Appendix A. Proof of Theorem 1

Theorem 1 *Suppose for objective function given by eq. (1), under the assumptions of μ-strong convexity, component-wise Lipschitz continuity of gradient, constant step size α random sampling without replacement for blocks and random sampling with replacement for mini-batches SAAG-II, converges linearly to optimal value p^*, for constant mini-batch size B, constant block size v, as given below,*

$$\mathbb{E}\left[f\left(w^k\right)-p^*\right] \leq \left(1-\frac{2v\alpha\mu}{p}\right)^k \left(f\left(w^0\right)-p^*\right)+\frac{pR_0^2}{2v\mu}\left(L\alpha\left(3+\frac{2B^2}{l^2}\right)-1+\frac{B}{l}\right) \quad (17)$$

where $\left\|\frac{1}{|B_i|}\sum_{h\in B_i}\left[L_h^{'}(w)\right]_{v_j} e_j\right\| \leq R_0,\ \forall\ w,i,j,B_i,v_j$ *and* $e_j(i)=\begin{cases}1 \text{ if } i\in v_j\\ 0 \text{ else}\end{cases}$.

Proof By definition of SAAG method,

$$f\left(w^{k+1}\right) \quad = f\left(w^k-\alpha G_{B_i,v_j}\right),$$

where $G_{B_i,v_j}=\left[\frac{1}{B}\sum_{h\in B_i}\left[L_h^{'}\left(w^k\right)\right]_{v_j} e_j-\frac{1}{l}\sum_{h\in B_i}\left[L_h^{'}\left(\tilde{w}\right)\right]_{v_j} e_j+\left[f^{'}(\tilde{w})\right]_{v_j} e_j\right]$ and

$e_j(i)=\begin{cases}1 \text{ if } i\in v_j\\ 0 \text{ else}\end{cases}$.

Using Lipschitz continuity of gradient, we have,

$$f\left(w^{k+1}\right) \leq f\left(w^k\right)-\alpha G_{B_i,v_j}^T\left[f^{'}(w^k)\right]_{v_j} e_j+\frac{L_{i,j}\alpha^2}{2}\left\|G_{B_i,v_j}\right\|^2$$

Subtracting optimal objective value p^* and taking expectation on both sides over mini-batches B_i, we have,

$$\mathbb{E}_i\left[f\left(w^{k+1}\right)-p^*\right] \leq f\left(w^k\right)-p^*-\alpha\mathbb{E}_i\left[G_{B_i,v_j}\right]^T\left[f^{'}(w^k)\right]_{v_j} e_j+\frac{L_{i,j}\alpha^2}{2}\mathbb{E}_i\left[\left\|G_{B_i,v_j}\right\|^2\right] \quad (18)$$

$$\begin{aligned}\text{Now, } \mathbb{E}_i\left[G_{B_i,v_j}\right] &= \frac{1}{B}\mathbb{E}_i\left[\sum_{h\in B_i}\left[L_h^{'}\left(w^k\right)\right]_{v_j} e_j\right]-\frac{1}{l}\mathbb{E}_i\left[\sum_{h\in B_i}\left[L_h^{'}\left(\tilde{w}\right)\right]_{v_j} e_j\right]+\left[f^{'}(\tilde{w})\right]_{v_j} e_j\\ &= \frac{1}{B}.\frac{1}{m}\sum_{i=1}^{m}\sum_{h\in B_i}\left[L_h^{'}\left(w^k\right)\right]_{v_j} e_j-\frac{1}{l}.\frac{1}{m}\sum_{i=1}^{m}\sum_{h\in B_i}\left[L_h^{'}\left(\tilde{w}\right)\right]_{v_j} e_j+\left[f^{'}(\tilde{w})\right]_{v_j} e_j\\ &= \frac{1}{l}\sum_{h=1}^{l}\left[L_h^{'}\left(w^k\right)\right]_{v_j} e_j-\frac{1}{ml}\sum_{h=1}^{l}\left[L_h^{'}\left(\tilde{w}\right)\right]_{v_j} e_j+\left[f^{'}(\tilde{w})\right]_{v_j} e_j\\ &= \left[f^{'}\left(w^k\right)\right]_{v_j} e_j-\frac{1}{m}\left[f^{'}\left(\tilde{w}\right)\right]_{v_j} e_j+\left[f^{'}(\tilde{w})\right]_{v_j} e_j,\end{aligned}$$

because of $mB=l$, and

$$\begin{aligned}\mathbb{E}_i\left[G_{B_i,v_j}\right]^T\left[f^{'}(w^k)\right]_{v_j} e_j &= \left(\left[f^{'}\left(w^k\right)\right]_{v_j} e_j+\left(1-\frac{1}{m}\right)\left[f^{'}\left(\tilde{w}\right)\right]_{v_j} e_j\right)^T\left[f^{'}(w^k)\right]_{v_j} e_j\\ &\leq \left\|\left[f^{'}\left(w^k\right)\right]_{v_j} e_j\right\|^2+\left(1-\frac{1}{m}\right)\left\|\left[f^{'}\left(w^k\right)\right]_{v_j} e_j\right\|\left\|\left[f^{'}\left(\tilde{w}\right)\right]_{v_j} e_j\right\|,\end{aligned}$$

because of using $\|ab\| \leq \|a\|\|b\|$.

