A Bregman Proximal Stochastic Gradient Method with Extrapolation for Nonconvex Nonsmooth Problems

Qingsong Wang[1], Zehui Liu[2], Chunfeng Cui[2*], Deren Han[2]

[1]School of Mathematics and Computational Science, Xiangtan University, China
[2]LMIB, School of Mathematical Sciences, Beihang University, China
nothing2wang@hotmail.com, {liuzehui, chunfengcui, handr}@buaa.edu.cn

Abstract

In this paper, we explore a specific optimization problem that involves the combination of a differentiable nonconvex function and a nondifferentiable function. The differentiable component lacks a global Lipschitz continuous gradient, posing challenges for optimization. To address this issue and accelerate the convergence, we propose a Bregman proximal stochastic gradient method with extrapolation (BPSGE), which only requires smooth adaptivity of the differentiable part. Under the variance reduction framework, we not only analyze the subsequential and global convergence of the proposed algorithm under certain conditions, but also analyze the sublinear convergence rate of the subsequence, and the complexity of the algorithm, revealing that the BPSGE algorithm requires at most O(epsilon^(-2)) iterations in expectation to attain an epsilon-stationary point. To validate the effectiveness of our proposed algorithm, we conduct numerical experiments on three real-world applications: graph regularized nonnegative matrix factorization (NMF), matrix factorization with weakly-convex regularization, and NMF with nonconvex sparsity constraints. These experiments demonstrate that BPSGE is faster than the baselines without extrapolation. The code is available at: https://github.com/nothing2wang/BPSGE-Algorithm.

Introduction

In this paper, we consider the following nonsmooth nonconvex optimization problem

$$\min_{x\in\overline{C}} \quad \Phi(x) := f(x) + h(x), \tag{1}$$

where $\overline{C}$ denotes the closure of C, which is a nonempty, convex, and open set in $\mathbb{R}^d$, f is a continuously differentiable (may be nonconvex) function which can be written as $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, and h is an extended valued function (maybe nonconvex), which is a regularizer promoting low-complexity structures such as sparsity (Donoho 1995; Fan and Li 2001; Zhang 2010) or non-negativity (Lee and Seung 1999; He et al. 2011) in the solution. Throughout this paper, the usual restrictive requirement of global Lipschitz continuity of the gradient of f is not needed (Lan 2020; Gillis 2020). Many applications can be categorized in the optimization problem (1), such as matrix and tensor factorization (Comon et al. 2008; Kolda and Bader 2009; Che and Wei 2020), supervised neural network model (Hasannasab et al. 2020), and Poisson deconvolution/inverse problems (Bolte et al. 2018).

Deterministic Bregman type methods. The Bregman proximal gradient (BPG) algorithm (Bolte et al. 2018) is a state-of-the-art method for addressing optimization problems characterized by the absence of global Lipschitz continuous gradients. It is simple and far-reaching, which substitutes the customary upper quadratic approximation of a smooth function with a more comprehensive proximity measure (Bauschke, Bolte, and Teboulle 2017). Zhang et al. (Zhang et al. 2019) proposed an extrapolation version of the BPG algorithm, denoted by BPGE, under the assumption that $h(x)$ is convex. Recently, Mukkamala et al. (Mukkamala et al. 2020) introduced an inertial variant of the BPG method, referred to as the CoCaIn BPG algorithm, which tunes the inertial parameter by convex-concave backtracking. Compared with the original BPG method (Bolte et al. 2018), the CoCaIn BPG algorithm exhibits superior numerical performance. Moreover, the BPG framework facilitates the discovery of novel techniques tailored to specific problem domains. For instance, Teboulle and Vaisbourd (Teboulle and Vaisbourd 2020) explored new decomposition settings of the nonnegative matrix factorization (NMF) problem with sparsity constraints. Additionally, an alternative perspective is to consider the BPG-type algorithm by the splitting method. Ahookhos et al. (Ahookhosh, Themelis, and Patrinos 2021) proposed a Bregman forward-backward splitting line search method, which demonstrates locally superlinear convergence to nonisolated local minima by leveraging the Bregman forward-backward envelope (Bauschke, Dao, and Lindstrom 2018; Laude, Ochs, and Cremers 2020). Furthermore, Wang et al. (Wang et al. 2022) investigated a Bregman and inertial extension of the forward-reflected-backward algorithm (Malitsky and Tam 2020) under relative smoothness conditions. Numerous other works have explored the Bregman gradient method framework to tackle the absence of global Lipschitz continuous gradients, including (Reem, Reich, and Pierro 2019; Zhao et al. 2022; Zhu et al. 2021; Dragomir et al. 2022).

Stochastic Bregman type methods. For large-scale datasets, the cost associated with computing the full gradi-

*Corresponding author.

Versions	Algorithm	$h(x)$	Inertial	Sequence convergence	Complexity
Deterministic	BPG (Bolte et al. 2018)	nonconvex	no	subsequential/global	-
	BPGE (Zhang et al. 2019)	convex	yes	subsequential/global	-
	CoCaIn BPG (Mukkamala et al. 2020)	weakly-convex	yes	subsequential/global	-
	BBPG (Teboulle and Vaisbourd 2020)	nonconvex	no	global	-
	i^*FRB (Wang et al. 2022)	nonconvex	no	global	-
	BELLA (Ahookhosh, Themelis, and Patrinos 2021)	nonconvex	no	subsequential/global	-
Stochastic	SMD (Zhang and He 2018)	convex	no	-	$\mathcal{O}(\varepsilon^{-2})$
	SVRAMD (Li et al. 2022)	convex	no	-	$\mathcal{O}(n\varepsilon^{-\frac{2}{3}} + \varepsilon^{-\frac{5}{3}})$
	BFinito (Latafat et al. 2022)	nonconvex	no	subsequential/global	-
	BPSG (Wang and Han 2023)	nonconvex	no	subsequential/global	-
	BPSGE (Algorithm 1)	weakly-convex	yes	subsequential/global	$\mathcal{O}(\varepsilon^{-2})$

Table 1: Summary of the properties of BPSGE (Algorithm 1) and several state-of-the-art methods. "Complexity" means the complexity (in expectation) to obtain an ε-stationary point (Definition 3) of Φ and "-" means not given.

ent can become prohibitively high. To address this concern, the full gradient can be replaced with a stochastic gradient estimator (Robbins and Monro 1951; Bottou 2010) which has yielded remarkable achievements in the field of machine learning (Lin, Li, and Fang 2020; Lan 2020). For the optimization problem without the global Lipschitz continuous gradients, Zhang and He (Zhang and He 2018) conducted a study on the non-asymptotic stationary convergence behaviour of stochastic mirror descent under certain conditions. In a similar vein, Li et al. (Li et al. 2022) proposed a simple and generalized algorithmic framework for incorporating variance reduction into adaptive mirror descent algorithms. However, both approaches (Zhang and He 2018; Li et al. 2022) require $h(\cdot)$ to be a convex function. To address the limitations posed by convexity assumptions, Latafat et al. (Latafat et al. 2022) introduced a Bregman incremental aggregated method that extends the Finito/MISO techniques(Defazio, Domke, and Caetano 2014; Mairal 2015) to non-Lipschitz and nonconvex scenarios. However, it is noteworthy that this approach is memory-intensive and demands periodic computation of the full gradient, which is expensive for large-scale problems. Furthermore, the analysis carried out by Latafat et al. (Latafat et al. 2022) only encompasses two specific variance reduction stochastic estimators, raising doubts about the generalizability of the convergence results to other estimators. To address the aforementioned concerns, Wang and Han (Wang and Han 2023) introduced a stochastic version of the BPG algorithm (Bolte et al. 2018), known as the Bregman proximal stochastic gradient (BPSG) method which does not assume $h(\cdot)$ is convex. Under a general framework of variance reduction, they also established the convergence properties of the generated sequence in terms of subsequential and global convergence.

