Supplementary Materials: Limited-memory Common-directions Method With Subsampled Newton Directions for Large-scale Linear Classification

Jui-Nan Yen
National Taiwan University
juinanyen@gmail.com

Chih-Jen Lin
National Taiwan University
cjlin@csie.ntu.edu.tw

A. Appendix

A. Proof of Theorem 1

To prove Theorem 1, we need the following lemma.

Lemma 1: Let Assumption 1 hold. Let $\mathbf{\Delta}_k = \mathbf{w}^* - \mathbf{w}_k$. Then for $H^* \equiv \nabla^2 f(\mathbf{w}^*)$, we have

$$\|\mathbf{g}_k + H^*\mathbf{\Delta}_k\|_2 \leq \frac{\hat{L}}{2}\|\mathbf{\Delta}_k\|_2^2.$$

Similarly, for H_k we have

$$\|\mathbf{g}_k + H_k\mathbf{\Delta}_k\|_2 \leq \frac{\hat{L}}{2}\|\mathbf{\Delta}_k\|_2^2.$$

Proof of Lemma 1

We follow the proof of Lemma 9 of Wang et al. [1].

First, we can write $\mathbf{g}_k \equiv \nabla f(\mathbf{w}_k)$ as

$$\begin{aligned}&\mathbf{g}_k = \nabla f(\mathbf{w}_k) - \nabla f(\mathbf{w}^*)\\ =&(\int_0^1 \nabla^2 f(\mathbf{w}^* + \tau(\mathbf{w}_k - \mathbf{w}^*))d\tau)(\mathbf{w}_k - \mathbf{w}^*)\\ =&-(\int_0^1 \nabla^2 f(\mathbf{w}^* + \tau(\mathbf{w}_k - \mathbf{w}^*))d\tau)\mathbf{\Delta}_k,\end{aligned} \tag{a.1}$$

where the first equality follows from the fact that $\nabla f(\mathbf{w}^*) = \mathbf{0}$.

Then from (a.1), it follows that

$$\begin{aligned}&\|\mathbf{g}_k + H^*\mathbf{\Delta}_k\|_2\\ =&\|[\nabla^2 f(\mathbf{w}^*) - \int_0^1 \nabla^2 f(\mathbf{w}^* + \tau(\mathbf{w}_k - \mathbf{w}^*))d\tau]\mathbf{\Delta}_k\|_2\\ \leq&\|\mathbf{\Delta}_k\|_2\| \int_0^1 [\nabla^2 f(\mathbf{w}^*) - \nabla^2 f(\mathbf{w}^* + \tau(\mathbf{w}_k - \mathbf{w}^*))]d\tau\|_2\\ \leq&\|\mathbf{\Delta}_k\|_2 \int_0^1 \|\nabla^2 f(\mathbf{w}^*) - \nabla^2 f(\mathbf{w}^* + \tau(\mathbf{w}_k - \mathbf{w}^*))\|_2 d\tau\\ \leq&\|\mathbf{\Delta}_k\|_2 \int_0^1 \tau\hat{L}\|\mathbf{w}_k - \mathbf{w}^*\|_2 d\tau\\ =&\frac{\hat{L}}{2}\|\mathbf{\Delta}_k\|_2^2.\end{aligned}$$

Similarly, we have $\|\mathbf{g}_k + H_k\mathbf{\Delta}_k\|_2 \leq \frac{\hat{L}}{2}\|\mathbf{\Delta}_k\|_2^2$, which completes the proof.

B. Proof of Theorem 1

Our proof is similar to Lemma 9 of Wang et al. [1]. But instead of $\bar{\nu}q_k(\mathbf{\Delta}_k/\bar{\nu})$, we use $\bar{\nu}q_k(\mathbf{\Delta}_k)$ in (a.4).

