

# Supplementary for “Cost-efficient SVRG with Arbitrary Sampling”

## A USEFUL DEFINITIONS AND LEMMAS

Table 2 lists the main symbols used in the paper.

Table 2: Summary of main notations.

Symbol	Definition
$N$	Number of worker nodes
$K$	Number of (outer) iterations
$T$	Maximum epoch length
$\mathbf{w}$	Global model (parameters)
$f_i$	Local loss function of node $i$
$\mathbf{g}_i(\mathbf{w})$	Gradient of node $i$ at point $\mathbf{w}$
$\mathbf{g}(\mathbf{w})$	Average gradient of all the nodes $(\sum_i \mathbf{g}_i(\mathbf{w})/N)$
$(\mathbf{x}_{ij}, y_{ij})$	Data sample $j$ of node $i$ and its label
$\mathcal{P}$	Sampling policy of the master node
$p_i$	Sampling probability of node $i$
$\alpha_k$	Step size (learning rate) at iteration $k$
$L$	Expected smoothness
$\xi_{k,t-1}$	Sampled node(s) at iteration $k$ , inner loop iteration $t$

As we recall, the random variables  $\xi_{k,t} \sim \mathcal{P}$  are all pairwise independent and identically distributed (i.i.d) and represent a sampling from the set  $[N]$  according to the distribution  $\mathcal{P}$ , taken at each inner loop iteration of Algorithm 1. For the sake of analysis, set

$$\mathbf{v}_{k,t} = \begin{cases} \tilde{\mathbf{w}}_k, & t = 0, \\ \mathbf{v}_{k,t-1} - \alpha_k \left( \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right), & 1 \leq t \leq T + 1, \end{cases}$$

where  $\tilde{\mathbf{w}}_k$  and  $\tilde{\mathbf{h}}_k = \sum_{i \in [N]} p_i \mathbf{h}_i(\tilde{\mathbf{w}}_k)$  are given by Algorithm 1, and  $\mathbf{h}_i(\mathbf{w}) := \mathbf{g}_i(\mathbf{w})/Np_i$ . Notice that these new variables  $\mathbf{v}_{k,t}$  have the same values as the variables  $\mathbf{w}_{k,t}$  in Algorithm 1. We will use the new variables in the convergence proof of the algorithm, since doing so will simplify some parts of the proof.

**Definition 1** (Sampling (Qian et al., 2019)). A mini-batch sampling  $\xi$  is a random set-valued mapping, with possible values being the subsets of  $[N]$ . Given a sampling  $\xi$ , we let  $p_i := \Pr(i \in \xi)$ . We say  $\xi$  is proper if  $p_i > 0$  for all  $i \in [N]$ .

**Lemma 3.** For any set of vectors  $\{\mathbf{w}_1, \dots, \mathbf{w}_n\} \subseteq \mathbb{R}^d$ , we have that  $\|\sum_{i=1}^n \mathbf{w}_i\|^2 \leq n \sum_{i=1}^n \|\mathbf{w}_i\|^2$ .

**Lemma 4** (Variance decomposition inequality). Let  $w$  be a random variable. Then

$$\mathbb{E} \left[ \|\mathbf{w} - \mathbb{E}[\mathbf{w}]\|^2 \right] = \mathbb{E} \left[ \|\mathbf{w}\|^2 \right] - \|\mathbb{E}[\mathbf{w}]\|^2 \leq \mathbb{E} \left[ \|\mathbf{w}\|^2 \right].$$

**Lemma 5.** Consider the variables  $\mathbf{v}_{k,t}$  above, for  $1 \leq t \leq T$ . We have that

$$\mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right\|^2 \mid \xi_{k,t-1} \right] \leq 4L (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)),$$

where  $L$  is the expected smoothness constant in Lemma 1.

