**Pure Mathematics** 理论数学, 2024, 14(5), 654-663
Published Online May 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.145218

Hans汉斯

# 一种非精确邻近梯度算法

**辜随佳，王湘美**

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2024年4月24日；录用日期：2024年5月22日；发布日期：2024年5月31日

## 摘　要

**邻近点算法(PPA)是求解非光滑优化问题的一种有效的迭代算法，对特殊结构问题的求解非常高效，但在实际问题中求解大规模可分离问题时花费很大。为解决上述问题且同时又保持PPA算法的优点，本文给出了一种非精确邻近梯度算法。该算法结合了线搜索法与邻近梯度下降算法的思想，在子问题的求解过程中采用近似的梯度，且不需要Lipschitz常数已知。基于以上思想，首先我们给出算法的伪代码，然后建立了算法收敛性的充分条件，最后证明在该条件下，算法迭代所产生序列的每个极限点是原问题的临界点。**

## 关键词

**邻近点算法，线搜索，收敛性分析，非精确梯度**

# An Inexact Proximal Gradient Algorithm

**Suijia Gu, Xiangmei Wang**

Faculty of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Apr. 24th, 2024; accepted: May 22nd, 2024; published: May 31st, 2024

## Abstract

**The Proximity Algorithm (PPA) is an effective iterative algorithm for solving non-smooth optimization problems, which is very efficient in solving special structural problems, but it is expensive to solve large-scale separable problems in practical problems. In order to solve the above problems and maintain the advantages of PPA algorithm, an inexact proximity gradient algorithm is proposed. The algorithm combines the ideas of the line search method and the proximity gradient descent algorithm, and adopts the approximate gradient in the solution of the sub-problem, and does not need the Lipschitz constant to be known. Based on the above ideas, firstly, we give the pseudocode of the algorithm, then establish the sufficient conditions for the convergence of the algorithm, and finally prove that under this condition, each limit point of the sequence generated by the algorithm iteration is the critical point of the original problem.**

文章引用: 辜随佳, 王湘美. 一种非精确邻近梯度算法[J]. 理论数学, 2024, 14(5): 654-663.
DOI: 10.12677/pm.2024.145218

**Keywords**

**Proximity Algorithm, Line Search, Convergence Analysis, Inexact Gradient**



## 1. 引言

本文考虑如下优化问题

$$\min_{x\in R^m} F(x), \tag{1}$$

其中 $F: R^m \to R\cup\{+\infty\}$ 是定义在欧氏空间 $R^m$ 上的下半连续凸函数，求解问题(1)的算法有很多，当问题(1)为一般优化问题且目标函数 $F$ 光滑时，可利用梯度法进行求解。当目标函数 $F$ 不具有光滑性时，可采用次梯度算法、邻近点算法等进行求解。当目标函数 $F$ 为大规模可分离优化问题时，即 $F(x):=f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x)$，其中 $n\geq 1$，其中每个 $f_i: R^m \to R\cup\{+\infty\}$ 可都是下半连续的凸函数且可微，这类优化问题可用一阶/二阶随机优化算法求解，如 Robbins 等人(1951)提出了随机梯度算法(SGD) [1]与基于 SGD 的随机方差减小的梯度下降算法[2]。二阶算法由于使用了更多信息，通常可得到更快的收敛速度，常用的算法有子样本牛顿算法(Newsamp) [3]，近似估计 Heesian 矩阵的 Lissa 算法[4]，对梯度与 Heesian 矩阵进行完全随机抽样的 SSN 算法[5]。此外，当函数包含不可微项时，即 $F(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x)+h(x)$，其中 $h: R^m \to R\cup\{+\infty\}$ 是定义在欧氏空间 $R^m$ 上的下半连续凸函数，这类问题的应用也较为广泛，如图像恢复[6]、信号处理[7]、机器学习[8]等。本文将在邻近梯度算法的基础上引入线搜索思想对该问题做系列分析，该算法的迭代格式如下

$$J_k := \operatorname{prox}_h\left(x^k-\nabla f\left(x^k\right)\right) := \arg\min_{y\in R^m} \nabla f\left(x^k\right)\left(y-x^k\right)+h(y)+\frac{1}{2}\left\|y-x^k\right\|^2,$$

$$x^{k+1}=x^k-\beta_k\left(x^k-J_k\right).$$

其中 $\beta_k>0$，称 $\operatorname{prox}_h$ 为邻近算子。

邻近点算法的思想可追溯到[9] [10]，1970 年，Martinet 最早提出了邻近点算法，1976 年，Rockafellar 在文[11] [12]利用它来求解极大单调算子的零点问题，1991 年，Guler 又进一步讨论了正常下半连续极小化问题的邻近点算法及其收敛性质。后 Beck 和 Teboulle [7]将邻近点算法与梯度下降算法相结合，这也就是邻近梯度算法，这类算法主要用于解决确定的复合优化问题，并且在目标函数凸或非凸时，分析了算法的收敛性和收敛速度。由于上述算法的迭代过程依赖可微函数的 Lipschitz 常数，因此 Bello 在文[13]中引入了线搜索的思想，提出了带线搜索的邻近梯度算法，该算法不仅保留了邻近梯度算法的优点，还不依赖于 Lipschitz 常数。在带有大规模可分离优化问题时，作者在文[14]中利用随机逼近(SA)方法解决优化问题(1)，并分析了随机梯度下降算法的收敛性。