First, from some simple calculations, we have

$$\begin{aligned}&\frac{1}{2}(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})^T H_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ =&\frac{1}{2}(-\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})^T H_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ &+ \mathbf{\Delta}_k^T H_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ =&\frac{1}{2}\mathbf{\Delta}_{k+1}^T H_k\mathbf{\Delta}_{k+1} - \frac{1}{2}\mathbf{\Delta}_k^T H_k\mathbf{\Delta}_k\\ &+ \mathbf{\Delta}_k^T H_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1}).\end{aligned} \tag{a.2}$$

Then recall the definition,

$$q_k(\mathbf{s}) = \frac{1}{2}\mathbf{s}^T H_k\mathbf{s} + \mathbf{g}_k^T\mathbf{s}.$$

Substituting (a.2) into the equation, we have

$$\begin{aligned}&q_k(\alpha_k\mathbf{u}_k)\\ =&q_k(\mathbf{w}_{k+1} - \mathbf{w}_k) = q_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ =&\frac{1}{2}(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})^T H_k(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ &+ \mathbf{g}_k^T(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1})\\ =&\frac{1}{2}\mathbf{\Delta}_{k+1}^T H_k\mathbf{\Delta}_{k+1} - \frac{1}{2}\mathbf{\Delta}_k^T H_k\mathbf{\Delta}_k\\ &+ (H_k\mathbf{\Delta}_k + \mathbf{g}_k)^T(\mathbf{\Delta}_k - \mathbf{\Delta}_{k+1}).\end{aligned} \tag{a.3}$$

Furthermore, from our assumption, we have

$$\nu(\alpha_k, \mathbf{u}_k) = \frac{q_k(\alpha_k\mathbf{u}_k)}{q_k(\mathbf{s}_k)} \geq \bar{\nu}.$$

Data sets	#instances	#features	sparsity	$\log_2(C_{\text{Best}})$
epsilon_normalized	400,000	2,000	1	3
HIGGS	11,000,000	28	0.92	-6
rcv1_test	677,399	47,236	1.5e-3	3
news20	19,996	1,355,191	3.4e-4	9
webspam_trigram	350,000	16,609,143	2.2e-4	2
yahoojp	176,203	832,026	1.6e-4	3
yahookr	460,554	3,052,939	1.1e-4	5
url_combined	2,396,130	3,231,961	3.6e-5	-4
avazu-site	25,832,830	999,962	1.5e-5	-5
kdda	8,407,752	20,216,830	1.8e-6	-4
kddb	19,264,097	29,890,095	9.8e-7	-2
kdd12	149,639,105	54,686,452	2.0e-7	-6

TABLE I: Data set statistics ordered by sparsity (#nnz/#instances/#features). C_{Best} is the regularization parameter selected by cross validation. It is worth noticing that for some sparse data sets there are large number of features not appearing in any of the training instances.

Thus,

$$\begin{aligned}
&q_k(\alpha_k \mathbf{u}_k) \le \bar{\nu} q_k(\mathbf{s}_k) \le \bar{\nu} q_k(\boldsymbol{\Delta}_k) \\
=&\bar{\nu}(\frac{1}{2}\boldsymbol{\Delta}_k^T H_k \boldsymbol{\Delta}_k + \mathbf{g}_k^T \boldsymbol{\Delta}_k) \\
=&\frac{\bar{\nu}}{2}(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k + \frac{\bar{\nu}}{2}\mathbf{g}_k^T\boldsymbol{\Delta}_k \\
=&\frac{\bar{\nu}-1}{2}(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k + \frac{\bar{\nu}-1}{2}\mathbf{g}_k^T\boldsymbol{\Delta}_k \\
&+ (H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k - \frac{1}{2}\boldsymbol{\Delta}_k^T H_k \boldsymbol{\Delta}_k.
\end{aligned} \tag{a.4}$$

Combing (a.3) and (a.4), we have

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\Delta}_{k+1}^T H_k \boldsymbol{\Delta}_{k+1} \\
\le&\frac{1}{2}\boldsymbol{\Delta}_k^T H_k \boldsymbol{\Delta}_k - (H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T(\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_{k+1}) \\
&+ \frac{\bar{\nu}-1}{2}(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k + \frac{\bar{\nu}-1}{2}\mathbf{g}_k^T\boldsymbol{\Delta}_k \\
&+ (H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k - \frac{1}{2}\boldsymbol{\Delta}_k^T H_k \boldsymbol{\Delta}_k \\
=&(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_{k+1} + \frac{\bar{\nu}-1}{2}(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k \\
&+ \frac{\bar{\nu}-1}{2}\mathbf{g}_k^T\boldsymbol{\Delta}_k \\
=&\frac{1-\bar{\nu}}{2}\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k + \frac{\bar{\nu}-1}{2}(H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k \\
&+ \frac{\bar{\nu}-1}{2}(H^*\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_k + (H_k\boldsymbol{\Delta}_k + \mathbf{g}_k)^T\boldsymbol{\Delta}_{k+1} \\
\le&\frac{1-\bar{\nu}}{2}\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k + \frac{(1-\bar{\nu})}{2}\frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^3 \\
&+ \frac{(1-\bar{\nu})}{2}\frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^3 + \frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^2\|\boldsymbol{\Delta}_{k+1}\|_2,
\end{aligned} \tag{a.5}$$

