

Communication-efficient Variance-reduced Stochastic Gradient Descent

Hossein S. Ghadikolaei and Sindri Magnússon

*School of EECS, KTH Royal Institute of Technology, Stockholm,
Sweden (emails: {hshokri, sindrim}@kth.se).*

Abstract: We consider the problem of communication efficient distributed optimization where multiple nodes exchange important algorithm information in every iteration to solve large problems. In particular, we focus on the stochastic variance-reduced gradient and propose a novel approach to make it communication-efficient. That is, we compress the communicated information to a few bits while preserving the linear convergence rate of the original uncompressed algorithm. Comprehensive theoretical and numerical analyses on real datasets reveal that our algorithm can significantly reduce the communication complexity, by as much as 95%, with almost no noticeable penalty. Moreover, it is much more robust to quantization (in terms of maintaining the true minimizer and the convergence rate) than the state-of-the-art algorithms for solving distributed optimization problems. Our results have important implications for using machine learning over internet-of-things and mobile networks.

Keywords: Distributed optimization, communication efficiency, SVRG, quantization.

1. INTRODUCTION

The recent success of artificial intelligence (AI) and large-scale machine learning is mainly due to the availability of big datasets and large platforms that can provide vast amounts of computational and communication resources for training. However, there are increasing demands to extend large-scale machine learning to a general networked AI where communication and energy resources may be rare and expensive commodities. Use cases include machine learning over internet-of-things, edge computing, and vehicular networks (Li et al., 2018; Jiang et al., 2019).

We address the problem of computational and communication-efficient distributed optimization. Consider optimizing a finite sum of differentiable functions $\{f_i : \mathbb{R}^d \mapsto \mathbb{R}\}_{i \in [N]}$ with corresponding gradients $\{g_i : \mathbb{R}^d \mapsto \mathbb{R}^d\}_{i \in [N]}$. Iteration k of a common solution algorithm involves finding the gradients $g_i(\mathbf{w}_k)$ at parameter \mathbf{w}_k , sending them back to a central controller (master node), updating the parameters to \mathbf{w}_{k+1} at the master node, and broadcasting it to start the next iteration. In a distributed optimization setting, one can divide data samples among N computational (worker) nodes for parallel computing. This approach improves privacy and obtains a good tradeoff between computational complexity and convergence rate, thanks to new algorithms like stochastic variance reduced gradient (SVRG) and its variants (Johnson and Zhang, 2013; Konečný et al., 2016). However, it raises a serious challenge of communication complexity, defined as the level of information exchange among the workers and the master node to ensure the convergence of the iterative solution algorithm (like SVRG) (Bottou et al., 2018). It has been shown in many applications with geographically-separated workers that the communication complexity dominates the runtime of a distributed optimization (Konečný et al., 2016).

This becomes particularly important when we implement machine learning and distributed optimization algorithms on bandwidth and battery-limited wireless networks. In that case, the latency (in remote industrial operation) or energy consumption (in low power internet-of-things) may render the ultimate solution and consequently the distributed algorithm useless.

There is a recent wave of attempts to address the communication complexity of distributed optimization. A prominent approach is to use a lossy-compression (realized through quantization with few bits) of the parameter and gradient vectors in every iteration to save the communication resources. This approach has been applied to both deterministic (Jordan et al., 2018; Magnússon et al., 2018, 2019) and stochastic (Bernstein et al., 2018; Wen et al., 2017; Sa et al., 2018) algorithms. However, the conventional wisdom, validated by empirical observations, is that there exists a precision-accuracy tradeoff: the fewer the number of bits per iteration the lower the accuracy of the final solution. Some recent studies, however, challenged that wisdom in the deterministic (Magnússon et al., 2019) and stochastic (Sa et al., 2018) settings, and showed that proper quantization approaches can maintain the convergence (both to the true minimizer and the convergence rate). Using similar ideas to (Magnússon et al., 2019), but in the stochastic gradient setting, we propose a novel adaptive quantization approach that theoretically ensures a linear convergence rate of a pseudo-contractive algorithm. The closest work to us is (Sa et al., 2018) where the authors proposed an adaptive quantization scheme to iterate based on low-precision \mathbf{w}_k while maintaining high-accuracy training performance. Our novel algorithm includes low-precision gradients (in uplink transmission), in addition to the low-precision parameters (in downlink transmission), into the iterations of SVRG. This is impor-

tant in many real applications as the uplink channel may have a much lower speed than the downlink channel (Furht and Ahson, 2016) in which sending low-precision gradients leads to a substantial performance gain. Handling this modification involves a non-trivial modification of the original SVRG algorithm using a memory unit, as described in Section 3.

In this paper, we focus on SVRG for the class of strongly convex and smooth functions (which appear in many applications including channel estimation, localization, resource allocation, and network optimization (Boyd and Vandenberghe, 2004)). Our proposed algorithm can recover the minimizer while maintaining the linear convergence rate with only a fixed number of bits for information exchange among the workers and master node. This is possible due to the fact that, with minor modification of the SVRG algorithm, the norm of the gradient and therefore the difference between $\mathbf{w}_k - \mathbf{w}_{k-1}$ gets smaller for larger k . Consequently, by properly tuning hyper-parameters of the quantization at every step, we can increase the quantization resolution without adding any additional bits. Comprehensive theoretical and numerical analyses on real datasets reveal that our algorithm can significantly reduce the communication complexity, by as much as 95%, with almost no noticeable penalty (neither in training nor in test). Our results also indicate that our quantized SVRG approach is much more robust to quantization (in terms of maintaining the true minimizer and the convergence rate) than other state-of-the-art algorithms for solving distributed optimization problems.