基于上述这些论文的启发，本文将提出求解大规模复合凸优化问题的带线搜索的非精确邻近梯度算法，在算法的迭代过程中，对目标函数中的可分离部分采用非精确梯度，非精确邻近梯度算法不仅保留了邻近梯度算法易计算的优点，还在迭代过程中减小了数据储存量，提高了算法的效率。

## 2. 预备知识

本节将介绍一些必要的预备知识，包括一些符号、定义以及算法的分析证明过程中所要用到的概念、引理。在文中，$R, R_{++}$ 分别表示实数集，正实数集。用小写字母表示欧式空间 $R^m$ 中的向量，例如，$v \in R^m$。空间 $R^m$ 中的内积和 2-范数分别用 $\langle \cdot, \cdot \rangle$ 和 $\|\cdot\|$ 表示。

设 $f: R^m \to R \cup \{+\infty\}$ 为定义在 $R^m$ 上的广义实值函数，其定义域为：

$$\operatorname{dom} f := \left\{ x \in R^m : f(x) < +\infty \right\},$$

若 $\operatorname{dom} f \neq \varnothing$，则称函数 $f$ 是真的。函数 $f$ 的上图定义为

$$\operatorname{epi} f := \left\{ (x, r) \in R^{m+1} : f(x) \leq r \right\}.$$

如果 $\operatorname{epi} f$ 在 $R^{m+1}$ 上是闭的，则称 $f$ 是下半连续的。假设 $f$ 为下半连续真凸函数，则函数 $f$ 在点 $x$ 处沿方向 $d \in R^m$ 的方向导数的定义为

$$f'(x; d) := \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t}.$$

$f$ 在 $x$ 处的次微分定义为

$$\partial f(x) := \left\{ v \in R^m : \langle v, y - x \rangle \leq f(y) - f(x) \ \forall y \in \operatorname{dom} f \right\}. \tag{2}$$

如果 $x \notin \operatorname{dom} f$，则 $\partial f(x) = \varnothing$。

**定义 1** 定义邻近算子 $\operatorname{prox}_h : R^m \to \operatorname{dom} h$ 如下：

$$\operatorname{prox}_h(z) = (Id + \partial h)^{-1}(z), \ z \in R^m,$$

即 $\operatorname{prox}_h(z) = \left\{ y \in R^m : y + \partial h(y) = z \right\}$。通过变形可得到其满足

$$z - \operatorname{prox}_h(z) \in \partial h\left(\operatorname{prox}_h(z)\right) \ \forall z \in R^m. \tag{3}$$

**引理 1** ([15], Proposition 17.2)设 $h: R^m \to R \cup \{+\infty\}$ 为一个下半连续真凸函数，则对任意 $x \in \operatorname{dom} h$ 与 $y \in R^m$ 满足

(1) $h'(x; y)$ 存在且 $h'(x; y) = \inf_{t \in R_{++}} \left( \frac{h(x + ty) - h(x)}{t} \right)$。

(2) $h'(x; y - x) + h(x) \leq h(y)$。

**引理 2** 设 $f, h: R^m \to R \cup \{+\infty\}$ 为下半连续真凸函数，则对任意的 $x \in \operatorname{dom} f$ 与 $y \in R^m$，下列结论成立：

(1) $\partial f(x)$ 为非空有界闭凸集。

(2) 若函数 $f$ 在点 $x$ 处可微，则有 $\partial f(x) = \{\nabla f(x)\}$，其中 $\nabla f$ 表示函数 $f$ 在点 $x$ 的梯度。

(3) 设 $x \in \operatorname{dom} f \cap \operatorname{dom} h$，则有 $\partial f(x) + \partial h(x) \subseteq \partial(f + h)(x)$。若 $x \in \operatorname{int}(\operatorname{dom} f) \cap \operatorname{dom} h$，则等号成立。

对于凸函数 $f$，如果 $0 \in \partial f(x_*)$，则 $x_*$ 是 $f$ 函数值的最小点，即 $f(x_*) = \min_{x \in R^m} f(x)$ 下面定义优化问题的 $\varepsilon$-近似解。

