Journal of Machine Learning Research 21 (2020) 1-53 Submitted 7/18; Revised 4/20; Published 10/20

A Unified q-Memorization Framework for Asynchronous Stochastic Optimization

Bin Gu JSGUBIN@GMAIL.COM
School of Computer & Software, Nanjing University of Information Science & Technology, China
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
JD Finance America Corporation, Mountain View, CA, 94043, USA

Wenhan Xian WEX37@PITT.EDU
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA

Zhouyuan Huo ZHHUO@GOOGLE.COM
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA

Cheng Deng CHDENG.XD@GMAIL.COM
School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, 710071, China

Heng Huang HENG.HUANG@PITT.EDU
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
JD Finance America Corporation, Mountain View, CA, 94043, USA

Editor: Mark Schmidt

Abstract

Asynchronous stochastic algorithms with various variance reduction techniques (such as SVRG, S2GD, SAGA and q-SAGA) are popular in solving large scale learning problems. Recently, Reddi et al. (2015) proposed an unified variance reduction framework (i.e., HSAG) to analyze the asynchronous stochastic gradient optimization. However, the HSAG framework cannot incorporate the S2GD technique, the analysis of the HSAG framework is limited to the SVRG and SAGA techniques on the smooth convex optimization. They did not analyze other important various variance techniques (*e.g.*, S2GD and q-SAGA) and other important optimization problems (*e.g.*, convex optimization with non-smooth regularization and non-convex optimization with cardinality constraint). In this paper, we bridge this gap by using an unified q-memorization framework for various variance reduction techniques (including SVRG, S2GD, SAGA, q-SAGA) to analyze asynchronous stochastic algorithms for three important optimization problems. Specifically, based on the q-memorization framework, **1**) we propose an asynchronous stochastic gradient hard thresholding algorithm with q-memorization (AsySGHT-qM) for the non-convex optimization with cardinality constraint, and prove that the convergence rate of AsySGHT-qM before reaching the inherent error induced by gradient hard thresholding methods is geometric. **2**) We propose an asynchronous stochastic proximal gradient algorithm (AsySPG-qM) for the convex optimization with non-smooth regularization, and prove that AsySPG-qM can achieve a linear convergence rate. **3**) We propose an asynchronous stochastic gradient descent algorithm (AsySGD-qM) for the general non-convex optimization problem, and prove that AsySGD-qM can achieve a sublinear convergence rate to stationary points. The experimental results on various large-scale datasets confirm the fast convergence of our AsySGHT-qM, AsySPG-qM and AsySGD-qM through concrete realizations of SVRG and SAGA.

Keywords: Stochastic optimization, q-memorization, asynchronous parallel computing, variance reduction, proximal operator, hard thresholding

1. Introduction

Large-scale learning problems are ubiquitous in the current era of big data. For example, Flickr (a public picture sharing site) daily received 1.8 million photos on average from February to March in 2012 (Michel, 2012; Wu et al., 2014). If we want to learn a classifier based on these cumulative photos (Wang et al., 2012), it is inevitable to design a large-scale learning algorithm to handle the massive amounts of photo data. Parallel computation and stochastic optimization are the dominant techniques for solving this kind of large scale learning problems.

Parallel Computation. Parallel computation techniques were recently proposed to address large-scale learning problems, benefiting from the popularity of multi-core processors and GPU-accelerators. Parallel computation techniques can be roughly divided into synchronous and asynchronous models, according to whether the reading or writing lock is used. As pointed out in many literatures (Lian et al., 2016; Liu and Wright, 2015; Zhao and Li, 2016; Lian et al., 2015), the synchronous parallel model reduces parallel efficiency, because all other computational resources need to wait the ongoing computational resource when reading or writing a variable. On the other hand, the asynchronous parallel model is much more efficient than the synchronous parallel model, because it keeps all computational resources busy all the time. It should be noted that, the convergence analysis for the asynchronous parallel algorithm is much more difficult than the one for the synchronous parallel algorithm, due to the *inconsistent reading*[1]. In this paper, we focus on the asynchronous parallel model on the parallel environment with shared memory (such as multi-core processors and GPU-accelerators).

Stochastic Optimization. Stochastic optimization is the other important big data computation technique. For the full gradient descent algorithm, the full gradient is used to update the solution, which is quite computational costly for large-scale data. Different to the full gradient descent algorithm, stochastic gradient descent (SGD) algorithm (Bottou, 2010) uses the stochastic gradient on a sample or a subset of samples to update the solution, instead of the full gradient. Thus, the SGD algorithm has a cheap computation for each iteration. However, compared to the linear convergence rate of the full gradient descent algorithm, the SGD algorithm has a low sublinear convergence rate $O(\frac{1}{T})$ due to the variance of stochastic gradients introduced by random sampling, where T is the iteration number.

To accelerate the SGD algorithm, there have been several variance reduction techniques proposed to reduce the variance of stochastic gradients. Basically, these variance reduction techniques use different strategies to combine the full gradient and the stochastic gradient. Specifically, the variance reduction techniques include SVRG (Johnson and Zhang, 2013), S2GD (Konečný and Richtárik, 2017), SAGA (Defazio et al., 2014) and q-SAGA (Hofmann et al., 2015), SAG (Schmidt et al., 2017). We give a detailed comparison of the representative unbiased variance reduction techniques (i.e., SVRG, S2GD, SAGA and q-SAGA) in Section

[1] Because the asynchronous parallel algorithm does not use the reading and writing locks, the variables read into the local memory may be inconsistent to the ones in shared memory. It is the so-called inconsistent reading.

2.1 and Table 2. From the comparison, we find that SVRG and S2GD have a low space cost and a high computational cost for each epoch. On the other hand, SAGA and q-SAGA have a low computational cost and a high space cost for each iteration. In addition, S2GD and q-SAGA are the general and adjustable versions to SVRG and SAGA respectively. Thus, each variance reduction technique has its specific merit. Note that, we only consider the unbiased variance reduction techniques in this paper. The SAG technique (Schmidt et al., 2017) uses biased stochastic gradients which is out of scope of this paper.

Asynchronous stochastic optimization algorithms with various variance reduction techniques have been proposed to solve large scale learning problems. Specifically, for smooth convex optimization problems, Zhao and Li (2016) proposed an asynchronous stochastic algorithm with SVRG, and proved its linear convergence rate. Mania et al. (2017) proposed a perturbed iterate framework to analyze the asynchronous stochastic SVRG algorithm with sparse gradients. Leblond et al. (2017) proposed an asynchronous SAGA algorithm, and proved its linear convergence rate. Huo and Huang (2017) extended the asynchronous stochastic SVRG algorithm to non-convex optimization problems and proved its sublinear convergence rate. Gu et al. (2018) proposed an asynchronous stochastic zeroth order gradient algorithm with SVRG technique to non-convex optimization problems and proved its sublinear convergence rate. For convex optimization problems with non-smooth regularization, Meng et al. (2017); Gu and Huo (2018) independently proposed asynchronous stochastic proximal gradient algorithms with SVRG, and proved their linear convergence rates. Pedregosa et al. (2017) proposed an asynchronous stochastic proximal gradient algorithm with SAGA, and proved its linear convergence rate. For non-convex optimization problems with cardinality constraint, Li et al. (2016) proposed asynchronous stochastic gradient hard thresholding algorithms with the SVRG and SAGA techniques, and proved the linear convergence rate to an approximately global optimum for the SVRG case. Gu et al. (2019) proposed asynchronous stochastic Frank-Wolfe algorithm and its SVRG variant, and proved their convergence rates. We also summarize these representative (asynchronous) stochastic gradient descent algorithms in Table 1.

As mentioned previously, different variance reduction techniques have their specific merit. Thus, it is highly desired to propose an unified variance reduction framework to asynchronous stochastic optimization. To the best of our knowledge, the only unified variance reduction framework for asynchronous stochastic optimization is (Reddi et al., 2015). Specifically, Reddi et al. (2015) proposed an unified variance reduction framework (i.e., HSAG) to analyze the asynchronous stochastic gradient algorithm for the smooth convex optimization, and proved its linear convergence rate. However, the HSAG framework cannot incorporate the S2GD technique, and the analysis for the HSAG framework is limited to the SVRG and SAGA techniques for smooth convex smooth optimization problems. They did not analyze other important variance reduction techniques (*e.g.*, S2GD and q-SAGA) and other important optimization problems, such as the non-convex optimization with cardinality constraint, the convex optimization with non-smooth regularization, and the general non-convex optimization problem.

To bridge this gap, we introduce a more unified and general variance reduction framework (i.e., q-memorization) (Hofmann et al., 2015) which is originally proposed for analyzing the sequential stochastic gradient algorithm for the smooth convex optimization. In this paper, we use the unified q-memorization framework to analyze asynchronous stochastic gra-

Table 1: Representative (asynchronous) stochastic gradient descent algorithms with various variance reduction techniques. (C, SC, NC, S, NS, RSS and RSC are the abbreviations of convex, strongly convex, non-convex, smooth, non-smooth, restricted strong smoothness and restricted strong convexity respectively.)

Reference	**Problem**	**Technique**	**Parallel**	**Asynchronous**
Johnson and Zhang (2013)	S&SC+NS	SVRG	No	No
Konečný and Richtárik (2017)	SC	S2GD	No	No
Defazio et al. (2014)	SC+NS	SAGA	No	No
Hofmann et al. (2015)	SC	q-SAGA	No	No
Schmidt et al. (2017)	SC	SAG	No	No
Reddi et al. (2016a)	S&NC	SVRG	No	No
Allen-Zhu and Hazan (2016)	S&NC	SVRG	No	No
Reddi et al. (2016b)	S&NC	SAGA	No	No
Lei et al. (2017)	S&NC	SCSG	No	No
Zhao and Li (2016)	S&SC	SVRG	Yes	Yes
Mania et al. (2017)	S&C/SC	SGD/SVRG	Yes	Yes
Huo and Huang (2017)	S&NC	SVRG	Yes	Yes
Leblond et al. (2017)	S&SC	SAGA	Yes	Yes
Meng et al. (2017); Gu and Huo (2018)	S&SC+NS	SVRG	Yes	Yes
Pedregosa et al. (2017)	S&SC+NS	SAGA	Yes	Yes
Li et al. (2016)	RSS&RSC+l_0	SVRG	Yes	Yes
Reddi et al. (2015)	S&SC	HSAG	Yes	Yes
Lian et al. (2015)	S&NC	SGD	Yes	Yes
Gu et al. (2018)	S&NC	SVRG	Yes	Yes
Gu et al. (2019)	NC	SGD/SVRG	Yes	Yes

dient algorithms for three classes of important optimization problems (i.e., the non-convex optimization problem with cardinality constraint, the convex optimization problem with non-smooth regularization and the general non-convex optimization problem). Specifically, based on the q-memorization framework,

1) we propose an asynchronous stochastic gradient hard thresholding algorithm with q-memorization (AsySGHT-qM) for the non-convex optimization problem with cardinality constraint. We prove that the convergence rate of AsySGHT-qM before reaching the inherent error induced by gradient hard thresholding methods is geometric.

2) we propose an asynchronous stochastic proximal gradient algorithm (AsySPG-qM) for the convex optimization problem with non-smooth regularization. We prove that AsySPG-qM can achieve a linear convergence rate.

3) we propose an asynchronous stochastic gradient descent algorithm (AsySGD-qM) for the general non-convex optimization problem. W prove that AsySGD-qM can achieve a sublinear convergence rate to stationary points.

The experimental results on various large-scale datasets confirm the fast convergence of our AsySGHT-qM, AsySPG-qM and AsySGD-qM through concrete realizations of SVRG and SAGA.

In the following, we give more details to the non-convex optimization problem with cardinality constraint, the convex optimization problem with non-smooth regularization, and general non-convex smooth optimization problem, which will be addressed in this paper.

1.1 Non-convex Optimization with Cardinality Constraint

Sparse learning plays an important role in machine learning. First, high-dimensional data (such as DNA microarray data, recommendation data, social network data and so on) are becoming increasingly available as data collection technique evolves. One truth is that most features in high-dimensional data are non-informative or noisy. Second, by substituting high dimensional sparse data by low-dimensional representations, the generalization ability of the models can be improved. Third, a sparse model can lead to a simplified decision rule which can lead to faster prediction which is especially important for large scale problems. Finally, sparse learning can lead to a model with better interpretation because a small set of important features is selected. Sparse learning is normally conducted by sparsity constraints on the model parameter. There are several sparse constraints, such as l_1-norm constraint (Tibshirani, 1996), $l_{1/2}$-norm constraint (Liang et al., 2013) and cardinality constraint (Régin, 1996), and so on. Among them, cardinality constraint is the intrinsic way for sparse learning. to stationary points

To implement the sparse learning, we consider the following generic non-convex optimization problem with cardinality constraint in this paper.

$$\min_{x \in \mathbb{R}^n} \quad \underbrace{\frac{1}{l} \sum_{i=1}^{l} f_i(x)}_{F(x)} \quad s.t. \quad \|x\|_0 \leq k \tag{1}$$

where $F(x)$ is a smooth and non-strongly convex function with the additive form $\frac{1}{l} \sum_{i=1}^{l} f_i(x)$, each function $f_i(x)$ is a smooth function. The formulation (1) covers many machine learning problems, such as sparsity-constrained linear regression model (Tropp and Gilbert, 2007), sparsity-constrained logistic regression model (Tropp and Gilbert, 2007), sparsity-constrained graphical model (Jalali et al., 2011).

Directly solving the problem (1) is NP-hard (Natarajan, 1995). Existing works (Yuan et al., 2014; Jain et al., 2014; Nguyen et al., 2017; Li et al., 2016; Shen and Li, 2018) try to obtain a good approximation of the global solution to (1). Specifically, Yuan et al. (2014) and Jain et al. (2014) proposed the gradient hard thresholding (GHT) algorithm which offer a fast and scalable batch algorithm. To further scale up the GHT algorithm, Nguyen et al. (2017) proposed the stochastic gradient hard thresholding (SGHT) algorithm. Li et al. (2016) and Shen and Li (2018) proposed the SGHT algorithm with SVRG, and prove its linear convergence rate before reaching the inherent error induced by GHT-style methods. In this paper, we design a new generalized variance reduction asynchronous stochastic gradient hard thresholding algorithm (AsySGHT-qM) based on the q-memorization framework.

1.2 Convex Optimization with Non-smooth Regularization

Many regularized empirical risk minimization problems (such as Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), regularized logistic regression (Lee et al., 2006)) consists of a finite sum of smooth convex functions and a convex (possibly non-smooth) regularized function. Formally, we present this kind of problems as a composite objective function as follows.

$$\min_{x\in\mathbb{R}^n} F(x) = \underbrace{\frac{1}{l}\sum_{i=1}^{l} f_i(x)}_{f(x)} + h(x) \tag{2}$$

where $f_i : \mathbb{R}^n \mapsto \mathbb{R}$ is a smooth and convex function, and $h : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is a convex but possibly non-smooth function. Without loss of generality, we further assume that there exists a partition $\{\mathcal{G}_1, \cdots, \mathcal{G}_k\}$ on n features (i.e., coordinates) of x, where each $\mathcal{G}_j$ is also called as block. Thus, we can write the function $h(x)$ as $h(x) = \sum_{j=1}^{k} h_{\mathcal{G}_j}(x_{\mathcal{G}_j})$.

To solve the problem (2) with $k = 1$, Xiao and Zhang (2014); Nitanda (2014) proposed proximal stochastic gradient algorithms with SVRG. Defazio et al. (2014) proposed a proximal stochastic gradient method with SAGA. To solve the problem (2) with $k > 1$, Hong et al. (2017) proposed a batch randomized block coordinate descent method which runs with full gradient on the randomized block coordinates. Zhao et al. (2014) proposed a double stochastic proximal gradient algorithm (DSPG) with SVRG. In addition to these sequential stochastic algorithms, Meng et al. (2017) and Gu and Huo (2018) independently proposed asynchronous stochastic proximal gradient algorithms with SVRG, and proved their linear convergence rates. Pedregosa et al. (2017) proposed an asynchronous stochastic proximal gradient algorithm with SAGA, and proved its linear convergence rate. In this paper, we design a new generalized variance reduction asynchronous stochastic proximal gradient algorithm (AsySPG-qM) based on the q-memorization framework.

1.3 General Non-Convex Smooth Optimization

Non-convex optimization has become increasingly popular in machine learning because of two inevitable trends, *i.e.*, robust learning and deep learning (Huo et al., 2018b,a). Take robust learning for example, we normally minimize an empirical or regularized risk problem with non-convex loss functions (*e.g.* ramp loss (Huang et al., 2014), sigmoid loss and correntropy induced loss (He et al., 2011)), instead of optimizing convex surrogate loss functions. Formally, we present this kind of problems as follows.

$$\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{l}\sum_{i=1}^{l} f_i(x) \tag{3}$$

where $f_i : \mathbb{R}^N \mapsto \mathbb{R}$ is a smooth and non-convex function. Note that one may absorb a regularizer in the definition of the function $f_i(x)$.

To solve the problem (3), Reddi et al. (2016a); Allen-Zhu and Hazan (2016) proposed SVRG algorithm to solve (3)), and proved the sublinear convergence rate of SVRG to stationary points under the general non-convex setting. Reddi et al. (2016b) proposed SAGA algorithm to solve (3) and proved the sublinear convergence rate of SVRG to stationary points under the general non-convex setting. Lei et al. (2017) proposed a variant of S2GD (stochastically controlled stochastic gradient (SCSG)) to solve (3), where the full gradient in S2GD is replaced by a mini-batch of samples). They proved the sublinear convergence rate of SCSG to stationary points under the general non-convex setting, and proved the linear convergence rate under the Polyak-Lojasiewicz condition. Lian et al. (2015) proposed an asynchronous parallel stochastic gradient algorithm to solve (3)), and proved the ergodic convergence rate under the general non-convex setting. Huo and Huang (2017); Gu et al. (2018) proposed the asynchronous stochastic SVRG algorithm to to solve (3)) and proved its sublinear convergence rate under the general non-convex setting. In this paper, we design a new generalized variance reduction asynchronous stochastic gradient descent algorithm (AsySPG-qM) based on the q-memorization framework.

1.4 Contributions

The main contributions of this paper are summarized as follows:

1. We analysis the limitations of the HSAG framework, and introduce a more unified and general variance reduction framework (i.e., q-memorization) (Hofmann et al., 2015) for analyzing the convergence rates of asynchronous stochastic algorithms.

2. Based on the q-memorization framework, we propose an asynchronous stochastic gradient hard thresholding algorithm with q-memorization (AsySGHT-qM) for the non-convex optimization problem with cardinality constraint. We prove that the convergence rate of AsySGHT-qM before reaching the inherent error induced by GHT-style methods is geometric. The experimental results on various large-scale datasets confirm the fast convergence of our AsySGHT-qM through concrete realizations of SVRG and SAGA.

3. Based on the q-memorization framework, we propose an asynchronous stochastic proximal gradient algorithm (AsySPG-qM) for the convex optimization problem with non-smooth regularization. We prove that AsySPG-qM achieves a linear convergence rate. The experimental results on various large-scale datasets confirm the fast convergence of our AsySPG-qM through concrete realizations of SVRG and SAGA.

4. Based on the q-memorization framework, we propose an asynchronous stochastic gradient descent algorithm (AsySGD-qM) for the general non-convex optimization problem. We prove that AsySGD-qM achieves a sublinear convergence rate to stationary points. The experimental results on various large-scale datasets confirm the fast convergence of our AsySGD-qM through concrete realizations of SVRG and SAGA.

1.5 Outline

We organize the rest of the paper as follows. In Section 2, we present a general variance reduction framework (i.e., q-memorization). In Section 3, we propose our AsySGHT-qM, and provide its linear convergence rate to reach the inherent error induced by GHT-style methods. In Section 4, we propose our AsySPG-qM algorithm, and provide its linear convergence rate. In Section 5, we propose our AsySGD-qM algorithm, and provide its sublinear convergence rate. In Section 6, we present the experimental results on a variety of datasets. In Section 7.2, we prove the convergence rates of AsySGHT-qM, AsySPG-qM and AsySGD-qM. Finally, we give some concluding remarks in Section 8.

1.6 Notations

In order to make notations easier to follow, we give a summary of the notations in the following list.

$\|x\|_0$ the number of nonzero entries in x.

$\|x\|_1 = \sum_{i=1}^n |x_i|$ the ℓ_1-norm of x.

$\mathbb{1}(m|t)$, $\mathbb{1}(m \not| t)$ $\mathbb{1}(m|t)$ is the boolean value determining whether t is divisible by m. $\mathbb{1}(m \not| t)$ is the inverse boolean value of $\mathbb{1}(m|t)$.

I_p the p-dimensional identity matrix.

$supp(x)$ the index set of nonzero entries of x.

$\nabla f_i(x)$ $\nabla f_i(x)$ is the gradient of the function $f_i(x)$ at the point x.

S_i, Φ_i S_i is the support of $\nabla f_i(x)$. Φ_i is the extended support of $\nabla f_i(x)$, which is the set of blocks that intersect S_i, and formally defined as $\Phi_i = \{S_i \cap \mathcal{G}', \mathcal{G}' \in \{\mathcal{G}_1, \cdots, \mathcal{G}_k\}\}$.

$\overline{\alpha} = \frac{1}{l}\sum_{k=1}^{l} \alpha_k$ the average gradient of α_k over $k = 1, \cdots, n$.

$\partial F(x)$ the set of all subgradinets of the function $F(x)$ at the point x.

$\overline{\mathcal{I}}$ the complementary set of $\mathcal{I}$, i.e., $\overline{\mathcal{I}} = \{1, 2, \ldots, n\} - \mathcal{I}$.

$(\alpha_i)_{\mathcal{I}}$, $\nabla_{\mathcal{I}} f_i(x)$ the vectors same with the vectors α_i and $\nabla f_i(x)$ respectively, except that the entries indexed by $\overline{\mathcal{I}}$ are zero.

$\mathcal{H}_k(\cdot)$ the hard thresholding operator that keeps the largest k entries in magnitude and sets the other entries equal to zero.

