Problem 1.1

I) $f(\vec{x}_2) \geq f(\vec{x}_1) + \nabla f(\vec{x}_1)^T(\vec{x}_2 - \vec{x}_1) + \frac{\mu}{2}\|\vec{x}_2 - \vec{x}_1\|_2^2$ ⊛

By Taylor expansion, there exist $\vec{z} \in \{t\vec{x}_1 + (1-t)\vec{x}_2 \text{ for } t \in [0,1]\}$ such that

$$f(\vec{x}_2) = f(\vec{x}_1) + \nabla f(\vec{x}_1)^T(\vec{x}_2 - \vec{x}_1) + \frac{1}{2}(\vec{x}_2 - \vec{x}_1)^T \nabla^2 f(\vec{z})(\vec{x}_2 - \vec{x}_1)$$

So if ⊛ holds, it means $(\vec{x}_2 - \vec{x}_1)^T \nabla^2 f(\vec{z})(\vec{x}_2 - \vec{x}_1) \geq \mu\|\vec{x}_2 - \vec{x}_1\|_2^2$ Since x_1, x_2 are chosen arbitrarily, this means $a \triangleq \frac{(x_2 - x_1)}{\|x_2 - x_1\|_2}$ and $a^T \nabla^2 f(z) a \geq \mu \Rightarrow \nabla^2 f(z) \geq mI$

Conversely:

$$f(\vec{x}_2) = f(\vec{x}_1) + \nabla f(\vec{x}_1)^T(\vec{x}_2 - \vec{x}_1) + \frac{1}{2}(\vec{x}_2 - \vec{x}_1)^T \underbrace{\nabla^2 f(\vec{z})}_{\geq \mu I}(\vec{x}_2 - \vec{x}_1)$$
$$\geq f(\vec{x}_1) + \nabla f(\vec{x}_1)^T(\vec{x}_2 - \vec{x}_1) + \frac{\mu}{2}\|\vec{x}_2 - \vec{x}_1\|_2^2$$

II) $$\left.\begin{array}{l} f(\vec{x}_2) \geq f(\vec{x}_1) + \nabla f(\vec{x}_1)^T(\vec{x}_2 - \vec{x}_1) + \frac{\mu}{2}\|\vec{x}_2 - \vec{x}_1\|_2^2 \\ f(\vec{x}_1) \geq f(\vec{x}_2) + \nabla f(\vec{x}_2)^T(\vec{x}_1 - \vec{x}_2) + \frac{\mu}{2}\|\vec{x}_2 - \vec{x}_1\|_2^2 \end{array}\right\} \oplus \Rightarrow (\nabla f(\vec{x}_2) - \nabla f(\vec{x}_1))^T(\vec{x}_2 - \vec{x}_1) \geq \mu\|\vec{x}_2 - \vec{x}_1\|_2^2$$

Conversely assume

$(\nabla f(\vec{x}_2) - \nabla f(\vec{x}_1))^T(\vec{x}_2 - \vec{x}_1) \geq \mu\|\vec{x}_2 - \vec{x}_1\|_2^2$ Define $g(\vec{x}) = f(\vec{x}) - \frac{\mu}{2}\|\vec{x}\|^2$

$$\Rightarrow ((\nabla f(\vec{x}_2) - \mu\vec{x}_2) - (\nabla f(x_1) - \mu\vec{x}_1))^T(\vec{x}_2 - \vec{x}_1) + \underbrace{\mu(\vec{x}_2 - \vec{x}_1)^T(\vec{x}_2 - \vec{x}_1)}\geq \underbrace{\mu\|\vec{x}_2 - \vec{x}_1\|_2^2}$$