**Lemma 6.** Consider the variables  $\mathbf{v}_{k,t}$  above, for  $1 \leq t \leq T$ , and suppose that  $\mathbf{w}^*$  is the minimizer of  $f$ , i.e., that  $\nabla f(\mathbf{w}^*) = 0$ . We have that

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{v}_{k,t} - \mathbf{w}^*\|^2 \mid \xi_{k,t-1} \right] &\leq \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 \\ &\quad + 4L\alpha_k^2 (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) - 2\alpha_k (f(\mathbf{v}_{k,t-1}) - f(\mathbf{w}^*)). \end{aligned}$$

## B PROOFS

### B.1 LEMMA 2

Since each function  $f_i$  is  $L_i$ -smooth, it follows that

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \mathbf{g}_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L_i} \|\mathbf{g}_i(\mathbf{y}) - \mathbf{g}_i(\mathbf{x})\|^2,$$

for any  $i$ ,  $\mathbf{x}$  and  $\mathbf{y}$ . If we set  $\mathbf{x} = \mathbf{w}^*$ ,  $\mathbf{y} = \mathbf{w}$ , and rearrange this inequality, we end up with

$$\|\mathbf{g}_i(\mathbf{w}) - \mathbf{g}_i(\mathbf{w}^*)\|^2 \leq 2L_i (f_i(\mathbf{w}) - f_i(\mathbf{w}^*) - \langle \mathbf{g}_i(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle).$$

Therefore, for any random variable  $\xi \sim \mathcal{P}$ ,

$$\begin{aligned} \mathbb{E} [\|\mathbf{h}_\xi(\mathbf{w}) - \mathbf{h}_\xi(\mathbf{w}^*)\|^2 \mid \xi] &= \mathbb{E} \left[ \left\| \frac{\mathbf{g}_\xi(\mathbf{w}) - \mathbf{g}_\xi(\mathbf{w}^*)}{Np_\xi} \right\|^2 \mid \xi \right] \\ &\leq \sum_{i \in [N]} \frac{2L_i}{N^2 p_i} (f_i(\mathbf{w}) - f_i(\mathbf{w}^*) - \langle \mathbf{g}_i(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle) \\ &\leq 2 \max_{1 \leq i \leq N} \left\{ \frac{L_i}{N^2 p_i} \right\} \sum_{i=1}^n f_i(\mathbf{w}) - f_i(\mathbf{w}^*) - \langle \mathbf{g}_i(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle \\ &= 2 \max_{i \in [N]} \left\{ \frac{L_i}{N p_i} \right\} (f(\mathbf{w}) - f(\mathbf{w}^*)), \end{aligned}$$

leading to  $L \leq \max_{i \in [N]} \{L_i/Np_i\}$ , where in the last step we used that  $f = \frac{1}{N} \sum_{i \in [N]} f_i$  and  $\mathbf{g}(\mathbf{w}^*) = \mathbf{0}$ .

### B.2 LEMMA 3

Write  $\|\sum_{i=1}^n w_i\|^2 = n^2 \|\frac{1}{n} \sum_{i=1}^n w_i\|^2$  and use Jensen's inequality with the convexity of the norm squared to conclude the lemma.

### B.3 LEMMA 5

Lemma 3 yields

$$\begin{aligned} \mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right\|^2 \mid \xi_{k,t-1} \right] &\leq 2\mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\mathbf{w}^*) \right\|^2 \mid \xi_{k,t-1} \right] \\ &\quad + 2\mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{w}^*) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right\|^2 \mid \xi_{k,t-1} \right] \\ &\stackrel{(a)}{\leq} 2\mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\mathbf{w}^*) \right\|^2 \mid \xi_{k,t-1} \right] \\ &\quad + 2\mathbb{E} \left[ \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{w}^*) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) \right\|^2 \mid \xi_{k,t-1} \right] \\ &\stackrel{(b)}{\leq} 4L (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)), \end{aligned}$$

where (a) is due to the variance decomposition inequality (Lemma 4) and the definition of  $\mathbf{h}(\tilde{\mathbf{w}})$ , and (b) is due to Lemma 1. For (a), recall that  $\tilde{\mathbf{h}}_k = \sum_i p_i \mathbf{h}_i(\tilde{\mathbf{w}}_k) = \mathbb{E}[\mathbf{h}_\xi(\tilde{\mathbf{w}}_k) \mid \xi]$ , for any random variable  $\xi \sim \mathcal{P}$ , which applies to  $\xi_{k,t-1}$ .

