A APPENDIX

A.1 PROOF OF THEOREM 1

Proof. First we show that the SMB step for each parameter group p can be expressed a special quasi-Newton update. For brevity, let us use s_k, s_k^t, g_k, g_k^t, and y_k instead of $s_{k,p}$, $s_{k,p}^t$, $g_{k,p}$, $g_{k,p}^t$, and $y_{k,p}$, respectively. Recalling the definitions of θ and δ given in (5), observe that

$$2\delta = \|s_k^t\|\|y_k\| + \frac{1}{\eta}\|s_k^t\|\|g_k\| - y_k^\top s_k^t = \alpha_k \left(\|g_k\|\|y_k\| + \frac{1}{\eta}\|g_k\|^2 + y_k^\top g_k \right) = \alpha_k \sigma,$$

and

$$\theta = \left(y_k^\top s_k^t + 2\delta\right)^2 - \|s_k^t\|^2\|y_k\|^2 = \alpha_k^2(\sigma - y_k^\top g_k)^2 - \alpha_k^2\|g_k\|^2\|y_k\|^2 = \alpha_k^2(\beta^2 - \|g_k\|^2\|y_k\|^2) = \alpha_k^2\gamma,$$

where

$$\sigma = \|g_k\|\|y_k\| + \frac{1}{\eta}\|g_k\|^2 + y_k^\top g_k,\ \beta = \sigma - y_k^\top g_k,\ \text{and}\ \gamma = (\beta^2 - \|g_k\|^2\|y_k\|^2).$$

Therefore, we have

$$c_g(\delta)g_k = -\frac{\|s_k^t\|^2}{2\delta}g_k = -\frac{\alpha_k^2\|g_k\|^2}{\alpha_k\sigma\gamma}\gamma g_k = -\alpha_k\frac{\|g_k\|^2}{\sigma\gamma}\gamma g_k,$$

$$\begin{aligned} c_y(\delta)y_k &= -\frac{\|s_k^t\|^2}{2\delta\theta}[-(y_k^\top s_k^t + 2\delta)(s_k^t)^\top g_k + \|s_k^t\|^2 y_k^\top g_k]y_k \\ &= -\frac{\|g_k\|^2}{\alpha_k\sigma\gamma}y_k[\alpha_k^2(\sigma - y_k^\top g_k)g_k^\top g_k + \alpha_k^2\|g_k\|^2 y_k^\top g_k] \\ &= -\alpha_k\frac{\|g_k\|^2}{\sigma\gamma}[\beta y_k g_k^\top + \|g_k\|^2 y_k y_k^\top]g_k, \end{aligned}$$

and

$$\begin{aligned} c_s(\delta)s_k^t &= -\frac{\|s_k^t\|^2}{2\delta\theta}[-(y_k^\top s_k^t + 2\delta)y_k^\top g_k + \|y_k\|^2(s_k^t)^\top g_k]s_k^t \\ &= -\frac{\|g_k\|^2}{\alpha_k\sigma\gamma}(-\alpha_k)g_k[-\alpha_k(\sigma - y_k^\top g_k)y_k^\top g_k - \alpha_k\|y_k\|^2 g_k^\top g_k] \\ &= -\alpha_k\frac{\|g_k\|^2}{\sigma\gamma}[\beta g_k y_k^\top + \|y_k\|^2 g_k g_k^\top]g_k. \end{aligned}$$

Now, it is easy to see that

$$\begin{aligned} s_k &= c_g(\delta)g_k + c_y(\delta)y_k + c_s(\delta)s_k^t \\ &= -\alpha_k\frac{\|g_k\|^2}{\sigma\gamma}\left[\gamma I + \beta y_k g_k^\top + \|g_k\|^2 y_k y_k^\top + \beta g_k y_k^\top + \|y_k\|^2 g_k g_k^\top\right]g_k. \end{aligned}$$

Thus, for each parameter group p, we define

$$H_{k,p} = \frac{\|g_{k,p}\|^2}{\sigma_p\gamma_p}\left[\gamma_p I + \beta_p y_{k,p}g_{k,p}^\top + \|g_{k,p}\|^2 y_{k,p}y_{k,p}^\top + \beta_p g_{k,p}y_{k,p}^\top + \|y_{k,p}\|^2 g_{k,p}g_{k,p}^\top\right], \tag{9}$$

where

$$\sigma_p = \|g_{k,p}\|\|y_{k,p}\| + \frac{1}{\eta}\|g_{k,p}\|^2 + y_{k,p}^\top g_{k,p},\ \beta_p = \sigma_p - y_{k,p}^\top g_{k,p},\ \text{and}\ \gamma_p = (\beta_p^2 - \|g_{k,p}\|^2\|y_{k,p}\|^2).$$

Now, assuming that we have the parameter groups $\{p_1, \ldots, p_n\}$, the SMB steps can be expressed as a quasi-Newton update given by

$$x_{k+1} = x_k - \alpha_k H_k g_k,$$

where

$$H_k = \begin{cases} I, & \text{if the Armijo condition is satisfied;} \\ \text{diag}(H_{k,p_1}, \ldots, H_{k,p_n}), & \text{otherwise.} \end{cases}$$

Here, I denotes the identity matrix, and $\text{diag}(H_{k,p_1}, \ldots, H_{k,p_n})$ denotes the block diagonal matrix with the blocks $H_{k,p_1}, \ldots, H_{k,p_n}$.

