Supplemental Material for "Stochastic L-BFGS Revisited: Improved Convergence Rates and Practical Acceleration Strategies"

S-1 Technical Lemmas

We first provide some technical lemmas that will be used in our subsequent proofs. Lemma S-1 states two classical inequalities of a strongly convex and smooth function, and its proof can be found in Boyd and Vandenberghe (2004, Chapter 9). Lemma S-2 states some results about importance sampling, and can be derived using definition of expectation and Lemma S-1. For a complete proof, see Xiao and Zhang (2014, Section 3). Lemma S-3 states a classical result on the distance between (uniformly) subsampled mean and (global) mean of a finite number of vectors. Its proof can be found in Konečný et al. (2016, Appendix B.A).

Lemma S-1. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth, then for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{\mu}{2}\left\|\mathbf{x}-\mathbf{x}^*\right\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\left\|\nabla f(\mathbf{x})\right\|^2, \tag{S-1}$$

$$\frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2}\left\|\mathbf{x}-\mathbf{x}^*\right\|^2, \tag{S-2}$$

where $\mathbf{x}^*$ denotes the unique minimizer of f on $\mathbb{R}^d$.

Lemma S-2. *Let f be defined as in* (1) *and satisfies Assumption 2. Define a distribution p with support $[n]$ such that $p_i = L_i/(n\overline{L})$, for any $i \in [n]$. Then for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_{i\sim p}\left[\frac{1}{np_i}\nabla f_i(\mathbf{x})\right] = \nabla f(\mathbf{x}), \tag{S-3}$$

$$\mathbb{E}_{i\sim p}\left[\left\|\frac{1}{np_i}\left(\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\right)\right\|^2\right] \leq 2\overline{L}(f(\mathbf{x}) - f(\mathbf{x}^*)), \tag{S-4}$$

where $\mathbf{x}^*$ denotes the unique minimizer of f on $\mathbb{R}^d$.

Lemma S-3. *Let $\{\mathbf{z}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ and define $\overline{\mathbf{z}} \triangleq 1/n\sum_{i=1}^n \mathbf{z}_i$. Uniformly sample a random subset $\mathcal{S}$ of $[n]$ with size b without replacement. Then*

$$\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{n}\sum_{i\in\mathcal{S}}^n \mathbf{z}_i - \overline{\mathbf{z}}\right\|^2\right] \leq \frac{n-b}{b(n-1)}\left(\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{z}_i\right\|^2\right). \tag{S-5}$$

S-2 Proof of Theorem 1

Proof. Fix an outer iteration s and consider an inner iteration t. Define $r \triangleq \lfloor(\sum_{i=0}^{s-1} T_i + t)/L\rfloor$,[11] then the iteration in (10) becomes

$$\widetilde{\mathbf{x}}_{s,t+1,r} = \widetilde{\mathbf{x}}_{s,t,r} - \eta\widetilde{\mathbf{v}}_{s,t,r}. \tag{S-6}$$

In addition, from Lemmas 1 and 3, we have

$$\mathbb{E}_{\mathcal{B}_{s,t}}\left[\widetilde{\mathbf{v}}_{s,t,r}\middle|\,\mathcal{F}_{s,t}\right] = \mathbf{H}_r^{1/2}\nabla f(\mathbf{x}_{s,t}) = \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}), \tag{S-7}$$

$$\begin{aligned}\mathbb{E}_{\mathcal{B}_{s,t}}\left[\left\|\widetilde{\mathbf{v}}_{s,t,r} - \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r})\right\|^2\middle|\,\mathcal{F}_{s,t}\right] &\leq \left\|\mathbf{H}^{1/2}\right\|^2\frac{4\overline{L}}{b}\left(f(\mathbf{x}_{s,t}) - f(\mathbf{x}^*) + f(\mathbf{x}^s) - f(\mathbf{x}^*)\right)\\ &\leq \frac{4\Gamma\overline{L}}{b}\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*) + \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right),\end{aligned} \tag{S-8}$$

where $r' \triangleq \lfloor sm/L\rfloor$. Next we express the distance between $\widetilde{\mathbf{x}}_{s,t+1,r}$ and $\widetilde{\mathbf{x}}_r^*$ as

$$\left\|\widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 = \left\|\widetilde{\mathbf{x}}_{s,t,r} - \eta\widetilde{\mathbf{v}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \tag{S-9}$$

$$= \left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 + 2\eta\left(\frac{\eta}{2}\left\|\widetilde{\mathbf{v}}_{s,t,r}\right\|^2 - \langle\widetilde{\mathbf{v}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\rangle\right). \tag{S-10}$$

[11] To avoid cluttered notations, we omit showing the dependence of r on s and t.

