CO 466/666: Continuous Optimization

Rui Gong

September 8, 2020

Acknowledgements

These notes are based on the CO466/666 lectures given by Professor *Levent Tunçel* in Fall 2020 at the University of Waterloo.

Contents

1 Introduction: Formulations, fundamental background and definitions

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

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

We also have

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

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

Definition 1.1: Global Minimizer

*We have $\overline{x} \in \mathbb{R}^n$ is a **global minimizer of (P)** if $x \in S$ and $f(x) \geq f(\overline{x})$, $\forall x \in S$.*

Remark. Sometimes we simply say $\overline{x}$ is a minimizer

Definition 1.2: Local Minimizers

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

$$f(x) \geqslant f(\overline{x}), \forall x \in S \cap U$$

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

$$f(x) > f(\overline{x}), \forall x \in (S \cap U) \setminus \{\overline{x}\}$$

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

Definition: 0

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

What kind of problems can be formulated as Continuous Optimization problems?
A: **Almost Everything**

Example 1.3: Fermat's Last Theorem

There do not exist positive integers x, y, z *and an integer* $n \geqslant 3$ *such that*

$$x^n + y^n = z^n$$

Consider,

$$\begin{aligned}
&(P): \\
&inf\ f(x) := (x_1^{x_4} + x_2^{x_4} - x_3^{x_4})^2 + sin^2(\pi x_1) + sin^2(\pi x_2) + sin^2(\pi x_3) + sin^2(\pi x_4) \\
&\quad s.t. \\
&\qquad g_1(x) := 1 - x_1 \leqslant 0 \\
&\qquad g_2(x) := 1 - x_2 \leqslant 0 \\
&\qquad g_3(x) := 1 - x_3 \leqslant 0 \\
&\qquad g_4(x) := 3 - x_4 \leqslant 0
\end{aligned}$$

Note: $(x_1^{x_4} + x_2^{x_4} - x_3^{x_4})^2 = 0$ *iff* $x_1^{x_4} + x_2^{x_4} - x_3^{x_4} = 0$. $f(x) = 0$ *requires all* sin *in the* $f(x)$ *is zero, so all* x*'s are integers.*
Conclusion: *The optimal pbjective value of* (P) *is zero and attained iff Fermat's Last theorem is false.*
We can show that (P) *has a sequence of feasible solutions* $\left\{x^{(k)}\right\}$ *such that*

$$f(x^{(k)}) \searrow 0$$

Since $f(x) \geqslant 0, \forall x \in \mathbb{R}^4$*, the optimal value of* (P) *is zero.*
FLT is true ifff (P) ***does not attain its optimal value (of zero)***

What does this example tell us?
Even when 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

$$sin(\pi x_1) = 0 \Leftrightarrow x_1 \in \mathbb{Z}$$

In Example 1.3, we have functions that are "highly nonlinear"

Example 1.4: Combinatorial Optimization,0,1, Interger Programming

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

$$\begin{aligned}(IP):\ Min &: c^T x \\ s.t.& \\ & Ax \leqslant b \\ & x \in \{0,1\}^n\end{aligned}$$

We can have $g(x) := Ax - b \leqslant 0$ *and* $h(x) := \{x_j(x_j - 1) = 0, \forall j \in \{1, 2, \ldots, n\}\}$. *This is problem with linear objective function, linear inequality constraints and only quadratic equations.*
Our continuous optimization problem is only mildly nonlinear.

Some conclusions from Example 1.3 and 1.4:
Continuous Optimization problems can be very hard even when

- 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 of Continuous Optimization Problems

Definition 1.5: Cone

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

$$\forall x \in K, \forall \lambda \in \mathbb{R}_+,\ \lambda x \in K$$

$\mathbb{R}_+$ is the set of all non-negative real numbers.

Definition 1.6: Convex set

*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 $\forall u, v, \in S, \forall \lambda [0,1], [\lambda u + (1-\lambda)v] \in S$)

Definition 1.7: 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

$$\begin{aligned} &inf\ f(x) \\ &s.t.\ g(x) \underset{K}{\preccurlyeq} 0 \Leftrightarrow -g(x) \in K \end{aligned}$$

where $K \subseteq \mathbb{R}^m$ is a convex cone and for $u, v \in R^m$, $u \underset{K}{\succcurlyeq} v$ means $(u - v) \in K$
This is at least as general as our original (P), the very first problem in the introduction.
Consider $K := \mathbb{R}^m_+ \oplus \{0\}, 0 \in \mathbb{R}^p, \ldots$

1.2 Derivatives

Definition 1.8: Directional Derivative

Directional derivative of $f : \mathbb{R}^n \to \mathbb{R}$ at $\overline{x} \in \mathbb{R}^n$ along the direction $d \in \mathbb{R}^n$ is

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

(Gâteaux (directional) derivative)

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

Definition 1.9: Differentiable

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

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

such A is called the derivative of f at $\overline{x}$ and is denoted by $Df(\overline{x})$ or $f'(\overline{x})$ (matrix representation of $Df(\overline{x})$).
We will also use

$$\nabla f(\overline{x}) := [f'(x)]^T$$

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

$$Df(\overline{x}) \in L(\mathbb{E}_1, \mathbb{E}_2), DF : E_1 \to L(E_1, E_2)$$

$$D^2 f(\overline{x}) \in L(\mathbb{E}_1, L(\mathbb{E}_1, \mathbb{E}_2)), D^2 F : E_1 \to L(\mathbb{E}_1, L(E_1, E_2))$$

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

Theorem 1.10: Taylor Theorem

Let $U \subseteq \mathbb{R}$ be open, $f :\to \mathbb{R}$ be a C^r (r times continuous and differentiable) 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 $\exists z \in (x, x + d)$ such that

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

1.3 Contraction and Fixed Points

Definition 1.11: Contraction Mapping

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

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

Theorem 1.12: 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

1. *(Existence and Uniqueness of solution-fixed point)*
   the mapping f has a unique fixed point $\overline{x} \in U$

2. *(Algorithm and convergence)*
   $\forall x^{(0)} \in U$, the sequence $\left\{x^{(k)}\right\}$ generated by

   $$x^{(k+1)} := f(x^{(k)}), k \in \{0, 1, 2, \ldots\} \Rightarrow \textbf{\textit{Fixed Point Iteration}}$$

   converges to $\overline{x}$. In particular,

   $$||x^{(k)} - \overline{x}|| \leqslant \lambda^k ||x^{(0)} - \overline{x}||, \forall k \in \{0, 1, 2, \ldots\}$$

10, September 2020

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

$$x^{(k+1)} := f(x^{(k)}),\ \forall k \in \mathbb{Z}_+$$

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

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

By induction on $k, \ldots$ we obtain

$$||x^{(k+1)} - x^{(k)}|| \leqslant \lambda^k ||x^{(1)} - x^{(0)}||, \forall k \in \mathbb{Z}_+ \ldots (eq.1)$$

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

$$\begin{aligned}
||x^{(m+k)} - x^{(m)}|| &= ||x^{(m+k)} - x^{(m+k-1)} + x^{(m+k-1)} - x^{(m+k-2)} + \ldots + x^{(m+1)} - x^{(m)}|| \\
&\leqslant \sum_{i=1}^{k} ||x^{(m+i)} - x^{(m+i-1)}|| \text{ By triangle inequality} \\
&\leqslant (\lambda^{m+k-1} + \lambda^{m+k-2} + \ldots + \lambda^m)||x^{(1)} - x^{(0)}||, by\ (eq.1) \\
&= \lambda^m (1 + \lambda + \lambda^2 + \ldots + \lambda^{k-1})||x^{(1)} - x^{(0)}|| \\
&= \frac{\lambda^m (1 - \lambda^k)}{1 - \lambda} ||x^{(1)} - x^{(0)}|| \\
&\leqslant \frac{\lambda^m}{1 - \lambda} ||x^{(1)} - x^{(0)}|| \to 0\ as\ m \to +\infty \text{ independent of k}
\end{aligned}$$

$\therefore \left\{x^{(k)}\right\}$ is a Cauchy sequence and hence it converges (U is complete). Let $\overline{x}$ be its limit. $\overline{x} \in U$ (it is closed).
$\forall k \in \mathbb{Z}_+$, we have

$$\begin{aligned}
||f(\overline{x}) - \overline{x}|| &\leqslant ||f(\overline{x}) - x^{(k)}|| + ||x^{(k)} - \overline{x}|| \\
&\leqslant \lambda \underbrace{||\overline{x} - x^{(k-1)}||}_{0} + \underbrace{||x^{(k)} - \overline{x}||}_{0}
\end{aligned}$$

As $k \to +\infty,\ RHS \to 0$. Thus, $f(\overline{x}) = \overline{x}$ (Existence proven)
Uniqueness: Suppoese $\exists \overline{x}, \overline{y} \in U,\ s.t. f(\overline{x}) = \overline{x}$ and $f(\overline{y}) = \overline{y}$. Then

$$\begin{aligned}
&||\overline{x} - \overline{y}|| = ||f(\overline{x}) - f(\overline{y})|| \leqslant \lambda ||\overline{x} - \overline{y}|| \\
&\Rightarrow (1 - \lambda)||\overline{x} - \overline{y}|| = 0 \underset{\lambda \in [0,1)}{\Rightarrow} \overline{x} = \overline{y}
\end{aligned}$$

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

$$||x^{(1)} - \overline{x}|| = ||f(x^{(0)}) - f(\overline{x})|| \leqslant \lambda ||x^{(0)} - \overline{x}||$$
$$\Rightarrow ||x^{(2)} - \overline{x}|| \leqslant \lambda^2 ||x^{(0)} - \overline{x}||$$
By induction on $k, \ldots$
$$||x^{(k)} - \overline{x}|| \leqslant \lambda^k ||x^{(0)} - \overline{x}||, \ \forall k \in \mathbb{Z}_+$$

as desired. □

Theorem 1.13: Brouwer's Fixed Point Thm[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 $\exists \overline{x} \in U$ s.t. $f(\overline{x}) = \overline{x}$

Theorem 1.14: 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 (2^U is the set of all subsets of U). If $Graph(f) := \left\{ \begin{pmatrix} x \\ v \end{pmatrix} \in U \oplus U : v \in f(x) \right\}$ is closed and $f(x)$ is nonempty and convex for every $x \in U$, then $\exists\, \overline{x} \in U$ s.t. $\overline{x} \in f(\overline{x})$

Theorem 1.15: Borsuk-Ulam Theorem[1930-1933]

Let $f : \{x \in \mathbb{R}^{n+1} : ||x||_2 = 1\} \to \mathbb{R}^n$ be continuous. Then $\exists\, \overline{x} \in \mathbb{R}^{n+1}$ s.t. $||\overline{x}||_2 = 1$ and $f(\overline{x}) = f(-\overline{x})$

Example

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

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

Theorem 1.16: Spectral Decomposition Theorem

For every $A \in \mathbb{S}^n, \exists\, Q \in \mathbb{R}^{n\times n}$ *orthoganal* $(Q^TQ = I)$ *such that*

$$A = QDQ^T, \text{ where } D \in \mathbb{R}^{n\times n} \text{ is a diagonal matrix.}$$

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 1.17

$A \in \mathbb{R}^{n\times n}$ *is* ***positive semidefinite*** *if* $h^TAh \geqslant 0,\ \forall h \in \mathbb{R}^n$*;*
such A *is* ***positive definite*** *if* $h^TAh > 0,\ \forall h \in \mathbb{R}^n \setminus \{0\}$

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

Notation.

- $\mathbb{S}^n_+ :=$ positive semidefinite matrices in $\mathbb{S}^n$,
- $\mathbb{S}^n_{++} :=$ positive definite matrices in $\mathbb{S}^n$.

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

Theorem 1.18: Choleski Decomposition Theorem

Let $A \in \mathbb{S}^n$*. Then,*

(a) A *is positive semidefinite iff* $\exists L \in \mathbb{R}^{n\times n}$ *lower triangular such that* $A = LL^T$

(b) A *is positive definite iff* $\exists L \in \mathbb{R}^{n\times n}$ *lower triangular and non-singular such that* $A = LL^T$

For (b), L *is non-singular makes sure there is no non-zero vector in the null space of* L*.*

15, September 2020

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

However, we have

Theorem 1.19

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

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

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

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

Then, we have

$$\begin{aligned}
&\|f(\overline{x} + d) - f(\overline{x}) - Df(\overline{x})d\|_2 \\
&\qquad = \|\int_0^1 [Df(\overline{x} + \alpha d) - Df(\overline{x})]d \, (\partial\alpha)\|_2 \\
&\qquad \leqslant \int_0^1 \|Df(\overline{x} + \alpha d) - Df(\overline{x})\|_2 * \|d\|_2 \, (\partial\alpha)
\end{aligned}$$

Prove the inequality later. First norm above is an operator 2-norm, the second one is a 2-norm on $\mathbb{R}^n$

$$\begin{aligned}
&\qquad \leqslant \int_0^1 L * \|d\|_2 * \|d\|_2 \alpha \, (\partial\alpha) \\
&\qquad = \frac{1}{2} L * \|d\|_2^2
\end{aligned}$$

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

(The estimate is $f(\overline{x}) + Df(\overline{x})d$)

$$h := \int_0^1 [Df(\overline{x} + \alpha d) - Df(\overline{x})]d \, (\partial\alpha)$$

$$\|h\|_2^2 = h^T h = h^T \int_0^1 [Df(\overline{x} + \alpha d) - Df(\overline{x})]d \, (\partial\alpha)$$

$$= \int_0^1 h^T [Df(\overline{x} + \alpha d) - Df(\overline{x})]d \, (\partial\alpha)$$

$$\leqslant \int_0^1 \|h\|_2 \|[Df(\overline{x} + \alpha d) - Df(\overline{x})]d\|_2 (\partial\alpha)$$

By Cauchy-Schwarz

$$\implies \|h\|_2 \leqslant \int_0^1 \|[Df(\overline{x} + \alpha d) - Df(\overline{x})]d\|_2 (\partial\alpha)$$

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

Theorem 1.20: Inverse Function Theorem

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

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

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

Theorem 1.21: Implicit Function Theorem

let $h : \mathbb{R}^n \to \mathbb{R}^p, h \in C^1$ in a neighborhood of $\overline{x} \in \mathbb{R}^n$ where $h(\overline{x}) = 0$. Suppoese $h'(\overline{x})$ has full row rank ($rank(h'(\overline{x})) = p \leqslant n$). Define a partition $[B|N]$ of columns of $h'(\overline{x})$:

$$h'(\overline{x}) =: [h'_B(\overline{x})|h'_N(\overline{x})]$$

where $h'_B(\overline{x}) \in \mathbb{R}^{p\times p}$, nonsingular, partition $\overline{x}$ and x with respect to the same $[B|N]$. Then, $\exists$ neighborhood U_B of $\overline{x}_B$ and U_N of $\overline{x}_N$ and a C^1 function $f : U_N \to U_B$ such that

- $f(\overline{x}_N) = \overline{x}_B$
- $h\begin{pmatrix} x_B \\ x_N \end{pmatrix} = 0 \Leftrightarrow x_B = f(x_N),\ \forall x_B \in U_B, x_N \in U_N$

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

Recall the very special case (e.g. equality constraints in a LP problem):

$$A \in \mathbb{R}^{p\times n},\ rank(A) = p, \text{ given } \begin{cases} Min\ c^T x \\ Ax = b \\ x \geqslant 0 \end{cases}$$

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

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

Lemma 1.22

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

$$D(f_2 \circ f_1)(\overline{x}) = Df_2(f_1(\overline{x})) \circ Df_1(\overline{x}), \forall \overline{x} \in U$$

Example: Line Search, directional derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ *differentiable on* $\mathbb{R}^n$ *is given also given are a current point* $\overline{x} \in \mathbb{R}^n$ *and a "search direction "* $d \in \mathbb{R}^n$.
We define $\phi : \mathbb{R} \to \mathbb{R}$ *by* $\phi(\alpha) := f(\overline{x} + \alpha d)$, *then*

$$\phi'(\alpha) =< \nabla f(\overline{x} + \alpha d), d >$$

. *If* f *is* C^2, *then* $\phi''(\alpha) = d^T \nabla^2 f(\overline{x} + \alpha d)d$. *Note*

- $\phi'(0) =< \nabla f(\overline{x}, d) >$
- $\phi''(0) = d^T \nabla^2 f(\overline{x})d$

Corollary 1.23

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

- $t(0) = 0$
- $t'(0) = 0$
- $h(\overline{x} + zd_z + t(d_z)) = 0, \ \forall d_z \in U$

