Supplementary Material for *Stochastic Adaptive Quasi-Newton Methods for Minimizing Expected Values*

Wenbo Gao [1] **Donald Goldfarb** [1] **Chaoxu Zhou** [1]

]

A. The Empirical Process Framework

We use the framework developed by Goldfarb, Iyengar, and Zhou (2017) for proving convergence of stochastic algorithms. These results originate in empirical process theory (W. van der Vaart & Wellner, 1996). The problem to be minimized has the form

$$\min_x F(x) = \mathbb{E}_\xi f(x, \xi)$$

We require the following assumptions on F, f for our analysis:

Assumptions:

1. There exist constants $L \geq \ell > 0$ such that for every $x \in \mathbb{R}^n$ and every realization of ξ, the Hessian of f with respect to x satisfies
$$\ell I \preceq \nabla_x^2 f(x, \xi) \preceq LI$$
That is, $f(x, \xi)$ is strongly convex for all ξ, with the eigenvalues of $\nabla_x^2 f(x, \xi)$ bounded below and above by ℓ and L, respectively.
2. $F_k(x)$ is standard self-concordant for every possible sampling $\xi_1, \ldots, \xi_{m_k}$.
3. There exist compact sets $\mathcal{D}_0$ and $\mathcal{D}$ with $x^* \in \mathcal{D}$ and $\mathcal{D}_0 \subseteq \mathcal{D}$, such that if x_0 is chosen in $\mathcal{D}_0$, then for all possible realizations of the samples $\xi_1, \ldots, \xi_{m_k}$ for every k, the sequence of iterates $\{x_k\}_{k=0}^\infty$ produced by the algorithm is contained within $\mathcal{D}$. We write $D = \sup\{\|x - y\| : x, y \in \mathcal{D}\}$ for the diameter of $\mathcal{D}$.

   Furthermore, we assume that the objective values and gradients are bounded:
$$u = \sup_\xi \sup_{x \in \mathcal{D}} f(x, \xi) < \infty$$
$$l = \inf_\xi \inf_{x \in \mathcal{D}} f(x, \xi) > -\infty$$
$$\gamma = \sup_\xi \sup_{x \in \mathcal{D}} \|\nabla f(x, \xi)\| < \infty$$

The key theorem of this framework is a concentration bound which limits the divergence of $F_k(x)$ from $F(x)$.

Theorem A.1 (Corollary 1, (Goldfarb et al., 2017)). *For any $\delta > 0$ and $0 < \epsilon < \min\{D, \frac{\delta}{2L}\}$, we have*

$$\mathbb{P}(\sup_{x \in \mathcal{D}} |F_k(x) - F(x)| \geq \delta) \leq 2n^{n/2} \frac{D^n}{\epsilon^n} \exp\left[-\frac{m_k(\delta - 2L\epsilon)^2}{2(u - l)^2}\right]$$

[1]Department of Industrial Engineering and Operations Research, Columbia University. Correspondence to: Chaoxu Zhou <cz2364@columbia.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Consequently, if $m_k \geq 3$, then we have

$$\mathbb{E} \sup_{x \in \mathcal{D}} |F_k(x) - F(x)| \leq C \sqrt{\frac{\log m_k}{m_k}}$$

and

$$\mathbb{E}|F_k(x_k^*) - F(x^*)| \leq C \sqrt{\frac{\log m_k}{m_k}}$$

where C is the constant

$$C = 4(|u| + |l|) n^{n/2} D^n \exp\left[-n\left(\log \frac{u-l}{2\sqrt{2}L}\right)\right] + (u-l)\sqrt{n+1}$$

We will also use Theorem A.1 to bound g_k and G_k. Assumption 1 implies that the partial derivatives $\frac{\partial F}{\partial x_i}(x)$ and $\frac{\partial^2 F}{\partial x_i x_j}(x)$ are also uniformly bounded for $x \in \mathcal{D}$. Hence, we can apply Theorem A.1 to each of the n entries $\frac{\partial F}{\partial x_i}$ of the gradient, and each of the n^2 entries $\frac{\partial^2 F}{\partial x_i \partial x_j}$ of the Hessian. Taking a union bound over the resulting $n^2 + n$ inequalities, we obtain the following concentration inequality for the sampled gradients and Hessians:

Corollary A.2. *For any $\delta > 0$ and $0 < \epsilon < \min\{D, \frac{\delta}{2L}\}$,*

$$\mathbb{P}(\sup_{x \in \mathcal{D}} \|G_k(x) - G(x)\| > \delta \text{ or } \sup_{x \in \mathcal{D}} \|g_k(x) - g(x)\| > \delta) \leq C_1 \epsilon^{-n} \exp\left[-C_2 m_k (\delta - C_3 \epsilon)^2\right]$$

where C_1, C_2, C_3 are constants depending only on F.

Recall the definitions of $\delta_k, \rho_k, \alpha_k$, and η_k above. In our analysis of stochastic methods, the gradients and Hessians are those of the *empirical* objective function. That is to say, $\rho_k = g_k^T H_k g_k$ and $\delta_k = \sqrt{d_k^T G_k d_k}$, where g_k and G_k are the gradient and Hessian of F_k.

We say that a constant c is *global* if it depends only on the properties of the function F, and is completely independent of the realization of the samples $\xi_1, \ldots, \xi_{m_k}$.

For convenience, we state again the main result for adaptive step sizes:

Theorem A.3 (Lemma 4.1, (Gao & Goldfarb, 2016))**.** *Let $\rho_k = \nabla f(x_k)^T H_k \nabla f(x_k)$. If α_k is chosen to be $\alpha_k = \frac{\rho_k}{\delta_k^2}$, then*

$$f(x_k + t_k d_k) \leq f(x_k) - \omega(\eta_k)$$

where $\eta_k = \frac{\rho_k}{\delta_k}$ and ω is the function $\omega(z) = z - \log(1+z)$.

B. Convergence of SA-GD

The SA-GD method corresponds to $H_k = I$. More generally, we may assume that the sequence of matrices H_k has bounded eigenvalues.

$$\lambda I \preceq H_k \preceq \Lambda I \qquad \text{for all } H_k \tag{1}$$

Since this condition is satisfied for L-BFGS, the following results also apply to L-BFGS (with slightly different constants).

Theorem B.1. *Let $\epsilon > 0$ be fixed. At each iteration, we draw m i.i.d samples $\xi_1, \ldots, \xi_m$, where the size of m satisfies*

$$\frac{\log m}{m} \leq \left(\frac{1-r}{4C}\right)^2 \epsilon^2$$

and C is the constant in Theorem A.1 and $r = 1 - \frac{\lambda^2 \ell^{3/2}}{(\sqrt{\ell}+\gamma)\Lambda^2 L}$. Then we have

$$\mathbb{E}F(x_{k+1}) - F(x^*) \leq \epsilon$$

when $k = \log(\epsilon^{-1} 2(u-l))/\log r$.