The rest of the paper is organized as follows. Section 2 presents the problem setting and basic assumptions. Section 3 provides our main theoretical results. We apply our algorithms on real datasets in Section 4, and then conclude the paper in Section 5. Due to a lack of space, we have moved all the proofs in the extended version of this paper (Ghadikolaei and Magnússon, 2019).

Notation: Normal font w or W , bold font small-case \mathbf{w} , bold-font capital letter \mathbf{W} , and calligraphic font \mathcal{W} denote scalar, vector, matrix, and set, respectively. We let $[N] = \{1, 2, \dots, N\}$ for any integer N . We denote by $\|\cdot\|$ the l_2 norm, by $\text{Conv}(\mathcal{W})$ the convex hull of set \mathcal{W} , by $\text{Proj}(\mathbf{w}, \mathcal{W})$ the projection of vector \mathbf{w} onto set \mathcal{W} , by $[\mathbf{w}]_i$ the i -th entry of vector \mathbf{w} , and by \mathbf{w}^T the transpose of \mathbf{w} .

2. SYSTEM MODEL

2.1 Problem Statement and Algorithm Model

Consider a network of N computational nodes (workers) that cooperatively solve a distributed computational problem involving an objective function $f(\mathbf{w})$. Let tuple $(\mathbf{x}_{ij}, \mathbf{y}_{ij})$ denote data sample j of worker $i \in [N]$, and \mathbf{w} denote the shared model parameter with dimension d . Our optimization problem is

$$\mathbf{w}^* \in \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) := \frac{1}{N} \sum_{i \in [N]} \underbrace{\frac{1}{N_i} \sum_{j \in [N_i]} f(\mathbf{w}; \mathbf{x}_{ij}, \mathbf{y}_{ij})}_{f_i(\mathbf{w})}, \quad (1)$$

where \mathbf{g}_i and \mathbf{g} denote the gradient of f_i and f , respectively, and N_i is the number of samples at node $i \in [N]$.

SVRG is among the state-of-the-art algorithms to balance computational loads of obtaining the full gradients $\{\mathbf{g}_i\}_i$ and the convergence rate of the distributed algorithm (Johnson and Zhang, 2013). The inner loop of SVRG (also called epoch) consists of T iterations, wherein the parameters will be sequentially updated using the gradient of f_ξ for a randomly chosen index ξ . In particular, in the beginning, the master node broadcasts parameter vector $\tilde{\mathbf{w}}_k$. It then broadcasts $\mathbf{w}_{k,t-1}$, realizes random variable $\xi \in [N]$, and receives $\mathbf{g}_\xi(\mathbf{w}_{k,t-1})$, in every inner-loop iteration t . At the end of the inner loop, the master node updates $\tilde{\mathbf{g}}_k$ and $\tilde{\mathbf{w}}_k$ for the next epoch, see Algorithm 1.

For the sake of mathematical analysis, we limit the class of objective functions to strongly convex and smooth function, though our approach is applicable to invex (Karimi et al., 2016) and multi-convex (Xu and Yin, 2013) structures (like a deep neural network training optimization problem) after some minor modifications.

Assumption 1. We assume that $f(\mathbf{w})$ is μ -strongly convex and its gradient terms $\{\mathbf{g}_i\}_i$ are L -Lipschitz $\forall i \in [N]$, i.e.,

$$(\mathbf{w} - \mathbf{y})^T (\mathbf{g}(\mathbf{w}) - \mathbf{g}(\mathbf{y})) \geq \mu \|\mathbf{w} - \mathbf{y}\|^2, \quad (2a)$$

$$\|\mathbf{g}_i(\mathbf{w}) - \mathbf{g}_i(\mathbf{y})\| \leq L \|\mathbf{w} - \mathbf{y}\|. \quad (2b)$$

Next, we define the quantization space and quantization operator.

2.2 Quantization

As stated, the SVRG iterations require an exchange of the current parameters and gradients among the workers and the master node. In practice (e.g., patrolling and surveillance, and smart manufacturing), the nodes must first represent real-valued vectors (\mathbf{w}_k and $\{\mathbf{g}_i\}_{i \in [N]}$) by a finite number of bits through the so-called quantization/compression operation. The current distributed optimization systems usually use single/double-precision floating-point (32/64 bits mapped according to the IEEE 754 format) to show every entry of those vectors and assume no gap between the quantized values and the original value (infinite number of bits assumption). In this work, we show that a very simple alternative mapping with much fewer bits can be used to maintain the convergence properties of unquantized SVRG.