A summary of the aforementioned algorithms is presented in Table 1.

Notwithstanding the notable advancements, there remain areas that warrant further improvement. Firstly, when $h(\cdot)$ is nonconvex, existing stochastic methods such as BFinito (Latafat et al. 2022) and BPSG (Wang and Han 2023) only analyzed the subsequence and global convergence of the proposed algorithms, yet the sublinear convergence rate of the subsequence and the complexity of the algorithms are unknown. Secondly, the accelerated versions of the Bregman stochastic gradient methods have not been taken into account. Numerically, the incorporation of accelerated techniques, such as the heavy ball (Polyak 1964) and the Nesterov acceleration (Nesterov 1983), with BPSG can further accelerate the numerical performance (Lin, Li, and Fang 2020).

In this paper, we introduce the Bregman proximal stochastic gradient method with extrapolation (BPSGE) to solve the nonconvex nonsmooth optimization problem (1). Our main contributions addressed in this article are as follows:

- With a general variance reduction framework (see Definition 4), we establish the sublinear convergence rate for the subsequence generated by the proposed algorithm.
- Under the assumption of $C = \mathbb{R}^d$, we show that the BPSGE algorithm requires at most $\mathcal{O}(\varepsilon^{-2})$ iterations in expectation to attain an ε-stationary point. Additionally, we establish the global convergence of the sequence generated by BPSGE, leveraging the Kurdyka-Łojasiewicz (KŁ) property.
- Numerical experiments are conducted on three distinct problems: graph regularized NMF, matrix factorization (MF) with weakly-convex regularization, and NMF with nonconvex sparsity constraints. The results of these experiments highlight the favourable performance and enhanced efficiency of the BPSGE algorithm when compared to corresponding algorithms without extrapolation.

The rest of this paper is organized as follows. Section provides some relevant definitions and results. We present a detailed formulation of the BPSGE algorithm and prove its convergence and convergence rate results in Section and Section , respectively. In Section we use three applications to compare BPSGE with several other algorithms. Finally, we draw conclusions in Section .

Preliminary

Definition 1. *((Auslender and Teboulle 2006; Bolte et al. 2018) kernel generating distance). Let C be a nonempty, convex, and open subset of $\mathbb{R}^d$. Associated with C, a function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ is called a kernel generating distance if it satisfies the following conditions:*

- *ψ is proper, lower semicontinuous, and convex, with dom $\psi \subset \overline{C}$ and dom $\partial\psi = C$.*

- *ψ is C^1 on int dom $\psi \equiv C$.*

The class of kernel-generating distances is denoted by $\mathcal{G}(C)$. Given $\psi \in \mathcal{G}(C)$, we define the proximity measure D_ψ : dom $\psi \times$ int dom $\psi \to \mathbb{R}_+$ by

$$D_\psi(x,y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle. \quad (2)$$

The proximity measures D_ψ is called Bregman distance (Bregman 1967). It measures the proximity of x and y.

Indeed, ψ is convex if and only if $D_\psi(x,y) \geq 0$ for any $x \in$ dom ψ and $y \in$ int dom ψ.

Definition 2. *((Mukkamala et al. 2020) $(\bar{L}, \underline{L})$-smooth adaptable) Given $\psi \in \mathcal{G}(C)$, let $f : \mathcal{X} \to (-\infty, +\infty]$ be a proper and lower semi-continuous function with $\operatorname{dom}\psi \subset \operatorname{dom} f$, which is continuously differentiable on C. We say (f,ψ) is $(\bar{L}, \underline{L})$- smooth adaptable on C if there exist $\bar{L} > 0$ and $\underline{L} \geq 0$ such that for any $x, y \in C$,*

$$f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq \bar{L} D_\psi(x,y), \quad (3)$$

and

$$-\underline{L} D_\psi(x,y) \leq f(x) - f(y) - \langle \nabla f(y), x-y\rangle. \quad (4)$$

If $\underline{L} = \bar{L}$, it recovers (Bolte et al. 2018, Definition 2.2). Suppose f is convex. If $\underline{L} = 0$, this definition recovers (Bauschke, Bolte, and Teboulle 2017, Lemma 1) and (Lu, Freund, and Nesterov 2018, Definition 1.1).

Definition 3. *((Lan 2020) ϵ-stationary point) Given $\epsilon > 0$, a point x^* is called an ϵ-stationary point of function $\Phi(x)$ if*

$$\textit{dist}(0, \partial\Phi(x^*)) \leq \epsilon.$$

Algorithm

The classic BPG algorithm (Bolte et al. 2018) is given by

$$x_{k+1} \in \operatorname*{argmin}_{x\in\bar{C}}\; h(x) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\eta_k} D_\psi(x, x_k),$$

where $\eta_k > 0$ is the stepsize. If $\psi = \frac{1}{2}\|\cdot\|^2$, it reduces to the proximal gradient method. When n is large, replacing the full gradient $\nabla f(x_k)$ by the stochastic gradient $\tilde{\nabla} f(x_k)$ can significantly reduce the computational cost, as shown in the BPSG algorithm (Wang and Han 2023).

In this section, we introduce the BPSGE algorithm, i.e., an extrapolation or acceleration version of the BPSG algorithm (Wang and Han 2023), to solve the nonconvex nonsmooth optimization problem (1). Before presenting the algorithm framework of BPSGE, we make the following assumptions.

Assumption 1. *We assume the following three conditions hold:*

- *$\psi \in \mathcal{G}(C)$ is a kernel generating distance given by Definition 1 with $\overline{C} = \overline{\text{dom } h}$.*
- *$h : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lower semicontinuous function with dom $h \cap C \neq \emptyset$.*
- *$f : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lower semicontinuous function with dom $\psi \subset$ dom f, and ψ is continuously differentiable on C.*

Assumption 1 is quite weak. In addition, we make the following assumptions.

