmatrix}, \nabla g_2(\bar{x}) = \begin{bmatrix}0\\-1\end{bmatrix}$

$\bar{\lambda}$ must be zero in this case; $\bar{\lambda} := 0, \bar{u}_1 := \bar{u}_2 := 1$ works.

Geometry of this problem is bad.

Example 61:

$$\begin{array}{rl} \min & f(x) := x_1 + x_2 \\ \text{s.t.} & g_1(x) := -x_1^3 \leq 0 \\ & g_2(x) := -x_2^3 \leq 0 \end{array} \tag{P}$$

Here $S = \mathbb{R}^2_+$.

$\bar{x} := \begin{bmatrix}0\\0\end{bmatrix}$, $\nabla g_1(\bar{x}) = \nabla g_2(\bar{x}) = \begin{bmatrix}0\\0\end{bmatrix}$, $\nabla f(\bar{x}) = \begin{bmatrix}1\\1\end{bmatrix}$. Again, we must set $\bar{\lambda} = 0$.

However, the geometry of S is nice! Nevertheless, the geometry of f', g', h' is determined by the original formulation (P) given by f, g, h. In this case, the formulation is bad! (P) is equivalent to an LP.

To have more useful results (than Theorem 59), we will look for necessary conditions in which $\bar{\lambda} > 0$.

3.1 The First-order Constraint Qualification at $\bar{x} \in S$

Let

$$D(\bar{x}) := \left\{ d \in \mathbb{R}^n : \begin{array}{ll} \langle \nabla g_i(\bar{x}), d \rangle \leq 0 & \forall i \in J(\bar{x}) \\ \langle \nabla h_i(\bar{x}), d \rangle = 0 & \forall i \in \{1, \dots, p\} \end{array} \right\}$$

Then, First-order Constraint Qualification holds at $\bar{x}$ if $\forall \bar{d} \in D(\bar{x})$ $\exists$ a sequence $\{d^{(k)}\}$ with $d^{(k)} \to \bar{d}$ such that $\exists \bar{\alpha}_k > 0$ and a $\mathcal{C}^1$ arc $t^{(k)} : [0, \bar{\alpha}_k) \to \mathbb{R}^n$ such that

$$\begin{cases} t^{(k)}(\alpha) \in S, & \forall \alpha \in [0, \bar{\alpha}) \\ t^{(k)}(0) = \bar{x} \\ (t^{(k)})'(0) = d^{(k)} \end{cases}$$

Informally, this means the polyhedral cone $D(\bar{x})$ is a reasonably good approximation to the set of feasible directions at $\bar{x}$.

In Example 60, $D(\bar{x}) = \text{span}\{e_1\}$. The Constraint Qualification looks ok for $d = e_1$ but fails for $d = -e_1$. Therefore, Constraint Qualification fails fails at $\bar{x}$.

In Example 61, $D(\bar{x}) = \mathbb{R}^2$. For $d = \begin{bmatrix}-1\\-1\end{bmatrix} \in D(\bar{x})$ the Constraint Qualification cannot be satisfied.

Lemma 62: Mangasarian-Fromowitz Constraint Qualification (1967)

Let $\bar{x} \in S$, $h, g \in \mathcal{C}^1$. If $h'(\bar{x})$ has rank p and $\exists \bar{d} \in \mathbb{R}^n$ such that

$$\begin{array}{ll} \langle \nabla g_i(\bar{x}), \bar{d} \rangle < 0 & \forall i \in J(\bar{x}), \\ \langle \nabla h_i(\bar{x}), \bar{d} \rangle = 0 & \forall i \in \{1, \dots, p\}, \end{array}$$

then the First-order Constraint Qualification holds at $\bar{x}$.

Proof:

Suppose the assumptions hold. Let $d \in D(\bar{x}), d^{(k)} := d + \frac{1}{k}\bar{d}$ $\forall k \in \mathbb{Z}_{++}$. Then $\langle \nabla g_i(\bar{x}), d^{(k)} \rangle < 0$ $\forall i \in J(\bar{x})$ and $h'(\bar{x})d^{(k)} = 0$ $\forall k \in \mathbb{Z}_{++}$. By Lemma 55, $\exists$ a suitable $\mathcal{C}^1$ arc $\hat{t}^{(k)}$ $\forall k \in \mathbb{Z}_{++}$. □

Corollary 63

Let $\bar{x} \in S$, $h, g \in \mathcal{C}^1$. If $\begin{bmatrix} g_J'(\bar{x}) \\ h'(\bar{x}) \end{bmatrix}$ has linearly independent rows, then the First-order Constraint Qualification holds at $\bar{x}$.

Proof:

Suppose $\bar{x} \in S$, and $g, h \in \mathcal{C}^1$. If $\exists \bar{d} \in \mathbb{R}^n$ satisfying $\begin{cases} g_J'(\bar{x})d < 0 \\ h'(\bar{x})d = 0 \end{cases}$, then we are done by Lemma 62.

Otherwise, by Lemma 57, $\begin{bmatrix} g_J'(\bar{x}) \\ h'(\bar{x}) \end{bmatrix}$ has linearly dependent rows. □

Corollary 64

If all constraints in (P) are linear (i.e., all functions $g_1, \dots, g_m, h_1, \dots, h_p$ are affine) then the First-order Constraint Qualification holds at every $x \in S$.

Proof:

Suppose the assumptions hold. Let $\bar{x} \in S$. For every $d \in D(\bar{x})$, set $\begin{cases} d^{(k)} := d \\ \hat{t}^{(k)} := \bar{x} + \alpha d \end{cases}$ $\forall k \in \mathbb{Z}_{++}$. □

Lemma 65

Let $A \in \mathbb{R}^{n \times q}, B \in \mathbb{R}^{n \times r}, c \in \mathbb{R}^n$. Then exactly one of the following systems has a solution:

1. $A^T d \leq 0, B^T d = 0, c^T d < 0$;
2. $c + Au + Bv = 0, u \geq 0$.