Definition 2. (Quantization Space and Quantizer). Let $\mathcal{R} := \mathcal{R}(\mathbf{c}, \mathbf{r}, \{b_i\}_{i \in [d]}) \subseteq \mathbb{R}^d$ for some positive vector \mathbf{r} be a d -dimensional lattice of 2^b points (quantization space) centered at $\mathbf{c} \in \mathbb{R}^d$ with 2^{b_i} points in coordinate $i \in [d]$ (where $b_i \in \mathbb{N}^+$) such that $\sum_{i \in [d]} b_i = b$. For any vector \mathbf{w} , we define quantizer $q: \mathbf{w} \mapsto \text{Proj}(\mathbf{w}, \mathcal{R})$. For notation simplicity, we assume that b/d is an integer.

Fig. 1 illustrates Definition 2, and Algorithm 1 shows the iterations of quantized SVRG assuming that $\mathcal{R}_{\mathbf{w},k}$ and $\mathcal{R}_{\mathbf{g},k}$ denote quantization grids for parameter \mathbf{w} and for gradient \mathbf{g} , respectively. Various definitions of the projection operator characterize various quantizers. For any point $\mathbf{w} \in \text{Conv}(\mathcal{R})$, let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{2^d}\}$ be the vertices of a d -dimensional cube in \mathcal{R} that contains \mathbf{w} . Next, we define unbiased random quantizer (URQ) as follows:

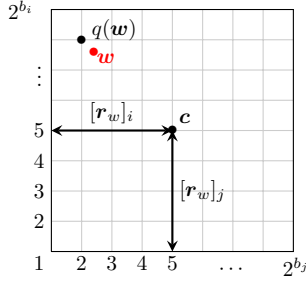


Fig. 1. Our quantization space \mathcal{R} , determined by a grid center \mathbf{c} , coverage radius in every coordinate $\{[r_w]_i\}_{i \in [d]}$, and number of bits $\{b_i\}_{i \in [d]}$. The quantizer maps $\mathbf{w} \in \text{Conv}(\mathcal{R})$ to one of its coordinates using mapping q .

Algorithm 1 Quantized SVRG

- 1: **Inputs:** Epoch length T , number of epochs K , N , step size sequence $(\alpha_k)_k$, number of bits
 - 2: **for** $k = 1, 2, \dots, K$ **do**
 - 3: $\tilde{\mathbf{g}}_k \leftarrow \frac{1}{N} \sum_{i \in [N]} \mathbf{g}_i(\tilde{\mathbf{w}}_k)$
 - 4: Set $\mathcal{R}_{w,k}$ and $\{\mathcal{R}_{\xi,k}\}_{\xi \in [N]}$
 - 5: $\mathbf{w}_{k,0} \leftarrow \tilde{\mathbf{w}}_k$
 - 6: **for** $t = 1, 2, \dots, T$ **do**
 - 7: Sample ξ uniformly from $[N]$
 - 8: Send $\mathbf{g}_\xi(\mathbf{w}_{k,t-1})$ and $q(\mathbf{g}_\xi(\tilde{\mathbf{w}}_k); \mathcal{R}_{\xi,k})$
 - 9: $\mathbf{u}_{k,t} \leftarrow \mathbf{w}_{k,t-1} - \alpha_k (\mathbf{g}_\xi(\mathbf{w}_{k,t-1}) - q(\mathbf{g}_\xi(\tilde{\mathbf{w}}_k)) + \tilde{\mathbf{g}}_k)$
 - 10: $\mathbf{w}_{k,t} \leftarrow q(\mathbf{u}_{k,t}; \mathcal{R}_{w,k})$
 - 11: Broadcast $\mathbf{w}_{k,t}$
 - 12: **end for**
 - 13: Sample ζ uniformly from $\{0, 1, \dots, T-1\}$
 - 14: $\tilde{\mathbf{w}}_{k+1} \leftarrow \mathbf{w}_{k,\zeta}$
 - 15: **end for**
 - 16: **Return:** $\tilde{\mathbf{w}}_{K+1}$
-

Example 3. (URQ). Consider Definition 2. For any $\mathbf{w} \in \text{Conv}(\mathcal{R})$, define URQ $q(\mathbf{w}; \mathcal{R}) = \mathbf{v}_i$ with probability p_i such that $\sum_i p_i = 1$ and $\sum_i p_i \mathbf{v}_i = \mathbf{w}$. By the definition, q has the following properties for $\mathbf{w} \in \text{Conv}(\mathcal{R})$:

- Unbiasedness: $\mathbb{E}[q(\mathbf{w}; \mathcal{R})] = \mathbf{w}$, and
- Error boundedness: $\|q(\mathbf{w}; \mathcal{R}) - \mathbf{w}\| \leq \max_{i,j} \|\mathbf{v}_i - \mathbf{v}_j\|$.

Let b_w and b_g be the number of bits used to quantize parameter and gradient vectors in the inner-loop of the quantized SVRG. The number of bits at every (outer-loop) iteration is $64dM + T(b_w + b_g)$, assuming double-precision floating points for the unquantized gradients of dimension d in the outer-loop. Different from (Sa et al., 2018), Algorithm 1 allows for the exchange of low-precision gradients in the inner-loop. To properly handle this extension, we need to modify the original unquantized SVRG algorithm using a memory unit, as described in the next section.

