CMO - Tutorial

Sheet 2 Solutions — E0230

For Tutors Only — Not For Distribution

1. *Convex function.* Consider the optimization problem

$$x^* = min_{x \in \Omega} f(x)$$

where f is a real-valued function and Ω is the feasible set. A set Ω is a *convex set* if for every $x_1, x_2 \in \Omega$ and every real number α, $0 < \alpha < 1$, the point $\alpha x_1 + (1-\alpha)x_2 \in \Omega$. A function f defined on a convex set Ω is said to be *convex* if for every $x_1, x_2 \in \Omega$ and every α, $0 \leq \alpha \leq 1$, the following holds

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$

We state the following properties of convex functions and the proofs of these statements will be discussed in a later lecture.

Proposition 1: Let $f \in C^1$. Then, f is convex over a convex set Ω if and only if

$$f(y) \geq f(x) + \nabla f(x)(y-x)$$

for all $x, y \in \Omega$

Proposition 2: Let $f \in C^2$. Then, f is convex over a convex set Ω containing an interior point if and only if the Hessian matrix of f, is positive semi-definite throughout Ω. [3]

Consider the function

$$f(x) = \frac{1}{2}||Ax+b||_2^2$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $m \geq n$.

(a) Show that f is a convex function

(b) Given that x and d in $\mathbb{R}^m$ compute the solution t^* to the 1-dimensional optimization problem

$$t^* = min_{t \in \mathbb{R}} f(x+td)$$

(c) State the condition for function f to have a unique solution.

(d) Assume that A satisfies the condition that you have identified in Part(c). Also, assume that $m \geq n$ and A has no non-singular values. Now, let us apply steepest descent with exact line search to the function f, what is the convergence rate of steepest descent algorithm starting from an arbitrary initial point.

Solution:

(a) $\nabla f(x) = A^T(Ax + b)$ and $\nabla^2 f(x) = A^T A$ is positive semi-definite for any A, so convex.

(b) $g(t) = f(x+td) = \frac{1}{2}||A(x+td)+b||_2^2 = \frac{1}{2}(t^2||Ad||_2^2 + 2t(Ad)^T(Ax+b) + ||Ax+b||_2^2)$. $g'(t) = t||Ad||_2^2 + (Ad)^T(Ax+b) = 0 \implies t^* = \frac{-(Ad)^T(Ax+b)}{||Ad||_2^2}$

(c) $A^T A$ is positive definite

(d) Convergence rate $= (\frac{r-1}{r+1})^2$ where $r = \frac{\lambda_1^2}{\lambda_n^2}$, where λ_1 is the largest singular value of A and λ_n is the smallest singular value of A, $\lambda_n \neq 0$.

2. *Convergence of steepest descent.* Suppose we use the method of steepest descent to minimize the quadratic function $f(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$ but we allow a tolerance $\pm\delta\alpha_k, \delta \geq 0$ in the line search. that is,

$$x_{k+1} = x_k - \alpha_k g_k,$$

where

$$(1-\delta)\overline{\alpha_k} \leq \alpha_k \leq (1+\delta)\overline{\alpha_k}$$

and $\overline{\alpha_k}$ minimizes $f(x_k - \alpha g_k)$ over α.

(a) Prove that the convergence rate of steepest descent with exact line search after T iterations, starting from an initial point x_0 is $f(x_T) \leq e^{-Tc} f(x_0)$ where, $c = \frac{(1-\delta^2)4aA}{(a+A)^2}$, a and A, are the smallest and largest eigen values of Q

(b) What is the range of values of δ that guarantees convergence of the algorithm

Solution:

(a) We know that

$$\frac{f(x_k) - f(x_{k+1})}{f(x_k)} = \frac{2\alpha_k(g_k^T Q g_k) - \alpha_k^2(g_k^T Q g_k)}{g_k^T Q^{-1} g_k}$$

where $g_k = Q(x_k - x^*)$. Assume $\alpha_k = (1+\delta)\overline{\alpha_k}$, where $\overline{\alpha_k} = \frac{(g_k^T g_k)^2}{g_k^T Q g_k}$

$$\begin{aligned}
\frac{f(x_k) - f(x_{k+1})}{f(x_k)} &= \frac{2(1+\delta)\overline{\alpha_k}(g_k^T Q g_k) - (1+\delta)^2\overline{\alpha_k}^2(g_k^T Q g_k)}{g_k^T Q^{-1} g_k} \\
&= \frac{2(1+\delta)\frac{(g_k^T g_k)^2}{g_k^T Q g_k} - (1+\delta)^2\frac{(g_k^T g_k)^2}{g_k^T Q g_k}}{g_k^T Q^{-1} g_k} \\
&= \frac{(1-\delta^2)(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^- 1 g_k)} \\
&\leq \frac{(1-\delta^2)4aA}{(a+A)^2}
\end{aligned}$$

$$\frac{f(x_{k+1})}{f(x_k)} \leq 1 - \frac{(1-\delta^2)4aA}{(a+A)^2}$$

Let $c = \frac{(1-\delta^2)4aA}{(a+A)^2}$ .We know that $(1-c) \leq e^{-c}$. Therefore, after T iterations,

$$f(x_T) \leq e^{-Tc} f(x_0)$$

(b) To calculate range of δ, $1 - \frac{(1-\delta^2)4aA}{(a+A)^2} < 1$. Simplifying this equation,

$$\frac{(1-\delta^2)4aA}{(a+A)^2} > 0$$

$$\implies \delta^2 < 1 \implies 0 < \delta < 1$$

3. *Constant step-size.* Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and its gradient is Lipchitz continuous with constant $L > 0$, ie. we have $||\nabla f(x) - \nabla f(y)||_2 \leq L||x-y||_2$ for any x, y. Then if we run steepest descent for T iterations with a fixed step size $\alpha_k = \alpha \leq 1/L$ for every iteration, then the steepest descent is guaranteed to converge with a rate proportional to $\mathcal{O}(1/T)$.

Solution: Lipschitz conditions on the gradient can be equivalently written as

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}||y-x||_2^2$$

Let $y = x_{k+1} = x_k - \alpha \nabla f(x)$

$$\begin{aligned} f(x_k - \alpha \nabla f(x)) &\leq f(x_k) + \nabla f(x_k)^T (x_k - \alpha \nabla f(x_k) - x_k) + \frac{L}{2}||x_k - \alpha \nabla f(x_k) - x_k||_2^2 \\ &= f(x_k) - (1 - \frac{L\alpha}{2})\alpha ||\nabla f(x_k)||_2^2 \end{aligned}$$

Applying $\alpha_k = \alpha \leq 1/L$, we get $-(1 - \frac{L\alpha}{2}) \leq \frac{L}{2L} - 1 \leq -1/2$. Thus,

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2}||\nabla f(x_k)||_2^2$$

Since f is convex, $f(y) \geq f(x) + \nabla f(x)^T (y-x)$, for all x, y in the feasible set. Therefore, if $f(x^*)$ is the optimal objective value $f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k)$.

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq -\nabla f(x_k)^T (x^* - x_k) - \frac{\alpha}{2}||\nabla f(x_k)||_2^2 \\ &\leq \frac{1}{2\alpha}(2\alpha \nabla f(x_k)^T (x_k - x^*) - \alpha^2 ||\nabla f(x_k)||_2^2 - ||x_k - x^*||_2^2 + ||x_k - x^*||_2^2) \\ &\leq \frac{1}{2\alpha}(||x_k - x^*||_2^2 - ||x_{k+1} - x^*||_2^2) \end{aligned}$$

