Advances in Applied Mathematics 应用数学进展, **2023, 12(9), 4090-4104**
Published Online September 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.129401

Hans汉斯

一种自适应选取步长的随机交替方向乘子法

李 静，薛 丹*

青岛大学数学与统计学院，山东 青岛

收稿日期：2023年8月19日；录用日期：2023年9月11日；发布日期：2023年9月19日

摘 要

本文研究了具有可分离变量的凸随机优化问题，提出了一种新的随机交替方向乘子(ADMM)算法。该算法是ADMM 与自适应选取步长的随机缩减梯度算法(SVRG-BB) 的结合，利用BB 步长实现了SVRG-ADMM方法自适应选取步长，而无需再使用递减步长或者手动调节步长。在一般的假设条件下，证明了算法的收敛性。最后给出相关数值实验表明了算法的有效性。

关键词

随机优化，交替方向乘子法，机器学习

A Stochastic Alternating Direction Multiplier Method with Adaptive Step Selection

Jing Li, Dan Xue*

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Aug. 19^{th}, 2023; accepted: Sep. 11^{th}, 2023; published: Sep. 19^{th}, 2023

* 通讯作者。

文章引用: 李静, 薛丹. 一种自适应选取步长的随机交替方向乘子法[J]. 应用数学进展, 2023, 12(9): 4090-4104.
DOI: 10.12677/aam.2023.129401

Abstract

This paper considers the problem of convex stochastic optimization with separable variables. We propose a stochastic alternating direction method of multipliers (ADMM) algorithm to solve this convex stochastic optimization problem. The algorithm can be roughly regarded as a combination of ADMM and adaptive step stochastic reduced gradient algorithm (SVRG-BB). BB step size is used to realize the adaptive step size selection by SVRG-ADMM method, without decreasing step size or manually adjusting step size. Under general assumptions, the convergence of the algorithm is proved. Finally, numerical experiments are given to show the effectiveness of the algorithm.

Keywords

Stochastic Optimization, Alternating Direction Method of Multipliers, Machine Learning

1. 介绍

随机优化是一类强大的优化工具，用于解决机器学习、模式识别和计算机视觉中的大规模问题 [1]。例如，随机梯度下降是解决以下优化问题的有效方法，这是机器学习的基础。本文考虑以下优化问题:

$$\begin{aligned} &\min_{x,y} f(x) + g(y) \\ &s.t.\ Ax + By = c, \end{aligned} \tag{1}$$

其中$x \in R^d, y \in R^p$, $f(x)$ 是光滑凸函数，$g(y)$ 是凸函数，但可能非光滑，并且有$A \in R^{d\times q}, B \in R^{q\times p}$，$c \in R^q$ 表示给定的矩阵和向量。

对于许多机器学习问题，$f(x)$通常是关于数据的损失函数，而$g(x)$ 是控制模型复杂性或提供关于解决方案的先验信息的正则化函数 [2]。对于这种问题，我们作如下假设：

$$f(x) = \mathbb{E}F(x;\xi), \tag{2}$$

其中$F(x;\xi)$是一个以随机数ξ作为索引的函数。对于传统的机器学习，数据通常是有限的。如果每个子函数都表示为$f_i(x)$，则$f(x)$具有以下形式 [3]:

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x), \tag{3}$$

其中n为样本总数，每个$f_i(x)$为对应于第i个数据样本的损失函数。当n有限时，最具代表性的例子是经验风险最小化问题。在实际应用中，n也可以是无限的，这是更普遍的情况。当n非常大时，获得精确的$f(x)$或其梯度$\nabla f(x)$在计算上可能是昂贵的，甚至在n无限时是几乎无法计算。为了处理这类大规模问题，常用的方法是使用一个或几个随机选择的样本来估计目标函数或全梯度的值，这种算法称为随机算法。随机梯度下降算法(stochastic gradient descent, SGD) [4]在每次迭代中只计算随机选取的一个或小批量样本的梯度，而不是计算所有样本的梯度，因此每次迭代的复杂度降低很多。尽管随机梯度算法具有可扩展性，但由于随机过程中存在方差，随机梯度比批处理梯度噪声更大 [5]。因此，随着随机梯度下降方法中的迭代，步长需要逐渐减小或者手动调节步长，导致收敛速度变慢。

为了降低方差，近期开发了几种减少方差的变体。Shalev - schwartz 和Zhang提出了随机双坐标上升(Stochastic Dual Coordinate Ascent, SDCA) [6]，并表明这类方法具有与SGD 相当或更好的理论保证。随机平均梯度法(SAG) [7] 记录之前计算的所有随机梯度，然后与新计算的随机梯度进行平均，作为下一个梯度估计，从而减小方差。Defazio 等人 [8]对SAG 进行了改进，提出了SAGA算法，SAGA算法的一个显著优点是可以直接求解非强凸问题，两种方法在适当条件下都可以实现线性收敛，但需要存储梯度信息，开销较大。Johnson和Zhang [9]提出了随机方差减少梯度(SVRG)算法，该算法通过周期性缓存全梯度而不存储梯度信息来减小方差，该算法有两个循环，外循环计算全梯度，内循环计算方差较小的随机梯度。但是SVRG在实际运用时，为了保证收敛率，需要逐渐减小的步长或者手动调节步长，这给研究带来了很多不便。为了步长方面的缺点，一些研究者将Barzilai-Borwein (BB) 方法 [10]与随机梯度方法相结合，实现了自动选择步长。[11]的研究包括SVRG-BB 和SGD-BB。 [12]提出了SVRG-TR-BB 和mSVRG-TR-BB。数值实验表明，结合BB步长的随机梯度方法具有良好的性能。