**定义 2** ($\varepsilon$-近似解)设 $f$ 为下半连续真凸函数，$\varepsilon > 0$。我们称点 $x_* \in \operatorname{dom} f$ 是优化问题 $\min_{x \in R^m} f(x)$ 的 $\varepsilon$-近似解，如果满足 $d_{\partial f(x_*)}(0) := \min_{v \in \partial f(x_*)} \|v\| \leq \varepsilon$。

对问题(1)的大规模可分离凸优化形式：

$$\min_{x\in R^m}\ \frac{1}{n}\sum_{i=1}^{n} f_i(x)+h(x).$$

我们总假设上述问题有解，解集记为 $S^*$，记 $\psi:=\{1,2,\cdots,n\}$ 为指标集，则 $f:=\frac{1}{n}\sum_{i\in\psi} f_i$。且满足

**假设(H)**

(C1) $h, f_i\ (\forall i\in\psi)$ 为下半连续真凸函数且 $\mathrm{dom}h\subseteq \mathrm{int}\left(\bigcap_{i\in\psi}\mathrm{dom}\, f\right)$。

(C2)函数 $f_i$ 在包含 $\mathrm{dom}h$ 的开集上是连续可微的。

**注：**由假设(H)和引理 2 (3)，对任意 $x\in\mathrm{dom}h$，有 $\partial(f+h)(x)$ 为有界闭凸集且 $\partial(f+h)(x)=\nabla f(x)+\partial h(x)$。

在下面的非精确邻近梯度算法中，我们将用近似梯度 $g$ 代替全梯度 $\nabla f$。下面考虑用随机抽样产生近似梯度：

$$g(x):=\frac{1}{|\xi|}\sum_{j\in\xi}\nabla f_j(x)\quad x\in\mathrm{dom}h,$$

其中 $\xi$ 和 $|\xi|$ 分别表示抽样样本和样本个数。以下引理说明，在抽样的样本数充分大时，$g$ 可在一定概率意义下达到预设的与全梯度的近似程度。

**引理 3** [16] 设 $\varepsilon_1,\lambda\in(0,1)$ 假设存在函数 $Q:R^n\to R$ 满足对任意 $i\in\psi$，有

$$\left\|\nabla f_i(x)\right\|\le Q(x)\quad \forall x\in R^n.$$

假设对指标集 $\psi$ 进行放回或者不放回的随机等可能抽样，且样本个数 $|\xi|$ 满足 $|\xi|\ge Q(x)^2\left(1+\sqrt{8\ln(1/\lambda)}\right)^2\Big/\varepsilon_1^2$，则有

$$\Pr\left(\left\|\nabla f(x)-g(x)\right\|\le\varepsilon_1\right)\ge 1-\lambda.$$

**定义 3** 设 $S$ 为 $R^m$ 的一个非空子集，序列 $\{x^k\}\subset R^m$。若存在一个数列 $\{\omega_k\}$ 满足 $\sum_{k=0}^{\infty}\omega_k<\infty$ 使对任意以 $x\in S$ 下不等式成立：

$$\text{得}\left\|x^{k+1}-x\right\|^2\le\left\|x^k-x\right\|^2+\omega_k\quad k\in N,$$

则称序列 $\{x^k\}$ quasi-Fejér 收敛到 $S$。

**引理 4** ([17], Theorem 4.1)若 $\{x^k\}$ quasi-Fejér 收敛到 $S$，则下列结论成立：

(1) 序列 $\{x^k\}$ 是有界的。

(2) 若序列 $\{x^k\}$ 有极限点 $x_*\in S$，则有 $\lim_{k\to\infty}x^k=x_*$。

**引理 5** 设 $\{a_k\},\{b_k\},\{c_k\}\subset[0,+\infty)$ 为非负序列，其中 $A=\sum_{k=0}^{\infty}a_k<+\infty$，$B=\sum_{k=0}^{\infty}b_k<+\infty$，且 $c_0>0$。如果存在常数 $\beta\le 1$，使得

$$c_{k+1}^2-c_k^2\le a_k+\beta c_k b_k\quad \forall k\in N,\tag{4}$$

则序列 $\{c_k\}$ 有界。

**证明：**注意到 $c_0>0$，所以存在常数 $M\ge 1$，使即 $c_0^2M^2+2c_0\beta BM+2A+c_0^2\ge 0$，即

$$c_0^2+A+c_0\beta BM\le\left(Mc_0\right)^2.\tag{5}$$

下面我们证明

$$c_k\le Mc_0\quad \forall k\in N.\tag{6}$$

显然当 $k=0$ 时，(6)成立。假设当 $k\le m$ 时(6)成立，即

$$c_k \le Mc_0 \quad \forall k \le m.$$

由(4)，我们有

$$\sum_{k=0}^{m}\left(c_{k+1}^2 - c_k^2\right) \le \sum_{k=0}^{m} a_k + \sum_{k=0}^{m} \beta c_k b_k \le A + \beta c_0 MB.$$

