Operations Research and Fuzziology 运筹与模糊学, 2024, 14(4), 342-357
Published Online August 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.144403

Hans汉斯

求解DC问题的一类随机优化算法

陈梦婷，裴训龙，李登辉

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2024年6月12日；录用日期：2024年8月7日；发布日期：2024年8月14日

摘　要

本文研究的是一类具有有限和形式的DC问题，其目标函数为具有有限和形式的光滑凸函数与连续凸函数之和再减去适当的闭凸函数的形式。传统的邻近DC算法(pDCA)在处理此类问题时，由于每一迭代步都需要对目标函数光滑部分的全梯度进行计算，从而导致计算成本较为昂贵，因此本文将随机梯度SARAH引入到pDCA中，提出了一种基于随机梯度SARAH的随机邻近DC算法(pDCA-SARAH)，并给出了该算法的具体迭代格式，以降低计算成本。在非凸情形下，本文针对pDCA-SARAH算法给出了收敛性及收敛率分析。具体的，本文给出了目标函数在期望意义下的下降量分析以及次线性收敛率的结果。最后，通过将pDCA-SARAH算法用于求解$l_{1\text{-}2}$正则化最小二乘问题，并与pDCA进行数值比较，展示了本文所提算法的高效性。

关键词

DC问题，随机梯度，$l_{1\text{-}2}$正则化最小二乘问题

A Class of Stochastic Optimization Algorithms for Solving DC Problems

Mengting Chen, Xunlong Pei, Denghui Li

School of Mathematics and Statistics, Nanjing University of Information Science and Technology (NUIST), Nanjing Jiangsu

Received: Jun. 12th, 2024; accepted: Aug. 7th, 2024; published: Aug. 14th, 2024

Abstract

In this paper, we study a class of DC problems with finite sum form, whose objective function is the sum of smooth convex function and continuous convex function with finite sum form, minus the appropriate closed convex function. When dealing with this kind of problems, the traditional proximal difference-of-convex algorithm (pDCA) needs to calculate the full gradient of the smooth

文章引用: 陈梦婷, 裴训龙, 李登辉. 求解 DC 问题的一类随机优化算法[J]. 运筹与模糊学, 2024, 14(4): 342-357.
DOI: 10.12677/orf.2024.144403

part of the objective function at each iterative step, so the computational cost is expensive. In this paper, stochastic gradient SARAH is introduced into pDCA, a stochastic proximal DC algorithm (pDCA-SARAH) based on stochastic gradient SARAH is proposed, and the specific iterative scheme of the algorithm is given to reduce the computational cost. In the non-convex case, the convergence and convergence rate analysis of pDCA-SARAH algorithm are given in this paper. Specifically, this paper gives the analysis of the decline of the objective function in the sense of expectation and the results of sublinear convergence rate. Finally, the pDCA-SARAH algorithm is applied to solve the $l_{1\text{-}2}$ regularized least square problem, and compared with pDCA, the efficiency of the proposed algorithm is demonstrated.

Keywords

DC Problems, Stochastic Gradient, $l_{1\text{-}2}$ Regularized Least Squares Problems

1. 引言

优化问题是指在既定约束条件下求解目标函数最值的问题，通过研究优化问题可实现提升效率，降低成本，实现资源最优配置与决策智能化等目的，因而在数学、计算机和金融等领域中至关重要。本文研究的是一类具有有限和形式的 DC (difference of convex, DC)问题，具体格式如下所示：

$$\min_x \ G(x)=\sum_{i=1}^{n} f_i(x)+h(x)-g(x), \tag{1.1}$$

其中 $f_i(x)\ (i=1,2,\cdots,n)$ 为光滑凸函数，且 $\nabla f_i(x)$ 是 Lipschitz 连续的，此外 $h(x)$ 是一个连续凸函数，$g(x)$ 是一个适当的闭凸函数。上述有限和形式的 DC 问题在许多应用领域都存在广泛应用，如机器学习[1]、压缩感知[2]、logistic 回归[3]和二分类问题[4]等。具体的，我们给出如下两个应用实例，其可表述为模型(1.1)的格式。

1) 正则化最小二乘问题[2]

最小二乘问题是一种常用的优化方法，其在高维数据分析、图像处理和金融风险建模等领域具有广泛应用，其中 l_{1-2} 正则化最小二乘问题结合了 l_1 范数和 l_2 范数的优点，通常用于求解一个既稀疏又保持一定稳定性的解。该方法在处理含有异常值或者噪声较大的数据时效果较好，其具体表述为如下优化模型：

$$\min_{x\in\mathbb{R}^n} F_{1-2}(x):=\frac{1}{2}\|Ax-b\|^2+\lambda\left(\|x\|_1-\|x\|\right),$$

其中矩阵 $A\in\mathbb{R}^{m\times n}$，向量 $b\in\mathbb{R}^m$，正则化参数 $\lambda>0$。

首先设 $B=A^{\mathrm{T}}A=(b_1\ b_2\ \cdots\ b_n)^{\mathrm{T}}$，$C=A^{\mathrm{T}}b=(c_1\ c_2\ \cdots\ c_n)^{\mathrm{T}}$，再令 $f(x)=\frac{1}{2}\|Ax-b\|^2$，$h(x)=\lambda\|x\|_1$，$g(x)=\lambda\|x\|$，$f_i(x)=\frac{1}{2}e_i^{\mathrm{T}}xb_ix-e_i^{\mathrm{T}}xc_i+\frac{1}{2n}b^{\mathrm{T}}b$，其中 e_i 表示第 i 列为 1 其余元素皆为 0 的向量，于是有 $f(x)=\sum_{i=1}^{n} f_i(x)$，从而上述优化问题可表述为模型(1.1)的格式。

2) 二分类问题[4]

二分类问题在机器学习等领域中，主要用于只有两种可能结果的决策场景，其通过建立一个模型并根据输入变量来预测输出变量的类别。该类问题在特征选择，医学影像分析等领域有着广泛应用，具体表述为如下优化模型：

$$\min_{\omega\in\mathbb{R}^n} F(\omega) := \frac{1}{n}\sum_{i=1}^{n} l\left(a_i^{\mathrm{T}}\omega, b_i\right) + \lambda\|\omega\|_1,$$

其中 $\{(a_i,b_i)\}_{i=1}^n \subset \mathbb{R}^n\times\{-1,1\}^n$ 是给定的训练数据集，$\lambda>0$ 为正则化参数，$l(\cdot,\cdot)$ 是非凸光滑损失函数。具体的，令 $f_i(x)=\frac{1}{n}\cdot l\left(a_i^{\mathrm{T}}\omega,b_i\right)$，$h(x)=\lambda\|\omega\|_1$，$g(x)=0$，则上述优化问题可转化为模型(1.1)的格式。

1.1. 研究现状

DC 问题的一个经典求解算法是 DCA (Difference of Convex Functions Algorithm)，由 Pham Dinh 在 1985 年首次提出[5]，自此学术界对 DC 问题的研究已接近 40 年。DCA 主要用于求解标准形式的 DC 规划问题，即目标函数为两个凸函数之差的形式，DC 问题的具体格式为

$$\min F(x) = f(x) - P(x).$$

具体的，DCA 的迭代格式为

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n}\left\{f(x) - P(x_k) - \langle y_k, x - x_k\rangle\right\},$$

其中 $y_k\in\partial P(x_k)$。DCA 的优势在于框架简单、收敛速度快、具有全局收敛性且能够处理大规模问题，但在处理一些 DC 问题时，有可能会遇到子问题比较复杂，没有显式解的情形，因此需要内部求解器对子问题进行求解。例如，针对如下形式的 DC 问题

$$\min F(x) = f(x) + h(x) - g(x),$$

若采用 DCA 对其进行求解，则其子问题将表述为

$$x_{k+1}\in\arg\min_{x\in\mathbb{R}^n}\left\{f(x) + h(x) - g(x_k) - \langle\xi^k, x\rangle\right\},$$