Theorem 66: First-order Necessary Conditions under Constraint Qualification (Karush 1939, Kuhn-Tucker 1951 (KKT Theorem))

Suppose $f, g, h \in \mathcal{C}^1$ and the First-order Constraint Qualification holds at $\bar{x} \in S$, a local minimizer for (P). Then, $\exists \binom{\bar{u}}{\bar{v}} \in \mathbb{R}^m \oplus \mathbb{R}^p$ such that

$$\begin{cases} \nabla f(\bar{x}) + [g'(\bar{x})]^T \bar{u} + [h'(\bar{x})]^T \bar{v} = 0 \\ \bar{u} \geq 0, \quad \bar{u}^T g(\bar{x}) = 0 \end{cases}$$

Proof:

Suppose the assumptions hold. Further suppose that $\exists d \in D(\bar{x})$ such that $\langle \nabla f(\bar{x}), d \rangle < 0$. Then by First-order Constraint Qualification, we can find $d^{(k)} \in \mathbb{R}^n$ such that $\langle \nabla f(\bar{x}), d^{(k)} \rangle < 0$ and $d^{(k)}$ is the first derivative of a feasible $\mathcal{C}^1$ arc $\hat{t}$ starting at $\bar{x}$. Defining $\phi(\alpha) := f(\hat{t}(\alpha))$ leads to $\phi'(0) = \langle \nabla f(\bar{x}), d^{(k)} \rangle < 0$. This leads to a contradiction to $\bar{x}$ being a local minimizer for (P). So, now we may assume the system

$$\begin{cases} \langle \nabla f(\bar{x}), d \rangle < 0 \\ g_J'(\bar{x})d \leq 0 \\ h'(\bar{x})d = 0 \end{cases}$$

has no solution.

By Lemma 65, $\exists \bar{u}_J \geq 0, \bar{v} \in \mathbb{R}^p$ such that $\nabla f(\bar{x}) + [g_J'(\bar{x})]^T \bar{u}_J + [h'(\bar{x})]^T \bar{v} = 0$. Setting $\bar{u}_i := 0$ $\forall i \in [m] \setminus J(\bar{x})$ yields the desired conclusion. □

Many algorithms for continuous optimization problems (and discrete optimization problems) are designed via these conditions.

KKT Conditions, KKT Triple

$$\left\{\begin{array}{rl} \left.\begin{array}{r} g(x) \leq 0 \\ h(x) = 0 \end{array}\right\} & \text{Primal feasibility} \\ \left.\begin{array}{r} \nabla f(x) + [g'(x)]^T u + [h'(x)]^T v = 0 \\ u \geq 0 \end{array}\right\} & \text{Dual feasibility} \\ u^T g(x) = 0 \Big\} & \text{Complementary Slackness} \end{array}\right\}$$

$\begin{pmatrix} \bar{x} \\ \bar{u} \\ \bar{v} \end{pmatrix}$ satisfying the above conditions (KKT conditions) is called a KKT triple.

Lagrangian: $\mathcal{L} : \mathbb{R}^n \oplus \mathbb{R}^m \oplus \mathbb{R}^p \to \mathbb{R}$

$$\boxed{\mathcal{L}(x, u, v) := f(x) + u^T g(x) + v^T h(x)}$$

$$\begin{aligned} &\nabla_x \mathcal{L}(x, u, v) = \nabla f(x) + [g'(x)]^T u + [h'(x)]^T v \\ &\nabla_u \mathcal{L}(x, u, v) = g(x) \\ &\nabla_v \mathcal{L}(x, u, v) = h(x) \end{aligned}$$

KKT COnditions can equivalently be stated as:

$$\left\{\begin{array}{l} \nabla_x \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) = 0 \\ \nabla_u \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) \leq 0 \\ \nabla_v \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) = 0 \\ \bar{u}^T \nabla_u \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) = 0 \\ \bar{u} \geq 0 \end{array}\right\}$$

This implies $\bar{x}$ satisfies First-order conditions for it to be a local minimizer of $\mathcal{L}(\cdot, \bar{u}, \bar{v})$ over $\mathbb{R}^n$; $\binom{\bar{u}}{\bar{v}}$ satisfies First-order conditions for it to be a local maximizer of $\mathcal{L}(\bar{x}, \cdot, \cdot)$ over $\mathbb{R}^m \oplus \mathbb{R}^p$. Therefore, $\begin{pmatrix} \bar{x} & \bar{u} & \bar{v} \end{pmatrix}^T$ satisfies the First-order conditions for it to be a saddle point of the Lagrangian.

Example 67:

Let $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p, c \in \mathbb{R}^n$ be given. Consider

$$\begin{array}{ll} \inf & f(x) := c^T x \\ \text{s.t.} & g(x) := -x \leq 0 \\ & h(x) := b - Ax = 0 \end{array} \tag{LP}$$

$$\begin{array}{ll} \sup & b^T v \\ \text{s.t.} & A^T v \leq c \end{array} \tag{LD}$$

Note:

$$\left\{\begin{array}{l} c + (-I)u + (-A^T)v = 0 \\ u \geq 0 \quad u^T x = 0 \end{array}\right\} \iff \left\{\begin{array}{l} A^T v \leq c \\ x^T(c - A^T v) \underset{\substack{\uparrow \\ c^T x = b^T y \\ \text{(using } Ax = b\text{)}}}{=} 0 \end{array}\right\}$$

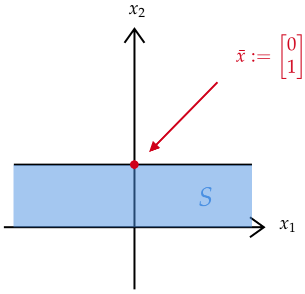
3.2 Second-order Conditions for Constrained Optimization

