Operations Research and Fuzziology 运筹与模糊学, 2024, 14(6), 676-688
Published Online December 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.146568

Hans汉斯

随机邻近牛顿型交替极小化算法

黄静仪，高任杰，李　磊

南京信息工程大学数学与统计学院，江苏　南京

收稿日期：2024年11月7日；录用日期：2024年12月10日；发布日期：2024年12月18日

摘　要

本文研究一类大规模有限和形式的非凸非光滑复合优化问题，其目标函数为适当下半连续凸函数与连续可微函数(有限和形式)之和。邻近交替线性极小化算法在求解这类问题上有显著优势，但考虑到针对大规模优化问题，该算法需在每一迭代计算全梯度，成本较高，故本文将目标函数光滑部分的随机梯度引入到已有算法，以降低算法的计算成本。此外，考虑到一阶算法在求解病态问题时速度较慢，故本文在算法中引入目标函数的二阶信息，提出了一种随机邻近牛顿型交替极小化算法。在适当的步长边界的基础上，我们建立了算法在期望意义下的全局收敛性。本文提出的随机邻近牛顿型交替极小化算法，不仅提升了算法在机器学习、统计学和图像处理等领域实际应用中的效率和实用性，也为非凸非光滑优化领域的理论发展提供了新的理论基础和算法框架。

关键词

非凸非光滑复合优化问题，全局收敛，随机梯度

Stochastic Proximity Newtonian Alternating Miniaturization Algorithm

Jingyi Huang, Renjie Gao, Lei Li

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

Received: Nov. 7th, 2024; accepted: Dec. 10th, 2024; published: Dec. 18th, 2024

Abstract

This article investigates a class of large-scale finite and structured nonconvex nonsmooth composite optimization problems, where the objective function is the sum of an appropriately lower semicontinuous convex function and a continuous differentiable function (finite and structured). The alternating linear minimization algorithm has significant advantages in solving such problems. However, considering the large-scale optimization problems, this algorithm requires computing the full gradient at each iteration, which is costly. Therefore, this article introduces the stochastic gradient of the smooth part of the objective function into the existing algorithm to reduce the

文章引用: 黄静仪, 高任杰, 李磊. 随机邻近牛顿型交替极小化算法[J]. 运筹与模糊学, 2024, 14(6): 676-688.
DOI: 10.12677/orf.2024.146568

computational cost. Additionally, considering that first-order algorithms are slow in solving ill-conditioned problems, this article incorporates the second-order information of the objective function into the algorithm and proposes a stochastic proximal Newton-type alternating minimization algorithm. Based on appropriate step size bounds, we establish the global convergence of the algorithm in the expected sense. The stochastic neighborhood Newton-type alternating minimization algorithm proposed in this paper not only improves the efficiency and practicability of the algorithm in practical applications in the fields of machine learning, statistics and image processing, but also provides a new theoretical foundation and algorithmic framework for the theoretical development in the field of nonconvex nonsmooth optimization.

Keywords

Nonconvex and Nonsmooth Composite Optimization Problem, Global Convergence, Stochastic Gradients

1. 引言

1.1. 研究背景及意义

随着信息技术的高速发展，许多学科领域涌现出了许多大规模优化问题，因此对于优化问题的高效求解，是当今研究的热点问题之一。本文研究的是一类具有有限和形式的非凸非光滑复合优化问题，该问题的具体格式如下：

$$\min \psi(x,y) = f(x) + g(y) + H(x,y), \tag{1.1}$$

其中 $f(x)$ 和 $g(y)$ 是适当下半连续凸函数， $H(x,y) = \sum h_i(x,y), i = 1,2,\cdots,n$ 是连续可微函数。这类模型在许多领域都有应用，包括机器学习、统计学和图像处理等。典型的例子包括非负或稀疏矩阵分解[1]、稀疏(PCA) [2] REF _Ref185269578 \r \h [3]、鲁棒 PCA [4]、最小二乘[5]和盲图像反卷积[6]等。下面我们给出稀疏非负矩阵分解(sparse-NMF)模型，该模型的具体形式为

$$\min \|A - XY\|_F^2,\ X, Y \geq 0,\ \|X_i\|_0 \leq s,\ \ i = 1,\cdots,r,$$

其中 X_i 表示 X 的第 i 列。在字典学习和稀疏编码中，X 被称为系数为 Y 的学习字典。在该模型中，X 上的稀疏性使用非凸 l_0 约束。模型可以表述为问题(1.1)的形式，具体地，我们假设 $f(x) = \delta_{\Omega_1}(X)$，

$\Omega_1 = \left\{X \geq 0, \|X_j\|_0 \leq s, j = 1,\cdots,r\right\}$， $g(Y) = \delta_{\Omega_2}(Y)$， $\Omega_2 = \{Y \geq 0\}$，

$H(X,Y) = \sum h_i(X,Y) = \sum \left(e_p^{\mathrm{T}}(A - XY)e_q\right)^2$，其中 e_j 是第 j 个元素为 1，其他元素均为 0 的向量。形如问题(1.1)的有限和形式的非凸非光滑复合优化问题在机器学习，图像处理和计算机视觉等领域中存在广泛应用，因此研究计算成本低、收敛速度快的算法是很有必要的。

1.2. 研究现状

针对非凸非光滑复合优化问题(1.1)，一个经典的求解算法是交替极小化算法(alternating minimization, AM) [7]，当函数凸且连续可微时，若其中一个参数固定时，关于另外一个参数它是严格凸的，那么用这种方法生成的序列的每一个极限点都是的极小点。AM 算法为一系列研究此问题的算法提供了理论或实

践上的支撑。然而在许多实际问题中，目标函数通常不满足上述较为严格的条件。为克服收敛性条件较强的缺陷，一些学者通过在 AM 算法的子问题中引入邻近项，提出了 PAM 算法[8]，在非凸情形下，当目标函数满足 KL 性质时，该算法的收敛性分析已被建立。然而，PAM 算法涉及两个函数和的邻近算子的计算，这通常是难以求解的。Bolte 等人[9]通过将目标函数中的光滑部分进行线性化处理，提出了邻近交替线性极小化算法(Proximal alternating linearized minimization, PALM)，其具体迭代格式如下所示：

$$x^{k+1}\in\arg\min\left\{f(x)+\left\langle\nabla_x H\left(x^k,y^k\right),x-x^k\right\rangle+\frac{c_k}{2\left\|x-x^k\right\|^2}\right\},$$

$$y^{k+1}\in\arg\min\left\{g(y)+\left\langle\nabla_y H\left(x^{k+1},y^k\right),y-y^k\right\rangle+\frac{d_k}{2\left\|y-y^k\right\|^2}\right\}.$$

需要说明的是，PALM 算法中子问题通常有显式解或易于求解。在非凸情形下，Bolte 等人在 KL 框架下证明了 PALM 算法的全局收敛性。本文正是围绕 PALM 算法的改进展开的。Bolte 等人的工作为本文的研究提供了算法设计的灵感，展示了如何通过交替优化策略来处理非凸非光滑问题，而且他们的全局收敛性证明为本文的算法分析提供了理论基础。

针对非凸非光滑复合优化问题(1.1)，当 n 很大的时候，PALM 算法在每次迭代计算成本将非常昂贵，因为它需要涉及到对全梯度 $\nabla H(x,y):=\frac{1}{n}\sum_{i=1}^{n}\nabla h_i(x,y)$ 的计算，这将导致较大的计算成本。为了克服这个问题，针对 $f=g=0$，仅涉及变量 x 的情形，随机梯度算法(SGD) [10]被提出，但是 SGD 要求步长随迭代进程不断衰减到 0 以弥补随机梯度带来的方差。为提高 SGD 算法的求解效率，一系列方差缩减的随机梯度算法被提出，包括 SVRG [11]，SAGA [12]，SARAH [13]等。随机梯度也被引入到邻近梯度算法，用以邻近梯度算法的降低计算成本，包括 prox-SGD，prox-SVRG，prox-SAGA [14]和 prox-SARAH [15]等。本文在处理大规模优化问题时，在设计算法的迭代步骤时引入随机梯度，大大降低计算成本。