其中 $\xi^k\in\partial g(x_k)$。由于子问题涉及到对函数 $f(x)+h(x)$ 的邻近算子的求解，这导致该子问题通常没有显式解，即便函数 $f(x)$ 和 $h(x)$ 的邻近算子是易于求解的。

针对 DCA 中子问题不易求解的情形，许多学者进一步将 DCA 与其他算法相结合，提出了一系列相关算法，如马尔科夫链随机 DCA (Markov chain stochastic DCA) [6]、求解球面上四次极小化问题的 DCA-Newton (A DCA-Newton method for quartic minimization over the sphere) [7]和邻近 DCA (proximal Difference of Convex Functions Algorithm, pDCA) [8]，其中 pDCA 的迭代格式如下所示：

$$\begin{aligned} x_{k+1} &= \arg\min_{x\in\mathbb{R}^n}\left\{h(x) + \langle\nabla f(x_k) - \xi^k, x\rangle + \frac{L}{2}\|x - x_k\|^2\right\} \\ &= \arg\min_{x\in\mathbb{R}^n}\left\{h(x) + \frac{L}{2}\left\|x - \left(x_k - \frac{1}{L}\left(\nabla f(x_k) - \xi^k\right)\right)\right\|^2\right\}. \end{aligned}$$

其中 $\xi^k\in\partial g(x_k)$，且由于 $h(x)$ 是适当闭凸函数，x_{k+1} 是唯一定义的。pDCA 在每一迭代步不仅需要对目标函数中的凹函数部分 $-g(x)$ 做线性优化，还需对目标函数中的光滑凸函数部分 $f(x)$ 做线性化。此时，如果函数 $h(x)$ 的邻近算子便于计算，pDCA 的子问题就很容易计算[8]。尽管 pDCA 中的子问题是易于求解的，但该算法的收敛速度较慢，这是因为当 $g(x)=0$ 时，pDCA 便退化为邻近梯度算法，而邻近梯度算

法只能达到$O(1/k)$的次线性收敛率，收敛速度较慢。许多学者利用外推技术[9]-[11]实现了 pDCA 的加速，具体地，Wen 等人[11]采用外推技术对 pDCA 进行加速，提出了带外推的邻近 DCA (proximal difference-of-convex algorithm with extrapolation, pDCAe)，该算法的具体迭代格式为

$$x_{k+1}=\arg\min_{y\in\mathbb{R}^n}\left\{h(y)+\left\langle\nabla f(y_k)-\xi^k, y\right\rangle+\frac{L}{2}\|y-y_k\|^2\right\},$$

其中$\xi^k\in\partial g(x_k)$, $y_k=x_k+\beta_k(x_k-x_{k-1})$, $\{\beta_k\}\subseteq[0,1)$且$\sup_k\beta_k<1$。在适当的假设条件下，pDCAe 的全局收敛性以及$O(1/k^2)$的收敛率已被证明[12]。

针对有限和形式的 DC 问题(1.1)，pDCA 在每一迭代步都需要对目标函数光滑部分的梯度$\nabla f_i(x)(i=1,2,\cdots,n)$进行计算，这将导致算法的计算相当昂贵。为降低计算成本，许多学者提出了一系列随机梯度算法，如随机梯度下降算法(stochastic gradient descent, SGD) [13]和小批量梯度下降算法(mini-batch gradient descent, Mini-batch) [14]，但由于方差的引入，这类算法为保证收敛性，其迭代步长需要随着迭代的进程衰减至 0，这导致算法的收敛速度变慢。为提高随机梯度算法的收敛速度，许多学者提出了一系列方差缩减的随机梯度估计方法，如随机方差缩减梯度算法(stochastic variance reduced gradient, SVRG) [15]、增量梯度算法(incremental gradient method, SAGA) [16]和随机递归梯度算法(stochastic recursive gradient, SARAH) [17]。特别的，SARAH 结合了 SVRG 和 SAGA 的思想，该算法与 SVRG 在内循环的更新方式不同，且相较于 SAGA 不需要存储旧的梯度信息，从而极大降低了存储成本[18]。这一随机梯度的方差随着迭代的进程不断衰减到 0，因此迭代步长可以取一个常数，这将使得算法的收敛速度较快。具体的，这类随机梯度算法与确定型一阶算法的收敛速度一致，在强凸情形下可达到期望意义下的线性收敛速度，在凸的情形下可以达到期望意义下的$O(1/k)$次线性收敛率。

1.2. 本文贡献

针对有限和形式的 DC 问题(1.1)，本文将方差缩减的随机梯度 SARAH 引入到 pDCA 中去，提出了一种基于随机梯度 SARAH 的随机邻近 DC 算法(pDCA-SARAH)，以降低计算成本。在非凸情形下，本文对算法的收敛性进行了理论分析，并通过数值实验说明了算法的高效性。具体的，本文的主要贡献包括：

在算法上，针对有限和形式的 DC 问题(1.1)，由于 pDCA 算法需在每次迭代过程中逐个计算$f_i(x)$的全梯度，从而当数据量 n 很大时，计算有限和部分的全梯度将会非常昂贵。为降低计算成本，本文将随机梯度 SARAH 引入到 pDCA 中，提出了基于 SARAH 的随机邻近 DC 算法(pDCA-SARAH)。该算法在每一轮外循环时首先计算一次全梯度，在每一轮内循环中随机抽取小批量的数据来计算随机梯度 SARAH，并用随机梯度 SARAH 来近似全梯度，以实现算法计算成本的降低。

在理论上，本文对 pDCA-SARAH 算法的收敛性及收敛率进行了分析，并在非凸情形下，详细地给出目标函数在期望意义下的下降量分析和次线性收敛率分析。

在数值上，本文通过将 pDCA-SARAH 算法应用于l_{1-2}正则化最小二乘问题的求解中，并通过与传统的 pDCA 进行比较，验证了本文所提算法理论上的收敛性和高效性。

1.3. 本文框架

本文框架如下，在第二节中，详细介绍了本文所涉及到的符号、定义以及相关引理。第三节中，本文提出了基于 SARAH 的随机邻近 DC 算法(pDCA-SARAH)，以解决传统的 pDCA 在处理形如问题(1.1)时所面临的计算成本昂贵的问题，并给出了 pDCA-SARAH 算法在期望意义下的下降量分析以及收

敛性分析。第四节中，将 pDCA-SARAH 算法应用于求解 l_{1-2} 正则化最小二乘问题来进行数值实验，并与 pDCA 进行对比，从数值上验证了本文所提出的 pDCA-SARAH 算法的高效性。第五节中，对本文做出总结。

2. 预备知识

为便于下文研究 pDCA-SARAH 算法的收敛性，本节将详细介绍本文所涉及到的符号、定义及相关引理。下面首先对本文出现的符号做出定义：记 $\mathbb{R}^n$ 为 n 维欧几里得空间；$\langle x,y\rangle$ 表示向量内积，其中 $x,y\in\mathbb{R}^n$；$\|x\|=\sqrt{\langle x,x\rangle}$ 为欧式距离；$E[\cdot]$ 为数学期望。

定义 2.1 (凸函数) [19]设函数 f 为适当函数，如果 $\mathrm{dom}\,f$ 是凸集，且

$$f\left(\theta x+(1-\theta)y\right)\le\theta f(x)+(1-\theta)f(y),$$

对所有的 $x,y\in\mathrm{dom}\,f,0\le\theta\le1$ 都成立，则称 f 是凸函数。

定义 2.2 (凸函数的次微分) [19] 设 $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ 为适当下半连续凸函数，定义域为 $\mathrm{dom}\,f:=\left\{x\in\mathbb{R}^n\middle|f(x)<+\infty\right\}$，$x$ 为定义域 $\mathrm{dom}\,f$ 中的一点。若向量 $u\in\mathbb{R}^n$ 满足

$$f(y)\ge f(x)+u^{\mathrm{T}}(y-x),\ \forall y\in\mathrm{dom}\,f,$$

则称 u 为函数 f 在点 x 处的一个次梯度。进一步地，函数 f 在点 x 处的次微分记作 $\partial f(x)$，定义为所有满足上述条件的向量 u 所构成的集合，即

$$\partial f(x)=\left\{u\middle|u\in\mathbb{R}^n,f(y)\ge f(x)+u^{\mathrm{T}}(y-x),\ \forall y\in\mathrm{dom}\,f\right\}.$$