We next show that the eigenvalues of the matrices H_k, $k \geq 1$, are bounded from above and below uniformly which is, of course, obvious when $H_k = I$. Using the Sherman-Morrison formula twice, one can see that for each parameter group p, the matrix $H_{k,p}$ is indeed the inverse of the positive semidefinite matrix

$$B_{k,p} = \frac{1}{\|g_{k,p}\|^2}(\sigma_p I - g_{k,p}y_{k,p}^\top - y_{k,p}g_{k,p}^\top),$$

and hence, it is also positive semidefinite. Therefore, it is enough to show the boundedness of the eigenvalues of $B_{k,p}$ uniformly on k and p.

Since $g_{k,p}y_{k,p}^\top + y_{k,p}g_{k,p}^\top$ is a rank two matrix, $\sigma_p/\|g_{k,p}\|^2$ is an eigenvalue of $B_{k,p}$ with multiplicity $n-2$. The remaining extreme eigenvalues are

$$\lambda_{max}(B_{k,p}) = \frac{1}{\|g_{k,p}\|^2}(\sigma_p + \|g_{k,p}\|\|y_{k,p}\| - y_{k,p}^\top g_{k,p}) \quad \text{and} \quad \lambda_{min}(B_{k,p}) = \frac{1}{\|g_{k,p}\|^2}(\sigma_p - \|g_{k,p}\|\|y_{k,p}\| - y_{k,p}^\top g_{k,p}),$$

with the corresponding eigenvectors $\|y_{k,p}\|g_{k,p} + \|g_{k,p}\|y_{k,p}$ and $\|y_{k,p}\|g_{k,p} - \|g_{k,p}\|y_{k,p}$, respectively.

Observe that,

$$\begin{aligned} \lambda_{min}(B_{k,p}) &= \frac{\sigma_p - \|g_{k,p}\|\|y_{k,p}\| - y_{k,p}^\top g_{k,p}}{\|g_{k,p}\|^2} \\ &= \frac{\|g_{k,p}\|\|y_{k,p}\| + \eta^{-1}\|g_{k,p}\|^2 + y_{k,p}^\top g_{k,p} - \|g_{k,p}\|\|y_{k,p}\| - y_{k,p}^\top g_{k,p}}{\|g_{k,p}\|^2} \\ &= \frac{\eta^{-1}\|g_{k,p}\|^2}{\|g_{k,p}\|^2} = \frac{1}{\eta} > 1. \end{aligned}$$

Thus, the smallest eigenvalue $B_{k,p}$ is bounded away from zero uniformly on k and p.

Now, by our assumption of Lipschitz continuity of the gradients, for any $x, y \in \mathbb{R}^n$ and ξ_k, we have

$$\|g(x, \xi_k) - g(y, \xi_k)\| \leq L\|x - y\|.$$

Thus, observing that $\|y_{k,p}\| = \|g_{k,p}^t - g_{k,p}\| \leq L\|x_{k,p}^t - x_{k,p}\| \leq \alpha_k L\|g_{k,p}\|$, we have

$$\begin{aligned} \lambda_{max}(B_{k,p}) &= \frac{\sigma_p + \|g_{k,p}\|\|y_{k,p}\| - y_{k,p}^\top g_{k,p}}{\|g_{k,p}\|^2} \\ &= \frac{\|g_{k,p}\|\|y_{k,p}\| + \eta^{-1}\|g_{k,p}\|^2 + y_{k,p}^\top g_{k,p} + \|g_{k,p}\|\|y_{k,p}\| + y_{k,p}^\top g_{k,p}}{\|g_{k,p}\|^2} \\ &= \frac{2\|g_{k,p}\|\|y_{k,p}\| + \eta^{-1}\|g_{k,p}\|^2}{\|g_{k,p}\|^2} \leq 2L\alpha_k + \frac{1}{\eta} \leq 2L\alpha_{max} + \eta^{-1}. \end{aligned}$$

This implies that the eigenvalues of $H_{k,p} = B_{k,p}^{-1}$ are bounded below by $1/(\eta^{-1} + 2L\alpha_{max})$ and bounded above by 1 uniformly on k and p. This result, together with our assumptions, shows that steps of the SMBi algorithm satisfy the conditions of Theorem 2.10 in (Wang et al., 2017) with $\underline{\kappa} = 1/(\eta^{-1} + 2L\alpha_{max})$ and $\overline{\kappa} = 1$ and Theorem 1 follows as a corollary. □