## B.4 LEMMA 6

The inner loop of Algorithm 1 yields

$$\begin{aligned}
\mathbb{E} [\|\mathbf{v}_{k,t} - \mathbf{w}^*\|^2 \mid \xi_{k,t-1}] &= \mathbb{E} \left[ \|\mathbf{v}_{k,t-1} - \alpha_k \left( \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right) - \mathbf{w}^*\|^2 \mid \xi_{k,t-1} \right] \\
&= \mathbb{E} \left[ \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 + \alpha_k^2 \left\| \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right\|^2 \right. \\
&\quad \left. - 2\alpha_k (\mathbf{v}_{k,t-1} - \mathbf{w}^*)^\top \left( \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right) \mid \xi_{k,t-1} \right] \\
&\stackrel{(a)}{\leq} \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 + 4L\alpha_k^2 (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) \\
&\quad - 2\alpha_k (\mathbf{v}_{k,t-1} - \mathbf{w}^*)^\top \mathbb{E} \left[ \mathbf{h}_\xi(\mathbf{v}_{k,t-1}) - \mathbf{h}_\xi(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \mid \xi_{k,t-1} \right] \\
&\stackrel{(b)}{=} \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 + 4L\alpha_k^2 (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) \\
&\quad - 2\alpha_k (\mathbf{v}_{k,t-1} - \mathbf{w}^*)^\top \mathbf{g}(\mathbf{v}_{k,t-1}) \\
&\stackrel{(c)}{\leq} \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 + 4L\alpha_k^2 (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) \\
&\quad - 2\alpha_k (f(\mathbf{v}_{k,t-1}) - f(\mathbf{w}^*)) ,
\end{aligned}$$

where (a) is due to Lemma 5, and the independence of  $\mathbf{w}_{k,t-1}$  from  $\xi_{k,t-1}$ , and (c) is due to the convexity of  $f$ . Equality (b) follows from

$$\mathbb{E} \left[ \mathbf{h}_{\xi_{k,t-1}}(\mathbf{v}_{k,t-1}) - \mathbf{h}_{\xi_{k,t-1}}(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \mid \xi_{k,t-1} \right] = \sum_{i=1}^N p_i \mathbf{h}_i(\mathbf{v}_{k,t-1}) = \mathbf{g}(\mathbf{w}_{k,t-1}) ,$$

where we have used the fact that  $\tilde{\mathbf{h}}_k = \mathbb{E}[\mathbf{h}_\xi(\tilde{\mathbf{w}}_k) \mid \xi]$ , where  $\xi \sim \mathcal{P}$  is a sampling of the set  $[N]$  according to the distribution  $\mathcal{P}$ .

## B.5 PROPOSITION 1

Using proof by contradiction, it is relatively easy to establish that  $p_i^*/L_i = a$  for all  $i \in [N]$  and some constant  $a > 0$ . Constraint  $\sum_i p_i^* = 1$  yields  $a = 1/\sum_i L_i = 1/NL$ , completing the proof.

## B.6 PROPOSITION 2

The proof sketch is similar to the convergence of the original SVRG algorithm (Johnson and Zhang, 2013). Recall that  $\zeta_k$  is uniformly sampled from  $\{1, \dots, T\}$ , and  $\tilde{\mathbf{w}}_{k+1} = \mathbf{w}_{k,\zeta_k} = \mathbf{v}_{k,\zeta_k}$ . That is,

$$\mathbb{E} [f(\tilde{\mathbf{w}}_{k+1})] = \frac{1}{T} \sum_{t=1}^T f(\mathbf{v}_{k,t}) . \tag{A.1}$$

