Accelerated Stochastic Optimization Methods under Quasar-convexity

Qiang Fu [1] **Dongchu Xu** [2] **Ashia Wilson** [3]

Abstract

Non-convex optimization plays a key role in a growing number of machine learning applications. This motivates the identification of specialized structure that enables sharper theoretical analysis. One such identified structure is quasar-convexity, a non-convex generalization of convexity that subsumes convex functions. Existing algorithms for minimizing quasar-convex functions in the stochastic setting have either high complexity or slow convergence, which prompts us to derive a new class of stochastic methods for optimizing smooth quasar-convex functions. We demonstrate that our algorithms have fast convergence and outperform existing algorithms on several examples, including the classical problem of learning linear dynamical systems. We also present a unified analysis of our newly proposed algorithms and a previously studied deterministic algorithm.

1. Introduction

Momentum is one of the most widely used techniques for speeding up the convergence rate of optimization methods. Many deterministic and stochastic momentum based algorithms have been proposed for optimizing (strongly) convex functions, e.g. accelerated gradient descent (AGD) (Nesterov, 1983; 2003; Beck & Teboulle, 2009), accelerated stochastic gradient descent (ASGD) (Ghadimi & Lan, 2012; 2016; Kulunchakov & Mairal, 2020), accelerated stochastic variance reduced gradient (ASVRG) methods and their related variants (Nitanda, 2016; Allen-Zhu, 2017; Kulunchakov & Mairal, 2020).

While much of our understanding of modern optimization algorithms relies on the ability to leverage the convexity of the objective function, a growing number of modern machine learning applications rely on non-convex optimization. Unfortunately, the theoretically guaranteed improvement for convex functions that accelerated algorithms have do not apply to many real-world scenarios. For many smooth non-convex optimization problems, we only have guarantees for finding stationary points instead of the global minimizer. However, some non-convex functions involved in several popular optimization problems such as low-rank matrix problems, deep learning and reinforcement learning, have special structure and exhibit convex-like properties (Ge et al., 2016; Bartlett et al., 2018; Mei et al., 2020), which makes it possible to find approximate global minimizers of these structured non-convex functions.

In this paper, we develop two accelerated stochastic optimization methods for optimizing quasar-convex functions. A quasar-convex function is parameterized by a constant $\gamma \in (0, 1]$. $\gamma = 1$ implies the function is star-convex, which is a relaxation of convexity (Nesterov & Polyak, 2006). Quasar-convexity was first proposed in Hardt et al. (2016). They prove that the objective of learning linear dynamical systems is quasar-convex under several mild assumptions. Zhou et al. (2019) and Kleinberg et al. (2018) also provide evidence to suggest that loss function of neural networks may conform to star-convexity in large neighborhoods of the minimizers. Several recent papers propose effective deterministic methods for minimizing L-smooth and γ-quasar-convex functions. While gradient descent (GD) and stochastic gradient descent (SGD) need $O(\gamma^{-1}\epsilon^{-1})$ and $O(\gamma^{-2}\epsilon^{-2})$ iterations to yield an ϵ-approximate solution Guminov & Gasnikov 2017; Gower et al. 2021, the algorithms developed by Guminov & Gasnikov (2017) and Hinder et al. (2020) need $O(\gamma^{-1}\epsilon^{-1/2})$ iterations and the algorithm developed by Nesterov et al. (2018) needs $O(\gamma^{-3/2}\epsilon^{-1/2})$ iterations. Hinder et al. (2020) also introduce a new metric in terms of the total number of function and gradient evaluations. In order to compute an ϵ-approximate solution, the method of Hinder et al. (2020) requires $O(\gamma^{-1}\epsilon^{-1/2}\log(\gamma^{-1}\epsilon^{-1}))$ total evaluations for γ-quasar-convex functions and $O(\gamma^{-1}\kappa^{1/2}\log(\gamma^{-1}\kappa)\log(\gamma^{-1}\epsilon^{-1}))$[1] total evaluations for μ-strongly γ-quasar-convex functions.

Many optimization problems in machine learning can be

[1]Sun Yat-sen University, Guangzhou, China [2]Harvard University, Cambridge, MA, USA [3]MIT, Cambridge, MA, USA. Correspondence to: Qiang Fu <fuqiang7@mail2.sysu.edu.cn>, Ashia Wilson <ashia07@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] $\kappa \triangleq L/\mu$ is the condition number.

expressed in the following format

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right], \tag{1}$$

which makes them particularly well-suited for stochastic optimization methods. When n is large, applying deterministic algorithms will lead to high computational cost due to the full gradient and function value access required per iteration. Therefore, motivated by ASGD, ASVRG in the convex setting as well as the contributions of Hinder et al. (2020), we propose both a quasar-accelerated stochastic gradient descent (QASGD) method and a quasar-accelerated stochastic variance reduced gradient (QASVRG) method for solving (1), where the objective function f is L-smooth and (strongly) quasar-convex. We also present a unified energy-based framework to analyze the convergence of these newly proposed accelerated algorithms, drawing inspiration from the unified analyses developed by Wilson et al. (2021) and Kulunchakov & Mairal (2020).

Our principal contributions are three-fold.

- *QASGD*: We introduce QASGD with momentum as the acceleration technique. Under a bounded gradient assumption 2.4, we prove that QASGD achieves convergence rates of $O\left(\frac{L}{t^2}+\frac{\sigma}{\gamma\sqrt{t}}+\frac{\epsilon}{2}\right)$ for general quasar-convex functions and $O\left((1+\gamma^2/16)^{-t}+\frac{\sigma^2}{\gamma^2 t}\right)$ for strongly quasar-convex functions, where ϵ comes from a binary line search. We empirically demonstrate that on learning time-invariant dynamical systems, QASGD outperforms several existing proposed methods.

- *QASVRG*: We introduce QASVRG in a mini-batch setting, which is an extension of Nitanda (2016) to quasar-convexity with momentum as the acceleration technique. Variance reduction and mini-batches are employed to compute the stochastic gradient per iteration. Under an interpolation assumption 2.7 and a compactness assumption 2.5, QASVRG achieves an overall complexity[2] of $\widetilde{O}\left(n+\min\left\{\frac{\kappa}{\gamma^2},\frac{n\sqrt{\kappa}}{\gamma}\right\}\right)$ and $\widetilde{O}\left(n+\min\left\{\frac{LR^2}{\gamma\epsilon},\frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right\}\right)$ for strongly quasar-convex functions and general quasar-convex functions. We also propose an alternative scheme for strongly quasar-convex functions with different parameter choice (Option II in Table 4), whose precise convergence rates are postponed to Theorem 3.6. These two schemes have different dependency on κ and ϵ and thus are suitable to different application scenarios. When n is large, our complexity is significantly lower than the complexity of AGD in Hinder et al. (2020).

- *Lyapunov analysis*: We present a unified analysis for our proposed algorithms under quasar-convexity and smoothness using a standard Lyapunov argument. Additionally, we incorporate the AGD method proposed in Hinder et al. (2020) in our energy-based framework, which we rename QAGD (quasar-accelerated gradient descent). Different from AGD in Hinder et al. (2020), QAGD admits the Bregman divergence which is more general than the Euclidean distance.

The remainder of this paper is organized as follows. Section 2 presents more details about quasar-convexity, related assumptions, and previously proposed methods. Section 3 presents the main algorithms of QAGD, QASGD and QASVRG for (strongly) quasar-convex functions and their convergence analysis. Section 4 describes our simulations verifying the effectiveness of our proposed algorithms.

Notation The following notation is used throughout the paper: $D_h(x,y)\triangleq h(x)-h(y)-\langle\nabla h(y),x-y\rangle$ denotes the Bregman divergence between $x,y\in\mathbb{R}^d$, where h is an arbitrary $\bar{\mu}$-strongly convex function. $[n]\triangleq\{1,2,...,n\}$ $\log^+()\triangleq\max\{\log(),1\}$, $\|\cdot\|\triangleq\|\cdot\|_2$, $\bar{a}_k\triangleq A_{k+1}-A_k$, $\bar{b}_k\triangleq B_{k+1}-B_k$, $\kappa\triangleq L/\bar{\mu}\mu$, $\mathcal{E}_k\triangleq f(y_k)-f(x^*)$. $a\simeq b$ signifies $a=O(b)$. $\langle,\rangle$ represents the inner product. $\mathcal{X}^*$ is the solution set of (1) which we assume is not empty, and a point x is an ϵ-approximate solution if $f(x)-f(x^*)\leq\epsilon$ for $x^*\in\mathcal{X}^*$. R denotes the upper bound of the initial distance such that $D_h(x^*,x_0)\leq R^2$. We assume $f(x^*)\geq 0$ without loss of generality. f is L-smooth, if $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$ for all $x,y\in\mathbb{R}^d$. $\mathcal{Q}_{\mu\gamma},\mathcal{F}_L$ respectively denote the set of μ-strongly γ-quasar-convex functions and the set of L-smooth functions, and $\mathcal{Q}_{\mu\gamma}$ reduces to the set of γ-quasar-convex functions when $\mu=0$. We use $O(\cdot)$ to hide constants and $\widetilde{O}(\cdot)$ to hide logarithmic factors and constants.

2. Background

There has been growing interest in exploiting structure present in large classes of non-convex functions. One such structure is quasar-convexity and strong quasar-convexity, defined as follows.

Definition 2.1 (Quasar-convexity)**.** Let $\gamma\in(0,1]$ and let x^* be a minimizer of the differentiable function $f:\mathbb{R}^d\to\mathbb{R}$. A function is γ-quasar-convex with respect to x^* if for all $x\in\mathbb{R}^d$,

$$f(x^*)\geq f(x)+\frac{1}{\gamma}\langle\nabla f(x),x^*-x\rangle. \tag{2}$$

For $\mu>0$, a function is μ-strongly γ-quasar-convex with respect to x^* if for all $x\in\mathbb{R}^d$,

$$f(x^*)\geq f(x)+\frac{1}{\gamma}\langle\nabla f(x),x^*-x\rangle+\frac{\mu}{2}\|x^*-x\|^2. \tag{3}$$

[2] Here we use overall complexity to denote the total number of function and gradient evaluations

2.1. Examples

We introduce a classical example in Hardt et al. (2016) of learning linear dynamical systems (LDS). Consider the following time-invariant linear dynamical system

$$h_{t+1} = Ah_t + Bx_t \tag{4a}$$
$$y_t = Ch_t + Dx_t + \xi_t, \tag{4b}$$

where $x_t \in \mathbb{R}, y_t \in \mathbb{R}$ are the input and output of time t; ξ_t is a random perturbation sampled i.i.d from a distribution; $h_t \in \mathbb{R}^d$ is the hidden state and $\Theta \triangleq (A, B, C, D) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times 1} \times \mathbb{R}^{1\times d} \times \mathbb{R}$ is the true parameter that we aim to learn. Assuming we have N pairs of training examples $S = \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ where each input sequence $x \in \mathbb{R}^T$ is sampled from a distribution and y is the corresponding output of the system above, we fit these training examples to the following model

$$\hat{h}_{t+1} = \hat{A}\hat{h}_t + \hat{B}x_t \tag{5a}$$
$$\hat{y}_t = \hat{C}\hat{h}_t + \hat{D}x_t, \tag{5b}$$

which is governed by $\Theta \triangleq (\hat{A}, \hat{B}, \hat{C}, \hat{D})$. According to the training examples and the model system, we consider the following optimization problem

$$\min \left\{ F(\hat{\Theta}) = \mathbb{E}_{\{x_t\},\{\xi_t\}} \left[\frac{1}{T} \sum_{t=1}^{T} \|\hat{y}_t - y_t\|^2 \right] \right\}. \tag{6}$$

Hardt et al. (2016) demonstrate that the objective function $F(\hat{\Theta})$ is weakly smooth and quasar-convex with respect to Θ under some mild conditions.

We introduce another example of generalized linear models (GLM). Consider the following square loss minimization problem

$$\min \left\{ f(w) := \mathbb{E}_{x\sim\mathcal{D}} \left[\frac{1}{2}(\sigma(w^\top x) - y)^2 \right] \right\} \tag{7}$$

where $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ is the link function; $x \in \mathbb{R}^d$ is i.i.d from $\mathcal{D}$ and there exists $w_* \in \mathbb{R}^d$ such that $y = \sigma({w_*}^\top x)$. The quasar-convex structure of $f(w)$ has been exploited in several literature (Foster et al., 2018; Ma, 2020; Wang & Wibisono, 2023).

2.2. Prior Deterministic Methods

Several deterministic first-order methods have been developed to minimize L-smooth γ-quasar-convex functions.

Guminov & Gasnikov (2017) prove that gradient descent achieves a convergence rate of $O(L/\gamma t)$. They also propose an accelerated algorithm achieving a convergence rate of $O(L/\gamma^2 t^2)$. This algorithm, however, depends on a low-dimensional subspace optimization method at each iteration, which is possibly prohibitively expensive to perform.

Hinder et al. (2020) propose a novel accelerated gradient method achieving a convergence rate of $O((1 - \gamma/\sqrt{2\kappa})^t)$ for strongly quasar-convex functions. Notably, when $\gamma = 1$, this rate matches the convergence rate achieved by Nesterov's AGD for strongly convex functions. The method introduced by Hinder et al. (2020) also achieves a convergence rate of $O(L/\gamma^2 t^2 + \epsilon/2)$ for general quasar-convex functions, which nearly matches the convergence rate achieved by Nesterov's AGD for convex functions when $\gamma = 1$. An additional factor ϵ (which can be made arbitrarily small) appears in the convergence rate due to a binary line search subroutine introduced in order to search for the appropriate momentum parameters. Notably, the momentum parameters classically chosen in accelerated gradient descent (Nesterov 1983) are not guaranteed to perform well under quasar-convexity. Compared with the low-dimensional subspace method in Guminov & Gasnikov (2017), the binary line search in Hinder et al. (2020)'s AGD achieves at most $O(\log(\gamma^{-1}\epsilon^{-1}))$ function and gradient evaluations, which can be considerably cheaper. Analogously, Bu & Mesbahi (2020) propose momentum-based accelerated algorithms relying on a subroutine but without complexity analysis of the subroutine. Hinder et al. (2020) also establishes a worst case complexity lower bound of $\Omega(\gamma^{-1}\epsilon^{-1/2})$ for any deterministic first-order methods applied to quasar-convex functions, and their methods are optimal up to a logarithmic factor. The complexity of methods in Guminov & Gasnikov (2017) conditionally matches this lower bound.

2.3. Prior Stochastic Methods

While deterministic accelerated methods for quasar-convex functions achieve fast convergence rates and near-optimal complexity, we focus on the development of stochastic methods to reduce the computational complexity when solving (1). When the objective f is γ-quasar-convex, Hardt et al. 2016 show that SGD achieves a convergence rate of $O(\Gamma/\gamma^2 t + \bar{\sigma}/\gamma\sqrt{t})$ under the assumptions of $\bar{\sigma}^2$-bounded variance and Γ-weak-smoothness.

Assumption 2.2 (Bounded Variance)**.** Suppose i is sampled i.i.d from $[n]$. For some constant $\bar{\sigma}$, we have

$$\mathbb{E}_i \left[\|\nabla f_i(x) - \nabla f(x)\|^2 \right] \le \bar{\sigma}^2.$$

The weak-smoothness assumption is milder than L-smoothness. Gower et al. (2021) propose a stochastic gradient method achieving a convergence rate of $O(\lambda^2/\gamma\sqrt{t})$ under L-smoothness and Assumption 2.3.

Assumption 2.3 (ER Condition)**.** Suppose i is sampled i.i.d from $[n]$. For some constants ρ and λ, we have

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \le 4\rho(f(x) - f(x^*)) + \|\nabla f(x)\|^2 + 2\lambda^2.$$

Method	Assumptions	Complexity
GD (Guminov & Gasnikov, 2017)	$f \in \mathcal{F}_L$	$\widetilde{O}\left(\frac{nLR^2}{\gamma\epsilon}\right)$
SGD (Gower et al., 2021)	$f \in \mathcal{F}_L$ & ER Condition	$\widetilde{O}\left(\frac{(R^2+\gamma^2\lambda^2)^2}{\gamma^2\epsilon^2}\right)$
	$f_i \in \mathcal{F}_L$ & Interpolation	$\widetilde{O}\left(\frac{LR^2}{\gamma^2\epsilon}\right)$
SGD (Jin, 2020)	$f \in \mathcal{F}_L$ & Bounded Variance	$\widetilde{O}\left(\frac{LR^2}{\gamma\epsilon} + \frac{\bar{\sigma}^2R^2}{\gamma^2\epsilon^2}\right)$
QAGD (Hinder et al., 2020)	$f \in \mathcal{F}_L$	$\widetilde{O}\left(\frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right)$
QASGD (Ours)	$f_i \in \mathcal{F}_L$ & Bounded Gradient	$\widetilde{O}\left(R\sqrt{\frac{L}{\epsilon}} + \frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)$
QASVRG (Ours)	$f_i \in \mathcal{F}_L$ & Interpolation & Compactness	$\widetilde{O}\left(n + \min\left\{\frac{LR^2}{\gamma\epsilon}, \frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right\}\right)$

Table 1. Comparison between some existing methods and our methods when $f \in \mathcal{Q}_{0\gamma}$

Compared with Hardt et al. (2016), the smoothness assumption in Gower et al. (2021) is stronger, but the assumption on the gradient estimate is weaker in a sense. Moreover, Gower et al. (2021) demonstrate that this rate can be improved to $O(L/\gamma^2 t)$ under Assumption 2.7. Under smoothness and bounded variance, Jin (2020) provides a sharper analysis of SGD compared with Gower et al. (2021) and extends the analysis to the non-smooth setting.