Algorithm 1: BPSGE: Bregman proximal stochastic gradient method with extrapolation

Input: Choose $\psi \in \mathcal{G}(C)$ with $C \equiv \operatorname{int}\operatorname{dom}\psi$ such that (f,ψ) is $(\bar{L}, \underline{L})$-smooth adaptable on C.
Initialization: $x_{-1} = x_0 \in \operatorname{int}\operatorname{dom}\psi$, $k_{\max}$, two constants δ, ϵ such that $0 < \epsilon < \delta < 1$.
Update: For $k = 0, 1, \ldots, k_{\max}$,

1: Compute an extrapolation parameter $\beta_k \in [0,1)$ such that

$$D_\psi(x_k, \bar{x}_k) \leq \frac{\delta - \epsilon}{1 + \underline{L}\eta_{k-1}} D_\psi(x_{k-1}, x_k), \quad (5)$$

where $\bar{x}_k = x_k + \beta_k(x_k - x_{k-1}) \in \operatorname{int}\operatorname{dom}\psi$.

2: Compute the stochastic gradient $\tilde{\nabla} f(\bar{x}_k)$ with the mini-batch B_k.

3: Set $0 < \eta_k \leq \min\{\eta_{k-1}, \bar{L}^{-1}\}$ and compute x_{k+1} by

$$x_{k+1} \in \operatorname*{argmin}_{x\in\bar{C}}\; h(x) + \langle \tilde{\nabla} f(\bar{x}_k), x - \bar{x}_k\rangle + \frac{1}{\eta_k} D_\psi(x, \bar{x}_k). \quad (6)$$

4: Stop if the stopping criterion is reached.

Assumption 2.
- *The function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ is σ-strongly convex on C. Let $\sigma = 1$ for simplicity.*
- *There exists $\alpha \in \mathbb{R}_+$ such that $h(\cdot) + \frac{\alpha}{2}\|\cdot\|^2$ is convex.*
- *The pair of functions (f,ψ) is $(\bar{L}, \underline{L})$-smooth adaptable on C.*

Now we describe the framework of the BPSGE algorithm as follows.

Remark 1.
- *If $\tilde{\nabla} f(\cdot) = \nabla f(\cdot)$ and $\beta_k = 0$, the BPSGE algorithm reduces to BPG algorithm (Bolte et al. 2018).*
- *If $\tilde{\nabla} f(\cdot) = \nabla f(\cdot)$, the BPSGE algorithm reduces to BPGE, which is a special case of CoCaIn BPG algorithm (Mukkamala et al. 2020) and is similar to the BPGE algorithm in (Zhang et al. 2019).*
- *If $\beta_k = 0$, BPSGE reduces to BPSG algorithm (Wang and Han 2023).*

It is not hard to verify that $\beta_k = 0$ always satisfies (5). If $\psi := \frac{1}{2}\|\cdot\|^2$, it follows from (5) that

$$\beta_k^2 \|x_k - x_{k-1}\|^2 \leq \frac{\delta - \epsilon}{1 + \underline{L}\eta_{k-1}} \|x_k - x_{k-1}\|^2. \quad (7)$$

Furthermore, it follows from $\underline{L} = \bar{L}$ and $\delta - \epsilon < 1$ that $\beta_k < 1/\sqrt{2}$. This insight helps us to choose a proper β_k in the numerical experiments.

An important property in our theoretical analysis is the variance reduction given as follows. This definition is similar to that in (Driggs et al. 2021; Wang and Han 2023).

Definition 4. *(Variance reduced stochastic gradient) We say a gradient estimator $\tilde{\nabla} f(\bar{x}_k)$ in Algorithm 1 is variance reduced with constants $V_1, V_2, V_\Gamma \geq 0$, and $\tau \in (0,1]$ if it satisfies the following three conditions:*

- *(MSE (mean squared error) bound): there exist two sequences of random variables $\{\Gamma_k\}_{k\geq 1}$, $\{\Upsilon_k\}_{k\geq 1}$, such that*

$$\begin{aligned}&\mathbb{E}_k[\|\tilde{\nabla} f(\bar{x}_k) - \nabla f(\bar{x}_k)\|_*^2]\\ \leq&\Gamma_k + V_1\left(\|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2\right),\end{aligned} \tag{8}$$

and

$$\begin{aligned}&\mathbb{E}_k[\|\tilde{\nabla} f(\bar{x}_k) - \nabla f(\bar{x}_k)\|_*]\\ \leq&\Upsilon_k + V_2\left(\|x_k - x_{k-1}\| + \|x_{k-1} - x_{k-2}\|\right),\end{aligned} \tag{9}$$

respectively. Here, $\|\cdot\|_*$ denotes the conjugate norm (Rockafellar 1970) of $\|\cdot\|$.

- *(Geometric decay): The sequence $\{\Gamma_k\}_{k\geq 1}$ satisfies the following inequality in expectation:*

$$\begin{aligned}&\mathbb{E}_k[\Gamma_{k+1}] \leq (1-\tau)\Gamma_k\\ &+ V_\Gamma\left(\|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2\right).\end{aligned} \tag{10}$$

- *(Convergence of estimator): If the sequence $\{x_k\}_{k=0}^{\infty}$ satisfies $\lim_{k\to\infty} \mathbb{E}\|x_k - x_{k-1}\|^2 \to 0$, then it follows that $\mathbb{E}\Gamma_k \to 0$ and $\mathbb{E}\Upsilon_k \to 0$.*

Remark 2.
- *In Proposition 1, we show SAGA and SARAH stochastic estimators satisfy Definition 4.*
- *If $\beta_k = 0$, the terms $\|x_{k-1} - x_{k-2}\|^2$ and $\|x_{k-1} - x_{k-2}\|$ in* (8), (9) *and* (10) *are not necessary.*
- *We do not require the stochastic gradient to be unbiased or the variance to be bounded.*

Convergence Analysis

This section presents a comprehensive analysis of the convergence properties. All proofs are placed in the appendix.

Subsequential Convergence Analysis

We first show the descent amount of $\Phi(x_k)$ as follows.

Lemma 1. *Suppose Assumptions 1-2 are satisfied and $\tilde{\nabla} f(\bar{x}_k)$ satisfies the variance reduction property defined by Definition 4. Let $\{x_k\}$ be the sequence generated by Algorithm 1. Then the following inequality holds for any $k > 0$,*

$$\begin{aligned}&\mathbb{E}_k[\Phi(x_{k+1})] + (1/\eta_k - \alpha - \gamma)\,\mathbb{E}_k[D_\psi(x_k, x_{k+1})]\\ &+ \frac{1}{2\tau\gamma}\mathbb{E}_k[\Gamma_{k+1}]\\ \leq&\Phi(x_k) + \frac{1}{2\tau\gamma}\Gamma_k + \left(\frac{\delta-\epsilon}{\eta_k} + \frac{\gamma}{2}\right) D_\psi(x_{k-1}, x_k)\\ &+ \frac{\gamma}{2} D_\psi(x_{k-2}, x_{k-1}).\end{aligned}$$

Here, $\gamma = \sqrt{2(V_\Gamma/\tau + V_1)}$, α is the weakly convex parameter in Assumption 2, δ and ϵ are introduced in (5), *and $V_1, V_2, V_\Gamma \geq 0$, $\tau \in (0, 1]$ are parameters in Definition 4.*

We introduce a new Lyapunov function and show it is monotonically decreasing in expectation.

