

求解非凸复合优化问题的随机 Douglas-Rachford分裂算法

李登辉, 孟士博, 陈梦婷

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10月8日; 录用日期: 2024年11月26日; 发布日期: 2024年12月5日

摘 要

本文研究了大规模非凸复合优化问题, 该问题的目标函数为非凸的光滑函数与凸(可能是非光滑的)函数之和, 其中非凸的光滑函数具有有限和形式。随着大规模非凸复合优化问题在机器学习、图像处理中广泛应用, 设计简洁、高效、计算成本低的优化算法来解决大规模非凸复合优化问题已成为当今研究热点之一。求解非凸复合优化问题的一个经典算法是Douglas-Rachford (DR) 分裂算法, 然而当实际问题规模较大时经典的DR分裂算法计算成本较大。据此本文提出了一种方差缩减的随机DR分裂算法(SDR), 将方差缩减的随机梯度引入到DR分裂算法中, 以减小算法的计算成本, 并基于Kurdyka-Łojasiewicz (KL) 框架对算法的收敛性进行了分析。具体的, 我们首先建立了Lyapunov函数的下降性, 然后建立了DR价值函数的相对误差界条件, 最后利用KL性质证明了算法的全局收敛性。此外, 通过将SDR算法用于求解 $\ell_1\ell_2$ 稀疏正则化Logistic回归问题, 并与PG、Prox-SARAH进行对比, 展示了本文所提出算法的优越性。

关键词

非凸复合优化问题, 方差缩减的随机梯度算法, Kurdyka-Łojasiewicz不等式,
Douglas-Rachford分裂算法

Stochastic Douglas-Rachford Splitting Algorithm for Solving Non-Convex Composite Optimization Problems

Denghui Li, Shibo Meng, Mengting Chen

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

Received: Oct. 8th, 2024; accepted: Nov. 26th, 2024; published: Dec. 5th, 2024

文章引用: 李登辉, 孟士博, 陈梦婷. 求解非凸复合优化问题的随机 Douglas-Rachford 分裂算法[J]. 运筹与模糊学, 2024, 14(6): 116-133. DOI: 10.12677/orf.2024.146516

Abstract

This paper investigates large-scale non-convex composite optimization problems, where the objective function is the sum of a nonconvex smooth function and a convex (possibly nonsmooth) function, with the nonconvex smooth function having a finite sum form. Given the widespread application of large-scale nonconvex composite optimization problems in machine learning and image processing, developing concise, efficient, and low-computational-cost optimization algorithms for solving such problems has become one of the current research hotspots. A classic algorithm for solving non-convex composite optimization problems is the Douglas-Rachford (DR) splitting algorithm. However, the classical DR splitting algorithm incurs significant computational costs when dealing with large-scale practical problems. Accordingly, this paper proposes a variance-reduced stochastic DR splitting algorithm (SDR), incorporating variance reduction stochastic gradients into the DR splitting algorithm to reduce computational costs, and analyzes the convergence of the algorithm based on the Kurdyka-Łojasiewicz (KL) framework. Moreover, we use the SDR algorithm to solve the $\ell_1\ell_2$ sparse logistic regression problem and compare its performance with PG and Prox-SARAH to show the efficiency of SDR.

Keywords

Non-Convex Composite Optimization Problem, Variance-Reduced Stochastic Gradient Algorithm, Kurdyka-Łojasiewicz Inequality, Douglas-Rachford Splitting Algorithm

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,随着大数据时代的到来,涌现出了大量的大规模非凸复合优化问题,其广泛应用于机器学习,图像处理,稀疏优化等领域因此,针对大规模非凸复合优化问题,设计更加高效、简洁、低计算成本的求解算法有着一定的研究意义和应用价值。本文,我们考虑如下形式的大规模非凸复合优化问题

$$\min_{x \in \mathbb{R}^d} F(x) = f(x) + h(x) = \sum_{i=1}^N f_i(x) + h(x). \quad (1)$$

其中 $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$ ($i=1,2,\dots,N$) 为连续可微的(可能非凸), $h: \mathbb{R}^d \rightarrow (-\infty, +\infty]$ 为适当下半连续凸函数(可能非光滑)。该复合优化问题有着广泛的应用背景,具体的在机器学习中的二分类 $\ell_1\ell_2$ -正则化逻辑回归问题,其一般形式为

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i a_i^T x)) + \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|^2,$$

其中 $a_i \in \mathbb{R}^d$, $b_i \in \{-1, 1\}$, $i \in \{1, 2, \dots, N\}$ 为样本的特征及相应的标签, $\lambda_1, \lambda_2 > 0$ 为正则化参数,用以控制所求变量的稀疏性。

1.1. 研究现状

针对复合优化问题(1)的研究已有很多,一个经典的求解算法为邻近梯度算法(proximal gradient, PG),也称向前向后分裂算法[1]:

$$x^{k+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ h(x) + f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha_k} \|x - x^k\|^2 \right\}.$$

该算法的优点在于充分利用了目标函数光滑部分和非光滑部分的结构, Beck 等人研究了在凸或非凸条件下, PG 算法在迭代点意义下的收敛性[2]。

为了解决目标函数两部分均不光滑的情况, Douglas 和 Rachford 提出了 Douglas-Rachford (DR) 分裂算法[3]。虽然在凸的情形下, DR 分裂算法[3]已经被充分的研究, 但针对问题非凸的情形, 相关的理论分析还不是很多。尽管如此, DR 分裂算法已经非常成功地应用于各种非凸问题[4]。针对非凸复合优化问题, 2015 年, Bot 等人[5]将惯性近似法和广义 DR 分裂算法相结合, 提出惯性 DR 分裂算法, 其相比于经典 DR 算法具有更快的收敛速度, 更好的适应性, 还可以避免局部最优解。2016 年, Li 等人[6]建立了 DR 分裂算法在非凸情形下的下降性, 并在其满足 Kurdyka-Łojasiewicz (KL) 性质时证明了算法的全局收敛性。

随着科技迅速发展, 实际应用中许多问题的规模通常很大, 即问题(1)中 N 很大的情形。针对这类问题, 许多经典的求解算法由于在内循环需要涉及到对光滑部分全梯度的计算, 导致了计算成本很高。为降低算法的计算成本, 针对光滑优化问题, 随机梯度下降算法(SGD) [7]和小批量梯度下降算法(Mini-batch) [8]被提出, 其在每一迭代步用一个样本的梯度或小批量样本的梯度近似全梯度的计算。然而, 由于方差的引入, 算法中的步长参数需要随迭代步不断衰减到 0 以保证其收敛性, 这导致了算法较慢的收敛速度。为提高随机算法的收敛速度, 一系列方差缩减的随机梯度被提出, 如 SAG [9]、SVRG [10]、SASG [11]、SCSG [12]、SARAH [13]、SPIDER [14]等。这一系列随机梯度的技巧被广泛的应用于许多一阶算法, 用以实现相关算法计算量的降低。针对本文所关注的大规模非凸复合优化问题, 相关的求解算法已有一些, 如邻近 SGD 算法(Proximal SGD)等[15] [16]。此外, 方差缩减的随机梯度技术也被用以和邻近梯度算法相结合, 如 Prox-SARAH [17]等。

1.2. 本文主要内容及结构

本文将方差缩减的随机梯度引入到 DR 分裂算法中, 针对大规模非凸复合优化问题提出随机 DR 分裂算法, 以降低经典 DR 分裂算法的计算成本, 更好的求解大规模优化问题。本文在 Kurdyka-Łojasiewicz (KL) 框架下, 给出算法的收敛性分析。具体的, 我们首先建立了 Lyapunov 函数在前后迭代步的下降性, 并建立 DR 价值函数的相对误差界条件, 最后利用 KL 性质证明了算法的全局收敛性。

本文框架如下, 在第二节中, 我们将给出本文所涉及到的符号、定义以及相关引理。第三节中, 我们提出了求解大规模非凸复合优化问题的随机 DR 分裂算法, 并在 KL 框架给出了算法的收敛性证明。为了说明所提出算法的优越性, 我们在第四节给出了相关的数值实验。在第五节, 对本文的工作做出总结。

2. 预备知识

2.1. 基本概念

本节将介绍讨论算法的收敛性所涉及的常用符号、基本定义及性质。首先对本论文所使用的符号做出定义: 记 $\mathbb{R}^d$ 为 d 维欧氏空间, $\|x\| = \sqrt{\langle x, x \rangle}$ 为欧氏距离, $\|x\|_1 = \sum_{i=1}^d |x_i|$ 为 1 范数, 对于一个非空闭集 $\Omega \subseteq \mathbb{R}^d$, x 到集合 Ω 的欧氏距离可被定义(2.1.1)为

$$\text{dist}(x, \Omega) := \inf_{y \in \Omega} \|x - y\|, \quad \forall x \in \mathbb{R}^d, \quad \forall \Omega \in \mathbb{R}^d.$$

定义 2.1.2 $\langle u, v \rangle$ 表示向量内积, 即

$$\langle u, v \rangle = \frac{1}{2} (\|u + v\|^2 - \|u\|^2 - \|v\|^2).$$

定义 2.1.3 (L -光滑[18] [19]) 我们称函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 是 L -光滑的, 若

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^d.$$

定义 2.1.4 (凸函数的次微分[19]) 设 $f: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ 是适当的下半连续函数, 对 $x \in \text{dom } f$, f 在点 x 处的次微分记作 $\partial f(x)$, 定义为所有满足下述条件的 $u \in \mathbb{R}^d$ 构成的集合

$$f(y) \geq f(x) + \langle u, y - x \rangle, \forall y \in \mathbb{R}^d.$$

若 $x \notin \text{dom } f$, 定义 f 在该点的次微分 $\partial f(x) = \emptyset$.