Defining $\widetilde{\boldsymbol{\delta}}_{s,t,r} \triangleq \widetilde{\mathbf{v}}_{s,t,r} - \nabla \widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r})$, we now bound

$$\frac{\eta}{2}\left\|\widetilde{\mathbf{v}}_{s,t,r}\right\|^2 - \langle \widetilde{\mathbf{v}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^* \rangle \leq -\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\right) - \left\langle \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^* \right\rangle - \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2. \tag{S-11}$$

This is because

$$\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) + \frac{\eta}{2}\left\|\widetilde{\mathbf{v}}_{s,t,r}\right\|^2 - \langle \widetilde{\mathbf{v}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^* \rangle + \left\langle \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^* \right\rangle + \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \tag{S-12}$$

$$= \widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \left\langle \widetilde{\mathbf{v}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} + \frac{\eta}{2}\widetilde{\mathbf{v}}_{s,t,r} - \widetilde{\mathbf{x}}_r^* \right\rangle + \left\langle \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^* \right\rangle + \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \tag{S-13}$$

$$\overset{(a)}{\leq} \widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \frac{\Gamma\overline{L}}{2}\eta^2\left\|\widetilde{\mathbf{v}}_{s,t,r}\right\|^2 - \left\langle \widetilde{\mathbf{v}}_{s,t,r} - \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^* \right\rangle + \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \tag{S-14}$$

$$= \widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \left\langle \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}), \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_{s,t,r} \right\rangle - \frac{\Gamma\overline{L}}{2}\left\|\eta\widetilde{\mathbf{v}}_{s,t,r}\right\|^2 - \left\langle \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}), \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^* \right\rangle + \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \tag{S-15}$$

$$\overset{(b)}{\leq} \widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) + \left\langle \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}), \widetilde{\mathbf{x}}_r^* - \widetilde{\mathbf{x}}_{s,t,r} \right\rangle + \frac{\gamma\overline{\mu}}{2}\left\|\widetilde{\mathbf{x}}_r^* - \widetilde{\mathbf{x}}_{s,t,r}\right\|^2 \tag{S-16}$$

$$\overset{(c)}{\leq} \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*), \tag{S-17}$$

where (a) follows from $\eta \leq 1/(\Gamma\overline{L})$, (b) follows from the $(\Gamma\overline{L})$-smoothness of $\widetilde{f}_r$ and (c) follows from the $(\gamma\overline{\mu})$-strong convexity of $\widetilde{f}_r$ (see Lemma 2). Substituting (S-11) into (S-10), we have

$$\left\|\widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \leq (1-\eta\gamma\overline{\mu})\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 - 2\eta\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\right) - 2\eta\left\langle \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^* \right\rangle \tag{S-18}$$

$$= (1-\eta\gamma\overline{\mu})\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 - 2\eta\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\right) + 2\eta^2\left\|\widetilde{\boldsymbol{\delta}}_{s,t,r}\right\|^2 - 2\eta\left\langle \widetilde{\boldsymbol{\delta}}_{s,t,r}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^* \right\rangle. \tag{S-19}$$

Taking expectation with respect to (w.r.t.) $\mathcal{B}_{s,t}$ and using (S-7) and (S-8), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{B}_{s,t}}\left[\left\|\widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \middle| \mathcal{F}_{s,t}\right] + 2\eta\mathbb{E}_{\mathcal{B}_{s,t}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_{s,t}\right] \\ &\qquad \leq (1-\eta\gamma\overline{\mu})\left\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 + \frac{8}{b}\Gamma\overline{L}\eta^2\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*) + \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right).\end{aligned} \tag{S-20}$$

By relaxing the factor $1-\eta\gamma\overline{\mu}$ to 1, we are able to telescope (S-20) over $t = 0, \ldots, m-1$ using the tower property of conditional expectations (Williams, 1991, Chapter 9.7) and obtain

$$\begin{aligned}&\mathbb{E}_{\mathcal{B}_{s,(m-1]}}\left[\left\|\widetilde{\mathbf{x}}_{s,m,r} - \widetilde{\mathbf{x}}_r^*\right\|^2 \middle| \mathcal{F}_s\right] + 2m\eta\left(1 - \frac{4}{b}\Gamma\overline{L}\eta\right)\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}_{\mathcal{B}_{s,(t-1]}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_{s,t-1}\right] \\ &\qquad \leq \left\|\widetilde{\mathbf{x}}^{s,r'} - \widetilde{\mathbf{x}}_{r'}^*\right\|^2 + \frac{8}{b}\Gamma\overline{L}\eta^2(1+m)\left(\widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right),\end{aligned} \tag{S-21}$$

where $\mathcal{B}_{s,(m-1]} \triangleq \{\mathcal{B}_{s,i}\}_{i=0}^{m-1}$. If we use option II to choose $\mathbf{x}^{s+1}$ (in line 20), we have $\widetilde{\mathbf{x}}^{s+1,r''} = 1/m\sum_{i=1}^{m}\widetilde{\mathbf{x}}_{s,t,r''}$, where $r'' \triangleq \lfloor (s+1)m/L \rfloor$. Using (7) and Jensen's inequality, we have

$$\frac{1}{m}\sum_{t=1}^{m}\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\right) = \frac{1}{m}\sum_{t=1}^{m}\left(\widetilde{f}_{r''}(\widetilde{\mathbf{x}}_{s,t,r''}) - \widetilde{f}_{r''}(\widetilde{\mathbf{x}}_{r''}^*)\right) \geq \widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}_{r''}(\widetilde{\mathbf{x}}_{r''}^*). \tag{S-22}$$