近年来，针对非凸非光滑复合优化问题(1.1)，已有学者将随机梯度策略引入到 PALM 算法中，提出了一系列随机的 PALM 算法。具体的，Xu 和 Yin [16]首先将简单随机梯度下降法(SGD)与 PALM 相结合，提出了块随机梯度法(BSG)并在对目标函数 ψ 的一些较为苛刻的假设条件下，证明了 BSG 方法的收敛性。基于此，Driggs 等人[17]提出了一种随机邻近交替线性极小化(SPRING)的方法，其中他们使用了方差减少的随机梯度算法，而不是 BSG 方法中使用的简单的 SGD 算法。数值实验表明，SPRING 的收敛速度比 BSG 方法快。值得注意的是，与 Xu 和 Yin 的先前工作相比，SPRING 的收敛性是在对目标函数更弱的假设下建立的。

然而，针对病态问题，上述一阶算法存在收敛速度较慢的问题。一个自然的想法是在算法中引入目标函数的二阶信息，以提升算法的求解速度。针对邻近梯度算法，Yang 等人[18]提出了一个随机外推拟牛顿算法(stochastic extra-step quasi-Newton method)，并证明了该算法在期望意义下的次线性收敛率。本文受到了他们方法的启发，在算法中通过残差方程引入目标函数的二阶信息，这些策略有助于提高算法的收敛速度。

本文扩展了 PALM 算法，通过引入随机梯度和二阶信息，提出了随机邻近牛顿型交替极小化算法。在计算效率和收敛速度上都有显著提升，这对于解决大规模优化问题尤为重要。

1.3. 本文的动机与贡献

考虑到一阶算法在处理病态问题时，收敛速度较慢，甚至有时不收敛。针对非凸非光滑复合优化问题(1.1)，我们希望引入高阶信息以提升一阶算法的收敛速度，此外引入随机梯度算法以实现计算量的降

低。具体的，本文将随机邻近牛顿型算法与邻近交替线性极小化算法相结合，提出了随机邻近牛顿型交替极小化算法，并对该算法的收敛性分析进行了证明。

1.4. 文章框架

本文框架如下，在第二节中，给出了本文所需要用到的基本符号、定义以及相关引理。在第三节中，我们给出了随机邻近牛顿型交替极小化算法的具体迭代格式，并给出算法在基本假设下的收敛性分析。在第四节，我们进行了总结。

2. 预备知识

纸型

为便于下文研究本文所提出算法的收敛性，本节将介绍文中涉及到的符号、定义以及相关引理。首先，我们对文中所涉及到的符号做出如下定义：用 $\langle\cdot,\cdot\rangle$ 和 $\|\cdot\| := \|\cdot\|_2$ 表示标准欧几里得内积和范数。对称正定 $n\times n$ 矩阵的集合用 S_{++}^n 表示。对于给定矩阵 $\Lambda \in S_{++}^n$，我们定义内积 $\langle x,y\rangle_\Lambda := \langle x,\Lambda y\rangle = \langle \Lambda x, y\rangle$，$\|x\|_\Lambda := \sqrt{\langle x,x\rangle_\Lambda}$。对于任意 $n\in\mathbb{N}$，我们令 $[n] := \{1,\cdots,n\}$，$[n]_0 = \{0\}\cup[n]$。设 $(\Omega,\mathrm{F},\mathrm{P})$ 为概率空间。我们将使用大写字母来描述随机变量 $X:\Omega\to\mathbb{R}^n$ 和 $Y:\Omega\to\mathbb{R}^n$。而小写字母通常保留给随机变量 $x:\Omega\to\mathbb{R}^n$ 和 $y:\Omega\to\mathbb{R}^n$。我们用 $L^p(\Omega) := L^p(\Omega,\mathrm{P})$，$p\in[1,\infty]$ 来表示 Ω 上的标准 L^p 空间。我们用 $X\in\mathrm{F}$ 表示 X 是可测量的。此外，$\sigma(X^1,\cdots,X^k)$ 表示由随机变量族 $X^1,\cdots,X^k$ 生成。对于随机变量 $X\in L^1(\Omega)$ 和子 σ 代数 $\mathrm{H}\subseteq\mathrm{F}$，给定 H 的 X 的条件期望记为 $\mathrm{E}[X\mid\mathrm{H}]$。我们使用缩写“a.e.”和“a.s.”，分别代表“几乎处处”和“几乎一定”。

定义 2.1 (凸函数的次微分) 设 $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ 为适当下半连续凸函数，对 $x\in\mathrm{dom}\,f$，函数 f 在点 x 处的次微分记作 $\partial f(x)$，定义为所有满足下述条件的 $u\in\mathbb{R}^n$ 所构成的集合：

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

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

定义 2.2 (L-光滑) 对于可微函数 $f:\mathbb{R}^n\to\mathbb{R}$，若存在常数 $L>0$，使其满足下列不等式

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

则称函数 f 是梯度 Lipschitz 连续的，其中 L 为 Lipschitz 常数。

引理 2.1 (下降引理) 函数 $f:\mathbb{R}^n\to\mathbb{R}$ 是可微的且 L-光滑($L>0$)，则它满足下列不等式：

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

3. 随机邻近牛顿型交替极小化算法及其收敛性分析

针对大规模非凸非光滑复合优化问题(1.1)，许多确定型一阶算法已被设计用以对其进行求解。但确定型算法由于计算量较大导致计算成本昂贵，如 PALM 算法在处理有限和形式的非凸非光滑复合优化问题时，需计算光滑部分函数的所有梯度，这将导致较大的计算成本。此外，由于一阶算法在求解病态问题时，求解速度较慢，甚至不收敛，因此将二阶信息引入到算法中，以实现算法的快速高效求解，也是本文需要考虑的一点。

具体的，在算法设计方面，为降低计算成本，提高算法的求解速度，本文将提出一种新颖的算法，即将随机梯度思想引入到 PALM 算法中，并借助广义残差方程引入目标函数的二阶信息，从而设计出随机邻近牛顿型交替极小化算法。在理论分析方面，本文在非凸情形下建立目标函数在期望意义下的下降性分析，进而证明出该算法收敛性。

3.1. 随机邻近牛顿型交替极小化算法

在这一小节，我们给出本文所提算法的具体结构。为更好的描述算法，我们首先给出相关符号解释。

首先，问题(1.1)最优性条件可以等价地改写为一个不动点方程，即点 $x,y\in\mathrm{dom}\,\psi$ 是(1.1)的稳定点当且仅当

$$\begin{cases} R^{\alpha_x}\left(x\right)=x-\mathrm{prox}_{\alpha_x f}\left(x-\alpha_x\nabla_x H\left(x,y\right)\right)=0,\ \alpha_x\in R^+, \\ R^{\alpha_y}\left(y\right)=y-\mathrm{prox}_{\alpha_y g}\left(y-\alpha_y\nabla_y H\left(x,y\right)\right)=0,\ \alpha_y\in R^+. \end{cases}$$

邻近算子是稳定非扩张的，即它是一个常数为 1 的全局 Lipschitz 连续函数，且满足：

$$\begin{cases} \left\|\mathrm{prox}_{\alpha_x f}\left(x_1\right)-\mathrm{prox}_{\alpha_x f}\left(x_2\right)\right\|^2\le\left\langle x_1-x_2,\mathrm{prox}_{\alpha_x f}\left(x_1\right)-\mathrm{prox}_{\alpha_x f}\left(x_2\right)\right\rangle,\ \forall x_1,x_2; \\ \left\|\mathrm{prox}_{\alpha_y g}\left(y_1\right)-\mathrm{prox}_{\alpha_y g}\left(y_2\right)\right\|^2\le\left\langle y_1-y_2,\mathrm{prox}_{\alpha_y g}\left(y_1\right)-\mathrm{prox}_{\alpha_y g}\left(y_2\right)\right\rangle,\ \forall y_1,y_2. \end{cases}$$

给定随机方向 $v_x,v_y\in\mathbb{R}^n$，我们考虑不动点方程的变体

$$\begin{cases} R_{v_x}^{\alpha_x}\left(x\right)=x-\mathrm{prox}_{\alpha_x f}\left(x-\alpha_x v_x\right); \\ R_{v_y}^{\alpha_y}\left(y\right)=y-\mathrm{prox}_{\alpha_y g}\left(y-\alpha_y v_y\right). \end{cases}$$

我们用

$$\begin{cases} p_{v_x}^{\alpha_x}\left(x\right)=\mathrm{prox}_{\alpha_x f}\left(x-\alpha_x v_x\right) \\ p_{v_y}^{\alpha_y}\left(y\right)=\mathrm{prox}_{\alpha_y g}\left(y-\alpha_y v_y\right) \end{cases}$$

