Problem Set 6

Yung-Hsu Tsui[*][‡]

July 30 , 2018

Ex 9.1

An unconstrained linear objective function will have the form $A\mathbf{x} + \mathbf{b}$, where A is a $m \times n$ matrix and **x**,**b** are vectors in $\mathbf{R}^n$. The necessary condition for a local minima to exist is that the first derivative has to be equal to zero at that point.

The first derivative, of the linear function is **A**. If A is zero then the linear function can be written as **b**, which is a constant, and if **A** is not zero then the minima doesn't exist.

Ex 9.2

Note that $A^T A$ is symmetric and positive definite(We already proved this in the 3 4th weeks!). Thus, as long as x satisfies the first order necessary condition, it is global minimizer of the objective function, $x^T A^T Ax - 2b^T Ax$. Then, the first order condition is (we take derivative w.r.t x), then

$$2A^T Ax - 2A^T b = 0$$

This results in $A^T Ax = A^T b$.

Ex 9.3

We were taught the following methods:

- Steepest Descent
- Gradient Descent
- Newton

[*] University of Chicago, Master of Art Program in Social Science, 1126 E. 59th Street, Chicago, Illinois, 60637, (773) 702-5079, yhtsui@uchicago.edu.

[†] I thank Jayhyung Kim for his significant help.

[‡] Due to the lack of time, Excercise 9.6-9.9 are not completed.

- Quasi-Newton
- Conjugate Gradient

The gradient descent method takes smallest amount of computational power per step, but it takes a larger number of steps to converge on average. The steepest descent method is very fast but it requires calculating a further optimization problem to identify the optimal α. Newton methods converge in fewer steps, but the steps are much more computationally expensive as they require calculation of Hessian at each step.Quasi-Newton methods try to address this problem by not requiring to calculate the Hessian at each step. Conjugate gradient method is the hybrid of these two methods.It takes n steps to solve an unconstrained quadratic optimization problem, whereas the steepest descent & gradient descent may take more steps.

If the dimension is not too big, and especially if the function is differentiable we can use Newton's method. It's often a good idea to use steepest descent to get a better starting x_0 for Newton's method or when the function is not differentiable. If the dimensionality is very large, we are forced to resort to conjugate gradient.

Ultimately using these methods are an art and intuition rather than a science and there is no method to rule them all, especially for non-linear optimization problems.

Ex 9.4
Let D is the derivative of $f(x)^T$, and λ is the eigenvalue of Q with $Df(x_0)^T = Qx_0 - b$. Then,

$$\begin{aligned}
x_1 =& x_0 - \frac{DD^T}{DQD^T} D^T \\
=& x_0 - \frac{DD^T}{D\lambda D^T} D^T \\
=& x_0 - \frac{1}{\lambda} D^T \\
=& x_0 - Q^{-1} D^T \\
=& x_0 - Q^{-1}(Qx_0 - b) \\
=& Q^{-1} b
\end{aligned}$$

Ex 9.5
Note that the updating rule of the Gradient Method results in $f(x_{k+1}) = f(x_k + \alpha_k \Delta f(x_k))$. Note that x_{k+1} is obtained by minimizing $f(x_{k+1})$ w.r.t α_k.

Then,

$$\frac{df(x_{k+1})}{d\alpha_k} = \Delta f(x_{k+1}) \Delta f(x_k) = 0$$

Note that the updating rules of x_k indicates that

$$x_{k+1} - x_k = -\alpha_k \Delta f(x_k)$$

As long as $\Delta f(x_k)$ and $\Delta f(x_{k+1})$ are orthogonal to each other, $x_{k+1} - x_k$ and $x_{k+2} - x_{k+1}$ are orthogonal.

Ex 9.10
See the following updating rule of the Newton iteration:

$$\begin{aligned} x1 =& x0 - (D^2(f(x_0)))^{-1} Df(x_0)^T \\ =& x0 - Q^{-1}(Qx_0 - b) \\ =& x0 - x_0 + Q^{-1}b \\ =& Q^{-1}b \end{aligned}$$

Thus, the iteration does not depend on the initial value, and converges in one go.