此外，若 $x\notin\mathrm{dom}\,f$，则定义 f 在该点的次微分 $\partial f(x)=\varnothing$。

定义 2.3 (梯度利普希兹连续) [20]对于可微函数 f，若对任意的 $x,y\in\mathrm{dom}\,f$，存在常数 $L>0$，使其满足不等式

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|,$$

则称函数 f 是梯度利普希兹连续的，简记为 L-光滑，其中 L 为 Lipschitz 常数。

引理 2.1 [20]对定义在凸集上的可微函数 f，f 是凸函数的充要条件为

$$f(y)\ge f(x)+\nabla f(x)^{\mathrm{T}}(y-x),\ \forall x,y\in\mathrm{dom}\,f.$$

引理 2.2 (下降引理) [21]若 $f:\mathbb{R}^n\to\mathbb{R}$ 连续可微且 L-光滑$(L>0)$，则 f 满足如下不等式：

$$f(y)\le f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|x-y\|^2,\ \forall x,y\in\mathbb{R}^n.$$

3. 算法及收敛性分析

3.1. 基于 SARAH 的随机邻近 DC 算法

本文主要研究的是一类具有有限和形式的 DC 问题，该问题的具体格式如下所示：

$$\min_x G(x)=\sum_{i=1}^{n}f_i(x)+h(x)-g(x),$$

其中 $f_i(x)\ (i=1,2,\cdots,n)$ 为光滑凸函数，且 $\nabla f_i(x)$ 是 Lipschitz 连续的，此外 $h(x)$ 是一个连续凸函数，$g(x)$ 是一个适当的闭凸函数。求解该类问题的一个经典算法是 pDCA，但由于 pDCA 每一迭代步都需要计算 $f_i(x)\ (i=1,2,\cdots,n)$ 的全梯度，因此针对大规模 DC 问题，若采用 pDCA 则计算成本较为昂贵，故本文在 pDCA 中引入随机梯度 SARAH 来降低计算成本。下面给出基于 SARAH 的随机邻近 DC 算法(pDCA-SARAH)的迭代格式。

基于 SARAH 的随机邻近 DC 算法(pDCA-SARAH)

for $s=0,1,\cdots,S-1$ do

$x_{s+1}^{0}=x_{s}^{\mathrm{T}}$

$V_{s+1}^{0}=\sum_{i=1}^{n}\nabla f_i\left(\hat{x}_s\right)$

for $t=0,1,\cdots,T-1$ do

随机选取小批量 $I_b\subset\{1,2,\cdots,n\}$，$|I_b|=b$

$V_{s+1}^{t}=\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)+V_{s+1}^{t-1}$

$x_{s+1}^{t+1}=\arg\min_{x}\left\{h(x)+\left\langle V_{s+1}^{t}-\xi_{s+1}^{t},\ x-x_{s+1}^{t}\right\rangle+\frac{1}{2\alpha}\left\|x-x_{s+1}^{t}\right\|^2\right\}$

end

$\hat{x}_{s+1}=x_{s+1}^{\mathrm{T}}$

end

输出：x_α 从 $\left\{\left\{x_{s+1}^{t+1}\right\}_{t=0}^{T}\right\}_{s=0}^{S-1}$ 随机均匀选择

注：在算法中，第 s 轮外循环输出的 x 更新值为 x_s^{T}，并将其赋值给第 $s+1$ 轮外循环中 x 的初始值，即 $x_{s+1}^{0}=x_s^{\mathrm{T}}$。然后在第 $s+1$ 轮外循环中进行 T 轮内循环，在其中第 t 轮内循环，由 $\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)+V_{s+1}^{t-1}$ 得到随机梯度 V_{s+1}^{t}，以此来近似全梯度并更新 x_{s+1}^{t}。

3.2. 算法收敛性及收敛率分析

本节将对 pDCA-SARAH 算法的收敛性以及收敛率进行分析。为了便于分析 pDCA-SARAH 算法的收敛性，当选取全梯度进行计算时，本文记 pDCA-SARAH 算法的子问题对应的精确解为

$$\bar{x}_{s+1}^{t+1}=\arg\min_{x}\left\{h(x)+\left\langle\nabla F\left(x_{s+1}^{t}\right)-\zeta_{s+1}^{t},\ x-x_{s+1}^{t}\right\rangle+\frac{1}{2\alpha}\left\|x-x_{s+1}^{t}\right\|^2\right\}.$$

此外，本文需要对优化问题的目标函数做出如下假设：

假设 1：有限和部分的 $f_i(x)\ (i=1,2,\cdots,n)$ 为光滑凸函数；

假设 2：$\nabla f_i(x)\ (i=1,2,\cdots,n)$ 是 Lipschitz 连续的；

假设 3：$h(x)$ 是一个连续凸函数，$g(x)$ 是一个适当的闭凸函数。

在进行收敛分析之前，首先需要证明引理 3.1 与 3.2。具体地，引理 3.1 给出了目标函数在前后迭代点处，期望意义下的下降量分析。

引理 3.1 (目标函数期望意义下的下降量)当假设 1~3 成立时，令 $\left\{x_s^t\right\}$ 是由 pDCA-SARAH 算法所产生的迭代序列，则有如下所示不等式成立：

$$E\left[G\left(x_{s+1}^{t+1}\right)\right]\le E\left[G\left(x_{s+1}^{t}\right)\right]+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)\cdot E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]$$
$$+\frac{1}{2\theta}E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]-\left(\frac{1}{2\alpha}-\theta\right)E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right],$$

其中 $\alpha>0$ 为迭代步长，$\theta>0$ 为常数。

证明。一方面，令 $F(x)=\sum_{i=1}^{n}f_i(x)$，由于 $\nabla f_i(x)\ (i=1,2,\cdots,n)$ 是 Lipschitz 连续的，于是根据引理 2.2

(下降引理)可得不等式(3.1)成立：

$$F\left(x_{s+1}^{t+1}\right)\le F\left(x_{s+1}^{t}\right)+\left\langle\nabla F\left(x_{s+1}^{t}\right),\, x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\frac{L}{2}\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}. \tag{3.1}$$

另一方面，因为函数 $h(x)$ 为凸函数，从而根据引理 2.1 可得，对 $\forall x, y\in \operatorname{dom} h$ 有 $h(y)\ge h(x)+\langle\partial h(x),\, y-x\rangle$，所以可以推导出如下两不等式，即

$$h\left(x_{s+1}^{t+1}\right)\le h\left(\overline{x}_{s+1}^{t+1}\right)+\left\langle\xi_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle,\quad \forall\xi_{s+1}^{t+1}\in\partial h\left(x_{s+1}^{t+1}\right), \tag{3.2}$$

和

$$h\left(\overline{x}_{s+1}^{t+1}\right)\le h\left(x_{s+1}^{t}\right)+\left\langle\overline{\xi}_{s+1}^{t+1},\, \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle,\quad \forall\overline{\xi}_{s+1}^{t+1}\in\partial h\left(\overline{x}_{s+1}^{t+1}\right). \tag{3.3}$$

下面对不等式(3.2)、(3.3)进一步处理。首先针对不等式(3.2)，根据算法的最优性条件可得

$$0\in\partial h\left(x_{s+1}^{t+1}\right)+V_{s+1}^{t}-\zeta_{s+1}^{t}+\frac{1}{\alpha}\left(x_{s+1}^{t+1}-x_{s+1}^{t}\right),$$