分别表示关于 x,y 的随机邻近梯度步。当 $v_x=\nabla_x H\left(x,y\right),v_y=\nabla_y H\left(x,y\right)$ 时，用 $p^{\alpha_x}\left(x\right),p^{\alpha_y}\left(y\right)$ 分别表示关于 x,y 的邻近梯度步。

下面给出算法的具体框架和步骤。

算法 1 随机邻近牛顿型交替极小化算法

初始化：选择正的步长 $\left\{\alpha_x^k\right\}_k$，$\left\{\alpha_{x,+}^k\right\}_k$，$\left\{\lambda_x^k\right\}_k$，$\left\{\beta_x^k\right\}_k$，点 $\left(x^0,y^0\right)\in\mathrm{dom}\,\psi$。

For $k=0,1,\cdots$ **do**

计算随机梯度 $v_x^k\approx\nabla_x H\left(x^k,y^k\right)$，残差

$$R_{v_x^k}^{\alpha_x^k}\left(x^k\right)=x^k-\mathrm{prox}_{\alpha_x^k f}\left(x^k-\alpha_x^k v_x^k\right). \tag{3.1}$$

以及方向 $d_x^k=-W_x^k R_{v_x^k}^{\alpha_x^k}$。计算 $\bar{x}^k=x^k+\beta_x^k d_x^k$ 及随机梯度 $v_{x,+}^k\approx\nabla_x H\left(\bar{x}^k,y^k\right)$，并执行更新

$$x^{k+1}=\mathrm{prox}_{\alpha_{x,+}^k f}\left(x^k+\lambda_x^k d_x^k-\alpha_{x,+}^k v_{x,+}^k\right). \tag{3.2}$$

计算随机梯度 $v_y^k\approx\nabla_y H\left(x^{k+1},y^k\right)$，以及残差

$$R_{v_y^k}^{\alpha_y^k}\left(y^k\right)=y^k-\mathrm{prox}_{\alpha_y^k g}\left(y^k-\alpha_y^k v_y^k\right). \tag{3.3}$$

和方向 $d_y^k=-W_y^k R_{v_y^k}^{\alpha_y^k}$。计算 $\bar{y}^k=y^k+\beta_y^k d_y^k$ 及随机梯度 $v_{y,+}^k\approx\nabla_y H\left(x^{k+1},\bar{y}^k\right)$，并执行更新

$$y^{k+1}=\mathrm{prox}_{\alpha_{y,+}^k g}\left(y^k+\lambda_y^k d_y^k-\alpha_{y,+}^k v_{y,+}^k\right). \tag{3.4}$$

本文算法根据残差方程的随机变体生成方向 d_x 和 d_y，基于此在算法中引入了目标函数的二阶信息。具体来说，本文考虑下述形式的方向

$$\begin{cases} d_x = -W_x R_{v_x}^{\alpha_x}(x); \\ d_y = -W_y R_{v_y}^{\alpha_y}(y). \end{cases}$$

其中矩阵 $W_x, W_y \in \mathbb{R}^{n\times n}$ 用以对基本随机邻近梯度方向 $-R_{v_x}^{\alpha_x}(x), -R_{v_y}^{\alpha_y}(y)$ 进行细化和改进，$v_x \approx \nabla_x H(x,y)$ 是对 $\nabla_x H(x,y)$ 的随机近似，$v_y \approx \nabla_y H(x,y)$ 是对 $\nabla_y H(x,y)$ 的随机近似。针对每一迭代步 x 的更新，我们首先通过 $\bar{x} = x + \beta d$ 计算一个新的点，然后执行额外的邻近梯度步骤，以获得下一个迭代点 x_+：

$$\begin{cases} \bar{x} = x + \beta_x d_x \\ x_+ = \mathrm{prox}_{\alpha_{x,+} f}\left(x + \lambda_x d_x - \alpha_{x,+} v_{x,+}\right). \end{cases}$$

其中 $\lambda_x, \beta_x \geq 0, \alpha_x, \alpha_{x,+} \in \mathbb{R}^+$ 是合适的步长参数，$v_{x,+} \in \mathbb{R}^n$ 是梯度 $\nabla_x H(x,y)$ 的随机近似。y 的更新过程与 x 的更新类似，这里不再赘述。

3.2. 基本假设

为了分析随机邻近牛顿型交替极小化算法的收敛性及收敛率分析，在这一小节我们首先给出保障算法收敛性的相关假设，具体如下。

假设 1 给定函数 $H(x,y): \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$，设

(A.1) $H(x,y)$ 关于 x, y 都连续可微，且梯度映射 $\nabla_x H(x,y)$ 和 $\nabla_y H(x,y)$ 在 $\mathbb{R}^n$ 上是 Lipschitz 连续的，模 $L_x, L_y \geq 1$。

(A.2) 目标函数 ψ 在 $\mathrm{dom}\, f \times \mathrm{dom}\, g$ 上下有界。

本文假设随机近似 v_x^k 和 $v_{x,+}^k$ 对应于随机向量 $V_x^k: \Omega \to \mathbb{R}^n$ 和 $V_{x,+}^k: \Omega \to \mathbb{R}^n$ 的实现，v_y^k 和 $v_{y,+}^k$ 对应于随机向量 $V_y^k: \Omega \to \mathbb{R}^n$ 和 $V_{y,+}^k: \Omega \to \mathbb{R}^n$ 的实现。此外，我们假设概率空间 $(\Omega, \mathrm{F}, \mathrm{P})$ 足够丰富，允许我们以统一的方式建模所涉及的随机过程。我们现在定义滤流

$$\mathrm{F}^k := \sigma\left(V_x^0, V_{x,+}^0, \cdots, V_x^k, V_y^0, V_{y,+}^0, \cdots, V_y^k\right),\ \mathrm{F}_+^k := \sigma\left(\mathrm{F}_x^k \cup \sigma\left(V_{x,+}^k\right) \cup \mathrm{F}_y^k \cup \sigma\left(V_{y,+}^k\right)\right).$$

用 $\left\{D_x^k\right\}_k$ 表示与算法中选择的方向 $\left\{d_x^k\right\}_k$ 相关的随机过程，$\left\{d_y^k\right\}_k$ 表示与算法中选择的方向 $\left\{d_y^k\right\}_k$ 相关的随机过程。我们给出一下随机假设。

假设 2 (B.1) 映射 $D_x^k: \Omega \to \mathbb{R}^n$ 和 $D_y^k: \Omega \to \mathbb{R}^n$ 对于所有 k 是可测的。

(B.2) 存在 $v_k > 0$，使得对于所有 $k \in \mathbb{N}$，都有

$$\mathrm{E}\left[\left\|D_y^k\right\|^2 \mid \mathrm{F}_+^{k-1}\right] \leq v_k^2 \cdot \mathrm{E}\left[\left\|R_{v_y^k}^{\lambda_k}\left(Y^k\right)\right\|^2 \mid \mathrm{F}_+^{k-1}\right] \quad \text{a.e.}$$

(B.3) 对于所有 $k \in \mathbb{N}$，$\mathrm{E}\left[V_x^k \mid \mathrm{F}_+^{k-1}\right] = \nabla_x H\left(X^k, Y^k\right)$，$\mathrm{E}\left[V_{x,+}^k \mid \mathrm{F}^k\right] = \nabla_x H\left(\bar{X}^k, Y^k\right)$，$\mathrm{E}\left[V_y^k \mid \mathrm{F}_+^{k-1}\right] = \nabla_y H\left(X^{k+1}, Y^k\right)$ 与 $\mathrm{E}\left[V_{y,+}^k \mid \mathrm{F}^k\right] = \nabla_y H\left(X^{k+1}, \bar{Y}^k\right)$ a.e.，且存在 $\sigma_k, \sigma_{k,+} > 0$，使得

$$\mathrm{E}\left[\left\|\nabla_x H\left(X^k, Y^k\right) - V^k\right\|^2 \mid \mathrm{F}_+^{k-1}\right] \leq \sigma_k^2 \text{ 和 } \mathrm{E}\left[\left\|\nabla_x H\left(\bar{X}^k, Y^k\right) - V_+^k\right\|^2 \mid \mathrm{F}^k\right] \leq \sigma_{k,+}^2 \quad \text{a.e},$$

$$\mathrm{E}\left[\left\|\nabla_y H\left(X^{k+1}, Y^k\right) - V^k\right\|^2 \mid \mathrm{F}_+^{k-1}\right] \leq \sigma_k^2 \text{ 和 } \mathrm{E}\left[\left\|\nabla_y H\left(X^{k+1}, \bar{Y}^k\right) - V_+^k\right\|^2 \mid \mathrm{F}^k\right] \leq \sigma_{k,+}^2 \quad \text{a.e}。$$