定义 2.1.5 (DR 价值函数) 令 $\gamma > 0$, 定义 DR 价值函数如下

$$D_\gamma(x, z, y) = f(x) + h(z) - \frac{1}{2\gamma}\|x - z\|^2 + \frac{1}{\gamma}\langle y - x, z - x \rangle. \quad (2)$$

再根据定义 2.1.2, 令 $u = y - x$, $v = z - x$, 函数 $D_\gamma(x, z, y)$ 可以等价表示为下面公式

$$D_\gamma(x, z, y) = f(x) + h(z) + \frac{1}{2\gamma}\|2x - z - y\|^2 - \frac{1}{2\gamma}\|y - x\|^2 - \frac{1}{\gamma}\|x - z\|^2, \quad (3)$$

再结合 $\langle a - b, c - d \rangle = \frac{1}{2}\|a - d\|^2 - \frac{1}{2}\|a - c\|^2 + \frac{1}{2}\|b - c\|^2 - \frac{1}{2}\|b - d\|^2$, 函数 $D_\gamma(x, z, y)$ 可以被表示成以下形式

$$D_\gamma(x, z, y) = f(x) + h(z) + \frac{1}{2\gamma}\|y - x\|^2 - \frac{1}{2\gamma}\|y - z\|^2. \quad (4)$$

2.2. 基本结论

引理 2.2.1 (下降引理[19]) 若函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 连续可微且 L -光滑 ($L > 0$), 则有

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

Kurdyka-Łojasiewicz (KL)不等式是证明非凸复合优化问题收敛性的一个重要性质, 满足这一性质的函数有很多, 如凸函数、幂函数、实解析函数等. 记 $p: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ 为正常的下半连续函数. 对 $-\infty < \eta_1 < \eta_2 \leq +\infty$, 定义

$$[\eta_1 < p < \eta_2] := \{x \in \mathbb{R}^d : \eta_1 < p(x) < \eta_2\}.$$

下面, 我们给出 KL 不等式的具体定义。

定义 2.2.1 [20] 若存在 $\eta \in (0, +\infty]$, $\bar{x} \in \text{dom } \partial f$ 的邻域 U 以及连续的凹函数 $\varphi: [0, \eta] \rightarrow \mathbb{R}^+$ 满足

- (i) $\varphi(0) = 0$;
- (iii) φ 在 $(0, \eta)$ 上是连续函数;
- (iii) 对任意的 $s \in (0, \eta)$, 有 $\varphi' > 0$;
- (iiii) 对任意的 $x \in U \cap [p(\bar{x}) < p < p(\bar{x} + \eta)]$, Kurdyka-Łojasiewicz 不等式成立, 即有

$$\varphi'(p(x) - p(\bar{x})) \text{dist}(0, \partial f(x)) \geq 1.$$

此时我们称函数 p 在点 $\bar{x}$ 处具有 Kurdyka-Łojasiewicz(KL)性质。

3. 算法及算法的收敛分析

本文将方差缩减的随机梯度算法引入到经典的 DR 分裂算法中, 提出了一种随机 DR 分裂算法, 该算法可以降低经典 DR 分裂算法的计算成本, 更高效的针对于大规模非凸复合优化问题的求解。

在本节中，我们首先给出求解大规模非凸复合优化问题的随机 DR 算法的迭代格式，然后在 KL 框架下，给出算法的收敛性分析。具体来说，我们首先建立了 Lyapunov 函数在前后迭代步的下降性质，并建立 DR 价值函数的相对误差界条件，最后基于函数之间的关系，我们利用 KL 性质证明了算法的全局收敛性。

3.1. 随机 DR 分裂算法

首先我们给出随机 DR 分裂算法的具体格式。

算法 1 求解非凸复合优化问题的随机 Douglas-Rachford 分裂算法

输入： $x^0, y^0 \in \mathbb{R}^d$ ， $\gamma \in (0, 1)$ 。

1: **For** $k = 0, 1, 2, \dots, K-1$ **do**

2: 令 n_k 均匀随机从 1 到 N 中取值，再随机从梯度 $\nabla f_i(x_k) (i = 1, 2, \dots, N)$ 中选取 n_k 个求平均值并赋值给 g_k ，

3: 计算 x^{k+1}

$$x^{k+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ \langle g_k, x - x^k \rangle + \frac{1}{2\gamma} \|x - y^k\|^2 \right\}, \quad (5)$$

4: 计算 z^{k+1}

$$\begin{aligned} z^{k+1} &= \arg \min_{z \in \mathbb{R}^d} \left\{ h(z) + \frac{1}{2\gamma} \|z - (2x^{k+1} - y^k)\|^2 \right\} \\ &= \arg \min_{z \in \mathbb{R}^d} \left\{ h(z) + \frac{1}{2\gamma} \|z - x^{k+1}\|^2 + \langle z - x^{k+1}, g_k \rangle \right\}, \end{aligned} \quad (6)$$

5: 计算 y^{k+1}

$$y^{k+1} = y^k + z^{k+1} - x^{k+1}, \quad (7)$$

6: **End for**

根据算法中 x^{k+1}, z^{k+1} 的迭代子问题，我们可以得到相应的最优性条件：

$$0 = g_k + \frac{1}{\gamma} (x^{k+1} - y^k). \quad (8)$$

和

$$0 \in \partial h(z^{k+1}) + \frac{1}{\gamma} (z^{k+1} - (2x^{k+1} - y^k)). \quad (9)$$

3.2. 收敛性分析

为分析算法的收敛性，我们首先给出一些假设条件。

假设 1 (1) $f_i(x): \mathbb{R}^d \rightarrow \mathbb{R} (i = 1, 2, \dots, N)$ 是 Lipschitz 连续可微的，Lipschitz 常数为 $L_i > 0$ ，且记 $L = \max_{i \in \{1, 2, \dots, N\}} \{L_i\}$ 。

(2) $h: \mathbb{R}^d \rightarrow \mathbb{R}$ 是适当下半连续凸函数，且在其定义域上是连续的。

(3) 函数 G 是强制的。

假设 2 (梯度方差缩减) 存在 $\rho \in (0, 1]$ ， $A_i \geq 0 (i = 1, 2, 3)$ 和随机向量 $\{\alpha_k^i\}_{i=1}^s$ 使得以下条件成立：

(1) 梯度估计的均方误差是有界的

$$\mathbb{E}_k \|g_k - \nabla f(x^k)\|^2 \leq \sigma_k^2 + A_1 \mathbb{E}_k [\|x^{k+1} - x^k\|^2], \quad (10)$$

$$\mathbb{E}_k \left\| g_k - \nabla f(x^k) \right\| \leq \xi_k + A_2 \mathbb{E}_k \left[\|x^{k+1} - x^k\| \right]. \quad (11)$$

其中 $\sigma_k^2 = \sum_{i=1}^S \|\alpha_k^i\|^2$, $\xi_k = \sum_{i=1}^S \|\alpha_k^i\|$.

(2) 序列 $\{\sigma_k^2\}_{k=0}^\infty$ 在几何上是衰减的, 满足

$$\mathbb{E}[\sigma_{k+1}^2] \leq (1-\rho)\sigma_k^2 + A_3 \mathbb{E}[\|x^{k+1} - x^k\|^2]. \quad (12)$$

(3) 对于任意序列 $\{x_k\}_{k=0}^\infty$ 满足 $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k - x_{k-1}\|^2] = 0$, 有下式成立

$$\sum_{k=1}^\infty \mathbb{E}[\sigma_k^2] < \infty, \sum_{k=1}^\infty \mathbb{E}[\xi_k] < \infty. \quad (13)$$

基于上述算法分析以及相关假设, 我们在本节将对随机 DR 算法的收敛性进行分析. 考虑到非凸问题的收敛性分析大部分是基于 Kurdyka-Łojasiewicz (KL) 框架给出的, 本文也采用这一框架对该算法进行收敛性分析. 具体的, 收敛性框架需要证明以下三个条件成立, 包括充分下降性条件、相对误差条件和函数连续性条件. 本文我们首先给出了 Lyapunov 函数在前后迭代步的充分下降性(引理 3.2), 并建立 DR 价值函数的相对误差界条件(引理 3.3), 最后本文利用 KL 性质给出了算法的全局收敛性(定理 3.1).

设 $D_\gamma(x^k, z^k, y^k)$, $D_\gamma(x^{k+1}, z^{k+1}, y^{k+1})$ 分别为函数 D_γ 第 k 次和第 $k+1$ 次的迭代步, 在下述引理中, 我们将给出该函数前后迭代步的关系.

