CS 726 Assignment 4

Ruochen Lin

March 13, 2018

1

Given $d_k = -\nabla f(x_k)$ and $\alpha \in (0, \frac{1}{L})$, we have

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

Also, equation (3.8) in the manuscript reads

$$f(x + \alpha d) \leqslant f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2} \|d\|^2 ;$$

by plugging in $d = -\nabla f(x_k)$ we have

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) - \alpha \left\|\nabla f(x_k)\right\|^2 + \frac{\alpha^2 L}{2} \left\|\nabla f(x_k)\right\|^2 \\ &= f(x_k) - \frac{\alpha(2 - \alpha L)}{2} \left\|\nabla f(x_k)\right\|^2 . \end{aligned} \tag{1}$$

1.1 General case

Rearrange equation (1) we have:

$$\left\|\nabla f(x_k)\right\|^2 \leqslant \frac{2}{\alpha(2 - \alpha L)} [f(x_k) - f(x_{k+1})]. \tag{2}$$

Sum both sides of equation (2) from 0 to n, we have

$$\begin{aligned} \sum_{k=0}^{n} \left\|\nabla f(x_k)\right\|^2 &\leqslant \frac{2}{\alpha(2 - \alpha L)} [f(x_0) - f(x_{n+1})] \\ &\leqslant \frac{2}{\alpha(2 - \alpha L)} [f(x_0) - f(x^*)] \end{aligned} \tag{3}$$

$$\Rightarrow \min_{k=0,\cdots,n} \left\|\nabla f(x_k)\right\| \leqslant \sqrt{\frac{2(f(x_0) - f(x^*))}{\alpha(n+1)(2 - \alpha L)}},$$

thus preserving the $\frac{1}{\sqrt{n}}$ convergence rate; note that if we plug $\alpha = \frac{1}{L}$ into the inequality above, we recover the results for steepest descent in the manuscript.

1.2 Convex case

If $f(x)$ is convex, then we have

$$\begin{aligned} f(x^*) &\geqslant f(x_k) + \nabla f(x_k)^T (x^* - x_k) \\ \Rightarrow f(x_k) &\leqslant f(x^*) - \nabla f(x_k)^T (x^* - x_k). \end{aligned}$$

Plug this into equation (1) we have

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x^*) - \nabla f(x_k)^T (x^* - x_k) - \frac{\alpha(2-\alpha L)}{2} \left\| \nabla f(x_k) \right\|^2 \\ &= f(x^*) - \frac{\alpha(2-\alpha L)}{2} \nabla f(x_k)^T [\frac{2}{\alpha(2-\alpha L)} (x^* - x_k) + \nabla f(x_k)] \\ &= f(x^*) - \frac{\alpha(2-\alpha L)}{2} \left(\left\| \nabla f(x_k) + \frac{x^* - x_k}{\alpha(2-\alpha L)} \right\|^2 - \left\| \frac{x^* - x_k}{\alpha(2-\alpha L)} \right\|^2 \right) \\ &= f(x^*) - \frac{1}{2\alpha(2-\alpha L)} \left(\left\| \alpha(2-\alpha L) \nabla f(x_k) + x^* - x_k \right\|^2 - \left\| x^* - x_k \right\|^2 \right) \\ &\leqslant f(x^*) - \frac{1}{2\alpha(2-\alpha L)} \left(\left\| \alpha \nabla f(x_k) + x^* - x_k \right\|^2 - \left\| x^* - x_k \right\|^2 \right) \\ &= f(x^*) - \frac{1}{2\alpha(2-\alpha L)} \left(\left\| x^* - x_{k+1} \right\|^2 - \left\| x^* - x_k \right\|^2 \right) \end{aligned}$$

$$\Longrightarrow \quad f(x_{k+1}) - f(x^*) \leqslant \frac{1}{2\alpha(2-\alpha L)} \left(\left\| x^* - x_k \right\|^2 - \left\| x^* - x_{k+1} \right\|^2 \right) \tag{4}$$

Sum inequality (4) from $k = 0$ to $n - 1$, we have

$$\begin{aligned} \sum_{k=0}^{n-1} (f(x_{k+1}) - f(x^*)) &\leqslant \frac{1}{2\alpha(2-\alpha L)} (\left\| x_0 - x^* \right\|^2 - \left\| x_n - x^* \right\|^2) \\ &\leqslant \frac{1}{2\alpha(2-\alpha L)} \left\| x_0 - x^* \right\|^2 \end{aligned} \tag{5}$$

Because $f(x_k)$ is decreasing,

$$f(x_n) - f(x^*) \leqslant \frac{1}{2n\alpha(2-\alpha L)} \left\| x_0 - x^* \right\|^2 = o(\frac{1}{n}).$$

1.3

2

2.1

There are infinite solutions, because for underdetermined (*i.e.* $n < d$) case like this, there is either no solution or infinite solutions.