于是

$$c_{m+1}^2 \le c_0^2 + A + \beta c_0 MB \le \left(Mc_0\right)^2.$$

由(5)可得 $c_m \le Mc_0$。由归纳法证明可得(6)。

## 3. 算法及收敛性分析

在对问题(1)中的大规模可分离部分进行求解时，为了保留邻近梯度算法的优点且不依赖 Lipschitz 常数，受[13] [14] [16]的启发，我们提出带线搜索的非精确邻近梯度算法。具体地，与[14]类似，我们在子问题求解过程中也采用的是非精确梯度，其中对梯度近似求解的思路参考了[16]，线搜索的思想主要来自([13], Lemma 3.2)。

### 3.1. 非精确邻近梯度算法

**Algorithm 1.** Inexact proximity gradient algorithm
**算法 1.** 非精确邻近梯度算法

Step 1. 给定 $\theta \in (0,1), x^0 \in \mathrm{dom}h, k = 0$。

Step 2. 令 $\beta = 1$，计算近似梯度 $g_k\left(x^k\right)$ 和

$$J_k := J\left(x^k, 1\right) = \mathrm{prox}_h\left(x^k - g_k\left(x^k\right)\right). \tag{7}$$

Step 3. 当

$$\begin{aligned}&\left(f+h\right)\left(x^k - \beta_k\left(x^k - J_k\right)\right)\\ &> \left(f+h\right)\left(x^k\right) - \beta_k\left[h\left(x^k\right) - h\left(J_k\right)\right] - \beta_k\left\langle g_k\left(x^k\right), x^k - J_k\right\rangle + \frac{\beta_k}{2}\left\|x^k - J_k\right\|^2,\end{aligned} \tag{8}$$

令 $\beta_k = \theta\beta_k$；否则，输出 $\beta_k$。

Step 4.令

$$x^{k+1} = x^k - \beta_k\left(x^k - J_k\right). \tag{9}$$

Step 5.若 $\left\|x^{k+1} - x^k\right\| \le \varepsilon_0$，输出 $x^{k+1}$；否则，令 $k := k+1$，返回 Step 2。

由算法 1 可知，若在迭代过程的第 2 步采用全梯度，则算法 1 为精确的线搜索邻近梯度算法，即与([13], Method 3)相同。下面的各类命题与定理中，都假设序列 $\left\{x^k\right\}$ 由算法 1 所产生。

**命题 1** 设 $k \in N$，$\beta_k$ 为算法 1 第 $k$ 次迭代的 Step 3 中生成的步长，则对任意 $x \in \mathrm{dom}h$，有：

$$\begin{aligned}\left\|x^{k+1} - x\right\|^2 - \left\|x^k - x\right\|^2 \le{}& 2\left[\left(f+h\right)\left(x^k\right) - \left(f+h\right)\left(x^{k+1}\right)\right] + 2\beta_k\left[\left(f+h\right)(x) - \left(f+h\right)\left(x^k\right)\right]\\ &+ 2\beta_k\left\|x - x^k\right\|\left\|\nabla f\left(x^k\right) - g_k\left(x^k\right)\right\|.\end{aligned} \tag{10}$$

**证明：**对任意 $x \in \mathrm{dom}h$，设

$$A_k := \left\|x^{k+1} - x^k\right\|^2 + \left\|x^k - x\right\|^2 - \left\|x^{k+1} - x\right\|^2 = 2\left\langle x^k - x^{k+1}, x^k - x\right\rangle.$$

由(9)可得

$$\begin{aligned}\frac{A_k}{2\beta_k}&=\frac{2}{2\beta_k}\left\langle x^k-x^{k+1},x^k-x\right\rangle=\left\langle x^k-J_k,x^k-x\right\rangle\\&=\left\langle x^k-J_k-g_k\left(x^k\right),x^k-x\right\rangle+\left\langle g_k\left(x^k\right),x^k-x\right\rangle\\&=\left\langle x^k-J_k-g_k\left(x^k\right),J_k-x\right\rangle+\left\langle x^k-J_k-g_k\left(x^k\right),x^k-J_k\right\rangle+\left\langle g_k\left(x^k\right),x^k-x\right\rangle\\&=\left\langle x^k-J_k-g_k\left(x^k\right),J_k-x\right\rangle+\left\|x^k-J_k\right\|^2-\left\langle g_k\left(x^k\right),x^k-J_k\right\rangle+\left\langle g_k\left(x^k\right),x^k-x\right\rangle.\end{aligned}$$

