Tutorial 7 : Second-order methods - Newton algorithm

Exercise 1 (Newton method for quadratic functions)**.** Consider a quadratic function $f(x) = \frac{1}{2}x^\top \mathbf{A}x + b^\top x + c$, where $\mathbf{A} \succ 0$.

1. What are the minimizers of f?
2. What is the pure Newton step for the function f?
3. How many steps does it take for the Newton method to reach the minimizers of f?

Exercise 2 (Divergence of pure Newton steps)**.** If one only performs pure Newton steps, the iterates might diverge, even if the function is convex. Consider:

$$f(x) = -\log x + x$$

Answer the following questions:

1. Is the function f convex? Is the function f strongly convex? What is its minimizer?
2. Write the pure Newton step for the function f.
3. What is the set of initialization x_0 so that pure Newton steps can converge to its minimizer? What is the convergence speed? Can we apply the theoretical results proved in the lecture?
4. Bonus: Answer the same questions with the function $f(x) = \log(\exp(x) + \exp(-x))$.

Exercise 3 (BFGS algorithm)**.** The BFGS algorithm is a quasi-Newton method, i.e., it shares a similar model to the Newton algorithm. Indeed, at the kth iteration, BFGS constructs a quadratic surrogate of f of the form:

$$m_k(p) = f(x_k) + \nabla f(x_k)^\top p + \frac{1}{2}p^\top \mathbf{B}_k p.$$

where $\mathbf{B}_k \succ 0$. We will update x_{k+1} as:

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \mathbf{B}_k^{-1}\nabla f(x_k). \\ \mathbf{B}_{k+1} &= (\mathbf{I} - \rho_k y_k s_k^\top)\mathbf{B}_k(\mathbf{I} - \rho_k s_k y_k^\top) + \rho_k y_k y_k^\top \\ s_k &= x_{k+1} - x_k, \qquad y_k = \nabla f(x_{k+1}) - \nabla f(x_k), \qquad \rho_k = \frac{1}{y_k^\top s_k}. \end{aligned}$$

Answer the following questions:

1. Prove that if $\mathbf{B}_k \succ 0$, then $p_k = -\mathbf{B}_k^{-1}\nabla f(x_k)$ is a descent direction, i.e., $p_k^\top \nabla f(x_k) < 0$.

2. Assume the step-size α_k satisfies the Wolfe condition, i.e.,:

$$\nabla f(x_{k+1})^\top p_k \geq c_2 \nabla f(x_k)^\top p_k, \quad c_2 \in (0,1),$$

prove that $\rho_k > 0$.

3. Prove that if $\mathbf{B}_k \succ 0$, then $\mathbf{B}_{k+1} \succ 0$.

Exercise 4 (How to derive B_{k+1} in BFGS algorithm?)**.** In this exercise, we will prove that:

$$\mathbf{B}_{k+1} \in \underset{\mathbf{B}\succ 0}{\operatorname{argmin}} \{\|\mathbf{B} - \mathbf{B}_k\|_{\mathbf{W}} \mid \mathbf{B}s_k = y_k\}$$

where $\mathbf{W}$ is any positive definite matrix satisfying $\mathbf{W}y_k = s_k$ and $\|\mathbf{A}\|_{\mathbf{W}} = \|\mathbf{W}^{\frac{1}{2}}\mathbf{A}\mathbf{W}^{\frac{1}{2}}\|_F$. Answer the following questions:

1. Let $\mathbf{W}^{1/2}\mathbf{B}_k\mathbf{W}^{1/2} = \hat{\mathbf{B}}_k$, $\mathbf{W}^{1/2}\mathbf{B}\mathbf{W}^{1/2} = \hat{\mathbf{B}}$, $\mathbf{W}^{1/2}y_k = \hat{y}_k, \mathbf{W}^{-1/2}s_k = \hat{s}_k$. Prove that our optimization problem can be reformulated as:

$$\arg\min\{\|\hat{\mathbf{B}} - \hat{\mathbf{B}}_k\|_F \mid \hat{\mathbf{B}} \succ 0, \hat{\mathbf{B}}\hat{\mathbf{s}}_k = \hat{\mathbf{s}}_k\}. \tag{1}$$

2. Let $u = \frac{\hat{s}_k}{\|\hat{s}_k\|} \in \mathbb{R}^d$ and $u_\perp \in \mathbb{R}^{d\times(d-1)}$ be any orthonormal complement of u (i.e., $(u \quad u_\perp) \in \mathbb{R}^{d\times d}$ is a unitary matrix). Prove the optimality condition for problem (1) can be written as

$$u_\perp^\top \hat{\mathbf{B}} u_\perp = u_\perp^\top \hat{\mathbf{B}}_k u_\perp.$$

3. Prove that the optimal solution of (1) is

$$\hat{\mathbf{B}} = uu^\top + (\mathbf{I} - uu^\top)\hat{\mathbf{B}}_k(\mathbf{I} - uu^\top).$$

4. Recover the original formula.

Exercise 5 (Fast update rule for BFGS)**.** In the BFGS algorithm, we only care about the product of $\mathbf{B}_k^{-1}$ with a vector. Therefore, it is better to update $\mathbf{B}_{k+1}^{-1}$ from $\mathbf{B}_k^{-1}$ directly. Prove that:

1. Sherman - Morrison – Woobury Formula: if $\mathbf{A} = \mathbf{B} + \mathbf{U}\mathbf{C}\mathbf{V}^\top$, then

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}^\top\mathbf{B}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{B}^{-1}$$

2. Prove that:

$$\mathbf{B}_{k+1}^{-1} = \mathbf{B}_k^{-1} + \frac{1}{s_k^\top y_k} s_k s_k^\top - \frac{1}{y_k^\top \mathbf{B}_k^{-1} y_k} \mathbf{B}_k^{-1} y_k y_k^\top \mathbf{B}_k^{-1}.$$

Exercise 6 (Another version of BFGS)**.** In reality, the most implemented version of BFGS uses a slightly different update rule from the one described in the previous exercise. Indeed, they choose to update $\mathbf{H}_k$, the inverse of $\mathbf{B}_k$ as:

$$\mathbf{H}_{k+1} \in \underset{\mathbf{H}\succ 0}{\operatorname{argmin}} \{\|\mathbf{H} - \mathbf{H}_k\|_{\mathbf{W}} \mid \mathbf{H}y_k = s_k\}$$

where $\mathbf{W}$ is any positive definite matrix satisfying $\mathbf{W}s_k = y_k$. The descent direction is then $\mathbf{H}_k\nabla f(x_k)$ (and not $\mathbf{B}_k^{-1}\nabla f(x_k)$). Can you derive the update rule of $\mathbf{H}_k$?