Abstractly, iterative algorithms search in a space feasible solutions to find the minimizer \mathbf{w}^* . Pseudo-contractive algorithms, which satisfy Assumption 1, inherit a nice property that $\|\mathbf{w}_k - \mathbf{w}^*\|$ and $\|\mathbf{g}(\mathbf{w}_k)\|$ get smaller with k . In other words, by every step, we can shrink the feasibility space. Therefore, using the same number of bits, we may reduce the distance among those quantization points, reducing $\|q(\mathbf{w}_k) - \mathbf{w}_k\|^2$. Under some technical

conditions, formally defined in Section 3, this error may become sufficiently small for a large k , and we may recover full precision of \mathbf{w}_k with a fixed number of bits. This is possible by an adaptive quantization grid, depicted in Fig. 1 for coordinate i and j . Given a proper center for the grid, \mathbf{c}_k , the grid mimics the local geometry of the optimization landscape at iteration k and ensures that \mathbf{w}_{k+1} (and ultimately \mathbf{w}^* as $k \rightarrow \infty$) remains inside the grid by proper choice of $[r_k]_i$ for every coordinate i .

3. MAIN RESULTS

In this section, we present a series of Propositions and Corollaries that characterize the convergence behavior of the quantized SVRG. We show that a naive quantization with few bits and a fixed grid may lead to a suboptimality gap, which can be eliminated by either increasing quantization budget (b_w and b_g) or using local geometry of the optimization landscape through our adaptive quantization scheme.

Proposition 4. Consider a URQ with fixed quantization grid, i.e., $\mathcal{R}_{w,k}$ and $\{\mathcal{R}_{g_\xi,k}\}_{\xi \in [N]}$ are fixed for all iterations k . Let $\alpha_k < 1/6L$ and $T > 1/(\mu\alpha_k(1 - 6L\alpha_k))$. Set $\Delta_k := \mathbb{E}[f(\tilde{\mathbf{w}}_k)] - f(\mathbf{w}^*)$, $\beta_{k,t} := \mathbb{E}[\|q(\mathbf{u}_{k,t}; \mathcal{R}_{w,k}) - \mathbf{u}_{k,t}\|^2]$, and $\delta_k := \mathbb{E}[\|q(\mathbf{g}_\xi(\tilde{\mathbf{w}}_k); \mathcal{R}_{g_\xi,k}) - \mathbf{g}_\xi(\tilde{\mathbf{w}}_k)\|^2]$. The iterates of Algorithm 1 satisfy for any $k \in [0, K-1]$

$$\Delta_{k+1} - \gamma_k \leq \sigma_k (\Delta_k - \gamma_k), \quad (3)$$

where $\sigma_k \in (0, 1)$ and

$$\sigma_k = \frac{\frac{1}{\mu T} + 3L\alpha_k^2}{\alpha_k - 3L\alpha_k^2}, \quad \gamma_k = \frac{3T\alpha_k^2\delta_k + \sum_{t=1}^T \beta_{k,t}}{2T\alpha_k - 12LT\alpha_k^2 - 2/\mu}.$$

Proposition 4 implies that, with the uniform bit allocation in URQ, the quantized SVRG (with fixed grids) converges at a similar rate as the original SVRG until it reaches an ambiguity ball, determined by the combined quantization errors of the gradients (in uplink) and parameters (in downlink), namely $3T\alpha_k^2\delta_k + \sum_{t=1}^T \beta_{k,t}$.

In the following, we show that we can gradually squeeze the quantization grids to constantly reduce $\beta_{k,t}$ and δ_k at every iteration k . Thereby, we may maintain linear convergence with a fixed and limited number of bits. To this end, we slightly modify the SVRG algorithm by adding one memory unit at the master node to save parameter and gradients of the last step. After the inner loop, if the new solution $\tilde{\mathbf{w}}_k$ has a higher gradient norm, we reject this solution and go to the inner loop by the previous one. Hereafter, we call this slightly modified SVRG algorithm as M-SVRG, and its quantized version by QM-SVRG. Now, one can observe that (details in (Ghadikolaei and Magnússon, 2019))

$$\begin{aligned} \|\tilde{\mathbf{w}}_{k+1} - \tilde{\mathbf{w}}_k\| &\leq \|\tilde{\mathbf{w}}_{k+1} - \mathbf{w}^*\| + \|\tilde{\mathbf{w}}_k - \mathbf{w}^*\| \\ &\leq \frac{\|\tilde{\mathbf{g}}_{k+1}\| + \|\tilde{\mathbf{g}}_k\|}{\mu} \leq \frac{2\|\tilde{\mathbf{g}}_k\|}{\mu} := r_{wk}, \end{aligned} \quad (4a)$$

$$\|\mathbf{g}_\xi(\tilde{\mathbf{w}}_{k+1}) - \mathbf{g}_\xi(\tilde{\mathbf{w}}_k)\| \leq \frac{2L\|\tilde{\mathbf{g}}_k\|}{\mu} := r_{gk}. \quad (4b)$$

These inequalities ensure that $\tilde{\mathbf{w}}_{k+1}$ (no matter its true value or quantized version), should be within r_{wk} distance of $\tilde{\mathbf{w}}_k$. Therefore, we can construct an adaptive grid size to

quantize parameter vector at the master node by setting $\mathcal{R}_{w,k} = (\tilde{\mathbf{w}}_k, r_{wk}, \{b_{wi}\}_i)$. Similarly, we can construct an adaptive grid size to quantize gradient vector at the workers by setting $\mathcal{R}_{g_\xi,k} = (\mathbf{g}_\xi(\tilde{\mathbf{w}}_k), r_{gk}, \{b_{gi}\}_i)$.

