TMA4180
Optimization I
Spring 2020

Norwegian University of Science and Technology
Department of Mathematical Sciences

Solutions to exercise set 7

1 Assume that $A \in \mathbb{R}^{m\times n}$ with $m < n$ is a matrix of full rank and that $b \in \mathbb{R}^m \setminus \{0\}$. Consider the optimization problem

$$\frac{1}{2}\|x\|^2 \to \min \qquad \text{subject to} \qquad Ax = b. \tag{1}$$

a) Formulate the KKT-conditions for this problem and show that the unique solution is given by

$$x^* = A^T(AA^T)^{-1}b.$$

Solution: We are now considering the problem

$$\min_{x\in\mathbb{R}^n} f(x), \text{ s.t. } c(x) = 0,$$

where

$$f(x) = \frac{1}{2}x^T x \text{ and } c(x) = Ax - b,$$

with $b \neq 0$. The Lagrangian is

$$\mathcal{L}(x,\lambda) = \frac{1}{2}x^T x - \lambda^T(Ax - b),$$

where $\lambda \in \mathbb{R}^m$. The KKT conditions become

$$\begin{aligned} \nabla\mathcal{L}(x,\lambda) = x - A^T\lambda = 0,\\ Ax - b = 0. \end{aligned}$$

Also, since A has full rank, then the LICQ hold everywhere, meaning the KKT conditions are necessary for minimizers. We therefore look for solutions that satisfy the KKT conditions. If $\lambda = 0$, then $x = 0$ and $Ax = 0$, meaning $Ax - b \neq 0$, so we must have $\lambda \neq 0$. The first condition then gives $x = A^T\lambda$, and inserting this into the second gives $AA^T\lambda = b$. Since A has full rank, AA^T is invertible and we have $\lambda = (AA^T)^{-1}b$, meaning $x = A^T(AA^T)^{-1}b$. Also, since $\nabla^2\mathcal{L}(x,\lambda) = \nabla^2 f(x) = I$, which is positive definite, this is a minimum.

b) Formulate the quadratic penalty method for this constrained optimization problem, and show that the unique minimizer with parameter $\mu > 0$ is given by

$$x_\mu := A^T\Big(\frac{1}{\mu}I + AA^T\Big)^{-1}b$$

with $I \in \mathbb{R}^{m\times m}$ denoting the identity matrix.

Solution: The quadratic penalty method considers the unconstrained optimization of

$$g(x) = f(x) + \frac{\mu}{2} c(x)^T c(x),$$

which in our case becomes

$$g(x) = \frac{1}{2} x^T x + \frac{\mu}{2} (Ax - b)^T (Ax - b).$$

Taking the gradient of this, we get

$$\begin{aligned} \nabla g(x) = x + \mu(A^T Ax - A^T b) = 0 \\ \Rightarrow \left(\frac{1}{\mu} I + A^T A \right) x = A^T b \\ \Rightarrow x = \left(\frac{1}{\mu} I + A^T A \right)^{-1} A^T b. \end{aligned}$$

This is, however, not the expression we were looking for. We can easily see that

$$A^T \left(\frac{1}{\mu} I + AA^T \right) = \left(\frac{1}{\mu} I + A^T A \right) A^T.$$

Multiplying both sides from the left by $\left(\frac{1}{\mu} I + A^T A \right)^{-1}$ and from the right by $\left(\frac{1}{\mu} I + AA^T \right)^{-1}$, we see that

$$\left(\frac{1}{\mu} I + A^T A \right)^{-1} A^T = A^T \left(\frac{1}{\mu} I + AA^T \right)^{-1},$$

meaning that we get

$$x = A^T \left(\frac{1}{\mu} I + AA^T \right)^{-1} b.$$

Another way of arriving at the desired expression is by use of the singular value decomposition of A, writing $A = U \Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices ($U^{-1} = U^T$ and $V^{-1} = V^T$) and $\Sigma \in \mathbb{R}^{m \times n}$ is the matrix containing the singular values of A along its diagonal. The singular values are all positive. We will write $I_{r \times r}$ for an $r \times r$ identity matrix. Now, we observe that

$$\left(\frac{1}{\mu}I_{n\times n}+A^TA\right)^{-1}A^T=\left(\frac{1}{\mu}I_{n\times n}+(U\Sigma V^T)^TU\Sigma V^T\right)^{-1}(U\Sigma V^T)^T$$
$$=\left(\frac{1}{\mu}I_{n\times n}+V\Sigma^TU^TU\Sigma V^T\right)^{-1}V\Sigma^TU^T$$
$$=\left(\frac{1}{\mu}I_{n\times n}+V\Sigma^T\Sigma V^T\right)^{-1}V\Sigma^TU^T$$
$$=\left(V\left(\frac{1}{\mu}I_{n\times n}+\Sigma^T\Sigma\right)V^T\right)^{-1}V\Sigma^TU^T$$
$$=V\left(\frac{1}{\mu}I_{n\times n}+\Sigma^T\Sigma\right)^{-1}V^TV\Sigma^TU^T$$
$$=V\left(\frac{1}{\mu}I_{n\times n}+\Sigma^T\Sigma\right)^{-1}\Sigma^TU^T$$
$$=V\Sigma\left(\frac{1}{\mu}I_{m\times m}+\Sigma\Sigma^T\right)^{-1}U^T$$
$$=V\Sigma U^TU\left(\frac{1}{\mu}I_{m\times m}+\Sigma\Sigma^T\right)^{-1}U^T$$
$$=A^T\left(\frac{1}{\mu}I_{m\times m}+U\Sigma\Sigma^TU^T\right)^{-1}$$
$$=A^T\left(\frac{1}{\mu}I_{m\times m}+U\Sigma V^TV\Sigma^TU^T\right)^{-1}$$
$$=A^T\left(\frac{1}{\mu}I_{m\times m}+AA^T\right)^{-1}.$$