ADMM是一种有效的优化工具，由于原问题具有变量可分离的优势，该模型在机器学习、模式识别、计算机视觉等诸多领域得到了广泛的应用 [13]。许多研究者致力于改进ADMM 方法，并开发了一些适合求解大规模优化问题的随机ADMM算法。Wang and Banerjee、Suzuki和Ouyang等人 [14] [15] 开发了一些在线或随机ADMM 算法。然而，这些算法存在收敛速度慢的缺点。近年来，为了提高算法的收敛速度，提出了一些加速的随机admm。S. shalev - schwartz等人 [16] 提出了SDCD-ADMM 算法，在适当条件下可以实现线性收敛。Zhong和Kwok提出的SAG-ADMM [17]算法通过计算平均梯度保证了收敛速度，降低了计算复杂度。Zheng和Kwok开发了SVRG-ADMM [18]，该算法在不需要为之前的梯度和对偶变量增加额外空间的情况下实现了较快的收敛速度。ASVRG-ADMM [19] [20] 与动量加速技术结合提出。这些也促使我们将随机算法集成到乘法器的交替方向方法(ADMM)中，以提高随机ADMM 算法的性能。在这些研究的启发下，我们提出了一种自适应选取步长的随机交替方向乘子法，结合SVRG-BB算法，用于求解大规模凸随机优化问题，命名为SVRG-BB-ADMM，在一般的假设条件下，我们证明了该算法具有线性收敛速度。数值结果表明，该算法是有效的。

2. 算法

在本节中，首先介绍了SVRG-BB算法，随后，我们提出了SVRG-BB-ADMM 算法，并解释了它的一些特性。SVRG-BB算法如下，

算法2.1 SVRG-BB

0. 给定$\tilde{x}_0$，初始步长η_0，更新频率m.
1. for $s = 0, 1, ...do$
2. 赋值$z_s = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{x}_s)$
3. if $s > 0$ then
4. $\eta_s = \frac{1}{m}\|\tilde{x}_s - \tilde{x}_{s-1}\|_2^2/(\tilde{x}_s - \tilde{x}_{s-1})^T(z_s - z_{s-1})$
5. end if
6. $x_0 = \tilde{x}_s$
7. for $t = 0, ..., m-1$ do
8. Randomly pick $i_t \in \{1, ..., n\}$
9. $x_{t+1} = x_t - \eta_s(\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x}_s) + z_s)$
10. end for
11. $\tilde{x}_{s+1} = x_m$
12. end for

在SVRG-BB算法中，使用BB方法计算步长η_s，实现了SVRG算法自适应选择步长。结合SVRG-BB算法，我们提出了如下的SVRG-BB-ADMM算法，

算法2.1 SVRG-BB-ADMM

0. 给定$\tilde{x}_0, \tilde{y}_0$, and $\tilde{\mu}_0 = \frac{1}{\rho}(A^T)^{\dagger}\nabla f(\tilde{x}_0)$，初始步长$\eta_0$，更新频率$m$，罚参数$\rho > 0$.
1. for $s = 0, 1, ...do$
2. 赋值$\tilde{x} = \tilde{x}_{s-1}; x_0 = \tilde{x}_{s-1}$;
3. $y_0 = \tilde{y}_{s-1}; \mu_0 = \tilde{\mu}_{s-1}$;
4. $z_s = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{x}_k)$
5. if $s > 0$
6. $\eta_s = \frac{1}{m}\|\tilde{x}_s - \tilde{x}_{s-1}\|_2^2/(\tilde{x}_s - \tilde{x}_{s-1})^T(z_s - z_{s-1})$
7. end if
8. for $t = 1, 2, ..., m$ do
9. $y_t = \arg\min_y \left(g(y) + \frac{\rho}{2}\|Ax_{t-1} + By - c + \mu_{t-1}\|^2\right)$
10. $x_t = \arg\min_x \left(\tilde{\nabla} f(x_{t-1})^T x + \frac{\rho}{2}\|Ax + By_t - c + \mu_{t-1}\|^2 + \frac{\|x - x_{t-1}\|_G^2}{2\eta_s}\right)$
11. $\mu_t = \mu_{t-1} + Ax_t + By_t - c$
12. end for
13. $\tilde{x}_{s+1} = x_m$; $\tilde{y}_{s+1} = y_m$; $\tilde{\mu}_s = \frac{1}{\rho}(A^T)^{\dagger}\nabla f(\tilde{x}_s)$;
14. end for

经典ADMM方法一般采用以下增广拉格朗日函数求解:

$$L_t(x,y,\mu)=f(x_t)+g(y)+\langle Ax+By-c,\mu\rangle+\frac{\rho}{2}\|Ax+By-c+\mu_{t-1}\|^2$$

其中, $\rho>0$ 是罚参数，μ 是对偶变量。在t次迭代中，ADMM 的更新方式如下:

$$x_t=\arg\min_x\left(f(x)+g(y_{t-1})+\langle\rho_{t-1},Ax+By_{t-1}-c\rangle+\frac{\rho}{2}\|Ax+By_{t-1}-c\|^2\right),$$

$$y_t=\arg\min_y\left(f(x_t)+g(y)+\langle\rho_{t-1},Ax_t+By-c\rangle+\frac{\rho}{2}\|Ax_t+By-c\|^2\right),$$

$$\mu_t=\mu_{t-1}+Ax_t+By_t-c.$$

考虑随机ADMM方法，定义近似增广拉格朗日量:

$$\hat{L}_t(x,y,\mu)=f(x_t)+\langle\nabla f(x_t),x-x_t\rangle+g(y)+\frac{\rho}{2}\|Ax+By-c+\frac{1}{\rho}\mu\|^2+\frac{1}{2\eta_{t+1}}\|x-x_t\|^2$$

