Advances in Applied Mathematics 应用数学进展, 2025, 14(3), 176-191
Published Online March 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.143104

Hans汉斯

基于随机镜像下降对称交替方向乘子法的非凸优化问题研究

王　爽

河北工业大学理学院，天津

收稿日期：2025年2月11日；录用日期：2025年3月4日；发布日期：2025年3月14日

摘　要

本文考虑求解一类可分的带有等式约束的非凸非光滑优化问题，其目标函数是由有限个光滑函数的加权平均与适当下半连续函数的和函数组成。交替方向乘子法能够充分利用模型的特点，是目前求解该类问题有效方法之一，但由于数据规模大，传统的交替方向乘子法效率低下。本文提出“随机镜像下降对称交替方向乘子法”，该算法引入随机方差缩减算子，通过随机选取梯度信息减少梯度计算量，从而提高算法效率；同时使用布雷格曼(Bregman)距离以保证子问题具有显示解，此外算法的对偶变量以对称形式进行更新，提高了算法的高效性和稳定性。理论结果表明，在目标函数满足半代数性质时，该算法生成的迭代序列全局收敛到原问题的驻点。同时数值实验结果验证了算法的有效性。

关键词

约束优化，非凸非光滑优化，随机方差缩减算子，全局收敛

The Research on Nonconvex Optimization Problems Based on Stochastic Mirror Descent Symmetric Alternating Direction Method of Multipliers

Shuang Wang

School of Science, Hebei University of Technology, Tianjin

Received: Feb. 11th, 2025; accepted: Mar. 4th, 2025; published: Mar. 14th, 2025

文章引用: 王爽. 基于随机镜像下降对称交替方向乘子法的非凸优化问题研究[J]. 应用数学进展, 2025, 14(3): 176-191.
DOI: 10.12677/aam.2025.143104

Abstract

This paper considers solving a class of separable nonconvex and nonsmooth optimization problems with equality constraints, where the objective function is composed of the weighted average of a finite number of smooth functions and the sum of another proper lower semicontinuous function. The Alternating Direction Method of Multipliers (ADMM) effectively leverages the characteristics of the model and is one of the popular and efficient methods for solving such problems. However, due to the large data scale, traditional ADMM suffers from low efficiency. This paper proposes the "Stochastic Mirror Descent Symmetric Alternating Direction Method of Multipliers", introduces a stochastic variance reduction operator, which reduces the gradient computation by randomly selecting gradient information, thereby improving the efficiency of the algorithm. Additionally, it uses the Bregman distance to ensure well-posed subproblems. Furthermore, the dual variables of the algorithm are updated symmetrically, which not only extends the applicability of the algorithm but also enhances its efficiency and stability. Theoretical results show that when the objective function satisfies the semi-algebraic property, the iterative sequence generated by the algorithm globally converges to a stationary point of the original problem. Numerical experiments further validate the effectiveness of the algorithm.

Keywords

Constrained Optimization, Nonconvex and Nonsmooth Optimization, Stochastic Variance Reduction Operator, Global Converges

1. 引言

本文考虑一类具有等式约束的有限和形式的非凸非光滑优化问题，在机器学习中被广泛应用，具体模型如下：

$$\min_{x\in\mathbb{R}^{d_1},y\in\mathbb{R}^{d_2}} f(\mathbf{x})+g(\mathbf{y}),\ \text{s.t. } A\mathbf{x}+B\mathbf{y}=b, \tag{1}$$

其中 $f:\mathbb{R}^{d_1}\to\mathbb{R}\cup\{+\infty\}$ 是适当下半连续函数， $g(\mathbf{y})=\frac{1}{m}\sum_{i=1}^{m}g_i(\mathbf{y})$ 其中每个分量函数 $g_i:\mathbb{R}^{d_2}\to\mathbb{R}$ 是光滑函数， $A\in\mathbb{R}^{d\times d_1}$， $B\in\mathbb{R}^{d\times d_2}$， $b\in\mathbb{R}^{d}$ 。

模型(1)广泛应用于统计学习[1]、计算机视觉[2]、3D CT 图像重建[3]等领域。给定一组训练样本 (a_i,b_i)，其中 $i=1,\cdots,n$， a_i 是输入数据，对应的标签 $b_i\in\{-1,1\}$ 。此时二分类任务[4]可以表示为：

$$\min_{y}\frac{1}{m}\sum_{i=1}^{m}g_i(y)+\lambda_1\|By\|_1,$$

其中取 g_i 为非凸 sigmoid 函数： $g_i(y)=\frac{1}{1+\exp(b_i a_i^{\mathrm{T}}y)}$ 。B 为给定的矩阵，引入变量 x，可将问题重新表述为(1)的形式，即 $\min_{x,y} f(x)+g(y)$，s.t. $x-By=0$，其中 $g(y)=1/m\sum_{i=1}^{m}g_i(y)$， $f(x)=\lambda_1\|x\|_1$ 。

近年来，交替方向乘子法(ADMM)在非凸和随机优化中得到了广泛的研究。文献[5] [6]中，研究了非

凸 Bregman ADMM 的收敛性。文献[7]中研究了对称形式 ADMM 的收敛性，然而在标准的凸性假设下并不收敛，但此文献验证了在确保其全局收敛的条件下对称 ADMM 比一般形式 ADMM 收敛更快，鉴于此，He 等人在文献[8]中提出了一种严格收缩的 Peaceman Rachford 分裂方法。需要指出的是，这些研究均基于确定性 ADMM 方法，即不涉及任何随机性。在处理 g 为有限求和的情况时，计算全梯度 ∇g 往往会非常耗时，导致方法效率降低。为了解决这一问题，研究者通过使用 ∇g 的随机估计来代替全梯度的计算，从而衍生出多个随机版本的 ADMM。随着大规模优化问题的出现随机梯度算法如 SAGA [9]、SVRG [10]和 SARAH [11]等推动了 ADMM 的进一步发展。在文献[12]中，作者将 ADMM 与 SAG 梯度估计算子进行结合，ADMM 与 SVRG 的结合可参考文献[13] [14]等。所有这些研究均在凸优化框架下对随机 ADMM 进行了分析。在[15]中，研究者探讨了使用三种不同梯度估计(SGD、SVRG、SAGA)的随机 ADMM 方法来解决非凸非光滑优化问题，随之，文章[16]中提出了框架形式的随机 ADMM 算法，对大规模非凸优化问题进行研究。

本文提出了“随机镜像下降对称交替方向乘子法(SMD-SADMM)”。首先，该算法通过引入随机方差缩减算子，通过随机选择梯度信息，有效地减少了计算全梯度的需求，特别对于处理大规模数据的优化问题，显著提高了算法的运行效率；其次，算法利用布雷格曼(Bregman)距离定义的邻近项取代二范数，这确保了子问题具有显示解，进而提高了算法的效率；最后，SMD-SADMM 采用了对偶变量的对称更新策略，有助于提升算法的收敛性，使得算法在处理非凸优化问题时表现出了更好的稳定性，从而为求解大规模非凸优化问题提供了一种稳健的解决方案。总体来说，SMD-SADMM 算法结合 Bregman 距离定义的邻近项、对偶变量以对称形式进行更新的迭代形式、方差缩减的随机技巧为解决现代大规模非凸优化问题提供了有力的工具。

2. 基本定义

2.1. 布雷格曼距离与勒让德函数

定义 2.1 [17] 设 $\Omega\subset\mathbb{R}^d$ 为一个非空开凸集，函数 $h:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ 如果满足以下性质：

(i) h 是适当下半连续凸函数，且 $dom\,h\subset\bar{\Omega}$，

(ii) h 在 $int\,dom\,h=\Omega$ 上是连续可微的，且 $dom\,\partial h=\Omega$，则称 h 为勒让德函数。

定义 2.2 [18] 设 h 为勒让德函数，定义与 h 相关的布雷格曼距离为 $D_h:dom\,h\times int\,dom\,h\to\mathbb{R}_+$，

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

2.2. Kurdyka-Lojasiewicz (KL)性质

定义 2.4 [19] 令 $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ 是一个适当下半连续函数，如果存在 $\eta\in(0,+\infty]$，$\bar{x}\in dom\,\partial f$ 的邻域 U，$\varphi:[0,\eta)\to\mathbb{R}_+$ 使得

(i) $\varphi(0)=0$，φ 在 $(0,\eta)$ 上是连续可微函数使得 $\varphi'>0$；

(ii) 对于任意 $x\in U\cap\left[x\in\mathbb{R}^d:f(\bar{x})<f(x)<f(\bar{x})+\eta\right]$，下列 KL 不等式成立：

$$\varphi'\left(f(x)-f(\bar{x})\right)dist\left(0,\partial f(x)\right)\geq 1,$$

则称函数 f 称为在 $\bar{x}\in dom\,\partial f$ 上具有 KL 性质。

注释 2.5 [19] 如下为定义 2.4 的相关说明：

(i) 定义 2.4 中的函数 φ 称为 f 的去奇异化函数；

(ii) 在 $dom\partial f$ 的每个点上满足 KL 不等式的适当下半连续函数称为 KL 函数；

(iii) 半代数函数满足 KL 不等式，其去奇异化函数的形式为 $\varphi(s)=as^{1-\theta}$，其中 $a>0$，$\theta\in[0,1)$ 称为该函数的 KL 指数。

