具有陈旧随机梯度的凸与非凸联邦学习：递减步长便足矣

Xinran Zheng XINRAN22@ILLINOIS.EDU
伊利诺伊大学厄巴纳-香槟分校
Tara Javidi TJAVIDI@UCSD.EDU
加利福尼亚大学圣迭戈分校
Behrouz Touri TOURI1@ILLINOIS.EDU
伊利诺伊大学厄巴纳-香槟分校

编辑： Steve Hanneke 和 Tor Lattimore

arXiv:2603.02639v1 [math.OC] 3 Mar 2026

摘要

我们提出了一个适用于延迟梯度模型的分布式随机优化一般框架。在此设定下，n 本地代理利用自身的数据和计算能力，协助中央服务器最小化由各智能体的局部成本函数构成的全局目标函数。每个智能体都可以传输其局部梯度的随机估计值，这些估计值可能存在偏差和延迟。尽管先前的工作曾提倡在存在延迟的情况下，为随机梯度下降（SGD）采用延迟自适应步长，但我们证明，预先选定的递减步长已足够，且其性能与自适应方案相当。此外，我们的分析表明，递减步长能够在非凸和强凸目标函数上达到随机梯度下降的最优收敛速率。

关键词： 联邦学习、延迟、分布式优化、投影随机梯度下降

1. 引言

分布式优化和联邦学习已成为现代学习与决策系统中的核心基础，在这些系统中，数据和计算天然分散于众多设备、组织或地理位置。在典型的服务器—客户端架构中，中央协调器旨在最小化由各局部成本函数汇总而成的全局目标函数，而n 智能体则利用自身数据计算更新量并为其做出贡献。这种范式能够实现可扩展性和隐私保护训练，但也带来了新的算法和理论挑战，因为通信Alistarh 等人 (2022年)、计算 Recht 等人 (2011)以及数据访问 McMahan 等人 (2017)；Fang 等人 (2022年)都并不完善。

本文聚焦于基于梯度下降的方法，因为这类方法简单且适用范围广，相关研究已十分丰富 Nedi´c 和 Ozdaglar (2009);Tsianos 等人(2012年);Nedi´c 和 Olshevsky(2014); Tatarenko 和 Touri(2017）。然而，在实际的分布式部署中，有两种现象不容忽视。

首先，智能体传输的梯度信息通常具有随机性——这是由数据采样、通信压缩/量化以及本地不精确计算所致——而且大多数现有分析都采用一个简化假设，即这些随机梯度是无偏的Alis- tarh 等人(2017）；Sra 等人(2016）；Ghadimi 等人(2016）。然而，有偏梯度替代量在实践中自然存在；一个典型例子是零阶优化，在这种方法中，智能体可以

© 2026年 X. Zheng、T. Javidi 和 B. Touri。

表1：与相关工作的假设比较。

Work	x的定义域	随机梯度	延迟
Ours	受约束的	有偏	缩放
Ghadimi 等人（2016）	受约束	无偏	0
Sra 等人（2016）	受约束的	无偏的	缩放后的
Zheng 等人 (2023年)	无约束的	有偏的	有界
Zheng 等人 (2024年)	无约束	有偏	有界

表 2：与相关工作的性能比较。

Work	非凸 $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2]$	强凸 $\mathbb{E}[\|\boldsymbol{x}(T)-\boldsymbol{x}^*\|^2]$	凸 $\mathbb{E}[f(\tilde{\boldsymbol{x}}(T))]-f^*$
Ours	$\mathcal{O}(1)$ 对于投影梯度	$\mathcal{O}(\frac{1}{T})$	$\mathcal{O}(\frac{\log T}{\sqrt{T}})$
Ghadimi 等人 (2016)	$\mathcal{O}(1)$ 对于投影梯度		$\mathcal{O}(\frac{\log T}{\sqrt{T}})$
Sra 等人 (2016年)			$\mathcal{O}(\frac{1}{\sqrt{T}})$
Zheng 等人 (2023年)	$\mathcal{O}(\frac{\log T}{\sqrt{T}})$		
Zheng 等人 (2024年)		$\mathcal{O}(\frac{1}{T})$	

只能查询函数值，并且必须通过随机扰动构造梯度估计，而这些估计通常有偏Tang 等人(2023）；Zheng 等人(2023）；Fang 等人(2022）；Malladi 等人(2023）。

其次，由于掉队节点、间歇性连接或异步通信，智能体更新经常会延迟 Reisizadeh 等人 (2022a);Yang 等人 (2022年); 因此，服务器可能会应用在较早迭代点上计算出的陈旧梯度信息。在以往的工作中，这类延迟通常被假设为一个随机变量，其上界为某个常数 Tang 等人(2023年);Zheng 等人 (2023年,2024年)，或者在时间上服从独立同分布的指数分布Reisizadeh 等人(2022a)。在Sra 等人 (2016年)中，一种采用延迟自适应步长的算法处理了较弱的缩放延迟假设。当存在约束时，SGD 的投影变体会进一步将这些影响与可行性考量耦合起来，因此，理解偏差和延迟如何共同影响收敛保证至关重要。

本文研究了温和延迟条件下，具有延迟近似梯度模型的约束随机梯度下降。具体而言，智能体可以传输其局部梯度的随机估计（可能存在偏差），而服务器则使用缩放延迟的梯度信息进行更新，也就是说，在时刻 t 使用的梯度不早于时刻 κt 计算，其中$\kappa \in (0,1)$，且延迟的二阶矩也有界。据我们所知，这是首项分析缩放延迟模型下带偏随机梯度估计器的约束随机梯度下降的工作。我们的结果传达了一个重要结论：无需采用延迟自适应方案；预先选定的标准递减步长已足以在与此前延迟自适应分析相同的延迟假设下实现最优性能（至多相差对数因子）Sra 等人 (2016)。

贡献 我们在温和延迟条件下，研究一种采用延迟的近似随机梯度进行分布式优化的一般框架。具体而言，我们假设延迟经过缩放（即，时刻 t 使用的梯度信息不早于时刻 κt 计算），且具有有界二阶矩。我们证明，对于**非凸**函数，其收敛性能

对于$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2]$，其中 $\boldsymbol{h}(t)$ 表示投影梯度，其收敛性能与无延迟的经典无偏随机梯度下降相当。对于**强凸** 函数，均方误差达到$\mathcal{O}(\frac{1}{T})$ 阶，同样与无延迟经典随机梯度下降的已知最优收敛速率相当。此外，对于一般**凸** 函数，我们在任意小的 $\varepsilon > 0$ 下得到误差界 $\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) = \mathcal{O}(\frac{1}{T^{\frac{1}{2}-\varepsilon}})$，与采用延迟自适应步长所获的误差界 $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ 相比，仅相差一个对数因子。

记号 我们用 $\mathbb{Z}^+$表示非负整数集 $\{0,1,\ldots\}$，用 $\mathbb{R}$表示实数，用 $\mathbb{R}^d$表示 d维实值向量的向量空间。我们用粗体小写字母表示向量。我们用 $\boldsymbol{I}_d$ 表示维数为 d的单位矩阵。我们用 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$表示零均值且协方差为 $\boldsymbol{\Sigma}$的多维高斯分布。我们用 $[n]$表示 $\{1,\ldots,n\}$。我们用 $\|\cdot\|$ 表示标准欧几里得范数，用 $\langle\cdot,\cdot\rangle$ 表示由 $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{y}$给出的标准内积。我们令 $\mathbb{S}_{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d \colon \|\boldsymbol{x}\| = 1\}$ 表示单位球面，$\mathrm{Unif}(\mathbb{S}_{d-1})$表示 $\mathbb{S}_{d-1}$上的均匀分布。对于任意随机变量 X，我们用 $\sigma(X)$表示由 X生成的 σ-代数。

2. 问题表述

我们考虑由中央服务器和 n智能体或工作节点共同执行的可分解代价函数优化，他们协同优化一个全局函数，该函数是 n个局部函数之和：

$$\min_{\boldsymbol{x}\in\mathcal{S}} f(\boldsymbol{x}) := \sum_{i=1}^{n} f_i(\boldsymbol{x}), \tag{1}$$

