Introduction to optimality conditions:
Optimality conditions for optimization over convex sets*

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

1 Special case: optimization over convex sets

Consider a constrained optimization problem

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x), \\ \text{subject to} \quad & x \in \Omega, \end{aligned} \tag{1}$$

where the *feasible set* $\Omega \subset \mathbb{R}^n$ is assumed to be non-empty and closed, and the *objective function* $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on Ω. Additionally, all the results in this note apart from Proposition 1, case 1, and Proposition 3, case 1, apply when Ω is *convex*.

Example 1. We have already encountered problems of the type (1) with a closed and convex Ω in this course:

- the trust-region subproblem

$$\begin{aligned} \underset{p\in\mathbb{R}^n}{\text{minimize}} \quad & m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p, \\ \text{subject to} \quad & \|p\| \le \Delta_k, \end{aligned}$$

  is of this class since the ball $\Omega = \{\, p \in \mathbb{R}^n \mid \|p\| \le \Delta_k \,\}$ is convex and closed;
- the problem leading to DFP/BFGS-type quasi-Newton update formulas:

$$\begin{aligned} \underset{B\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & f(B) = \frac{1}{2}\|B - B_k\|^2, \\ \text{subject to} \quad & B - B^T = 0, \\ & Bs_k = y_k, \end{aligned}$$

  where $B_k \in \mathbb{R}^{n\times n}$, $s_k, y_k \in \mathbb{R}^n$ are given, and $\|\cdot\|$ is typically a scaled Frobenius norm of the matrix. (Convince yourself that the set $\{\, B \in \mathbb{R}^{n\times n} \mid B^T = B, Bs_k = y_k \,\}$ is convex and closed. More generally, check that the solution set of any system of linear equations and non-strict inequalities is convex and closed.)

* Section 1 of this note is based on Section 4.4 in "Introduction to continuous optimization" by N. Andréasson, AE, M. Patriksson, E. Gustavsson, M. Önnheim: Studentlitteratur (2013), 2nd ed.

In the theory of unconstrained optimization we have started with the first order necessary optimality conditions, which could be succinctly stated as $\nabla f(x) = 0$ (see Theorem 2.2 in N&W). This statement expresses the fact that at any point of local minimum $x \in \mathbb{R}^n$, there should be no direction $p \in \mathbb{R}^n$, along which we could move and decrease the function (*direction of descent*). In the absence of constraints we are allowed to move along any direction, and thus we could always take the *steepest descent direction* $p = -\nabla f(x)$, unless the gradient vanishes at x.

The situation is drastically different in the presence of constraints. Indeed, if we take x to be on the boundary of Ω, that is $x \in \Omega \setminus \text{interior}(\Omega)$, then it may happen that we cannot take even the smallest step along some directions $p \in \mathbb{R}^n$ without leaving Ω, that is, $\forall$"small"$\delta > 0 : x + \delta p \notin \Omega$. Therefore these directions may still be directions of descent for f and yet not prevent $x \in \Omega$ from being a point of local minimum for f over Ω.

Definition 1 (Feasible direction). *Direction $p \in \mathbb{R}^n$ is a* feasible direction *at $x \in \Omega \subset \mathbb{R}^n$ if small steps along p do not take us outside of Ω.*

Formally, $p \in \mathbb{R}^n$ is a feasible direction at $x \in \Omega \subset \mathbb{R}^n$ if there is $\hat{\delta} > 0$ such that for all $0 < \delta < \hat{\delta}$ the inclusion $x + \delta p \in \Omega$ holds.