Use the inequality for  $\mathbb{E} [\|\mathbf{v}_{k,t} - \mathbf{w}^*\|^2 \mid \xi_{k,t-1}]$  in Lemma 6, together with the tower law  $\mathbb{E} [X] = \mathbb{E} [\mathbb{E} [X \mid Y]]$ , and sum them over the  $T$  iterations of one epoch:

$$\begin{aligned}
\sum_{t=1}^{T+1} \mathbb{E} [\|\mathbf{v}_{k,t} - \mathbf{w}^*\|^2] &\leq \sum_{t=1}^{T+1} \mathbb{E} \left[ \|\mathbf{v}_{k,t-1} - \mathbf{w}^*\|^2 \right] \\
&\quad + \mathbb{E} \left[ \sum_{t=1}^{T+1} 4L\alpha_k^2 (f(\mathbf{v}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) - 2\alpha_k (f(\mathbf{v}_{k,t-1}) - f(\mathbf{w}^*)) \right] \\
&\stackrel{(A.1)}{=} \sum_{t=0}^T \mathbb{E} \left[ \|\mathbf{v}_{k,t} - \mathbf{w}^*\|^2 \right] \\
&\quad + \mathbb{E} [4LT\alpha_k^2 (f(\tilde{\mathbf{w}}_{k+1}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) - 2T\alpha_k (f(\tilde{\mathbf{w}}_{k+1}) - f(\mathbf{w}^*)) \\
&\quad + 4L\alpha_k^2 (f(\mathbf{w}_{k,0}) + f(\tilde{\mathbf{w}}_k) - 2f(\mathbf{w}^*)) - 2\alpha_k (f(\mathbf{w}_{k,0}) - f(\mathbf{w}^*))] .
\end{aligned}$$

Cancelling similar terms from both sides, noting that  $\mathbf{v}_{k,0} = \mathbf{w}_{k,0} = \tilde{\mathbf{w}}_k$ , and setting  $\Delta_k := \mathbb{E}[f(\tilde{\mathbf{w}}_k)] - f(\mathbf{w}^*)$ , yields

$$0 \leq \mathbb{E}[\|\mathbf{v}_{k,T} - \mathbf{w}^*\|^2] \leq \mathbb{E}[\|\tilde{\mathbf{w}}_k - \mathbf{w}^*\|^2] + 4LT\alpha_k^2(\Delta_{k+1} + \Delta_k) - 2T\alpha_k\Delta_{k+1} + 8L\alpha_k^2\Delta_k - 2\alpha_k\Delta_k \\ \stackrel{(a)}{\leq} \frac{2\Delta_k}{\mu} + 4LT\alpha_k^2(\Delta_{k+1} + \Delta_k) - 2T\alpha_k\Delta_{k+1} + (8L\alpha_k^2 - 2\alpha_k)\Delta_k.$$

where (a) follows from strong convexity of  $f$ . After rearranging, we end up with

$$\Delta_{k+1} \leq \left( \frac{\frac{2}{\mu} + 4LT\alpha_k^2 + 8L\alpha_k^2 - 2\alpha_k}{2T\alpha_k - 4LT\alpha_k^2} \right) \Delta_k = \left( \frac{\frac{1}{\mu T\alpha_k} + 2L\alpha_k + \frac{4L\alpha_k - 1}{T}}{1 - 2L\alpha_k} \right) \Delta_k, \quad (\text{A.2})$$

assuming  $\alpha_k \in (0, 1/2L)$  to ensure a positive denominator. Forcing the contraction factor to be in interval  $(0, 1)$  yields  $T > 0$  and  $\alpha < 1/2L$  to meet the lower limit and

$$T > \frac{1 + \mu\alpha_k(4L\alpha_k - 1)}{\mu\alpha_k(1 - 4L\alpha_k)}, \quad (\text{A.3})$$

to meet the upper limit. Moreover,  $T > 0$  in (A.3) implies  $\alpha_k < 1/4L$ , and therefore  $4L\alpha_k - 1 \leq 0$ , which allows us to use the bounds

$$T > \frac{1}{\mu\alpha_k(1 - 4L\alpha_k)}, \quad \text{and} \quad (\text{A.4})$$

$$\Delta_{k+1} \leq \left( \frac{\frac{1}{\mu T\alpha_k} + 2L\alpha_k}{1 - 2L\alpha_k} \right) \Delta_k. \quad (\text{A.5})$$

These bounds are comparable to the ones in the original SVRG paper (Johnson and Zhang, 2013), with the difference being that here  $L$  represents the expected, rather than the maximum, smoothness constant.