For matrices H_k with bounded eigenvalues, η_k can readily be bounded in terms of the empirical gradients, and the sequence $\{\eta_k\}_{k=0}^{\infty}$ is bounded.

Theorem B.2. *There exists a global constant $\Gamma = \frac{\gamma}{\sqrt{\ell}}$ such that $\eta_k \leq \Gamma$ for all k. Furthermore, $\eta_k \geq \frac{\lambda}{\Lambda\sqrt{L}}\|g_k\|$ for all k.*

Proof. By Assumption 1 (strong convexity), G_k satisfies $\ell I \preceq G_k \preceq LI$. Thus, from the definition of η_k, we have

$$\eta_k = \frac{g_k^T H_k g_k}{\sqrt{g_k^T H_k G_k H_k g_k}} \leq \frac{\|g_k\|\|H_k g_k\|}{\sqrt{\ell}\|H_k g_k\|} = \frac{1}{\sqrt{\ell}}\|g_k\|$$

By Assumption 3, we find that $\|g_k\| = \|g_k(x_k)\| \leq \gamma$. Hence, we may take $\Gamma = \frac{\gamma}{\sqrt{\ell}}$. We also find that

$$\eta_k = \frac{g_k^T H_k g_k}{\sqrt{g_k^T H_k G_k H_k g_k}} \geq \frac{\lambda}{\Lambda\sqrt{L}}\|g_k\|$$

□

Lemma B.3. *The empirical objective function $F_k(x)$ satisfies*

$$F_k(x_{k+1}) - F_k(x_k^*) \leq r(F_k(x_k) - F_k(x_k^*))$$

for the global constant $r = 1 - \frac{\ell}{(1+\Gamma)L} < 1$.

Proof. Observe that the function $\omega(z)$ satisfies $\omega(z) \geq \frac{1}{2}(1+\Gamma)^{-1}z^2$ for all $z \in [0, \Gamma]$. Also, recall that the strongly convex function F_k satisfies $\|g_k(x)\|^2 \geq 2\ell(F_k(x) - F_k(x_k^*))$. By Theorem A.3 and Theorem B.2, we find that

$$\begin{aligned} F_k(x_{k+1}) - F_k(x_k^*) \leq F_k(x_k) - F_k(x_k^*) - \omega(\eta_k) &\leq F_k(x_k) - F_k(x_k^*) - \frac{1}{2}(1+\Gamma)^{-1}\eta_k^2 \\ &\leq F_k(x_k) - F_k(x_k^*) - \frac{1}{2}(1+\Gamma)^{-1}\frac{\lambda^2}{\Lambda^2 L}\|g_k\|^2 \\ &\leq \left(1 - \frac{\lambda^2\ell}{(1+\Gamma)\Lambda^2 L}\right)(F_k(x_k) - F_k(x_k^*)) \end{aligned}$$

Thus, we may take $r = 1 - \frac{\lambda^2\ell}{(1+\Gamma)\Lambda^2 L}$. For SA-GD in particular, $\lambda = \Lambda = 1$, so $r = 1 - \frac{\ell}{(1+\Gamma)L}$. □

We are now ready to prove Theorem B.1.

Proof. By Lemma B.3, we calculate that

$$\begin{aligned} F_k(x_{k+1}) - F_k(x_k^*) &\leq r(F_k(x_k) - F_k(x_k^*)) \\ &= r(F_{k-1}(x_k) - F_{k-1}(x_{k-1}^*)) \\ &\quad + r(F_k(x_k) - F(x_k) - F_{k-1}(x_k) + F(x_k)) \\ &\quad + r(F_{k-1}(x_{k-1}^*) - F(x^*) - F_k(x_k^*) + F(x^*)) \\ &\leq r(F_{k-1}(x_k) - F_{k-1}(x_{k-1}^*)) \\ &\quad + r(\sup_{x\in\mathcal{D}}|F_k(x) - F(x)| + \sup_{x\in\mathcal{D}}|F_{k-1}(x) - F(x)|) \\ &\quad + r(|F_k(x_k^*) - F(x^*)| + |F_{k-1}(x_{k-1}^*) - F(x^*)|) \end{aligned}$$

By iterating this expansion, we find that

$$\begin{aligned} F_k(x_{k+1}) - F_k(x_k^*) \leq\ & r^k(F_0(x_1) - F_0(x_0^*)) \\ & + \sum_{j=1}^{k} r^j(\sup_{x\in\mathcal{D}} |F_{k+1-j}(x) - F(x)| + \sup_{x\in\mathcal{D}} |F_{k-j}(x) - F(x)|) \\ & + \sum_{j=1}^{k} r^j(|F_{k+1-j}(x_{k+1-j}^*) - F(x^*)| + |F_{k-j}(x_{k-j}^*) - F(x^*)|) \end{aligned}$$

Decompose $F_k(x_{k+1}) - F_k(x_k^*)$ as

$$F_k(x_{k+1}) - F_k(x_k^*) = F(x_{k+1}) - F(x^*) + [F_k(x_{k+1}) - F(x_{k+1})] + [F(x^*) - F_k(x_k^*)]$$

We can move the terms in square brackets to the right hand side, and upper bound them, to obtain

$$\begin{aligned} F(x_{k+1}) - F(x^*) \leq\ & r^k(F_0(x_1) - F_0(x_0^*)) \\ & + \sup_{x\in\mathcal{D}} |F_k(x) - F(x)| \\ & + \sum_{j=1}^{k} r^j(\sup_{x\in\mathcal{D}} |F_{k+1-j}(x) - F(x)| + \sup_{x\in\mathcal{D}} |F_{k-j}(x) - F(x)|) \\ & + |F_k(x_k^*) - F(x^*)| \\ & + \sum_{j=1}^{k} r^j(|F_{k+1-j}(x_{k+1-j}^*) - F(x^*)| + |F_{k-j}(x_{k-j}^*) - F(x^*)|) \end{aligned} \tag{2}$$

Suppose that we draw a constant number of samples $m_k = m$ at each iteration. Taking expectations on both sides of equation (2) and applying the concentration bound of Theorem A.1, we obtain

$$\begin{aligned} \mathbb{E}F(x_{k+1}) - F(x^*) &\leq r^k(u-l) + 2C\sqrt{\frac{\log m}{m}}\sum_{j=0}^{k} r^j \\ &\leq r^k(u-l) + \frac{2C}{1-r}\sqrt{\frac{\log m}{m}} \end{aligned}$$

In order to obtain an ϵ-optimal solution, we may use sufficiently large samples, and take sufficiently many iterations, so that

$$\begin{aligned} r^k(u-l) &\leq \frac{\epsilon}{2} \\ \frac{2C}{1-r}\sqrt{\frac{\log m}{m}} &\leq \frac{\epsilon}{2} \end{aligned}$$

This yields the given bounds on m and k in Theorem B.1. □

In particular, it suffices to take $m = O(\epsilon^{-2}\log\epsilon^{-1})$ and $k = O(\log\epsilon^{-1})$.