2. Unified Variance Reduction Framework

Stochastic gradient algorithm uses the updating rule $x^+ \leftarrow x - \gamma \nabla f_i(x)$ which minimizes the problem $f(x) = \frac{1}{l}\sum_{i=1}^{l} f_i(x)$, where γ is the step size, i is an index of the sample selected uniformly at random and x^+ denotes the updated solution x after one algorithm iteration. However, as mentioned previously, the standard stochastic gradient algorithm has a low convergence due to the variance of stochastic gradients introduced by random sampling. To reduce the variance of stochastic gradients, various variance reduction techniques and frameworks have been proposed to accelerate the stochastic gradient algorithm. The updating rule of these variance reduction techniques have the following form.

$$x^+ \leftarrow x - \gamma v, \quad \text{where } \; v = \nabla f_i(x) - \alpha_i + \overline{\alpha} \tag{4}$$

where α_i denotes the outdated gradient on the i-th sample. Different variance reduction techniques have different strategies to update the outdated gradient α_i. Note that we consider the unbiased variance reduction techniques such that $\mathbb{E}v = \nabla f(x)$.

In this section, we first present the representative variance reduction techniques, and then present the HSAG framework. Finally, we introduce a more general variance reduction framework (i.e., q-memorization) which will be used to analyze our AsySGD-qM, AsySPG-qM and AsySGHT-qM.

2.1 Representative Unbiased Variance Reduction Techniques

In this subsection, we give the descriptions of representative unbiased variance reduction techniques (i.e., SVRG, S2GD, SAGA, and q-SAGA).

As mentioned above, different variance reduction techniques have different strategies to update the outdated gradient α_i. Technically, we define two elements (i.e, index set J and iteration number of one epoch) to define the different variance reduction techniques. Specifically, . We present the different variance reduction techniques as follows.

1. **SVRG**: SVRG updates all α_i after m iterations, where m is fixed. Thus, the index set $J=\emptyset$ for the iterations $\{1,\cdots,m-1\}$, and $J=\{1,\cdots,l\}$ for m-th iteration.

2. **S2GD**: S2GD updates all α_i after t iterations, where $t\in\{1,\cdots,m\}$ is a random variable obeying the distribution $\mathbf{P}(t)=\frac{(1-\nu\gamma)^{m-t}}{\beta}$, where ν is a nonnegative constant not greater than the strong convexity parameter of the objective function, γ is the steplength parameter, and $\beta=\sum_{t=1}^{m}(1-\nu\gamma)^{m-t}$. Thus, the index set $J=\emptyset$ for the iterations $\{1,\cdots,t-1\}$, and $J=\{1,\cdots,l\}$ for t-th iteration.

3. **SAGA**: SAGA updates α_i with $\nabla f_i(x)$ for each iteration, where i is the index of the sample used in (4). Thus, the size of the index set J is one.

4. **q-SAGA**: q-SAGA randomly selects an index set $J\subseteq\{1,\cdots,l\}$ such that $|J|=q$ for each iteration. q-SAGA updates α_J with $\nabla f_J(x)$ for each iteration.

We also summarize these variance reduction techniques in Table 2.

Table 2: Representative unbiased variance reduction techniques. (J is the index set for updating the stochastic gradient, $\mathbf{P}(t)$ denotes the probability of the value t.)

Technique	Iteration number for an epoch	Index set J for each iteration	Space cost for an epoch	Time cost for an epoch
SVRG	m	$J=\{1,\cdots,l\}\vee J=\emptyset$	$O(N)$	$O(Nl)$
S2GD	$t\in\{1,\cdots,m\}$, $\mathbf{P}(t)=\frac{(1-\nu\gamma)^{m-t}}{\beta}$	$J=\{1,\cdots,l\}\vee J=\emptyset$	$O(N)$	$O(Nl)$
SAGA	1	$\lvert J\rvert=1\wedge J\subseteq\{1,\cdots,l\}$	$O(Nl)$	$O(N)$
q-SAGA	1	$\lvert J\rvert=q\wedge J\subseteq\{1,\cdots,l\}$	$O(Nl)$	$O(qN)$

2.2 HSAG Framework

Reddi et al. (2015) proposed an unified variance reduction framework (i.e., HSAG) to analyze the asynchronous stochastic gradient algorithm, and proved its linear convergence rate. The HSAG framework is presented in Algorithm 1. Because the iteration numbers of an epoch for SAGA and SVRG are different, the HSAG framework uses two different rules to update $\{\alpha_i^t\}_{i=1}^l$. Specifically, the SAGA and q-SAGA techniques correspond to the rule in the case of $i \in J$ in (5), and the SVRG technique corresponds to the rule in the case of $i \notin J$ in (5). Because the HSAG framework uses the above two different updating rules, the HSAG framework is not unified enough. In addition, the HSAG framework assumes that the iteration number for an epoch is fixed. Thus, the HSAG framework cannot incorporate the S2GD technique, where the iteration number for an epoch is a random variable obeying a distribution as discussed in Section 2.1. Thus, the HSAG framework is not general enough.

Because the HSAG framework is not unified and general enough as mentioned above, the analysis for the HSAG framework of (Reddi et al., 2015) is limited to the SVRG and SAGA techniques for the convex and smooth optimization problems. They did not analyze other important variance reduction techniques (e.g. S2GD and q-SAGA) and other important optimization problems, such as the convex optimization problem (2) with non-smooth regularization and the non-convex optimization problem (1) with cardinality constraint. In this paper, we will try to address these challenges.

Algorithm 1 HSAG($\{\alpha_i^t\}_{i=1}^l$, J, m, t)

```
Input: $\{\alpha_i^t\}_{i=1}^l$, the index set $J$, the iteration number for an epoch $m$, and the current iteration number $t$.
Output: $\{\alpha_i^{t+1}\}_{i=1}^l$.
for $i = 1, 2, \cdots, l$ do
    Update
    $$\alpha_i^{t+1} = \begin{cases} \mathbb{1}(i_t = i)\nabla f_i(x^t) + \mathbb{1}(i_t \neq i)\alpha_i^t & \text{if } i \in J \\ \mathbb{1}(m|t)\nabla f_i(x^t) + \mathbb{1}(m \not| t)\alpha_i^t & \text{if } i \notin J \end{cases} \quad (5)$$
end for
```

2.3 Unified q-Memorization Framework

As mentioned above, the HSAG framework is not unified and general enough. In this paper, we introduce a more unified and general variance reduction framework (i.e., q-memorization) which is originally proposed by Hofmann et al. (2015) to analyze the sequential stochastic gradient algorithm for the convex and smooth optimization problems.

To formulate the variance reduction techniques such as SVRG, SAGA and q-SAGA, Hofmann et al. (2015) proposed an unified q-memorization framework. In this paper, we present the unified q-memorization framework in Definition 1 by highlighting its two conditions. As mentioned above, S2GD cannot be included in the HSAG framework (Reddi et al., 2015). According to the two conditions in Definition 1, we show that S2GD can be viewed as a special case of the unified q-memorization framework which was not discussed in (Hofmann et al., 2015).

Definition 1 (Unified q-Memorization) *A stochastic gradient algorithm is satisfied with the unified q-memorization framework, if each iteration of the algorithm satisfies the following two conditions:*

1. *The algorithm selects a random index set $J \subseteq \{1, \cdots, l\}$ in each iteration to update α_i as*

$$\alpha_i^{t+1} = \begin{cases} \nabla f_i(x^t) & \textit{if } \ i \in J \\ \alpha_i^t & \textit{otherwise} \end{cases} \tag{6}$$

2. *Any α_i has a same probability $\frac{q}{l}$ to be updated. It means that, $\forall i_1, i_2$, and $i_1 \neq i_2$, we have that $\sum_{J \ni i_1} \boldsymbol{P}\{J\} = \sum_{J \ni i_2} \boldsymbol{P}\{J\} \overset{\text{def}}{=} \frac{q}{l}$.*

It is easy to verify that SAGA and q-SAGA satisfy the two conditions of Definition 1. In the following, we give a brief explanation to show how SVRG and S2GD satisfy the two conditions of Definition 1.

1. **SVRG**: As mentioned in Section 2.1, standard SVRG updates all α_i after m iterations, where m is fixed. It is not compatible with the 1st condition of Definition 1 (i.e., randomly select a index set J). Same with (Hofmann et al., 2015), we give a variant of SVRG to adapt Definition 1. Specifically, fixing $q > 0$ and generating a random variables r from a uniform distribution on the interval $[0, 1]$. If $r < \frac{q}{l}$, we choose $J = \{1, \cdots, l\}$. Otherwise, we choose $J = \emptyset$. Thus, any α_i has a same probability of $\frac{q}{l}$ of being updated. For simplicty, we also use SVRG to refer to the q-memorization variant of SVRG in this paper. That means SVRG standards for the q-memorization variant of SVRG if SVRG is used with q-memorization. Otherwise, SVRG standards for the classical variant.

2. **S2GD**: As mentioned in Section 2.1, standard S2GD updates all α_i after t iterations, where $t \in \{1, \cdots, m\}$ and $\mathbf{P}(t) = \frac{(1-\nu\gamma)^{m-t}}{\beta}$. It is also not compatible with the randomness of the index set J in the 1st condition of Definition 1. Thus, we give a variant of S2GD. Specifically, fixing $\widehat{q} > 0$ and generating a random variables r from a uniform distribution on the interval $[0, 1]$. If $r < \frac{\widehat{q}t}{ml}$, we choose $J = \{1, \cdots, l\}$. Otherwise, we choose $J = \emptyset$. Thus, any α_i has a same probability of $\frac{q}{l} \overset{\text{def}}{=} \frac{\widehat{q}\sum_{t=1}^{m} t(1-\nu\gamma)^{m-t}}{\beta ml} = \frac{\widehat{q}}{\beta ml}\left(\frac{m-1}{\nu\gamma} - \frac{(1-\nu\gamma)\big(1-(1-\nu\gamma)^{m-1}\big)}{\nu^2\gamma^2}\right)$ of being updated.

3. Asynchronous Stochastic Gradient Hard Thresholding Algorithm with Generalized Variance Reduction

In this section, to solve the non-convex optimization problem (1) with cardinality constraint, we first propose our AsySGHT-qM algorithm based on the unified q-memorization framework. Then, we prove the linear convergence rate of AsySGHT-qM to an approximately global optimum.

3.1 AsySGHT-qM

Based on the unified q-memorization framework, we propose our AsySGHT-qM algorithm in this section. AsySGHT-qM is also designed for the parallel environment with shared memory, such as multi-core processors and GPU-accelerators, but it can also work in the parallel environment with distributed memory.

In the parallel environment with shared memory, all cores in CPU or GPU can read and write the vectors x and α_i in shared memory simultaneously without any lock. AsySGHT-qM is to parallelly and repeatedly read and update the vectors x and α_i in shared memory. Specifically, all cores repeat the following steps independently and concurrently without any lock:

1. **Read:** Read the vectors x and α_i from shared memory to local memory without reading lock. We use $\widehat{x}$ and $\widehat{\alpha}_i$ to denote their values respectively.

2. **Compute:** Randomly pick $i \in \{1, ..., l\}$ with equal probability, and locally compute $\widehat{v} = \nabla f_i(\widehat{x}) - \widehat{\alpha}_i + \frac{1}{l}\sum_{j=1}^{l} \widehat{\alpha}_j$.

3. **Update:** Update the vector x in shared memory as $x \leftarrow \mathcal{H}_k\left(\widehat{x} - \gamma\widehat{v}\right)$, and the vectors α_i in shared memory without writing lock.

Based on the three above steps, we present our AsySGHT-qM in Algorithm 2.

Algorithm 2 Generalized Variance Reduction Asynchronous Stochastic Gradient Hard Thresholding Algorithm (AsySGHT-qM)

```
Input: Steplength $\gamma$, parameter $q$.
Initialize shared variables $x$, and $\{\alpha_i\}_{i=1}^{l}$.
keep doing in parallel
    Inconsistent read of $x$ and $\alpha_i$ as $\widehat{x}$ and $\widehat{\alpha}_i$ respectively.
    Randomly pick $i \in \{1, ..., l\}$ with equal probability.
    Compute $\widehat{v} = \nabla f_i(\widehat{x}) - \widehat{\alpha}_i + \frac{1}{l}\sum_{j=1}^{l} \widehat{\alpha}_j$.
    Update $x \leftarrow \mathcal{H}_k\left(\widehat{x} - \gamma\widehat{v}\right)$ using coordinate atomic operation.
    Update $\{\alpha_i\}_{i=1}^{l}$ using coordinate-wise atomic operation based on the $q$-memorization framework.
end parallel loop
Output: $x$.
```

3.2 Convergence Analysis of AsySGHT-qM

In this section, we first give several preliminaries, and then prove the convergence rate of AsySGHT-qM.

3.2.1 Preliminaries

In this subsection, we introduce the conditions of restricted strong smoothness (RSS) for the functions $f_i(x)$ and the restricted strong convexity (RSC) for the function $F(x)$, which are widely used in the analysis for the non-convex optimization problem (2) with cardinality constraint (Li et al., 2016; Shen and Li, 2018; Nguyen et al., 2017). We also discuss the global time counter t, coordinate-wise atomic write operation and x^t which are important for analyzing the convergence of asynchronous parallel algorithms.

1. **RSS:** For the differentiable functions $f_i(x)$, we assume that the function $f_i(x)$ is with the restricted strong smoothness as follows.

   Assumption 1 (RSS) *The differentiable function $f_i(x)$ is restricted ρ_s^+-strongly smooth at sparsity level s if there exists a generic constant $\rho_s^+ > 0$ such that for any x and y with $\|x-y\|_0 \leq s$, we have*

   $$f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x-y \rangle + \frac{\rho_s^+}{2} \|x-y\|^2 \tag{7}$$

2. **RSC:** For the convex functions $F(x)$, we assume that the function $F(x)$ is with the restricted strong convexity as follows.

   Assumption 2 (RSC) *The convex function $F(x)$ restricted ρ_s^--strongly convex at sparsity level s if there exists a generic constant $\rho_s^- > 0$ such that for any x and y with $\|x-y\|_0 \leq s$, we have*

   $$F(x) \geq F(y) + \langle \nabla F(y), x-y \rangle + \frac{\rho_s^-}{2} \|x-y\|^2 \tag{8}$$

3. **Global Time Counter t:** As discussed in (Leblond et al., 2017), formalizing the meaning of x_t and $\widehat{x}_t$ highlights a subtle but important difficulty arising when analyzing randomized parallel algorithms: what is the meaning of t? Thus, the global time counter t plays an important role in the convergence rate analyses of AsySGHT-qM. In this paper, we follow the strategy of "after read" labeling proposed in (Leblond et al., 2017), in which we update our iterate counter t as each core finishes reading the parameters x and α_i. This means that $\widehat{x}_t$ is the $(t+1)$th fully completed read. The advantage of this approach is that it guarantees both that the i_t are uniformly distributed and that i_t and $\widehat{x}_t$ are independent.

4. **Coordinate-Wise Atomic Write Operation:** Because AsySGHT-qM does not use any locks in the write operation. Thus, while x in shared memory is currently updated by a thread as shown in the line 7 of Algorithm 2, x may be overwritten by other threads. The phenomenon of overwriting is also called as write-write conflict which means that the uncommitted data is overwritten by other interleaved execution of transactions. In this paper we use compare-and-swap to implement coordinate-wise atomic write operation which can avoid the phenomenon of overwriting and is used in Leblond et al. (2017); Mania et al. (2017); Pedregosa et al. (2017). Note that hardware provides write operations such that they will be successfully recorded in shared memory at some point. Atomic write on float or doubles can be strictly enforced through compare-and-swap operations.

5. x^t: Because x^t updated in shared memory may be inconsistent with the ideal one computed in the local memory due to the coordinate-wise writing which will make the convergence analysis more difficult. To address this issue, we only consider the ideal ones x^t in the analysis which is defined as

$$x^{t+1} \leftarrow \mathcal{H}_k\left(\widehat{x}^t - \gamma \widehat{v}^t\right) \tag{9}$$

It is noted that, if T is the last iteration of AsySGHT-qM, x^T is exactly what is stored in shared memory. Thus, although the ideal ones x^t is considered in the analysis, we can still build the convergence rate for AsySGHT-qM. Similarly, we define $\alpha_i = \nabla f_i(x^u)$, where u is the last iteration number for updating α_i.

3.2.2 Convergence Analysis

Because AsySGHT-qM is an asynchronous parallel stochastic algorithm without any lock, the inconsistent reading could arise to the vectors x and $\{\alpha_i\}_{i=1}^l$ in shared memory, which makes the convergence analysis of AsySGHT-qM more challenging. To address this challenge, we provide the relationships between x and $\widehat{x}$, and between α_i and $\widehat{\alpha}_i$ as follows, which are also summarized in Fig. 1.

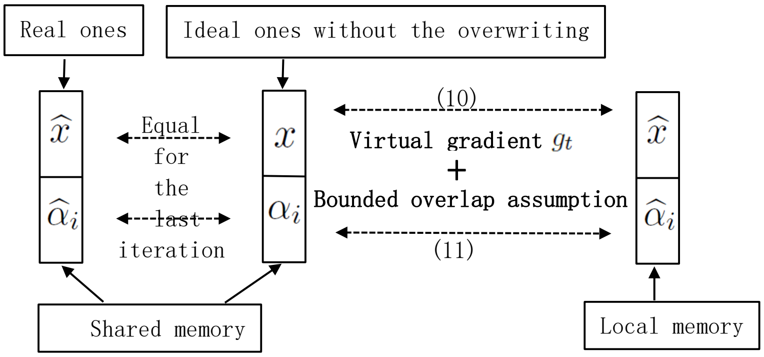


Figure 1: The relationships between x and $\widehat{x}$, and between α_i and $\widehat{\alpha}_i$.

Before building the relationships between the ideal x^t (α_i^t) and $\widehat{x}^t$ ($\widehat{\alpha}_i^t$), we first define a virtual gradient g_t for the t-th iteration of AsySGHT-qM, which will be used to build the relationships between the ideal x^t (α_i^t) and $\widehat{x}^t$ ($\widehat{\alpha}_i^t$). Let x^t be a sparse vector with $\|x^t\|_0 \leq k$ and define $\mathcal{I}^t = supp(x^t)$. For the t-th iteration of AsySGHT-qM, we define the virtual gradient g_t as

$$g_t = \frac{1}{\gamma}\left(x^t - x^{t+1}\right) \tag{10}$$

where $\mathcal{I}^{t+1}$ denotes the support set of x^{t+1}. $x^t - (x^t)_{\mathcal{I}^{t+1}}$ is a subset of nonzero elements of x^t. Note that the virtual gradient g_t is only used for the analysis, not computed in the implementation. We can provide an upper bound of $\|g_t\|^2$ based on $\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2$ (i.e., Lemma 2). The detailed proof of Lemma 2 is provided in Section 7.2.

Lemma 2 *If $\left\|(\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \leq c\left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2$, where $c \geq 0$ is constant, the virtual gradient g_t can be bounded by $\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2$ for any $\mathcal{I} \supseteq \mathcal{I}^t \cup \mathcal{I}^*$ with a coefficient $\varsigma = \frac{4n+4c}{\gamma^2} + 2$ as follows.*

$$\|g_t\|^2 \leq \varsigma \left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2 \tag{11}$$

In addition to the virtual gradient g_t, we also give the bounded overlap assumption (i.e., Assumption 3). In the asynchronous computation, the iteration t_1 and t_2 overlap if they are processed simultaneously. We make an assumption of bounded overlap (i.e., Assumption 3) which is widely used in the asynchronous parallel analysis (Mania et al., 2017; Lian et al., 2015; Huo and Huang, 2017; Zhao and Li, 2016).

Assumption 3 (Bounded overlap) *We assume that there exists a bound τ on the maximum difference of the numbers of iterations that overlap. The bound τ means that iteration t_1 cannot overlap with iteration t_2 for $t_1 \geq t_2 + \tau + 1$.*

Based on the bounded overlap assumption (Assumption 3) and the virtual gradient g_t, we have the relationship between the ideal x^t and $\widehat{x}^t$ as follows.

$$\widehat{x}^t - x^t = \gamma \sum_{u=(t-\tau)_+}^{t-1} S_u^t g_u \tag{12}$$

where S_u^t are $n \times n$ diagonal matrices with terms in $\{1, 0\}$ (please refer to (Leblond et al., 2017) for more details). According to the the definitions of global time counter t and x^t as mentioned previously, we know that 0 denotes that every update in $\widehat{x}^t$ is already in x^t. Conversely, 1 denotes that some updates might be late. $\widehat{x}^t$ may be lacking some updates from the "past" in some sense, whereas according to our rule of global time counter defined by the strategy of "after read" labeling proposed in (Leblond et al., 2017), it cannot contain updates from the "future". Based on (12), we have the relationship between the ideal α_i^t and $\widehat{\alpha}_i^t$ as follows if $\|x^{t-1} - \widehat{x}^{t-1}\|_0 \leq s$.

$$\begin{aligned} \left\|\alpha_i^t - \widehat{\alpha}_i^t\right\|^2 = \left\|\nabla f_i(x^{t-1}) - \nabla f_i(\widehat{x}^{t-1})\right\|^2 \leq (\rho_s^+)^2 \left\|x^{t-1} - \widehat{x}^{t-1}\right\|^2 \\ = \quad (\rho_s^+\gamma)^2 \left\|\sum_{k=(t-\tau-1)_+}^{t-2} S_k^{t-1} g_k\right\|^2 \end{aligned} \tag{13}$$

Based on the relationship between the ideal x^t and $\widehat{x}^t$ as (12), and the relationship between the ideal α_i^t and $\widehat{\alpha}_i^t$ as (13), we can prove that the convergence rate of AsySGHT-qM before reaching the inherent error induced by GHT-style methods is geometric as shown in Theorem 3. The detailed proofs are provided in Section 7.2. Corollary 6 provides the convergence result of sequential version of AsySGHT-qM (i.e., $\tau = 0$, denoted by SGHT-qM).