因此可对不等式(3.2)中的 $\left\langle\xi_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle$ 项做如下变换(其中 $\xi_{s+1}^{t+1}\in\partial h\left(x_{s+1}^{t+1}\right)$)：

$$\begin{aligned}
&\left\langle\xi_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&=\left\langle-V_{s+1}^{t}+\zeta_{s+1}^{t}-\frac{1}{\alpha}\left(x_{s+1}^{t+1}-x_{s+1}^{t}\right),\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&=\left\langle-V_{s+1}^{t}+\zeta_{s+1}^{t},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle-\frac{1}{\alpha}\left\langle x_{s+1}^{t+1}-x_{s+1}^{t},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&=\left\langle-V_{s+1}^{t}+\zeta_{s+1}^{t},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\frac{1}{\alpha}\left\langle x_{s+1}^{t}-x_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle.
\end{aligned}$$

接着对上述变换结果做进一步的化简，由于

$$\langle a-b,\, c-d\rangle=\frac{1}{2}\left(\|a-d\|^{2}-\|a-c\|^{2}+\|b-c\|^{2}-\|b-d\|^{2}\right),$$

从而其中的 $\left\langle x_{s+1}^{t}-x_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle$ 项可以转化为

$$\left\langle x_{s+1}^{t}-x_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle=\frac{1}{2}\left(\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right),$$

再将该式代入上述变换结果，可得

$$\begin{aligned}
\left\langle\xi_{s+1}^{t+1},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle=&\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right)\\
&+\left\langle-V_{s+1}^{t}+\zeta_{s+1}^{t},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle.
\end{aligned} \tag{3.4}$$

类似地，对于不等式(3.3)，由算法的最优性条件可得

$$0\in\partial h\left(\overline{x}_{s+1}^{t+1}\right)+\nabla F\left(x_{s+1}^{t}\right)-\zeta_{s+1}^{t}+\frac{1}{\alpha}\left(\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right),$$

从而可化简不等式(3.3)得到

$$\left\langle\partial h\left(\overline{x}_{s+1}^{t+1}\right),\, \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle=\left\langle-\nabla F\left(x_{s+1}^{t}\right)+\zeta_{s+1}^{t},\, \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle-\frac{1}{\alpha}\left\|\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}. \tag{3.5}$$

下面分别将(3.4)式代入(3.2)式，(3.5)式代入(3.3)式，则有如下两个不等式成立：

$$h\left(x_{s+1}^{t+1}\right)\le h\left(\overline{x}_{s+1}^{t+1}\right)+\left\langle-V_{s+1}^{t}+\zeta_{s+1}^{t},\, x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right);$$

$$h\left(\overline{x}_{s+1}^{t+1}\right)\le h\left(x_{s+1}^{t}\right)+\left\langle -\nabla F\left(x_{s+1}^{t}\right)+\zeta_{s+1}^{t},\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle-\frac{1}{\alpha}\left\|\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}.$$

对上面两不等式求和并整理，具体过程为

$$\begin{aligned}
h\left(x_{s+1}^{t+1}\right)&\le h\left(x_{s+1}^{t}\right)+\left\langle -V_{s+1}^{t}+\zeta_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right)+\zeta_{s+1}^{t},\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&\quad+\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}-\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right)-\frac{1}{\alpha}\left\|\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}\\
&=h\left(x_{s+1}^{t}\right)+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&\quad+\left\langle \zeta_{s+1}^{t},\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right)\\
&=h\left(x_{s+1}^{t}\right)+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&\quad-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right),
\end{aligned}$$

从而可以得到不等式(3.6)成立：

$$\begin{aligned}
h\left(x_{s+1}^{t+1}\right)&\le h\left(x_{s+1}^{t}\right)-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right)\\
&\quad+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle.
\end{aligned}\tag{3.6}$$

此外，又因为$g\left(x\right)$是凸函数，所以根据引理 2.1 可以得到

$$g\left(x_{s+1}^{t}\right)+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\le g\left(x_{s+1}^{t+1}\right).\tag{3.7}$$

接着再对(3.1)式、(3.6)式与(3.7)式求和，得到不等式(3.8)如下所示：

$$\begin{aligned}
&F\left(x_{s+1}^{t+1}\right)+h\left(x_{s+1}^{t+1}\right)+g\left(x_{s+1}^{t}\right)+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&\le F\left(x_{s+1}^{t}\right)+\left\langle \nabla F\left(x_{s+1}^{t}\right),\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+h\left(x_{s+1}^{t}\right)+\frac{L}{2}\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}\\
&\quad+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle \zeta_{s+1}^{t},\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&\quad-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right)+g\left(x_{s+1}^{t+1}\right),
\end{aligned}\tag{3.8}$$

化简(3.8)式中$\left\langle \nabla F\left(x_{s+1}^{t}\right),\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle$，具体为

$$\begin{aligned}
&\left\langle \nabla F\left(x_{s+1}^{t}\right),\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle+\left\langle -\nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle\\
&=\left\langle \nabla F\left(x_{s+1}^{t}\right),\ x_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle-\left\langle \nabla F\left(x_{s+1}^{t}\right),\ \overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\rangle+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&=\left\langle \nabla F\left(x_{s+1}^{t}\right),\ x_{s+1}^{t+1}-x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}+x_{s+1}^{t}\right\rangle+\left\langle -V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&=\left\langle \nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle,
\end{aligned}$$

设$G\left(x\right)=F\left(x\right)+h\left(x\right)-g\left(x\right)$，将上述化简结果代入不等式(3.8)，则可以得到

$$\begin{aligned}
G\left(x_{s+1}^{t+1}\right)&\le G\left(x_{s+1}^{t}\right)+\frac{L}{2}\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^{2}+\left\langle \nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\\
&\quad-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^{2}\right),
\end{aligned}$$

其中α是步长，则有$\alpha>0$，故$\frac{1}{2\alpha}\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^{2}>0$，于是放缩该不等式可以得到

$$G\left(x_{s+1}^{t+1}\right)\le G\left(x_{s+1}^{t}\right)+\frac{L}{2}\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2+\left\langle\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle$$
$$-\frac{1}{2\alpha}\left(\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2+\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^2\right).$$

根据上述不等式，可得 $G\left(x_{s+1}^{t+1}\right)$ 的期望满足如下不等式：

$$\begin{aligned}E\left[G\left(x_{s+1}^{t+1}\right)\right]\le E\left[G\left(x_{s+1}^{t}\right)\right]+E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\right]\\+\left(\frac{L}{2}-\frac{1}{2\alpha}\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]-\frac{1}{2\alpha}E\left[\left\|\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right].\end{aligned}\tag{3.9}$$

下面处理不等式(3.9)，首先由 $\langle a,b\rangle\le\frac{L}{2}\|a\|^2+\frac{1}{2L}\|b\|^2$ 与 $\langle a,b\rangle=\langle b,a\rangle$ 可得

$$\left\langle\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle=\frac{\theta}{2}\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^2+\frac{1}{2\theta}\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2.\tag{3.10}$$

进一步化简(3.10)式，针对其中的项 $\frac{\theta}{2}\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^2$，根据范数的三角不等式，即 $\|a+b\|\le\|a\|+\|b\|$，$\forall a,b\in\mathbb{R}^n$，可以得到

$$\|a+b\|^2\le\left(\|a\|+\|b\|\right)^2=\|a\|^2+2\|a\|\cdot\|b\|+\|b\|^2\le2\|a\|^2+2\|b\|^2,$$

于是有

$$\frac{\theta}{2}\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\|^2=\frac{\theta}{2}\left\|x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}+x_{s+1}^{t}-x_{s+1}^{t}\right\|^2\le\theta\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2+\theta\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^2.$$