$(\nabla g(\vec{x}_2) - \nabla g(\vec{x}_1))^T(\vec{x}_2 - \vec{x}_1) \geq 0$ ⊛

Define $h(t) = g(\vec{x}_2 + t(\vec{x}_1 - \vec{x}_2))$

then $h'(t) - h'(0) = \nabla g(\vec{x}_2 + t(\vec{x}_1 - \vec{x}_2))^T(\vec{x}_1 - \vec{x}_2) - \nabla g(\vec{x}_2)(\vec{x}_1 - \vec{x}_2)$

$$= \frac{1}{t}\left(\nabla g(\vec{x}_2 + t(\vec{x}_1 - \vec{x}_2))^T - \nabla g(\vec{x}_2)\right)(t(\vec{x}_1 - \vec{x}_2)) \geq 0 \quad ⊛⊛$$

$$g(x_1) = h(1) = h(0) + \int_0^1 h'(t) \overset{⊛⊛}{\geq} h(0) + h'(0)$$
$$= g(\vec{x}_2) + \nabla g(\vec{x}_2)(\vec{x}_1 - \vec{x}_2)$$

$$\Rightarrow f(\vec{x}_1) - \frac{\mu}{2}\|x_1\|^2 \geq f(\vec{x}_2) - \frac{\mu}{2}\|x_2\|^2 + (\nabla f(\vec{x}_2) - \mu\vec{x}_2)(\vec{x}_1 - \vec{x}_2)$$

$$\Rightarrow f(\vec{x}_1) \geq f(\vec{x}_2) + \nabla f(\vec{x}_2)(\vec{x}_1 - \vec{x}_2) + \frac{\mu}{2}\|\vec{x}_1 - \vec{x}_2\|_2^2$$

III - a) $f(x) \geq f(y) + \nabla f(y)^T (x-y) + \frac{\mu}{2}\|x-y\|^2 \quad \forall x,y$

Define $g(z) = f(z) - \nabla f(x)^T z$ $\qquad \nabla g(z) = 0 \quad \nabla f(z) - \nabla f(x) = 0 \qquad z^* = x$

we show that $\quad g(z) \geq g(y) + \nabla g(y)^T (z-y) + \frac{\mu}{2}\|z-y\|^2 \quad (*)$

note that by expanding $g(\cdot)$ we have $(*)$ is equivalent to:

$f(z) - \nabla f(x)^T z \geq f(y) - \nabla f(x)^T y + (\nabla f(y) - \nabla f(x))^T (z-y) + \frac{\mu}{2}\|z-y\|^2$

$\equiv \quad f(z) \geq f(y) + \nabla f(y)^T (z-y) + \frac{\mu}{2}\|z-y\|^2$ which we have assumed to hold.

So $(*)$ holds.

$\Rightarrow \min_z g(z) \geq g(y) - \nabla g(y)^T y + \min_z \overbrace{\{\nabla g(y)^T z + \frac{\mu}{2}\|z-y\|^2\}}^{A(z)} \qquad \operatorname{argmin}_z (A(z)) = y - \frac{\nabla g(y)}{\mu}$

$\overset{z^*=x}{\Rightarrow} \quad f(x) - \nabla f(x)^T x \geq f(y) - \nabla f(x)^T y - \nabla g(y)^T y + \left[-\frac{1}{\mu}\|\nabla g(y)\|^2 + \nabla g(y)^T y + \frac{\mu}{2}\frac{1}{\mu^2}\|\nabla g(y)\|^2\right]$

$\overset{x=x^*}{\Rightarrow}: \quad f^* \geq f(y) - \frac{1}{2\mu}\|\nabla g(y)\|^2$

III - b) In III-a we showed that by defining $g(z) = f(z) - \nabla f(x)^T z$ we have

$g(z) \geq g(y) + \nabla g(y)^T (z-y) + \frac{\mu}{2}\|z-y\|^2$ Then by minimizing with respect to z we have:

$f(x) - \nabla f(x)^T x \geq f(y) - \nabla f(x)^T y - \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2$

So $f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2$

from strict convexity

$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|x-y\|^2$

$(+)$

$\frac{\mu}{2}\|x-y\|^2 \leq \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2 \quad \Rightarrow \quad \|\nabla f(y) - \nabla f(x)\|_2 \geq \mu\|x-y\|_2$

III - c) From Cauchy-Schwarz inequality we have:

$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \leq \|\nabla f(x_2) - \nabla f(x_1)\|_2 \, \|x_2 - x_1\|_2$

$\overset{\text{III-a}}{\leq} \frac{1}{\mu}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2$

where the last step holds according to the last part of the problem.