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

Proof. Let $h, \overline{x}$ and z be as in the assumptions. Using the partition $[B|N]$, define

$$z =: \begin{bmatrix} z_B \\ z_N \end{bmatrix} (\text{recall} h'(\overline{x}) = [h'_B(\overline{x})|h'_N(\overline{x})])$$

let

$$U := \{d_z \in \mathbb{R}^q : (\overline{x}_N + z_N d_z) \in U_N\}$$

Define t by

$$t_N(d_z) := 0, \ t_B(d_z) := f(\overline{x}_N + z_N d_z) - \overline{x}_B - z_B d_z$$

Thus,

$$\begin{aligned} h(\overline{x} + zd_z + t(d_z)) &= h \begin{bmatrix} \overline{x}_B + z_B d_z + f(\overline{x}_N + z_N d_z) - \overline{x}_B - z_B d_z \\ \overline{x}_N + z_N d_z + 0 \end{bmatrix} \\ &= h \begin{bmatrix} f(\overline{x}_N + z_N d_Z) \\ \overline{x}_N + z_N d_z \end{bmatrix} && = 0 \text{ By theorem 21} \end{aligned}$$

Also,

$$\begin{aligned} t(0) &= f(\overline{x}_N) - \overline{x}_B = 0 \\ t'_N(0) &= 0 \\ t'_B(0) &= f'(\overline{x}_N)z_N - z_B \text{ (By chain rule 22)} \\ &= -[h'_B(\overline{x})]^{-1}h'_N(\overline{x})z_N - z_B \\ &= [h'_B(\overline{x})]^{-1}\underbrace{[-h'_N(\overline{x})z_N - h'_B(\overline{x})z_B]}_{-h'(\overline{x})z=0} \\ &= 0 \end{aligned}$$

□

Question: What does the size of the neighborhood depend on? Note: In LPs $t(d_z) := 0 \forall d_z \in \mathbb{R}^q$

Corollary 1.24

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

- $\hat{t}(0) = \overline{x}$
- $h(\hat{t}(\lambda)) = 0, \ \forall \in [0, \overline{\lambda})$
- $\hat{t}'(0) = d$

Proof. In the statement of Corollary 23, plug in $z := d$ and then using the resulting t, $\hat{t}(\lambda) := \overline{x} + \lambda d + t(\lambda)$ where think λ as "d_z" and d as z. □

17, September 2020

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 $\overline{y} \in \mathbb{R}^p$ be given then

$$h^{-1}(\overline{y}) = \{x \in \mathbb{R}^n : Ax = \overline{y} + b\}$$

Theorem 1.25: Sard's Theorem, Morse-Sard Theorem

Let $h : \mathbb{R}^n \to \mathbb{R}^p$, where $p \leqslant n, h \in C^r$ with $r \geqslant n - p + 1$. Then the p-dimensional Lebesgue measure of

$$\{y \in \mathbb{R}^p : y \textit{ is not a regular value}\} \textit{ is zero}$$

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

2 Unconstrained Continuous Optimization

$$\begin{aligned} (P)\ inf\ f(x) \quad & f:\mathbb{R}^n \to \mathbb{R} \\ s.t.\ g(x) \leqslant 0 \quad & g:\mathbb{R}^n \to \mathbb{R}^m \\ h(x) = 0 \quad & h:\mathbb{R}^n \to \mathbb{R}^p \end{aligned}$$

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

Here, we assume $S = \mathbb{R}^n$

Theorem 2.1: First-order necessary conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and $S = \mathbb{R}^n$. Then, $\overline{x} \in \mathbb{R}^n$ is a local minimum for $(P) \Rightarrow f'(\overline{x}) = 0$
$\overline{x}$ is a stationary point of f

Proof. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is C^1, $S = \mathbb{R}^n$, and $\overline{x} \in \mathbb{R}^n$ is a local minimizer for (P).
For the sake of seeking a contradiction, suppose $f'(\overline{x}) \neq 0$. Then, $\exists d \in \mathbb{R}^n$ such that $\langle f'(\overline{x}), d\rangle < 0$(e.g. let $A \in \mathbb{S}^n_{++}$, and set $d := -Af'(\overline{x})$). Consider $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(\alpha) := f(\overline{x} + \alpha d)$. Then,

$$\phi'(0) = \langle f'(\overline{x}), d\rangle < 0$$

Thus, for all sufficiently small, positive α, $f(\overline{x}+\alpha d) < f(\overline{x})$. Therefore, $\overline{x}$ 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 2.2

$d \in \mathbb{R}^n$ is a descent direction for f at $\overline{x} \in \mathbb{R}^n$, if $\langle f'(\overline{x}), d\rangle < 0$.
$d \in \mathbb{R}^n$ is a improving direction for f at $\overline{x} \in \mathbb{R}^n$, if $f(\overline{x} + \alpha d) < f(\overline{x})$, $\forall \alpha > 0$ and sufficiently small.

Theorem 2.3: Second-Order necessary conditions

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

Proof. Suppose $\overline{x}$ is a local minimizer for (P). Since f is C^2 by theorem 27, $f'(\overline{x}) = 0$. Suppose for the sake of contradiction that $\nabla^2 f(\overline{x}) \notin \mathbb{S}^n_+$. Since $f \in C^2$, $\nabla^2 f(\overline{x}) \in \mathbb{S}^n$. So, $\exists d \in \mathbb{R}^n$ such that $d^T \nabla^2 f(\overline{x}) d < 0$. Define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(\alpha) := f(\overline{x} + \alpha d)$. THen $\phi'(0) = \left\langle \underbrace{\nabla f(\overline{x})}_{=0}, d \right\rangle = 0$, $\phi''(0) = d^T \nabla^2 f(\overline{x}) d < 0$
Therefore, for all $\epsilon > 0$ and sufficiently small $f(\overline{x} + \epsilon d) < f(\overline{x})$ which contradicts the fact that $\overline{x}$ is a local minimizer for (P). □

Definition 2.4

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

Theorem 2.5: Taylor's Theorem– implicit remainder version

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

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

where $R(\overline{x}, *) : \mathbb{R}^n \to \mathbb{R}$ such that

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

Theorem 2.6: Second order sufficient conditions

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

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

$$\underbrace{\delta}_{\lambda_n(\nabla^2 f(\overline{x}))} := min \left\{ d^T \nabla^2 f(\overline{x}) d : \|d\|_2 = 1 \right\} > 0$$

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

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

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

Proposition 2.7

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2 and consider $\tilde{f}(x) := f(x) + c^T x$, where $c \in \mathbb{R}^n$ is given. Then for almost all $c \in \mathbb{R}^n$,
$\tilde{f}'(\overline{x}) = 0 \Rightarrow \nabla^2 f(\overline{x})$ is nonsingular.

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

Definition 2.8

$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is* <u>*convex*</u> *if* $epi(f) := \left\{ \begin{pmatrix} \mu \\ x \end{pmatrix} \in \mathbb{R} \oplus \mathbb{R}^n : f(x) \leqslant \mu \right\}$ *is convex*

Theorem 2.9

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 minimzer of* (P)*.*

2.1 Unconstrained Continuous Optimization and Affine Subspace Constraints

One of the most popular form of continuous optimization problems is

$$(P) = \begin{cases} inf\ f(x),\ A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p \\ s.t.\ Ax = b \end{cases}$$

At a first glance (and strictly speaking) (P) does not belong to the class of Unconstrained continuous optimizating problems.
We may assume $rank(A) = p$; otherwise

- We easily prove $Ax = b$ has no solution $\implies$ (P) is infeasible
- Or we easily find all redundant equations and $\overline{x} \in \mathbb{R}^n$ s.t. $A\overline{x} = b$

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

$$[A_B|A_N] := A,\ \begin{bmatrix} x_B \\ x_N \end{bmatrix} := x$$

Then

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

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

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

Thus, (P) is equivalent to

$$(\tilde{P})\ inf\ \tilde{f}(\overline{x}),\ x \in \mathbb{R}^{n-p}$$

any algorithm from any starting point $x^0 \in \mathbb{R}^{n-p}$.
Another equivalent approach:
Let $\overline{x} \in S(i.e.,\ A\overline{x} = b)$. Then,

$$S = \{\overline{x} + u :\ u \in Null(A)\}$$

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

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

where $\hat{f} : \mathbb{R}^{n-p} \to \mathbb{R}$ is defined as $\hat{f}(v) = f(\overline{x} + Zv)$ In applications, with either of these two approaches, we must be very careful about exploiting sparsity as well as making sure we can efficitently and accurately evalueate all ingredients of the algorithms we choose to use on such problems.

2.2 A small preview and dipping our toes into some applciations

Some other ways of dealing with constrained optimization problems using Unconstrained optimization algorithms:

1. Form the Lagrangian for (P):

   $$l(x, v) := f(x) + v^T(b - Ax)$$

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

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

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

   where $\beta, \alpha \in \mathbb{R}$ suitably defined, $\eta \in \mathbb{R}_{++}$ is a penalty parameter (think about spline regression, smoother, etc)

In compressed sensing and related applications one seeks a solution of

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

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

$$inf\{f(x) + \eta_1||x||_1 + \eta_2||Ax = b||_x^\alpha\}$$

where $\eta_1, \eta_2, \alpha \in \mathbb{R}$ are usually fixed.
We can generalize such approaches to matrix variables. Very many interesint applications in Machine Learning, AI and modern Data Science. In many of these applications, we want to find a low-rank solution.

$$E.g.\ min\{rank(x) :\ A(x) = b\}$$

where $A : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ linear, $b \in \mathbb{R}^p$ both A, b are given.

2.2.1 Prototype lowrank approximation problem

Given $A \in \mathbb{R}_+^{m\times n}$ (both $m\&n$ are huge), we want to find matrices $u \in \mathbb{R}_+^{m\times k}, V \in \mathbb{R}_+^{n\times k}$ such taht $A = UV^T$ 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 $Q_1 \in \mathbb{R}^{m\times n}, Q_2 \in \mathbb{R}^{n\times n}$ are orthogonal and $D \in \mathbb{R}^{m\times n}$ diagonal. Let's assume $m \leqslant n$, then

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

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

Theorem 2.10

Every $A \in \mathbb{R}^{m\times n}$ has a SVD. Requiring U, V to be nonnegative, makes the problem hard. Let p be an upper bound on k (taking p as $(nm+1)$ suffices, but in pratice, better guesses can help).
Suppose our guess for the min. 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 matrtices 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 fixe, and U_- denotes the $\mathbb{R}^{m\times p}$ matrix with obnly negative entries of U.

2.3 Classical Algorithmic Approaches to Unconstrained Continuous Option

1. 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

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

$$min\ m_k(x^{(k)} + d)$$
$$s.t.\ d \in \text{Trust Region (around } x^{(k)})$$

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

$$min\ m_k(d) := f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), d \right\rangle + \frac{1}{2} d^T B_k d$$
$$s.t.\ ||d||_2 \leqslant S_k$$

Let $\overline{d}$ denotes an optimal solution of this trust-region subproblem. If $x^{(k)} + d$ satisfies certain criteria, then set $x^{(k+1)} := x^{(k)} + \overline{d}$; otherwise either modify σ_k, or the step size, ...
Depending on how well we did with teh latest σ_k choose a suitable value for σ_{K+1} and repeat. (Size of the Trust-Region is being adjusted.)

2.4 Search direction +line search strategies

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

For convex optimization problems and also near local minimizers of nonvoncex problems we want $\alpha_k \approx 1$ with thise Newton diredtion $\longrightarrow$ Superlinear or quadratic convergence.

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

2. Inexact Line Search
   Armijo-Goldstein[1966-67] conditions, (or Wolfe[1969] conditions)
   Choose $\alpha > 0$ so that
   $$f(x^{(k)} + \alpha d^{(k)}) \leqslant f(x^{(k)}) + c_1 * \alpha \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$
   sufficiently good rate for the decrease in the objective function and
   $$\left\langle \nabla f(x^{(k)} + \alpha d^{(k)}), d^{(k)} \right\rangle \geqslant c_2 \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$
   (step size should not be too small)
   where constants c_1, c_2 satisfy $0 < c_1 < c_2 < 1$

3. Strong Wolfe Conditions
   $$f(x^{(k)} + \alpha d^{(k)}) \leqslant f(x^{(k)}) + c_1 \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$
   and
   $$|\left\langle \nabla f(x^{(k)} + \alpha d^{(k)}), d^{(k)} \right\rangle| \leqslant c_2 |\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle|$$
   The second condition disallows $|\left\langle \nabla f(x^{(k)} + \alpha d^{(k)}), d^{(k)} \right\rangle|$ being too large and positive.

Lemma 2.11

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1, and $d \in \mathbb{R}^n$ be a descent direction at $\overline{x} \in \mathbb{R}^n$ for f. Suppose f is bounded from below on the ray $\{\overline{x} + \alpha d : \alpha \in \mathbb{R}_+\}$. Then $\forall 0 < c_1 < c_2 < 1$, $\exists$ step lengths $\alpha > 0$ satisfying Armijo-Goldstein-Wolfe as well as strong Wolfe conditions.
With $\phi(\alpha) := f(\overline{x} + \alpha d), 0 < c_1 < c_2 < 1$, choose $\alpha > 0$ such that

$$\textit{Armijo-Goldstein-Wolfe} = \begin{cases} \phi(\alpha) \leqslant \phi(0) + c_1 \alpha \phi'(0) \\ \phi'(\alpha) \geqslant c_2 \phi'(0) \end{cases}$$

$$\textit{Strong Wolfe} = \begin{cases} \phi(\alpha) \leqslant \phi(0) + c_1 \alpha \phi'(0) \\ |\phi'(\alpha)| \leqslant c_2 |\phi'(0)| \end{cases}$$

Proof. Suppose the stated assumptions hold. We adopt the above mentioned notation with ϕ. Then $\phi(\alpha)$ is bounded from below on $\{\alpha \in \mathbb{R} : \alpha \geqslant 0\}$. Since $c_1 \in (0, 1)$ and

$$\phi'(0) = \langle \nabla f(\overline{x}), d \rangle < 0 d \text{ is a descent direction for } f$$

the ray $\{\phi(0) + (c_1 \phi'(0))\alpha : \alpha \geqslant 0\}$ is unbounded below and therefore, intersects the graph of ϕ at least once for $\alpha > 0$. Let $\overline{\alpha} > 0$ denote the smallest value of α which the ray intersects the graph of ϕ. Then,

$$\phi(\overline{\alpha}) = \phi(0) + \overline{\alpha} c_1 \phi'(0) \ldots (*)$$

Thus, the first condition of A-G-W holds on $(0, \overline{\alpha}]$
By the mean value theorem, $\exists \hat{\alpha} \in (0, \overline{x})$ such that

$$\phi(\overline{\alpha}) - \phi(0) = \overline{\alpha}\phi'(\hat{\alpha})$$

Therefore,

$$\phi(\overline{\alpha}) - \phi(0) = \overline{\alpha}\phi'(\hat{\alpha}) = \overline{\alpha}c_1\phi'(0) > c_2\overline{\alpha}\phi'(0)$$

by $(*)$ and $c_2 > c_1 \& \phi'(0) > 0$
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 $\hat{\alpha}$ □

Read about Backtracking Line Search in the textbook.