C. Convergence of SA-BFGS

Our goal in this section is to prove that SA-BFGS converges superlinearly with probability 1.

Theorem C.1. *Suppose that we draw m_k samples on the k-th step, where m_k^{-1} converges R-superlinearly to 0. Then SA-BFGS converges to the optimal solution x^* almost surely.*

Our arguments closely follow the proofs given in (Powell, 1976) and (Griewank & Toint, 1982) for the deterministic BFGS method.

Along the way, we will also consider the behavior of SA-BFGS when ϵ-optimality suffices, and m_k is held constant. Note that the results preceding Lemma C.10 do not depend on any particular choice of sample sizes m_k.

We introduce the following assumption in this section:

4. The Hessian $G(x)$ is Lipschitz continuous with constant L_H.

The adaptive step size is known to satisfy the Armijo-Wolfe conditions in the deterministic setting. A similar property holds for the empirical objective functions.

Theorem C.2 (Theorem 6.2, (Gao & Goldfarb, 2016))**.** *The adaptive step size t_k satisfies the Armijo condition for $\alpha = \frac{1}{2}$, for the empirical objective function $F_k(x)$.*

Recall that the SA-BFGS algorithm performs a BFGS update at step k only if t_k satisfies the Wolfe condition. If t_k does not satisfy the Wolfe condition, then we take a SA-GD step instead. In this case, the direction is $-g_k$ and the step size is the adaptive step size for SA-GD.

We use $q(j)$ to denote the the index of the j-th BFGS step, or equivalently, the index at which the j-th BFGS update is performed. The steps $\{q(j)\}_{j=1}^{\infty}$ where we perform BFGS updates will be referred to as *update times*. Later on, we will see that if m_k grows at a sufficient rate, then all $q(j)$ exist with probability 1.

The following technical lemma is used in the analysis of BFGS; it can also be found in (Byrd et al., 1987) and (Powell, 1976).

Lemma C.3. *Let $k = q(j)$ be an update time. Let $\overline{G}_k = \int_0^1 G_k(x_k + \tau s_k) d\tau$, and let θ_k denote the angle between the vectors $-g_k$ and s_k. Then*

1. *$y_k = \overline{G}_k s_k$, and $s_k^T y_k \leq L\|s_k\|^2$.*
2. *$\|s_k\| \leq \frac{1}{\ell}\|g_k\| \cos\theta_k$*
3. *If the Wolfe condition is satisfied on step k, then $\langle y_k, s_k\rangle \geq (1-\beta)\langle -g_k, s_k\rangle$ and $\|s_k\| \geq \frac{(1-\beta)}{L}\|g_k\|\cos\theta_k$.*

Proof. The first statement follows from the definition $y_k = g_k(x_{k+1}) - g_k(x_k)$. Since $G_k(x) \preceq LI$ for all x, we also have $\overline{G}_k \preceq LI$, and hence $s_k^T y_k = s_k^T \overline{G}_k s_k \leq L\|s_k\|^2$.

The second statement follows from the Armijo condition (Theorem C.2) and Taylor's theorem. Let $\overline{x}$ be a point on the line $[x_k, x_{k+1}]$ with $F_k(x_{k+1}) = F_k(x_k) + \langle g_k, s_k\rangle + \frac{1}{2}s_k^T G_k(\overline{x}) s_k$. Since $F_k(x_{k+1}) - F_k(x_k) \leq \frac{1}{2}\langle g_k, s_k\rangle$, we have $\frac{1}{2}\langle -g_k, s_k\rangle \geq \frac{1}{2}s_k^T G_k(\overline{x}) s_k \geq \frac{1}{2}m\|s_k\|^2$ as desired.

The Wolfe condition implies that $\langle y_k, s_k\rangle = \langle g_k(x_{k+1}) - g_k(x_k), s_k\rangle \geq (1-\beta)\langle -g_k, s_k\rangle$. Writing $\langle -g_k, s_k\rangle = \|g_k\|\|s_k\|\cos\theta_k$, we have $L\|s_k\|^2 \geq (1-\beta)\|g_k\|\|s_k\|\cos\theta_k$, which gives the last statement. □

The next result is the key technical lemma in proving that SA-BFGS converges R-linearly. Its proof is identical to the deterministic case (Powell, 1976).

Lemma C.4. *There exists a global constant c such that*

$$\prod_{j=1}^{k} \frac{\|g_{q(j)}\|^2}{\langle -g_{q(j)}, s_{q(j)}\rangle} \leq c^k$$

Proof. By considering the BFGS update formula, we have

$$\mathrm{Tr}(B_{j+1}) = \mathrm{Tr}(B_j) - \frac{s_j^T B_j^2 s_j}{s_j^T B_j s_j} + \frac{y_j^T y_j}{s_j^T y_j}$$

Recall from Lemma C.3 that $y_j = \overline{G}_j s_j$. Therefore, writing $z_j = \overline{G}_j^{1/2} s_j$, we have

$$\frac{y_j^T y_j}{s_j^T y_j} = \frac{z_j^T \overline{G}_j z_j}{z_j^T z_j} \leq L$$

where the last inequality follows from Assumption 1. Let $c_1 = \mathrm{Tr}(B_0) + kL$. The BFGS formula implies that $\mathrm{Tr}(B_{q(k+1)}) \leq \mathrm{Tr}(B_0) + kL \leq c_1 k$, and since $B_{q(k+1)}$ is positive definite, we also have

$$\sum_{j=1}^{k} \frac{s_{q(j)}^T B_{q(j)}^2 s_{q(j)}}{s_{q(j)}^T B_{q(j)} s_{q(j)}} \leq \mathrm{Tr}(B_0) + kL \leq c_1 k$$

Observe that $s_j^T B_j^2 s_j = t_j^2 \|g_j\|^2$ and that $s_j^T B_j s_j = t_j \langle -g_j, s_j \rangle$. By the arithmetic mean-geometric mean (AM-GM) inequality,

$$\prod_{j=1}^{k} \frac{t_{q(j)} \|g_{q(j)}\|^2}{\langle -g_{q(j)}, s_{q(j)} \rangle} \leq c_1^k \tag{3}$$

Next, we use the recursive formula for the determinant:

$$\det(B_{j+1}) = \frac{y_j^T s_j}{s_j^T B_j s_j} \det(B_j)$$

Since the Wolfe condition is satisfied, we have

$$y_j^T s_j = (g_j(x_{j+1}) - g_j(x_j))^T s_j \geq (1-\beta)\langle -g_j, s_j \rangle$$

