Advances in Applied Mathematics 应用数学进展, 2024, 13(6), 2641-2652
Published Online June 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.136252

Hans汉斯

一类带偶次惩罚范数的非凸函数及周期ADMM算法的收敛性分析

宋政纲

兰州理工大学理学院，甘肃 兰州

收稿日期：2024年5月17日；录用日期：2024年6月11日；发布日期：2024年6月18日

摘　要

在机器学习以及其它相关领域中，针对非凸函数的优化问题，目前存在的算法理论上对非凸函数的收敛和全局稳定性无法得到有效保证。本文提出将*Lp*范数(*p*为偶数)引入到非凸函数中，并在此基础上设计一种周期交替方向乘子(Periodic Alternating Direction Method of Multipliers, PADMM)的优化算法，用于此类非凸函数收敛性分析。我们证明在惩罚参数足够大的情况下，带偶次惩罚范数的非凸函数必收敛，并且收敛到全局最小值。此外，PADMM算法不对变量更新的先后顺序作特殊要求，这一特性大大增强了PADMM算法在处理各类非凸函数优化问题时的普适性。

关键词

机器学习，非凸函数，*Lp*范数，交替方向乘子

Convergence Analysis of a Class of Nonconvex Functions with Even-Powered Penalty Norms and the Periodic ADMM Algorithm

Zhenggang Song

School of Science, Lanzhou University of Technology, Lanzhou Gansu

Received: May 17th, 2024; accepted: Jun. 11th, 2024; published: Jun. 18th, 2024

Abstract

In machine learning and other related fields, for the optimization problem of non-convex func-

文章引用: 宋政纲. 一类带偶次惩罚范数的非凸函数及周期 ADMM 算法的收敛性分析[J]. 应用数学进展, 2024, 13(6): 2641-2652. DOI: 10.12677/aam.2024.136252

tions, the existing algorithms cannot effectively guarantee the convergence and global stability of non-convex functions in theory. In this paper, the *Lp* norm (*p* is even) is introduced into the non-convex function, and on this basis, an optimization algorithm of Periodic Alternating Direction Method of Multipliers (PADMM) is designed for the convergence analysis of such non-convex functions. We prove that when the penalty parameter is large enough, the nonconvex function with even penalty norm will converge and converge to the global minimum. In addition, the PADMM algorithm does not impose special requirements on the order of variable updating, which greatly enhances the universality of the PADMM algorithm in dealing with various non-convex function optimization problems.

Keywords

Machine Learning, Nonconvex Function, *Lp* Norm, Alternating Direction Multiplier

1. 引言

非凸函数优化问题在多个科学与工程领域都有着广泛的应用，[1]中给出非凸函数在机器学习、人工智能等领域的应用。[2]提出大部分非凸优化问题，由于具有局部极值点多、曲率变化复杂、数据维数庞大等特点，使得其求解难度尤为突出。这就要求研究者必须从非凸函数模型和求解算法方面设计出更为泛化且鲁棒的优化策略。

近年来，研究者提出了一系列用于分析非凸优化问题的算法。如[3]-[7]中的随机梯度下降及其变体、动量法、信赖域法、拟牛顿法、基于概率和迭代全局搜索技术等。这些算法均适应于处理大规模的非凸优化问题，在特定的条件下可提高局部收敛速度和准确性。然而这些算法普遍面临参数调整的挑战、可能会陷入局部最优、在处理大规模或复杂问题时收敛速度慢，及对问题特性敏感，导致实际性能与理论预期存在差距。[8]中交替方向乘子法(Alternating Direction Method of Multipliers, ADMM)是在 20 世纪 70 年代初提出的用于解决非线性椭圆型偏微分方程的算法，此后 ADMM 逐渐被应用于更广泛的优化领域，尤其是那些具有可分离结构的非凸优化问题。ADMM 算法相对于传统的原对偶型算法(如[9] [10]中的对偶上升算法或乘数法)收敛速度更快，也特别适合并行实现。然而在某些问题中 ADMM 的迭代解可能出现振荡，即在一组解附近来回波动，而非直接向最优解收敛，这可能需要额外的技术来稳定迭代过程，尽管有大量的研究文献将 ADMM 算法应用于非凸优化问题的实践案例，但对该算法在非凸优化情境下的理论理解和分析仍有较大的局限性。例如，对于大部分的收敛性分析只能针对某些特殊结构的非凸优化问题。[11]表明只有目标函数和约束条件满足一定的附加条件时，ADMM 算法才会线性收敛。[12]研究发现对于多块可分离非凸优化问题和某些病态问题，原始 ADMM 算法可能会发散。因此大部分的研究为了保证非凸问题的收敛性，需要对非凸问题的目标函数和约束条件做出限制，这就可能导致对原始的优化问题产生破坏。[13]提出了多块 ADMM 算法，其将复杂的优化问题分解成多个相对独立的子问题，每个子问题所涉及的变量块较少，可进行单独求解多块 ADMM 算法。在大规模优化问题和分布式计算环境下有明显的优势。[14] [15]提出了全局优化 ADMM 算法，其在原始的 ADMM 算法的基础上引入新的算法设计和分析方法，如光滑技术、正则化技术等，确保非凸函数的全局收敛性。然而以往文献中的 ADMM 算法及其变体对非凸优化问题的收敛性分析非常有限，主要体现在：收敛性条件严格、全局最优