Proposition 5. Consider a URQ with adaptive quantization grids with $\mathcal{R}_{w,k} = (\tilde{\mathbf{w}}_k, r_{wk})$ and $\mathcal{R}_{g_\xi,k} = (\mathbf{g}_\xi(\tilde{\mathbf{w}}_k), r_{gk})$ in iteration k where r_{wk} and r_{gk} are defined in (4a) and (4b). Set $\Delta_k := \mathbb{E}[f(\tilde{\mathbf{w}}_k)] - f(\mathbf{w}^*)$. Consider uniform bit allocation to all coordinates and also $b_w = b_g = b$, so $b_{xi} = b_{gj} = b/d$ for all $i, j \in [d]$. Let $\alpha_k < \frac{1}{6L}$ and

$$\frac{b}{d} \geq \left\lceil \log_2 \left(1 + \sqrt{\frac{4Ld(1+3L^2\alpha_k^2)}{\mu^2\alpha_k(1-6L\alpha_k)}} \right) \right\rceil, \text{ and}$$

$$T > \frac{1}{\mu\alpha_k(1-6L\alpha_k) - \frac{4L}{\mu}(1+3L^2\alpha_k^2) \frac{d}{(2^{b/d}-1)^2}}.$$

The iterates of the modified version of Algorithm 1 satisfy

$$\Delta_{k+1} \leq \sigma_k \Delta_k, \text{ for any } k \in [0, K-1], \quad (5)$$

where $\sigma_k \in (0, 1)$ and $\sigma_k = \frac{\left(\frac{1}{T} + 3\mu L\alpha_k^2 + \frac{4L}{\mu} \frac{(1+3L^2\alpha_k^2)d}{(2^{b/d}-1)^2}\right)}{\mu(\alpha_k - 3L\alpha_k^2)}$.

Corollary 6. To ensure a contraction constant of at most $\bar{\sigma} < 1$, the minimum number of bits per coordinate and the epoch length should satisfy

$$\frac{b}{d} \geq \left\lceil \log_2 \left(1 + \sqrt{\frac{4Ld(1+3L^2\alpha_k^2)}{\mu^2\alpha_k(\bar{\sigma} - 3L\alpha_k\bar{\sigma} - 3L\alpha_k)}} \right) \right\rceil, \text{ and}$$

$$T > \frac{1}{\mu\alpha_k(\bar{\sigma} - 3L\alpha_k\bar{\sigma} - 3L\alpha_k) - (1+3L^2\alpha_k^2) \frac{4Ld}{\mu(2^{b/d}-1)^2}}.$$

Proposition 5 indicates that our proposal recovers linear convergence rate to the optimal solution with a contraction factor of σ_k . From, Corollary 6, a sufficient condition on scaling of the number of bits per dimension (b/d) is $\log_2(\sqrt{d})$, which is an almost negligible: increasing d from 10 to 1000 may lead to a penalty of 3 additional bits. Moreover, compared to the deterministic approaches, we observe an extra $\log_2(\sqrt{d})$ penalty, which is due to the stochastic gradient terms in the inner loop of Algorithm 1.

4. EXPERIMENTAL RESULTS

4.1 Settings

In this section, we numerically characterize the convergence of our quantized SVRG algorithm on some real-world datasets. For the quantization for each coordinate i of vector \mathbf{v} , we map $[\mathbf{v}]_i$ onto one of the two closest vertices in the quantization grid with probabilities that are inversely proportional to their distance to $[\mathbf{v}]_i$ Sa et al. (2018), leading to an unbiased quantization. We set $\lambda = 0.1$ and consider a logistic ridge regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} \ln \left(1 + e^{-\mathbf{w}^T \mathbf{x}_i y_i} \right) + \lambda \|\mathbf{w}\|_2^2,$$

where each summand is $f_i(\mathbf{w})$. Define $\mathbf{z}_i := \mathbf{x}_i y_i$. 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 $L \geq \text{eig}_{\max}(\nabla^2 f(\mathbf{w}))$ and $\mu \leq \text{eig}_{\min}(\nabla^2 f(\mathbf{w}))$ inequalities. We have,

- Smoothness parameter L :

$$\text{eig}_{\max}(\nabla^2 f(\mathbf{w})) \leq \frac{1}{4N} \sum_{i \in [N]} \|\mathbf{z}_i\|_2^2 + 2\lambda := L.$$

- Strong conv. parameter: $\text{eig}_{\min}(\nabla^2 f(\mathbf{w})) \geq 2\lambda := \mu$.