Assume $f, g, h \in \mathcal{C}^2$,

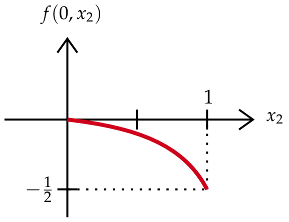
$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned} \tag{P}$$

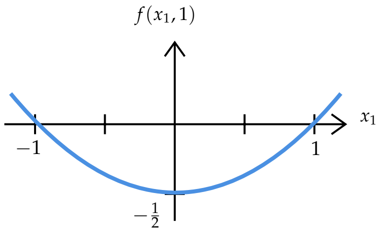
Example 68:

$$\begin{aligned} \inf \quad & f(x) := \tfrac{1}{2}x_1^2 - \tfrac{1}{2}x_2^2 \\ \text{s.t.} \quad & g_1(x) := x_2 - 1 \leq 0 \\ & g_2(x) := -x_2 \leq 0 \end{aligned}$$



$\bar{x}$ is the unique minimizer of (P). $J(\bar{x}) = \{1\}$.





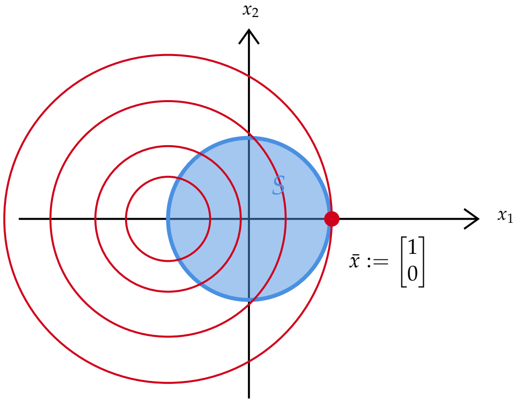
$\nabla f(\bar{x}) = \begin{bmatrix}0\\-1\end{bmatrix}$, $\nabla g_1(\bar{x}) = \begin{bmatrix}0\\1\end{bmatrix}$.

KKT conditions hold at $\bar{x}$ with $\bar{u} := \begin{bmatrix}1\\0\end{bmatrix}$. $\nabla^2 f(\bar{x}) = \begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}$, $\nabla^2 g_1(\bar{x}) = \nabla^2 g_2(\bar{x}) = 0$.

$\nabla^2 f(\bar{x})$ is not positive semidefinite. However, it is positive semidefinite in the $\underbrace{\text{appropriate linear subspace}}_{\{d \,:\, g_J'(\bar{x})d = 0,\ h'(\bar{x})d = 0\}}$ $\big(\underbrace{\text{tangent}}_{d_2 = 0}$ to $\underbrace{\text{the active constraints}}_{x_2 = 1}\big)$

Example 69:

$$\begin{aligned} \inf \quad & f(x) := -\tfrac{1}{2}(x_1 + 1)^2 - \tfrac{1}{2}x_2^2 \\ \text{s.t.} \quad & g_1(x) := \tfrac{1}{2}x_1^2 + \tfrac{1}{2}x_2^2 - \tfrac{1}{2} \leq 0 \end{aligned}$$



$\bar{x}$ is the unique optimal solution.

$\nabla f(\bar{x}) = \begin{bmatrix}-2\\0\end{bmatrix}$, $\nabla g_1(\bar{x}) = \begin{bmatrix}1\\0\end{bmatrix}$. KKT conditions are satisfied at $\bar{x}$ with $\bar{u} := 2$.

$\nabla^2 f(\bar{x}) = \begin{bmatrix}-1 & 0\\0 & -1\end{bmatrix}$, $\nabla^2 g_1(\bar{x}) = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$. $\nabla^2 f(\bar{x})$ is not positive semidefinite; but

$$\nabla^2_{xx}\mathcal{L}(\bar{x},\bar{u}) = \nabla^2 f(\bar{x}) + 2\nabla^2 g_1(\bar{x}) = -I + 2I = I$$

is positive semidefinite.

Second-order Constraint Qualification (at $\bar{x} \in S$) holds if

$$\left\{\begin{matrix} g_J'(\bar{x})d = 0 \\ h'(\bar{x})d = 0 \end{matrix}\right\} \implies \begin{matrix} \exists\bar{\alpha} \text{ and a } \mathcal{C}^2 \text{ arc } \hat{t} : [0,\bar{\alpha}) \to \mathbb{R}^n \text{ such that} \\ \left\{\begin{array}{ll} \hat{t}(0) = \bar{x} & \\ \hat{t}'(0) = d & \\ g_J(\hat{t}(\alpha)) = 0 & \forall\alpha \in [0,\bar{\alpha}) \\ h(\hat{t}(\alpha)) = 0 & \forall\alpha \in [0,\bar{\alpha}) \end{array}\right. \end{matrix}$$

Theorem 70: Second-order necessary conditions

Suppose $\bar{x} \in S$ is a local minimizer for (P) and second-order Constraint Qualification holds at $\bar{x}$. Then if $\begin{pmatrix}\bar{x} & \bar{u} & \bar{v}\end{pmatrix}^T$ is a KKT triple, we have

$$\left.\begin{matrix} g_J'(\bar{x})d = 0 \\ h'(\bar{x})d = 0 \end{matrix}\right\} \implies d^T\left[\nabla^2_{xx}\mathcal{L}(\bar{x},\bar{u},\bar{v})\right]d \geq 0.$$

Corollary 71

Suppose $\bar{x} \in S$ is a local minimizer for (P) and the first-order & second-order Constraint Qualifications hold at $\bar{x}$. Then, $\exists\bar{u} \in \mathbb{R}^m, \bar{v} \in \mathbb{R}^p$ such that

$$\nabla f(\bar{x}) + [g'(\bar{x})]^T\bar{u} + [h'(\bar{x})]^T\bar{v} = 0, \quad \bar{u} \geq 0, \quad \bar{u}^T g(\bar{x}) = 0,$$

and $\nabla^2_{xx}\mathcal{L}(\bar{x},\bar{u},\bar{v})$ is positive semidefinite on $\left\{ d \in \mathbb{R}^n : \begin{matrix} g_J'(\bar{x})d = 0 \\ h'(\bar{x})d = 0 \end{matrix} \right\}$.

