arXiv:2305.04736v2 [math.OC] 5 Jun 2023

加速 类凸性下的随机优化方法

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

摘要

非凸优化在日益增多的机器学习应用中发挥着关键作用。这促使人们识别能够实现更精确理论分析的特殊结构。其中一种已识别的结构是拟凸性，它是凸性的非凸推广，并涵盖凸函数。在随机设置下用于最小化拟凸函数的现有算法，要么复杂度较高，要么收敛较慢，因此我们推导出一类新的随机方法，用于优化光滑拟凸函数。我们证明，我们的算法收敛速度快，并且在多个示例中优于现有算法，其中包括经典的线性动力系统学习问题。我们还对新提出的算法与此前研究过的一种确定性算法进行了统一分析。

1. 引言

动量是加快优化方法收敛速度最常用的技术之一。人们提出了许多基于动量的确定性和随机算法，用于优化（强）凸函数，例如加速梯度下降（AGD）（Nes-terov，1983年；2003年；Beck & Teboulle，2009年）、加速随机梯度下降（ASGD）（Ghadimi& Lan，2012年；2016年；Kulunchakov& Mairal，2020年）、加速随机方差缩减梯度（ASVRG）方法及其相关变体（Nitanda，2016年；Allen-Zhu，2017年；Kulun-chakov & Mairal，2020年）。

尽管我们对现代优化算法的许多理解都依赖于利用目标函数凸性的能力，但越来越多的现代机器学习应用依赖于非凸优化。遗憾的是，加速算法对凸函数所能提供的理论保证的改进，并不适用于许多现实场景。对于许多光滑的非凸优化问题，我们只能保证找到驻点，而非全局极小化点。然而，一些涉及若干热门优化问题（如低秩矩阵问题、深度学习和强化学习）的非凸函数具有特殊结构，并呈现出类似凸性的性质（Ge 等，2016年；Bartlett 等，2018年；Mei 等，2020），因此有可能找到这些具有结构的非凸函数的近似全局最小值点。

本文提出了两种加速随机优化方法，用于优化拟凸函数。拟凸函数由常数 $\gamma \in (0,1]$ 参数化。 $\gamma = 1$ 意味着该函数是星形凸的，这是对凸性的放宽（Nesterov & Polyak，2006）。拟凸性最早由 Hardt 等人（2016）提出。他们证明，在若干温和的假设下，学习线性动力系统的目标函数是拟凸的。Zhou 等人（2019）和 Kleinberg et al.（2018）也提供了证据，表明神经网络的损失函数在极小值点的大邻域内可能符合星形凸性。近期有多篇论文提出了有效的确定性方法，用于最小化 L-光滑且 γ-拟凸的函数。梯度下降（梯度下降）和随机梯度下降（随机梯度下降）分别需要 $O(\gamma^{-1}\epsilon^{-1})$ 次和 $O(\gamma^{-2}\epsilon^{-2})$ 次迭代，才能得到一个 ϵ-近似解；而 Guminov & Gasnikov（2017）和 Hinder 等人（2020）提出的算法需要 $O(\gamma^{-1}\epsilon^{-1/2})$ 次迭代，Nesterov 等人（2018）提出的算法则需要 $O(\gamma^{-3/2}\epsilon^{-1/2})$ 次迭代。Hinder 等人（2020）还引入了一种新指标，以函数评估和梯度评估的总次数为依据。为计算一个 ϵ-近似解，Hinder 等人（2020）的方法在 γ-拟凸函数上需要总计 $O(\gamma^{-1}\epsilon^{-1/2}\log(\gamma^{-1}\epsilon^{-1}))$ 次评估，在 μ-强 γ-拟凸函数上则需要总计 $O(\gamma^{-1}\kappa^{1/2}\log(\gamma^{-1}\kappa)\log(\gamma^{-1}\epsilon^{-1}))$[1] 次评估。

机器学习中的许多优化问题可以

[1]中国广州中山大学 [2]美国马萨诸塞州剑桥市哈佛大学 [3]美国马萨诸塞州剑桥市麻省理工学院。联系邮箱：Qiang Fu <fuqiang7@mail2.sysu.edu.cn>，Ashia Wilson <ashia07@mit.edu>。

论文集： 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年，作者所有。

[1] $\kappa \triangleq L/\mu$ 是条件数。

表示为以下形式

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

这使其特别适用于随机优化方法。当 n 较大时，由于每次迭代都需要访问全梯度和函数值，采用确定性算法将导致较高的计算成本。因此，受凸情形下的加速随机梯度下降、ASVRG 以及Hinderet al.（2020年）的研究成果启发，我们提出了两种方法，用于求解（1）：类凸加速随机梯度下降（QASGD）方法，以及类凸加速随机方差缩减梯度（QASVRG）方法。其中，目标函数 f 是 L-光滑且（强）类凸的。我们还提出了一个统一的基于能量的框架，用于分析这些新提出的加速算法的收敛性；该框架借鉴了Wilson et al.（2021年）和Kulunchakov &Mairal（2020）提出的统一分析方法。

我们的主要贡献有三方面。

- *QASGD*：我们将带动量的 QASGD 作为加速技术。在有界梯度假设 2.4下，我们证明 QASGD 对一般拟凸函数和 $O\left(\frac{L}{t^2}+\frac{\sigma}{\gamma\sqrt{t}}+\frac{\epsilon}{2}\right)$ 具有 $O\left((1+\gamma^2/16)^{-t}+\frac{\sigma^2}{\gamma^2 t}\right)$ 的收敛速率，其中 ϵ 来自二分线搜索。我们通过实验证明，在学习时不变动力系统时，QASGD 的表现优于现有的几种方法。

- *QASVRG*：我们在小批量设置下引入了 QASVRG，这是 Nitanda （2016年）方法在具有动量这一加速技术的类凸性情形下的扩展。我们采用方差缩减和小批量方法来计算每次迭代中的随机梯度。在插值假设 2.7 和紧性假设 2.5下，QASVRG 对于强拟星凸函数和一般拟凸函数分别实现了总体复杂度[2] 为 $\widetilde{O}\left(n+\min\left\{\frac{\kappa}{\gamma^2},\ \frac{n\sqrt{\kappa}}{\gamma}\right\}\right)$ 以及 $\widetilde{\mathcal{O}}\left(n+\min\left\{\frac{LR^2}{\gamma\epsilon},\ \frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right\}\right)$。我们还针对强拟星凸函数提出了另一种具有不同参数选择的方案（表 4中的选项 II），其精确的收敛速率将在定理 3.6中给出。这两种方案对 κ 和 ϵ具有不同的依赖关系，因而适用于不同的应用场景。当 n 较大时，我们的复杂度显著低于 Hinder etal.（2020年）中的 AGD 复杂度。

[2]这里我们用总体复杂度来表示函数和梯度评估的总次数。

- 李雅普诺夫分析：我们使用标准的李雅普诺夫论证，在类凸性和光滑性条件下对所提出的算法进行了统一分析。此外，我们将 Hinder et al.（2020年）提出的 AGD 方法纳入基于能量的分析框架，并将其重新命名为 QAGD（类凸加速梯度下降）。与 Hinder et al.（2020年）中的 AGD 不同，QAGD 支持比欧几里得距离更一般的 Bregman 散度。

本文其余部分安排如下。第 2 节将进一步介绍类凸性、相关假设以及此前提出的方法。第 3 节介绍针对（强）拟凸函数的 QAGD、QASGD 和 QASVRG 主要算法及其收敛性分析。第 4 节介绍用于验证所提算法有效性的模拟。

记号 本文始终使用以下记号：$D_h(x,y)\triangleq h(x)-h(y)-\langle\nabla h(y)\ x-y\rangle$ 表示 x与 $y\in\mathbb{R}^d$之间的 Bregman 散度，其中 h 是任意的 $\bar{\mu}$-强凸函数。 $[n]\triangleq\{1,2,...,n\}\log^+()\triangleq\max\{\log()\ 1\}$, $\|\cdot\|\triangleq\|\cdot\|_2$, $\bar{a}_k\triangleq A_{k+1}-A_k$, $\bar{b}_k\triangleq B_{k+1}-B_k$, $\kappa\triangleq L/\bar{\mu}\mu$, $\mathcal{E}_k\triangleq f(y_k)-f(x^*)$。 $a\simeq b$ 表示 $a=O(b)$。 $\langle,\rangle$ 表示内积。 $\mathcal{X}^*$ 是 (1) 的解集，我们假设它非空；若点 x满足 $f(x)-f(x^*)\leq\epsilon$for$x^*\in\mathcal{X}^*$，则称为一个 ϵ-近似解。 R 表示初始距离的上界，使得 $D_h(x^*,x_0)\leq R^2$。不失一般性，我们假设 $f(x^*)\geq 0$。若 $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$ for all $x,y\in\mathbb{R}^d$，则称 f 是 L-光滑的。 $\mathcal{Q}_{\mu\gamma},\mathcal{F}_L$ 分别表示 μ-强 γ-拟凸函数集合和 L-光滑函数集合；当 $\mu=0$时， $\mathcal{Q}_{\mu\gamma}$ 退化为 γ-拟凸函数集合。我们使用 $O(\cdot)$隐藏~常数，并使用 $O(\cdot)$隐藏对数因子和常数。

2. 背景

人们越来越关注如何利用大类非凸函数中存在的结构。其中一种结构是准星凸性和强准星凸性，定义如下。

定义 2.1 （准星凸性）。 设 $\gamma\in(0,1]$，并令 x^*为可微函数 $f\colon\mathbb{R}^d\to\mathbb{R}$.的极小化点。若对所有$x\in\mathbb{R}^d$，函数相对于 x^* 是 γ-准星凸的，

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

对于 $\mu>0$，如果对于所有 $x\in\mathbb{R}^d$，函数相对于 x^* 是 μ-强 γ-拟凸的，

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

2.1. 示例

我们引入一个Hardtet al. （2016年）中关于学习线性动力系统（LDS）的经典示例。考虑以下时不变线性动力系统：

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

其中， $x_t \in \mathbb{R}, y_t \in \mathbb{R}$ 分别是时刻 t 的输入和输出； ξ_t 是从某个分布中独立同分布采样得到的随机扰动； $h_t \in \mathbb{R}^d$ 是隐藏状态，而 $\Theta \triangleq (A, B, C, D) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times 1} \times \mathbb{R}^{1\times d} \times \mathbb{R}$ 是我们旨在学习的真实参数。假设我们有 N 对训练样本 $S = \{(x^{(1)}, y^{(1)}) ...,(x^{(N)}, y^{(N)})\}$，其中每个输入序列 $x \in \mathbb{R}^T$ 都从某个分布中采样得到， y 是上述系统相应的输出；我们将这些训练样本拟合到以下模型中

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

其由 $\Theta \triangleq (\hat{}A, \hat{}B, \hat{}C, \hat{}D)$ 控制。根据训练样本和模型系统，我们考虑以下优化问题

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

Hardt etal. (2016年) 证明，在某些温和条件下，目标函数$\hat{}F(\Theta)$ 相对于Θ是弱光滑且拟凸的。

我们再介绍一个广义线性模型（GLM）的例子。考虑以下平方损失最小化问题

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

其中， $\sigma(\cdot)$: $\mathbb{R} \to \mathbb{R}$ 是链接函数； $x \in \mathbb{R}^d$ 是来自 $\mathcal{D}$ 的独立同分布样本，并且存在 $w_* \in \mathbb{R}^d$ 使得 $y = \sigma(w_*{}^\top x)$。多个文献已利用 $f(w)$ 的拟凸结构（Fosteret al.，2018年；Ma，2020；Wang &Wibisono，2023年）。