III.d) $f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2}\|x_2 - x_1\|_2^2$
$r(x_2) \geq r(x_1) + \nabla r(x_1)^T (x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|_2^2$ (+) $g(x) \triangleq f(x) + r(x)$, $\mu' = \mu + L$

$$g(x_2) \geq g(x_1) + \nabla g(x_1)^T (x_2 - x_1) + \frac{\mu'}{2}\|x_2 - x_1\|_2^2$$

so $g(x)$ μ'- strongly convex

1.5) $\forall x, y$ $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2 \rightarrow$ L-Smooth

$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \mu\|x-y\|^2 \rightarrow \mu$- strongly convex

Define $g(x) \triangleq f(x) - \frac{\mu}{2}\|x\|^2$ $\quad \nabla g(x) = \nabla f(x) - \mu x$

$$g(x) - g(y) = f(x) - f(y) - \frac{\mu}{2}\|x\|^2 + \frac{\mu}{2}\|y\|^2 \leq \nabla f(x)^T(x-y) - \frac{\mu}{2}\|x-y\|^2 + \frac{\mu}{2}(\|y\|^2 - \|x\|^2)$$

from strong convexity of $f(x)$

$$= (\nabla f(x) - \mu x)^T (x-y) = \nabla g(x)^T (x-y) \Rightarrow g(x) \text{ is convex.}$$

$$\|\nabla g(x) - \nabla g(y)\|_2^2 = \|\nabla f(x) - \nabla f(y) - \mu(x-y)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|^2 + \mu^2\|x-y\|^2 - 2\mu(\nabla f(x) - \nabla f(y))^T(x-y)$$

(problem 1.2-c)
$$\leq \|\nabla f(x) - \nabla f(y)\|^2 - \frac{2\mu}{L}\|\nabla f(x) - \nabla f(y)\|^2 + \mu^2\|x-y\|^2$$

L-smooth $f(x)$
$$\leq (L-\mu)^2\|x-y\|^2$$

So $g(x)$ is convex & $L-\mu$ smooth

From problem 1.2-c : $(\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2$

~~problem 1.2-c~~

So $(\nabla f(x) - \nabla f(y) - \mu(x-y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y) - \mu(x-y)\|^2$

$$(L-\mu)(\nabla f(x) - \nabla f(y))^T(x-y) - \mu(L-\mu)\|x-y\|^2 \geq \|\nabla f(x) - \nabla f(y)\|^2 + \mu^2\|x-y\|^2 - 2\mu(\nabla f(x) - \nabla f(y))^T(x-y)$$

$$\Rightarrow (L+\mu)(\nabla f(x) - \nabla f(y))^T(x-y) \geq \mu L\|x-y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2$$

$$\Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{\mu L}{L+\mu}\|x-y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2$$

Scanned with CamScanner

1.3 Rates of Convergence

In asymptotic analysis, convergence rates allow to compare the speed at which different algorithms approach some limiting value. Rates of convergence are typically determined by the amount of information about the function f is used in the updating process of the sequence. More simple algorithms tend to converge more slowly than those that require more information about f.

If a sequence x_n converges to x_∞, we say that the convergence is linear if there exists an $r \in (0,1)$ such that

$$\frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} \leq r \qquad \text{for all } n \text{ sufficiently large}$$

example: $x_n = 1 + \left(\frac{1}{2}\right)^n$ converges linearly to $x_\infty = 1$ $\rightarrow$ $\dfrac{\left(\frac{1}{2}\right)^{n+1}}{\frac{1}{2}} = \frac{1}{2}$

In linear converge, $x_n - x_\infty = O(r^k)$. This means that for an accuracy of ϵ we need $O\left(\frac{\log \epsilon}{\log r}\right) = O(-\log \epsilon)$ iterations. The linear terminology comes from log-plot of $x_n - x_\infty$ with respect to n.