We can deduce the following corollary from Proposition 2:

**Corollary 1.** *To ensure a contraction factor of at most  $\sigma_{\max} < 1$ , the step size and the maximum epoch length should satisfy*

$$\alpha_k \leq \frac{\sigma_{\max}}{2L(1 + \sigma_{\max})}, \quad \text{and} \quad T \geq \frac{1}{\min_{k \in \mathbb{N}} \mu\alpha_k(\sigma_{\max} - 2L\alpha_k\sigma_{\max} - 2L\alpha_k)}. \quad (\text{A.6})$$

## B.7 COROLLARY 1

From Proposition 2, we set

$$\frac{\frac{1}{\mu T\alpha_k} + 2L\alpha_k}{1 - 2L\alpha_k} \leq \sigma_{\max}$$

and obtain

$$T \geq \frac{1}{\mu\alpha_k(\sigma_{\max} - 2L\alpha_k\sigma_{\max} - 2L\alpha_k)}.$$

Equation (A.6) immediately follows. Notice that positivity of the denominator implies

$$\alpha_k < \frac{\sigma_{\max}}{2L(1 + \sigma_{\max})},$$

for some positive  $\sigma_{\max} < 1$ . This condition automatically implies  $\alpha_k < 1/4L$  (from  $\sigma/(1 + \sigma)$  being monotonically increasing for  $0 < \sigma \leq 1$ ), under which Proposition 2 holds.

## B.8 OPTIMIZATION PROBLEM (3)

First, we can rewrite (3) as

$$\underset{p_1, p_2, \dots, p_N}{\text{minimize}} \quad T \sum_{i \in [N]} c_i p_i, \quad (\text{A.7a})$$

$$\text{subject to} \quad \sum_{i \in [N]} p_i = 1, \quad (\text{A.7b})$$

$$p_i \geq \frac{4L_i}{N} \max \left\{ \alpha, \frac{1}{4L_{\max}} \right\}, \quad \forall i \in [N]. \quad (\text{A.7c})$$

To this end, notice that Constraint (3c) is equivalent to  $p_i \geq 4\alpha L_i/N$  for all  $i \in [N]$ . Moreover,

$$\frac{x_1}{y_1} \geq \frac{x_2}{y_2} \iff \frac{x_1}{x_1 + y_1} \geq \frac{x_2}{x_2 + y_2} \quad (\text{A.8})$$

for all  $x_1, x_2 \geq 0$  and  $y_1, y_2 > 0$  implies that constraint (3d) is equivalent to  $L_{\max} \geq \max_i \{L_i/N p_i\}$  and therefore to  $p_i \geq L_i/N L_{\max}$  for all  $i \in [N]$ , concluding the transformation.

Now, let  $\beta_i := 4L_i \max\{\alpha, 1/4L_{\max}\}/N$  and  $\bar{L} := \sum_{i=1}^N L_i/N$ . Feasibility of (A.7) is ensured if and only if  $\sum_{i \in [N]} \beta_i \leq 1$ , since the constraint (A.7c) is equivalent to  $\beta_i \leq p_i$ , and we should have  $\sum_i p_i = 1$ . Consequently,

$$\max \left\{ 4\alpha \bar{L}, \frac{\bar{L}}{L_{\max}} \right\} = \sum_{i \in [N]} \beta_i \leq 1. \quad (\text{A.9})$$

Recall that  $\bar{L} \leq L_{\max}$ , so that  $\bar{L}/L_{\max} \leq 1$ . Since  $\alpha \leq 1/4\bar{L}$ , it follows that (A.9) is satisfied, meaning that the problem is always feasible. Let  $(p_1, \dots, p_N)$  be the probability distribution given in (4), and notice that it satisfies  $p_i = \beta_i$  for every  $i \neq j$ , and  $p_j = 1 - \sum_{i \neq j} \beta_i \geq \beta_j$ . That is, it satisfies the constraint (A.7c), meaning that it is a feasible solution. We will now show that it is also an optimal solution to the problem.