Exercise 1 (Feasible directions for linear constraints). Suppose that all equality and inequality constraints are *linear*, that is, $c_i(x) = a_i^T x + b_i$, $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $i \in \mathcal{I} \cup \mathcal{J}$, see eq. (12.1) in N&W. Show that the set of feasible directions $p \in \mathbb{R}^n$ at $x \in \Omega$ is

$$\{\, p \in \mathbb{R}^n \mid a_i^T p = 0, i \in \mathcal{E}, \quad a_i^T p \geq 0, i \in \mathcal{A}(x) \cap \mathcal{I} \,\},$$

where $\mathcal{A}(x)$ is the set of *active* or *binding* constraints at x, see Definition 12.1 in N&W.

With this definition we are ready to prove the following version of the first order optimality conditions for constrained problems.

Proposition 1 (First order necessary optimality conditions). *Consider a set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*.*

1. *Then for every feasible direction $p \in \mathbb{R}^n$ at x^* it holds that $\nabla f^T(x^*)p \geq 0$.*
2. *If, additionally, Ω is convex then*

$$\nabla f(x^*)^T(x - x^*) \geq 0, \quad \forall x \in \Omega. \tag{2}$$

Proof. 1. The proof is along the lines of Theorem 2.2 in N&W. For the sake of contradiction, let p be a feasible direction at x^* but $\nabla f(x^*)^T p < 0$. Owing to the continuity of $\nabla f(\cdot)$ around x^*, the same inequality holds at all nearby points, that is, $\nabla f(x^* + \delta p)^T p < 0$, for all $0 \leq \delta < \bar{\delta}$. Furthermore, the point $x_\delta = x^* + \delta p$ is feasible, for all $0 < \delta < \hat{\delta}$, owing to Definition 1.

Utilizing the first order Taylor series expansion we conclude that for every $0 < \delta < \min\{\bar{\delta}, \hat{\delta}\}$ we have the inclusion $x_\delta \in \Omega$ and the strict inequality

$$f(x_\delta) = f(x^*) + \nabla f(x^* + t_\delta \delta p)^T p < f(x^*),$$

where $0 < t_\delta < 1$. Since $x_\delta \to x^*$ as $\delta \to 0$ the point x^* cannot be a point of local minimum for f over Ω.

2. It suffices to show that $p_x = x - x^*$, $x \in \Omega$, is a feasible direction at x^*. Indeed, $\forall 0 < \lambda < 0$ we have

$$x^* + \lambda p_x = x^* + \lambda(x - x^*) = \lambda x + (1 - \lambda)x^* \in \Omega,$$

owing to the convexity of Ω.

□

As before, we will refer to the points satisfying the first order necessary conditions as the stationary points.

For a given point $\bar{x} \in \Omega$ let us consider the following linearization the original problem (1):

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & m(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}), \\ \text{subject to} & x \in \Omega. \end{array} \tag{3}$$

One can formulate the equivalent characterization of local optimality over convex sets in terms of the linearized problem (3).

Corollary 1. *Under the assumptions of Proposition 1, case 2, the locally optimal solution x^* to* (1) *must be a globally optimal solution to the linearized problem* (3)*, where the linearization point is $\bar{x} = x^*$.*

Proof. For any $x \in \Omega$ we have

$$m(x) = f(x^*) + \underbrace{\nabla f(x^*)^T (x - x^*)}_{\geq 0, \text{ see (2)}} \geq f(x^*) = m(x^*).$$

□

In fact the problem (3) has been utilized for algorithmic purposes. Indeed, since the linearization point $\bar{x}$ is feasible, it holds that any solution $\bar{x}^*$ to this convex optimization problem, whenever exists (for example when Ω is bounded), satisfies $m(\bar{x}^*) \leq m(\bar{x}) = f(\bar{x})$. Thus if we were able to find a solution to (3) such that $m(\bar{x}^*) < f(\bar{x})$ then the direction $p = \bar{x}^* - \bar{x}$ satisfies $\nabla f(\bar{x})^T p < 0$ and thus is a feasible descent direction for f at $\bar{x}$. One may then perform linesearch along such a direction and compute a better feasible point than $\bar{x}$. Such a step is the basis of some early first order algorithms for constrained optimization (Frank–Wolfe or conditional gradient algorithm). For the algorithm based on these ideas to be efficient the subproblem (3) should be easier to solve than the original problem (1). This is for example the case when Ω is a solution set for a system of linear equations and inequalities.