2.2. 先前的确定性方法

已有多种确定性一阶方法被开发出来，用于最小化 L-光滑 γ-拟凸函数。

Guminov & Gasnikov （2017）证明，梯度下降的收敛速率为 $O(L/\gamma t)$。他们还提出了一种加速算法，其收敛速率为$O(L/\gamma^2 t^2)$。然而，该算法在每次迭代中都依赖于一种低维子空间优化方法，而执行该方法的成本可能高得令人难以承受。

Hinder 等人（2020）提出了一种新型加速梯度方法，对于强拟星凸函数，其收敛速率为 $O((1-\gamma/\sqrt{2\kappa})^t)$。值得注意的是，当 $\gamma = 1$ 时，该速率与 Nesterov 针对强凸函数的 AGD 所达到的收敛速率相匹配。Hinder 等人（2020）提出的方法对于一般拟凸函数也能达到 $O(L/\gamma^2 t^2 + \epsilon/2)$ 的收敛速率；当 $\gamma = 1$ 时，该速率几乎与 Nesterov 针对凸函数的 AGD 所达到的收敛速率相匹配。由于引入了二分线搜索子程序来搜索适当的动量参数，收敛速率中额外出现了一个因子 ϵ （该因子可以任意小）。值得注意的是，在加速梯度下降中经典选取的动量参数（Nesterov 1983）并不能保证在类凸性下表现良好。与 Guminov & Gasnikov （2017）中的低维子空间方法相比，Hinder 等人（2020）的 AGD 中的二分线搜索至多需要进行 $O(\log(\gamma^{-1}\epsilon^{-1}))$ 次函数评估和梯度评估，成本可能显著更低。类似地，Bu & Mesbahi（2020）提出了基于动量的加速算法，这些算法依赖于一个子程序，但未对该子程序进行复杂度分析。Hinder 等人（2020）还为应用于拟凸函数的任意确定性一阶方法建立了 $\Omega(\gamma^{-1}\epsilon^{-1/2})$ 的最坏情况复杂度下界；他们的方法除对数因子外是最优的。Guminov & Gasnikov（2017）中方法的复杂度在有条件的情况下达到了这一下界。

2.3. 先前的随机方法

虽然针对拟凸函数的确定性加速方法能够实现快速收敛率和近乎最优的复杂度，但我们着重开发随机方法，以降低求解（1）时的计算复杂度。当目标函数 f 是 γ-拟凸时，Hardtet al.2016 证明，在O（$\Gamma/\gamma^2 t + \bar{\sigma}/\gamma\sqrt{t}$）有界方差和 $\bar{\sigma}^2$-弱光滑性假设下，随机梯度下降的收敛率为 Γ。

假设 2.2（有界方差）。 假设 i 从 $[n]$中独立同分布地采样。对于某个常数 $\bar{\sigma}$，有

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

弱光滑性假设比 L-光滑性更宽松。 Gower 等人 （2021年）提出了一种随机梯度方法，在 O（$\lambda^2/\gamma\sqrt{t}$）的 L-光滑性和假设2.3下，实现了收敛率。

假设 2.3 （误差残差条件）。假设 i 从 $[n]$中独立同分布地采样。对于某些常数 ρ 和 λ，有

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

方法	假设	复杂度
梯度下降（Guminov &Gasnikov，2017年）	$f \in \mathcal{F}_L$	$\widetilde{O}\left(\frac{nLR^2}{\gamma\epsilon}\right)$
随机梯度下降（Gower 等人，2021年）	$f \in \mathcal{F}_L$ &误差残差条件	$\widetilde{O}\left(\frac{(R^2+\gamma^2\lambda^2)}{\gamma^2\epsilon^2}^2\right)$
	$f_i \in \mathcal{F}_L$ &插值	$\widetilde{O}\left(\frac{LR^2}{\gamma^2\epsilon}\right)$
随机梯度下降 (Jin,2020)	$f \in \mathcal{F}_L$ &有界方差	$\widetilde{O}\left(\frac{LR^2}{\gamma\epsilon}+\frac{\bar{\sigma}^2R^2}{\gamma^2\epsilon^2}\right)$
QAGD (Hinder et al., 2020)	$f \in \mathcal{F}_L$	$\widetilde{O}\left(\frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right)$
QASGD（我们的方法）	$f_i \in \mathcal{F}_L$ &有界梯度	$\widetilde{O}\left(R\sqrt{\frac{L}{\epsilon}}+\frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)$
QASVRG（我们的方法）	$f_i \in \mathcal{F}_L$ &插值&紧性	$\widetilde{O}\left(n+\min\left\{\frac{LR^2}{\gamma\epsilon},\frac{nR}{\gamma}\sqrt{\frac{L}{\epsilon}}\right\}\right)$

表 1. 在 $f \in \mathcal{Q}_{0\gamma}$时对比一些现有方法与我们的方法

与Hardt et al.（2016年）相比，Gower 等人（2021年）的光滑性假设更强，但对梯度估计的假设在某种意义上更弱。此外，Gower 等人（2021年）证明，在假设O（$L/\gamma^2 t$）下，该收敛速率可以改进为2.7。在光滑性和方差有界的条件下，Jin （2020）对随机梯度下降进行了比Gower 等人（2021年）更精细的分析，并将分析扩展到非光滑情形。

有几种加速随机方法，在目标函数为凸时，理论上能够达到优于随机梯度下降的最坏情形收敛速率。在凸情形下，目标函数通常包含一个正则化项，该项是凸下半连续的，但不一定光滑。Ghadimi 和 Lan（2016）以及 Kulunchakov 和 Mairal（2020）提出了近端加速随机梯度下降，在 L-光滑性和 σ^2-有界方差假设下，对于一般凸函数和强凸函数，分别达到 $O(L/t^2+\sigma/\sqrt{t})$和$O((1-1/\sqrt{\kappa})^t+\bar{\sigma}^2/t)$的收敛率。方差缩减是一种有效提升收敛率的技术。Allen-Zhu（2017）和 Kulunchakov & Mairal（2020）提出了加速近端随机方差缩减梯度法，对于L-光滑（强）凸函数，其收敛率保证分别为 $O((1-\min\{1/\sqrt{3\kappa n},1/\sqrt{2n}\})^t)$和 $O(Ln/t^2)$。Nitanda（2016）提出了加速小批量随机方差缩减梯度法，用于最小化不带正则化项的（强）凸有限和目标函数。该算法采用多阶段方案，对于 L-光滑（强）凸函数，收敛率分别为 $\widetilde{O}(n+\min\{\kappa,n\sqrt{\kappa}\})$和 $\widetilde{O}(n+\min\{L/\epsilon,n\sqrt{L/\epsilon}\})$。相比之下，这些方法在更新随机方差缩减梯度法的固定锚点和控制方差方面采用了不同的策略。

2.4. 动机

受上述加速随机方法的启发，我们在不同假设条件下，将 Kulunchakov &Mairal(2020年) 提出的加速随机梯度下降和 Nitanda (2016年) 提出的 AMSVRG 推广至（强）拟凸情形。在本小节中，

在本文中，我们将讨论这些假设，以及它们与先前假设体系的比较方式。不同于 Ku-lunchakov & Mairal(2020，我们不考虑函数值和梯度的随机扰动，因为给定 x^*在扰动后可能不再是全局最小值点。此外，我们在提出的每种方法中都引入了二分线搜索（ Hinder et al.，2020），以寻找适当的动量参数。

对于 QASGD，我们的关键假设是有界梯度假设，这是非凸情形下随机梯度下降标准收敛性分析中常用的假设（Hazan &Kale，2014；Rakhlin et al.，2011；Recht et al.，2011年；Nemirovski et al.，2009）。由于强拟星凸函数具有特殊结构，其梯度并不有界，因此我们将这一假设推广如下。

假设 2.4（有界梯度）。假设 $f \in \mathcal{Q}_{\mu\gamma}$ 和 i 是从 $[n]$中独立同分布抽样得到的。对于某些 $\sigma \geq 0$和 $x^* \in \mathcal{X}^*$，有

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

在一般的类凸性下，该假设将退化为标准的有界梯度假设。与误差残差条件相比，有界梯度假设更强。若选择逻辑函数作为链接函数，例（7）满足该假设（$\mu=0$）。对于 $\mu>0$，我们考虑一个拟星凸有限和 $f=\sum_{i=1}^n f_i$ wherex^* 是 f 和 $\mathbb{E}_i[\|\nabla f_i\|^2] \leq \sigma^2$ 的极小化点。Then$g(x) = f(x) + \frac{\mu}{2}\|x-x^*\|^2$是强拟星凸的，并满足该假设。尽管某些类凸函数本身并不满足假设 2.4，但在假设 2.5 下，该假设在实践中将成立；假设 2.5 也由 Bottou 与 Le Cun（2005）、G¨urb¨uzbalaban 等（2015）和 Nitanda（2016）提出，用于分析增量方法和随机方法。我们在注 2.6 中总结了上述四个假设之间的关系。

假设 2.5（紧性）。存在一个紧集 $\mathcal{C} \subseteq \mathbb{R}^d$，其中包含通过某种优化算法生成的迭代点。

备注 2.6。假设 2.3 (ER)、假设2.2 (BV)、假设2.4 (BG) 与假设 2.5(紧性) 之间的关系如下所示。



关于 QASVRG，我们将在下一节证明，Nitanda(2016年) 提出的梯度方差上界也能给出拟凸函数的梯度方差上界（命题 3.4），前提是问题（ f_i（1）中的每个 L_i均为光滑函数，并满足以下插值假设。

假设 2.7 (插值)。存在 $x^* \in \mathcal{X}^*$，使得对于所有 $i \in [n]$

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

插值假设在过参数化机器学习模型中很常见，近年来受到广泛关注（Zhou et al., 2019年；Ma 等人，2018年；Vaswani 等人，2019年；Gower 等人，2021年）。如果模型具有足够多的参数，就能够完全插值带标签的训练数据。特别地，当 $\xi_t = 0$ 时，例（6）满足插值假设，因为 Θ 也是由每个训练样本生成的目标函数的全局最小值点。对于每个 $x \sim \mathcal{D}$，若 $y = \sigma(w_*^{\mathsf{T}} x)$，例（7）也满足该假设。令 $L = \max_i\{L_i\}$；为简洁起见，我们假设每个 f_i 都是 L-光滑的。此外，Nitanda（2016）仅针对凸函数和强凸函数给出了一种参数选择。本文在强准星凸性条件下提供了两种参数选择（选项 I 和 II）。类似地，表 4 中的选项 II 与一般拟星凸函数的参数选择相同。选项 I 与 Nitanda（2016）的 AMSVRG 的直接扩展略有不同。第 3.2 小节将对这两种参数选择作更深入的技术比较。

3. 算法

为了求解（1），QAGD、QASGD 和 QASVRG 需要从预言机获取梯度或梯度估计，我们将其记为 ∇_k。本文考虑以下梯度（估计）：

- **全梯度**：$\nabla_k = \nabla f(x_{k+1})$。问题（1）变为确定性的。
- **随机梯度**：$\nabla_k = \nabla f_i(x_{k+1})$ 与索引 i 从 $[n]$ 中独立同分布地采样。我们有 $\mathbb{E}_i[\nabla_k] = \nabla f(x_{k+1})$, 其中 $\mathbb{E}_i$ 表示关于索引 i 的期望。
- **小批量随机方差缩减梯度**：$\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(\tilde{x}) + \nabla f(\tilde{x})$，其中 $\tilde{x}$ 是每个阶段固定的锚点；$I_k = \{i_1, i_2, ..., i_{b_k}\}$ 从 $[n]$ 中独立同分布地抽取，且满足 $f_{I_k} \triangleq \frac{1}{b_k}\sum_{j=1}^{b_k} f_{i_j}$。批量大小为 $|I_k| = b_k$。我们有 $\mathbb{E}_{I_k}[\nabla_k] = \nabla f(x_{k+1})$，其中 $\mathbb{E}_{I_k}$ 表示关于小批量 I_k 的期望。