3. 随机镜像下降对称交替方向乘子法及其收敛性

在本节中，我们首先给出随机镜像下降对称交替方向乘子法，然后对其进行收敛性分析，首先我们给出以下假设。

假设 3.1

(i) $\inf\left\{f(x)+g(y): x\in\mathbb{R}^{d_1}, y\in\mathbb{R}^{d_2}\right\}=\underline{\Phi}>-\infty$；

(ii) $f:\mathbb{R}^{d_1}\to\mathbb{R}\cup\{+\infty\}$ 是适当的下半连续函数，且 $g:\mathbb{R}^{d_2}\to\mathbb{R}$ 是一个 L_g 光滑函数，矩阵 B 是列满秩的；

(iii) h 在任意有界区间上是 μ_h 强凸的。

3.1. 随机镜像下降对称交替方向乘子法算法

算法 1：随机镜像下降对称交替方向乘子法。

1. 输入 $\beta>\dfrac{-2r+L_g+\dfrac{V_\Upsilon}{\rho}+V_1+2+\dfrac{8}{s+1}\sqrt{\dfrac{V_\Upsilon}{\rho}+V_1}}{\left(\dfrac{2}{s+1}-1\right)\lambda_{\min}\left(B^{\mathrm{T}}B\right)}$。$s,r\in(0,1)$，$\mu>0$，并初始化 x^0,y^0,λ^0；

2. 当初始条件满足执行：

$$x^{k+1}=\arg\min_x\left\{f(x)+\left(\lambda^k\right)^{\mathrm{T}}\left(Ax+By^k-b\right)+\frac{\beta}{2}\left\|Ax+By^k-b\right\|^2+\mu D_h\left(x,x^k\right)\right\},\tag{2}$$

$$\lambda^{k+\frac{1}{2}}=\lambda^k+s\beta\left(Ax^{k+1}+By^k-b\right),\tag{3}$$

$$y^{k+1}=\arg\min_y\left\{\left\langle\tilde{\nabla}g\left(y^k\right),y-y^k\right\rangle+\left(\lambda^{k+\frac{1}{2}}\right)^{\mathrm{T}}\left(Ax^{k+1}+By-b\right)+\frac{\beta}{2}\left\|Ax^{k+1}+By-b\right\|^2+\frac{r}{2}\left\|y-y^k\right\|^2\right\},\tag{4}$$

$$\lambda^{k+1}=\lambda^{k+\frac{1}{2}}+\beta\left(Ax^{k+1}+By^{k+1}-b\right),\tag{5}$$

3. 当终止条件满足，执行 x^{K+1}，y^{K+1}。

注：$\tilde{\nabla}$ 是具有方差缩减的随机梯度估计算子(见定义 3.1)。

定义 3.1 方差缩减的随机梯度估计算子

记 $\mathbb{E}_k$ 为算法 1 中随机变量前 k 次迭代的条件期望，对于常数 $V_1,V_2,V_\Upsilon\geq0$ 以及 $\rho\in(0,1]$，若下述条件成立，则称梯度估计 $\tilde{\nabla}$ 为方差缩减随机梯度估计算子：

(1) (均方误差有界)存在随机变量序列 $\{\Upsilon_k\}_{k\geq1}$ 以及随机向量 v_k^i，其中 $\Upsilon_k=\sum_{i=1}^{s}\left\|v_k^i\right\|^2$ 使得

$$\mathbb{E}_k\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2\leq\Upsilon_k+V_1\left(\mathbb{E}_k\left\|y^{k+1}-y^k\right\|^2+\left\|y^k-y^{k-1}\right\|^2\right),$$

以及存在 $\Gamma_k=\sum_{i=1}^{s}\left\|v_k^i\right\|$ 使得 $\mathbb{E}_k\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|\leq\Gamma_k+V_2\left(\mathbb{E}_k\left\|y^{k+1}-y^k\right\|+\left\|y^k-y^{k-1}\right\|\right)$ 成立。

(2) (几何迭代) $\mathbb{E}_k\Upsilon_{k+1}\leq(1-\rho)\Upsilon_k+V_\Upsilon\left(\mathbb{E}_k\left\|y^{k+1}-y^k\right\|^2+\left\|y^k-y^{k-1}\right\|^2\right)$；

(3) (估计量的收敛性)如果$\left\{y^k\right\}_{k\in N}$满足$\lim\limits_{k\to+\infty}\mathbb{E}\left\|y^k-y^{k-1}\right\|^2=0$则有$\mathbb{E}\Upsilon_k\to 0$且$\mathbb{E}\Gamma_k\to 0$。

注释 3.2 SAGA 与 SARAH 方差缩减随机梯度估计参数取值情况[16] [20]：

SAGA 梯度估计作为一个有效的方差缩减随机梯度估计算子，其表达式如下：

$$\tilde{\nabla}_{\mathrm{SAGA}}g\left(y^k\right)=\frac{1}{b}\sum_{j\in B_k}\left(\nabla g_j\left(y^k\right)-\nabla g_j\left(\varphi_k^j\right)\right)+\frac{1}{m}\sum_{i=1}^{m}\nabla g_i\left(\varphi_k^i\right),$$

其中B_k是从所有包含b个元素的子集中均匀随机选择的小批量集合，子集的元素包含在$\{1,2,\cdots,m\}$中。根据定义 3.1 计算可知 SAGA 相应参数为$V_1=V_2=0$，$\rho=\dfrac{b}{2m}$以及$V_\Upsilon=\left(1+\dfrac{2m}{b}\right)L_g^2$。

另一个常见的随机梯度估计为 SARAH 随机梯度估计，

$$\tilde{\nabla}_{\mathrm{SARAH}}g\left(y^k\right)=\begin{cases}\nabla g\left(y^k\right), & \text{以概率}\dfrac{1}{p},\\ \dfrac{1}{b}\sum\limits_{j\in B_k}\left(\nabla g_j\left(y^k\right)-\nabla g_j\left(y^{k-1}\right)\right)+\tilde{\nabla}_{\mathrm{SARAH}}g\left(y^{k-1}\right), & \text{否则.}\end{cases}$$

SARAH 梯度估计相应参数取值为$V_1=V_\Upsilon=L_g^2$，$V_2=L_g$，$\rho=\dfrac{1}{p}$。

3.2. 随机镜像下降对称交替方向乘子法算法的全局收敛性分析

定理 3.1 [21] [上鞅收敛定理] 设$\mathbb{E}_k$为随机镜像下降对称交替方向乘子法算法前k次迭代的条件期望。设$b\in\mathbb{R}$，$\{U_k\}$和$\{V_k\}$分别为取值于$[b,+\infty)$和$[0,+\infty)$的随机变量序列，且U_k和V_k仅依赖于算法的前k次迭代。若对于所有$k\in\mathbb{N}$，

$$\mathbb{E}_kU_{k+1}+V_k\le U_k$$

成立，则几乎必然有$\sum\limits_{k=0}^{+\infty}V_k<+\infty$，并且$U_k$几乎必然收敛到$[b,+\infty)$上的一个随机变量。

定义第$k\in\mathbb{N}$次迭代的李雅普诺夫函数如下：

$$\begin{aligned}&\Psi_k\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)\\&=\Phi_k\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\Upsilon_{k+1}\\&\quad+\frac{4}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\left\|\tilde{\nabla}g\left(y^k\right)-g\left(y^k\right)\right\|^2,\end{aligned}$$

其中$\Upsilon_k,V_1,V_2,V_r,\rho$是方差缩减梯度估计算子相应的随机变量和常数(参考定义 3.1)，李雅普诺夫函数中具体参数以及函数表示如下：

$$\Phi_k\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)=\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)+\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2+G\left\|y^k-y^{k-1}\right\|^2,\ t_1=\frac{1}{\frac{V_\Upsilon}{\rho}+V_1},$$

$$\begin{aligned}G=&-\frac{4r^2}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}-\frac{L_g}{2}+r-\frac{1}{2t_1}+\left(-\frac{1}{2}\beta+\frac{\beta}{s+1}\right)\lambda_{\min}\left(B^{\mathrm{T}}B\right)\\&-\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\left(\frac{V_\Upsilon}{\rho}+V_1\right),\end{aligned}$$

$$\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)=f\left(x^k\right)+g\left(y^k\right)+\left(\lambda^k\right)^{\mathrm{T}}\left(Ax^k+By^k-b\right)+\frac{\beta}{2}\left\|Ax^k+By^k-b\right\|^2.$$

为了便于叙述将李雅普诺夫函数 $\Psi_k\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)$ 记为 Ψ_k，$\Phi_k\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)$ 记为 $\Phi\left(X^k\right)$ 并将 $\left\{\left(x^k,y^k,\lambda^k,x^{k-1},y^{k-1}\right)\right\}_{k\in N}$ 记为 $\left\{X^k\right\}_{k\in N}$。