Lemma 2. *Suppose the same conditions with Lemma 1 hold, and the stepsize satisfies*

$$\eta_k \leq \min\left\{\eta_{k-1}, \frac{1}{\bar{L}}, \frac{1-\delta}{\alpha + 2\gamma}\right\},\ \forall\, k > 0. \tag{11}$$

Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by BPSGE (Algorithm 1) and define the following Lyapunov sequence

$$\begin{aligned}\Psi_{k+1} :=&\eta_k(\Phi(x_{k+1}) - \mathcal{V}(\Phi)) + t_k D_\psi(x_k, x_{k+1})\\ &+ \eta_k\left(\frac{\gamma}{2} + \frac{\epsilon}{3\eta_k}\right) D_\psi(x_{k-1}, x_k) + \frac{\eta_k \Gamma_{k+1}}{2\tau\gamma},\end{aligned} \tag{12}$$

where $t_k = 1 - \eta_k\alpha - \eta_k\gamma - \epsilon/3$. Then, for all $k \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}_k[\Psi_{k+1}] \leq \Psi_k - \frac{\epsilon}{3}(&\mathbb{E}_k[D_\psi(x_k, x_{k+1})]\\ &+ D_\psi(x_{k-1}, x_k) + D_\psi(x_{k-2}, x_{k-1})).\end{aligned} \tag{13}$$

Based on Lemma 2, we get the subsequential convergence of BPSGE.

Theorem 1. *Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by BPSGE algorithm. Then, the following statements hold.*

- *The sequence $\{\mathbb{E}[\Psi_k]\}_{k\in\mathbb{N}}$ is nonincreasing.*
- *$\sum_{k=1}^{+\infty} \mathbb{E}[D_\psi(x_{k-1}, x_k)] < +\infty$, and the sequence $\{\mathbb{E}[D_\psi(x_{k-1}, x_k)]\}$ converges to zero.*
- *$\min_{1\leq k\leq K} \mathbb{E}[D_\psi(x_{k-1}, x_k)] \leq 3\Psi_1/(\epsilon K)$.*

Global Convergence Analysis

Now we show the convergence of the whole sequence to a stochastic stationary point under more conditions. Consider the case $C = \mathbb{R}^d$. We require the following additional assumptions.

Assumption 3.
- *$\nabla f_i(x)$ is Lipschitz continuous with constant $M_1 > 0$ on any bounded subset of $\mathbb{R}^d$.*
- *$\nabla\psi$ is Lipschitz continuous with constant $M_2 > 0$ on any bounded subset of $\mathbb{R}^d$.*

From Assumption 3 (1), the function $\nabla f(x)$ is also Lipschitz continuous with constant $M_1 > 0$ on any bounded subset of $\mathbb{R}^d$. Furthermore, we show SAGA (Defazio, Bach, and Lacoste-Julien 2014) and SARAH (Nguyen et al. 2017) satisfy the following proposition.

Proposition 1. *Assume Assumption 3 (1) holds and the mini-batch B_k is uniform randomly chosen from $\{1, \ldots, n\}$ with cardinality b, i.e., $b = |B_k|$.*

- *The SAGA gradient estimator (Defazio, Bach, and Lacoste-Julien 2014)*

$$\tilde{\nabla}^{SAGA} f(\bar{x}_k) = \frac{1}{b}\Big(\sum_{j\in B_k} \nabla f_j(\bar{x}_k) - g_k^j\Big) + \frac{1}{n}\sum_{i=1}^{n} g_k^i,$$

$$g_{k+1}^i = \begin{cases}\nabla f_i(\bar{x}_k), & \text{if } i \in B_k,\\ g_k^i, & \text{otherwise.}\end{cases}$$

is variance reduced with

$$\Gamma_{k+1} := \frac{1}{bn}\sum_{i=1}^{n} \|\nabla f_i(\bar{x}_k) - \nabla f_i(z_k^i)\|_*^2,$$

$$\Upsilon_{k+1} := \frac{1}{\sqrt{bn}}\sum_{i=1}^{n} \|\nabla f_i(\bar{x}_k) - \nabla f_i(z_k^i)\|_*,$$

where $z_k^i = x_k$ *if* $i \in B_k$ *and* $z_k^i = z_{k-1}^i$ *otherwise. The constants* $\tau = \frac{b}{2n}$, $V_\Gamma = \frac{2b+4n}{b^2} M_1^2$, $V_1 = V_2 = M_1$ $V_1 = M_1^2, V_2 = M_1$.

- *The SARAH gradient estimator (Nguyen et al. 2017)*

$$\begin{aligned}&\tilde{\nabla}^{SARAH} f(\bar{x}_k)\\ =&\begin{cases}\nabla f(\bar{x}_k), & \textit{w.p. } \frac{1}{p},\\ \frac{1}{b}\big(\sum\limits_{j\in B_k} \nabla f_j(\bar{x}_k) - \nabla f_j(\bar{x}_{k-1})\big) & \\ \quad + \tilde{\nabla}^{SARAH} f(\bar{x}_{k-1}), & \textit{otherwise.}\end{cases}\end{aligned}$$

is variance reduced with

$$\begin{aligned}\Gamma_{k+1} =& \|\tilde{\nabla}^{SARAH} f(\bar{x}_k) - \nabla f(\bar{x}_k)\|_*^2,\\ \Upsilon_{k+1} =& \|\tilde{\nabla}^{SARAH} f(\bar{x}_k) - \nabla f(\bar{x}_k)\|_*,\end{aligned}$$

and constants $\tau = \frac{1}{p}, V_1 = V_\Gamma = 2M_1^2, V_2 = 2M_1$. *Here "w.p.* $\frac{1}{p}$*" means with probability* $\frac{1}{p} \in (0, 1]$.

Corollary 1. *If* $C = \mathbb{R}^d$*, from Lemma 2, it shows that*

$$\begin{aligned}\mathbb{E}_k[\Psi_{k+1}] \leq \Psi_k - \frac{\epsilon}{6}(&\mathbb{E}_k\|x_{k+1} - x_k\|^2\\ &+ \|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2).\end{aligned}$$

Now we prove the following bound for $\partial\Phi(x_k)$.

Lemma 3. *Suppose that Assumptions 1-3 hold and the stepsize* η_k *satisfies* $0 < \eta \leq \eta_k$[1] *and* (11)*. Let* $\{x_k\}_{k\in\mathbb{N}}$ *be a bounded sequence generated by the BPSGE algorithm. Define*

$$w_{k+1} := \nabla f(x_{k+1}) - \tilde{\nabla} f(\bar{x}_k) + \frac{1}{\eta_k}(\nabla\psi(\bar{x}_k) - \nabla\psi(x_{k+1})).$$

Then, we have $w_{k+1} \in \partial\Phi(x_{k+1})$ *and*

$$\begin{aligned}&\mathbb{E}_k\|w_{k+1}\| \leq \rho(\mathbb{E}_k\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\\ &+ \|x_{k-1} - x_{k-2}\|) + \Upsilon_k,\end{aligned}$$

where $\rho = \max\{M_1 + \frac{M_2}{\eta}, V_2 + \beta_k M_1 + \frac{\beta_k M_2}{\eta}, V_2\}$.