29, Sept 2020

2.5 Convergence Properties of Descent Algorithms

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

- $\{x \in \mathbb{R}^n : f(x) \leqslant \beta\}$ 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 literature use contour of f)

Consider a descent algorithm:
Start with $x^{(0)} \in \mathbb{R}^n$, at each iteration k, choose $d^{(k)} \in \mathbb{R}^n$ s.t. $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle < 0$ and choose $\alpha_k > 0,\ x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}$
Recall the geometric factL

$$\forall u, v \in \mathbb{R}^n,\ \langle u, v \rangle = ||u||_2 ||v||_2 cos(\theta)$$

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

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

Theorem 2.12: Zoutendijk[1970], Wolfe[1969]

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

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

Asuume ∇f is Lipschitz continuous on N with Lipschitz constant $L \in \mathbb{R}_{++}$. Thenm every descent algorithm following Armijo-Goldstein-Wolfe consditions 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

$$\left\langle \nabla f(x^{(k+1)}), d^{(k)} \right\rangle \geqslant c_2 \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$

$$\implies \left\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), d^{(k)} \right\rangle \geqslant (c_2 - 1) \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle \ (eq.2)$$

Due to the fact that we are working with a descent algorithm ($\left\langle \nabla f(x^{(k+1)}), d^{(k)} \right\rangle < 0, \forall k$) and the first condition of A-G-W, $\{x^{(k)} \subset N\}$. Since ∇f is Lipschitz cont on N with Lipschitz constant L,

$$\begin{aligned}
&\left\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), d^{(k)} \right\rangle \\
&\leqslant ||\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})||_2 ||d^{(k)}||_2 \\
&\leqslant \alpha_k L ||d^{(k)}||_2^2 \ (eq.3) \\
&\implies \alpha_k \geqslant \frac{(c_2 - 1) \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle}{L ||d^{(k)}||_2^2}
\end{aligned}$$

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

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

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

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

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

Since f is bounded from below, $[f(x^{(0)}) - f(x^{(k)})]$ 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:

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

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

2.6 A General Conversation about Convergence

Ex:

$$f : \mathbb{R} \to \mathbb{R}, \; f(x) = \frac{1}{4}x^4 - 5x$$

Then f is convex, global minimizer is unique and attained at $\overline{x} = 3\sqrt{5}$ (irrational, even thought the data are integers)
$\implies$ We cannot expect finite algorithm in the worst case (machine has finite precisions)
We will generate a sequence $x^{(1)}, x^{(2)}, \ldots$ And we hope for

- $\forall x^{(0)}, x^{(k)} \to \overline{x}$ a global minimizer
- $\forall x^{(0)}, x^{(k)} \to \overline{x}$ a local minimizer
- $\forall x^{(0)}$ all limit points of $\{x^{(k)}\}$ are global (local) minimizers or $f(x^{(k)}) \to -\infty$
- $\forall x^{(0)}$ all limit points of $\{x^{(k)}\}$ satisfy second-order necessary conditions
- $\forall x^{(0)}$ all limit points of $\{x^{(k)}\}$ satisfy first-order necessary conditions
- $\forall x^{(0)}$, $\lim_{k\to\infty} ||\nabla f(x^{(k)})|| = 0$

Locally, replace "$\forall x^{(0)} \in \mathbb{R}^n$" by $x^{(0)} \in B(\overline{x}, \eta)$ and hope that second-order sufficient condition holds.

How fast does it converge? $\epsilon_k := ||x^{(k)} - \overline{x}||$
Ex:

- $\epsilon_k := (0.1)^k \to 10^{-1}, 10^{-2}, \ldots$ linear converge
- $\epsilon_k := (0.9)^k \to 0.9, 0.82, 0.0729, \ldots$ linear converge
- $\epsilon_k := (0.1)^{2^k} \to 10^{-2}, 10^{-4}, 10^{-8}, \ldots$ quadratic converge
- $\epsilon_k := (0.9)^{2^k} \to 0.81, 0.65, 0.43, 0.185, 0.034, \ldots$ quadratic converge

- $\epsilon_k := (0.1)^{3^k} \to \ldots$ cubic converge

Definition 2.13

If $\epsilon \searrow 0$, and $\epsilon_{k+1} \leqslant \beta\ (\epsilon_k)^p$ for some $p \geqslant 1$ and $\beta(\beta < 1$ if $p = 1)$, and for all sufficiently large k, then we say $\epsilon_k \to 0$ 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 superlinear.

- *Q-linear: Q-order 1*
- *Q-quadratic: Q-order 2*

e.g.:$\epsilon_k = (\frac{1}{k})^k, k \in \mathbb{Z}_{++}$, 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 $\eta_i := sup\{\epsilon_k : k \geqslant i\}$. Then

$$\lim_{k\to\infty} sup\{\epsilon_k\} := \lim_{i\to\infty}\{\eta_i\}$$

Definition 2.14: I

$\epsilon_k \searrow 0$, and $\lim_{k\to\infty} sup\{\epsilon_k^{\frac{1}{q^k}}\} < 1, \forall 0 < q < p,\ p > 1$, then $\epsilon_k \to 0$ with R-order (at least) p.

This is the same as

$$\lim_{k\to\infty} sup\{\frac{1}{q^k} log(\epsilon_k)\} < 0$$

Proposition 2.15

1. *If $x^{(k)} \to \overline{x}$ with Q-order p (R-order p) so does $\{x^{(k+l)}\}$ for all xied $l \in \mathbb{Z}_+$*
2. *If $\epsilon_k \searrow 0$ with Q-order p and $0 < \eta_k \leqslant \epsilon_k, \forall k \in \mathbb{Z}_{++}$ then $\eta_k \searrow 0$ with R-order p*

October 1, 2020

2.7 Fast Local Convergence of Newton's Method

This goes to Kantorovich. In addition to his funcdational work on the convergence theory of Newton's Method, Kantorovich also made significant contributions to functional analysis and operation theory.

Lemma 2.16: L

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

Proof. A, B as above, then

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

, and

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

If $C \in \mathbb{R}^{n\times n}$ nonsingular such that $||C||_2 \leqslant \frac{1}{3}$ then $(I-C)$ is invertible and $(I-C)^{-1} = I + C + C^2 + \ldots = \sum_{k=0}^{\infty} C^k$

$$\implies ||(I-C)^{-1}||_2 \leqslant \sum_{k=0}^{\infty}(\frac{1}{3})^k = \frac{1}{2/3} = \frac{3}{2}$$

Then, $C := A^{-1}(A-B)$, (then $B = A(I-C)$), B is invertible

$$B^{-1} = (I-C)^{-1}A^{-1}$$

and

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

□

Lemma 2.17: L

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

$$||g(y) - g(x) - \nabla g(x)(y-x)||_2 \leqslant \frac{L}{2}||y-x||_2^2, \forall x, y \in D$$

Proof. We already proved this as a part of THM19 □

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

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

Theorem 2.18

Let $f:\mathbb{R}^n\to\mathbb{R},\ f\in C^2,\ x^{(0)}\in\mathbb{R}^n,\ L\geqslant 1$.
Assme $\nabla f(\overline{x})=0,\ \nabla^2 f(\overline{x})$ is nonsingular, $\nabla^2 f\in Lip(L)$ in an open neighborhood of $\overline{x}$. Then exists an open neighborhood N_1 of $\overline{x}$ such that $\forall x^{(0)}\in N_1$, Newton's Method converges to $\overline{x}$ linearly and the method is locally Q-quadratically convergent (there exists an open neighborhood $N_2\subseteq N_1$ of $\overline{x}$ such that $\forall x^{(0)}\in N_2, ||x^{(k+1)}-\overline{x}||_2\leqslant$ constant $||x^{(k)}-\overline{x}||_2^2, \forall k\in\mathbb{Z}_+$)
Moreover, $||\nabla f(x^{(k)})||$ also converges to zero in N_1, locally Q-quadratically. ($\forall x^{(0)}\in N_2,\ ||\nabla f(x^{(k+1)})||_2\leqslant$ constant $||\nabla f(x^{(k)})||_2^2,\ \forall k\in\mathbb{Z}_+$)

Proof. Suppose assumptions hold.

$$\sigma=||[\nabla^2 f(\overline{x})]^{-1}||_2,\ \text{choose } \eta>0,\ \text{such that } \mathbb{B}\ :=B(\overline{x},\eta)=\{x\in\mathbb{R}^n:||x-\overline{x}||_2<\eta\}$$

$\nabla^2 f\in Lip(L)$ on $\mathbb{B}$ and $\eta\leqslant\dfrac{1}{3\sigma l}$

Then $\forall x\in\mathbb{B}$

$$||\nabla^2 f(x)-\nabla^2 f(\overline{x})||_2\leqslant L||x-\overline{x}||_2\leqslant L\eta\leqslant\frac{1}{3\sigma}\ldots\text{eq.4}$$

Therefore, by Lemma 41 (with $A:=\nabla^2 f(\overline{x}),\ B:=\nabla^2 f(x), x\in\mathbb{B}$), $\nabla^2 f(x)$ is nonsingular $\forall x\in\mathbb{B}$, thus, Newton's Method is well-defined for $\{x^{(k)}\}\subseteq B$
We prove by induction on k.
let $x^{(0)}\in\mathbb{B}$(in general, $(k)\in\mathbb{B}$), then

$$\begin{aligned}
||x^{(k+1)}-\overline{x}||_2 &= ||x^{(k)}-[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})-\overline{x}||_2\\
&= ||[\nabla^2 f(x^{(k)})]^{-1}[0-\nabla f(x^{(k)})-\nabla^2 f(x^{(k)})(\overline{x}-x^{(k)})]||_2\\
&\leqslant ||[\nabla^2 f(x^{(k)})]^{-1}||_2||[\nabla f(\overline{x})-\nabla f(x^{(k)})-\nabla^2 f(x^{(k)})(\overline{x}-x^{(k)})]||_2\\
&\leqslant \frac{3\gamma}{2}*\frac{L}{2}||x^{(k)}-\overline{x}||_2^2=\frac{3\gamma L}{4}||x^{(k)}-\overline{x}||_2^2 \text{ by Lemma 41,42}
\end{aligned}$$

Also, if $x^{(k)}\in\mathbb{B}$, then we know $||x^{(k)}-\overline{x}||_2^2\leqslant\frac{1}{3\sigma L}||x^{(k)}-\overline{x}||_2$

$$\begin{aligned}
||x^{(k+1)}-\overline{x}||_2 &\leqslant \frac{3\sigma L}{4}\frac{1}{3\sigma L}||x^{(k)}-\overline{x}||_2\\
&= \frac{1}{4}||x^{(k)}-\overline{x}||_2\leftarrow \text{ linear}
\end{aligned}$$

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

$$\begin{aligned}
||\nabla f(x^{(k+1)})||_2 &= ||\nabla f(x^{(k+1)})-\nabla f(x^{(k)})-\nabla^2 f(x^{(k)})d||_2 \ d \text{ is defined by Newton}\\
&\leqslant \frac{1}{2}||d||_2^2\\
&= \frac{L}{2}||[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})||_2^2 \text{ eq 5}\\
&\leqslant \frac{L}{2}||[\nabla^2 f(x^{(k)})]^{-1}||_2^2||\nabla f(x^{(k)})||_2^2\\
&\leqslant \frac{9\sigma^2 L}{8}||\nabla f(x^{(k)})||_2^2 \text{ by } x^{(k)}\in\mathbb{B}\text{, eq5 and lemma 41}
\end{aligned}$$

Now, we proved,

$$\forall x^{(0)} \in \mathbb{B}, ||x^{(1)} - \overline{x}||_2 \leqslant \frac{1}{4}||x^{(0)} - \overline{x}||_2$$
$$x^{(1)} \in \mathbb{B}, ||x^{(1)} - \overline{x}||_2 \leqslant \frac{3\sigma L}{4}||x^{(0)} - \overline{x}||_2^2$$
$$||\nabla f(x^{(1)})||_2 \leqslant \frac{9\sigma^2 L}{8}||\nabla f(x^{(0)})||_2^2$$

By induction on k, we estabilish the desired equations on $x^{(k)}$ from (eq.5)

$$\begin{aligned}
||\nabla f(x^{(k+1)})||_2 &\leqslant \frac{L}{2}||[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})||_2^2 \\
&= \frac{L}{2}||x^{(k+1)} - x^{(k)}||_2||[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})||_2 \\
&\leqslant \frac{L}{2} * \frac{2}{3\sigma L} * \frac{3\sigma}{2}||\nabla f(x^{(k)})||_2 \text{ by } x^{(k+1)}, x^{(k)} \in \mathbb{B}, \text{lemma 41} \\
&= \frac{1}{2}||\nabla f(x^{(k)})||_2
\end{aligned}$$

Therefore, $\forall x^{(0)} \in \mathbb{B}$

$$||x^{(k)} - \overline{x}||_2 \to 0 \text{ Q-linearly, and locally Q-quadratically}$$

$$||\nabla f(x^{(k)})||_2 \to 0 \text{ Q-linearly, and locally Q-quadratically}$$

□

Note this proof can be applied to nonlinear equations.
$g : \mathbb{R}^n \to \mathbb{R}^n,\ g \in C^1$ an open and convex set $D \subseteq \mathbb{R}^n$

$$\exists \overline{x} \in D,\ s/t\ g(\overline{x}) = 0,\ \nabla g(\overline{x}) \text{ is nonsingular}$$
$$\nabla g \in Lip(L) \text{ on } D$$
$$x^{(0)} \in N \subseteq D$$
$$x^{(k+1)} := x^{(k)} - [\nabla g(x^{(k)})]^{-1} g(x^{(k)}), \forall k \in \mathbb{Z}_{++}$$

2.7.1 Potential Problems with Newton's Method

1. Fails if $\nabla^2 f(x^{(k)})$ is singular (or very ill-conditioned)
2. $x^{(k+1)}$ is not the local minimizer of the local quadratic model $\tilde{f}$ for f if $\nabla^2 f(x^{(k)})$ is not postive definite ($\tilde{f}$ is an approxiamtetion model at $x^{(k)}$) using gradient and Hessian
3. Not globally convergent in general
4. May not even provide descent in general.

Possible Remedies:

- To address (1) and (2), modify $\nabla^2 f(x^{(k)})$, if necessary, to a "nearby" systemtric positive definite matrix B_k
- Together with the above remedy, use Armijos-Goldstein-Wolfe or strong Wolfe based line searches to address (3) and (4)
- Still some advantages
    1. evalueate Hessians at every iteration
    2. We must provide n-by-n linear systems of equations in every iterations

For some problems, evaluating the Hessian is very largely extra work compared to f, ∇f.
Also, in some cases Automake differential via a smalle number of $\nabla f()$ evaluataions suffice(chapter 8).

2.8 Quasi-Newton Method

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

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

Since $B_k \in S^n_{++}$, $\tilde{f}$ has a unique global minimizer at

$$\overline{d} = -B_k^{-1} \nabla f(x^{(k)})$$

Now we can do a line search and full $x^{(k+1)}$, then we have $f(x^{(k+1)})\&\nabla f(x^{(k+1)})$
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 $\nabla^2 f$

$$s^{(k)} := x^{(k+1)} - x^{(k)} \text{ (primnal step at iteration } k)$$
$$y^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \text{ (dual step at iteration } k)$$

Magical Soltion:BFGS :

$$B_{k+1} = B_k - \frac{1}{{s^{(k)}}^T B^{(k)} s^{(k)}} * B^{(k)} s^{(k)} {s^{(k)}}^T B^{(k)} + \frac{y^{(k)} {y^{(k)}}^T}{{y^{(k)}}^T s^{(k)}}$$

October 06,20 Note: By Theorem 19, we have

$$y^{(k)} = \left[\int_0^1 \nabla^2 f(x^{(k)} + \alpha s^{(k)}) \partial \alpha \right] s^{(k)}$$

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

$$y^{(k)} = B_{k+1} s^{(k)} \text{ secant equation}$$

Be enforcing this equation on B_{k+1}, we can incorporate new "secant" information about $\nabla^2 f$ If $B_{k+1} \succ 0$ satisfies the secant equation, then

$$\left\langle y^{(k)}, s^{(k)} \right\rangle = \left\langle B_{k+1} s^{(k)}, s^{(k)} \right\rangle > 0 \text{ since } s^{(k)} \neq 0, B_{k+1} \succ 0$$

Notice that $\left\langle y^{(k)}, s^{(k)} \right\rangle$ is positively proportional to:

$$\begin{aligned} \left\langle y^{(k)}, s^{(k)} \right\rangle &= \left\langle \nabla f(x^{(k+1)}), d^{(k)} \right\rangle - \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle \\ &= \phi'(\alpha_k) - \phi'(0) > 0 \text{ if we use A-G-W or strong Wolfe based line-search} \end{aligned}$$

The condition

$$\left\langle y^{(k)}, s^{(k)} \right\rangle > 0$$

is called the curvature condition.