假设 2 (B.3)中 V_x^k 和 $V_{x,+}^k$，V_y^k 和 $V_{y,+}^k$ 的条件在随机优化中很常见[19]。第二个假设要求所选方向 $\left\{d_x^k\right\}_k$ 和随机过程 $\left\{D_x^k\right\}_k$ 与随机非光滑残差 $R_{v_x^k}^{\lambda_k}\left(X^k\right)$，$k \in \mathbb{N}$ 有关，并且所选方向 $\left\{d_y^k\right\}_k$ 和随机过程 $\left\{D_y^k\right\}_k$ 与随机非光滑残差 $R_{v_y^k}^{\lambda_k}\left(Y^k\right)$，$k \in \mathbb{N}$ 相关。在假设(B.1)下，算法的设计意味着 $\left\{\left(X^k, Y^k\right)\right\}_k$ 和 $\left\{\left(X^{k+1}, Y^k\right)\right\}_k$ 适用于滤流 F^k，$\left\{\left(\bar{X}^k, Y^k\right)\right\}_k$ 和 $\left\{\left(X^{k+1}, \bar{Y}^k\right)\right\}_k$ 适用于滤流 F_+^k。即我们有

$\left\{\left(\bar{X}^k, Y^k\right)\right\}_k, \left\{\left(X^{k+1}, \bar{Y}^k\right)\right\}_k \in \mathrm{F}^k$ 并且 $\left\{\left(X^{k+1}, Y^{k+1}\right)\right\}_k, \left\{\left(X^{k+2}, Y^{k+1}\right)\right\}_k \in \mathrm{F}_+^k$， $\forall k \geq 0$。

3.3. 算法的收敛性分析

本节对提出的算法的收敛性展开分析，首先给出 4 个辅助引理以建立目标函数 ψ 的近似下降性，最后在定理 1 中证明了算法的收敛性。下面首先给出引理 1，在该引理中给出了目标函数 f 的近似下降性。

引理 1：假设 1 成立，$\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ 为随机邻近牛顿型交替极小化算法生成的序列，则有

$$\begin{aligned} f\left(x^{k+1}\right) - f\left(x^k\right) \leq & \left\langle \frac{\lambda_x^k}{\alpha_{x,+}^k} d_x^k - v_{x,+}^k, x^{k+1} - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle + \left\langle v_x^k, p^{\alpha_x^k}\left(x^k\right) - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle \\ & + \left\langle \nabla_x H\left(x^k, y^k\right), x^k - p^{\alpha_x^k}\left(x^k\right)\right\rangle + \frac{1}{2\alpha_{x,+}^k}\left\| x^k - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\|^2 \end{aligned} \tag{3.5}$$

证明：由迭代 $x^{k+1} = \mathrm{prox}_{\alpha_{x,+}^k f}\left(x^k + \lambda_x^k d_x^k - \alpha_{x,+}^k v_{x,+}^k\right)$ 的最优性条件可得

$$0 \in \partial f\left(x^{k+1}\right) + v_{x,+}^k + \frac{1}{\alpha_{x,+}^k}\left(x^{k+1} - x^k - \lambda_x^k d_x^k\right),$$

进一步有

$$\frac{1}{\alpha_{x,+}^k}\left(x^k + \lambda_x^k d_x^k - x^{k+1} - \alpha_{x,+}^k v_{x,+}^k\right) \in \partial f\left(x^{k+1}\right). \tag{3.6}$$

利用函数 f 的凸性可知，

$$f\left(x^{k+1}\right) - f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right) \leq \left\langle \partial f\left(x^{k+1}\right), x^{k+1} - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle. \tag{3.7}$$

结合公式(3.5)和(3.6)可得：

$$\begin{aligned} f\left(x^{k+1}\right) - f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right) & \leq \left\langle \frac{\lambda_x^k}{\alpha_{x,+}^k} d_x^k - v_{x,+}^k, x^{k+1} - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle + \frac{1}{\alpha_{x,+}^k}\left\langle x^k - x^{k+1}, x^{k+1} - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle \\ & = \left\langle \frac{\lambda_x^k}{\alpha_{x,+}^k} d_x^k - v_{x,+}^k, x^{k+1} - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle + \frac{1}{2\alpha_{x,+}^k}\left\| x^k - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\|^2 \end{aligned} \tag{3.8}$$

其中等式是由 $\langle a-b, c-d\rangle = \frac{1}{2}\left(\|a-b\|^2 - \|a-c\|^2 + \|b-c\|^2 - \|b-d\|^2\right)$ 得到的。

由 $p_{v_x^k}^{\alpha_x^k}\left(x^k\right) = \mathrm{prox}_{\alpha_x^k f}\left(x^k - \alpha_x^k v_x^k\right)$ 可得，

$$\frac{1}{\alpha_x^k}\left(x^k - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right) - v_x^k \in \partial f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right), \tag{3.9}$$

又由函数 f 的凸性可知

$$f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right) - f\left(p^{\alpha_x^k}\left(x^k\right)\right) \leq \left\langle \partial f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right), p_{v_x^k}^{\alpha_x^k}\left(x^k\right) - p^{\alpha_x^k}\left(x^k\right)\right\rangle. \tag{3.10}$$

因此

$$\begin{aligned} f\left(p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right) - f\left(p^{\alpha_x^k}\left(x^k\right)\right) \leq & \left\langle v_x^k, p^{\alpha_x^k}\left(x^k\right) - p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\rangle + \frac{1}{2\alpha_x^k}\left\| x^k - p^{\alpha_x}\left(x^k\right)\right\|^2 \\ & - \frac{1}{2\alpha_x^k}\left\| p_{v_x^k}^{\alpha_x^k}\left(x^k\right) - x^k\right\|^2 - \frac{1}{2\alpha_x^k}\left\| p_{v_x^k}^{\alpha_x^k}\left(x^k\right) - p^{\alpha_x}\left(x^k\right)\right\|^2. \end{aligned} \tag{3.11}$$

由 $p^{\alpha_x^k}\left(x^k\right)$ 的定义可得

$$\frac{1}{\alpha_x^k}\left(x^k-p^{\alpha_x^k}\left(x^k\right)\right)-\nabla_x H\left(x^k,y^k\right)\in\partial f\left(p^{\alpha_x^k}\left(x^k\right)\right), \tag{3.12}$$

又由函数 f 的凸性可知

$$f\left(p^{\alpha_x^k}\left(x^k\right)\right)-f\left(x^k\right)\le\left\langle\partial f\left(p^{\alpha_x^k}\left(x^k\right)\right),\ p^{\alpha_x^k}\left(x^k\right)-x^k\right\rangle, \tag{3.13}$$

由(3.11)、(3.12)得

$$\begin{aligned}f\left(p^{\alpha_x^k}\left(x^k\right)\right)-f\left(x^k\right)\le&\left\langle\nabla_x H\left(x^k,y^k\right),x^k-p^{\alpha_x^k}\left(x^k\right)\right\rangle+\frac{1}{2\alpha_x^k}\left\|x^k-x^k\right\|^2\\&-\frac{1}{2\alpha_x^k}\left\|p^{\alpha_x^k}\left(x^k\right)-x^k\right\|^2-\frac{1}{2\alpha_x^k}\left\|p^{\alpha_x^k}\left(x^k\right)-x^k\right\|^2.\end{aligned} \tag{3.14}$$