Datasets: For the binary classification, we consider from the UCI repository the Individual Household Electric Power Consumption dataset, which has 2,075,259 samples (\mathbf{x}_i, y_i) of dimension $d = 9$. We apply a hard threshold technique on the value of one output to make it a binary class. For the multiclass classification task, we use MNIST dataset, which has 60,000 training samples (\mathbf{x}_i, y_i) of dimension $d = 784$ and 10 classes corresponding to handwritten digits. We applied the one-versus-all technique to solve 10 independent binary classification problems (using logistic ridge regression) and obtain 10 optimal classifiers, $\mathbf{w}^{(l)}$, one for each digit. Given a test data \mathbf{x} , we choose label l that maximizes $(\mathbf{w}^{(l)})^T \mathbf{x}$.

Benchmark Algorithms: We have implemented SVRG, modified SVRG (M-SVRG), gradient descent (GD) (Bottou et al., 2018), stochastic gradient descent (SGD) (Bottou et al., 2018), and stochastic average gradient (SAG) (Schmidt et al., 2017). We should highlight that we count outer-loops of SVRG (also its variants) as the iterations. In terms of bits, following is the number of bits required to run one iteration:

$$\text{SGD and SAG} = 128d, \quad \text{GD} = 64d(1+N), \quad \text{and}$$

$$\text{SVRG and M-SVRG} = 64dN + 192dT,$$

where $192d$ corresponds to sending two gradients and one parameter vectors of size d with N workers.

We have also implemented the quantized version of M-SVRG using fixed lattice (QM-SVRG-F) adaptive lattice (QM-SVRG-A) as well as that of benchmark algorithms (GD, SGD, and SAG). We also implement QM-SVRG-F+ and QM-SVRG-A+ in which we quantize both $\mathbf{g}_\xi(\mathbf{w}_{k,t-1})$ and $\mathbf{g}_\xi(\tilde{\mathbf{w}}_k)$ using grid $\mathcal{R}(g_\xi, k)$ in the inner loop. For the benchmarks, we apply quantizer with a fixed lattice to gradient and parameter vectors, as of the QM-SVRG-F algorithm. Following is the number of bits required to run one iteration:

$$\text{Q-SGD and Q-SAG} = b_w + b_g, \quad \text{Q-GD} = b_w + b_g N,$$

$$\text{QM-SVRG-F and QM-SVRG-A} = 64dN + 64dT + (b_w + b_g)T,$$

$$\text{QM-SVRG-F+ and QM-SVRG-A+} = 64dN + (b_w + b_g)T.$$

Performance Measures: We have considered the convergence of loss function (zero-order stopping criterion) and the gradient norm (first-order stopping criterion) for the training data, and the F1-score of \mathbf{w}_k on the test dataset to assess the generalization performance. All the simulation codes are available at (Ghadikolaei and Magnússon, 2019).

4.2 Results and Discussions

Fig. 2 illustrates the theoretical bounds, derived in Section 3. The top figure depicts the theoretical lower bound on the epoch size T that ensures a contraction factor of at most $\bar{\sigma}$. First of all, depending on the target contraction factor and the number of bits per coordinate b/d , the feasibility region of α_k varies. For example, to ensure $\bar{\sigma} = 0.2$ (a

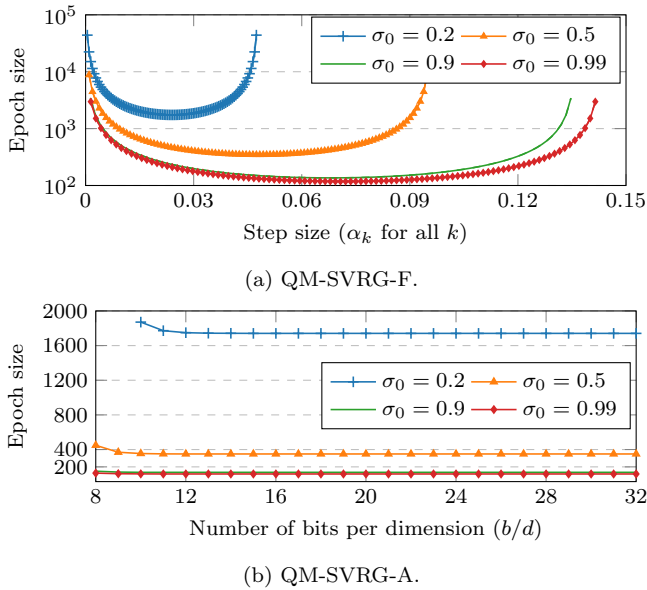


Fig. 2. Sufficient conditions (lower bounds) on the minimum epoch size T against (a) step-size α_k and (b) number of bits per dimension b/d to ensure contraction factor $\bar{\sigma}$ for the Individual Household Electric Power Consumption dataset. In (a), values of QM-SVRG-A with $b/d = 10$ coincide with the corresponding values of QM-SVRG-F, given in this subfigure.

very good contraction factor), we need a high resolution quantizer (10 bits per coordinate) and very small step-size of $\alpha_k < 0.047$. However, $\bar{\sigma} = 0.9$ is attainable by 8 bits per coordinate and α_k as large as 0.124. The bottom figure shows the same y-axis for the number of bits. Clearly, increasing the number of bits reduces the required T to ensure contraction $\bar{\sigma}$ but in a saturating fashion. In particular, there is no difference in terms of epoch size requirement as well as convergence (as we shall see later on) between $b/d = 15$ and the usual $b/d = 64$. Another point is that the lower the target contraction factor $\bar{\sigma}$, the higher b/d and minimum T . Notice that unlike QM-SVRG-F, our QM-SVRG-A algorithm can recover the optimal solution using a finite number of bits per iteration. In the following, we show that the theoretical bounds to ensure convergence are only sufficient conditions and may be very conservative, and we may be able to quantize in practice well beyond those bounds.