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

Solve the optimization problem

$$(P_1)\ Min\ ||B-H||_F$$
$$s.t.\ Bs=y,\ B\in\mathbb{R}^{n\times n}$$

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

$$||A||_F:=\left(\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}^2\right)^{\frac{1}{2}}=[Tr(A^TA)]^{\frac{1}{2}}=[vec(A)^Tvec(A)]^{\frac{1}{2}}$$

where $||\dot{|}|_F$ represents Frobenius norm.
(P_1) always has a unique solution $\overline{B}$

$$Z:=\overline{B}-H$$

Note $\overline{B}s=y\Leftrightarrow\overline{B}s-Hs=y-Hs=:r$

With this change of variable and definitions, (P_1) is equivalent to

$$(P_2)\ Min\ ||Z||_F$$
$$s.t.\ Zs=r$$

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

$$(P_3)\ Min\ ||\tilde{z}||_F$$
$$\tilde{z}e_1=\frac{1}{\beta}r$$

$$\implies \tilde{z}=\begin{bmatrix}\frac{1}{\beta}r & 0 & 0 & \dots & 0\end{bmatrix}$$

Using our definitions, we compute:

$$z=\tilde{z}Q=\frac{1}{\beta}re_1^TQ=\frac{1}{\beta^2}rs^TQ^TQ=\frac{1}{\beta^2}rs^T$$
$$zs=r\implies\frac{1}{\beta^2}r(s^Ts)=r\implies s^Ts=\beta^2\text{ unless }r=0\text{, in which case }z=0$$

Therefore, the unique optimal solution of (P_1) is :

$$z=\frac{rs^T}{s^Ts}=\frac{1}{s^Ts}(y-Hs)s^T$$

Theorem 2.19: (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*

$$B := H + \frac{1}{s^T s}(y - Hs)s^T \leftarrow \textit{Good Broyden}$$

Setting $v := H^T y$ *and*

$$B := H + \frac{1}{v^T s}(y - Hs)v^T \leftarrow \textit{Bad Broyden}$$

leads to "Broyden's Second Method"

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

Consider

$$\begin{aligned} Min\ &||B-H||_F \\ &Bs = y \\ &B = B^T \\ &B \in \mathbb{R}^{n\times n} \end{aligned}$$

Theorem 2.20: Powell[1970]

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

$$B := H + \frac{1}{s^T s}\left[(y-Hs)s^T + s(y-Hs)^T - (y-Hs)^T sss^T\right]$$

In the above formula, 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 systemtric, positive definite, provided $H \succ 0$ and $y^T s > 0$.
We consider solving

$$\begin{aligned} (P_w)\ Min\ &||W^{\frac{1}{2}}(B-H)W^{\frac{1}{2}}||_F \\ s.t.&Bs = y \\ &B \in \mathbb{S}^n \end{aligned}$$

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

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

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

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

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

Note that

$$B^{-1} = H^{-1} - \frac{H^{-1}yy^T H^{-1}}{y^T H^{-1} y} + \frac{ss^T}{y^T s}$$

Next, consider

$$(P_w^{BFGS})\ Min\ ||W^{-\frac{1}{2}}(B-H)W^{-\frac{1}{2}}||_F$$
$$s.t.\ By = s$$
$$B \in \mathbb{S}^n$$

Theorem 2.21

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 thaty $Wy^{(k)} = s^{(k)}$, the unique solution of (P_W^{BFGS}) is

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

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

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

$$H - \frac{Hss^T H}{s^T Hs} + \frac{yy^T}{y^T s}$$

October 08,2020

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

with

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

DFP: Solve

$$\begin{aligned} Min \ &||W^{(\frac{1}{2})}(B-H)W^{(\frac{1}{2})}||_F \\ s.t. &Bs = y \\ &B \in \mathbb{S}^n \end{aligned}$$

BFGS: Solve

$$\begin{aligned} Min \ &||W^{-\frac{1}{2}}(B-H^{-1})W^{-\frac{1}{2}}|| \\ s.t. \ &By = s \\ &B \in \mathbb{S}^n \end{aligned}$$

then the inverse of the solution is the BFGS estimate of the Hessian $\nabla^2 f$
P and D are convex sets

$$u \in P \Leftrightarrow u^{-1} \in D$$

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

Broyden's convex class $\{\lambda B^{DFP} + (1-\lambda)B^{BFGS}\}$

2.8.1 Convergence Results

1. Global Convergence

   (a) Powell[1972]: If f is strictly convex (f is convex and $f(\lambda u + (1-\lambda)v) < \lambda f(u) + (1-\lambda)f(v), \forall \lambda \in (0,1)$ and $u \neq v$),

   $$\{x \in \mathbb{R}^n : f(x) \leqslant f(x^{(0)})\} \text{ is compact,}$$

   $f \in C^2$, and exact line search is used then Quasi-Newton methjod based on DFP converges.

   (b) Dixon[1972]: If exact line search is used then DFP, 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 (3) holds for all of Broyden's convex class, except DFP(i.e.$\lambda \in [0,1)$)

   (e) it seems that DFP is worse than BFGS in practice

2. Local Convergence: Assume

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

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

   (b) Broyden, Dennis, More[1973]: If we use $\alpha_k = 1,\ \forall k \in \mathbb{Z}_+$ and for suitably small $\epsilon > 0, \sigma > 0$, we have $||x^{(0)} - \overline{x}|| \leqslant \epsilon$ and $||B_0 - \nabla^2 f(\overline{x})|| \leqslant \sigma$ then

   $$x^{(k)} \to \overline{x} \text{ Q-superlinearly}$$

   (c) Powerll[1976]: Assumptions as in 1.(a), 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]:2.(c) applies to every update in Broyden's Convex Class, except DFP.

2.8.2 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)

which 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 $r \in \{10, 11, \ldots 23\}$
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:$H = LL^T$. We would like Choleski decomposition of B^{BFGS}

Lemma 2.22

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,

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

where $\beta := \sqrt{\frac{y^T s}{s^T Hs}}$

Proof. Just a computation □

So, B is written as

$$(L + uv^T)(L^T + vu^T) \leftarrow \text{ Not a Choleski decomposition}$$

However, we can recover a Choleski factorization $\overline{LL}^T$ of B as follows:

Remark. For every orthoganal 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, focus on vu^T

2.8.3 Using Givens' Rotations(James Wallace Givens' Jr[1958])

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

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

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & c & s \\ 0 & 0 & 0 & \dots & -s & c \end{bmatrix} \begin{bmatrix} \\ \\ \\ v \\ \\ \\ \end{bmatrix} = \begin{bmatrix} * \\ * \\ . \\ . \\ . \\ * \\ 0 \end{bmatrix}$$

Then,

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

Keep doing this, we find $Q_1 \in \mathbb{R}^{n \times n}$ orthoganal such that

$$Q_1(L^T + vu^t) = A + \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} u^T \end{bmatrix} = A'$$

where A and A' are Upper Hessenbergs
Nexty, we apply $(n-1)$ special orthogonal matrices (Givens' Rotations), to zero-out the nonozeroes below the diagonal.

$$\begin{bmatrix} c & s & 0 & \dots & 0 \\ -s & c & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

, then

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & c & s & 0 & \dots & 0 \\ 0 & -s & c & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$\rightarrow$ orthogonal matrix $Q_2 \in \mathbb{R}^{n\times n}$ such that

$$(Q_2Q_1)(L^T + vu^T) =: \overline{L}^T \leftarrow \text{ Choleski factor of } B$$

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

Oct,20,2020

2.9 Conjugate Gradient Methods

Let $C \subseteq \mathbb{R}^n$ be a convex set. Note that every C^2 function $f : C \to \mathbb{R}$ with $\nabla^2 f(x) \succ 0, \forall x \in C$, is strictly convex on C
On strictly convex quadratic functions $f : \mathbb{R}^n \to \mathbb{R}$ ($f(x) := \gamma + c^T x + \frac{1}{2} x^T H x$, 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: $f(x) := \gamma + c^T x + \frac{1}{2} x^T D x$. D is diagonal and positive definite. In this case, the problem

$$inf\ f(x),\ x \in \mathbb{R}^n$$

is separable.
Coordinate Descent solves this problem in n iterations.
Now, consider an arbitrary $H \in \mathbb{S}^n_{++}$ with $f(x) := \gamma + c^T x + \frac{1}{2} x^T H x$. 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 15, Spectral Decomposition Theorem).
Then upon defining $v := Q^T x$, we have

$$\begin{aligned} f(x) &= \gamma + c^T x + \frac{1}{2} x^T Q D Q^T x \\ &= \gamma + c^T Q v + \frac{1}{2} v^T D v \end{aligned}$$

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 2.23

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

$$u^T H v = 0$$

Observation 1

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

Lemma 2.24

Let $H \in \mathbb{S}^n_{++}$, suppose that $d^{(1)}, d^{(2)}, \ldots, d^{(k)} \in \mathbb{R}^n \setminus \{0\}$ are pairwise H-conjugate. Then $\{d^{(1)}, d^{(2)}, \ldots, d^{(k)}\}$ is linearly independent.

Proof. Let H and $d^{(1)}, d^{(2)}, \ldots, d^{(k)}$ be as in the statement of the lemma. Then,

$$H^{\frac{1}{2}} d^{(1)}, H^{\frac{1}{2}} d^{(2)}, \ldots, H^{\frac{1}{2}} d^{(k)} \in \mathbb{R}^n \setminus \{0\}$$

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

Theorem 2.25

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, $D := [d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}] \in \mathbb{R}^{n\times n}$. 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 thorem. Then, D is nonsingular, by Lemma 50. Moreover,

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

$(ij)^{th}$ entry of $D^T HD = \left\langle d^{(i)}, Hd^{(j)}\right\rangle = \begin{cases} 0, \ if\ i \neq j \\ > 0, \ if\ i = j \end{cases}$ Therefore, $\hat{f}$ is separable. □

Corollary 2.26

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 subspace

$$\{x^{(0)} + \sum_{i=0}^{j-1} \mu_i d^{(i)} : \mu_i \in \mathbb{R}\}, \ \forall j \in \{1, 2, \ldots, b\}$$

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

Proof. Follows from the last theorem. □

2.9.1 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 $d^{(k)}$)

$$\begin{aligned} \text{If } \nabla f(x^{(k)}) = 0, \text{ set } x^{(k+1)} &:= x^{(k)} \\ \text{Else } x^{(k+1)} &:= x^{(k)} + \alpha_k d^{(k)} \\ d^{(k+1)} &:= -\nabla f(x^{(k+1)}) + \beta_k d^{(k)} \\ \text{where } \beta_k &:= \frac{\left\langle \nabla f(x^{(k+1)}), Hd^{(k)} \right\rangle}{\left\langle d^{(k)}, Hd^{(k)} \right\rangle} \end{aligned}$$

note that the α_k above is obtained by exact line search.

Theorem 2.27

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 "$\nabla f(x^{(k+1)}) = 0$" $\rightarrow$ then, we are done!
- or "$d^{(0)}, d^{(1)}, \ldots, d^{(k+1)}$ are all nonozero and pairwise H-conjugate"$\rightarrow$ 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 $\{x^{(0)} + \sum_{i=0}^{k} \mu_i d^{(i)} : \mu \in \mathbb{R}^{k+1}\}$. Then,

$$\left\langle \nabla f(x^{(k+1)}), d^{(j)} \right\rangle = 0, \ \forall j \in \{0, 1, \ldots, k\}$$

Since $\left\langle \nabla f(x^{(k+1)}), d^{(k)} \right\rangle = 0, \ d^{(k+1)} = -\nabla f(x^{(k+1)}) + \beta_k d^{(k)} \neq 0.$
Next, we prove $\left\langle d^{(k+1)}.Hd^{(j)} \right\rangle = 0, \forall j \in \{0, 1, \ldots, k\}$
By definition of β_k,

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

Consider $d^{(j)}, \ j \in \{0, 1, \ldots, k-1\}$.

$$x^{(j+1)} = x^{(j)} + \alpha_j d^{(j)}$$

, and $\alpha_j > 0$ since

$$\left\langle \nabla f(x^{(j)}), d^{(j)} \right\rangle = \left\langle \nabla f(x^{(j)}), -\nabla f(x^{(j)}) + \beta_{j-1} d^{(j-1)} \right\rangle = -||\nabla f(x^{(j)})||_2^2 < 0$$

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

$$\begin{aligned} \text{Since } \nabla f(x^{(j+1)}) &\in \ span\{d^{(j)}, d^{(j+1)}\} \\ \text{and } \nabla f(x^{(j)}) &\in \ span\{d^{(j-1)}, d^{(j)}\} \\ \text{we have } Hd^{(j)} &\in \ span\{d^{(j-1)}, d^{(j)}, d^{(j+1)}\} \end{aligned}$$

Then,

$$\begin{aligned}\left\langle d^{(k+1)}, Hd^{(j)}\right\rangle &= \left\langle -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}, Hd^{(j)}\right\rangle \\ &= \left\langle -\nabla f(x^{(k+1)}), Hd^{(j)}\right\rangle = 0\end{aligned}$$

This finishes the inductive step. □

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

What if f is not quadratic?

2.9.2 Nonlinear Conjugate Gradient

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

$$\beta_k = \begin{cases} \frac{\left\langle x^{(k+1)}, y^{(k)} \right\rangle}{\left\langle d^{(k)}, y^{(k)} \right\rangle} \text{ Sorensen-Wolfe (SW), Hestenes-Stiefd} \\ \frac{\left\langle f(x^{(k+1)}), \nabla f(x^{(k+1)}) \right\rangle}{\left\langle f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle} \text{ Fletcher-Reeves} \\ \frac{\left\langle x^{(k+1)}, y^{(k)} \right\rangle}{\left\langle f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle} \text{Polak-Ribiere} \end{cases}$$

- 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

$$\begin{aligned} \beta_{k+1} &:= \left\langle y^{(k)} - 2\frac{||y^{(k)}||_2^2}{\langle d^{(k)}, y^{(k)} \rangle} d^{(k)}, \frac{\nabla f(x^{(k+1)})}{\langle d^{(k)}, y6(k) \rangle} \right\rangle \\ &= SW + 2\frac{||y^{(k)}||_2^2}{\langle d^{(k)}, y^{(k)} \rangle} \left\langle d^{(k)}, \nabla f(x^{(k+1)}) \right\rangle \end{aligned}$$

2.9.3 Preconditional Conjugate Gradient

Let L be lower triangular such taht

$$LL^T \approx \nabla^2 f(x^{(k)})$$

(e.g. "approximate" possibly "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])

Oct 22,2020

3 Constrained Optimization

3.1 Back to 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 C^1.

$$\begin{aligned}(P) = inf\ f(x)\\ s.t.\ g(x) \leqslant 0\\ h(x) = 0\end{aligned}$$

$S := \{x \in \mathbb{R}^n : g(x) \leqslant 0, h(x) = 0\}$.
For $\overline{x} \in S$, $J(\overline{x}) := \{i : g_i(\overline{x}) = 0\} \leftarrow$ active constraints at $\overline{x}$; tight constraints at $\overline{x}$ $J := J(\overline{x})$, then g_J is the corresponding "subfunction". $d \in \mathbb{R}^n$ is a feasible direction for (P) at $\overline{x}$, if $\exists \overline{\alpha} > 0$ such that $(\overline{x} + \alpha d) \in S,\ \forall \alpha \in [0, \overline{\alpha})$.