结合公式(3.7)，(3.10)和(3.13)可得(3.14)。

类似的，对于变量 y，有以下结论。

引理 2：假设 1 成立，$\left\{\left(x^k,y^k\right)\right\}_{k\in\mathbb{N}}$ 为随机邻近牛顿型交替极小化算法生成的序列，则有

$$\begin{aligned}g\left(y^{k+1}\right)-g\left(y^k\right)\le&\left\langle\frac{\lambda_y^k}{\alpha_{y,+}^k}d_y^k-v_{y,+}^k,y^{k+1}-p_{v_y^k}^{\alpha_y^k}\left(y^k\right)\right\rangle+\left\langle v_y^k,p^{\alpha_y^k}\left(y^k\right)-p_{v_y^k}^{\alpha_y^k}\left(y^k\right)\right\rangle\\&+\left\langle\nabla_y H\left(x^{k+1},y^k\right),y^k-p^{\alpha_y^k}\left(y^k\right)\right\rangle+\frac{1}{2\alpha_{y,+}^k}\left\|y^k-p_{v_y^k}^{\alpha_y^k}\left(y^k\right)\right\|^2\end{aligned} \tag{3.15}$$

基于引理 1 和引理 2，可以建立 ψ 的近似下降性。首先，由 ψ 定义可得

$$\begin{aligned}&\psi\left(x^{k+1},y^{k+1}\right)-\psi\left(x^k,y^k\right)\\=&f\left(x^{k+1}\right)+g\left(y^{k+1}\right)+H\left(x^{k+1},y^{k+1}\right)-f\left(x^k\right)-g\left(y^k\right)-H\left(x^k,y^k\right)\\=&f\left(x^{k+1}\right)+H\left(x^{k+1},y^k\right)-\left(f\left(x^k\right)+H\left(x^k,y^k\right)\right)\\&+g\left(y^{k+1}\right)+H\left(x^{k+1},y^{k+1}\right)-\left(g\left(y^k\right)+H\left(x^{k+1},y^k\right)\right),\end{aligned} \tag{3.16}$$

下面首先给出 $f\left(x^{k+1}\right)+H\left(x^{k+1},y^k\right)-\left(f\left(x^k\right)+H\left(x^k,y^k\right)\right)$ 的近似下降性。

引理 3：假设 1 成立，$\left\{\left(x^k,y^k\right)\right\}_{k\in N}$ 为随机邻近牛顿型交替极小化算法生成的序列，则有

$$\begin{aligned}&f\left(x^{k+1}\right)+H\left(x^{k+1},y^k\right)-\left(f\left(x^k\right)+H\left(x^k,y^k\right)\right)\\\le&\frac{\alpha_{x,+}^k}{2}\left(L_x\beta_x^k+\frac{\lambda_x^k}{\alpha_{x,+}^k}\right)^2\left\|d_x^k\right\|^2+\frac{\alpha_{x,+}^k}{2}\left\|\nabla_x H\left(\bar{x}^k,y^k\right)-v_{x,+}^k\right\|^2\\&+\left\langle\nabla_x H\left(\bar{x}^k,y^k\right)-v_{x,+}^k,\alpha_{x,+}^k\left(\nabla_x H\left(x^k,y^k\right)-\nabla_x H\left(\bar{x}^k,y^k\right)\right)+\lambda_x^k d_x^k\right\rangle\\&+\frac{\alpha_x^k}{2}\left\|\nabla_x H\left(x^k,y^k\right)-v_x^k\right\|^2+\left(\frac{1}{2\alpha_{x,+}^k}-\frac{1}{2\alpha_x^k}\right)\left\|x^k-p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\|^2\\&+\left(\frac{L_x}{2}-\frac{1}{2\alpha_{x,+}^k}\right)\left\|x^{k+1}-x^{k2}\right\|-\frac{1}{2\alpha_x^k}\left\|x^k-p^{\alpha_x^k}\left(x^k\right)\right\|^2.\end{aligned} \tag{3.17}$$

证明：关于 H 的部分我们使用下降引理可得

$$H\left(x^{k+1},y^{k}\right)\le H\left(x^{k},y^{k}\right)+\left\langle \nabla_{x}H\left(x^{k},y^{k}\right),x^{k+1}-x^{k}\right\rangle+\frac{L_{x}}{2}\left\|x^{k+1}-x^{k}\right\|^{2},\tag{3.18}$$

则由(3.5)和(3.18)可得：

$$\begin{aligned}
&f\left(x^{k+1}\right)+H\left(x^{k+1},y^{k}\right)-\left(f\left(x^{k}\right)+H\left(x^{k},y^{k}\right)\right)\\
\le&\left\langle\frac{\lambda_{x}^{k}}{\alpha_{x,+}^{k}}d_{x}^{k}-v_{x,+}^{k},x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle+\left\langle v_{x}^{k},p^{\alpha_{x}^{k}}\left(x^{k}\right)-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle\\
&+\left\langle\nabla_{x}H\left(x^{k},y^{k}\right),x^{k}-p^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle+\left\langle\nabla_{x}H\left(x^{k},y^{k}\right),x^{k+1}-x^{k}\right\rangle\\
&+\left(\frac{1}{2\alpha_{x,+}^{k}}-\frac{1}{2\alpha_{x}^{k}}\right)\left\|x^{k}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}-\frac{1}{2\alpha_{x,+}^{k}}\left\|x^{k}-x^{k+1}\right\|^{2}-\frac{1}{2\alpha_{x,+}^{k}}\left\|x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}\\
&-\frac{1}{2\alpha_{x}^{k}}\left\|p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)-p^{\alpha_{x}}\left(x^{k}\right)\right\|^{2}-\frac{1}{2\alpha_{x}^{k}}\left\|p^{\alpha_{x}^{k}}\left(x^{k}\right)-x^{k}\right\|^{2}+\frac{Lx}{2}\left\|x^{k+1}-x^{k}\right\|^{2},
\end{aligned}$$

整理可得

$$\begin{aligned}
&f\left(x^{k+1}\right)+H\left(x^{k+1},y^{k}\right)-\left(f\left(x^{k}\right)+H\left(x^{k},y^{k}\right)\right)\\
\le&\left\langle\frac{\lambda_{x}^{k}}{\alpha_{x,+}^{k}}d_{x}^{k}-v_{x,+}^{k},x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle+\left\langle v_{x}^{k},p^{\alpha_{x}^{k}}\left(x^{k}\right)-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle\\
&+\left\langle\nabla_{x}H\left(x^{k},y^{k}\right),x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)+p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)-p^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle\\
&+\left(\frac{1}{2\alpha_{x,+}^{k}}-\frac{1}{2\alpha_{x}^{k}}\right)\left\|x^{k}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}-\frac{1}{2\alpha_{x,+}^{k}}\left\|x^{k}-x^{k+1}\right\|^{2}-\frac{1}{2\alpha_{x,+}^{k}}\left\|x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}\\
&-\frac{1}{2\alpha_{x}^{k}}\left\|p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)-p^{\alpha_{x}}\left(x^{k}\right)\right\|^{2}-\frac{1}{2\alpha_{x}^{k}}\left\|p^{\alpha_{x}^{k}}\left(x^{k}\right)-x^{k}\right\|^{2}+\frac{Lx}{2}\left\|x^{k+1}-x^{k}\right\|^{2}\\
\le&\left\langle\nabla_{x}H\left(x^{k},y^{k}\right)+\frac{\lambda_{x}^{k}}{\alpha_{x,+}^{k}}d_{x}^{k}-v_{x,+}^{k},x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle\\
&+\left\langle\nabla_{x}H\left(x^{k},y^{k}\right)-v_{x}^{k},p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)-p^{\alpha_{x}^{k}}\left(x^{k}\right)\right\rangle+\left(\frac{1}{2\alpha_{x,+}^{k}}-\frac{1}{2\alpha_{x}^{k}}\right)\left\|x^{k}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}\\
&+\left(\frac{L_{x}}{2}-\frac{1}{2\alpha_{x,+}^{k}}\right)\left\|x^{k+1}-x^{k}\right\|^{2}-\frac{1}{2\alpha_{x,+}^{k}}\left\|x^{k+1}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}\\
&-\frac{1}{2\alpha_{x}^{k}}\left\|p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)-p^{\alpha_{x}}\left(x^{k}\right)\right\|^{2}-\frac{1}{2\alpha_{x}^{k}}\left\|p^{\alpha_{x}^{k}}\left(x^{k}\right)-x^{k}\right\|^{2}.
\end{aligned}\tag{3.19}$$