在算法中，我们使用下面的公式来更新x_t,y_t和μ_t:

$$y_t=\arg\min_y\left(g(y)+\frac{\rho}{2}\|Ax_{t-1}+By-c+\mu_{t-1}\|^2\right),$$

$$x_t=\arg\min_x\left({v_{t-1}}^Tx+\frac{\rho}{2}\|Ax+By_t-c+\mu_{t-1}\|^2+\frac{\|x-x_{t-1}\|_G^2}{2\eta_s}\right),$$

$$\mu_t=\mu_{t-1}+Ax_t+By_t-c.$$

其中，$v_{t-1}=\tilde{\nabla}f(x_{t-1})=\nabla f_{i_t}(x_{t-1})-\nabla f_{i_t}(\tilde{x}+z_s$

当A^TA很大时，该矩阵的存储可能仍然存在问题。为改善该问题，常用的方法是线性化[?],令$G=\gamma I-\eta\rho A^TA$，为确保$G\succeq I$，取$\gamma\geqslant\gamma_{min}\equiv\eta\rho\|A^TA\|+1$。

对于变量x, $\|x\|$其是ℓ_2-范数, 且$\|x\|_G=\sqrt{x^TGx}$.

3. 收敛性分析

本部分首先给出算法的收敛性分析。

假设3.1 设每个f_i是连续的、可微凸函数，且梯度为L-李普希茨连续的，即，存在$L_i>0$ 使得对任意的x_i,x_j，有下式成立

$$f_i(x_j)\leqslant f_i(x_i)+\nabla f_i(x_i)^T(x_j-x_i)+\frac{L_i}{2}\|x_i-x_j\|^2$$

假设3.2 设(x_*, y_*)是原问题的最优解，μ_*是对应的对偶解。在最有情况下，有下式成立

$$\nabla f(x_*) + \rho A^T \mu_* = 0,$$

$$g'(y_*) + \rho B^T \mu_* = 0$$

$$Ax_* + By_* = c$$

假设3.3 函数f强凸，即对所有的x_i, x_j，有

$$f(x_i) \geqslant f(x_j) + \nabla f(x_j)^T (x_i - x_j) + \frac{\lambda_f}{2}\|x_i - x_j\|^2$$

假设3.4 设矩阵A行满秩。

引理3.1 步长η_s的上界为$\frac{1}{m\lambda_f}$.

证明 利用函数f的强凸性，易得到BB步长上界如下式所示，

$$\begin{aligned}\eta_s &= \frac{1}{m} \cdot \frac{\|\tilde{x}_s - \tilde{x}_{s-1}\|_2^2}{(\tilde{x}_s - \tilde{x}_{s-1})^T (z_s - z_{s-1})} \\ &\leqslant \frac{1}{m} \cdot \frac{\|\tilde{x}_s - \tilde{x}_{s-1}\|_2^2}{\lambda_f \|\tilde{x}_s - \tilde{x}_{s-1}\|_2^2} \\ &= \frac{1}{m\lambda_f}\end{aligned}$$

引理3.2 $\tilde{\nabla} f(x_{t-1})$的方差上界为：

$$\begin{aligned}&E\|\tilde{\nabla} f(x_{t-1}) - \nabla f(x_{t-1})\|^2 \\ &\quad \leqslant 4L_{max}(f(x_{t-1}) + f(\tilde{x}) - 2f(x_*) - \nabla f(x_*)^T (x_{t-1} + \tilde{x} - 2x_*))\end{aligned}$$

其中，$L_{max} = max_i L_i$.

证明

$$\begin{aligned}&E\|\tilde{\nabla} f(x_{t-1}) - \nabla f(x_{t-1})\|^2 \\ &\quad = E\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x}) - (\nabla f(x_{t-1}) - \nabla f(\tilde{x}))\|^2 \\ &\quad = E\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})\|^2 - \|\nabla f(x_{t-1}) - \nabla f(\tilde{x})\|^2 \\ &\quad \leqslant E\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})\|^2 \\ &\quad \leqslant 2E\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(x_*)\|^2 + 2E\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x_*)\|^2 \\ &\quad = 2\sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(x_{t-1}) - \nabla f_i(x_*)\|^2 + E\|\nabla f_i(\tilde{x}) - f_i(x_*)\|^2 \\ &\quad \leqslant 4L_{max}(f(x_{t-1}) + f(\tilde{x}) - 2f(x_*) - \nabla f(x_*)^T (x_{t-1} + \tilde{x} - 2x_*))\end{aligned}$$

其中，第三个等式中，运用了公式$E\|x_i - Ex_i\|^2 = E\|x_i\|^2 - \|Ex_i\|^2$，第二个不等式运用了$\|a+b\|^2 \leqslant 2\|a\|^2 + \|b\|^2$，最后一个不等式采用了 [22]中$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(x_*)\|_2^2 \leqslant 2L_{max}(f(x) - f(x_*) - \nabla f(x_*)^T(x - x_*))$.