Lemma 3.1

If $d \in \mathbb{R}^n$ is a feasible direction for (P) at $\overline{x}$, then

$$\langle \nabla g_i(\overline{x}), d\rangle \leqslant 0,\ \forall i \in J \text{ and } h'(\overline{x})d = 0$$

Recall:
Corollary 24: Assume h and $\overline{x}$ are as described in Theorem 21. Let $d \in \mathbb{R}^n$ such that $h'(\overline{x})d = 0$. Then there exists $\overline{\lambda} > 0$ and a C^1 arc(directed curve) $\hat{t}$ with the properties

- $\hat{t}(0) = \overline{x}$
- $h(\hat{t}(\lambda)) = 0,\ \forall \in [0, \overline{\lambda})$
- $\hat{t}'(0) = d$

Lemma 3.2

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

Proof. Assignment 4
Sketch:Apply Corollary 24, to determine $\overline{\alpha} > 0$(and to prove its existence) note that $\forall i \in \{1, 2, \ldots, m\} \setminus J,\ g_i(\overline{x}) < 0$(by definition of $J = J(\overline{x})$) □

Corollary 3.3

If $\overline{x} \in S$ *is a local min of* (P) *and* $h'(\overline{x})$ *has rank* p*, then* $\nexists d \in \mathbb{R}^n$ *satisfying*
$$\begin{cases} \langle \nabla g_i(\overline{x}), d\rangle < 0, \forall i \in J(\overline{x}) \\ h'(\overline{x})d = 0 \\ \langle \nabla f(\overline{x}), d\rangle < 0 \end{cases}$$

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

Lemma 3.4: 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:*

1. $A^T d < 0,\ B^T d = 0$
2. $Au + Bv = 0,\ u \geqslant 0, u \neq 0$

Proof. Suppose (1) has a solution $\overline{d} \in \mathbb{R}^n$ and (2) has a solution $(\overline{u}, \overline{v})$ $\overline{u} \in \mathbb{R}^q,\ \overline{v} \in \mathbb{R}^r$. Then,
$$0 = A\overline{u} + B\overline{v} \implies 0 = \underbrace{\overline{d}^T A\overline{u}}_{<0} + \underbrace{\overline{d}^T B\overline{v}}_{0} < 0$$
a contradiction.
Suppose (2) does not have a solution. Consider for the LP
$$\begin{aligned} (LP)\ Max\ & \mathbb{1}^T u \\ & Au + Bv = 0 \\ & u \geqslant 0 \end{aligned}$$

$$\begin{aligned} (LD)\ Min\ & 0^T d \\ & A^T d \geqslant \mathbb{1} \\ & B^T d = 0 \end{aligned}$$

(LD) is equivalent to $Min\{0^T d : A^T d \leqslant -\mathbb{1}, B^T d = 0\}$ Since (2) has no solution and $\overline{u} := 0, \overline{v} := 0$ 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 $\overline{d}$. Therefore, system (1) has a solution. □

Where we used

Theorem 3.5: Strong Duality Theorem of Linear Programming

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

Theorem 3.6: Karushp[1939],FritzJohn[1948]

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

$$\begin{cases} \overline{\lambda}\nabla f(\overline{x}) + \sum_{i=1}^m \overline{u}_i \nabla g_i(\overline{x}) + \sum_{i=1}^p \overline{v}_i \nabla h_i(\overline{x}) = 0 \\ \sum_{i=1}^m \overline{u}_i g_i(\overline{x}) = 0 \end{cases}$$

Consider the second condition. Since $\overline{u} \geqslant 0$ and $g(\overline{x}) \leqslant 0$, this condition is equivalent to $\forall i \in \{1, 2, \ldots, m\}$, either $g_i(\overline{x}) = 0$ or $\overline{u}_i = 0$ (possibly both). (Complementary Slackness Conditions, or Complementarity Conditions)

Proof. Suppose $\overline{x} \in S$ is a local minimizer for (P). If $h'(\overline{x})$ does not have rank p, then $\exists \overline{v} \in \mathbb{R}^p \setminus \{0\}$ such that $\overline{v}^T h'(\overline{x}) = 0^T$. So, we may set $\overline{\lambda} := 0$ and $\overline{u} := 0$, and we are done.
Otherwise ($rank(h'(\overline{x})) = p$), by corollary 56, the system

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

has no solution.
Thus, by Lemma57, $\exists \overline{\lambda} \in \mathbb{R}_+, \overline{u}_J \geqslant 0, \overline{v} \in \mathbb{R}^p$ such that $\begin{pmatrix} \overline{\lambda} \\ \overline{u}_J \end{pmatrix} \neq 0$ and $\overline{\lambda}\nabla f(\overline{x}) + \sum_{i \in J(\overline{x})} \overline{u}_i \nabla g_i(\overline{x}) + \sum_{i=1}^p \overline{v}_i \nabla h_i(\overline{x}) = 0$ □

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

Example 3.7

check the handwritten notes

Example 3.8

check the handwritten notes

Oct 27,2020

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

3.1.1 The First-order Constraint Qualification (at $\overline{x} \in S$)

Let

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

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

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

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

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

In example 61, $D(\overline{x}) = \mathbb{R}^2$. For $d := \begin{pmatrix} -1 \\ -1 \end{pmatrix} \in D(\overline{x})$ the CQ cannot be satisfied.

Lemma 3.9: Mangasarian-Fromiwitz CQ[1967]

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

$$\begin{cases} \left\langle \nabla g_i(\overline{x}), \overline{d} \right\rangle < 0, \forall i \in J(\overline{x}) \\ \left\langle \nabla h_i(\overline{x}), \overline{d} \right\rangle = 0, \forall i \in \{1, 2, \dots, p\} \end{cases}$$

then the First-order CQ holds at $\overline{x}$

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

Corollary 3.10

Let $\overline{x} \in S,\ h, g \in C^1$. If $\begin{pmatrix} g'_J(\overline{x}) \\ h'(\overline{x}) \end{pmatrix}$ has linearly independent rows, then the First-order CQ holds at $\overline{x}$

Proof. Suppose $\overline{x} \in S$, and $g, h \in C^1$. If $\exists\, \overline{d} \in \mathbb{R}^n$ satisfying

$$\begin{cases} g'_J(\overline{x})d < 0 \\ h'(\overline{x})d = 0 \end{cases}$$

then we are done by Lemma 62. Otherwise, by Lemma 57, $\begin{pmatrix} g'_J(\overline{x}) \\ h'(\overline{x}) \end{pmatrix}$ has linearly dependent rows □

Corollary 3.11

If all constraints in (P) are linear (i.e. all functions $g_1, g_2, \ldots, g_m, h_1, h_2, \ldots, h_p$ are affine) then the First-order CQ holds at every $x \in S$.

Proof. Suppose the assumptions hold. Let $\overline{x} \in S$. For every $d \in D(\overline{x})$, set

$$\begin{cases} d^{(k)} := d \\ \hat{t}^{(k)}(\alpha) := \overline{x} + \alpha d \end{cases} \quad \forall k \in \mathbb{Z}_{++}$$

□

Lemma 3.12

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. $A^T d \leqslant 0, B^T d = 0, c^T d < 0$
2. $c + Au + Bv = 0,\ u \geqslant 0$

Proof. Assignment 4 □

Theorem 3.13: First-order Necessary Conditions under CQ

[Karush 1939, Kuhn-Tucker 1951(KKT theorem)]
Suppose $f, g, h \in C^1$ and the First-order CQ holds at $\overline{x} \in S$, a local minimizer for (P). Then, $\exists \begin{bmatrix} \overline{u} \\ \overline{v} \end{bmatrix} \in \mathbb{R}^m \oplus \mathbb{R}^p$ such that

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

Proof. Suppose the assumptions hold. Further suppose that $\exists\, d \in D(\overline{x})$ such that $\langle \nabla f(\overline{x}), d\rangle < 0$. Then by First-order CQ, we can find $d^{(k)} \in \mathbb{R}^n$ such that $\left\langle \nabla f(\overline{x}), d^{(k)} \right\rangle < 0$ and $d^{(k)}$ is the first derivative of a feasible C^1 arc $\hat{t}$ starting at $\overline{x}$. Defining

$$\phi(\alpha) := f(\hat{t}(\alpha)) \text{ leads to } \phi'(0) = \left\langle \nabla f(\overline{x}), d^{(k)} \right\rangle < 0$$

This leads to a contradiction to $\overline{x}$ being a local min. for (P). So, now we may assume the system

$$\begin{cases} \langle \nabla f(\overline{x}), d \rangle < 0 \\ g'_J(\overline{x})d \leqslant 0 \\ h'(\overline{x})d = 0 \end{cases}$$

has no solution.
By Lemma 65, $\exists \overline{u}_J \geqslant 0, \overline{v} \in \mathbb{R}^p$ such that

$$\nabla f(\overline{x}) + [g'_J(\overline{x})]^T \overline{u}_J + [h'(\overline{x})]^T \overline{v} = 0$$

Setting $\overline{u}_i := 0, \forall i \in \{1, 2, \ldots, m\} \setminus J(\overline{x})$ yields the desired conclusion. □

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

$$\begin{cases} \left.\begin{matrix} g(x) \leqslant 0 \\ h(x) = 0 \end{matrix}\right\} \text{ Primal feasibility} \\ \left.\begin{matrix} \nabla f(x) + [g'(x)]^T u + [h'(x)]^T v = 0 \\ u \geqslant 0 \end{matrix}\right\} \text{ Dual feasiblity} \\ u^T g(x) = 0 \} \text{ Complementary Slackness} \end{cases}$$

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

Oct 29,2020

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

$$l(x,u,v) := f(x) + u^T g(x) + v^T h(x)$$

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

KKT Conditions can equivalently be stated as:

$$\begin{cases}\nabla_x l(\overline{x},\overline{u},\overline{v}) = 0 \\ \nabla_u l(\overline{x},\overline{u},\overline{v}) \leqslant 0 \\ \nabla_v l(\overline{x},\overline{u},\overline{v}) = 0 \\ \overline{u}^T \nabla_u l(\overline{x},\overline{u},\overline{v}) = 0 \\ \overline{u} \geqslant 0\end{cases}$$

Where $\overline{x}$ satisfies First-order conditions for it to be a local minimizer of $l(\cdot,\overline{u},\overline{v})$ over $\mathbb{R}^n$;
$\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}$ satisfies First-order conditions for it to be a local maximizer of $l(\overline{x},\cdot,\cdot)$ over $\mathbb{R}^m \oplus \mathbb{R}^p$

Therefore, $\begin{pmatrix}\overline{x}\\ \overline{u}\\ \overline{v}\end{pmatrix}$ satisfies the First-order conditions for it to be a saddle point of the Lagrangian.

Example 3.14

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

$$(LP) \quad \begin{aligned} inf\ & f(x) := c^T x \\ & g(x) := -x \leqslant 0 \\ & h(x) := b - Ax = 0\end{aligned}$$

$$(LD) \quad \begin{aligned} sup\ & b^T v \\ & A^T v \leqslant c\end{aligned}$$

Note:

$$\left\{\begin{aligned} & c + (-I)u + (-A^T)v = 0 \\ & u \geqslant 0,\ u^T x = 0\end{aligned}\right\} \iff \left\{\begin{aligned} & A^T v \leqslant c \\ & x^T(c - A^T v) = 0,\ c^T x = b^T v \text{ using } Ax = b\end{aligned}\right\}$$

3.1.2 Second-order Conditions for Constrained Optimization

$$\begin{aligned}(P)\ inf\ & f(x) \\ & g(x) \leqslant 0, \text{ Assume } f, g, h \in C^2 \\ & h(x) = 0\end{aligned}$$

Example 3.15

$$inf\ f(x) := \frac{1}{2}{x_1}^2 - \frac{1}{2}{x_2}^2$$
$$g_1(x) := x_2 - 1 \leqslant 0$$
$$g_2(x) := -x_2 \leqslant 0$$

$\overline{x}$ is the unique minimizer of (P). $J(\overline{x}) = \{1\}$
$\nabla f(\overline{x}) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $\nabla g_1(\overline{x}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. *KKT conditions hold at $\overline{x}$ with $\overline{u} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,* $\nabla^2 f(\overline{x}) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, $\nabla^2 g_1(\overline{x}) = \nabla^2 g_2(\overline{x}) = 0$
$\nabla^2 f(\overline{x})$ is not positive semidefinite. However, it is positive semidefinite in the approximate linear subspace $\{d : g_J'(\overline{x})d = 0, h'(\overline{x})d = 0\}$ (tangent ($d_2 = 2$) to the active constraints ($x_2 = 1$))

Example 3.16

$$inf\ f(x) := -\frac{1}{2}(x_1 + 1)^2 - \frac{1}{2}{x_2}^2$$
$$g_1(x) := \frac{1}{2}{x_1}^2 + \frac{1}{2}{x_2}^2 - \frac{1}{2} \leqslant 0$$

$\overline{x}$ is the unique optimal solution.
$\nabla f(\overline{x}) = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$, $\nabla g_1(\overline{x}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. *KKT conditions are satisfied at $\overline{x}$ with $\overline{u} := 2$*
$\nabla^2 f(\overline{x}) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$, $\nabla^2 g_1(\overline{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. *$\nabla^2 f(\overline{x})$ is not positive semidefinite; but,* $\nabla^2_{xx} l(\overline{x}, \overline{u}) = \nabla^2 f(\overline{x}) + 2\nabla^2 g_1(\overline{x}) = -I + 2I = I$ *is positive semidefinite.*

Second-order CQ(at $\overline{x} \in S$)hold if

$$\left\{\begin{matrix} g_J'(\overline{x})d = 0 \\ h'(\overline{x})d = 0 \end{matrix}\right\} \implies \exists \overline{\alpha} > 0 \text{ and a } C^2 \text{ arc } \hat{t} : [0, \overline{\alpha}) \to \mathbb{R}^n \text{ such that}$$

$$\left\{\begin{matrix} \hat{t}(0) = \overline{x} \\ \hat{t}'(0) = d \\ g_J(\hat{t}(\alpha)) = 0 \\ h(\hat{t}(\alpha)) = 0 \end{matrix}\right\}, \ \forall \alpha \in [0, \overline{\alpha})$$

Theorem 3.17: Second-Order necessary conditions

Suppose $\overline{x} \in S$ is a local minimzier for (P) and second-order CQ holds at $\overline{x}$.
Then, if $\begin{pmatrix} \overline{x} \\ \overline{u} \\ \overline{v} \end{pmatrix}$ is a KKT triple, we have

$$\left\{\begin{matrix} g_J'(\overline{x})d = 0 \\ h'(\overline{x})d = 0 \end{matrix}\right\} \implies d^T[\nabla^2_{xx} l(\overline{x}, \overline{u}, \overline{v})]d \geqslant 0$$

Corollary 3.18

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

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

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

Theorem 3.19

Suppose $g, h \in C^2, \overline{x} \in S$.
If $\begin{bmatrix} g_J'(\overline{x}) \\ h'(\overline{x}) \end{bmatrix}$ has linearly independent rows, then the First-order as well as Second-order CQs hold at $\overline{x}$ (Use the Implicit Function Theorem (Theorem 21)).

November 3, 2020

Theorem 3.20: Second Order Sufficiency Condition

Suppose $\begin{pmatrix}\overline{x}\\ \overline{u}\\ \overline{v}\end{pmatrix}$ *is a KKT triple for* (P) *and*

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

Then, $\overline{x}$ *is a strict local minimzier of* (P)

Strict Complementarity

Let $\begin{pmatrix}\overline{x}\\ \overline{u}\\ \overline{v}\end{pmatrix}$ be a KKT triple for (P). We say that $\begin{pmatrix}\overline{x}\\ \overline{u}\\ \overline{v}\end{pmatrix}$ satisfies strict complementarity (or, equivalently $\overline{x}$ and $\overline{u}$ are strictly complementary) if for every $i \in \{1, 2, \ldots, n\}$ exactly one of the following holds $\begin{cases} g_i(\overline{x}) = 0\\ \overline{u}_i = 0 \end{cases}$

Recall:Since we have a KKT triple, we already have $\forall i \in \{1, 2, \ldots, m\}$ at least one of $g_i(\overline{x})$, $\overline{u}_i$ is zero.

When the KKT triple satisfies strict complementarity the statement of the last theorem and its proof simplify.

Theorem 3.21: 2nd-Order Suff. Condition when Strict Complementary

Suppose $\begin{pmatrix}\overline{x}\\ \overline{u}\\ \overline{v}\end{pmatrix}$ *is strictly complementary KKT triple for* (P) *and*

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

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

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

Theorem 3.22

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

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

Then, $\exists \overline{\rho} \geqslant 0$ *such that*

$$\forall \rho \geqslant \overline{\rho}, (B + \rho A A^T) \in \mathbb{S}^n_{++}$$

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

3.1.3 Augmented Lagrangians

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

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

Theorem 3.23

Suppose $\begin{pmatrix} \overline{x} \\ \overline{u} \\ \overline{v} \end{pmatrix}$ *satisfies the second-order sufficiency conditions for being a strict local minimzier for* (P)*. Suppose strict complementarity holds at* $\begin{pmatrix} \overline{x} \\ \overline{u} \\ \overline{v} \end{pmatrix}$*. Then,* $\exists \overline{\rho} \geqslant 0$ *and* $\overline{\sigma} \geqslant 0$ *such that* $\forall \rho \geqslant \overline{\rho}, \sigma \geqslant \overline{\sigma}$*,* $\overline{x}$ *is a strict local minimizer of* $l_{\rho,\sigma}(\cdot, \overline{u}, \overline{v})$*. Furthermore,* $\begin{pmatrix} \overline{u} \\ \overline{v} \end{pmatrix}$ *is a global maximizer of* $l_{\rho,\sigma}(\overline{x}, \cdot, \cdot)$

November 5, 2020

3.1.4 Algorithm 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 $\xi_i, i \in \{1, 2, , \ldots, m\}$

$$\begin{aligned} &inf\ f(x) \\ &s.t.\ g_i(k) + \xi_i^2 = 0,\ i \in \{1, 2, \ldots, m\} \\ &\quad h_i(x) = 0,\ i \in \{1, 2, \ldots, p\} \end{aligned}$$

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

Let

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

where, for $w \in \mathbb{R}^m$, $[w]_+ \in \mathbb{R}^m$ is defined by for each $j \in \{1, 2, \ldots, m\}$, $max\{0, w_j\}$.
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:

$$\begin{aligned} x^{(k+1)} &:= \underset{x \in R^n}{argmin}\{L_{\rho_k}(x, u^{(k)}, v^{(k)})\} \\ u^{(k+1)} &:= [u^{(k)} + \rho_k g(x^{(k+1)})]_+ \\ v^{(k+1)} &:= v^{(k)} + \rho_k h(x^{(k+1)}) \end{aligned}$$

Update ρ_k to ρ_{k+1}

How do we choose ρ_k

- Preset strategy (e.g. $\rho_k := \beta^k$, where $\beta > 1$ constant)
- Adaptive (if $g(x^{(k)})$ is "approx. $\leqslant 0$" 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 3.24: Bertsekas[1982]

$(P)\{inf\ f(x); g(x) \leqslant 0\}$. Suppose $\overline{x} \in \mathbb{R}^n$ is a local minmizer for (P); $f, g \in C^2$ and $\nabla^2 f, \nabla^2 g_i (i \in \{1, 2, \dots, m\}) \in Lip$ in a neighborhood of $\overline{x}$. Further assume second-order sufficiency conditions hold at $\overline{x}$ with Lagrange multipliers $\overline{u} \geqslant 0, \nabla g_J(\overline{x})$ has full column rank, strict complementarity holds at $\begin{pmatrix} \overline{x} \\ \overline{u} \end{pmatrix}$.