For convex objective functions the condition (2) is also sufficient for global optimality over Ω.

Proposition 2 (Necessary and sufficient optimality conditions for convex problems). *Suppose that $\Omega \subset \mathbb{R}^n$ is a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable on Ω. Then $x^* \in \Omega$ is a point of global minimum for f over Ω* if and only if (2) *holds.*

Proof. The necessity of (2) has been established in Proposition 1, thus is only remains to show sufficiency. For differentiable convex functions we can write, for any $x \in \Omega$,

$$f(x) \geq f(x^*) + \underbrace{\nabla f(x^*)^T (x - x^*)}_{\geq 0} \geq f(x^*),$$

where the first inequality is established in Proposition 3, "Basic tools" note. Thus f attains its smallest value over Ω at $x^* \in \Omega$. □

The *variational inequality* (2) may still be difficult to check in practical situations as it involves infinitely many inequalities, one for every $x \in \Omega$. It is possible to convert it into a system of equations, which may be easier to verify. In order to do this we need a concept of Euclidean projection.

Proposition 3. *Let $\Omega \subset \mathbb{R}^n$ be a non-empty and closed set, and $z \in \mathbb{R}^n$ be an arbitrary point. We consider the optimization problem*

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\textit{minimize}} & \dfrac{1}{2}\|x - z\|_2^2, \\ \textit{subject to} & x \in \Omega. \end{array} \tag{4}$$

1. *There is at least one globally optimal solution to* (4).
2. *If, additionally, the set Ω is convex, then such a solution is unique.*

Proof. 1. The triangle inequality implies that $\|x - z\|_2 \geq \|x\|_2 - \|z\|_2$, and thus the objective function in (4) is coercive. Therefore, the claim follows from Theorem 2 in the Note on "Lower semi-continuity, compactness, and existence of solutions."

2. Suppose that $x_1 \in \Omega$ and $x_2 \in \Omega$ are both closest to z; then for every $0 < \lambda < 1$ we have the inequality

$$\|z - [\lambda x_1 + (1 - \lambda)x_2]\|_2^2 \geq \|z - x_1\|_2^2 = \|z - x_2\|_2^2,$$

since $\lambda x_1 + (1 - \lambda)x_2 \in \Omega$. On the other hand,

$$\begin{aligned} &\|z - [\lambda x_1 + (1 - \lambda)x_2]\|_2^2 \\ &= \lambda^2 \|z - x_1\|_2^2 + 2\lambda(1 - \lambda)(z - x_1)^T (z - x_2) + (1 - \lambda)^2 \|z - x_2\|_2^2 \\ &\leq \lambda^2 \|z - x_1\|_2^2 + 2\lambda(1 - \lambda)\|z - x_1\|_2 \|z - x_2\|_2 + (1 - \lambda)^2 \|z - x_2\|_2^2 \\ &= \|z - x_1\|^2 = \|z - x_2\|^2, \end{aligned}$$

where the equality sign in the Cauchy–Schwarz inequality is only possible when $z - x_1 = \alpha(z - x_2)$ for some $\alpha \geq 0$. In view of $\|z - x_1\| = \|z - x_2\|$ either $\alpha = 1$ or $\|z - x_1\| = \|z - x_2\| = 0$. In either case, $x_1 = x_2$, as claimed. □

For a non-empty, closed, and convex Ω we will write $\pi_\Omega[z]$ to denote the unique solution of the problem (4).

Proposition 4 (First order necessary optimality conditions and gradient projection). *Consider a convex set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*. Then for every $\alpha > 0$ it holds that*

$$\pi_\Omega[x^* - \alpha \nabla f(x^*)] = x^*, \tag{5}$$

that is, x^* is a fixed point of the map $x \mapsto \pi_\Omega[x - \alpha \nabla f(x)]$.