引理 3.1 若假设 1 与假设 2 成立, $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ 是由随机 DR 算法生成的序列. 对于任意 $k \in \mathbb{N}$, 我们有

$$\begin{aligned} & \mathbb{E}[D_\gamma(x^{k+1}, z^{k+1}, y^{k+1}) - D_\gamma(x^k, z^k, y^k)] \\ & \leq \left(\frac{1}{2\theta} + 4 + 3\gamma \right) \left[\sigma_k^2 + A_1 \left(\mathbb{E}[\|x^{k+1} - x^k\|^2] \right) \right] + (4 + 3\gamma) \left[\sigma_{k-1}^2 + A_1 \left(\mathbb{E}[\|x^k - x^{k-1}\|^2] \right) \right] \\ & \quad + (4L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\theta}{2} - 4 \right) \|x^{k+1} - x^k\|^2. \end{aligned}$$

证明: 针对 $D_\gamma(x^{k+1}, z^{k+1}, y^{k+1}) - D_\gamma(x^k, z^k, y^k)$, 做如下处理

$$\begin{aligned} & D_\gamma(x^{k+1}, z^{k+1}, y^{k+1}) - D_\gamma(x^k, z^k, y^k) \\ & = (D_\gamma(x^{k+1}, z^k, y^k) - D_\gamma(x^k, z^k, y^k)) + (D_\gamma(x^{k+1}, z^{k+1}, y^k) - D_\gamma(x^{k+1}, z^k, y^k)) \\ & \quad + (D_\gamma(x^{k+1}, z^{k+1}, y^{k+1}) - D_\gamma(x^{k+1}, z^{k+1}, y^k)). \end{aligned}$$

由公式(4)可得:

$$\begin{aligned} & D_\gamma(x^{k+1}, z^k, y^k) - D_\gamma(x^k, z^k, y^k) \\ & = f(x^{k+1}) + \frac{1}{2\gamma} \|y^k - x^{k+1}\|^2 - \left[f(x^k) + \frac{1}{2\gamma} \|y^k - x^k\|^2 \right]. \end{aligned} \quad (14)$$

由 $f_i (i \in \{1, 2, \dots, N\})$ 的梯度 Lipschitz 连续性和引理 2.2.1 可知

$$f(x^{k+1}) - f(x^k) \leq \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2. \quad (15)$$

结合公式(2)~(4), (14), (15)可以计算化简得出

$$\begin{aligned}
& D_{\gamma}(x^{k+1}, z^k, y^k) - D_{\gamma}(x^k, z^k, y^k) \\
& \leq \frac{1}{2\gamma} \|y^k - x^{k+1}\|^2 - \frac{1}{2\gamma} \|y^k - x^{k+1} + x^{k+1} - x^k\|^2 + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\
& = -\frac{1}{\gamma} \langle y^k - x^{k+1}, x^{k+1} - x^k \rangle - \frac{1}{2\gamma} \|x^{k+1} - x^k\|^2 + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2,
\end{aligned}$$

其中等式是由 $\|a+b\|^2 = \|a\|^2 + 2\langle a, b \rangle + \|b\|^2$ 得到的。又由(8)可以得到以下不等式

$$\begin{aligned}
& D_{\gamma}(x^{k+1}, z^k, y^k) - D_{\gamma}(x^k, z^k, y^k) \\
& \leq \langle \nabla f(x^k) - g_k, x^{k+1} - x^k \rangle - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{k+1} - x^k\|^2.
\end{aligned} \tag{16}$$

利用 $\langle a, b \rangle \leq \frac{1}{2\beta} \|a\|^2 + \frac{\beta}{2} \|b\|^2$ ($\beta > 0$) 可得到

$$\begin{aligned}
& D_{\gamma}(x^{k+1}, z^k, y^k) - D_{\gamma}(x^k, z^k, y^k) \\
& \leq \frac{1}{2\beta} \|\nabla f(x^k) - g_k\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} \right) \|x^{k+1} - x^k\|^2.
\end{aligned} \tag{17}$$

接下来处理 z 的前后代步关系。首先根据公式(3)可得

$$\begin{aligned}
& D_{\gamma}(x^{k+1}, z^{k+1}, y^k) - D_{\gamma}(x^{k+1}, z^k, y^k) \\
& = h(z^{k+1}) + \frac{1}{2\gamma} \|z^{k+1} - (2x^{k+1} - y^k)\|^2 - \frac{1}{\gamma} \|x^{k+1} - z^{k+1}\|^2 \\
& \quad - \left[h(z^k) + \frac{1}{2\gamma} \|z^k - (2x^{k+1} - y^k)\|^2 - \frac{1}{\gamma} \|x^{k+1} - z^k\|^2 \right],
\end{aligned}$$

进一步结合迭代步(6)和(7)，有下式成立

$$D_{\gamma}(x^{k+1}, z^{k+1}, y^k) - D_{\gamma}(x^{k+1}, z^k, y^k) \leq -\frac{1}{\gamma} \|y^{k+1} - y^k\|^2 + \frac{1}{\gamma} \|x^{k+1} - z^k\|^2. \tag{18}$$

下面处理 $\frac{1}{\gamma} \|x^{k+1} - z^k\|^2$ 这一项。由(8)可得

$$\begin{cases} 0 = g_k + \frac{1}{\gamma}(x^{k+1} - y^k), \\ 0 = g_{k-1} + \frac{1}{\gamma}(x^k - y^{k-1}). \end{cases}$$

因此有

$$y^k - y^{k-1} = x^{k+1} - x^k + \gamma(g_k - g_{k-1}) \tag{19}$$

所以

$$\begin{aligned}
& \|y^k - y^{k-1}\| \\
& \leq \|x^{k+1} - x^k\| + \gamma \|g_k - g_{k-1}\| \\
& = \|x^{k+1} - x^k\| + \gamma \|g_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1}) - g_{k-1}\| \\
& \leq \|x^{k+1} - x^k\| + \gamma \|g_k - \nabla f(x^k)\| + \gamma \|\nabla f(x^{k-1}) - g_{k-1}\| + \gamma L \|x^k - x^{k-1}\|.
\end{aligned} \tag{20}$$

其中最后一个不等式利用公式 $\|x+y\| \leq \|x\| + \|y\|$ 和定义 2.1.3。进而可得

$$\begin{aligned}
\|y^k - y^{k-1}\|^2 &\leq \|x^{k+1} - x^k\|^2 + \gamma^2 \|g_k - \nabla f(x^k)\|^2 + \gamma^2 \|g_{k-1} - \nabla f(x^{k-1})\|^2 \\
&\quad + \gamma^2 L^2 \|x^k - x^{k-1}\|^2 + 2\gamma \langle x^{k+1} - x^k, g_k - \nabla f(x^k) \rangle \\
&\quad + 2\gamma \langle x^{k+1} - x^k, g_{k-1} - \nabla f(x^{k-1}) \rangle + 2\gamma \langle x^{k+1} - x^k, L(x^k - x^{k-1}) \rangle \\
&\quad + 2\gamma^2 \langle g_k - \nabla f(x^k), g_{k-1} - \nabla f(x^{k-1}) \rangle \\
&\quad + 2\gamma^2 \langle g_k - \nabla f(x^k), L(x^k - x^{k-1}) \rangle \\
&\quad + 2\gamma^2 \langle g_{k-1} - \nabla f(x^{k-1}), L(x^k - x^{k-1}) \rangle.
\end{aligned}$$

进一步利用不等式 $2\langle a, b \rangle \leq \|a\|^2 + \|b\|^2$ 可得

$$\begin{aligned}
\|y^k - y^{k-1}\|^2 &\leq (1+3\gamma) \|x^{k+1} - x^k\|^2 + (\gamma+3\gamma^2) \|g_k - \nabla f(x^k)\|^2 \\
&\quad + (\gamma+3\gamma^2) \|g_{k-1} - \nabla f(x^{k-1})\|^2 + (\gamma L^2 + 3\gamma^2 L^2) \|x^k - x^{k-1}\|^2.
\end{aligned} \tag{21}$$

由不等式 $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ 可得

$$\begin{aligned}
\|g_k - g_{k-1}\|^2 &= \|g_k - \nabla f(x^k) + \nabla f(x^{k-1}) - g_{k-1} + \nabla f(x^{k-1}) - \nabla f(x^k)\|^2 \\
&\leq 3\|g_k - \nabla f(x^k)\|^2 + 3\|\nabla f(x^{k-1}) - g_{k-1}\|^2 + 3\|\nabla f(x^{k-1}) - \nabla f(x^k)\|^2.
\end{aligned} \tag{22}$$