定理 3.2 设假设 3.1 成立。令 $\left\{\left(x^k,y^k,\lambda^k\right)\right\}_{k\in\mathbb{N}}$ 是由随机镜像下降对称交替方向乘子法算法生成的序列，并假设该序列是有界的。那么：

(i) 序列 $\left\{\Psi_k\right\}_{k\in\mathbb{N}}$ 在期望意义下是单调非增的。特别地，对于任意 $k\in\mathbb{N}$，$t_1>0$，

$$\tilde{\eta}=G-\frac{4\left(L_g+r\right)^2}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}-\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\left(\frac{V_{\Upsilon}}{\rho}+V_1\right)>0,$$

成立：

$$\mathbb{E}_k\Psi_{k+1}\le\Psi_k-\tilde{\eta}\mathbb{E}_k\left\|y^k-y^{k-1}\right\|^2+\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2,$$

(ii) 迭代点间距平方的期望是可和的，即，$\sum_{k=0}^{+\infty}\mathbb{E}\left\|x^k-x^{k-1}\right\|^2<+\infty$，$\sum_{k=0}^{+\infty}\mathbb{E}\left\|y^k-y^{k-1}\right\|^2<+\infty$。此外，当 $k\to+\infty$ 时，有 $\mathbb{E}\left\|x^k-x^{k-1}\right\|\to0$，$\mathbb{E}\left\|y^k-y^{k-1}\right\|\to0$ 且 $\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|\to0$。

(iii) $\sum_{k=1}^{+\infty}\left\|x^k-x^{k-1}\right\|^2<+\infty,\left\|x^k-x^{k-1}\right\|\to0,\sum_{k=1}^{+\infty}\left\|y^k-y^{k-1}\right\|^2<+\infty,\left\|y^k-y^{k-1}\right\|\to0$ 成立，并有 Φ^* 取值于 $[\underline{\Phi},\infty)$ 使得 $\lim_{k\to+\infty}\mathbb{E}\Psi_k=\lim_{k\to+\infty}\mathbb{E}\Phi\left(X^k\right)=\Phi^*$。

证明：(i) 结合迭代(3)和(5)，有：

$$Ax^{k+1}+By^k-b=\frac{1}{(s+1)\beta}\left(\lambda^{k+1}-\lambda^k\right)-\frac{1}{s+1}\left(By^{k+1}-By^k\right),\tag{6}$$

$$Ax^{k+1}+By^{k+1}-b=\frac{1}{(s+1)\beta}\left(\lambda^{k+1}-\lambda^k\right)+\frac{s}{s+1}\left(By^{k+1}-By^k\right).\tag{7}$$

基于迭代(2)，

$$\mathcal{L}_\beta\left(x^{k+1},y^k,\lambda^k\right)+\mu D_h\left(x^{k+1},x^k\right)\le\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^k\right).\tag{8}$$

又 g 是 L_g 梯度利普希茨连续的，从而有

$$\begin{aligned}g\left(y^{k+1}\right)&\le g\left(y^k\right)+\left\langle\nabla g\left(y^k\right),y^{k+1}-y^k\right\rangle+\frac{L_g}{2}\left\|y^k-y^{k+1}\right\|^2\\&=g\left(y^k\right)+\left\langle\nabla g\left(y^k\right)-\tilde{\nabla}g\left(y^k\right),y^{k+1}-y^k\right\rangle+\left\langle\tilde{\nabla}g\left(y^k\right),y^{k+1}-y^k\right\rangle+\frac{L_g}{2}\left\|y^k-y^{k+1}\right\|^2.\end{aligned}\tag{9}$$

显然，

$$\begin{aligned}\mathcal{L}_\beta\left(x^{k+1},y^k,\lambda^k\right)=\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^k\right)+g\left(y^k\right)-g\left(y^{k+1}\right)+\left\langle\lambda^k,By^k-By^{k+1}\right\rangle\\+\frac{\beta}{2}\left\|Ax^{k+1}+By^k-b\right\|^2-\frac{\beta}{2}\left\|Ax^{k+1}+By^{k+1}-b\right\|^2,\end{aligned}$$

结合公式(9)，可以得到：

$$\begin{aligned}\mathcal{L}_\beta\left(x^{k+1},y^k,\lambda^k\right)\geq\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^k\right)+\left\langle\nabla g\left(y^k\right)-\tilde{\nabla}g\left(y^k\right),y^k-y^{k+1}\right\rangle\\-\frac{L_g}{2}\left\|y^k-y^{k+1}\right\|^2+\left\langle\tilde{\nabla}g\left(y^k\right),y^k-y^{k+1}\right\rangle+\left\langle\lambda^k,By^k-By^{k+1}\right\rangle\\+\frac{\beta}{2}\left\|Ax^{k+1}+By^k-b\right\|^2-\frac{\beta}{2}\left\|Ax^{k+1}+By^{k+1}-b\right\|^2.\end{aligned}\tag{10}$$

利用迭代(4)的最优性条件：

$$0\in\tilde{\nabla}g\left(y^k\right)+B^{\mathrm{T}}\lambda^{k+\frac{1}{2}}+\beta B^{\mathrm{T}}\left(Ax^{k+1}+By^{k+1}-b\right)+r\left(y^{k+1}-y^k\right),$$

可得

$$\left\langle\tilde{\nabla}g\left(y^k\right),y^k-y^{k+1}\right\rangle=\left\langle\lambda^{k+\frac{1}{2}},By^{k+1}-By^k\right\rangle+r\left\|y^{k+1}-y^k\right\|^2+\beta\left\langle Ax^{k+1}+By^{k+1}-b,By^{k+1}-By^k\right\rangle,\tag{11}$$

代入公式(10)，

$$\begin{aligned}&\mathcal{L}_\beta\left(x^{k+1},y^k,\lambda^k\right)\\&\geq\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^k\right)+\left\langle\nabla g\left(y^k\right)-\tilde{\nabla}g\left(y^k\right),y^k-y^{k+1}\right\rangle+\left(r-\frac{L_g}{2}\right)\left\|y^k-y^{k+1}\right\|^2\\&+\left\langle By^k-By^{k+1},\lambda^k-\lambda^{k+\frac{1}{2}}+\beta\left(Ax^{k+1}-b\right)+\frac{\beta}{2}\left(By^{k+1}+By^k\right)-\beta\left(Ax^{k+1}+By^{k+1}-b\right)\right\rangle,\end{aligned}$$

同时由于

$$\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^k\right)=\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^{k+1}\right)+\left\langle\lambda^k-\lambda^{k+1},Ax^{k+1}+By^{k+1}-b\right\rangle.\tag{12}$$

将公式(8) (11) (12)相加，并由 h 是 μ_h 强凸可得：

$$\begin{aligned}&\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^{k+1}\right)\\&\leq\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)-\frac{\mu\mu_h}{2}\left\|x^{k+1}-x^k\right\|+\left\langle\nabla g\left(y^k\right)-\tilde{\nabla}g\left(y^k\right),y^{k+1}-y^k\right\rangle\\&+\left(\frac{L_g}{2}-r\right)\left\|y^{k+1}-y^k\right\|^2+\frac{1}{(s+1)\beta}\left\|\lambda^{k+1}-\lambda^k\right\|^2+\left(\frac{\beta}{2}-\frac{\beta}{s+1}\right)\left\|By^{k+1}-By^k\right\|^2.\end{aligned}\tag{13}$$

将迭代(5)代入迭代(4)，并取一阶最优性条件得到 $0=\tilde{\nabla}g\left(y^k\right)+B^{\mathrm{T}}\lambda^{k+1}+r\left(y^{k+1}-y^k\right)$。

由此可得

$$\begin{aligned}\left\|\lambda^{k+1}-\lambda^k\right\|&\leq\frac{1}{\sqrt{\lambda_{\min}\left(B^{\mathrm{T}}B\right)}}\left\|B^{\mathrm{T}}\left(\lambda^{k+1}-\lambda^k\right)\right\|\\&\leq\frac{1}{\sqrt{\lambda_{\min}\left(B^{\mathrm{T}}B\right)}}\left[\left\|\tilde{\nabla}g\left(y^{k-1}\right)-\nabla g\left(y^{k-1}\right)\right\|+\left\|\nabla g\left(y^{k-1}\right)-\nabla g\left(y^k\right)\right\|\right.\\&\left.+\left\|\nabla g\left(y^k\right)-\tilde{\nabla}g\left(y^k\right)\right\|+r\left\|y^k-y^{k-1}\right\|+r\left\|y^{k+1}-y^k\right\|\right].\end{aligned}\tag{14}$$