Any other feasible probability distribution  $(p'_1, \dots, p'_N)$  can be obtained as  $p'_i = p_i + e_i$ , where  $e_i \geq 0$  for  $i \neq j$ , and  $e_j = -\sum_{i \neq j} e_i \leq \beta_j - p_j$ , due to the constraints  $p'_i \geq \beta_i$ , and  $\sum_i p'_i = 1 = \sum_i p_i$ . However, the cost associated with the distribution  $(p'_1, \dots, p'_N)$  is

$$\sum_i p'_i c_i = \sum_i p_i c_i + e_j c_j + \sum_{i \neq j} e_i c_i \geq \sum_i p_i c_i + e_j c_j - e_j c_{\min} = \sum_i p_i c_i,$$

since  $c_i \geq c_{\min}$ , and  $c_j = c_{\min}$ . That is, the cost is not lower than the one associated with the distribution  $(p_1, \dots, p_N)$ . This shows that the probability distribution given by (4) is an optimal solution.

## B.9 REMARK 1

It is straightforward to show that all main lemmas and proof steps would remain the same for the mini-batch SVRG-AS+ algorithm.

## C ADDITIONAL DISCUSSIONS AND RESULTS

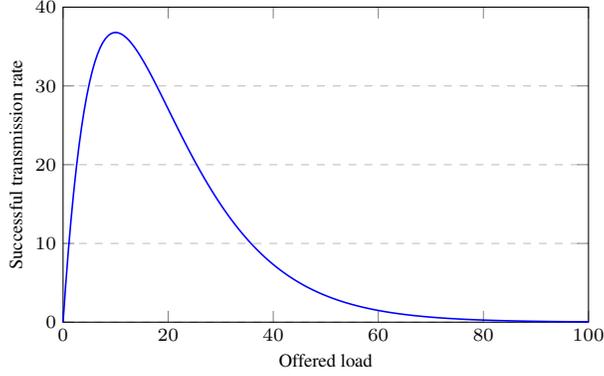
### C.1 RATE MODEL FOR SHARED WIRELESS CHANNEL

Define the so-called offered load  $\ell > 0$  as the average number of packets (gradients) that the workers inject to the wireless channel (Bertsekas et al., 2004). The successful transmission rate model follows  $r = r_0 \ell \exp\{-\ell/r_1\}$  for some positive constants  $r_0$  and  $r_1$  describing the wireless channel and communication protocol (Bertsekas et al., 2004). Figure A.1 shows the success rate for  $r_0 = r_1 = 10$ . This rate model implies that the channel can handle the incoming traffic when almost idle. After a certain point, it becomes ‘‘congested’’ and increasing the number of transmitted packets (offered loads) would only add to the congestion and packet drops, leading to a lower success rate.

### C.2 MINIMUM LATENCY SVRG-AS+ FOR SECTION 4.2

Define the so-called offered load  $\ell > 0$  as the average number of packets (gradients) that the workers inject to the wireless channel (Bertsekas et al., 2004). The successful transmission rate model follows  $r = r_0 \ell \exp\{-\ell/r_1\}$  for some positive constants  $r_0$  and  $r_1$  describing the wireless channel and communication protocol (Bertsekas et al., 2004). By the definition of  $p_i$ ,  $\ell = \sum_i p_i$  for mini-batch SVRG-AS+. With the natural assumption of fixed packet size for every gradient vector,<sup>1</sup> latency is inversely proportional to the success rate. Assuming a fixed  $\mathbf{p}$  for all iterations, the expected cost of

<sup>1</sup>Assuming 32 bits per coordinate,  $\mathbf{g}$  has  $32d$  bits. Compression algorithms, reviewed in Section 2, reduce this number.

Figure A.1: Illustration of rate model for  $r_0 = r_1 = 10$ .