Therefore, we have

$$\frac{1}{m}\sum_{t=1}^{m}\mathbb{E}_{\mathcal{B}_{s,(t-1]}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_{s,t-1}\right] \geq \mathbb{E}_{\mathcal{B}_{s,(m-1]}}\left[\widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}_{r''}(\widetilde{\mathbf{x}}_{r''}^*)\middle| \mathcal{F}_s\right]. \tag{S-23}$$

Alternatively, if we use option I to determine $\mathbf{x}^{s+1}$ (in line 19), we still have (S-23) (with inequality replaced by equality). If we further use (S-1) in Lemma S-1 to bound $\|\widetilde{\mathbf{x}}^{s,r'} - \widetilde{\mathbf{x}}^*_{r'}\|^2$ in (S-21), we have

$$2m\eta\left(1-\frac{4}{b}\Gamma\overline{L}\eta\right)\mathbb{E}_{\mathcal{B}_{s,(m-1]}}\left[\widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}_{r''}(\widetilde{\mathbf{x}}^*_{r''})\middle|\mathcal{F}_s\right] \leq \left(\frac{8}{b}\Gamma\overline{L}\eta^2(1+m)+\frac{2}{\gamma\overline{\mu}}\right)\left(\widetilde{f}_{r'}(\widetilde{x}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^*_{r'})\right). \tag{S-24}$$

Using (7) again and rearranging, we have

$$\mathbb{E}\left[f(\mathbf{x}^{s+1}) - f(\mathbf{x}^*)\middle|\mathcal{F}_s\right] \leq \rho(f(\mathbf{x}^s) - f(\mathbf{x}^*)). \tag{S-25}$$

Taking expectation on both sides and we complete the proof. □

S-3 Proof of Lemma 3

First from (S-4) we immediately have $\mathbb{E}_{\mathcal{B}_{s,t}}\left[\mathbf{v}_{s,t}|\mathcal{F}_{s,t}\right] = \nabla f(\mathbf{x}_{s,t})$. Then

$$\mathbb{E}_{\mathcal{B}_{s,t}}\left[\left\|\mathbf{v}_{s,t} - \nabla f(\mathbf{x}_{s,t})\right\|^2|\mathcal{F}_{s,t}\right] \tag{S-26}$$

$$= \mathbb{E}_{\mathcal{B}_{s,t}}\left[\left\|\frac{1}{b}\sum_{j=1}^{b}\left(1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^s)\right) + \nabla f(\mathbf{x}^s) - \nabla f(\mathbf{x}_{s,t})\right)\right\|^2\middle|\mathcal{F}_{s,t}\right] \tag{S-27}$$

$$\overset{(a)}{=} \frac{1}{b^2}\sum_{j=1}^{b}\mathbb{E}_{i_j}\left[\left\|1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^s)\right) + \nabla f(\mathbf{x}^s) - \nabla f(\mathbf{x}_{s,t})\right\|^2\middle|\mathcal{F}_{s,t}\right] \tag{S-28}$$

$$\overset{(b)}{\leq} \frac{1}{b^2}\sum_{j=1}^{b}\mathbb{E}_{i_j}\left[\left\|1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^s)\right)\right\|^2\middle|\mathcal{F}_{s,t}\right] \tag{S-29}$$

$$= \frac{1}{b^2}\sum_{j=1}^{b}\mathbb{E}_{i_j}\left[\left\|1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^*)\right) + 1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}^*) - \nabla f_{i_j}(\mathbf{x}^s)\right)\right\|^2\middle|\mathcal{F}_{s,t}\right] \tag{S-30}$$

$$\overset{(c)}{\leq} \frac{2}{b^2}\sum_{j=1}^{b}\mathbb{E}_{i_j}\left[\left\|1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^*)\right)\right\|^2\middle|\mathcal{F}_{s,t}\right] + \mathbb{E}_{i_j}\left[\left\|1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}^s) - \nabla f_{i_j}(\mathbf{x}^*)\right)\right\|^2\middle|\mathcal{F}_{s,t}\right] \tag{S-31}$$

$$\overset{(d)}{\leq} \frac{2}{b^2}\sum_{j=1}^{b}2\overline{L}(f(\mathbf{x}_{s,t}) - f(\mathbf{x}^*)) + 2\overline{L}(f(\mathbf{x}^s) - f(\mathbf{x}^*)) \tag{S-32}$$

$$= \frac{4\overline{L}}{b}\left(f(\mathbf{x}_{s,t}) - f(\mathbf{x}^*) + f(\mathbf{x}^s) - f(\mathbf{x}^*)\right), \tag{S-33}$$

where (a) follows from the independence of i_j and $i_{j'}$ when $j \neq j'$ and

$$\mathbb{E}_{i_j}\left[1/(np_{i_j})\left(\nabla f_{i_j}(\mathbf{x}_{s,t}) - \nabla f_{i_j}(\mathbf{x}^s)\right) + \nabla f(\mathbf{x}^s) - \nabla f(\mathbf{x}_{s,t})\middle|\mathcal{F}_{s,t}\right] = 0, \tag{S-34}$$

(b) follows from (S-3) and $\mathbb{E}\left[\left\|\mathbf{a} - \mathbb{E}[\mathbf{a}]\right\|^2\middle|\mathcal{G}\right] \leq \mathbb{E}\left[\left\|\mathbf{a}\right\|^2\middle|\mathcal{G}\right]$ almost surely, for any σ-algebra $\mathcal{G}$, (c) follows from $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, and (d) follows from (S-4).