利用 g 是 L_g 光滑的，从而有：

$$\begin{aligned}\left\|\lambda^{k+1}-\lambda^k\right\|^2\leq\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)}&\left[\left\|\tilde{\nabla}g\left(y^{k-1}\right)-\nabla g\left(y^{k-1}\right)\right\|^2+\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2\right.\\&\left.+\left(L_g+r\right)^2\left\|y^k-y^{k-1}\right\|^2+r^2\left\|y^k-y^{k+1}\right\|^2\right].\end{aligned}\tag{15}$$

将公式(15)代入公式(13)并结合对于任意 $t_1>0$，可得

$$\begin{aligned}&\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^{k+1}\right)\\&\le\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)-\frac{\mu\mu_h}{2}\left\|x^{k+1}-x^k\right\|^2+G_1\left\|y^{k+1}-y^k\right\|^2+\frac{4\left(L_g+r\right)^2}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\left\|y^k-y^{k-1}\right\|^2\\&+\left[\frac{t_1}{2}+\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right]\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2+\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\left\|\tilde{\nabla}g\left(y^{k-1}\right)-\nabla g\left(y^{k-1}\right)\right\|^2.\end{aligned}\tag{16}$$

其中，$G_1=\dfrac{4r^2}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}+\dfrac{L_g}{2}-r+\dfrac{1}{2t_1}-\left(-\dfrac{1}{2}\beta+\dfrac{\beta}{s+1}\right)\lambda_{\min}\left(B^{\mathrm{T}}B\right)$。结合定义 3.1 中的均方误差有界性：

$$\begin{aligned}&\mathbb{E}_k\mathcal{L}_\beta\left(x^{k+1},y^{k+1},\lambda^{k+1}\right)+\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\mathbb{E}_k\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2+\frac{\mu\mu_h}{2}\mathbb{E}_k\left\|x^{k+1}-x^k\right\|^2\\&+G\mathbb{E}_k\left\|y^{k+1}-y^k\right\|^2+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\mathbb{E}_k\Upsilon_{k+1}\\&\le\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)+\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\left\|\tilde{\nabla}g\left(y^{k-1}\right)-\nabla g\left(y^{k-1}\right)\right\|^2+\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2\\&+G\left\|y^k-y^{k-1}\right\|^2+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\Upsilon_k-\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2\\&-\left[G-\frac{4\left(L_g+r\right)^2}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}-\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\left(\frac{V_\Upsilon}{\rho}+V_1\right)\right]\left\|y^k-y^{k-1}\right\|^2,\end{aligned}\tag{17}$$

结合

$$\begin{aligned}\Psi_k=&\mathcal{L}_\beta\left(x^k,y^k,\lambda^k\right)+\frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\left\|\tilde{\nabla}g\left(y^{k-1}\right)-\nabla g\left(y^{k-1}\right)\right\|^2+\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2\\&+G\left\|y^k-y^{k-1}\right\|^2+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)(s+1)\beta}\right)\Upsilon_k,\end{aligned}\tag{18}$$

有以下不等式成立：

$$\mathbb{E}_k\left[\Psi_{k+1}+\tilde{\eta}\left\|y^k-y^{k-1}\right\|^2+\frac{\mu\mu_h}{2}\left\|x^k-x^{k-1}\right\|^2\right]\le\Psi_k.\tag{19}$$

(ii) 对公式(19)两边取期望，得到

$$\mathbb{E}\Psi_{k+1}\le\mathbb{E}\Psi_k-\tilde{\eta}\mathbb{E}\left\|y^k-y^{k-1}\right\|^2-\frac{\mu\mu_h}{2}\mathbb{E}\left\|x^k-x^{k-1}\right\|^2.$$

将上述不等式中 k 从零到 $T-1$ 求和，由于 $\underline{\Phi}=\inf\left\{f(x)+g(y)\right\}>-\infty$，从而有 $\underline{\Phi}=\inf\left\{\Phi\left(X^k\right)\right\}>-\infty$，

$$\frac{\mu\mu_h}{2}\sum_{k=0}^{T-1}\mathbb{E}\left\|x^k-x^{k-1}\right\|^2+\tilde{\eta}\sum_{k=0}^{T-1}\mathbb{E}\left\|y^k-y^{k-1}\right\|^2\le\mathbb{E}\Psi_0-\mathbb{E}\Psi_T\le\Psi_0-\underline{\Phi}<+\infty,$$

令 $T\to+\infty$，可得序列 $\mathbb{E}\left\|x^k-x^{k+1}\right\|^2$ 和 $\mathbb{E}\left\|y^k-y^{k-1}\right\|^2$ 均为可和的，且

$$\lim_{k\to\infty}\mathbb{E}\left\|x^k-x^{k-1}\right\|=0,\ \lim_{k\to\infty}\mathbb{E}\left\|y^k-y^{k-1}\right\|=0.$$

从关系式(15)与方差缩减算子中的均方误差有界性可得

$$\mathbb{E}\left\|\lambda^{k+1}-\lambda^{k}\right\|^{2} \le \frac{4}{\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\mathbb{E}\Big[\Upsilon_{k-1}+2V_1\left\|y^{k}-y^{k-1}\right\|^{2}+V_1\left\|y^{k-1}-y^{k-2}\right\|^{2}$$
$$+V_1\left\|y^{k+1}-y^{k}\right\|^{2}+\left(L_g+r\right)^{2}\left\|y^{k}-y^{k-1}\right\|^{2}+r^{2}\left\|y^{k}-y^{k+1}\right\|^{2}\Big].$$

因此有 $\lim_{k\to\infty}\mathbb{E}\left\|\lambda^{k}-\lambda^{k-1}\right\|=0$。

(iii) 由(i)可得

$$\mathbb{E}_k\left[\Psi_{k+1}+\tilde{\eta}\left\|y^{k}-y^{k-1}\right\|^{2}+\frac{\mu\mu_h}{2}\left\|x^{k}-x^{k-1}\right\|^{2}\right]\le\Psi_k.$$

上鞅收敛定理表明 $\sum_{k=1}^{+\infty}\left\|x^{k}-x^{k-1}\right\|^{2}<+\infty,\sum_{k=1}^{+\infty}\left\|y^{k}-y^{k-1}\right\|^{2}<+\infty$ 几乎必然成立。因此 $\lim_{k\to\infty}\left\|x^{k}-x^{k-1}\right\|=0,\lim_{k\to\infty}\left\|y^{k}-y^{k-1}\right\|=0$ 几乎必然成立。上鞅收敛定理还保证了 Ψ_k 几乎必然收敛到一个有限值 Φ^{*}，并有 $\lim_{k\to\infty}\mathbb{E}\Psi_k=\lim_{k\to\infty}\mathbb{E}\Phi\left(X^{k}\right)=\Phi^{*}\in\left[\underline{\Phi},\infty\right)$ 成立。 □

命题 3.3 设假设 3.1 成立。令 $\left\{\left(x^{k},y^{k},\lambda^{k}\right)\right\}_{k\in\mathbb{N}}$ 为随机镜像下降对称交替方向乘子法算法生成的序列，且假设该序列有界。对于所有 $k\ge0$，定义如下向量：

$$\zeta_x^{k}=A^{\mathrm{T}}\left(\lambda^{k}-\lambda^{k-1}\right)+\beta A^{\mathrm{T}}\left(By^{k}-By^{k-1}\right)+\mu\mu_h\left(x^{k}-x^{k-1}\right)-\mu\left(\nabla h\left(x^{k}\right)-\nabla h\left(x^{k-1}\right)\right),\tag{20}$$

$$\zeta_y^{k}=\left(\nabla g\left(y^{k}\right)-\tilde{\nabla}g\left(y^{k}\right)\right)-\left(B^{\mathrm{T}}\lambda^{k+1}-B^{\mathrm{T}}\lambda^{k}\right)+\frac{1}{s+1}\left(B^{\mathrm{T}}\lambda^{k}-B^{\mathrm{T}}\lambda^{k-1}\right)$$
$$+\frac{(s+1)\beta-\beta}{s+1}\left(B^{\mathrm{T}}By^{k}-B^{\mathrm{T}}By^{k-1}\right)+2G\left(y^{k}-y^{k-1}\right)-r\left(y^{k+1}-y^{k}\right),\tag{21}$$

$$\zeta_\lambda^{k}=\frac{1}{(s+1)\beta}\left(\lambda^{k}-\lambda^{k-1}\right)+\frac{s}{s+1}\left(By^{k}-By^{k-1}\right),\ \zeta_{x'}^{k}=-\mu\mu_h\left(x^{k}-x^{k-1}\right),\ \zeta_{y'}^{k}=-2G\left(y^{k}-y^{k-1}\right),\tag{22}$$

对于 $\left(\zeta_x^{k},\zeta_y^{k},\zeta_\lambda^{k},\zeta_{x'}^{k},\zeta_{y'}^{k}\right)$，有以下性质成立：