where the last inequality follows from Lemma 1.

From Assumption 1 and (a.5), we then have

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\Delta}_{k+1}^T H^* \boldsymbol{\Delta}_{k+1} \\
\le&\frac{1}{2}\boldsymbol{\Delta}_{k+1}^T H_k \boldsymbol{\Delta}_{k+1} + \frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2\|\boldsymbol{\Delta}_{k+1}\|_2^2 \\
\le&\frac{1-\bar{\nu}}{2}\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k + \frac{(1-\bar{\nu})\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^3 \\
&+ \frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^2\|\boldsymbol{\Delta}_{k+1}\|_2 + \frac{\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2\|\boldsymbol{\Delta}_{k+1}\|_2^2.
\end{aligned} \tag{a.6}$$

Furthermore, since

$$\|\boldsymbol{\Delta}_k\|_2\|\boldsymbol{\Delta}_{k+1}\|_2 \le \frac{1}{2}\|\boldsymbol{\Delta}_k\|_2^2 + \frac{1}{2}\|\boldsymbol{\Delta}_{k+1}\|_2^2,$$

we have

$$\|\boldsymbol{\Delta}_k\|_2^2\|\boldsymbol{\Delta}_{k+1}\|_2 \le \frac{1}{2}\|\boldsymbol{\Delta}_k\|_2^3 + \frac{1}{2}\|\boldsymbol{\Delta}_k\|_2\|\boldsymbol{\Delta}_{k+1}\|_2^2. \tag{a.7}$$

From (a.6) and (a.7), it follows that

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\Delta}_{k+1}^T H^* \boldsymbol{\Delta}_{k+1} \\
\le&\frac{1-\bar{\nu}}{2}\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k + \frac{(1-\bar{\nu})\hat{L}}{2}\|\boldsymbol{\Delta}_k\|_2^3 \\
&+ \frac{\hat{L}}{4}\|\boldsymbol{\Delta}_k\|_2^3 + \frac{3\hat{L}}{4}\|\boldsymbol{\Delta}_k\|_2\|\boldsymbol{\Delta}_{k+1}\|_2^2 \\
\le&\frac{1-\bar{\nu}}{2}\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k + \frac{(3-2\bar{\nu})\hat{L}}{4\lambda_{\min}(H^*)^{\frac{3}{2}}}(\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k)^{\frac{3}{2}} \\
&+ \frac{3\hat{L}}{4\lambda_{\min}(H^*)^{\frac{3}{2}}}(\boldsymbol{\Delta}_k^T H^* \boldsymbol{\Delta}_k)^{\frac{1}{2}}(\boldsymbol{\Delta}_{k+1}^T H^* \boldsymbol{\Delta}_{k+1}).
\end{aligned} \tag{a.8}$$

This states that when $\|\boldsymbol{\Delta}_k\|_2$ is small enough, $\frac{1}{2}\boldsymbol{\Delta}_{k+1}^T H^* \boldsymbol{\Delta}_{k+1}$ will converge linearly with rate $1-\bar{\nu}$, which completes the proof.