Theorem 3 *Let x^* denote any optimal sparse solution of the problem (1) with $\|x^*\|_0 \leq k^*$. We define $\mathcal{I}^* = supp(x^*)$ which denotes the support of x^*. Assume that the function F satisfies the RSC condition and the functions f_i satisfy the RSS condition with*

$s = 2k + k^*$. *Define* $\widetilde{\mathcal{I}} = supp(\mathcal{H}_k(\nabla F(x^*)) \cup supp(x^*))$. *Let* $\alpha = 1 + \frac{2\sqrt{k^*}}{k-k^*}$, $\beta \in (0,1)$, $\varrho = \alpha(1+\beta)\left(1 - \gamma\frac{\rho_s^-}{2}\right)$, $\max\{\varrho, 1-\frac{q}{l}\} < \rho < 1$, $\nu = \frac{2}{1 - \frac{6\varsigma\tau^2(\rho_s^+)^2\gamma^2}{\rho^{\tau+1}} - \frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau^3}{l\rho^{2\tau+2}}}$, $\omega = \alpha\left((1+\frac{1}{\beta})\gamma^2\tau\varsigma + 2(1+\beta)\varsigma\tau\rho_s^+\gamma^3(1+3\rho_s^+\gamma)\right)$ *and* $\Gamma = \nu\omega\rho^{-(\tau+1)} + \frac{12\nu\alpha(1+\beta)\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\rho^{-(2\tau+2)}$, *where* ς *is define in Lemma 2. Suppose the nonnegative steplength parameter* γ *satisfies* $\gamma \leq \frac{1}{\sqrt{\frac{6\varsigma\tau^2(\rho_s^+)^2}{\rho^{\tau+1}} + \frac{12\varsigma q(\rho_s^+)^2\tau^3}{l\rho^{2\tau+2}}}}$ *and* $-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2 + \Gamma)\left(1 + \frac{\rho_s^+}{1 - \frac{1-\frac{q}{l}}{\rho}}\right) \leq 0$. *For AsySGHT-qM, under Assumptions 1, 2 and 3, we have that*

$$\begin{aligned} & \mathbb{E}F(x^t) - F(x^*) \qquad (14) \\ \leq \;& \rho^t \frac{(2\rho - \varrho)\|x_0 - x^*\|^2}{\left(1 - \frac{\varrho}{\rho}\right)(\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)} \\ & + \frac{3(2\alpha(1+\beta)\gamma^2 + \Gamma)}{(1-\rho)(1-\varrho)(\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2 \end{aligned}$$

Remark 4 *The convergence rate of Theorem 3 is built on the iteration number* t*, instead of the epoch number like in Theorem 4.1. of (Li et al., 2016) (the result for SVRG). Thus, although the parameter of* ρ *may be close to* 1*, we can still obtain a linear convergence rate* $\mathbb{E}F(x^t) - F(x^*) \leq (\frac{5}{6})^{t/m}\frac{(2\rho-\varrho)\|x_0-x^*\|^2}{\left(1-\frac{\varrho}{\rho}\right)(\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)} + \frac{3(2\alpha(1+\beta)\gamma^2+\Gamma)}{(1-\rho)(1-\varrho)(\alpha(1+\beta)\gamma-24\alpha(1+\beta)\gamma^2-12\Gamma)}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2$ *before reaching the inherent error, where* $m = \log_\rho(\frac{5}{6})$ *is the iteration number of each epoch, and* $\frac{3(2\alpha(1+\beta)\gamma^2+\Gamma)}{(1-\rho)(1-\varrho)(\alpha(1+\beta)\gamma-24\alpha(1+\beta)\gamma^2-12\Gamma)}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2$ *is the inherent error induced by GHT-style methods. To the best of our knowledge, there is still no work that has proven the convergence rate of asynchronous stochastic gradient hard thresholding algorithm with SAGA. Our result in Theorem (9) is the first one to provide the convergence rate of asynchronous stochastic gradient hard thresholding algorithm with SAGA because AsySGHT-qM covers the case of SAGA.*

Remark 5 *For the smooth convex problem* $\min_{x\in\mathbb{R}^n} F(x)$*, we have* $\nabla_{\widetilde{\mathcal{I}}}F(x^*) = \boldsymbol{0}$ *if* x^* *is a global minimizer of* $\min_{x\in\mathbb{R}^n} F(x)$*. However, the problem (1) considered in this paper is with a cardinality constraint. Thus,* x^* *is normally not a global minimizer of* $\min_{x\in\mathbb{R}^n} F(x)$*, and the second term on the R.H.S of (14) is nonzero, which is same with the results of the convergence rates in (Li et al., 2016; Shen and Li, 2018).*

Corollary 6 *Let* x^* *denote any optimal sparse solution of the problem (1) with* $\|x^*\|_0 \leq k^*$*. We define* $\mathcal{I}^* = supp(x^*)$ *which denotes the support of* x^**. Assume that the function* F *satisfies the RSC condition and the functions* f_i *satisfy the RSS condition with* $s = 2k + k^*$*. Define* $\widetilde{\mathcal{I}} = supp(\mathcal{H}_k(\nabla F(x^*)) \cup supp(x^*))$*. Let* $\alpha = 1 + \frac{2\sqrt{k^*}}{k-k^*}$, $\beta \in (0,1)$, $\varrho = \alpha(1+\beta)\left(1 - \gamma\frac{\rho_s^-}{2}\right)$, $\max\{\varrho, 1-\frac{q}{l}\} < \rho < 1$*, where* ς *is define in Lemma 2. Suppose the nonnegative steplength parameter* γ *satisfies* $\gamma \leq \frac{1}{24(1+\beta)\left(1+\frac{\rho_s^+}{1-\frac{1-\frac{q}{l}}{\rho}}\right)}$*. For SGHT-qM, under Assumptions 1 and 2, we have that*

$$\mathbb{E}F(x^t) - F(x^*) \qquad (15)$$

$$
\begin{aligned}
\leq \quad & \rho^t \frac{(2\rho - \varrho)\left\|x_0 - x^*\right\|^2}{\alpha\left(1 - \frac{\varrho}{\rho}\right)\left((1+\beta)\gamma - 24(1+\beta)\gamma^2\right)} \\
& + \frac{6(1+\beta)\gamma}{(1-\rho)(1-\varrho)((1+\beta) - 24(1+\beta)\gamma)} \mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}} F(x^*)\right\|^2
\end{aligned}
$$

4. Asynchronous Stochastic Proximal Gradient Algorithm with Generalized Variance Reduction

As mentioned above, AsySGHT-qM is designed to solve the non-convex optimization problem (1) with cardinality constraint. In this section, to solve the convex optimization problem (2) with non-smooth regularization, we first propose our AsySPG-qM algorithm based on the unified q-memorization framework. Then, we prove the linear convergence rate of our AsySPG-qM. Note that, our AsySPG-qM algorithm follows from (Pedregosa et al., 2017), effectively utilizes the sparsity of the data, and is much more efficient than the traditional asynchronous stochastic proximal gradient algorithms (Meng et al., 2017; Gu and Huo, 2018) for the sparse data.

4.1 Sparse Proximal Updating

In this section, we first discuss the necessity of the sparse proximal updating for the asynchronous parallel stochastic proximal gradient descent algorithm, and then present the sparse proximal updating which will be used in our AsySPG-qM algorithm.

Due to the fact that the sparse data widely exists in the real world applications, the gradient $\nabla f_i(x)$ on the i-th training sample would be sparse correspondingly. For example, if the function $f_i(x)$ (e.g. square loss, logistic loss) is with the form of $f_i'(\langle a_i, x\rangle)$, it is easily to derive that the gradient $\nabla f_i(x)$ has the same sparsity with the data a_i. Thus, how to leverage this sparsity is crucial for the efficiency of our AsySPG-qM algorithm.

To utilize the q-memorization variance reduction technique, we only need to read and update the blocks of coefficients that intersect with the support of the current partial gradient as shown in (4). Thus, some blocks might be read and updated more frequently than others, which leads to an unbalanced number of updates per block. Following the works of (Leblond et al., 2017), we define a block-wise reweighting matrix D_i on the the average gradient $\overline{\alpha}$ to counterbalance the frequency of updating each block. Specifically, the new sparse stochastic gradient with variance reduction technique can be formulated as

$$
v = \nabla f_i(x) - \alpha_i + D_i \overline{\alpha} \tag{16}
$$

It is time to give the definition of diagonal matrix D_i. First, we let $d_{\mathcal{G}'} = \frac{n}{n_{\mathcal{G}'}}$ if $n_{\mathcal{G}'} > 0$ and 0 otherwise, where $n_{\mathcal{G}'}$ is the number of times that $\mathcal{G}' \in \Phi_i$. Based on $d_{\mathcal{G}'}$, we define the diagonal matrix D_i as $[D_i]_{\mathcal{G}',\mathcal{G}'} = d_{\mathcal{G}'} I_{|\mathcal{G}'|}$ if $\mathcal{G}' \in \Phi_i$ and $0 \cdot I_{|\mathcal{G}'|}$ otherwise.

The traditional proximal gradient algorithm requires computing the proximal operator of $h(x)$ as $\text{Prox}_{\gamma h}(x') = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2\gamma}\left\|x - x'\right\|^2 + h(x)$, which involves a full pass on the coordinates. Obviously, operating on all coordinates is a bottleneck for designing a sparse proximal gradient algorithm. To address this issue, we replace $h(x)$ by a block-wise reweighted function $\varphi_i(x) = \sum_{\mathcal{G}' \in \Phi_i} d_{\mathcal{G}'} h_{\mathcal{G}'}(x)$, similar with the work of (Pedregosa et al., 2017). Thus, we compute the new sparse proximal operator $\text{Prox}_{\gamma\varphi_i}(x')$ instead of $\text{Prox}_{\gamma h}(x')$. Note that, it is easy to verify that $\mathbb{E}\varphi_i(x) = h(x)$.

4.2 AsySPG-qM

AsySPG-qM is designed for the parallel environment with shared memory, such as multi-core processors and GPU-accelerators, but it can also work in the parallel environment with distributed memory.

In parallel environment with shared memory, all cores in CPU or GPU can read and write the vectors x and the historical gradients $\{\alpha_i\}_{i=1}^{l}$ in the shared memory simultaneously without any lock. AsySPG-qM is to parallelly and repeatedly read and update the vectors x and the historical gradients $\{\alpha_i\}_{i=1}^{l}$ in shared memory. Specifically, all cores repeat the following steps independently and concurrently without any lock:

1. **Read:** Read the vectors x and α_i from shared memory to local memory without reading lock. We use $\widehat{x}$ and $\widehat{\alpha}_i$ to denote their values respectively.

2. **Compute:** Randomly pick $i \in \{1, ..., l\}$ with equal probability, and locally compute $[\widehat{v}]_{\Phi_i} = \nabla f_i([\widehat{x}]_{\Phi_i}) - [\widehat{\alpha}_i]_{\Phi_i} + D_i[\widehat{\overline{\alpha}}]_{\Phi_i}$, and $\Delta x_{\Phi_i} = [\widehat{x}]_{\Phi_i} - [\text{Prox}_{\varphi_i}([\widehat{x}]_{\Phi_i} - \gamma[\widehat{v}]_{\Phi_i})]_{\Phi_i}$.

3. **Update:** Update the vector x and the historical gradients $\{\alpha_i\}_{i=1}^{l}$ in shared memory without writing lock.

Based on the three above steps, we provide two versions (i.e., the analyzed version and the implementation version) of our AsySPG-qM in Algorithm 3.

4.3 Convergence Analysis of AsySPG-qM

In this section, we first give several preliminaries, and then prove the convergence rate of our AsySPG-qM.

4.3.1 Preliminaries

In this subsection, we introduce the Lipschitz smoothness for the function $f_i(x)$ and the strong convexity for the function $F(x)$, which are widely used in the optimization analysis (Mania et al., 2017; Lian et al., 2015; Huo and Huang, 2017; Zhao and Li, 2016).

1. **Lipschitz smooth:** For the smooth functions $f_i(x)$, we have the Lipschitz constant L for $\nabla f_i(x)$ as follows.

   Assumption 4 *L is the Lipschitz constant for $\nabla f_i(x)$ ($\forall i \in \{1, \cdots, l\}$) in (2). Thus, $\forall x$ and $\forall y$, L-Lipschitz smooth can be presented as*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\| \tag{17}$$

   Equivalently, L-Lipschitz smooth can also be written as the formulation (18).

$$f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \tag{18}$$

Algorithm 3 Asynchronous Stochastic Proximal Gradient Algorithm with Generalized Variance Reduction (AsySPG-qM) (**Left**: analyzed version. **Right**: implementation version)

```
Input: Steplength $\gamma$, parameter $q$.
Initialize shared variables $x$ and $\{\alpha_i\}_{i=1}^{l}$.
keep doing in parallel
    Inconsistent read of $x$ and $\alpha_i$ as $\widehat{x}$ and $\widehat{\alpha}_i$ respectively.
    Pick $i$ randomly from $l$ samples.
    Let $\Phi_i$ be the extended support of $\nabla f_i(x)$ in $\{\mathcal{G}_1, \cdots, \mathcal{G}_k\}$.
    $[\widehat{\overline{\alpha}}]_{\Phi_i} = \frac{1}{l}\sum_{k=1}^{l}[\widehat{\alpha}_k]_{\Phi_i}$.
    $[\widehat{v}]_{\Phi_i} = \nabla f_i([\widehat{x}]_{\Phi_i}) - [\widehat{\alpha}_i]_{\Phi_i} + D_i[\widehat{\overline{\alpha}}]_{\Phi_i}$.
    $\Delta x_{\Phi_i} = [\widehat{x}]_{\Phi_i} - [\mathrm{Prox}_{\varphi_i}([\widehat{x}]_{\Phi_i} - \gamma[\widehat{v}]_{\Phi_i})]_{\Phi_i}$.
    for $\mathcal{G}' \in \Phi_i$ do
        for $b \in \mathcal{G}'$ do
            $[x]_b \leftarrow [x]_b + [\Delta x]_b$.
        end for
    end for
    Pick a subset $J$ of the size of $q$ randomly from $l$ samples.
    Let $S_i$ be the support of $\nabla f_i(\widehat{x})$.
    for $i \in J$ do
        for $b \in S_i$ do
            $[\alpha_i]_b \leftarrow [\nabla f_i(\widehat{x})]_b$.
        end for
    end for
end parallel loop
Output: $x$.
```

```
Input: Steplength $\gamma$, parameter $q$.
Initialize shared variables $x$ and $\{\alpha_i\}_{i=1}^{l}$.
keep doing in parallel
    Inconsistent read of $x$ and $\alpha_i$ as $\widehat{x}$ and $\widehat{\alpha}_i$ respectively.
    Pick $i$ randomly from $l$ samples.
    Let $\Phi_i$ be the extended support of $\nabla f_i(x)$ in $\{\mathcal{G}_1, \cdots, \mathcal{G}_k\}$.
    $[\widehat{\overline{\alpha}}]_{\Phi_i}$=inconsistent read of $\overline{\alpha}$ on $\Phi_i$.
    $[\widehat{v}]_{\Phi_i} = \nabla f_i([\widehat{x}]_{\Phi_i}) - [\widehat{\alpha}_i]_{\Phi_i} + D_i[\widehat{\overline{\alpha}}]_{\Phi_i}$.
    $\Delta x_{\Phi_i} = [\widehat{x}]_{\Phi_i} - [\mathrm{Prox}_{\varphi_i}([\widehat{x}]_{\Phi_i} - \gamma[\widehat{v}]_{\Phi_i})]_{\Phi_i}$.
    for $\mathcal{G}' \in \Phi_i$ do
        for $b \in \mathcal{G}'$ do
            $[x]_b \leftarrow [x]_b + [\Delta x]_b$.
        end for
    end for
    Pick a subset $J$ of the size of $q$ randomly from $l$ samples.
    Let $S_i$ be the support of $\nabla f_i(\widehat{x})$.
    for $i \in J$ do
        for $b \in S_i$ do
            $[\alpha_i]_b \leftarrow [\nabla f_i(\widehat{x})]_b$.
            $[\overline{\alpha}]_b \leftarrow [\overline{\alpha}]_b + \frac{1}{l}([\nabla f_i(\widehat{x})]_b - [\widehat{\alpha}_i]_b)$.
        end for
    end for
end parallel loop
Output: $x$.
```

2. **Strong convexity:** For the convex functions $f(x)$, we assume that the function $f(x)$ is μ-strongly convex as follows.

   Assumption 5 (Strong convexity) *The convex function $F(x)$ has the condition of strong convexity with parameter $\mu > 0$, which means that, $\forall x$ and $\forall y$, we have*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2 \tag{19}$$

We use the notation $\kappa = \frac{L}{\mu}$ to denote the condition number for the L-Lipschitz smooth and μ-strongly convex function. In addition, we also follow the global time counter t and the coordinate-wise atomic write operation as described in Section 3.2.1.

4.3.2 Convergence Analysis

As mentioned above, AsySPG-qM is an asynchronous parallel stochastic algorithm without any lock, the inconsistent reading could arise to the vectors x and $\{\alpha_i\}_{i=1}^l$ in shared memory. In other words, some components of x (α_i) in shared memory are different to the ones of $\widehat{x}$ ($\widehat{\alpha}_i$) in the local memory, which makes the convergence analysis of AsySPG-qM more challenging. To address this challenge, we build the relationships between x and $\widehat{x}$, and between α_i and $\widehat{\alpha}_i$ similar to AsySGHT-qM.

First, we can define x and α_i as the ideal x and α_i similar to AsySGHT-qM. Then, we define a virtual gradient g_t for the t-th iteration of AsySPG-qM, which will be used to build the relationships between the ideal x^t (α_i^t) and $\widehat{x}^t$ ($\widehat{\alpha}_i^t$). Specifically, the virtual gradient g_t is defined as

$$g_t = \frac{1}{\gamma}\left(x^t - \text{Prox}_{\gamma h}\left(x^t - \gamma\widehat{v}^{t+1}\right)\right) \tag{20}$$

Because $x^{t+1} = \arg\min_x \frac{1}{2\gamma}\|x - (x^t - \gamma\widehat{v}^{t+1})\| + h(x)$, based on the optimality condition, we have that $\frac{1}{\gamma}\left(-x^{t+1} + (x^t - \gamma\widehat{v}^{t+1})\right) \in \partial h(\widehat{x}^{t+1})$. Thus, we have

$$g_t = \frac{1}{\gamma}\left(x^t - x^{t+1}\right) = \widehat{v}^{t+1} + \xi^{t+1} \tag{21}$$

where $\xi^{t+1} \in \partial h(\widehat{x}^{t+1})$. Note that the virtual gradient g_t is only used for the analysis, not computed in the implementation.

Based on (12), we have the relationship between the ideal α_i^t and $\widehat{\alpha}_i^t$ as follows.

$$\left\|\alpha_i^t - \widehat{\alpha}_i^t\right\|^2 = \left\|\nabla f_i(x^{t-1}) - \nabla f_i(\widehat{x}^{t-1})\right\|^2 \leq L^2\left\|x^{t-1} - \widehat{x}^{t-1}\right\|^2 = L^2\gamma^2\left\|\sum_{k=(t-\tau-1)_+}^{t-2} S_k^{t-1} g_k\right\|^2 \tag{22}$$

Based on the relationship between x and $\widehat{x}$ as (12), and the relationship between α_i and $\widehat{\alpha}_i$ as (22), we can prove the linear convergence of AsySPG-qM in Theorem 9. Before presenting Theorem 9, we first give the definition of block sparsity in Definition 7 which will be used in Theorem 9.

Definition 7 (Block Sparsity) *Let $\triangle = \max_{j=1,\ldots,k} \frac{|\{i:\mathcal{G}_j \in \Phi_i\}|}{l}$ which is the normalized maximum number of data points that share a specfic block in their extend support. We call $\triangle$ as the block sparsity of the training set.*

Remark 8 *According to Definition 7, we have that $\frac{1}{l} \leq \triangle \leq 1$. Specifically, if a block appears in all Φ_i, we have that $\triangle = 1$. If there are not two Φ_i sharing a same block, we have that $\triangle = 1$.*

Theorem 9 *Suppose $\tau \leq \frac{1}{10\sqrt{\triangle}}$, the nonnegative steplength parameter $\gamma = \frac{a}{L}$ with $a \leq \min\left\{\frac{3}{248}, \frac{2\kappa}{11\tau}\right\}$, and $q \leq 5\sqrt{l}$. For AsySPG-qM, under Assumptions 3, 4 and 5, we have that $\mathbb{E}\|x^t - x^*\| \leq (1-\rho(a))^t\widetilde{C}_0$, where $\rho(a) = \frac{1}{5}\min\{\frac{q}{l}, \frac{a}{\kappa}\}$, and $\widetilde{C}_0$ is a constant independent of t.*

Theorem 9 can be proved in a similar way than the proof of ProxASAGA in (Pedregosa et al., 2017). We provide a brief proof to Theorem 9 in Section 7.2. Next, we provide a variant of Theorem 9 for a SAGA version of AsySPG-qM (*i.e.*, $q = 1$) in Corollary 10.

Corollary 10 *Suppose $\tau \leq \frac{1}{10\sqrt{\triangle}}$, the nonnegative steplength parameter $\gamma = \frac{a}{L}$ with $a \leq \min \frac{1}{36} \left\{1, \frac{6\kappa}{\tau}\right\}$. For SAGA version of AsySPG-qM (i.e., $q = 1$), under Assumptions 3, 4 and 5, we have that $\mathbb{E}\|x^t - x^*\| \leq (1 - \rho(a))^t \widetilde{C}_0$, where $\rho(a) = \frac{1}{5}\min\{\frac{1}{l}, \frac{a}{\kappa}\}$, and $\widetilde{C}_0$ is a constant independent of t.*

The conclustion of Corollary 10 is exactly same to the conclusion of ProxASAGA in (Pedregosa et al., 2017).

5. Asynchronous Stochastic Gradient Descent Algorithm with Generalized Variance Reduction

In this section, to solve the general non-convex smooth optimization problem (3), we first propose our AsySGD-qM algorithm based on the unified q-memorization framework. Then, we prove the convergence rates of our AsySGD-qM.

5.1 AsySGD-qM

Similar to AsySGHT-qM and AsySPG-qM, AsySGD-qM is to parallelly and repeatedly read and update the vectors x and the historical gradients $\{\alpha_i\}_{i=1}^l$ in shared memory. Specifically, all cores repeat the following steps independently and concurrently without any lock:

1. **Read:** Read the vectors x and α_i from shared memory to local memory without reading lock. We use $\widehat{x}$ and $\widehat{\alpha}_i$ to denote their values respectively.

2. **Compute:** Randomly pick $i \in \{1, ..., l\}$ with equal probability, and locally compute $\widehat{v} = \nabla f_i(\widehat{x}) - \widehat{\alpha}_i + \frac{1}{l}\sum_{j=1}^{l} \widehat{\alpha}_j$.