(i) $\left(\zeta_x^{k},\zeta_y^{k},\zeta_\lambda^{k},\zeta_{x'}^{k},\zeta_{y'}^{k}\right)\in\partial\Phi\left(X^{k}\right)$，存在正常数 P，使得对于任意 $k\ge1$，有：

$$\mathbb{E}_k\left\|\left(\zeta_x^{k},\zeta_y^{k},\zeta_\lambda^{k},\zeta_{x'}^{k},\zeta_{y'}^{k}\right)\right\|\le P\mathbb{E}_k\left(\left\|y^{k+1}-y^{k}\right\|+\left\|y^{k}-y^{k-1}\right\|+\left\|\lambda^{k+1}-\lambda^{k}\right\|\right.$$
$$\left.+\left\|\lambda^{k}-\lambda^{k-1}\right\|+\left\|x^{k}-x^{k-1}\right\|\right)+\Gamma_k,$$

(ii) $\mathbb{E}\mathrm{dist}\left(0,\partial\Phi\left(X^{k}\right)\right)\to0$。

证明：(i) 由 Φ 的定义，对于任意 $k\ge1$有：

$$\partial_x\Phi^{k}=\partial f\left(x^{k}\right)+A^{\mathrm{T}}\lambda^{k}+\beta A^{\mathrm{T}}\left(Ax^{k}+By^{k}-b\right)+\mu\mu_h\left(x^{k}-x^{k-1}\right),\tag{23}$$

$$\partial_y\Phi^{k}=\nabla g\left(y^{k}\right)+B^{\mathrm{T}}\lambda^{k}+\beta B^{\mathrm{T}}\left(Ax^{k}+By^{k}-b\right)+2G\left(y^{k}-y^{k-1}\right),\tag{24}$$

$$\partial_\lambda\Phi^{k}=Ax^{k}+By^{k}-b,\partial_{x'}\Phi^{k}=-\mu\mu_h\left(x^{k}-x^{k-1}\right),\partial_{y'}\Phi^{k}=-2G\left(y^{k}-y^{k-1}\right).\tag{25}$$

由(2)的一阶最优性条件可得 $-A^{\mathrm{T}}\lambda^{k-1}-\beta A^{\mathrm{T}}\left(Ax^{k}+By^{k-1}-b\right)-\mu\left(\nabla h\left(x^{k}\right)-\nabla h\left(x^{k-1}\right)\right)\in\partial f\left(x^{k}\right)$，将其代入公式(23)可得 $\zeta_x^{k}\in\partial_x\Phi\left(X^{k}\right)$。同样，结合 $0=\tilde{\nabla}g\left(y^{k}\right)+B^{\mathrm{T}}\lambda^{k+1}+r\left(y^{k+1}-y^{k}\right)$ 和公式(7)，可以得到 $\zeta_y^{k}\in\partial_y\Phi\left(X^{k}\right)$。通过(6)，也可以得到 $\zeta_\lambda^{k}\in\partial_\lambda\Phi\left(X^{k}\right)$，从而有 $\left(\zeta_x^{k},\zeta_y^{k},\zeta_\lambda^{k},\zeta_{x'}^{k},\zeta_{y'}^{k}\right)\in\partial\Phi\left(X^{k}\right)$。

根据 h 是 L_h 光滑的，结合定义 3.1 中的均方误差(MSE)有界以及等式(23)~(25)，我们可以推导出以

下式子成立：

$$\mathbb{E}_k\left\|\zeta_x^k,\zeta_y^k,\zeta_\lambda^k,\zeta_{x'}^k,\zeta_{y'}^k\right\|\le P\mathbb{E}_k\left(\left\|\lambda^{k+1}-\lambda^k\right\|+\left\|\lambda^k-\lambda^{k-1}\right\|+\left\|y^{k+1}-y^k\right\|+\left\|y^k-y^{k-1}\right\|+\left\|x^k-x^{k-1}\right\|\right),$$

其中

$$P=\max\left\{\|B\|,2\mu\mu_h+\mu_hL_h,V_2+r,\|A\|+\frac{\|B\|}{s+1}+\frac{1}{(s+1)\beta},\|\beta A^{\mathrm{T}}B\|+\frac{s\beta}{s+1}\|B^{\mathrm{T}}B\|+4\|G\|+\frac{s}{s+1}\|B\|+V_2\right\}.$$

(ii) 由于

$$\mathbb{E}_k\left\|\zeta_x^k,\zeta_y^k,\zeta_\lambda^k,\zeta_{x'}^k,\zeta_{y'}^k\right\|\le P\mathbb{E}_k\left(\left\|\lambda^{k+1}-\lambda^k\right\|+\left\|\lambda^k-\lambda^{k-1}\right\|+\left\|y^{k+1}-y^k\right\|+\left\|y^k-y^{k-1}\right\|+\left\|x^k-x^{k-1}\right\|\right).$$

由引理 3.2，得 $\mathbb{E}\left\|x^k-x^{k-1}\right\|\to 0$，$\mathbb{E}\left\|y^k-y^{k-1}\right\|\to 0$，$\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|\to 0$，因此结论成立。 □

记由随机镜像下降对称交替方向乘子法算法生成的序列 $\left\{X^k\right\}_{k\in\mathbb{N}}$ 的极限点集合为 Ω^*，即：

$$\Omega^*=\left\{X^*:\text{存在递增的整数序列}\left\{k_q\right\}_{q\in\mathbb{N}},q\to+\infty\text{ 时 }X^{k_q}\to X^*\right\}$$

定理 3.4 设假设 3.1 成立。令 $\left\{\left(x^k,y^k,\lambda^k\right)\right\}_{k\in\mathbb{N}}$ 为随机镜像下降对称交替方向乘子法算法生成的序列，且假设该序列有界。则以下结论成立：

(i) Ω^* 非空，几乎必然是紧的且连通的。此外，$\operatorname{dist}\left(X^k,\Omega^*\right)\to 0$ 几乎必然成立；

(ii) 对于所有 $X^*\in\Omega^*$，有 $\mathbb{E}\operatorname{dist}\left(0,\partial\Phi\left(X^*\right)\right)=0$，并且 $\mathbb{E}\Phi\left(X^*\right)=\Phi^*$。

证明：(i) 详细证明参考文献[22]。

(ii) 取任意点 $X^*\in\Omega^*$，存在子序列 $\left\{X^{k_q}\right\}_{q\in\mathbb{N}}$ 满足：当 $q\to+\infty$ 时，$X^{k_q}\to X^*$。由于 f 是适当下半连续函数，因此：

$$\liminf_{q\to+\infty} f\left(x^{k_q}\right)\ge f\left(x^*\right).$$

将 $x=x^*$ 代入迭代公式(2)，结合等式(6)和 $\left\|\lambda^{k_q}-\lambda^{k_q-1}\right\|\to 0$，可以得到 $Ax^{k_q}+By^{k_q-1}-b\to 0$。令 $k=k_q-1$，并取极限 $q\to+\infty$ 可得 $\limsup_{q\to+\infty} f\left(x^{k_q}\right)\le f\left(x^*\right)$。结合 $\liminf_{q\to+\infty} f\left(x^{k_q}\right)\ge f\left(x^*\right)$ 可得当 $q\to+\infty$ 时 $f\left(x^{k_q}\right)\to f\left(x^*\right)$。又 g 是连续函数，得 $\lim_{q\to+\infty}\Phi\left(X^{k_q}\right)=\Phi\left(X^*\right)$。由命题 3.3 和 $\partial\Phi$ 的闭性，可以得到：$\mathbb{E}\operatorname{dist}\left(0,\partial\Phi\left(X^*\right)\right)=0$。

最后，证明 Φ 在 Ω^* 上具有常数期望值。取任意点 $X^*\in\Omega^*$，存在子序列 $\left\{X^{k_q}\right\}_{q\in\mathbb{N}}$ 满足当 $q\to+\infty$ 时 $X^{k_q}\to X^*$。根据引理 3.2，$\lim_{k\to+\infty}\mathbb{E}\Phi\left(X^k\right)=\Phi^*$，这意味着 $\lim_{q\to+\infty}\mathbb{E}\Phi\left(X^{k_q}\right)=\Phi^*$。结合 $q\to+\infty$ 时 $\Phi\left(X^{k_q}\right)\to\Phi\left(X^*\right)$，对于任意 $X^*\in\Omega^*$，有 $\mathbb{E}\Phi\left(X^*\right)=\Phi^*$。 □

定理 3.5 设假设 3.1 成立。令 $\{(x^k,y^k,\lambda^k)\}_{k\in\mathbb{N}}$ 为随机镜像下降对称交替方向乘子法算法生成的序列，且假设该序列有界。设 Φ 为一个半代数函数，则存在常数 k_1，$a>0$，$\theta\in[0,1)$ 和一个去奇异化函数 $\phi_0=ar^{1-\theta}$ 使得以下不等式成立：

$$\phi_0'\left(\mathbb{E}\left[\Phi\left(X^k\right)-\Phi_k^*\right]\right)\mathbb{E}\operatorname{dist}\left(0,\partial\Phi\left(X^k\right)\right)\ge 1,\quad\forall k\ge k_1,$$