B. Proof of Proposition 1

To solve the maximization in (6), we write down the first-order condition for $\alpha^* \equiv \operatorname{argmin}_\alpha q_k(\alpha \mathbf{u})$,

$$\alpha^* \mathbf{u}^T H \mathbf{u} + \mathbf{g}^T \mathbf{u} = 0.$$

This tells us

$$\alpha^* = \frac{-\mathbf{g}^T\mathbf{u}}{\mathbf{u}^T H \mathbf{u}}. \tag{a.9}$$

Plugging (a.9) into $q_k(\alpha^*\mathbf{u})$ gives us

$$\begin{aligned} q_k(\alpha^*\mathbf{u}) &= \frac{1}{2}\alpha^{*2}\mathbf{u}^T H\mathbf{u} + \alpha^*\mathbf{g}^T\mathbf{u} \\ &= -\frac{1}{2}\frac{(\mathbf{g}^T\mathbf{u})^2}{\mathbf{u}^T H\mathbf{u}} = -\frac{1}{2}\frac{(\mathbf{s}^T H\mathbf{u})^2}{\mathbf{u}^T H\mathbf{u}}, \end{aligned} \tag{a.10}$$

where the second equality comes from (4).

Furthermore, also from (4), we have

$$q_k(\mathbf{s}) = \frac{1}{2}\mathbf{s}^T H\mathbf{s} + \mathbf{g}^T\mathbf{s} = -\frac{1}{2}\mathbf{s}^T H\mathbf{s}. \tag{a.11}$$

Combining (a.11) and (a.10), we have

$$\mu(\mathbf{u}) = \frac{q_k(\alpha^*\mathbf{u})}{q_k(\mathbf{s})} = \frac{(\mathbf{s}^T H\mathbf{u})^2}{(\mathbf{u}^T H\mathbf{u})(\mathbf{s}^T H\mathbf{s})}. \tag{a.12}$$

From (a.12) and (8), one can see that we have

$$\tau_H(\mathbf{u},\mathbf{s})^2 = \frac{(\mathbf{s}^T H\mathbf{u})^2}{(\mathbf{u}^T H\mathbf{u})(\mathbf{s}^T H\mathbf{s})} = \mu(\mathbf{u}),$$

which completes the proof.

C. Proof of Theorem 2

For the first part, we have

$$\min_{i,\alpha} q_k(\alpha\mathbf{p}_i) = \min_{i,\alpha} q_k(P_k(\alpha\mathbf{e}_i)) \geq \min_{\mathbf{t}} q_k(P_k\mathbf{t}),$$

where $\mathbf{e}_i$ denotes the vector with a 1 in the ith coordinate and 0's elsewhere. Since $q_k(\mathbf{s}) < 0$, this tells us

$$\begin{aligned} \max_i \mu(\mathbf{p}_i) &= \frac{\min_{i,\alpha} q_k(\alpha\mathbf{p}_i)}{q_k(\mathbf{s})} \\ \leq &\frac{\min_{\mathbf{t}} q_k(P_k\mathbf{t})}{q_k(\mathbf{s})} = \mu(P_k), \end{aligned}$$

which completes the proof.

For the second part, we first define

$$M \equiv (P_k^T H P_k). \tag{a.13}$$

By our assumption, the columns of P_k are linearly independent, so we have $P_k\mathbf{u} \neq \mathbf{0}$ for all $\mathbf{u} \neq \mathbf{0}$. This combined with the fact that H is positive definite tells us that

$$\mathbf{u}^T M\mathbf{u} = \mathbf{u}^T(P_k^T H P_k)\mathbf{u} = (\mathbf{u}^T P_k^T)H(P_k\mathbf{u}) > 0.$$

This means M is also positive definite, and thus invertible.

For $\mathbf{t}^* = \operatorname{argmin}_{\mathbf{t}} q_k(P_k\mathbf{t})$, it should satisfy the first-order condition

$$(P_k^T H P_k)\mathbf{t}^* + P_k^T\mathbf{g} = \mathbf{0}.$$

As a result, we have

$$q_k(P_k\mathbf{t}^*) = -\frac{1}{2}(P_k^T\mathbf{g})^T M^{-1}(P_k^T\mathbf{g}). \tag{a.14}$$

Similarly, for $\alpha_i^* = \operatorname{argmin}_{\alpha} q_k(\alpha\mathbf{p}_i)$, we have

$$(\mathbf{p}_i^T H\mathbf{p}_i)\alpha_i^* + \mathbf{p}_i^T\mathbf{g} = 0,$$

and

$$\min_{\alpha} q_k(\alpha\mathbf{p}_i) = -\frac{1}{2}(\mathbf{p}_i^T\mathbf{g})^T(M_{i,i})^{-1}(\mathbf{p}_i^T\mathbf{g}).$$

As a result, we have

$$\operatorname*{mean}_i[\min_{\alpha} q_k(\alpha\mathbf{p}_i)] = -\frac{1}{2}\frac{(P_k^T\mathbf{g})^T D^{-1}(P_k^T\mathbf{g})}{m}, \tag{a.15}$$

where $D \equiv \operatorname{diag}(M)$.