结合公式(7), (19)~(21)可得

$$\begin{aligned}
\frac{1}{\gamma} \|x^{k+1} - z^k\|^2 &= \frac{1}{\gamma} \|x^{k+1} - x^k\|^2 + \frac{1}{\gamma} \|x^k - z^k\|^2 + \frac{2}{\gamma} \langle x^{k+1} - x^k, x^k - z^k \rangle \\
&\leq \frac{1}{\gamma} \|x^{k+1} - x^k\|^2 + \left(\frac{1}{\gamma} + 3\right) \|x^{k+1} - x^k\|^2 + (1+3\gamma) \|g_k - \nabla f(x^k)\|^2 \\
&\quad + (1+3\gamma) \|g_{k-1} - \nabla f(x^{k-1})\|^2 + (L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2 \\
&\quad + \frac{2}{\gamma} \langle x^{k+1} - x^k, x^k - x^{k+1} - \gamma(g_k - g_{k-1}) \rangle \\
&\leq \frac{1}{\gamma} \|x^{k+1} - x^k\|^2 + \left(\frac{1}{\gamma} + 3\right) \|x^{k+1} - x^k\|^2 + (1+3\gamma) \|g_k - \nabla f(x^k)\|^2 \\
&\quad + (1+3\gamma) \|g_{k-1} - \nabla f(x^{k-1})\|^2 + (L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2 - \frac{2}{\gamma} \|x^{k+1} - x^k\|^2 \\
&\quad + 2\langle x^k - x^{k+1}, g_k - g_{k-1} \rangle.
\end{aligned}$$

结合该式和不等式 $2\langle a, b \rangle \leq \|a\|^2 + \|b\|^2$, 公式(22)可得

$$\begin{aligned}
\frac{1}{\gamma} \|x^{k+1} - z^k\|^2 &\leq 4\|x^{k+1} - x^k\|^2 + (4+3\gamma) \|g_k - \nabla f(x^k)\|^2 + (4+3\gamma) \|g_{k-1} - \nabla f(x^{k-1})\|^2 \\
&\quad + (4L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2.
\end{aligned}$$

利用该式和公式(18)可推出

$$\begin{aligned}
&D_\gamma(x^{k+1}, z^{k+1}, y^k) - D_\gamma(x^{k+1}, z^k, y^k) \\
&\leq -\frac{1}{\gamma} \|y^{k+1} - y^k\|^2 + 4\|x^{k+1} - x^k\|^2 + (4+3\gamma) \|g_k - \nabla f(x^k)\|^2 \\
&\quad + (4+3\gamma) \|g_{k-1} - \nabla f(x^{k-1})\|^2 + (4L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2.
\end{aligned} \tag{23}$$

最后处理 y 的前后迭代步关系。根据(2)可得

$$D_{\gamma}(x^{k+1}, z^{k+1}, y^{k+1}) - D_{\gamma}(x^{k+1}, z^{k+1}, y^k) = \frac{1}{\gamma} \langle y^{k+1} - y^k, z^{k+1} - x^{k+1} \rangle.$$

再由公式(7)整理可得

$$D_{\gamma}(x^{k+1}, z^{k+1}, y^{k+1}) - D_{\gamma}(x^{k+1}, z^{k+1}, y^k) = \frac{1}{\gamma} \|y^{k+1} - y^k\|^2. \quad (24)$$

所以结合 x, z, y 的前后迭代步关系, 即公式(17) (23) (24)可以得证函数 D_{γ} 的前后迭代步的关系, 即

$$\begin{aligned} & D_{\gamma}(x^{k+1}, z^{k+1}, y^{k+1}) - D_{\gamma}(x^k, z^k, y^k) \\ & \leq \left(\frac{1}{2\beta} + 4 + 3\gamma \right) \|\nabla f(x^k) - g_k\|^2 + (4 + 3\gamma) \|g_{k-1} - \nabla f(x^{k-1})\|^2 \\ & \quad + (4L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} - 4 \right) \|x^{k+1} - x^k\|^2. \end{aligned} \quad (25)$$

此时由于使用了随机梯度, 所以接下来对公式(25)左右两边同时求期望, 并利用假设 2, 可得

$$\begin{aligned} & \mathbb{E} \left[D_{\gamma}(x^{k+1}, z^{k+1}, y^{k+1}) - D_{\gamma}(x^k, z^k, y^k) \right] \\ & \leq \left(\frac{1}{2\beta} + 4 + 3\gamma \right) \left[\sigma_k^2 + A_1 \left(\mathbb{E} \left[\|x^{k+1} - x^k\|^2 \right] \right) \right] \\ & \quad + (4 + 3\gamma) \left[\sigma_{k-1}^2 + A_1 \left(\mathbb{E} \left[\|x^k - x^{k-1}\|^2 \right] \right) \right] \\ & \quad + (4L^2 + 3\gamma L^2) \|x^k - x^{k-1}\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} - 4 \right) \|x^{k+1} - x^k\|^2, \end{aligned} \quad (26)$$

得证。

为了得到算法的下降性, 本文定义如下的 Lyapunov 函数

$$T^{k+1} = D_{\gamma}(x^{k+1}, z^{k+1}, y^{k+1}) + a\sigma_{k+1}^2 + b\sigma_k^2 + c\|x^{k+1} - x^k\|^2,$$

其中 a, b, c 的取值分别为

$$a = \frac{b + \frac{1}{2\beta} + 4 + 3\gamma}{\rho} = \frac{(4 + 3\gamma)(1 + \rho) \cdot 2\beta + \rho}{2\beta\rho^2}, b = \frac{4 + 3\gamma}{\rho}, c = 4L^2 + 3\gamma L^2 + \frac{4 + 3\gamma}{\rho} A_1.$$

引理 3.2 若假设 1 与假设 2 成立, $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ 是由随机 DR 算法生成的序列, 步长 γ 满足

$$\gamma \in \left(0, \frac{\sqrt{B^2 + \frac{24(\rho^2 + \rho + 1)A_1}{\rho^2}} - B}{48(\rho^2 + \rho + 1)A_1} \cdot \rho^2 \right), \text{ 其中 } B = L + \beta + 8 + \left(\frac{9\rho + 8}{\rho^2} + \frac{1}{\beta} + 8 \right) A_1. \text{ 对于任意 } k \in \mathbb{N}, \text{ 我们有}$$

$$\mathbb{E}[T^{k+1}] \leq \mathbb{E}[T^k] - B_1 \mathbb{E}\|x^{k+1} - x^k\|^2,$$

$$\text{其中 } B_1 = \frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} - 4 - aA_1 - \left(\frac{1}{2\beta} + 4 + 3\gamma \right) A_1 > 0.$$

证明: 由引理 3.1 和函数 T 的定义可得到以下不等式

$$\begin{aligned}
\mathbb{E}[T^{k+1}] &= \mathbb{E}\left[D_\gamma(x^{k+1}, z^{k+1}, y^{k+1})\right] + a\sigma_{k+1}^2 + b\sigma_k^2 + c\mathbb{E}\|x^{k+1} - x^k\|^2 \\
&\leq \mathbb{E}\left[D_\gamma(x^k, z^k, y^k)\right] + a\sigma_k^2 + b\sigma_{k-1}^2 + c\mathbb{E}\|x^k - x^{k-1}\|^2 \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} - 4 - aA_1 - \left(\frac{1}{2\beta} + 4 + 3\gamma\right)A_1\right)\mathbb{E}\|x^{k+1} - x^k\|^2 \\
&= \mathbb{E}[T^k] - B_1\mathbb{E}\|x^{k+1} - x^k\|^2,
\end{aligned} \tag{27}$$

由于 γ 满足 $\gamma \in \left(0, \frac{\sqrt{B^2 + \frac{24(\rho^2 + \rho + 1)A_1}{\rho^2}} - B}{48(\rho^2 + \rho + 1)A_1} \cdot \rho^2\right)$, 其中 $B = L + \beta + 8 + \left(\frac{9\rho + 8}{\rho^2} + \frac{1}{\beta} + 8\right)A_1$, 所以

$$B_1 = \frac{1}{2\gamma} - \frac{L}{2} - \frac{\beta}{2} - 4 - aA_1 - \left(\frac{1}{2\beta} + 4 + 3\gamma\right)A_1 > 0, \text{ 得证。}$$

本文将基于 KL 框架分析算法的收敛性, 上文已得到充分下降性, 接下来我们要证明相对误差条件。

引理 3.3 若假设 1 和假设 2 成立, 令 $\{(x^{k+1}, z^{k+1}, y^{k+1})\}$ 是由随机 DR 算法生成的序列, 那么存在 $\omega^{k+1} \in \partial D_\gamma(x^{k+1}, z^{k+1}, y^{k+1})$, 使得

$$\mathbb{E}\|\omega^{k+1}\| \leq \left(2L + \frac{1}{\gamma} + 2A_2\right)\mathbb{E}\|x^{k+1} - x^k\| + \left(A_2 + \frac{2}{\gamma}\right)\mathbb{E}\|x^{k+2} - x^{k+1}\| + 2\mathbb{E}\xi_k + \mathbb{E}\xi_{k+1}.$$

证明: 令

$$\begin{cases} \omega_x^{k+1} = \nabla f(x^{k+1}) - \frac{1}{\gamma}(y^{k+1} - x^{k+1}) \\ \omega_z^{k+1} = \frac{1}{\gamma}(2x^{k+1} - y^k - z^{k+1}) + \frac{1}{\gamma}(y^{k+1} - z^{k+1}) \\ \omega_y^{k+1} = \frac{1}{\gamma}(z^{k+1} - x^{k+1}). \end{cases}$$