由(3)与(7)可得： $x^k-J_k-g_k\left(x^k\right)\in\partial h\left(J_k\right)$

又 $f$ 是光滑的凸函数，所以

$$\begin{aligned}f(x)-f(y)&\geq\left\langle\nabla f(y),x-y\right\rangle\\&=\left\langle\nabla f(y)-g(y),x-y\right\rangle+\left\langle g(y),x-y\right\rangle\\&\geq-\|x-y\|\left\|\nabla f(y)-g(y)\right\|+\left\langle g(y),x-y\right\rangle.\end{aligned}$$

结合算法的第 3 步，并令 $y:=x^k$ ，得

$$\begin{aligned}\frac{A_k}{2\beta_k}&\geq h\left(J_k\right)-h(x)+\frac{1}{\beta_k}\left[\left(f+h\right)\left(x^{k+1}\right)-\left(f+h\right)\left(x^k\right)\right]+h\left(x^k\right)-h\left(J_k\right)+\frac{1}{2}\left\|x^k-J_k\right\|^2\\&\quad+f\left(x^k\right)-f(x)-\left\|x-x^k\right\|\left\|\nabla f\left(x^k\right)-g_k\left(x^k\right)\right\|\\&=\left[\left(f+h\right)\left(x^k\right)-\left(f+h\right)(x)\right]+\frac{1}{\beta_k}\left[\left(f+h\right)\left(x^{k+1}\right)-\left(f+h\right)\left(x^k\right)\right]+\frac{1}{2}\left\|x^k-J_k\right\|^2\\&\quad-\left\|x-x^k\right\|\left\|\nabla f\left(x^k\right)-g_k\left(x^k\right)\right\|.\end{aligned}$$

于是

$$\begin{aligned}\left\|x^{k+1}-x\right\|^2-\left\|x^k-x\right\|^2\leq\left\|x^{k+1}-x^k\right\|^2+2\beta_k\left[\left(f+h\right)(x)-\left(f+h\right)\left(x^k\right)\right]+2\left[\left(f+h\right)\left(x^k\right)-\left(f+h\right)\left(x^{k+1}\right)\right]\\-\beta_k\left\|x^k-J_k\right\|^2+2\beta_k\left\|x-x^k\right\|\left\|\nabla f\left(x^k\right)-g_k\left(x^k\right)\right\|.\end{aligned}$$

又由于 $x^k-x^{k+1}=\beta_k\left(J_k-x^k\right)$ ，且 $\beta_k\in(0,1)$ ，则 $\beta_k^2\leq\beta_k$ 。即 $\left\|x^k-x^{k+1}\right\|^2-\beta_k\left\|J_k-x^k\right\|^2\leq0$ 。

从而(10)得证。

特别地，当 $x=x^k$ 时，由(10)有

$$\left(f+h\right)\left(x^k\right)-\left(f+h\right)\left(x^{k+1}\right)\geq\frac{1}{2}\left\|x^{k+1}-x\right\|^2\geq0. \tag{11}$$

从而序列 $\left\{\left(f+h\right)\left(x^k\right)\right\}$ 单调减少，这说明算法 1 为下降算法。

关于算法 1，有以下收敛性结论。

**定理 1** 设解集 $S_*\neq\varnothing$ ，$\left\{x^k\right\}$ 和 $\left\{\beta_k\right\}$ 为算法 1 所产生的序列。记 $d_0:=\operatorname{dist}\left(x^0,S_*\right)$ ，假设近似梯度 $g_k$ 连续且存在 $a\geq1$ 满足

$$\Delta:=\sum_{k=0}^{\infty}\left\|\nabla f\left(x^k\right)-g_k\left(x^k\right)\right\|<+\infty, \tag{12}$$

则序列 $\left\{x^k\right\}$ 收敛于 $S_*$ 中一点，即存在 $\bar{x}\in S_*$ ，使得

$$\lim_{k\to\infty}x^k=\bar{x}. \tag{13}$$

**证明：** 在(7)中令 $x=x_*\in S_*$ ，有

$$\left\|x^{k+1}-x_*\right\|^2-\left\|x^k-x_*\right\|^2\le 2\left[(f+h)(x^k)-(f+h)(x^{k+1})\right]+2\beta_k\left\|x_*-x^k\right\|\cdot\left\|\nabla f(x^k)-g_k(x^k)\right\|,\ \forall k\in N. \quad (14)$$

由于序列$\left\{(f+h)(x^k)\right\}$单调递减，则

$$\sum_{k=0}^{\infty}\left[(f+h)(x^k)-(f+h)(x^{k+1})\right]\le(f+h)(x^0)-(f+h)(x_*)<+\infty.$$