From (a.14) and (a.15), it follows that

$$\begin{aligned} &\frac{q_k(P_k\mathbf{t}^*)}{\operatorname{mean}_i[\min_{\alpha} q_k(\alpha\mathbf{p}_i)]} \\ =&\frac{(D^{-\frac{1}{2}}P_k^T\mathbf{g})^T(D^{\frac{1}{2}}M^{-1}D^{\frac{1}{2}})(D^{-\frac{1}{2}}P_k^T\mathbf{g})}{(D^{-\frac{1}{2}}P_k^T\mathbf{g})^T(D^{-\frac{1}{2}}P_k^T\mathbf{g})}m \\ \geq& m\lambda_{\min}(G^{-1}) = m/\lambda_{\max}(G), \end{aligned} \tag{a.16}$$

where $G \equiv D^{-\frac{1}{2}}MD^{-\frac{1}{2}}$.

By the definition of G, D, M, and (8), we have

$$G_{i,j} = \frac{M_{i,j}}{M_{i,i}^{\frac{1}{2}}M_{j,j}^{\frac{1}{2}}} = \frac{\mathbf{p}_i^T H\mathbf{p}_j}{\|H^{\frac{1}{2}}\mathbf{p}_i\|_2\|H^{\frac{1}{2}}\mathbf{p}_j\|_2} = \tau_H(\mathbf{p}_i,\mathbf{p}_j).$$

Especially, we have $G_{i,i} = 1$. The Gershgorin circle theorem and the definition of ζ in (11) give us

$$\begin{aligned} &\lambda_{\max}(G) \leq \max_i \Sigma_{j=1}^m |G_{ij}| \\ =&1 + (m-1)\max_i \operatorname*{mean}_{j:j\neq i}[|G_{ij}|] = 1 + (m-1)\zeta. \end{aligned}$$

This combined with (a.16) gives us

$$\frac{q_k(P_k\mathbf{t}^*)}{\operatorname{mean}_i[\min_{\alpha} q_k(\alpha\mathbf{p}_i)]} \geq \frac{m}{1+(m-1)\zeta} = \Gamma. \tag{a.17}$$

As a result,

$$\begin{aligned} \mu(P_k) &= \frac{\min_{\mathbf{t}} q_k(P_k\mathbf{t})}{q_k(\mathbf{s})} = \frac{q_k(P_k\mathbf{t}^*)}{q_k(\mathbf{s})} \\ \geq&\Gamma \operatorname*{mean}_i[\frac{\min_{\alpha} q_k(\alpha\mathbf{p}_i)}{q_k(\mathbf{s})}] = \Gamma \operatorname*{mean}_i \mu(\mathbf{p}_i), \end{aligned}$$

where the first equality comes from $q_k(\mathbf{s}) < 0$ and the inequality comes from (a.17). This completes our proof.

D. Proof of Theorem 3

To prove Theorem 3, we need the following lemma and its corollary.

Lemma 2: Given $0 \leq \beta \leq 1$ and $\mathbf{a}, \mathbf{b}, \mathbf{c} \in R^n$, which satisfy

$$a_i b_i > 0 \qquad and \qquad c_i = \frac{1}{\beta/a_i + (1-\beta)/b_i} \tag{a.18}$$

for all i, we have

$$\frac{|\mathbf{a}^T\mathbf{b}|}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} \leq \min\{\frac{|\mathbf{a}^T\mathbf{c}|}{\|\mathbf{a}\|_2\|\mathbf{c}\|_2}, \frac{|\mathbf{b}^T\mathbf{c}|}{\|\mathbf{b}\|_2\|\mathbf{c}\|_2}\}. \tag{a.19}$$

Proof of Lemma 2

First, we show that it is sufficient to prove the result for rational a_i, b_i, c_i and β. For real a_i, b_i, c_i, and β, since they are respectively limits of rational a_i, b_i, c_i, and β, by the fact that all the math components used are continuous, the result in (a.19) still holds.