则我们有 $\omega^{k+1} = (\omega_x^{k+1}, \omega_z^{k+1}, \omega_y^{k+1}) \in \partial D_\gamma(x^{k+1}, z^{k+1}, y^{k+1})$ 。由公式(12)与定义 2.1.3 可得

$$\begin{aligned}
\|\omega_x^{k+1}\| &= \left\| \nabla f(x^{k+1}) - \frac{1}{\gamma}(y^{k+1} - x^{k+1}) \right\| \\
&= \left\| \nabla f(x^{k+1}) - \nabla f(x^k) + \nabla f(x^k) - g_k + g_k - \frac{1}{\gamma}(y^{k+1} - x^{k+1}) \right\| \\
&\leq L\|x^{k+1} - x^k\| + \|\nabla f(x^k) - g_k\| + \frac{1}{\gamma}\|y^{k+1} - y^k\|.
\end{aligned}$$

结合公式(19)整理可得

$$\|\omega_x^{k+1}\| \leq L\|x^{k+1} - x^k\| + \frac{1}{\gamma}\|x^{k+2} - x^{k+1}\| + \|\nabla f(x^k) - g_k\| + \|g_{k+1} - g_k\|. \tag{28}$$

整理可得

$$\begin{aligned}
\|\omega_z^{k+1}\| &= \frac{1}{\gamma}\|y^{k+1} - y^k\| \\
&\leq \frac{1}{\gamma}\|x^{k+1} - x^k\| + \|g_{k+1} - g_k\|.
\end{aligned} \tag{29}$$

针对 ω_y^{k+1} , 有

$$\begin{aligned}\|\omega_y^{k+1}\| &= \frac{1}{\gamma} \|y^{k+1} - y^k\| \\ &\leq \frac{1}{\gamma} \|x^{k+2} - x^{k+1}\| + \|g_{k+1} - g_k\|,\end{aligned}\quad (30)$$

其中不等式是由公式(19)得到的。由公式(28), (29), (30)可得

$$\begin{aligned}\|\omega^{k+1}\| &\leq \left(L + \frac{1}{\gamma}\right) \|x^{k+1} - x^k\| + \frac{2}{\gamma} \|x^{k+2} - x^{k+1}\| + \|\nabla f(x^k) - g_k\| + \|g_{k+1} - g_k\| \\ &\leq \left(L + \frac{1}{\gamma}\right) \|x^{k+1} - x^k\| + \frac{2}{\gamma} \|x^{k+2} - x^{k+1}\| + \|\nabla f(x^k) - g_k\| \\ &\quad + \|g_{k+1} - \nabla f(x^{k+1}) + \nabla f(x^{k+1}) - \nabla f(x^k) + \nabla f(x^k) - g_k\| \\ &\leq \left(2L + \frac{1}{\gamma}\right) \|x^{k+1} - x^k\| + 2\|\nabla f(x^k) - g_k\| \\ &\quad + \|\nabla f(x^{k+1}) - g_{k+1}\| + \frac{2}{\gamma} \|x^{k+2} - x^{k+1}\|.\end{aligned}$$

对上式左右两边同时求期望可得

$$\begin{aligned}\mathbb{E}\|\omega^{k+1}\| &\leq \left(2L + \frac{1}{\gamma}\right) \mathbb{E}\|x^{k+1} - x^k\| + 2\mathbb{E}\|\nabla f(x^k) - g_k\| \\ &\quad + \mathbb{E}\|\nabla f(x^{k+1}) - g_{k+1}\| + \frac{2}{\gamma} \mathbb{E}\|x^{k+2} - x^{k+1}\| \\ &\leq \left(2L + \frac{1}{\gamma} + 2A_2\right) \mathbb{E}\|x^{k+1} - x^k\| \\ &\quad + \left(A_2 + \frac{2}{\gamma}\right) \mathbb{E}\|x^{k+2} - x^{k+1}\| + 2\mathbb{E}\xi_k + \mathbb{E}\xi_{k+1},\end{aligned}$$

其中第二个不等式是由公式(11)得到的。相对误差界条件得证。

为便于收敛性分析, 我们用 $S(\mu^0)$ 表示由初始值 μ^0 产生的序列 $\{\mu^k = (x^k, y^k, z^k)\}_{k \in \mathbb{N}}$ 的聚点集。

$$S(\mu^0) := \{x^* \mid \text{存在一个子列 } \{\mu^{k_j}\}_{j=0}^{\infty} \text{ 使得 } \lim_{j \rightarrow \infty} \mu^0 = \bar{\mu}\}.$$

命题 3.1 假设 1, 2 成立, 令 $\{\mu^k = (x^k, y^k, z^k)\}_{k \in \mathbb{N}}$ 是由随机 DR 分裂算法所生成的迭代序列, 且该序列有界, 其初始点为 $\mu^0 = (x^0, y^0, z^0)$ 。则可以得到以下结论:

- (1) $\lim_{k \rightarrow \infty} \mathbb{E}[T_k] = \lim_{k \rightarrow \infty} \mathbb{E}[D_\gamma(\mu^k)] = \lim_{k \rightarrow \infty} \mathbb{E}[F(y^k)] = F^*$, 其中 $F^* \in [F, \infty)$;
- (2) $\lim_{k \rightarrow \infty} \mathbb{E}[\text{dist}(0, \partial D_\gamma(\mu^k))] = 0$;
- (3) 对于所有的 $\bar{\mu} = (\bar{x}, \bar{y}, \bar{z}) \in S(\mu^0)$ 有集合 $S(\mu^0) \neq \emptyset$ 和 $\mathbb{E}[\text{dist}(0, \partial D_\gamma(\bar{\mu}))] = 0$;
- (4) $\text{dist}(\mu^k, S(\mu^0)) \xrightarrow{a.s.} 0$, 当 $k \rightarrow \infty$;
- (5) $S(\mu^0)$ 是几乎处处紧的且联通的;
- (6) 对于所有的 $\bar{\mu} \in S(\mu^0)$, 有 $\mathbb{E}[D_\gamma(\bar{\mu})] = F^*$ 。

证明: (1) 根据引理 3.2, 我们知道 $\mathbb{E}[T_k]$ 是收敛的且 $\lim_{k \rightarrow \infty} \mathbb{E}[\|x^{k+1} - x^k\|^2] = 0$ 。进而我们能得到

$\lim_{k \rightarrow \infty} \mathbb{E}[\|x^{k+1} - x^k\|^2] = 0$, $\lim_{k \rightarrow \infty} \mathbb{E}[\|z^{k+1} - y^{k+1}\|^2] = 0$, 根据 x^{k+1} 的更新, 我们有 $z^{k+1} - z^k = x^{k+1} - x^k + x^{k-1} - x^k + y^{k+1} - y^k$, 进而可推导出 $\lim_{k \rightarrow \infty} \mathbb{E}[\|z^{k+1} - z^k\|^2] = 0$ 。同理我们可以得到 $\lim_{k \rightarrow \infty} \mathbb{E}[\|\mu^{k+1} - \mu^k\|^2] = 0$ 。所以我们可以推出

$$\lim_{k \rightarrow \infty} \mathbb{E}[T_k] = \lim_{k \rightarrow \infty} \mathbb{E}[D_\gamma(\mu^k)] = \lim_{k \rightarrow \infty} \mathbb{E}[F(y^k)] = F^*.$$

(2) 通过引理 3.3, 我们可以推导出 $w_k \in \partial D_\gamma(\mu^k)$, 存在一个常数 b_2 使得

$$\mathbb{E}[\|w^k\|] \leq \left(2L + \frac{1}{\gamma} + 2A_2\right) \mathbb{E}\|x^k - x^{k-1}\| + \left(A_2 + \frac{2}{\gamma}\right) \mathbb{E}\|x^{k+1} - x^k\| + 2\mathbb{E}\xi_{k-1} + \mathbb{E}\xi_k.$$

由于 $\lim_{k \rightarrow \infty} \mathbb{E}[\|x^{k+1} - x^k\|^2] = 0$, 易得 $\lim_{k \rightarrow \infty} \mathbb{E}[\|y^k - y^{k-1}\|] = \lim_{k \rightarrow \infty} \mathbb{E}[\|y^{k+1} - y^k\|] = 0$ 。

又因为假设 2, 故我们有 $\lim_{k \rightarrow \infty} \mathbb{E}[\text{dist}(0, \partial D_\gamma(\mu^k))] = 0$ 。

(3) 对任意 $\bar{\mu} = (\bar{x}, \bar{y}, \bar{z}) \in S(\mu^0)$, 让 $\{\mu^{k_j}\}_{j=0}^\infty$ 为 $\{\mu^k\}_{k=0}^\infty$ 的收敛到 $\bar{\mu}$ 的子序列, 我们有

$\lim_{j \rightarrow \infty} h(z^{k_j+1}) = h(\bar{z})$ 。根据 f 的连续性, 我们可以得到 $\lim_{j \rightarrow \infty} D_\gamma(\mu^{k_j}) = D_\gamma(\bar{\mu})$ 。根据引理 3.3 和上面 (2) 的分析我们得到了 $w^{k_j} \in \partial D_\gamma(\mu^{k_j})$ 和 $\lim_{j \rightarrow \infty} \mathbb{E}[\|w^{k_j}\|] = 0$, 通过进一步利用次微分的闭性, 我们有 $\lim_{j \rightarrow \infty} \mathbb{E}[\|w^{k_j}\|] = \mathbb{E}[\text{dist}(0, \partial D_\gamma(\bar{\mu}))] = 0$ 。

