Supplementary Materials for "A Study on Conjugate Gradient Stopping Criteria in Truncated Newton Methods for Linear Classification"

Leonardo Galli Marco Sciandrone Chih-Jen Lin

In this supplementary material we will first discuss a numerical issue on which we may come across to satisfy the angle condition below (1.2). In the second section we will focus on the local convergence analysis of the truncated Newton method obtained by employing different termination criteria for the inner conjugate method.

1 Angle condition

We recall here the definition of the Hessian and of the gradient of the objective function f

$$\begin{aligned} \boldsymbol{g} &:= \nabla f(\boldsymbol{w}) = \boldsymbol{w} + C\sum_{i=1}^{l} \xi'(y_i \boldsymbol{w}^\intercal \boldsymbol{x}_i) y_i \boldsymbol{x}_i, \\ H &:= \nabla^2 f(\boldsymbol{w}) = I + CX^\intercal DX, \end{aligned} \tag{1.1}$$

where I is the identity matrix, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l]^\intercal$ is the data matrix and D is a diagonal matrix with $D_{ii} = \xi''(y_i \boldsymbol{w}^\intercal \boldsymbol{x}_i)$. We also recall the angle condition

$$-\boldsymbol{g}_k^\intercal \boldsymbol{s}_k \geq c\|\boldsymbol{g}_k\|\|\boldsymbol{s}_k\| \tag{1.2}$$

that is always satisfied by the truncated Newton method defined in Section 2 of the original paper. Note in fact that global convergence is always obtained: the following discussion is only pointing out an undesired issue that might come out in practice. In the end, we recall the definition of the cosine between $-\boldsymbol{g}_k$ and $\boldsymbol{s}_k$

$$\cos(-\boldsymbol{g}_k, \boldsymbol{s}_k) = \frac{-\boldsymbol{g}_k^\intercal \boldsymbol{s}_k}{\|\boldsymbol{g}_k\|\|\boldsymbol{s}_k\|}. \tag{1.3}$$

When the Hessian of the objective function is positive definite and we approximate $\boldsymbol{s}_k \approx {H_k}^{-1}\boldsymbol{g}_k$ we have that

$$\begin{aligned} \boldsymbol{g}_k^\intercal \boldsymbol{s}_k &\approx -\boldsymbol{g}_k^\intercal {H_k}^{-1}\boldsymbol{g}_k \leq -\frac{\|\boldsymbol{g}_k\|^2}{\lambda_M(\boldsymbol{g}_k)} \\ \|\boldsymbol{s}_k\| &\approx \|H^{-1}\boldsymbol{g}_k\| \leq \frac{\|\boldsymbol{g}_k\|}{\lambda_m(H_k)}, \end{aligned}$$

where $\lambda_m(\cdot), \lambda_M(\cdot)$ respectively compute the minimum and the maximum eigenvalue of a matrix. If now we have that there exist two positive constants, M and m, such that $M \geq \lambda_M(\nabla^2 f(\boldsymbol{w}_k)) \geq \lambda_m(\nabla^2 f(\boldsymbol{w}_k)) \geq m$, from the inequalities above we also get that

$$-\boldsymbol{g}_k^\intercal \boldsymbol{s}_k \geq \frac{m}{M}\|\boldsymbol{g}_k\|\|\boldsymbol{s}_k\|. \tag{1.4}$$

For the linear classification problem, from (1.1) we have

$$1 \leq \lambda_m(H_k) \leq \lambda_M(H_k) \leq 1 + C\lambda_M(X^\intercal DX).$$

This means that when the regularization parameter C is large (e.g., $C = 100C_{\text{best}}$), (1.4) shows that the lower bound of the cosine value in (1.3) we can derive is smaller. Thus the angel condition (1.2) is harder to be satisfied. Note that equation (1.2) is a property needed in the convergence proof, which requires the existence of a constant c. in fact, there is no control on the right-hand-side of (1.2), while what we hope is that the left-hand-side of (1.2) is large so that it can be easily satisfied. Unfortunately (1.4) shows that for large C, the left-hand-side of (1.2) tends to be not that large.

In conclusion, even if theoretically convergence is not an issue, the above discussion still gives us a hint on what might actually happen when C is large.