We can further assume that c_i are integers, as multiplying $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ with a constant does not affect the result.

Next, we show that we can further assume

$$a_i > 0,\ b_i > 0 \quad \text{and} \quad c_i = \frac{1}{\beta/a_i + (1-\beta)/b_i} = 1 \quad \text{(a.20)}$$

for all i without loss of generality. We prove this by showing that if there exists a counterexample $\mathbf{a}, \mathbf{b}, \mathbf{c}$ which satisfies (a.18) but not (a.19), then we can construct a counterexample $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\mathbf{c}}$ which also satisfies (a.20).

Assume that $\mathbf{a}, \mathbf{b}, \mathbf{c}$ is a counterexample, where c_i are integers for all i. We can define a helper function $h : \{0, \ldots, n\} \to Z$ (the set of integers),

$$h(x) = \Sigma_{k=1}^{x} c_k^2.$$

and a matrix $Q \in R^{h(n)\times n}$, where

$$Q_{ij} = \begin{cases} 1/c_j, & \text{if } h(j-1) < i \le h(j) \\ 0, & \text{otherwise.} \end{cases} \quad \text{(a.21)}$$

We then construct $\tilde{\mathbf{a}} = Q\mathbf{a}, \tilde{\mathbf{b}} = Q\mathbf{b}, \tilde{\mathbf{c}} = Q\mathbf{c}$.

We will show that this construction satisfies (a.20). From (a.18), we know that a_j, b_j and c_j should have the same sign. Thus from (a.18) and (a.21), we know that

$$\tilde{a}_i = \frac{a_j}{c_j} > 0,$$

$$\tilde{b}_i = \frac{b_j}{c_j} > 0,$$

and

$$\tilde{c}_i = \frac{1}{c_j/c_j} = 1 = \frac{1}{c_j(\beta/a_j + (1-\beta)/b_j)} = \frac{1}{\beta/\tilde{a}_i + (1-\beta)/\tilde{b}_i},$$

so $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\mathbf{c}}$ satisfy (a.20).

Furthermore, from (a.21), we have

$$Q^T Q = I.$$

Thus for any vectors $\mathbf{u}$ and $\mathbf{v}$, we have

$$\mathbf{u}^T\mathbf{v} = (Q\mathbf{u})^T(Q\mathbf{v}),$$

which means we have $\|\tilde{\mathbf{a}}\|_2 = \|\mathbf{a}\|_2$, $|\tilde{\mathbf{a}}^T\tilde{\mathbf{b}}| = |\mathbf{a}^T\mathbf{b}|$, etc.

Consequently, if $\mathbf{a}, \mathbf{b}, \mathbf{c}$ is a counterexample for Lemma 2, then $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\mathbf{c}}$ should also be a counterexample for Lemma 2. Therefore, in the rest of the proof, we can assume that (a.20) holds.

From (a.20), we have

$$c_i = \frac{1}{\beta/a_i + (1-\beta)/b_i} = 1,$$

which means if a_i are sorted in the ascending order, then b_i are sorted in the descending order.

As a result, according to the rearrangement inequality, for any permutation σ, we have

$$\Sigma_i a_{\sigma(i)} b_i \ge \Sigma_i a_i b_i.$$

This gives us

$$\frac{(\Sigma_i a_i)(\Sigma_i b_i)}{n} = \frac{\Sigma_i \Sigma_j a_{g(i,j)} b_j}{n} \ge \Sigma_i a_i b_i, \quad \text{(a.22)}$$

where $g(i,j) = ((i+j) \mod n) + 1$.

Furthermore, since x^2 is a convex function, we have

$$\sqrt{\frac{\Sigma_i b_i^2}{n}} \ge \frac{\Sigma_i b_i}{n},$$

which implies

$$\frac{\sqrt{n}}{\Sigma_i b_i} \ge \frac{1}{\sqrt{\Sigma_i b_i^2}}. \quad \text{(a.23)}$$

We then have

$$\begin{aligned} \frac{|\mathbf{a}^T\mathbf{c}|}{\|\mathbf{a}\|_2\|\mathbf{c}\|_2} &= \frac{\Sigma_i a_i}{\sqrt{\Sigma_i a_i^2}\sqrt{n}} \\ &= \frac{1}{\sqrt{\Sigma_i a_i^2}}\left(\frac{\Sigma_i a_i \Sigma_i b_i}{n}\right)\left(\frac{\sqrt{n}}{\Sigma_i b_i}\right) \\ &\ge \frac{\Sigma_i a_i b_i}{\sqrt{\Sigma_i a_i^2}\sqrt{\Sigma_i b_i^2}} = \frac{|\mathbf{a}^T\mathbf{b}|}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}, \end{aligned}$$

where the inequality follows from (a.22) and (a.23).