$K$  iterations of SVRG-AS+ is then  $Kr^{-1}$ , after which  $\Delta_K \leq \sigma^K \Delta_0$ . Ensuring  $\Delta_k \leq \epsilon_1$  for some constant  $\epsilon_1 > 0$  implies  $\sigma \leq (\epsilon_1/\Delta_0)^{1/K}$ , where we have assumed  $\alpha_1 = \alpha_2 = \dots = \alpha_K = \alpha$ . Now, using the definition of  $\sigma$  in Proposition 2, definition of  $L$  in Lemma (2), and equivalence in (A.8), we can show that  $\Delta_K \leq \epsilon_1$  is equivalent to

$$p_i \geq \frac{2\alpha L_i}{N\epsilon_2} \text{ for every } i \in [N], \text{ where } \epsilon_2 = \left( \frac{(\epsilon_1/\Delta_0)^{1/K}}{1 + (\epsilon_1/\Delta_0)^{1/K}} \right) \left( 1 + \frac{1}{\mu T \alpha} \right) - \frac{1}{\mu T \alpha}.$$

Now, we can formulate our minimum latency SVRG-AS+ problem as

$$\begin{aligned} & \underset{p_1, p_2, \dots, p_N}{\text{minimize}} && KT \frac{\exp \left\{ \frac{\sum_{i \in [N]} p_i}{r_1} \right\}}{r_0 \sum_{i \in [N]} p_i}, \\ & \text{subject to} && p_i \in \left[ \frac{2L_i}{N} \max \left\{ \frac{\alpha}{\epsilon_2}, \frac{1}{2L_{\max}} \right\}, 1 \right], \forall i \in [N]. \end{aligned}$$

### C.3 SMOOTHNESS AND STRONG CONVEXITY PARAMETERS FOR LOGISTIC REGRESSION

Consider function

$$f_i(\mathbf{w}) = \frac{1}{M_i} \sum_{j \in [M_i]} \ln \left( 1 + e^{-\mathbf{w}^T \mathbf{x}_{ij} y_{ij}} \right).$$

Define  $\mathbf{z}_{ij} := \mathbf{x}_{ij} y_{ij}$ . Let  $\text{eig}_{\max}(\mathbf{X})$  and  $\text{eig}_{\min}(\mathbf{X})$  denote the largest and smallest eigenvalues of matrix  $\mathbf{X}$ . Now, we characterize the geometry of our problem using  $L_i \geq \text{eig}_{\max}(\nabla^2 f_i(\mathbf{w}))$  and  $\mu \leq \text{eig}_{\min}(\nabla^2 f(\mathbf{w}))$  inequalities. We have,

- Smoothness parameter  $L_i$ :

$$\begin{aligned} \text{eig}_{\max}(\nabla^2 f_i(\mathbf{w})) &\leq \frac{1}{M_i} \sum_{j \in [M_i]} \text{eig}_{\max}(\nabla^2 f_i(\mathbf{w})) \\ &\leq \frac{1}{4M_i} \sum_{j \in [M_i]} \|\mathbf{z}_{ij}\|_2^2 := L_i. \end{aligned}$$

- Strong convexity parameter  $\mu$ :

$$\text{eig}_{\min}(\nabla^2 f(\mathbf{w})) \geq \frac{1}{M_i} \sum_{j \in [M_i]} \text{eig}_{\min}(\nabla^2 f_i(\mathbf{w})) := \mu.$$

### C.4 MINI-BATCH SVRG-AS+ ALGORITHM

Algorithm 2 shows SVRG-AS+ with mini-batch updates.