Theorem 72

Suppose $g, h \in \mathcal{C}^2$, $\bar{x} \in S$. If $\begin{bmatrix} g_J'(\bar{x}) \\ h'(\bar{x}) \end{bmatrix}$ has linearly independent rows, then the First-order as well as Second-order Constraint Qualifications hold at $\bar{x}$.

Use the Implicit Function Theorem (Theorem 21).

Theorem 73: Second Order Sufficiency Condition

Suppose $\begin{pmatrix} \bar{x} & \bar{u} & \bar{v} \end{pmatrix}^T$ is a KKT triple for (P) and

$$\left.\begin{aligned} g_J'(\bar{x})d &\leq 0 \\ h'(\bar{x})d &= 0 \\ \bar{u}_J^T g_J'(\bar{x})d &= 0 \\ d &\neq 0 \end{aligned}\right\} \implies d^T \nabla_{xx}^2 \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) d > 0.$$

Then, $\bar{x}$ is a strict local minimizer of (P).

3.3 Strict Complementarity

Let $\begin{pmatrix} \bar{x} & \bar{u} & \bar{v} \end{pmatrix}^T$ be a KKT triple for (P). We say that $\begin{pmatrix} \bar{x} & \bar{u} & \bar{v} \end{pmatrix}^T$ satisfies **strict complementarity** (or, equivalently $\bar{x}$ and $\bar{u}$ are **strict complementary**) if the following holds $\begin{cases} g_i(\bar{x}) = 0 \\ \bar{u}_i = 0 \end{cases}$

Recall: Since we have a KKT triple, we already have $\forall i \in [m]$, at least one of $g_i(\bar{x})$, $\bar{u}_i$ is zero. When the KKT triple satisfies strict complementarity, the statement of the last theorem and its proof simplify.

Theorem 74: Second Order Sufficiency Condition when strict complementarity holds

Suppose $\begin{pmatrix} \bar{x} & \bar{u} & \bar{v} \end{pmatrix}^T$ is a strictly complementary KKT triple for (P) and

$$\left.\begin{aligned} \begin{bmatrix} g_J'(\bar{x}) \\ h'(\bar{x}) \end{bmatrix} d &= 0 \\ d &\neq 0 \end{aligned}\right\} \implies d^T \nabla_{xx}^2 \mathcal{L}(\bar{x}, \bar{u}, \bar{v}) d > 0.$$

Then $\bar{x}$ is a strict local minimizer of (P).

In a proof of Theorem 74 and in some similar situations, the following fact is useful.

Lemma 75

Let $A \in \mathbb{R}^{n \times q}$, $B \in \mathbb{S}^n$ such that

$$\left.\begin{aligned} A^T d &= 0 \\ d &\neq 0 \end{aligned}\right\} \implies d^T B d > 0.$$

Then, $\exists \bar{\rho} \geq 0$ such that

$$\forall \rho \geq \bar{\rho}, \quad (B + \rho A A^T) \in \mathbb{S}_{++}^n.$$

When (P) is a convex optimization problem (e.g., S is a convex set and f is a convex function on S), every local minimizer of (P) is a global minimizer of (P) and our results above can be made "global".

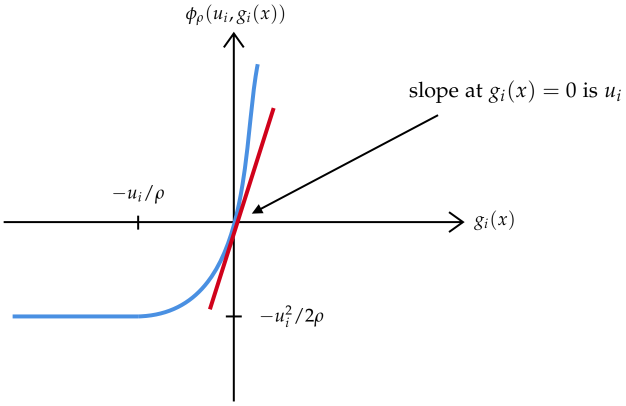
3.4 Augmented Lagrangians

Let $\rho > 0, \sigma > 0$.

$$\begin{aligned}\mathcal{L}_{\rho,\sigma}(x,u,v) &:= \inf_{\substack{y:\ y \geq g(x)\\ z:\ z=h(x)}} \left\{ f(x) + u^T y + v^T z + \frac{1}{2}\rho y^T y + \frac{1}{2}\sigma z^t z \right\}\\ &= f(x) + v^T h(x) + \frac{\sigma}{2}\|h(x)\|_2^2 + \sum_{i=1}^m \underbrace{\inf_{y_i \geq g_i(x)} \left\{ u_i y_i + \frac{1}{2}\rho y_i^2 \right\}}_{=:\phi_\rho(u_i, g_i(x))}\end{aligned}$$

Example:

Consider the case $u_i > 0$ (recall $\rho > 0$)



Theorem 76

Suppose KKT-triple $\begin{pmatrix}\bar{x} & \bar{u} & \bar{v}\end{pmatrix}^T$ satisfies the Second-order Sufficiency conditions for being a strict local minimizer for (P). Suppose strict complementary holds at $\begin{pmatrix}\bar{x} & \bar{u} & \bar{v}\end{pmatrix}^T$. Then $\exists \bar{\rho} \geq 0$ and $\bar{\sigma} \geq 0$ such that $\forall \rho \geq \bar{\rho}, \sigma \geq \bar{\sigma}$, $\bar{x}$ is a strict local minimizer of $\mathcal{L}_{\rho,\sigma}(\cdot, \bar{u}, \bar{v})$. Furthermore, $\begin{pmatrix}\bar{u} & \bar{v}\end{pmatrix}^T$ is a global maximizer of $\mathcal{L}_{\rho,\sigma}(\bar{x}, \cdot, \cdot)$.