3.1. Quasar 加速算法

我们将算法 1 作为一个通用框架引入，该框架通过不同的参数选择整合了 QAGD、QASGD 以及 QASVRG 的单个阶段。基于算法 1，我们还引入了算法 2 所描述的多阶段 QASVRG。值得注意的是，我们在附录中提供了有关 Bisearch（算法 1 第 3 行）的更多信息，包括完整的算法，以及 Hinder 等人（2020）给出的相应复杂度分析（算法 3、引理 A.5）。我们的方法所保证的性能依赖于内部假设 3.1。关系式 (8) 要求 h 是 $\bar{\mu}$-强凸的；关系式 (9) 则是利用 Bregman 散度作为距离对 μ-强 γ-类凸性的一种推广，在接下来的分析中，我们将用它替代关系式 (3)。当 $h = \frac{1}{2}\|\cdot\|^2$ 时，该关系式与 (3) 完全相同。在线性动力系统的欧几里得空间中，只要 f 关于 x^* 是 μ-强 γ-类凸的，关系式 (10) 就成立（Hinder 等人（2020），推论 1）。

假设 3.1。假设 f_i 对每个$i \in [n]$均可微。对于某些 $\bar{\mu} > 0$、$0 < \gamma \leq 1$ 和 $\mu \geq 0$，并且对所有 x，$y \in \mathbb{R}^d$，要求

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

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

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

3.2. 收敛性分析

我们使用以下李雅普诺夫函数，对 QAGD、QASGD 和 QASVRG 进行统一分析：

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

其中 A_k 和 B_k 是正的不减序列，表2、表3 和表4 中所示的参数选择与它们密切相关。基于李雅普诺夫分析推导出的收敛率，我们得出了欧几里得空间中每种方法的复杂度上界（$h = \frac{1}{2}\|\cdot\|^2$）。由于 QAGD 的收敛结果已经

算法 1 $(A_k, B_k, \tilde{y}_0, t, \epsilon)$

要求： h 满足 $D_h(x,y) \geq \frac{\bar{\mu}}{2}\|x-y\|^2$; $\tilde{f} \in \mathcal{F}_L$; $f \in \mathcal{Q}_{\mu\gamma}$.

初始化 $x_0 = z_0 = y_0 = \tilde{y}_0$ 并指定 $\theta_k \triangleq (\nabla_k, \alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon})$

对于 $k = 0, \ldots, t-1$ **执行**

$\tau_k \leftarrow$ Bisearch($\tilde{f}, y_k, z_k, b, c, \tilde{\epsilon}$) (搜索 $\tau_k \in [0,1]$ 满足 $\tau_k\langle \nabla \tilde{f}(d_k), y_k - z_k\rangle - b\|d_k - z_k\|^2 \leq c(\tilde{f}(y_k) - \tilde{f}(d_k)) + \tilde{\epsilon}$ 其中 $d_k := \tau_k y_k + (1-\tau_k)z_k$; 详见算法 3 了解更多详情)

$x_{k+1} \leftarrow (1-\tau_k)z_k + \tau_k y_k$ (耦合步骤)

$z_{k+1} \leftarrow \arg\min_{z\in\mathbb{R}^d} \{\langle \nabla_k, z - z_k\rangle + \beta_k D_h(z, z_k) + \alpha_k D_h(z, x_{k+1})\}$ (镜像下降步骤)

$y_{k+1} \leftarrow \arg\min_{y\in\mathbb{R}^d} \left\{\rho_k\langle \nabla_k, y - x_{k+1}\rangle + \frac{1}{2}\|y - x_{k+1}\|^2\right\}$ (梯度下降步骤)

结束循环

输出 y_t

算法 2 $(A_k, B_k, b_k, \tilde{y}_0, p, q, \epsilon)$

要求： $D_h(x,y) \geq \frac{\bar{\mu}}{2}\|x-y\|^2$; $f_i \in \mathcal{F}_L$; $f \in \mathcal{Q}_{\mu\gamma}$; $q \leq \frac{1}{4}$; $p \leq \frac{\gamma\mu}{16}$

初始化 $y_0 = \tilde{y}_0$

对于s=0 ,1,... **执行**

$$t \leftarrow \begin{cases} \sqrt{\frac{17LD_h(x^*,y_s)\gamma^2\mu q f(y_s)}{}} & \mu = 0 \text{或 } \mu > 0 \text{(选项 II)} \\ \text{对数}^{1+\frac{\gamma}{\sqrt{8}}}\left(\frac{2}{\gamma q}\right), & \mu > 0 \text{（优化 I）} \end{cases}$$

$y_s \leftarrow$ 算法 1$(A_k, B_k, y_0, \lceil t\rceil, \epsilon)$ (指定 $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ 其中 $|I_k| = b_k$)

$y_0 \leftarrow y_s$

结束循环

输出 y_s

已在 Hinderet al. (2020) 中建立，因此本小节不再提供 QAGD 的收敛性分析。相反，我们在 Lyapunov框架下对 QAGD 进行收敛性分析，并获得与 Hinderet al.(2020) 相吻合的结果。相关证明和参数选择见附录 C 和 D。

定理3.2 (QASGD)。 假设满足假设*3.1* 和假设*2.4* ， $D_h(x^*, z_0) \leq R^2$， $f_i \in \mathcal{F}_L$ 对所有$i \in [n]$ 并任取 $\tilde{y}_0 \in \mathbb{R}^d$。则采用*1* 号算法，并选取 $\nabla_k = \nabla f_i(x_{k+1})$ 和 A_k， B_k， θ_k 表*3*中指定的参数，满足

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

推论3.3。 考虑在定理*3.2*相同假设下的 *QASGD*。则 *QASGD* 达到 $\mathbb{E}[f(y_t) - f(x^*)] \leq \epsilon$的总体复杂度上界为

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

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

分别对应 $\mu = 0$和 $\mu > 0$ 。

以下命题表明，QASVRG 的梯度估计方差会与目标函数同样快速地降低。这是我们证明中的一项关键技巧，用于控制随机梯度方差。该命题也由Nitanda (2016)提出，他们假设 f_i 是凸且光滑的。本文通过使用假设 f_i ，规避了对2.7的凸性要求。命题3.4 的证明推迟至附录 B。

命题3.4 (方差上界)。 假设*2.7*成立，$\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(\tilde{x}) + \nabla f(\tilde{x})$ 且$f_i \in \mathcal{F}_L$ 对于每个 $i \in [n]$，则可得到以下不等式

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

其中 $|I_k| = b_k$ 和 $\mathbb{E}_{I_k}$ 表示关于小批量 I_k的期望。

定理 *3.5*（*QASVRG*（单阶段））。假设假设 *3.1* 和假设 *2.7* 成立， $D_h(x^*, z_0) \leq R^2, f_i \in \mathcal{F}_L$ 对每个 $i \in [n]$, 并任意选择 $\tilde{y}_0 \in \mathbb{R}^d$。那么，算法 *1* 采用表中指定的 $\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ 和 A_k, B_k, b_k, θ_k

4 满足

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

其中 $p \leq \frac{\gamma\mu}{16}$ 由用户指定。

在假设 2.5 下，如果由算法2 生成的 $\{y_s\}$ 被限制在一个紧集内，则 $\{D_h(x^*, y_s)\}$ 可以被某个常数一致地界定。因此，我们可以将 Bregman 散度隐藏在 $O(\cdot)$和 $O(\cdot)$中。请注意，只有在 $\mu = 0$时我们才需要这一假设。根据定理3.5，单阶段 QASVRG 存在偏差，这意味着无法通过单阶段 QASVRG 生成一个 ϵ-近似解。例如，当 $\mu = 0$时，偏差的上界为 $O\left(\left(\frac{p}{\gamma} + \epsilon\right) f(y_0)\right)$；但当 $t = \Omega\left(\sqrt{\frac{LR^2}{\gamma^2 f(y_0)}}\right)$时，最优性差距的期望 $\mathcal{E}_t$ 可以在每个阶段以较小的p和 q递减。因此，我们需要$O(\log(1/\epsilon))$个阶段来生成一个 ϵ-近似解。

推论 3.6。 在假设 *2.5* 以及定理*3.5*中的相同假设下，算法 *2* 达到 $\mathbb{E}\left[f(y_s) - f(x^*)\right] \leq \epsilon$ 所需的总体复杂度上界为

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

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

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

分别对应于 $\mu = 0$ 和 $\mu > 0$ （最后两个界），其中 $q \in (0,\ 1/4]$。

证明概要 我们用于推导各算法收敛率和复杂度的方法是统一的。我们先计算每个 Lyapunov 阶段之间的差值，然后取条件期望，从而得到 $\mathbb{E}[E_{k+1} - E_k]$的上界。最后，我们对 k 求和，以得出各算法的收敛率及相应的迭代复杂度。将该复杂度与 Bisearch 的复杂度结合起来，即可得出各推论中的总体复杂度。更多细节请参见附录C和D。

QASVRG 的最后两个界对应于表 4 中两种不同的参数选择（选项 I 和选项 II）。选项 I 得出的复杂度界对 κ 的依赖更不利，而选项 II 得出的复杂度界对 ϵ的依赖更不利。这表明，选项 I 在条件良好的问题（例如 $\kappa\epsilon < 1$ ）上表现更好，而选项 II 在条件较差的问题（例如 $\kappa\epsilon \gg 1$）上表现更好。

QASGD 和 QASVRG 的复杂度界中包含额外的对数因子，这是由 Bisearch 引起的。 Hinder et al.(2020证明，若该子程序中涉及的函数是 L-光滑的，则 Bisearch 的复杂度至多相差一个对数因子。在随机设置下，若涉及的函数是单个f_i 或一个由 f_{I_k}组成的小批量，则由于采用了均匀采样，我们需要假设对所有 $f_i \in \mathcal{F}_L$ i 均成立。

我们在表1中总结了我们的方法和一些现有方法，包括它们各自的假设和复杂度上界。总而言之，当 n 较大时，QASGD 和 QASVRG 的复杂度上界都优于 QAGD；在更强的假设下，QASGD 比随机梯度下降具有更快的收敛速度和更低的复杂度。尽管 QASVRG 每个阶段都需要访问完整梯度和函数值，因此其计算开销可能高于随机梯度下降，但从理论上讲，它的收敛速度比随机梯度下降更快。

4. 仿真

在本节中，我们使用合成数据集，在欧几里得空间中评估我们的方法在例子（6）上的表现，其中每个输入序列 $x^{(i)} \sim \mathcal{N}(0, 1)$ 按坐标生成。与 Hardt et al. (2016年)不同，我们在训练前生成 N 个训练样本和随机扰动 ξ_t，而不是在每次迭代时生成新数据和随机扰动。因此，例子（6）可以改写为

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

其中，上标 (i) 表示输出是使用 i^{th} 训练数据 $(x^{(i)}, y^{(i)})$ 生成的。与 Hardt 等人（2016）类似，我们实验中的实际目标函数为$F(\hat{\Theta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T-T_1}\sum_{t>T_1}\|\hat{y}_t^{(i)} - y_t^{(i)}\|^2\right]$，其中$T_1 = T/4$。我们按照 Hardt 等人（2016）的方法，使用参数 $N = 5000, d = 20$, $T = 500$生成真实的动力系统和数据。遵循 Hinder 等人（2020）的方法，我们通过扰动真实系统的参数来生成初始迭代点 (A_0, C_0, D_0)，并保持 A_0 的谱半径严格小于 1。我们选择随机种子值属于 $\{0, 12, 24, 36, 48\}$，以生成五个真实的 LDS 实例及其初始化。本节仅展示 $\{0, 24, 48\}$ 的模拟结果，其余结果见 F 节。注意，B 不是可训练参数，因为 B 是已知的。正如 Hardt 等人（2016）所述，精确计算 LDS 目标函数的 quasar 凸性参数 γ，甚至对其进行估计，都是难以处理的。因此，我们对我们的方法进行评估。