There are several accelerated stochastic methods that can theoretically achieve better worst-case convergence rates than SGD when the objective is convex. In the convex setting, the objective function usually includes a regularizer term, which is convex lower semi-continuous and not necessarily smooth. Ghadimi & Lan (2016) and Kulunchakov & Mairal (2020) propose proximal ASGD which achieves convergence rates of $O(L/t^2 + \sigma/\sqrt{t})$ and $O((1-1/\sqrt{\kappa})^t + \sigma^2/t)$ for general convex and strongly convex functions respectively under L-smoothness and the σ^2-bounded variance assumption. Variance reduction is a powerful technique to achieve a better convergence rate. Allen-Zhu (2017) and Kulunchakov & Mairal (2020) propose accelerated proximal SVRG with a convergence rate guarantee of $O((1-\min\{1/\sqrt{3\kappa n}, 1/\sqrt{2n}\})^t)$ and $O(Ln/t^2)$ for L-smooth (strongly) convex functions. Nitanda (2016) proposes accelerated mini-batch SVRG methods for minimizing (strongly) convex finite sum without regularizer. This algorithm is a multi-stage scheme achieving convergence rates of $\widetilde{O}(n + \min\{\kappa, n\sqrt{\kappa}\})$ and $\widetilde{O}(n + \min\{L/\epsilon, n\sqrt{L/\epsilon}\})$ for L-smooth (strongly) convex functions. By contrast, the methods they use to update the fixed anchor point of SVRG and control the variance are different.

2.4. Motivation

Inspired by the accelerated stochastic methods discussed above, we extend ASGD of Kulunchakov & Mairal (2020) and AMSVRG of Nitanda (2016) to the (strongly) quasar-convex setting under different assumptions. In this subsection we will discuss these assumptions and how they are compared to prior sets of assumptions. Different from Kulunchakov & Mairal (2020), we do not consider random perturbations of the function value and gradient given x^* may not be the global minimizer after perturbation. Furthermore, binary line search (Hinder et al., 2020) is incorporated into each of our proposed methods for finding the appropriate momentum parameters.

For QASGD, our key assumption is the bounded gradient assumption, which is a frequently used assumption in the standard convergence analysis of SGD in the non-convex setting (Hazan & Kale, 2014; Rakhlin et al., 2011; Recht et al., 2011; Nemirovski et al., 2009). Due to the special structure of strongly quasar-convex functions whose gradient is not bounded, we generalize this assumption as follows.

Assumption 2.4 (Bounded Gradient)**.** Suppose $f \in \mathcal{Q}_{\mu\gamma}$ and i is sampled i.i.d from $[n]$. For some $\sigma \geq 0$ and $x^* \in \mathcal{X}^*$, we have

$$\mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] \leq \sigma^2 + 2\mu^2\|x^* - x\|^2.$$

This assumption will reduce to the standard bounded gradient assumption under general quasar-convexity. Compared with ER condition, the bounded gradient assumption is stronger. Example (7) satisfies this assumption ($\mu = 0$) if we choose the link functions to be logistic. For $\mu > 0$, we consider a quasar-convex finite sum $f = \sum_{i=1}^n f_i$ where x^* is the minimizer of f and $\mathbb{E}_i[\|\nabla f_i\|^2] \leq \sigma^2$. Then $g(x) = f(x) + \frac{\mu}{2}\|x - x^*\|^2$ is strongly quasar-convex and satisfies this assumption. While some quasar-convex functions intrinsically do not satisfy Assumption 2.4, it will hold in practice under Assumption 2.5, which was also proposed in Bottou & Le Cun (2005), Gürbüzbalaban et al. (2015) and Nitanda (2016) to analyze the incremental and stochastic methods. We summarize the relation of four assumptions above in Remark 2.6.

Assumption 2.5 (Compactness)**.** There exists a compact set $\mathcal{C} \subseteq \mathbb{R}^d$ containing iterates generated by some optimization algorithm.

Remark 2.6. The relation of Assumption 2.3 (ER), Assumption 2.2 (BV), Assumption 2.4 (BG) and Assumption 2.5 (Compactness) is illustrated as follows.

$$\begin{tikzcd} \boxed{\text{Compactness}} \arrow[r] & \boxed{\text{BG}} \arrow[r] & \boxed{\text{BV}} \arrow[r] & \boxed{\text{ER}} \end{tikzcd}$$

For QASVRG, we will prove in the next section that the upper bound of the gradient variance introduced by Nitanda (2016) also upper bounds the gradient variance of quasar-convex functions (Proposition 3.4) provided that each f_i in problem (1) is L_i-smooth and satisfies the following interpolation assumption.

Assumption 2.7 (Interpolation)**.** There exists $x^* \in \mathcal{X}^*$ such that for all $i \in [n]$

$$\min_{x \in \mathbb{R}^d} f_i(x) = f_i(x^*).$$

The interpolation assumption is commonly observed in the over-parameterized machine learning models and has attracted much attention recently (Zhou et al., 2019; Ma et al., 2018; Vaswani et al., 2019; Gower et al., 2021). If the model is sufficiently over-parameterized, it can interpolate the labelled training data completely. Particularly, example (6) satisfies the interpolation assumption when $\xi_t = 0$, as Θ is also the global minimizer of the objective function generated by each training example. Example (7) also satisfies this assumption given that $y = \sigma(w_*^\intercal x)$ for each $x \sim \mathcal{D}$. Let $L = \max_i\{L_i\}$; we assume throughout that each f_i is L-smooth for brevity. Moreover, Nitanda (2016) only presents one parameter choice for both convex and strongly convex functions. In this paper, we provide two parameter choices (Option I and II) under strong quasar-convexity. Similarly, Option II in Table 4 is identical to the parameter choice of general quasar-convex functions. Option I is a slightly different method from the direct extension of Nitanda (2016)'s AMSVRG. More technical comparison between these two parameter choices are provided in subsection 3.2.

3. Algorithms

In order to solve (1), QAGD, QASGD and QASVRG need to access the gradient or the gradient estimate from the oracle, which we denote ∇_k. In this paper, we consider the following gradient (estimates):

- **Full Gradient**: $\nabla_k = \nabla f(x_{k+1})$. Problem (1) becomes deterministic.
- **Stochastic Gradient**: $\nabla_k = \nabla f_i(x_{k+1})$ with the index i sampled i.i.d from $[n]$. We have $\mathbb{E}_i[\nabla_k] = \nabla f(x_{k+1})$, where $\mathbb{E}_i$ denotes the expectation with respect to the index i.
- **Mini-batch SVRG**: $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(\tilde{x}) + \nabla f(\tilde{x})$, where $\tilde{x}$ is the anchor point fixed per stage; $I_k = \{i_1, i_2, ..., i_{b_k}\}$ is sampled i.i.d from $[n]$ with $f_{I_k} \triangleq \frac{1}{b_k}\sum_{j=1}^{b_k} f_{i_j}$. Batchsize $|I_k| = b_k$. We have $\mathbb{E}_{I_k}[\nabla_k] = \nabla f(x_{k+1})$, where $\mathbb{E}_{I_k}$ denotes the expectation with respect to the mini-batch I_k.

3.1. Quasar-accelerated Algorithms

We introduce Algorithm 1 as a general framework incorporating QAGD, QASGD and a single stage of QASVRG with different parameter choices. Based on Algorithm 1, we also introduce the multi-stage QASVRG as described in Algorithm 2. Notably, we provide more information about Bisearch (line 3 of Algorithm 1) in the Appendix including the whole algorithm and the corresponding complexity analysis obtained from Hinder et al. (2020) (Algorithm 3, Lemma A.5). The guaranteed performance of our methods relies on the internal assumption 3.1. Relation (8) requires h is $\bar{\mu}$-strongly convex; relation (9) is a generalization of μ-strongly γ-quasar-convexity using Bregman divergence as the distance, which we will substitute (3) with in the following analysis. When $h = \frac{1}{2}\|\cdot\|^2$, this relation will be identical to (3). Relation (10) holds in the Euclidean setting given that f is μ-strongly γ-quasar-convex with respect to x^* (Hinder et al. (2020), Corollary 1).

Assumption 3.1. Suppose f_i is differentiable for each $i \in [n]$. For some $\bar{\mu} > 0$, $0 < \gamma \le 1$ and $\mu \ge 0$, and for all $x, y \in \mathbb{R}^d$, require

$$D_h(x,y) \ge \frac{\bar{\mu}}{2}\|x-y\|^2, \tag{8}$$

$$f(x^*) \ge f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle + \mu D_h(x^*, x), \tag{9}$$

$$f(x) \ge f(x^*) + \frac{\gamma\mu}{2-\gamma} D_h(x^*, x). \tag{10}$$

3.2. Convergence Analysis

We develop a unified analysis of QAGD, QASGD and QASVRG using the following Lyapunov function:

$$E_k \triangleq A_k(f(y_k) - f(x^*)) + B_k D_h(x^*, z_k), \tag{11}$$

where A_k and B_k are positive non-decreasing sequences that the parameter choices shown in Table 2, Table 3 and Table 4 are highly related to. Based on the convergence rates derived by Lyapunov analysis, we deduce the complexity upper bound of each method in the Euclidean setting ($h = \frac{1}{2}\|\cdot\|^2$). Since the convergence results of QAGD have

```
**Algorithm 1** $(A_k, B_k, \tilde{y}_0, t, \epsilon)$
**Require:** $h$ satisfies $D_h(x,y) \geq \frac{\bar{\mu}}{2}\|x-y\|^2$; $\tilde{f} \in \mathcal{F}_L$; $f \in \mathcal{Q}_{\mu\gamma}$.
Initialize $x_0 = z_0 = y_0 = \tilde{y}_0$ and specify $\theta_k \triangleq (\nabla_k, \alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon})$
for $k = 0, ..., t-1$ do
    $\tau_k \leftarrow$ Bisearch$(\tilde{f}, y_k, z_k, b, c, \tilde{\epsilon})$    (search $\tau_k \in [0,1]$ satisfying $\tau_k\langle\nabla\tilde{f}(d_k), y_k - z_k\rangle - b\|d_k - z_k\|^2 \leq c(\tilde{f}(y_k) - \tilde{f}(d_k)) + \tilde{\epsilon}$ with $d_k := \tau_k y_k + (1-\tau_k)z_k$; see Algorithm 3 for more details)
    $x_{k+1} \leftarrow (1-\tau_k)z_k + \tau_k y_k$    (coupling step)
    $z_{k+1} \leftarrow \arg\min_{z\in\mathbb{R}^d} \{\langle\nabla_k, z - z_k\rangle + \beta_k D_h(z, z_k) + \alpha_k D_h(z, x_{k+1})\}$    (mirror descent step)
    $y_{k+1} \leftarrow \arg\min_{y\in\mathbb{R}^d} \{\rho_k\langle\nabla_k, y - x_{k+1}\rangle + \frac{1}{2}\|y - x_{k+1}\|^2\}$    (gradient descent step)
end for
output $y_t$
```

```
**Algorithm 2** $(A_k, B_k, b_k, \tilde{y}_0, p, q, \epsilon)$
**Require:** $D_h(x,y) \geq \frac{\bar{\mu}}{2}\|x-y\|^2$; $f_i \in \mathcal{F}_L$; $f \in \mathcal{Q}_{\mu\gamma}$; $q \leq \frac{1}{4}$; $p \leq \frac{\gamma\bar{\mu}}{16}$
Initialize $y_0 = \tilde{y}_0$
for s=0,1,... do
    $t \leftarrow \begin{cases} \sqrt{\frac{17LD_h(x^*,y_s)}{\gamma^2\bar{\mu}qf(y_s)}}, & \mu = 0 \text{ or } \mu > 0 \text{ (Opt II)} \\ \log_{1+\frac{\gamma}{\sqrt{8\kappa}}}\left(\frac{2}{\gamma q}\right), & \mu > 0 \text{ (Opt I)} \end{cases}$
    $y_s \leftarrow$ Algorithm 1 $(A_k, B_k, y_0, \lceil t\rceil, \epsilon)$    (specify $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ where $|I_k| = b_k$)
    $y_0 \leftarrow y_s$
end for
output $y_s$
```

already been established in Hinder et al. (2020), we will not provide the convergence analysis of QAGD in this subsection. Instead, we make convergence analysis of QAGD in Lyapunov framework and obtain results matching Hinder et al. (2020). Relevant proofs and parameter choices are provided in Appendix C and D.

Theorem 3.2 (QASGD)**.** *Suppose Assumption 3.1 and Assumption 2.4 hold, $D_h(x^*, z_0) \leq R^2$, $f_i \in \mathcal{F}_L$ for all $i \in [n]$ and choose any $\tilde{y}_0 \in \mathbb{R}^d$. Then Algorithm 1 with the choices of $\nabla_k = \nabla f_i(x_{k+1})$ and A_k, B_k, θ_k specified in Table 3 satisfies*

$$\mathbb{E}\left[\mathcal{E}_t\right] \simeq \begin{cases} \dfrac{LR^2}{t^2} + \dfrac{\sigma R}{\gamma\sqrt{t}} + \dfrac{\epsilon}{2}, & \mu = 0, \\ \left(1 + \min\left\{\dfrac{\gamma^2\bar{\mu}^2}{16}, \dfrac{1}{2}\right\}\right)^{-t} E_0 + \dfrac{\sigma^2}{\gamma^2 t}, & \mu > 0. \end{cases}$$

Corollary 3.3. *Consider QASGD under the same assumption in Theorem 3.2. Then the overall complexity of QASGD to achieve $\mathbb{E}\left[f(y_t) - f(x^*)\right] \leq \epsilon$ is upper bounded by*

$$O\left(R\sqrt{\frac{L}{\epsilon}}\log^+\left(\frac{LR^2}{\gamma\epsilon}\right) + \frac{\sigma^2 R^2}{\gamma^2\epsilon^2}\log^+\left(\frac{LR^2}{\gamma\epsilon}\right)\right),$$

$$O\left(\frac{1}{\gamma^2}\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right)\log\left(\frac{\mathcal{E}_0}{\gamma\epsilon}\right) + \frac{\sigma^2}{\gamma^2\epsilon}\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$$

for $\mu = 0$ and $\mu > 0$ respectively.

The following proposition shows the variance of the gradient estimate of QASVRG reduces as fast as the objective, which is a key technique in our proof to control the stochastic gradient variance. This proposition is also proposed in Nitanda (2016) where they assume f_i is convex and smooth. In this paper, we circumvent the convexity of f_i by using Assumption 2.7. The proof of Proposition 3.4 is postponed to Appendix B.

Proposition 3.4 (Variance upper bound)**.** *Suppose Assumption 2.7 holds, $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(\tilde{x}) + \nabla f(\tilde{x})$ and $f_i \in \mathcal{F}_L$ for each $i \in [n]$, then we obtain the following inequality*

$$\begin{aligned} &\mathbb{E}_{I_k}\|\nabla_k - \nabla f(x_{k+1})\|^2 \\ &\quad \leq 4L\frac{n - b_k}{b_k(n-1)}\left(f(x_{k+1}) - f(x^*) + f(\tilde{x}) - f(x^*)\right), \end{aligned}$$

where $|I_k| = b_k$ and $\mathbb{E}_{I_k}$ denotes the expectation with respect to the mini-batch I_k.

Theorem 3.5 (QASVRG (Single-stage))**.** *Suppose Assumption 3.1 and Assumption 2.7 hold, $D_h(x^*, z_0) \leq R^2$, $f_i \in \mathcal{F}_L$ for each $i \in [n]$ and choose any $\tilde{y}_0 \in \mathbb{R}^d$. Then Algorithm 1 with the choices of $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ and A_k, B_k, b_k, θ_k specified in Table*

4 satisfies

$$\mathbb{E}\left[\mathcal{E}_t\right] \simeq \begin{cases} \dfrac{LR^2}{\gamma^2 t^2} + \left(\dfrac{p}{\gamma} + \epsilon\right) f(y_0), & \mu = 0, \\ \left(1 + \dfrac{\gamma}{\sqrt{8\kappa}}\right)^{-t} E_0 + \dfrac{\mathcal{E}_0}{2}, & \mu > 0 \; (\mathit{Opt\ I}), \\ \dfrac{LR^2}{\gamma^2 t^2} + \left(\dfrac{p}{\gamma} + \epsilon\right) f(y_0), & \mu > 0 \; (\mathit{Opt\ II}), \end{cases}$$

where $p \le \frac{\gamma\bar{\mu}}{16}$ is user specified.

Under Assumption 2.5, $\{D_h(x^*, y_s)\}$ can be uniformly bounded by some constant if $\{y_s\}$ generated by Algorithm 2 are restricted to a compact set. Thus we can hide the Bregman divergence inside $O(\cdot)$ and $\widetilde{O}(\cdot)$. Note that we only need this assumption when $\mu = 0$. According to Theorem 3.5, single-stage QASVRG is biased which means that an ϵ-approximate solution can not be generated via single-stage QASVRG. For instance, the bias is upper bounded by $O\left(\left(\frac{p}{\gamma} + \epsilon\right) f(y_0)\right)$ when $\mu = 0$, but the expectation of the optimality gap $\mathcal{E}_t$ can shrink at each stage with small p and q when $t = \Omega\left(\sqrt{\frac{LR^2}{\gamma^2 f(y_0)}}\right)$. Consequently we need $O(\log(1/\epsilon))$ stages to generate an ϵ-approximate solution.