再将该不等式代入(3.10)式，可以得到

$$\left\langle\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t},\ x_{s+1}^{t+1}-\overline{x}_{s+1}^{t+1}\right\rangle\le\theta\left\|x_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2+\theta\left\|x_{s+1}^{t}-\overline{x}_{s+1}^{t+1}\right\|^2+\frac{1}{2\theta}\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2.$$

将上述不等式代入不等式(3.9)，则可得不等式(3.11)，即

$$\begin{aligned}E\left[G\left(x_{s+1}^{t+1}\right)\right]\le E\left[G\left(x_{s+1}^{t}\right)\right]+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)\cdot E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]\\+\frac{1}{2\theta}E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]-\left(\frac{1}{2\alpha}-\theta\right)E\left[\left\|\overline{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right],\end{aligned}\tag{3.11}$$

引理 3.1 证毕。

为了证明 pDCA-SARAH 算法的收敛性，我们还需要对该算法中采用的随机梯度的方差进行估计，具体见引理 3.2。

引理 3.2 当假设 1-3 成立时，令 $\left\{x_s^t\right\}$ 是由 pDCA-SARAH 算法所产生的迭代序列，则有

$$E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]\le E\left[\left\|\nabla F\left(x_{s+1}^{0}\right)-V_{s+1}^{0}\right\|^2\right]+\frac{L^2}{b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right],$$

其中 b 是随机抽取的小批量数据集的长度，$L>0$ 为 Lipschitz 常数。

证明。 首先对(3.11)式中 $E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]$ 做进一步处理，具体过程为

$$\begin{aligned}&E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]\\&=E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)+\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}+V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&= E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\|^{2}\right]+E\left[\left\|\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\|^{2}\right]+E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^{2}\right]\\
&\quad+2E[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle+\left\langle\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1},\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\\
&\quad+\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle].
\end{aligned}
$$

从而可得(3.12)式，即

$$
\begin{aligned}
&E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^{2}\right]\\
&= E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\|^{2}\right]+E\left[\left\|\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\|^{2}\right]+E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^{2}\right]\\
&\quad+2E[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle+\left\langle\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1},\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\\
&\quad+\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle].
\end{aligned}
\tag{3.12}
$$

对(3.12)式的项：$2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle+\left\langle\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1},\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\right]$，根据 $V_{s+1}^{t-1}-V_{s+1}^{t}=\nabla F\left(x_{s+1}^{t-1}\right)-\nabla F\left(x_{s+1}^{t}\right)$ 与 $\langle a,b\rangle=\langle b,a\rangle$，可将该项化简为

$$
\begin{aligned}
&2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle+\left\langle\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1},\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\right]\\
&=2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle-\left\langle\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1},\ \nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\rangle\right]\\
&=2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle-\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\rangle\right]\\
&=0.
\end{aligned}
$$

类似的，化简等式(3.12)中的另一项 $2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\right]$，具体过程为

$$
\begin{aligned}
&2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ V_{s+1}^{t-1}-V_{s+1}^{t}\right\rangle\right]\\
&=2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t-1}\right)-\nabla F\left(x_{s+1}^{t}\right)\right\rangle\right]\\
&=-2E\left[\left\langle\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right),\ \nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\rangle\right]\\
&=-2E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\|^{2}\right].
\end{aligned}
$$

再将上述两项的化简结果代入(3.12)式，整理后可得如下不等式：

$$
E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^{2}\right]\le E\left[\left\|\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\|^{2}\right]+E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^{2}\right]-E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\|^{2}\right],
$$

其中 $-E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-\nabla F\left(x_{s+1}^{t-1}\right)\right\|^{2}\right]<0$，则可以由放缩法得到不等式(3.13)成立，即

$$
E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^{2}\right]\le E\left[\left\|\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\|^{2}\right]+E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^{2}\right].
\tag{3.13}
$$

此外，根据 pDCA-SARAH 算法梯度更新公式 $V_{s+1}^{t}=\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)+V_{s+1}^{t-1}$ 可得

$$
V_{s+1}^{t-1}-V_{s+1}^{t}=V_{s+1}^{t-1}-\left(\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)+V_{s+1}^{t-1}\right)=-\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right),
$$

则有 $\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^{2}=\left\|\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)\right\|^{2}$，又因 $\nabla f_i\left(x_{s+1}^{t}\right)$ 是 Lipschitz 连续的，故有

$$E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^2\right]=E\left[\left\|\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)\right\|^2\right]\le\frac{L^2}{b}E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t-1}\right\|^2\right].$$

再将不等式 $E\left[\left\|V_{s+1}^{t-1}-V_{s+1}^{t}\right\|^2\right]\le\frac{L^2}{b}E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t-1}\right\|^2\right]$ 代入(3.13)式，于是有

$$E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]\le E\left[\left\|\nabla F\left(x_{s+1}^{t-1}\right)-V_{s+1}^{t-1}\right\|^2\right]+\frac{L^2}{b}E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t-1}\right\|^2\right],$$

进一步的，以此类推，可以得到不等式

$$E\left[\left\|\nabla F\left(x_{s+1}^{t}\right)-V_{s+1}^{t}\right\|^2\right]\le E\left[\left\|\nabla F\left(x_{s+1}^{0}\right)-V_{s+1}^{0}\right\|^2\right]+\frac{L^2}{b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right].\tag{3.14}$$

引理 3.2 证毕。

基于引理 3.1 以及引理 3.2，本文给出了 pDCA-SARAH 算法的次线性收敛率结果。为便于衡量 pDCA-SARAH 算法的收敛速度，我们定义 DC 问题的广义邻近梯度映射为 $G_\eta\left(x_{s+1}^{t}\right)=\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}$。下面，我们给出 pDCA-SARAH 算法的次线性收敛率结果。

定理 3.1 (次线性收敛率)当假设 1~3 成立时，令 $\left\{x_s^t\right\}$ 是由 pDCA-SARAH 算法所产生的迭代序列，则有如下所示不等式成立：

$$E\left[\left\|G_\eta\left(\tilde{x}_{s+1}\right)\right\|^2\right]=\frac{1}{k}\sum_{s=0}^{S-1}\sum_{t=0}^{T-1}E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]\le\frac{2\alpha}{k\left(1-2\alpha\theta\right)}E\left[G\left(\tilde{x}_0\right)-G^*\right],$$

其中 $k=ST$，表示总迭代次数，$\alpha>0$ 为迭代步长，$\theta>0$ 为常数。

证明。将不等式(3.14)代入不等式(3.11)，则有

$$\begin{aligned}E\left[G\left(x_{s+1}^{t+1}\right)\right]\le{}&E\left[G\left(x_{s+1}^{t}\right)\right]+\frac{1}{2\theta}E\left[\left\|\nabla F\left(x_{s+1}^{0}\right)-V_{s+1}^{0}\right\|^2\right]+\frac{L^2}{2\theta b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]\\&+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]-\left(\frac{1}{2\alpha}-\theta\right)E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right].\end{aligned}$$

因为 pDCA-SARAH 算法的梯度更新公式为 $V_{s+1}^{t}=\frac{1}{b}\sum_{i\in I_b}\left(\nabla f_i\left(x_{s+1}^{t}\right)-\nabla f_i\left(x_{s+1}^{t-1}\right)\right)+V_{s+1}^{t-1}$，且该算法的第 $s+1$ 轮外循环中初始点的全梯度记作 V_{s+1}^{0}，即 $\nabla F\left(x_{s+1}^{0}\right)=V_{s+1}^{0}$，于是有 $\left\|\nabla F\left(x_{s+1}^{0}\right)-V_{s+1}^{0}\right\|^2=0$ 成立，因此可以更新上述不等式得到

$$\begin{aligned}E\left[G\left(x_{s+1}^{t+1}\right)\right]\le{}&E\left[G\left(x_{s+1}^{t}\right)\right]+\frac{L^2}{2\theta b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]\\&+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]-\left(\frac{1}{2\alpha}-\theta\right)E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right].\end{aligned}\tag{3.15}$$