其中，$\mathcal{S} \subset \mathbb{R}^d$ 是一个闭凸集。我们假设每个局部函数 f_i（及其任何梯度信息）只能由智能体 i访问。中央服务器无法直接访问这些f_i；相反，它维护全局决策变量 $\boldsymbol{x}$，并在每个时刻 t向所有智能体广播 $\boldsymbol{x}(t)$，同时请求在 $\boldsymbol{x}(t)$处计算的梯度信息。智能体 i返回的梯度记为 $\boldsymbol{g}_i(\boldsymbol{x},\ \boldsymbol{\xi})$，它取决于 $\boldsymbol{x}$ 以及一个与 $\boldsymbol{x}$独立的随机变量 $\boldsymbol{\xi}$。此外，由于计算与通信延迟，服务器可能会收到并应用在较早迭代点计算出的陈旧估计。

3. 主要结果

本节将介绍一类用于求解问题 (1) 的算法的一般框架。这些算法在离散时间时刻 $t \in \mathbb{Z}^+$。

在时刻 $t = 0$，中央服务器在任意点 $\boldsymbol{x}(0) = \boldsymbol{x}_0 \in \mathcal{S}$处生成一个初始猜测。然后，对于 $t \in \mathbb{Z}^+$，中央服务器向每个智能体发送 $\boldsymbol{x}(t)$ 和时间戳 t。与此同时，每个智能体根据其收到的部分 $\boldsymbol{x}$ 计算或估计 f_i 的梯度，并将梯度信息连同相应的原始时间戳发送回中央服务器。与此同时，中央服务器接收来自每个智能体的梯度信息，并跟踪来自智能体 i 的最新梯度信息 $\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)),\ \boldsymbol{\xi}(\tau_i(t)))$。这里，$\boldsymbol{x}(\tau_i(t))$ 是中央服务器在时刻 $\tau_i(t)$发出的信息，表示截至时刻 t 从智能体 i 收到的最新梯度信息。如果中央服务器从未收到过来自智能体 i 的任何梯度信息，我们

令 $\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)),\boldsymbol{\xi}(\tau_i(t)))=0$，且 $\tau_i(t)=-1$。然后，中央服务器计算全局代价函数的一个近似梯度，如下所示

$$\boldsymbol{g}(t)=\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)),\boldsymbol{\xi}(\tau_i(t))), \tag{2}$$

并将优化变量更新为

$$\boldsymbol{x}(t+1)=\Pi_{\mathcal{S}}[\boldsymbol{x}(t)-\eta(t)\boldsymbol{g}(t)], \tag{3}$$

其中 $\Pi_{\mathcal{S}}:\ \mathbb{R}^d\to\mathcal{S}$ 是集合 $\mathcal{S}$上的投影算子，$\eta(t)$是适当的步长。

3.1. 分析

在此，我们给出一般框架的收敛保证方面的主要结果。为此，我们对目标函数、梯度估计器和通信延迟作出以下假设。

第一个假设针对目标函数。

假设1（目标函数假设）我们假设 $f^*:=\inf_{\boldsymbol{x}\in\mathcal{S}}f(\boldsymbol{x})>-\infty$，且$f$ 是连续可微的。此外，对于 $i\in[n]$，$f_i:\mathcal{S}\to\mathbb{R}$ 是 L-光滑的，即对所有 $\boldsymbol{x}$、$\boldsymbol{y}\in\mathcal{S}$，

$$\|\nabla f_i(\boldsymbol{x})-\nabla f_i(\boldsymbol{y})\|\le L\|\boldsymbol{x}-\boldsymbol{y}\|.$$

第二个假设关于梯度估计器。为讨论这一点，对于由式（3）生成的估计值（随机过程）$\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}^+}$，我们定义随机过程 $\{\tilde{\boldsymbol{g}}_i(t)\}_{t\in\mathbb{Z}^+}$ 为$\tilde{\boldsymbol{g}}_i(t):=\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{x}(t),\boldsymbol{\xi}(t))|\sigma(\boldsymbol{x}(t))]$，对每个 $i\in[n]$ 均成立。下面的假设确立了 $\boldsymbol{g}_i(\cdot)$ 与 $\nabla f_i(\cdot)$ 之间的关系。

假设2（梯度估计器假设）我们假设，对于由式（3）生成的随机过程$\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}^+}$，存在 $G\in(0,\infty)$，使得 $\mathbb{E}[\|\boldsymbol{g}_i(\boldsymbol{x}(t),\boldsymbol{\xi}(t))\|^2|\sigma(\boldsymbol{x}(t))]\le G$ 对所有 $i\in[n]$ 均成立。此外，对于期望梯度估计 $\{\tilde{\boldsymbol{g}}_i(t)\}_{t\in\mathbb{Z}^+}$，我们假设对任意 t_1、$t_2\in\mathbb{Z}^+$，都有

$$\|\tilde{\boldsymbol{g}}_i(t_1)-\tilde{\boldsymbol{g}}_i(t_2)\|\le L\|\boldsymbol{x}(t_1)-\boldsymbol{x}(t_2)\|.$$

最后，我们假设对于某个非负序列 $\{q(t)\}_{t\in\mathbb{Z}^+}$，$\|\tilde{\boldsymbol{g}}_i(t)-\nabla f_i(\boldsymbol{x}(t))\|\le q(t)$ 成立。我们将 $q(t)$ 称为梯度估计器偏差。

备注 1 正如一项密切相关的工作中所讨论的 *Zheng* 等人 （*2024*年），假设 *2* 被一大类梯度估计器所满足。其中包括无偏随机梯度（含 $q(t)=0$），以及许多使用随机扰动的零阶梯度估计器。例如，在 *Tang* 等人 （*2023*年）；*Zheng* 等人 （*2023*年）中，使用了带有高斯扰动的两点梯度估计器，即 f_i 在 $\boldsymbol{x}(t)$ 处的梯度估计为

$$\boldsymbol{g}_i(\boldsymbol{x}(t),\boldsymbol{z}(t))=\frac{f_i(\boldsymbol{x}(t)+u(t)\boldsymbol{z}(t))-f_i(\boldsymbol{x}(t)-u(t)\boldsymbol{z}(t))}{2u(t)}\boldsymbol{z}(t), \tag{4}$$

其中 $u(t)$ 是平滑半径，而 $\boldsymbol{z}(t) \sim \underline{\mathcal{N}}(\mathbf{0}, \boldsymbol{I}_d)$。已有研究表明，该梯度估计器的偏差有界于 $L\sqrt{d}u(t)$，因此可以通过选择 $u(t)$ 来控制收敛速率。另一个例子是基于在单位球上进行平滑的两点梯度估计器，见 *Shamir* (*2017*)

$$\boldsymbol{g}_i(\boldsymbol{x}(t), \boldsymbol{z}(t)) = d\frac{f_i(\boldsymbol{x}(t) + u(t)\boldsymbol{z}(t)) - f_i(\boldsymbol{x}(t) - u(t)\boldsymbol{z}(t))}{2u(t)}\boldsymbol{z}(t),$$

其中 $\boldsymbol{z}(t) \sim$ 均匀分布$(\mathbb{S}_{d-1})$，其偏差有界于 $Lu(t)$。

在接下来的分析中，为简洁起见，当不需要考虑随机变量 $\boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{\xi})$ 的依赖关系时，我们将其记为 $\boldsymbol{g}_i(\boldsymbol{x})$。根据假设 2，我们可以直接推导出全局梯度估计量的二阶矩有界。

引理 2 对于 $\boldsymbol{g}(t)$ 中给出的 (2)，有 $\mathbb{E}[\|\boldsymbol{g}(t)\|^2] \leq n^2G$ 对所有 $t \in \mathbb{Z}^+$均成立。