2 Local Convergence

In this section we will derive local convergence properties for all the combination of *ratios* and forcing sequences presented in the original paper. We first recall all the *ratios*:

- residual:

$$ratio = \frac{\|\boldsymbol{g}_k + H_k \boldsymbol{s}_k^j\|}{\|\boldsymbol{g}_k\|}. \tag{2.5}$$

- residual_{l1}:

$$ratio = \frac{\|\boldsymbol{g}_k + H_k \boldsymbol{s}_k^j\|_1}{\|\boldsymbol{g}_k\|_1}. \tag{2.6}$$

- quadratic:

$$\begin{aligned} Q(\boldsymbol{s}) &:= \boldsymbol{s}^\intercal \boldsymbol{g}_k + \frac{1}{2}\boldsymbol{s}^T H_k \boldsymbol{s}, \\ ratio &= \frac{(Q_j - Q_{j-1})}{Q_j/j}, \end{aligned} \tag{2.7}$$

where $Q_j := Q(\boldsymbol{s}_k^j)$ and $Q_{j-1} := Q(\boldsymbol{s}_k^{j-1})$.

Now we recall all the forcing sequences:

- adaptive:

$$\eta_k = \min\{c_1; c_2\|\boldsymbol{g}_k\|^{c_3}\}, \tag{2.8}$$

η_k / *ratio*	(2.8)	(2.9)	(2.10)
(2.5)	SL $(1+c_3)$	SL $(1+c_3)$	L
(2.6)	SL $(1+c_3)$	SL $(1+c_3)$	L
(2.7)	SL $(1+\frac{c_3}{2})$	SL $(1+\frac{c_3}{2})$	L

Table 1: In this table we distinguish between the termination criteria that are able to obtain q-Super-Linear (SL) convergence from those that only achieve q-Linear (L) convergence. In addition, we specify which order a SL convergence is able to obtain.

- adaptive$_{l1}$:

$$\eta_k = \min\{c_1; c_2\|\boldsymbol{g}_k\|_1^{c_3}\}, \tag{2.9}$$

- constant:

$$\eta_k = c_0 \qquad \text{with } c_0 \in (0,1). \tag{2.10}$$

In Table 1 we present all the local convergence results and in the following we will show how to obtain them. The basic convergence results are given by Theorem 2.2 and 3.3 from [1]. We recall the statement here, see [1] for the proof.

Theorem 2.1. *If $\boldsymbol{w}_0$ is sufficiently close to $\boldsymbol{w}^*$, the ratio employed is* (2.5) *and $0 \le \eta_k < \eta_{\max} < 1$, then $\{\boldsymbol{w}_k\}$ converges q-linearly to $\boldsymbol{w}^*$. If $\lim_{k\to\infty} \eta_k = 0$ then the convergence is q-superlinear. If $\eta_k = O(\|\boldsymbol{g}_k\|^{c_3})$ with $0 < c_3 \le 1$, then the convergence is of order at least $(1+c_3)$.*

From this result we can fill two cells of Table 1, in particular those in which *ratio* is (2.5) and η_k are (2.8) and (2.10). Now thanks to the fact that $\|\boldsymbol{v}\|_1 \le \sqrt{n}\|\boldsymbol{v}\|_2$, for any vector $\boldsymbol{v}$ of dimension n, if we employ (2.9) in η_k we still have that $\eta_k = O(\|\boldsymbol{g}_k\|^r)$. Thus, if the *ratio* is (2.5) and η_k is (2.9) we still obtain superlinear convergence of order $(1+c_3)$.

If we now employ (2.6) we also need the other side of the norm equivalence. In particular, thanks to $\|\boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_1 \le \sqrt{n}\|\boldsymbol{v}\|_2$ we get that

$$\begin{aligned} \|H_k\boldsymbol{s}_k^j + \boldsymbol{g}_k\|_2 &\le \|H_k\boldsymbol{s}_k^j + \boldsymbol{g}_k\|_1 \\ &\le \eta_k\|\boldsymbol{g}_k\|_1 \\ &\le \sqrt{n}\eta_k\|\boldsymbol{g}_k\|_2. \end{aligned} \tag{2.11}$$

Thus, employing (2.6) as the *ratio*, thanks to (2.11), it is possible to obtain same local convergence results as the one achieved by using (2.5).

To fill the last row of Table 1, we need some additional results. The following derivation is based on the results that can be found in Section 4.3 from [3]. Assume that iteration k of the truncated Newton method computes a step $\boldsymbol{s}_k$ that satisfies

$$\frac{(Q_j - Q_{j-1})}{Q_j/j} \le \eta_k \tag{2.12}$$

for a specified value of η_k. Thus from now on, *ratio* is (2.7) and η_k is unspecified. Thanks to CG method properties, we have that Q_j is monotonically decreasing and

$$Q_1 = -\frac{1}{2}\frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T H_k \boldsymbol{g}_k} < 0.$$

Thus, we have that $(Q_j - Q_{j-1}) < 0$ and $Q_j < 0$ for every $j \ge 1$. For this reason (2.12) is equivalent to

$$Q_{j-1} - Q_j \le \eta_k \cdot \frac{-Q_j}{j}. \tag{2.13}$$

From now on, for vectors and matrices $\|\cdot\|$ indicate the l2-norm.