3.4.1 Algorithms from Augmented Lagrangians

There are many ways to design algorithms based on Augmented Lagrangians. Let us put (P) into an equality form using new variables $\xi_i, i \in [m]$:

$$\begin{array}{lll} \inf & f(x) & \\ \text{s.t.} & g_i(x) + \xi_i^2 = 0 & i \in [m] \\ & h_i(x) = 0 & i \in [p] \end{array}$$

$$L_\rho\left(\begin{bmatrix} x \\ \xi \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right) = f(x) + u^T g(x) + \sum_{i=1}^m u_i \xi_i^2 + v^T h(x) + \frac{\rho}{2}\|\cdots\|_2^2$$

Let

$$\begin{aligned} L_\rho(x,u,v) &:= \inf_{\xi \in \mathbb{R}^m} \left\{ L_\rho\left(\begin{bmatrix} x \\ \xi \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right) \right\} \\ &= f(x) + \frac{1}{2}\rho\left[g(x) + \frac{u}{\rho}\right]_+^T \left[g(x) + \frac{u}{\rho}\right] - \frac{1}{2\rho}u^T u + v^T h(x) + \frac{1}{2}\rho\|h(x)\|_2^2 \end{aligned}$$

where, for $w \in \mathbb{R}^m$, $[w]_+ \in \mathbb{R}^m$ is defined by for each $j \in [m]$, $\max\{0, w_j\}$.

When L_ρ is differentiable in x,

$$\nabla_x L_\rho(x,u,v) = \nabla f(x) + \nabla g(x)[u+g(x)]_+ + \nabla h(x)[v+\rho h(x)].$$

Algorithm Choose $x^{(0)}, u^{(0)}, v^{(0)}, \rho_0 > 0; k := 0$. At iteration k, DO:

$$\begin{aligned} x^{(k+1)} &\underset{\substack{\uparrow \\ \text{perhaps only approximately}}}{:=} \operatorname*{argmin}_{x\in\mathbb{R}^n}\{L_{\rho_k}(x,u^{(k)},v^{(k)})\} \\ u^{(k+1)} &:= [u^{(k)} + \rho_k g(x^{(k+1)})]_+ \\ v^{(k+1)} &:= v^{(k)} + \rho_k h^{x^{(k+1)}} \\ &\text{update } \rho_k \text{ to } \rho_{k+1} \end{aligned}$$

How do we choose ρ_k?

- Present strategy (e.g., $\rho_k := \beta^k$, where $\beta > 1$ constant).
- Adaptive (if $g(x^{(k)})$ is "approximately ≤ 0" and $h(x) \approx 0$ then keep ρ_k the same; otherwise, increase ρ_k).

Now, let us consider (P) in pure inequality form.

Theorem 77: Bertsekas (1982)

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \end{aligned} \tag{P}$$

Suppose $\bar{x} \in \mathbb{R}^n$ is a local minimizer for (P); $f, g \in \mathcal{C}^2$ and $\nabla^2 f, \nabla^2 g_i$ $(i \in [m]) \in$ Lip in a neighborhood of $\bar{x}$. Further assume 2nd-order Sufficiency conditions hold at $\bar{x}$ with Lagrange multipliers $\bar{u} \geq 0$, $\nabla g_J(\bar{x})$ has full column rank, strict complementarity holds at $\begin{pmatrix}\bar{x} & \bar{u}\end{pmatrix}^T$.

Then, $\forall U \subset \mathbb{R}^m$ bounded, $\exists \bar{\rho} > 0$ such that $\rho > \bar{\rho}$ implies $L_\rho(\cdot, u)$ for $u \in U$ has a local minimizer $x(u,\rho)$ and $\exists$ a constant $M > 0$ such that

$$\begin{aligned} \|x(u,\rho) - \bar{x}\| &\leq \frac{M}{\rho}\|u - \bar{u}\|, \\ \|[u + \rho \cdot g(x(u,\rho))]_+ - \bar{u}\| &\leq \frac{M}{\rho}\|u - \bar{u}\|. \end{aligned}$$

Therefore, if we choose $\rho > M$, then we get at least Q-linear convergence of $u^{(k)}$s, and at least R-linear convergence of $x^{(k)}$s. If $\rho_k \to +\infty$ fast, we get Q-superlinear convergence of $u^{(k)}$s. If f, g_i are convex, then we get global convergence.

3.5 Method of Multipliers

Consider

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & Ax = b \end{aligned} \tag{P}$$

where $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$ given.

$$\mathcal{L}_\rho = f(x) + v^T(\overbrace{Ax - b}^{h(x)}) + \frac{\rho}{2}\|Ax - b\|_2^2$$

Algorithm Choose $x^{(0)} \in \mathbb{R}^n$, $v^{(0)} \in \mathbb{R}^p$, $\rho_0 > 0$. At iteration k, DO:

$$\begin{aligned} x^{(k+1)} &:= \operatorname*{argmin}_{x \in \mathbb{R}^n} \mathcal{L}_{\rho_k}(x, v^{(k)}) \\ v^{(k+1)} &:= v^{(k)} + \rho_k(Ax^{(k+1)} - b) \\ &\text{update } \rho_k \text{ to } \rho_{k+1} \end{aligned}$$

Suppose f is $\mathcal{C}^1$. Then, KKT conditions:

$$\begin{array}{ll} Ax = b & \text{primal feasibility} \\ \nabla f(x) + A^T v = 0 & \text{dual feasibility} \end{array}$$

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \mathcal{L}_{\rho_k}(x, v^{(k)})$$

$$\implies \nabla_x \mathcal{L}_{\rho_k}(x^{(k+1)}, v^{(k)}) = 0 = \nabla f(x^{(k+1)}) + A^T \underbrace{[v^{(k)} + \overset{\text{"dual step size"}}{\overset{\downarrow}{\rho_k}} (Ax^{(k+1)} - b)]}_{=v^{(k+1)}}$$

$$\iff \nabla f(x^{(k+1)}) + A^T v^{(k+1)} = 0$$

At the end of each iteration, $x^{(k)}, v^{(k)}$ satisfy dual feasibility. Algorithm strives to achieve primal feasibility.