最后一个假设涉及通信延迟。

假设 3（关于通信延迟的假设）我们假设延迟 $t - \tau_i(t)$ 在不同时间和智能体之间相互独立。我们还假设延迟的二阶矩有界，即存在一个 $0 < C < \infty$ 使得 $\mathbb{E}[(t - \tau_i(t))^2] \leq C$ 对所有 $t \in \mathbb{Z}^+$ 和所有 $i \in [n]$均成立。此外，我们假设对所有 $t \in \mathbb{Z}^+$，存在一个 $\kappa \in (0,1)$ 使得 $t - \tau_i(t) \leq (1-\kappa)t$，即 $\tau_i(t) \geq \kappa t$。

备注3 请注意，假设*3* 远弱于所有延迟均由一个常数上界约束的假设*Tang* 等人 *(2023*年*); Zheng* 等人 *(2023*年，*2024*年）。假设*3* 受到*Sra* 等人 *(2016*）在假设*2.5(B)*中所述时延模型的启发。我们将在后文证明，在这一较弱假设下，我们可以实现与有界延迟假设下相同的时间收敛速率 t。

在假设1-3的条件下，我们得到支撑本文分析的关键引理，该引理量化了延迟对期望距离的影响，即 $\boldsymbol{x}(t)$ 与 $\boldsymbol{x}(\tau_i(t))$。

引理4 假设*1-3* 成立。令 $p(t) = \eta(\lceil \kappa t \rceil)$。则对任意 $i \in [n]$ 以及任意$t \in \mathbb{Z}^+$，有

$$\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}(\tau_i(t))\|^2] \leq n^2GCp^2(t),$$

and

$$\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(t) - \tilde{\boldsymbol{g}}_i(\tau_i(t))\|^2] \leq n^2GCL^2p^2(t).$$

同样，$\mathbb{E}[\|\nabla f(\boldsymbol{x}(t)) - \nabla f(\boldsymbol{x}(\tau_i(t)))\|^2] \leq n^2GCL^2p^2(t)$。

证明 我们有

$$\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t))\|^2]=\mathbb{E}[\|\sum_{\tau=1}^{t-\tau_i(t)}(\boldsymbol{x}(t-\tau+1)-\boldsymbol{x}(t-\tau))\|^2]$$
$$=\mathbb{E}[\|\sum_{\tau=1}^{t-\tau_i(t)}\left(\Pi_{\mathcal{S}}[\boldsymbol{x}(t-\tau)-\eta(t-\tau)\boldsymbol{g}(t-\tau)]-\Pi_{\mathcal{S}}[\boldsymbol{x}(t-\tau)]\right)\|^2]$$
$$\leq\mathbb{E}[(\sum_{\tau=1}^{t-\tau_i(t)}\left\|\Pi_{\mathcal{S}}[\boldsymbol{x}(t-\tau)-\eta(t-\tau)\boldsymbol{g}(t-\tau)]-\Pi_{\mathcal{S}}[\boldsymbol{x}(t-\tau)]\right\|)^2]$$
$$\overset{(a)}{\leq}\mathbb{E}[(\sum_{\tau=1}^{t-\tau_i(t)}\|\eta(t-\tau)\boldsymbol{g}(t-\tau)\|)^2]\overset{(b)}{\leq}\mathbb{E}[\sum_{\tau=1}^{t-\tau_i(t)}\eta^2(t-\tau)\sum_{\tau=1}^{t-\tau_i(t)}\|\boldsymbol{g}(t-\tau)\|^2]$$
$$\overset{(c)}{\leq}\mathbb{E}\left[\mathbb{E}[\sum_{\tau=1}^{t-\tau_i(t)}\eta^2(t-\tau)\sum_{\tau=1}^{t-\tau_i(t)}\|\boldsymbol{g}(t-\tau)\|^2|\tau_i(t)]\right]$$

其中，(a) 由投影的非扩张性质推出（见 Nesterov 等人 (2018)中的推论 2.2.3），(b) 由柯西-施瓦茨不等式推出，(c) 则由全期望定律推出。但是，

$$\mathbb{E}\left[\mathbb{E}[\sum_{\tau=1}^{t-\tau_i(t)}\eta^2(t-\tau)\sum_{\tau=1}^{t-\tau_i(t)}\|\boldsymbol{g}(t-\tau)\|^2|\tau_i(t)]\right]\overset{(a)}{\leq}n^2G\,\mathbb{E}[(t-\tau_i(t))\sum_{\tau=1}^{t-\tau_i(t)}\eta^2(t-\tau)]$$
$$\overset{(b)}{\leq}n^2G\sqrt{\mathbb{E}[(t-\tau_i(t))^2]\,\mathbb{E}[(\sum_{\tau=1}^{t-\tau_i(t)}\eta^2(t-\tau))^2]}\overset{(c)}{\leq}n^2G\sqrt{C\,\mathbb{E}[\left((t-\tau_i(t))\eta^2(\tau_i(t))\right)^2]}$$
$$\overset{(d)}{\leq}n^2G\sqrt{C\,\mathbb{E}[(t-\tau_i(t))^2]\eta^4(\lceil\kappa t\rceil)}\overset{(e)}{\leq}n^2GC\eta^2(\lceil\kappa t\rceil)=n^2GCp^2(t),$$

其中，(a) 由引理 2推出，(b) 由柯西-施瓦茨不等式推出，(c)、(d) 和 (e) 使用假设 3，而 (c) 和 (d) 利用了步长序列的递减性质。

对于第二个不等式，根据假设 2，我们有

$$\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(t)-\tilde{\boldsymbol{g}}_i(\tau_i(t))\|^2]\leq L^2\,\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t))\|^2]\leq n^2GCL^2p^2(t).$$

最后一个不等式的推导与上述不等式类似。 ■

3.1.1. 主要结果

现在我们可以给出主要结果，刻画在我们的假设下（3）的收敛速率。第一个结果确立了（3）对于一般（非凸）函数 f 的收敛性质。

定理 5 令 $\{\boldsymbol{h}(t)\}_{t\in\mathbb{Z}^+}$ 定义为

$$\boldsymbol{h}(t):=\frac{1}{\eta(t)}\left(\boldsymbol{x}(t)-\Pi_{\mathcal{S}}[\boldsymbol{x}(t)-\eta(t)\nabla f(\boldsymbol{x}(t))]\right).$$

令步长序列 $\{\eta(t)\}$ 是一个递减序列，满足 极限$_{t\to\infty}$ $\frac{\eta(\lceil \kappa t \rceil)}{\eta\ t}$ $() = \bar{\eta} > 0$ 其中$\bar{\eta} > 0$ 是一个常数，且 $\sum_{t=0}^{\infty} \eta(t) = \infty$。在假设 *1-3*下，有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2] = \mathcal{O}\left(\frac{\sum_{t=0}^{T}\left(\eta(t)+\eta^2(t)+\eta(t)q^2(t)\right)}{T\eta(T)}\right).$$

特别地，对于 $\eta(t) = \frac{\eta_0}{(t+1)^\alpha}$，若某个常数 $\eta_0 > 0$ 和 $\alpha \in (0,1)$满足条件，且对任意非递增的$\{q(t)\}$，我们有 $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2] = \mathcal{O}(1)$。

备注 6 与*Ghadimi* 等人（*2016*）类似，这里我们讨论约束优化；因此，在非凸情形下，我们通过投影梯度映射 $\boldsymbol{h}(t)$而非 $\nabla f(\boldsymbol{x}(t))$来衡量驻点性。此外，我们可以按照*Ghadimi* 等人（*2016*）中推论 *2* 的相同方式，将结果改写如下。若 s 是一个取非负整数值、概率质量函数为 $\mathbb{P}(s=t) = \frac{\eta(t)}{\sum_{t=0}^{T}\eta(t)}$ 的随机变量，对任意 $t = 0, \ldots\ldots, T$，我们有

$$\mathbb{E}_s[\|\boldsymbol{h}(s)\|^2] = \mathcal{O}\left(\frac{\sum_{t=0}^{T}\left(\eta(t)+\eta^2(t)+\eta(t)q^2(t)\right)}{\sum_{t=0}^{T}\eta(t)}\right).$$