3. **Update:** Update the vector x and the historical gradients $\{\alpha_i\}_{i=1}^l$ in shared memory without writing lock.

Based on the above three steps, we present our AsySGD-qM in Algorithm 4.

5.2 Convergence Analysis of AsySGD-qM

In this section, we first provide the sublinear convergence rate of AsySGD-qM, then provide the linear convergence rate of AsySGD-qM under different assumptions. In the analysis of this section, we follow the global time counter t, the coordinate-wise atomic write operation (please refer to Section 3.2.1), the relationship between x and $\widehat{x}$ as (12) and the relationship between α_i and $\widehat{\alpha}_i$ as (22) where g is replaced by $\widehat{v}$.

5.2.1 Sublinear Convergence Rate

Under Assumptions 3 and 4, we can prove the sublinear convergence of AsySGD-qM in Theorem 11. The detailed proofs are provided in Section 7.2.

Algorithm 4 Generalized Variance Reduction Asynchronous Stochastic Gradient Descent Algorithm (AsySGD-qM)

```
Input: Steplength $\gamma$, parameter $q$.
Initialize shared variables $x$, and $\{\alpha_i\}_{i=1}^l$.
keep doing in parallel
    Inconsistent read of $x$ and $\alpha_i$ as $\widehat{x}$ and $\widehat{\alpha}_i$ respectively.
    Randomly pick $i \in \{1, ..., l\}$ with equal probability.
    Compute $\widehat{v} = \nabla f_i(\widehat{x}) - \widehat{\alpha}_i + \frac{1}{l}\sum_{j=1}^l \widehat{\alpha}_j$.
    Update $x \leftarrow x - \gamma \widehat{v}$ using coordinate atomic operation.
    Update $\{\alpha_i\}_{i=1}^l$ using coordinate-wise atomic operation based on the $q$-memorization
        framework.
end parallel loop
Output: (in theory): $x^s$, $s$ is randomly chosen from $\{0, \ldots, T\}$, where both of the total
    iteration number $T$ and the vector $x^t$ are defined in Section 3.2.1.
Output: (in practice): $x$.
```

Theorem 11 *Under Assumptions 3, 4 and* $\gamma < \frac{1}{\sqrt{6\tau^2L^2 + \frac{12qL^2\tau^2}{l}}}$. *Let* $a = \frac{2}{1-6\tau^2L^2\gamma^2 - \frac{12qL^2\gamma^2\tau^3}{l}}$, $c_T = 0$, $c_t = aL^3\gamma^2 + 2aL^2c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + a\tau^2L^4\gamma^3 + c_{t+1}\frac{l-1}{l}(1+\gamma\beta)$, $\Gamma_t = \frac{\gamma}{2} - 2a\left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2L^2\gamma^3}{2}\right)$. *For AsySGD-qM, we have that*

$$\mathbb{E}\|\nabla f(x^s)\|^2 \leq \frac{f(x^0) - f(x^*)}{T\min_{0\leq t\leq T-1}\Gamma_t} \tag{23}$$

where T is the total iteration number.

5.2.2 Linear Convergence Rate

We firstly give the assumption of α-Polyak-Lojasiewicz in Assumption 6.

Assumption 6 *Let $\alpha > 0$ and f be a differentiable function. The function f is α-Polyak-Lojasiewicz. Specifically, α-Polyak-Lojasiewicz can be presented as, for every x, we have*

$$\frac{1}{2}\|\nabla f(x)\| \geq \alpha(f(x) - f(x^*)) \tag{24}$$

Based on the assumption of α-Polyak-Lojasiewicz and Theorem 11, we can obtain the linear convergence rate for AsySGD-qM as follows.

Theorem 12 *Under Assumptions 3, 4, 6 and* $\gamma < \frac{1}{\sqrt{6\tau^2L^2 + \frac{12qL^2\tau^2}{l}}}$. *Let* $a = \frac{2}{1-6\tau^2L^2\gamma^2 - \frac{12qL^2\gamma^2\tau^3}{l}}$, $c_T = 0$, $c_t = aL^3\gamma^2 + 2aL^2c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + a\tau^2L^4\gamma^3 + c_{t+1}\frac{l-1}{l}(1+\gamma\beta)$, $\Gamma_t = \frac{\gamma}{2} - 2a\left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2L^2\gamma^3}{2}\right)$. *For AsySGD-qM, let* $T > \frac{\alpha}{\min_{0\leq t\leq T-1}\Gamma_t}$, *and* $\rho = \frac{\alpha}{T\min_{0\leq t\leq T-1}\Gamma_t} < 1$ *we have that*

$$\mathbb{E}f(x^s) - f(x^*) \leq \rho\left(f(x^0) - f(x^*)\right) \tag{25}$$

where T is the total iteration number.

Remark 13 *If we call AsySGD-qM for multiple times, we can obtain the multi-stage version of AsySGD-qM which has the linear convergence rate as proved in Theorem 12.*

6. Experiments

In this section, we first give the experimental setup, then present our experimental results and discussion.

6.1 Experimental Setup

6.1.1 Design of Experiments

In the experiments, we verify the fast convergence of our AsySGHT-qM, AsySPG-qM and AsySGD-qM through concrete realizations of asynchronous SVRG and SAGA. In the following, we give the design of experiments for the AsySGHT-qM, AsySPG-qM and AsySGD-qM algorithms respectively.

AsySGHT-qM: For the experiments to the AsySPG-qM algorithm, we consider the sparse logistic regression (26) with cardinality constraint for binary classification problem.

$$\min_x \frac{1}{l}\sum_{i=1}^{l} \log(1+e^{-b_i x^T a_i}) \quad s.t. \quad \|x\|_0 \leq k \tag{26}$$

where $a_i \in \mathbb{R}^n$ is the input of a sample, and $b_i \in \{+1,-1\}$ is the label of a sample. To verify the fast convergence of our AsySGHT-qM, we compare the following three asynchronous stochastic gradient hard thresholding algorithms.

1. AsySGHT: AsySGHT without any variance reduction technique (Nguyen et al., 2017).
2. AsySGHT-qM(SVRG): Our AsySGHT-qM with SVRG.
3. AsySGHT-qM(SAGA): Our AsySGHT-qM with SAGA.

Specifically, we observe the convergence of the objective function (26) w.r.t. the running time and iterations, respectively, for the three implementations of asynchronous stochastic gradient hard thresholding algorithms, to verify the fast convergence rate of our AsySGHT-qM. In addition, we test the speedup of our AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA), to show that asynchronous parallel computation can achieve a near-linear speedup.

AsySPG-qM: For the experiments to the AsySPG-qM algorithm, we consider the logistic regression (27) with l_1-norm and l_2-norm for binary classification problem.

$$\min_x \frac{1}{l}\sum_{i=1}^{l} \log(1+e^{-b_i x^T a_i}) + \lambda_1\|x\|_2^2 + \lambda_2\|x\|_1 \tag{27}$$

where $a_i \in \mathbb{R}^n$ is the input of a sample, $b_i \in \{+1,-1\}$ is the label of a sample, and λ_1 and λ_2 are two regularization parameters. For the implementation of AsySPG-qM with the dense norm $\|x\|_2^2$, we treat $\lambda_1\|x\|_2^2 + \lambda_2\|x\|_1$ as the function of $h(x)$. Thus, we can use the sparse proximal operator $\text{Prox}_{\gamma h}(x') = \text{argmin}_{x\in\mathbb{R}^n} \frac{1}{2\gamma}\left\|x - x'\right\|^2 + h(x)$ to update the solution. To verify the fast convergence of our AsySPG-qM, we compare the following three asynchronous stochastic proximal gradient algorithms.

1. AsySPG: AsySPG without any variance reduction technique.
2. AsySPG-qM(SVRG): Our AsySPG-qM with SVRG.
3. AsySPG-qM(SAGA): Our AsySPG-qM with SAGA, which is exactly same with the AsySPG with SAGA.

Specifically, we observe the convergence of the objective function of (27) w.r.t. the running time and iterations, respectively, for the three implementations of asynchronous stochastic proximal gradient algorithms, to verify the fast convergence rate of our AsySPG-qM. In addition, we test the speedup of our AsySPG-qM(SVRG) and AsySPG-qM(SAGA), to show that asynchronous parallel computation can achieve a near-linear speedup.

AsySGD-qM: For the experiments to the AsySGD-qM algorithm, we consider the non-convex correntropy induced loss (Feng et al., 2015; He et al., 2011) for robust regression.

$$\min_x \frac{1}{l} \sum_{i=1}^{l} \frac{\sigma^2}{2} \left(1 - e^{-\frac{(b_i - x^T a_i)^2}{\sigma^2}} \right) \tag{28}$$

where $a_i \in \mathbb{R}^n$ is the input of a sample, $b_i \in \mathbb{R}$ is the label of a sample. To verify the fast convergence of our AsySGD-qM, we compare the following four asynchronous stochastic proximal gradient algorithms.

1. AsySGD: AsySGD without any variance reduction technique.
2. AsySGD-qM(SVRG): Our AsySGD-qM with SVRG.
3. AsySGD-qM(SAGA): Our AsySGD-qM with SAGA, which is exactly same with the AsySGD with SAGA.

Specifically, we observe the convergence of the objective function of (28) w.r.t. the running time and iterations, respectively, for the three implementations of asynchronous stochastic gradient descent algorithms, to verify the fast convergence rate of our AsySGD-qM. In addition, we test the speedup of our AsySGD-qM(SVRG) and AsySGD-qM(SAGA), to show that asynchronous parallel computation can achieve a near-linear speedup.

6.1.2 Implementations

We implement our AsySGHT-qM, AsySPG-qM and AsySGD-qM (including SVRG and SAGA) using C++, where the shared memory parallel computation is handled via OpenMP (Chandra, 2001). Note that the SVRG implementations for AsySGHT-qM and AsySPG-qM follow the unified q-Memorization framework as mentioned in Definition 1, where q is set as 10. In the experiments, the steplength γ for all compared methods is selected from $\{10^2, 10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ according to the optimal value of objective funtion reached in a fixed number of iterations, and the number of iterations for each epoch is the number of training samples. For the experiments of AsySPG-qM, we fix the parameter of λ_1 to 10^{-6}. Our experiments are performed on a 32-core two-socket Intel Xeon E5-2699 machine where each socket has 16 cores.

6.1.3 Datasets

Table 3 summarizes the six large-scale real-world binary classification datasets (i.e., the A1a, Covtype, Phishing, Criteo, Kdd2012, and URL datasets) used in our experiments. They are from LIBSVM website(Chang and Lin, 2011)[2].

Table 3: Summary of large-scale real-world binary classification dasetsets used in the experiments.

Dataset	Features	Samples	Sparsity
A1a	123	30,956	0.58%
Covtype	54	581,012	22.00%
Phishing	68	11,055	44.12%
Criteo	1,000,000	45,840,617	0.004%
Kdd2012	54,686,452	149,639,105	0.00002%
URL	3,231,961	2,396,130	0.004%

6.2 Experimental Results and Discussion

In this section, we present the experimental results and discussion to AsySGHT-qM, AsySPG-qM and AsySGD-qM respectively.

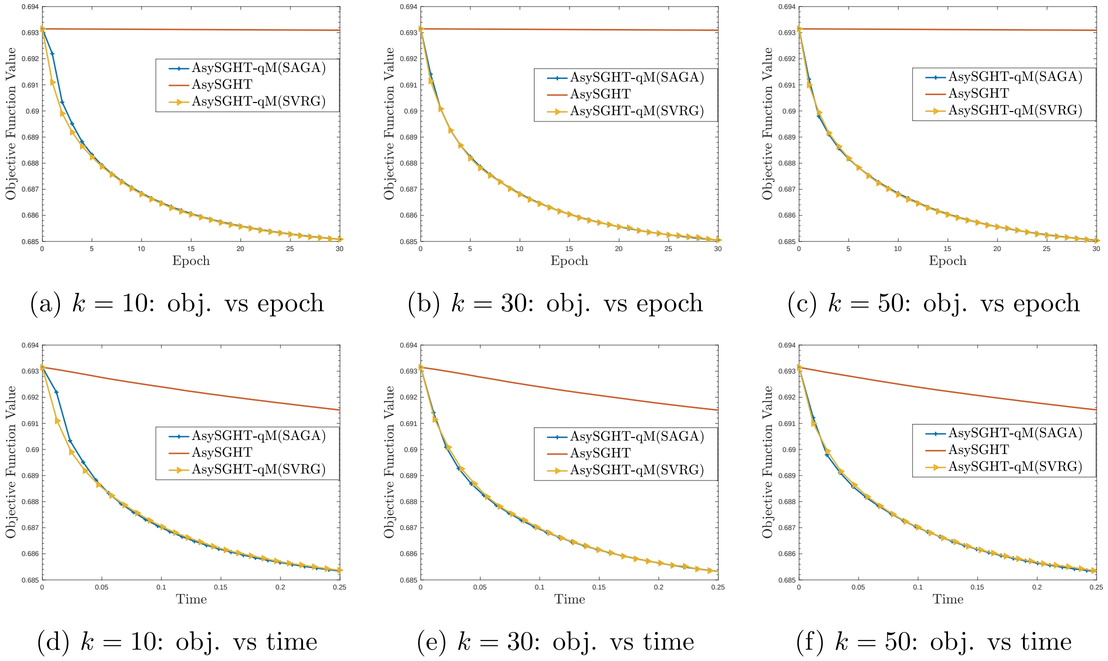
6.2.1 AsySGHT-qM

Figure 2 presents the convergence of objective values of AsySGHT, AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA) w.r.t. the epoch and running time on the Ala dataset, where k is set as 10, 30 and 50 respectively. Figures 3 and 4 present the convergence of objective values of AsySGHT, AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA) w.r.t. the epoch and running time on the Covtype and Phishing datasets, where k is set as 10, 20 and 30 respectively. Note that the convergence of AsySGHT, AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA) with different values of k (i.e., $k = 10, 20, 30$ or $k = 10, 30, 50$) are different slightly. The results show that AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA) have similar convergence on the Ala dataset. The reason is that the Ala dataset has high sparseness which makes the hard thresholding operators based on variance reduced gradients of SVRG and SAGA on different samples of Ala dataset have similar values. Finally, the results of Figures 2, 3 and 4 confirm the fast convergence of our AsySGHT-qM.

To estimate the scalability of our AsySGHT-qM, we perform AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA) on 1, 2, 4, 8 and 16 cores to observe the speedup. Figures 5 and 6 present the objective values of AsySGHT-qM(SVRG) and AsySGHT-qM(SAGA), respectively, w.r.t. the epoch and running time on the A1a, Covtype and Phishing datasets. The results show that AsySGHT-qM can have a near-linear speedup on a parallel system with shared memory. This is because that we do not use any lock in the implementation of AsySGHT-qM.

To verify the effect of q to the convergence rate of AsySGHT-qM in Theorem 3, we test the convergence of AsySGHT-qM (including AsySGHT-qM(SVRG) and AsySGHT-

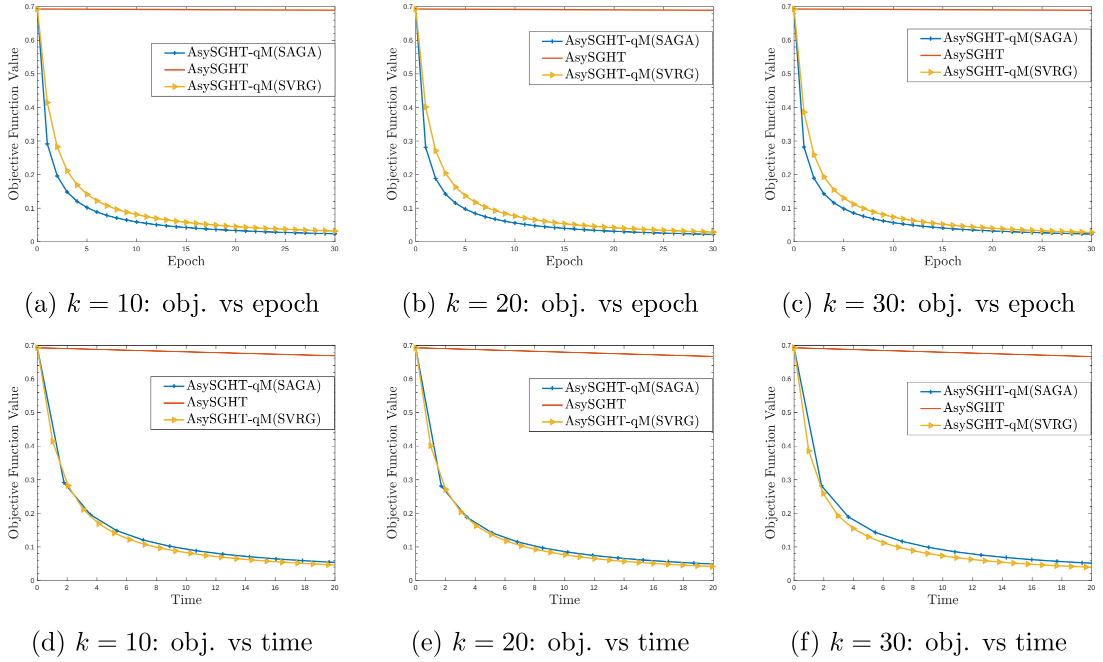
[2] The datasets are available at: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.



(a) $k = 10$: obj. vs epoch (b) $k = 30$: obj. vs epoch (c) $k = 50$: obj. vs epoch

(d) $k = 10$: obj. vs time (e) $k = 30$: obj. vs time (f) $k = 50$: obj. vs time

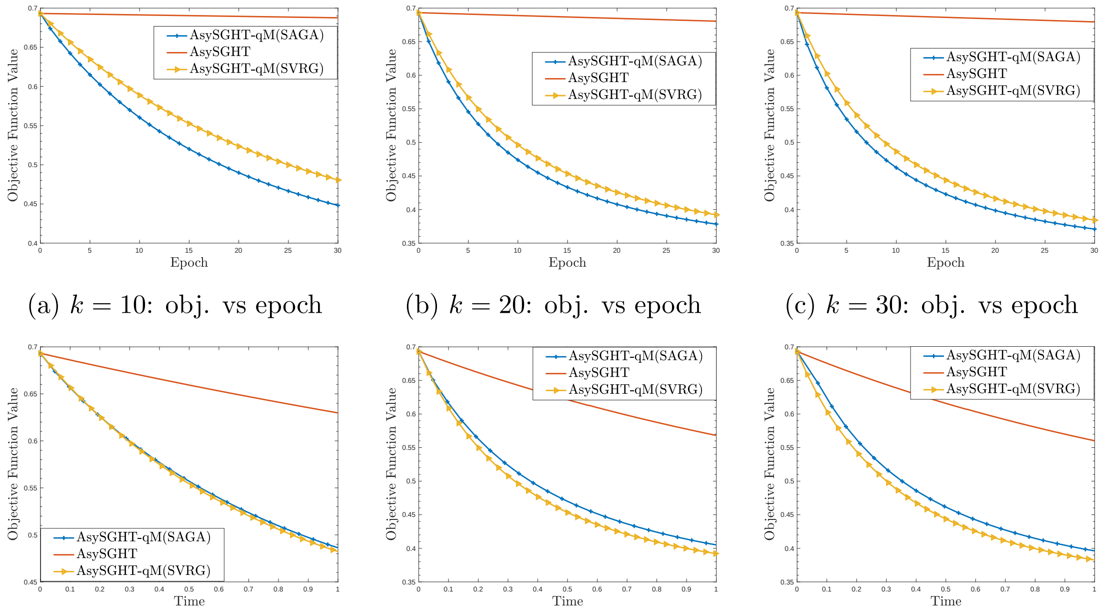
Figure 2: Convergence of AsySPG-qM and AsySPG on Ala dataset.



(a) $k = 10$: obj. vs epoch (b) $k = 20$: obj. vs epoch (c) $k = 30$: obj. vs epoch

(d) $k = 10$: obj. vs time (e) $k = 20$: obj. vs time (f) $k = 30$: obj. vs time

Figure 3: Convergence of of AsySPG-qM and AsySPG on Covtype dataset.



(a) $k = 10$: obj. vs epoch

(b) $k = 20$: obj. vs epoch

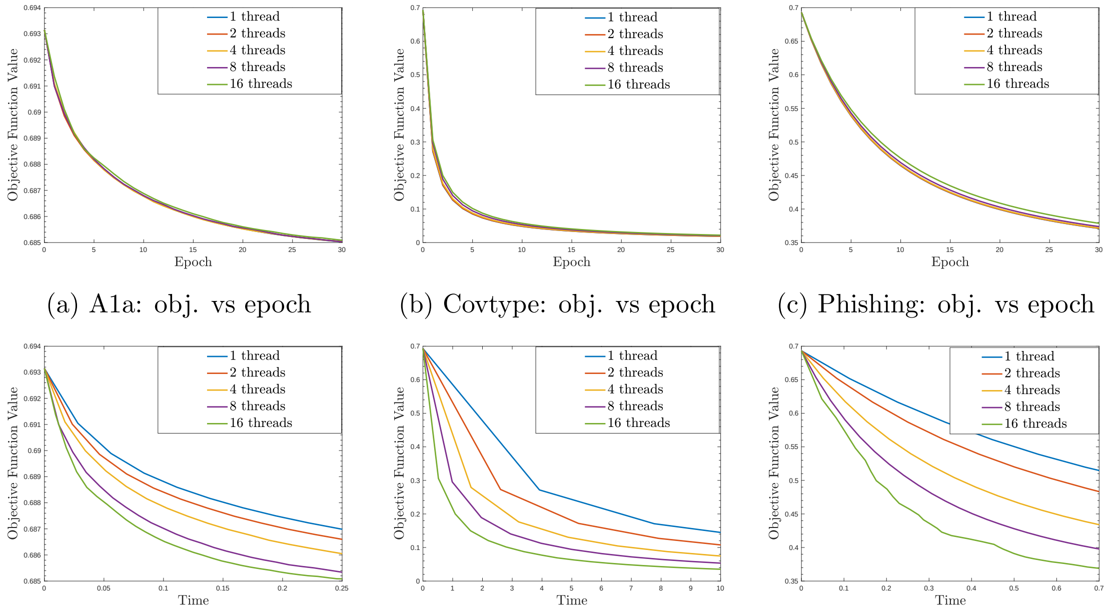
(c) $k = 30$: obj. vs epoch

(d) $k = 10$: obj. vs time

(e) $k = 20$: obj. vs time

(f) $k = 30$: obj. vs time

Figure 4: Convergence of of AsySPG-qM and AsySPG on Phishing dataset.