if instead

$$\lim_{n \to \infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} = 0 \rightarrow$$ we say the convergence is superlinear, faster than linear convergence.

example: $x_n = 1 + \left(\frac{1}{n}\right)^n$

$$\lim_{n \to \infty} \frac{\left(\frac{1}{n}\right)^{n+1}}{\left(\frac{1}{n}\right)^n} = \frac{1}{n} \rightarrow$$ Converges super linearly to 1

If $x_n - x_\infty = o(r^{e^k})$ then to achieve accuracy of ϵ we need $O(\log \log 1/\epsilon)$ which is much smaller (faster) than linear convergence

On the other hand, the sequence converges <u>sublinearly</u>, which means it is slower tha linear convergence, if:

$$\lim_{n\to\infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} = 1$$

example: $x_n = 1 + \frac{1}{n}$ converges sublinearly to 1. $\lim_{n\to\infty} \frac{\frac{1}{n+1}}{\frac{1}{n}} = 1$

For sublinear convergence, $x_n - x_\infty = O(1/k)$ which means that for a accuracy ϵ for x_n we need $k = O(1/\varepsilon)$ iterations. For very small ϵ, this number is exponentially big.

Finally, we say a sequence has <u>quadratic</u> convergence if there some constant $0 \leq M \leq \infty$, such that

$$\lim_{n\to\infty} \frac{\| x_{n+1} - x_\infty \|}{\| x_n - x_\infty \|^2} \leq M < \infty$$

Quadratic convergence is faster than the previous cases and is generally considered desirable, if possible to achieve.

example: $x_n = 1 + \left(\frac{1}{n}\right)^{2^n}$ converges quadratically to 1.

$$\lim_{n\to\infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|^2} = \frac{\left(\frac{1}{n+1}\right)^{2^{n+1}}}{\left(\frac{1}{n}\right)^{(2^n)2}} = \left(\frac{n}{n+1}\right)^{2^{n+1}} \leq 1$$

Problem 1.2

Given $f: R^d \to R$ is L-smooth.

We have $\forall x_1, x_2 \in \mathbb{R}^d \; \|\nabla f(x_2) - \nabla f(x_1)\|_2 \le L\|x_2 - x_1\|_2$. (2)

Sol: (a) Let $g(t) = f((1-t)x_1 + t x_2) \Rightarrow \begin{cases} g(0) = f(x_1) \\ g(1) = f(x_2) \end{cases}$

$$f(x_2) - f(x_1) = \int_0^1 g'(t)\,dt = \int_0^1 \nabla f((1-t)x_1 + t x_2)^T (x_2 - x_1)\,dt$$

Now, we consider $\left| f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \right|$

$$= \left| \int_0^1 \{\nabla f((1-t)x_1 + t x_2)^T (x_2 - x_1) - \nabla f(x_1)^T (x_2 - x_1)\} dt \right|$$

$$\le \int_0^1 \left| (\nabla f((1-t)x_1 + t x_2)^T - \nabla f(x_1)^T)(x_2 - x_1) \right| dt$$

(Cauchy-Schwarz)

$$\le \int_0^1 \|\nabla f((1-t)x_1 + t x_2) - \nabla f(x_1)\| \, \|x_2 - x_1\| \, dt$$

(from (2))

$$\le \int_0^1 Lt \, \|x_2 - x_1\|^2 dt$$

$$= \frac{L}{2}\|x_2 - x_1\|^2.$$

$$\therefore \underbrace{f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \le L/2 \|x_2 - x_1\|^2}_{(A)}. \quad \text{QED}$$