Therefore,

$$\det(B_{q(k+1)}) \geq \det(B_0) \prod_{j=1}^{k} \frac{1-\beta}{t_{q(j)}}$$

By the AM-GM inequality applied to the eigenvalues of $B_{q(k+1)}$, we find that $\det(B_{q(k+1)}) \leq (c_1 k/n)^n \leq c_2^k$ for a global constant c_2. Hence, $\prod_{j=1}^{k} \frac{1-\beta}{t_{q(j)}} \leq c_2^k$. Multiplying this together with inequality (3), and taking $c = \frac{c_1}{(1-\beta)c_2}$, we find that

$$\prod_{j=1}^{k} \frac{\|g_{q(j)}\|^2}{\langle -g_{q(j)}, s_{q(j)} \rangle} \leq c^k$$

as desired. □

Lemma C.5. *At least $\frac{1}{2}k$ of the angles $\theta_{q(1)}, \ldots, \theta_{q(k)}$ satisfy $\cos^2 \theta_{q(j)} > (\ell/c)^2$, where c is the constant of Lemma C.4.*

Proof. By Lemma C.3, $\|s_j\| \leq \frac{1}{\ell}\|g_j\| \cos\theta_j$. Substituting this in Lemma C.4 yields

$$c^k \geq \prod_{j=1}^{k} \frac{\|g_{q(j)}\|^2}{\langle -g_{q(j)}, s_{q(j)} \rangle} \geq \prod_{j=1}^{k} \frac{\ell}{\cos^2 \theta_{q(j)}} = \ell^{k+1} \prod_{j=1}^{k} \frac{1}{\cos^2 \theta_{q(j)}}$$

Hence, $\prod_{j=1}^{k} \cos^2 \theta_{q(j)} \geq (\ell/c)^k$. It follows that at least $\frac{1}{2}k$ of the angles must satisfy $\cos^2 \theta_{q(j)} \geq (\ell/c)^2$. □

We can proceed to show that stochastic adaptive BFGS converges R-linearly. The argument proceeds by showing that if k is not an update time, then SA-BFGS inherits the Q-linear convergence rate of SA-GD, and if $k = q(j)$, then we can measure the decrement with Lemma C.4.

Lemma C.6. *If k is not an update time, then*

$$F_k(x_{k+1}) - F_k(x_k^*) \leq r(F_k(x_k) - F_k(x_k^*))$$

where $r = 1 - \frac{\ell^{3/2}}{(\sqrt{\ell}+\gamma)L}$.

Proof. This follows from Lemma B.3 for SA-GD. □

Lemma C.7. *Let $k = q(j)$. Then*

$$F_k(x_{k+1}) - F_k(x_k^*) \leq \left(1 - (1-\beta)\ell L^{-1}\cos^2\theta_k\right)\left(F_k(x_k) - F_k(x_k^*)\right)$$

Proof. Since the adaptive step size t_k satisfies the Armijo condition for $\alpha = \frac{1}{2}$, we have

$$F_k(x_{k+1}) - F_k(x_k) \leq \frac{1}{2}\langle g_k, s_k\rangle = -\frac{1}{2}\|g_k\|\|s_k\|\cos\theta_k$$

Using Lemma C.3, we rewrite $\|s_k\|$ in terms of $\|g_k\|, \cos\theta_k$ to obtain

$$F_k(x_{k+1}) - F_k(x_k) \leq -\frac{1}{2}(1-\beta)L^{-1}\|g_k\|^2\cos^2\theta_k$$

Since $\|g_k\|^2 \geq 2\ell(F_k(x_k) - F_k(x_k^*))$, we rearrange to obtain

$$F_k(x_{k+1}) - F_k(x_k^*) \leq (1 - (1-\beta)\ell L^{-1}\cos^2\theta_k)(F_k(x_k) - F_k(x_k^*))$$

□

Theorem C.8. *Suppose that we draw samples of size m_k at step k, where m_k^{-1} converges superlinearly to 0. With probability 1, SA-BFGS converges R-linearly.*

Proof. Let $\nu = \max\{1 - (1-\beta)\ell L^{-1}(\ell/c)^2, r\} < 1$. Let $\mathbb{I}_1(k)$ be the 0-1 indicator variable for the event that k is a BFGS update time, and let $\mathbb{I}_2(k)$ be the indicator for the event that k is a BFGS update time *and* $\cos^2\theta_k \geq (\ell/c)^2$. Combining Lemma C.6 and Lemma C.7 by using these indicator variables, we have

$$\begin{aligned}
F_k(x_{k+1}) - F_k(x_k^*) &\leq (1 - (1-\beta)\ell L^{-1}\cos^2\theta_k)^{\mathbb{I}_1(k)} r^{1-\mathbb{I}_1(k)}(F_k(x_k) - F_k(x_k^*)) \\
&\leq (1 - (1-\beta)\ell L^{-1}(\ell/c)^2)^{\mathbb{I}_2(k)} r^{1-\mathbb{I}_1(k)}(F_k(x_k) - F_k(x_k^*)) \\
&\leq \nu^{\mathbb{I}_2(k)+1-\mathbb{I}_1(k)}(F_k(x_k) - F_k(x_k^*))
\end{aligned}$$

For any $t \leq k$, let $b(t) = \sum_{j=0}^{t}\mathbb{I}_1(j)$. Rewritten with indicators, Lemma C.5 states that $\sum_{j=0}^{t}\mathbb{I}_2(j) \geq \frac{1}{2}b(t)$. Therefore

$$\sum_{j=0}^{k}(\mathbb{I}_2(j) + 1 - \mathbb{I}_1(j)) \geq k - \frac{1}{2}b$$

Define $\mathbb{I}_3(k) = \mathbb{I}_2(k) + 1 - \mathbb{I}_1(k)$. Iterating the above expansion, we have

$$\begin{aligned}
F_k(x_{k+1}) - F_k(x_k^*) &\leq \nu^{\mathbb{I}_3(k)}(F_k(x_k) - F_k(x_k^*)) \\
&\leq \nu^{\mathbb{I}_3(k)}(F_{k-1}(x_k) - F_{k-1}(x_{k-1}^*) + (F_k(x_k) - F_{k-1}(x_k)) + (F_{k-1}(x_{k-1}^*) - F_k(x_k^*)) \\
&\leq \nu^{\sum_{i=0}^{k}\mathbb{I}_3(i)}(F_0(x_0) - F_0(x_0^*)) \\
&\quad + \sum_{j=1}^{k}\nu^{\sum_{i=j}^{k}\mathbb{I}_3(i)}[\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| + \sup_{x\in\mathcal{D}}|F_{j-1}(x) - F(x)|] \\
&\quad + \sum_{j=1}^{k}\nu^{\sum_{i=j}^{k}\mathbb{I}_3(i)}[|F_j(x_j^*) - F(x^*)| + |F_{j-1}(x_{j-1}^*) - F(x^*)|] \\
&\leq \nu^{k-b/2}(F_0(x_0) - F_0(x_0^*)) \\
&\quad + 2\sum_{0\leq j\leq k-b/2}\nu^{k-b/2-j}(\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| + |F_j(x_j^*) - F(x^*)|) \\
&\quad + 2\sum_{j>k-b/2}^{k}(\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| + |F_j(x_j^*) - F(x^*)|)
\end{aligned}$$