Fig. 3 illustrates the convergence of our performance measures in the presence of quantization. From Fig. 3a, QM-SVRG-A+ can maintain the convergence to the optimal solution even with a very severe quantization: 3 bits per dimension that is equivalent to 95% compression in the parameter update and gradient reports. Moreover, this is obtained by $T = 8$, well below the values with theoretical guarantees (usually $T > 100$, see Fig. 2). The QM-SVRG-F algorithm (as well as Q-GD, Q-SAG, and Q-SGD), however, cannot keep up with such a severe quantization and maintain convergence, even to a small ambiguity ball. This problem may be addressed only by increasing the number of bits, as we can observe in Fig. 3b. Some algorithms, such as Q-SAG and Q-GD, may be more sensitive and need more quantization bits to recover their convergence. By adding more bits, these algorithms converge at a similar

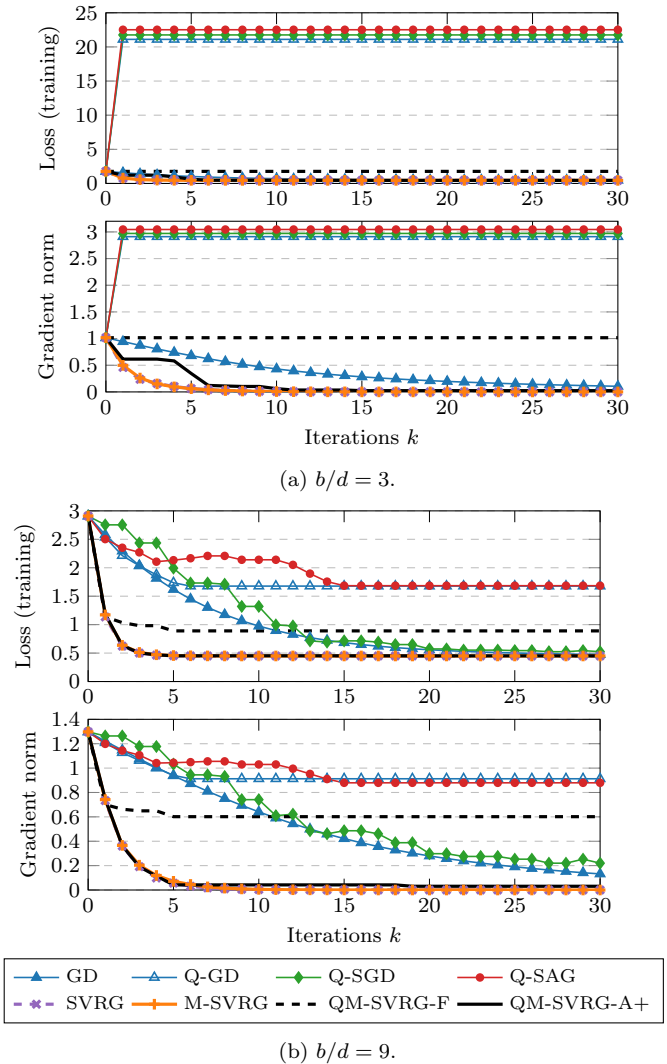


Fig. 3. Convergence results for Individual Household Electric Power Consumption dataset with $T = 8$ and $\alpha_k = 0.2$ for all k . Legend Q-x stands for quantized version of algorithm x.

rate of their unquantized version to an ambiguity ball, whose radius determined by the quantization error. Another observation is the noticeable superior performance of the quantized version of M-SVRG, compared to that of other approaches.

To assess the impact of higher dimensions and harder learning tasks, we consider a multiclass classification task using the MNIST dataset. Fig. 4 illustrates the training performance of the classifier for digit 9 with 7 and 10 bits per dimension. Similar to previous cases, the QM-SVRG-A solver can preserve the convergence rate with a severe quantization, thanks to our proposed adaptive grid structure, while other approaches fail to retrieve the optimal solution. To better understand the performance of our final solution on all digits, we have reported in Table 1 the F1-score, averaged over all classes. F1-score is computed assuming digit 9 is the class 1 while all other digits are class -1. M-SVRG slightly outperforms SVRG (not shown in the table) at both $b/d = 7$ and $b/d = 10$. We should highlight that we did not try to optimize hyper-parameters to achieve a better F1-score in