S-4 Proof of Lemma 4

For any $\mathbf{x} \in \mathbb{R}^d$ and $\mathcal{T} \subseteq [n]$ with cardinality b_{H},

$$\overline{\mu}_{b_{\mathrm{H}}}\mathbf{I} \preceq \nabla^2 f_{\mathcal{T}}(\mathbf{x}) \preceq \overline{L}_{b_{\mathrm{H}}}\mathbf{I}, \tag{S-35}$$

Define

$$\mathbf{V}_k \triangleq \mathbf{I} - \frac{\mathbf{y}_k {\mathbf{s}_k}^T}{{\mathbf{y}_k}^T \mathbf{s}_k} \quad \text{and} \quad \mathbf{Q}_k \triangleq \frac{\mathbf{s}_k {\mathbf{s}_k}^T}{{\mathbf{y}_k}^T \mathbf{s}_k}. \tag{S-36}$$

Fix any $r \in \mathbb{N}$. Since

$$\mathbf{y}_k = \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)\mathbf{s}_k, \tag{S-37}$$

we have

$$\mathbf{V}_k = \mathbf{I} - \frac{\nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)\mathbf{s}_k {\mathbf{s}_k}^T}{{\mathbf{s}_k}^T \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)\mathbf{s}_k} \tag{S-38}$$

$$= \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)^{1/2} \left(\mathbf{I} - \frac{\widetilde{\mathbf{s}}_k \widetilde{\mathbf{s}}_k^T}{\widetilde{\mathbf{s}}_k^T \widetilde{\mathbf{s}}_k} \right) \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)^{-1/2}, \tag{S-39}$$

where $\widetilde{\mathbf{s}}_k \triangleq \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)^{1/2}\mathbf{s}_k$. Hence

$$\|\mathbf{V}_k\| \leq \left\| \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)^{1/2} \right\| \left\| \mathbf{I} - \frac{\widetilde{\mathbf{s}}_k \widetilde{\mathbf{s}}_k^T}{\|\widetilde{\mathbf{s}}_k\|^2} \right\| \left\| \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)^{-1/2} \right\| \tag{S-40}$$

$$\leq \overline{L}_{b_{\mathrm{H}}}^{1/2} \overline{\mu}_{b_{\mathrm{H}}}^{-1/2} = \kappa_{b_{\mathrm{H}}}^{1/2}. \tag{S-41}$$

Similarly,

$$\|\mathbf{Q}_k\| = \frac{\|\mathbf{s}_k\|^2}{{\mathbf{s}_k}^T \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)\mathbf{s}_k} \leq \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}}. \tag{S-42}$$

By (5), we have

$$\left\| \mathbf{H}_r^{(k)} \right\| \leq \|\mathbf{V}_k\|^2 \left\| \mathbf{H}_r^{(k-1)} \right\| + \|\mathbf{Q}_k\| \leq \kappa_{b_{\mathrm{H}}} \left\| \mathbf{H}_r^{(k-1)} \right\| + \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}}. \tag{S-43}$$

Now, apply (S-43) repeatedly over $k = r - M' + 1, \ldots, r$ and we have

$$\|\mathbf{H}_r\| = \left\| \mathbf{H}_r^{(r)} \right\| \leq \kappa_{b_{\mathrm{H}}}^{M'} \left\| \mathbf{H}_r^{(r-M')} \right\| + \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}} \sum_{i=0}^{M'-1} \kappa_{b_{\mathrm{H}}}^i \tag{S-44}$$

$$\overset{\text{(a)}}{\leq} \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}} \left(\kappa_{b_{\mathrm{H}}}^{M'} + \frac{\kappa_{b_{\mathrm{H}}}^{M'} - 1}{\kappa_{b_{\mathrm{H}}} - 1} \right) \tag{S-45}$$

$$\leq \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}} \kappa_{b_{\mathrm{H}}}^{M'} \left(1 + \frac{1}{\kappa_{b_{\mathrm{H}}} - 1} \right) \tag{S-46}$$

$$\overset{\text{(b)}}{\leq} \frac{\kappa_{b_{\mathrm{H}}}^{M+1}}{\overline{\mu}_{b_{\mathrm{H}}}(\kappa_{b_{\mathrm{H}}} - 1)}, \tag{S-47}$$

where (a) follows from

$$\mathbf{H}_r^{(r-M')} = \frac{{\mathbf{s}_r}^T \mathbf{y}_r}{{\mathbf{y}_r}^T \mathbf{y}_r} \mathbf{I} \tag{S-48}$$