下面针对不等式(3.15)中的项：$\frac{L^2}{2\theta b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]$，进行 $t=0,1,\cdots,T-1$ 的累和，得到的结果如下所示：

$$\begin{aligned}&\sum_{t=0}^{T-1}\left\{\frac{L^2}{2\theta b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]\right\}\\&=\left(\frac{TL^2}{2\theta b}+\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)\cdot\left(\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]\right)+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{T-1}-x_{s+1}^{T}\right\|^2\right].\end{aligned}$$

只要满足 $\frac{TL^2}{2\theta b}+\frac{L}{2}-\frac{1}{2\alpha}+\theta\le 0,\ \frac{L}{2}-\frac{1}{2\alpha}+\theta<0$，则有如下不等式成立：

$$\frac{L^2}{2\theta b}\sum_{i=0}^{t-1}E\left[\left\|x_{s+1}^{i+1}-x_{s+1}^{i}\right\|^2\right]+\left(\frac{L}{2}-\frac{1}{2\alpha}+\theta\right)E\left[\left\|x_{s+1}^{t}-x_{s+1}^{t+1}\right\|^2\right]<0,$$

从而可以化简不等式(3.15)得到

$$\left(\frac{1}{2\alpha}-\theta\right)E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]\le E\left[G\left(x_{s+1}^{t}\right)\right]-E\left[G\left(x_{s+1}^{t+1}\right)\right].$$

类似的，对上述不等式进行 $t=0,1,\cdots,T-1$ 的累和，可以得到不等式(3.16)。即

$$\left(\frac{1}{2\alpha}-\theta\right)\sum_{t=0}^{T-1}E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]\le E\left[G\left(x_{s+1}^{0}\right)\right]-E\left[G\left(x_{s+1}^{T}\right)\right].\tag{3.16}$$

接着再对不等式(3.16)进行 $s=0,1,\cdots,S-1$ 的累和，其中右式的累和结果为

$$E\left[G\left(\tilde{x}_0\right)-G^*\right]=E\left[G\left(x_1^0\right)\right]-E\left[G\left(x_1^T\right)\right]+E\left[G\left(x_2^0\right)\right]-E\left[G\left(x_2^T\right)\right]+\cdots+E\left[G\left(x_S^0\right)\right]-E\left[G\left(x_S^T\right)\right],$$

将其记作 $E\left[G\left(\tilde{x}_0\right)-G^*\right]$，因此不等式(3.16)的累和结果为

$$\left(\frac{1}{2\alpha}-\theta\right)\sum_{s=0}^{S-1}\sum_{t=0}^{T-1}E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]\le E\left[G\left(\tilde{x}_0\right)-G^*\right].$$

令 $k=ST$，$E\left[\left\|G_\eta\left(\tilde{x}_{s+1}\right)\right\|^2\right]=\frac{1}{k}\sum_{s=0}^{S-1}\sum_{t=0}^{T-1}E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]$，结合上述累和结果可得

$$E\left[\left\|G_\eta\left(\tilde{x}_{s+1}\right)\right\|^2\right]=\frac{1}{k}\sum_{s=0}^{S-1}\sum_{t=0}^{T-1}E\left[\left\|\bar{x}_{s+1}^{t+1}-x_{s+1}^{t}\right\|^2\right]\le\frac{2\alpha}{k\left(1-2\alpha\theta\right)}E\left[G\left(\tilde{x}_0\right)-G^*\right].$$

定理 3.1 证毕，说明 pDCA-SARAH 算法在期望意义下是收敛的。

4. 数值实验

在本节中，我们将本文所提出的 pDCA-SARAH 算法用于求解 l_{1-2} 正则化最小二乘问题，并将该算法与 pDCA 进行比较，说明本文所提算法的高效性。本节所有的实验均在 64 位 PC 上使用 Matlab 2016a 进行的，该 PC 配备有 2.7 GHz 和 8 G RAM。

本节主要考虑如下所示的 l_{1-2} 正则化最小二乘问题：

$$\min_{x\in\mathbb{R}^n}\ F_{1-2}\left(x\right):=\frac{1}{2}\left\|Ax-b\right\|^2+\lambda\left(\left\|x\right\|_1-\left\|x\right\|\right),\tag{4.1}$$

其中 $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$ 且正则化参数 $\lambda>0$。为将问题(4.1)转化成本文所研究的具有有限和形式的 DC 问题(1.1)，首先令 $f\left(x\right)=\frac{1}{2}\left\|Ax-b\right\|^2$，$h\left(x\right)=\lambda\left\|x\right\|_1$，$g\left(x\right)=\lambda\left\|x\right\|$，通过对 $f\left(x\right)$ 简单的整理，便可以得到

$$f\left(x\right)=\frac{1}{2}\left\|Ax-b\right\|^2=\frac{1}{2}\left(Ax-b\right)^{\mathrm{T}}\left(Ax-b\right)=\frac{1}{2}\left(x^{\mathrm{T}}A^{\mathrm{T}}Ax-2x^{\mathrm{T}}A^{\mathrm{T}}b+b^{\mathrm{T}}b\right)$$

进一步，我们令 $B=A^{\mathrm{T}}A=\left(b_1\ b_2\ \cdots\ b_n\right)^{\mathrm{T}}$，$C=A^{\mathrm{T}}b=\left(c_1\ c_2\ \cdots\ c_n\right)^{\mathrm{T}}$，将其代入上式，可得

$$f\left(x\right)=\frac{1}{2}\left(x^TBx-2x^TC+b^Tb\right),$$

其中 $B\in\mathbb{R}^{n\times n}$, $C\in\mathbb{R}^n$，$b^{\mathrm{T}}b$ 为一常数。

进一步的，令 $b_i\ \left(i=1,2,\cdots,n\right)$ 是由矩阵 B 第 i 行构成的行向量，$c_i\ \left(i=1,2,\cdots,n\right)$ 为矩阵 C 的第 i 行

对应元素，e_i $(i=1,2,\cdots,n)$ 为第 i 列为 1 且其余元素皆为 0 的行向量，并设

$$f_i(x)=\frac{1}{2}e_i^{\mathrm{T}}xb_ix-e_i^{\mathrm{T}}xc_i+\frac{1}{2n}b^{\mathrm{T}}b,$$

于是有 $f(x)=\sum_{i=1}^{n}f_i(x)$ 成立，从而问题(4.1)可以转化为具有有限和形式的 DC 问题(1.1)。

为了衡量算法的收敛程度，下面给出相对误差[22]的定义为

$$e(k):=\frac{F_{1-2}(x^k)-F_{1-2}^{\min}}{F_{1-2}(x^0)-F_{1-2}^{\min}},$$

其中 $F_{1-2}(x^k)$ 与 $F_{1-2}(x^0)$ 分别为目标函数 $F_{1-2}(x)$ 在 x^k 与 x^0 处的函数值，$F_{1-2}^{\min}$ 为最小函数值。接着令 $T(k)$ 为得到 x^k 的总计算时间，从而定义每个时间步长下的最小相对误差为

$$E(t):=\min\left\{e(k)\middle|k\in\left\{i\middle|T(i)\le t\right\}\right\}.$$

当最小相对误差 $E(t)$ 越接近 0 时，目标函数 $F_{1-2}(x)$ 越接近最优解。

针对数值实验的参数设置，首先随机生成矩阵 $A\in\mathbb{R}^{m\times n}$，向量 $b\in\mathbb{R}^m$，并控制变量，依次选取正则化参数 $\lambda=0.5$ 与 $\lambda=0.3$，用于比较不同正则化参数条件下 pDCA-SARAH 算法和 pDCA 的数值效果。针对问题规模设置，本节考虑如下四个三元数组：

$$(n,\ m,\ s)=(3000,\ 900,\ 180);\ (n,\ m,\ s)=(5000,\ 1500,\ 300);$$

$$(n,\ m,\ s)=(8000,\ 2400,\ 480);\ (n,\ m,\ s)=(10000,\ 3000,\ 600).$$

对以上四个数组 $(n,\ m,\ s)$ 分别生成 10 组独立实验并计算出每组的平均最小相对误差，以此来比较 pDCA 与 pDCA-SARAH 算法的数值效果。此外，我们统一规定两算法的外循环最大迭代次数为 10,000，并根据问题规模 $(n,\ m,\ s)$ 分别设置最大迭代时间为 100 秒、300 秒、700 秒和 1000 秒。其次设置 pDCA-SARAH 算法内循环迭代次数为 2，每轮内循环随机抽取的小批量数据个数为 3。