Summing over T iterations, we get

$$\begin{aligned}
f(x_T) - f(x^*) &\leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \\
&\leq \frac{1}{2T\alpha}\sum_{k=0}^{T}(||x_k - x^*||_2^2 - ||x_{k+1} - x^*||_2^2) \\
&\leq \frac{1}{2T\alpha}(||x_0 - x^*||_2^2) \\
&\leq \mathcal{O}(1/T)
\end{aligned}$$

4. *Convergence rate.* Suppose an iterative algorithm of the form

$$x_{k+1} = x_k + \alpha_k d_k$$

is applied to quadratic problem with matrix Q, where α_k is the minimum point of the line search, d_k is a vector satisfying $d_k^T g_k < 0$ and $(d_k^T g_k)^2 \geq \beta(d_k^T Q d_k)(g_k^T Q^{-1} g_k)$, where $0 < \beta \leq 1$. This corresponds to a steepest descent algorithm with 'sloppy' choice of direction. Estimate the rate of convergence of this algorithm.

Solution: Let us consider a quadratic function $f(x_k) = \frac{1}{2}x_k^T Q x_k - x_k^T b$.

$$\begin{aligned}
f(x_k + \alpha_k d_k) &= \frac{1}{2}(x_k + \alpha_k d_k)^T Q(x_k + \alpha_k d_k) - (x_k + \alpha_k d_k)^T b \\
&= \frac{1}{2}[x_k^T Q x_k + 2x_k^T Q \alpha_k d_k + \alpha_k^2 d_k^T Q d_k] - x_k^T b - \alpha_k d_k^T b
\end{aligned}$$

Differentiating w.r.t α_k,

$$\frac{d}{d\alpha_k} f(x_k + \alpha_k d_k) = x_k^T Q d_k + \alpha_k d_k^T Q d_k - d_k^T b = 0$$

$$\alpha_k = -\frac{d_k^T g_k}{d_k^T Q d_k}$$

where $g_k = Qx_k - b$. We know that

$$\begin{aligned}
E(x_k) &= \frac{1}{2}(x_k - x^*)^T Q(x_k - x^*) \\
&= \frac{1}{2}[x_k^T Q x_k - 2x_k^T Q x^* + x^{*T} Q x^*] \\
E(x_{k+1}) &= \frac{1}{2}(x_{k+1} - x^*)^T Q(x_{k+1} - x^*) \\
&= \frac{1}{2}(x_k + \alpha_k d_k - x^*)^T Q(x_k + \alpha_k d_k - x^*) \\
&= \frac{1}{2}[x_k^T Q x_k + 2x_k^T Q \alpha_k d_k - 2x_k^T Q x^* + \alpha_k^2 d_k^T Q d_k - 2\alpha_k d_k^T Q x^* + x^{*T} Q x^*]
\end{aligned}$$

Now we have,

$$\frac{E(x_{k+1}) - E(x_k)}{E(x_k)} = \frac{2\alpha_k g_k^T d_k + \alpha_k^2 d_k^T Q d_k}{g_k^T Q^{-1} g_k}$$

Substituting value of α in the above equation,

$$\begin{aligned}\frac{E(x_{k+1})-E(x_k)}{E(x_k)} &= \frac{-\frac{2(d_k^T g_k)^2}{d_k^T Q d_k}+\frac{(d_k^T g_k)^2}{d_k^T Q d_k}}{g_k^T Q^{-1} g_k} \\ &= \frac{-(d_k^T g_k)^2}{(g_k^T Q^{-1} g_k)(d_k^T Q d_k)} \\ &\leq -\beta \\ \frac{E(x_{k+1})}{E(x_k)} &\leq 1-\beta\end{aligned}$$

5. *Minimizing quadratic functions.* Consider the function $f(x) = \sum_{i=1}^{d} i x_i^2 - b^T x$ where $b \in \mathbb{R}^d$.

   (a) Find x^*, the global minimum of x. Justify your answer.