In the last inequality, we have simply split the sums into two sums, one running over the indices $0 \leq j \leq k - b/2$ and the other over $k - b/2 < j \leq k$. Writing the left side as

$$F_k(x_{k+1}) - F_k(x_k^*) = F(x_{k+1}) - F(x^*) + (F_k(x_{k+1}) - F(x_{k+1})) + (F(x^*) - F_k(x_k^*))$$

we can move terms to the right to obtain

$$\begin{aligned} F(x_{k+1}) - F(x^*) \leq \nu^{k-b/2}(F_0(x_0) - F_0(x_0^*)) \\ + \sup_{x \in \mathcal{D}} |F_k(x) - F(x)| + |F_k(x_k^*) - F(x^*)| \\ + 2 \sum_{0 \leq j \leq k-b/2} \nu^{k-b/2-j} (\sup_{x \in \mathcal{D}} |F_j(x) - F(x)| + |F_j(x_j^*) - F(x^*)|) \\ + 2 \sum_{j > k-b/2}^{k} (\sup_{x \in \mathcal{D}} |F_j(x) - F(x)| + |F_j(x_j^*) - F(x^*)|) \end{aligned}$$

Taking expectations, and applying Theorem A.1 on the right, we have

$$\mathbb{E}F(x_{k+1}) - F(x^*) \leq \nu^{k-b/2}(u - l) + 4C \sum_{0 \leq j \leq k-b/2} \nu^{k-b/2-j} \sqrt{\frac{\log m_j}{m_j}} + 4C \sum_{j > k-b/2}^{k} \sqrt{\frac{\log m_j}{m_j}} \tag{4}$$

Our choice of m_j satisfies $m_j = \Omega(\nu^{-2j})$, so $\sqrt{\frac{\log m_j}{m_j}} = O(\nu^j \sqrt{j})$. Hence, by bounding each term with a multiple of $\nu^{k-b/2}$, we may find a global constant ϕ, with $1 > \phi > \nu$, and a global constant c_3, such that

$$\mathbb{E}F(x_{k+1}) - F(x^*) \leq c_3 \phi^{k-b/2}$$

Clearly $b \leq k$, and thus we find that

$$\mathbb{E}F(x_{k+1}) - F(x^*) \leq c_3 \phi^{k/2}$$

Now, fix any constant φ with $\phi < \varphi < 1$. By Markov's inequality,

$$\mathbb{P}(F(x_k) - F(x^*) \geq \varphi^{k/2}) \leq \frac{\mathbb{E}(F(x_k) - F(x^*))}{\varphi^{k/2}} \leq c_3 \left(\frac{\phi}{\varphi}\right)^{k/2}$$

Since $\sum_{k=0}^{\infty} \left(\frac{\phi}{\varphi}\right)^{k/2} < \infty$, the Borel-Cantelli Lemma implies that the sequence of events A_k with

$$A_k = \{F(x_k) - F(x^*) > \varphi^{k/2}\}$$

occurs finitely often with probability 1. Therefore, with probability 1, SA-BFGS converges R-linearly. □

Before proceeding further, let us digress briefly to consider the behavior of SA-BFGS when we are satisfied with an ϵ-optimal solution, and wish to hold the number of samples constant.

Lemma C.9. *Let $\epsilon > 0$. Suppose we draw m i.i.d samples at each step, where $m = O(\epsilon^2(\log \epsilon^{-1})^3)$. Then SA-BFGS converges in expectation to an ϵ-optimal solution after k steps, where $k = O(\epsilon^{-1})$.*

Proof. Note that equation (4) in the proof of Theorem C.8 holds in the absence of any assumptions on the sample sizes m_k. Suppose that we take $m_k = m$. Then we have

$$\begin{aligned} \mathbb{E}F(x_{k+1}) - F(x^*) &\leq \nu^{k-b/2}(u - l) + 4C \sum_{0 \leq j \leq k-b/2} \nu^{k-b/2-j} \sqrt{\frac{\log m_j}{m_j}} + 4C \sum_{j > k-b/2}^{k} \sqrt{\frac{\log m_j}{m_j}} \\ &\leq \nu^{k/2}(u - l) + 4C \sqrt{\frac{\log m}{m}} \left(\frac{1}{1 - \nu} + k/2\right) \end{aligned}$$

Therefore, in order to obtain an ϵ-optimal solution from SA-BFGS, we may take

$$\nu^{k/2}(u-l) \leq \frac{\epsilon}{2}$$
$$\frac{4C}{1-r}\sqrt{\frac{\log m}{m}}\left(\frac{1}{1-\nu}+k/2\right) \leq \frac{\epsilon}{2}$$

Thus, it suffices to take $k = \log(\epsilon^{-1}2(u-l))/\log\nu$. Substituting this value of k into the second inequality, we see that it suffices to take $m = O(\epsilon^2(\log\epsilon^{-1})^3)$. □

We now concern ourselves with R-superlinear convergence to the true optimal solution. Henceforth, we assume that the sample sizes grow so that m_k^{-1} converges R-superlinearly to 0.

Lemma C.10. *We have $\sum_{k=0}^{\infty}\omega(\eta_k) < \infty$ with probability 1. In particular, $\eta_k \to 0$ almost surely.*

Proof. By Theorem A.3, we find that

$$\begin{aligned}
F_k(x_{k+1}) &\leq F_k(x_k) - \omega(\eta_k) \\
&= F_{k-1}(x_k) + (F_k(x_k) - F_{k-1}(x_k)) - \omega(\eta_k) \\
&\leq F_0(x_0) + \sum_{j=1}^{k}(F_j(x_j) - F_{j-1}(x_j)) - \sum_{j=0}^{k}\omega(\eta_j) \\
&\leq F_0(x_0) + \sum_{j=1}^{k}\sup_{x\in\mathcal{D}}|F_j(x) - F_{j-1}(x)| - \sum_{j=0}^{k}\omega(\eta_j) \\
&\leq F_0(x_0) + 2\sum_{j=1}^{k}\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| - \sum_{j=0}^{k}\omega(\eta_j) \\
&\leq F_0(x_0) + 2\sum_{j=1}^{\infty}\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| - \sum_{j=0}^{k}\omega(\eta_j)
\end{aligned}$$

Let $Y = \sum_{j=1}^{\infty}\sup_{x\in\mathcal{D}}|F_j(x) - F(x)|$. By the monotone convergence theorem and Theorem A.1, we have

$$\mathbb{E}Y = \sum_{j=1}^{\infty}\mathbb{E}\sup_{x\in\mathcal{D}}|F_j(x) - F(x)| \leq C\sum_{j=1}^{\infty}\sqrt{\frac{\log m_j}{m_j}}$$

By our choice of m_j, the latter sum is finite. This implies that $\mathbb{P}(Y < \infty) = 1$. Since $F_k(x)$ is bounded below on $\mathcal{D}$ by Assumption 3, we necessarily have $\sum_{k=0}^{\infty}\omega(\eta_k) < \infty$ whenever $Y < \infty$. Thus $\eta_k \to 0$ with probability 1. □

Theorem C.11. *Fix any $\beta < 1$. With probability 1, there exists a finite index k_0 such that the Wolfe condition is satisfied for all $k \geq k_0$.*

Proof. This follows from Theorem 6.3 in (Gao & Goldfarb, 2016), for any realization of the empirical objective functions $F_0, F_1, \ldots$ such that $\eta_k \to 0$. By Lemma C.10, the event $\eta_k \to 0$ occurs with probability 1. □

In particular, this implies that with probability 1, there exists a finite time k_0 after which every step is a BFGS step, and BFGS updates are always performed.