Lemma 2.1. *If G is symmetric and positive-definite then*

$$y^T G^2 y \le \|G\| y^T G y. \tag{2.14}$$

Proof. We have

$$\begin{aligned} y^T G^2 y &= (G^{\frac{1}{2}}y)^T G (G^{\frac{1}{2}}y) \\ &\le \lambda_{\max}(G^{\frac{1}{2}}y)^T(G^{\frac{1}{2}}y) \\ &= \|G\| y^T G y, \end{aligned}$$

where $\lambda_{\max}$ is the highest eigenvalue of G. □

Lemma 2.2. *Let $\boldsymbol{s}^*$ be the point that minimizes $Q(\boldsymbol{s})$. Then, for any $\boldsymbol{s}$ we have*

$$\|H_k\boldsymbol{s} + \boldsymbol{g}_k\|^2 \le 2\|H_k\| \cdot (Q(\boldsymbol{s}) - Q(\boldsymbol{s}^*)) \tag{2.15}$$

Proof. From Lemma 2.1 we have

$$\begin{aligned} \|H_k\boldsymbol{s} + \boldsymbol{g}_k\|^2 &= (H_k\boldsymbol{s} + \boldsymbol{g}_k)^T(H_k\boldsymbol{s} + \boldsymbol{g}_k) \\ &= (H_k\boldsymbol{s} - H_k\boldsymbol{s}^*)^T(H_k\boldsymbol{s} - H_k\boldsymbol{s}^*) \\ &= (\boldsymbol{s} - \boldsymbol{s}^*)^T H_k^2 (\boldsymbol{s} - \boldsymbol{s}^*) \\ &\le \|H_k\|(\boldsymbol{s}^T H_k \boldsymbol{s} - 2\boldsymbol{s}^{*T} H_k \boldsymbol{s} + \boldsymbol{s}^{*T} H_k \boldsymbol{s}^*) \\ &= \|H_k\|\Big((\boldsymbol{s}^T H_k \boldsymbol{s} + 2\boldsymbol{s}^T\boldsymbol{g}_k) \\ &\quad - (\boldsymbol{s}^{*T} H_k \boldsymbol{s}^* + 2\boldsymbol{s}^{*T}\boldsymbol{g}_k)\Big) \\ &= 2\|H_k\|(Q(\boldsymbol{s}) - Q(\boldsymbol{s}^*)), \end{aligned}$$

where second and fourth equalities follow from $\boldsymbol{g}_k = -H_k\boldsymbol{s}^*$. □

LEMMA 2.3. *If we employ* (2.12) *as the inner stopping criterion of the CG procedure, we get*

$$\|H_k\boldsymbol{s}_k^j + \boldsymbol{g}_k\| \le L\sqrt{\eta_k}\|\boldsymbol{g}_k\|, \tag{2.16}$$

where $L = \sqrt{\frac{K\|H_k\|\cdot\|H_k^{-1}\|}{1-K}}$.

Proof. From (6.18) of [2] we get that

$$K(Q(\boldsymbol{s}_k^j) - Q(\boldsymbol{s}_k^*)) \ge Q(\boldsymbol{s}_k^{j+1}) - Q(\boldsymbol{s}_k^*),$$

where $K = \left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2$, $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the highest and lowest eigenvalues. Thus, moving $KQ(\boldsymbol{s}_k^*)$ on the right side and subtracting $KQ(\boldsymbol{s}_k^{j+1})$ from both sides of the above inequality we get

$$\begin{aligned}&KQ(\boldsymbol{s}_k^j) - KQ(\boldsymbol{s}_k^{j+1}) \ge \\ &-KQ(\boldsymbol{s}_k^{j+1}) + KQ(\boldsymbol{s}_k^*) + Q(\boldsymbol{s}_k^{j+1}) - Q(\boldsymbol{s}_k^*) = \\ &(1-K)\left(Q(\boldsymbol{s}_k^{j+1}) - Q(\boldsymbol{s}_k^*)\right),\end{aligned}$$

which means that

$$Q(\boldsymbol{s}_k^{j+1}) - Q(\boldsymbol{s}_k^*) \le \frac{K}{1-K}\left(Q(\boldsymbol{s}_k^j) - Q(\boldsymbol{s}_k^{j+1})\right). \tag{2.17}$$