解的不确定性、参数调优困难、理论分析缺失。[16] [17]提出在进行全局收敛性分析时，对于任何已知的方法，都需要对算法所产生的序列设定一些无法直接通过计算过程来检验的条件。

虽然对于非凸优化问题一般没有全局最优解的有效求法，[18]中提出可以通过增加适当的惩罚项改变搜索空间的几何特征，使得一些连续的非凸函数收敛到全局最优解，本文受[18]的启发在非凸函数的变体中添加了惩罚项。[19] [20]提出了 *Lasso* 和 *Ridge* 正则化方法，其在处理非凸函数时可以约束模型复杂度，生成稀疏解并且可提升优化过程的稳定性和有效性。[21]针对非凸优化问题提出了使用非凸惩罚项和范数进行优化的方法，使得非凸函数的收敛性较好，但是难以找到全局最优解，可能会陷入局部最优解。[22]提出了一种非凸鲁棒主成分分析法，其在特定条件下可能发现一个局部极小点，这个局部极小点比使用传统的凸优化方法找到的局部极小点表现得更好，更加的接近全局极小值点。然而这些方法一般不能保证全局最优性。

本文在[19] [20]的基础上提出了将 Lp 范数(p 为偶数)引入到非凸函数中作为惩罚项，通过设计 PADMM 算法进行非凸优化函数的理论分析，通过理论分析证明我们提出的方法在惩罚参数足够大的情况下，非凸函数必定会收敛，并且收敛到全局极值点。

2. 周期交替方向乘子法

考虑以下非凸问题：

$$\begin{aligned}&\min_{x} h_1(x)+h_2(x)+\cdots+h_K(x)+g(x)\\&\text{s.t. } x\in X\end{aligned} \tag{1}$$

其中 $h_i(x)$ 可以是光滑凸函数也可以是光滑非凸函数， $g(x)$ 是非光滑凸函数，X 为闭凸集。

在实际问题分析中，(1)中的 $h_i(x)$ 需要进行单独处理。为了便于分析，可在(1)的基础上引入一组新的变量 $(x_1,x_2,\cdots,x_K)$，则(1)可重新表述为(2)：

$$\begin{aligned}&\min_{x_0} h_1(x_1)+h_2(x_2)+\cdots+h_K(x_K)+g(x_0)\\&\text{s.t. } x_k=x_0,\forall k=1,2,\cdots,K\\&x_0\in X\end{aligned} \tag{2}$$

(2)中所有子问题共享变量 x_0，每个子问题负责优化自身的局部目标函数 $h_K(x_K)$，所有子问题的解满足一致性条件 $x_k=x_0$。通过每个局部函数的最优达到整体最优。在重新引入变量 $(x_1,x_2,\cdots,x_K)$ 后问题的维数增加到了 K，(2)和(1)相比较增大了问题求解的迭代次数，但(2)确保了每个局部函数 $h_K(x_K)$ 的独立性，即每个分布式节点可以独立地处理各自的变量 x_k。

(2)的拉格朗日函数可由(3)给出：

$$\acute{L}(\{x_k\},x_0;\mu)=\sum_{k=1}^{K}h_k(x_k)+g(x_0)+\sum_{k=1}^{K}\langle\mu_k\cdot(x_k-x_0)\rangle \tag{3}$$

在(2)的基础上添加惩罚项 $\sum_{k=1}^{K}\frac{\xi_k}{p}\|x_k-x_0\|_p^p$，(2)的增广拉格朗日函数由(4)给出：

$$L(\{x_k\},x_0;\mu)=\acute{L}(\{x_k\},x_0;\mu)+\sum_{k=1}^{K}\frac{\xi_k}{p}\|x_k-x_0\|_p^p \tag{4}$$

其中：$\mu=(\mu_1,\mu_2,\cdots,\mu_k)\in R^k$ 是拉格朗日乘子，$\xi_k\in R$ 是惩罚参数，$p\in R$ 且 p 为偶数，$\|x\|_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}$。