**Algorithm 2** Mini-batch SVRG-AS+

---

```

1: Inputs: Maximum epoch length  $T$ , number of epochs  $K$ ,  $N$ , step size sequence  $(\alpha_k)_k$ , and
   probabilities  $p_1, \dots, p_N$ .
2: for  $k = 1, 2, \dots, K - 1$  do
3:    $\tilde{\mathbf{h}}_k \leftarrow \sum_{i \in [N]} p_i \mathbf{h}_i(\tilde{\mathbf{w}}_k)$ 
4:    $\mathbf{w}_{k,0} \leftarrow \tilde{\mathbf{w}}_k$ 
5:   Sample  $\zeta := \zeta_k$  uniformly from  $\{1, 2, \dots, T\}$ 
6:   for  $t = 1, 2, \dots, \zeta$  do
7:     Every worker  $i$  with probability  $p_i$ , independent of other workers, computes  $\mathbf{g}_i(\mathbf{w})$  and
        $\mathbf{h}_i(\mathbf{w}) := \mathbf{g}_i(\mathbf{w})/Np_i$ , and sends it to the master node, for both  $\mathbf{w} = \mathbf{w}_{k,t-1}$  and  $\mathbf{w} = \tilde{\mathbf{w}}_k$ .

8:     Define  $\xi := \{i \mid \mathbf{h}_i(\mathbf{w}_{k,t-1}) \text{ is sampled}\}$ .
9:     Compute  $\mathbf{w}_{k,t} \leftarrow \mathbf{w}_{k,t-1} - \alpha_k \sum_{i=1}^N \mathbb{1}_{i \in \xi} \left( \mathbf{h}_i(\mathbf{w}_{k,t-1}) - \mathbf{h}_i(\tilde{\mathbf{w}}_k) + \tilde{\mathbf{h}}_k \right)$ 
10:    Broadcast  $\mathbf{w}_{k,t}$ 
11:  end for
12:   $\tilde{\mathbf{w}}_{k+1} \leftarrow \mathbf{w}_{k,\zeta}$ 
13: end for
14: Return:  $\tilde{\mathbf{w}}_K$ 

```

---

Table 3: F1-score of the MNIST test dataset and cost of training the model using a distributed algorithm. Results are for the two stragglers cost model with  $(\alpha_k = 0.2)_k$ ,  $T = 15$ , and 20 iterations.

$N$	SVRG		SVRG-AS+	
	F1-score	cost (x1000)	F1-score	cost (x1000)
10	0.864	103.5	0.859	43.5
50	0.860	269.2	0.861	25.9
100	0.841	476.3	0.839	20.0

### C.5 EXTRA EXPERIMENTS ON MNIST DATASET

To evaluate the performance of our final solution on all digits, we have reported in Table 3 the F1-score, averaged over all classes. In all cases, the convergence of SVRG-AS+ was as fast as that of SVRG. We should highlight that we did not try to optimize hyper-parameters to achieve a better F1-score in our experiments.

To further improve the robustness to the stragglers, one may exploit the inherent redundancy of big datasets (Ghadikolaei et al., 2019). In particular, increasing the number of nodes may raise the correlation among the local private datasets, leading to lower dependency on the data stored in a single node. In the example of handwritten digit classification, we can expect less sensitivity of the training task to the lack of information from one straggler node if the information in its local dataset (relevant to the training task) can be captured by others. In the case of the MNIST dataset, when we increase the number of nodes to more than 10, samples of every class will be found in more than one node. Consequently, we can even further reduce  $p_i$  for straggler nodes and even ignore their inputs for the outer loop, knowing that their contributions to the training task can be replaced by others with low cost. Using a grid search, we can reduce the cost when  $N = 100$  (last row of Table 3) from 476,330 to 20015, leading to 96% less costs to achieve the same solution. We leave the formal design of this extension of cost-efficient SVRG-AS+ as a future work.

### C.6 COMMENTS ON CIFAR10 EXPERIMENTS

Our optimization problems and convergence bounds are obtained for strongly convex surface, such as (3) and (5). To run non-convex experiments on CIFAR10, we had to hand-tune the constraints, meaning that the step-size  $\alpha$  and maximum epoch length  $T$  were manually tuned, as is common for most gradient descent methods.

### C.7 COST REDUCTION DUE TO EARLY CUT OF INNER-LOOP

In theory, cutting the inner loop early in this way cannot account for more than half of the reduction in cost (and in practice much less than that). Indeed, every inner loop requires  $N$  communications to initialize the parameters, and our version performs on average  $T/2$  inner loop iterations instead of  $T$ . Since each inner loop iteration requires one communication (considering only the uplink), the average communications required for each epoch is  $N + T/2$  for our version, instead of the original  $N + T$ . As can be seen,  $N + T/2$  is more than half of  $N + T$ . Extra saving in the experiments comes from a network-aware optimal sampling policy.