Thereby, we have $x_\mu=A^T\left(\frac{1}{\mu}I_{m\times m}+AA^T\right)^{-1}b$. The fact that

$$\left(\frac{1}{\mu}I_{n\times n}+\Sigma^T\Sigma\right)^{-1}\Sigma^T=\Sigma\left(\frac{1}{\mu}I_{m\times m}+\Sigma\Sigma^T\right)^{-1}$$

can be checked by writing the product componentwise.

c) Now consider the optimization problem

$$\frac{1}{2}\|x\|^2\to\min\qquad\text{subject to}\qquad\frac{1}{2}\|Ax-b\|^2\le\varepsilon$$

for some $\varepsilon>0$, and denote its solution by $\hat{x}_\varepsilon$. Show that either $\frac{1}{2}\|b\|^2\le\varepsilon$ (in which case $\hat{x}_\varepsilon=0$), or there exists $\mu>0$ such that $\hat{x}_\varepsilon=x_\mu$.

Solution: We now consider the problem

$$\min_{x\in\mathbb{R}^n}f(x),\ \text{s.t. } c(x)\ge 0,$$

where

$$f(x)=\frac{1}{2}x^Tx\text{ and }c(x)=\epsilon-\frac{1}{2}||Ax-b||^2,$$

The KKT conditions for this problem are

$$\begin{aligned}\nabla\mathcal{L}(x,\lambda) = x + \lambda(A^TAx - A^Tb) = 0\\ \lambda(\epsilon - \frac{1}{2}||Ax-b||^2) = 0\\ \epsilon - \frac{1}{2}||Ax-b||^2 \geq 0\\ \lambda \geq 0.\end{aligned}$$

With $\lambda = 0$, we get $x = 0$. For the third condition to hold, we must have $\epsilon \geq ||b||^2/2$. This is then a valid KKT point. Also, we have $\nabla^2\mathcal{L}(x,0) = I$, which is positive definite, so it is a minimum.
If $\lambda \neq 0$, we get, as in the previous exercise, that

$$\hat{x}_e = A^T\left(\frac{1}{\lambda}I_{m\times m} + AA^T\right)^{-1} b.$$

Here, λ must satisfy the condition that $\lambda > 0$ and λ must solve

$$\epsilon - \frac{1}{2}\left|\left|\left(AA^T\left(\frac{1}{\lambda}I_{m\times m} + AA^T\right)^{-1} - I_{m\times m}\right)b\right|\right|^2 = 0.$$

We can show that such a λ exists; since f is coercive and Ω is closed, there must exist a global minimizer. Since the LICQ holds everywhere, the KKT conditions are necessary for a minimum, and since, if $\epsilon < \frac{1}{2}||b||^2$, our candidate $\hat{x}_e$ is the only KKT point, it must be the global minimum, and thereby have a λ satisfying the above conditions. Thus, by taking $\mu = \lambda$, we get $\hat{x}_\epsilon = x_\mu$.

2 Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function, and let x_k and x_{k+1} $\mathbb{R}^n$ be two iterates of some iterative algorithm with $x_k \neq x_{k+1}$.

Define

$$s_k := x_{k+1} - x_k, \qquad y_k := \nabla f(x_{k+1}) - \nabla f(x_k).$$

The BFGS update for $H_k \approx (\nabla^2 f(x_k))^{-1}$ is

$$H_{k+1} = (I - \rho_k s_k y_k^\top)H_k(I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top,$$

where $\rho_k = \frac{1}{y_k^\top s_k}$.

a) Show that if H_k is symmetric positive definite and $y_k^\top s_k > 0$, then H_{k+1} is symmetric positive definite as well.
Hint: Try to show that $u^\top H_{k+1}u > 0$ for all nonzero u.
Solution: Define $V_k = I - \rho_k y_k s_k^\top$, Then $H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top$ and it is clear that H_{k+1} is symmetric if H_k is. Following the hint, we consider

$$\begin{aligned}u^\top H_{k+1}u =& u^\top V_k^\top H_k V_k u + \rho_s u^\top s_k s_k^\top u\\ =& (V_k u)^\top H_k (V_k u) + \rho_k (s_k^\top u)^2\end{aligned}$$

Since H_k is positive definite and $\rho_k = \frac{1}{y_k^\top s_k} \geq 0$, each of the two terms are nonnegative, and we can conclude that $u^\top H_{k+1}u \geq 0$ for all u. Furthermore, if $u^\top H_{k+1}u$ is equal to zero, then $s_k^\top u = 0$ and $V_k u = 0$. Finally $V_k u = u - \rho_k(s_k^\top u)y_k$, so $s_k^\top u =$ and $V_k u = 0$ implies $u = 0$. In conclusion, $u^\top H_{k+1}u > 0$ for all nonzero u, which implies that H_{k+1} is positive definite.

b) Show that if f is strongly convex[1] then $y_k^\top s_k > 0$ for any choice of nonequal x_k and x_{k+1}.
Solution: By Taylor's theorem, we have

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k) = \nabla^2 f(z)(x_{k+1} - x_k) = \nabla^2 f(z) s_k$$

where z is some point on the line segment joining x_k and x_{k+1}. Since f is strictly convex, we have

$$y_k^\top s_k = s_k^\top \nabla^2 f(z) s_k \geq m \|s_k\|^2 > 0.$$

(The last inequality is because $s_k \neq 0$).

[1] A C^2 function f is strongly convex if there exists an $m > 0$ such that $p^\top \nabla^2 f(x) p \geq m \|p\|^2$ for all $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$.