and

$$\frac{{\mathbf{s}_r}^T \mathbf{y}_r}{{\mathbf{y}_r}^T \mathbf{y}_r} = \frac{\widetilde{\mathbf{s}}_r^T \widetilde{\mathbf{s}}_r}{\widetilde{\mathbf{s}}_r^T \nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r) \widetilde{\mathbf{s}}_r} \leq \frac{1}{\overline{\mu}_{b_{\mathrm{H}}}}, \tag{S-49}$$

and (b) follows from $\kappa_{b_{\mathrm{H}}} \geq 1$ and $M' \leq M$. To show $\gamma = 1/(M+1)\overline{L}_{b_{\mathrm{H}}}$, it suffices to show

$$\|\mathbf{B}_r\| \leq (M+1)\overline{L}_{b_{\mathrm{H}}}, \tag{S-50}$$

where $\mathbf{B}_r \triangleq \mathbf{H}_r^{-1}$. To begin with, we rewrite (5) as Nocedal and Wright (2006, Chapter 6)

$$\mathbf{B}_r^{(k)} = \mathbf{B}_r^{(k-1)} - \frac{\mathbf{B}_r^{(k-1)} \mathbf{s}_k {\mathbf{s}_k}^T \mathbf{B}_r^{(k-1)}}{{\mathbf{s}_k}^T \mathbf{B}_r^{(k-1)} \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{s}_k^T \mathbf{y}_k}, \tag{S-51}$$

where $\mathbf{B}_r^{(k)} \triangleq \left(\mathbf{H}_r^{(k)}\right)^{-1}$, for each $k \in [r-M'+1:r]$. Since the update rule (5) preserves positive definiteness of $\mathbf{H}_r^{(k)}$ (Nocedal and Wright, 2006), we have that $\mathbf{B}_r^{(k)} \succ 0$ for all $k \in [r-M'+1:r]$. Define $\widehat{\mathbf{s}}_k \triangleq \left(\mathbf{B}_r^{(k)}\right)^{1/2} \mathbf{s}_k$, then

$$\left\|\mathbf{B}_r^{(k-1)} - \frac{\mathbf{B}_r^{(k-1)}\mathbf{s}_k{\mathbf{s}_k}^T\mathbf{B}_r^{(k-1)}}{{\mathbf{s}_k}^T\mathbf{B}_r^{(k-1)}\mathbf{s}_k}\right\| \leq \left\|\left(\mathbf{B}_r^{(k-1)}\right)^{1/2}\right\|^2 \left\|\mathbf{I} - \frac{\widehat{\mathbf{s}}_k\widehat{\mathbf{s}}_k^T}{\|\widehat{\mathbf{s}}_k\|^2}\right\| = \left\|\mathbf{B}_r^{(k-1)}\right\|. \tag{S-52}$$

By (S-51), we have that

$$\left\|\mathbf{B}_r^{(k)}\right\| \leq \left\|\mathbf{B}_r^{(k-1)}\right\| + \left\|\frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{s}_k^T\mathbf{y}_k}\right\| \leq \left\|\mathbf{B}_r^{(k-1)}\right\| + \overline{L}_{b_{\mathrm{H}}}, \tag{S-53}$$

since

$$\left\|\frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{s}_k^T\mathbf{y}_k}\right\| = \frac{\|\mathbf{y}_k\|^2}{\mathbf{s}_k^T\mathbf{y}_k} = \frac{\widetilde{\mathbf{s}}_k^T\nabla^2 f_{\mathcal{T}_r}(\overline{\mathbf{x}}_r)\widetilde{\mathbf{s}}_k}{\widetilde{\mathbf{s}}_k^T\widetilde{\mathbf{s}}_k} \leq \overline{L}_{b_{\mathrm{H}}}. \tag{S-54}$$

Then we have

$$\|\mathbf{B}_r\| = \left\|\mathbf{B}_r^{(r)}\right\| \leq \left\|\mathbf{B}_r^{(r-M')}\right\| + M'\overline{L}_{b_{\mathrm{H}}} \overset{\text{(a)}}{\leq} (M+1)\overline{L}_{b_{\mathrm{H}}}, \tag{S-55}$$

where (a) follows from (S-48) and (S-54).

S-5 Proof of Proposition 2

Our proof leverages a refined telescoping of (S-20). For each $t \in [m]$, we multiply both sides of (S-20) by $(1-\eta\gamma\overline{\mu})^{m-t}$ and obtain

$$\begin{aligned}
&(1-\eta\gamma\overline{\mu})^{m-t}\left(\mathbb{E}_{\mathcal{B}_{s,t}}\left[\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\|^2 \middle| \mathcal{F}_{s,t}\right] + 2\eta\mathbb{E}_{\mathcal{B}_{s,t}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_{s,t}\right]\right) \\
&\quad \leq (1-\eta\gamma\overline{\mu})^{m-t}\left((1-\eta\gamma\overline{\mu})\|\widetilde{\mathbf{x}}_{s,t-1,r} - \widetilde{\mathbf{x}}_r^*\|^2 + \frac{8}{b}\Gamma\overline{L}\eta^2\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t-1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*) + \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right)\right).
\end{aligned} \tag{S-56}$$