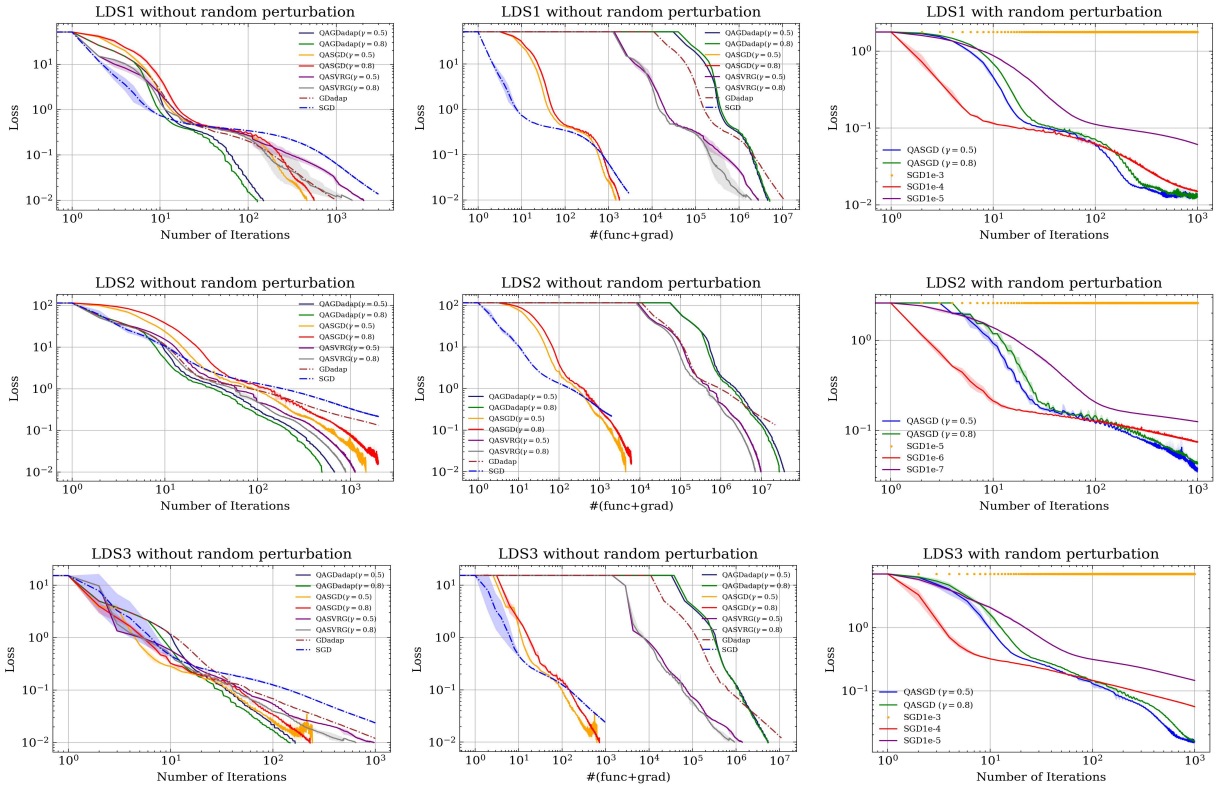

图*1*。 在三个不同的线性动力系统实例上的评估。对于线性动力系统1、线性动力系统2和线性动力系统3，我们分别为随机梯度下降选择 $\epsilon = 10^{-2}$，步长为 5×10^{-5}， 1×10^{-6}， 1×10^{-4}；为 QASGD 选择 $L = 1 \times 10^{6}$， 1×10^{8}， 1×10^{5}；为 QASVRG 选择 $L = 3 \times 10^{4}$， 1×10^{6}， 1×10^{4}。第三列中的平线表示在此步长设置下，损失会发散到无穷大。

对于 $\gamma \in \{0.5, 0.8\}$。我们观察到，不精确的 γ 不会影响所有相关方法的性能。如表3所示， η 涉及 QASGD 的参数选择；在我们的模拟中，我们将其选为 $\min \left\{ \frac{1}{L}, \frac{\sqrt{2}\gamma\|z_0\|}{\sigma\ t+1}(\ )^{3/2} \right\}$。虽然假设 2.4 中的 $\bar{\sigma}$ 取决于紧集和初始化，但通过在具有 $\sigma \in \{1, 10, 10^2, 10^3\}$的三个 LDS 实例上评估 QASGD，我们观察到，对于所有模拟，选择 $\sigma = 1$都是一种稳健的选择（更多详情请参见附录F中的图3）。根据 Hardt 等人（2016）的分析， $F(\Theta)$ 是 L-弱光滑的，而 $F(\Theta)$ 是否为 L-光滑仍属未知。鉴于我们的方法的参数选择涉及 L，我们对 QASGD 和 QASVRG 的 L 值进行微调，并在每个实例中通过大量网格搜索为 SGD 选择最佳步长。对于 QAGD 和梯度下降，我们采用与 Hinder 等人（2020）相同的自适应步长。我们考虑随机噪声 $\xi_t \sim \mathcal{N}(0, 10^{-2})$ 或 $\xi_t = 0$ 对真实系统输出的扰动。如果 $\xi_t \sim \mathcal{N}(0, 10^{-2})$，则插值假设会被违反，因为 Θ 不再是每个训练样本所生成目标函数的全局最小值点。因此，在这种情况下，我们只评估 SGD 和 QASGD。在算法 2 中，计算 t 可能很困难，尤其是在 L 未知时。因此，我们可以具体规定……

化 t 相对较大（我们选择 $t = 10^4$），并在算法 1 中采用适当的重启方案，以提升 QASVRG 的性能。遵循 Nitanda (2016 年) 的做法，当关系 $\langle \nabla_k, y_{k+1} - y_k \rangle > 0$ 成立时，我们中断算法1 并返回 y_s，然后开始下一阶段。由于我们通过 ˆ生成初始迭代点 $\rho(A_0) < 1$，训练期间没有使用 Hardt et al. (2016年提出的梯度裁剪或投影。图1中的误差条是通过对每种随机算法运行三次所得结果取平均值生成的，并逐点选取最大值和最小值分别作为上误差条和下误差条。

图 1 中的模拟结果验证了我们的方法，并展示了我们的方法在收敛速度和总体复杂度方面的优越性。此外，QASGD 和 QASVRG 对随机梯度的随机采样均具有鲁棒性。代码可在 h ttps://github.co m/QiangFu09/Stochastic-quasar-convex-acceleration 获取。我们的模拟中存在一个有趣的现象。当 t 较大时，QASGD 的收敛速率与 SGD 的速率相匹配，但图 1 仍然显示出 QASGD 相较于 QASVRG 的显著优越性。事实上，QASGD 具有 $O\left(\frac{1}{t^2} + \frac{1}{\sqrt{t}} + \frac{\epsilon}{2}\right)$

这表明存在一个由 $O\left(\frac{1}{t^2}\right)$ 主导收敛的快速初始阶段。我们推测，在我们的大多数模拟中，$\overline{\text{Q}}$ASGD 并未逃离初始阶段，因此能够实现快速收敛。上述模拟结果促使我们重新思考：QASGD 是否确实需要有界梯度假设，因为 LDS 的目标函数并不满足这一假设，但 QASGD 在 LDS 上的表现仍然良好且稳健。我们认为，该假设用于理论分析，但可能并非必要。

5. 讨论

本文提出了 QASGD 和 QASVRG，在相应假设下实现了快速收敛和较低复杂度。我们在统一框架下提出这些算法，并通过单一的能量分析建立了 QAGD、QASGD 和 QASVRG 的收敛率与复杂度。最后，我们简要讨论了若干可能的未来研究方向。

首先，我们为 QASGD 引入了有界梯度假设，但这一假设能否在一定程度上放宽，仍有待研究。鉴于 Gower 等人(2021年) 和 Jin (2020) 分别在误差残差条件和有界方差假设下证明了随机梯度下降在拟凸函数上的收敛，因此，探究能否在 QASGD 中采用这些较弱的假设，将会很有意义。

其次，Hinderet al.(2020) 基于其为一阶确定性方法建立的下界，证明了 QAGD 近乎最优。未来，我们希望针对不同假设下应用于拟凸函数的一阶随机方法，建立最坏情形复杂度下界。我们预计，这类界也将证明我们的方法近乎最优。

此外，高阶方法通常能带来更快的收敛速度。Nesterov & Polyak (2006) 提出了用于优化星形凸函数的三次正则化牛顿法（$\gamma = 1$）。因此，我们也有兴趣探究能否使用高阶方法来提升我们方法的收敛速度。

最后，我们希望在机器学习中发掘准凸函数的更多应用。

致谢

我们谨向匿名评审者致谢，感谢他们提出的建设性意见；感谢 Kyurae Kim（宾夕法尼亚大学）就引文提供的有益讨论；并感谢 Oliver Hinder（匹兹堡大学）和 Nimit Sohoni（斯坦福大学）分享线性动力系统（LDS）的代码。

参考文献

Allen-Zhu, Z. Katyusha：首个直接加速随机梯度方法。机器学习研究杂志，18(1):8194–8244，2017。

Bartlett, P., Helmbold, D., and Long, P. 通过深度残差网络进行带恒等初始化的梯度下降，可以高效学习正定线性变换。载于国际机器学习会议，第 521–530 页。PMLR，2018年。

Beck, A. and Teboulle, M. 用于线性逆问题的快速迭代收缩阈值算法。*SIAM* 影像科学杂志，2(1):183–202，2009年。

Bottou, L. 和 Le Cun, Y. 超大规模数据集的在线学习。商业与工业中的应用随机模型，21(2):137–151，2005年。doi: https://doi.org/10. 1002/asmb.538。

U
R
L
h
t
t
p
s://
o
n
l
inelibrary.wiley.com/doi/abs/10.1002/asmb.538。

Bu, J. 和 Mesbahi, M. 关于非凸优化中 Nesterov 加速方法的札记：一种弱估计序列方法。*arXiv* 预印本 *arXiv:2006.08548*，2020年。Dua, D. 和 Graff, C. UCI 机器学习库，2017年。URLhttp://archive.ics.uci.edu/ml。Foster, D. J.、Sekhari, A. 和 Sridharan, K. 非凸学习与优化中梯度的均匀收敛。神经信息处理系统进展，31，2018年。Ge, R.、Lee, J. D. 和 Ma, T. 矩阵补全不存在伪局部极小值。神经信息处理系统进展，29，2016年。Ghadimi, S. 和 Lan, G. 强凸随机复合优化的最优随机逼近算法 I：通用算法框架。*SIAM*优化期刊，22(4):1469–1492，2012年。Ghadimi, S. 和 Lan, G. 非凸非线性及随机规划的加速梯度方法。数学规划，156(1):59–99，2016年。Gower, R.、Sebbouh, O. 和 Loizou, N. 结构化非凸函数的随机梯度下降：学习率、小批量处理与插值。载于国际人工智能与统计会议，第 1315–1323 页。PMLR，2021年。Guminov, S. 和 Gasnikov, A. 加速方法用于 α-弱拟凸问题。*arXiv* 预印本 *arXiv:1710.00797*，2017年。G¨urb¨uzbalaban, M.、Ozdaglar, A. 和 Parrilo, P. 一种全局收敛的增量牛顿法。数学规划，151(1):283–313，2015年。