针对不同的正则化参数 λ 与问题规模 $(n,\ m,\ s)$，分别将最小相对误差 $E(t)=10^{-6}$ 时，pDCA 与 pDCA-SARAH 算法迭代所用时长记录在表 1 与表 2 中。从表中不难看出，当正则化参数 λ 确定时，对本文所规定的四组不同的问题规模 $(n,\ m,\ s)$，就两算法的最小相对误差 $E(t)$ 达到相同阈值 1×10^{-6} 所需的迭代时间而言，pDCA-SARAH 算法均比 pDCA 短，且问题规模越大，pDCA-SARAH 算法的优势便越显著，这说明了 pDCA-SARAH 算法在处理大规模 DC 问题时的优越性。

Table 1. When $\lambda=0.5$, $E(t)=10^{-6}$, comparison between two algorithms with respect to iteration times in problem (4.1)

表 1. $\lambda=0.5$, $E(t)=10^{-6}$ 时，问题(4.1)中两算法迭代时间对比

案例	问题规模			迭代时间(秒)	
	n	m	s	pDCA	pDCA-SARAH
1	3000	900	180	99.02	89.12
2	5000	1500	300	255.32	232.79
3	8000	2400	480	664.42	622.51
4	10,000	3000	600	982.70	836.98

Table 2. When $\lambda = 0.3$, $E(t) = 10^{-6}$, comparison between two algorithms with respect to iteration times in problem (4.1)
表 2. $\lambda = 0.3$, $E(t) = 10^{-6}$ 时，问题(4.1)中两算法迭代时间对比

案例	问题规模			迭代时间(秒)	
	n	m	s	pDCA	pDCA-SARAH
1	3000	900	180	85.88	79.27
2	5000	1500	300	267.69	227.04
3	8000	2400	480	672.10	606.27
4	10000	3000	600	907.02	837.88

此外，针对不同问题规模(n, m, s)与正则化参数λ，我们还在图 1 和图 2 中给出了 pDCA 与 pDCA-SARAH 算法的迭代收敛结果。具体的，图 1 给出了正则化参数$\lambda = 0.5$时，pDCA 与 pDCA-SARAH 算法最小相对误差$E(t)$的下降比较，从图中可知 pDCA-SARAH 算法的收敛速度比 pDCA 更快。类似的，针对正则化参数$\lambda = 0.3$的情形，由图 2 亦可以得出上述结论。因此，在处理大规模 DC 问题时，就达到相同最小相对误差$E(t)$所需时间而言，pDCA-SARAH 算法要优于 pDCA。

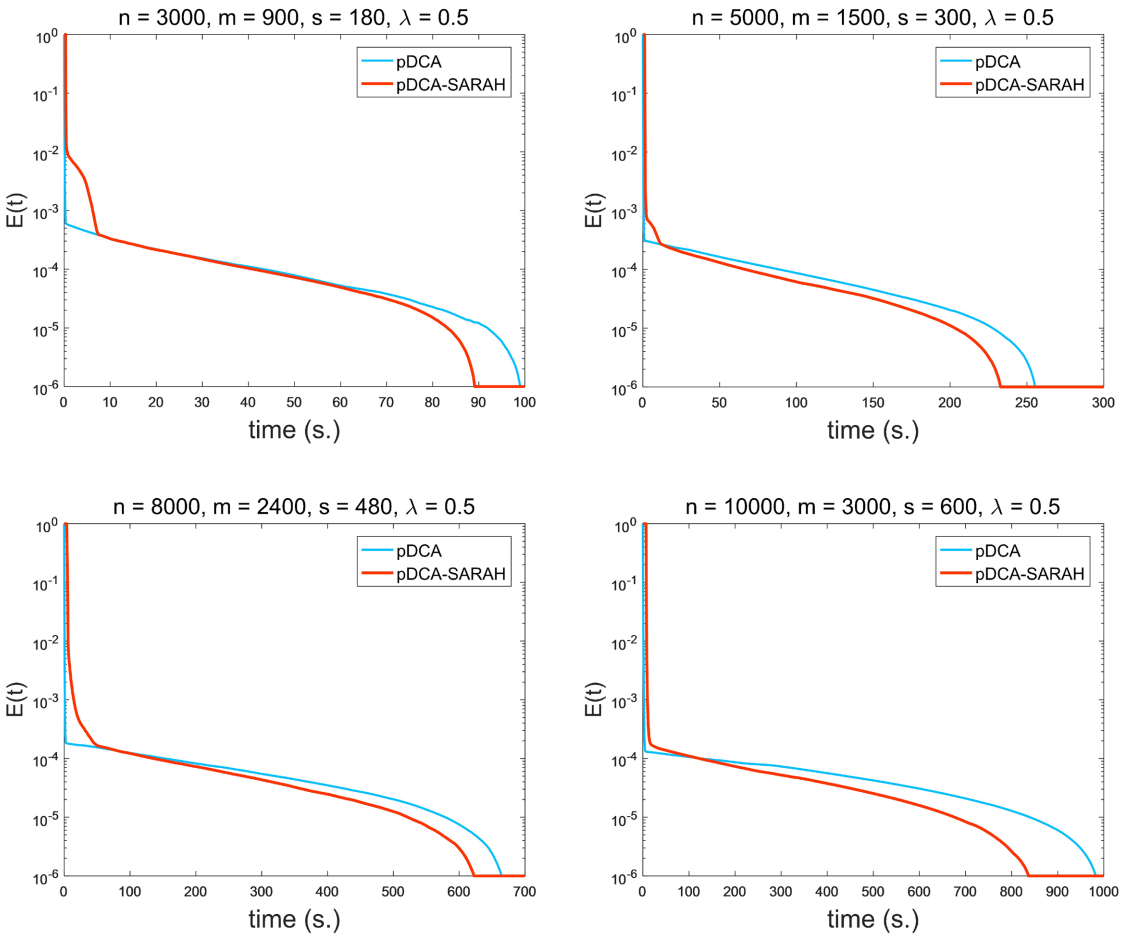

Figure 1. When $\lambda = 0.5$, comparison of relative error reduction $E(t)$ in pDCA and pDCA-SARAH algorithms
图 1. $\lambda = 0.5$时，pDCA 与 pDCA-SARAH 算法最小相对误差 $E(t)$下降比较