(4) 和 (5) 对于任意满足 $\|\mu^k - \mu^{k-1}\| \xrightarrow{a.s.} 0, k \rightarrow \infty$ 条件的序列成立。这一结论可由 [20, Lemma 5, Remark 5] 和 [21, Lemma 4.3 (5) (6)] 证到。

(6) 从 (1) 的证明中, 我们得到了 $\lim_{k \rightarrow \infty} \mathbb{E}[D_\gamma(\mu^k)] = F^*$, 则可以推导出对任意收敛于 $\bar{\mu} \in S(\mu^0)$ 的序列 $\{\mu^{k_j}\}_{j=0}^\infty$ 都满足 $\lim_{k \rightarrow \infty} \mathbb{E}[D_\gamma(\mu^k)] = F^*$ 。因为函数 f 的连续性以及 $\lim_{j \rightarrow \infty} h(z^{k_j+1}) = h(\bar{z})$, 我们有 $\lim_{j \rightarrow \infty} \mathbb{E}[D_\gamma(\mu^{k_j})] = \mathbb{E}[D_\gamma(\bar{\mu})] = \mathbb{E}[F(\bar{y})]$, 所以, 这可以推导出 D_γ 在 $S(\mu^0)$ 上的期望为常数。

充分下降性和相对误差条件已被建立, 接下来我们要证明随机 DR 分裂算法在 KL 性质下的全局收敛性。对于任意满足 p 矩矩 ($p \geq 1$) 有限的随机变量 X , 我们定义 $|X|_p = (\mathbb{E}[X^p])^{1/p}, p \geq 1$ 。则我们可以得到 $|X|_p \leq |X|_q, \forall 1 \leq p \leq q$ 。

引理 3.4 [21] 在假设 1, 假设 2 成立的条件下, 令 $\{\mu^k = (x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ 是由随机 DR 分裂算法产生的序列, 如果序列 $\{\mu^k\}_{k \in \mathbb{N}}$ 是有界的, 并且函数 D_γ 是一个半代数函数, 假设 D_γ 满足 $\varphi(t) = at^{1-\theta}$ 的 KL 性质, 即存在 $k_0 > 0$, 使得

$$\varphi'(E[D_\gamma^k - F_k^*])E[\text{dist}(0, \partial D_\gamma^k)] \geq 1, k > k_0.$$

其中 $\{F_k^*\}_{k=0}^\infty$ 是收敛到 $\mathbb{E}[D_\gamma(\bar{\mu})]$ 的非递减序列。

定理 3.1 (随机 DR 分裂算法的全局收敛) 在假设 1, 假设 2 成立的条件下, 令 $\{\mu^k = (x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ 是由随机 DR 分裂算法产生的序列且假设其有界。假设 D_γ 满足 KL 性质, 则 $\sum_{k=0}^\infty \mathbb{E}\|x^{k+1} - x^k\| < \infty$ 成立, 且算法全局收敛。

证明: 由 D_γ 满足 KL 性质可知, 存在常数 $k_1 > 0$ 和函数 $\varphi(x) = ax^{1-\theta}, \theta \in (0, 1)$ 满足

$$\varphi'(\mathbb{E}[D_\gamma^k - F_k^*])\mathbb{E}[\text{dist}(0, \partial D_\gamma^k)] \geq 1,$$

其中 $\{F_k^*\}_{k \in \mathbb{N}}$ 是一个收敛到 $\mathbb{E}[D_\gamma(\bar{\mu})]$ 的非减序列, $\bar{\mu}$ 为序列 $\{\mu^k = (x^k, y^k, z^k)\}_{k \in \mathbb{N}}$ 的子列的极限点。当 $\theta \in (0, \frac{1}{2})$ 时, D_γ^k 满足指数为 $\frac{1}{2}$ 时的 KL 性质。因此只需分析 $\theta \in [\frac{1}{2}, 1)$ 的情况。

由假设 2 可得

$$\mathbb{E}[\xi_k] = \mathbb{E}\left[\sum_{i=1}^s \|\alpha_k^i\|\right] \leq \sqrt{s} \left(\mathbb{E}\left[\sum_{i=1}^s \|\alpha_k^i\|^2\right] \right)^{1/2} \leq \sqrt{s} (\mathbb{E}[\sigma_k^2])^{1/2} = \sqrt{s} |\sigma_k|_2.$$

由上式以及性质 2.2.1, 还有 $\{F_s^*\}$ 的非减性整理 KL 不等式可得

$$\begin{aligned} & \mathbb{E}[\text{dist}(0, \partial D_\gamma^k)] \\ & \leq \left(A_2 + \frac{2}{\gamma}\right) \mathbb{E}\|x^{k+1} - x^k\| + B_2 \mathbb{E}\|x^k - x^{k-1}\| + 2\mathbb{E}\xi_{k-1} + \mathbb{E}\xi_k \\ & \leq \left(A_2 + \frac{2}{\gamma}\right) \mathbb{E}|x^{k+1} - x^k|_2 + B_2 \mathbb{E}|x^k - x^{k-1}|_2 + 2\sqrt{s} |\sigma_{k-1}|_2 + \sqrt{s} |\sigma_k|_2, \end{aligned} \quad (31)$$

其中 $B_2 = 2L + \frac{1}{\gamma} + 2A_2$ 。

在 $\rho \in (0, 1)$ 时有以下不等式成立

$$\begin{aligned} |\sigma_k|_2 & \leq \sqrt{(1-\rho)|\sigma_{k-1}|_2^2 + A_3|x^k - x^{k-1}|_2^2} \\ & \leq \sqrt{(1-\rho)} |\sigma_{k-1}|_2 + \sqrt{A_3} |x^k - x^{k-1}|_2 \\ & \leq \left(1 - \frac{\rho}{2}\right) |\sigma_{k-1}|_2 + \sqrt{A_3} |x^k - x^{k-1}|_2, \end{aligned}$$

其中第二步是由不等式 $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ 得到。进而可以推出以下不等式

$$|\sigma_{k-1}|_2 \leq \frac{2\sqrt{A_3}}{\rho} |x^k - x^{k-1}|_2 + \frac{2}{\rho} (|\sigma_{k-1}|_2 - |\sigma_k|_2), |\sigma_k|_2 \leq \frac{2\sqrt{A_3}}{\rho} |x^{k+1} - x^k|_2 + \frac{2}{\rho} (|\sigma_k|_2 - |\sigma_{k+1}|_2).$$

由此将上式左右两边分别乘 $2\sqrt{s}$, $\sqrt{s}$ 作为系数可以计算得到

$$2\sqrt{s} |\sigma_{k-1}|_2 \leq \frac{4\sqrt{sA_3}}{\rho} |x^k - x^{k-1}|_2 + \frac{4\sqrt{s}}{\rho} (|\sigma_{k-1}|_2 - |\sigma_k|_2), \quad (32)$$

$$\sqrt{s} |\sigma_k|_2 \leq \frac{2\sqrt{sA_3}}{\rho} |x^{k+1} - x^k|_2 + \frac{2\sqrt{s}}{\rho} (|\sigma_k|_2 - |\sigma_{k+1}|_2). \quad (33)$$

所以综合公式(31), (32), (33)可以计算得出此不等式

$$\begin{aligned} \mathbb{E}[\text{dist}(0, \partial D_\gamma^k)] & \leq \left(B_2 + \frac{4\sqrt{sA_3}}{\rho}\right) |x^k - x^{k-1}|_2 + \left(A_2 + \frac{2}{\gamma} + \frac{2\sqrt{sA_3}}{\rho}\right) |x^{k+1} - x^k|_2 \\ & \quad + \frac{4\sqrt{s}}{\rho} (\sigma_{k-1}^2 - \sigma_k^2) + \frac{2\sqrt{s}}{\rho} (\sigma_k^2 - \sigma_{k+1}^2). \end{aligned}$$

我们定义一个新的序列 C_k

$$C_k := \left(B_2 + \frac{4\sqrt{sA_3}}{\rho} \right) \|x^k - x^{k-1}\|_2 + \left(A_2 + \frac{2}{\gamma} + \frac{2\sqrt{sA_3}}{\rho} \right) \|x^{k+1} - x^k\|_2 \\ + \frac{4\sqrt{s}}{\rho} (\sigma_{k-1}^2 - \sigma_k^2) + \frac{2\sqrt{s}}{\rho} (\sigma_k^2 - \sigma_{k+1}^2).$$

记 $M = \max \left\{ \left(B_2 + \frac{4\sqrt{sA_3}}{\rho} \right), \left(A_2 + \frac{2}{\gamma} + \frac{2\sqrt{sA_3}}{\rho} \right) \right\}$, 则有

$$C_k \leq M \left[\|x^{k+1} - x^k\|_2 + \|x^k - x^{k-1}\|_2 + (\sigma_{k-1}^2 - \sigma_k^2) + (\sigma_k^2 - \sigma_{k+1}^2) \right]. \quad (34)$$