   (b) How many iterations will the steepest descent algorithm with exact line-search take to reach to a point whose function value is ϵ close to $f(x^*)$, starting from the initial point.

   (c) Now if you apply steepest descent with heavy ball method, given by the following equation

   $$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

   where $\alpha = \frac{4}{\sqrt{M}+\sqrt{m}}$, $\beta = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}$, M and m are the largest and smallest eigen value of $\nabla^2 f(x)$, calculate the number of iterations required to reach a point whose function value is ϵ close to $f(x^*)$, starting from the initial point.

Solution:

(a) $f(x) = \sum_{i=1}^{d} i x_i^2 - b^T x = \frac{1}{2} X^T A X - b^T X$ where $A = diag(1, 2, ..., d)$ and $X \in \mathbb{R}^d$. $\nabla f(x) = A^T X - b$. $X^* = A^{-1} b$.

(b) We have $f(x_k) - f(x^*) \leq \epsilon$ after atmost

$$\frac{log((f(x_0) - f(x^*))/\epsilon)}{log(1/c)}$$

iterations, where $c = 1 - m/M$. $log(1/c) = -log(1 - m/M) \approx m/M$.

(c) For heavy ball method, $f(x_k) - f(x^*) \leq \epsilon$ after atmost

$$\frac{log(\epsilon/(f(x_0) - f(x^*)))}{log(c)}$$

iterations, where $c = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}$

6. *Backtracking line search.* Consider the problem of inexact line search for minimising a function $f : \mathbb{R}^d \to \mathbb{R}$ along the descent direction $u \in \mathbb{R}$, i.e, $\min_{t>0} f(x+tu)$. The Armio-Goldstein condition of inexact line search states that t should satisfy $f(x+tu) \leq f(x) + \alpha t \nabla f(x)^T u$, for a given constant $\alpha \in \left(0, \frac{1}{2}\right)$.

   (a) Suppose there exists $m, M \in \mathbb{R}_+$ such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in Dom(f)$. Show that the Armio-Goldstein condition is satisfied if
   $$0 \leq t \leq -\frac{\nabla f(x)^T u}{M||u||_2^2}.$$

   (b) Let $\bar{t} = \min\{t : f(x+tu) = f(x) + \alpha t \nabla f(x)^T u\}$. In the backtracking line search algorithm, an initial value of $t_0 = 1$ is chosen and for some $\beta \in (0,1)$, t is repeatedly updated as $t_k \leftarrow \beta t_{k-1}$ until it satisfies $t_k \leq \bar{t}$. Provide a bound of the number of updates k required, in terms of β and $\bar{t}$.

Solution:

(a) Let $t \geq 0$. For some $z \in \mathbb{R}^d$, we have
$$\begin{aligned}
f(x+tu) &= f(x) + t\nabla f(x)^T u + \frac{1}{2}(tu)^T \nabla^2 f(z)(tu) \\
&\leq f(x) + t\nabla f(x)^T u + \frac{1}{2}(tu)^T MI(tu) \qquad (\text{since } \nabla^2 f(z) \preceq MI) \\
&= f(x) + t\nabla f(x)^T u + \frac{M}{2} t^2 u^T u.
\end{aligned}$$
So, for the given line search condition to be satisfied, it is sufficient to have
$$\begin{aligned}
f(x) + t\nabla f(x)^T u + \frac{M}{2} t^2 u^T u &\leq f(x) + \alpha t \nabla f(x)^T u \\
\text{i.e., } \frac{M}{2} t u^T u &\leq (\alpha - 1)\nabla f(x)^T u \\
\text{i.e., } t &\leq -2(1-\alpha)\frac{\nabla f(x)^T u}{M||u||_2^2}.
\end{aligned}$$
Now, since $\alpha < \frac{1}{2}$, $1-\alpha > \frac{1}{2}$ and $2(1-\alpha) > 1$. So, for the above condition to hold, it is sufficient that
$$t \leq -\frac{\nabla f(x)^T u}{M||u||_2^2}.$$