该结果表明 $\mathbb{E}[\|\boldsymbol{h}(s)\|^2]$ 收敛到0附近的一个邻域，其半径由分子中乘在 $\sum_{t=0}^{T}\eta(t)$ 前的系数决定。如下文证明所示，该系数等于 $\frac{5}{2}n^2G$。这与 *Ghadimi* 等人 *(2016)* 的推论 *2* 一致，该推论指出，邻域大小与随机梯度的二阶矩成正比，并随批量大小增加而反比减小。

证明 (定理 *5*的证明)

对于任意 $\boldsymbol{v} \in \mathbb{R}^d$，我们记 $P_{\mathcal{S}}(\boldsymbol{x},\ \boldsymbol{v},\ t) = \frac{1}{\eta\ t()}(\boldsymbol{x} - \Pi_{\mathcal{S}}[\boldsymbol{x} - \eta(t)\boldsymbol{v}])$。因此，我们的目标是为 $\boldsymbol{h}(t) = P_{\mathcal{S}}(\boldsymbol{x}(t)\ \nabla f(\boldsymbol{x}(t))\ t)$ 建立一个界。

如 Ghadimi 等人 (2016) $P_{\mathcal{S}}(\boldsymbol{x},\ \boldsymbol{v},\ t)$ 所证明的，对于任意 $\boldsymbol{v},\ \boldsymbol{v}_1,\ \boldsymbol{v}_2 \in \mathbb{R}^d$和 $t \in \mathbb{Z}^+$，具有以下性质：

$$\|P_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{v},t)\|^2 \le \langle \boldsymbol{v}, P_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{v},t)\rangle \le \|\boldsymbol{v}\|^2, \tag{5}$$

$$\|P_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{v}_1,t) - P_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{v}_2,t)\| \le \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|. \tag{6}$$

为简化记号，稍微滥用一下记法，记 $P_{\mathcal{S}}(\boldsymbol{x}(t)\ \boldsymbol{g}(t)\ t)$为 $P_{\mathcal{S}}(\boldsymbol{x}(t)\ \boldsymbol{g}(t))$。利用$L$-光滑条件 f，式（3）蕴含

$$f(\boldsymbol{x}(t+1)) \le f(\boldsymbol{x}(t)) - \eta(t)\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}(\boldsymbol{x}(t),\boldsymbol{g}(t))\rangle + \frac{L}{2}\eta^2(t)\|P_{\mathcal{S}}(\boldsymbol{x}(t),\boldsymbol{g}(t))\|^2.$$

对两边取期望值，我们得到

$$\mathbb{E}[f(\boldsymbol{x}(t+1))] \le \mathbb{E}[f(\boldsymbol{x}(t))] + \eta(t)\,\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}(\boldsymbol{x}(t),\boldsymbol{g}(t))\rangle] + \frac{L}{2}\eta^2(t)\,\mathbb{E}[\|P_{\mathcal{S}}(\boldsymbol{x}(t),\boldsymbol{g}(t))\|^2]. \tag{7}$$

对于右侧的第二项（7），通过加减中间项，我们得到

$$
\begin{aligned}
&\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}(\boldsymbol{x}(t), \boldsymbol{g}(t))\rangle] \\
=&\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}(\boldsymbol{x}(t), \boldsymbol{g}(t)) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(\tau_i(t))\right)\rangle] \\
&+\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(\tau_i(t))\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(t)\right)\rangle] \\
&+\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(t)\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}(t))\right)\rangle] \\
&+\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}(t))\right)\rangle].
\end{aligned} \tag{8}
$$

为了对式（8）中的每一项进行界定，利用不等式 $(a+b)^2 \leq 2(a^2+b^2)$，可得

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] =& \mathbb{E}[\|\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}(t))\|^2] \leq 2\,\mathbb{E}[\|\sum_{i=1}^{n} (\nabla f_i(\boldsymbol{x}(t)) - \tilde{\boldsymbol{g}}_i(t))\|^2] + 2\,\mathbb{E}[\|\sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(t)\|^2] \\
\overset{\text{(a)}}{\leq}& 2\,\mathbb{E}[(\sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}(t)) - \tilde{\boldsymbol{g}}_i(t)\|)^2] + 2\,\mathbb{E}[\|\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{x}(t))\|^2] \overset{\text{(b)}}{\leq} 2n^2(q^2(t)+G),
\end{aligned} \tag{9}
$$

其中，(a) 是由范数的三角不等式以及对任意随机向量 $\|\mathbb{E}[\boldsymbol{v}]\|^2 \leq \mathbb{E}[\|\boldsymbol{v}\|^2]$ $\boldsymbol{v}$ 得出的。此外，注意到 $\pm\langle \boldsymbol{u},\ \boldsymbol{v}\rangle \leq \frac{1}{2}(a\|\boldsymbol{u}\|^2 + \frac{1}{a}\|\boldsymbol{v}\|^2)$ 对任意 $a>0$ 和所有向量 $\boldsymbol{u},\ \boldsymbol{v} \in \mathbb{R}^d$ 均成立。

因此，根据（9）、（6）以及引理2，式（8）中的第一项可以界定为

$$
\begin{aligned}
&\mathbb{E}[-\langle \nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}(\boldsymbol{x}(t), \boldsymbol{g}(t)) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(\tau_i(t))\right)\rangle] \\
\leq& \frac{1}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2}\,\mathbb{E}[\|P_{\mathcal{S}}(\boldsymbol{x}(t), \boldsymbol{g}(t)) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(\tau_i(t))\right)\|^2] \\
\leq& \frac{1}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2}\,\mathbb{E}[\|\boldsymbol{g}(t) - \sum_{i=1}^{n} \tilde{\boldsymbol{g}}_i(\tau_i(t))\|^2] \\
\leq& n^2q^2(t) + n^2G + \frac{1}{2}n^2G = n^2q^2(t) + \frac{3}{2}n^2G.
\end{aligned}
$$

根据（9）、（6）以及引理4，式（8）中的第二项可以界定为

$$\begin{aligned}
&\mathbb{E}[-\langle\nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(\tau_i(t))\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t)\right)\rangle] \\
\leq&\frac{p(t)}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2p(t)}\,\mathbb{E}[\|P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(\tau_i(t))\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t)\right)\|^2] \\
\leq&\frac{p(t)}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2p(t)}\,\mathbb{E}[\|\sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(\tau_i(t)) - \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t)\|^2] \\
\leq&n^2p(t)q^2(t) + n^2Gp(t) + \frac{1}{2}n^4GCL^2p(t).
\end{aligned}$$

使用与上述类似的论证，根据（9）、（6）和假设2，（8）中的第三项可界定为

$$\begin{aligned}
&\mathbb{E}[-\langle\nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t)\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}(t))\right)\rangle] \\
\leq&\frac{1}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2}\,\mathbb{E}[\|P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t)\right) - P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}(t))\right)\|^2] \\
\leq&\frac{1}{2}\,\mathbb{E}[\|\nabla f(\boldsymbol{x}(t))\|^2] + \frac{1}{2}\,\mathbb{E}[\|\sum_{i=1}^{n}\tilde{\boldsymbol{g}}_i(t) - \sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}(t))\|^2] \\
\leq&n^2q^2(t) + n^2G + \frac{1}{2}n^2q^2(t) = \frac{3}{2}n^2q^2(t) + n^2G.
\end{aligned}$$

根据（5），（8）中的最后一项可界定为

$$\mathbb{E}[-\langle\nabla f(\boldsymbol{x}(t)), P_{\mathcal{S}}\left(\boldsymbol{x}(t), \sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}(t))\right)\rangle] \leq -\,\mathbb{E}[\|P_{\mathcal{S}}(\boldsymbol{x}(t), \nabla f(\boldsymbol{x}(t)))\|^2] = -\,\mathbb{E}[\|\boldsymbol{h}(t)\|^2].$$

于是，（7）变为

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{x}(t+1))] \leq \mathbb{E}[f(\boldsymbol{x}(t))] - \eta(t)\ \mathbb{E}[\|\boldsymbol{h}(t)\|^2] \\
&+ \frac{5}{2}n^2G\eta(t) + \frac{1}{2}n^2GL\eta^2(t) + n^2G(1 + \frac{1}{2}n^2CL^2)\eta(t)p(t) + \frac{5}{2}n^2\eta(t)q^2(t) + n^2\eta(t)p(t)q^2(t).
\end{aligned}$$