Telescope (S-56) over $t = 1, \ldots, m$ and we have

$$\begin{aligned}
&2c\eta\left(1 - \frac{4}{b}\frac{\Gamma\overline{L}\eta}{1-\eta\gamma\overline{\mu}}\right)\frac{1}{c}\sum_{t=1}^{m}(1-\eta\gamma\overline{\mu})^{m-t}\mathbb{E}_{\mathcal{B}_{s,(t]}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_s\right] \\
&\quad \leq (1-\eta\gamma\overline{\mu})^m\left\|\widetilde{\mathbf{x}}^{s,r'} - \mathbf{x}_{r'}^*\right\|^2 + \frac{8}{b}\Gamma\overline{L}\eta^2\left((1-\eta\gamma\overline{\mu})^m + c\right)\left(\widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right).
\end{aligned} \tag{S-57}$$

Now we consider using option IV to choose $\mathbf{x}^{s+1}$. Since $0 < \beta \leq 1-\eta\gamma\overline{\mu}$, using (7) and Jensen's inequality in a similar manner as in (S-22) and taking expectations, we have

$$\frac{1}{c}\sum_{t=1}^{m}(1-\eta\gamma\overline{\mu})^{m-t}\mathbb{E}_{\mathcal{B}_{s,(t]}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle| \mathcal{F}_s\right] \geq \mathbb{E}_{\mathcal{B}_{s,(m]}}\left[\widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}(\widetilde{\mathbf{x}}_{r''}^*)\middle| \mathcal{F}_s\right]. \tag{S-58}$$

Alternatively, if $\mathbf{x}^{s+1}$ is determined using option III, the definition of distribution q still yields (S-58). Finally, using (S-1) in Lemma S-1 to bound $\|\widetilde{\mathbf{x}}^{s,r'} - \widetilde{\mathbf{x}}_{r'}^*\|^2$ in (S-57), we have

$$\begin{aligned}
&2c'\eta\left(1 - \frac{4}{b}\frac{\Gamma\overline{L}\eta}{1-\eta\gamma\overline{\mu}}\right)\mathbb{E}_{\mathcal{B}_{s,(m]}}\left[\widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}(\widetilde{\mathbf{x}}_{r''}^*)\middle| \mathcal{F}_s\right] \\
&\qquad \leq \left(\frac{8}{b}\Gamma\overline{L}\eta^2(1+c') + \frac{2}{\gamma\overline{\mu}}\right)\left(\widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right).
\end{aligned} \tag{S-59}$$

Taking expectation on both sides, rearranging and using (7), we reach (31).

S-6 Proof of Proposition 3

The subsampled gradient strategy essentially introduces errors in $\{\mathbf{g}_s\}_{s\geq 0}$. Therefore, we explicit model this error by $\mathbf{e}_s \triangleq \widetilde{\mathbf{g}}_s - \mathbf{g}_s$. We first bound the second moment of $\mathbf{e}_s$. Since $\mathbf{g}_s = \nabla f(\mathbf{x}^s)$, by Lemma S-3 and $\widetilde{b}_s \geq nS^2\alpha_s/(S^2\alpha_s + (n-1)\xi^2\overline{\rho}^{2s})$, we have for any $s \in (S]$,

$$\mathbb{E}_{\widetilde{\mathcal{B}}_s}\left[\|\mathbf{e}_s\|^2\right] \leq \frac{n-\widetilde{b}_s}{\widetilde{b}_s(n-1)}\alpha_s \leq \frac{\xi^2}{S^2}\overline{\rho}^{2s}. \tag{S-60}$$

As a result,

$$\mathbb{E}_{\widetilde{\mathcal{B}}_s}\left[\|\mathbf{e}_s\|\right] \leq \frac{\xi}{S}\overline{\rho}^{s}. \tag{S-61}$$

The introduction of random sets $\{\widetilde{\mathcal{B}}_s\}_{s\geq 0}$ requires us to redefine the filtration $\{\mathcal{F}_{s,t}\}_{s\geq 0, t\in(m-1]}$ as

$$\mathcal{F}_{s,t} \triangleq \sigma\left(\{\tau_j\}_{j=0}^{s-1} \cup \{\widetilde{\mathcal{B}}_j\}_{j\in(s]} \cup \{\mathcal{B}_{i,j}\}_{i\in(s-1],j\in(m-1]} \cup \{\mathcal{B}_{s,j}\}_{j=0}^{t-1} \cup \{\mathcal{T}_j\}_{j=0}^{\lfloor (sm+t)/L\rfloor}\right). \tag{S-62}$$