Hardt, M.、Ma, T. 和 Recht, B. 梯度下降可以学习线性动力系统。*arXiv* 预印本 *arXiv:1609.05191*，2016年。

Hazan, E. 和 Kale, S. 突破遗憾最小化壁垒：随机强凸优化的最优算法。机器学习研究期刊，15(1):2489–2512，2014年。

Hinder, O.、Sidford, A. 和 Sohoni, N. 最小化星形凸函数及其他问题的近乎最优方法。载于学习理论会议，第1894–1938页。PMLR，2020年。

Jin, J. 关于拟凸优化的一阶方法收敛性。*arXiv*预印本 *arXiv:2010.04937*，2020年。Kleinberg, B., Li, Y., and Yuan, Y. 另一种视角：随机梯度下降何时会逃离局部极小值？载于国际机器学习会议，第 2698–2707 页。PMLR，2018年。Kulunchakov, A. and Mairal, J. 随机复合优化的估计序列：方差缩减、加速和抗噪鲁棒性。机器学习研究杂志，21(1):6184–6235，2020年。Ma, S., Bassily, R., and Belkin, M. 插值的力量：理解随机梯度下降在现代过参数化学习中的有效性。载于国际机器学习会议，第 3325–3334 页。PMLR，2018年。Ma, T. 为什么局部方法能解决非凸问题？2020年。Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. 关于 Soft max策略梯度方法的全局收敛速率。载于国际机器学习会议，第 6820–6829 页。PMLR，2020年。Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. 面向随机规划的鲁棒随机逼近方法。*SIAM*优化期刊，19(4):1574–1609，2009年。Nesterov, Y. 一种求解凸规划问题的方法，其收敛率为 $o(1/k^2)$。苏联科学院院刊，269:543–547，1983年。Nesterov, Y. 凸优化导论：基础课程，第 87 卷。Springer科学与商业媒体，2003年。Nesterov, Y. and Polyak, B. T. 牛顿法的三次正则化及其全局性能。数学规划，108(1):177–205，2006年。

Nesterov, Y., Gasnikov, A., Guminov, S., andDvurechensky, P. 用于凸和非凸优化问题的带线搜索原始-对偶加速梯度下降。*arXiv*预印本 *arXiv:1809.05895*，2018年。

Nitanda, A. 用于最小化有限和的加速随机梯度下降。载于人工智能与统计，第195–203页。PMLR，2016年。

Rakhlin, A.、Shamir, O. 和 Sridharan, K. 使梯度下降在强凸随机优化中达到最优。*arXiv* 预印本 *arXiv:1109.5647*，2011年。

Recht, B., Re, C., Wright, S. 和 Niu, F. Hogwild!：一种无需锁的随机梯度下降并行化方法。神经信息处理系统进展，24，2011年。

Schmidt, M. 和 Roux, N. L. 强增长条件下随机梯度下降的快速收敛。*arXiv* 预印本 *arXiv:1308.6370*，2013年。Vaswani, S.、Bach, F. 和 Schmidt, M. 过参数化模型中随机梯度下降的快速与更快收敛，以及加速感知机。载于第*22*届人工智能与统计国际会议，第 1195–1204 页。PMLR，2019年。Wang, J.-K. 和 Wibisono, A. 非凸优化中类星凸函数的连续化加速。*arXiv* 预印本 *arXiv:2302.07851*，2023年。Wilson, A. C.、Recht, B. 和 Jordan, M. I. 优化中加速方法的李雅普诺夫分析。机器学习研究期刊，22:113–1，2021年。Zhou, Y.、Yang, J.、Zhang, H.、Liang, Y. 和 Tarokh, V. 通过星形凸路径，随机梯度下降在深度学习中收敛至全局最小值。*arXiv* 预印本 *arXiv:1901.00451*，2019年。

A. 有用的引理和假设

引理 A.1 (三点恒等式)。 对于所有 $x \in dom\ h$ 和 $y, z \in int(dom\ h)$

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

其中 $D_h(x,\ y) = h(x) - h(y) - \langle \nabla h(y)\ x - y\rangle$。

引理 A.2 (Fenchel-Young 不等式)。 对于所有 $x,\ y \in \mathbb{R}^d$ 和 $p \neq 0,$ 有

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

其中 $\|\cdot\|_*$ 表示 $\|\cdot\|$的对偶范数。

引理 A.3 (Hinder et al. (2020)，引理 2)。 设 $f\colon \mathbb{R}^d \to \mathbb{R}$ 可微，且设 $y,\ z \in \mathbb{R}^d$。对于 $\tau \in \mathbb{R}$ 定义$x_\tau \triangleq \tau y + (1-\tau)z$。对于任意 $c \geq 0$存在 $\tau \in [0,1]$使得

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

事实上，我们可以在一定程度上放宽条件（14），并且可以找到一个合适的动量参数 τ 满足

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

对于 $b,\ c,\ \tilde{\epsilon} \geq 0$. 若存在满足条件（$\tau$14），则意味着存在 τ 满足条件（1 5).

引理 A.4。 若 $f\colon \mathbb{R}^d \to \mathbb{R}$ 是 L-光滑的，则对于任意 $x, y \in \mathbb{R}^d$

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

N现在我们介绍一种算法，以有效地搜索任意 L-光滑函数的“良好” τ。

引理 A.5 (Hinder et al. (2020), C.2)。 对于 L-光滑 $f\colon\ \mathbb{R}^d \to \mathbb{R},\ x,\ v \in \mathbb{R}^n$ 以及标量 $b,\ c,\ \tilde{\epsilon} \geq 0,$ 算法 (3)计算出满足 $\alpha \in [0,1]$ *(15)* 的量，且至多需要

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

函数和梯度评估，其中 $\log^+(\cdot) \triangleq \max\{\log(\cdot)\ 1\}$。

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

要求： f**是** L**-光滑的；** $z, y \in \mathbb{R}^d$**；** $c \geq 0$**；如果提供了“猜测”** $\in [0,1]$。**“猜测”可以是凸性下的动量参数或其他参数。1：定义** $g(\alpha) \triangleq f(\tau y + (1-\tau)z)$**和** $p \triangleq b\|z-y\|^2$**2：如果提供了猜测且** cg**(猜测)** $+$ **猜测** $\cdot\ (g'$**(猜测)** $-$ **猜测** $\cdot\ p)$ $\leq cg$**(1)，则返回猜测 3：如果** g'**(1)** $\leq p + \tilde{\epsilon}$**，则返回 1；4：否则，如果** $c = 0$ **或** g**(0)** $\leq g$**(1)** $+ \tilde{\epsilon}/c$**，则返回 0；5：** $\delta \leftarrow 1 - g'$**(1)**$/L\|z-y\|^2$**6：lo**$\leftarrow 0$**, hi**$\leftarrow \delta, \tau \leftarrow \delta$**7：当** $cg(\tau) + \tau(g'(\tau) - \tau p) > cg$**(1)** $+ \tilde{\epsilon}$**时执行 8：** $\tau \leftarrow$ **(lo** $+$**hi)**$/2$**9：如果** $g(\tau) \leq g(\delta)$**，则 hi** $\leftarrow \tau$**10：否则 lo**$\leftarrow \tau$**11：循环结束 输出** τ

命题 A.6. 根据假设 *3.1*，我们有 $\kappa \geq \frac{\gamma}{2-\gamma}$，其中 $\kappa = \frac{L}{\mu\mu}$且 $\mu > 0$。 —

证明。

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

因此，我们有 $\frac{\gamma\mu}{2-\gamma} \le \frac{L}{\bar{\mu}}$。此外，我们有 $\sqrt{\kappa} \ge \sqrt{\frac{\gamma}{2-\gamma}} \ge \sqrt{\gamma^2} = \gamma$。 □

引理 A.7. 假设条件 2.7 成立，且每个 $f_i \in \mathcal{F}_L$。如果 $\nabla_k = \nabla f_i(x_{k+1}) - \nabla f_i(y_k) + \frac{1}{n}\sum_{i=1}^n f_i(y_k)$，我们可以得到

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

证明。由于 f 是 L-光滑的，对于任意 x，$y \in \mathbb{R}^d$ 我们有

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

令 $\nabla g(\tilde{x}) = 0$，且 $\tilde{x} = y - \frac{1}{L}\nabla f(y)$ 是 $g(x)$ 的极小化点。并且我们有

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

由于 $\nabla f(x^*) = 0$，我们有

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

根据（18），我们可以将 $\mathbb{E}\left[\|\nabla_k - \nabla f(x_{k+1})\|^2\right]$ 上界估计如下

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

其中，第一个不等式用到了关系 $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$（对所有随机变量 X均成立）；第二个不等式用到了关系 $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$。 □

引理 A.8。 设 $\{\xi_i\}_{i=1}^n$ 是 $\mathbb{R}^d$ 中的一组向量，$\bar{\xi}$ 表示 $\{\xi_i\}_{i=1}^n$的平均值。设 I 表示一个均匀随机变量，它代表 $\{1,2,...,n\}$ 的一个子集，且该子集的大小为 b。那么可得：

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

我们请读者参阅 补充材料中 Nitanda （2016年，其中给出了上述引理的证明细节 .

B. 命题3.4 的证明

基于引理 A.7 和引理A.8，我们证明 3.4。

证明。令 $\nabla_k^i = \nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x})$ 且 $\nabla_k = \frac{1}{b_k}\sum_{i\in I_k}\nabla_k^i$。由引理 (A.8)，可得

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

其中，不等式由引理（A.7）得出。 □

C. 收敛速率证明

C.1. QAGD 的证明

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

表 *2.* QAGD 的参数选择

我们首先使用李雅普诺夫函数 1， $\nabla_k = \nabla f(x_{k+1})$ 以及表 2 中指定的参数，开始执行算法 11：**情形 1：** $\mu = 0$ 对于 $\nabla_k = \nabla f(x_{k+1})$ ，我们首先使用李雅普诺夫函数 1 开始执行算法 11：

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

第二个不等式由 Fenchel-Young 不等式推出，最后一个不等式由 f 的类凸性以及（15）（引理 A.3）推出。根据表 2中总结的参数选择，我们得到以下界：

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

最后一个不等式由（16）得出。事实上，（20）可由梯度下降、 $\rho_k = 1/L$ 以及 L-光滑性推出。

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

取 β_k 和 A_k，我们有

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

因此，我们得到最终界：

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

对上述不等式两边求和，我们得到

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

情形 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)}{\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 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^*)) \\
&\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 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^*)) \\
&\le \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}$$

第一个等式和第三个等式由引理 A.1 得出；第二个等式由镜像下降得出。第一个不等式由 h的强凸性得出，第二个不等式则由（15）（引理 A.3）得出。取¯为 α_k 和 β_k，我们有 $\alpha_k B_k/\beta_k = b_k$，这解释了第一个不等式。此外，取 B_k 并结合观察 A.6，我们有