其中 Φ_k^* 是一个单调递增的序列，收敛于某个 $\mathbb{E}\Phi\left(X^*\right)$，其中 $X^*\in\Omega^*$。

定理 3.6 设假设 3.1 成立。令 $\left\{\left(x^k,y^k,\lambda^k\right)\right\}_{k\in\mathbb{N}}$ 为随机镜像下降对称交替方向乘子法算法生成的序列，且假设该序列有界。设 Φ 为一个半代数函数，则有

$$\sum_{k=0}^{+\infty}\mathbb{E}\left\|x^{k+1}-x^k\right\|<+\infty,\quad \sum_{k=0}^{+\infty}\mathbb{E}\left\|y^{k+1}-y^k\right\|<+\infty,\quad \sum_{k=l}^{+\infty}\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|<+\infty.$$

证明：根据引理 3.2，$\lim\limits_{k\to+\infty}\mathbb{E}\Psi_k=\lim\limits_{k\to+\infty}\mathbb{E}\Phi\left(X^k\right)=\Phi^*$ 成立。因此，我们需要考虑以下两种情况。

第一种情况，即存在整数 $\bar{k}$，使得对于任意 $k\geq\bar{k}$，有 $\mathbb{E}\Psi_k=\Phi^*$ 成立。因此，对于任意 $k\geq\bar{k}$，由詹森不等式可得

$$\mathbb{E}\left\|x^k-x^{k+1}\right\|\leq\sqrt{\mathbb{E}\left\|x^k-x^{k+1}\right\|^2}=0,\mathbb{E}\left\|y^k-y^{k+1}\right\|\leq\sqrt{\mathbb{E}\left\|y^k-y^{k+1}\right\|^2}=0,\mathbb{E}\left\|\lambda^k-\lambda^{k+1}\right\|\leq\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k+1}\right\|^2}=0.$$

此时结论自然成立。

另一种情况，即对于所有 $k\geq0$，都有 $\mathbb{E}\Psi_k>\Phi^*$。命题 3.3 给出了 $\mathbb{E}\mathrm{dist}\left(0,\partial\Phi\left(X^k\right)\right)$ 的一个上界：

$$\begin{aligned}&\mathbb{E}\mathrm{dist}\left(0,\partial\Phi\left(X^k\right)\right)\leq\mathbb{E}\left\|\zeta_x^k,\zeta_y^k,\zeta_\lambda^k,\zeta_{x'}^k,\zeta_{y'}^k\right\|\\&\leq P\mathbb{E}\left(\left\|y^{k+1}-y^k\right\|+\left\|y^k-y^{k-1}\right\|+\left\|\lambda^{k+1}-\lambda^k\right\|+\left\|\lambda^k-\lambda^{k-1}\right\|+\left\|x^k-x^{k-1}\right\|\right)+\mathbb{E}\Gamma_k\\&\leq P\left(\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|\lambda^{k+1}-\lambda^k\right\|^2}+\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}+\sqrt{s\mathbb{E}\Upsilon_k}\right.,\end{aligned}\tag{26}$$

其中，最后一个不等式是由 $\mathbb{E}\Gamma_k=\mathbb{E}\sum\limits_{i=1}^{s}\left\|v_k^i\right\|\leq\mathbb{E}\sqrt{s\sum\limits_{i=1}^{s}\left\|v_k^i\right\|^2}\leq\sqrt{s\mathbb{E}\Upsilon_k}$ 得到的。下面基于 $\Upsilon_{k-1}\geq0$，$V_\Upsilon\geq0$，以及 $\rho\in(0,1]$，$\sqrt{1-\rho}=1-\dfrac{\rho}{2}-\sum\limits_{k=2}^{+\infty}\dfrac{(2k-3)!!}{(2k)!!}\rho^k$ 以及估计量的几何衰减性质对 $\sqrt{\mathbb{E}\Upsilon_k}$ 进行更精确的估计，

$$\begin{aligned}\sqrt{\mathbb{E}\Upsilon_k}&\leq\sqrt{(1-\rho)\mathbb{E}\Upsilon_{k-1}+V_\Upsilon\mathbb{E}\left[\left\|y^k-y^{k-1}\right\|^2+\left\|y^{k-1}-y^{k-2}\right\|^2\right]}\\&\leq\left(1-\frac{\rho}{2}\right)\sqrt{\mathbb{E}\Upsilon_{k-1}}+\sqrt{V_\Upsilon}\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\sqrt{V_\Upsilon}\sqrt{\mathbb{E}\left\|y^{k-1}-y^{k-2}\right\|^2},\end{aligned}$$

进而 $\sqrt{\mathbb{E}\Upsilon_k}\leq\dfrac{2}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_k}-\sqrt{\mathbb{E}\Upsilon_{k+1}}\right)+\dfrac{2\sqrt{V_\Upsilon}}{\rho}\left(\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\right)$。

将公式 $\sqrt{\mathbb{E}\Upsilon_k}\leq\dfrac{2}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_k}-\sqrt{\mathbb{E}\Upsilon_{k+1}}\right)+\dfrac{2\sqrt{V_\Upsilon}}{\rho}\left(\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\right)$ 代入公式(32)中，得到：

$$\begin{aligned}&\mathbb{E}\mathrm{dist}\left(0,\partial\Phi\left(X^k\right)\right)\\&\leq P\sqrt{\mathbb{E}\left\|\lambda^{k+1}-\lambda^k\right\|^2}+\left(P+\frac{2\sqrt{sV_\Upsilon}}{\rho}\right)\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+P\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}\\&+P\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}+\left(P+\frac{2\sqrt{sV_\Upsilon}}{\rho}\right)\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_k}-\sqrt{\mathbb{E}\Upsilon_{k+1}}\right).\end{aligned}$$

令

$$\begin{aligned}K_1&=P+\frac{2\sqrt{sV_\Upsilon}}{\rho},\\C_k&=\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_k}-\sqrt{\mathbb{E}\Upsilon_{k+1}}\right)+K_1\left(\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\right.\\&\left.+\sqrt{\mathbb{E}\left\|\lambda^{k+1}-\lambda^k\right\|^2}+\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}\right),\end{aligned}$$

可得存在一个正常数 P_1，使得对于任意 $k \geq k_0$，

$$C_k \leq P_1\left(\sqrt{\mathbb{E}\Upsilon_k}-\sqrt{\mathbb{E}\Upsilon_{k+1}}\right)+P_1\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+P_1\sqrt{\mathbb{E}\left\|x^{k+1}-y^{k+1}\right\|^2}-2P_1\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2},$$

此外可以得到 $\mathbb{E}\operatorname{dist}\left(0,\partial\Phi\left(X^k\right)\right)\leq C_k$。

根据引理 3.5，对于任意 $k \geq k_1$，$\phi_0'\left(\mathbb{E}\left[\Phi\left(X^k\right)-\Phi_k^*\right]\right)C_k \geq 1$，根据 C_k 的定义有：

$$\begin{aligned}C_k &\geq c_1\left(\sqrt{\mathbb{E}\Upsilon_k}+\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|\lambda^{k+1}-\lambda^k\right\|^2}\right.\\&\quad\left.+\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}\right)\\&\geq c_1\left(\sqrt{\mathbb{E}\Upsilon_k}+\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\right),\end{aligned}$$

其中 $c_1>0$。结合 $\mathbb{E}\left\|x^k-x^{k-1}\right\|^2\to 0$，$\mathbb{E}\left\|y^k-y^{k-1}\right\|^2\to 0$，$\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2\to 0$ 且 $\theta\in[0,1)$，可知存在正常数 k_2 和正常数 c_2,c_3 使得对于所有 $k\geq k_2$，有：

$$\begin{aligned}&\left(\mathbb{E}\left[\frac{4}{(s+1)\beta\lambda_{\min}}\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\Upsilon_{k+1}\right]\right)^\theta\\&\leq\left(\mathbb{E}\left[\frac{4}{(s+1)\beta\lambda_{\min}}\Upsilon_k+\frac{4V_1}{(s+1)\beta\lambda_{\min}}\left(\left\|y^{k+1}-y^k\right\|^2+\left\|y^k-y^{k-1}\right\|^2\right)+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\Upsilon_{k+1}\right]\right)^\theta\\&\leq\left(\mathbb{E}\left[\left(\frac{4}{(s+1)\beta\lambda_{\min}}+\frac{1-\rho}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\right)\Upsilon_k\right.\right.\\&\quad\left.\left.+\left(\frac{4V_1}{(s+1)\beta\lambda_{\min}}+\frac{V_\Upsilon}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\right)\left(\left\|y^{k+1}-y^k\right\|^2+\left\|y^k-y^{k-1}\right\|^2\right)\right]\right)^\theta\\&\leq c_2\left(\sqrt{\mathbb{E}\Upsilon_k}+\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}\right)^\theta\leq\frac{c_2c_3}{c_1^\theta}C_k,\end{aligned}$$