由(12)与引理 5 (用$(f+h)(x^k)-(f+h)(x^{k+1})$，$\left\|\nabla f(x^k)-g_k(x^k)\right\|$和$\left\|x^k-x_*\right\|$代替$a_k,b_k$和$c_k$)可知序列$\left\{\left\|x^k-x_*\right\|\right\}$有界，从而有

$$\begin{aligned}&2\sum_{k=0}^{\infty}\left[(f+h)(x^k)-(f+h)(x^{k+1})\right]+2\sum_{k=0}^{\infty}\beta_k\left\|x_*-x^k\right\|\cdot\left\|\nabla f(x^k)-g_k(x^k)\right\|\\&\le 2\left[(f+h)(x^0)-(f+h)(x_*)\right]+2\sum_{k=0}^{\infty}\beta_k\left\|x_*-x^k\right\|\cdot\left\|\nabla f(x^k)-g_k(x^k)\right\|<+\infty.\end{aligned}$$

结合(14)与定义 3 可知序列$\left\{x^k\right\}$ quasi-Fejér 收敛到$S_*$。又由引理 4 知序列$\left\{x^k\right\}$有界，假设$\bar{x}$为$\left\{x^k\right\}$的一个聚点，下面证明$\bar{x}\in S_*$。令

$$\hat{\beta}_k:=\frac{\beta_k}{\theta}>\beta_k>0,$$

$$\hat{y}_k:=x^k-\hat{\beta}_k(x^k-J_k)=(1-\hat{\beta}_k)x^k+\hat{\beta}_kJ_k. \quad (15)$$

由算法 1 的 step3 可知

$$(f+h)(\hat{y}_k)>(f+h)(x^k)-\hat{\beta}_k\left[h(x^k)-h(J_k)\right]-\hat{\beta}_k\left\langle g_k(x^k),x^k-J_k\right\rangle+\frac{\hat{\beta}_k}{2}\left\|x^k-J_k\right\|^2.$$

结合(2)和(15)可得

$$\begin{aligned}0&>(f+h)(x^k)-(f+h)(\hat{y}_k)-\hat{\beta}_k\left[h(x^k)-h(J_k)\right]-\hat{\beta}_k\left\langle g_k(x^k),x^k-J_k\right\rangle+\frac{\hat{\beta}_k}{2}\left\|x^k-J_k\right\|^2\\&\ge\left\langle g_k(\hat{y}_k),x^k-\hat{y}_k\right\rangle+\left\langle\nabla f(\hat{y}_k)-g_k(\hat{y}_k),x^k-\hat{y}_k\right\rangle+h(x^k)-(1-\hat{\beta}_k)h(x^k)-\hat{\beta}_kh(J_k)\\&\quad-\hat{\beta}_k\left[h(x^k)-h(J_k)\right]-\hat{\beta}_k\left\langle g_k(x^k),x^k-J_k\right\rangle+\frac{\hat{\beta}_k}{2}\left\|x^k-J_k\right\|^2\\&=\hat{\beta}_k\left\langle g_k(\hat{y}_k)-g_k(x^k),x^k-J_k\right\rangle+\hat{\beta}_k\left\langle\nabla f(\hat{y}_k)-g_k(\hat{y}_k),x^k-J_k\right\rangle+\frac{\hat{\beta}_k}{2}\left\|x^k-J_k\right\|^2.\end{aligned}$$

整理可得

$$\frac{1}{2}\left\|x^k-J_k\right\|\le\left\|g_k(\hat{y}_k)-g_k(x^k)\right\|+\left\|\nabla f(\hat{y}_k)-g_k(\hat{y}_k)\right\|. \quad (16)$$

由于算子$\mathrm{prox}_h(\cdot)$的非扩张性，由(7)可知$\left\|J_k-J_0\right\|\le\left\|x^k-x^0\right\|+\left\|g(x^k)-g(x^0)\right\|$。

由(15)可知当$\beta_k\to 0$时，有$\lim_{k\to\infty}\left\|\hat{y}_k-x^k\right\|=0$，由$g$的连续性可知当$k\to\infty$时$\left\|g_k(\hat{y}_k)-g_k(x^k)\right\|\to 0$。结合引理 3 可知$\left\|\nabla f(\hat{y}_k)-g_k(\hat{y}_k)\right\|\to 0$，这表明了

$$\lim_{k\to\infty}\left\|x^k-J_k\right\|=0. \quad (17)$$

注意$\bar{x}$为$\left\{x^k\right\}$的一个聚点，即存在子序列$\left\{x^{k_j}\right\}$收敛到$\bar{x}$，则$\left\{J_{k_j}\right\}$也收敛到$\bar{x}$，则又由引理 3 有

$$\lim_{j\to\infty}\left\|g_{k_j}(x^{k_j})-\nabla f(J_{k_j})\right\|=0. \quad (18)$$