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

$$||x(u,\rho) - \overline{x}|| \leqslant \frac{M}{\rho}||u - \overline{u}||,$$
$$||[u + \rho g(x(u,\rho))]_+ - \overline{u}|| \leqslant \frac{M}{\rho}||u - \overline{u}||$$

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.

3.1.5 Method of Multipliers

Given $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$, consider

$$(P): \begin{matrix} inf\ f(x) \\ Ax = b \end{matrix}$$

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

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

$$\left\{ \begin{array}{l} x^{(k+1)} := \underset{x \in R^n}{argmin}\ l_{\rho_k}(x, v^{(k)}) \\ v^{(k+1)} := v^{(k)} + \rho_k(Ax^{(k+1)} - b) \\ \text{Update } \rho_k \text{ to } \rho_{k+1} \end{array} \right\}$$

Suppose f is C^1. Then, KKT conditions:

$$Ax = b \leftarrow \text{primal Feasibility}$$
$$\nabla f(x) + A^T v = 0 \leftarrow \text{dual feasibility}$$

$$x^{(k+1)} = \underset{x \in R^n}{argmin}\ l_{\rho_k}(x, v^{(k)})$$

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

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

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

3.1.6 Alternating Direction Method of Multiplier(ADMM)

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

$$(P): \begin{array}{l} inf\ f_1(x) + f_2(\xi) \\ \quad A_1 x + A_2 \xi = b \end{array}$$

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

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

$$\begin{cases} x^{(k+1)} := \underset{x \in R^{n_1}}{argmin}\ l_{\rho_k}\left(\begin{bmatrix} x \\ \xi^{(k)} \end{bmatrix}, v^{(k)}\right) \\ \xi^{(k+1)} := \underset{\xi \in R^{n_2}}{argmin}\ l_{\rho_k}\left(\begin{bmatrix} x^{(k+1)} \\ \xi \end{bmatrix}, v^{(k)}\right) \\ v^{(k+1)} := v^{(k)} + \rho_k(A_1 x^{(k+1)} + A_2 \xi^{(k+1)} - b) \\ \text{Update } \rho_k \text{ to } \rho_{k+1} \end{cases}$$

Nov 10,2020

In our illustrations 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:

$$f(x) = \sum_{l=1}^{L} f_l(x_l)$$

which separate into $L \geqslant 2$ subfunctions.

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

3.2 Projection and Different Methods

3.2.1 Proximal Operator

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

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

is closed and convex.
Proximal operator of f is $prox_f : \mathbb{R}^n \to \mathbb{R}^n$,

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

Consider the continuous optimization problem:

$$\begin{cases} inf\ f(x) \\ s.t.\ x \in S \end{cases}$$

, 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:

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

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 unconstrained convex optimization problem

$$(\tilde{P})\ \underset{x \in R^n}{inf}\ \tilde{f}(x)$$

We do not even need $\tilde{f}$ to be C^1.
$h \in \mathbb{R}^n$ is a subgradient of $\tilde{f}$ at $\overline{x} \in \mathbb{R}^n$ if

$$\tilde{f}(x) \geqslant \tilde{f}(\overline{x}) + h^T(x - \overline{x}), \forall x \in \mathbb{R}^n$$

$$\underbrace{\partial \tilde{f}(\overline{x})}_{\text{subdifferential of } \tilde{f} \text{ at } \overline{x}} := \{h \in \mathbb{R}^n : h \text{ is a subgradient of } \tilde{f} \text{ at } \overline{x}\}$$

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

$$0 \in \partial \tilde{f}(\tilde{x})$$

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

$$\begin{cases} x^{(k+1)} := prox_{\lambda f}(x^{(k)}) \\ k := k + 1 \end{cases}$$

In fact, $prox_{\lambda f}(\cdot) = \underbrace{(I + \lambda \partial f)^{-1}}_{\text{Resolvent operator}}(\cdot)$. The interpretation of Resolvent operator connects proximal point algorithms to Fixed Point Theory (More on this in CO463/663).

3.2.2 Closest Points and Projections

Theorem 3.25: Kolmogorov Criteria

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

$$(z - proj(z|S))^T(x - proj(z|S)) \leqslant 0, \forall x \in S$$

Proof. See the proof of Corallary 111 in co255 Lecture notes. □

$proj(z|S)$ is the unique optimal solution of

$$inf\{||x - z||_2^2 \colon x \in S\}$$

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

$$\underbrace{S^*}_{\text{dual cone of } S} := \{s \in \mathbb{R}^n : x^T s \geqslant 0, \forall x \in S\}$$

Theorem 3.26: Moreau Decomposition

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

In the above, $\overline{y} = proj(-z|S^*)$
Therefore, $\forall z \in \mathbb{R}^n$ can be expressed as

$$z = proj(z|S) - proj(-z|S^*)$$

Recall, $\forall z \in \mathbb{R}^n$,

$$z = [z]_+ - [-z]_+$$

3.2.3 A Stochastic Descent Algorithm

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

$$Ax \leqslant b \Leftrightarrow \langle a_i, x\rangle \leqslant b_i, \forall i \in \{1, 2, \ldots, m\}$$

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

$$\begin{cases} \text{Choose } i \in \{1, 2, \ldots, m\} \text{ uniformly, randomly} \\ x^{(k+1)} := \text{ closest point in } \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\} \text{ to } x^{(k)} \\ k := k + 1 \end{cases}$$

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

Nov 12,2020 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 half space $\{x \in \mathbb{R}^n : \langle a_i, x \rangle \leqslant b_i\}$; 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 $||a_i||_2 = 1, \forall i$.

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

Theorem 3.27: Hoffman[1952]

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

$$\underset{x: Ax \leqslant b}{Min} ||x - \tilde{x}||_2 \leqslant L_A ||[A\tilde{x} - b]_+||_2$$

i.e.

$$dist(\tilde{x}, S) \leqslant L_A * dist(b - A\tilde{x}, \mathbb{R}^m_+)$$

, 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 problem exist.

Theorem 3.28: Leventhal-Lewis[2010]

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

$$E\left[(dist(x^{(k+1)}, S))^2 | x^{(k)}\right]) \leqslant (1 - \frac{1}{m * L_A^2})(dist(x^{(k)}, S))^2$$

Proof. Suppose $S \neq \phi$, let $k \in \mathbb{Z}_+, i \in \{1, 2, \ldots, m\}$. Note

$$\left[dist(x^{(k+1)}, S)\right]^2 = ||x^{(k+1)} - proj(x^{(k+1)}|S)||_2^2$$

, and

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

Thus,

$$\begin{aligned}
\left[dist(x^{(k+1)}, S)\right]^2 &\leqslant ||x^{(k+1)} - proj(x^{(k)|S)})||_2^2 \\
&= ||x^{(k)} - [\langle a_i, x^{(k)} \rangle - b_i]_+ a_i - proj(x^{(k)}|S)||_2^2 \\
\left[dist(x^{(k+1)}, S)\right]^2 &\leqslant ||x^{(k)} - proj(x^{(k)}|S))||_2^2 + [\langle a_i, x^{(k)} \rangle - b_i]_+^2 \\
&\quad - 2[\langle a_i, x^{(k)} \rangle - b_i]_+ \langle a_i, x^{(k)} - proj(x^{(k)}|S) \rangle \\
&\leqslant [dist(x^{(k),S})]^2 - [\langle a_i, x^{(k)} \rangle - b_i]_+^2 \\
\underbrace{\Longrightarrow}_{\text{Taking expectation over all } i \in \{1, 2, \ldots, m\}} & \quad E\left[(dist(x^{(k+1)}, S))^2 | x^{(k)}\right] \leqslant [dist(x^{(k)}|S)]^2 - \frac{1}{m} ||[Ax^{(k)} - b]_+||_2^2
\end{aligned}$$

Note: $\langle a_i, x^{(k)} - proj(x^{(k)}|S) \rangle = \langle a_i, x^{(k)} \rangle - b_i - \left(\langle a_i, proj(x^{(k)}|S) \rangle - b_i\right)$
Now, we apply Theorem 80 to the second term in the RHS to get

$$-\frac{1}{m} ||[Ax^{(k)} - b]_+||_2^2 \leqslant -\frac{1}{m * L_A^2} dist(x^{(k)}|S)^2$$

Therefore,

$$E\left[(dist(x^{(k+1)},S))^2|x^{(k)}\right] \leqslant (1-\frac{1}{m*L_A^2})(dist(x^{(k),S}))^2$$

□

The underlying algorithm has its roots in the algorithm of Kaczmarz from 1930's (for solving systems of linear equations).
We doscissed 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 $\inf_{x\in R^n} f(x) := \sum_{i=1}^m f_i(x)$.
In (SGD) we randomly choose $i \in \{1,2,\ldots,m\}$,

$$x^{(k+1)} := x^{(k)} - \alpha_k \nabla f_i(x^{(k)})$$

Note that in our Randomized Kaczmarz Algorithm we used the probability distribution: $p_i = \frac{1}{m}\forall i$.
If we hadn't normalized $||a_i||_2 = 1, \forall i$, we should 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 separability)
- Easy to code, easy to modify
- Easy to analyze
- Can try to strengthen by utilizing second-order info.

Nov 17, 2020

3.2.4 Sequential Quadratic Programming(SQP)

$$(P) = \begin{array}{r} inf\ f(x) \\ s.t.\ g(x) \leqslant 0 \\ h(x) = 0 \end{array}$$

Given current iterate $x^{(k)}$ (not necessary feasible) and estimates $u^{(k)}$, $v^{(k)}$ of Lagrange multipliers (dual variables), and $B_k \approx \nabla^2_{xx} l(x^{(u)}, u^{(k)}, v^{(k)})$, construct an approxiamting

$$(QP)_k \begin{array}{r} inf\ f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), d \right\rangle + \frac{1}{2} d^T B_k d \\ g(x^{(k)}) + \nabla g(x^{(k)})^T d \leqslant 0 \\ h(x^{(k)}) + \nabla h(x^{(k)})^T d = 0 \\ d \in \mathbb{R}^n \end{array}$$

Start with $x^{(0)}$, $u^{(0)}$, $v^{(0)}$, B_0.
At iteration k, solve $(QP)_k$ for $d \in \mathbb{R}^n$ to determine the search direction, or the step.
How do we update $u^{(k)}$, $v^{(k)}$, B_k?
How do we make sure we make progress towards satisfying all the constraints?
We can merge many ideas here to design SQP based algorithms.
Let $\overline{d} \in \mathbb{R}^n$ be an optimal solution of $(QP)_k$. We may update by $x^{(k+1)} := x^{(k)} + \overline{d}$ or $x^{(k+1)} := x^{(k)} + \alpha \overline{d}$ and determine α by a line search using a "merit function" or a "potential function". E.g.

$$\phi_\mu(x) := f(x) + \mu ||[g(x)]_+|| + \mu ||h(x)||$$

for $\mu > 0$; or we may use a Trust-Region approach.
We may update B_k using a Quasi-Newton type approach, where

$$y^{(k)} := \nabla_x l(x^{(k+1)}, u^{(k+1)}, v^{(k+1)}) - \nabla_x l(x^{(k)}, u^{(k+1)}, v^{(k+1)})$$

We may update $u^{(k)} \to u^{(k+1)}$, $v^{(k)} \to v^{(k+1)}$ as in the Augmented Lagrangian based algorithms (or in some other way which still takes into consideration individual entries of $g(x^{(k+1)})$ and $h(x^{k+1})$)

3.2.5 Penalty and Barrier Methods, Modern Interio-Point Methods

In many of the approaches we discussed during the recent lectures, we used Lagrange multipliers or dual variables or "penalties" to "move" the constraints into the objective function of (P) and "convert" the constrained continuous optimization problem at hand to an unconstrained optimization problem.
Suppose (P) is a convex optimization problem. Then under some mild assumptions, we can express (P) in the following conic form:

$$(CP) := \left\{ \begin{array}{r} inf\ \langle c, x \rangle \\ s.t.\ Ax = b \\ x \in K \end{array} \right\}$$

where $K \subset \mathbb{R}^n$ is closed convex cone with nonempty interior, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$,$c \in \mathbb{R}^n$ are all given.

$$\underset{\text{Dual cone of } K}{K^*} := \{s \in \mathbb{R}^n : \langle x, s\rangle \geqslant 0, \forall x \in K\}$$

We define the dual of (CP) as:

$$(CD) := \begin{Bmatrix} sup\ \langle b, y\rangle \\ A^T y + s = c \\ s \in K^* \end{Bmatrix}$$

Theorem 3.29: Weak Duality Theorem for Conic Optimization

For every $\overline{x} \in \mathbb{R}^n$ that is feasible for (CP) and for every $(\overline{y}, \overline{s}) \in \mathbb{R}^m \oplus \mathbb{R}^n$ that is feasible for (CD), we have $\langle c, \overline{x}\rangle \geqslant \langle b, \overline{y}\rangle$. Moreover, if $\langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle$, then $\overline{x}$ is optimal for (CP) and $(\overline{y}, \overline{s})$ is optimal for (CD).

Proof. Let $\overline{x}, (\overline{y}, \overline{s})$ be feasible solutions to $(CP), (CD)$ respectively. Then

$$\begin{aligned} \langle c, \overline{x}\rangle - \langle b, \overline{y}\rangle &= \left\langle A^T\overline{y} + \overline{s}, \overline{x}\right\rangle - \langle b, \overline{y}\rangle \\ &= \left\langle \overline{y}, \underset{=b}{A\overline{x}}\right\rangle + \langle \overline{s}, \overline{x}\rangle - \langle b, \overline{y}\rangle \\ &= \left\langle \underset{\in K^*}{\overline{s}}, \underset{\in K}{\overline{x}}\right\rangle \underbrace{\geqslant}_{\text{by defn of } K^*} 0 \end{aligned}$$

Applying the first part to $\overline{x}$ and every feasible solution (y, s) of (CD) estabilishes that $(\overline{y}, \overline{s})$ is optimal for (CD). Applying the first part to $(\overline{y}, \overline{s})$ and every .... □

Suppose $F : int(K) \to \mathbb{R}$ has the following properties:

- $F \in C^3$
- $\forall\ \{x^{(k)}\} \subset int(K)$ such that $x^{(k)} \to \overline{x} \in bd(K)$, $F(x^{(k)}) \to \infty$
- $|D^3F(x)[d, d, d]| \leqslant 2(D^2F(x)[d, d])^{\frac{3}{2}}, \forall x \in int(K), \forall d \in \mathbb{R}^n$
- $F(tx) = F(x) - \theta ln(t), \forall x \in int(K), \forall t \in \mathbb{R}_{++}$, for some $\theta \geqslant 1$

Such an F is called a Logarithmically Homogeneous Self-concordant barrier for K.
Recall, in Theorem 43, we needed $D^2f \in Lip(L)$.

$$|D^3F(x)[d_1, d_2, d_2]| = \lim_{t\to 0} \frac{1}{t}\left|D^2F(x + td_1)[d_2, d_2] - D^2F(x)[d_2, d_2]\right| \leqslant L\|d_1\|\|d_2\|^2$$

In the theory of self-concordant functions, we are replacing 2-norms with local norms defined by $D^2F(X)$.
For every $\mu > 0$, we define

$$(P_\mu) := \begin{Bmatrix} inf\ \langle c, x\rangle + \mu F(x) \\ s.t.\ Ax = b \end{Bmatrix} \text{ Define } F(x) := \infty, \forall x \in \mathbb{R}^n \setminus int(K)$$

Some examples of LHSCBs:
LP:

$$K := \mathbb{R}^n_+, \theta = n,\ F(x) := \begin{cases} -\sum_{j=1}^n \ln(x_j),\ x \in \mathbb{R}^n_{++} \\ \infty,\ \text{otherwise} \end{cases}$$

Semidefinite Programming:

$$K := \mathbb{S}^n_+, \theta = n,\ F(x) := \begin{cases} -\ln det(X),\ x \in \mathbb{S}^n_{++} \\ \infty,\ \text{otherwise} \end{cases}$$

$$K := \left\{ \begin{pmatrix} t \\ x \end{pmatrix} \in \mathbb{R} \oplus \mathbb{R}^n : t \geqslant \|x\|_2 \right\} \text{ (Second-Oder Cone)}$$

$$F(t,x) := \begin{cases} -\ln(t^2 - \|x\|_2^2),\ \|x\|_2 < t \\ \infty,\ \text{otherwise} \end{cases}$$

Taking direct sums of Second Order Cones leads to Second-Order Cone Programming (SOCP) problems.