引理3.3 当$0 \leqslant \eta \leqslant \min\{\frac{1}{\lambda_f}, \frac{1}{\lambda_f}\}$时，有下式成立，

$$\begin{aligned} & f(x)+q_t(x-x_t) \\ & \geqslant f(x_t)+\varphi_t^T(x-x_{t-1})+\frac{\eta_s}{2}\|\varphi_t\|_{G^{-1}}^2+(\tilde{\nabla}f(x_{t-1}))-\nabla f(x_{t-1}))^T(x_t-x) \end{aligned}$$

证明 首先，考虑x_t的更新:

$$\tilde{\nabla}f(x_{t-1})^T x+\frac{\rho}{2}\|Ax+By_t-c+\mu_{t-1}\|^2+\frac{\|x-x_{t-1}\|_G^2}{2\eta_s}$$

其中$\tilde{\nabla}f(x_{t-1})=\nabla f_{i_t}(x_{t-1})-\nabla f_{i_t}(\tilde{x}+z_s$，求$x$在$x_t$处的导数得到：

$$\varphi_t+\frac{1}{\eta_s}G(x_t-x_{t-1})=0$$

其中，

$$\varphi_t=\tilde{\nabla}f(x_{t-1})+q_t$$

$$q_t=\rho A^T(Ax_t+By_t-c+\mu_{t-1})$$

因此，迭代可以重写为如下形式：

$$x_t=x_{t_1}-\eta_s G^{-1}\varphi_t$$

考虑下式，由f的凸性及其梯度李普希兹连续性有

$$\begin{aligned} & f(x)+q_t(x-x_t) \\ & \geqslant f(x_{t-1})+\nabla f(x_{t-1})^T(x-x_{t-1})+q_t(x-x_t) \\ & \geqslant f(x_t)-\nabla f(x_{t-1})^T(x_t-x_{t-1})-\frac{L_f}{2}\|x-x_{t-1}\|^2+\nabla f(x_{t-1})^T(x_t-x_{t-1})+q_t(x-x_t) \\ & \geqslant f(x_t)-\nabla f(x_{t-1})^T(x_t-x_{t-1})-\frac{L_f}{2}\|x-x_{t-1}\|_G^2+\nabla f(x_{t-1})^T(x_t-x_{t-1})+q_t(x-x_t) \\ & = f(x_t)-\nabla f(x_{t-1})^T(x_t-x_{t-1})-\frac{\eta_s^2 L_f}{2}\|\varphi_t\|_{G^{-1}}^2+\nabla f(x_{t-1})^T(x_t-x_{t-1})+q_t(x-x_t) \\ & = f(x_t)+\nabla f(x_{t-1})^T(x-x_t)+(\varphi_t-\tilde{\nabla}f(x_{t-1}))^T(x-x_t)-\frac{\eta_s^2 L_f}{2}\|\varphi_t\|_{G^{-1}}^2 \\ & = f(x_t)+\varphi_t^T(x-x_{t-1}+x_{t-1}-x_t)+(\tilde{\nabla}f(x_{t-1}))-\nabla f(x_{t-1}))^T(x_t-x)-\frac{\eta_s^2 L_f}{2}\|\varphi_t\|_{G^{-1}}^2 \\ & = f(x_t)+\varphi_t^T(x-x_{t-1})+\eta_s\|\varphi_t\|_{G^{-1}}^2+(\tilde{\nabla}f(x_{t-1}))-\nabla f(x_{t-1}))^T(x_t-x)-\frac{\eta_s^2 L_f}{2}\|\varphi_t\|_{G^{-1}}^2 \\ & \geqslant f(x_t)+\varphi_t^T(x-x_{t-1})+\frac{\eta_s}{2}(2-\eta_s L_f)\|\varphi_t\|_{G^{-1}}^2+(\tilde{\nabla}f(x_{t-1}))-\nabla f(x_{t-1}))^T(x_t-x) \\ & \geqslant f(x_t)+\varphi_t^T(x-x_{t-1})+\frac{\eta_s}{2}\|\varphi_t\|_{G^{-1}}^2+(\tilde{\nabla}f(x_{t-1}))-\nabla f(x_{t-1}))^T(x_t-x) \end{aligned}$$

引理3.4 由于g是凸函数，有下式成立，

$$E[g(y_t)-g(y_*)-g'(y_*)^T(y_t-y_*)+(B^T\rho(\mu_t-\mu_*))^T(y_t-y_*)]$$
$$\leqslant \frac{\rho}{2}E[\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2+\|\mu_t-\mu_{t-1}\|^2]$$

证明 由于g是凸函数，我们有

$$\begin{aligned}
g(y_t)-g(y_*)&\leqslant g'(y_t)^T(y_t-y_*)\\
&=-(\rho B^T(Ax_{t-1}+By_t-c+\mu_{t-1}))^T(y_t-y_*)\\
&=-(\rho B^T(Ax_{t-1}+By_t-c+\mu_t-Ax_t-By_t+c))^T(y_t-y_*)\\
&=-(\rho B^T\mu_t)^T(y_t-y_*)+\rho A^TB(x_t-x_{t-1})^T(y_t-y_*)\\
&=-(\rho B^T\mu_t)^T(y_t-y_*)+\frac{\rho}{2}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2)\\
&\quad+\frac{\rho}{2}(\|Ax_t+By_t-c\|^2-\|Ax_{t-1}+By_t-c\|^2)\\
&\leqslant-(\rho B^T\mu_t)^T(y_t-y_*)+\frac{\rho}{2}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2)\\
&\quad+\frac{\rho}{2}\|\mu_t-\mu_{t-1}\|^2
\end{aligned}$$

因此，可以得到

$$g(y_t)-g(y_*)+(\rho B^T\mu_t)^T(y_t-y_*)$$
$$\leqslant\frac{\rho}{2}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2)+\frac{\rho}{2}\|\mu_t-\mu_{t-1}\|^2$$

考虑不等式左边有

$$\begin{aligned}
&g(y_t)-g(y_*)+(\rho B^T\mu_t)^T(y_t-y_*)\\
&=g(y_t)-g(y_*)+(\rho B^T(\mu_t-\mu_*))^T(y_t-y_*)+(\rho B^T\mu_*)^T(y_t-y_*)\\
&=g(y_t)-g(y_*)+(\rho B^T(\mu_t-\mu_*))^T(y_t-y_*)-g'(y_*)^T(y_t-y_*)
\end{aligned}$$