定义 $\bar{C}_k=\frac{4}{(s+1)\beta\lambda_{\min}}\left\|\tilde{\nabla}g\left(y^k\right)-\nabla g\left(y^k\right)\right\|^2+\frac{1}{\rho}\left(\frac{t_1}{2}+\frac{8}{(s+1)\beta\lambda_{\min}\left(B^{\mathrm{T}}B\right)}\right)\Upsilon_{k+1}$，由于 $\left(\mathbb{E}\bar{C}_k\right)^\theta$ 比 C_k 小，从而存在常数 $\hat{c}<\frac{c_2c_3}{c_1^\theta}$ 使得 $\frac{\hat{c}a(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi\left(X^k\right)-\Phi_k^*\right]\right)^\theta+\left(\mathbb{E}\bar{C}_k\right)^\theta}\geq 1$，已知对任意的 $a,b>0$，$\theta\in[0,1)$，有 $(a+b)^\theta\leq a^\theta+b^\theta$ 成立，且 $\frac{\hat{c}a(1-\theta)C_k}{\left(\mathbb{E}\left[\Psi_k-\Phi_k^*\right]\right)^\theta}=\frac{\hat{c}a(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi\left(X^k\right)-\Phi_k^*+\bar{C}_k\right]\right)^\theta}\geq\frac{\hat{c}a(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi\left(X^k\right)-\Phi_k^*\right]\right)^\theta+\left(\mathbb{E}\bar{C}_k\right)^\theta}\geq 1$，因此存在 $\phi=\hat{c}ar^{1-\theta}$ 与 $l=\max\{k_0,k_1,k_2\}$，使得

$$\phi'\left(\mathbb{E}\left[\Psi_k-\Phi_k^*\right]\right)C_k\geq 1,\ \forall k\geq l. \tag{27}$$

结合 ϕ 的凹性，以及 Φ_k^* 是单调递增的，

$$\begin{aligned}&\phi\left(\mathbb{E}\left[\Psi_k-\Phi_k^*\right]\right)-\phi\left(\mathbb{E}\left[\Psi_{k+1}-\Phi_{k+1}^*\right]\right)\\&\geq\phi'\left(\mathbb{E}\left[\Psi_k-\Phi_k^*\right]\right)\mathbb{E}\left[\Psi_k-\Phi_k^*+\Phi_{k+1}^*-\Psi_{k+1}\right]\\&\geq\phi'\left(\mathbb{E}\left[\Psi_k-\Phi_k^*\right]\right)\mathbb{E}\left[\Psi_k-\Psi_{k+1}\right],\end{aligned}\tag{28}$$

令 $\Delta_{p,q}=\phi\left(\mathbb{E}\left[\Psi_p-\Phi_p^*\right]\right)-\phi\left(\mathbb{E}\left[\Psi_q-\Phi_q^*\right]\right)$，结合公式(19) (27)和(28)，

$$\Delta_{k,k+1}C_k\geq\tilde{\eta}\mathbb{E}\left\|y^k-y^{k-1}\right\|^2+\frac{\mu\mu_h}{2}\mathbb{E}\left\|x^k-x^{k-1}\right\|^2\geq\tilde{K}\left(\mathbb{E}\left\|y^k-y^{k-1}\right\|^2+\mathbb{E}\left\|x^k-x^{k-1}\right\|^2\right),$$

其中 $\tilde{K}=\min\left\{\tilde{\eta},\frac{\mu\mu_h}{2}\right\}$。因此，

$$2\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2+\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\leq2\sqrt{\frac{1}{\tilde{K}}\Delta_{k,k+1}C_k}\leq\frac{C_k}{2P_1}+\frac{2P_1\Delta_{k,k+1}}{\tilde{K}}.\tag{29}$$

将不等式(29)中 k 从 l 到 K 进行求和得到：

$$\begin{aligned}&2\sum_{k=l}^{K}\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}+2\sum_{k=l}^{K}\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}\\&\leq\sum_{k=l}^{K}\left(\frac{1}{2}\sqrt{\mathbb{E}\left\|y^{k+1}-y^k\right\|^2}+\frac{1}{2}\sqrt{\mathbb{E}\left\|x^{k+1}-x^k\right\|^2}-\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}\right)\\&\quad+\frac{1}{2}\left(\sqrt{\mathbb{E}\Upsilon_l}-\sqrt{\mathbb{E}\Upsilon_{K+1}}\right)+\frac{2P_1}{\tilde{K}}\phi\left(\mathbb{E}\left[\Psi_l-\Phi_l^*\right]\right).\end{aligned}$$

因此，

$$\begin{aligned}&\sum_{k=l}^{K}\mathbb{E}\left\|x^k-x^{k-1}\right\|+\sum_{k=l}^{K}\mathbb{E}\left\|y^k-y^{k-1}\right\|+\sum_{k=l}^{K}\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|\\&\leq\sum_{k=l}^{K}\sqrt{\mathbb{E}\left\|x^k-x^{k-1}\right\|^2}+\sum_{k=l}^{K}\sqrt{\mathbb{E}\left\|y^k-y^{k-1}\right\|^2}+\sum_{k=l}^{K}\sqrt{\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|^2}\\&\leq\frac{1}{2}\sqrt{\mathbb{E}\left\|y^{K+1}-y^K\right\|^2}-\frac{1}{2}\sqrt{\mathbb{E}\left\|y^l-y^{l-1}\right\|^2}+\frac{1}{2}\sqrt{\mathbb{E}\left\|x^{K+1}-x^K\right\|^2}\\&\quad-\frac{1}{2}\sqrt{\mathbb{E}\left\|x^l-x^{l-1}\right\|^2}+\frac{1}{2}\left(\sqrt{\mathbb{E}\Upsilon_l}-\sqrt{\mathbb{E}\Upsilon_{K+1}}\right)+\frac{2P_1}{\tilde{K}}\phi\left(\mathbb{E}\left[\Psi_l-\Phi_l^*\right]\right),\end{aligned}\tag{30}$$

公式(30)中第一个不等式由詹森不等式得出。令 $K\to+\infty$，

$$\sum_{k=l}^{+\infty}\mathbb{E}\left\|x^k-x^{k-1}\right\|<+\infty,\ \sum_{k=l}^{+\infty}\mathbb{E}\left\|y^k-y^{k-1}\right\|<+\infty,\ \sum_{k=l}^{+\infty}\mathbb{E}\left\|\lambda^k-\lambda^{k-1}\right\|<+\infty.$$ □

4. 数值实验

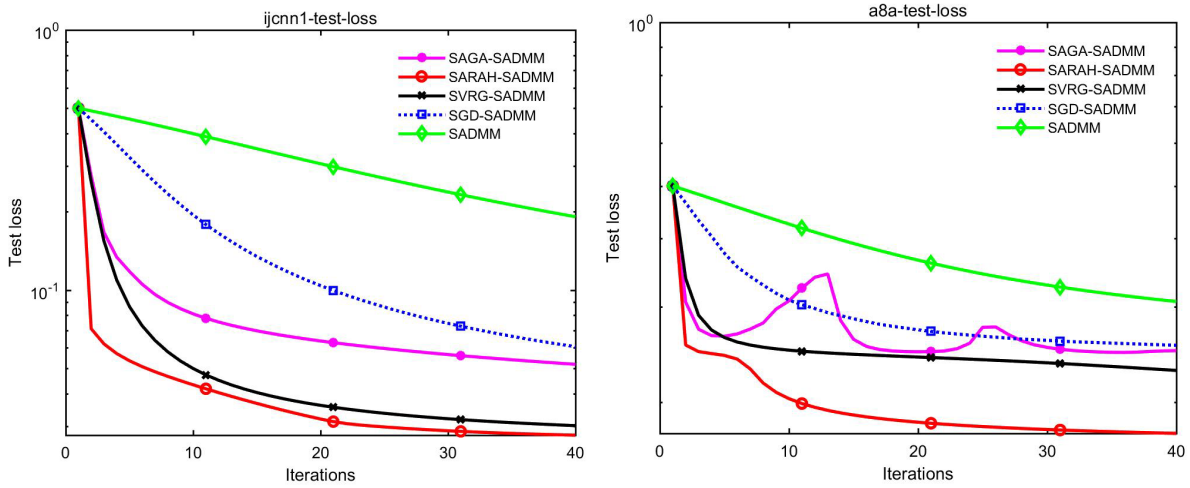
本节中，我们研究算法 1 在图引导融合 lasso 问题上的数值性能。数值实验在 MATLAB R2017a 环境下，配置 Intel Core i7-13700H 处理器(2.40 GHz)和 16 GB 内存的 64 位电脑上进行。确定性对称 ADMM 记为 SADMM，并分别将使用 SGD、SARAH、SAGA、SVRG 方差缩减随机梯度估计算子的 SADMM 分别记为 SGD-SADMM、SARAH-SADMM、SAGA-SADMM、SVRG-SADMM。