Corollary C.12. *With probability 1, we have $\sum_{k=0}^{\infty}\|x_k - x^*\| < \infty$.*

Proof. This follows from Theorem C.8. Let $\{x_k\}_{k=0}^{\infty}$ be any instance of the algorithm where $F(x_k) \leq F(x^*) + \varphi^{k/2}$ for all $k \geq k_0$, for some index k_0. Since $F(x)$ is strongly convex,

$$\|x_k - x^*\| \leq \frac{2}{\ell}(F(x_k) - F(x^*)) \leq \frac{2}{\ell}\varphi^{k/2}$$

for all $k \geq k_0$. Hence $\sum_{k=0}^{\infty}\|x_k - x^*\| < \infty$. By Theorem C.8, this occurs with probability 1. □

Let us define $e_k = \max\{\|x_k - x^*\|, \|x_{k+1} - x^*\|\}$. Corollary C.12 implies that $\sum_{k=0}^{\infty} e_k < \infty$.

Next, we perform a detailed analysis of the evolution of H_{k+1}. By applying Corollary A.2, we can use a modified form of the classical argument ((Griewank & Toint, 1982)) on a path-by-path basis.

Corollary C.13. *Let $\sigma_k = m_k^{-2/5}$. By taking $\delta = \sigma_k$ in Corollary A.2, we can find global constants c_4 and $\omega < 1$ such that*

$$\mathbb{P}(\sup_{x\in\mathcal{D}} \|G_k(x) - G(x)\| > \sigma_k \text{ or } \sup_{x\in\mathcal{D}} \|g_k(x) - g(x)\| > \sigma_k) \leq c_4\omega^k$$

Hence, with probability 1, there exists an index k_0 such that for all $k \geq k_0$, we have both $\sup_{x\in\mathcal{D}} \|G_k(x) - G(x)\| < \sigma_k$ and $\sup_{x\in\mathcal{D}} \|g_k(x) - g(x)\| < \sigma_k$.

By construction, $\{\sigma_k\}$ converges to 0 at a R-superlinear rate.

Proof. The first part follows by Corollary A.2. Taking $\epsilon = \frac{\delta}{2L+1}$, our probability bound is

$$\mathbb{P}(\sup_{x\in\mathcal{D}} \|G_k(x) - G(x)\| > \sigma_k \text{ or } \sup_{x\in\mathcal{D}} \|g_k(x) - g(x)\| > \sigma_k) \leq C_1 \exp(\frac{2}{5} n \log m_k - C_2(1 - \frac{C_3}{2L+1})^2 m_k^{1/5})$$

Since $\frac{m_k^{1/5}}{\log m_k} \to 0$ and $m_k = \Omega(k^5)$ by construction, we can find the desired $\omega < 1$. The second statement then follows immediately from the Borel-Cantelli Lemma. □

Let Ω denote the space of paths where $\sum_{k=0}^{\infty} e_k < \infty$ and for some k_0, $\sup_{x\in\mathcal{D}} \|G_k(x) - G(x)\| \leq \sigma_k$ and $\sup_{x\in\mathcal{D}} \|g_k(x) - g(x)\| \leq \sigma_k$ for all $k \geq k_0$. By Corollary C.12 and Corollary C.13, $\mathbb{P}(\Omega) = 1$. Henceforth, we restrict our analysis to the paths belonging to Ω.

The BFGS algorithm is invariant under a linear change of variables, so without loss of generality, we may assume that $G(x^*) = I$. This corresponds to the change of variables $\widetilde{F}(y) = F(G(x^*)^{-1/2}y), y = G(x^*)^{1/2}x$. Define two 'hypothetical' updates:

$$\widehat{B}_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{G_k(x^*) s_k s_k^T G_k(x^*)}{s_k^T G_k(x^*) s_k}$$
$$\widetilde{B}_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{G(x^*) s_k s_k^T G(x^*)}{s_k^T G(x^*) s_k}$$

Lemma C.14. *We have*

$$\|\widetilde{B}_{k+1} - I\|_F^2 \leq \|B_k - I\|_F^2$$

and

$$\|\widetilde{H}_{k+1} - I\|_F^2 \leq \|H_k - I\|_F^2$$

Proof. For brevity, we write $s = s_k, B = B_k, H = H_k$. By a routine calculation (see §4 of (Griewank & Toint, 1982)), we have

$$\|\widetilde{B}_{k+1} - I\|_F^2 - \|B_{k+1} - I\|_F^2 = -\left[\left(1 - \frac{s^T B^2 s}{s^T B s}\right)^2 + 2\left(\frac{s^T B^3 s}{s^T B s} - \left(\frac{s^T B^2 s}{s^T B s}\right)^2\right)\right]$$

and

$$\|\widetilde{H}_{k+1} - I\|_F^2 - \|H_{k+1} - I\|_F^2 = -\left[\left(1 - \frac{s^T H s}{s^T s}\right)^2 + 2\left(\frac{s^T H^2 s}{s^T s} - \left(\frac{s^T H s}{s^T s}\right)^2\right)\right]$$

The Cauchy-Schwarz inequality implies that the latter terms in the brackets are non-positive, which gives the desired result. □

Lemma C.15. *Every path in Ω satisfies*

$$\|B_{k+1} - \widetilde{B}_{k+1}\| \leq O(e_k + \sigma_k)$$

and

$$\|H_{k+1} - \widetilde{H}_{k+1}\| \leq (\|H_k - I\| + 1)O(e_k + \sigma_k)$$

Proof. We again write $s = s_k, y = y_k, B = B_k, H = H_k$ for brevity.

We can bound the difference $\|B_{k+1} - \widehat{B}_{k+1}\|$, as both updates are performed with sampled gradients, and then use Corollary C.13 to bound $\|\widehat{B}_{k+1} - \widetilde{B}_{k+1}\}$.