取期望即可得到结果。

引理3.5

$$E[-(Ax_t+By_t-c)^T\rho(\mu_t-\mu_*)]=\frac{\rho}{2}E[\|\mu_{t-1}-\mu_*\|^2-\|\mu_t-\mu_*\|^2-\|\mu_t-\mu_{t-1}\|^2]$$

证明 由迭代$\mu_t=\mu_{t-1}+Ax_t+By_t-c$，我们有

$$\begin{aligned}
-(Ax_t+By_t-c)^T\rho(\mu_t-\mu_*)&=\rho(\mu_{t-1}-\mu_t)^T(\mu_t-\mu_*)\\
&=\frac{\rho}{2}(\|\mu_{t-1}-\mu_*\|^2-\|\mu_t-\mu_*\|^2-\|\mu_t-\mu_{t-1}\|^2)
\end{aligned}$$

取期望即可得到结果。

定理3.6 令

$$\varepsilon=\frac{\|G+\eta\rho A^TA\|}{\lambda_f\eta m(1-4\eta L_{max})}+\frac{4\eta(m+1)L_{max}}{m(1-4\eta L_{max})}+\frac{\lambda_f\|A^\dagger(A^\dagger)^T\|}{\rho m(1-4\eta L_{max})}.$$

选取$0<\eta<\min\{\frac{1}{L_f},\frac{1}{4L_{max}}\}$, 并且迭代的次数足够大，使得$\varepsilon<1$. 那么有

$$\mathbb{E}R(\tilde{x}_s,\tilde{y}_s)\leqslant\varepsilon^sR(\tilde{x}_0,\tilde{y}_0).$$

证明 利用引理3.3，我们有

$$\begin{aligned}\|x_t-x_*\|_G^2&=\|x_{t-1}-\eta_sG^{-1}\varphi_t-x_*\|_G^2\\&=\|x_{t-1}-x_*\|_G^2+\eta_s^2\|\varphi_t\|_{G^{-1}}^2-2\eta_s(x_{t-1}-x_*)^T\varphi_t\\&\leqslant\|x_{t-1}-x_*\|_G^2-2\eta_s(f(x_t)-f(x_*))-2\eta_s(v_{t-1}-\nabla f(x_{t-1}))^T(x_t-x_*)\\&\quad+2\eta_sq_t^T(x_*-x_t)\end{aligned}$$

考虑$-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(x_t-x_*)$. 定义如下凸函数

$$\chi_t(x)=\frac{\rho}{2}\|Ax+By_t-c+\mu_{t-1}\|^2+\frac{1}{2\eta_s}\|x-x_{t-1}\|_{G-I}^2$$

及

$$\overline{x}=prox_{\eta_s\chi_t}(x_{t-1}-\eta_s\nabla f(x_{t-1})),$$

其中，$prox_{\eta_s\chi_t}(y)=\min_x\eta_s\chi_t(x)+\frac{1}{2}\|x-y\|^2$ 是近端算子。注意到:

$$x_t=prox_{\eta_s\chi_t}(x_{t-1}-\eta_s\tilde{\nabla}f(x_{t-1})),$$

由于

$$\begin{aligned}x_t&=\arg\min_x\left(\tilde{\nabla}f(x_{t-1})^Tx+\frac{\rho}{2}\|Ax+By_t-c+\mu_{t-1}\|^2+\frac{\|x-x_{t-1}\|_G^2}{2\eta_s}\right)\\&=\arg\min_x\left(\eta_s\tilde{\nabla}f(x_{t-1})^Tx+\frac{\eta_s\rho}{2}\|Ax+By_t-c+\mu_{t-1}\|^2+\frac{\|x-x_{t-1}\|_{G-I}^2}{2}+\frac{\|x-x_{t-1}\|^2}{2}\right)\\&=\arg\min_x\left(\eta_s\chi_t(x)+\frac{1}{2}\|x-(x_{t-1}-\eta_s\tilde{\nabla}f(x_{t-1}))\|^2\right)\end{aligned}$$

可以得到

$$\begin{aligned}&-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(x_t-x_*)\\&\quad=-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(x_t-\overline{x})-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(\overline{x}-x_*)\\&\quad\leqslant2\eta_s\|\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1})\|\|x_t-\overline{x}\|-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(\overline{x}-x_*)\\&\quad\leqslant2\eta_s\|\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1})\|\|(x_{t-1}-\eta_s\tilde{\nabla}f(x_{t-1}))-(x_{t-1}-\eta_s\nabla f(x_{t-1}))\|\\&\qquad-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(\overline{x}-x_*)\\&\quad=2\eta_s^2\|\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1})\|^2-2\eta_s(\tilde{\nabla}f(x_{t-1})-\nabla f(x_{t-1}))^T(\overline{x}-x_*)\end{aligned}$$

在第一个不等式中使用柯西-施瓦茨不等式, 在第二个不等式中，我们利用了近端算子的性质。综合以上结果，可得

$$\begin{aligned}&\|x_t - x_*\|_G^2 - 2\eta_s q_t^T(x_* - x_t)\\&\leqslant \|x_{t-1} - x_*\|_G^2 - 2\eta_s(f(x_t) - f(x_*)) + 2\eta_s^2\|\tilde{\nabla}f(x_{t-1}) - \nabla f(x_{t-1})\|^2\\&\quad - 2\eta_s(\tilde{\nabla}f(x_{t-1}) - \nabla f(x_{t-1}))^T(\overline{x} - x_*)\end{aligned}$$