3.6 Alternating Direction Method of Multipliers (ADMM)

We will again illustrate the algorithm for a special form of (P). Let $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ be $\mathcal{C}^1$ functions. $A_1 \in \mathbb{R}^{p \times n_1}, A_2 \in \mathbb{R}^{p \times n_2}, b \in \mathbb{R}^p$ be given.

$$\begin{array}{ll} \inf & f_1(x) + f_2(\xi) \\ \text{s.t.} & A_1 x + A_2 \xi = b \end{array} \tag{P}$$

$$\mathcal{L}_{\rho_k}\left(\begin{bmatrix} x \\ \xi \end{bmatrix}, v\right) = f_1(x) + f_2(\xi) + v^T(A_1 x + A_2 \xi - b) + \frac{\rho_k}{2} \|A_1 x + A_2 \xi - b\|_2^2$$

Algorithm Choose $x^{(0)} \in \mathbb{R}^{n_1}, \xi^{(0)} \in \mathbb{R}^{n_2}, v^{(0)} \in \mathbb{R}^p, \rho_0 > 0$. At iteration k, DO:

$$\begin{aligned} x^{(k+1)} &:= \operatorname*{argmin}_{x \in \mathbb{R}^{n_1}} \mathcal{L}_{\rho_k}\left(\begin{bmatrix} x \\ \xi^{(k)} \end{bmatrix}, v^{(k)}\right) \\ \xi^{(k+1)} &:= \operatorname*{argmin}_{x \in \mathbb{R}^{n_2}} \mathcal{L}_{\rho_k}\left(\begin{bmatrix} x^{(k+1)} \\ \xi \end{bmatrix}, v^{(k)}\right) \\ v^{(k+1)} &:= v^{(k)} + \rho_k\big(A_1 x^{(k+1)} + A_2 \xi^{(k+1)} - b\big) \\ &\text{update } \rho_k \text{ to } \rho_{k+1} \end{aligned}$$

In our illustration of the ADMM algorithm, we had a continuous optimization problem which was separable with respect to x and ξ: $f_1(x) + f_2(\xi)$. Of course, this approach easily extends to objective functions $f(x) = \sum_{\ell=1}^{L} f_\ell(x_\ell)$ which separate into $L \geq 2$ subfunctions.

3.7 Proximal Point Methods

There is a more general framework which unifies algorithms inspired by Augmented Lagrangians, ADMM, Douglas-Rachford splitting methods, operator splitting methods, Dykstra's alternating projections, Spingarn's method of partial inverses, Bregman iterations: Proximal Point Method(s)

Proximal Operator

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Suppose

$$\text{epi}(f) = \left\{ \begin{pmatrix} \mu \\ x \end{pmatrix} \in \mathbb{R} \oplus \mathbb{R}^n : f(x) \leq \mu \right\}$$

is closed and convex.

Proximal operator of f is $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$,

$$\text{prox}_f(z) := \underset{x \in \mathbb{R}^n}{\text{argmin}} \left\{ f(x) + \frac{1}{2}\|x - z\|_2^2 \right\}.$$

Consider the continuous optimization problem:

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & x \in S \end{array} \tag{P}$$

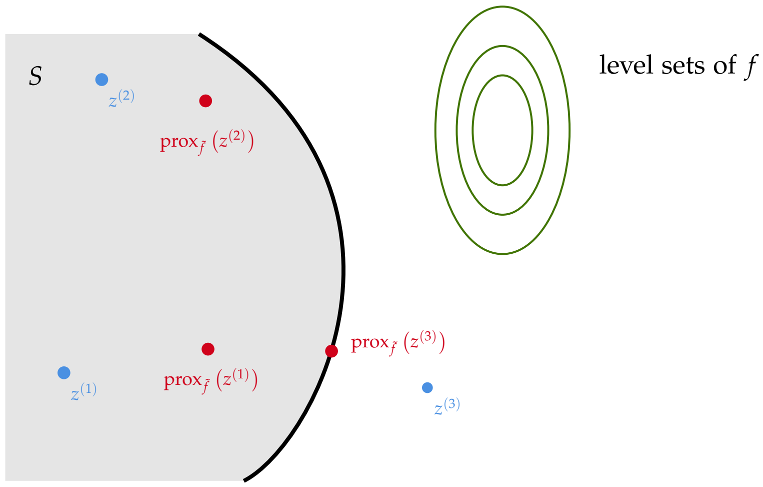
where $S \subseteq \mathbb{R}^n$ is a convex set, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a convex function.