As usual, we define $\mathcal{F}_s \triangleq \mathcal{F}_{s,0}$. In the sequel, we follow the naming convention in coordinate transformation framework as in Section 4.1. In particular, we define $\widetilde{\mathbf{e}}_{s,r'} \triangleq \mathbf{H}_{r'}^{1/2}\mathbf{e}_s$. Now, define $\widetilde{\mathbf{v}}'_{s,t,r} \triangleq \widetilde{\mathbf{v}}_{s,t,r} + \widetilde{\mathbf{e}}_{s,r'}$, then from (S-7) and (S-8), we have

$$\mathbb{E}_{\mathcal{B}_{s,t}}\left[\widetilde{\mathbf{v}}'_{s,t,r}\middle|\mathcal{F}_{s,t}\right] = \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) + \widetilde{\mathbf{e}}_{s,r'}, \tag{S-63}$$

$$\mathbb{E}_{\mathcal{B}_{s,t}}\left[\left\|\widetilde{\mathbf{v}}'_{s,t,r} - \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r})\right\|^2\middle|\mathcal{F}_{s,t}\right] \leq \frac{4\Gamma\overline{L}}{b}\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*) + \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right) + \|\widetilde{\mathbf{e}}_{s,r'}\|^2, \tag{S-64}$$

Following a similar argument as in the proof of Theorem 1, we have

$$\|\widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^*\|^2 \leq (1-\eta\gamma\overline{\mu})\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\|^2 - 2\eta\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\right) + 2\eta^2\left\|\widetilde{\boldsymbol{\delta}}'_{s,t,r}\right\|^2 - 2\eta\left\langle\widetilde{\boldsymbol{\delta}}'_{s,t,r}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\right\rangle, \tag{S-65}$$

where $\widetilde{\boldsymbol{\delta}}'_{s,t,r} \triangleq \widetilde{\mathbf{v}}'_{s,t,r} - \nabla\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r})$. Taking expectation w.r.t. $\mathcal{B}_{s,t}$ and using (S-63) and (S-64), we have

$$\begin{aligned}\mathbb{E}_{\mathcal{B}_{s,t}}\left[\|\widetilde{\mathbf{x}}_{s,t+1,r} - \widetilde{\mathbf{x}}_r^*\|^2\middle|\mathcal{F}_{s,t}\right] + 2\eta\mathbb{E}_{\mathcal{B}_{s,t}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t+1,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle|\mathcal{F}_{s,t}\right] &\leq (1-\eta\gamma\overline{\mu})\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\|^2 \\ &+ \frac{8}{b}\Gamma\overline{L}\eta^2\left(\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*) + \widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right) + 2\eta^2\|\widetilde{\mathbf{e}}_{s,r'}\|^2 - 2\eta\langle\widetilde{\mathbf{e}}_{s,r'}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\rangle.\end{aligned} \tag{S-66}$$

Using Cauchy-Schwartz inequality, we have

$$-2\eta\langle\widetilde{\mathbf{e}}_{s,r'}, \widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\rangle \leq 2\eta\|\widetilde{\mathbf{e}}_{s,r'}\|\|\widetilde{\mathbf{x}}_{s,t,r} - \widetilde{\mathbf{x}}_r^*\| \leq 2\eta\|\widetilde{\mathbf{e}}_{s,r'}\|\left\|\mathbf{H}_r^{-1/2}\right\|\|\mathbf{x}_{s,t} - \mathbf{x}^*\| \leq 2\eta B\gamma^{-1/2}\|\widetilde{\mathbf{e}}_{s,r'}\|. \tag{S-67}$$

Substituting (S-67) into (S-66), and using the telescoping techniques in Section S-5, we have

$$\begin{aligned}&2c\eta\left(1 - \frac{4}{b}\frac{\Gamma\overline{L}\eta}{1-\eta\gamma\overline{\mu}}\right)\frac{1}{c}\sum_{t=1}^{m}(1-\eta\gamma\overline{\mu})^{m-t}\mathbb{E}_{\mathcal{B}_{s,(t]}}\left[\widetilde{f}_r(\widetilde{\mathbf{x}}_{s,t,r}) - \widetilde{f}_r(\widetilde{\mathbf{x}}_r^*)\middle|\mathcal{F}_s\right] \leq (1-\eta\gamma\overline{\mu})^m\left\|\widetilde{\mathbf{x}}^{s,r'} - \mathbf{x}_{r'}^*\right\|^2 \\ &+ \frac{8}{b}\Gamma\overline{L}\eta^2\left((1-\eta\gamma\overline{\mu})^m + c\right)\left(\widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right) + 2\eta\left((1-\eta\gamma\overline{\mu})^m + c\right)\left(B\gamma^{-1/2}\|\widetilde{\mathbf{e}}_{s,r'}\| + \eta\|\widetilde{\mathbf{e}}_{s,r'}\|^2\right).\end{aligned} \tag{S-68}$$