Nov 19,2020 Each of these cones is a pointed, closed convex cone. A convex set is pointed if it does not contain whole line(s).

Theorem 3.30

If $K \subset \mathbb{R}^n$ is a pointed closed convex cone with nonempty interior, then so is its dual K^*.

- $\overline{x} \in \mathbb{R}^n$ is a Slater point for (CP) if $A\overline{x} = b$ and $\overline{x} \in int(K)$
- $(\overline{y}, \overline{s}) \in \mathbb{R}^m \oplus \mathbb{R}^n$ is a Slater point for (CD) if $A^T\overline{y} + \overline{s} = c$ and $\overline{s} \in int(K^*)$

The conditions above are the CQs for (CP) and (CD)

Theorem 3.31: Strong Duality Theorem for Conic Optimization

Suppose (CP) has a slater point and the objective function of (CP) is bounded from below over its feasible region. Then, (CD) has an optimal solution and the optimal objective values of (CP) and (CD) are the same.

Observation 2: Remark

The dual of (CD) is equivalent to (CP)

So we can swap $(CP) \leftarrow (CD)$ in Theorem 83

Corollary 3.32

Suppose both (CP) and (CD) have Slater points. Then, both (CP) and (CD) have optimal solutions and the optimal objective values of (CP) and (CD) are the same.

Suppose $A \in \mathbb{R}^{m \times n}$ has full row rank ($rank(A) = m$), K is a pointed closed convex cone with nonempty interio, F is a θ-LHSCB for K; $b \subseteq \mathbb{R}^m$, $c \in \mathbb{R}^n$ are given so that (CP) and (CD) have Slater points.
Recall the family of problems:$\mu > 0$

$$(CP_\mu) := \begin{Bmatrix} inf & \langle c, x \rangle + \mu F(x) \\ & Ax = b \end{Bmatrix}$$

Necessary and sufficient conditions for optimality:

$$Central - Path := \begin{Bmatrix} Ax = b, x \in int(K) \\ A^T y - \mu \nabla F(x) = c \end{Bmatrix}$$

For every $\mu > 0$, the above system has a unique solution $(x(\mu), y(\mu))$. In fact, $s(\mu) := \underbrace{-\mu \nabla F(x(\mu))}_{\in int(K^*)}$ yields a solution $(y(\mu), s(\mu))$ of (CD).

Observation 3: Remark

Let $x \in int(K)$, F *be a* $\theta - LHSCB$ *for* K. *Then* $\forall t > 0, F(tx) = F(x) - \theta ln(t)$
$\Rightarrow \begin{cases} \nabla F(tx) = \frac{1}{t}\nabla F(x) \\ \langle \nabla F(tx), x \rangle = -\frac{\theta}{t} \end{cases} \Rightarrow \langle -\nabla F(x), x \rangle = \theta$

We have from (Central-Path),

$$A^T y - \mu \nabla F(x(\mu)) = c \Rightarrow \left\langle x(\mu), A^T y(\mu) \right\rangle - \mu \left\langle \nabla F(x(\mu)), x(\mu) \right\rangle = \langle c, x(\mu) \rangle \Leftrightarrow \langle b, y(\mu) \rangle + \theta\mu = \langle c, x(\mu) \rangle$$

We can see that as $\mu \searrow 0$, $\langle a, x(\mu) \rangle$ and $\langle b, y(\mu) \rangle$ converge to the optimal objective values of (CP) and (CD)

Say $x^{(k)} \in int(K) \cap S$ is a very good approximation of $x(\mu_k)$ (where $x(\mu_k)$) is the optimal solution of (CP_{μ_k})). Let

$$\mu_{k+1} := \left(1 - \frac{0.1}{\sqrt{\theta}}\right) \mu_k$$

so that $x^{(k)}$ is a good approximation of $x(\mu_{k+1})$. Then taking one Newton step (or a similar move) $x^{(k+1)} := x^{(k)} + \alpha d$, $x^{(k+1)}$ becomes a very good approximation of $x(\mu_{k+1})$. Continuing, we obtain in k iterations, a feasible solution $x^{(k)}$ of (CP) such that

$$\mu_k = \left(1 - \frac{0.1}{\sqrt{\theta}}\right)^k \mu_0$$

Therefore, in $O(\sqrt{\theta} \ln (\frac{\theta \mu_0}{\epsilon}))$ iterations, we have $\overline{x}$ feasible in (CP) such that $\langle c, \overline{x} \rangle$ is within ϵ of the optimal objective value of (CP)
We can also design algorithms which utilize the dual problem (CD) more than we did.
Legendre-Fenchel Conjugate of F

$$F_* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, F_*(s) := \sup_{x \in R^n} \{- \langle s, x \rangle - F(x)\}, dom(F) := \{x : F(x) < \infty\}$$

By definition, $\forall \overline{x} \in dom(F)$, we have

$$- \langle s, \overline{x} \rangle - F(\overline{x}) \leqslant \sup_{x \in R^n} \{- \langle s, x \rangle - F(x)\}, dom(F) := \{x : F(x) < \infty\}, \forall s \in \mathbb{R}^n$$

Proposition 3.33: Fenchel-Young Inequality

For every $x \in dom(F)$ and $s \in \mathbb{R}^n$, we have

$$F(x) + F_*(s) \geqslant - \langle s, x \rangle$$

Theorem 3.34: Nesterov & Nemisovski[1994]

Let $K \subset \mathbb{R}^n$ be a pointed, closed convex cone with nonempty interior. If F is a $\theta - LHSCB$ for K, then F_* is a $\theta - LHSCB$ for K^*

Some Practical Issues

- We assumed having available $x^{(0)}$ such that $Ax^{(0)} = b, x^{(0)} \in int(K)$. In practice, we should be able to start from infeasible points $x^{(0)}$ (hence, infeasible-start algorithms). Let $e \in int(K)$, consider the auxiliary problem

$$(CP_{aux}) := \left\{ \begin{array}{ll} inf\ z & \\ & Ax + (b - Ae)z = b \\ & x \in K \\ & z \in \mathbb{R}_+ \end{array} \right\}$$

where $\begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix} := \begin{pmatrix} e \\ 1 \end{pmatrix}$ is a feasible solution (CP_{aux}). In fact, $\begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix} \in int(K \oplus \mathbb{R}_+)$
We may use a two-phase approach (compute a Slater point in Phase I, then initiate Phase II

to solve (CP)).
However, successful practical algorithms take a combines approach and strive to reduce both infeasibility and μ in a controlled way.

- The updates $\mu_{k+1} \leftarrow (1 - \frac{constant}{\sqrt{\theta}})\mu_k$ are too conservative in practice. We use much more aggressive strategies to decrease μ in practical algorithms.

Nov 24, 2020

Practical Issues continued:

In each iteration of an interior-point algorithm, we solve (perhaps appriximately) a linear system of equations

$$(AH^{-1}A^T)d_y = r, \text{ or } \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

where r, r_1, r_2 are given (easily computed), d_x and d_y are unknowns (leading to the seach direction in the current iteration) and H is either $\nabla^2 F(x^{(k)})$ or $[\nabla^2 F_*(s^{(k)})]^{-1}$ or some symmetric positive definite matrix related to these.

Current best upper bounds on the iteration complexity of interior-point algorithms are $\Omega(\sqrt{\theta} ln(\frac{1}{\epsilon}))$ to obtain an ϵ-optimal solution.

However, in practice, infeasible-start interior-point algorithms require 10-80 iterations to obtain a solution that is 10^{-9}-optimal on well-posed instaces of convex optimization problems.

A meta theorem of ipm practice:

Given a well-posed instance of a convex optimization problem, if we can perform one iteration of the ipm in a reasonable amount of time, we can solve the isntance ”””’

We can use the NEOs server for optimization

and/or cvx

and/or DDS

Leading commercial conic optimization solver MOSEK

Some interior-point algorithms can attain quadratic convergence (locally) or near-quadratic super-linear convrgence.

What if our problem instances are so huge that we cannot even perform a single iteration of an interior-point algorithm in a reasonable amount of time. Consider First-Order Algorithms

Aside:We also have soem techniques to address this within an ipm framework.

3.3 First-Order Methods

3.3.1 Worst-Case Computational Compelxity of First-Order Methods

Suppose $S = \mathbb{R}^n$, $f \in C^1$ and convex with Lipshitz continuous gradients, Lipschitz constant L. Consider the class of algorithms which generate a set of iterates with the properly that

$$x^{(k)} \in x^{(0)} + span\left\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\right\} ...(P.1)$$
$$\forall k \in \mathbb{Z}_{++}$$

Let's build a family of functions $\{f_e\}$ for which we can prove a lower bound on the number of iterations required by any algorithms with $(P.1)$ to compute an approximate minimizer.
Fix $L > 0$. Consider $\forall l \in \{1, 2 \ldots, n\}, f : \mathbb{R}^n \to \mathbb{R}$,

$$f_l(x) := \frac{L}{4}\left\{\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{l-1}(x_i - x_{i+1})^2 + x_l^2\right] - x_1\right\}$$

Note: f_l is a quadratic function $\forall l$. In fact,

$$f_l(x) = \frac{L}{8}x^T A_l x - \frac{L}{4}x_1,\ A_l \in \mathbb{S}^n$$

$$\nabla^2 f_l(x) = \frac{L}{4}A_l, \forall l$$

, where

$$A_l = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & \vdots \\ 0 & -1 & 2 & \ldots & 0 & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & \ldots & -1 & 2 & 0 \\ 0 & \ldots & \ldots & \ldots & 0 & 0 \end{pmatrix} \succcurlyeq 0 (\succ 0 \text{ if } l = n)$$

In fact,

$$0 \preccurlyeq \nabla^2 f_l(x) \preccurlyeq L.I, \forall l$$

Theorem 3.35

Let $n \geqslant 3$ be an odd integer. Then, for every $k \in \{1, 2, \ldots, \frac{n-1}{2}\}$ and for every $x^{(0)} \in \mathbb{R}^n$, there eixsts a C^∞ convex function f with $\nabla f \in Lip(L)$ such that every first-order gradient algrotihm obeying property $(P.1)$, we have

$$f(x^{(k)}) - \underline{f} \geqslant \frac{3L}{32(k+1)^2}\|x^{(0)} - \overline{x}\|_2^2$$
$$\|x^{(k)} - \overline{x}\|_2^2 \geqslant \frac{1}{8}\|x^{(0)} - \overline{x}\|_2^2$$

where $\overline{x} \in \mathbb{R}^n$ is the unique minimizer of f and $\underline{f} := f(\overline{x})$

Nov 26,2020

More preparations for proving Theorem 90:

The minimizers of our family $\{f_l\}$ of fucntions satisfy $A_l x = e_1$. Therefore, a minimizer of f is defined by

$$\overline{x}_j := \begin{cases} 1 - \frac{j}{l+1}, \text{ if } j \in \{1, 2, \dots, l\} \\ 0, \text{ otherwise} \end{cases}$$

leading to $\underline{f_l} = f_l(\overline{x}) = -\frac{L}{8}\left(1 - \frac{1}{l+1}\right)$

Note: By perturbing f_l slightly, we can make $\overline{x}$ the unique minimizer

We may assume $x^{(0)} = 0$ (do the corresponding "shift" to functions f_l). Then,

$$\begin{aligned} \|x^{(0)} - \overline{x}\|_2^2 &= \sum_{j=1}^{l} \left(1 - \frac{j}{l+1}\right)^2 = l - \frac{2}{l+1}\frac{l(l+1)}{2} + \frac{l(l+1)(2l+1)}{6(l+1)^2} \\ &= \frac{l * (2l+1)}{6(l+1)} < \frac{2L(l+1)}{6(l+1)} = \frac{1}{3}l \end{aligned}$$

Since $x^{(0)} = 0$, the affine subspace in $(P.1)$ for our $\{f_l\}$ becomes

$$span\left\{\nabla f_l(0), \nabla f_l(x^{(1)}), \dots, \nabla f_l(x^{(k-1)})\right\} \underbrace{=}_{induction} \mathbb{R}^k \oplus \{\underbrace{0}_{\in \mathbb{R}^{n-l}}\}$$

Using the above ingredients, we can prove Theorem 90. Can we do better if we restrict f to a nicer class of convex functions? (Our worst-case family $\{f_l\}$ was made up from convex quadratic functions, but $\lambda_{min}(\nabla^2 f(x)) = 0$).

Definition 3.36

Let $\mu \in \mathbb{R}_{++}$. Then, $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$ is called μ-strongly convex, if $\forall x, y \in \mathbb{R}^n$

$$f(y) \geqslant f(x) + \langle \nabla f(x), (y - x) \rangle + \frac{1}{2}\mu \|y - x\|_2^2$$

f is called strongly convex if $\exists \mu \in \mathbb{R}_{++}$ such that f is μ-strongly convex.

Proposition 3.37

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and let $\mu \in \mathbb{R}_{++}$. Then TFAE

- *f is μ-strongly convex;*
- $\forall \lambda \in [0, 1], \forall x, y \in \mathbb{R}^n$,

$$\lambda f(x) + (1 - \lambda) f(y) \geqslant f(\lambda x + (1 - \lambda) y) + \frac{\mu}{2}\lambda(1 - \lambda)\|y - x\|_2^2$$

- $\forall x, y \in \mathbb{R}^n$, $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant \mu \|y - x\|_2^2$

Proposition 3.38

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2, let $\mu \in \mathbb{R}_{++}$. Then, f is μ-strongly convex if and only if $\nabla^2 f(x) \succcurlyeq \mu I, \forall\, x \in \mathbb{R}^n$

An extension of the family $\{f_l\}$ of convex quadratic functions above (used in the proof of Theorem 90) can be used to prove:

Theorem 3.39

Let $n \geqslant 3$ be an integer. Then, for every $x^{(0)} \in \mathbb{R}^n$ and for every pair of constants $L > \mu > 0$ there exists a C^∞ function f which is μ-strongly convex, $\nabla f \in Lip(L)$ such that every First-Order algorithm obeying property $(P.1)$ generates a sequence $\{x^{(k)}\}$ satisfying:

$$f(x^{(k)}) - \underline{f} \geqslant \frac{\mu}{2}\left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k} \|x^{(0)} - \overline{x}\|_2^2$$

$$\|x^{(k)} - \overline{x}\|_2^2 \geqslant \left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k} \|x^{(0)} - \overline{x}\|$$

3.3.2 Optimal First-Order Methods

These typically use estimating (or auxiliary) sequences. Let $x^{(0)} \in \mathbb{R}^n$, choose $\alpha_0 \in (0,1),\ y^{(0)} := x^{(0)}$.

$$\text{Iteration } k: \begin{cases} \text{Evaluate } f(y^{(k)}), \nabla f(y^{(k)}) \\ x^{(k+1)} := y^{(k)} - \frac{1}{L}\nabla f(y^{(k)}) \\ \text{Compute } \alpha_{k+1} \text{ by solving } \alpha_{k+1}^2 = (1-\alpha_{k+1}){\alpha_k}^2 + \frac{\mu}{L}\alpha_{k+1} \\ \text{Set } \beta_k := \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}} \\ y^{(k+1)} := x^{(k+1)} + \beta_k\left(x^{(k+1)} - x^{(k)}\right) \end{cases}$$

What is all this in terms of $\{x^{(k)}\}$?

$$x^{(k+1)} = (1+\beta_{k-1})x^{(k)} - \beta_{k-1}x^{(k-1)} - \frac{1}{L}\nabla f\left((1+\beta_{k-1})x^{(k)} - \beta_{k-1}x^{(k)}\right)$$

$$\beta_{k-1} = \frac{1-\alpha_{k-1}}{\alpha_{k-1} + \frac{\alpha_k}{\alpha_{k-1}}},\ \alpha_{k+1} = \sqrt{\underbrace{\frac{\mu}{L}\alpha_{k+1} + \alpha_k^2(1-\alpha_{k+1})}_{\in[\frac{\mu}{L},\alpha_k^2]}}$$

If we choose $\alpha_0 := \sqrt{\mu/L}$, then the algorithm simplifies: $\alpha_k = \alpha_0, \forall k$, and $\beta_k := \frac{1-\alpha_0}{1+\alpha_0}, \forall k$. Thus,

$$\begin{cases} x^{(k+1)} := y^{(k)} - \frac{1}{L}\nabla f(y^{(k)}) \\ y^{(k+1)} := \frac{2}{1+\alpha_0}x^{(k+1)} - \frac{1-\alpha_0}{1+\alpha_0}x^{(k)} \end{cases}$$
$$\Leftrightarrow x^{(k+1)} := \left[\frac{2}{1+\alpha_0}x^{(k)} - \frac{1-\alpha_0}{1+\alpha_0}x^{(k-1)}\right] - \frac{1}{L}\nabla f\left(\left[\frac{2}{1+\alpha_0}x^{(k)} - \frac{1-\alpha_0}{1+\alpha_0}x^{(k-1)}\right]\right)$$

Theorem 3.40

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and μ-strongly convex, $\nabla f \in Lip(L)$, where $L > \mu > 0$. Suppose in the above algorithm $\alpha_0 \geqslant \sqrt{\mu/L}$ is chosen. Then, the iterates $\{x^{(k)}\}$ satisfy:

$$f(x^{(k)}) - \underline{f} \leqslant min\left\{\left(1 - \sqrt{\mu/L}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2}\right\}\left(f(x^{(0)}) - \underline{f} + \frac{\gamma_0}{2}\|x^{(0)} - \overline{x}\|_2^2\right)$$

where $\gamma_0 := \frac{\alpha_0(\alpha_0 L - \mu)}{1-\alpha_0}$

If $\alpha_0 = \sqrt{\mu/L}$, then $\gamma_0 = \frac{\sqrt{L\mu} - \mu}{\sqrt{L/\mu} - 1} = \mu$

December 1,2020

Can we relate the conlcusions of Theorem 94&95?