$$\implies \mathbb{E}_i \left[G_{B_i,v_j}\right]^T \left[f'(w^k)\right]_{v_j} e_j = \left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 + \left(1 - \frac{1}{m}\right) R_0^2, \tag{19}$$

[because of taking $\left\| \frac{1}{|B_i|} \sum_{h \in B_i} \left[L_h'(w)\right]_{v_j} e_j \right\| \leq R_0 \quad \forall\, w, i, j, B_i, v_j$].

$$\begin{aligned}
\mathbb{E}_i \left[\left\|G_{B_i,v_j}\right\|^2\right] &= \mathbb{E}_i \left[\left\| \left[\frac{1}{B} \sum_{h \in B_i} \left[L_h'\left(w^k\right)\right]_{v_j} e_j - \frac{1}{l} \sum_{h \in B_i} \left[L_h'\left(\tilde{w}\right)\right]_{v_j} e_j + \left[f'(\tilde{w})\right]_{v_j} e_j \right] \right\|^2 \right] \\
&\leq 2\mathbb{E}_i \left[\left\| \left[\frac{1}{B} \sum_{h \in B_i} \left[L_h'\left(w^k\right)\right]_{v_j} e_j - \frac{1}{l} \sum_{h \in B_i} \left[L_h'\left(\tilde{w}\right)\right]_{v_j} e_j \right] \right\|^2 \right] + 2\mathbb{E}_i \left[\left\| \left[f'(\tilde{w})\right]_{v_j} e_j \right\|^2 \right],
\end{aligned}$$

because of using $\|a+b\|^2 \leq 2\|a\|^2 + \|b\|^2$.

$$\begin{aligned}
\mathbb{E}_i \left[\left\|G_{B_i,v_j}\right\|^2\right] &\leq 4\mathbb{E}_i \left[\left\| \left[\frac{1}{B} \sum_{h \in B_i} \left[L_h'\left(w^k\right)\right]_{v_j} e_j \right] \right\|^2 \right] \\
&\quad + 4\mathbb{E}_i \left[\left\| \left[\frac{B}{l} . \frac{1}{B} \sum_{h \in B_i} \left[L_h'\left(\tilde{w}\right)\right]_{v_j} e_j \right] \right\|^2 \right] + 2\mathbb{E}_i \left[\left\| \left[f'(\tilde{w})\right]_{v_j} e_j \right\|^2 \right], \\
\implies \mathbb{E}_i \left[\left\|G_{B_i,v_j}\right\|^2\right] &= 4R_0^2 + \frac{4B^2}{l^2} R_0^2 + 2R_0^2 = 6R_0^2 + \frac{4}{m^2} R_0^2
\end{aligned} \tag{20}$$

Substituting the values of eqs. (19) and (20) in eq. (18) and taking $\underset{i,j}{max}\, L_{i,j} = L$, we have,

$$\begin{aligned}
\mathbb{E}_i \left[f\left(w^{k+1}\right) - p^*\right] &\leq f\left(w^k\right) - p^* - \alpha \left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 - \alpha \left(1 - \frac{1}{m}\right) R_0^2 + \frac{L\alpha^2}{2} \left(6R_0^2 + \frac{4}{m^2} R_0^2\right) \\
&\leq f\left(w^k\right) - p^* - \alpha \left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 + \alpha R_0^2 \left(L\alpha \left(3 + \frac{2}{m^2}\right) - 1 + \frac{1}{m}\right)
\end{aligned}$$

Taking expectation over the blocks v_j, we have,

$$\mathbb{E}_{i,j} \left[f\left(w^{k+1}\right) - p^*\right] \leq f\left(w^k\right) - p^* - \alpha \mathbb{E}_j \left[\left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 \right] + \alpha R_0^2 \left(L\alpha \left(3 + \frac{2}{m^2}\right) - 1 + \frac{1}{m}\right)$$

Since $\mathbb{E}_j \left[\left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 \right] = \frac{1}{s} \sum_{j=1}^{s} \left\| \left[f'\left(w^k\right)\right]_{v_j} e_j \right\|^2 = \frac{1}{s} \left\|f'(w^k)\right\|^2$ and because of strong convexity $f\left(w^k\right) - p^* \leq \frac{1}{2\mu} \left\|f'(w^k)\right\|^2$.

$$\begin{aligned}
\implies \mathbb{E}_{i,j} \left[f\left(w^{k+1}\right) - p^*\right] &\leq f\left(w^k\right) - p^* - \frac{2\alpha\mu}{s} \left(f\left(w^k\right) - p^*\right) + \alpha R_0^2 \left(L\alpha \left(3 + \frac{2}{m^2}\right) - 1 + \frac{1}{m}\right) \\
&\leq \left(1 - \frac{2\alpha\mu}{s}\right) \left(f\left(w^k\right) - p^*\right) + \alpha R_0^2 \left(L\alpha \left(3 + \frac{2}{m^2}\right) - 1 + \frac{1}{m}\right)
\end{aligned}$$