3. 非凸 PADMM 算法

为了使原始变量和对偶变量的更新次序有更大的选择性，本节提出一种灵活的 PADMM 算法。设 $x_0,\cdots,x_K$ 的索引为 $k=0,\cdots,K$，令 $D^T=\left(0,1,2,\cdots,K\right)$ 为第 T 次更新的变量集。

PADMM 算法如下：

首先令 $D^1=\left(0,1,2,\cdots,K\right)$，$T=1,2,\cdots$。在初次更新时所有的变量均参与更新。

如果 $T+1\geq 2$，任选 $D^T\subseteq\left(0,1,2,\cdots,K\right)$。在第二次之后的更新是从 $x_0,\cdots,x_K$ 中任选几个变量进行更新，即不必对全部变量进行更新，对变量的更新顺序也没有要求。

当 $k=0$ 时，x_0 的第 $T+1$ 次迭代 x_0^{T+1} 计算方式由(5)给出：

$$x_0^{T+1}=\arg\min_{x_0\in X} L\left(\left\{x_k^T\right\},x_0;\mu^T\right) \tag{5}$$

否则：

$$x_0^{T+1}=x_0^T \tag{6}$$

当 $k\neq 0$ 时，x_k 的第 $T+1$ 次迭代 x_k^{T+1} 计算方式由(7)给出：

$$x_k^{T+1}=\arg\min_{x_k} g_k\left(x_k\right)+\left\langle\mu_k^T,x_k-x_0^{T+1}\right\rangle+\frac{\xi_k}{p}\left\|x_k-x_0^{T+1}\right\|_p^p \tag{7}$$

第 k 个拉格朗日乘子 μ_k 第 $T+1$ 次更新迭代在 x_0^{T+1} 和 x_k^{T+1} 的基础上得到，μ_k^{T+1} 的计算方式由(8)给出：

$$\mu_k^{T+1}=\mu_k^T+\xi_k\left(\left|x_k^{T+1}-x_0^{T+1}\right|\right)^{p-1} \tag{8}$$

否则：

$$x_k^{T+1}=x_k^{T+1},\quad \mu_k^{T+1}=\mu_k^T \tag{9}$$

假定存在一个正周期 M，我们规定 $\bigcup_{i=1}^M D^{T+i}=\left(0,1,2,\cdots,K\right)$，即在一个周期内每个变量至少更新一次。

上述变量更新的迭代计算中，我们规定 $\mu_k^{T+1}=\mu_k^T+\xi_k\left(\left|x_k^{T+1}-x_0^{T+1}\right|\right)^{p-1}$，其余变量都是通过极小化目标函数得到的。

4. 收敛性分析

为了对 PADMM 算法进行理论分析，我们给出下面假设。

假设 1) $h_k\left(x\right)$ 满足利普希茨条件条件即存在 L_k 使得(10)成立。

$$\left\|\nabla h_k\left(x_k\right)-\nabla h_k\left(y_k\right)\right\|\leq L_k\left\|x_k-y_k\right\| \tag{10}$$

假设 2) $f\left(x\right)$ 在定义域中存在下界。

$$f\left(x\right)>-\infty \tag{11}$$

假设 3) 对于所有的迭代次数 k，(7)为模数为 $m_k\left(\xi_k\right)$ 的强凸函数，并且对于所有的迭代次数 k 有 $\xi_k m_k\left(\xi_k\right)>pL_k^2$。

令 T_k 表示在更新迭代 $T+1$ 之前，x_k 最后一次被更新的迭代索引。即：

$$T\left(k\right)=\max\left(a\,|\,a\leq T,k\in D^a\right),\quad T\left(0\right)=\max\left(a\,|\,a\leq T,0\in D^a\right)。\ k=1,2,\cdots,K$$

令：$\hat{x}_0^{T+1}=\arg\min_{x_0\in X} L\left(\left\{x_k^T\right\},x_0;\mu^T\right)$

$$\hat{x}_k^{T+1}=\arg\min_{x_k} g_k\left(x_k\right)+\left\langle\mu_k^T,x_k-\hat{x}_0^{T+1}\right\rangle+\frac{\xi_k}{p}\left\|x_k-\hat{x}_0^{T+1}\right\|_p^p$$

$$\hat{\mu}_k^{T+1}=\mu_k^T+\xi_k\left(\hat{x}_k^{T+1}-\hat{x}_0^{T+1}\right)$$

$$\tilde{x}_k^{T+1}=\arg\min_{x_k} g_k\left(x_k\right)+\left\langle\mu_k^T,x_k-x_0^T\right\rangle+\frac{\xi_k}{p}\left\|x_k-x_0^T\right\|_p^p$$