Take $\Delta = G_k(x^*)s - y$. By Lemma C.3, we can write $y = G_k(\widehat{x})s$ for some $\widehat{x}$ on the line segment $[x_k, x_{k+1}]$, and we deduce that:

1. $\ell\|s\|^2 \leq y^T s \leq L\|s\|^2$
2. $\|\Delta\| \leq L_H e_k \|s\|$.
3. $\frac{y^T\Delta}{s^T y} \leq L L_H e_k$

Hence, writing $\frac{1}{s^T y + \Delta^T s} = \frac{1}{s^T y} - \frac{y^T\Delta}{s^T y + y^T\Delta}$, we have

$$\begin{aligned}
\|B_{k+1} - \widehat{B}_{k+1}\| &= \left\| \frac{yy^T}{s^T y} - \frac{(y+\Delta)(y+\Delta)^T}{(y+\Delta)^T s} \right\| \\
&= \left\| -\frac{y\Delta^T + \Delta y^T + \Delta\Delta^T}{s^T y} + \frac{y^T\Delta(yy^T + y\Delta^T + \Delta y^T + \Delta\Delta^T)}{s^T y + y^T\Delta} \right\| \\
&\leq O(e_k)
\end{aligned}$$

Next, write $\widehat{y} = G_k(x^*)s$ and $\widetilde{y} = G(x^*)s$. Since our path lies in Ω, we know that $\|G_k(x^*) - G(x^*)\| \leq \sigma_k$. Let $\Delta = \widehat{y} - \widetilde{y}$, so $\|\Delta\| \leq \sigma_k\|s\|$, and perform the same calculation as above to obtain

$$\begin{aligned}
\|\widehat{B}_{k+1} - \widetilde{B}_{k+1}\| &= \left\| -\frac{\widetilde{y}\Delta^T + \Delta\widetilde{y}^T + \Delta\Delta^T}{s^T\widetilde{y}} + \frac{\widetilde{y}^T\Delta(\widetilde{y}\widetilde{y}^T + \widetilde{y}\Delta^T + \Delta\widetilde{y}^T + \Delta\Delta^T)}{s^T\widetilde{y} + \widetilde{y}^T\Delta} \right\| \\
&\leq O(\sigma_k)
\end{aligned}$$

Hence, $\|B_{k+1} - \widetilde{B}_{k+1}\| \leq O(e_k + \sigma_k)$.

A similar calculation holds for H.

$$\begin{aligned}
\|H_{k+1} - \widehat{H}_{k+1}\| = \| &\frac{ss^T}{(y+\Delta)^T s} - \frac{ss^T}{s^T y} \\
&+ \left(\frac{s(y+\Delta)^T}{(y+\Delta)^T s} - \frac{sy^T}{s^T y}\right) H + H\left(\frac{(y+\Delta)s^T}{(y+\Delta)^T s} - \frac{ys^T}{s^T y}\right) \\
&+ \frac{s(y+\Delta)^T H(y+\Delta)s^T}{((y+\Delta)^T s)^2} - \frac{sy^T Hys^T}{(s^T y)^2}\|
\end{aligned}$$

It is elementary, though tedious, to verify that $\frac{ss^T}{(y+\Delta)^T s} - \frac{ss^T}{s^T y} \leq O(e_k)$ and that the other terms are bounded by $O(\|H\|e_k)$. The same calculation shows that $\|\widehat{H}_{k+1} - \widetilde{H}_{k+1}\| \leq O(\sigma_k + \|H\|\sigma_k)$. Thus, we have $\|H_{k+1} - \widetilde{H}_{k+1}\| \leq (\|H_k - I\| + 1)O(e_k + \sigma_k)$. □

Corollary C.16. *By Lemma C.15, Lemma C.14, and the triangle inequality,*

$$\|B_{k+1} - I\| \leq \|B_{k+1} - \widetilde{B}_{k+1}\| + \|\widetilde{B}_{k+1} - I\| \leq \|B_k - I\| + O(e_k + \sigma_k)$$

and

$$\|H_{k+1} - I\| \leq \|H_{k+1} - \widetilde{H}_{k+1}\| + \|\widetilde{H}_{k+1} - I\| \leq (\|H_k - I\| + 1)O(e_k + \sigma_k)$$

A lemma of Griewank and Toint shows that this forces the convergence of $\{\|B_k - I\|\}$ and $\{\|H_k - I\|\}$.

Lemma C.17 (Lemma 3.3 of (Griewank & Toint, 1982)). *Let $\{\phi_k\}$ and $\{\delta_k\}$ be sequences of non-negative numbers such that $\phi_{k+1} \leq (1+\delta_k)\phi_k + \delta_k$ and $\sum_{k=1}^{\infty} \delta_k < \infty$. Then $\{\phi_k\}$ converges.*

In our case, we take $\delta_k = e_k + \sigma_k$, as $\sum_{k=0}^{\infty}(e_k + \sigma_k) < \infty$ by Corollary C.12 and Corollary C.13.

Following §4 of (Griewank & Toint, 1982), our previous results yield the Dennis-Moré ((Dennis Jr. & Moré, 1974)) condition:

$$\lim_{k\to\infty} \frac{\|(B_k - I)s_k\|}{\|s_k\|} = 0$$

It only remains to show that this implies R-superlinear convergence in the stochastic setting. Since $I = G(x^*)$, we have

$$\begin{aligned}
\|B_k s_k - G(x^*)s_k\| &= \| - g_k - G(x^*)s_k + g_k(x_{k+1}) - g_k(x_{k+1})\| \\
&= \|g_k(x_{k+1}) - g_k - G(x^*)s_k - g_k(x_{k+1})\| \\
&= \| \int_0^1 (G_k(x_k + \tau s_k) - G(x^*))s_k d\tau - g_k(x_{k+1})\| \\
&= \| \int_0^1 (G(x_k + \tau s_k) - G(x^*))s_k d\tau + \int_0^1 (G_k(x + \tau s_k) - G(x_k + \tau s_k))s_k d\tau - g_k(x_{k+1})\| \\
&\geq \|g_k(x_{k+1})\| - (L_H e_k + \sigma_k)\|s_k\|
\end{aligned}$$

and therefore $\frac{\|g_k(x_{k+1})\|}{\|s_k\|} \to 0$. By Assumption 1, the empirical objective function $F_k(x)$ is strongly convex, and therefore

$$\frac{\|g_k(x_{k+1})\|}{\|s_k\|} \geq \frac{|\|g_k(x_{k+1}) - g_k(x^*)\| - \|g_k(x^*) - g(x^*)\||}{\|x_{k+1} - x^*\| + \|x_k - x^*\|} \tag{5}$$

To complete the analysis, let $a_k = \frac{\|g_{k+1}\|}{\|s_k\|}, b_k = \|g_k(x^*) - g(x^*)\|$, and $z_k = \|x_k - x^*\|$. Our above results show that $a_k \to 0$, and $b_k \leq \sigma_k$ tends to 0 R-superlinearly. For convenience, we assume without loss of generality that $\{b_k\}$ converges Q-superlinearly, by replacing $\{b_k\}$ by the Q-superlinear sequence bounding σ_k if necessary.