Similarly, we have

$$\frac{|\mathbf{b}^T\mathbf{c}|}{\|\mathbf{b}\|_2\|\mathbf{c}\|_2} \ge \frac{|\mathbf{a}^T\mathbf{b}|}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2},$$

so the proof is complete.

Corollary 1: Lemma 2 can be extended to cases where $a_i = b_i = c_i = 0$ for some i is allowed since we can reduce them to lower-dimensional cases by removing these dimensions.

Proof of Theorem 3

From the definition of $\bar{H}$, we have

$$\beta\bar{H}^{-\frac{1}{2}}\bar{H}_i\bar{H}^{-\frac{1}{2}} + (1-\beta)\bar{H}^{-\frac{1}{2}}\bar{H}_j\bar{H}^{-\frac{1}{2}} = I,$$

so we know that $\bar{H}^{-\frac{1}{2}}\bar{H}_i\bar{H}^{-\frac{1}{2}}$ and $\bar{H}^{-\frac{1}{2}}\bar{H}_j\bar{H}^{-\frac{1}{2}}$ share the same set of eigenvectors.

In addition, since they are both positive definite, we can write their eigendecomposition as

$$\bar{H}^{-\frac{1}{2}}\bar{H}_i\bar{H}^{-\frac{1}{2}} = U\Lambda_i U^{-1} \quad \text{(a.24)}$$

and

$$\bar{H}^{-\frac{1}{2}}\bar{H}_i\bar{H}^{-\frac{1}{2}} = U\Lambda_j U^{-1}, \quad \text{(a.25)}$$

where U is an orthogonal matrix, and Λ_i and Λ_j are diagonal matrices satisfying

$$\beta\Lambda_i + (1-\beta)\Lambda_j = I. \quad \text{(a.26)}$$

Let

$$\begin{aligned}\mathbf{a} &= \Lambda_i^{-1} U^{-1}\mathbf{v},\\ \mathbf{b} &= \Lambda_j^{-1} U^{-1}\mathbf{v},\\ \mathbf{c} &= U^{-1}\mathbf{v},\end{aligned} \tag{a.27}$$

where $\mathbf{v}$ is any non-zero vector. Since

$$\begin{aligned}a_k &= (\Lambda_i^{-1})_{k,k}(U^{-1}\mathbf{v})_k,\\ b_k &= (\Lambda_j^{-1})_{k,k}(U^{-1}\mathbf{v})_k,\\ c_k &= (U^{-1}\mathbf{v})_k,\end{aligned}$$

and

$$a_k b_k = (\Lambda_i^{-1})_{k,k}(\Lambda_j^{-1})_{k,k} c_k^2,$$

from (a.26), we know that they either satisfy $a_k = b_k = c_k = 0$ or (a.18). As a result, Corollary 1 gives us

$$\frac{|\mathbf{a}^T\mathbf{b}|}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} \le \min\{\frac{|\mathbf{a}^T\mathbf{c}|}{\|\mathbf{a}\|_2\|\mathbf{c}\|_2}, \frac{|\mathbf{b}^T\mathbf{c}|}{\|\mathbf{b}\|_2\|\mathbf{c}\|_2}\}. \tag{a.28}$$

From (a.24) and (a.25), we have

$$\bar{H}^{\frac{1}{2}} {\bar{H}_i}^{-1} \bar{H}^{\frac{1}{2}} = U\Lambda_i^{-1}U^{-1} \tag{a.29}$$

and

$$\bar{H}^{\frac{1}{2}} {\bar{H}_j}^{-1} \bar{H}^{\frac{1}{2}} = U\Lambda_j^{-1}U^{-1}. \tag{a.30}$$