$$\tilde{\mu}_k^{T+1}=\mu_k^T+\xi_k\left(\tilde{x}_k^{T+1}-x_0^T\right)$$

下面证明对偶变量的连续变化量的大小上限是由原变量的变化量大小决定的。

定理 1：在假设 1)、2)、3)成立的情况下 $\forall k=1,2,\cdots,K$ 有以下的条件成立。

$$L_k^2\left\|x_k^{T+1}-x_k^T\right\|^2\geq\left\|\mu_k^{T+1}-\mu_k^T\right\|^2 \tag{12}$$

$$L_k^2\left\|\hat{x}_k^{T+1}-x_k^T\right\|^2\geq\left\|\hat{\mu}_k^{T+1}-\mu_k^T\right\|^2 \tag{13}$$

$$L_k^2\left\|\tilde{x}_k^{T+1}-x_k^T\right\|^2\geq\left\|\tilde{\mu}_k^{T+1}-\mu_k^T\right\|^2 \tag{14}$$

证明：在此只对(12)展开证明，(13)、(14)同理可得。

$\forall k\in D^{T+1}$ 从 x_k 开始更新迭代，我们可以得到：

$$\nabla h_k\left(x_k^{T+1}\right)+\mu_k^T+\xi_k\left(\left|x_k^{T+1}-x_0^{T+1}\right|\right)^{p-1}=0$$

又因为 $\mu_k^{T+1}=\mu_k^T+\xi_k\left(x_k^{T+1}-x_0^{T+1}\right)^{p-1}$，可得：

$$\nabla h_k\left(x_k^{T+1}\right)=-\mu_k^{T+1}$$

由假设 1)可得：

$$\left\|\mu_k^{T+1}-\mu_k^T\right\|=\left\|\mu_k^{T+1}-\mu_k^{T(k)}\right\|=\left\|\nabla h_k\left(x_k^{T+1}\right)-\nabla h_k\left(x_k^{T(k)}\right)\right\|$$

由利普希茨连续条件可得：

$$\left\|\nabla h_k\left(x_k^{T+1}\right)-\nabla h_k\left(x_k^{T(k)}\right)\right\|\leq L_k\left\|x_k^{T+1}-x_k^{T(k)}\right\|=L_k\left\|x_k^{T+1}-x_k^T\right\|$$

即：

$$\left\|\mu_k^{T+1}-\mu_k^T\right\|\leq L_k\left\|x_k^{T+1}-x_k^T\right\|$$

证毕。

定理 2：对于 PADMM 算法可以得到以下结论：

$$L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)$$
$$\leq\sum_{k\in D^{T+1}}\left(\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{p}\right)\left\|x_k^{T+1}-x_k^T\right\|^2-\frac{m}{p}\left(x_0^{T+1}-x_0^T\right)$$

证明：

$$
\begin{aligned}
&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T};\mu^{T}\right)\\
&=\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left\langle\mu^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\\
&\quad-\left(\sum_{k=1}^{K}h_k\left(x_k^{T}\right)+g\left(x_0^{T}\right)\right)-\left(\sum_{k=1}^{K}\left\langle\mu^{T}\cdot\left(x_k^{T}-x_0^{T}\right)\right\rangle+\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T}-x_0^{T}\right\|_p^p\right)\\
&=\left(\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left\langle\mu^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle\right)+\left(\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\\
&\quad-\left(\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)\right)-\left(\sum_{k=1}^{K}\left\langle\mu^{T}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\\
&\quad+\left(\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left\langle\mu^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle\right)+\left(\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\\
&\quad-\left(\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)\right)-\left(\sum_{k=1}^{K}\left\langle\mu^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\\
&=\left(L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)\right)\\
&\quad+\left(L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T};\mu^{T}\right)\right)
\end{aligned}
$$

因为：

$$
\begin{aligned}
&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)\\
&=\sum_{k=1}^{K}\left\langle\mu^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle-\sum_{k=1}^{K}\left\langle\mu^{T}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle\\
&=\sum_{k=1}^{K}\left\langle\left(\mu^{T+1}-\mu^{T}\right)\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle
\end{aligned}
$$

又因为：

$$
x_k^{T+1}-x_0^{T+1}=\left(\left|\frac{\mu^{T+1}-\mu^{T}}{\xi_k}\right|\right)^{\frac{1}{p-1}}
$$

所以：

$$
L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)=\sum_{k\in D^{T+1}}\frac{1}{\xi_k^{p-1}}\left(\left|\mu^{T+1}-\mu^{T}\right|\right)^{\frac{p}{p-1}}
$$

又因为：

$$
\begin{aligned}
&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T};\mu^{T}\right)\\
&=L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T+1};\mu^{T}\right)+L\left(\left\{x_k^{T}\right\},x_0^{T+1};\mu^{T}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T};\mu^{T}\right)
\end{aligned}
$$