利用公式(3.18)和不等式 $2\langle a,b\rangle\le\alpha_{x}^{k}\|a\|^{2}+\frac{1}{\alpha_{x}^{k}}\|b\|^{2}$ 可得

$$\begin{aligned}
&f\left(x^{k+1}\right)+H\left(x^{k+1},y^{k}\right)-\left(f\left(x^{k}\right)+H\left(x^{k},y^{k}\right)\right)\\
\le&\frac{\alpha_{x,+}^{k}}{2}\left\|\nabla_{x}H\left(x^{k},y^{k}\right)+\frac{\lambda_{x}^{k}}{\alpha_{x,+}^{k}}d_{x}^{k}-v_{x,+}^{k}\right\|^{2}+\frac{\alpha_{x}^{k}}{2}\left\|\nabla_{x}H\left(x^{k},y^{k}\right)-v_{x}^{k}\right\|^{2}\\
&+\left(\frac{1}{2\alpha_{x,+}^{k}}-\frac{1}{2\alpha_{x}^{k}}\right)\left\|x^{k}-p_{v_{x}^{k}}^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}+\left(\frac{L_{x}}{2}-\frac{1}{2\alpha_{x,+}^{k}}\right)\left\|x^{k+1}-x^{k}\right\|^{2}-\frac{1}{2\alpha_{x}^{k}}\left\|x^{k}-p^{\alpha_{x}^{k}}\left(x^{k}\right)\right\|^{2}
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{\alpha_{x,+}^k}{2}\left\|\nabla_x H\left(x^k,y^k\right)-\nabla_x H\left(\bar{x}^k,y^k\right)+\frac{\lambda_x^k}{\alpha_{x,+}^k}d_x^k\right\|^2+\frac{\alpha_{x,+}^k}{2}\left\|\nabla_x H\left(\bar{x}^k,y^k\right)-v_{x,+}^k\right\|^2\\
&+\left\langle\nabla_x H\left(\bar{x}^k,y^k\right)-v_{x,+}^k,\alpha_{x,+}^k\left(\nabla_x H\left(x^k,y^k\right)-\nabla_x H\left(\bar{x}^k,y^k\right)\right)+\lambda_x^k d_x^k\right\rangle\\
&+\frac{\alpha_x^k}{2}\left\|\nabla_x H\left(x^k,y^k\right)-v_x^k\right\|^2+\left(\frac{1}{2\alpha_{x,+}^k}-\frac{1}{2\alpha_x^k}\right)\left\|x^k-p_{v_x^k}^{\alpha_x^k}\left(x^k\right)\right\|^2\\
&+\left(\frac{L_x}{2}-\frac{1}{2\alpha_{x,+}^k}\right)\left\|x^{k+1}-x^k\right\|^2-\frac{1}{2\alpha_x^k}\left\|x^k-p^{\alpha_x^k}\left(x^k\right)\right\|^2,
\end{aligned}
$$

其中第二个不等式利用了 $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$。进一步，根据 $H(x,y)$ 关于变量 x 连续可微且 $\bar{x}^k=x^k+\beta_x^k d_x^k$ 可得(3.17)。

类似地，我们可以建立 $g\left(y^{k+1}\right)+H\left(x^{k+1},y^{k+1}\right)-\left(g\left(y^k\right)+H\left(x^{k+1},y^k\right)\right)$ 的近似下降性。

引理 4：假设 1 成立，$\left\{\left(x^k,y^k\right)\right\}_{k\in N}$ 为随机邻近牛顿型交替极小化算法生成的序列，则有

$$
\begin{aligned}
&g\left(y^{k+1}\right)+H\left(x^{k+1},y^{k+1}\right)-\left(g\left(y^k\right)+H\left(x^{k+1},y^k\right)\right)\\
&\le\frac{\alpha_{y,+}^k}{2}\left(L_y\beta_y^k+\frac{\lambda_y^k}{\alpha_{y,+}^k}\right)^2\left\|d_y^k\right\|^2+\frac{\alpha_{y,+}^k}{2}\left\|\nabla_y H\left(x^{k+1},\bar{y}^k\right)-v_{y,+}^k\right\|^2\\
&+\left\langle\nabla_y H\left(x^{k+1},\bar{y}^k\right)-v_{y,+}^k,\alpha_{y,+}^k\left(\nabla_y H\left(x^{k+1},y^k\right)-\nabla_y H\left(x^{k+1},\bar{y}^k\right)\right)+\lambda_y^k d_y^k\right\rangle\\
&+\frac{\alpha_y^k}{2}\left\|\nabla_y H\left(x^{k+1},y^k\right)-v_y^k\right\|^2+\left(\frac{1}{2\alpha_{y,+}^k}-\frac{1}{2\alpha_y^k}\right)\left\|y^k-p_{v_y^k}^{\alpha_y^k}\left(y^k\right)\right\|^2\\
&+\left(\frac{L_y}{2}-\frac{1}{2\alpha_{y,+}^k}\right)\left\|y^{k+1}-y^k\right\|^2-\frac{1}{2\alpha_y^k}\left\|y^k-p^{\alpha_y^k}\left(y^k\right)\right\|^2.
\end{aligned}
\tag{3.20}
$$

基于引理 1~4，下面给出算法的收敛性分析。

定理 1：令随机过程 $\left\{\left(X^k,Y^k\right)\right\}_k$ 为算法生成的序列。假设 1 和假设 2 成立，且步长 $\left\{\alpha_x^k\right\}_k,\left\{\alpha_{x,+}^k\right\}_k,\left\{\lambda_x^k\right\}_k,\left\{\beta_x^k\right\}_k$ 以及 $\left\{\alpha_y^k\right\}_k,\left\{\alpha_{y,+}^k\right\}_k,\left\{\lambda_y^k\right\}_k,\left\{\beta_y^k\right\}_k$ 满足对于所有的 $k\in\mathbb{N}$ 和某个 $\bar{\rho}\in(0,1)$ 有以下式子成立

$$
\begin{aligned}
&\alpha_{x,+}^k\le\frac{1}{L_x},\ \alpha_x^k\le\frac{(1-\bar{\rho})\alpha_{x,+}^k}{1+\left(v_x^k\right)^2\left(\alpha_x^k+L_x\beta_x^k\alpha_{x,+}^k\right)^2};\\
&\alpha_{y,+}^k\le\frac{1}{L_y},\ \alpha_y^k\le\frac{(1-\bar{\rho})\alpha_{y,+}^k}{1+\left(v_y^k\right)^2\left(\alpha_y^k+L_y\beta_y^k\alpha_{y,+}^k\right)^2}.
\end{aligned}
\tag{3.21}
$$

则若 $\sum\alpha_x^k=\infty$，$\sum\alpha_x^k\left(\sigma_x^k\right)^2<\infty$，$\sum\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2<\infty$，$\sum\alpha_y^k=\infty$，$\sum\alpha_y^k\left(\sigma_y^k\right)^2<\infty$，有 $\liminf_{k\to\infty}E\left[\left\|F^1\left(X^k\right)\right\|^2\right]=0$，$\liminf_{k\to\infty}E\left[\left\|F^1\left(Y^k\right)\right\|^2\right]=0$，$\liminf_{k\to\infty}\left\|F^1\left(X^k\right)\right\|=0$，$\liminf_{k\to\infty}\left\|F^1\left(Y^k\right)\right\|=0$ a.s.，并且 $\lim_{k\to\infty}E\left\|X^{k+1}-X^k\right\|=0$，$\lim_{k\to\infty}E\left\|Y^{k+1}-Y^k\right\|=0$。

证明：由 $\mathrm{F}_+^{k-1}\subset\mathrm{F}^k$ 与 $X^k,D_x^k,\bar{X}^k,Y^k,D_y^k,\bar{Y}^k\in\mathrm{F}^k$ 可得下述结果几乎处处成立

$$
\begin{aligned}
&E\left[\left\langle\nabla_x H\left(\bar{X}^k,Y^k\right)-V_{x,+}^k,\alpha_{x,+}^k\left(\nabla_x H\left(X^k,Y^k\right)-\nabla_x H\left(\bar{X}^k,Y^k\right)\right)+\lambda_x^k D_x^k\right\rangle|\,\mathrm{F}_+^{k-1}\right]\\
&=E\left[\left\langle E\left[\nabla_x H\left(\bar{X}^k,Y^k\right)-V_{x,+}^k|\,\mathrm{F}^k\right],\alpha_{x,+}^k\left(\nabla_x H\left(X^k,Y^k\right)-\nabla_x H\left(\bar{X}^k,Y^k\right)\right)+\lambda_x^k D_x^k\right\rangle|\,\mathrm{F}_+^{k-1}\right]\\
&=0,
\end{aligned}
$$

$$
\begin{aligned}
&E\left[\left\langle \nabla_y H\left(X^{k+1},\bar{Y}^k\right)-V_{y,+}^k,\alpha_{y,+}^k\left(\nabla_y H\left(X^{k+1},Y^k\right)-\nabla_y H\left(X^{k+1},\bar{Y}^k\right)\right)+\lambda_y^k D_y^k\right\rangle \mid \mathrm{F}_+^{k-1}\right]\\
&=E\left[\left\langle E\left[\nabla_y H\left(X^{k+1},\bar{Y}^k\right)-V_{y,+}^k \mid \mathrm{F}^k\right],\alpha_{y,+}^k\left(\nabla_y H\left(X^{k+1},Y^k\right)-\nabla_y H\left(X^{k+1},\bar{Y}^k\right)\right)+\lambda_y^k D_y^k\right\rangle \mid \mathrm{F}_+^{k-1}\right]\\
&=0,
\end{aligned}
$$