Corollary 3.6. *Under Assumption 2.5 and the same assumptions in Theorem 3.5, the overall complexity required for Algorithm 2 to achieve $\mathbb{E}\left[f(y_s) - f(x^*)\right] \le \epsilon$ is upper bounded by*

$$O\left(\left(n + \frac{nLR^2}{\gamma\epsilon n + \gamma\sqrt{\epsilon L R^2}} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right) \log\left(\frac{1}{\epsilon}\right)\right),$$

$$O\left(\left(n + \frac{n\kappa}{\gamma^2 n + \gamma\sqrt{\kappa}} \log\left(\frac{2}{\gamma q}\right) \log^+\left(\frac{\kappa^{7/6}}{\gamma}\right)\right) \log\left(\frac{1}{\epsilon}\right)\right),$$

$$O\left(\left(n + \frac{n\kappa}{\gamma^2 n + \gamma^{3/2}\sqrt{\kappa}} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right) \log\left(\frac{1}{\epsilon}\right)\right)$$

for $\mu = 0$ and $\mu > 0$ (the last two bounds) respectively, where $q \in (0, 1/4]$.

Proof Sketch The method we use to derive the convergence rates and complexity for each algorithm is unified. We take the difference between each Lyapunov stage, and then take the conditional expectation to obtain the upper bound on $\mathbb{E}[E_{k+1} - E_k]$. Finally we sum over k to conclude the convergence rates and a subsequent iteration complexity for each algorithm. Combining this complexity with that of Bisearch, we conclude the overall complexity in each corollary. For more details, see Appendix C and D.

The last two bounds of QASVRG correspond to two different choices of parameters in Table 4 (Option I and II). The complexity bound derived with Option I has a more unfavorable dependency on κ while the complexity bound derived with Option II has a more unfavorable dependency on ϵ. This suggests Option I performs better on well-conditioned problems e.g. $\kappa\epsilon < 1$ and Option II performs better on ill-conditioned problems e.g. $\kappa\epsilon \gg 1$.

In the complexity bounds of QASGD and QASVRG, extra logarithmic factors are included, which comes from Bisearch. Hinder et al. (2020) prove that the complexity of Bisearch is at most a logarithmic factor given that the function involved in this subroutine is L-smooth. In the stochastic setting, where the functions involved are single f_i or a mini-batch of f_{I_k}, we need to assume $f_i \in \mathcal{F}_L$ for all i due to the uniform sampling.

We summarize our methods and some existing methods in Table 1, including their corresponding assumptions and complexity upper bounds. To summarize, both QASGD and QASVRG achieve better complexity upper bounds than QAGD when n is large, and QASGD enjoys a faster convergence rate and lower complexity than SGD under a stronger assumption. While QASVRG has the potential to be more computationally expensive than SGD due to the full gradient and function value access once a stage, it enjoys a theoretically faster convergence rate than SGD.

4. Simulations

In this section, we evaluate our methods on example (6) in the Euclidean setting using synthetic dataset where each input sequence $x^{(i)} \sim \mathcal{N}(0, 1)$ coordinate-wise. Different from Hardt et al. (2016), we generate N training examples and random perturbations ξ_t before training instead of generating fresh data and random perturbations at each iteration. Thus example (6) can be reformulated as

$$\min\left\{F(\hat{\Theta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{y}_t^{(i)} - y_t^{(i)}\right\|^2\right]\right\},$$

where the superscript (i) represents that the output is generated using i^{th} training data $(x^{(i)}, y^{(i)})$. Similar to Hardt et al. (2016), the actual objective in our experiments is $F(\hat{\Theta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T-T_1}\sum_{t>T_1}\|\hat{y}_t^{(i)} - y_t^{(i)}\|^2\right]$, where $T_1 = T/4$. We generate the true dynamical system and data the same way as in Hardt et al. (2016) using parameters $N = 5000, d = 20, T = 500$. Following Hinder et al. (2020), we generate the initial iterate $(\hat{A}_0, \hat{C}_0, \hat{D}_0)$ by perturbing the parameters of the true system and keep the spectral radius of $\hat{A}_0$ strictly less than 1. We choose the value of random seed to be in $\{0, 12, 24, 36, 48\}$ for generating five true LDS instances and their initialization. We only present simulation results of $\{0, 24, 48\}$ in this sections and the remaining results are provided in section F. Note that $\hat{B}$ is not a trainable parameter since B is known. As is described in Hardt et al. (2016), it is intractable to calculate the precise value of quasar-convexity parameter γ of LDS objective or even estimate it. Thus we evaluate our methods

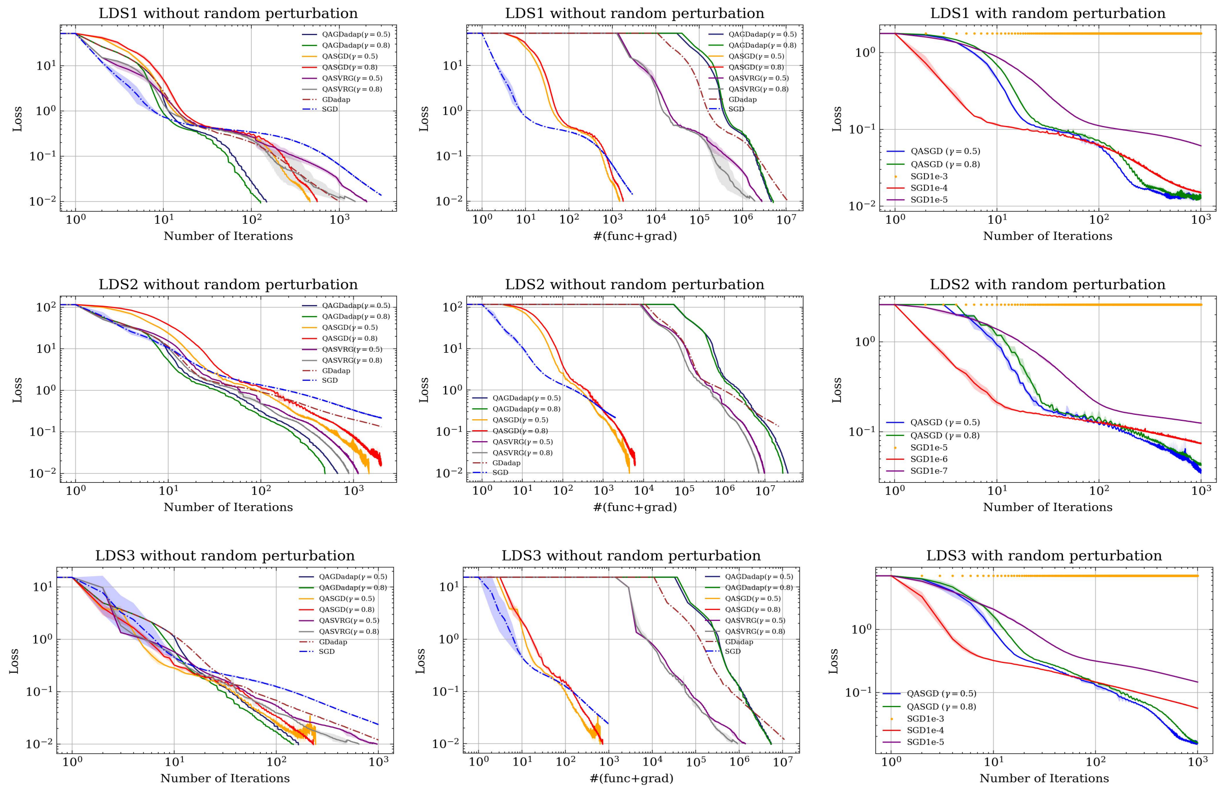

Figure 1. Evaluation on three different LDS instances. We choose $\epsilon = 10^{-2}$, the stepsize to be $5 \times 10^{-5}, 1 \times 10^{-6}, 1 \times 10^{-4}$ for SGD, $L = 1 \times 10^6, 1 \times 10^8, 1 \times 10^5$ for QASGD and $L = 3 \times 10^4, 1 \times 10^6, 1 \times 10^4$ for QASVRG in LDS1, LDS2 and LDS3. The flat line in the third column means the loss blows up to infinity with this choice of stepsize.

with $\gamma \in \{0.5, 0.8\}$. We observe that the imprecise γ does not affect the performance of all methods involved. As is shown in Table 3, η is involved in the parameter choice of QASGD, which we choose to be $\min\left\{\frac{1}{L}, \frac{\sqrt{2}\gamma\|z_0\|}{\sigma(t+1)^{3/2}}\right\}$ in our simulations. While σ in Assumption 2.4 relies on the compact set and initialization, we observe that choosing $\sigma = 1$ is a robust choice for all of our simulations by evaluating QASGD on three LDS instances with $\sigma \in \{1, 10, 10^2, 10^3\}$ (See Figure 3 in Appendix F for more details). According to the analysis in Hardt et al. (2016), $F(\hat{\Theta})$ is L-weakly smooth, and it is still unknown whether $F(\hat{\Theta})$ is L-smooth. Given that the parameter choice of our methods involves L, we fine-tune the value of L for QASGD and QASVRG and choose the best stepsize for SGD by extensive grid search in each instance. We use the adaptive stepsize for QAGD and GD the same way as in Hinder et al. (2020). We consider the random noise $\xi_t \sim \mathcal{N}(0, 10^{-2})$ or $\xi_t = 0$ perturbing the output of the true systems. If $\xi_t \sim \mathcal{N}(0, 10^{-2})$, the interpolation assumption will be violated since Θ is no longer the global minimizer of the objective generated by each training example. Thus we only evaluate SGD and QASGD in this case. In Algorithm 2, it may be difficult to calculate t especially when L is unknown. Therefore we can specify t to be relatively large (we choose $t = 10^4$) and use an appropriate restart scheme in Algorithm 1 to boost the performance of QASVRG. Following Nitanda (2016), when the relation $\langle \nabla_k, y_{k+1} - y_k \rangle > 0$ holds, we break Algorithm 1 to return y_s and start the next stage. Since we generate the initial iterate with $\rho(\hat{A}_0) < 1$, we don't use gradient clipping or projection proposed in Hardt et al. (2016) during training. We generate the error bar in Figure 1 by averaging the results obtained from running each stochastic algorithm three times and choose the maximum and minimum value pointwise to be the upper bar and lower bar.

The simulation results in Figure 1 validate our methods and show the superiority of our methods in terms of the convergence speed and the overall complexity. In addition, both QASGD and QASVRG are robust to the random sampling of the stochastic gradient. Code is available at https://github.com/QiangFu09/Stochastic-quasar-convex-acceleration. There is an interesting phenomenon in our simulations. While the convergence rate of QASGD matches the rate of SGD when t is large, Figure 1 still shows the substantial superiority of QASGD over QASVRG. In fact, QASGD enjoys a convergence rate of $O\left(\frac{1}{t^2} + \frac{1}{\sqrt{t}} + \frac{\epsilon}{2}\right)$

indicating rapid initial phase where $O\left(\frac{1}{t^2}\right)$ dominates the convergence. We speculate that QASGD in most of our simulations does not escape the initial phase and thus enjoys a fast convergence. The simulation results above lead to a reconsideration about whether we need the bounded gradient assumption in QASGD or not, as the objective of LDS does not satisfy this assumption but the performance of QASGD on LDS is still favorable and robust. We believe that this assumption is used in the theoretical analysis but may not be necessary.

5. Disscussion

In this paper, we propose QASGD and QASVRG achieving fast convergence and low complexity under their corresponding assumptions. We present our algorithms in a unified framework using a single energy-based analysis to establish the convergence rates and complexity of QAGD, QASGD and QASVRG. We close with a brief discussion of some possible future work.

First, we introduced the bounded gradient assumption for QASGD, but it remains to be seen whether we can weaken this assumption to some extent. Given that Gower et al. (2021) and Jin (2020) establish the convergence of SGD for quasar-convex functions under the ER condition and bounded variance assumption respectively, it would be of interest to see whether it is possible to apply these weaker assumptions in QASGD.

Second, QAGD are proven near-optimal in Hinder et al. (2020) based on the lower bound they establish for first-order deterministic methods. In future work we hope to establish a worst case complexity lower bound for first-order stochastic methods applied to quasar-convex functions under different assumptions. We expect such bounds will prove that our methods are nearly optimal as well.

Moreover, a higher order method usually leads to a better convergence rate. Nesterov & Polyak (2006) propose the cubic regularized Newton method to optimize star-convex functions ($\gamma = 1$). Therefore, we are also interested in whether it is possible to use higher order methods to improve the convergence of our methods.

Finally, we hope to exploit more applications of quasar-convex functions in machine learning.

Acknowledgements

We would like to thank our anonymous reviewers for constructive comments, thank Kyurae Kim (University of Pennsylvania) for helpful discussions on citations and thank Oliver Hinder (University of Pittsburgh) and Nimit Sohoni (Stanford University) for sharing codes of LDS.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Bartlett, P., Helmbold, D., and Long, P. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pp. 521–530. PMLR, 2018.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bottou, L. and Le Cun, Y. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005. doi: https://doi.org/10.1002/asmb.538. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.538.

Bu, J. and Mesbahi, M. A note on nesterov's accelerated method in nonconvex optimization: a weak estimate sequence approach. *arXiv preprint arXiv:2006.08548*, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

Gower, R., Sebbouh, O., and Loizou, N. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Guminov, S. and Gasnikov, A. Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*, 2017.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A globally convergent incremental newton method. *Mathematical Programming*, 151(1):283–313, 2015.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Hinder, O., Sidford, A., and Sohoni, N. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pp. 1894–1938. PMLR, 2020.

Jin, J. On the convergence of first order methods for quasar-convex optimization. *arXiv preprint arXiv:2010.04937*, 2020.

Kleinberg, B., Li, Y., and Yuan, Y. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Kulunchakov, A. and Mairal, J. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *The Journal of Machine Learning Research*, 21(1):6184–6235, 2020.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Ma, T. Why do local methods solve nonconvex problems?, 2020.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Nesterov, Y. and Polyak, B. T. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient descent with line search for convex and nonconvex optimization problems. *arXiv preprint arXiv:1809.05895*, 2018.

Nitanda, A. Accelerated stochastic gradient descent for minimizing finite sums. In *Artificial Intelligence and Statistics*, pp. 195–203. PMLR, 2016.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Recht, B., Re, C., Wright, S., and Niu, F. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

Wang, J.-K. and Wibisono, A. Continuized acceleration for quasar convex functions in non-convex optimization. *arXiv preprint arXiv:2302.07851*, 2023.

Wilson, A. C., Recht, B., and Jordan, M. I. A lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22:113–1, 2021.

Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

A. Helpful Lemmas and Assumptions

Lemma A.1 (Three-point identity)**.** *For all $x \in$ dom h and $y, z \in$ int(dom h)*

$$D_h(x,y) - D_h(x,z) = -\langle \nabla h(y) - \nabla h(z), x - y\rangle - D_h(y,z), \tag{12}$$

where $D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle$.

Lemma A.2 (Fenchel-Young inequality)**.** *For all $x, y \in \mathbb{R}^d$ and $p \neq 0$, we have*

$$\langle x, y\rangle + \frac{1}{p}\|y\|^p \geq -\frac{p-1}{p}\|x\|_*^{\frac{p}{p-1}}, \tag{13}$$

where $\|\cdot\|_*$ denotes the conjugate norm of $\|\cdot\|$.

Lemma A.3 (Hinder et al. (2020), Lemma 2)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and let $y, z \in \mathbb{R}^d$. For $\tau \in \mathbb{R}$ define $x_\tau \triangleq \tau y + (1-\tau)z$. For any $c \geq 0$ there exists $\tau \in [0,1]$ such that*

$$\tau\langle \nabla f(x_\tau), y - z\rangle \leq c\left(f(y) - f(x_\tau)\right). \tag{14}$$

In fact, we can slacken condition (14) to some extent, and we can find an appropriate momentum parameter τ satisfying

$$\tau\langle \nabla f(x_\tau), y - z\rangle - \tau^2 b\|y - z\|^2 \leq c\left(f(y) - f(x_\tau)\right) + \tilde{\epsilon}, \tag{15}$$

for $b, c, \tilde{\epsilon} \geq 0$. Existence of τ satisfying condition (14) implies the existence of τ satisfying condition (15).

Lemma A.4. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then for any $x, y \in \mathbb{R}^d$*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{16}$$

Now We introduce an algorithm to effectively search the "good" τ for any L-smooth functions.

Lemma A.5 (Hinder et al. (2020), C.2)**.** *For L-smooth $f : \mathbb{R}^d \to \mathbb{R}$, $x, v \in \mathbb{R}^n$ and scalars $b, c, \tilde{\epsilon} \geq 0$, Algorithm (3) computes $\alpha \in [0,1]$ satisfying (15) with at most*

$$6 + 3\left\lceil \log_2^+\left((4+c)\min\left\{\frac{2L^3}{b^3}, \frac{L\|y - z\|^2}{2\tilde{\epsilon}}\right\}\right)\right\rceil$$

function and gradient evaluations, where $\log^+(\cdot) \triangleq \max\{\log(\cdot), 1\}$.