所以：

$$
\begin{aligned}
&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right)-L\left(\left\{x_k^{T}\right\},x_0^{T};\mu^{T}\right)\\
&\le\sum_{k=1}^{K}\left(\left\langle\nabla_{x_k}L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T}\right),\left(x_k^{T+1}-x_0^{T}\right)\right\rangle-\frac{m_k\left(\xi_k\right)}{p}\left\|x_k^{T+1}-x_k^{T}\right\|_p^p\right)\\
&\quad+\left\langle\tau_{x_0}^{T+1},\left(x_0^{T+1}-x_0^{T}\right)\right\rangle-\frac{m}{p}\left\|x_0^{T+1}-x_0^{T}\right\|_p^p
\end{aligned}
$$

令 $\tau_{x_0}^{T+1}\in\partial_{x_0}\left(L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)\right)$ 是 x_0 的次导数，$\sigma=\begin{cases}1, & 0\in D^{T+1}\\ 0, & 0\notin D^{T+1}\end{cases}$，$m=\sum_{k=1}^{K}\xi_k$。

由凸函数的性质可得：

$$\begin{aligned}&\sum_{k=1}^{K}\left(\left\langle\nabla_{x_k}L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^T\right),\left(x_k^{T+1}-x_0^T\right)\right\rangle-\frac{m_k\left(\xi_k\right)}{p}\left\|x_k^{T+1}-x_k^T\right\|_p^p\right)+\left\langle\tau_{x_0}^{T+1},\left(x_0^{T+1}-x_0^T\right)\right\rangle-\left\|x_0^{T+1}-x_0^T\right\|_p^p\\&=\sum_{k\in D^{T+1}}\left(\left(\left\langle\nabla_{x_k}L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^T\right),\left(x_k^{T+1}-x_0^T\right)\right\rangle-\frac{m_k\left(\xi_k\right)}{p}\left\|x_k^{T+1}-x_k^T\right\|_p^p\right)\right.\\&\quad\left.+\sigma\left(\left\langle\tau_{x_0}^{T+1},\left(x_0^{T+1}-x_0^T\right)\right\rangle\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|_p^p\right)\right)\\&\le-\sum_{k\in D^{T+1}}\frac{m_k\left(\xi_k\right)}{p}\left\|x_k^{T+1}-x_k^T\right\|_p^p-\sigma\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|_p^p\end{aligned}$$

综上可得：

$$\begin{aligned}&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)\\&\le-\sum_{k\in D^{T+1}}\frac{m_k\left(\xi_k\right)}{p}\left\|x_k^{T+1}-x_k^T\right\|^p-\sigma\frac{m}{p}\left\|x_k^{T+1}-x_k^T\right\|_p^p+\sum_{k\in D^{T+1}}\frac{1}{\xi_k}\left\|x_0^{T+1}-x_0^T\right\|_p^p\\&\le\sum_{k\in D^{T+1}}\left(\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{p}\right)\left\|x_k^{T+1}-x_k^T\right\|_p^p-\sigma\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|_p^p\end{aligned}$$

证毕。

由定理 2 可知当 $\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{P}\le0$，即 $pL_k^2\le\xi_k m_k\left(\xi_k\right)$ 时，

$\sum_{k\in D^{T+1}}\left(\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{p}\right)\left\|x_k^{T+1}-x_k^T\right\|_p^p-\sigma\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|_p^p$，也就是说当选取较大的惩罚参数 ξ_k 时，即可保证增广拉格朗日函数是递减的。

下面证明构造的增广拉格朗日函数是收敛的。

定理 3：对于增广拉格朗日函数有以下的极限存在：

$$\lim_{T\to\infty}L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)\ge-\infty$$

证明：

$$\begin{aligned}&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)\\&=\sum_{k=1}^{K}h_k\left(x_k^{T+1}\right)+g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left\langle\mu_k^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\sum_{k=1}^{K}\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\\&=g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left(h_k\left(x_k^{T+1}\right)+\left\langle\mu_k^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\end{aligned}$$

又由定理 1 可知：

$$\nabla h_k\left(x_k^{T+1}\right)=-\mu_k^{T+1}$$

所以有：

$$
\begin{aligned}
&g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left(h_k\left(x_k^{T+1}\right)+\left\langle\mu_k^{T+1}\cdot\left(x_k^{T+1}-x_0^{T+1}\right)\right\rangle+\frac{\xi_k}{p}\left\|x_k^{T+1}-x_0^{T+1}\right\|_p^p\right)\\
&=g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}\left(h_k\left(x_k^{T+1}\right)+\left\langle\nabla h_k\left(x_k^{T+1}\right)\cdot\left(x_0^{T+1}-x_k^{T+1}\right)\right\rangle+\frac{\xi_k}{p}\left\|x_k^{T+1}x_0^{T+1}\right\|_p^p\right)\\
&\geq g\left(x_0^{T+1}\right)+\sum_{k=1}^{K}h_k\left(x_0^{T+1}\right)=f\left(x_0^{T+1}\right)
\end{aligned}
$$