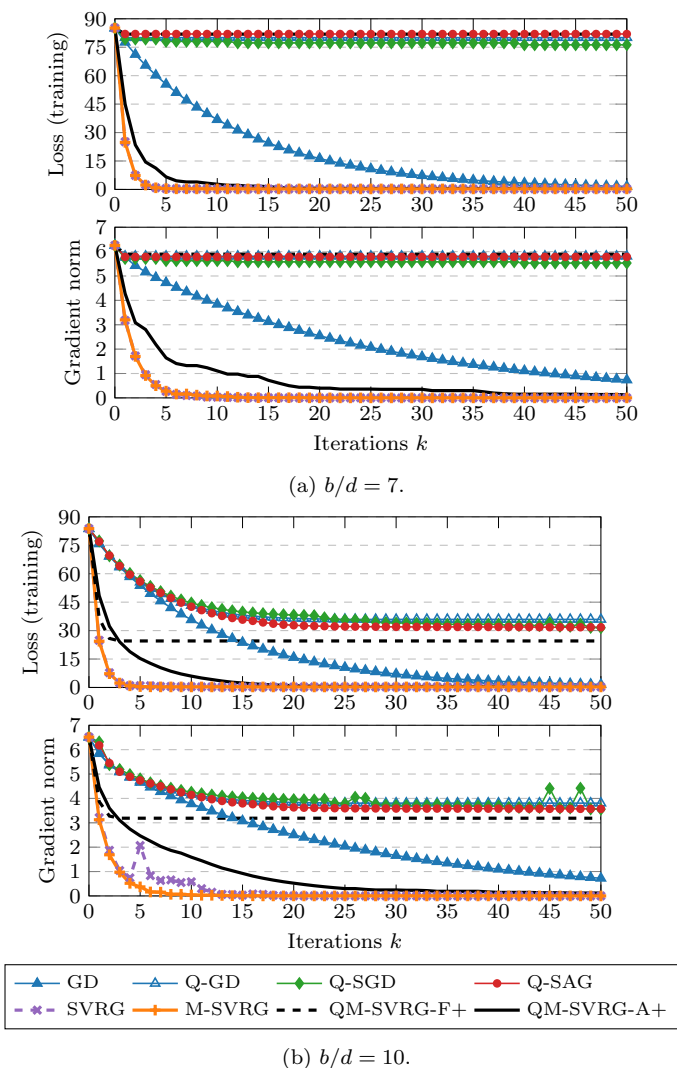


Fig. 4. Convergence results for the MNIST dataset (digit 9) with $T = 15$ and $(\alpha_k = 0.2)_k$.

Table 1. F1-score for the MNIST test dataset with $(\alpha_k = 0.2)_k$, $T = 15$, and 50 iterations. Table shows the performance of unquantized and quantized algorithms. “Q-F” and “Q-A” are QM-SVRG-F+ and QM-SVRG-A+.

b/d	GD	M-SVRG	Q-GD	Q-SGD	Q-SAG	Q-F	Q-A
7	0.775	0.841	0.127	0.101	0.130	0.139	0.806
10	0.780	0.841	0.248	0.402	0.168	0.280	0.838

our experiments. This table, along with Fig. 4, indicates a very good performance of our approach compared to the state-of-the-art unquantized and quantized baselines, in terms of training and generalization performance as well as robustness to smaller b/d .

5. CONCLUSIONS

We have proposed an adaptive quantization method for SVRG algorithm to maintain its convergence property by communicating only a few bits per iterations. We have theoretically derived the convergence of our novel algorithm and evaluated its performance on a variety of machine learning tasks. Our results indicated that our approach provides a significant compression and is more

robust than the quantized versions of other state-of-the-art algorithms for solving distributed optimization problems. Our approach has an immediate use case in the emerging field of machine learning over wireless networks (e.g., internet-of-things) and edge computing.

REFERENCES

- Bernstein, J., Wang, Y.X., Azizzadenesheli, K., and Anandkumar, A. (2018). signSGD: compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*.
- Bottou, L., Curtis, F., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223–311.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.
- Furht, B. and Ahson, S.A. (2016). *Long Term Evolution: 3GPP LTE radio and cellular technology*. Crc Press.
- Ghadikolaei, H.S. and Magnússon, S. (2019). Communication-efficient variance-reduced stochastic gradient descent: Extended version. [Online]. Available: <https://github.com/hshokrig/papers-QM-SVRG>.
- Jiang, X. et al. (2019). Low-latency networking: Where latency lurks and how to tame it. *Proc. IEEE*, 107(2), 280–306.
- Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in NIPS*, 315–323.
- Jordan, M.I., Lee, J.D., and Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526), 668–681.
- Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–811. Springer.
- Konečný, J. et al. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.
- Li, H., Ota, K., and Dong, M. (2018). Learning IoT in edge: Deep learning for the internet of things with edge computing. *IEEE Network*, 32(1), 96–101.
- Magnússon, S., Enyioha, C., Li, N., Fischione, C., and Tarokh, V. (2018). Convergence of limited communication gradient methods. *IEEE Trans. Automat. Contr.*, 63(5), 1356–1371.
- Magnússon, S., S. Ghadikolaei, H., and Li, N. (2019). On maintaining linear convergence of distributed learning and optimization under limited communication. *arXiv preprint arXiv:1902.11163*.
- Sa, D. et al. (2018). High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*.
- Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2), 83–112.
- Wen, W. et al. (2017). Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in NIPS*, 1509–1519.
- Xu, Y. and Yin, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imag. Sciences*, 6(3), 1758–1789.