Algorithm 3 Bisearch($f, y, z, b, c, \tilde{\epsilon}$,[guess])

```
Require: $f$ is $L$-smooth; $z, y \in \mathbb{R}^d$; $c \geq 0$; "guess" $\in [0,1]$ if provided. "guess" can be the momentum parameter under convexity or other.
Define $g(\alpha) \triangleq f(\tau y + (1-\tau)z)$ and $p \triangleq b\|z - y\|^2$
if guess provided and $cg(\text{guess}) + \text{guess} \cdot (g'(\text{guess}) - \text{guess} \cdot p) \leq cg(1)$ then return guess
if $g'(1) \leq p + \tilde{\epsilon}$ then return 1;
else if $c = 0$ or $g(0) \leq g(1) + \tilde{\epsilon}/c$ then return 0;
$\delta \leftarrow 1 - g'(1)/L\|z - y\|^2$
lo$\leftarrow 0$, hi$\leftarrow \delta$, $\tau \leftarrow \delta$
while $cg(\tau) + \tau(g'(\tau) - \tau p) > cg(1) + \tilde{\epsilon}$ do
    $\tau \leftarrow (\text{lo} + \text{hi})/2$
    if $g(\tau) \leq g(\delta)$ then hi $\leftarrow \tau$
    else lo$\leftarrow \tau$
end while
output $\tau$
```

Proposition A.6. *Based on Assumption 3.1, we have $\kappa \geq \frac{\gamma}{2-\gamma}$, where $\kappa = \frac{L}{\hat{\mu}\mu}$ and $\mu > 0$.*

Proof.

$$\frac{\gamma\mu}{2-\gamma} D_h(x^*, x) \le f(x) - f(x^*) \le \frac{L}{2}\|x^* - x\|^2 \le \frac{L}{\bar{\mu}} D_h(x^*, x)$$

Thus, we have $\frac{\gamma\mu}{2-\gamma} \le \frac{L}{\bar{\mu}}$. Furthermore, we have $\sqrt{\kappa} \ge \sqrt{\frac{\gamma}{2-\gamma}} \ge \sqrt{\gamma^2} = \gamma$. □

Lemma A.7. *Suppose Assumption 2.7 holds, and each $f_i \in \mathcal{F}_L$. If $\nabla_k = \nabla f_i(x_{k+1}) - \nabla f_i(y_k) + \frac{1}{n}\sum_{i=1}^n f_i(y_k)$, we can obtain*

$$\mathbb{E}_i\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] \le 4L\left(f(x_{k+1}) - f(x^*) + f(y_k) - f(x^*)\right) \tag{17}$$

Proof. Since f is L-smooth, for any $x, y \in \mathbb{R}^d$ we have

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \triangleq g(x)$$

Let $\nabla g(\tilde{x}) = 0$, and $\tilde{x} = y - \frac{1}{L}\nabla f(y)$ is the minimizer of $g(x)$. And we have

$$f(x^*) \le f(x) \le g(\tilde{x}) = f(y) - \frac{1}{2L}\|\nabla f(y)\|^2 \Rightarrow \|\nabla f(y)\|^2 \le 2L(f(y) - f(x^*))$$

Since $\nabla f(x^*) = 0$, we have

$$\|\nabla f(y) - \nabla f(x^*)\| \le 2L(f(y) - f(x^*)) \tag{18}$$

By (18), we can upper bound $\mathbb{E}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right]$ as follows

$$\begin{aligned}
\mathbb{E}_i\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] &= \mathbb{E}_i\left[\|\nabla f_i(x_{k+1}) - \nabla f_i(y_k) - \mathbb{E}_i[\nabla f_i(x_{k+1}) - \nabla f_i(y_k)]\|^2\right] \\
&\le \mathbb{E}_i\left[\|\nabla f_i(x_{k+1}) - \nabla f_i(y_k)\|^2\right] \\
&= \mathbb{E}_i\left[\|\nabla f_i(x_{k+1}) - \nabla f(x^*) + \nabla f(x^*) - \nabla f_i(y_k)\|^2\right] \\
&\le 2\mathbb{E}_i\left[\|\nabla f_i(x_{k+1}) - \nabla f(x^*)\|^2\right] + 2\mathbb{E}_i\left[\|\nabla f(x^*) - \nabla f_i(y_k)\|^2\right] \\
&\le 4L\mathbb{E}_i[f_i(x_{k+1}) - f_i(x^*)] + 4L\mathbb{E}_i[f_i(y_k) - f_i(x^*)] \\
&= 4L\left(f(x_{k+1}) - f(x^*) + f(y_k) - f(x^*)\right),
\end{aligned}$$

where the first inequality uses the relation $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$ for all random variable X; the second inequaity uses the relation $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. □

Lemma A.8. *Let $\{\xi_i\}_{i=1}^n$ be a set of vectors in $\mathbb{R}^d$ and $\bar{\xi}$ denote an average of $\{\xi_i\}_{i=1}^n$. Let I denote a uniform random variable representing a subset of $\{1, 2, ..., n\}$ with its size equal to b. Then, it follows that,*

$$\mathbb{E}_I\left\|\frac{1}{b}\sum_{i\in I}\xi_i - \bar{\xi}\right\|^2 = \frac{n-b}{b(n-1)}\mathbb{E}_i\|\xi_i - \bar{\xi}\|^2.$$

We orient our readers to the supplementary materials of Nitanda (2016) for proof details of the above lemma.

B. Proof of Proposition 3.4

Based on Lemma A.7 and Lemma A.8, we prove 3.4.

Proof. Let $\nabla_k^i = \nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x})$ and $\nabla_k = \frac{1}{b_k}\sum\limits_{i\in I_k}\nabla_k^i$. Using Lemma (A.8), we have

$$\mathbb{E}_{I_k}\|\nabla_k - \nabla f(x_{k+1})\|^2 = \frac{1}{b}\frac{n-b_k}{n-1}\mathbb{E}_i\left[\|\nabla_k^i - \nabla f(x_{k+1})\|^2\right] \le 4L\frac{n-b_k}{b_k(n-1)}\left(f(x_{k+1}) - f(x^*) + f(\tilde{x}) - f(x^*)\right), \tag{19}$$

where the inequality follows from Lemma (A.7). □

C. Proofs of Convergence Rates

C.1. Proof of QAGD

QAGD
γ-quasar-convex $(\mu = 0)$
$A_k = \frac{\bar{\mu}\gamma^2}{4L}(k+1)^2,\ B_k = 1$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{\bar{a}_k}, \frac{1}{L}, f, 0, \frac{\gamma A_k}{\bar{a}_k}, \frac{\gamma \epsilon}{2}\right)$
μ-strongly γ-quasar-convex $(\mu > 0)$
$A_k = (1 + \gamma/2\sqrt{\kappa})^k,\ B_k = \mu A_k$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\gamma\mu, \frac{\gamma\mu B_k}{b_k}, \frac{1}{L}, f, \frac{\gamma\bar{\mu}\mu}{2}, \frac{\gamma A_k}{\bar{a}_k}, 0\right)$

Table 2. Parameter choices for QAGD

we begin with Algorithm 1, $\nabla_k = \nabla f(x_{k+1})$ and parameters specified in Table 2 using Lyapunov function (11):

Case 1: $\mu = 0$ For $\nabla_k = \nabla f(x_{k+1})$, we begin with Algorithm 1 using Lyapunov function (11):

$$\begin{aligned}
E_{k+1} - E_k &\overset{(12)}{=} -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), x^* - z_{k+1}\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(8)}{\leq} \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), x^* - z_k\rangle + \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{\bar{\mu}}{2}\|z_{k+1} - z_k\|^2 \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(13)}{\leq} \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), x^* - z_k\rangle + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \frac{\tau_k}{\beta_k}\langle \nabla f(x_{k+1}), y_k - z_k\rangle + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) \\
&\leq \frac{\gamma}{\beta_k}(f(x^*) - f(x_{k+1})) + \frac{1}{\beta_k}(c(f(y_k) - f(x_{k+1})) + \tilde{\epsilon}) + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k))
\end{aligned}$$

The second inequality follows from the Fenchel-Young inequality, and the last inequality follows from the quasar-convexity of f and (15) (Lemma A.3). With the choice of parameter summarized in Table 2, we obtain the following bound:

$$\begin{aligned}
E_{k+1} - E_k &\leq \frac{\gamma}{\beta_k}(f(x^*) - f(x_{k+1})) + \frac{1}{\beta_k}(c(f(y_k) - f(x_{k+1})) + \tilde{\epsilon}) + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) \\
&= \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \frac{(A_{k+1} - A_k)}{2}\epsilon \\
&\overset{(16)}{\leq} \left(\frac{1}{2\bar{\mu}\beta_k^2} - \frac{A_{k+1}}{2L}\right)\|\nabla f(x_{k+1})\|^2 + \frac{(A_{k+1} - A_k)}{2}\epsilon
\end{aligned}$$

The last inequality follows from (16). In fact, (20) is implied by the gradient descent with $\rho_k = 1/L$ and L-smoothness.

$$f(y_{k+1}) - f(x_{k+1}) \leq \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L}{2}\|y_{k+1} - x_{k+1}\|^2 = -\frac{1}{2L}\|\nabla f(x_{k+1})\|^2 \tag{20}$$

With the choice of β_k and A_k, we have

$$\frac{1}{2\bar{\mu}\beta_k^2} = \frac{\bar{a}_k^2}{2\bar{\mu}\gamma^2} = \frac{\bar{\mu}\gamma^2}{32L^2}(2k+3)^2 \leq \frac{\bar{\mu}\gamma^2}{8L^2}(k+2)^2 = \frac{A_{k+1}}{2L}$$

Thus, we obtain the final bound:

$$E_{k+1} - E_k \leq \left(\frac{1}{2\bar{\mu}\beta_k^2} - \frac{A_{k+1}}{2L}\right)\|\nabla f(x_{k+1})\|^2 + \frac{(A_{k+1}-A_k)}{2}\epsilon \leq \frac{(A_{k+1}-A_k)}{2}\epsilon$$

By summing both sides of the inequality above, we obtain

$$\begin{aligned}
f(y_t) - f(x^*) &\leq A_t^{-1}\left(A_0(f(y_0) - f(x^*)) + D_h(x^*, z_0)\right) + \frac{\epsilon}{2} \\
&\overset{(16)}{\leq} A_t^{-1}\left(\frac{\bar{\mu}\gamma^2}{8}\|y_0 - x^*\|^2 + D_h(x^*, z_0)\right) + \frac{\epsilon}{2} \\
&\overset{(8)}{\leq} A_t^{-1}\left(\left(\frac{\gamma^2}{4} + 1\right) D_h(x^*, z_0)\right) + \frac{\epsilon}{2} \\
&\leq \frac{8LD_h(x^*, z_0)}{\gamma^2\bar{\mu}(t+1)^2} + \frac{\epsilon}{2} \leq \frac{8LR^2}{\gamma^2\bar{\mu}(t+1)^2} + \frac{\epsilon}{2}
\end{aligned}$$

Case 2: $\mu > 0$

$$\begin{aligned}
E_{k+1} - E_k &\overset{(12)}{=} \bar{b}_k D_h(x^*, z_{k+1}) - B_k\langle\nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \bar{b}_k D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}\langle\nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), x^* - z_{k+1}\rangle \\
&\quad - B_k D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(12)}{=} \left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k}\right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}\left(D_h(x^*, x_{k+1}) - D_h(z_{k+1}, x_{k+1})\right) + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), x^* - z_k\rangle \\
&\quad + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(8)}{\leq} \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), x^* - z_k\rangle \\
&\quad + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), x^* - x_{k+1}\rangle \\
&\quad + \frac{B_k}{\beta_k}\left(c(f(y_k) - f(x_{k+1})) + b\|x_{k+1} - z_k\|^2\right) + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), z_k - z_{k+1}\rangle + \left(\frac{\bar{\mu}\alpha_k B_k}{2\beta_k} - \frac{\bar{\mu}B_k}{2}\right)\|z_{k+1} - z_k\|^2 \\
&\quad + \frac{B_k}{\beta_k}\langle\nabla f(x_{k+1}), x^* - x_{k+1}\rangle + A_k(f(y_k) - f(x_{k+1})) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*))
\end{aligned}$$

The first equality and the third equality follows from Lemma A.1; the second equality follows from mirror descent. The first inequality follows from the strong convexity of h, and the second inequality follows from (15) (Lemma A.3). With the choice of α_k and β_k, we have $\alpha_k B_k/\beta_k = \bar{b}_k$, which explains the first inequality. Moreover, with the choice of B_k and Observation A.6, we have

$$\frac{\alpha_k}{\beta_k} = \frac{\bar{b}_k}{B_k} = \frac{\gamma}{2\sqrt{\kappa}} \leq \frac{\gamma}{2}\sqrt{\frac{2-\gamma}{\gamma}} \leq \frac{\gamma}{2}\sqrt{\frac{1}{\gamma^2}} = \frac{1}{2}.$$

Combined with the initial bound and the relation above, we obtain the following bound:

$$\begin{aligned}
E_{k+1} - E_k \leq\ & \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k} \langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle + \left(\frac{\bar{\mu} \alpha_k B_k}{2\beta_k} - \frac{\bar{\mu} B_k}{2} \right) \| z_{k+1} - z_k \|^2 \\
& + \frac{B_k}{\beta_k} \langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + A_k(f(y_k) - f(x_{k+1})) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
\leq\ & \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k} \langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - \frac{\bar{\mu} B_k}{4} \| z_{k+1} - z_k \|^2 + \frac{B_k}{\beta_k} \langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle \\
& + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \bar{a}_k (f(x_{k+1}) - f(x^*)) \\
\overset{(13)(9)}{\leq}\ & \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu} \beta_k^2} \| \nabla f(x_{k+1}) \|^2 + \frac{\gamma B_k}{\beta_k} (f(x^*) - f(x_{k+1}) - \mu D_h(x^*, x_{k+1})) \\
& + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \bar{a}_k (f(x_{k+1}) - f(x^*)) \\
\leq\ & \frac{B_k}{\bar{\mu} \beta_k^2} \| \nabla f(x_{k+1}) \|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
\overset{(20)}{\leq}\ & \left(\frac{B_k}{\bar{\mu} \beta_k^2} - \frac{A_{k+1}}{2L} \right) \| \nabla f(x_{k+1}) \|^2 \\
\leq\ & 0
\end{aligned}$$

The third inequality follows from the Fenchel-Young inequality of h and the strong quasar-convexity of f. The fifth inequality follows from L-smoothness of f. With the choice of B_k, β_k and A_k, we have

$$\frac{B_k}{\bar{\mu} \beta_k^2} = \frac{(\gamma / 2\sqrt{\kappa})^2 A_k}{\gamma^2 \bar{\mu} \mu} = \frac{A_k}{4L} \leq \frac{A_{k+1}}{2L},$$

which explains the last inequality. Therefore, we obtain

$$f(y_t) - f(x^*) \leq \left(1 + \frac{\gamma}{2\sqrt{\kappa}} \right)^{-t} (f(y_0) - f(x^*) + \mu D_h(x^*, z_0)) = \left(1 + \frac{\gamma}{2\sqrt{\kappa}} \right)^{-t} E_0$$

C.2. Proof of Theorem 3.2

QASGD
γ**-quasar-convex** $(\mu = 0)$
$A_k = \eta(k+1)^2, \eta = \min\left(\frac{\bar{\mu}}{L}, \sqrt{\frac{4D_h(x^*, z_0)}{\sigma^2}} \frac{\gamma}{(t+1)^{3/2}} \right), B_k = 1$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{\bar{a}_k}, 0, f_i, 0, \frac{\gamma A_k}{\bar{a}_k}, \frac{\gamma \epsilon}{2} \right)$
μ**-strongly** γ**-quasar-convex** $(\mu > 0)$
Step 1: Choose A_k, B_k, θ_k as follows until convergence
$A_k = \left(1 + \min\left\{ \gamma^2 \bar{\mu}^2 / 16, 1/2 \right\} \right)^k, B_k = \mu A_k$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\frac{\gamma\mu}{2}, \frac{\gamma \mu B_k}{2\bar{b}_k}, 0, f_i, \frac{\gamma \bar{\mu} \mu}{4}, \frac{\gamma A_k}{2\bar{a}_k}, 0 \right)$
Step 2: Restart and choose A_k, B_k, θ_k as follows
$A_k = \frac{\gamma^2 \bar{\mu}^2}{36} \left(k + \max\{48/\gamma^2 \bar{\mu}^2, 5\} \right)^2, B_k = \mu A_k$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\frac{\gamma\mu}{2}, \frac{\gamma \mu B_k}{2\bar{b}_k}, 0, f_i, \frac{\gamma \bar{\mu} \mu}{4}, \frac{\gamma A_k}{2\bar{a}_k}, 0 \right)$

Table 3. Parameter choices for QASGD

Note that y_k is identical to x_k since $\rho_k = 0$. Thus we can substitute y_k in Lyapunov functions (11) with x_k. We begin with Algorithm 1, $\nabla_k = \nabla f_i(x_{k+1})$ and parameters specified in Table 3 using Lyapunov function (11):