由假设可知 $f\left(x_0^{T+1}\right)\geq-\infty$，即 $\lim_{T\to\infty}L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)\geq-\infty$。所以构造的增广拉格朗日函数是收敛的。

证毕。

下面证明算法收敛于平稳解集合。

定理 4：根据 PADMM 算法可知，对于所有的 $k=1,2,\cdots,K$ 有：

$\lim_{T\to\infty}\left\|x_k^{T+1}-x_0^{T+1}\right\|=0$ 成立。

证明：根据定理 2 有：

$$
\begin{aligned}
&L\left(\left\{x_k^{T+1}\right\},x_0^{T+1};\mu^{T+1}\right)-L\left(\left\{x_k^T\right\},x_0^T;\mu^T\right)\\
&\leq\sum_{k\in D^{T+i}}\left(\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{p}\right)\left\|x_k^{T+1}-x_k^T\right\|_p^p-\sigma\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|_p^p\\
&=\sum_{i=1}^{M}\sum_{k\in D^{T+i}}\left(\frac{L_k^2}{\xi_k}-\frac{m_k\left(\xi_k\right)}{p}\right)\left\|x_k^{T+1}-x_k^T\right\|_p^p-\frac{m}{p}\left\|x_0^{T+1}-x_0^T\right\|
\end{aligned}
$$

如果 $k\neq0$，$k\notin D^{T+i}$，则有 $x_k^{T+i}-x_k^{T+i-1}$，利用 k 在 $[T,T+M]$ 至少更新一次以及定理 3，对于 $k=1,2,\cdots,K$ 可得：

$$\left\|x_0^{T+1}-x_0^{T(k)}\right\|\to0,\ \left\|x_k^{T+1}-x_k^{T(k)}\right\|\to0$$

由定理 1 可知对于 $k=1,2,\cdots,K$ 有 $\left\|\mu_k^{T+1}-\mu_k^{T(k)}\right\|\to0$，根据 PADMM 算法的迭代步骤，当 $\left\|\mu_k^{T+1}-\mu_k^{T(k)}\right\|\to0$ 时，可得 $\left\|x_k^{T+1}-x_0^{T+1}\right\|\to0$。

证毕。

定理 5：假设 $\left(\left\{x_k^*\right\},x_0^*,x_k^*\right)$ 是 PADMM 算法的全局极限点，那么有以下式子成立：

$$\nabla h_k\left(x_k^*\right)+x_k^*=0$$

$$x_0^*\in\arg\min_{x\in X}g\left(x\right)+\sum_{k=1}^{K}\left\langle x_0^*,x_k^*-x\right\rangle$$

$$x_k^*=x_0^*$$

$$\forall k=1,2,\cdots,K$$

证明：对于 $k\in D^{T+1}$，$k\neq0$ 有：

$$\nabla h_k\left(x_k^{T+1}\right)+\mu_k^T+\xi_k\left(x_k^{T+1}-x_0^{T+1}\right)=0$$

假设 $0\in D^{T+1}$，令 $\theta^{T+1}\in\partial g\left(x_0^{T+1}\right)$，可得：

$$\left\langle\left(x-x_0^{T+1}\right)\cdot\left(\theta^{T+1}-\sum_{k=1}^{K}\left(\mu_k^T-\xi_k\left(x_0^{T+1}-x_0^T\right)\right)\right)\right\rangle\geq0$$

因为：

$$g(x)-g\left(x_0^{T+1}\right)=\theta^{T+1}\left(x-x_0^{T+1}\right)$$

所以：

$$\left\langle\left(x-x_0^{T+1}\right),\left(\theta^{T+1}-\sum_{k=1}^{K}\left(\mu_k^T-\xi_k\left(x_0^{T+1}-x_0^T\right)\right)\right)\right\rangle$$

$$=g(x)-g\left(x_0^{T+1}\right)+\left\langle\left(x-x_0^{T+1}\right),\sum_{k=1}^{K}\left(\xi_k\left(x_0^{T+1}-x_0^T\right)-\mu_k^T\right)\right\rangle$$

根据 PADMM 算法迭代的规则，当 $k\neq 0$ 时对于所有的 T 有：

$$\nabla h_k\left(x_k^{\varphi(k)}\right)+\mu_k^{\varphi(k)}=0,\varphi(k)\in[T,T+M]$$