其中我们用到了 $V_{x,+}^k$ 与 $V_{y,+}^k$ 分别是 $\nabla_x H\left(\bar{X}^k,Y^k\right)$ 与 $\nabla_y H\left(X^{k+1},\bar{Y}^k\right)$ 的无偏估计。基于此，将公式(3.17)与公式(3.20)相加，求条件期望，并利用假设 2 可得以下结果几乎处处成立：

$$
\begin{aligned}
&E\left[\psi\left(X^{k+1},Y^{k+1}\right)\mid \mathrm{F}_+^{k-1}-\psi\left(X^k,Y^k\right)\right]\\
&\le \frac{\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2}{2}+\frac{1}{2}\left(L_x-\frac{1}{\alpha_{x,+}^k}\right)E\left[\left\|X^{k+1}-X^k\right\|^2\mid \mathrm{F}_+^{k-1}\right]-\frac{1}{2\alpha_x^k}\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2\\
&+\frac{\alpha_x^k\left(\sigma_x^k\right)^2}{2}+\frac{1}{2}\left[\frac{1}{\alpha_{x,+}^k}-\frac{1}{\alpha_x^k}+\left(v_x^k\right)^2\alpha_{x,+}^k\left(L_x\beta_x^k+\frac{\lambda_x^k}{\alpha_{x,+}^k}\right)^2\right]E\left[\left\|F_{V_x^k}^{\alpha_x^k}\left(X^k\right)\right\|^2\mid \mathrm{F}_+^{k-1}\right] \qquad (3.23)\\
&+\frac{\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2}{2}+\frac{1}{2}\left(L_y-\frac{1}{\alpha_{y,+}^k}\right)E\left[\left\|Y^{k+1}-Y^k\right\|^2\mid \Phi_+^{k-1}\right]-\frac{1}{2\alpha_y^k}\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2\\
&+\frac{\alpha_y^k\left(\sigma_y^k\right)^2}{2}+\frac{1}{2}\left[\frac{1}{\alpha_{y,+}^k}-\frac{1}{\alpha_y^k}+\left(v_y^k\right)^2\alpha_{y,+}^k\left(L_y\beta_y^k+\frac{\lambda_y^k}{\alpha_{y,+}^k}\right)^2\right]E\left[\left\|F_{V_y^k}^{\alpha_y^k}\left(Y^k\right)\right\|^2\mid \Phi_+^{k-1}\right].
\end{aligned}
$$

结合公式(3.21)与(3.23)可得

$$
\begin{aligned}
&E\left[\psi\left(X^{k+1},Y^{k+1}\right)\mid \mathrm{F}_+^{k-1}\right]-\psi\left(X^k,Y^k\right)\\
&\le \frac{\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2}{2}+\frac{\alpha_x^k\left(\sigma_x^k\right)^2}{2}-\frac{1}{2\alpha_x^k}\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2-\frac{\bar{\rho}}{2\alpha_x^k}E\left[\left\|F_{V_x^k}^{\alpha_x^k}\left(X^k\right)\right\|^2\mid \mathrm{F}_+^{k-1}\right]\\
&+\frac{\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2}{2}+\frac{\alpha_y^k\left(\sigma_y^k\right)^2}{2}-\frac{1}{2\alpha_y^k}\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2-\frac{\bar{\rho}}{2\alpha_y^k}E\left[\left\|F_{V_y^k}^{\alpha_y^k}\left(Y^k\right)\right\|^2\mid \mathrm{F}_+^{k-1}\right].
\end{aligned}
$$

由假设 1 可知存在 $\psi^*\in\mathbb{R}$ 满足 $\psi(x,y)\ge\psi^*$ 对于所有的 $(x,y)\in\mathbb{R}^n\times\mathbb{R}^n$。对上式左右两边求全期望并移项可得

$$
\begin{aligned}
&\frac{E\left[\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2+\bar{\rho}\left\|F_{V_x^k}^{\alpha_x^k}\left(X^k\right)\right\|^2\right]}{\alpha_x^k}+\frac{E\left[\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2+\bar{\rho}\left\|F_{V_y^k}^{\alpha_y^k}\left(Y^k\right)\right\|^2\right]}{\alpha_y^k}\\
&\le 2\left[\psi\left(x^k,y^k\right)-\psi\left(x^{k+1},y^{k+1}\right)\right]+\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2+\alpha_x^k\left(\sigma_x^k\right)^2+\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2+\alpha_y^k\left(\sigma_y^k\right)^2.
\end{aligned}
$$

对上式左右两边同时累和可得对于所有的 $K\in\mathbb{N}$ 有

$$
\begin{aligned}
&\sum_{k=0}^{K}\frac{E\left[\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2+\bar{\rho}\left\|F_{V_x^k}^{\alpha_x^k}\left(X^k\right)\right\|^2\right]}{\alpha_x^k}+\frac{E\left[\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2+\bar{\rho}\left\|F_{V_y^k}^{\alpha_y^k}\left(Y^k\right)\right\|^2\right]}{\alpha_y^k}\\
&\le 2\left[\psi\left(x^0,y^0\right)-\psi^*\right]+\sum_{k=0}^{K}\left[\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2+\alpha_x^k\left(\sigma_x^k\right)^2+\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2+\alpha_y^k\left(\sigma_y^k\right)^2\right].
\end{aligned}
$$

因为 $L_x,L_y\ge 1$，则 $\alpha_{x,+}^k\le 1,\alpha_x^k\le 1$ 并且 $\alpha_{y,+}^k\le 1,\alpha_y^k\le 1$。又映射 $\delta\mapsto\delta^{-1}\left\|F^{\frac{1}{\delta}}(x)\right\|$ 是一个关于 δ 的递减函

数[20]，因此，对于所有的 $k\in\mathbb{N}$，有

$$E\left[\left\|F^1\left(X^k\right)\right\|^2\right]\le\left(\alpha_x^k\right)^{-2}E\left[\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2\right],\ E\left[\left\|F^1\left(Y^k\right)\right\|^2\right]\le\left(\alpha_y^k\right)^{-2}E\left[\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2\right],$$

进而我们可以得到 $\sum\alpha_x^kE\left[\left\|F^1\left(X^k\right)\right\|^2\right]<\infty$，$\sum\alpha_y^kE\left[\left\|F^1\left(Y^k\right)\right\|^2\right]<\infty$。又因为 $\sum\alpha_x^k=\infty$，$\sum\alpha_y^k=\infty$，因此 $\liminf_{k\to\infty}E\left[\left\|F^1\left(X^k\right)\right\|^2\right]=0$，$\liminf_{k\to\infty}E\left[\left\|F^1\left(Y^k\right)\right\|^2\right]=0$。又根据 Borel-Cantelli 引理[21]可知：当 $\sum\alpha_x^kE\left[\left\|F^1\left(X^k\right)\right\|^2\right]<\infty$ 与 $\sum\alpha_y^kE\left[\left\|F^1\left(Y^k\right)\right\|^2\right]<\infty$ 几乎处处成立时，可以推出 $\liminf_{k\to\infty}\left\|F^1\left(X^k\right)\right\|=0$，$\liminf_{k\to\infty}\left\|F^1\left(Y^k\right)\right\|=0$ 必定成立。