给定一组训练样本 $\left\{\left(a_i,b_i\right)\right\}_{i=1}^n$，其中 $a_i\in\mathbb{R}^m$，$b_i\in\{-1,+1\}$，$i\in\{1,2,\cdots,n\}$，图引导的融合 lasso 如下：$\min_y\frac{1}{n}\sum_{i=1}^{n}g_i\left(y\right)+\lambda_1\left\|By\right\|_1$，其中：$g_i\left(y\right)=\frac{1}{1+\exp\left(b_ia_i^\top y\right)}$ 是非凸非光滑的 sigmoid 损失函数，λ_1 是正则化

参数。矩阵 B 的形式为 $B=[G;I]$，其中 G 是通过稀疏逆协方差矩阵估计得到的[4]。在实验中，设定 $g(y)=\frac{1}{n}\sum_{i=1}^{n}g_i(y)$，$f(By)=\lambda_1\|By\|_1$。令 $n=1000$, $s=0.95$，$\beta=1$，$\mu=r=0.05$，$\lambda_1=1\mathrm{e}-5$。使用两个公开数据集[3]，如表 1 所示。并在图 1、图 2 分别给出了损失函数随迭代次数与迭代时间变化关系图。

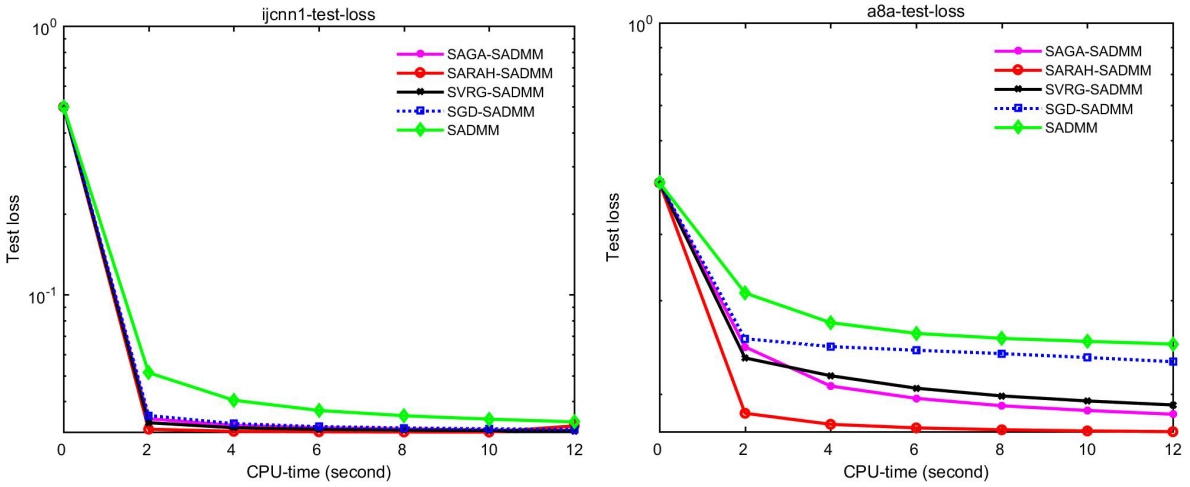
Table 1. Datasets for graph-guided fused lasso
表 1. Graph-guided fused lasso 数据集

数据集	训练集	测试集	分类
a8a	11,348	11,348	2
Ijcnn1	17,500	17,500	2



(a) ijcnn1　　(b) a8a

Figure 1. Relationship between iteration number and loss function variation
图 1. 迭代次数与损失函数变化关系图



(a) ijcnn1　　(b) a8a

Figure 2. Relationship between CPU-time and loss function variation
图 2. 迭代时间与损失函数变化关系图

在图 1 中，我们给出了不同方法在前 40 次迭代下损失函数的测试结果，结果显示在相同迭代次数下 SARAH-ADMM 算法损失函数的下降量最大。图 2 展示了 SADMM 与几种随机算法在相同时间内损失函数变化情况，我们可以观察到：SARAH-SADMM、SVRG-SADMM 在相同时间内损失函数下降最大，而 SAGA-SADMM、SGD-SADMM 的表现相似，且都比 SADMM 效果显著。从而可得随机形式 SADMM 在图 lasso 问题上效果明显优于确定形式的 SADMM。至此实验阐述了不同类型的方差缩减算子与对称 ADMM 在四个公开数据集上的数值表现，并且基于问题的特殊性将 Bregman 距离选取为二范数形式即可得到问题的显示解，相比于其他类型 legendre 函数选取的方式简单高效。

5. 结论

本文提出的“随机镜像下降对称交替方向乘子法”为求解带有等式约束的非凸非光滑优化问题提供了一种高效稳定的方案。理论分析表明，在目标函数满足半代数性质的条件下，算法生成的迭代序列全局收敛到原问题的驻点。数值实验进一步验证了该算法在实际应用中的高效性与稳定性。在之后的研究中考虑将广义惯性步加入文中的算法中，观察不同惯性步参数对数值效果的影响，进而与文中算法的数值效果进行对比。

参考文献

[1] Hsieh, C., Sustik, M., Dhillon, I. and Ravikumar, P. (2014) Quadratic Approximation for Sparse Inverse Covariance Estimation. *Journal of Machine Learning Research*, **15**, 2911-2947.

[2] Papanikolopoulos, N.P. and Khosla, P.K. (1993) Adaptive Robotic Visual Tracking: Theory and Experiments. *IEEE Transactions on Automatic Control*, **38**, 429-445. https://doi.org/10.1109/9.210141

[3] Kazem, R., Suad, J. and Abdulbaqi, H. (2021) Super-Resolution Using 3D Convolutional Neural Networks in CT Scan Image of COVID19. *Turkish Journal of Computer and Mathematics Education*, **12**, 4408-4415.

[4] Friedman, J., Hastie, T. and Tibshirani, R. (2007) Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**, 432-441. https://doi.org/10.1093/biostatistics/kxm045

[5] Xu, J. and Chao, M. (2021) An Inertial Bregman Generalized Alternating Direction Method of Multipliers for Nonconvex Optimization. *Journal of Applied Mathematics and Computing*, **68**, 1-27. https://doi.org/10.1007/s12190-021-01590-1

[6] Yin, J., Tang, C., Jian, J. and Huang, Q. (2024) A Partial Bregman ADMM with a General Relaxation Factor for Structured Nonconvex and Nonsmooth Optimization. *Journal of Global Optimization*, **89**, 899-926. https://doi.org/10.1007/s10898-024-01384-2

[7] Glowinski, R., Karkkainen, T. and Majava, K. (2003) On the Convergence of Operator-Splitting Methods. *Proceedings of CIMNE* 2003: *Numerical Methods for Scientific Computing, Variational Problems and Applications*, Barcelona, Spain. 67-79.

[8] He, B., Liu, H., Wang, Z. and Yuan, X. (2014) A Strictly Contractive Peaceman—Rachford Splitting Method for Convex Programming. *SIAM Journal on Optimization*, **24**, 1011-1040. https://doi.org/10.1137/13090849x

[9] Schmidt, M., Le Roux, N. and Bach, F. (2016) Erratum To: Minimizing Finite Sums with the Stochastic Average Gradient. *Mathematical Programming*, **162**, 113-113. https://doi.org/10.1007/s10107-016-1051-1

[10] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. *Advances in Neural Information Processing Systems*, **26**, 315-323.

[11] Nguyen, L., Liu, J. and Scheinberg, K. (2017) SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. *International Conference on Machine Learning*, **70**, 2613-2621.

[12] Zhong, W. and Kwok, J. (2014) Fast Stochastic Alternating Direction Method of Multipliers. *International Conference on Machine Learning*, **32**, 46-54.

[13] Zheng, S. and Kwok, J. (2016) Fast-and-Light Stochastic ADMM. *International Joint Conference on Artificial Intelligence*, **25**, 2407-2613.

[14] Liu, Y., Shang, F. and Cheng, J. (2017) Accelerated Variance Reduced Stochastic ADMM. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 2287-2293. https://doi.org/10.1609/aaai.v31i1.10843

[15] Huang, F. and Chen, S. (2018) Mini-Batch Stochastic ADMMs for Nonconvex Nonsmooth Optimization. arXiv: 1802.03284v3.

[16] Bian, F., Liang, J. and Zhang, X. (2021) A Stochastic Alternating Direction Method of Multipliers for Non-Smooth and Non-Convex Optimization. *Inverse Problems*, **37**, Article ID: 075009. https://doi.org/10.1088/1361-6420/ac0966

[17] Bauschke, H.H., Bolte, J. and Teboulle, M. (2017) A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. *Mathematics of Operations Research*, **42**, 330-348. https://doi.org/10.1287/moor.2016.0817

[18] Bregman, L.M. (1967) The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics*, **7**, 200-217. https://doi.org/10.1016/0041-5553(67)90040-7

[19] Attouch, H., Bolte, J., Redont, P. and Soubeyran, A. (2010) Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz Inequality. *Mathematics of Operations Research*, **35**, 438-457. https://doi.org/10.1287/moor.1100.0449

[20] 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

[21] Davis, D. (2016) The Asynchronous PALM Algorithm for Nonsmooth Nonconvex Problems. arXiv: 1604.00526.

[22] 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