Sol: (b) Let $z = x_2 - \frac{1}{L}(\nabla f(x_2) - \nabla f(x_1))$

$$f(x_1) - f(x_2) = f(x_1) - f(z) + f(z) - f(x_2)$$

$$\le \nabla f(x_1)^T (x_1 - z) \quad \text{(Convexity)}$$

(from (A))

$$\nabla f(x_2)^T (z - x_2) + L/2 \|z - x_2\|^2$$

$$\le \nabla f(x_1)^T (x_1 - z) - \frac{1}{L} \nabla f(x_2)^T (\nabla f(x_2) - \nabla f(x_1)) + \frac{1}{2L} \|\nabla f(x_2) - \nabla f(x_1)\|^2$$

$$= \nabla f(x_1)^T (x_1 - x_2) + \nabla f(x_1)^T (x_2 - z) - 1/L \, \nabla f(x_2)^T (\nabla f(x_2) - \nabla f(x_1)) + \frac{1}{2L} \|f(x_2) - \nabla f(x_1)\|^2$$

(Note: $x_2 - z = 1/L(\nabla f(x_2) - \nabla f(x_1))$)

$$= \nabla f(x_1)^T(x_1-x_2) - \frac{1}{L}\|\nabla f(x_2)-\nabla f(x_1)\|^2 + \frac{1}{2L}\|\nabla f(x_2)-\nabla f(x_1)\|^2$$

$$\therefore \underbrace{f(x_2) \geq f(x_1) + \nabla f(x_1)^T(x_2-x_1)}_{B} + \frac{1}{2L}\|\nabla f(x_2)-\nabla f(x_1)\|^2.$$

QED.

Sol: (c) Since B is true for any x_1, x_2, we have

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^T(x_1-x_2) + \frac{1}{2L}\|\nabla f(x_2)-\nabla f(x_1)\|^2 \quad C$$

Adding B and C, we obtain

$$0 \geq \nabla f(x_1)^T(x_2-x_1) - \nabla f(x_2)^T(x_2-x_1) + \frac{1}{L}\|\nabla f(x_2)-\nabla f(x_1)\|^2$$

$$\Rightarrow (\nabla f(x_2) - \nabla f(x_1))^T(x_2-x_1) \geq \frac{1}{L}\|\nabla f(x_2)-\nabla f(x_1)\|^2.$$

QED

Problem 1.4

Study Group #3

1

minimize $\frac{1}{M}\sum_{i=1}^{M} f_i(x_i)$

subjected to $Ax=b$

for $b \in \mathbb{R}^{P\times M}$ and $x=[x_1,\dots,x_M]^T$

solution

a) $M = 1000$ (or $M=10^3$)

* $\{x \mid Ax=b\} = \{Fz+\hat{x} \mid z \in \mathbb{R}^{M-P}\}$ (1)

where $F \in \mathbb{R}^{M\times(M-P)}$ is a matrix and $\hat{x} \in \mathbb{R}^M$ is a vector that parametrize the affine set given in (1).

The constrained optimization problem becomes

* unconstrained as follows:

minimize $\tilde{f}(z) = f(Fz+\hat{x})$

* From its solution z^*, we can find the solution of the equality constrained problem as $x^* = Fz^* + \hat{x}$

b) $M = 10^9$

The same method above could be used.

References

Stephen Boyd, convex optimization, © 2004
L. Vandenberghe page 523, section 10.1.2

Problem 14 (continuation) | Study group #3 | 2

c) Yes. Newton's method can be used for $M = 10^9$.

1) minimize $\hat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 (f(x)) v$ (1)

subject to $A(x+v) = b$

⇒ were the objective function is replaced with its second order approximation.

2) The KKT matrix is given by:

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix} \quad (2)$$

3) $Ax^* = b$ and $\nabla f^*(x) + A^T v^* = 0$ (3)

4) substituting $x + \Delta x_{nt}$ for x^* and w for v^*, we get:

$A(x + \nabla x_{int}) = b$, and

$\nabla f(x + \Delta x_{nt}) + A^T w \approx \nabla f(x) + \nabla^2 f(x) \Delta x_{nt} + A^T w = 0$

5) using $Ax = b$ and $A \Delta x_{nt} = 0$

$\nabla^2 f(x) \Delta x_{nt} + A^T w = -\nabla f(x)$

d) Adding $r(x)$ to $f(x)$ can be similarly solved by Newton's method iff $r(x)$ is convex.

References

Stephen Boyd & L. Vandenberghe, Convex optimization, © 2004 pages 525-529, section 10.2