对上述不等式从 $t = 0$，……，T 求望远镜求和，可得

$$\begin{aligned}
&\sum_{t=0}^{T}\eta(t)\,\mathbb{E}[\|\boldsymbol{h}(t)\|^2] \leq \mathbb{E}[f(\boldsymbol{x}(0))] - f^* + \frac{5}{2}n^2G\sum_{t=0}^{T}\eta(t) + \frac{1}{2}n^2GL\sum_{t=0}^{T}\eta^2(t) \\
&+n^2G(1 + \frac{1}{2}n^2CL^2)\sum_{t=0}^{T}\eta(t)p(t) + \frac{5}{2}n^2\sum_{t=0}^{T}\eta(t)q^2(t) + n^2\sum_{t=0}^{T}\eta(t)p(t)q^2(t).
\end{aligned}$$

令 $V(T) \coloneqq \sum_{t=0}^{T}\eta(t)\ \mathbb{E}[\|\boldsymbol{h}(t)\|^2]$。由于 $p(t) = \eta(\lceil\kappa t\rceil) = \mathcal{O}(\eta(t))$，我们有

$$V(T) = \mathcal{O}\left(\sum_{t=0}^{T}\left(\eta(t) + \eta^2(t) + \eta(t)q^2(t)\right)\right). \tag{10}$$

接下来，我们为 $M(T) := \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2]$ 建立一个上界。对于任意 $\theta \in (0,\ 1)$，定义

$$M_\theta(T) = \left[\frac{1}{T+1}\sum_{t=0}^{T}\left(\mathbb{E}[\|\boldsymbol{h}(t)\|^2]\right)^\theta\right]^{\frac{1}{\theta}}.$$

注意，根据 Hölder 不等式（Folland （1999年）中的定理 6.2），对于任意 $p,\ q > 1$ 满足 $\frac{1}{p}+\frac{1}{q}=1$，以及非负序列 $\{a_t\}_{t=0}^T$ 和 $\{b_t\}_{t=0}^T$，有

$$\left(\sum_{t=0}^{T} a_t b_t\right)^q \le \left(\sum_{t=0}^{T} a_t{}^p\right)^{\frac{q}{p}} \sum_{t=0}^{T} b_t{}^q. \tag{11}$$

令 $a_t = \left(\eta_t{}^{1}()\right)^\theta$，$b_t = \left(\eta(t)\,\mathbb{E}[\|\boldsymbol{h}(t)\|^2]\right)^\theta$，以及 $(p,\ q) = \left(\frac{1}{1-\theta}, \frac{1}{\theta}\right)$。注意到 $\sum_{t=0}^{T} b_t^{\frac{1}{\theta}} = V(T)$，并利用（11），可得

$$\begin{aligned} M_\theta(T) &= \left(\frac{1}{T+1}\sum_{t=0}^{T} a_t b_t\right)^{\frac{1}{\theta}} \le (T+1)^{-\frac{1}{\theta}} \left(\sum_{t=0}^{T} a_t^{\frac{1}{1-\theta}}\right)^{\frac{1-\theta}{\theta}} \sum_{t=0}^{T} b_t^{\frac{1}{\theta}} \\ &= (T+1)^{-\frac{1}{\theta}} \left(\sum_{t=0}^{T}\left(\frac{1}{\eta(t)}\right)^{\frac{\theta}{1-\theta}}\right)^{\frac{1-\theta}{\theta}} V(T). \end{aligned}$$

采用与 Reisizadeh 等人（2023年）中命题 1 的证明类似的论证，可以证明，对于任意 $\theta \in$（0，1），

$$\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2] \le \left[\sum_{t=0}^{T}\left(\mathbb{E}[\|\boldsymbol{h}(t)\|^2]\right)^\theta\right]^{\frac{1}{\theta}}.$$

现在我们可以将 $M(T)$ 的上界写为

$$M(T) \le (T+1)^{\frac{1}{\theta}-1} M_\theta(T) \le (T+1)^{-1}\left(\sum_{t=0}^{T}\left(\frac{1}{\eta(t)}\right)^{\frac{\theta}{1-\theta}}\right)^{\frac{1-\theta}{\theta}} V(T). \tag{12}$$

注意到 $\{\eta(t)\}$ 是一个递减序列，因此有

$$\lim_{\theta\to 1}\left(\sum_{t=0}^{T}\left(\frac{1}{\eta(t)}\right)^{\frac{\theta}{1-\theta}}\right)^{\frac{1-\theta}{\theta}} = \max_{0\le t\le T}\frac{1}{\eta(t)} = \frac{1}{\eta(T)}.$$

由于（12）对所有 $\theta \in (0,\ 1)$ 均成立，$M(T) = \mathcal{O}\left(T^{-1}\eta^{1}(T)\,V(T)\right)$。最后，回顾 $V(T)$ 在（10）中的上界，可得

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\boldsymbol{h}(t)\|^2] = \mathcal{O}\left(\frac{\sum_{t=0}^{T}\left(\eta(t)+\eta^2(t)+\eta(t)q^2(t)\right)}{T\eta(T)}\right).$$

■

我们的第二个结果为强凸成本函数 f 建立了收敛保证。

定理 7 对于 μ-强凸函数 f，即对所有 $\boldsymbol{x}$、$\boldsymbol{y} \in \mathcal{S}$，$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$，令 $\boldsymbol{x}^* = \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x})$。在假设 1-3 下，若 $\{\eta(t)\}$ 是一个递减序列，且 $\lim_{t\to\infty} \eta(\lceil \kappa t \rceil)\, \eta(t) = \bar{\eta} > 0$，其中 $\bar{\eta} > 0$ 为常数，$\sum_{t=0}^{\infty} \eta(t) = \infty$，$\sum_{t=0}^{\infty} \eta^2(t) < \infty$，并且 $\sum_{t=0}^{\infty} \eta(t) q^2(t) < \infty$，则有 $\lim_{t\to\infty} \mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] = 0$。

此外，如果 $\eta(t) = \frac{\eta_0}{t+1}$ 对某个常数 $\eta_0 > 0$成立，且 $q(t) = \frac{q_0}{(t+1)^\beta}$ 对某些常数 $q_0 > 0$和 $\beta \geq \frac{1}{2}$成立，那么 $\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] = \mathcal{O}(\frac{1}{t})$。

证明 (定理7的证明概要

我们首先为 $v(t)$ 建立一个递归不等式：$= \mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2]$。

由（3），利用投影的非扩张性质（参见 Nesterov 等人(2018年）可得

$$\begin{aligned}\|\boldsymbol{x}(t+1) - \boldsymbol{x}^*\|^2 =& \|\Pi_{\mathcal{S}}[\boldsymbol{x}(t) - \eta(t)\boldsymbol{g}(t)] - \Pi_{\mathcal{S}}[\boldsymbol{x}^*]\|^2 \leq \|\boldsymbol{x}(t) - \boldsymbol{x}^* - \eta(t)\boldsymbol{g}(t)\|^2 \\ =& \|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 + \eta^2(t)\|\boldsymbol{g}(t)\|^2 - 2\eta(t)\langle \boldsymbol{x}(t) - \boldsymbol{x}^*, \boldsymbol{g}(t)\rangle.\end{aligned}$$

以下论证与 Zheng 等人 (2024年中定理 1 的证明基本一致，因此完整细节留待附录说明。 ■

推论 8 对于 μ-强凸函数 f，在假设 1-3成立的条件下，若对所有 t，$\eta(t) = \eta$ 在$\eta \in (0, \frac{1}{\mu})$ 内成立，且 $q(t) = q > 0$，则有

$$\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] \leq (1 - \mu\eta)^t\, \mathbb{E}[\|\boldsymbol{x}(0) - \boldsymbol{x}^*\|^2] + n^2 G(1 + 2\sqrt{C})\frac{\eta}{\mu} + 2n^4 C G L^2 \frac{\eta^2}{\mu^2} + 2n^2 \frac{q^2}{\mu^2}.$$