Similarly, we show the bound for $\mathbb{E}[\mathrm{dist}(0, \partial\Phi(x_{k+1}))^2]$.

Lemma 4. *Under the same conditions as in Lemma 3, there exists a constant* $\bar{\rho} > 0$ *such that*

$$\begin{aligned}&\mathbb{E}[\mathrm{dist}(0, \partial\Phi(x_{k+1}))^2] \leq \bar{\rho}\mathbb{E}[\|x_{k+1} - x_k\|^2\\ &+ \|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2] + 3\mathbb{E}\Gamma_k.\end{aligned}$$

Using Lemma 4, we show the $\mathcal{O}(1/\epsilon^2)$ complexity in expectation to obtain an ϵ-stationary point.

Theorem 2. *Assume that Assumptions 1-3 hold, and the stepsize satisfies* $0 < \eta \leq \eta_k$ *and* (11)*. Let* $\{x_k\}_{k\in\mathbb{N}}$ *be a bounded sequence generated by the BPSGE algorithm. Then there exists some* $0 < \sigma < \epsilon/6$ *such that*

$$\begin{aligned}\mathbb{E}[\mathrm{dist}(0, \partial\Phi(x_{\hat{k}}))^2] \leq& \frac{\bar{\rho}}{(\epsilon/6 - \sigma)K}(\mathbb{E}\Psi_1 + \frac{\epsilon/2 - 3\sigma}{\tau\bar{\rho}}\mathbb{E}\Gamma_1)\\ =&\mathcal{O}(1/K),\end{aligned}$$

where $\hat{k}$ *is drawn from* $\{2, \ldots, K + 1\}$*. In other words, it takes at most* $\mathcal{O}(1/\epsilon^2)$ *iterations in expectation to obtain an* ϵ*-stationary point (Definition 3) of* Φ.

[1] η is the lower bound of η_k.

We define the set of limit points of $\{x_k\}_{k=0}^\infty$ as

$$\begin{aligned}\omega(x_0) := \{&x : \exists \text{ an increasing sequence of integers } \{k_l\}_{l\in\mathbb{N}}\\ &\text{s.t. } x_{k_l} \to x \text{ as } l \to \infty\}.\end{aligned}$$

Now we get the properties of limit points of $\{x_k\}$ as follows.

Lemma 5. *Suppose that Assumptions 1 to 3 hold, the step* η_k *satisfies* $0 < \eta \leq \eta_k$ *and* (11)*. Then the following statements hold.*

- $\sum_{k=0}^\infty \|x_{k+1} - x_k\|^2 < +\infty$ *a.s., and* $\lim_{k\to+\infty} \|x_{k+1} - x_k\| \to 0$ *a.s.*
- $\mathbb{E}[\Phi(x_k)] \to \Phi_*$*, where* $\Phi_* \in [\bar{\Phi}, +\infty)$ *with* $\bar{\Phi} := \inf_x \Phi(x)$*, and* $\mathbb{E}\Phi(x_*) = \Phi_*$ *for all* $x_* \in \omega(x_0)$.
- $\mathbb{E}[\mathit{dist}(0, \partial\Phi(x_k))] \to 0$*. Moreover, the set* $\omega(x_0)$ *is nonempty, and* $\mathbb{E}[\mathit{dist}(0, \partial\Phi(x_*))] = 0$ *for all* $x_* \in \omega(x_0)$.
- $\mathit{dist}(x_k, \omega(x_0)) \to 0$ *a.s., and* $\omega(x_0)$ *is a.s. compact and connected.*

We further show the whole sequence convergence under the KŁ property.

Theorem 3. *Suppose that Assumptions 1-3 hold, the step* η_k *satisfies* $0 < \eta \leq \eta_k$ *and* (11)*. Let* $\{x_k\}_{k\in\mathbb{N}}$ *be a bounded sequence generated by the BPSGE algorithm. If the optimization function* $\Phi(x)$ *is a semialgebraic function that satisfies the KŁ property with exponent* $\theta \in [0, 1)$ *(see Lemma 6 in Appendix), then either the point* x_k *is a critical point after a finite number of iterations or the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *almost surely satisfies the finite length property in expectation, namely,*

$$\sum_{k=0}^{+\infty} \mathbb{E}\|x_{k+1} - x_k\| < +\infty.$$

Numerical Experiments

In this section, we present our numerical study on the practical performance of the proposed BPSGE algorithm with three different stochastic gradient estimators. For all numerical experiments, $C = \mathbb{R}^d$. The experiments are implemented in Matlab 2020b and conducted on a computer with AMD Ryzen 5 5600x 6-Core 3.7 GHz and 48GB memory.

For the stochastic algorithms, we repeat all numerical experiments 10 times and report their average performance. All the initial points are generated by the uniform distribution between 0 and 0.1 and are the same for all algorithms. Inspired by (7), we set $\beta^k = 0.6\frac{k-1}{k+2}$ for simplicity[2] in the BPGE and BPSGE-SGD/SAGA/SARAH algorithms. The stepsize is set as $\eta_k = \min(\eta_{k-1}, L_k^{-1})$, where L_k is the approximated Lipschitz constant estimated by the power method to $\bar{V}_k$ and $\bar{U}_k$ for U- and V-update, respectively. This choice of stepsize is the same for all compared algorithms.

[2] The inequality (5) in Algorithm 1 is required for our convergence analysis. It is time-consuming to check this inequality in numerical experiments. Due to this reason, we directly set $\beta^k = 0.6\frac{k-1}{k+2}$. Our numerical experiments show that BPSGE always converges with this β^k.

Dataset	BPG	BPSG			BPGE	BPSGE		
		-SGD	-SARAH	-SAGA		-SGD	-SARAH	-SAGA
COIL20	75.56	79.30	85.47	85.83	85.33	87.48	89.37	**89.97**
PIE	77.21	84.85	86.03	86.33	84.27	87.79	88.45	**88.83**
COIL100	73.06	80.70	82.23	82.46	76.25	82.24	84.02	**84.29**
TDT2	68.08	82.01	85.04	85.22	77.92	85.21	87.42	**87.54**

Table 2: Comparison of clustering accuracy (%) on four datasets by graph regularized NMF.