定义函数 $\varphi(t) = at^{1-\theta}$, $\theta \in \left[\frac{1}{2}, 1 \right)$ 。对于所有的 $k \geq k_1$, D_γ^k 满足 KL 不等式, 所以

$$\varphi'(\mathbb{E}[D_\gamma^k - F_k^*]) = a(1-\theta)(\mathbb{E}[D_\gamma^k - F_k^*])^{-\theta} \\ \Rightarrow (\mathbb{E}[D_\gamma^k - F_k^*])^\theta \leq a(1-\theta)C_k.$$

而对于所有的 $k \geq k_2$, 结合假设 2 可知存在常数 $c > 0$ 满足以下不等式

$$\left(\mathbb{E} \left[a\sigma_k^2 + b\sigma_{k-1}^2 + c\|x^k - x^{k-1}\|^2 \right] \right)^\theta \leq O \left((\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\| + \sigma_{k-1}^2)^\theta \right) \\ \leq cC_k.$$

又因为 $(a+b)^\theta \leq a^\theta + b^\theta$, $\theta \in \left[\frac{1}{2}, 1 \right)$, 因此可以得到对于所有的 $k \geq k_3 = \max\{k_1, k_2\}$, 以下不等式成立

$$(\mathbb{E}[T_\gamma^k - F_k^*])^\theta = \left(\mathbb{E} \left[D_\gamma(x^k, z^k, y^k) + a\sigma_k^2 + b\sigma_{k-1}^2 + c\|x^k - x^{k-1}\|^2 - F_k^* \right] \right)^\theta \\ \leq \left(\mathbb{E}[D_\gamma^k - F_k^*] \right)^\theta + \left(\mathbb{E} \left[a\sigma_k^2 + b\sigma_{k-1}^2 + c\|x^k - x^{k-1}\|^2 \right] \right)^\theta \\ \leq a(1-\theta)C_k + cC_k.$$

进而可以得到以下不等式

$$(\mathbb{E}[T_\gamma^k - F_k^*])^\theta \leq ad(1-\theta)C_k, \quad \forall k > k_2, \quad (35)$$

其中 d 满足 $ad(1-\theta) \geq a(1-\theta) + c$ 。定义函数 $\phi(t) = adt^{1-\theta}$, $\theta \in \left[\frac{1}{2}, 1 \right)$, 由(35)可得

$$\phi'(\mathbb{E}(T^k - F_k^*)) \geq C_k^{-1}.$$

对于所有的 $k \in \mathbb{N}$, 我们定义 $\Delta_{k,k+1} = \phi(\mathbb{E}[T_\gamma^k - F_k^*]) - \phi(\mathbb{E}[T_\gamma^{k+1} - F_{k+1}^*])$ 。

由函数 ϕ 的凹性可得

$$\Delta_{k,k+1} \geq \phi'(\mathbb{E}[T_\gamma^k - F_k^*]) (\mathbb{E}[T_\gamma^{k+1} - T_\gamma^k + F_k^* - F_{k+1}^*]) \\ \geq \phi'(\mathbb{E}[T_\gamma^k - F_k^*]) \mathbb{E}[T_\gamma^{k+1} - T_\gamma^k] \\ \geq \frac{B_1 \mathbb{E}\|x^{k+1} - x^k\|^2}{C_k}.$$

综合上式和(34), 整理可得以下不等式

$$9\mathbb{E}\|x^{k+1}-x^k\|^2 \leq \frac{9M}{B_1}\Delta_{k,k+1}\left[\left|x^{k+1}-x^k\right|_2+\left|x^k-x^{k-1}\right|_2+\frac{4\sqrt{s}}{\rho M}(\sigma_{k-1}^2-\sigma_k^2)+\frac{2\sqrt{s}}{\rho M}(\sigma_k^2-\sigma_{k+1}^2)\right].$$
 又因为

$\mathbb{E}X \leq \sqrt{\mathbb{E}X^2}$, 并且对于任意的 $u, v \geq 0$, $2\sqrt{uv} \leq u+v$, 因此对上式左右两边同时开根号可得, 对于所有的 $k \geq k_3 = \max\{k_1, k_2\}$,

$$3\left|x^{k+1}-x^k\right|_2 \leq \frac{9M}{4B_1}\Delta_{k,k+1}+\left|x^{k+1}-x^k\right|_2+\left|x^k-x^{k-1}\right|_2 \\ +\frac{4\sqrt{s}}{\rho M}(\sigma_{k-1}^2-\sigma_k^2)+\frac{2\sqrt{s}}{\rho M}(\sigma_k^2-\sigma_{k+1}^2).$$

化简可得

$$2\left|x^{k+1}-x^k\right|_2 \leq \frac{9M}{4B_1}\Delta_{k,k+1}+\left|x^k-x^{k-1}\right|_2 \\ +\frac{4\sqrt{s}}{\rho M}(\sigma_{k-1}^2-\sigma_k^2)+\frac{2\sqrt{s}}{\rho M}(\sigma_k^2-\sigma_{k+1}^2).$$

最后对上式左右两边同时累加化简可以得到

$$\sum_{k \geq k_3}\left|x^{k+1}-x^k\right|_2 \leq \frac{9M}{4B_1}\left[\phi(T_{k_3}-F_{k_3}^*)-\phi(T_\infty-F_\infty^*)\right]+\left|x^{k_3}-x^{k_3-1}\right|_2 \\ +\frac{4\sqrt{s}}{\rho M}(\sigma_{k_3-1}^2-\sigma_\infty^2)+\frac{2\sqrt{s}}{\rho M}(\sigma_{k_3}^2-\sigma_\infty^2) \\ < \infty.$$

因此可以得到

$$\sum_{k \geq k_3}\mathbb{E}\|x^{k+1}-x^k\| = \sum_{k \geq k_3}\left|x^{k+1}-x^k\right|_1 \leq \sum_{k \geq k_3}\left|x^{k+1}-x^k\right|_2 < \infty.$$

进而可知序列 $\{x^k\}_{k \in \mathbb{N}}$ 柯西收敛。

由公式(20)可得

$$\|y^{k+1}-y^k\| \leq \|x^{k+2}-x^{k+1}\|+\gamma\|g_{k+1}-\nabla f(x^{k+1})\|+\gamma\|\nabla f(x^k)-g_k\|+\gamma L\|x^{k+1}-x^k\|.$$

对上式左右两边同时求期望可得

$$\mathbb{E}\|y^{k+1}-y^k\| \leq \mathbb{E}\|x^{k+2}-x^{k+1}\|+\gamma\mathbb{E}\|g_{k+1}-\nabla f(x^{k+1})\| \\ +\gamma\mathbb{E}\|\nabla f(x^k)-g_k\|+\gamma L\mathbb{E}\|x^{k+1}-x^k\| \\ \leq (\gamma A_2+1)\mathbb{E}\|x^{k+2}-x^{k+1}\|+(\gamma A_2+\gamma L)\mathbb{E}\|x^{k+1}-x^k\| \\ +\gamma\mathbb{E}[\xi_{k+1}]+\gamma\mathbb{E}[\xi_k].$$

又因为 $\sum_{k \geq k_3}\mathbb{E}\|x^{k+1}-x^k\| < \infty$, 对上式左右两边同时累和可得

$$\sum_{k \geq k_3}\mathbb{E}\|y^{k+1}-y^k\| \leq (\gamma A_2+1)\sum_{k \geq k_3}\mathbb{E}\|x^{k+2}-x^{k+1}\|+(\gamma A_2+\gamma L)\sum_{k \geq k_3}\mathbb{E}\|x^{k+1}-x^k\| \\ +\gamma\sum_{k \geq k_3}\mathbb{E}[\xi_{k+1}]+\gamma\sum_{k \geq k_3}\mathbb{E}[\xi_k] \\ < \infty.$$

由公式(7)可得

$$\begin{aligned}\|z^{k+1} - z^k\| &= \|y^{k+1} - y^k + x^{k+1} - y^k + y^{k-1} - x^k\| \\ &\leq \|y^{k+1} - y^k\| + \|y^k - y^{k-1}\| + \|x^{k+1} - x^k\|.\end{aligned}$$

由对上式左右两边同时求期望并累和可得

$$\begin{aligned}\sum_{k \geq k_3} \mathbb{E} \|z^{k+1} - z^k\| &= \sum_{k \geq k_3} \mathbb{E} \|y^{k+1} - y^k + x^{k+1} - y^k + y^{k-1} - x^k\| \\ &\leq \sum_{k \geq k_3} \mathbb{E} \|y^{k+1} - y^k\| + \sum_{k \geq k_3} \mathbb{E} \|y^k - y^{k-1}\| + \sum_{k \geq k_3} \mathbb{E} \|x^{k+1} - x^k\| \\ &< \infty.\end{aligned}$$

所以, 序列 $\{x^k\}, \{y^k\}, \{z^k\}$ ($k \in \mathbb{N}$) 均柯西收敛。因此算法生成序列 $\{(x^k, z^k, y^k)\}$ ($k \in \mathbb{N}$) 柯西收敛, 即该算法全局收敛。