(a) A1a: obj. vs epoch

(b) Covtype: obj. vs epoch

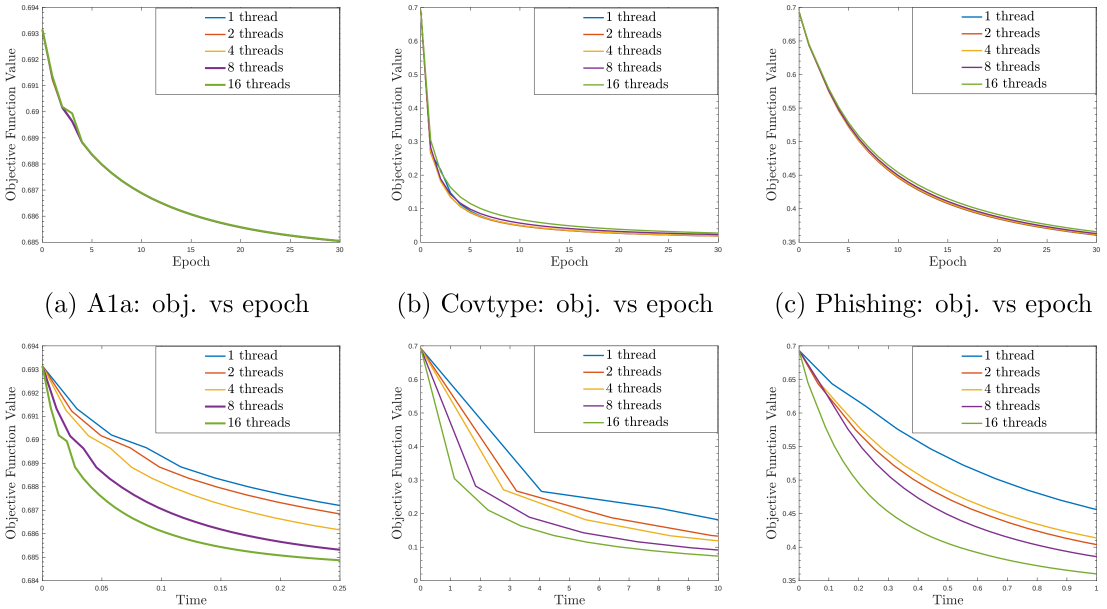
(c) Phishing: obj. vs epoch

(d) A1a: obj. vs time

(e) Covtype: obj. vs time

(f) Phishing: obj. vs time

Figure 5: Speedup results of AsySGHT-qM(SVRG).



(a) A1a: obj. vs epoch (b) Covtype: obj. vs epoch (c) Phishing: obj. vs epoch

(d) A1a: obj. vs time (e) Covtype: obj. vs time (f) Phishing: obj. vs time

Figure 6: Speedup results of AsySGHT-qM(SAGA).

qM(SAGA)) with q = 1, 5, 10 and 20 in Figure 7. The results show that AsySGHT-qM has a faster convergence rate if q is larger. The experimental results support the theoretical result in Theorem 3.

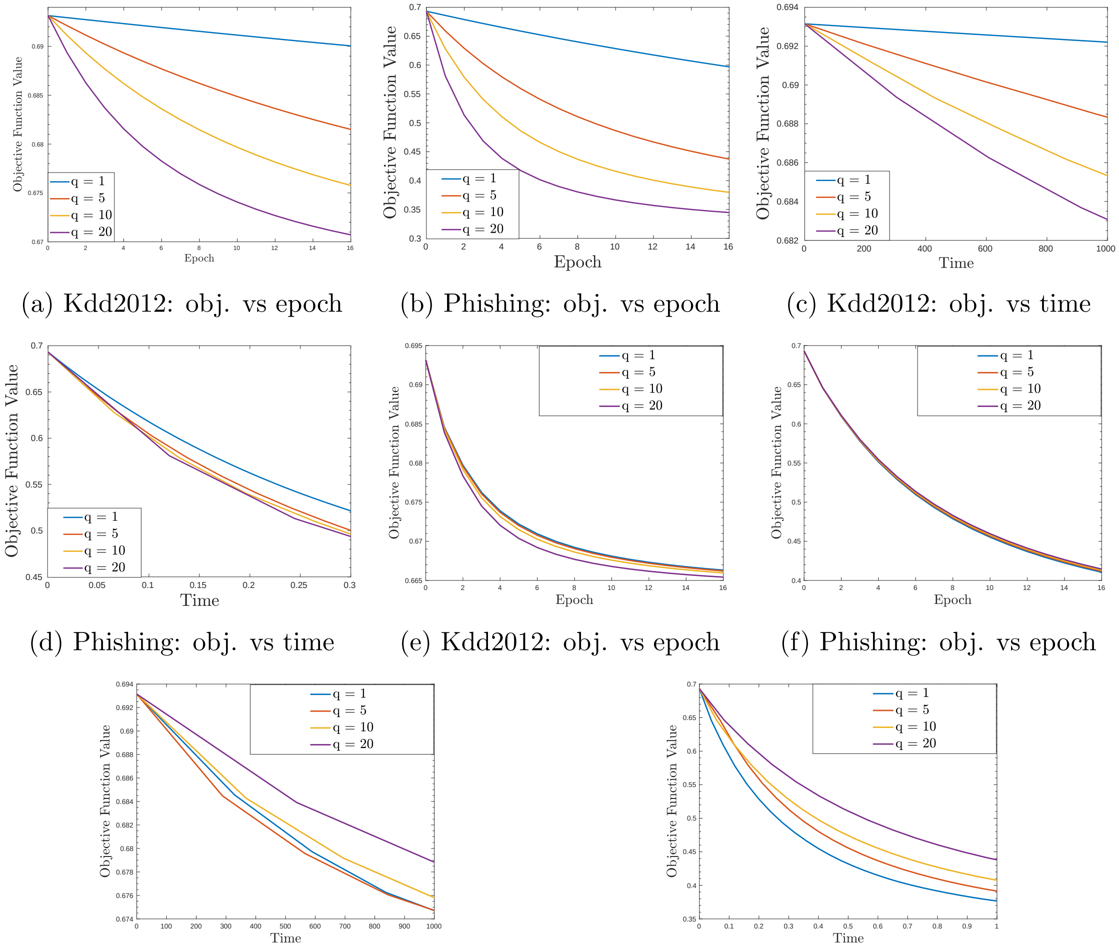
6.2.2 AsySPG-qM

Figure 8 presents the convergence of objective values of and AsySPG, AsySPG-qM(SVRG) and AsySPG-qM(SAGA) w.r.t. the epoch and running time on the Criteo dataset, where $\lambda_2 = 10^{-3}$, 10^{-4} and 10^{-5}. Figure 9 presents the convergence of objective values of and AsySPG, AsySPG-qM(SVRG) and AsySPG-qM(SAGA) w.r.t. the epoch and running time on the URL dataset, where $\lambda_2 = 10^{-3}$, 10^{-4} and 10^{-5}. The results confirm the fast convergence of our AsySPG-qM no matter what the value of λ_2 is..

To estimate the scalability of AsySPG-qM, we perform AsySPG-qM(SVRG) and AsySPG-qM(SAGA) on 1, 2, 4, 8 and 16 cores to observe the speedup. Figure 12 presents the objective values of AsySPG-qM(SVRG) and AsySPG-qM(SAGA) w.r.t. the epoch and running time on the Criteo and URL datasets, which show that AsySPG-qMR can have a near-linear speedup on a parallel system with shared memory. Similarly to AsySGHT-qM, this is because that we do not use any lock in the implementation of AsySPG-qM.

6.2.3 AsySGD-qM

Figure 11 presents the convergence of objective values of and AsySGD, AsySGD-qM(SVRG) and AsySGD-qM(SAGA) w.r.t. the epoch and running time on the A1a, Covtype, Criteo and URL dataset. The results confirm the fast convergence of our AsySGD-qM.



(a) Kdd2012: obj. vs epoch

(b) Phishing: obj. vs epoch

(c) Kdd2012: obj. vs time

(d) Phishing: obj. vs time

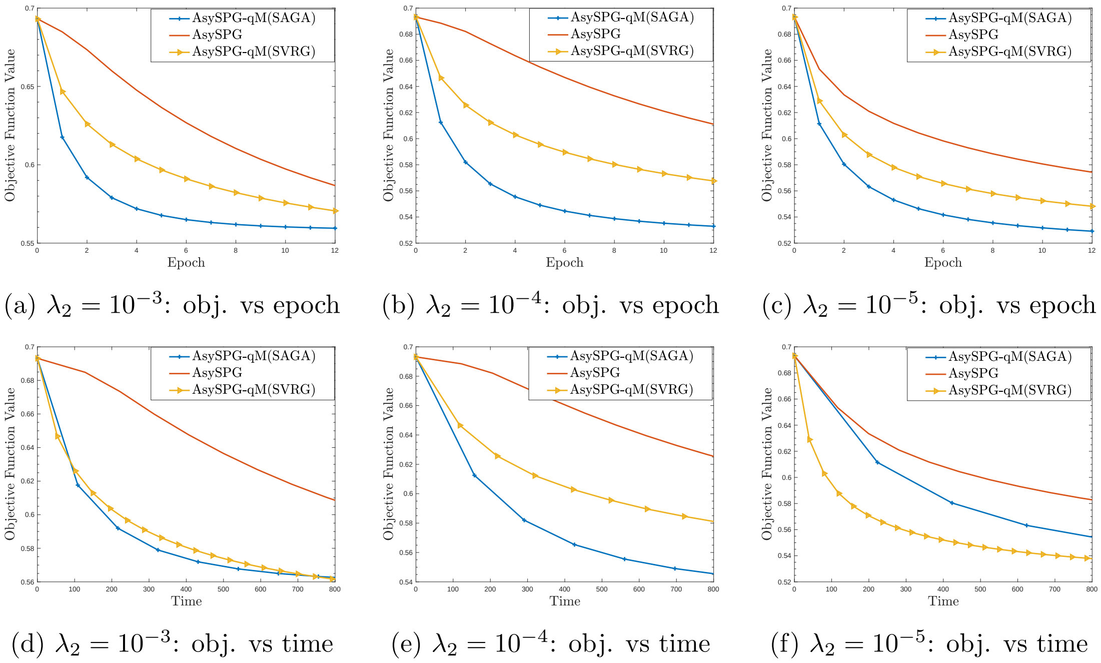
(e) Kdd2012: obj. vs epoch

(f) Phishing: obj. vs epoch

(g) Kdd2012: obj. vs time

(h) Phishing: obj. vs time

Figure 7: Convergence results of AsySGHT-qM with different values of q. (a)-(d) AsySGHT-qM(SVRG). (e)-(h) AsySGHT-qM(SAGA).



(a) $\lambda_2 = 10^{-3}$: obj. vs epoch

(b) $\lambda_2 = 10^{-4}$: obj. vs epoch

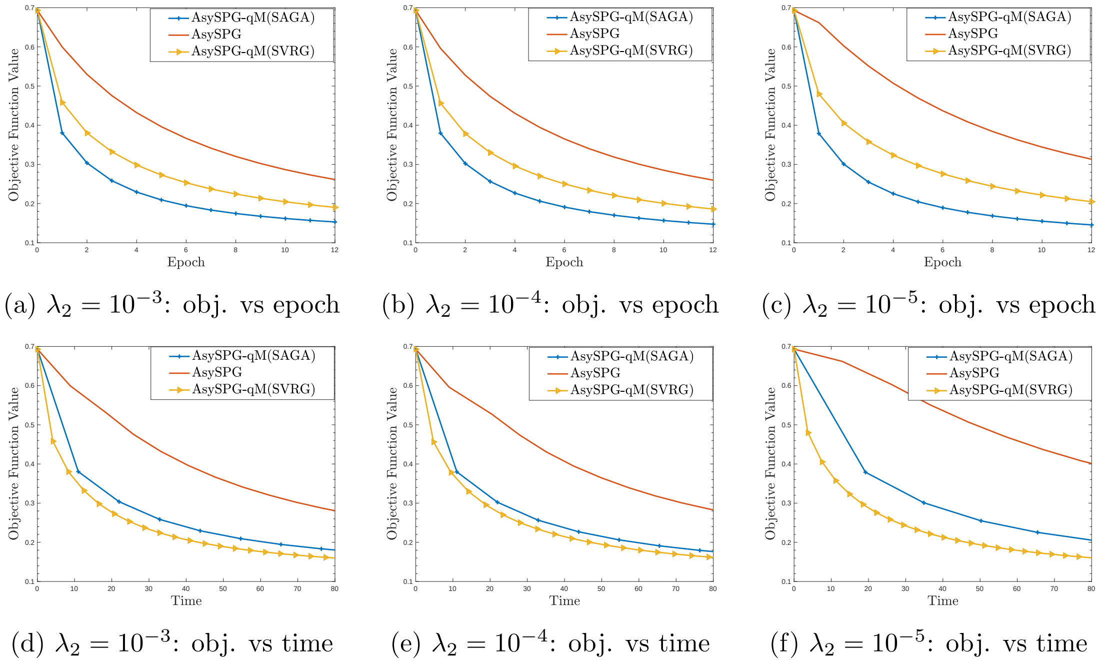
(c) $\lambda_2 = 10^{-5}$: obj. vs epoch

(d) $\lambda_2 = 10^{-3}$: obj. vs time

(e) $\lambda_2 = 10^{-4}$: obj. vs time

(f) $\lambda_2 = 10^{-5}$: obj. vs time

Figure 8: Convergence of AsySPG-qM and AsySPG on the Criteo dataset.



(a) $\lambda_2 = 10^{-3}$: obj. vs epoch

(b) $\lambda_2 = 10^{-4}$: obj. vs epoch

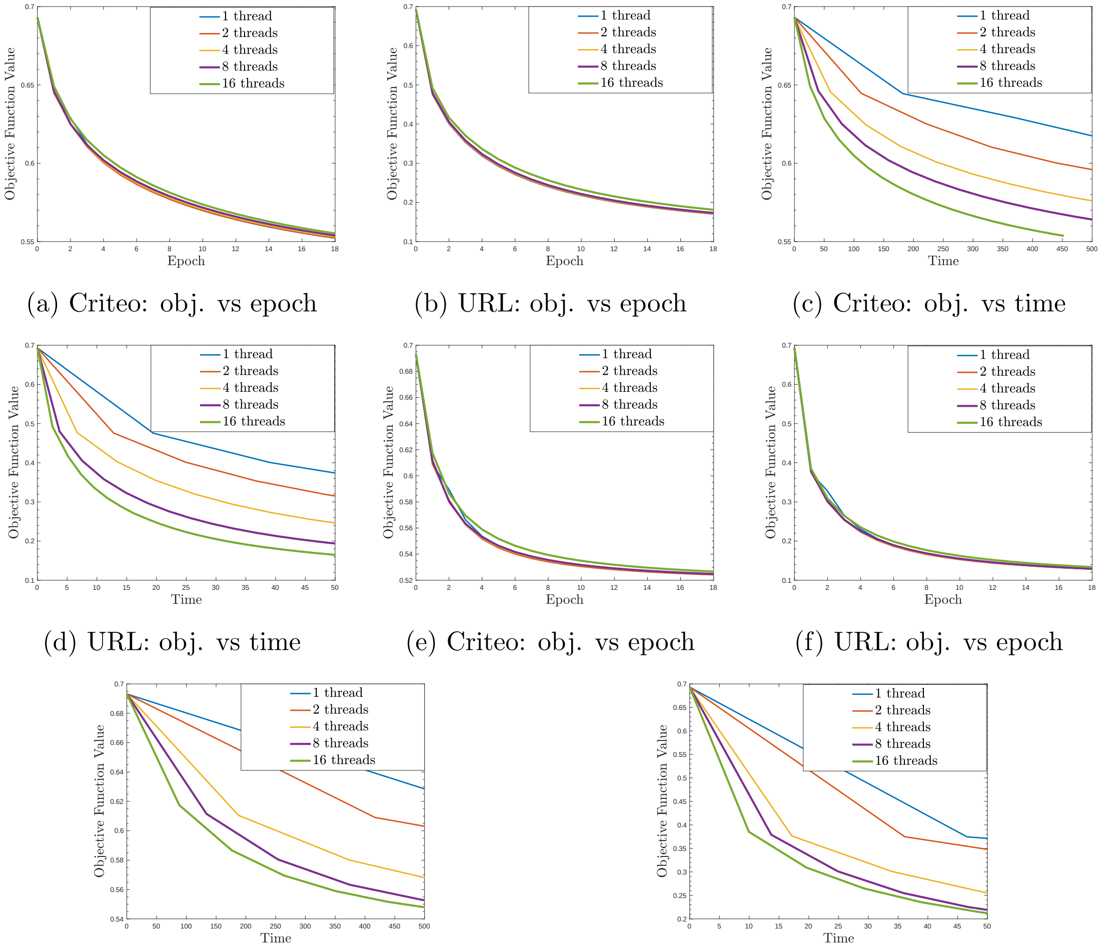
(c) $\lambda_2 = 10^{-5}$: obj. vs epoch

(d) $\lambda_2 = 10^{-3}$: obj. vs time

(e) $\lambda_2 = 10^{-4}$: obj. vs time

(f) $\lambda_2 = 10^{-5}$: obj. vs time

Figure 9: Convergence of AsySPG-qM and AsySPG on the URL dataset.



(a) Criteo: obj. vs epoch

(b) URL: obj. vs epoch

(c) Criteo: obj. vs time

(d) URL: obj. vs time

(e) Criteo: obj. vs epoch

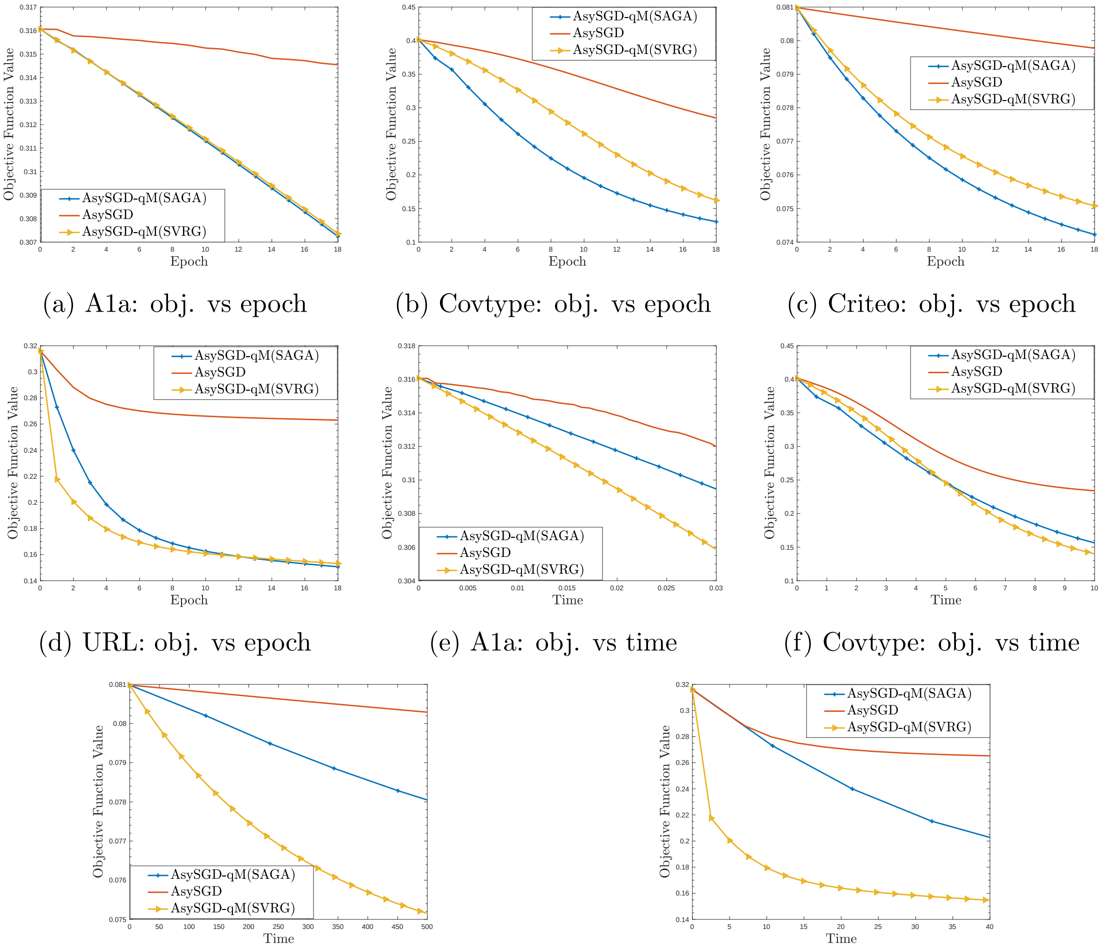
(f) URL: obj. vs epoch

(g) Criteo: obj. vs time

(h) URL: obj. vs time

Figure 10: Speedup results of AsySPG-qM. (a)-(d) AsySPG-qM(SVRG). (e)-(h) AsySPG-qM(SAGA).

To estimate the scalability of AsySGD-qM, we perform AsySGD-qM(SVRG) and AsySGD-qM(SAGA) on 1, 2, 4, 8 and 16 cores to observe the speedup. Figure 12 presents the objective values of AsySGD-qM(SVRG) and AsySPG-qM(SAGA) w.r.t. the epoch and running time on the Criteo and URL datasets, which show that AsySPG-qMR can have a near-linear speedup on a parallel system with shared memory. Similarly to AsySGD-qM, this is because that we do not use any lock in the implementation of AsySGD-qM.