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

结合初始界以及上述关系，我们得到以下界：

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

h 的 Fenchel-Young 不等式和 f的强准星凸性给出了第三个不等式。第五个不等式由 f的 L-光滑性得出。取 B_k, β_k和 A_k，我们有

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

这解释了最后一个不等式。因此，我们得到

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

C.2. 定理3.2的证明

QASGD
γ-类凸$(\mu=0)$
$A_k=\eta(k+1)^2, \eta=$最小化 $\left(\frac{\bar\mu}{L}, \sqrt{\frac{4D_h(x^*,\sigma^2 z_0}{\quad}}\frac{\gamma}{\quad})(t+1)^{3/2}\right), B_k=1$ $\tilde{}(\alpha_k,\beta_k,\rho_k,f,b,c,\tilde\epsilon) \leftarrow \left(0,\frac{\gamma}{\bar a_k},0,f_i,0,\frac{\gamma A_k}{\bar a_k},\frac{\gamma\epsilon}{2}\right)$
μ- γ-强拟凸$(\mu>0)$
步骤 1：选择 A_k, B_k, θ_k，如下所示，直到收敛
$A_k=\left(1+\text{最小化}\ \{\gamma^2\bar\mu^2/16, 1/2\}\right)^k, B_k=\mu A_k$ $(\alpha_k,\beta_k,\rho_k,\tilde f,b,c,\tilde\epsilon) \leftarrow \left(\frac{\gamma\mu}{2},\frac{\gamma\mu B_k}{2\bar b_k},0,f_i,\frac{\gamma\bar\mu\mu}{4},\frac{\gamma A_k}{2\bar a_k},0\right)$
步骤 2：重新开始并按如下方式选择 A_k，B_k，θ_k
$A_k=\frac{\gamma^2\bar\mu^2}{36}\left(k+\text{最大值}\{48/\gamma^2\bar\mu^2,5\}\right)^2, B_k=\mu A_k$ $(\alpha_k,\beta_k,\rho_k,\tilde f,b,c,\tilde\epsilon) \leftarrow \left(\frac{\gamma\mu}{2},\frac{\gamma\mu B_k}{2\bar b_k},0,f_i,\frac{\gamma\bar\mu\mu}{4},\frac{\gamma A_k}{2\bar a_k},0\right)$

表3. QASGD 的参数选择

请注意，y_k 与 x_k 相同，因为 $\rho_k=0$。因此，我们可以将李雅普诺夫函数（11）中的 y_k 替换为 x_k。我们首先使用李雅普诺夫函数（1，$\nabla_k=\nabla f_i(x_{k+1})$）以及表3 中指定的参数：11：

情形 1： $\mu = 0$

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

其中 $\mathcal{A}_k \triangleq \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k))$。第二个等式由镜像下降得到；第一个不等式由 h的强凸性得到；第二个不等式由 Fenchel-Young 不等式得到，最后一个不等式则由（15）（引理 A.3）得到。我们对初始界两边取期望，得到

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

将上述不等式两边相加，我们得到

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

情形 2： $\mu > 0$

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

¯第二个等式由镜像下降得出。取 β_k， α_k 并结合观察 A.6，我们有 $b_k = \alpha_k B_k/\beta_k$以及 $\alpha_k/\beta_k \le 1/2$。因此，我们得到如下界：

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

其中 $\mathcal{A}_k \triangleq \bar{a}_k(f(x_{k+1}) - f(x^*)) + A_k(f(x_{k+1}) - f(x_k))$。第二个不等式由（15）（引理 A.3）得出；第二个不等式由三角不等式得出，最后一个不等式则由 Fenchel-Young 不等式得出。对上述不等式两边取期望，得到

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

接下来，我们证明 $4\mu^2 B_k/\bar{\mu}^2\beta_k^2 \le \gamma\mu B_k/2\beta_k$。只需证明 $4\mu/\bar{\mu}^2\beta_k \le \gamma/2$。

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

因此，我们得到最终的界：

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

对上述不等式两边求和，并使用记号 $\text{II} \triangleq \min\{\gamma^2\bar{\mu}^2/16, 1/2\}$，可得

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

在步骤 1 中，运行算法直至收敛，并令 x_0 为步骤 1 的最后一次迭代。假设 $\mathbb{E}[f(x_0) - f(x^*) + \mu D_h(x^*, z_0)] \le \frac{\bar{\mu}\sigma^2}{4\mu}$，然后使用步骤 2 中的参数和记号 $m = \max\left\{\frac{48}{\gamma^2\bar{\mu}^2}, 5\right\}$ 重新启动算法。于是可得

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

此外，我们需要验证步骤 2 中的参数满足两个关键关系：$\alpha_k/\beta_k \le 1/2$ 和 $4\mu/\bar{\mu}^2\beta_k^2 \le \gamma/2$，这对于得到最终界 $E_{k+1} - E_k$ 至关重要。

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

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

C.3. 定理3.5的证明

QASVRG
γ-类凸$(\mu = 0)$
$A_k = \frac{\gamma^2\bar{\mu}}{16L}(k+1)^2,\ B_k = 1$ 批量大小 $b_k = \left\lceil \frac{8n}{\gamma\ n-1}(2k+3)\ ()p+\gamma\bar{\mu}(2k+3)\right\rceil, p \le \frac{\gamma\bar{\mu}}{16}$ $\tilde{}(\alpha_k, \beta_k, \rho_k, f, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{2\bar{a}_k}, \frac{1}{L}, f_{I_k}, 0, \frac{\gamma A_k}{2\bar{a}_k}, \frac{\gamma\epsilon f(y_0)}{2}\right)$
μ-强凸- γ-准凸$(\mu > 0)$
选项 I
$A_k = (1+\gamma/\sqrt{8\kappa})^k,\ B_k = \mu A_k$ 批量大小 $b_k = \left\lceil \frac{8n}{\gamma\ n-1}(\sqrt{8\kappa}+\gamma)\ ()+8(\sqrt{8\kappa}+\gamma)\right\rceil$ $(\alpha_k, \beta_k, \rho_k, \tilde{f}, b, c, \tilde{\epsilon}) \leftarrow \left(\frac{\gamma\mu}{2}, \frac{\gamma\mu B_k}{2\bar{b}_k}, \frac{1}{L}, f_{I_k}, \frac{\gamma\bar{\mu}\mu}{4}, \frac{\gamma A_k}{2\bar{a}_k}, 0\right)$
选项 II
$A_k = \frac{\gamma^2\bar{\mu}}{16L}(k+1)^2,\ B_k = 1$ 批量大小 $b_k = \left\lceil \frac{8n}{\gamma\ n-1}(2k+3)\ ()p+\gamma\bar{\mu}(2k+3)\right\rceil, p \le \frac{\gamma\bar{\mu}}{16}$ $\tilde{}(\alpha_k, \beta_k, \rho_k, f, b, c, \tilde{\epsilon}) \leftarrow \left(0, \frac{\gamma}{2\bar{a}_k}, \frac{1}{L}, f_{I_k}, 0, \frac{\gamma A_k}{2\bar{a}_k}, \frac{\gamma\epsilon f(y_0)}{2}\right)$

表 4. QASVRG 的参数选择

我们首先介绍算法 1，$\nabla_k = \nabla f_{I_k}(x_{k+1}) - \nabla f_{I_k}(y_0) + \nabla f(y_0)$ 并使用表 4 中指定的参数，以及李雅普诺夫函数（11）：

情形 1： $\mu = 0$

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

第一个等式由镜像下降得到；第三个等式由动量步得到；第一个不等式由 Fenchel-Young 不等式得到，最后一个不等式由（15）（引理 A.3）得到。然后，我们对随机变量 I_{i_j} 的历史取期望，其规模等于 b_i，其中 $i = 0$，1，……，$t-1$ 以及 $j = 1$，2，……，$\binom{n}{b_i}$，从而得到

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

第一个不等式由 $\mathbb{E}\left[\|\nabla_k\|^2\right] = \mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right]$ 得出，因为 ∇_k 对应于 k^{th} 迭代中的小批量；第二个不等式则由 $\mathbb{E}_{I_k}\left[\|\nabla_k\|^2\right] = \mathbb{E}_{I_k}\left[\|\nabla_k - \nabla f\ x_{k+1}\ \|^2\right] +$ f 的类凸性以及关系式 ( ) ( ) $\|\nabla f\ x_{k+1}\ \|^2 \quad L \quad 1,$ 得出。根据 -光滑性和算法中的梯度下降，我们得到以下关系：

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

利用上述关系，我们得到以下界：

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

我们使用记号 $\delta_k \triangleq \frac{n-b_k}{b_k(n-1)}$。第三个不等式由关系 $1/\bar{\mu}\beta_k^2 \leq A_{k+1}/L$ 得出。实际上，取 β_k 和 A_k，我们有

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

其中 $\eta = \gamma^2\bar{\mu}/16L$。最后一个不等式由命题 3.1（19）得出。接下来我们证明关系式 $4L\delta_k/\beta_k \leq p$。只需证明 $\delta_k \leq \gamma p/8L\bar{a}_k$。

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

因此我们有 $4A_{k+1}\delta_k \leq pA_{k+1}\beta_k/L$，由此得到 $\mathbb{E}[E_{k+1}-E_k$ 的最终界].

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

第三个不等式由两个关系式 $\bar{a}_k \geq pA_{k+1}\beta_k/L$ 和 $A_{k+1}\beta_k/L \leq 8\bar{a}_k/\gamma\bar{\mu}$ 得出，最后一个不等式由 $f(x^*) \geq 0$ 得出。取 A_k 和 β_k，我们有

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

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

将关于最终界的不等式两边求和，得到

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

当 $t \geq \left\lceil \sqrt{\frac{17LD_h(x^*,z_0)\gamma^2\bar{\mu}qf(y_0)}{}} \right\rceil$ 时，我们有

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

其中，$q + 8p/\gamma\bar{\mu} + \epsilon < 1$ 在选择 q 和 p 的情况下。接下来，我们给出全局阶段的收敛率。假设 y_s 是阶段 s 的输出。如果在每个阶段都有 $t \geq \left\lceil \sqrt{\frac{17LD_h}{(x^*,\ y_s)\gamma^2\bar{\mu}qf(y_s)}} \right\rceil$，则

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

情形 2： $\mu > 0$（选项 I）使用记号 $d_{k+1} = A_{k+1}(f(y_{k+1}) - f(x^*)) - A_k(f(y_k) - f(x^*))$，我们得到 n

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

注意到 $\xi_k = \frac{B_k}{\beta_k}\langle \nabla f(y_0) - \nabla f_{I_k}(y_0)\ x_{k+1} - z_k\rangle$，并且 $\mathbb{E}[\xi_k] = 0$。第一个不等式由函数 μ 的 h-强凸性推出；第二个不等式由式（15）（引理 A.3）推出，最后一个不等式则由关系 $\alpha_k/\beta_k \leq 1/2$ 和三角不等式推出。然后，我们对随机变量 I_{i_j} 的历史取期望，其大小为 b_i，其中 $i = 0$、1、.....、$t-1$ 以及 $j = 1$、2、.....、$\binom{n}{b_i}$。

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

第一个不等式由函数 f 的一致拟凸性推出；第二个不等式由式（21）推出；第三个不等式由关系 $B_k/\bar{\mu}\beta_k^2 \leq A_{k+1}/2L$ 推出，最后一个不等式由命题 3.1（19）推出。我们验证上述关系的正确性。取 A_k、B_k、α_k、β_k，并结合观察 A.6，可得

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

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

接下来我们证明 $4A_{k+1}\delta_k \leq \bar{a}_k/2$。只需证明 $\delta_k \leq \bar{a}_k/8A_{k+1}$。

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

因此，我们得到了 $\mathbb{E}[E_{k+1}-E_k]$ 的最终界。

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

将上述不等式两边求和，得到

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

当 $t \ge \log_{1+\gamma/\sqrt{8\kappa}}(2/\gamma q)$时，我们有

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

其中 $q+1/2<1$ 在选择 q时。接下来我们得出全局阶段的收敛速率。假设 y_s 是阶段 s的输出。如果每个阶段都满足 $t \ge \log_{1+\gamma/\sqrt{8\kappa}}(2/\gamma q)$，则有