Indicator function of S:

$$\delta(x|S) := \begin{cases} 0, & \text{if } x \in S \\ +\infty, & \text{otherwise.} \end{cases}$$

Define $\tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that $\tilde{f}(x) := f(x) + \delta(x|S)$. Then $\tilde{f}$ is convex and (P) is equivalent to the constrained convex optimization problem

$$\inf_{x \in \mathbb{R}^n} \tilde{f}(x) \tag{$\tilde{\text{P}}$}$$



S
$z^{(2)}$
$\text{prox}_{\tilde{f}}\left(z^{(2)}\right)$
level sets of f
$\text{prox}_{\tilde{f}}\left(z^{(3)}\right)$
$z^{(1)}$
$\text{prox}_{\tilde{f}}\left(z^{(1)}\right)$
$z^{(3)}$

We do not even need $\tilde{f}$ to be $\mathcal{C}^1$. $h \in \mathbb{R}^n$ is a subgradient of $\tilde{f}$ ar $\bar{x} \in \mathbb{R}^n$ if $\tilde{f}(x) \geq \tilde{f}(\bar{x}) + h^T(x - \bar{x})$ $\forall x \in \mathbb{R}^n$.

$$\underset{\substack{\uparrow \\ \text{subdiffential of } \tilde{f} \text{ at } \bar{x}}}{\partial \tilde{f}(\bar{x})} := \{h \in \mathbb{R}^n : h \text{is a subgradient of } \tilde{f} \text{ at } \bar{x}\}$$

(P) is equivalent to: find $\tilde{x} \in \mathbb{R}^n$ such that $0 \in \partial \tilde{f}(\tilde{x})$.

Algorithm (Proximal point algorithm): Choose $x^{(0)} \in \mathbb{R}^n, \lambda \in \mathbb{R}_{++}$. At iteration k, DO

$$x^{(k+1)} := \text{prox}_{\lambda f}(x^{(k)})$$
$$k := k + 1$$

In fact, $\text{prox}_{\lambda f}(\cdot) = \underbrace{(I + \lambda \partial f)^{-1}}_{\text{Resolvent operator}}(\cdot)$. This interpretation connected proximal point algorithms to Fixed Point Theory. More on this in CO 463.

3.8 Closest Points and Projections

Theorem 78: Kolmogorov Criterion

Let $S \subseteq \mathbb{R}^n$ be a nonempty closed convex set, and let $z \in \mathbb{R}^n$. Then the closest point $\operatorname{proj}(z|S)$ exists and is unique and it satisfies

$$(z - \operatorname{proj}(z|S))^T (z - \operatorname{proj}(z|S)) \leq 0 \quad \forall x \in S.$$

Proof:
See the proof of Corollary 111 in CO 255 Lecture Notes. □

$\operatorname{proj}(z|S)$ is the unique optimal solution of $\inf \{ \|x - z\|_2^2 : x \in S \}$.

A very useful characterization of the closest point (projection) applies to the case when S is a convex cone.

$$\underset{\substack{\uparrow \\ \text{dual cone (of } S)}}{S^*} := \{ s \in \mathbb{R}^n : x^T s \geq 0 \quad \forall x \in S \}.$$

Theorem 79: Moreau Decomposition

Let $S \subseteq \mathbb{R}^n$ be a nonempty closed convex cone and $z \in \mathbb{R}^n$. Then, $\bar{z} = \operatorname{proj}(z|S)$ if and only if $\bar{z} \in S$ and $\exists \bar{y} \in S^*$ such that $z = \bar{z} - \bar{y}$ and $\bar{z}^T \bar{y} = 0$.

In the above, $\bar{y} = \operatorname{proj}(-z|S^*)$.

Therefore, $\forall z \in \mathbb{R}^n$ can be expressed as $z = \operatorname{proj}(z|S) - \operatorname{proj}(-z|S^*)$.

Recall, $\forall z \in \mathbb{R}^n, z = [z]_+ - [-z]_+$.

3.9 A Stochastic Descent Algorithm

Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ be given. We want to find $\bar{x} \in \mathbb{R}^n$ such that $A\bar{x} \leq b$.

$$Ax \leq b \iff \langle a_i, x \rangle \leq b_i \qquad \forall i \in [m]$$

Choose $x^{(0)}$. At iteration k, DO

$$\begin{aligned} &\text{choose } i \in [m] \text{ uniformly randomly} \\ &x^{(k+1)} := \text{closest point in } \{x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i\} \text{ to } x^{(k)} \\ &k := k + 1 \end{aligned}$$

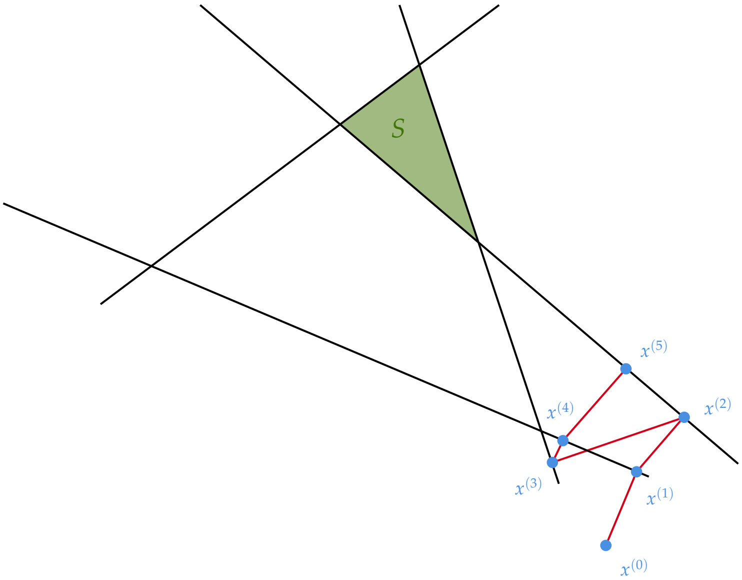
$S := \{x \in \mathbb{R}^n : Ax \leq b\}$.

Note that

$$x^{(k+1)} = x^{(k)} - \frac{[\langle a_i, x^{(k)} \rangle - b_i]_+}{\|a_i\|^2} a_i.$$

I.e., $x^{(k+1)} = x^{(k)}$ if $x^{(k)}$ lies in the halfspace $\{x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i\}$; otherwise, $x^{(k+1)}$ is the orthogonal projection of $x^{(k)}$ on the hyperplane $\{x \in \mathbb{R}^n : \langle a_i, x \rangle = b_i\}$.