Case 1: $\mu = 0$

$$
\begin{aligned}
E_{k+1} - E_k &= -\left\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \right\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(x_{k+1}) - f(x^*)) - A_k(f(x_k) - f(x^*)) \\
&= \frac{1}{\beta_k} \left\langle \nabla_k, x^* - z_{k+1} \right\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(x_{k+1}) - f(x^*)) - A_k(f(x_k) - f(x^*)) \\
&\overset{(8)}{\le} \frac{1}{\beta_k} \left\langle \nabla_k, x^* - z_k \right\rangle + \frac{1}{\beta_k} \left\langle \nabla_k, z_k - z_{k+1} \right\rangle - \frac{\bar{\mu}}{2} \|z_{k+1} - z_k\|^2 + A_{k+1}(f(x_{k+1}) - f(x^*)) \\
&\quad - A_k(f(x_k) - f(x^*)) \\
&\le \frac{1}{\beta_k} \left\langle \nabla_k, x^* - z_k \right\rangle + \frac{1}{2\bar{\mu}\beta_k^2} \|\nabla_k\|^2 + A_{k+1}(f(x_{k+1}) - f(x^*)) - A_k(f(x_k) - f(x^*)) \\
&= \frac{1}{\beta_k} \left\langle \nabla_k, x^* - x_{k+1} \right\rangle + \frac{\tau_k}{\beta_k} \left\langle \nabla_k, x_k - z_k \right\rangle + \frac{1}{2\bar{\mu}\beta_k^2} \|\nabla_k\|^2 + \mathcal{A}_k \\
&\le \frac{1}{\beta_k} \left\langle \nabla_k, x^* - x_{k+1} \right\rangle + \frac{1}{\beta_k} (c(f_i(x_k) - f_i(x_{k+1})) + \tilde{\epsilon}) + \frac{1}{2\bar{\mu}\beta_k^2} \|\nabla_k\|^2 + \mathcal{A}_k
\end{aligned}
$$

where $\mathcal{A}_k \triangleq \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k))$. The second equality follows from mirror descent; the first inequality follows from the strong convexity of h; the second inequality follows from the Fenchel-Young inequality, and the last inequality follows from (15) (Lemma A.3). We take the expectation of both sides of the initial bound and obtain

$$
\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] &\le \frac{1}{\beta_k} \left\langle \nabla f(x_{k+1}), x^* - x_{k+1} \right\rangle + \frac{c}{\beta_k}(f(x_k) - f(x_{k+1})) + \frac{1}{2\bar{\mu}\beta_k^2} \mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{\bar{a}_k}{2}\epsilon \\
&\quad + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k)) \\
&\le \frac{\gamma}{\beta_k}(f(x^*) - f(x_{k+1})) + A_k(f(x_k) - f(x_{k+1})) + \frac{1}{2\bar{\mu}\beta_k^2} \mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{\bar{a}_k}{2}\epsilon \\
&\quad + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k)) \\
&\le \frac{1}{2\bar{\mu}\beta_k^2} \mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{\bar{a}_k}{2}\epsilon
\end{aligned}
$$

By summing both sides of the inequality above, we obtain

$$
\begin{aligned}
\mathbb{E}[f(x_t) - f(x^*)] &\le A_t^{-1} \left(A_0(f(x_0) - f(x^*)) + D_h(x^*, z_0) + \sum_{k=0}^{t-1} \frac{1}{2\bar{\mu}\beta_k^2} \mathbb{E}\left[\|\nabla_k\|^2\right] \right) + \frac{\epsilon}{2} \\
&\le A_t^{-1} \left(2D_h(x^*, z_0) + \sum_{k=0}^{t-1} \frac{1}{2\bar{\mu}\beta_k^2} \mathbb{E}\left[\|\nabla_k\|^2\right] \right) + \frac{\epsilon}{2} \\
&\le \frac{2D_h(x^*, z_0)}{\eta(t+1)^2} + \frac{\sigma^2\eta}{\gamma^2\bar{\mu}}(t+1) + \frac{\epsilon}{2} \\
&\le \frac{2LD_h(x^*, z_0)}{\bar{\mu}(t+1)^2} + \frac{2\sigma}{\gamma\bar{\mu}} \sqrt{\frac{D_h(x^*, z_0)}{t+1}} + \frac{\epsilon}{2} \le \frac{2LR^2}{\bar{\mu}(t+1)^2} + \frac{2\sigma R}{\gamma\bar{\mu}\sqrt{t+1}} + \frac{\epsilon}{2}
\end{aligned}
$$

Case 2: $\mu > 0$

$$
\begin{aligned}
E_{k+1} - E_k &= \bar{b}_k D_h(x^*, z_{k+1}) - B_k \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle - B_k D_h(z_{k+1}, z_k) \\
&\quad + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k)) \\
&= \bar{b}_k D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k} \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1} \rangle + \frac{B_k}{\beta_k} \langle \nabla_k, x^* - z_{k+1} \rangle \\
&\quad - B_k D_h(z_{k+1}, z_k) + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k)) \\
&\overset{(12)}{=} \left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k} \right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k} \left(D_h(x^*, x_{k+1}) - D_h(z_{k+1}, x_{k+1}) \right) + \frac{B_k}{\beta_k} \langle \nabla_k, x^* - z_k \rangle \\
&\quad + \frac{B_k}{\beta_k} \langle \nabla_k, z_k - z_{k+1} \rangle - B_k D_h(z_{k+1}, z_k) + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k))
\end{aligned}
$$

The second equality follows from mirror descent. With the choice of β_k, α_k and Observation A.6, we have $\bar{b}_k = \alpha_k B_k/\beta_k$ and $\alpha_k/\beta_k \le 1/2$. Therefore, We obtain the following bound:

$$
\begin{aligned}
E_{k+1} - E_k &\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{\tau_k B_k}{\beta_k}\langle \nabla_k, x_k - z_k\rangle \\
&\quad + \frac{B_k}{\beta_k}\langle \nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 + \mathcal{A}_k \\
&\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{B_k}{\beta_k}(c(f_i(x_k) - f_i(x_{k+1})) \\
&\quad + b\|z_k - x_{k+1}\|^2) + \frac{B_k}{\beta_k}\langle \nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 + \mathcal{A}_k \\
&\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}B_k}{4}\|z_{k+1} - z_k\|^2 + \frac{B_k}{\beta_k}\langle \nabla_k, z_k - z_{k+1}\rangle + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle \\
&\quad + A_k(f_i(x_k) - f_i(x_{k+1})) + \mathcal{A}_k \\
&\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \mathcal{A}_k + A_k(f_i(x_k) - f_i(x_{k+1}))
\end{aligned}
$$

where $\mathcal{A}_k \triangleq \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k))$. The second inequality follows from (15) (Lemma A.3); the second inequality follows from the triangle inequality and the last inequality follows from the Fenchel-Young inequality. We take the expectation of both sides of the inequality above and obtain

$$
\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] &\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{B_k}{\beta_k}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \bar{a}_k(f(x_{k+1}) - f(x^*)) \\
&\le \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{\gamma B_k}{\beta_k}(f(x^*) - f(x_{k+1}) - \mu D_h(x^*, x_{k+1})) \\
&\quad + \bar{a}_k(f(x_{k+1}) - f(x^*)) \\
&\le -\frac{\gamma\mu B_k}{2\beta_k} D_h(x^*, x_{k+1}) + \frac{2\mu^2 B_k}{\bar{\mu}\beta_k^2}\|x^* - x_{k+1}\|^2 + \frac{\sigma^2 B_k}{\bar{\mu}\beta_k^2} \\
&\le \left(\frac{4\mu^2 B_k}{\bar{\mu}^2\beta_k^2} - \frac{\gamma\mu B_k}{2\beta_k}\right) D_h(x^*, x_{k+1}) + \frac{\sigma^2 B_k}{\bar{\mu}\beta_k^2}
\end{aligned}
$$

Next We prove that $4\mu^2 B_k/\bar{\mu}^2\beta_k^2 \le \gamma\mu B_k/2\beta_k$. It suffices to prove that $4\mu/\bar{\mu}^2\beta_k \le \gamma/2$.

$$
\frac{4\mu}{\bar{\mu}^2\beta_k} = \frac{8\bar{a}_k}{\gamma\bar{\mu}^2 A_k} \le \frac{8}{\gamma\bar{\mu}^2}\frac{\gamma^2\bar{\mu}^2}{16} = \frac{\gamma}{2}
$$

Thus we obtain the final bound:

$$
\mathbb{E}[E_{k+1} - E_k] \le \frac{\sigma^2 B_k}{\bar{\mu}\beta_k^2}
$$

By summing both sides of the inequality above and using the notation $\mathrm{II} \triangleq \min\{\gamma^2\bar{\mu}^2/16, 1/2\}$, we obtain

$$
\begin{aligned}
\mathbb{E}[f(x_t) - f(x^*)] &\le A_t^{-1}\left(f(x_0) - f(x^*) + \mu D_h(x^*, z_0) + \sum_{k=0}^{t-1}\frac{\sigma^2 B_k}{\bar{\mu}\beta_k^2}\right) \\
&\le A_t^{-1}\left(f(x_0) - f(x^*) + \mu D_h(x^*, z_0) + \sum_{k=0}^{t-1}\frac{4\sigma^2(A_{k+1} - A_k)^2}{\gamma^2\bar{\mu}\mu A_k}\right) \\
&\le A_t^{-1}\left(f(x_0) - f(x^*) + \mu D_h(x^*, z_0) + \sum_{k=0}^{t-1}\frac{4\sigma^2 q^2 A_k}{\gamma^2\bar{\mu}\mu}\right) \\
&\le A_t^{-1}\left(f(x_0) - f(x^*) + \mu D_h(x^*, z_0) + \frac{4\sigma^2 q A_t}{\gamma^2\bar{\mu}\mu}\right) \\
&= (1 + \mathrm{II})^{-t} E_0 + \frac{\bar{\mu}\sigma^2}{4\mu}
\end{aligned}
$$

In Step 1, the algorithm is run until convergence and we let x_0 be the last iterate of the Step 1. Assume $\mathbb{E}[f(x_0) - f(x^*) + \mu D_h(x^*, z_0)] \leq \frac{\bar{\mu}\sigma^2}{4\mu}$, and we restart the algorithm using the parameters in Step 2 and the notation $m = \max\left\{\frac{48}{\gamma^2\bar{\mu}^2}, 5\right\}$. Then We obtain

$$\begin{aligned}
\mathbb{E}[f(x_t) - f(x^*)] &\leq A_t^{-1}\left(f(x_0) - f(x^*) + \mu D_h(x^*, z_0) + \sum_{k=0}^{t-1} \frac{4\sigma^2(A_{k+1} - A_k)^2}{\gamma^2\bar{\mu}\mu A_k}\right) \\
&\leq A_t^{-1}\left(\frac{\bar{\mu}\sigma^2}{4\mu} + \sum_{k=0}^{t-1} \frac{\bar{\mu}\sigma^2(2k+2m+1)^2}{9\gamma^2\mu(k+m)^2}\right) \\
&\leq A_t^{-1}\left(\frac{\bar{\mu}\sigma^2}{4\mu} + \frac{\bar{\mu}\sigma^2 t}{\gamma^2\mu}\right) \\
&\leq \frac{9\sigma^2}{\gamma^2\bar{\mu}\mu(t+m)^2} + \frac{36\sigma^2}{\gamma^2\bar{\mu}\mu(t+m)}
\end{aligned}$$

Additionally, we need to verify that the parameters in Step 2 satisfy two essential relations, $\alpha_k/\beta_k \leq 1/2$ and $4\mu/\bar{\mu}^2\beta_k^2 \leq \gamma/2$, which are key to obtaining the final bound of $E_{k+1} - E_k$.

$$\frac{\alpha_k}{\beta_k} = \frac{A_{k+1} - A_k}{A_k} = \frac{2k+2m+1}{(k+m)^2} \leq \frac{2m+1}{m^2} \leq \frac{11}{25} \leq \frac{1}{2}$$

$$\frac{4\mu}{\bar{\mu}^2\beta_k^2} \leq \frac{8(A_{k+1} - A_k)}{\gamma\bar{\mu}^2 A_k} \leq \frac{8}{\gamma\bar{\mu}^2}\frac{2m+1}{m^2} \leq \frac{8}{\gamma\bar{\mu}^2}\frac{3}{m} \leq \frac{8}{\gamma\bar{\mu}^2}\frac{\gamma^2\bar{\mu}^2}{16} = \frac{\gamma}{2}$$

C.3. Proof of Theorem 3.5

QASVRG
γ**-quasar-convex** $(\mu = 0)$
$A_k = \frac{\gamma^2\bar{\mu}}{16L}(k+1)^2$, $B_k = 1$ Batchsize $b_k = \left\lceil \frac{\gamma\bar{\mu}n(2k+3)}{2(n-1)p+\gamma\bar{\mu}(2k+3)} \right\rceil, p \leq \frac{\gamma\bar{\mu}}{16}$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{2\bar{a}_k}, \frac{1}{L}, f_{I_k}, 0, \frac{\gamma A_k}{2\bar{a}_k}, \frac{\gamma\epsilon f(y_0)}{2}\right)$
μ**-strongly** γ**-quasar-convex** $(\mu > 0)$
Option I
$A_k = (1 + \gamma/\sqrt{8\kappa})^k$, $B_k = \mu A_k$ Batchsize $b_k = \left\lceil \frac{8n(\sqrt{8\kappa}+\gamma)}{\gamma(n-1)+8(\sqrt{8\kappa}+\gamma)} \right\rceil$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\frac{\gamma\mu}{2}, \frac{\gamma\mu B_k}{2\bar{b}_k}, \frac{1}{L}, f_{I_k}, \frac{\gamma\bar{\mu}\mu}{4}, \frac{\gamma A_k}{2\bar{a}_k}, 0\right)$
Option II
$A_k = \frac{\gamma^2\bar{\mu}}{16L}(k+1)^2$, $B_k = 1$ Batchsize $b_k = \left\lceil \frac{\gamma\bar{\mu}n(2k+3)}{2(n-1)p+\gamma\bar{\mu}(2k+3)} \right\rceil, p \leq \frac{\gamma\bar{\mu}}{16}$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{2\bar{a}_k}, \frac{1}{L}, f_{I_k}, 0, \frac{\gamma A_k}{2\bar{a}_k}, \frac{\gamma\epsilon f(y_0)}{2}\right)$

Table 4. Parameter choices for QASVRG

We begin with Algorithm 1, $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ and parameters specified in Table 4 using Lyapunov function (11):

case 1: $\mu = 0$

$$\begin{aligned}
E_{k+1} - E_k &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k}\langle \nabla_k, x^* - z_{k+1}\rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\le \frac{1}{\beta_k}\langle \nabla_k, x^* - z_k\rangle + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{\tau_k}{\beta_k}\langle \nabla_k, y_k - z_k\rangle + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) \\
&\le \frac{1}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{1}{\beta_k}\left(c(f_{I_k}(y_k) - f_{I_k}(x_{k+1})) + \tilde{\epsilon}\right) + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) + \frac{\tau_k}{\beta_k}\langle \nabla f(y_0) - \nabla f_{I_k}(y_0), y_k - z_k\rangle
\end{aligned}$$

The first equality follows from the mirror descent; the third equality follows from the momentum step; the first inequality follows from the Fenchel-Young inequality, and the last inequality follows from (15) (Lemma A.3). Then we take the expectation with respect to the history of random variable I_{i_j} with its size equal to b_i, where $i = 0, 1, ..., t-1$ and $j = 1, 2, ..., \binom{n}{b_i}$, and we obtain

$$\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] &\le \frac{1}{\beta_k}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \frac{c}{\beta_k}(f(y_k) - f(x_{k+1})) + \frac{1}{2\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right] + \bar{a}_k(f(x_{k+1}) - f(x^*)) \\
&\quad + A_{k+1}\mathbb{E}\left[f(y_{k+1}) - f(x_{k+1})\right] + A_k(f(x_{k+1}) - f(y_k)) + \bar{a}_k f(y_0)\epsilon \\
&\le \bar{a}_k(f(x^*) - f(x_{k+1})) + \frac{1}{2\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] + A_{k+1}\mathbb{E}\left[f(y_{k+1}) - f(x_{k+1})\right] \\
&\quad + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + \bar{a}_k f(y_0)\epsilon
\end{aligned}$$

The first inequality follows from $\mathbb{E}\left[\|\nabla_k\|^2\right] = \mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right]$, as ∇_k corresponds to the mini-batch in the k^{th} iteration; the second inequality follows from the quasar-convexity of f and the relation $\mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right] = \mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] + \|\nabla f(x_{k+1})\|^2$. By L-smoothness and the gradient descent in Algorithm 1, we have the following relation:

$$\begin{aligned}
\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] &\le \mathbb{E}[\langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle] + \frac{L}{2}\mathbb{E}\left[\|y_{k+1} - x_{k+1}\|^2\right] \\
&= -\frac{1}{L}\|\nabla f(x_{k+1})\|^2 + \frac{1}{2L}\mathbb{E}\left[\|\nabla_k\|^2\right] \\
&= -\frac{1}{2L}\|\nabla f(x_{k+1})\|^2 + \frac{1}{2L}\mathbb{E}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right]
\end{aligned} \tag{21}$$

Using the relation above, we obtain the following bound:

$$\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] &\le \bar{a}_k(f(x^*) - f(x_{k+1})) + \frac{1}{2\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] + A_{k+1}\mathbb{E}\left[f(y_{k+1}) - f(x_{k+1})\right] \\
&\quad + \frac{1}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + \bar{a}_k f(y_0)\epsilon \\
&\le \bar{a}_k(f(x^*) - f(x_{k+1})) + \left(\frac{1}{2\bar{\mu}\beta_k^2} + \frac{A_{k+1}}{2L}\right)\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] \\
&\quad + \left(\frac{1}{2\bar{\mu}\beta_k^2} - \frac{A_{k+1}}{2L}\right)\|\nabla f(x_{k+1})\|^2 + \bar{a}_k f(y_0)\epsilon \\
&\le \bar{a}_k(f(x^*) - f(x_{k+1})) + \frac{A_{k+1}}{L}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] + \bar{a}_k f(y_0)\epsilon \\
&\overset{(19)}{\le} \bar{a}_k(f(x^*) - f(x_{k+1})) + 4A_{k+1}\delta_k(f(x_{k+1}) - f(x^*)) + 4A_{k+1}\delta_k(f(y_0) - f(x^*)) + \bar{a}_k f(y_0)\epsilon
\end{aligned}$$