Rearrange inequality (5) to obtain

$$\ell z_{k+1} = \ell\|x_{k+1} - x^*\| \leq \|g_k(x_{k+1}) - g_k(x^*)\| \leq a_k(z_{k+1} + z_k) + b_k$$

Eventually, $a_k < \frac{1}{2}\ell$, as $a_k \to 0$. Beyond that point, we find that

$$z_{k+1} \leq \frac{a_k}{\ell - a_k} z_k + b_k \leq \frac{2}{\ell} a_k z_k + b_k \tag{6}$$

Let $c_k = \max\{a_k z_k, b_k\}$. Clearly $z_{k+1} \leq (2 + \frac{2}{\ell})c_k$, so it suffices to prove that $\{c_k\}$ converges superlinearly. There are two cases to consider. If $c_{k+1} = a_{k+1}z_{k+1}$, then

$$\frac{c_{k+1}}{c_k} = \frac{a_{k+1}z_{k+1}}{c_k} \leq a_{k+1}\frac{(2 + \frac{2}{\ell})c_k}{c_k} = \left(2 + \frac{2}{\ell}\right) a_{k+1}$$

and $a_k \to 0$. Otherwise, if $c_{k+1} = b_{k+1}$, then

$$\frac{c_{k+1}}{c_k} = \frac{b_{k+1}}{c_k} \leq \frac{b_{k+1}}{b_k}$$

and by construction, $\{b_k\}$ converges to 0 superlinearly, so $\frac{b_{k+1}}{b_k} \to 0$.

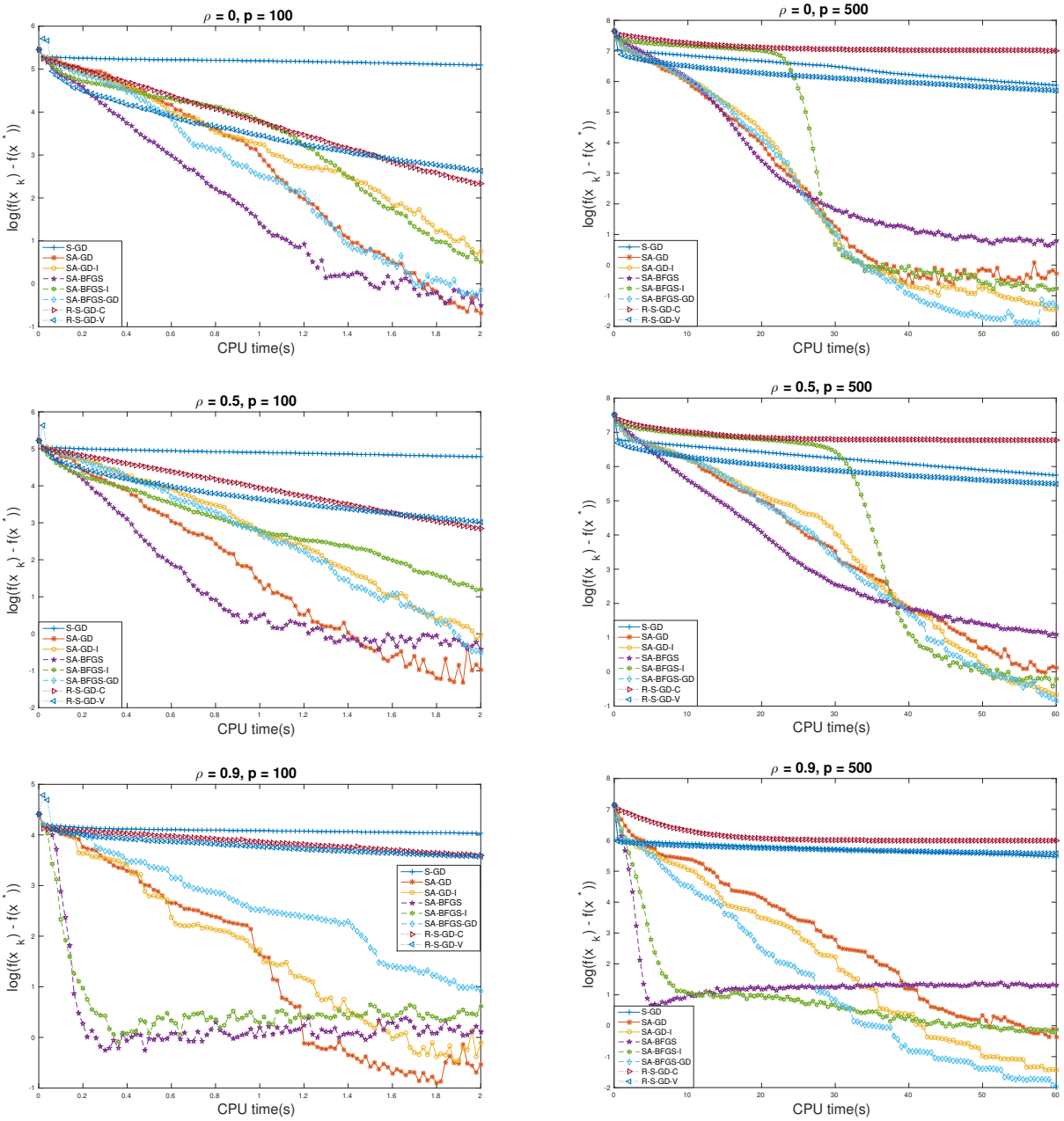
This proves that z_k converges R-superlinearly, and completes the proof of Theorem C.1.

D. Additional Experiments

To complement the numerical experiments for general stochastic optimization problems, we provide additional results for ERM (empirical risk minimization) problems. We compare all the algorithms in section 8 on ridge regression problems, that is,

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\|w\|_2^2,$$

where we set $n = 10^6$, $X_i \sim N(0, \Sigma(\rho))$, $\Sigma(\rho) = (1-\rho^2)I_p + \rho^2 J$ (here J is the all-ones matrix), β is a fixed p dimensional vector and $\lambda = 1$. We test problems of size $p = 100, 500$ and $\rho = 0, 0.5, 0.9$. From the figures, we



may draw similar conclusions as to those in section 6 for the methods that use an adaptive step length. One

interesting finding in this set of experiments is that the robust SGD methods do not work well especially for $p = 500$.

References

Byrd, Richard H., Nocedal, Jorge, and Yuan, Ya-Xiang. Global convergence of a class of quasi-Newton methods on convex problems. *Siam. J. Numer. Anal.*, (5):1171–1190, 1987.

Dennis Jr., John E. and Moré, Jorge J. Characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.*, 28(106):549–560, 1974.

Gao, Wenbo and Goldfarb, Donald. Quasi-Newton methods: Superlinear convergence without line search for self-concordant functions. *in review. arXiv:1612.06965*, 2016.

Goldfarb, Donald, Iyengar, Garud, and Zhou, Chaoxu. Linear convergence of stochastic Frank-Wolfe variants. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1066–1074, 2017.

Griewank, Andreas and Toint, Philippe L. Local convergence analysis for partitioned quasi-Newton updates. *Numer. Math.*, 39:429–448, 1982.

Powell, Michael J. D. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In Cottle, Richard and Lemke, C.E. (eds.), *Nonlinear Programming*, volume IX. SIAM-AMS Proceedings, 1976.

W. van der Vaart, Aad. and Wellner, Jon A. *Weak Convergence and Empirical Processes*. Springer New York, 1996.