Theorem

Assume ... then the sequence of iterates $\{x^{(k)}\}$ *generated by every First-Order Algorithm satisfy*

$$f(x^{(k)}) - \underline{f} \geqslant \ldots \|x^{(0)} - \overline{x}\|_2^2 \text{ and } \|x^{(k)} - \overline{x}\|^2 \geqslant \ldots \|x^{(0)} - \overline{x}\|_2^2$$

Theorem

Assume ... then there exists a First-Order Algorithm whose iterates $\{x^{(k)}\}$ *satisfy*

$$f(x^{(k)}) - \underline{f} \leqslant \ldots \left(f(x^{(0)}) - \underline{f} + \frac{\gamma_0}{2}\|x^{(0)} - \overline{x}\|_2^2\right)$$

We will use the following:

Lemma 3.41

For every $\alpha \in (-1, 1)$,

$$\alpha - \frac{\alpha^2}{2(1-|\alpha|)} \leqslant ln(1+\alpha) \leqslant \alpha$$

Let's try to answer the questions: Given $\epsilon > 0$,

1. What is the lower bound on the number of iterations required to obtain $x^{(k)}$ such that $f(x^{(k)}) - \underline{f} \leqslant \epsilon$?

2. What is the upper bound on the number of iterations of the algorithm we described (Nesterov's Algorithm) which guarantees that $f(x^{(k)}) - \underline{f} \leqslant \epsilon$?

Q.1 From Theorem 94: Note $f(x^{(k)}) - f$ is not part of the iff

$$\frac{\mu}{2}\left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k} \|x^{(0)} - \overline{x}\|_2^2 \leqslant f(x^{(k)}) - \underline{f} \leqslant \epsilon \text{ iff}$$

$$ln(\mu) + 2k ln\left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right) + ln\left(\|x^{(0)} - \overline{x}\|_2^2\right) \leqslant ln(\epsilon) + ln(2)$$

$$\Rightarrow k \geqslant \left(\sqrt{L/\mu} - 1\right)\left[ln(1/\epsilon) + ln(\|x^{(0)} - \overline{x}\|_2^2) + ln(\mu) - ln(2)\right] \text{ We used lemma 96}$$

$$\Rightarrow k = \Omega\left[\sqrt{L/\mu}\left(ln(1/\epsilon) + ln(\|x^{(0)} - \overline{x}\|_2^2) + ln(\mu) -\right)\right]$$

Q.2 From Theorem 95 (for k large enough):

$$f(x^{(k)}) - \underline{f} \leqslant \left(1 - \sqrt{\mu/L}\right)^k \left(f(x^{(0)}) - \underline{f} + \frac{\gamma_0}{2}\|x^{(0)} - \overline{x}\|_2^2\right) \leqslant \epsilon$$

$$\text{iff } \ln \epsilon \geqslant k \ln\left(1 - \sqrt{\mu/L}\right) + \ln\left(\underbrace{f(x^{(0)}) - \underline{f}}_{1)} + \frac{\gamma_0}{2}\underbrace{\|x^{(0)} - \overline{x}\|_2^2}_{2)}\right)$$

Can we express $1)$ in terms of $2)$?

Lemma 3.42

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and convex. Suppose f has a unique minimizer $\overline{x}$ and $\nabla f \in Lip(L)$. Then,

$$f(x) - \underline{f} \leqslant L\|x - \overline{x}\|_2^2, \forall x \in \mathbb{R}^n$$

Proof. Suppose all the asuumptions in the statement of the lemma hold. Since f is C^1 and convex,

$$f(\overline{x}) \geqslant f(x) + \langle \nabla f(x), \overline{x} - x\rangle, \forall x \in \mathbb{R}^n \tag{3.1}$$

$$\Rightarrow \forall x \in \mathbb{R}^n f(x) - \underline{f} \leqslant \|\nabla f(x)\|_2 \|x - \overline{x}\|_2 \text{ Used Cauchy-Schwarz Inequality} \tag{3.2}$$

$$\nabla f \in Lip(L) \Rightarrow \|\nabla f(x) - \underbrace{\nabla f(\overline{x})}_{=0}\|_2 \leqslant L\|x - \overline{x}\|_2, \forall x \in \mathbb{R}^n \tag{3.3}$$

By $(2.2), (2.3)$,

$$f(x) - \underline{f} \leqslant L\|x - \overline{x}\|_2^2, \forall x \in \mathbb{R}^n$$

□

Back to answering Q.2:
To guarantee $f(x^{(k)}) - \underline{f} \leqslant \epsilon$, it suffices to ensure (we used lemma 97):

$$-k \ln\left(1 - \sqrt{\mu/L}\right) \geqslant \ln(1/\epsilon) + \ln\left(L + \frac{\gamma_0}{2}\right) + \ln\left(\|x^{(0)} - x\|_2^2\right)$$

Thus, by Lemma 96, it suffices to ensure

$$k \geqslant \sqrt{L/\mu}\left[ln(1/\epsilon) + ln\left(\|x^{(0)} - \overline{x}\|_2^2\right) + ln(L + \gamma_0/2)\right]$$

Therefore,

$$O\left(\sqrt{L/\mu}\left[ln(1/\epsilon) + ln\left(\|x^{(0)} - \overline{x}\|_2^2\right) + ln(L + \mu/2)\right]\right) \text{ Assuming } \alpha_0 = \sqrt{\mu/L}$$

iterations of Nesterov's Algorithm suffices.

How about Nonsmooth Convex Optimization?
$f : \mathbb{R}^n \to \mathbb{R}^n$, convex. Recall, subgradient of f at $\overline{x} \in \mathbb{R}^n$ is $h \in \mathbb{R}^n$ such that $f(x) \geqslant f(\overline{x}) + h^T(x - \overline{x}), \forall x \in \mathbb{R}^n$

Theorem 3.43

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then, $\overline{x} \in \mathbb{R}^n$ is a minimizer of f iff $0 \in \partial f(\overline{x})$

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, $\overline{x} \in \mathbb{R}^n$. Then,

$$0 \in \partial f(\overline{x}) \Leftrightarrow f(x) \geqslant f(\overline{x}), \forall x \in \mathbb{R}^n$$

□

Assumptions for iteration complexity lower bound
f is convex, has a unique minimizer $\overline{x}$, $\|x^{(0)} - \overline{x}\|_2 \leqslant R$, f is Lipschitz on $B(\overline{x}, R)$ with Lipschitz constant L.

Remark. (P.2) at each iteration k, an algorithm inputs $x^{(k)}$, gets $f(x^{(k)}), h^{(k)} \in \partial f(x^{(k)})$ generates $x^{(k+1)} \in x^{(0)} + span\{h^{(0)}, h^{(1)}, \ldots, h^{(k)}\}$

Theorem 3.44

For every choice of integer $n \geqslant 2$, and integer $k \in \{0, 1, \ldots, n-1\}$, $x^{(0)} \in \mathbb{R}^n, R > 0, L > 0$, there exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with a unique minimizer $\overline{x} \in B(x^{(0)}, R)$ such that f is Lipschitz continuous on $B(\overline{x}, R)$ with Lipshitz constant L and

$$f(x^{(k)}) - \underline{f} \geqslant \frac{L\,R}{2(2 + \sqrt{k+1})}$$

for every First-Order algorithm obeying property (P.2)

Proof. Just give an idea.
Consider the family of functions

$$f_k : \mathbb{R}^n \to \mathbb{R}, k \in \{1, 2, \ldots, n\}$$

defined by

$$f_k(x) := \mu_1 \left\| \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \right\|_\infty + \frac{\mu_2}{2}\|x\|_2^2$$

for suitable $\mu_1, \mu_2 > 0$ □

December 4,2020

Well,trivially, $\overline{x} := 0$ is the unique minimizer $\forall k$, and the nonsmooth part of f_k, even though tries to hide some information, it does not provide enough of a challenge against fast convergence. Next,let's try

$$f_k(x) := \mu_1 \underset{j\in\{1,2,\ldots,k\}}{max}\{x_j\} + \frac{\mu_2}{2}\|x\|_2^2$$

$\forall k \in \{1, 2, \ldots, n\}$, where $\mu_1, \mu_2 > 0$ to be chosen.

$$[\overline{x}(k)]_j := \begin{cases} -\frac{\mu_1}{\mu_2 k}, \text{ if } j \in \{1, 2, \ldots, k\} \\ 0, \text{ otherwise} \end{cases}$$

is the unique minimizer of f_k.

$$\partial f_k(x) = \mu_2 x + \mu_1 conv\{e_j : j \in J(x)\}, \forall k \in \{1, 2, \ldots, n\}$$

where $J(x) := \{j : x_j = \underset{i\in\{1,2,\ldots,k\}}{maximum}\{x_i\}\}$

$$\underline{f_k} := f_k(\overline{x}(k)) = -\frac{\mu_1^2}{\mu_2 k} + \frac{\mu_2}{2} k \frac{\mu_1^2}{\mu_2^2 k^2} = -\frac{\mu_1^2}{2\mu_2 k}$$

Claim: f_k is Lipschitz continuous on $B(\overline{x}(k), R), \forall k \in \{1, 2, \ldots, k\}$
Proof of claim: Let $y \in \mathbb{R}^n$ and $u \in \partial f_k(y)$. Then

$$f_k(y) - f_k(x) \leqslant \langle u, y - x\rangle \leqslant \|u\|_2\|y - x\|_2, \forall x \in \mathbb{R}^n$$

Thus,

$$|f_k(y) - f_k(x)| \leqslant \|u\|_2\|y - x\|_2, \forall x, y \in \mathbb{R}^n$$

And we know that

$$\|u\|_2 \leqslant \mu_1\sqrt{k} + \mu_2\|\overline{x}(k)\|_2 + \mu_2\|x - \overline{x}(k)\|_2 \leqslant \left(\sqrt{k} + \frac{1}{\sqrt{k}}\right)\mu_1 + \mu_2\|x - \overline{x}(k)\|_2$$

Therefore, for every $x, y \in B(\overline{x}(k), R)$

$$|f_k(y) - f_k(x)| \leqslant \underbrace{\left[\left(\sqrt{k} + \frac{1}{\sqrt{k}}\right) + R\mu_2\right]}_{\text{Lipschitz constant}}\|y - x\|_2$$

Choosing $\mu_1 := \frac{\sqrt{k+1}}{2+\sqrt{k+1}}L$, $\mu_2 := \frac{L}{(2+\sqrt{k+1})R}$, with the above ingredients leads to a proof of Theorem 99.

3.3.3 An Optimal Subgradient Algorithm

Let $S \subseteq \mathbb{R}^n$ be a closed convex set. Let $f : |R^n \to \mathbb{R}$ be a convex function. Consider the continuous optimization problem:

$$\begin{cases} inf\ f(x) \\ s.t.\ x \in S \end{cases}$$

Assume: $\forall x \in \mathbb{R}^n$, we can efficitently compute

$$proj(x|S) = argmin\{\|x - y\|_2 : y \in S\}$$

Algorithm:
Initialization: Choose $x^{(0)} \in S$, and a sequence $\{\alpha_k\} \subset \mathbb{R}$ such that $\alpha_k > 0, \forall k \in \mathbb{Z}_+, \alpha_k \searrow 0, \sum_{k=0}^{\infty} \alpha_k = \infty$
At iteration k:(we have $x^{(k)}$)
Compute $f(x^{(k)}), d^{(k)} \in \partial f(x^{(k)})$,

$$x^{(k+1)} := proj\left(x^{(k)} - \alpha_k \frac{d^{(k)}}{\|d^{(k)}\|_2} | S\right)$$

Theorem 3.45

let $S \subseteq \mathbb{R}^n$ be a closed convex set. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $\overline{x} := argmin\{f(x) : x \in S\}$ (a unique minimizer). Suppose $\exists R > 0$ such that f is Lipschitz continuous on $B(\overline{x}, R)$, with Lipschitz constant L. Then the above algorithm generates a sequence $\{x^{(k)}\}$ satisfying

$$f(x^{(k)}) - \underline{f} \leqslant L \frac{R^2 + \sum_{i=0}^{k} \alpha_i^2}{2\sum_{i=0}^{k} \alpha_i}, \forall k \in \mathbb{Z}_+$$

Proof. Suppose all the assumptions in the statement of the theorem hold. Denote the distance between the i^{th} iterate and the minimizer by δ_i. $\delta_i := \|x^{(i)} - \overline{x}\|_2$. Then,

$$\delta_{i+1}^2 = \left\| proj\left(x^{(i)} - \alpha_i \frac{d^{(i)}}{\|d^{(i)}\|_2} | S\right) - \overline{x} \right\|_2^2 \leqslant \left\| x^{(i)} - \overline{x} - \alpha_i \frac{d^{(i)}}{\|d^{(i)}\|_2} \right\|_2^2 \text{ By Theorem 78}$$

$$= \delta_i^2 - 2\alpha_i \frac{\left\langle x^{(i)} - \overline{x}, d^{(i)} \right\rangle}{\|d^{(i)}\|} + \alpha_i^2$$

$$\Rightarrow \delta_0^2 + \sum_{i=0}^{k} \alpha_i^2 \geqslant \underbrace{\delta_{k+1}^2 + 2\sum_{i=0}^{k} \alpha_i \frac{\left\langle x^{(i)} - \overline{x}, d^{(i)} \right\rangle}{\|d^{(i)}\|}}_{\text{We can show that this is bounded below by } \frac{2}{L}(f(x^{(k)}) - \underline{f}) \sum_{i=0}^{k} \alpha_i}$$

□

Suppose we want to stop after iteration K.

Then let's choose $\alpha_i := \frac{R}{\sqrt{K+1}}, \forall i \in \{1, 2, \ldots, K\}$. Thus,

$$\frac{R^2 + \sum_{i=0}^{K} \alpha_i^2}{2\sum_{i=0}^{K} \alpha_i} = \frac{2R^2}{2R\sqrt{K+1}} = \frac{R}{\sqrt{K+1}}$$
$$\Rightarrow f(x^{(K)}) - \underline{f} \leqslant \frac{L\ R}{\sqrt{K+1}}$$

which meets the lower bound from Theorem 99
If we want $\epsilon = \frac{LR}{\sqrt{K+1}} (\Leftrightarrow k+1 = \frac{L^2R^2}{\epsilon^2})$
If L, R both $O(1)$, this iteration complexity bound is $O(\frac{1}{\epsilon^2})$.