We use the notation $\delta_k \triangleq \frac{n-b_k}{b_k(n-1)}$. The third inequality follows from the relation $1/\bar{\mu}\beta_k^2 \leq A_{k+1}/L$. Actually, with the choice of β_k and A_k, we have

$$\frac{1}{\bar{\mu}\beta_k^2} = \frac{4\bar{a}_k^2}{\bar{\mu}\gamma^2} = \frac{4\eta^2(2k+3)^2}{\bar{\mu}\gamma^2} \leq \frac{16\eta^2(k+2)^2}{\bar{\mu}\gamma^2} = \frac{A_{k+1}}{L},$$

where $\eta = \gamma^2\bar{\mu}/16L$. The last inequality follows from Proposition 3.1 (19). Next we prove the relation $4L\delta_k/\beta_k \leq p$. It suffices to prove $\delta_k \leq \gamma p/8L\bar{a}_k$.

$$\delta_k = \frac{n-b_k}{b_k(n-1)} \leq \frac{2np}{2(n-1)p+\gamma\bar{\mu}(2k+3)} \frac{2(n-1)p+\gamma\bar{\mu}(2k+3)}{\gamma\bar{\mu}n(2k+3)} = \frac{2p}{\gamma\bar{\mu}(2k+3)} = \frac{\gamma p}{8L\bar{a}_k}$$

Thus we have $4A_{k+1}\delta_k \leq pA_{k+1}\beta_k/L$, by which we obtain the final bound of $\mathbb{E}[E_{k+1}-E_k]$.

$$\begin{aligned}
\mathbb{E}[E_{k+1}-E_k] &\leq \bar{a}_k(f(x^*)-f(x_{k+1})) + 4A_{k+1}\delta_k(f(x_{k+1})-f(x^*)) + 4A_{k+1}\delta_k(f(y_0)-f(x^*)) + \bar{a}_k f(y_0)\epsilon \\
&\leq \left(\bar{a}_k - \frac{pA_{k+1}\beta_k}{L}\right)(f(x^*)-f(x_{k+1})) + \frac{pA_{k+1}\beta_k}{L}(f(y_0)-f(x^*)) + \bar{a}_k f(y_0)\epsilon \\
&\leq \frac{8p\bar{a}_k}{\gamma\bar{\mu}}(f(y_0)-f(x^*)) + \bar{a}_k f(y_0)\epsilon \\
&\leq \frac{8p\bar{a}_k}{\gamma\bar{\mu}} f(y_0) + \bar{a}_k f(y_0)\epsilon
\end{aligned}$$

The third inequality follows from two relations, $\bar{a}_k \geq pA_{k+1}\beta_k/L$ and $A_{k+1}\beta_k/L \leq 8\bar{a}_k/\gamma\bar{\mu}$ and the last inequality follows from $f(x^*) \geq 0$. With the choice of A_k and β_k, we have

$$\frac{pA_{k+1}\beta_k}{L} = \frac{\gamma p A_{k+1}}{2L\bar{a}_k} \leq \frac{\gamma^2\bar{\mu}(k+2)^2}{32L(2k+3)} \leq \frac{\gamma^2\bar{\mu}(2k+3)}{32L} = \frac{\bar{a}_k}{2} \leq \bar{a}_k$$

$$\frac{A_{k+1}\beta_k}{L} = \frac{\gamma A_{k+1}}{2L\bar{a}_k} = \frac{\gamma(k+2)^2}{2L(2k+3)} \leq \frac{\gamma(2k+3)}{2L} = \frac{8\bar{a}_k}{\gamma\bar{\mu}}$$

Summing both sides of the inequality about the final bound, we obtain

$$\begin{aligned}
\mathbb{E}[f(y_t)-f(x^*)] &\leq A_t^{-1}\left(A_0(f(y_0)-f(x^*)) + D_h(x^*,z_0)\right) + A_t^{-1}\sum_{k=0}^{t-1}\frac{8p\bar{a}_k}{\gamma\bar{\mu}} f(y_0) + f(y_0)\epsilon \\
&\leq \frac{17LD_h(x^*,z_0)}{\gamma^2\bar{\mu}(t+1)^2} + \left(\frac{8p}{\gamma\bar{\mu}}+\epsilon\right) f(y_0) \\
&\leq \frac{17LR^2}{\gamma^2\bar{\mu}(t+1)^2} + \left(\frac{8p}{\gamma\bar{\mu}}+\epsilon\right) f(y_0)
\end{aligned}$$

When $t \geq \left\lceil \sqrt{\frac{17LD_h(x^*,z_0)}{\gamma^2\bar{\mu}qf(y_0)}} \right\rceil$, we have

$$\mathbb{E}[f(y_t)-f(x^*)] \leq \left(q + \frac{8p}{\gamma\bar{\mu}} + \epsilon\right) f(y_0),$$

where $q + 8p/\gamma\bar{\mu} + \epsilon < 1$ with the choice of q and p. Next we conclude the convergence rate of global stages. Suppose y_s is the output of stage s. If $t \geq \left\lceil \sqrt{\frac{17LD_h(x^*,y_s)}{\gamma^2\bar{\mu}qf(y_s)}} \right\rceil$ at each stage, we have

$$\mathbb{E}[f(y_s)-f(x^*)] \leq \left(q + \frac{8p}{\gamma\bar{\mu}} + \epsilon\right)^{s+1} f(y_0)$$

Case 2: $\mu > 0$ (Option I) Using the notation $d_{k+1} = A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*))$, we obtain

$$\begin{aligned}
E_{k+1} - E_k = &\bar{b}_k D_h(x^*, z_{k+1}) - B_k\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) + d_{k+1} \\
\leq &\bar{b}_k D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - z_{k+1}\rangle \\
&- \frac{\bar{\mu} B_k}{2}\|z_{k+1} - z_k\|^2 + d_{k+1} \\
= &\left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k}\right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}(D_h(x^*, x_{k+1}) - D_h(z_{k+1}, x_{k+1})) + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle \\
&+ \frac{B_k(1-\tau_k)}{\beta_k}\langle \nabla_k, y_k - z_k\rangle + \frac{B_k}{\beta_k}\langle \nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu} B_k}{2}\|z_{k+1} - z_k\|^2 + d_{k+1} \\
\leq &\frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{B_k}{\beta_k}\langle \nabla_k, z_k - z_{k+1}\rangle \\
&- \frac{\bar{\mu} B_k}{2}\|z_{k+1} - z_k\|^2 + \frac{B_k}{\beta_k}\left(c(f_{I_k}(y_k) - f_{I_k}(x_{k+1})) + b\|x_{k+1} - z_k\|^2\right) + d_{k+1} + \xi_k \\
\leq &\frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle \nabla_k, x^* - x_{k+1}\rangle + \frac{B_k}{\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + \frac{cB_k}{\beta_k}(f_{I_k}(y_k) - f_{I_k}(x_{k+1})) + d_{k+1} + \xi_k
\end{aligned}$$

Note that $\xi_k = \frac{B_k}{\beta_k}\langle \nabla f(y_0) - \nabla f_{I_k}(y_0), x_{k+1} - z_k\rangle$, and $\mathbb{E}[\xi_k] = 0$. The first inequality follows from the μ-strong convexity of function h; the second inequality follows from (15) (Lemma A.3), and the last inequality follows from the triangle inequality with the relation $\alpha_k/\beta_k \leq 1/2$. Then we take the expectation with respect to the history of random variable I_{i_j} with its size equal to b_i, where $i = 0, 1, ..., t-1$ and $j = 1, 2, ..., \binom{n}{b_i}$.

$$\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] = &\frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right] + \frac{cB_k}{\beta_k}(f(y_k) - f(x_{k+1})) \\
&+ \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] \\
\leq &\frac{(\alpha_k - \gamma\mu)B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{\gamma B_k}{\beta_k}(f(x^*) - f(x_{k+1})) + \frac{cB_k}{\beta_k}(f(y_k) - f(x_{k+1})) \\
&+ \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right] + \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] \\
\leq &\left(\frac{\gamma B_k}{\beta_k} - \bar{a}_k\right)(f(x^*) - f(x_{k+1})) + \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] + \frac{B_k}{\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 \\
&+ \frac{A_{k+1}}{2L}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] - \frac{A_{k+1}}{2L}\|\nabla f(x_{k+1})\|^2 \\
\leq &\left(\frac{\gamma B_k}{\beta_k} - \bar{a}_k\right)(f(x^*) - f(x_{k+1})) + \frac{A_{k+1}}{L}\mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right] \\
\leq &\bar{a}_k(f(x^*) - f(x_{k+1})) + 4A_{k+1}\delta_k(f(x_{k+1}) - f(x^*)) + 4A_{k+1}\delta_k(f(y_0) - f(x^*))
\end{aligned}$$

The first inequality follows from the uniform quasar-convexity of function f; the second inequality follows from (21); the third inequality follows from the relation $B_k/\bar{\mu}\beta_k^2 \leq A_{k+1}/2L$, and the last inequality follows from Proposition 3.1 (19). We testify the correctness of relations above. With the choice of $A_k, B_k, \alpha_k, \beta_k$ and Observation A.6, we have

$$\frac{\alpha_k}{\beta_k} = \frac{\bar{b}_k}{B_k} = \frac{\gamma}{\sqrt{8\kappa}} \leq \frac{\gamma}{\sqrt{8}}\sqrt{\frac{2-\gamma}{\gamma}} \leq \frac{\gamma}{\sqrt{8}}\sqrt{\frac{1}{\gamma^2}} \leq \frac{1}{2}.$$

$$\frac{B_k}{\bar{\mu}\beta_k^2} = \frac{4(A_{k+1} - A_k)^2}{\gamma^2\bar{\mu}\mu A_k} = \frac{4A_k(\gamma/\sqrt{8\kappa})^2}{\gamma^2\bar{\mu}\mu} = \frac{A_k}{2L} \leq \frac{A_{k+1}}{2L}$$

Next we prove $4A_{k+1}\delta_k \leq \bar{a}_k/2$. It suffices to prove $\delta_k \leq \bar{a}_k/8A_{k+1}$.

$$\delta_k = \frac{n - b_k}{b_k(n-1)} \leq \frac{\gamma}{8(\sqrt{8\kappa} + \gamma)} = \frac{\bar{a}_k}{8A_{k+1}}$$

Thus we obtain the final bound of $\mathbb{E}[E_{k+1} - E_k]$.

$$\begin{aligned}\mathbb{E}[E_{k+1} - E_k] &\leq \bar{a}_k(f(x^*) - f(x_{k+1})) + 4A_{k+1}\delta_k(f(x_{k+1}) - f(x^*)) + 4A_{k+1}\delta_k(f(y_0) - f(x^*)) \\ &\leq \frac{\bar{a}_k}{2}(f(y_0) - f(x^*))\end{aligned}$$

Summing both sides of the inequality above, we obtain

$$\begin{aligned}\mathbb{E}[f(y_t) - f(x^*)] &\leq \left(1 + \frac{\gamma}{\sqrt{8\kappa}}\right)^{-t}(f(y_0) - f(x^*) + \mu D_h(x^*, z_0)) + \frac{1}{2}(f(y_0) - f(x^*)) \\ &\leq \left(1 + \frac{\gamma}{\sqrt{8\kappa}}\right)^{-t}\frac{2}{\gamma}(f(y_0) - f(x^*)) + \frac{1}{2}(f(y_0) - f(x^*))\end{aligned}$$

When $t \geq \log_{1+\gamma/\sqrt{8\kappa}}(2/\gamma q)$, we have

$$\mathbb{E}[f(y_t) - f(x^*)] \leq \left(q + \frac{1}{2}\right)(f(y_0) - f(x^*)),$$

where $q + 1/2 < 1$ with the choice of q. Next we conclude the convergence rate of global stages. Suppose y_s is the output of stage s. If $t \geq \log_{1+\gamma/\sqrt{8\kappa}}(2/\gamma q)$ at each stage, we have

$$\mathbb{E}[f(y_s) - f(x^*)] \leq \left(q + \frac{1}{2}\right)^{s+1}(f(y_0) - f(x^*)) = \left(q + \frac{1}{2}\right)^{s+1}\mathcal{E}_0.$$

For $\mu > 0$ (Option II), our analysis is identical to case 1.

D. Proofs of Complexity

In this section, we analyze the overall complexity of QASGD and QASVRG. As the complexity of QAGD has been analyzed in Hinder et al. (2020), we will not provide the related proof. Analogously, we apply the same metric as Hinder et al. (2020) propose, which is the total number of function and gradient evaluations. Roughly speaking, the overall complexity is the multiplication of the number of iterations and the number of function and gradient evaluations per iteration. We consider the worst case of Bisearch where $0, 1$ and "guess" do not meet our conditions, and we need to do line search at each iteration. In this situation, we access the gradient (estimate) in Bisearch that can be directly used in the subsequent updates at each iteration, i.e., no additional access to ∇_k is required in the mirror descent step and the gradient descent step. The number of f_i involved in Bisearch affects the complexity. While QAGD needs all f_i in Bisearch, QASGD and QASVRG only need a single f_i and a mini-batch of f_i respectively. Next we present the analysis of the complexity of Bisearch per iteration for QASGD and QASVRG in the Euclidean setting where $D_h(x, y) = \frac{1}{2}\|x - y\|^2$, and present the proof of Theorem 3.5 and Theorem 3.6. Lemma A.5 implies that Bisearch needs $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right)$ function and gradient evaluations per iteration for a single function. As $\tilde{\epsilon}$ and b can not be simultaneously non-zero, we have two different situations corresponding to different complexity.

$$O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = \begin{cases} O\left(\log^+\left((1+c)\frac{L^3}{b^3}\right)\right) & \mu > 0, \\ O\left(\log^+\left((1+c)\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}\right)\right) & \mu = 0. \end{cases}$$

When $\mu = 0$, the key to our proof is to bound $\|y_k - z_k\|$.

D.1. Proof of Corollary 3.3

Case 1: $\mu = 0$

By the proof of 3.2, we have $\mathbb{E}[E_{k+1}-E_k] \le \frac{\sigma^2}{2\beta_k^2} + \frac{\bar{a}_k\epsilon}{2}$. Assuming $\|x^*-z_0\| \le R$, we have the following relation.

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\|x^*-z_k\|^2 &\le A_0(f(x_0)-f(x^*)) + \frac{1}{2}\|x^*-z_0\|^2 + \sum_{j=0}^{k-1}\frac{\sigma^2}{2\beta_j^2} + \frac{A_k\epsilon}{2} \\
&\le \|x^*-z_0\|^2 + \sum_{j=0}^{k-1}\frac{\sigma^2\eta^2}{2\gamma^2}(2j+3)^2 + \frac{\eta\epsilon}{2}(k+1)^2 \\
&\le 5R^2 + \frac{(k+1)^2}{2L}\epsilon
\end{aligned}$$

The third inequality follows from $\eta = \min\left(\frac{1}{L}, \sqrt{\frac{2\|x^*-z_0\|^2}{\sigma^2}}\frac{\gamma}{(t+1)^{3/2}}\right)$. Combining the analysis above, we have $\mathbb{E}\|x^*-z_k\|^2 \le 10R^2 + \frac{(k+1)^2}{L}\epsilon$ and $\mathbb{E}\|x^*-z_k\| \le \sqrt{\mathbb{E}\|x^*-z_k\|^2} \le \sqrt{10R^2+\frac{(k+1)^2}{L}\epsilon}$ by Jensen's inequality. Thus we obtain

$$\mathbb{E}\|z_k - z_{k-1}\| = \mathbb{E}\left\|\frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \le \mathbb{E}\|x^*-z_k\| + \mathbb{E}\|x^*-z_{k-1}\| \le 2\sqrt{10R^2+\frac{(k+1)^2}{L}\epsilon}.$$

As the stepsize of gradient descent is 0, y_k is identical to x_k, and we solely need to bound $\|x_k-z_k\|$. By the definition of z_k and x_k, we have

$$\begin{aligned}
\mathbb{E}\|x_k-z_k\| &= \mathbb{E}\left\|(1-\tau_{k-1})z_{k-1} + \tau_{k-1}x_{k-1} - z_{k-1} + \frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \\
&= \mathbb{E}\left\|\tau_{k-1}(x_{k-1}-z_{k-1}) + \frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \\
&\le \tau_{k-1}\mathbb{E}\|x_{k-1}-z_{k-1}\| + \mathbb{E}\left\|\frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \\
&\le \mathbb{E}\|x_{k-1}-z_{k-1}\| + 2\sqrt{10R^2+\frac{(k+1)^2}{L}\epsilon} \\
&\le \mathbb{E}\|x_{k-1}-z_{k-1}\| + 2\sqrt{10}R + 2(k+1)\sqrt{\frac{\epsilon}{L}}
\end{aligned}$$

The last inequality follows from the relation $\sqrt{a^2+b^2} \le a+b$ for $a,b \ge 0$. By the proof of Theorem 3.2, we have

$$\mathbb{E}[f(x_t)-f(x^*)] \le \frac{LR^2}{(t+1)^2} + \frac{2\sigma}{\gamma}\frac{R}{\sqrt{2(t+1)}} \le \frac{LR^2+\frac{\sqrt{2}\sigma R}{\gamma}}{\sqrt{t+1}}.$$

Suppose $k \le k_{\max} = \left\lfloor \frac{4\left(LR^2+\frac{\sqrt{2}\sigma R}{\gamma}\right)^2}{\epsilon^2} \right\rfloor$, and we obtain

$$\begin{aligned}
\mathbb{E}\|x_k-z_k\| &\le 2\sqrt{10}Rk + \sqrt{\frac{\epsilon}{L}}k(k+3) \\
&\le 2\sqrt{10}Rk + 4\sqrt{\frac{\epsilon}{L}}k^2 \\
&\le 8\sqrt{10}R\frac{\left(LR^2+\frac{\sqrt{2}\sigma R}{\gamma}\right)^2}{\epsilon^2} + \frac{64\left(LR^2+\frac{\sqrt{2}\sigma R}{\gamma}\right)^4}{\sqrt{L}\epsilon^{\frac{7}{2}}} \\
&\le 8\sqrt{10}R\frac{(LR^2+\sqrt{2}\sigma R)^2}{\gamma^2\epsilon^2} + \frac{64\left(LR^2+\sqrt{2}\sigma R\right)^4}{\sqrt{L}\gamma^4\epsilon^{\frac{7}{2}}} \\
&= O\left(\frac{L^{7/2}R^8}{\gamma^4\epsilon^{7/2}}\right)
\end{aligned}$$

The first inequality follows from the relation $k+3 \leq 4k$ for $k \geq 1$. By Markov's inequality, $P_r(\|x_k - z_k\| \geq k_{\max}\mathbb{E}\|x_k - z_k\|) \leq \frac{1}{k_{\max}}$, which implies $\|x_k - z_k\| \leq k_{\max}\mathbb{E}\|x_k - z_k\| \leq O\left(\frac{L^{11/2}R^{12}}{\gamma^6\epsilon^{11/2}}\right)$ with probability at least $1 - \frac{1}{k_{\max}}$. Thus we have $\frac{L\|x_k - z_k\|^2}{\tilde{\epsilon}} \leq O\left(\frac{L^{12}R^{24}}{\gamma^{13}\epsilon^{12}}\right)$. Besides, we have $c+1 = \frac{\gamma(k+1)^2}{2k+3} + 1 \leq \gamma k + 2 \leq \frac{4(LR^2+\sqrt{2}\sigma R)^2}{\gamma\epsilon^2} + 2 = O\left(\frac{L^2R^4}{\gamma\epsilon^2}\right)$. Then we obtain the following upper bound of the term inside $\log^+()$:

$$\begin{aligned}(1+c)\min\left\{\frac{L\|x_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\} &= (1+c)\frac{L\|x_k - z_k\|^2}{\tilde{\epsilon}} \\ &\leq O\left(\frac{L^{14}R^{28}}{\gamma^{14}\epsilon^{14}}\right)\end{aligned}$$

Thus we have $O\left(\log^+\left((1+c)\min\left\{\frac{L\mathbb{E}\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) \leq O(\log^+(LR^2\gamma^{-1}\epsilon^{-1}))$. In the case of $\mu = 0$, QASGD needs $O\left(\sqrt{\frac{LR^2}{\epsilon}} + \frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)$ iterations to generate an ϵ-approximate solution under expectation. Therefore, the overall complexity of QASGD is upper bounded by $(\mu = 0)$ is $O\left(\left(\sqrt{\frac{LR^2}{\epsilon}} + \frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)\log^+\left(\frac{LR^2}{\gamma\epsilon}\right)\right)$ with high probability.