Proof. Owing to Proposition 1 we know that (2) holds at x^*. Also, $x^* = \pi_\Omega[z]$ if and only if

$$\nabla_x \left[\frac{1}{2} \|x - z\|_2^2 \right] \Big|_{x=x^*}^T (x - x^*) = (x^* - z)^T (x - x^*) \geq 0, \quad \forall x \in \Omega, \tag{6}$$

owing to the convexity of the objective function in (4) and Proposition 2. It only remains to substitute $z = x^* - \alpha \nabla f(x^*)$ into (6) to conclude the proof. □

Exercise 2. When we discussed the solution of the trust-region problem, we have not established the necessity of the conditions (4.8) in N&W for the optimality, only their sufficiency. Utilizing Proposition 4 and the explicit characterization of the projection onto the trust-region show that every locally optimal solution to the trust-region problem must satisfy the first order conditions (4.8a)–(4.8b) in N&W for some non-negative scalar λ.

Exercise 3. Prove the following finite-dimensional geometric version of Hahn–Banach theorem: every non-empty closed convex set $\Omega \subset \mathbb{R}^n$ and a point $z \in \mathbb{R}^n \setminus \Omega$ may be separated with a hyperplane. That is, there is $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, such that

$$a^T z < b \leq a^T x, \quad \forall x \in \Omega.$$

Thus a is a normal for the said hyperplane. Hint: consider the optimality conditions (2) for the projection problem (4).

Proposition 4 may in rare circumstances be utilized for algorithmic purposes as well. Suppose that Ω is such that $\pi_\Omega[\cdot]$ is easy to evaluate. Further assume that it is possible to select $\alpha > 0$ such that the function $F(x) := \pi_\Omega[x - \alpha \nabla f(x)]$ is a *contraction*, that is, $\exists 0 < \gamma < 1$ such that $\forall x, y \in \Omega$ it holds that $\|F(x) - F(y)\| \leq \gamma \|x - y\|$. Then the Banach fixed point theorem tells us that the iteration $x_{k+1} = F(x_k)$ converges to the unique fixed point of $F(\cdot)$. By the construction of $F(\cdot)$, this fixed point satisfies the necessary optimality conditions (5).

Finally, let us consider yet another equivalent characterization of optimality over convex sets.

Definition 2 (Normal cone). *For a convex set Ω and $\bar{x} \in \Omega$ let us consider the set*

$$N_\Omega[\bar{x}] = \{\, q \in \mathbb{R}^n \mid q^T(x - \bar{x}) \le 0, \forall x \in \Omega \,\}, \tag{7}$$

that is, a set of directions forming an angle of at least $\pi/2$ with feasible directions $p = x - \bar{x}$ at $\bar{x}$. Often one defines $N_\Omega[\bar{x}] = \emptyset$ for $\bar{x} \notin \Omega$. $N_\Omega[\bar{x}]$ is called the normal cone for Ω at $\bar{x}$.

Exercise 4. Show that for any $\bar{x} \in \Omega$ the set $N_\Omega[\bar{x}]$ is non-empty, convex, closed, and is a cone. The latter property is defined as follows: $q \in N_\Omega[\bar{x}] \implies \alpha q \in N_\Omega[\bar{x}]$, for any $\alpha > 0$.

Proposition 5. *Consider a convex set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*. Then $-\nabla f(x^*) \in N_\Omega[x^*]$.*

Proof. $-\nabla f(x^*)^T(x - x^*) \le 0$, $\forall x \in \Omega$ by Proposition 1, case 1. □

2 Even more special case: optimization over hyperplanes

Consider now the situation when the set Ω is defined by the linear equality constraints only, i.e.

$$\Omega = \{\, x \in \mathbb{R}^n \mid a_i^T x = b_i, i \in \mathcal{E} \,\}. \tag{8}$$

In particular Ω is closed and convex, and therefore for a point $x^* \in \Omega$ to be a point of local minimum for a continuously differentiable function f it must satisfy the variational inequality (2).