Applying this inequality recursively and taking expectation, after simplifying we have,

$$\mathbb{E}_{i,j}\left[f\left(w^{k+1}\right)-p^*\right] \leq \left(1-\frac{2\alpha\mu}{s}\right)^{k+1}\left(f\left(w^0\right)-p^*\right)+\alpha R_0^2\left(L\alpha\left(3+\frac{2}{m^2}\right)-1+\frac{1}{m}\right)\sum_{t=0}^{k}\left(1-\frac{2\alpha\mu}{s}\right)^t$$

Since $\sum_{t=0}^{k} r^t \leq \sum_{t=0}^{\infty} r^t = \frac{1}{1-r}, \quad 1>r>0,$

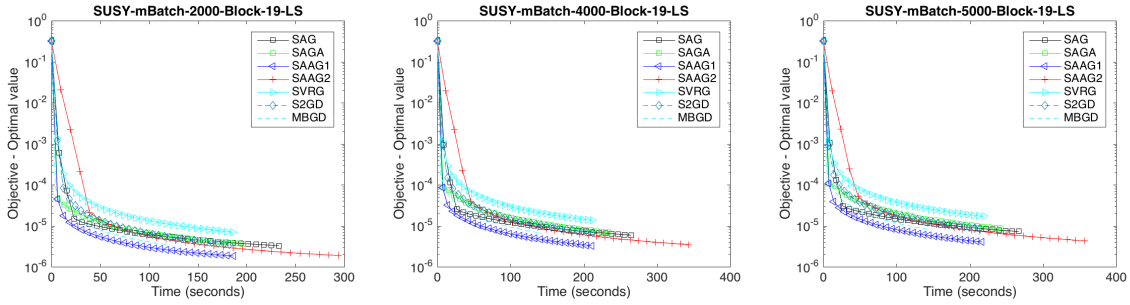
$$\begin{aligned}\implies \mathbb{E}_{i,j}\left[f\left(w^{k+1}\right)-p^*\right] &\leq \left(1-\frac{2\alpha\mu}{s}\right)^{k+1}\left(f\left(w^0\right)-p^*\right)+\alpha R_0^2\left(L\alpha\left(3+\frac{2}{m^2}\right)-1+\frac{1}{m}\right)\cdot\frac{s}{2\alpha\mu}\\ &\leq \left(1-\frac{2\alpha\mu}{s}\right)^{k+1}\left(f\left(w^0\right)-p^*\right)+\frac{sR_0^2}{2\mu}\left(L\alpha\left(3+\frac{2}{m^2}\right)-1+\frac{1}{m}\right)\end{aligned}$$

$$\implies \mathbb{E}_{i,j}\left[f\left(w^{k}\right)-p^*\right] \leq \left(1-\frac{2v\alpha\mu}{p}\right)^{k}\left(f\left(w^0\right)-p^*\right)+\frac{pR_0^2}{2v\mu}\left(L\alpha\left(3+\frac{2B^2}{l^2}\right)-1+\frac{B}{l}\right) \tag{21}$$

[because of $p=vs$, $l=Bm$ and replacing $k+1$ with k]
Thus, algorithm converges linearly. ■

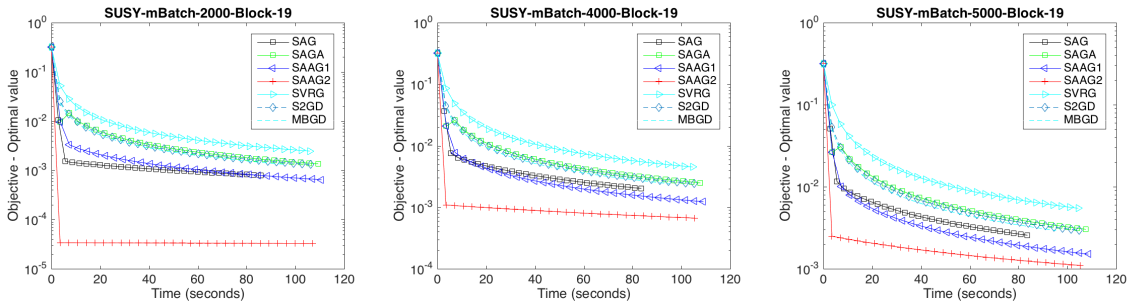
Appendix B. More Experiments

In the following experiments the difference of objective function and optimal value is plotted against the training time of different methods. As it is clear from the figures, SAAGs outperform other methods, similar to results against number of epochs. SAAG-I outperforms other methods using backtracking line search but SAAG-II outperforms other methods using constant step size.



(a) (b) (c)

Figure 1: Sub-optimality against training time with line search method and SUSY dataset



(a) (b) (c)

Figure 2: Sub-optimality against training time with constant step size using Lipschitz constant and SUSY dataset