2.2

Instead of minimizing the function $f_0(x) = \frac{1}{n}\|Ax - b\|^2$, we opt to minimize the function $f(x) = \|Ax - b\|^2$ to get rid of the cumbersome coefficient $\frac{1}{n}$. Since they only differ in a factor of constant, they'll have the same convergence peoperties, the same minimizer, and even the same minimum 0; only that the error in f_0 is 10 times smaller than that of f:

$$f_0(x) \leqslant \epsilon \ \Leftrightarrow f(x) \leqslant n\epsilon.$$

We first note that $f(x)$ (and $f_0(x)$, of course) is convex, but not strongly convex. To prove this, we first write $f(x)$ in the form of a quadratic function:

$$\begin{aligned} f(x) &= x^T A^T Ax - 2b^T Ax + b^T b, \\ \nabla f(x) &= 2A^T Ax - 2A^T b = 2A^T(Ax - b). \end{aligned}$$

$f(x)$ is convex because

$$\begin{aligned} &f(y) - f(x) - \nabla f(x)^T(y - x) \\ =&y^T A^T Ay - 2b^T Ay - x^T A^T Ax + 2b^T Ax \\ &- 2x^T A^T A(y - x) + 2b^T A(y - x) \\ =&y^T A^T Ay - 2x^T A^T Ay + x^T A^T Ax \\ =&\|Ax - Ay\|^2 \geqslant 0 \end{aligned}$$

$$\implies f(y) \geqslant f(x) + \nabla f(x)(y - x);$$

it's not strongly convex because $A^T A$ is singular and must have 0 as at least one of its eigenvalues: $\text{rank}(A^T A) = \text{rank(A)} = n < d$, but $A^T A$ is a $d \times d$ matrix; in other words, $A^T A$ is positive semidefinite.

In order to make $f(x)$ fit into our analysis in class, we define $A' = 2A^T A$, so that $f(x) = \frac{1}{2}x^T A' x - 2b^T Ax + b^T b$. Suppose the largest eigenvalue of A' is L, and x^* is a minimizer of f, then

$$\begin{aligned} f(x_K) - f(x^*) = f(x_K) &\leqslant \frac{L}{2K} \left\| x_0 - x^* \right\|^2 \\ &= \frac{L}{2K} \left\| x^* \right\|^2 ; \end{aligned}$$

if we require the error within K steps in $f(x)$ to be smaller than $n\epsilon$, then

$$K \leqslant \frac{L}{2n\epsilon} \left\| x^* \right\|^2 .$$

Here are some notes:

- Since $f(x)$ has infinite minimizers, our iteration does not necessarily lead to $x_k \to x^*$; however, $\|x^*\|$ can still be used to bound our error; in fact, the error can be bounded by $\left\| x^\dagger \right\| = \min_{\{x\,:\,Ax=b\}} \|x\|$, with $x^\dagger$ being the minimizer of $f(x)$ that has the smallest Euclidean distance from origin.

- The relationship between the spectra of A' and $A^T A$ is the following: $\lambda_i(A') = 2\lambda_i(A^T A)$; thus if we define L' as the largest eigenvalue of $A^T A$, then we should replace the Ls in our inequality with $2L'$.

- Writing $f(x)$ in its quadratic form is only of conceptual use to us; we would never want of calculate $A^T A$, which has the complexity of about $O(nd^2)$, in practice. Evaluating $f(x) = \|Ax - b\|^2$ and $\nabla f(x) = 2A^T(Ax - b)$ each costs us $O(nd)$, and doing a exact line serch would have similar time complexity, if we alway evaluate expressions like $x^T A^T Ax$ as $(Ax)^T(Ax)$.

- If n is small compared to d, then we probably can afford to evaluate the matrix product AA^T, which costs $O(n^2 d)$, at the beginning of the program. By doing so, we can transform the problem into one with much nicer properties: Minimizing $g(t) = \left\| AA^T t - b \right\|^2$. Because AA^T is a $n \times n$ matrix with rank n, it is now invertible and thus is positive definite, (instead of being positive semi-definite, as we've seen above,) making $g(t)$ strictly convex. Now that $g(t)$ is a strictly convex function, we can yield much faster convergence with descent methods: given that the condition number of $A^T A$ is κ, after K iterations we would have

$$g(t_K) \leqslant (1 - \frac{1}{\kappa})^K \|b\|^2$$

$$\Rightarrow \quad k \geqslant \frac{\ln \frac{n\epsilon}{\|b\|^2}}{\ln(1-\frac{1}{\kappa})},$$

if we want the error in $g(t)$ to be smaller than $n\epsilon$. Finally, we can get the corresponding minimizer in x-space with

$$x = A^T t.$$

The solution we get from this algorithm is also the $x^\dagger$ we mentioned above, namely the solution that's closest to origin.