Let $L = \{\, p \in \mathbb{R}^n \mid p = x - x^*, x \in \Omega \,\} = \Omega - x^*$, and $\hat{L} = \{\, p \in \mathbb{R}^n \mid a_i^T p = 0 \,\}$. Owing to the linearity of the constraints, we have the inclusions $\Omega - x^* \subset \hat{L}$ and $x^* + \hat{L} \subset \Omega$; thus $L = \hat{L}$. In particular, L is a linear subspace of $\mathbb{R}^n$. Therefore if $p \in L$ then also $-p \in L$, and thus both $\nabla f(x^*)^T p \ge 0$ and $\nabla f(x^*)^T(-p) \ge 0$. We conclude that $\nabla f(x^*)^T p = 0$ for all $p \in L$.

Let us now form a matrix A with rows a_i^T, $i \in \mathcal{E}$. Then L is precisely the nullspace of A, and we have shown that $\nabla f(x^*)$ is orthogonal to all vectors in $\mathrm{null}(A)$. Thus $\nabla f(x^*) \in \mathrm{null}(A)^\perp = \mathrm{range}(A^T)$ (basic linear algebra results). We have established the following characterization of points of local minimum.

Proposition 6 (First order necessary optimality conditions for linear equality constraints). *Consider a set Ω given by* (8) *and an arbitrary continuously differentiable function f. Then $x^* \in \Omega$ is a point of local minimum for f over Ω only if $\nabla f(x^*) \in span[a_i, i \in \mathcal{E}]$. That is, there is a vector of* Lagrange multipliers *$\lambda \in \mathbb{R}^{|\mathcal{E}|}$, where $|\mathcal{E}|$ is the number of equality constraints, such that*

$$\nabla f(x^*) = \sum_{i \in \mathcal{E}} \lambda_i a_i = \sum_{i \in \mathcal{E}} \lambda_i \nabla c_i. \tag{9}$$

We will see that the characterization of the type (9) can be further generalized to include non-linear equality and inequality constraints under some technical regularity conditions; such a generalization is known as the Karush–Kuhn–Tucker (KKT) theorem. Equation (9) is simply a specialization of KKT conditions for linear equality constraints, in the same way as the conditions (4.8a)–(4.8b) in N&W is the specialization of the KKT conditions for the trust-region subproblem with one inequality constraint $\|p\| \leq \Delta_k$.

Example 2. Using Proposition 6 we can derive DFP (or BFGS) quasi-Newton algorithm. To simplify the notation we define $C = B - B_k$ and until the end of the derivation we will omit the quasi-Newton iteration index k. Remember that $B_k = B_k^T$. Consider an arbitrary symmetric and positive definite matrix W and the optimization problem

$$\begin{aligned}
\underset{C \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & f(C) = \frac{1}{2}\text{trace}[WCW^TC^T] = \frac{1}{2}\sum_{\alpha=1}^{n}[CWCW^T]_{\alpha\alpha}, \\
\text{subject to} \quad & g_{ij}(C) = C_{ij} - C_{ji} = 0, \qquad i,j = 1,\ldots,n, \\
& h_i(C) = \sum_{j=1}^{n} C_{ij}s_j - \bar{y}_i = 0, \qquad i = 1,\ldots,n,
\end{aligned}$$

where $\bar{y}_i = y_i - \sum_{j=1}^{n}[B_k]_{ij}s_j$. Thus we have a minimization problem in n^2 optimization variables with $n^2 + n$ linear equality constraints. As a result, Proposition 6 applies. Let us denote the Lagrange multipliers corresponding to the matrix symmetry constraints with η_{ij}, $i,j = 1,\ldots,n$, and the ones corresponding to the secant equation with μ_i, $i = 1,\ldots,n$. It remains to calculate the derivatives of the objective function and the constraints with respect to B_{ij} (gradients of the constraints w.r.t. optimization variables result in the vectors a_i in the notation of Proposition 6).