(b) If $\bar{t} \geq 1$, then $t_0 \leq \bar{t}$ and hence number of steps required is $k = 0$. Suppose $\bar{t} < 1$. For $t_k \leq \bar{t}$, we need $\beta^k t_0 \leq \bar{t}$, i.e.,
$$\begin{aligned}
k \log \beta &\leq \log \bar{t} \\
\text{i.e., } k \log \frac{1}{\beta} &\geq \log \frac{1}{\bar{t}} \\
\text{i.e., } k &\geq \frac{\log\left(\frac{1}{\bar{t}}\right)}{\log\left(\frac{1}{\beta}\right)} \text{ is sufficient}
\end{aligned}$$

Therefore, the number of iterations required is at most $\left\lceil \frac{\log\left(\frac{1}{t}\right)}{\log\left(\frac{1}{\beta}\right)} \right\rceil$.

7. *In-exact line search.* Consider a quadratic function given by

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

Its one-dimensional minimizer along the ray $x_k + \alpha_k d_k$ is given by

$$\alpha_k = -\frac{\nabla f(x_k)^T d_k}{d_k^T Q d_k}$$

where d_k is the descent direction. Suppose, there exists $m, M > 0$ such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in Dom(f)$. Show that the one-diamensional minimizer of f satisfies the Goldstein condition given by.

$$f(x_k) + (1-c)\alpha_k \nabla f(x_k)^T d_k \leq f(x_k + \alpha_k d_k) \leq f(x_k) + c\alpha_k \nabla f(x_k)^T d_k$$

with $0 < c < 1/2$.

Solution: By Taylor's theorem we have

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) \\ &\geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}||y-x||_2^2 \end{aligned}$$

where $x = x_k$ and $y = x_k + \alpha_k d_k$. Setting the gradient w.r.t $y = 0$, we get $y - x = -\frac{1}{m}\nabla f(x)$

$$\begin{aligned} f(y) &\geq f(x) - \frac{1}{m}\nabla f(x)^T \nabla f(x) + \frac{m}{2m^2}\nabla f(x)^T \nabla f(x) \\ &\geq f(x) - \frac{1}{2m}\nabla f(x)^T \nabla f(x) \\ &\geq f(x) - \frac{(1-c)}{m}\nabla f(x)^T \nabla f(x) \\ &\geq f(x) + (1-c)\alpha_k \nabla f(x)^T d_k \end{aligned}$$

since $c < 1/2$, $-(1-c) < -1/2$ and $-\frac{1}{m}\nabla f(x) = \alpha_k d_k$. Similarly, the other inequality can be proved.

8. *Steepest descent.* Consider the steepest descent method with exact line search applied to convex quadratic function.

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

Suppose that the initial point x_0 is such that $x_0 = x^* + ucI$ where $x^* = arg\,min_x f(x)$, c is a constant, u is an eigen vector of Q and I is identity matrix. Then how many steps does steepest descent take to reach x^*, starting from x_0.

Solution: One step. We have $\nabla f(x_0) = Qx_0 - b = Q(x_0 - x^*) = \lambda(x_0 - x^*)$, where λ is the eigen value corresponding to eigen vector of Q which is parallel to $x_0 - x^*$. $\alpha_0 = 1/\lambda$. Substituting this in $x_1 = x_0 - \alpha_0 \nabla f(x_0) = x_0 - \frac{\lambda(x_0 - x^*)}{\lambda} = x^*$.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. USA: Cambridge University Press, 2004. ISBN: 0521833787.

[2] Roger Fletcher. *Practical Methods of Optimization*. Second. New York, NY, USA: John Wiley & Sons, 1987.

[3] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Springer Publishing Company, Incorporated, 2015. ISBN: 3319188410.

[4] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. second. New York, NY, USA: Springer, 2006.