Graph Regularized NMF for Clustering

Previous studies (Cai et al. 2011; Shahnaz et al. 2006; Ahmed et al. 2021) show that NMF is powerful for clustering problems, especially in document clustering and image clustering tasks. We consider the graph regularized NMF problem for clustering proposed in (Cai et al. 2011) which is defined as

$$\min_{U\in\mathbb{R}_+^{m\times r},V\in\mathbb{R}_+^{r\times d}} \frac{1}{2}\|M-UV\|_F^2+\frac{\mu_0}{2}\mathrm{Tr}(U^TLU), \quad (14)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, L denotes the graph Laplace matrix, and μ_0 is a positive parameters. This model can help to distinguish anomalies from normal observation (Ahmed et al. 2021). We first define the following kernel-generating distance, i.e.,

$$\begin{aligned}\psi_1(U,V) &:= \left(\|U\|_F^2/2+\|V\|_F^2/2\right)^2,\\ \psi_2(U,V) &:= \|U\|_F^2/2+\|V\|_F^2/2,\end{aligned} \quad (15)$$

which is designed to allow for closed-form update in the BPSGE algorithm. Let

$$\begin{aligned}f(U,V) &:= \frac{1}{2}\|M-UV\|_F^2+\frac{\mu_0}{2}\mathrm{Tr}(U^TLU),\\ h(U,V) &:= I_{U\geq 0}+I_{V\geq 0},\\ \psi(U,V) &:= 3\psi_1(U,V)+(\|M\|_F+\mu_0\|L\|_F)\psi_2(U,V),\end{aligned} \quad (16)$$

where $I_{U\geq 0}$ is the indicator function. Now we give the closed form of (U_{k+1},V_{k+1}) as follows.

Proposition 2. *With the above defined f,h,ψ in* (16)*, the update steps* (6) *in each iteration are given by*

$$U_{k+1}=t\Pi_+(-P_k), \quad V_{k+1}=t\Pi_+(-Q_k),$$

where

$$\begin{aligned}P_k &= \eta_k\tilde{\nabla}_U f(\bar{U}_k,\bar{V}_k)-\nabla_U\psi(\bar{U}_k,\bar{V}_k),\\ Q_k &= \eta_k\tilde{\nabla}_V f(\bar{U}_k,\bar{V}_k)-\nabla_V\psi(\bar{U}_k,\bar{V}_k).\end{aligned} \quad (17)$$

Here, $\Pi_+(\cdot)$ is the projection onto the nonnegative orthant, and $t\geq 0$ satisfies

$$\begin{aligned}&3\left(\|\Pi_+(-P_k)\|_F^2+\|\Pi_+(-Q_k)\|_F^2\right)t^3\\ &+(\|M\|_F+\mu_0\|L\|_F)t-1=0.\end{aligned}$$

We use four datasets *COIL20*, *PIE*, *COIL100*, and *TDT2* [3] to illustrate the numerical performance of the proposed algorithm. In this numerical experiment, we let $\mu_0=100$, and $r=20$ for *COIL20*, $r=68$ for *PIE*, $r=100$ for *COIL100*, and $r=30$ for *TDT2*, respectively. Here we conduct experiments with 50 epochs using the minibatch subsampling ratio 5%.

After solving (14) by Proposition 2, we compute the clustering label by implementing the K-means method (Cai 2011) on U. All results are presented in Table 2. From this table, it shows that the extrapolation technique can improve the numerical performance. In addition, the stochastic algorithms can get better numerical results than their deterministic versions. Furthermore, the variance reduction stochastic gradient estimators can get the best performance in the stochastic framework.

MF with Weakly-convex Regularization

Consider the following optimization problem with a weakly-convex regularization (Yin et al. 2015; Ma, Lou, and Huang 2017)

$$\min_{U,V}\frac{1}{2}\|M-UV\|_F^2+\lambda_1\|U\|_1-\frac{\lambda_2}{2}\|U\|_F^2, \quad (18)$$

where $U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times d}$, $\|U\|_1:=\sum_{i,j}|U_{i,j}|$ and λ_1,λ_2 are two positive parameters. The term $\lambda_1\|U\|_1-\frac{\lambda_2}{2}\|U\|_F^2$ is a λ_2-weakly convex function.

Now we give the closed-form of (U_{k+1},V_{k+1}) for the problem (18) in the following proposition.

Proposition 3. *If $f(U,V):=\frac{1}{2}\|M-UV\|_F^2$, $h(U,V):=\lambda_1\|U\|_1-\frac{\lambda_2}{2}\|U\|_F^2$, $\psi(U,V):=3\psi_1(U,V)+\|M\|_F\psi_2(U,V)+\frac{\eta\lambda_2}{2}\|U\|_F^2$ (where $\eta=\eta_k$ in the k-th iteration), we have the update steps for solving* (6) *in each iteration are*

$$U_{k+1}=t\mathcal{S}_{\lambda_1\eta_k}(-P_k), \quad V_{k+1}=-tQ_k,$$

where P_k and Q_k are defined by (17)*, $t\geq 0$ satisfies*

$$3(\|\mathcal{S}_{\lambda_1\eta_k}(-P_k)\|_F^2+\|-Q_k\|_F^2)t^3+\|M\|_Ft-1=0,$$

and $\mathcal{S}_{\lambda_1\eta_k}(\cdot)$ is the soft-thresholding operator[4] *(Donoho 1995).*

We use two real datasets[5], i.e., *ORL* with the size of 4096×400 and *Yale-B* with the size of 2414×1024, to illustrate the numerical performance. We let $\lambda_1=0.05$ and $\lambda_2=0.02$ for all numerical experiments, and let $r=25$ for *ORL* dataset and $r=49$ for *Yale-B* dataset, respectively. We conduct experiments with 200 epochs using the minibatch subsampling ratio 5%.

[3]http://www.cad.zju.edu.cn/home/dengcai/Data/data.html

[4]For any $y\in\mathbb{R}^d$, $S_\tau(y)=\arg\min_{x\in\mathbb{R}^d}\{\tau\|x\|_1+\frac{1}{2}\|x-y\|^2\}=\max\{|y|-\tau,0\}\mathrm{sign}(y)$.

[5]http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

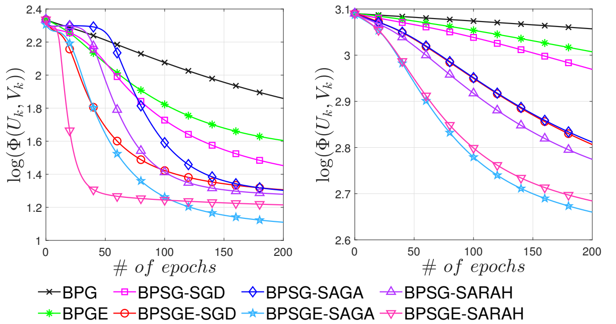


Figure 1: Numerical experiment results on *ORL* and *Yale-B* datasets for problem (18). Left: *ORL* with $r = 25$. Right: *Yale-B* with $r = 49$.

Fig. 1 indicates that the extrapolation-based algorithms perform better than those without extrapolation. Furthermore, the variance reduced stochastic gradient algorithms (with SAGA/SARAH) show better performance compared to SGD-based methods.