(a) A1a: obj. vs epoch

(b) Covtype: obj. vs epoch

(c) Criteo: obj. vs epoch

(d) URL: obj. vs epoch

(e) A1a: obj. vs time

(f) Covtype: obj. vs time

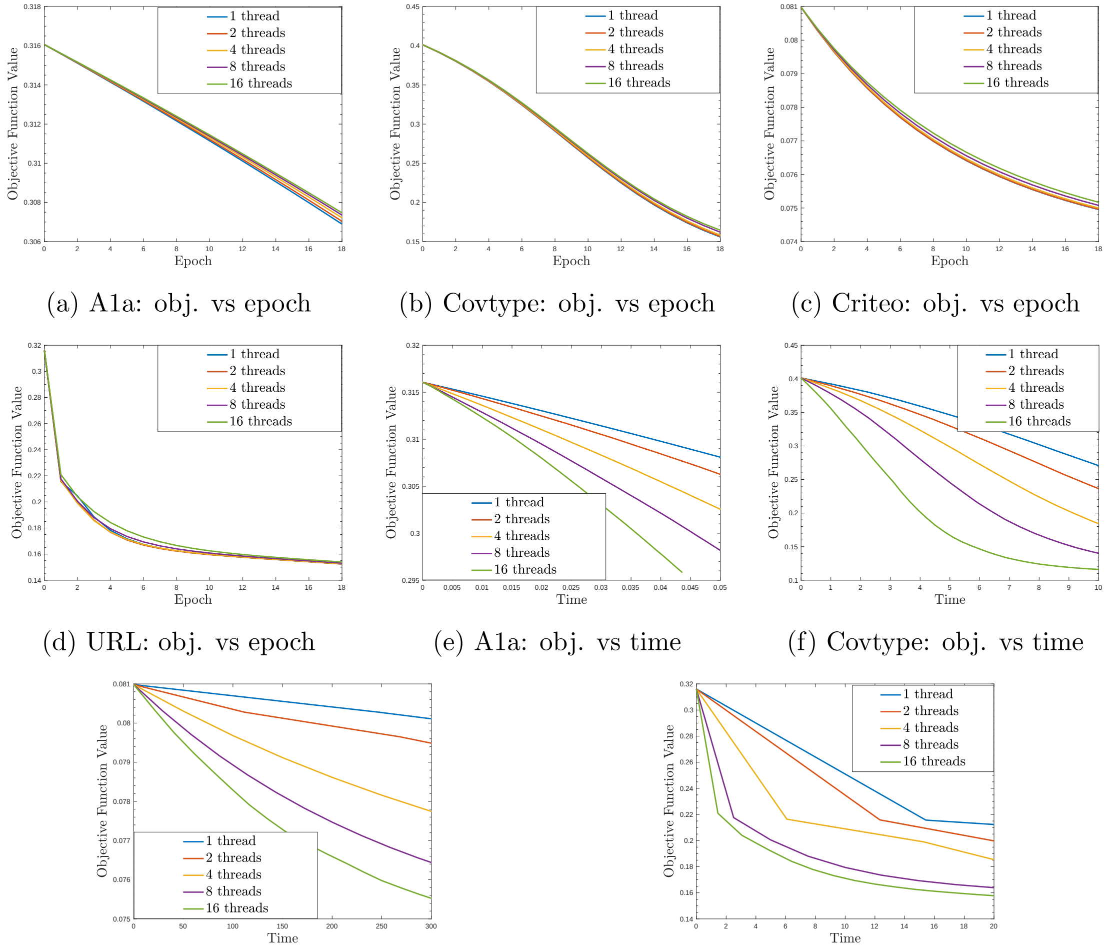
(g) Criteo: obj. vs time

(h) URL: obj. vs time

Figure 11: Convergence of AsySGD-qM and AsySGD on different datasets.

7. Proofs to Theorems 3, 9 and 11

In this section, we first provide the proof to Theorem 3, then give a brief proof to Theorem 9. Finally, we provide the proof to Theorem 11.



(a) A1a: obj. vs epoch

(b) Covtype: obj. vs epoch

(c) Criteo: obj. vs epoch

(d) URL: obj. vs epoch

(e) A1a: obj. vs time

(f) Covtype: obj. vs time

(g) Criteo: obj. vs time

(h) URL: obj. vs time

Figure 12: Speedup results of AsySGD-qM. (a)-(d) AsySGD-qM(SVRG). (e)-(h) AsySGD-qM(SAGA).

7.1 AsySGHT-qM

In this section, we provide the convergence analysis for AsySGHT-qM. Specifically, we first give the upper bounds to $\|g_t\|^2$, $\mathbb{E}\left\|\alpha_{i_t}^t - \nabla f_{i_t}(x^*)\right\|^2$, $\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2$ and $\mathbb{E}\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2$ in Lemmas 2, 14, 15 and 16 respectively. Then, based on Lemma 14, 15 and 16, we give the recursive relationship between $\mathbb{E}\left\|x^{t+1} - x^*\right\|^2$ and $\mathbb{E}\left\|x_t - x^*\right\|^2$ in Theorem 18. Finally, based on Theorem 18, we prove the linear convergence of AsySGHT-qM to an approximately global optimum in Theorem 3.

We first provide the proof to Lemma 2 as follows.

Proof Firstly, we have that

$$\begin{aligned} g_t &= \frac{1}{\gamma}\left(x^t - x^{t+1}\right) = \frac{1}{\gamma}\left(x^t - \mathcal{H}_k\left(x^t - \gamma\widehat{v}^{t+1}\right)\right) = \frac{1}{\gamma}\left(x^t - \left(x^t - \gamma(\widehat{v}^{t+1})\right)_{\mathcal{I}^{t+1}}\right) \\ &= \frac{1}{\gamma}\left(x^t - (x^t)_{\mathcal{I}^{t+1}}\right) + (\widehat{v}^{t+1})_{\mathcal{I}^{t+1}} = \frac{1}{\gamma}(x^t)_{\overline{\mathcal{I}^{t+1}}} + (\widehat{v}^{t+1})_{\mathcal{I}^{t+1}} \end{aligned} \tag{29}$$

Let $\widehat{\mathcal{I}^{t+1}}$ be the subset of $\mathcal{I}^{t+1}$ such that $(x^t)_i = 0, \forall i \in \widehat{\mathcal{I}^{t+1}}$ and $(x^t)_i \neq 0, \forall i \in \mathcal{I}^{t+1} - \widehat{\mathcal{I}^{t+1}}$. Thus, we have the upper bound of $\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}}\right\|^2$ as follows.

$$\begin{aligned} &\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}} + (\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \leq \left|\overline{\mathcal{I}^{t+1}}\right| \max_{i \in \overline{\mathcal{I}^{t+1}}} \left|(x^t)_i + (\widehat{v}^{t+1})_i\right|^2 \qquad (30) \\ \leq\ & \left|\overline{\mathcal{I}^{t+1}}\right| \min_{i \in \mathcal{I}^{t+1}} \left|(x^t)_i + (\widehat{v}^{t+1})_i\right|^2 \leq \left|\overline{\mathcal{I}^{t+1}}\right| \min_{i \in \widehat{\mathcal{I}^{t+1}}} \left|(\widehat{v}^{t+1})_i\right|^2 \\ \leq\ & \left|\overline{\mathcal{I}^{t+1}}\right| \left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 \end{aligned}$$

In addition, we have that

$$\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}} + (\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \geq \frac{1}{2}\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}}\right\|^2 - \left\|(\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \tag{31}$$

Based on (30) and (31), we have that

$$\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \leq 2\left|\overline{\mathcal{I}^{t+1}}\right| \left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 + 2\left\|(\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \tag{32}$$

Based on the above inequalities, we have that

$$\begin{aligned} &\|g_t\|^2 \leq \left\|\frac{1}{\gamma}(x^t)_{\overline{\mathcal{I}^{t+1}}} + (\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 \leq \frac{2}{\gamma^2}\left\|(x^t)_{\overline{\mathcal{I}^{t+1}}}\right\|^2 + 2\left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 \qquad (33) \\ \leq\ & \left(\frac{4\left|\overline{\mathcal{I}^{t+1}}\right|}{\gamma^2} + 2\right)\left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 + \frac{4}{\gamma^2}\left\|(\widehat{v}^{t+1})_{\overline{\mathcal{I}^{t+1}}}\right\|^2 \\ \leq\ & \left(\frac{4\left|\overline{\mathcal{I}^{t+1}}\right|}{\gamma^2} + 2\right)\left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 + \frac{4c}{\gamma^2}\left\|(\widehat{v}^{t+1})_{\mathcal{I}^{t+1}}\right\|^2 \\ \leq\ & \left(\frac{4n+4c}{\gamma^2} + 2\right)\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2 \end{aligned}$$

This completes the proof. ■

Lemma 14 *Suppose that the functions $f_i(x)$ satisfy the RSS condition with $s = 2k + k^*$. For AsySGHT-qM, we have that*

$$\mathbb{E}_{i_t}\left\|(\alpha_{i_t}^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 \le \frac{4\rho_s^+}{l}\sum_{u=1}^{t-1}\left(1-\frac{q}{l}\right)^{t-u-1} e(x^u) + 4\rho_s^+\left(1-\frac{q}{l}\right)^t e(x^0) \tag{34}$$

where $e(x^u) = \mathbb{E}F(x^u) - F(x^*)$.

Proof Firstly, we have that

$$\mathbb{E}_{i_t}\mathbb{E}\left\|(\alpha_{i_t}^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 = \frac{1}{l}\sum_{i=1}^{l}\mathbb{E}\left\|(\alpha_{i}^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} f_{i}(x^*)\right\|^2 \tag{35}$$

$$\begin{aligned}
&= \frac{1}{l}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}}\mathbb{E}\left(\alpha_i^t - \nabla f_i(x^*)\right)_d^2 \\
&= \frac{1}{l}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}}\mathbb{E}\sum_{u=0}^{t-1}\mathbf{1}_{\{u_{i,d}^t=u\}}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2 \\
&= \frac{1}{l}\sum_{u=0}^{t-1}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}}\mathbb{E}\mathbf{1}_{\{u_{i,d}^t=u\}}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2
\end{aligned}$$

where $u_{i,d}^t$ denote the time of the iterate last used to write the $[\alpha_i^u]_d$. We consider the two cases $u > 0$ and $u = 0$ as follows.

For $u > 0$, we have that

$$\begin{aligned}
& \mathbb{E}\left(\mathbf{1}_{\{u_{i,d}^t=u\}}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2\right) \qquad (36)\\
\le\ & \mathbb{E}\left(\mathbf{1}_{\{i_u=i\}}\mathbf{1}_{\{i_v\neq i,\forall v\ s.t.\ u+1\le v\le t-1\}}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2\right) \\
\le\ & P\{i_u = i\}P\{i_v \neq i, \forall v\ s.t.\ u+1 \le v \le t-1\}\mathbb{E}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2 \\
\le\ & \frac{q}{l}\left(1-\frac{q}{l}\right)^{t-u-1}\mathbb{E}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2
\end{aligned}$$

where i_u and i_v denote that the random picked index uniformly at random in $\{1, ..., l\}$ for the u-th and v-th iterations respectively, $P\{i_u = i\} = \frac{q}{l}$ and $P\{i_v \neq i\} = 1 - \frac{q}{l}$ are obtained from the condition 2 of Definition 1, the second inequality holds because i_v is independent to x^u.

For $u = 0$, we have that

$$\begin{aligned}
& \mathbb{E}\left(\mathbf{1}_{\{u_{i,d}^0=0\}}\left(\nabla f_i(x^0) - \nabla f_i(x^*)\right)_d^2\right) \qquad (37)\\
\le\ & \mathbb{E}\left(\mathbf{1}_{\{i_v\neq i,\forall v\ s.t.\ 0\le v\le t-1\}}\left(\nabla f_i(x^0) - \nabla f_i(x^*)\right)_d^2\right) \\
\le\ & P\{i_v \neq i, \forall v\ s.t.\ 0 \le v \le t-1\}\mathbb{E}\left(\nabla f_i(x^0) - \nabla f_i(x^*)\right)_d^2 \\
\le\ & \left(1-\frac{q}{n}\right)^t\mathbb{E}\left(\nabla f_i(x^0) - \nabla f_i(x^*)\right)_d^2
\end{aligned}$$

Substituting (36) and (37) into (35), we have that

$$\begin{aligned}
& \mathbb{E}\left\|(\alpha_{i_t}^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 \\
\leq \; & \frac{1}{l}\sum_{u=0}^{t-1}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}} \mathbb{E}\mathbf{1}_{\{u_{i,d}^t=u\}}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2 \\
\leq \; & \frac{1}{l}\sum_{u=1}^{t-1}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}} \left(\frac{q}{l}\left(1-\frac{q}{l}\right)^{t-u-1}\mathbb{E}\left(\nabla f_i(x^u) - \nabla f_i(x^*)\right)_d^2\right) \\
& + \frac{1}{l}\sum_{i=1}^{l}\sum_{d\in\mathcal{I}}\left(\left(1-\frac{q}{l}\right)^{t-1}\mathbb{E}\left(\nabla f_i(x^0) - \nabla f_i(x^*)\right)_d^2\right) \\
= \; & \frac{1}{l}\sum_{u=1}^{t-1}\left(1-\frac{q}{l}\right)^{t-u-1}\mathbb{E}\left\|\nabla_{\mathcal{I}} f_i(x^u) - \nabla_{\mathcal{I}} f_i(x^*)\right\|^2 + \left(1-\frac{q}{l}\right)^{t-1}\mathbb{E}\left\|\nabla_{\mathcal{I}} f_i(x^0) - \nabla_{\mathcal{I}} f_i(x^*)\right\|^2 \\
\leq \; & \frac{4\rho_s^+}{l}\sum_{u=1}^{t-1}\left(1-\frac{q}{l}\right)^{t-u-1} e(x^u) + 4\rho_s^+\left(1-\frac{q}{l}\right)^{t} e(x^0)
\end{aligned} \tag{38}$$

where the second inequality uses (36) and (37), the last inequality uses (8.16) in (Li et al., 2016). This completes the proof. ■

Lemma 15 *Suppose that the functions $f_i(x)$ satisfies the RSS condition with $s = 2k + k^*$. For AsySGHT-qM, we have that*

$$\begin{aligned}
& \mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 \\
\leq \; & \frac{12\rho_s^+}{l}\sum_{u=1}^{t-1}\left(1-\frac{q}{l}\right)^{t-u-1} e(x^u) + 12\rho_s^+\left(1-\frac{q}{l}\right)^{t} e(x^0) + 12e(x^t) + 3\mathbb{E}\left\|\nabla_{\mathcal{I}} F(x^*)\right\|^2
\end{aligned} \tag{39}$$

where $e(x^u) = \mathbb{E}F(x^u) - F(x^*)$.

Proof Define $v^{t+1} = \nabla f_{i_t}(x^t) - \alpha_{i_t}^t + \frac{1}{l}\sum_{i=1}^{l}\alpha_i^t$, we first give the upper bound to $\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2$.

$$\begin{aligned}
& \mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 = \mathbb{E}\left\|\nabla_{\mathcal{I}} f_{i_t}(x^t) - (\alpha_{i_t}^t)_{\mathcal{I}} + \frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}}\right\|^2 \\
= \; & \mathbb{E}\left\|\nabla_{\mathcal{I}} f_{i_t}(x^t) - \nabla_{\mathcal{I}} f_{i_t}(x^*) - (\alpha_{i_t}^t)_{\mathcal{I}} + \nabla_{\mathcal{I}} f_{i_t}(x^*) + \frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} F(x^*) + \nabla_{\mathcal{I}} F(x^*)\right\|^2 \\
\leq \; & 3\mathbb{E}\left\|\nabla_{\mathcal{I}} f_{i_t}(x^*) - (\alpha_{i_t}^t)_{\mathcal{I}} + \frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} F(x^*)\right\|^2 \\
& + 3\mathbb{E}\left\|\nabla_{\mathcal{I}} f_{i_t}(x^t) - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 + 3\mathbb{E}\left\|\nabla_{\mathcal{I}} F(x^*)\right\|^2 \\
\leq \; & 3\mathbb{E}\left\|(\alpha_{i_t}^t)_{\mathcal{I}} - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 + 3\mathbb{E}\left\|\nabla_{\mathcal{I}} f_{i_t}(x^t) - \nabla_{\mathcal{I}} f_{i_t}(x^*)\right\|^2 + 3\mathbb{E}\left\|\nabla_{\mathcal{I}} F(x^*)\right\|^2
\end{aligned} \tag{40}$$

$$\leq \frac{12\rho_s^+}{l}\sum_{u=1}^{t-1}\left(1-\frac{q}{l}\right)^{t-u-1}e(x^u)+12\rho_s^+\left(1-\frac{q}{l}\right)^t e(x^0)+12e(x^t)+3\mathbb{E}\left\|\nabla_{\mathcal{I}}F(x^*)\right\|^2$$

where the first inequality uses $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$, the second inequality follows from $\mathbb{E}\left\|x-\mathbb{E}x\right\|^2 \leq \mathbb{E}\left\|x\right\|^2$, and the third inequality uses Lemma 14. This completes the proof. ■

Lemma 16 *Suppose that the functions $f_i(x)$ satisfies the RSS condition with $s=2k+k^*$. For AsySGHT-qM, we have that*

$$\begin{aligned}
&\mathbb{E}\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2 && (41)\\
\leq\ & 6\varsigma\tau(\rho_s^+)^2\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2+\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
&+2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2
\end{aligned}$$

Proof Next, we have that

$$\begin{aligned}
&\mathbb{E}\left\|(\widehat{v}^{t+1})_{\mathcal{I}}-(v^{t+1})_{\mathcal{I}}\right\|^2 && (42)\\
=\ & \mathbb{E}\left\|\nabla_{\mathcal{I}}f_{i_t}(\widehat{x}^t)-(\widehat{\alpha}_{i_t}^t)_{\mathcal{I}}+\frac{1}{l}\sum_{i=1}^{l}(\widehat{\alpha}_i^t)_{\mathcal{I}}-\nabla_{\mathcal{I}}f_{i_t}(x^t)+(\alpha_{i_t}^t)_{\mathcal{I}}-\frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}}\right\|^2\\
\leq\ & 3\mathbb{E}\underbrace{\left\|\nabla_{\mathcal{I}}f_{i_t}(\widehat{x}^t)-\nabla_{\mathcal{I}}f_{i_t}(x^t)\right\|^2}_{Q_1}+3\mathbb{E}\underbrace{\left\|(\alpha_{i_t}^t)_{\mathcal{I}}-(\widehat{\alpha}_{i_t}^t)_{\mathcal{I}}\right\|^2}_{Q_2}+3\mathbb{E}\underbrace{\left\|\frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}}-\frac{1}{l}\sum_{i=1}^{l}(\widehat{\alpha}_i^t)_{\mathcal{I}}\right\|^2}_{Q_3}
\end{aligned}$$

where the second inequality uses $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$. We will give the upper bounds for the expectations of Q_1, Q_2 and Q_3 respectively.

$$\begin{aligned}
&\mathbb{E}Q_1=\mathbb{E}\left\|\nabla_{\mathcal{I}}f_{i_t}(\widehat{x}^t)-\nabla_{\mathcal{I}}f_{i_t}(x^t)\right\|^2\leq(\rho_s^+)^2\mathbb{E}\left\|\widehat{x}^t-x^t\right\|^2=(\rho_s^+)^2\gamma^2\mathbb{E}\left\|\sum_{u=(t-\tau)_+}^{t-1}S_u^t g_u\right\|^2\\
\leq\ & \tau(\rho_s^+)^2\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|S_u^t g_u\right\|^2\leq\varsigma\tau(\rho_s^+)^2\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2 && (43)
\end{aligned}$$

where the first inequality uses the L-Lipschitz continuity, the second inequality uses $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$, the third inequality uses the upper bound of virtual gradient (i.e., Assumption 2).

$$\mathbb{E}Q_2=\mathbb{E}\left\|(\alpha_{i_t}^t)_{\mathcal{I}}-(\widehat{\alpha}_{i_t}^t)_{\mathcal{I}}\right\|^2=\frac{1}{l}\sum_{i=1}^{l}\mathbb{E}\left\|(\alpha_i^t)_{\mathcal{I}}-(\widehat{\alpha}_i^t)_{\mathcal{I}}\right\|^2 \qquad (44)$$

$$\begin{aligned}
&= \frac{1}{l}\sum_{i=1}^{l}\mathbb{E}\sum_{u=0}^{t-1}\sum_{d\in\mathcal{I}}\mathbf{1}_{\{u_{i,d}^{t}=u\}}\left(\alpha_i^u-\widehat{\alpha}_i^u\right)_d^2 \\
&= \sum_{u=0}^{t-1}\frac{1}{l}\sum_{i=1}^{l}\mathbb{E}\sum_{d\in\mathcal{I}}\mathbf{1}_{\{u_{i,d}^{t}=u\}}\left(\alpha_i^u-\widehat{\alpha}_i^u\right)_d^2 = \sum_{u=(t-\tau)_+}^{t-1}\frac{1}{l}\sum_{i=1}^{l}\frac{q}{l}\mathbb{E}\left\|(\alpha_i^u)_{\mathcal{I}}-(\widehat{\alpha}_i^u)_{\mathcal{I}}\right\|^2 \\
&= \frac{q}{l}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\alpha_i^u)_{\mathcal{I}}-(\widehat{\alpha}_i^u)_{\mathcal{I}}\right\|^2 = \frac{q}{l}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|\nabla_{\mathcal{I}}f_i(\widehat{x}^{u-1})-\nabla_{\mathcal{I}}f_i(x^{u-1})\right\|^2 \\
&\leq \frac{q(\rho_s^+)^2}{l}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|\widehat{x}^{u-1}-x^{u-1}\right\|^2 = \frac{q(\rho_s^+)^2\gamma^2}{l}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|\sum_{k=(u-\tau-1)_+}^{u-2}S_k^{u-1}g_k\right\|^2 \\
&\leq \frac{\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2
\end{aligned}$$

where $u_{i,d}^t$ denotes the time of the iterate last used to write the $[\widehat{\alpha}_i^u]_d$, the sixth equality holds because $\alpha_i^u=\widehat{\alpha}_i^u$ if $u<t-\tau$, according to Assumption 3.

$$\begin{aligned}
\mathbb{E}Q_3 = \mathbb{E}&\left\|\frac{1}{l}\sum_{i=1}^{l}(\alpha_i^t)_{\mathcal{I}}-\frac{1}{l}\sum_{i=1}^{l}(\widehat{\alpha}_i^t)_{\mathcal{I}}\right\|^2 \\
\leq\ & \frac{1}{l}\sum_{i=1}^{l}\mathbb{E}\left\|(\alpha_i^t)_{\mathcal{I}}-(\widehat{\alpha}_i^t)_{\mathcal{I}}\right\|^2 \\
\leq\ & \frac{\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2
\end{aligned}\tag{45}$$

where the first inequality uses $\|\sum_{i=1}^{n}a_i\|^2\leq n\sum_{i=1}^{n}\|a_i\|^2$, the second inequality uses (44).

$$\begin{aligned}
&\mathbb{E}\left\|(\widehat{v}^{t+1})_{\mathcal{I}}\right\|^2 \leq 2\mathbb{E}\left\|(\widehat{v}^{t+1})_{\mathcal{I}}-(v^{t+1})_{\mathcal{I}}\right\|^2+2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 \\
\leq\ & 6\mathbb{E}Q_1+6\mathbb{E}Q_2+6\mathbb{E}Q_3+2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 \\
\leq\ & 6\varsigma\tau(\rho_s^+)^2\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2+\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2 \\
& +2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2
\end{aligned}\tag{46}$$

where the second inequality uses Lemma 14. This completes the proof. ■

Lemma 17 (Li et al. (2016)) *For $k>k^*$ and for any vector $x\in\mathbb{R}^N$, we have*

$$\|\mathcal{H}_k(x)-x^*\|\leq\left(1+\frac{2\sqrt{k^*}}{k-k^*}\right)\|x-x^*\|\tag{47}$$

Lemma 17 is proved in (Li et al., 2016).