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

对于 $\mu>0$ （选项 II），我们的分析与情形 1 完全相同。

D. 复杂度证明

在本节中，我们分析 QASGD 和 QASVRG 的总体复杂度。由于 Hinder 等人（2020）已经分析了 QAGD 的复杂度，因此我们不再提供相关证明。类似地，我们采用 Hinder 等人（2020）提出的同一指标，即函数评估和梯度评估的总次数。粗略地说，总体复杂度是迭代次数与每次迭代中函数评估和梯度评估次数的乘积。我们考虑 Bisearch 的最坏情况，即 0、1 和“猜测值”均不满足我们的条件，因此每次迭代都需要进行线搜索。在这种情况下，我们在 Bisearch 中访问的梯度（估计值）可以直接用于后续的更新，也就是说，镜像下降步骤和梯度下降步骤都不需要额外访问 ∇_k。Bisearch 中涉及的 f_i 数量会影响复杂度。QAGD 在 Bisearch 中需要所有 f_i，而 QASGD 和 QASVRG 则分别只需要一个 f_i 和一个 f_i 小批量。接下来，我们分析欧几里得空间中 QASGD 和 QASVRG 每次迭代的 Bisearch 复杂度，其中 $D_h(x,y) = \frac{1}{2}\|x-y\|^2$，并给出定理 3.5 和定理 3.6 的证明。引理 A.5 表明，对于单个函数，Bisearch 每次迭代需要 $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k-z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right)$ 次函数评估和梯度评估。由于 $\tilde{\epsilon}$ 和 b 不可能同时非零，因此我们有两种对应不同复杂度的情况。

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

当 $\mu=0$ 时，我们证明的关键是对 $\|y_k-z_k\|$ 给出界。

D.1. 推论的证明 3.3

情形 1： $\mu=0$

根据 3.2 的证明，我们 有 $\mathbb{E}[E_{k+1}-E_k] \le \frac{\sigma^2}{2\beta_k^2} + \frac{\bar{a}_k\epsilon}{2}$。假设 $\|x^*-z_0\| \le R$，我们有以下关系。

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

第三个不等式由 $\eta = \min\left(\frac{1}{L},\ \sqrt{\frac{2\|x^*-z_0\|^2}{\sigma^2}}\frac{\gamma}{(t+1)^{3/2}}\right)$ 得出。结合上述分析，我们有 $\mathbb{E}\|x^*-z_k\|^2 \le 10R^2 + \left(\frac{k+1}{L}\right)^2\epsilon$，并根据詹森不等式得到 $\mathbb{E}\|x^*-z_k\| \le \sqrt{\mathbb{E}\|x^*-z_k\|^2} \le \sqrt{10R^2 + \left(\frac{k+1}{L}\right)^2\epsilon}$。因此，我们得到

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

由于梯度下降的步长为 0， y_k 与 x_k相同，我们只需对 $\|x_k - z_k\|$给出界。根据z_k 和 x_k的定义，我们有

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

最后一个不等式 f 由关系 $\sqrt{a^2+b^2} \le a + b$ 可得 $a,\ b \ge 0$。根据定理3.2的证明，我们有

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

假设 $k \le k_{\max} = \left\lfloor \frac{4\left(LR^2 + \frac{\sqrt{2}\sigma R}{\gamma}\right)^2}{\epsilon^2} \right\rfloor$，我们得到

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

第一个不等式由关系 $k+3 \leq 4k$（当 $k \geq 1$ 时）得出。根据马尔可夫不等式，$P_r(\|x_k - z_k\| \geq k_{\max}\mathbb{E}\|x_k - z_k\|) \leq \frac{1}{k_{\max}}$，这意味着 $\|x_k - z_k\| \leq k_{\max}\mathbb{E}\|x_k - z_k\| \leq O\left(\frac{L^{11/2}R^{12}}{\gamma^6\epsilon^{11/2}}\right)$ 以至少 $1 - \frac{1}{k_{\max}}$ 的概率成立。因此我们有 $\frac{L\|x_k - z_k\|^2}{\tilde{\epsilon}} \leq O\left(\frac{L^{12}R^{24}}{\gamma^{13}\epsilon^{12}}\right)$。此外，我们有 $c+1 = \frac{\gamma(k+1)^2}{2k+3} + 1 \leq \gamma k + 2 \leq \frac{4)(LR^2+\sqrt{2}\sigma R)}{\gamma\epsilon^2}^2 + 2 = O\left(\frac{L^2R^4}{\gamma\epsilon^2}\right)$。于是得到 $\log^+()$ 中项的如下上界：

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

因此，我们有 $O\left(\log^+\left((1+c)\min\left\{\frac{L\mathbb{E}\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) \leq O(\log^+(LR^2\gamma^{-1}\epsilon^{-1}))$。在 $\mu = 0$ 的情况下，QASGD 需要 $O\left(\sqrt{\frac{LR^2}{\epsilon}} + \frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)$ 次迭代，才能在期望意义下生成一个 ϵ-近似解。因此，QASGD 的总体复杂度以高概率被 ($\mu = 0$) 上界所界定，为 $O\left(\left(\sqrt{\frac{LR^2}{\epsilon}} + \frac{\sigma^2R^2}{\gamma^2\epsilon^2}\right)\log^+\left(\frac{LR^2}{\gamma\epsilon}\right)\right)$。

情形 2： $\mu > 0$

当 $\tilde{\epsilon} = 0$ 和 $b > 0$ 时，我们有 $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left((1+c)\frac{L^3}{b^3}\right)\right)$。在步骤 1 中，$b = \frac{\gamma\mu}{4}$ 和 $c = \frac{8}{\gamma}$。于是我们有

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

因此，我们有 $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right)\right)$。根据定理 3.2 的证明，QASGD 在步骤 1 中需要 $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f\ x^*}{\gamma\epsilon}\right)\right)()$ 次迭代，步骤 1 的复杂度为 $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f\ x^*}{\gamma\epsilon}\right)()\right.$ $\left.\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right)\right)$。在步骤 2 中，$b = \frac{\gamma\mu}{4}$ 和 $c = \frac{\gamma}{2}(k+48/\gamma^2\ ^2)(2k+96/\gamma^2+1) \leq \frac{\gamma}{2}(k+48/\gamma^2)^2 \leq \frac{k+48}{2\gamma}$。我们可以对步骤 2 的收敛速率给出上界：

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

假设 $k \leq \left\lfloor\frac{90\sigma^2}{\gamma^2\mu\epsilon}\right\rfloor$。我们有 $c+1 \leq \frac{C_3}{\gamma^3\epsilon}$，其中 $C_3 = \frac{45\sigma^2}{\mu} + 25$。

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

因此，$O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O\left(\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$。QASGD 在步骤 2 中需要 $O\left(\frac{\sigma^2}{\gamma^2\epsilon}\right)$ 次迭代，步骤 2 的复杂度为 $O\left(\frac{\sigma^2}{\gamma^2\epsilon}\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\right)$。总之，QASGD 的总体复杂度（$\mu > 0$）上界为 $O\left(\frac{1}{\gamma^2}\log\left(\frac{f(x_0)-f\ x^*}{\gamma\epsilon}\right)()\right)\log^+\left(\frac{\kappa^{3/4}}{\gamma}\right) + \frac{\sigma^2}{\gamma^2\epsilon}\log^+\left(\frac{\kappa^{2/3}}{\gamma\epsilon^{1/6}}\right)\Bigg)$。

D.2. 推论的证明3.6

情形 1： $\mu = 0$

根据定理 3.3 的证明，我们有 $\mathbb{E}[E_{k+1} - E_k] \leq \bar{a}_k\left(\frac{1}{2} + \epsilon\right)f(y_0)$。假设 $\|x^* - z_0\| \leq R$，则有以下关系。

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

因此我们有 ave $\mathbb{E}\|x^* - z_k\|^2 \le 2R^2 + \frac{3\gamma^2}{16L}(k+1)^2 f(y_0) + \frac{\gamma^2}{8L}(k+1)^2 f(y_0)\epsilon$ 并根据詹森不等式

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

根据 y_k 和 z_k的定义，我们有

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

假设 $f(y_0) \ge \epsilon$ 且 $k \le \left\lfloor \sqrt{\frac{17LR^2}{2\gamma^2 q f(y_0)}} \right\rfloor \le k_{\max} = \left\lfloor \sqrt{\frac{17LR^2}{2\gamma^2 q\epsilon}} \right\rfloor$，则可得

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

根据马尔可夫不等式，$P_r(\|y_k - z_k\| \ge k_{\max}\mathbb{E}\|y_k - z_k\|) \le \frac{1}{k_{\max}}$，这意味着以至少 $1 - \frac{1}{k_{\max}}$ 的概率有 $\|y_k - z_k\| \le k_{\max}\mathbb{E}\|x_k - z_k\| \le O\left(\frac{LR^3}{q^{3/2}\gamma^3\epsilon}\right)$。因此我们有 $\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}} \le O\left(\frac{L^3R^6}{q^3\gamma^7\epsilon^4}\right)$。此外，我们有 $c+1 = \frac{\gamma}{2}\frac{k+1}{2k+3}^2() + 1 \le \frac{\gamma k}{2} + \frac{3}{2} \le \frac{R\sqrt{17L}}{2\sqrt{2q\epsilon}} + \frac{3}{2} = O\left(\frac{L^{1/2}R}{q^{1/2}\epsilon^{1/2}}\right)$。于是可得对 $\log^+()$ 中项的如下上界：

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

因此，以高概率有 $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) \le O\left(\log^+\left(\frac{L^{1/2}R}{q^{1/2}\gamma\epsilon^{9/14}}\right)\right)$。由于我们需要在每个阶段访问在 y_0 处计算的全梯度和函数值，以及小批量的梯度和函数值来

计算随机方差缩减G 和 $\tilde{\epsilon}$，生成 ϵ-近似解的 QASVRG ($\mu = 0$) 的总体复杂度为

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

其中 $t = \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q f\ y_0}}() \right\rceil \le \sqrt{\frac{17LR^2}{2\gamma^2 q f\ y_0}}()$ 是每个阶段的最大迭代次数，而 $O(\log(\epsilon^{-1}))$ 是阶段数。注意，根据假设 $D_h(x^*, y_s)$，R^2 在假设 2.5下有一致上界，该上界已包含在上面的界中。

情形 2： $\mu > 0$

对于选项 II，根据假设 $t = \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q f\ y_0}}() \right\rceil \le \left\lceil \sqrt{\frac{17L\|x^* - z_0\|^2}{2\gamma^2 q \mathcal{E}_0}} \right\rceil \le \sqrt{\frac{17L(2-\gamma)}{\gamma^3 q\mu}} \le \sqrt{\frac{34\kappa}{\gamma^3 q}}$ 3.1 中的最后一个关系，我们有 3.1。因此，QASVRG（选项 II）生成一个 ϵ-近似解的总体复杂度为

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

其中，$O(\log(\epsilon^{-1}))$ 表示阶段数。对于选项 I，$\tilde{\epsilon} = 0$，$b = \frac{\gamma\mu}{4}$ 和 $c = \sqrt{2\kappa}$。由此得到以下关系：

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

因此，我们有 $O\left(\log^+\left((1+c)\min\left\{\frac{L\|y_k - z_k\|^2}{\tilde{\epsilon}}, \frac{L^3}{b^3}\right\}\right)\right) = O(\log^+(\kappa^{7/6}\gamma^{-1}))$。在每个阶段，QASVRG（选项 II）都会运行，直到满足 $(1 + \gamma/\sqrt{8\kappa})^{-t} \le \frac{\gamma q}{2}$。令 $(1 + \gamma/\sqrt{8\kappa})^{-t} \le e^{-\frac{\gamma t}{\sqrt{8\kappa}+\gamma}} \le \frac{\gamma q}{2}$，则有 $t \ge \frac{\sqrt{8\kappa}+\gamma}{\gamma}\log\left(\frac{2}{\gamma q}\right)$，以及$t = \left\lceil \log_{1+\frac{\gamma}{\sqrt{8\kappa}}}\left(\frac{2}{\gamma q}\right)\right\rceil \le \frac{\sqrt{8\kappa}+\gamma}{\gamma}\log\left(\frac{2}{\gamma q}\right) \le \frac{5\sqrt{\kappa}}{\gamma}\log\left(\frac{2}{\gamma q}\right)$。因此，QASVRG（选项I）生成一个 ϵ-近似解的总体复杂度为