由无偏估计，有$E[\tilde{\nabla}f(x_{t-1})] = \nabla f(x_t)$。取期望，我们有

$$\begin{aligned}&E[\|x_t - x_*\|_G^2 - 2\eta_s q_t^T(x_* - x_t)]\\&\leqslant \|x_{t-1} - x_*\|_G^2 - 2\eta_s(Ef(x_t) - f(x_*)) + 2\eta_s^2 E\|\tilde{\nabla}f(x_{t-1}) - \nabla f(x_{t-1})\|^2\\&\leqslant \|x_{t-1} - x_*\|_G^2 - 2\eta_s(Ef(x_t) - f(x_*))\\&\quad + 8\eta_s^2 L_{max}(f(x_{t-1}) + f(\tilde{x}) - 2f(x_*) - \nabla f(x_*)^T(x_{t-1} + \tilde{x} - 2x_*))\end{aligned}$$

移项并取期望，我们得到

$$\begin{aligned}&2\eta_s E[f(x_t) - f(x_*) - q_t^T(x_* - x_t)]\\&\leqslant E\|x_{t-1} - x_*\|_G^2 - E\|x_t - x_*\|_G^2 + 8\eta^2 L_{max}(f(x_{t-1}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} - x_*))\\&\quad + 8\eta^2 L_{max}(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*))\end{aligned}$$

利用最优性条件$\nabla f(x_*) + \rho A^T\mu_* = 0$, $\alpha_t = \rho A^T(Ax_t + By_t - c + \mu_{t-1}) = \rho A^T\mu_t$，令$\sigma_t = \rho(\mu_t - \mu_*)$，可得

$$\begin{aligned}&2\eta_s E[f(x_t) - f(x_*) - \alpha_t^T(x_* - x_t)]\\&= 2\eta_s E[f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (\rho A^T\mu_*)^T(x_t - x_*) - (\rho A^T\mu_t)^T(x_* - x_t)]\\&= 2\eta_s E[f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (A^T\sigma_t)^T(x_* - x_t)]\end{aligned}$$

因此，我们有

$$\begin{aligned}&2\eta_s E[f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (A^T\sigma_t)^T(x_* - x_t)]\\&\leqslant E\|x_{t-1} - x_*\|_G^2 - E\|x_t - x_*\|_G^2 + 8\eta_s^2 L_{max}(f(x_{t-1}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} - x_*))\\&\quad + 8\eta^2 L_{max}(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*))\end{aligned}$$

对$t = 1, ..., m$,求和，利用$2\eta_s(1 - 4\eta L_{max}) \leqslant 2\eta_s$，和$x_0 = \tilde{x}_{s-1}$可得

$$\begin{aligned}&2\eta_s(1 - 4\eta_s L_{max})\sum_{i=0}^{m} E[f(x_i) - f(x_*) - \nabla f(x_*)^T(x_i - x_*)] - 2\eta_s E\sum_{i=0}^{m}(A^T\sigma_i)^T(x_* - x_i)\\&\leqslant \|\tilde{x}_{s-1} - x_*\|_G^2 - E\|x_m - x_*\|_G^2 + 8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_{s-1} - x_*))\\&\leqslant \|\tilde{x}_{s-1} - x_*\|_G^2 + 8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_{s-1} - x_*))\end{aligned}$$

同时，我们有

$$\begin{aligned}&-(A^T\sigma_t)^T(x_*-x_t)+(B^T\sigma_t)^T(y_*-y_t)-(Ax_t+By_t-c)^T\sigma_t\\&=-(Ax_*+By_*-c)^T\sigma_t+(Ax_t-Ax_t+By_t-By_t)^T\sigma_t\\&=0\end{aligned}$$

定义$R(x,y)=f(x)-f(x_*)-\nabla f(x_*)^T(x-x_*)+g(y)-g(y_*)-g'(y_*)^T(y-y_*)$. 接下来，考虑以下过程

$$\begin{aligned}&2\eta_s(1-4\eta L_{max})mER(\tilde{x}_s,\tilde{y}_s)\\&\leqslant\|\tilde{x}_{s-1}-x_*\|_G^2+8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\eta_s\rho\|A\tilde{x}_{s-1}+By_*-c\|^2+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&=\|\tilde{x}_{s-1}-x_*\|_G^2+8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\eta_s\rho\|A\tilde{x}_{s-1}-Ax_*\|^2+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&=\|\tilde{x}_{s-1}-x_*\|_{G+\eta_s\rho A^TA}^2+8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&\leqslant\|G+\eta_s\rho A^TA\|\|\tilde{x}_{s-1}-x_*\|^2+8\eta_s^2(m+1)L_{max}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&\leqslant_s\left(\frac{2\|G+\eta_s\rho A^TA\|}{\lambda_f}+8\eta_s^2(m+1)L_{max}\right)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&\leqslant\left(\frac{2\|G+\eta_s\rho A^TA\|}{\lambda_f}+8\eta_s^2(m+1)L_{max}\right)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\&\quad+\left(\frac{2\|G+\eta_s\rho A^TA\|}{\lambda_f}+8\eta_s^2(m+1)L_{max}\right)(g(\tilde{y}_{s-1})-g(y_*)-g'(y_*)^T(\tilde{y}_{s-1}-y_*))\\&\quad+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&=\left(\frac{2\|G+\eta_s\rho A^TA\|}{\lambda_f}+8\eta_s^2(m+1)L_{max}\right)R(\tilde{x}_{s-1},\tilde{y}_{s-1})+\eta_s\rho\|\tilde{\mu}_{s-1}-\mu_*\|^2\end{aligned}$$