2.3

$$l_\mu(x) = \frac{1}{n}\|Ax-b\|^2 + \mu\|x\|^2,$$
$$\nabla l_\mu(x) = \frac{2}{n}A^T(Ax-b) + 2\mu x = (\frac{2}{n}A^T A + \mu I)x - \frac{2}{n}A^T b.$$

At the minimizer of l_μ, we have

$$\nabla l_\mu(x^{(\mu)}) = 0$$
$$\Rightarrow \quad \big(\frac{2}{n}A^T A + 2\mu I\big)x^{(\mu)} = \frac{2A^T b}{n},$$
$$x^{(\mu)} = \big(A^T A + n\mu I\big)^{-1} A^T b.$$

Here $A^T A + n\mu I$ is invertible because $A^T A$ is positive semidefinite, as we've shown above, and $n\mu I$ is positive definite, making the sum positive definite and thus invertible.

2.4

$$\begin{aligned} l_\mu(x) &= \frac{2}{n}\|Ax-b\|^2 + \mu\|x\|^2 \\ &= x^T(\frac{1}{n}A^T A + \mu I)x - \frac{2}{n}b^T Ax + \|b\|^2. \end{aligned}$$

If we define $\tilde{A} = \frac{1}{n}A^T A + \mu I$, and its condition number $\tilde{\kappa}$, then

$$\begin{aligned} l_\mu(x_k) - l_\mu(x^\mu) &\leqslant (1-\frac{1}{\tilde{\kappa}})^k(l_\mu(x_0) - l_\mu(x^{(\mu)})) \\ &= (1-\frac{1}{\tilde{\kappa}})^k(\frac{\|b\|^2}{n} - l_\mu(x^{(\mu)})). \end{aligned}$$

If we desire the left-hand side of the inequality above to be no larger than ϵ after K steps, then

$$\begin{aligned}
(1-\frac{1}{\tilde{\kappa}})^K(\frac{\|b\|^2}{n}-l_\mu(x^{(\mu)})) &\leqslant \epsilon \\
K\ln(1-\frac{1}{\tilde{\kappa}}) &\leqslant \ln\frac{\epsilon}{\frac{\|b\|^2}{n}-l_\mu(x^{(\mu)})}, \\
K &\geqslant \frac{\ln\frac{\epsilon}{\frac{\|b\|^2}{n}-l_\mu(x^{(\mu)})}}{\ln(1-\frac{1}{\tilde{\kappa}})}.
\end{aligned}$$

2.5

$$\begin{aligned}
\frac{1}{n}\|A\hat{x}-b\|^2 &= l_\mu(\hat{x})-\mu\|\hat{x}\|^2 \\
&\leqslant \epsilon + l_\mu(x^{(\mu)})-\mu\|\hat{x}\|^2\,;
\end{aligned}$$

plug in the expression for $x^{(\mu)}$, we have

$$\begin{aligned}
\frac{1}{n}\|A\hat{x}-b\|^2 &\leqslant \epsilon+\frac{1}{n}\left\|A(A^TA+n\mu I)^{-1}A^Tb-b\right\|^2+\mu\left\|(A^TA+n\mu I)^{-1}A^Tb\right\|^2-\mu\|\hat{x}\|^2 \\
&\leq \epsilon+\frac{1}{n}\left\|A(A^TA+n\mu I)^{-1}A^Tb-b\right\|^2+\mu\left\|(A^TA+n\mu I)^{-1}A^Tb\right\|^2.
\end{aligned}$$

3

Given symmetric positive definite matrix A, if for a set of vectors $\{p_k\}$ we have

$$p_i^TAp_j=0,\ \text{if } i\neq j,$$

then we define

$$P=\begin{bmatrix}p_0, & p_1, & \dots, & p_l\end{bmatrix}$$

and

$$\Sigma=P^TAP=\begin{bmatrix}\sigma_0 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_l\end{bmatrix}\succ 0,$$

in which $\sigma_i=p_i^TAp_i>0$ because A is positive definite.

We can prove the theorem by contradiction: if $\{p_k\}$ is linearly dependent, *i.e.* $\exists x \neq 0$ such that $Px = \sum_{i=0}^{l} x_i p_i = 0$, then, on one hand,

$$x^T \Sigma x = x^T P^T A P x = (Px)^T A (Px) = 0;$$

on the other, since Σ is positive definite,

$$x^T \Sigma x > 0,$$

which contradicts the equation above. Hence we've shown that $\{p_k\}$ can only be linearly independent if they're conjugate with respect to symmetric possitive definte matrix A.