由公式(3.21)和(3.23)可得

$$\begin{aligned}
&\frac{1}{2}\left(\frac{1}{\alpha_{x,+}^k}-L_x\right)E\left[\left\|X^{k+1}-X^k\right\|^2\mid\mathrm{F}_+^{k-1}\right]+\frac{1}{2}\left(\frac{1}{\alpha_{y,+}^k}-L_y\right)E\left[\left\|Y^{k+1}-Y^k\right\|^2\mid\mathrm{F}_+^{k-1}\right]\\
&\le\psi\left(X^k,Y^k\right)-E\left[\psi\left(X^{k+1},Y^{k+1}\right)\mid\mathrm{F}_+^{k-1}\right]\\
&\quad+\frac{\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2}{2}+\frac{\alpha_x^k\left(\sigma_x^k\right)^2}{2}-\frac{1}{2\alpha_x^k}\left\|F^{\alpha_x^k}\left(X^k\right)\right\|^2-\frac{\bar{\rho}}{2\alpha_x^k}E\left[\left\|F_{V_x^k}^{\alpha_x^k}\left(X^k\right)\right\|^2\mid\mathrm{F}_+^{k-1}\right]\\
&\quad+\frac{\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2}{2}+\frac{\alpha_y^k\left(\sigma_y^k\right)^2}{2}-\frac{1}{2\alpha_y^k}\left\|F^{\alpha_y^k}\left(Y^k\right)\right\|^2-\frac{\bar{\rho}}{2\alpha_y^k}E\left[\left\|F_{V_y^k}^{\alpha_y^k}\left(Y^k\right)\right\|^2\mid\Phi_+^{k-1}\right]\\
&\le\psi\left(X^k,Y^k\right)-E\left[\psi\left(X^{k+1},Y^{k+1}\right)\mid\Phi_+^{k-1}\right]\\
&\quad+\frac{\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2}{2}+\frac{\alpha_x^k\left(\sigma_x^k\right)^2}{2}+\frac{\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2}{2}+\frac{\alpha_y^k\left(\sigma_y^k\right)^2}{2}.
\end{aligned}$$

记 $M=\min\left\{\frac{1}{2}\left(\frac{1}{\alpha_{x,+}^k}-L_x\right),\frac{1}{2}\left(\frac{1}{\alpha_{y,+}^k}-L_y\right)\right\}$，由上式可得

$$\begin{aligned}
&E\left[\left\|X^{k+1}-X^k\right\|^2\mid\mathrm{F}_+^{k-1}\right]+E\left[\left\|Y^{k+1}-Y^k\right\|^2\mid\mathrm{F}_+^{k-1}\right]\\
&\le\frac{1}{M}\left[\psi\left(X^k,Y^k\right)-E\left[\psi\left(X^{k+1},Y^{k+1}\right)\mid\mathrm{F}_+^{k-1}\right]\right]\\
&\quad+\frac{1}{2M}\left[\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2+\alpha_x^k\left(\sigma_x^k\right)^2+\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2+\alpha_y^k\left(\sigma_y^k\right)^2\right].
\end{aligned}$$

对上式左右两边求全期望并累和可得

$$\begin{aligned}
&\sum_{k=0}^{\infty}E\left\|X^{k+1}-X^k\right\|^2+\sum_{k=0}^{\infty}E\left\|Y^{k+1}-Y^k\right\|^2\\
&\le\frac{1}{M}\left[\psi\left(X^0,Y^0\right)-\psi^*\right]\\
&\quad+\frac{1}{2M}\sum_{k=0}^{\infty}\left[\alpha_{x,+}^k\left(\sigma_{x,+}^k\right)^2+\alpha_x^k\left(\sigma_x^k\right)^2+\alpha_{y,+}^k\left(\sigma_{y,+}^k\right)^2+\alpha_y^k\left(\sigma_y^k\right)^2\right]\\
&<\infty,
\end{aligned}$$

由此我们可以得到 $\lim_{k\to\infty}E\left\|X^{k+1}-X^k\right\|^2=0$，$\lim_{k\to\infty}E\left\|Y^{k+1}-Y^k\right\|^2=0$。进一步可以得到 $\lim_{k\to\infty}E\left\|X^{k+1}-X^k\right\|=0$，$\lim_{k\to\infty}E\left\|Y^{k+1}-Y^k\right\|=0$。

4. 总结

本文提出一种随机邻近牛顿型交替极小化算法用以求解大规模有限和形式的非凸非光滑复合优化问题，相较于邻近交替线性极小化算法 PALM，本文采用的随机梯度算法可大大降低计算成本。此外，本文通过残差方程引入目标函数的二阶信息以提升算法的收敛速度。在目标函数非凸的情形下，本文给出了算法的收敛性分析，建立了随机邻近牛顿型交替极小化算法在期望意义下的收敛性。

参考文献

[1] Kawasumi, R. and Takeda, K. (2023) Automatic Hyperparameter Tuning in Sparse Matrix Factorization. *Neural Computation*, **35**, 1086-1099. https://doi.org/10.1162/neco_a_01581

[2] d'Aspremont, A., El Ghaoui, L., Jordan, M.I. and Lanckriet, G.R.G. (2007) A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, **49**, 434-448. https://doi.org/10.1137/050645506

[3] Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286. https://doi.org/10.1198/106186006x113430

[4] Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011) Robust Principal Component Analysis? *Journal of the ACM*, **58**, 1-37. https://doi.org/10.1145/1970392.1970395

[5] Aravkin, A. and Davis, D. (2020) Trimmed Statistical Estimation via Variance Reduction. *Mathematics of Operations Research*, **45**, 292-322. https://doi.org/10.1287/moor.2019.0992

[6] Patrizio, C. and Karen, E. (2017) Blind Image Deconvolution: Theory and Applications. CRC Press, 12-21.

[7] Wang, Q. and Han, D. (2024) Stochastic Gauss-Seidel Type Inertial Proximal Alternating Linearized Minimization and Its Application to Proximal Neural Networks. *Mathematical Methods of Operations Research*, **99**, 39-74. https://doi.org/10.1007/s00186-024-00851-6

[8] 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

[9] 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

[10] Nguyen, L.M., Scheinberg, K. and Takáč, M. (2020) Inexact SARAH Algorithm for Stochastic Optimization. *Optimization Methods and Software*, **36**, 237-258. https://doi.org/10.1080/10556788.2020.1818081

[11] Pan, H. and Zheng, L. (2022) N-SVRG: Stochastic Variance Reduction Gradient with Noise Reduction Ability for Small Batch Samples. *Computer Modeling in Engineering & Sciences*, **131**, 493-512. https://doi.org/10.32604/cmes.2022.019069

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

[13] Wang, Z. and Wen, B. (2022) Proximal Stochastic Recursive Momentum Algorithm for Nonsmooth Nonconvex Optimization Problems. *Optimization*, **73**, 481-495. https://doi.org/10.1080/02331934.2022.2112191

[14] Huang, F., Gu, B., Huo, Z., Chen, S. and Huang, H. (2019) Faster Gradient-Free Proximal Stochastic Methods for Nonconvex Nonsmooth Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 1503-1510. https://doi.org/10.1609/aaai.v33i01.33011503

[15] Pham, N.H., *et al*. (2020) ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization. *Journal of Machine Learning Research*, **21**, 1-48.

[16] Xu, Y. and Yin, W. (2015) Block Stochastic Gradient Iteration for Convex and Nonconvex Optimization. *SIAM Journal on Optimization*, **25**, 1686-1716. https://doi.org/10.1137/140983938

[17] Driggs, D., Tang, J., Liang, J., *et al*. (2020) SPRING: A Fast Stochastic Proximal Alternating Method for Non-Smooth Non-Convex Optimization.

[18] Yang, M., Milzarek, A., Wen, Z. and Zhang, T. (2021) A Stochastic Extra-Step Quasi-Newton Method for Nonsmooth Nonconvex Optimization. *Mathematical Programming*, **194**, 257-303. https://doi.org/10.1007/s10107-021-01629-y

[19] Bollapragada, R., Byrd, R.H. and Nocedal, J. (2018) Exact and Inexact Subsampled Newton Methods for Optimization. *IMA Journal of Numerical Analysis*, **39**, 545-578. https://doi.org/10.1093/imanum/dry009

[20] Nesterov, Y. (2012) Gradient Methods for Minimizing Composite Functions. *Mathematical Programming*, **140**, 125-161. https://doi.org/10.1007/s10107-012-0629-5

[21] Durrett, R. (2019) Probability: Theory and Examples. Cambridge University Press. https://doi.org/10.1017/9781108591034