若对所有 $t, \eta(t) = \eta$ 在 $\eta \in (0, \frac{1}{\mu})$内成立，且 $q(t) = \frac{q_0}{(t+1)^\beta}$对 $\beta > 0$成立，则有

$$\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] = \mathcal{O}(t^{-2\beta}) + n^2 G(1 + 2\sqrt{C})\frac{\eta}{\mu} + 2n^4 C G L^2 \frac{\eta^2}{\mu^2}.$$

注 9 首先，在较弱的延迟假设（假设 3）下，定理 7 恢复了有界延迟结果中 Zheng 等人 $\mathcal{O}(\frac{1}{T})$ 得到的随机梯度下降最优速率 Zheng 等人 (2024年)。其次，推论 8重现了 Zheng 等人 Zheng 等人 (2024年) 中定理 2 的速率，该速率适用于常数步长和常数偏差；此外还表明，当步长为常数且偏差递减时，均方误差会收敛到一个邻域，其半径与具体的偏差序列无关 $\{q(t)\}$；这种收敛是亚指数的，但又是超多项式的。

我们的最后一个结果研究凸（但不一定是强凸）函数的主要动力学（3） f。

定理10 对于凸函数 f，除假设 1-3 外，假设存在 $R \in (0, \infty)$，使得 $\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] \leq R$ 对于 $\boldsymbol{x}^* \in \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x})$ 以及所有 t 均成立（如果 $\mathcal{S}$ 有界，这在所有实际情形中自然成立）。令 $\{\eta(t)\}$ 为递减序列，使得 $\lim_{t\to\infty} \eta(\lceil \kappa t \rceil)\, \eta(t) = \bar{\eta} > 0$，其中 $\bar{\eta} > 0$ 为常数，且 $\sum_{t=0}^{\infty} \eta(t) = \infty$。令 $\tilde{\boldsymbol{x}}(T) = \sum_{t=0}^{T} \frac{\eta(t)}{\sum_{t=0}^{T} \eta(t)} \boldsymbol{x}(t)$，则有

$$\mathbb{E}[f(\tilde{\boldsymbol{x}}(T))] - f^* = \mathcal{O}\left(\frac{\sum_{t=0}^{T}\left(\eta^2(t) + \eta(t)q(t)\right)}{\sum_{t=0}^{T} \eta(t)}\right).$$

特别地，若 $\eta(t) = \frac{\eta_0}{\sqrt{t+1}}$对某个常数 $\eta_0 > 0$成立，且 $q(t) = \frac{q_0}{(t+1)^\beta}$ 对某个常数 $q_0 > 0$ 和 $\beta \geq \frac{1}{2}$成立，则 $\mathbb{E}[f(\tilde{\boldsymbol{x}}(T))] - f^* = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) = \mathcal{O}\left(\frac{1}{T^{\frac{1}{2}-\varepsilon}}\right)$ 对任意 $0 < \varepsilon < \frac{1}{2}$均成立。

注 11 请注意，$\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] \leq R$ 是一个较弱的假设；具体而言，延迟自适应工作 *Sra* 等人 *(2016)* 中的假设 *2.3* 严格更强。定理 *10*表明，对于凸目标函数，我们的收敛速率 $\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$ 与经典随机梯度下降的最优速率相匹配 *Garrigos* 和 *Gower(2023*年*)*，与 *Sra* 等人 *(2016)* 提出的延迟自适应方法相比，差异至多为一个对数因子。

4. 结论

总之，我们针对具有延迟和近似（可能存在偏差）梯度模型的分布式随机优化，构建了一个统一框架，并表明在较弱的缩放延迟条件下，简单地预先选定递减步长即可实现近乎最优的性能。我们的结果恢复了非凸和强凸目标函数的经典随机梯度下降收敛速率，并且对于凸目标函数，在对数因子范围内达到延迟自适应保证，从而阐明在此情形下，实现最优收敛性能并不需要延迟自适应。除了重新审视步长选择的作用外，我们的分析还清晰刻画了偏差、随机性、约束与缩放延迟之间的相互作用，为设计稳健的分布式学习系统提供了实用指导，并指出了若干未来研究方向，例如在不含对数项的情况下推导更紧的凸情形界，以及推广到更一般的网络化架构或完全去中心化架构。

致谢

本研究获得了美国国家科学基金会 CNS-2148313 项目、AutoCOMBOT 多学科大学研究计划、美国空军科学研究办公室 FA9550-23-1-0057 项目以及美国空军科学研究办公室 FA9550-24-1-0129 项目的资助。

参考文献

Dan Alistarh、Demjan Grubic、Jerry Li、Ryota Tomioka和Milan Vojnovic。QSGD：通过梯度量化与编码实现通信高效的随机梯度下降。神经信息处理系统进展，30，2017年。Wenzhi Fang、Ziyi Yu、Yuning Jiang、Yuanming Shi、Colin N Jones和Yong Zhou。面向联邦学习的通信高效随机零阶优化。*IEEE* 信号处理汇刊，70:5058–5073，2022年。Gerald B Folland。实分析：现代技术及其应用，第40卷。John Wiley & Sons，1999年。Guillaume Garrigos和Robert M Gower。《（随机）梯度方法收敛定理手册》。*arXiv*预印本*arXiv:2301.11235*，2023年。Saeed Ghadimi、Guanghui Lan和Hongchao Zhang。面向非凸随机复合优化的小批量随机逼近方法。《数学规划》，155(1):267–305，2016年。

Sadhika Malladi、Tianyu Gao、Eshaan Nichani、Alex Damian、Jason D Lee、Danqi Chen和Sanjeev Arora。仅用前向传播微调语言模型。神经信息处理系统进展，36:53038–53075，2023年。Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson和Blaise Aguera y Arcas。基于去中心化数据的深度网络通信高效学习。载于人工智能与统计，第1273–1282页。PMLR，2017年。Angelia Nedi´c和Alex Olshevsky。时变有向图上的分布式优化。*IEEE*自动控制汇刊，60(3):601–615，2014年。Angelia Nedi´c和Asuman Ozdaglar。多智能体优化的分布式次梯度法。*IEEE*自动控制汇刊，54(1):48–61，2009年。Arkadi Nemirovski、Anatoli Juditsky、Guanghui Lan和Alexander Shapiro。随机规划的鲁棒随机逼近方法。*SIAM*优化期刊，19(4):1574–1609，2009年。Yurii Nesterov等。凸优化讲义，第137卷。施普林格，2018年。Benjamin Recht、Christopher Re、Stephen Wright和Feng Niu。Hogwild!：一种并行化随机梯度下降的无锁方法。神经信息处理系统进展，24，2011年。Amirhossein Reisizadeh、Isidoros Tziotis、Hamed Hassani、Aryan Mokhtari和Ramtin Pedarsani。抗掉队联邦学习：利用统计精度与系统异构性之间的相互作用。*IEEE*信息论选题期刊，3(2):197–205，2022a年。Hadi Reisizadeh、Behrouz Touri和Soheil Mohajer。信息不完美共享下时变图上的分布式优化。*IEEE*自动控制汇刊，68(7):4420–4427，2022b年。Hadi Reisizadeh、Behrouz Touri和Soheil Mohajer。Dimix：面向松散智能体的递减混合。*SIAM*优化期刊，33(2):978–1005，2023年。Herbert Robbins和David Siegmund。非负几乎上鞅的收敛定理及其若干应用。载于统计学中的优化方法，第233–257页。爱思唯尔，1971年。Ohad Shamir。带两点反馈的赌博机与零阶凸优化的最优算法。机器学习研究期刊，18(1):1703–1713，2017年。Suvrit Sra、Adams Wei Yu、Mu Li和Alex Smola。Adadelay：延迟自适应分布式随机优化。载于人工智能与统计，第957–965页。PMLR，2016年。Yujie Tang、Zhaolin Ren和Na Li。面向协作多智能体系统的零阶反馈优化。*Automatica*，148:110741，2023年。Tatiana Tatarenko和Behrouz Touri。非凸分布式优化。*IEEE*自动控制汇刊，62(8):3744–3757，2017年。