With either option III or IV and (S-1), we have

$$\begin{aligned}&2c'\eta\left(1 - \frac{4}{b}\frac{\Gamma\overline{L}\eta}{1-\eta\gamma\overline{\mu}}\right)\mathbb{E}_{\mathcal{B}_{s,(m]}}\left[\widetilde{f}_{r''}(\widetilde{\mathbf{x}}^{s+1,r''}) - \widetilde{f}(\widetilde{\mathbf{x}}_{r''}^*)\middle|\mathcal{F}_s\right] \\ &\quad\leq\left(\frac{8}{b}\Gamma\overline{L}\eta^2(1+c') + \frac{2}{\gamma\overline{\mu}}\right)\left(\widetilde{f}_{r'}(\widetilde{\mathbf{x}}^{s,r'}) - \widetilde{f}_{r'}(\widetilde{\mathbf{x}}_{r'}^*)\right) + 2\eta(1+c')\left(B\gamma^{-1/2}\|\widetilde{\mathbf{e}}_{s,r'}\| + \eta\|\widetilde{\mathbf{e}}_{s,r'}\|^2\right).\end{aligned} \tag{S-69}$$

Taking expectation on both sides and using (7), we have

$$\mathbb{E}\left[f(\mathbf{x}^{s+1}) - f(\mathbf{x}^*)\right] \leq \overline{\rho}\mathbb{E}\left[f(\mathbf{x}^s) - f(\mathbf{x}^*)\right] + \left(1 + \frac{1}{c'}\right) \frac{b}{b - 4\Gamma\overline{L}\eta/(1-\eta\gamma\overline{\mu})} \left(B\gamma^{-1/2}\mathbb{E}\left[\|\widetilde{\mathbf{e}}_{s,r'}\|\right] + \eta\mathbb{E}\left[\|\widetilde{\mathbf{e}}_{s,r'}\|^2\right]\right). \tag{S-70}$$

From (S-60) and (S-61), we have

$$\mathbb{E}\left[\|\widetilde{\mathbf{e}}_{s,r'}\|\right] \leq \frac{\Gamma^{1/2}\xi}{S}\overline{\rho}^s, \tag{S-71}$$

$$\mathbb{E}\left[\|\widetilde{\mathbf{e}}_{s,r'}\|^2\right] \leq \frac{\Gamma\xi^2}{S^2}\overline{\rho}^{2s}. \tag{S-72}$$

Substituting (S-71) and (S-72) into (S-70), we have

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}^{s+1}) - f(\mathbf{x}^*)\right] &\leq \overline{\rho}\mathbb{E}\left[f(\mathbf{x}^s) - f(\mathbf{x}^*)\right] + \left(1 + \frac{1}{c'}\right) \frac{b}{b - 4\Gamma\overline{L}\eta/(1-\eta\gamma\overline{\mu})} \left(\frac{\kappa_{\mathrm{H}}^{1/2}B\xi}{S}\overline{\rho}^s + \frac{\eta\Gamma\xi^2}{S^2}\overline{\rho}^{2s}\right) \\ &\leq \overline{\rho}\mathbb{E}\left[f(\mathbf{x}^s) - f(\mathbf{x}^*)\right] + \left(1 + \frac{1}{c'}\right) \frac{b}{b - 4\Gamma\overline{L}\eta/(1-\eta\gamma\overline{\mu})} \left(\kappa_{\mathrm{H}}^{1/2}B + \eta\Gamma\xi\right) \frac{\xi}{S}\overline{\rho}^s \end{aligned} \tag{S-73}$$

Applying (S-73) recursively and we complete the proof.

S-7 Future Work and Open Problems

For future work, we propose to pursue from the following two directions:

- Based on our novel coordinate transformation framework, *proximal* and momentum-based stochastic L-BFGS algorithms can be developed and analyzed. The proximal variant can greatly enable our algorithm to handle composite (convex) objective function, thereby greatly extending the applications of our method. The accelerated variant can potentially improve the linear convergence rate of our algorithm, and thus reduce the total computational complexity.
- The linear convergence of the strategy in Section 6.3 can be analyzed theoretically. In Figure 4, we observed the linear convergence empirically. This indicates that linear convergence of Algorithm 1 may still hold under such a strategy.

Besides future work, there also exists an open problem. Although we have improved the linear convergence rate and computational complexity of Algorithm 1 as compared to those in Moritz et al. (2016) and Gower et al. (2016), it seems our improved complexity in (29) is still inferior to the computational complexity of SVRG. In SVRG, the complexity is $O((n+\kappa)d\log(1/\epsilon))$ (Xiao and Zhang, 2014), so our complexity (29) has an additional multiplicative factor κ_{H}. This contradicts the experimental results in Moritz et al. (2016), Gower et al. (2016) and Section 7.2, where stochastic L-BFGS-type algorithms have been repeatedly shown to outperform their first-order counterparts. A careful analysis will reveal that the additional κ_{H} arises from the uniform spectral bound of the metric matrices $\{\mathbf{H}_r\}_{r\geq 0}$. This uniform bound is effectively a worst-case bound, and does not reflect the *local curvature information* contained in recent iterates at any time (s,t). Since the judicious use of curvature information serves as a very important reason for the fast convergence of stochastic quasi-Newton algorithms, such information should also be reflected in theoretical analysis as well (possibly in an adaptive spectral bound for $\{\mathbf{H}_r\}_{r\geq 0}$). We also believe an effective adaptive bound can potentially improve the complexity result (29). *In short, how to obtain a (computational) complexity bound of stochastic L-BFGS algorithm that is better (or at least as good as) that of SVRG will be an interesting problem to consider.*