因为我们假设A的行秩是满的，通过更新规则$\tilde{\mu}_{s-1}=-\frac{1}{\rho}(A^T)^\dagger\nabla f(\tilde{x}_{s-1})$(证明见文献 [17])，可得

$$\begin{aligned}&\|\tilde{\mu}_{s-1}-\mu_*\|^2\\&=\frac{1}{\rho^2}\|\nabla f(\tilde{x}_{s-1})-\nabla f(x_*)\|_{A^\dagger(A^\dagger)^T}^2\\&\leqslant\frac{2L_f\|A^\dagger(A^\dagger)^T\|}{\rho^2}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\end{aligned}$$

因此，结合以上结果，我们有

$$\begin{aligned} & 2\eta_s m(1-4\eta_s L_{max})ER(\tilde{x}_s,\tilde{y}_s) \\ \leqslant & \left(\frac{2\|G+\eta_s\rho A^TA\|}{\lambda_f}+8\eta_s^2(m+1)L_{max}+\frac{2L_f\|A^\dagger(A^\dagger)^T\|}{\rho^2}\right)R(\tilde{x}_{s-1},\tilde{y}_{s-1}) \end{aligned}$$

令$\varepsilon=\frac{\|G+\eta_s\rho A^TA\|}{\lambda_f\eta_s m(1-4\eta_s L_{max})}+\frac{4\eta_s(m+1)L_{max}}{m(1-4\eta_s L_{max})}+\frac{\lambda_f\|A^\dagger(A^\dagger)^T\|}{\rho m(1-4\eta_s L_{max})}$, 我们有

$$ER(\tilde{x}_s,\tilde{y}_s)\leqslant\varepsilon^s R(\tilde{x}_0,\tilde{y}_0)$$

算法的收敛性得证。

4. 数值实验

本节我们在一些标准数据集上进行数值实验，使用几种随机优化算法求解逻辑回归问题，并在不同参数下比较结果，提供数值实验结果来证明SVRG-BB-ADMM方法的有效性。

我们在广义lasso模型

$$\min_x\sum_{i=1}^n\ell_i(x)+\lambda\|Ax\|_1$$

上进行实验，其中ℓ_i是样本i上的逻辑损失，A是稀疏矩阵。在实验中，我们使用graph-guided融合lasso，其中集合$A=[G;I]$，G为稀疏逆协方差估计得到的图的稀疏模式。对于ADMM公式，我们引入了变量y和约束$Ax=y$。实验在四个标准数据集上进行(见表 1)。这些数值实验是在LIBSVM 网站上获得的标准真实数据集上进行的，表4.1提供了数据集的详细描述。包含训练大小n、特征d 和正则化参数λ。所有实验均在Matlab 2020b 中进行，实验在一台安装了Windows 11操作系统的笔记本电脑上进行，该电脑CPU 为英特尔酷睿i5-12500H，主频为2.5GHz，内存为16GB。

Table 1. Parameter of data sets

表 1. 数据集参数

Dataset	n	d	λ
$ijcnn1$	49,990	22	10^{-3}
$w8a$	49,749	300	10^{-3}
$rcv1.binary$	20,242	47,236	10^{-2}
$news20$	15,935	62,061	10^{-5}

实验结果

通过在不同的标准数据集上的数值实验(见图 1)可以看出，我们的算法与现存的一些随机交替方向乘子法相比具有一定的竞争力，在解决大规模机器学习的实际应用中具有一定优势。

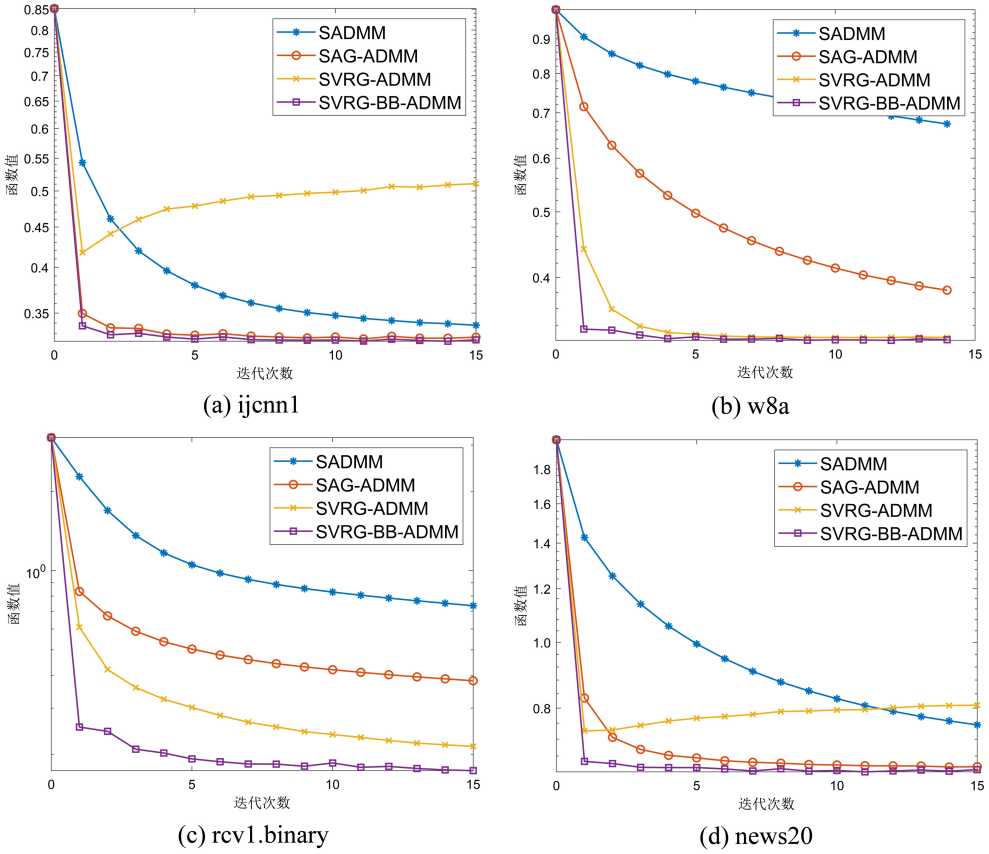

Figure 1. Comparison of numerical results of SADMM, SAG-ADMM, SVRG-ADMM, and SVRG-BB-ADMM methods on different data sets