Konstantinos I Tsianos、Sean Lawlor 和 Michael G Rabbat。基于共识的分布式优化：大规模机器学习中的实际问题与应用。载于*2012*年第*50*届 *Allerton* 通信、控制与计算年会（*Allerton*），第 1543–1550 页。IEEE，2012年。

Zhaohui Yang、Mingzhe Chen、Kai-Kit Wong、H Vincent Poor 和 Shuguang Cui。面向 6G 的联邦学习：应用、挑战与机遇。工程，8：33–41，2022年。

Xinran Zheng、Tara Javidi 和 Behrouz Touri。采用递减步长和平滑半径的协作多智能体系统零阶非凸优化。*IEEE* 控制系统快报，7：2785–2790，2023年。

Xinran Zheng、Tara Javidi 和 Behrouz Touri。可分解代价函数近似延迟梯度下降的一般框架。*IEEE* 控制系统快报，8：1229–1234，2024年。

附录 A. 引理的证明 2

证明 利用2) 中 $\boldsymbol{g}(t)$ 以及假设 2 可得

$$\mathbb{E}[\|\boldsymbol{g}(t)\|^2] = \mathbb{E}\left[\left\|\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))\right\|^2\right] \leq n\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))\|^2\right] \leq n^2 G.$$

■

附录 B. 定理的证明 7

证明 我们首先为 $v(t)$ 建立一个递归不等式：$= \mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2]$。

由（3），利用投影的非扩张性质（见 Nesterov 等人(2018) 中的推论 2.2.3），可得

$$\begin{aligned}\|\boldsymbol{x}(t+1) - \boldsymbol{x}^*\|^2 =& \|\Pi_{\mathcal{S}}[\boldsymbol{x}(t) - \eta(t)\boldsymbol{g}(t)] - \Pi_{\mathcal{S}}[\boldsymbol{x}^*]\|^2 \leq \|\boldsymbol{x}(t) - \boldsymbol{x}^* - \eta(t)\boldsymbol{g}(t)\|^2 \\ =& \|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 + \eta^2(t)\|\boldsymbol{g}(t)\|^2 - 2\eta(t)\langle \boldsymbol{x}(t) - \boldsymbol{x}^*, \boldsymbol{g}(t)\rangle.\end{aligned}$$

对两边取期望值，并利用引理2，可得

$$v(t+1) \leq v(t) + n^2 G\eta^2(t) - 2\eta(t)\,\mathbb{E}[\langle \boldsymbol{x}(t) - \boldsymbol{x}^*, \boldsymbol{g}(t)\rangle].$$

将上述不等式右侧的最后一项改写为

$$
\begin{aligned}
&-\mathbb{E}[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \boldsymbol{g}(t)\rangle] = -\mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))\rangle\right] \\
=&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t)), \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\rangle\right] - \sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(\tau_i(t))-\boldsymbol{x}^*, \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\rangle\right] \\
&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(\tau_i(t))-\tilde{\boldsymbol{g}}_i(t)\rangle\right] - \sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(t)-\nabla f_i(\boldsymbol{x}(t))\rangle\right] \\
&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \nabla f_i(\boldsymbol{x}(t))\rangle\right]. \qquad (13)
\end{aligned}
$$

为了界定式（13）中的第一项，利用柯西-施瓦茨不等式和引理 4中的第一个不等式，可得

$$
\begin{aligned}
&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t)), \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\rangle\right] \le \sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t))\|\|\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\|\right] \\
\le&\sum_{i=1}^n \sqrt{\mathbb{E}\left[\|\boldsymbol{x}(t)-\boldsymbol{x}(\tau_i(t))\|^2\right]}\cdot\sqrt{\mathbb{E}\left[\|\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\|^2\right]} \\
\le&\sum_{i=1}^n \sqrt{n^2GCp^2(t)\cdot G} = n^2\sqrt{C}Gp(t).
\end{aligned}
$$

对于式（13）中的第二项，利用全期望公式，可得

$$
\begin{aligned}
&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(\tau_i(t))-\boldsymbol{x}^*, \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\rangle\right] \\
=&-\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\langle \boldsymbol{x}(\tau_i(t))-\boldsymbol{x}^*, \boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))\rangle|\sigma(\boldsymbol{x}(\tau_i(t)))\right]\right] \\
=&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(\tau_i(t))-\boldsymbol{x}^*, \mathbb{E}\left[\boldsymbol{g}_i(\boldsymbol{x}(\tau_i(t)))-\tilde{\boldsymbol{g}}_i(\tau_i(t))|\sigma(\boldsymbol{x}(\tau_i(t)))\right]\rangle\right] = 0.
\end{aligned}
$$

为了界定（13）中的第三项，利用 $ab \le \frac{1}{2}\left(a^2+b^2\right)$ 和引理4，可得

$$
\begin{aligned}
&-\sum_{i=1}^n \mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(\tau_i(t))-\tilde{\boldsymbol{g}}_i(t)\rangle\right] \\
\le&\frac{1}{2}\sum_{i=1}^n\left(\frac{\mu}{2n}\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}^*\|^2]+\frac{2n}{\mu}\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(\tau_i(t))-\tilde{\boldsymbol{g}}_i(t)\|^2]\right) \le \frac{1}{2}\left(\frac{\mu}{2}v(t)+2n^4CGL^2\frac{p^2(t)}{\mu}\right).
\end{aligned}
$$

为了界定（13）中的第四项，利用 $ab \leq \frac{1}{2}(a^2+b^2)$ 和假设2，可得

$$\begin{aligned}&-\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(t)-\nabla f_i(\boldsymbol{x}(t))\rangle\right]\\ \leq&\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\mu}{2n}\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}^*\|^2]+\frac{2n}{\mu}\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(t)-\nabla f_i(\boldsymbol{x}(t))\|^2]\right)\leq\frac{1}{2}\left(\frac{\mu}{2}v(t)+2n^2\frac{q^2(t)}{\mu}\right).\end{aligned}$$

为了对（13）中的最后一项进行界定，利用 f的强凸性，我们有

$$-\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \nabla f_i(\boldsymbol{x}(t))\rangle\right]=-\mathbb{E}[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \nabla f(\boldsymbol{x}(t))\rangle]\leq-\mu\,\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}^*\|^2]=-\mu v(t).$$

利用上述不等式对（13）中的各项进行界定，可得关于 v（t）的以下递归不等式

$$v(t+1)\leq(1-\mu\eta(t))v(t)+n^2G\eta^2(t)+2n^2\sqrt{C}G\eta(t)p(t)+2n^4CGL^2\frac{\eta(t)p^2(t)}{\mu}+2n^2\frac{\eta(t)q^2(t)}{\mu}. \tag{14}$$

由于 p（t）$=\mathcal{O}$（η（t））， 如果 $\sum_{t=0}^{\infty}\eta$（t）$=\infty$， $\sum_{t=0}^{\infty}\eta^2$（$t$）$<\infty$，且 $\sum_{t=0}^{\infty}\eta$（t）q^2（t）$<\infty$，利用 Robbins-Siegmund 定理的确定性版本Robbins 和 Siegmund（1971），我们可以得出

$$\lim_{t\to\infty}v(t)=\lim_{t\to\infty}\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}^*\|^2]=0.$$

若我们选择 $\eta(t)=\frac{\eta_0}{t+1}$，其中 $\eta_0>0$，且 $q(t)=\frac{q_0}{(t+1)^\beta}$，其中常数 $q_0>0$ 和 $\beta\geq\frac{1}{2}$，则根据 Reisizadeh 等人的引理5（2022b），或参照 Nemirovski 等人（2009）第2.1节中的推导采用类似论证，可得 $v(t)=\mathbb{E}[\|\boldsymbol{x}(t)-\boldsymbol{x}^*\|^2]=\mathcal{O}(\frac{1}{t})$。■