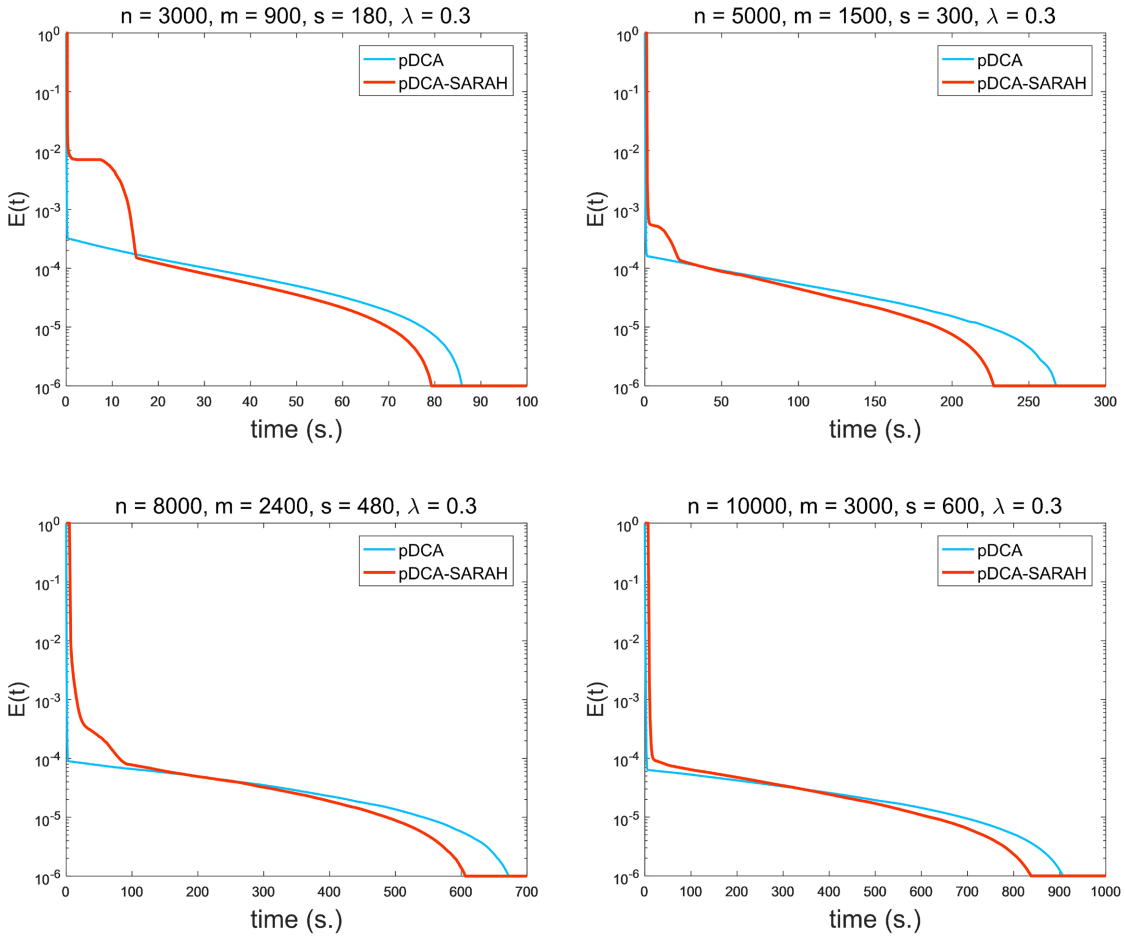

Figure 2. When $\lambda = 0.3$, comparison of relative error reduction $E(t)$ in pDCA and pDCA-SARAH algorithms
图 2. $\lambda = 0.3$ 时，pDCA 与 pDCA-SARAH 算法最小相对误差 $E(t)$下降比较

5. 总结

本文将随机梯度 SARAH 与 pDCA 相结合，提出基于 SARAH 的随机邻近 DC 算法(pDCA-SARAH)，并将其应用于求解一类具有有限和形式的 DC 问题。

首先，本文提出 pDCA-SARAH 算法，该算法通过在内循环中抽取小批量的数据来计算随机梯度，并以此来近似全梯度，以实现计算成本的降低。其次，本文对 pDCA-SARAH 算法的收敛性及收敛率进行了分析，并在非凸情形下，详细地给出了目标函数在期望意义下的下降量分析和次线性收敛率分析。最后，通过数值实验，验证了 pDCA-SARAH 算法在处理大规模 DC 问题时的高效性。

参考文献

[1] Le Thi, H.A., Le, H.M. and Pham Dinh, T. (2014) Feature Selection in Machine Learning: An Exact Penalty Approach Using a Difference of Convex Function Algorithm. *Machine Learning*, **101**, 163-186. https://doi.org/10.1007/s10994-014-5455-y

[2] Yin, P., Lou, Y., He, Q. and Xin, J. (2015) Minimization of ℓ_{1-2} for Compressed Sensing. *SIAM Journal on Scientific Computing*, **37**, A536-A563. https://doi.org/10.1137/140952363

[3] Le Thi, H.A., Le, H.M., Phan, D.N. and Tran, B. (2020) Stochastic DCA for Minimizing a Large Sum of DC Functions with Application to Multi-Class Logistic Regression. *Neural Networks*, **132**, 220-231. https://doi.org/10.1016/j.neunet.2020.08.024

[4] Pham, N.H., Nguyen, L.M., *et al*. (2020) ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Non-Convex Optimization. *Journal of Machine Learning Research*, **21**, 4455-4502.

[5] Dinh, T.P. (1986) Methods of Subgradients. North-Holland Mathematics Studies.

[6] Luu, H.P.H., Le, H.M. and Le Thi, H.A. (2024) Markov Chain Stochastic DCA and Applications in Deep Learning with PDEs Regularization. *Neural Networks*, **170**, 149-166. https://doi.org/10.1016/j.neunet.2023.11.032

[7] Hu, S. and Yan, Z. (2024) Quadratic Growth and Linear Convergence of a DCA Method for Quartic Minimization over the Sphere. *Journal of Optimization Theory and Applications*, **201**, 378-395. https://doi.org/10.1007/s10957-024-02401-w

[8] Gotoh, J., Takeda, A. and Tono, K. (2017) DC Formulations and Algorithms for Sparse Optimization Problems. *Mathematical Programming*, **169**, 141-176. https://doi.org/10.1007/s10107-017-1181-0

[9] Nesterov, Y. (2004) Introductory Lectures on Convex Optimization: a Basic Course. Kluwer Academic Publishers.

[10] Polyak, B.T. (1964) Some Methods of Speeding up the Convergence of Iteration Methods. *USSR Computational Mathematics and Mathematical Physics*, **4**, 1-17. https://doi.org/10.1016/0041-5553(64)90137-5

[11] Wen, B., Chen, X. and Pong, T.K. (2017) A Proximal Difference-Of-Convex Algorithm with Extrapolation. *Computational Optimization and Applications*, **69**, 297-324. https://doi.org/10.1007/s10589-017-9954-1

[12] Gao, L. and Wen, B. (2022) Convergence Rate Analysis of an Extrapolated Proximal Difference-of-Convex Algorithm. *Journal of Applied Mathematics and Computing*, **69**, 1403-1429. https://doi.org/10.1007/s12190-022-01797-w

[13] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**, 400-407. https://doi.org/10.1214/aoms/1177729586

[14] Li, M., Zhang, T., Chen, Y. and Smola, A.J. (2014) Efficient Mini-Batch Training for Stochastic Optimization. *Proceedings of the* 20*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 661-670. https://doi.org/10.1145/2623330.2623612

[15] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. *News in Physiological Sciences*, **1**, 315-323.

[16] Defazio, A., Bach, F. and Lacoste-Julien, S. (2014) SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. In: Ghahramani, Z., Ed., *Advances in Neural Information Processing Systems*, MIT Press, 1646-1654.

[17] Nguyen, L.M., Liu, J., Scheinberg, K., *et al*. (2017) SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 2613-2621.

[18] 董萍. 求解大规模优化问题的 Gauss-Seidd 型惯性邻近交替线性极小化算法[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2023.

[19] 刘浩洋. 最优化计算方法[M]. 北京: 高等教育出版社, 2020.

[20] 刘浩洋. 最优化: 建模、算法与理论[M]. 北京: 高等教育出版社, 2020.

[21] Nesterov, Y., *et al*. (2018) Lectures on Convex Optimization: Volume 137. Springer.

[22] Liu, T. and Takeda, A. (2022) An Inexact Successive Quadratic Approximation Method for a Class of Difference-of-Convex Optimization Problems. *Computational Optimization and Applications*, **82**, 141-173. https://doi.org/10.1007/s10589-022-00357-z