We multiply both sides of i^{th} inequality by $\frac{1}{\|a_i\|_2}$. Thus, we may assume $\|a_i\|_2 = 1$ for all i.



Since $\|a_i\|_2 = 1\ \forall i \in [m]$, we have $\|A\|_F^2 = m$.

Theorem 80: Hoffman (1952)

Let $A \in \mathbb{R}^{m\times n}$. Then $\exists$ a constant L_A such that $\forall b \in \mathbb{R}^m$ for which $\{x \in \mathbb{R}^n : Ax \le b\} \neq \varnothing$, and $\forall \tilde{x} \in \mathbb{R}^n$,

$$\min_{x:\ Ax\le b} \|x - \tilde{x}\|_2 \le L_A \big\|[A\tilde{x} - b]_+\big\|_2.$$

The last inequality is equivalent to $\text{dist}(\tilde{x}, S) \le L_A \cdot \text{dist}(b - A\tilde{x}, \mathbb{R}^m_+)$. L_A is sometimes called the Lipschitz bound of A.

These type of results are also called "error bounds" in the literature. Generalizations to various classes of convex optimization problems exist.

Theorem 81: Leventhal-Lewis (2010)

Suppose $S \neq \varnothing$. Then the above algorithm converges linearly in expectation. In particular, $\forall k \in \mathbb{Z}_+$,

$$\mathbb{E}\Big[\big(\text{dist}(x^{(k+1)}, S)\big)^2 | x^{(k)}\Big] \le \left(1 - \frac{1}{mL_A^2}\right)\big(\text{dist}(x^{(k)}, S)\big)^2.$$

Proof:

Suppose $S \neq \varnothing$, let $k \in \mathbb{Z}_+, i \in [m]$. Note $[\text{dist}(x^{(k+1)}, S)]^2 = \big\|x^{(k+1)} - \text{proj}(x^{(k+1)}|S)\big\|_2^2$, and

$$\big\|x^{(k+1)} - \underbrace{\text{proj}(x^{(k)}|S)}_{\text{some point in } S}\big\|_2^2 \ge \big\|x^{(k+1)} - \underbrace{\text{proj}(x^{(k+1)}|S)}_{\substack{\text{closest point}\\ \text{to } x^{(k+1)} \text{ in } S}}\big\|_2^2$$

Thus,

$$\begin{aligned}
[\text{dist}(x^{(k+1)}, S)]^2 &\le \big\|x^{(k+1)} - \text{proj}(x^{(k)}|S)\big\|_2^2 \\
&= \big\|x^{(k)} - \big[\langle a_i, x^{(k)}\rangle - b_i\big]_+ a_i - \text{proj}(x^{(k)}|S)\big\|_2^2 \\
&\le \big\|x^{(k)} - \text{proj}(x^{(k)}|S)\big\|_2^2 + [\langle a_i, x^{(k)}\rangle - b_i]_+^2 - 2[\langle a_i, x^{(k)}\rangle - b_i]_+ \quad \underbrace{\langle a_i, x^{(k)} - \text{proj}(x^{(k)}|S)\rangle}_{=\langle a_i, x^{(k)}\rangle - b_i - \left(\langle a_i, \text{proj}(x^{(k)}|S) - b_i\rangle\right)} \\
&\le [\text{dist}(x^{(k)}, S)]^2 - [\langle a_i, x^{(k)}\rangle - b_i]_+^2
\end{aligned}$$

Take expectation over all $i \in [m]$, we have

$$\mathbb{E}\left[\left(\text{dist}(x^{(k+1)}, S)\right)^2 | x^{(k)}\right] \leq \left(\text{dist}(x^{(k)}, S)\right)^2 - \frac{1}{m}\|[Ax^{(k)} - b]_+\|_2^2$$

Now, we apply Theorem 90 to the second term in the RHS to get

$$-\frac{1}{m}\|[Ax^{(k)} - b]_+\|_2^2 \leq -\frac{1}{mL_A^2} \text{dist}(x^{(k)}|S)^2.$$

Therefore,

$$\mathbb{E}\left[\left(\text{dist}(x^{(k+1)}, S)\right)^2 | x^{(k)}\right] \leq \left(1 - \frac{1}{mL_A^2}\right)\left(\text{dist}(x^{(k)}, S)\right)^2$$

as desired. □

The underlying algorithm has its roots in the algorirhtm of Kaczmarz from 1930's (for solving systems of linear equations). We discussed Randomized Kaczmarz algorithm for systems of linear inequalities.

In the above algorithm and its analysis we illustrated some of the fundamental ingredients for Stochastic Gradient Descent (SGD) applied to $\inf_{x \in \mathbb{R}^n} f(x) := \sum_{i=1}^m f_i(x)$.

In (SGD) we randomly choose $i \in [m]$,

$$x^{(k+1)} := x^{(k)} - \underset{\substack{\uparrow \\ \text{step size (in ML, it is usually called the learning rate)}}}{\alpha_k} \nabla f_i(x^{(k)})$$

$$\vdots$$

Note that in our Randomized Kaczmarz Algorithm we used the probability distribution: $p_i = \frac{1}{m}$ for all i. If we hadn't normalized $\|a_i\|_2 = 1$ for all i, we would have chosen instead: $p_i = \frac{\|a_i\|_2^2}{\|A\|_F^2}$ $\forall i$.

Convergence speed may be very very slow on many instances. Why should we use it? (More like, when should we use it?)

- very very large instances (big data)
- highly parallelizable (if ∃ enough separablility)
- easy to code, easy to modify
- easy to analyze
- can try to strengthen by utilizing second-order info.
- ...

4

Sequential Quadratic Programming

Index