对于 $\varphi(0)\in[T,T+M]$，可得：

$$g(x)-g\left(x_0^{\varphi(0)}\right)+\left\langle\left(x-x_0^{\varphi(0)}\right),\sum_{k=1}^{K}\left(\xi_k\left(x_0^{\varphi(0)}-x_0^{\varphi(0)-1}\right)-\mu_k^{\varphi(0)-1}\right)\right\rangle\geq 0$$

根据定理 4，可得：

$$\left\|x_k^{\varphi(k)}-x_k^{T+1}\right\|\to 0,\quad \left\|x_0^{\varphi(0)}-x_0^{T+1}\right\|\to 0$$

$$\left\|\mu_k^{T+1}-\mu_k^{\varphi(k)}\right\|\to 0,\quad \left\|\mu_k^{T+1}-\mu_k^{\varphi(0)-1}\right\|\to 0$$

又因为：

$$\left\|x_k^{T+1}-x_k^T\right\|\to 0,x_0^{T+1}\to x_0^*,x_k^{T+1}\to x_k^*,\mu_k^{T+1}\to\mu_k^*,$$

对 $\nabla h_k\left(x_k^{\varphi(k)}\right)+\mu_k^{\varphi(k)}=0$ 求极限得：

$$\nabla h_k\left(x_k^*\right)+x_k^*=0$$

对 $g(x)-g\left(x_0^{\varphi(0)}\right)+\left\langle\left(x-x_0^{\varphi(0)}\right)+\sum_{k=1}^{K}\left(\xi_k\left(x_0^{\varphi(0)}-x_0^{\varphi(0)-1}\right)-\mu_k^{\varphi(0)-1}\right)\right\rangle\geq 0$ 求极限可得：

$$g(x)-g\left(x_0^*\right)+\sum_{k=1}^{K}\left\langle x-x_0^*,-\mu_k^*\right\rangle\geq 0$$

由于 $\left\|\mu_k^{T+1}-\mu_k^T\right\|\to 0$ 对所有的 k 都成立，可得：

$$x_k^*=x_0^*$$

证毕。

5. PADMM 算法实例应用

下面我们用 PADMM 算法求解以下非凸函数的最优值。

$$\min_x f(x)=\sin x+\sin 3x$$

$$\text{s.t}\ -2\pi\leq x\leq 2\pi$$

图 1 为函数 $f(x)=\sin x+\sin 3x$ 的走势图，从图中可以看出 $f(x)$ 是一个非凸函数，在定义域内函数有增有减。$f(x)$ 的最小值是−1.56。图 2 是 PADMM 算法的部分迭代过程图，可以看出迭代点沿着函数负梯度的方向下降，并且在极值点附近迭代步长逐渐减小。从图 3 中可以看出 $f(x)$ 随着迭代次数的增加，函数值逐渐稳定到−1.56，对比图 1 和图 3 可以看出 PADMM 算法可有效的解决非凸函数优化问题。