Theorem 18 *Assume that the function F satisfies the RSC condition and the functions f_i satisfy the RSS condition with $s = 2k + k^*$. For AsySGHT-qM, we have that*

$$\begin{aligned}
& \mathbb{E}\left\|x^{t+1}-x^*\right\|^2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (48)\\
\leq\; & \varrho\mathbb{E}\left\|x^{t}-x^*\right\|^2 + \alpha(1+\beta)\frac{12\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
& +\alpha\left((1+\frac{1}{\beta})\gamma^2\tau\varsigma + 2(1+\beta)\varsigma\tau\rho_s^+\gamma^3(1+3\rho_s^+\gamma)\right)\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
& +2\alpha(1+\beta)\gamma^2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 - \alpha(1+\beta)\gamma e(x^t)
\end{aligned}$$

where $\varrho = \alpha(1+\beta)\left(1-\gamma\rho_s^-\right)$ and $e(x^u) = \mathbb{E}F(x^u) - F(x^*)$.

Proof According to (7.24) in (Li et al., 2016), we have the the upper bound to $-\mathbb{E}\left\langle(\widehat{v}^{t+1})_{\mathcal{I}}, x^t - x^*\right\rangle$.

$$-\mathbb{E}\left\langle(\widehat{v}^{t+1})_{\mathcal{I}}, x^t - x^*\right\rangle \leq -e(x^t) + \varsigma\rho_s^+\tau\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2 \tag{49}$$

Let $\widetilde{x}^{t+1} = \widehat{x}^t - \gamma\widehat{v}^{t+1}$. We have that

$$\begin{aligned}
& \mathbb{E}\left\|\widetilde{x}^{t+1}-x^*\right\|^2 = \mathbb{E}\left\|\widehat{x}^{t}-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad (50)\\
=\; & \mathbb{E}\left\|\widehat{x}^{t}-x^t\right\|^2 + \mathbb{E}\left\|x^t-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2 + 2\langle\widehat{x}^t - x^t, x^t - \gamma(\widehat{v}^{t+1})_{\mathcal{I}} - x^*\rangle\\
\leq\; & (1+\frac{1}{\beta})\mathbb{E}\left\|\widehat{x}^{t}-x^t\right\|^2 + (1+\beta)\mathbb{E}\left\|x^t-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2\\
=\; & (1+\frac{1}{\beta})\gamma^2\mathbb{E}\left\|\sum_{u=(t-\tau)_+}^{t-1}S_u^t g_u\right\|^2 + (1+\beta)\mathbb{E}\left\|x^t-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2\\
\leq\; & (1+\frac{1}{\beta})\gamma^2\tau\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|g_u\right\|^2 + (1+\beta)\mathbb{E}\left\|x^t-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2\\
\leq\; & (1+\frac{1}{\beta})\gamma^2\tau\varsigma\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2 + (1+\beta)\mathbb{E}\left\|x^t-\gamma(\widehat{v}^{t+1})_{\mathcal{I}}-x^*\right\|^2\\
\leq\; & (1+\beta)\mathbb{E}\left\|x^{t}-x^*\right\|^2 + (1+\beta)\frac{12\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
& +\left((1+\frac{1}{\beta})\gamma^2\tau\varsigma + 6(1+\beta)\varsigma\tau(\rho_s^+)^2\gamma^4 + 2(1+\beta)\varsigma\rho_s^+\tau\gamma^3\right)\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
& +2(1+\beta)\gamma^2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 - 2(1+\beta)\gamma e(x^t)
\end{aligned}$$

where the first inequality uses Lemma 16 and (49). In addition, according to the RSC condition, we have that

$$e(x^t) = \mathbb{E}F(x^t) - F(x^*) \geq \frac{\rho_s^-}{2}\mathbb{E}\left\|x^t - x^*\right\|^2 \tag{51}$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{k-k^*}$, according to (50), (51) and Lemma 17, we have that

$$\begin{aligned}
& \mathbb{E}\left\|x^{t+1} - x^*\right\|^2 \qquad (52)\\
\leq\ & \alpha(1+\beta)\left(1-\gamma\rho_s^-\right)\mathbb{E}\left\|x^t - x^*\right\|^2 + \alpha(1+\beta)\frac{12\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
& +\alpha\left((1+\frac{1}{\beta})\gamma^2\tau\varsigma + 2(1+\beta)\varsigma\tau\rho_s^+\gamma^3(1+3\rho_s^+\gamma)\right)\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
& +2\alpha(1+\beta)\gamma^2\mathbb{E}\left\|(v^{t+1})_{\mathcal{I}}\right\|^2 - \alpha(1+\beta)\gamma e(x^t)
\end{aligned}$$

This completes the proof. ∎

Based on Theorem 18, we provide the proof to Theorem 3.

Proof We will first give the upper bound of $\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2$. Let $0<\rho<1$, we have that

$$\begin{aligned}
& \sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2 \qquad (53)\\
\leq\ & \sum_{u=0}^{t}\rho^{t-u}\Bigg[6\varsigma\tau(\rho_s^+)^2\gamma^2\sum_{k=(u-\tau)_+}^{u-1}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
& +\frac{12(1+\beta)\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{k_1=(u-\tau)_+}^{u-1}\sum_{k_2=(k_1-\tau-1)_+}^{k_1-2}\mathbb{E}\left\|(\widehat{v}^{k_2+1})_{\mathcal{I}}\right\|^2 + 2\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2\Bigg]\\
=\ & \sum_{u=0}^{t}\rho^{t-u}6\varsigma\tau(\rho_s^+)^2\gamma^2\sum_{k=(u-\tau)_+}^{u-1}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
& +\sum_{u=0}^{t}\rho^{t-u}\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau}{l}\sum_{k_1=(u-\tau)_+}^{u-1}\sum_{k_2=(k_1-\tau-1)_+}^{k_1-2}\mathbb{E}\left\|(\widehat{v}^{k_2+1})_{\mathcal{I}}\right\|^2 + 2\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2\\
\leq\ & 6\varsigma\tau^2(\rho_s^+)^2\gamma^2\sum_{u=0}^{t}\rho^{t-(u+\tau+1)}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
& +\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau^3}{l}\sum_{u=0}^{t}\rho^{t-(u+2\tau+2)}\mathbb{E}\left\|(\widehat{v}^{k_2+1})_{\mathcal{I}}\right\|^2 + 2\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2\\
=\ & 6\varsigma\tau^2(\rho_s^+)^2\gamma^2\rho^{-(\tau+1)}\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2
\end{aligned}$$

$$+\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau^3}{l}\rho^{-(2\tau+2)}\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2+2\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2$$

where the first inequality uses Lemma 16. Based on (53), if $\gamma \leq \frac{1}{\sqrt{\frac{6\varsigma\tau^2(\rho_s^+)^2}{\rho^{\tau+1}}+\frac{12\varsigma q(\rho_s^+)^2\tau^3}{l\rho^{2\tau+2}}}}$ and $\nu=\frac{2}{1-\frac{6\varsigma\tau^2(\rho_s^+)^2\gamma^2}{\rho^{\tau+1}}-\frac{12\varsigma q(\rho_s^+)^2\gamma^2\tau^3}{l\rho^{2\tau+2}}}$, the upper bound of $\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2$ can be obtained as follows.

$$\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2 \leq \nu\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2 \tag{54}$$

Define $a^t=\mathbb{E}\left\|x^t-x^*\right\|^2$, and $\mathcal{L}^t=\sum_{u=0}^{t}\rho^{t-u}a^u$, we have that

$$\begin{aligned}
&\mathcal{L}_{t+1} \qquad\qquad (55)\\
=\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}a^{u+1}\\
\leq\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}\Bigg[\varrho a^u+\underbrace{\alpha\left((1+\frac{1}{\beta})\gamma^2\tau\varsigma+2(1+\beta)\varsigma\tau\rho_s^+\gamma^3(1+3\rho_s^+\gamma)\right)}_{\omega}\sum_{k=(u-\tau)_+}^{u-1}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
&+\frac{12\alpha(1+\beta)\varsigma\alpha q(\rho_s^+)^2\gamma^4\tau}{l}\sum_{v=(u-\tau)_+}^{u-1}\sum_{k=(v-\tau-1)_+}^{v-2}\mathbb{E}\left\|(\widehat{v}^{k+1})_{\mathcal{I}}\right\|^2\\
&+2\alpha(1+\beta)\gamma^2\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2-\alpha(1+\beta)\gamma e(x^u)\Bigg]\\
\leq\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}\left[\varrho a^u+2\alpha(1+\beta)\gamma^2\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2-\alpha(1+\beta)\gamma e(x^u)\right]\\
&+\omega\rho^{-(\tau+1)}\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
&+\frac{12\alpha(1+\beta)\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\rho^{-(2\tau+2)}\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(\widehat{v}^{u+1})_{\mathcal{I}}\right\|^2\\
\leq\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}\left[\varrho a^u+2\alpha(1+\beta)\gamma^2\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2-\alpha(1+\beta)\gamma e(x^u)\right]\\
&+\underbrace{\left(\nu\omega\rho^{-(\tau+1)}+\frac{12\nu\alpha(1+\beta)\varsigma q(\rho_s^+)^2\gamma^4\tau}{l}\rho^{-(2\tau+2)}\right)}_{\Gamma}\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2\\
=\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}\left[\varrho a^u+(2\alpha(1+\beta)\gamma^2+\Gamma)\mathbb{E}\left\|(v^{u+1})_{\mathcal{I}}\right\|^2-\alpha(1+\beta)\gamma e(x^u)\right]\\
\leq\;& \rho^{t+1}a^0+\sum_{u=0}^{t}\rho^{t-u}\Bigg[\varrho a^u+\frac{12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)}{l}\sum_{k=1}^{u-1}\left(1-\frac{q}{l}\right)^{u-k-1}e(x^k) \qquad\qquad (56)
\end{aligned}$$

$$
\begin{aligned}
& +12\rho_s^+ \left(1-\frac{q}{l}\right)^u (2\alpha(1+\beta)\gamma^2+\Gamma)e(x^0) + 12(2\alpha(1+\beta)\gamma^2+\Gamma)e(x^u) \\
& +3(2\alpha(1+\beta)\gamma^2+\Gamma)\mathbb{E}\left\|\nabla_{\mathcal{I}}F(x^*)\right\|^2 - \alpha(1+\beta)\gamma e(x^u)\Big] \\
\leq\ & \rho^{t+1}a^0 + \varrho\mathcal{L}_t - (\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)e(x^t) \\
& +3(2\alpha(1+\beta)\gamma^2+\Gamma)\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2
\end{aligned}
$$

where the first inequality uses Theorem 18, the third inequality uses (54), the fourth inequality uses Lemma 15, the fifth inequality holds by appropriately choosing γ such that the terms related to $e(x^u)$ $(u=0,\cdots,t-1)$ are negative, because the signs related to the lowest orders of $e(x^u)$ $(u=0,\cdots,t-1)$ are negative. In the following, we give the detailed analysis for how to choose γ such that the terms related to $e(x^u)$ $(u=0,\cdots,t-1)$ are negative. We first consider $u=0$. Assume that $C(e(x^0))$ is the coefficient term of $e(x^0)$ in (56), we have that

$$
\begin{aligned}
& C(e(x^0)) \qquad\qquad (57) \\
=\ & -\alpha\gamma\rho^t + 12(2\alpha(1+\beta)\gamma^2+\Gamma)\rho^t + 12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)\sum_{u=0}^{t}\rho^{t-u}\left(1-\frac{q}{l}\right)^u \\
=\ & \rho^t\left(-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma) + 12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)\sum_{u=0}^{t}\rho^{-u}\left(1-\frac{q}{l}\right)^u\right) \\
=\ & \rho^t\left(-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma) + 12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)\frac{1-\left(\frac{1-\frac{q}{l}}{\rho}\right)^t}{1-\frac{1-\frac{q}{l}}{\rho}}\right) \\
\overset{1-\frac{q}{l}<\rho}{\leq}\ & \rho^t\left(-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma)\left(1+\frac{\rho_s^+}{1-\frac{1-\frac{q}{l}}{\rho}}\right)\right)
\end{aligned}
$$

Based on (57), we can carefully choose γ such that $-\alpha\gamma+12(2\alpha(1+\beta)\gamma^2+\Gamma)\left(1+\frac{\rho_s^+}{1-\frac{1-\frac{q}{l}}{\rho}}\right)\leq 0$.

Assume that $C(e(x^u))$ is the coefficient term of $e(x^u)$ $(u=1,\cdots,t-1)$ in the big square brackets of (55), we have that

$$
\begin{aligned}
& C(e(x^u)) \qquad\qquad (58) \\
=\ & -\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma) + \frac{12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)}{l}\sum_{v=u+1}^{t-1}\left(1-\frac{q}{l}\right)^{v-u-1} \\
=\ & -\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma) + \frac{12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)}{l}\frac{1-(1-\frac{q}{l})^{(t-u-1)_+}}{1-\left(1-\frac{q}{l}\right)} \\
=\ & -\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma) + \frac{12\rho_s^+(2\alpha(1+\beta)\gamma^2+\Gamma)}{l}\frac{1-(1-\frac{q}{l})^{(t-u-1)_+}}{\frac{q}{l}}
\end{aligned}
$$

$$\leq \quad -\alpha\gamma + 12\left(1+\frac{\rho_s^+}{q}\right)(2\alpha(1+\beta)\gamma^2+\Gamma)$$

where the inequality holds due to $0 < (1-\frac{q}{l})^{(t-u-1)_+} \leq 1$. Based on (58), we can carefully choose γ such that $-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma)\left(1+\frac{\rho_s^+}{q}\right) \leq 0$.

In addition, assume that $C(e(x^t))$ is the coefficient term of $e(x^t)$ in (56), we have that $-\alpha\gamma + 12(2\alpha\gamma^2+\Gamma) \leq 0$. Combing the two above cases, we have that the terms related to $e(x^u)$ $(u = 0, \cdots, t-1)$ are negative if $-\alpha\gamma + 12(2\alpha(1+\beta)\gamma^2+\Gamma)\left(1+\frac{\rho_s^+}{1-\frac{1-\frac{q}{l}}{\rho}}\right) \leq 0$.

Thus, based on (55), we have that

$$\begin{aligned}
& (\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)e(x^t) \qquad (59)\\
\leq\ & (\alpha(1+\beta)\gamma - 24\alpha(1+\beta)\gamma^2 - 12\Gamma)e(x^t) + \mathcal{L}_{t+1}\\
\leq\ & \rho^{t+1}a^0 + \varrho\mathcal{L}_t + 3(2\alpha(1+\beta)\gamma^2+\Gamma)\sum_{u=0}^{t}\rho^{t-u}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2\\
\leq\ & \varrho^{t+1}\mathcal{L}_0 + \rho^{t+1}a^0\sum_{k=0}^{t+1}\left(\frac{\varrho}{\rho}\right)^k + \frac{3(2\alpha(1+\beta)\gamma^2+\Gamma)}{1-\rho}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2\sum_{k=0}^{t}\varrho^k\\
\overset{\varrho<\rho}{\leq}\ & \varrho^{t+1}a_0 + \rho^{t+1}\frac{a^0}{1-\frac{\varrho}{\rho}} + \frac{3(2\alpha(1+\beta)\gamma^2+\Gamma)}{(1-\rho)(1-\varrho)}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2\\
\leq\ & \rho^{t+1}\frac{(2-\frac{\varrho}{\rho})a^0}{1-\frac{\varrho}{\rho}} + \frac{3(2\alpha(1+\beta)\gamma^2+\Gamma)}{(1-\rho)(1-\varrho)}\mathbb{E}\left\|\nabla_{\widetilde{\mathcal{I}}}F(x^*)\right\|^2
\end{aligned}$$

From (60), we have the conclusion of Theorem 3. This completes the proof. ■

7.2 AsySPG-qM

Lemma 19 *For AsySPG-qM, under Assumptions 3, 4 and 5, we have that*

$$\mathbb{E}\left\|\widehat{\alpha}_{i_t}^t - \nabla f_{i_t}(x^*)\right\| \leq \frac{2qL}{l}\sum_{u=1}^{t=1}\left(1-\frac{q}{l}\right)^{(t-2\tau-u-1)_+} B_f(\widehat{x}_u, x^*) + \left(1-\frac{q}{l}\right)^{(t-\tau)_+}\widetilde{e}_0 \qquad (60)$$

where $B_f(\widehat{x}_u, x^*) = f(\widehat{x}_u) - f(x^*) - \langle\nabla f(x^*), \widehat{x}_u - x^*\rangle$, *and* $\widetilde{e}_0 = \mathbb{E}\|\alpha_i^0 - \nabla f_i(x^*)\|$.

The Lemma 19 can be got similarly to Lemma 2 of (Leblond et al., 2017). Based on Lemma 19, we provide the brief proof to Theorem 3.

Proof Define $a^t = \mathbb{E}\left\|x^t - x^*\right\|^2$, according to Lemma 19 and (50) in (Pedregosa et al., 2017) we have that

$$\begin{aligned}
& a^{t+1} \qquad (61)\\
\leq\ & \left(1-\frac{\gamma\mu}{2}\right)a^t + \frac{4\gamma^2L}{\beta}\left(1-\frac{q}{l}\right)^{(t-\tau)_+}\widetilde{e}_0 + \gamma^2\left[\beta - 1 + \sqrt{\triangle}\tau\right]\mathbb{E}\|g_t\|^2
\end{aligned}$$

$$+\left[\gamma^2\sqrt{\triangle}+\gamma^3\mu(1+\sqrt{\triangle}\tau)\right]\sum_{u=(t-\tau)_+}^{t}\mathbb{E}\|g_u\|^2-2\gamma\mathbb{E}B_f(\widehat{x}_t,x^*)$$
$$+\frac{4\gamma^2L}{\beta}\mathbb{E}B_f(\widehat{x}_t,x^*)+\frac{4\gamma^2qL}{\beta l}H^t$$

where $H^t=\sum_{u=1}^{t=1}\left(1-\frac{q}{l}\right)^{(t-2\tau-u-1)_+}B_f(\widehat{x}_u,x^*)$. Define $\mathcal{L}^t=\sum_{u=0}^{t}(1-\rho)^{t-u}a^u$, we have that

$$\begin{aligned}
&\mathcal{L}^{t+1} \qquad (62)\\
\leq\ & (1-\rho)^{t+1}a^0+\sum_{u=0}^{t}(1-\rho)^{t-u}\Bigg[(1-\frac{\gamma\mu}{2})a^u+\frac{4\gamma^2L}{\beta}(1-\frac{q}{l})^{(u-\tau)_+}\widetilde{e}_0\\
&+\gamma^2(\beta-1+\sqrt{\triangle}\tau)\mathbb{E}\|g_u\|^2+\left(\gamma^2\sqrt{\triangle}+\gamma^3\mu(1+\sqrt{\triangle}\tau)\right)\sum_{v=(u-\tau)_+}^{u}\mathbb{E}\|g_v\|^2\\
&-2\gamma\mathbb{E}B_f(\widehat{x}_u,x^*)+\frac{4\gamma^2L}{\beta}\mathbb{E}B_f(\widehat{x}_u,x^*)+\frac{4\gamma^2qL}{\beta l}H^u\Bigg]
\end{aligned}$$

After regrouping similar terms, we get

$$\mathcal{L}^{t+1}\leq(1-\rho)^{t+1}(a^0+A\widetilde{e}_0)+(1-\frac{\gamma\mu}{2})\mathcal{L}^t+\sum_{u=0}^{t}s_u^t\mathbb{E}\|g_u\|^2+\sum_{u=1}^{t}r_u^t\mathbb{E}B_f(\widehat{x}_u,x^*) \qquad (63)$$

where s_u^t, r_u^t and A are defined as follows.

$$\begin{aligned}
s_u^t &\leq (1-\rho)^{t-u}\left[\gamma^2(\beta-1+\sqrt{\triangle}\tau)+\tau(1-\rho)^{-\tau}(\gamma^2\sqrt{\triangle}+\gamma^3\mu(1+\sqrt{\triangle}\tau))\right] && (64)\\
r_u^t &\leq (1-\rho)^{t-u}\left[-2\gamma+\frac{4\gamma^2L}{\beta}+\frac{4L\gamma^2q}{l\beta}(1-\rho)^{-2\tau-1}\left(2\tau+\frac{1}{1-\tilde{\rho}}\right)\right] && (65)\\
A &= \frac{4\gamma^2L}{\beta}(1-\rho)^{-\tau-1}\left(\tau+1+\frac{1}{1-\tilde{\rho}}\right) && (66)
\end{aligned}$$

where $\tilde{\rho}=\frac{1-q/l}{1-\rho}$. Now, provided that we can prove that under certain conditions the s_u^t and r_u^t terms are all negative (and that the A term is not too big), we can drop them from the right-hand side of (63) which will allow us to finish the proof.

