8 Supplementary Material

8.1 Proof of Lemma 4

We factorize and bound $\|\hat{\mathbf{H}}_{n,k}^{-1}\nabla R_n(\mathbf{x}_m) - \mathbf{H}_n^{-1}\nabla R_n(\mathbf{x}_m)\|$ as

$$\|\hat{\mathbf{H}}_{n,k}^{-1}\nabla R_n(\mathbf{x}_m) - \mathbf{H}_n^{-1}\nabla R_n(\mathbf{x}_m)\| \leq \|\mathbf{I} - \hat{\mathbf{H}}_{n,k}^{-1}\mathbf{H}_n\|\|\mathbf{H}_n^{-1}\nabla R_n(\mathbf{x}_m)\|. \tag{21}$$

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

$$\|\mathbf{I} - \hat{\mathbf{H}}_{n,k}^{-1}\mathbf{H}_n\| = \|\mathbf{I} - \mathbf{U}[(\hat{\mathbf{\Sigma}}_k + cV_n\mathbf{I})^{-1} \times (\mathbf{\Sigma} + cV_n\mathbf{I})]\mathbf{U}^T\|, \tag{22}$$

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

$$\|\mathbf{I} - \hat{\mathbf{H}}_{n,k}^{-1}\mathbf{H}_n\| = \left|\frac{\mu_{k+1}}{cV_n}\right|. \tag{23}$$

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 $\|\hat{\mathbf{H}}_{n,k}^{-1}\nabla R_n(\mathbf{x}_m) - \mathbf{H}_n^{-1}\nabla R_n(\mathbf{x}_m)\| \leq \epsilon_n\|\mathbf{H}_n^{-1}\nabla R_n(\mathbf{x}_m)\|$. The Newton decrement of the k-TAN iterate $\lambda_n(\mathbf{x}_n)$ is bounded by*

$$\lambda_n(\mathbf{x}_n) \leq \frac{\left[(1+\epsilon_n)\lambda_n(\mathbf{x}_m)^2 + \epsilon_n\lambda_n(\mathbf{x}_m)\right]}{(1-(1+\epsilon_n)\lambda_n(\mathbf{x}_m))^2} \quad w.h.p \tag{24}$$

Lemma 6 provides a bound on the Newton decrement of the iterate $\mathbf{x}_n$ computed from the k-TAN update in (6) in terms of Newton decrement of the previous iterate $\mathbf{x}_m$ and the error ϵ_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 $S_n(\mathbf{x}) = R_n(\mathbf{x}) - R_n(\mathbf{x}_n^*)$ in terms of the Newton decrement parameter $\lambda_n(\mathbf{x})$. Consider the result from [22, Theorem 4.1.11],

$$\begin{aligned}\lambda_n(\mathbf{x}) - \ln(1+\lambda_n(\mathbf{x})) &\leq R_n(\mathbf{x}) - R_n(\mathbf{x}_n^*) \\ &\leq -\lambda_n(\mathbf{x}) - \ln(1-\lambda_n(\mathbf{x}))\,.\end{aligned} \tag{25}$$

Consider the Taylor's expansion of $\ln(1+a)$ for $a = \lambda_n(\mathbf{x})$ to obtain the lower bound on $\lambda_n(\mathbf{x})$,

$$\lambda_n(\mathbf{x}) \geq \ln(1+\lambda_n(\mathbf{x})) + \frac{1}{2}\lambda_n(\mathbf{x})^2 - \frac{1}{3}\lambda_n(\mathbf{x})^3. \tag{26}$$

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

$$\frac{1}{6}\lambda_n(\mathbf{x})^2 \leq \frac{1}{2}\lambda_n(\mathbf{x})^2 - \frac{1}{3}\lambda_n(\mathbf{x})^3 \tag{27}$$

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

$$-\lambda_n(\mathbf{x}) - \ln(1-\lambda_n(\mathbf{x})) \leq \lambda_n(\mathbf{x})^2 \tag{28}$$

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

$$\frac{1}{6}\lambda_n(\mathbf{x})^2 \leq S_n(\mathbf{x}) \leq \lambda_n(\mathbf{x})^2. \tag{29}$$

Now, consider the bound for Newton decrement of the k-TAN iterate $\lambda_n(\mathbf{x}_n)$ from (24). As we assume that $\lambda_n(\mathbf{x}_m) < 1/4$, we have

$$\lambda_n(\mathbf{x}_n) \leq \frac{4}{(3-\epsilon_n)^2}\left[(1+\epsilon_n)\lambda_n(\mathbf{x}_m)^2 + \lambda_n(\mathbf{x}_m)\epsilon_n\right]. \tag{30}$$

We substitute this back into the upper bound in (29) for $\mathbf{x} = \mathbf{x}_n$ to obtain

$$S_n(\mathbf{x}_n) \leq \lambda_n(\mathbf{x}_n)^2 \tag{31}$$

$$\begin{aligned}&\leq \frac{16}{(3-\epsilon_n)^4}\left[(1+\epsilon_n)\lambda_n(\mathbf{x}_m)^2 + \lambda_n(\mathbf{x}_m)\epsilon_n\right]^2 \\ &= \frac{16}{(3-\epsilon_n)^4}[(1+\epsilon_n)^2\lambda_n(\mathbf{x}_m)^4 \\ &\quad + 2\epsilon_n(1+\epsilon_n)\lambda_n(\mathbf{x}_m)^3 + \epsilon_n^2\lambda_n(\mathbf{x}_m)^2].\end{aligned} \tag{32}$$

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

$$\begin{aligned}S_n(\mathbf{x}_n) \leq \frac{16}{(3-\epsilon_n)^4}&[36(1+\epsilon_n)^2S_n(\mathbf{x}_m)^2 \\ &+ 30\epsilon_n(1+\epsilon_n)S_n(\mathbf{x}_m)^{3/2} + 6\epsilon_n^2S_n(\mathbf{x}_m)].\end{aligned} \tag{33}$$

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 put the problem 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.

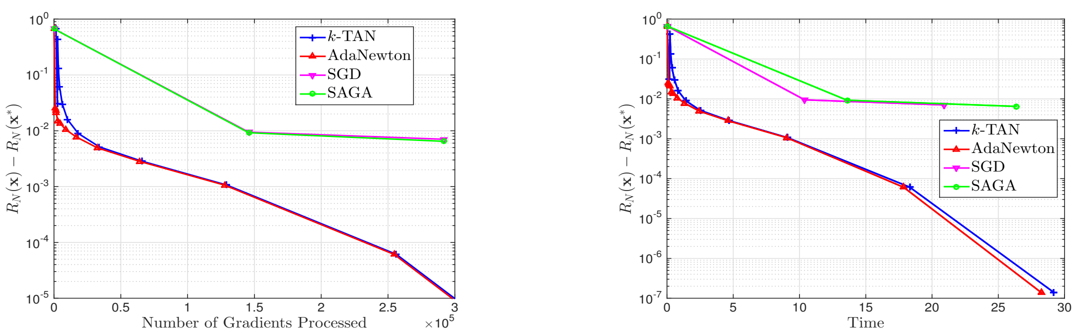

Figure 5: Convergence of k-TAN, AdaNewton, SGD, and SAGA in terms of number of processed gradients (left) and runtime (right) for the BIO protein homology classification problem.