附录C. 推论证明8

证明 若令 $\eta(t)=\eta$ 且 $q(t)=q$ 对所有 t，(14) 可写为

$$v(t+1)\leq(1-\mu\eta)v(t)+n^2G(1+2\sqrt{C})\eta^2+2n^4CGL^2\frac{\eta^3}{\mu}+2n^2\frac{\eta q^2}{\mu}.$$

令 $r(\eta,\ q):=n^2G(1+2\sqrt{C})\eta^2+2n^4CGL^2\frac{\eta^3}{\mu}+2n^2\frac{\eta q^2}{\mu}$，那么，当 $\mu\eta<1$ 时，我们有

$$v(t)\leq(1-\mu\eta)^tv(0)+r(\eta,q)\sum_{s=0}^{t-1}(1-\mu\eta)^s\leq(1-\mu\eta)^tv(0)+r(\eta,q)\frac{1}{\mu\eta}.$$

因此，当 $0<\eta<\frac{1}{\mu}$ 时， $\boldsymbol{x}(t)$以指数速度收敛到 $\boldsymbol{x}^*$ 的 r-邻域，其中$r=r(\eta,q)\frac{1}{\mu\eta}=n^2G(1+2\sqrt{C})\frac{\eta}{\mu}+2n^4CGL^2\frac{\eta^2}{\mu^2}+2n^2\frac{q^2}{\mu^2}$。

如果我们令 $\eta(t) = \eta$ 对所有 t 且 $\{q(t)\}$ 为形如 $q(t) = \frac{q_0}{(t+1)^\beta}$的递减序列，(14)变为

$$v(t+1) \leq (1-\mu\eta)v(t) + n^2G(1+2\sqrt{C})\eta^2 + 2n^4CGL^2\frac{\eta^3}{\mu} + 2n^2\frac{\eta}{\mu}q^2(t).$$

设 $r'(\eta) := n^2G(1+2\sqrt{C})\eta^2 + 2n^4CGL^2\frac{\eta^3}{\mu}$。那么，当 $\mu\eta < 1$ 时，我们有

$$\begin{aligned} v(t) \leq & (1-\mu\eta)^t v(0) + r'(\eta)\sum_{s=0}^{t-1}(1-\mu\eta)^s + 2n^2\frac{\eta}{\mu}\sum_{s=0}^{t-1}(1-\mu\eta)^{t-1-s}q^2(s) \\ \leq & (1-\mu\eta)^t v(0) + r'(\eta)\frac{1}{\mu\eta} + 2n^2\frac{\eta}{\mu}\sum_{s=0}^{t-1}(1-\mu\eta)^{t-1-s}q^2(s). \end{aligned}$$

根据Reisizadeh 等人(2022年b中的引理5，我们有 $\sum_{s=0}^{t-1}(1-\mu\eta)^{t-1-s}q^2(s) = \mathcal{O}(t^{-2\beta})$。因此，当$0 < \eta < \frac{1}{\mu}$， $\boldsymbol{x}(t)$ 收敛到 r'-邻域 $\boldsymbol{x}^*$，其中 $r' = r'(\eta)\frac{1}{\mu\eta} = n^2G(1+2\sqrt{C})\frac{\eta}{\mu} + 2n^4CGL^2\frac{\eta^2}{\mu^2}$。 —

■

附录D。定理10的证明

证明取一个 $x^* \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{S}} f(\boldsymbol{x})$。我们仍采用 $v(t) = \mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2]$的动态过程，并分别对（13）中的每一项进行界定，但处理方式略有不同。

我们用证明定理 7 的相同方法来界定 (13) 中的前两项。

利用柯西-施瓦茨不等式、引理 4，以及 $\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] \leq R$ 对所有 t 均成立这一假设，我们可以将 (13) 中的第三项界定为

$$\begin{aligned} & -\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}(t) - \boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(\tau_i(t)) - \tilde{\boldsymbol{g}}_i(t)\rangle\right] \\ \leq & \sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\| \cdot \|\tilde{\boldsymbol{g}}_i(\tau_i(t)) - \tilde{\boldsymbol{g}}_i(t)\|] \\ \leq & \sum_{i=1}^{n}\sqrt{\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2]} \cdot \sqrt{\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(\tau_i(t)) - \tilde{\boldsymbol{g}}_i(t)\|^2]} \leq n^2\sqrt{RGC}Lp(t). \end{aligned}$$

利用柯西-施瓦茨不等式、假设2以及 $\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2] \leq R$对所有 t成立这一假设，我们可以将式（13）中的第四项界定为

$$\begin{aligned} & -\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}(t) - \boldsymbol{x}^*, \tilde{\boldsymbol{g}}_i(t) - \nabla f_i(\boldsymbol{x}(t))\rangle\right] \\ \leq & \sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\| \cdot \|\tilde{\boldsymbol{g}}_i(t) - \nabla f_i(\boldsymbol{x}(t))\|] \\ \leq & \sum_{i=1}^{n}\sqrt{\mathbb{E}[\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2]} \cdot \sqrt{\mathbb{E}[\|\tilde{\boldsymbol{g}}_i(t) - \nabla f_i(\boldsymbol{x}(t))\|^2]} \leq n\sqrt{R}q(t). \end{aligned}$$

利用 f的凸性，我们可以将式（13）中的最后一项界定为

$$-\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \nabla f_i(\boldsymbol{x}(t))\rangle\right] = -\mathbb{E}[\langle \boldsymbol{x}(t)-\boldsymbol{x}^*, \nabla f(\boldsymbol{x}(t))\rangle] \le f(\boldsymbol{x}^*) - \mathbb{E}[f(\boldsymbol{x}(t))].$$

利用上述不等式对（13）中的各项进行界定，可得以下关于$\eta(t)\left(\mathbb{E}[f(\boldsymbol{x}(t))]-f(\boldsymbol{x}^*)\right)$：

$$\begin{aligned}2\eta(t)\left(\mathbb{E}[f(\boldsymbol{x}(t))]-f(\boldsymbol{x}^*)\right) \le & v(t)-v(t+1)+n^2G\eta^2(t)+2n^2\sqrt{CG}(\sqrt{G}+L\sqrt{R})\eta(t)p(t)\\ & +2n\sqrt{R}\eta(t)q(t).\end{aligned}$$

通过望远镜求和，我们得到

$$\begin{aligned}2\sum_{t=0}^{T}\eta(t)\left(\mathbb{E}[f(\boldsymbol{x}(t))]-f(\boldsymbol{x}^*)\right) \le & v(0)+n^2G\sum_{t=0}^{T}\eta^2(t)+2n^2\sqrt{CG}(\sqrt{G}+L\sqrt{R})\sum_{t=0}^{T}\eta(t)p(t)\\ & +2n\sqrt{R}\sum_{t=0}^{T}\eta(t)q(t).\end{aligned}$$

将两边除以 $2\sum_{t=0}^{T}\eta(t)$，得到

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{t=0}^{T}\eta(t)f(\boldsymbol{x}(t))}{\sum_{t=0}^{T}\eta(t)}\right]-f(\boldsymbol{x}^*) \le & \frac{1}{2\sum_{t=0}^{T}\eta(t)}\left(v(0)+n^2G\sum_{t=0}^{T}\eta^2(t)\right.\\ & \left.+2n^2\sqrt{CG}(\sqrt{G}+L\sqrt{R})\sum_{t=0}^{T}\eta(t)p(t)+2n\sqrt{R}\sum_{t=0}^{T}\eta(t)q(t)\right).\end{aligned}$$

令 $\tilde{\boldsymbol{x}}(T)=\sum_{t=0}^{T}\frac{\eta(t)}{\sum_{t=0}^{T}\eta(t)}\boldsymbol{x}(t)$。由于 f 是凸的，我们有

$$\mathbb{E}[f(\tilde{\boldsymbol{x}}(T))]-f^* \le \mathbb{E}\left[\frac{\sum_{t=0}^{T}\eta(t)f(\boldsymbol{x}(t))}{\sum_{t=0}^{T}\eta(t)}\right]-f^* = \mathcal{O}\left(\frac{\sum_{t=0}^{T}\left(\eta^2(t)+\eta(t)q(t)\right)}{\sum_{t=0}^{T}\eta(t)}\right).$$

■