故在(3)中令$z=x^{k_j}-g_{k_j}(x^{k_j})$可得到

$$x^{k_j}-J_{k_j}-g_{k_j}\left(x^{k_j}\right)+\nabla f\left(J_{k_j}\right)\in\partial h\left(J_{k_j}\right)+\nabla f\left(J_{k_j}\right)=\partial(f+h)\left(J_{k_j}\right).$$

令 $j\to\infty$，由(17)及(18)可得 $0\in\partial(f+h)(\bar{x})$，从而 $\bar{x}\in S_*$，由引理 4 可得(13)。

### 3.2. 算法 1 的迭代复杂度

在本小节中，我们将分析算法 1 的迭代复杂性。下面的定理表示当线搜索步长 $\{\beta_k\}$ 有正的下界时，函数值的收敛速度为 $o\left(k^{-1}\right)$，这与([13], Method3)中的(精确)邻近梯度算法的迭代复杂度类似。

**定理 2** 设定理 1 中的假设成立，且满足 $\inf_{k\in N}\beta_k\geq\beta>0$，则下列估计式成立：

$$\lim_{k\to\infty}k\left[(f+h)\left(x^k\right)-\min_{x\in R^m}(f+h)(x)\right]=0.$$

**证明**：由定理 1 可设

$$\lim_{k\to\infty}x^k=x_*\in S_*,$$

因此对任意 $\varepsilon>0$，存在 $K>0$ 使得

$$\left\|x^k-x_*\right\|\leq\varepsilon\text{ 和 }(f+h)\left(x^k\right)-(f+h)(x_*)\leq\varepsilon\quad\forall k\geq K. \tag{19}$$

由命题 1 (10) (令 $x=x_*$)，对任意 $l\in N$ 有

$$\begin{aligned}0&\geq(f+h)(x_*)-(f+h)\left(x^l\right)\\&\geq\frac{1}{2\beta_l}\left(\left\|x^{l+1}-x_*\right\|^2-\left\|x^l-x_*\right\|^2+2\left[(f+h)\left(x^{l+1}\right)-(f+h)\left(x^l\right)\right]-2\beta_l\left\|x_*-x^l\right\|\cdot\left\|\nabla f\left(x^l\right)-g_k\left(x^l\right)\right\|\right)\\&\geq\frac{1}{\beta_l}\left(\left\|x^{l+1}-x_*\right\|^2-\left\|x^l-x_*\right\|^2+2\left[(f+h)\left(x^{l+1}\right)-(f+h)\left(x^l\right)\right]-2\beta_l\left\|x_*-x^l\right\|\cdot\left\|\nabla f\left(x^l\right)-g_k\left(x^l\right)\right\|\right).\end{aligned}$$

上式不等式对 $l=K,K+1,\cdots,K+k-1$ 求和得

$$\begin{aligned}&k(f+h)(x_*)-\sum_{l=K}^{K+k-1}(f+h)\left(x^l\right)\\&\geq\frac{1}{2\beta_l}\left(\left\|x^{K+k}-x_*\right\|^2-\left\|x^K-x_*\right\|^2+2\left[(f+h)\left(x^{K+k}\right)-(f+h)\left(x^K\right)\right]\right.\\&\quad\left.-\sum_{l=K}^{K+k-1}2\beta_l\left\|x_*-x^l\right\|\cdot\left\|\nabla f\left(x^l\right)-g_k\left(x^l\right)\right\|\right).\end{aligned} \tag{20}$$

即注意到 $\left\{(f+h)\left(x^k\right)\right\}$ 单调减少，我们有

$$\sum_{l=K}^{K+k-1}(f+h)\left(x^l\right)\leq k(f+h)\left(x^K\right),\ (f+h)\left(x^{K+k}\right)-(f+h)\left(x^K\right)\geq(f+h)(x_*)-(f+h)\left(x^K\right).$$

这和(20)一起表明

$$\begin{aligned}k\left[(f+h)(x_*)-(f+h)\left(x^{K+k}\right)\right]&\geq-\frac{1}{2\beta_l}\left\|x^K-x_*\right\|^2-\frac{1}{\beta_l}\left[(f+h)\left(x^K\right)-(f+h)(x_*)\right]\\&\quad-\sum_{l=K}^{K+k-1}\left\|x_*-x^l\right\|\cdot\left\|\nabla f\left(x^l\right)-g_k\left(x^l\right)\right\|\\&\geq-\left(\frac{\varepsilon^2}{2\beta}+\frac{\varepsilon}{\beta}+\varepsilon\Delta\right),\end{aligned}$$