Ex 9.12

$$\begin{aligned} Ax =& \lambda x \\ (B - \mu I)x =& \lambda x \\ Bx =& (\lambda + \mu)x \end{aligned}$$

Ex 9.15

$$\begin{aligned} & (A + BCD)\left[A^{-1} - A^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1}\right] \\ =& \left\{I - B\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1}\right\} + \left\{BCDA^{-1} - BCDA^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1}\right\} \\ =& \left\{I + BCDA^{-1}\right\} - \left\{B\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1} + BCDA^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1}\right\} \\ =& I + BCDA^{-1} - \left(B + BCDA^{-1}B\right)\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1} \\ =& I + BCDA^{-1} - BC\left(C^{-1} + DA^{-1}B\right)\left(C^{-1} + DA^{-1}B\right)^{-1}DA^{-1} \\ =& I + BCDA^{-1} - BCDA^{-1} \\ =& I. \end{aligned}$$

Ex 9.16

Substituting

$$x = \frac{v - As}{\|s\|^2}, y = s$$

into the Sherman-Morrison-Woodbury formula yields

$$\begin{aligned}A^{-1}-\frac{A^{-1}xy^TA^{-1}}{1+y^TA^{-1}x}=&A^{-1}-\frac{A^{-1}(v-As)/\|s\|^2s^TA^{-1}}{1+s^TA^{-1}(v-As)/\|s\|^2}\\=&A^{-1}-\frac{(A^{-1}v-s)s^TA^{-1}/\|s\|^2}{1+s^TA^{-1}v/\|s\|^2-1}\\=&A^{-1}+\frac{(s-A^{-1}v)s^TA^{-1}}{s^TA^{-1}v}\end{aligned}$$

Ex 9.18
Note that

$$\begin{aligned}\phi_k(\alpha)=&f(x_k+\alpha_kd_k)\\=&\frac{1}{2}x_k^TQx_k+\alpha_k(d^k)^TQx_k+\frac{\alpha_k^2}{2}(d^k)^TQd^k-x_k^Tb-\alpha_k(d^k)^Tb\end{aligned}$$

As long as Q is positive definite,i,e, $(d^k)^TQd^k$ is positive, we can find minimizer of this function with respect to α_k. Thus, taking derivative results in;

$$\begin{aligned}0=&\frac{\partial\phi_k(\alpha_k)}{\partial\alpha_k}\\=&\alpha_k(d^k)^TQd^k-(d^k)^Tb+(d^k)^TQx_k\end{aligned}$$

Then,

$$\begin{aligned}\alpha_k(d^k)^TQd^k=&(d^k)^Tb+(d^k)^TQx_k\\=&(d^k)^T(b-Qx_k)\\=&(d^k)^Tr_k\end{aligned}$$

Thus, dividing $(d^k)^TQd^k$ on both sides results in the answer.

Ex 9.20
Define the basis $W_i:=\{\tilde{r}^0,\tilde{r}^0,...,\tilde{r}^{i-1}\}$ where $\tilde{r}^k:=b-Qx^k$. By applying Gram-Schmit process, we can construct

$$r^k:=\tilde{r}^k-\sum_{j=0}^{k-1}\frac{(r^j,\tilde{r}^k)_Q}{\|r^j\|_A^2}r^j$$

Note that $W_i=span\{r^0,...,r^{i-1}\}$ because every r^i can be written as a linear combination of $\tilde{r}^0,...,\tilde{r}^i$. Also, note that with this setting we can deduce that the Conjugate Gradient method in each iteration is solving the following minimization problem;

$$\min_{x\in x^0+W_i}f(x)$$

where the minimizer of this problem is x^i.

Thus, the following problem results in $t = 0$ such that

$$\min_t f(x^i + t\tilde{r}^j)$$

with $j < i$.
Thus, the first order condition of this problem is

$$Df(x^i)\tilde{r}^j = 0$$

implying the following;

$$\begin{aligned} 0 =& Df(x^i)\tilde{r}^j \\ =& (Qx^i - b)^T\tilde{r}^j \\ =& -\tilde{r}^i\tilde{r}^j \end{aligned}$$

$\forall j < i$.