图 1. SADMM, SAG-ADMM, SVRG-ADMM和SVRG-BB-ADMM在不同数据集上的数值结果比较

5. 总结

在本文中，我们开发了一种新的随机算法，SVRG-BB-ADMM算法利用随机缩减梯度方法减小了随机梯度的方差，并运用BB方法自适应选择步长。通过理论证明，可以看出该算法具有良好的收敛性与稳定性。在标准数据集上的数值实验表明算法在解决大规模问题时有不错的表现。

基金项目

国家自然科学基金[批准号11601252]，山东省自然科学基金[ZR2020MA045]。

参考文献

[1] Che, M.L., Qi, L.Q. and Wei, Y.M. (2015) Positive-Definite Tensors to Nonlinear Complementarity Problems. *Journal of Optimization Theory and Applications*, **168**, 475-487.

https://doi.org/10.1007/s10957-015-0773-1

[2] Schmidt, M.W., Le Roux, N. and Bach, F.R. (2013) Minimizing Finite Sums with the Stochastic Average Gradient. *Mathematical Programming*, **162**, 83-112. https://doi.org/10.1007/s10107-016-1030-6

[3] Mairal, J., Bach, F.R., Ponce, J. and Sapiro, G. (2009) Online Dictionary Learning for Sparse Coding. *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, 14-18 June 2009, 689-696. https://doi.org/10.1145/1553374.1553463

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

[5] Bottou, L., Curtis, F.E. and Nocedal, J. (2016) Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, **60**, 223-311. https://doi.org/10.1137/16M1080173

[6] Shalev-Shwartz, S. and Zhang, T. (2013) Stochastic Dual Coordinate Ascent Methods for Regularized Loss. *Journal of Machine Learning Research*, **14**, 567-599.

[7] Le Roux, N., Schmidt, M.W. and Bach, F.R. (2012) A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets. NIPS.

[8] Defazio, A., Bach, F.R. and Lacoste-Julien, S. (2014) SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. NIPS.

[9] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. NIPS.

[10] Barzilai, J. and Borwein, J.M. (1988) Two-Point Step Size Gradient Methods. *IMA Journal of Numerical Analysis*, **8**, 141-148. https://doi.org/10.1093/imanum/8.1.141

[11] Tan, C.H., Ma, S.Q., Dai, Y.H. and Qian, Y.Q. (2016) Barzilai-Borwein Step Size for Stochastic Gradient Descent. NIPS.

[12] Yu, T.T., Liu, X.W., Dai, Y.H. and Sun, J. (2021) Stochastic Variance Reduced Gradient Methods Using a Trust-Region-Like Scheme. *Journal of Scientific Computing*, **87**, Article No. 5. https://doi.org/10.1007/s10915-020-01402-x

[13] Boyd, S.P., Parikh, N., Chu, E.K.-W., Peleato, B. and Eckstein, J. (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[14] Suzuki, T. (2013) Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method. *Proceedings of the 30th International Conference on Machine Learning*, **28**, 392-400. https://dl.acm.org/doi/10.5555/3042817.3042863

[15] Ouyang, H., He, N., Tran, L. and Gray, A.G. (2013) Stochastic Alternating Direction Method of Multipliers. *Proceedings of the 30th International Conference on Machine Learning*, **28**, 80-88. https://dl.acm.org/doi/10.5555/3042817.3042828

[16] Suzuki, T. (2014) Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers. *Proceedings of the 31st International Conference on Machine Learning*, **32**, 736-744. https://dl.acm.org/doi/10.5555/3044805.3044889

[17] Zhong, W.L. and Kwok, J. (2013) Fast Stochastic Alternating Direction Method of Multipliers. *Proceedings of the 31st International Conference on Machine Learning*, **32**, 46-54.

[18] Zheng, S. and Kwok, J.T.-Y. (2016) Fast-and-Light Stochastic ADMM. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2407-2413. https://dl.acm.org/doi/10.5555/3060832.3060958

[19] Liu, Y.Y., Shang, F.H. and Cheng, J. (2017) Accelerated Variance Reduced Stochastic ADMM. *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, San Francisco, California, 2287-2293. https://dl.acm.org/doi/10.5555/3298483.3298569

[20] Liu, Y.Y., Shang, F.H., Liu, H.Y., Kong, L., Jiao, L.C. and Lin, Z.C. (2020) Accelerated Variance Reduction Stochastic ADMM for Large-Scale Machine Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 4242-4255. https://doi.org/10.1109/TPAMI.2020.3000512

[21] Zhang, X.Q., Burger, M. and Osher, S. (2010) A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration. *Journal of Scientific Computing*, **46**, 20-46. https://doi.org/10.1007/s10915-010-9408-8

[22] Xiao, L. and Zhang, T. (2014) A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM Journal on Optimization*, **24**, 2057-2075. https://doi.org/10.1137/140961791