NMF with Nonconvex Sparsity Constraints

We continue to study the NMF problem with nonconvex sparsity constraints given by

$$\min_{U,V} \left\{ \frac{1}{2}\|M - UV\|_F^2 : \|U_{:,i}\|_0 \le s_1, \|V_{j,:}\|_0 \le s_2 \right\}, \quad (19)$$

where $U \in \mathbb{R}_+^{m\times r}, V \in \mathbb{R}_+^{r\times d}$, $r > 0$, $i, j \in \{1, 2 \ldots, r\}$, $U_{:,i}$ denotes the i-th column of U and $\|U_{:,i}\|_0$ denotes the number of non-zero entries of $U_{:,i}$. Similarly, $V_{j,:}$ is the j-th row of V. Now, we denote

$$\begin{aligned} f(U,V) :=& \frac{1}{2}\|M - UV\|_F^2, \\ h(U,V) :=& I_{U\ge 0} + I_{\|U_{:,1}\|_0 \le s_1} + \cdots + I_{\|U_{:,r}\|_0 \le s_1} \\ &+ I_{V\ge 0} + I_{\|V_{1,:}\|_0 \le s_2} + \cdots + I_{\|V_{r,:}\|_0 \le s_2}, \\ \psi(U,V) :=& 3\psi_1(U,V) + \|M\|_F \psi_2(U,V). \end{aligned}$$

Here, ψ_1 and ψ_2 are given by (15). The closed-form of (U_{k+1}, V_{k+1}) is the same as that of (Mukkamala and Ochs 2019, Proposition D.8) and is shown as follows.

Proposition 4. *Given the optimization problem* (19) *with the above defined* $f(\cdot)$, $h(\cdot)$ *and* $\psi(\cdot)$, *the update* (6) *in the BPSGE algorithm are given by*

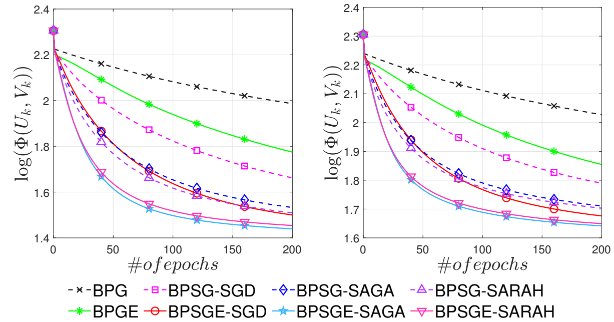
$$U_{k+1} = t\mathcal{H}_{s_1}(\Pi_+(-P_k)), \quad V_{k+1} = t\mathcal{H}_{s_2}(\Pi_+(-Q_k)),$$

where P_k *and* Q_k *are defined by* (17), $t \ge 0$ *and satisfies*

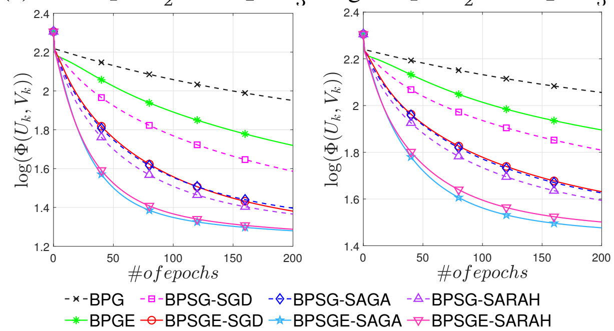
$$3(\|\mathcal{H}_{s_1}(\Pi_+(-P_k))\|_F^2 + \|\mathcal{H}_{s_2}(\Pi_+(-Q_k))\|_F^2)t^3 + \|M\|_F t - 1 = 0.$$

Here, $\Pi_+(\cdot)$ *is the projection on the nonnegative space,* $\mathcal{H}_s(\cdot)$ *is the hard-thresholding operator*[6] *(Luss and Teboulle 2013).*

Fig. 2 shows that the BPGE algorithm outperforms the BPG algorithm, thanks to its extrapolation technique. The adaptive stepsize and the variance reduction techniques also prove to be effective in the stochastic framework.



(a) Left: $s_1 = \frac{m}{2}$ and $s_2 = \frac{d}{3}$. Right: $s_1 = \frac{m}{2}$ and $s_2 = \frac{d}{5}$.



(b) Left: $s_1 = \frac{m}{2}$ and $s_2 = \frac{d}{2}$. Right: $s_1 = \frac{m}{3}$ and $s_2 = \frac{d}{2}$.

Figure 2: Numerical experiments for MF problem (19).

Conclusion

This paper presented a Bregman proximal stochastic gradient descent algorithm with extrapolation (BPSGE) for the objective function that lacks a global Lipschitz continuous gradient. Under certain suitable conditions, we established the subsequential convergence, demonstrated that the subgradient of the objective function exhibits a sublinear convergence rate, and established the global convergence of the sequence. At last, we conducted numerical experiments on three specific applications and demonstrated the superior performance of the BPSGE algorithm.

Acknowledgments

The authors are grateful to the anonymous referees for their valuable comments and suggestions. This research is supported by the National Natural Science Foundation of China (NSFC) grants 12126608, 12131004, 12126603, the R&D project of Pazhou Lab (Huangpu) (Grant no. 2023K0603), and the Fundamental Research Funds for the Central Universities (Grant No. YWF-22-T-204).

[6]Let $y \in \mathbb{R}^d$ and $|y_1| \ge |y_2| \ge \cdots \ge |y_d|$. Then the hard-thresholding operator is given by

$$\mathcal{H}_s(y) = \arg\min_{x\in\mathbb{R}^d} \{\|x - y\|^2 : \|x\|_0 \le s\} = \begin{cases} y_i, & i \le s, \\ 0, & \text{otherwise}. \end{cases}$$

where $s > 0$ and the operations are applied element-wise.

References

Ahmed, I.; Hu, X. B.; Acharya, M. P.; and Ding, Y. 2021. Neighborhood Structure Assisted Non-negative Matrix Factorization and Its Application in Unsupervised Point-wise Anomaly Detection. *Journal of Machine Learning Research*, 22: 34:1–34:32.

Ahookhosh, M.; Themelis, A.; and Patrinos, P. 2021. A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima. *SIAM Journal on Optimization*, 31(1): 653–685.

Auslender, A.; and Teboulle, M. 2006. Interior Gradient and Proximal Methods for Convex and Conic Optimization. *SIAM Journal on Optimization*, 16(3): 697–725.

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(2): 330–348.

Bauschke, H. H.; Dao, M. N.; and Lindstrom, S. B. 2018. Regularizing with Bregman-Moreau Envelopes. *SIAM Journal on Optimization*, 28(4): 3208–3228.

Bolte, J.; Sabach, S.; Teboulle, M.; and Vaisbourd, Y. 2018. First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. *SIAM Journal on Optimization*, 28(3): 2131–2151.

Bottou, L. 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. In *19th International Conference on Computational Statistics, COMPSTAT*, 177–186.

Bregman, L. 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(3): 200–217.

Cai, D. 2011. Litekmeans: the fastest MATLAB implementation of kmeans. http://www.cad.zju.edu.cn/home/dengcai/Data/Clustering.html. Accessed: 2011-12-10.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph Regularized Non-negative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8): 1548–1560.