In addition, since $U^{-1} = U^T$, we have

$$\mathbf{x}^T\mathbf{y} = (U\mathbf{x})^T(U\mathbf{y}).$$

Thus from (a.27), (a.29), and (a.30), we know that for

$$\begin{aligned}\tilde{\mathbf{a}} &\equiv U\mathbf{a} = \bar{H}^{\frac{1}{2}} {\bar{H}_i}^{-1} \bar{H}^{\frac{1}{2}}\mathbf{v},\\ \tilde{\mathbf{b}} &\equiv U\mathbf{b} = \bar{H}^{\frac{1}{2}} {\bar{H}_j}^{-1} \bar{H}^{\frac{1}{2}}\mathbf{v},\\ \tilde{\mathbf{c}} &\equiv U\mathbf{c} = \mathbf{v},\end{aligned} \tag{a.31}$$

we have $\|\tilde{\mathbf{a}}\|_2 = \|\mathbf{a}\|_2$, $|\tilde{\mathbf{a}}^T\tilde{\mathbf{b}}| = |\mathbf{a}^T\mathbf{b}|$, etc. Consequently, $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}$, and $\tilde{\mathbf{c}}$ should also satisfy (a.28).

Picking $\mathbf{v} = -\bar{H}^{-\frac{1}{2}}\mathbf{g}$, then from (a.31) and the definition of $\bar{\mathbf{s}}_i, \bar{\mathbf{s}}_j, \bar{\mathbf{s}}$ in Theorem 3, we have

$$\begin{aligned}\tilde{\mathbf{a}} &= -\bar{H}^{\frac{1}{2}} {\bar{H}_i}^{-1}\mathbf{g} = \bar{H}^{\frac{1}{2}}\bar{\mathbf{s}}_i,\\ \tilde{\mathbf{b}} &= -\bar{H}^{\frac{1}{2}} {\bar{H}_j}^{-1}\mathbf{g} = \bar{H}^{\frac{1}{2}}\bar{\mathbf{s}}_j,\\ \tilde{\mathbf{c}} &= -\bar{H}^{-\frac{1}{2}}\mathbf{g} = \bar{H}^{\frac{1}{2}}\bar{\mathbf{s}}.\end{aligned}$$

As a result, (a.28) now becomes

$$\tau_{\bar{H}}(\bar{\mathbf{s}}_i, \bar{\mathbf{s}}_j) \le \min\{\tau_{\bar{H}}(\bar{\mathbf{s}}_i, \bar{\mathbf{s}}), \tau_{\bar{H}}(\bar{\mathbf{s}}_j, \bar{\mathbf{s}})\},$$

which completes our proof.

E. Further Connections

We can further connect the average strength of the subsampled Newton directions $\mathrm{mean}_i\, \mu(\bar{\mathbf{s}}_i)$ with their similarity ζ, as long as we are willing to make more assumptions.

As one can see, a noteworthy analogy to (13) is

$$\tau_H(\bar{\mathbf{s}}_i, \bar{\mathbf{s}}_j) \le \min\{\sqrt{\mu(\bar{\mathbf{s}}_i)}, \sqrt{\mu(\bar{\mathbf{s}}_j)}\}, \tag{a.32}$$

where $\bar{H}$ is replaced by H and Proposition 1 is applied.

If one believes (a.32) will hold on average, or, to be more precise, for every i we have

$$\operatorname*{mean}_{j:j\neq i} \tau_H(\bar{\mathbf{s}}_i, \bar{\mathbf{s}}_j) \le \operatorname*{mean}_{j:j\neq i} \sqrt{\mu(\bar{\mathbf{s}}_j)},$$

then we have

$$\begin{aligned}\zeta &\le \max_i \operatorname*{mean}_{j:j\neq i} \sqrt{\mu(\bar{\mathbf{s}}_j)} \le \sqrt{\max_i \operatorname*{mean}_{j:j\neq i} \mu(\bar{\mathbf{s}}_j)}\\ &\le \sqrt{\frac{m}{m-1} \operatorname*{mean}_i \mu(\bar{\mathbf{s}}_i)}.\end{aligned}$$

As a result, whenever the directions are close to each other so the similarity ζ is large, the average strength $\mathrm{mean}_i\, \mu(\bar{\mathbf{s}}_i)$ should also be large, which means the directions are strong. This avoids the weakness of gradient directions, where the directions can be both weak and similar.