其中最后一个不等式因为(12)，(19)和$\inf_{k\in N}\beta_k \geq \beta$。由于$\varepsilon > 0$的任意性，我们有

$$\limsup_{k\to\infty} k\left[\left(f+h\right)\left(x^k\right)-\left(f+h\right)\left(x_*\right)\right] \leq 0.$$

又$\left(f+h\right)\left(x^k\right)-\left(f+h\right)\left(x_*\right)\geq 0$，所以定理结论成立。

下列命题表明在$\nabla f_i$ $(i\in\psi)$的 Lipschitz 连续假设条件下，$\{\beta_k\}$有正的下界。证明思路与([13], Proposition 5.4)类似，所以我们省略了它的证明。

**命题 2** 设$\{\beta_k\}$为线搜索 1 的算法 1 所产生的序列，设$x_*$是$\{x^k\}$的极限点，即$\lim_{k\to\infty} x^k = x_*$，近似梯度序列$\{g_k\}$为抽样产生。若对每个$i\in\psi$，$\nabla f_i$在点$x_*$是常数为$L_i\left(x_*\right)>0$的局部 Lipschitz 连续的，则

$$\liminf_{k\to\infty} \beta_k \geq \min\left\{1, \frac{\theta}{2L\left(x_*\right)}\right\},$$

其中$L\left(x_*\right) := \max_{i\in\psi} L_i\left(x_*\right)$。

## 4. 结论

本文介绍了一种求解大规模复合凸优化的非精确邻近梯度算法，算法不依赖可微函数的 Lipschitz 常数，并分析了算法的收敛性。接下来，我们可以构造求解大规模优化问题的非精确求解其他方法，得到高效的算法。

## 基金项目

国家自然科学基金(No. 12161017)；贵州省科技厅科技计划项目(No. ZK[2022]110)。

## 参考文献

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

[2] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. *Advances in Neural Information Processing Systems*, **1**, 315-323.

[3] Erdogdu, M.A. and Montanari, A. (2015) Convergence Rates of Sub-Sampled Newton Methods. *International Conference on Neural Information Processing Systems*, MIT Press, 28.

[4] Agarwal, N., Bullins, B. and Hazan, E. (2017) Second-Order Stochastic Optimization for Machine Learning in Linear Time. *Journal of Machine Learning Research*, **18**, 1-14.

[5] Moreau, J.J. (1962) Fonctions convexes duales et points proximaux dans un espacehilbertien. *Comptes rendushebdomadaires des séances de l'Académie des Sciences*, **255**, 2897-2899.

[6] Yuan, X. (2012) Alternating Direction Methods for Sparse Covariance Selection. *Journal of Scientific Computing*, **51**, 261-273. https://doi.org/10.1007/s10915-011-9507-1

[7] Beck, A. and Teboulle, M. (2009) Gradient-Based Algorithms with Applications to Signal Recovery. *Convex Optimization in Signal Processing and Communications*, 42-88. https://doi.org/10.1017/CBO9780511804458.003

[8] Machart, P., Anthoine, S. and Baldassarre, L. (2012) Optimal Computational Tradeoff of Inexact Proximal Methods.

[9] Moreau, J.J. (1965) Proximité et dualité dans un espacehilbertien. *Bulletin de la Société mathématique de France*, **93**, 273-299. https://doi.org/10.24033/bsmf.1625

[10] Krasnoselkii, M.A. (1957) Two Observations about the Method of Successive Approximations, Uspehi Math. *Nauk*, **10**, 131-140.

[11] Rockafellar, R.T. (1976) Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, **14**, 877-898. https://doi.org/10.1137/0314056

[12] Rockafellar, R.T. (1976) Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming. *Mathematics of Operations Research*, **1**, 97-116. https://doi.org/10.1287/moor.1.2.97

[13] Bello Cruz, J.Y. and Nghia, T.T.A. (2016) On the Convergence of the Forward-Backward Splitting Method with Line-

searches. *Optimization Methods and Software*, **31**, 1209-1238. https://doi.org/10.1080/10556788.2016.1214959

[14] Polyak, B.T. and Juditsky, A. B. (1992) Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, **30**, 838-855. https://doi.org/10.1137/0330046

[15] Bauschke, H.H. and Combettes, P.L. (2011) Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York. https://doi.org/10.1007/978-1-4419-9467-7

[16] Roosta-Khorasani, F. and Mahoney, M.W. (2019) Sub-Sampled Newton Methods. *Mathematical Programming*, **174**, 293-326. https://doi.org/10.1007/s10107-018-1346-5

[17] Iusem, A.N., Svaiter, B.F. and Teboulle, M. (1994) Entropy-Like Proximal Methods in Convex Programming. *Mathematics of Operations Research*, **19**, 790-814. https://doi.org/10.1287/moor.19.4.790