Case 2: $\mu > 0$

As $\tilde{\epsilon} = 0$ and $b > 0$, we have $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left((1+c)\frac{L^3}{b^3}\right)\right)$. In Step 1, $b = \frac{\gamma\mu}{4}$ and $c = \frac{8}{\gamma}$. Then we have

$$(1+c)\frac{L^3}{b^3} \leq \frac{9}{\gamma}\frac{64L^3}{\gamma^3\mu^3} = \frac{576\kappa^3}{\gamma^4}$$

Thus we have $O\left(\log^+\left((1+c)\min\left\{\frac{L\mathbb{E}\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right)\right)$. By the proof of Theorem 3.2, QASGD needs $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f(x^*))}{\gamma\epsilon}\right)\right)$ iterations in Step 1, and the complexity of Step 1 is $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f(x^*))}{\gamma\epsilon}\right)\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right)\right)$. In step 2, $b = \frac{\gamma\mu}{4}$ and $c = \frac{\gamma(k+48/\gamma^2)^2}{2(2k+96/\gamma^2+1)} \leq \frac{\gamma(k+48/\gamma^2)}{2} \leq \frac{k+48}{2\gamma}$. We can upper bound the convergence rate of Step 2:

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{9\sigma^2}{\gamma^2\mu(t+48/\gamma^2)^2} + \frac{36\sigma^2}{\gamma^2\mu(t+48/\gamma^2)} \leq \frac{45\sigma^2}{\gamma^2\mu(t+48/\gamma^2)}.$$

Suppose $k \leq \left\lfloor\frac{90\sigma^2}{\gamma^2\mu\epsilon}\right\rfloor$. We have $c+1 \leq \frac{C_3}{\gamma^3\epsilon}$, where $C_3 = \frac{45\sigma^2}{\mu} + 25$.

$$(1+c)\frac{L^3}{b^3} \leq \frac{C_3}{\gamma^3\epsilon}\frac{64\kappa^3}{\gamma^3} = \frac{64C_3\kappa^3}{\gamma^6\epsilon}$$

Thus $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$. QASGD needs $O\left(\frac{\sigma^2}{\gamma^2\epsilon}\right)$ iterations in Step 2, and the complexity of Step 2 is $O\left(\frac{\sigma^2}{\gamma^2\epsilon}\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$. In summary, the overall complexity of QASGD $(\mu > 0)$ is upper bounded by $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f(x^*))}{\gamma\epsilon}\right)\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right) + \frac{\sigma^2}{\gamma^2\epsilon}\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$.

D.2. Proof of Corollary 3.6

Case 1: $\mu = 0$

By the proof of Theorem 3.3, we have $\mathbb{E}[E_{k+1} - E_k] \leq \bar{a}_k\left(\frac{1}{2}+\epsilon\right)f(y_0)$. Assuming $\|x^* - z_0\| \leq R$, we have the following relation.

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|x^* - z_k\|^2 &\leq A_0(f(y_0) - f(x^*)) + \frac{1}{2}\|x^* - z_0\|^2 + A_k\left(\frac{1}{2}+\epsilon\right)f(y_0) \\ &\leq \|z_0 - x^*\|^2 + \frac{3\gamma^2}{32L}(k+1)^2 f(y_0) + \frac{\gamma^2}{16L}(k+1)^2 f(y_0)\epsilon \\ &\leq R^2 + \frac{3\gamma^2}{32L}(k+1)^2 f(y_0) + \frac{\gamma^2}{16L}(k+1)^2 f(y_0)\epsilon\end{aligned}$$

Thus we have $\mathbb{E}\|x^* - z_k\|^2 \leq 2R^2 + \frac{3\gamma^2}{16L}(k+1)^2 f(y_0) + \frac{\gamma^2}{8L}(k+1)^2 f(y_0)\epsilon$ and by Jensen's inequality

$$\begin{aligned}\mathbb{E}\|\nabla_{k-1}\| = \beta_{k-1}\mathbb{E}\|z_k - z_{k-1}\| &\leq \beta_{k-1}\mathbb{E}(\|x^* - z_k\| + \|x^* - z_{k-1}\|) \\ &\leq \frac{16L}{\gamma(2k+1)}\sqrt{2R^2 + \frac{3\gamma^2}{16L}(k+1)^2 f(y_0) + \frac{\gamma^2}{8L}(k+1)^2 f(y_0)\epsilon}.\end{aligned}$$

By the definition of y_k and z_k, we have

$$\begin{aligned}\mathbb{E}\|y_k - z_k\| &= \mathbb{E}\left\|x_k - \frac{1}{L}\nabla_{k-1} - z_{k-1} + \frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \\ &= \mathbb{E}\left\|(1-\tau_{k-1})z_{k-1} + \tau_{k-1}y_{k-1} - \frac{1}{L}\nabla_{k-1} - z_{k-1} + \frac{1}{\beta_{k-1}}\nabla_{k-1}\right\| \\ &\leq \tau_{k-1}\mathbb{E}\|y_{k-1} - z_{k-1}\| + \left|\frac{1}{\beta_{k-1}} - \frac{1}{L}\right|\mathbb{E}\|\nabla_{k-1}\| \\ &\leq \mathbb{E}\|y_{k-1} - z_{k-1}\| + \left(\frac{1}{\beta_{k-1}} + \frac{1}{L}\right)\mathbb{E}\|\nabla_{k-1}\| \\ &\leq \mathbb{E}\|y_{k-1} - z_{k-1}\| + \frac{2k+9}{8L}\mathbb{E}\|\nabla_{k-1}\| \\ &\leq \mathbb{E}\|y_{k-1} - z_{k-1}\| + \frac{8}{\gamma}\sqrt{2R^2 + \frac{3\gamma^2}{16L}(k+1)^2 f(y_0) + \frac{\gamma^2}{8L}(k+1)^2 f(y_0)\epsilon} \\ &\leq \mathbb{E}\|y_{k-1} - z_{k-1}\| + \frac{8\sqrt{2}R}{\gamma} + 2\sqrt{3}(k+1)\sqrt{\frac{f(y_0)}{L}} + 2\sqrt{2}(k+1)\sqrt{\frac{f(y_0)\epsilon}{L}}\end{aligned}$$

Suppose $f(y_0) \geq \epsilon$ and $k \leq \left\lfloor\sqrt{\frac{17LR^2}{2\gamma^2 q f(y_0)}}\right\rfloor \leq k_{\max} = \left\lfloor\sqrt{\frac{17LR^2}{2\gamma^2 q\epsilon}}\right\rfloor$, and we obtain

$$\begin{aligned}\mathbb{E}\|y_k - z_k\| &\leq \frac{8\sqrt{2}Rk}{\gamma} + \sqrt{3}k(k+3)\sqrt{\frac{f(y_0)}{L}} + \sqrt{2}k(k+3)\sqrt{\frac{f(y_0)\epsilon}{L}} \\ &\leq \frac{8\sqrt{2}Rk}{\gamma} + 4\sqrt{3}k^2\sqrt{\frac{f(y_0)}{L}} + 4\sqrt{2}k^2\sqrt{\frac{f(y_0)\epsilon}{L}} \\ &\leq \frac{8\sqrt{17q}R^2}{\gamma^2 q}\sqrt{\frac{L}{f(y_0)}} + \frac{34\sqrt{3}R^2}{\gamma^2 q}\sqrt{\frac{L}{f(y_0)}} + \frac{34\sqrt{2}R^2}{\gamma^2 q}\sqrt{\frac{L\epsilon}{f(y_0)}} \\ &\leq \frac{108R^2L^{1/2}}{\gamma^2\epsilon^{1/2}q} + \frac{34\sqrt{2}R^2L^{1/2}}{\gamma^2 q} \\ &\leq \frac{176R^2L^{1/2}}{\gamma^2\epsilon^{1/2}q} = O\left(\frac{L^{1/2}R^2}{q\gamma^2\epsilon^{1/2}}\right)\end{aligned}$$

By Markov's inequality, $P_r(\|y_k - z_k\| \geq k_{\max}\mathbb{E}\|y_k - z_k\|) \leq \frac{1}{k_{\max}}$, which implies $\|y_k - z_k\| \leq k_{\max}\mathbb{E}\|x_k - z_k\| \leq O\left(\frac{LR^3}{q^{3/2}\gamma^3\epsilon}\right)$ with probability at least $1 - \frac{1}{k_{\max}}$. Thus we have $\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}} \leq O\left(\frac{L^3R^6}{q^3\gamma^7\epsilon^4}\right)$. Besides, we have $c + 1 = \frac{\gamma(k+1)^2}{2(2k+3)} + 1 \leq \frac{\gamma k}{2} + \frac{3}{2} \leq \frac{R}{2}\sqrt{\frac{17L}{2q\epsilon}} + \frac{3}{2} = O\left(\frac{L^{1/2}R}{q^{1/2}\epsilon^{1/2}}\right)$. Then we obtain the following upper bound of the term inside $\log^+()$:

$$(1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\} = (1+c)\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}} \leq O\left(\frac{L^{7/2}R^7}{q^{7/2}\gamma^7\epsilon^{9/2}}\right)$$

Thus we have $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) \leq O\left(\log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)$ with high probability. As we need to access the full gradient and function value evaluated at y_0 per stage and the gradient and function value of mini-batch to calculate SVRG and $\tilde{\epsilon}$, the overall complexity of QASVRG ($\mu = 0$) to generate an ϵ-approximate solution is

$$\begin{aligned}
&O\left(\left(2n + \sum_{k=0}^{t-1} b_k \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&= O\left(\left(2n + \sum_{k=0}^{t-1} \frac{\gamma n(2k+3)}{2(n-1)p + \gamma(2k+3)} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&\leq O\left(\left(2n + \frac{\gamma n t(2t+1)}{2(n-1)p + \gamma(2t+1)} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&\leq O\left(\left(n + \frac{nLR^2}{\gamma\epsilon n + \gamma\sqrt{\epsilon L R^2}} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right)
\end{aligned}$$

where $t = \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q f(y_0)}} \right\rceil \leq \sqrt{\frac{17LR^2}{2\gamma^2 q f(y_0)}}$ is the maximum number of iterations per stage, and $O(\log(\epsilon^{-1}))$ is the number of stages. Note that $D_h(x^*, y_s)$ is uniformly bounded by R^2 under Assumption 2.5, which is in the bound above.

Case 2: $\mu > 0$

For Option II, we have $t = \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q f(y_0)}} \right\rceil \leq \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q \mathcal{E}_0}} \right\rceil \leq \sqrt{\frac{17L(2-\gamma)}{\gamma^3 q\mu}} \leq \sqrt{\frac{34\kappa}{\gamma^3 q}}$ using the last relation in Assumption 3.1. Thus the overall complexity of QASVRG (Option II) to generate an ϵ-approximate solution is

$$\begin{aligned}
&O\left(\left(2n + \sum_{k=0}^{t-1} b_k \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&= O\left(\left(2n + \sum_{k=0}^{t-1} \frac{\gamma n(2k+3)}{2(n-1)p + \gamma(2k+3)} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&\leq O\left(\left(2n + \frac{\gamma n t(2t+1)}{2(n-1)p + \gamma(2t+1)} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&\leq O\left(\left(n + \frac{n\kappa}{\gamma^2 n + \gamma^{3/2}\sqrt{\kappa}} \log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right)
\end{aligned}$$

where $O(\log(\epsilon^{-1}))$ is the number of stages. For Option I, $\tilde{\epsilon} = 0, b = \frac{\gamma\mu}{4}$ and $c = \sqrt{2\kappa}$. Then we obtain the following relation:

$$(1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\} = (1+c)\frac{L^3}{b^3} = (1+\sqrt{2\kappa})\frac{64L^3}{\gamma^3\mu^3} = (1+\sqrt{2\kappa})\frac{64\kappa^3}{\gamma^3}.$$

Thus we have $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O(\log^+(\kappa^{7/6}\gamma^{-1}))$. At each stage, QASVRG (Option II) is run until $(1+\gamma/\sqrt{8\kappa})^{-t} \leq \frac{\gamma q}{2}$. Let $(1+\gamma/\sqrt{8\kappa})^{-t} \leq e^{-\frac{\gamma t}{\sqrt{8\kappa}+\gamma}} \leq \frac{\gamma q}{2}$, and we have $t \geq \frac{\sqrt{8\kappa}+\gamma}{\gamma}\log\left(\frac{2}{\gamma q}\right)$, and $t = \left\lceil \log_{1+\frac{\gamma}{\sqrt{8\kappa}}}\left(\frac{2}{\gamma q}\right) \right\rceil \leq \frac{\sqrt{8\kappa}+\gamma}{\gamma}\log\left(\frac{2}{\gamma q}\right) \leq \frac{5\sqrt{\kappa}}{\gamma}\log\left(\frac{2}{\gamma q}\right)$. Thus the overall complexity of QASVRG (Option I) to generate an ϵ-approximate solution is

$$\begin{aligned}
O\left(\left(n + \sum_{k=0}^{t-1} b_k \log^+\left(\frac{\kappa^{7/6}}{\gamma}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) &= O\left(\left(n + \sum_{k=0}^{t-1} \frac{8n(\sqrt{8\kappa}+\gamma)}{\gamma(n-1) + 8(\sqrt{8\kappa}+\gamma)} \log^+\left(\frac{\kappa^{7/6}}{\gamma}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&= O\left(\left(n + \frac{40nt\sqrt{\kappa}}{\gamma(n-1) + 8(\sqrt{8\kappa}+\gamma)} \log^+\left(\frac{\kappa^{7/6}}{\gamma}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right) \\
&\leq O\left(\left(n + \frac{n\kappa}{\gamma^2 n + \gamma\sqrt{\kappa}} \log\left(\frac{2}{\gamma q}\right)\log^+\left(\frac{\kappa^{7/6}}{\gamma}\right)\right)\log\left(\frac{1}{\epsilon}\right)\right)
\end{aligned}$$

where $O(\log(\epsilon^{-1}))$ is the number of stages.

E. Theoretical Extension

We analyze QASGD under a restrictive condition: strong growth condition (SGC), which is formally formulated in the following. This condition has been proposed in Schmidt & Roux (2013), Vaswani et al. (2019), and Gower et al. (2021). Schmidt & Roux (2013) derive optimal convergence rates for SGD under SGC for convex and strongly convex functions.

Assumption E.1 (SGC)**.** Suppose i is sampled i.i.d from $[n]$. For some constant ρ and $x^* \in \mathcal{X}^*$, we have

$$E_i\left[\|\nabla f_i(x)\|^2\right] \le \rho \|\nabla f(x)\|^2.$$

If $\nabla f(x) = 0$, then $\nabla f_i(x) = 0$ under SGC, which implies the interpolation assumption. We derive better convergence rates for QASGD under SGC for $f \in \mathcal{Q}_{\mu\gamma}$.