$$
\begin{aligned}
[WCW^TC^T]_{\alpha\beta} &= \sum_{\gamma,\delta,\epsilon=1}^{n} W_{\alpha\gamma}C_{\gamma\delta}W_{\epsilon\delta}C_{\beta\epsilon}, \\
\frac{1}{2}\mathrm{trace}[WCW^TC^T] &= \frac{1}{2}\sum_{\alpha,\gamma,\delta,\epsilon=1}^{n} W_{\alpha\gamma}C_{\gamma\delta}W_{\epsilon\delta}C_{\alpha\epsilon}, \\
\frac{\partial f}{\partial C_{ij}} &= \underbrace{\frac{1}{2}\sum_{\alpha,\epsilon=1}^{n} W_{\alpha i}W_{\epsilon j}C_{\alpha\epsilon} + \frac{1}{2}\sum_{\gamma,\delta=1}^{n} W_{i\gamma}C_{\gamma\delta}W_{j\delta}}_{\text{remember: } W = W^T} \\
&= \sum_{\gamma,\delta=1}^{n} W_{i\gamma}C_{\gamma\delta}W_{\delta j} = [WCW]_{ij}, \\
\frac{\partial g_{k\ell}}{\partial C_{ij}} &= \begin{cases} 1, & \text{if } (i,j) = (k,\ell), \\ -1, & \text{if } (i,j) = (\ell,k), \\ 0, & \text{otherwise,} \end{cases} \\
\frac{\partial h_k}{\partial C_{ij}} &= \begin{cases} s_j, & \text{if } k = i, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_{k,\ell=1}^{n} \eta_{k\ell}\frac{\partial g_{k\ell}}{\partial C_{ij}} &= \eta_{ij} - \eta_{ji}, \\
\sum_{k=1}^{n} \mu_k \frac{\partial h_k}{\partial C_{ij}} &= \mu_i s_j,
\end{aligned}
$$

and the feasibility of C and optimality conditions (9) are expressed as

$$
\begin{aligned}
C_{ij} - C_{ji} &= 0, && i,j = 1,\dots,n, \\
\sum_{j=1}^{n} C_{ij}s_j &= \bar{y}_i, && i = 1,\dots,n, \\
[WCW]_{ij} &= \eta_{ij} - \eta_{ji} + \mu_i s_j, && i,j = 1,\dots,n.
\end{aligned}
$$

In matrix-vector notation we write:

$$
\begin{aligned}
C - C^T &= 0, \\
Cs &= \bar{y}, \\
WCW &= \eta - \eta^T + \mu s^T,
\end{aligned}
$$

where we have introduced a matrix η with elements η_{ij} and a vector μ with elements μ_i. Thus we end up with a system of $2n^2+n$ linear algebraic equations in $2n^2+n$ unknowns C, η, μ - it can easily be solved numerically, but in this case we can also solve it analytically with some manipulations.

Since W is positive definite (non-singular) symmetric, we can solve the last equation for C and then transpose it:

$$\begin{aligned} C &= W^{-1}(\eta - \eta^T + \mu s^T)W^{-1}, \\ C^T &= W^{-1}(\eta^T - \eta + s\mu^T)W^{-1}. \end{aligned}$$

Remebering that $C - C^T = 0$ we get

$$\begin{aligned} W^{-1}(2\eta - 2\eta^T + \mu s^T - s\mu^T)W^{-1} &= 0, \\ \eta - \eta^T &= \frac{1}{2}[s\mu^T - \mu s^T]. \end{aligned}$$

