8 Supplementary Material

8.1 Proof of Lemma 4

We factorize and bound \( \|H_{n,k}^{-1} \nabla R_n(x_m) - H_n^{-1} \nabla R_n(x_m)\| \) as

\[
\|H_{n,k}^{-1} \nabla R_n(x_m) - H_n^{-1} \nabla R_n(x_m)\| \leq (21) \\
\|I - H_{n,k}^{-1} H_n\| \|H_n^{-1} \nabla R_n(x_m)\|. 
\]

Thus, it remains to bound \( \|I - H_{n,k}^{-1} H_n\| \) by some \( \epsilon_n \). To do so, consider that we can factorize \( H_n = U(\Sigma + cV_n)I^T \) and \( H_{n,k}^{-1} \) as in (8). We can then expand \( \|I - H_{n,k}^{-1} H_n\| \) as

\[
\|I - H_{n,k}^{-1} H_n\| = \|I - U([\Sigma_k + cV_n]I)^{-1} \times ([\Sigma_k + cV_n]I)U^T \| 
\]

where \( \Sigma_k \in \mathbb{R}^{p \times p} \) is the truncated eigenvalue matrix \( \Sigma_k \) with zeros padded for the last \( p-k \) diagonal entries. Observe that the first \( k \) entries of the product \( ([\Sigma_k + cV_n]I)^{-1} \times ([\Sigma_k + cV_n]I) \) are equal to 1, while the last \( p-k \) entries are equal to \((\mu_j + cV_n)/cV_n) \). Thus, we have that

\[
\|I - H_{n,k}^{-1} H_n\| = \|U[H_{n,k}^{-1} - I]U^T\| = \left| \frac{\mu_{k+1}}{cV_n} \right|. 
\]

(23)

8.2 Proof of Lemma 5

To begin, recall the result from Lemma 4 in (18). From this, we use the following result from [25, Lemma 6], which present here as a lemma.

Lemma 6 Consider the k-TAN step where

\[
\|H_{n,k}^{-1} \nabla R_n(x_m) - H_n^{-1} \nabla R_n(x_m)\| \leq \epsilon_n \|H_n^{-1} \nabla R_n(x_m)\|. 
\]

The Newton decrement of the k-TAN iterate \( \lambda_n(x_n) \) is bounded by

\[
\lambda_n(x_n) \leq \left\{ \left( 1 + \epsilon_n \right) \lambda_n(x_m)^2 + \epsilon_n \lambda_n(x_m) \right\} \left( 1 - (1 + \epsilon_n) \lambda_n(x_m)^2 \right) \\
\text{w.h.p.} 
\]

(24)

Lemma 6 provides a bound on the Newton decrement of the iterate \( x_n \) computed from the k-TAN update in [6] in terms of Newton decrement of the previous iterate \( x_m \) and the error \( \epsilon_n \) incurred from the truncation of the Hessian. We proceed in a manner similar to [16, Proposition 4] by finding upper and lower bounds for the sub-optimality \( S_n(x) = R_n(x) - R_n(x^*) \) in terms of the Newton decrement parameter \( \lambda_n(x) \). Consider the result from [22, Theorem 4.1.11],

\[
\lambda_n(x) - \ln (1 + \lambda_n(x)) \leq R_n(x) - R_n(x^*) \leq -\lambda_n(x) - \ln (1 - \lambda_n(x)). 
\]

Consider the Taylor’s expansion of \( \ln(1 + a) \) for \( a = \lambda_n(x) \) to obtain the lower bound on \( \lambda_n(x) \),

\[
\lambda_n(x) \geq \ln (1 + \lambda_n(x)) + \frac{1}{2} \lambda_n(x)^2 - \frac{1}{3} \lambda_n(x)^3. 
\]

(26)

Assume that \( x \) is such that \( 0 < \lambda_n(x) < 1/4 \). Then the expression in (26) can be rearranged and bounded as

\[
\frac{1}{6} \lambda_n(x)^2 \leq \frac{1}{2} \lambda_n(x) - \frac{1}{3} \lambda_n(x)^3 
\]

(27)

Now, consider the Taylor’s expansion of \( \ln(1 - a) \) for \( a = \lambda_n(x) \) in a similar manner to obtain for \( \lambda_n(x) < 1/4 \), from [5, Chapter 9.6.3],

\[
-\lambda_n(x) - \ln (1 - \lambda_n(x)) \leq \lambda_n(x)^2 \quad \text{(28)}
\]

Using these bounds with the inequalities in (25) we obtain the upper and lower bounds on \( S_n(x) \) as

\[
\frac{1}{6} \lambda_n(x)^2 \leq S_n(x) \leq \lambda_n(x)^2.
\]

(29)

Now, consider the bound for Newton decrement of the k-TAN iterate \( \lambda_n(x_n) \) from (24). As we assume that \( \lambda_n(x_m) < 1/4 \), we have

\[
\lambda_n(x_n) \leq \frac{4}{(3 - \epsilon_n)^2} \left[ (1 + \epsilon_n) \lambda_n(x_m)^2 + \lambda_n(x_m) \epsilon_n \right].
\]

(30)

We substitute this back into the upper bound in (29) for \( x = x_n \) to obtain

\[
S_n(x_n) \leq \lambda_n(x_n)^2 \leq \frac{16}{(3 - \epsilon_n)^4} \left[ (1 + \epsilon_n) \lambda_n(x_m)^2 + \lambda_n(x_m) \epsilon_n \right]^2 \\
= \frac{16}{(3 - \epsilon_n)^4} \left[ (1 + \epsilon_n) \lambda_n(x_m)^4 \right] \\
+ 2 \epsilon_n (1 + \epsilon_n) \lambda_n(x_m)^3 + \epsilon_n^2 \lambda_n(x_m)^2. 
\]

(32)

Consider also from (29) that we can upper bound the Newton decrement as \( \lambda_n(x_m)^2 \leq 6S_n(x_m) \). We plug this back into (32) to obtain a final bound for sub-optimality as

\[
S_n(x_n) \leq \frac{16}{(3 - \epsilon_n)^4} \left[ 36 (1 + \epsilon_n)^2 S_n(x_m)^2 \right] \quad \text{(33)}
\]

\[ + 30 \epsilon_n (1 + \epsilon_n) S_n(x_m)^3/2 + 6 \epsilon_n^2 S_n(x_m)]. \]

8.3 Additional Experiments

In Figure 5, we show results on the BIO dataset used for protein homology classification in KDD Cup 2004. The dimensions are \( N = 145751 \) and \( p = 74 \). In this setting, the number of samples is very large and the dimension is very small. Observe in Figure 5 that both k-TAN and AdaNewton greatly outperform the first order methods, due to the reduced cost in Hessian computation that comes from adaptive sample size. However, because \( p \) is small, the additional gain from the truncating in the inverse in k-TAN does not provide significant benefit relative to AdaNewton.
Figure 5: Convergence of $k$-TAN, AdaNewton, SGD, and SAGA in terms of number of processed gradients (left) and runtime (right) for the BIO protein homology classification problem.