We consider s_u^t here. Firstly, we assume $\rho\leq\frac{q}{4l}$, and $\tau\leq\frac{1}{10\sqrt{\triangle}}\leq\frac{1}{10}\sqrt{l}$, thus we have that

$$\frac{1}{1-\tilde{\rho}}=\frac{1}{1-\frac{1-q/l}{1-\rho}}=\frac{1-\rho}{q/l-\rho}=\frac{1-\frac{q}{4l}}{q/l-\frac{q}{4l}}\leq\frac{4l}{3q} \qquad (67)$$

$$(1-\rho)^{-k\tau-1}\leq\frac{1}{1-\rho(k\tau+1)}\leq\frac{1}{1-\frac{q}{4l}(k\frac{1}{10}\sqrt{l}+1)}=\frac{1}{1-\frac{kq}{40\sqrt{l}}-\frac{q}{4l}} \qquad (68)$$

If $k\in\{0,1,2\}$ and $q\leq\frac{5l}{\sqrt{l}-5}\leq5\sqrt{l}$, we have that $(1-\rho)^{-k\tau-1}\leq\frac{4}{3}$. Thus, setting $\beta=\frac{1}{2}$, we have

$$s_u^t\ \leq\ (1-\rho)^{t-u}\gamma^2\left[-\frac{1}{2}+\sqrt{\triangle}\tau+\frac{4}{3}(\sqrt{\triangle}\tau+\gamma\mu\tau(1+\sqrt{\triangle}\tau))\right] \qquad (69)$$

$$\leq \ (1-\rho)^{t-u}\gamma^2\left[-\frac{1}{2}+\frac{1}{10}+\frac{4}{30}+\gamma\mu\tau\frac{4}{3}\frac{11}{10}\right]$$

Thus, the condition under which all s_u^t are negative boils down to $\gamma\mu\tau \leq \frac{2}{11}$.

Next, we consider r_u^t as follows.

$$\begin{aligned} r_u^t \ &\leq \ (1-\rho)^{t-u}\left[-2\gamma+8\gamma^2 L+\frac{8\gamma^2 qL}{l}\frac{4}{3}\left(\frac{\sqrt{l}}{5}+\frac{4l}{3q}\right)\right] \qquad (70) \\ &\leq \ (1-\rho)^{t-u}\left[-2\gamma+8\gamma^2 L+\frac{8\gamma^2 qL}{l}\frac{4}{3}\left(\frac{l}{q}+\frac{4l}{3q}\right)\right] \end{aligned}$$

Thus, the condition under which all r_u^t are negative boils down to $\gamma \leq \frac{3}{124L}$. If we add $\gamma B_f(\widehat{x}_t, x^*)$ to both sides of (63), r_t^t is replaced by $r_t^t+\gamma$, which makes for a slightly worse bound of γ as $\gamma \leq \frac{3}{248L}$ to ensure linear convergence.

Thus, if $\gamma = \frac{a}{L}$, $a \leq \min\left\{\frac{3}{248}, \frac{2\kappa}{11\tau}\right\}$ and $q \leq 5\sqrt{l}$, we have that

$$\mathcal{L}^{t+1} \quad \leq \quad \mathcal{L}^{t+1}+\gamma B_f(\widehat{x}_t, x^*) \leq (1-\rho)^{t+1}(a^0+A\widetilde{e}_0)+(1-\frac{\gamma\mu}{2})\mathcal{L}^t \qquad (71)$$

By unrolling the recursion (71), we can carefully combine the effect of the geometric term $(1-\rho)$ with the one of $(1-\frac{\gamma\mu}{2})$ which yields the overall rate:

$$\frac{\gamma\mu}{2}\mathbb{E}\|\widehat{x}_t - x^*\| \leq \gamma B_f(\widehat{x}_t, x^*) \leq (1-\rho^*)^{t+1}\widetilde{C}_0 \qquad (72)$$

where $\rho^* = \frac{1}{5}\min\{\frac{q}{l}, \frac{a}{\kappa}\}$, $\widetilde{C}_0 = \frac{21l^2\kappa}{a\gamma}\left(\|x_0 - x^*\|^2 + \frac{1}{5L^2}\sum_{i=1}^{l}\|\alpha - \nabla f_i(x^*)\|^2\right)$. This completes the proof. ■

7.3 AsySGD-qM

In this section, we provide the convergence analysis for AsySGD-qM under the smooth assumption (i.e., Assumption 4). Firstly, we give an upper bound to $\sum_{t=0}^{T-1} c_t\mathbb{E}\left\|(\widehat{v}^t)\right\|^2$ in Lemma 20.

Lemma 20 *Under Assumptions 3 and 4 and* $\gamma < \frac{1}{\sqrt{6\tau^2L^2+\frac{12qL^2\tau^2}{l}}}$. *Let* $\{c_t\}_{t=0}^{T}$ *be a monotonically decreasing sequence and* $a = \frac{2}{1-6\tau^2L^2\gamma^2-\frac{12qL^2\gamma^2\tau^3}{l}}$. *For AsySGD-qM, we have that*

$$\sum_{t=0}^{T-1} c_t\mathbb{E}\left\|(\widehat{v}^t)\right\|^2 \leq a\sum_{t=0}^{T-1}\mathbb{E}c_t\left\|(v^t)\right\|^2 \qquad (73)$$

Proof Similar to (73), we can obtain

$$\begin{aligned} &\mathbb{E}\left\|(\widehat{v}^t)\right\|^2 \qquad (74) \\ \leq \ & 6\tau L^2\gamma^2\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\left\|(\widehat{v}^{u+1})\right\|^2+\frac{12\varsigma qL^2\gamma^2\tau}{l}\sum_{u=(t-\tau)_+}^{t-1}\sum_{k=(u-\tau-1)_+}^{u-2}\mathbb{E}\left\|\widehat{v}^{k+1}\right\|^2+2\mathbb{E}\left\|v^{t+1}\right\|^2 \end{aligned}$$

By summing the the inequality (74) over $t = 0, \ldots, T-1$, we obtain

$$\begin{aligned}
& \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|(\widehat{v}^t)\right\|^2 \qquad (75) \\
\leq \;& 6\tau L^2 \gamma^2 \sum_{t=0}^{T-1} c_t \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E} \left\|(\widehat{v}^{u+1})\right\|^2 + \frac{12qL^2\gamma^2\tau}{l} \sum_{t=0}^{T-1} c_t \sum_{u=(t-\tau)_+}^{t-1} \sum_{k=(u-\tau-1)_+}^{u-2} \mathbb{E} \left\|\widehat{v}^{k+1}\right\|^2 \\
& +2 \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|v^t\right\|^2 \\
\leq \;& 6\tau^2 L^2 \gamma^2 \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|(\widehat{v}^t)\right\|^2 + \frac{12\varsigma qL^2\gamma^2\tau^3}{l} \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|\widehat{v}^t\right\|^2 + 2 \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|v^t\right\|^2
\end{aligned}$$

According to (75), we have that

$$\left(1 - 6\tau^2 L^2 \gamma^2 - \frac{12qL^2\gamma^2\tau^3}{l}\right) \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|(\widehat{v}^t)\right\|^2 \leq 2 \sum_{t=0}^{T-1} c_t \mathbb{E} \left\|v^t\right\|^2 \qquad (76)$$

If $\gamma < \frac{1}{\sqrt{6\tau^2 L^2 + \frac{12qL^2\tau^2}{l}}}$, we have that $1 - 6\tau^2 L^2 \gamma^2 - \frac{12\varsigma qL^2\gamma^2\tau^3}{l} > 0$. Thus, we can have the conclusion. This completes the proof. ■

Based on Lemma 20, we provide the proof to Theorem 11.

Proof We define the Lyapunov function as $R^t = \mathbb{E}\left(f(x^t) + \frac{c_t}{l}\sum_{i=1}^{l} \|x^t - \alpha_i^t\|^2\right)$. Firstly, we give the upper bound to $\mathbb{E}f(x^{t+1})$ as follows.

$$\begin{aligned}
& \mathbb{E}f(x^{t+1}) \qquad (77) \\
\leq \;& \mathbb{E}\left(f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2\right) \\
= \;& \mathbb{E}\left(f(x^t) - \gamma \langle \nabla f(x^t), \widehat{v}^t \rangle + \frac{L\gamma^2}{2}\|\widehat{v}^t\|^2\right) \\
= \;& \mathbb{E}f(x^t) - \gamma \mathbb{E}\langle \nabla f(x^t), \nabla f(\widehat{x}^t) \rangle + \frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2 \\
= \;& \mathbb{E}f(x^t) - \frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(x^t)\|^2 + \|\nabla f(\widehat{x}^t)\|^2 - \|\nabla f(x^t) - \nabla f(\widehat{x}^t)\|^2\right) + \frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2 \\
\leq \;& \mathbb{E}f(x^t) - \frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(x^t)\|^2 + \|\nabla f(\widehat{x}^t)\|^2\right) + \frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2 + \frac{\gamma L^2}{2}\mathbb{E}\|x^t - \widehat{x}^t\|^2 \\
\leq \;& \mathbb{E}f(x^t) - \frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(x^t)\|^2 + \|\nabla f(\widehat{x}^t)\|^2\right) + \frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2 + \frac{\tau L^2\gamma^3}{2} \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}\|\widehat{v}^u\|^2
\end{aligned}$$

Secondly, we give the upper bound to $\frac{1}{l}\sum_{i=1}^{l} \mathbb{E}\|x^{t+1} - \alpha_i^{t+1}\|^2$ as follows.

$$\frac{1}{l}\sum_{i=1}^{l} \mathbb{E}\|x^{t+1} - \alpha_i^{t+1}\|^2 = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{l}\mathbb{E}\|x^{t+1} - x^t\|^2 + \frac{l-1}{l}\mathbb{E}\|x^{t+1} - \alpha_i^t\|^2\right) \qquad (78)$$

$$
\begin{aligned}
&= \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{l}\mathbb{E}\|x^{t+1}-x^t\|^2+\frac{l-1}{l}\mathbb{E}\left(\|x^{t+1}-x^t\|^2+\|x^t-\alpha_i^t\|^2+2\langle x^{t+1}-x^t,x^t-\alpha_i^t\rangle\right)\right)\\
&= \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{l}\mathbb{E}\|x^{t+1}-x^t\|^2+\frac{l-1}{l}\mathbb{E}\left(\|x^{t+1}-x^t\|^2+\|x^t-\alpha_i^t\|^2-2\gamma\langle\widehat{v}^t,x^t-\alpha_i^t\rangle\right)\right)\\
&\leq \frac{1}{l}\sum_{i=1}^{l}\left(\frac{\gamma^2}{l}\mathbb{E}\|\widehat{v}^t\|^2+\frac{l-1}{l}\mathbb{E}\left(\gamma^2\|\widehat{v}^t\|^2+\|x^t-\alpha_i^t\|^2+\frac{\gamma}{\beta}\|\widehat{v}^t\|^2+\gamma\beta\|x^t-\alpha_i^t\|^2\right)\right)\\
&= \frac{\gamma^2}{l}\mathbb{E}\|\widehat{v}^t\|^2+\frac{l-1}{l}\left(\gamma^2+\frac{\gamma}{\beta}\right)\mathbb{E}\|\widehat{v}^t\|^2+\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2\\
&= \left(\gamma^2+\frac{(l-1)\gamma}{\beta l}\right)\mathbb{E}\|\widehat{v}^t\|^2+\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2
\end{aligned}
$$

Based on (77) and (78), we have that

$$
\begin{aligned}
&R^{t+1}=\mathbb{E}\left(f(x^{t+1})+\frac{c_{t+1}}{l}\sum_{i=1}^{l}\|x^{t+1}-\alpha_i^{t+1}\|^2\right) \qquad (79)\\
\leq\;& \mathbb{E}f(x^t)-\frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(x^t)\|^2+\|\nabla f(\widehat{x}^t)\|^2\right)+\frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2+\frac{\tau L^2\gamma^3}{2}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\|\widehat{v}^u\|^2\\
&+c_{t+1}\left(\gamma^2+\frac{(l-1)\gamma}{\beta l}\right)\mathbb{E}\|\widehat{v}^t\|^2+c_{t+1}\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2
\end{aligned}
$$

By summing the the inequality (79) over $t=0,\ldots,T-1$, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}R^{t+1} \qquad (80)\\
\leq\;& \sum_{t=0}^{T-1}\Bigg(\mathbb{E}f(x^t)-\frac{\gamma}{2}\mathbb{E}\left(\|\nabla f(x^t)\|^2+\|\nabla f(\widehat{x}^t)\|^2\right)+\frac{L\gamma^2}{2}\mathbb{E}\|\widehat{v}^t\|^2+\frac{\tau L^2\gamma^3}{2}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\|\widehat{v}^u\|^2\\
&+c_{t+1}\left(\gamma^2+\frac{(l-1)\gamma}{\beta l}\right)\mathbb{E}\|\widehat{v}^t\|^2+c_{t+1}\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2\Bigg)\\
=\;& \sum_{t=0}^{T-1}\mathbb{E}f(x^t)-\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla f(x^t)\|^2+\|\nabla f(\widehat{x}^t)\|^2\right)+\frac{\tau L^2\gamma^3}{2}\sum_{t=0}^{T-1}\sum_{u=(t-\tau)_+}^{t-1}\mathbb{E}\|\widehat{v}^u\|^2\\
&+\sum_{t=0}^{T-1}\left(\frac{L\gamma^2}{2}+c_{t+1}\left(\gamma^2+\frac{(l-1)\gamma}{\beta l}\right)\right)\mathbb{E}\|\widehat{v}^t\|^2+\sum_{t=0}^{T-1}c_{t+1}\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2\\
\leq\;& \sum_{t=0}^{T-1}\mathbb{E}f(x^t)-\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla f(x^t)\|^2+\|\nabla f(\widehat{x}^t)\|^2\right)+\sum_{t=0}^{T-1}c_{t+1}\frac{l-1}{l^2}(1+\gamma\beta)\sum_{i=1}^{l}\mathbb{E}\|x^t-\alpha_i^t\|^2
\end{aligned}
$$

$$
\begin{aligned}
& + \sum_{t=0}^{T-1} \left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2 L^2 \gamma^3}{2} \right) \mathbb{E}\|\widehat{v}^t\|^2 \\
\leq \;& \sum_{t=0}^{T-1} \mathbb{E} f(x^t) - \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}\left(\|\nabla f(x^t)\|^2 + \|\nabla f(\widehat{x}^t)\|^2 \right) + \sum_{t=0}^{T-1} c_{t+1} \frac{l-1}{l^2}(1+\gamma\beta) \sum_{i=1}^{l} \mathbb{E}\|x^t - \alpha_i^t\|^2 \\
& + \sum_{t=0}^{T-1} a \left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2 L^2 \gamma^3}{2} \right) \mathbb{E}\|v^t\|^2 \\
\leq \;& \sum_{t=0}^{T-1} \mathbb{E} f(x^t) - \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}\left(\|\nabla f(x^t)\|^2 + \|\nabla f(\widehat{x}^t)\|^2 \right) + \sum_{t=0}^{T-1} c_{t+1} \frac{l-1}{l^2}(1+\gamma\beta) \sum_{i=1}^{l} \mathbb{E}\|x^t - \alpha_i^t\|^2 \\
& + \sum_{t=0}^{T-1} a \left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2 L^2 \gamma^3}{2} \right) \left(2\mathbb{E}\|\nabla f(x^t)\|^2 + \frac{2L^2}{l} \sum_{i=1}^{l} \mathbb{E}\|x^t - \alpha_i^t\|^2 \right) \\
= \;& \sum_{t=0}^{T-1} \mathbb{E} f(x^t) - \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\widehat{x}^t)\|^2 \\
& + \sum_{t=0}^{T-1} \frac{1}{l} \underbrace{\left(aL^3\gamma^2 + 2aL^2 c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + a\tau^2 L^4 \gamma^3 + c_{t+1}\frac{l-1}{l}(1+\gamma\beta) \right)}_{c_t} \sum_{i=1}^{l} \mathbb{E}\|x^t - \alpha_i^t\|^2 \\
& - \sum_{t=0}^{T-1} \underbrace{\left(\frac{\gamma}{2} - 2a\left(\frac{L\gamma^2}{2} + c_{t+1}\left(\gamma^2 + \frac{(l-1)\gamma}{\beta l}\right) + \frac{\tau^2 L^2 \gamma^3}{2} \right) \right)}_{\Gamma_t} \mathbb{E}\|\nabla f(x^t)\|^2 \\
= \;& \sum_{t=0}^{T-1} R^t - \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\widehat{x}^t)\|^2 - \sum_{t=0}^{T-1} \Gamma_t \mathbb{E}\|\nabla f(x^t)\|^2
\end{aligned}
$$

where the first inequality uses (79), the third inequality uses Lemma 20, the fourth inequality uses Lemma 2 of Reddi et al. (2016b). If we carefully choose γ such that $\Gamma_t > 0$ over $t = 0, \ldots, T-1$, we have that

$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{R^0 - R^T}{T \min_{0 \leq t \leq T-1} \Gamma_t} = \frac{f(x^0) - f(x^*)}{T \min_{0 \leq t \leq T-1} \Gamma_t} \tag{81}
$$

This completes the proof. ■

8. Conclusion

In this paper, we introduced an unified q-memorization framework for various variance reduction techniques (including SVRG, S2GD, SAGA, q-SAGA) to analyze asynchronous stochastic algorithms for three important optimization problems. Specifically, based on the q-memorization framework, **1**) we propose an asynchronous stochastic gradient hard thresholding algorithm with q-memorization (AsySGHT-qM) for the non-convex optimization with cardinality constraint, and prove that the convergence rate of AsySGHT-qM before reaching the inherent error induced by GHT-style methods is geometric. **2**) We propose an asynchronous stochastic proximal gradient algorithm (AsySPG-qM) for the convex optimization with non-smooth regularization, and prove that AsySPG-qM can achieve a linear convergence rate. **3**) We propose an asynchronous stochastic gradient descent algorithm (AsySGD-qM) for the general non-convex optimization problem, and prove that AsySGD-qM can achieve a sublinear convergence rate. The experimental results on various large-scale datasets confirm the fast convergence of our AsySGHT-qM, AsySPG-qM and AsySGD-qM through concrete realizations of SVRG and SAGA.

Besides the non-convex optimization problem with cardinality constraint, the convex optimization problem with non-smooth regularization and the general non-convex smooth optimization problem, we believe that our analysis scheme based on the q-memorization framework can be extended to the asynchronous stochastic algorithms for other optimization problems.

Acknowledgments

B. Gu was partially supported by he National Natural Science Foundation of China (No: 62076138), the Natural Science Foundation (No. BK20161534), Six talent peaks project (No. XYDXX-042) and the 333 Project (No. BRA2017455) in Jiangsu Province and the National Natural Science Foundation of China (No: 61573191).

References

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707, 2016.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Rohit Chandra. *Parallel programming in OpenMP*. Morgan kaufmann, 2001.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, and Johan AK Suykens. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16:993–1034, 2015.

Bin Gu and Zhouyuan Huo. Asynchronous doubly stochastic group regularized learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, 2018.

Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning*, pages 1807–1816, 2018.

Bin Gu, Wenhan Xian, and Heng Huang. Asynchronous stochastic frank-wolfe algorithms for nonconvex optimization. In *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 2019.

Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8): 1561–1576, 2011.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163 (1-2):85–114, 2017.

Xiaolin Huang, Lei Shi, and Johan AK Suykens. Ramp loss linear programming support vector machine. *The Journal of Machine Learning Research*, 15(1):2185–2211, 2014.

Zhouyuan Huo and Heng Huang. Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization. In *AAAI*, pages 2043–2049, 2017.

Zhouyuan Huo, Bin Gu, and Heng Huang. Training neural networks using features replay. In *Advances in Neural Information Processing Systems*, pages 6659–6668, 2018a.

Zhouyuan Huo, Bin Gu, Heng Huang, et al. Decoupled parallel backpropagation with convergence guarantee. In *International Conference on Machine Learning*, pages 2098–2106, 2018b.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Ali Jalali, Christopher C Johnson, and Pradeep K Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems*, pages 1935–1943, 2011.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Jakub Konečný and Peter Richtárik. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017. ISSN 2297-4687. doi: 10.3389/fams.2017.00009. URL `https://www.frontiersin.org/article/10.3389/fams.2017.00009`.

Rémi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Asaga: Asynchronous parallel saga. In *Artificial Intelligence and Statistics*, pages 46–54, 2017.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l_1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

Xingguo Li, Raman Arora, Han Liu, Jarvis Haupt, and Tuo Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *arXiv preprint arXiv:1605.02711*, 2016.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. pages 3054–3062, 2016.

Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, and Hai Zhang. Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification. *BMC bioinformatics*, 14(1):198, 2013.

Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

Qi Meng, Wei Chen, Jingcheng Yu, Taifeng Wang, Zhiming Ma, and Tie-Yan Liu. Asynchronous stochastic proximal optimization algorithms with variance reduction. pages 2329–2335, 2017.

F Michel. How many photos are uploaded to flickr every day and month? *Papadimitriou, S., & Sun, J.(2008). Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining*, 2012.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Trans. Information Theory*, 63(11): 6869–6895, 2017.

Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

Fabian Pedregosa, Rémi Leblond, and Simon Lacoste-Julien. Breaking the nonsmooth barrier: A scalable parallel method for composite optimization. pages 56–65, 2017.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems*, pages 2647–2655, 2015.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016a.

Sashank J. Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Fast incremental method for smooth nonconvex optimization. pages 1971–1977, 2016b.

Jean-Charles Régin. Generalized arc consistency for global cardinality constraint. In *Proceedings of the thirteenth national conference on Artificial intelligence-Volume 1*, pages 209–215. AAAI Press, 1996.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018. URL `http://jmlr.org/papers/v18/16-299.html`.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

Gang Wang, Derek Hoiem, and David Forsyth. Learning image similarity from flickr groups using fast kernel machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2177–2188, 2012.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 127–135, 2014.

Shen-Yi Zhao and Wu-Jun Li. Fast asynchronous parallel stochastic gradient descent: A lock-free approach with convergence guarantee. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Tuo Zhao, Mo Yu, Yiming Wang, Raman Arora, and Han Liu. Accelerated mini-batch randomized block coordinate descent method. In *Advances in neural information processing systems*, pages 3329–3337, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.