4. 数值实验

本节, 我们应用本文提出的随机 DR 分裂算法(SDR)来求解稀疏 Logistic 回归问题, 并将该算法与经典的邻近梯度算法 PG 以及 Prox-SARAH 进行比较, 以阐述本文所提出的算法的高效性。

首先, 我们考虑带有 $l_1 l_2$ 正则化的稀疏 Logistic 回归模型, 如今已被广泛应用于二分类问题。该模型的具体表述如下:

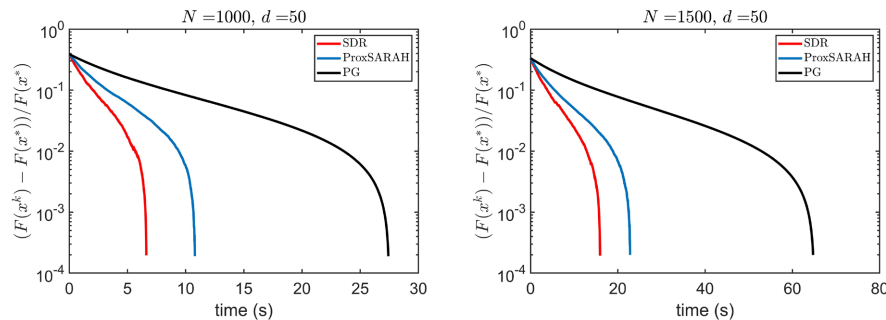
$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i a_i^T x)) + \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|^2,$$

其中 $a_i \in \mathbb{R}^d, b_i \in \{-1, 1\}, i \in \{1, 2, \dots, N\}$ 为样本的特征以及相应的标签, $\lambda_1, \lambda_2 > 0$ 为正则化参数, 用以控制变量的稀疏性。通过令 $f(x) = \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i a_i^T x)), h(x) = \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|^2$, 上述稀疏 Logistic 回归问题即可转化为问题(1)。

在实验中, 我们随机生成特征矩阵 $A = (a_1, a_2, \dots, a_N) \in \mathbb{R}^{d \times N}$, 并计算出对应的标签

$b = (b_1, b_2, \dots, b_N)^T \in \mathbb{R}^N$, 稀疏正则化参数设置为 $\lambda_1 = \lambda_2 = 10^{-5}$ 。针对问题规模, 本文选择样本量为 $N = 1000$ 和 $N = 1500$, 以及 $d = 50$ 和 $d = 100$ 。两两组合共构成 4 种不同的数据规模设置, 以对比分析算法在不同规模下的性能表现。此外, 我们将算法的最大迭代次数统一设置为 20,000, 并以相对误差 $\frac{F(x) - F(x^*)}{F(x^*)} \leq 2 \times 10^{-4}$ 作为算法的停机准则, 其中 $F(x^*)$ 为最小的函数值。

数值实验结果见下图 1, 针对不同的样本规模, 我们提出的算法 SDR 总是能够在规定的最大迭代次数内达到收敛, 且收敛所用的时间远少于 PG 和 Prox-SARAH, 收敛速度最快。此外, 可以看到当数据规模增大时, 算法达到收敛需要的时间也显著增大, 而对比而言, SDR 因规模增大而增加的时间也是最少的, 充分展示了我们所提出的算法的优越性。



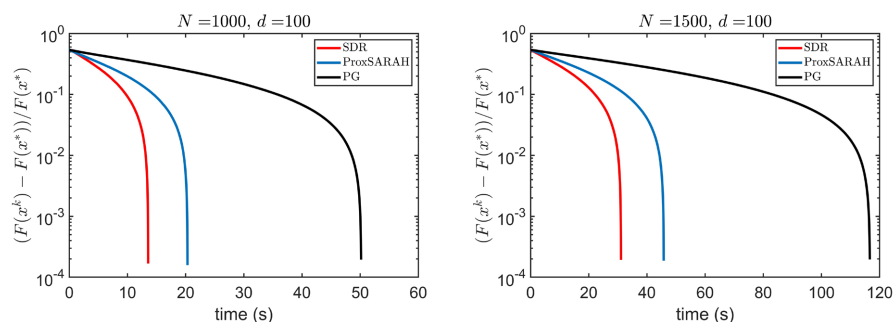


Figure 1. Comparison of SDR, PG and Prox-SARAH under different data scales

图 1. 不同数据规模下 SDR 算法与 PG、Prox-SARAH 算法对比

5. 总结

针对大规模非凸复合优化问题, 本文将方差缩减的随机梯度算法引入到 DR 分裂算法中, 提出随机 DR 分裂算法, 可以更高效, 更稳定的解决大规模优化问题, 降低算法的计算成本, 并在 KL 不等式的前提下, 证明随机 DR 分裂算法的全局收敛性。具体地, 我们首先建立 Lyapunov 函数在算法前后迭代步的下降性, 再建立 DR 价值函数的相对误差界条件, 并最终基于 KL 性质建立算法的全局收敛性。此外, 通过数值实验, 验证了随机 DR 分裂算法(SDR)的优越性。

参考文献

- [1] Lions, P.L. and Mercier, B. (1979) Splitting Algorithms for the Sum of Two Nonlinear Operators. *SIAM Journal on Numerical Analysis*, **16**, 964-979. <https://doi.org/10.1137/0716071>
- [2] Beck, A. and Teboulle, M. (2009) Gradient-Based Algorithms with Applications to Signal-Recovery Problems. In: *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, 42-88. <https://doi.org/10.1017/cbo9780511804458.003>
- [3] Douglas, J. and Rachford, H.H. (1956) On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables. *Transactions of the American Mathematical Society*, **82**, 421-439. <https://doi.org/10.1090/s0002-9947-1956-0084194-4>
- [4] Aragón Artacho, F.J., Borwein, J.M. and Tam, M.K. (2013) Recent Results on Douglas-Rachford Methods for Combinatorial Optimization Problems. *Journal of Optimization Theory and Applications*, **163**, 1-30. <https://doi.org/10.1007/s10957-013-0488-0>
- [5] Boţ, R.I., Csetnek, E.R. and Hendrich, C. (2015) Inertial Douglas-Rachford Splitting for Monotone Inclusion Problems. *Applied Mathematics and Computation*, **256**, 472-487. <https://doi.org/10.1016/j.amc.2015.01.017>
- [6] Li, G. and Pong, T.K. (2015) Douglas-Rachford Splitting for Nonconvex Optimization with Application to Nonconvex Feasibility Problems. *Mathematical Programming*, **159**, 371-401. <https://doi.org/10.1007/s10107-015-0963-5>
- [7] 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>
- [8] Li, M., Zhang, T., Chen, Y. and Smola, A.J. (2014). Efficient Mini-Batch Training for Stochastic Optimization. *Proceedings of the 20th 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>
- [9] Schmidt, M., Le Roux, N. and Bach, F. (2016) Minimizing Finite Sums with the Stochastic Average Gradient. *Mathematical Programming*, **162**, 83-112. <https://doi.org/10.1007/s10107-016-1030-6>
- [10] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. *Advances in Neural Information Processing Systems*, **2013**, Article 26.
- [11] Doucet, A., Godsill, S. and Andrieu, C. (2000) On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing*, **10**, 197-208. <https://doi.org/10.1023/a:1008935410038>
- [12] Nesterov, Y. (2012) Gradient Methods for Minimizing Composite Functions. *Mathematical Programming*, **140**, 125-161. <https://doi.org/10.1007/s10107-012-0629-5>

-
- [13] Nguyen, L.M., van Dijk, M., Phan, D.T., Nguyen, P.H., Weng, T. and Kalagnanam, J.R. (2022) Finite-Sum Smooth Optimization with Sarah. *Computational Optimization and Applications*, **82**, 561-593. <https://doi.org/10.1007/s10589-022-00375-x>
 - [14] Fang, C., Li, C.J., Lin, Z. and Zhang, T. (2018) Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator.
 - [15] Ghadimi, S. and Lan, G. (2013) Stochastic First and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, **23**, 2341-2368. <https://doi.org/10.1137/120880811>
 - [16] Ghadimi, S. and Lan, G. (2015) Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Programming. *Mathematical Programming*, **156**, 59-99. <https://doi.org/10.1007/s10107-015-0871-8>
 - [17] Pham, N.H., Nguyen, L.M., Phan, D.T., *et al.* (2020) Prox SARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization. *Journal of Machine Learning Research*, **21**, 1-48.
 - [18] Rockafellar, R.T. and Wets, R.J.B. (2009) Variational Analysis. Springer Science & Business Media.
 - [19] Nesterov, Y. (2018) Lectures on Convex Optimization. Springer.
 - [20] Bolte, J., Sabach, S. and Teboulle, M. (2013) Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems. *Mathematical Programming*, **146**, 459-494. <https://doi.org/10.1007/s10107-013-0701-9>
 - [21] Driggs, D., Tang, J., Liang, J., Davies, M. and Schönlieb, C. (2021) A Stochastic Proximal Alternating Minimization for Nonsmooth and Nonconvex Optimization. *SIAM Journal on Imaging Sciences*, **14**, 1932-1970. <https://doi.org/10.1137/20m1387213>