Theorem E.2 (QASGD under SGC)**.** *Suppose Assumption 3.1 and Assumption E.1 hold, $D_h(x^*, z_0) \le R^2$, $f \in \mathcal{F}_L$, and choose any $\tilde{y}_0 \in \mathbb{R}^d$. Then Algorithm 1 with the choices of $\nabla_k = \nabla f_i(x_{k+1})$ and A_k, B_k, θ_k specified in Table 5 satisfies*

$$\mathbb{E}[E_{k+1} - E_k] \le \begin{cases} \dfrac{\bar{a}_k \epsilon}{2}, & \mu = 0, \\ 0, & \mu > 0. \end{cases} \tag{22}$$

Summing both sides of (22), we conclude the convergence rate as follows:

$$\mathbb{E}[f(y_t) - f(x^*)] \simeq \begin{cases} \dfrac{L\rho R^2}{\gamma^2 t^2} + \dfrac{\epsilon}{2}, & \mu = 0, \\ \left(1 + \dfrac{\gamma}{2\rho\sqrt{\kappa}}\right)^{-t} E_0, & \mu > 0. \end{cases} \tag{23}$$

QASGD under SGC
γ**-quasar-convex** $(\mu = 0)$
$A_k = \frac{\bar{\mu}\gamma^2}{4\rho L}(k+1)^2$, $B_k = 1$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{\bar{a}_k}, \frac{1}{\rho L}, f, 0, \frac{\gamma A_k}{\bar{a}_k}, \frac{\gamma\epsilon}{2}\right)$
μ**-strongly** γ**-quasar-convex** $(\mu > 0)$
$A_k = (1 + \gamma/2\rho\sqrt{\kappa})^k$, $B_k = \mu A_k$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\gamma\mu, \frac{\gamma\mu B_k}{\bar{b}_k}, \frac{1}{\rho L}, f, \frac{\gamma\bar{\mu}\mu}{2}, \frac{\gamma A_k}{\bar{a}_k}, 0\right)$

Table 5. Parameter choices for QASGD under SGC

Proof.

$$\begin{aligned}
E_{k+1} - E_k &\overset{(12)}{=} -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k} \langle \nabla_k, x^* - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(8)}{\le} \frac{1}{\beta_k} \langle \nabla_k, x^* - z_k \rangle + \frac{1}{\beta_k} \langle \nabla_k, z_k - z_{k+1} \rangle - \frac{\bar{\mu}}{2} \|z_{k+1} - z_k\|^2 \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(13)}{\le} \frac{1}{\beta_k} \langle \nabla_k, x^* - z_k \rangle + \frac{1}{2\bar{\mu}\beta_k^2} \|\nabla_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&= \frac{1}{\beta_k} \langle \nabla_k, x^* - x_{k+1} \rangle + \frac{\tau_k}{\beta_k} \langle \nabla_k, y_k - z_k \rangle + \frac{1}{2\bar{\mu}\beta_k^2} \|\nabla_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\
&\quad + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k))
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[E_{k+1} - E_k] &\leq \frac{\gamma}{\beta_k}(f(x^*) - f(x_{k+1})) + \frac{1}{\beta_k}(c(f(y_k) - f(x_{k+1})) + \tilde{\epsilon}) + \frac{1}{2\bar{\mu}\beta_k^2}\mathbb{E}\left[\|\nabla_k\|^2\right] \\
&\quad + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] + (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(y_k)) \\
&\leq \frac{\rho}{2\bar{\mu}\beta_k^2}\|\nabla f(x_{k+1})\|^2 + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] + \frac{(A_{k+1} - A_k)}{2}\epsilon \\
&\overset{(16)}{\leq} \left(\frac{\rho}{2\bar{\mu}\beta_k^2} - \frac{A_{k+1}}{2\rho L}\right)\|\nabla f(x_{k+1})\|^2 + \frac{(A_{k+1} - A_k)}{2}\epsilon \leq \frac{A_{k+1} - A_k}{2}\epsilon
\end{aligned}
$$

Case 2: $\mu > 0$

$$
\begin{aligned}
E_{k+1} - E_k &\overset{(12)}{=} \bar{b}_k D_h(x^*, z_{k+1}) - B_k\langle\nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) \\
&\quad - A_k(f(y_k) - f(x^*)) \\
&= \bar{b}_k D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}\langle\nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(12)}{=} \left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k}\right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k}\left(D_h(x^*, x_{k+1}) - D_h(z_{k+1}, x_{k+1})\right) + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - z_k\rangle \\
&\quad + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle - B_k D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\overset{(8)}{\leq} \left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k}\right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - z_k\rangle \\
&\quad + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\leq \left(\bar{b}_k - \frac{\alpha_k B_k}{\beta_k}\right) D_h(x^*, z_{k+1}) + \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) - \frac{\bar{\mu}\alpha_k B_k}{2\beta_k}\|z_{k+1} - x_{k+1}\|^2 + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - x_{k+1}\rangle \\
&\quad + \frac{B_k}{\beta_k}\left(c(f_i(y_k) - f_i(x_{k+1})) + b\|x_{k+1} - z_k\|^2\right) + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{2}\|z_{k+1} - z_k\|^2 \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle + \left(\frac{\bar{\mu}\alpha_k B_k}{2\beta_k} - \frac{\bar{\mu}B_k}{2}\right)\|z_{k+1} - z_k\|^2 + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - x_{k+1}\rangle \\
&\quad + A_k(f(y_k) - f(x_{k+1})) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*))
\end{aligned}
$$

The first equality and the third equality follows from Lemma A.1; the second equality follows from mirror descent. The first inequality follows from the strong convexity of h, and the second inequality follows from (15) (Lemma A.3). With the choice of α_k and β_k, we have $\alpha_k B_k/\beta_k = \bar{b}_k$, which explains the last inequality. Moreover, with the choice of B_k and Observation A.6, we have

$$
\frac{\alpha_k}{\beta_k} = \frac{\bar{b}_k}{B_k} = \frac{\gamma}{2\sqrt{\kappa}} \leq \frac{\gamma}{2}\sqrt{\frac{2-\gamma}{\gamma}} \leq \frac{\gamma}{2}\sqrt{\frac{1}{\gamma^2}} = \frac{1}{2}.
$$

Combined with the initial bound and the relation above, we obtain the following bound:

$$
\begin{aligned}
E_{k+1} - E_k &\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle + \left(\frac{\bar{\mu}\alpha_k B_k}{2\beta_k} - \frac{\bar{\mu}B_k}{2}\right)\|z_{k+1} - z_k\|^2 + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - x_{k+1}\rangle \\
&\quad + A_k(f(y_k) - f(x_{k+1})) + A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*)) \\
&\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\beta_k}\langle\nabla_k, z_k - z_{k+1}\rangle - \frac{\bar{\mu}B_k}{4}\|z_{k+1} - z_k\|^2 + \frac{B_k}{\beta_k}\langle\nabla_k, x^* - x_{k+1}\rangle \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \bar{a}_k(f(x_{k+1}) - f(x^*)) \\
&\overset{(13)(9)}{\leq} \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu}\beta_k^2}\|\nabla_k\|^2 + \frac{\gamma B_k}{\beta_k}(f_i(x^*) - f_i(x_{k+1}) - \mu D_h(x^*, x_{k+1})) \\
&\quad + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \bar{a}_k(f(x_{k+1}) - f(x^*))
\end{aligned}
$$

Taking the expectation, we obtain

$$\begin{aligned}\mathbb{E}[E_{k+1} - E_k] &\leq \frac{\alpha_k B_k}{\beta_k} D_h(x^*, x_{k+1}) + \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}\left[\|\nabla_k\|^2\right] + \frac{\gamma B_k}{\beta_k}(f(x^*) - f(x_{k+1}) - \mu D_h(x^*, x_{k+1}))) \\ &\quad + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] + \bar{a}_k(f(x_{k+1}) - f(x^*)) \\ &\leq \frac{B_k}{\bar{\mu}\beta_k^2}\mathbb{E}\left[\|\nabla_k\|^2\right] + A_{k+1}\mathbb{E}[f(y_{k+1}) - f(x_{k+1})] \\ &\leq \left(\frac{\rho B_k}{\bar{\mu}\beta_k^2} - \frac{A_{k+1}}{2\rho L}\right)\mathbb{E}\left[\|\nabla_k\|^2\right] \leq 0\end{aligned}$$

□

F. Additional Simulation Results

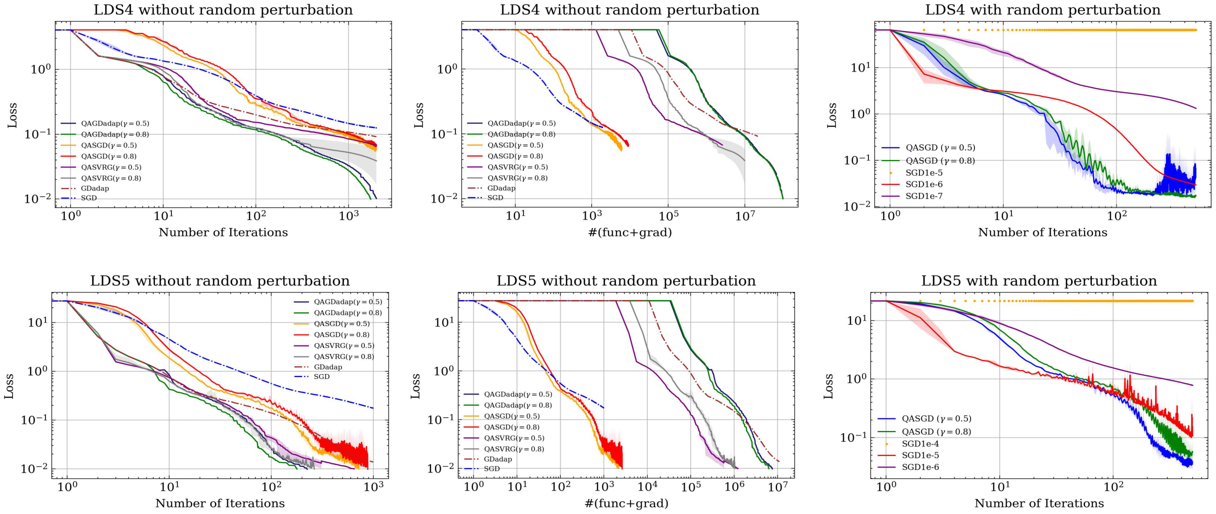

Figure 2. Evaluation on two different LDS instances with random seed in $\{12, 36\}$. We choose $\epsilon = 10^{-2}$, the stepsize to be $1 \times 10^{-6}, 1 \times 10^{-5}$ for SGD, $L = 1 \times 10^7, 5 \times 10^6$ for QASGD and $L = 3 \times 10^6, 1 \times 10^5$ for QASVRG in LDS4 and LDS5. The flat line in the third column means the loss blows up to infinity with this choice of stepsize.

We provide a contrived experiment by constructing an objective satisfying all the assumptions required. Consider the following optimization problem

$$\min_{x \in \mathbb{R}^d}\left[f(x) = \frac{1}{n}\sum_{i=1}^{n} g_\gamma(b_i a_i^\top x) + \frac{\mu}{2}\|x\|^2\right], \quad g_\gamma(x) = \begin{cases}\frac{x^\gamma - 1}{\gamma} + \frac{1}{2}, & x \geq 1, \\ \frac{x^2}{2}, & 0 \leq x \leq 1, \\ 0, & x \leq 0,\end{cases} \tag{24}$$

where $(a_i, b_i)_{i=1,\ldots,n}$ is training data with $a_i \in \mathbb{R}^d$ and $b_i \in \{+1, -1\}$; $\mu \geq 0$, and $f(x)$ satisfies Assumption 2.4. $f(x)$ is μ-strongly γ-quasar-convex and L-smooth by properties of quasar-convex functions introduced in (Hinder et al. (2020), D.3), where $L = \sum_{i=1}^{n} \|a_i\|/n + 0.5\mu$. We choose $\gamma \in \{0.5, 0.8\}$ and normalize each a_i for simplicity so that $L = 1 + 0.5\mu$. Note that each $g_\gamma(b_i a_i^\top x)$ has at least one common minimizer. Therefore, Assumption 2.7 is also satisfied by f. We use the following multi-classification dataset from Dua & Graff (2017), which we treat as binary classification datasets. We have $n = 1372$ and $d = 4$ according to the dataset. We set $\epsilon = 10^{-2}$ when $\mu = 0$, and set $\epsilon = 10^{-3}$ when $\mu > 0$. We generate the error bar the same way as simulations in section 4. Figure 4 shows that QASGD enjoys faster convergence than SGD while QASVRG enjoys fast convergence and lower complexity than QAGD and GD. When $\mu = 0.002$, Figure 4 also shows the superiority of QASVRG (Option I) in terms of convergence speed and complexity given $\kappa\epsilon \approx 0.5 < 1$. We also compare our methods with GD, QAGD and SGD on solving empirical risks of GLM with logistic link function

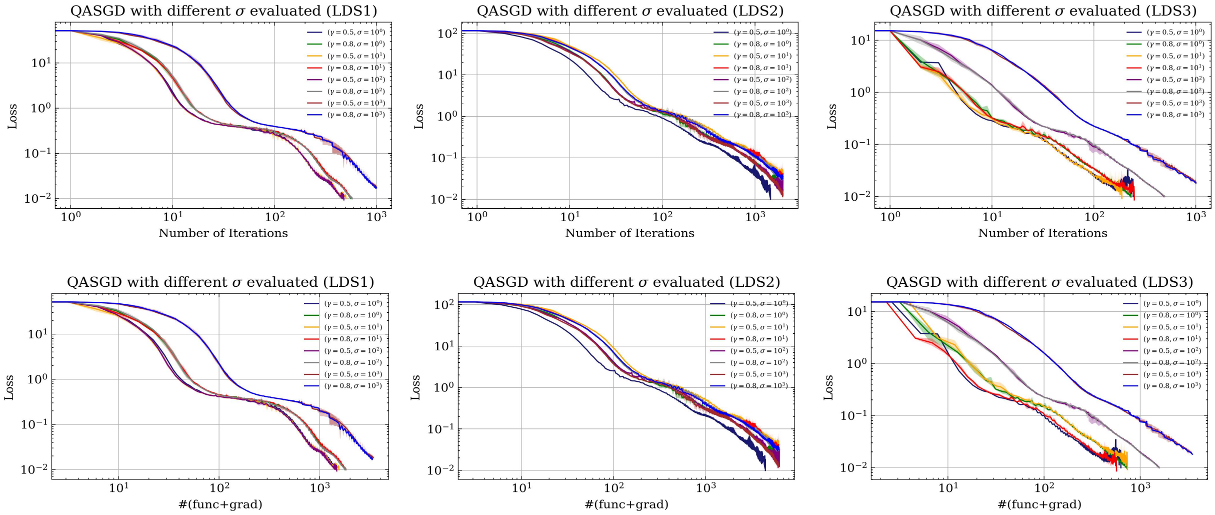

Figure 3. Evaluation of QASGD with different σ on three LDS instances

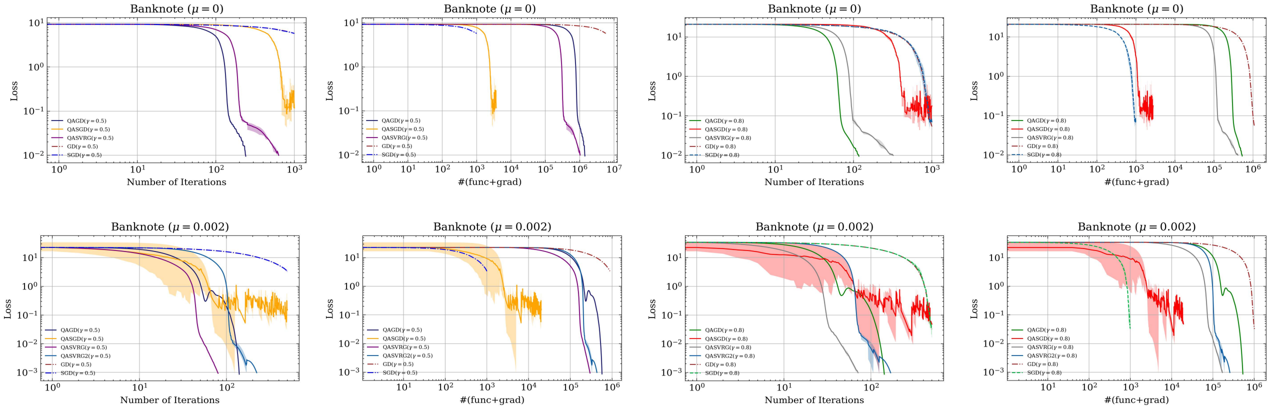

Figure 4. Evaluation of each algorithm on problem (24)

$\sigma(z) = (1 + \exp(-z))^{-1}$. Consider the following optimization problem

$$\min\left\{ f(w) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\sigma\left(w^{\top}x_i\right) - y_i\right)^2\right]\right\}, \tag{25}$$

where $x_i \sim \mathcal{N}(0, I)$, $w^* \sim \mathcal{N}(0, I)$ and $y_i = \sigma\left(w_*^{\top}x_i\right)$ for each $i \in [n]$. In our experiment, we choose $n = 5000$, $d = 50$ and the initial iterate $w_0 \sim \mathcal{N}(0, 100I)$. Since it is intractable to compute the parameter of quasar-convexity γ and smoothness L, we evaluate our methods with $\gamma = 0.5$ and $L = 10^5$ by extensive grid search.

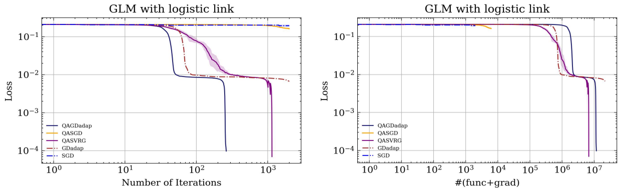

Figure 5. Evaluation of each algorithm on problem (25)