Che, M.; and Wei, Y. 2020. Multiplicative Algorithms for Symmetric Nonnegative Tensor Factorizations and Its Applications. *Journal of Scientific Computing*, 83(3): 53.

Comon, P.; Golub, G. H.; Lim, L.; and Mourrain, B. 2008. Symmetric Tensors and Symmetric Tensor Rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3): 1254–1279.

Defazio, A.; Bach, F. R.; and Lacoste-Julien, S. 2014. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *Advances in Neural Information Processing Systems 27*, 1646–1654.

Defazio, A.; Domke, J.; and Caetano, T. S. 2014. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, 1125–1133.

Donoho, D. L. 1995. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3): 613–627.

Dragomir, R.; Taylor, A. B.; d'Aspremont, A.; and Bolte, J. 2022. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194(1): 41–83.

Driggs, D.; Tang, J.; Liang, J.; Davies, M. E.; and Schönlieb, C. 2021. A Stochastic Proximal Alternating Minimization for Nonsmooth and Nonconvex Optimization. *SIAM Journal on Imaging Sciences*, 14(4): 1932–1970.

Fan, J.; and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360.

Gillis, N. 2020. *Nonnegative Matrix Factorization*. SIAM.

Hasannasab, M.; Hertrich, J.; Neumayer, S.; Plonka, G.; Setzer, S.; and Steidl, G. 2020. Parseval Proximal Neural Networks. *Journal of Fourier Analysis and Applications*, 26(59).

He, Z.; Xie, S.; Zdunek, R.; Zhou, G.; and Cichocki, A. 2011. Symmetric Nonnegative Matrix Factorization: Algorithms and Applications to Probabilistic Clustering. *IEEE Transactions on Neural Networks*, 22(12): 2117–2131.

Kolda, T. G.; and Bader, B. W. 2009. Tensor Decompositions and Applications. *SIAM Review*, 51(3): 455–500.

Lan, G. 2020. *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer.

Latafat, P.; Themelis, A.; Ahookhosh, M.; and Patrinos, P. 2022. Bregman Finito/MISO for Nonconvex Regularized Finite Sum Minimization without Lipschitz Gradient Continuity. *SIAM Journal on Optimization*, 32(3): 2230–2262.

Laude, E.; Ochs, P.; and Cremers, D. 2020. Bregman Proximal Mappings and Bregman-Moreau Envelopes Under Relative Prox-Regularity. *Journal of Optimization Theory and Applications*, 184(3): 724–761.

Lee, D. D.; and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 788–791.

Li, W.; Wang, Z.; Zhang, Y.; and Cheng, G. 2022. Variance reduction on general adaptive stochastic mirror descent. *Machine Learning*, 111: 4639–4677.

Lin, Z.; Li, H.; and Fang, C. 2020. *Accelerated Optimization for Machine Learning*. Springer.

Lu, H.; Freund, R. M.; and Nesterov, Y. E. 2018. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1): 333–354.

Luss, R.; and Teboulle, M. 2013. Conditional Gradient Algorithms for Rank-One Matrix Approximations with a Sparsity Constraint. *SIAM Review*, 55(1): 65–98.

Ma, T.; Lou, Y.; and Huang, T. 2017. Truncated l_{1-2} Models for Sparse Recovery and Rank Minimization. *SIAM Journal on Imaging Sciences*, 10(3): 1346–1380.

Mairal, J. 2015. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2): 829–855.

Malitsky, Y.; and Tam, M. K. 2020. A Forward-Backward Splitting Method for Monotone Inclusions Without Cocoercivity. *SIAM Journal on Optimization*, 30(2): 1451–1472.

Mukkamala, M. C.; and Ochs, P. 2019. Beyond Alternating Updates for Matrix Factorization with Inertial Bregman Proximal Gradient Algorithms. In *Advances in Neural Information Processing Systems 32*, 4268–4278.

Mukkamala, M. C.; Ochs, P.; Pock, T.; and Sabach, S. 2020. Convex-Concave Backtracking for Inertial Bregman Proximal Gradient Algorithms in Nonconvex Optimization. *SIAM Journal on Mathematics of Data Science*, 2(3): 658–682.

Nesterov, Y. E. 1983. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2): 372–376.

Nguyen, L. M.; Liu, J.; Scheinberg, K.; and Takác, M. 2017. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. In *Proceedings of the 34th International Conference on Machine Learning*, 2613–2621.

Polyak, B. T. 1964. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5): 1–17.

Reem, D.; Reich, S.; and Pierro, A. R. D. 2019. A Telescopic Bregmanian Proximal Gradient Method Without the Global Lipschitz Continuity Assumption. *Journal of Optimization Theory and Applications*, 182(3): 851–884.

Robbins, H.; and Monro, S. 1951. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3): 400–407.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton, NJ, USA: Princeton University Press.

Shahnaz, F.; Berry, M. W.; Pauca, V. P.; and Plemmons, R. J. 2006. Document clustering using nonnegative matrix factorization. *Information Processing and Management*, 42(2): 373–386.

Teboulle, M.; and Vaisbourd, Y. 2020. Novel Proximal Gradient Methods for Nonnegative Matrix Factorization with Sparsity Constraints. *SIAM Journal on Imaging Sciences*, 13(1): 381–421.

Wang, Q.; and Han, D. 2023. A Bregman Stochastic Method for Nonconvex Nonsmooth Problem Beyond Global Lipschitz Gradient Continuity. *Optimization Methods and Software*, 38(5): 914–946.

Wang, Z.; Themelis, A.; Ou, H.; and Wang, X. 2022. A mirror inertial forward-reflected-backward splitting: Global convergence and line search extension beyond convexity and Lipschitz smoothness. *arXiv: 2212.01504*.

Yin, P.; Lou, Y.; He, Q.; and Xin, J. 2015. Minimization of l_{1-2} for Compressed Sensing. *SIAM Journal on Scientific Computing*, 37(1): A536–A563.

Zhang, C.-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942.

Zhang, S.; and He, N. 2018. On the Convergence Rate of Stochastic Mirror Descent for Nonsmooth Nonconvex Optimization. *arXiv: 1806.04781*.

Zhang, X.; Barrio, R.; Carballo, M. A. M.; Jiang, H.; and Cheng, L. 2019. Bregman Proximal Gradient Algorithm With Extrapolation for a Class of Nonconvex Nonsmooth Minimization Problems. *IEEE Access*, 7: 126515–126529.

Zhao, J.; Dong, Q.; Rassias, M. T.; and Wang, F. 2022. Two-step inertial Bregman alternating minimization algorithm for nonconvex and nonsmooth problems. *Journal of Global Optimization*, 84(4): 941–966.

Zhu, D.; Deng, S.; Li, M.; and Zhao, L. 2021. Level-Set Subdifferential Error Bounds and Linear Convergence of Bregman Proximal Gradient Method. *Journal of Optimization Theory and Applications*, 189(3): 889–918.