We can substitute the latter equation back into the expression for C to obtain

$$C = \frac{1}{2}W^{-1}(s\mu^T + \mu s^T)W^{-1},$$

We now use the secant equation:

$$\begin{aligned} \bar{y} = Cs &= \frac{1}{2}W^{-1}(s\mu^T + \mu s^T)W^{-1}s, \\ 2W\bar{y} &= (s\mu^T + \mu s^T)W^{-1}s, \\ \mu(s^TW^{-1}s) &= 2W\bar{y} - s\mu^TW^{-1}s, \\ \mu &= [2W\bar{y} - s\mu^TW^{-1}s]/(s^TW^{-1}s), \end{aligned}$$

which gives us μ if we know $\mu^TW^{-1}s = s^TW^{-1}\mu$. To obtain the latter quantity we proceed further:

$$\begin{aligned} W^{-1}\mu &= [2\bar{y} - W^{-1}s\mu^TW^{-1}s]/(s^TW^{-1}s), \\ s^TW^{-1}\mu &= 2s^T\bar{y}/(s^TW^{-1}s) - \mu^TW^{-1}s, \\ s^TW^{-1}\mu &= s^T\bar{y}/(s^TW^{-1}s). \end{aligned}$$

Finally, the last expression is used to obtain the final expression for μ:

$$\begin{aligned} \mu &= 2W\bar{y}/(s^TW^{-1}s) - ss^T\bar{y}/(s^TW^{-1}s)^2 \\ &= \frac{2}{s^TW^{-1}s}W\bar{y} - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}s \end{aligned}$$

Therefore, the final expression for C is:

$$\begin{aligned} C &= \frac{1}{2}W^{-1}\left[\frac{2}{s^TW^{-1}s}s\bar{y}^TW - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}ss^T + \frac{2}{s^TW^{-1}s}W\bar{y}s^T - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}ss^T\right]W^{-1} \\ &= \frac{1}{s^TW^{-1}s}W^{-1}s\bar{y}^T + \frac{1}{s^TW^{-1}s}\bar{y}s^TW^{-1} - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}W^{-1}ss^TW^{-1}. \end{aligned}$$

Note that this expression is valid for an arbitrary symmetric positive definite matrix W, and it defines the only stationary point for our optimization problem.

With some more linear algebra one can show that the optimization problem is convex (this is the only place where positive definiteness, and not just non-singularity and symmetry of W is utilized) so that this is the globally optimal solution to the problem thanks to Proposition 2.

We can now select W satisfying the equation $Wy = s$, so that $W^{-1}s = y$. This results in

$$C = \frac{1}{s^T y} y\bar{y}^T + \frac{1}{s^T y}\bar{y}y^T - \frac{s^T\bar{y}}{(s^T y)^2} yy^T.$$

We now recall that $B = B_k + C$, $s = s_k$, $y = y_k$, $\bar{y} = y_k - B_k s_k$, $B_k^T = B_k$:

$$\begin{aligned} B &= B_k + \frac{1}{s_k^T y_k} y_k[y_k - B_k s_k]^T + \frac{1}{s_k^T y_k}[y_k - B_k s_k]y^T - \frac{s_k^T[y_k - B_k s_k]}{(s_k^T y_k)^2} y_k y_k^T \\ &= B_k + \frac{1}{s_k^T y_k} y_k y_k^T - \frac{1}{s_k^T y_k} y_k s_k^T B_k - \frac{1}{s_k^T y_k} B_k s_k y_k^T + \frac{s_k^T B_k s_k}{(s_k^T y_k)^2} y_k y_k^T \\ &= \left[I - \frac{1}{s_k^T y_k} y_k s_k^T\right] B_k \left[I - \frac{1}{s_k^T y_k} s_k y_k^T\right] + \frac{1}{s_k^T y_k} y_k y_k^T, \end{aligned}$$

which is precisely formula (6.13) for DFP Hessian update in N&W.