Now since $Q(\boldsymbol{s}_k^j)$ is monotonically decreasing as j increases we get that

$$-Q(\boldsymbol{s}_k^*) \ge -Q(\boldsymbol{s}_k^j) \quad \forall j. \tag{2.18}$$

At the solution $\boldsymbol{s}_k^*$, from $\boldsymbol{s}_k^* = -H_k^{-1}\boldsymbol{g}_k$, we have

$$Q(\boldsymbol{s}_k^*) = \boldsymbol{s}_k^{*T}\boldsymbol{g}_k + \frac{1}{2}\boldsymbol{s}_k^{*T}H_k\boldsymbol{s}_k^* = -\frac{1}{2}\boldsymbol{g}_k^T H_k^{-1}\boldsymbol{g}_k.$$

Thus, together with (2.18), we get

$$-Q(\boldsymbol{s}_k^j) \le -Q(\boldsymbol{s}_k^*) \le \frac{1}{2}\|H_k^{-1}\|\cdot\|\boldsymbol{g}_k\|^2. \tag{2.19}$$

Finally from Lemma 2.2, (2.17), (2.13), (2.19) we get

$$\begin{aligned}\|H_k\boldsymbol{s}_k^j + \boldsymbol{g}_k\|^2 &\le 2\|H_k\|\cdot(Q(\boldsymbol{s}_k^j) - Q(\boldsymbol{s}_k^*)) \\ &\le \frac{2K\|H_k\|}{1-K}\left(Q(\boldsymbol{s}_k^{j-1}) - Q(\boldsymbol{s}_k^j)\right) \\ &\le \eta_k\cdot\frac{2K\|H_k\|}{1-K}\cdot\frac{-Q(\boldsymbol{s}_k^j)}{j} \\ &\le \eta_k\cdot\frac{K\|H_k\|\cdot\|H_k^{-1}\|}{1-K}\cdot\|\boldsymbol{g}_k\|^2 \\ &= \eta_k\cdot L^2\cdot\|\boldsymbol{g}_k\|^2,\end{aligned}$$

where $L^2 = \frac{K\|H_k\|\cdot\|H_k^{-1}\|}{1-K}$ and the last inequality follows from the fact that $j \ge 1$. ∎

Now thanks to (2.16) we can extend Theorem 2.1 to obtain Theorem 2.2. In fact, if in (2.16) we define $\bar{\eta}_k := L\sqrt{\eta_k}$ we obtain again (2.5)

$$\|H_k\boldsymbol{s}_k^j + \boldsymbol{g}_k\| \le \bar{\eta}_k\|\boldsymbol{g}_k\|,$$

where this time the forcing sequence used in the original termination criteria is under the sign of a root. Thus, we can apply Theorem 2.1 to prove the following.

THEOREM 2.2. *If* $\boldsymbol{w}_0$ *is sufficiently close to* $\boldsymbol{w}^*$, *the ratio employed is* (2.7) *and* $0 \le \eta_k < \eta_{\max} < 1$, *then* $\{\boldsymbol{w}_k\}$ *converges q-linearly to* $\boldsymbol{w}^*$. *If* $\lim_{k\to\infty}\eta_k = 0$ *then the convergence is q-superlinear. If* $\eta_k = O(\|\nabla f(\boldsymbol{w}_k)\|^{c_3})$ *with* $0 < c_3 \le 2$, *then the convergence is of order at least* $(1+\frac{c_3}{2})$.

Note that the only difference with Theorem 2.1 is the fact that in the new bound that correlates the norm of the residual to the norm of the gradient (2.16), the forcing sequence is under the sign of a root. For this reason when we use (2.7), to obtain the same order of local convergence we now need to use an exponent c_3 which is the double of the one that we would have needed by using (2.5). For the rest, the last row of Table 1 can be filled in the same way as (2.5).

3 Complete Experimental Results

See `https://www.csie.ntu.edu.tw/~cjlin/papers/inner_stopping`.

References

[1] R. S. Dembo and S. C. Eisenstat and T. Steihaug, *Inexact newton methods*, SIAM Journal on Numerical analysis, 8 (1982), pp. 400–408.

[2] M. R. Hestenes, *Conjugate direction methods in optimization*, Springer-Verlag, Berlin, (1978).

[3] S. G. Nash, *Truncated-Newton methods*, Report STAN-CS-82-906, Computer Science Department, Stanford University, (1982).