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

其中 $O(\log(\epsilon^{-1}))$ 表示阶段数。

E. 理论扩展

我们在一个严格条件下分析 QASGD：强增长条件（SGC），其正式定义如下。该条件由 Schmidt & Roux (2013年)、Vaswani 等人 (2019 年) 以及 Gower 等人 (2021 年) 提出。Schmidt & Roux (2013 年) 推导出了 SGC 条件下，针对凸函数和强凸函数的随机梯度下降最优收敛率。

假设 E.1 (SGC)。 假设 i 从 $[n]$ 中独立同分布地抽样。对于某个常数 ρ 和 $x^* \in \mathcal{X}^*$，有

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

如果 $\nabla f(x) = 0$，那么在 SGC 下，$\nabla f_i(x) = 0$，这意味着插值假设。我们推导出了 $f \in \mathcal{Q}_{\mu\gamma}$ 时 QASGD 在 SGC 下更好的收敛率。

定理 E.2 (SGC 下的 QASGD)。 假设 *3.1* 和假设 *E.1* 成立，$D_h(x^*, z_0) \leq R^2$，$f \in \mathcal{F}_L$，并任取 $\tilde{y}_0 \in \mathbb{R}^d$。则算法 *1* 采用 $\nabla_k = \nabla f_i(x_{k+1})$ 和 A_k，B_k，θ_k （见表 *5* ）中指定的取值，满足

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

对(22)两边求和，可得如下收敛速率：

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

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

表 *5.* SGC 下 QASGD 的参数选择

证明。

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

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

情形 2： $\mu>0$

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

第一个等式和第三个等式由引理 A.1 得出；第二个等式由镜像下降得出。第一个不等式由 h 的强凸性得出，第二个不等式由（15）（引理 A.3）得出。取$^-$，并令 α_k 和 β_k，则有 $\alpha_kB_k/\beta_k=b_k$，这解释了最后一个不等式。此外，取 B_k，由观察 A.6，可得

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

结合初始界以及上述关系，我们得到以下界：

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

取期望后，我们得到

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

□

F. 其他模拟结果

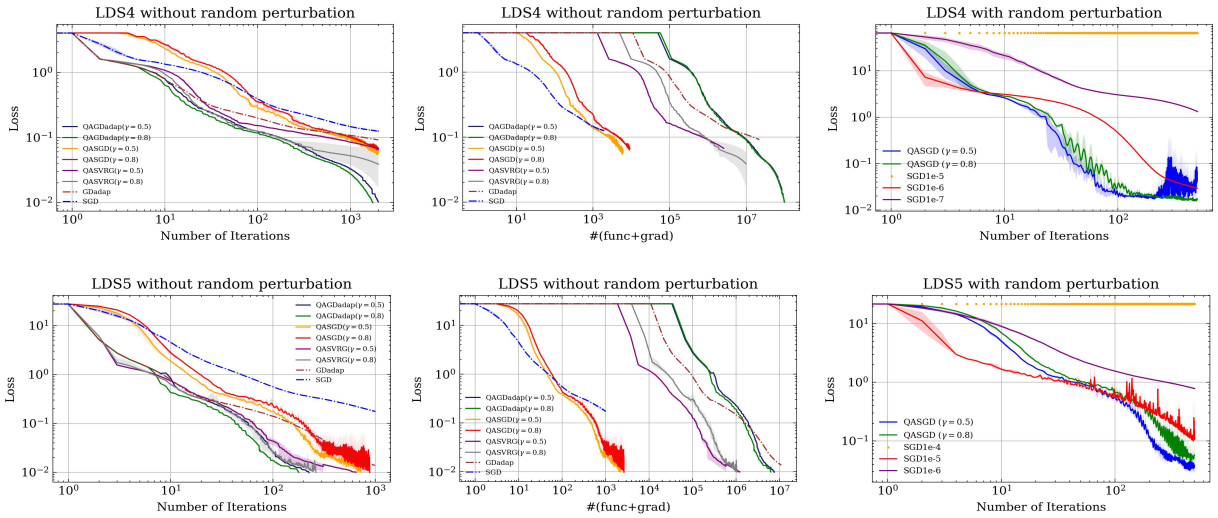

图 2 $\{12, 36\}$ $\epsilon = 10^{-2}$ 1×10^{-6} $1\times$.使用随机种子，对两个不同的 LDS 10^{-5} $L = 1\times10^7$ 5×10^6 $L = 3\times10^6$ 1×10^5 4 5 实例进行评估。我们选择，SGD 的步长为，QASGD 的步长为，QASVRG 在 LDS 和 LDS 中的步长为。第三列中的平线表示在此步长选择下，损失会发散至无穷大。

我们通过构造一个满足所有必要假设的目标函数，给出一个人为设计的实验。考虑以下优化问题

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

其中，$(a_i, b_i)_{i=1,\ldots,n}$是满足 $a_i \in \mathbb{R}^d$和 $b_i \in \{+1, -1\}$的训练数据； $\mu \ge 0$，且 $f(x)$满足假设 2.4。根据 Hinder 等人（2020）D.3 中介绍的拟凸函数的性质， $f(x)$是μ-强 γ-拟凸且 L-光滑的，其中 $L = \sum_{i=1}^n \|a_i\|/n + 0.5\mu$。为简化起见，我们选择 $\gamma \in \{0.5, 0.8\}$ 并对每个 a_i 进行归一化，使得 $L = 1 + 0.5\mu$。注意，每个$g_\gamma(b_i a_i^\mathsf{T} x)$至少有一个公共极小点。因此， f也满足假设 2.7。我们使用 Dua 和 Graff（2017）提供的以下多分类数据集，并将其视为二分类数据集。根据该数据集，我们有 $n = 1372$ 和 $d = 4$ 。当 $\mu = 0$时，我们设置 $\epsilon = 10^{-2}$ ；当 $\mu > 0$时，我们设置 $\epsilon = 10^{-3}$ 。我们按照第 4 节中模拟所采用的相同方式生成误差条。图 4 显示，QASGD 的收敛速度快于随机梯度下降，而 QASVRG 收敛迅速，且复杂度低于 QAGD 和梯度下降。当 $\mu = 0.002$时，图 4 还显示，在给定 $\kappa\epsilon \approx 0.5 < 1$的情况下，QASVRG（选项 I）在收敛速度和复杂度方面具有优势。我们还比较了我们的方法与梯度下降、QAGD 和随机梯度下降在求解具有逻辑链接函数的广义线性模型经验风险时的表现

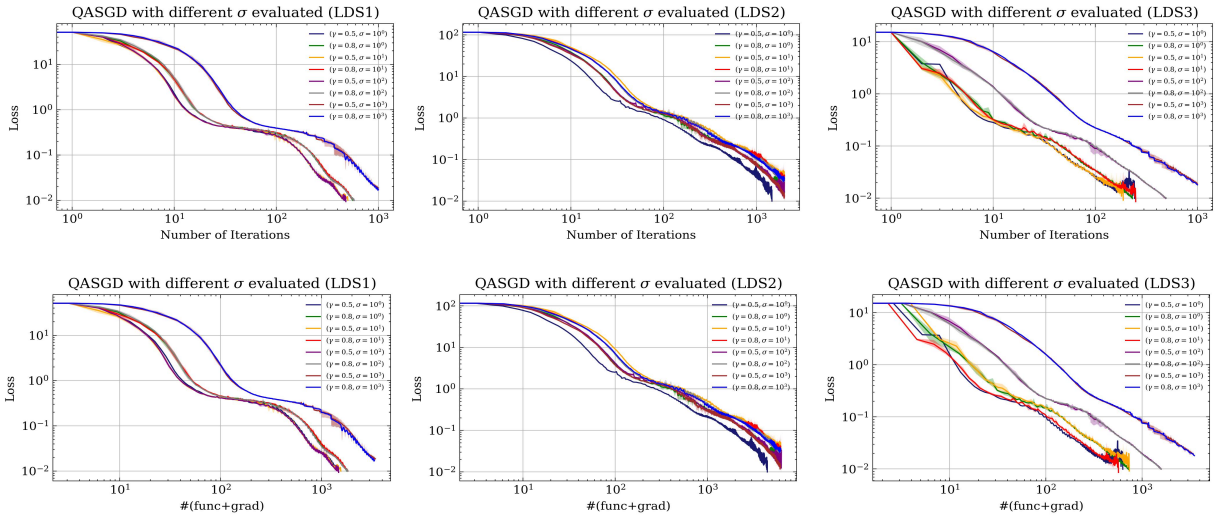

图 *3.* 在三个 LDS 实例上评估采用不同 σ 的 QASGD。

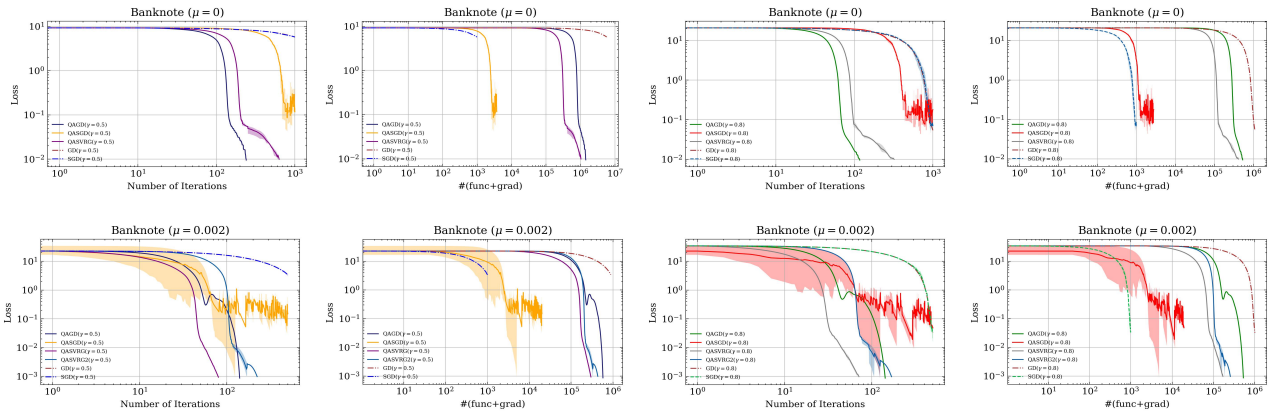

图 *4.* 各算法在问题 (24) 上的评估

$\sigma(z) = (1+\exp(-z))^{-1}$。考虑以下优化问题

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

其中 $x_i \sim \mathcal{N}(0, I)$、$w^* \sim \mathcal{N}(0, I)$，且对每个 $i \in [n]$ 均有 $y_i = \sigma\left(w_*^{\mathsf{T}}x_i\right)$。在我们的实验中，我们选择 $n = 5000$、$d = 50$ 和初始迭代点 $w_0 \sim \mathcal{N}(0, 100I)$。由于无法计算类凸性的参数 γ 和光滑性 L，我们通过广泛的网格搜索，使用 $\gamma = 0.5$ 和 $L = 10^5$ 对我们的方法进行评估。

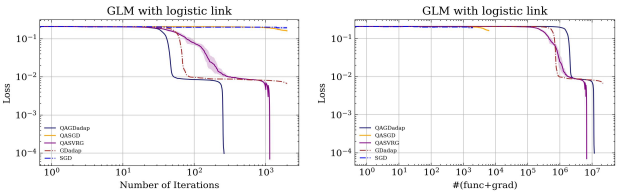

图 *5*。 各算法在问题（25）上的评估