4

Lemma: if A is symmetric, then any polynomial of A is also symmetric.

Proof: Suppose A is an $n \times n$ symmetric matrix, then

$$\begin{aligned}
(A^k)_{ij} &= \sum_{x_1=1}^{n} \sum_{x_2=1}^{n} \cdots \sum_{x_{k-1}=1}^{n} A_{ix_1} A_{x_1 x_2} \cdots A_{x_{k-1} j} \\
&= \sum_{x_1=1}^{n} \sum_{x_2=1}^{n} \cdots \sum_{x_{k-1}=1}^{n} A_{x_1 i} A_{x_2 x_1} \cdots A_{j x_{k-1}} \\
&= (A^k)_{ji}.
\end{aligned}$$

And clearly the sum of symmetric matrices is also symmetric; and a polynomial of A to the kth power is a sum of $k+1$ symmetric matrices, which is symmetric, and thus $[I + P_k(A)A]^T = [I + P_k(A)A]$.

With the result from the textbook that $P_k(A)v_i = P_k(\lambda_i)v_i$, we have:

$$\begin{aligned}
[I + P_k(A)A]^T A [I + P_k(A)A] v_i &= [I + P_k(A)A] A [v_i + P_k(A) A v_i] \\
&= [I + P_k(A)A] A [(1 + \lambda_i P_k(\lambda_i)) v_i] \\
&= (1 + \lambda_i P_k(\lambda_i)) [I + P_k(A)A] A v_i \\
&= \lambda_i (1 + \lambda_i P_k(\lambda_i)) [I + P_k(A)A] v_i \\
&= \lambda_i (1 + \lambda_i P_k(\lambda_i))^2 v_i.
\end{aligned}$$

Hence $(v_i, \lambda_i(1 + \lambda_i P_k(\lambda_i))^2)$ is an eigenpair of $[I + P_k(A)A]^T A [I + P_k(A)A]$, given that (v_i, λ_i) is one for A.

5

For the given problem, the solution is

$$x^* = \begin{bmatrix} 1 - \frac{1}{n+1} \\ 1 - \frac{2}{n+1} \\ \vdots \\ 1 - \frac{n}{n+1} \end{bmatrix},$$

and x_k can have non-zero entries only in the first k spots, given the starting point $x_0 = 0$. Thus

$$\begin{aligned}
\left\|x_0 - x^*\right\|^2 = \left\|x^*\right\|^2 &= \sum_{j=1}^{n} \left(1 - \frac{j}{n+1}\right)^2 \\
&= \sum_{j=1}^{n} \left(1 - \frac{2j}{n+1} + \frac{j^2}{(n+1)^2}\right) \\
&= n - \frac{2}{n+1} \sum_{j=1}^{n} j + \frac{1}{(n+1)^2} \sum_{j=1}^{n} j^2 \\
&= n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{n(n+1)(2n+1)}{6} \\
&= n - n + \frac{2n+1}{n+1} \cdot \frac{n}{6} \\
&\leqslant \frac{2n}{6} = \frac{n}{3}
\end{aligned}$$

and

$$
\begin{aligned}
\left\|x_k - x^*\right\|^2 &\geqslant \left\| \left[0, \quad \cdots, \quad 0, \quad 1-\tfrac{k+1}{n+1}, \quad \cdots, 1-\tfrac{n}{n+1}\right]^T \right\| \\
&= \sum_{j=k+1}^{n} \left(1 - \frac{j}{n+1}\right)^2 = \frac{\sum_{j=k+1}^{n}(n+1-j)^n}{(n+1)^2} \\
&= \frac{1}{(n+1)^2}\left((n-k)^2 + (n-k-1)^2 + \cdots + 2^2 + 1^2\right) \\
&= \frac{1}{(n+1)^2} \cdot \frac{(n-k)(n-k+1)(2n-2k+1)}{6} \\
&\geqslant \frac{1}{(n+1)^2} \cdot \frac{(n-k)(n-k)(2n-2k)}{6} \\
&= \frac{1}{(n+1)^2} \cdot \frac{(n-k)^3}{3} = \frac{(n-k)^3}{3(n+1)^2} \\
&\geqslant \frac{(n-k)^3}{n(n+1)^2}\left\|x_0 - x^*\right\|^2,
\end{aligned}
$$

in the last step of which we invoked the inequality we proved for $\|x_0 - x^*\|^2$.