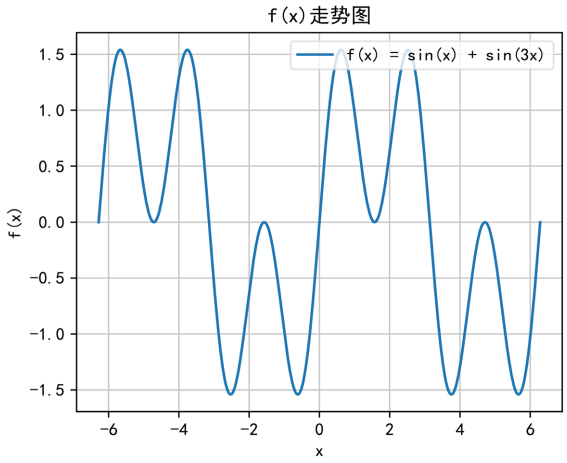

Figure 1. Trend graph of $f(x)$

图 1. $f(x)$走势图

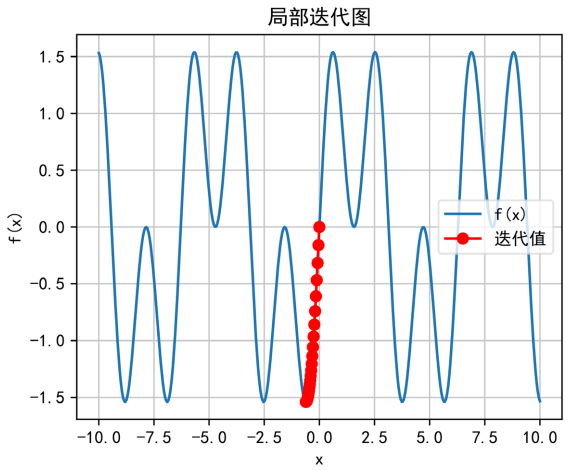

Figure 2. Local iterative graph

图 2. 局部迭代图

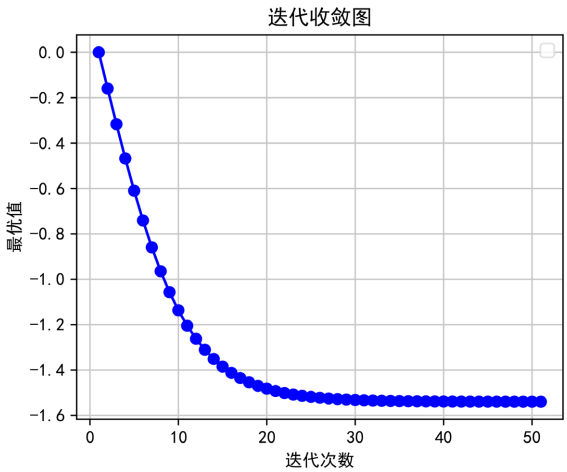

Figure 3. Iteration convergence diagram

图 3. 迭代收敛图

6. 结论

本文提出了一类带偶次惩罚范数的非凸函数和 PADMM 算法，证明了在惩罚参数足够大的情况下，带偶次惩罚范数的非凸函数在 PADMM 算法的求解中是收敛的。并且非凸函数的解收敛于平稳集，在存在极小值的情况下带偶次惩罚范数的非凸函数会收敛到全局极小值。PADMM 算法在求解非凸问题时不用考虑参数更新的顺序，这样可加速函数求解的收敛速度，同时更有普适性。

我们只考虑了偶次范数情况，对于一般的范数并没有进行理论分析，未来的一个研究方向是将偶次范数推广到整个实数空间中。

参考文献

[1] Jain, P. and Kar, P. (2017) Non-Convex Optimization for Machine Learning. *Foundations and Trends® in Machine Learning*, **10**, 142-336. https://doi.org/10.1561/2200000058

[2] Du, S., Lee, J., Li, H., *et al*. (2019) Gradient Descent Finds Global Minima of Deep Neural Networks. *Proceedings of the* 36*th International Conference on Machine Learning*, Long Beach, 28 May 2019, 1675-1685.

[3] Mignacco, F. and Urbani, P. (2022) The Effective Noise of Stochastic Gradient Descent. *Journal of Statistical Mechanics*: *Theory and Experiment*, **2022**, Article 083405. https://doi.org/10.1088/1742-5468/ac841d

[4] Huang, F., Gao, S., Pei, J., *et al*. (2022) Accelerated Zeroth-Order and First-Order Momentum Methods from Mini to Minimax Optimization. *Journal of Machine Learning Research*, **23**, 1616-1685.

[5] Shani, L., Efroni, Y. and Mannor, S. (2020) Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5668-5675. https://doi.org/10.1609/aaai.v34i04.6021

[6] Krutikov, V., Tovbis, E., Stanimirović, P. and Kazakovtsev, L. (2023) On the Convergence Rate of Quasi-Newton Methods on Strongly Convex Functions with Lipschitz Gradient. *Mathematics*, **11**, Article 4715. https://doi.org/10.3390/math11234715

[7] Glowinski, R. and Marroco, A. (1975) Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-Dualité d'une classe de problèmes de Dirichlet non linéaires. Revue française d'automatique, informatique, recherche opérationnelle. *Analyse Numérique*, **9**, 41-76. https://doi.org/10.1051/m2an/197509r200411

[8] Gabay, D. and Mercier, B. (1976) A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximation. *Computers & Mathematics with Applications*, **2**, 17-40. https://doi.org/10.1016/0898-1221(76)90003-1

[9] Bertsekas, D.P. (2014) Constrained Optimization and Lagrange Multiplier Methods. Academic Press.

[10] Jakovetic, D., Bajovic, D., Xavier, J. and Moura, J.M.F. (2020) Primal-Dual Methods for Large-Scale and Distributed Convex Optimization and Data Analytics. *Proceedings of the IEEE*, **108**, 1923-1938. https://doi.org/10.1109/jproc.2020.3007395

[11] Ma, S. (2015) Alternating Proximal Gradient Method for Convex Minimization. *Journal of Scientific Computing*, **68**, 546-572. https://doi.org/10.1007/s10915-015-0150-0

[12] Chen, C., He, B., Ye, Y. and Yuan, X. (2014) The Direct Extension of ADMM for Multi-Block Convex Minimization Problems Is Not Necessarily Convergent. *Mathematical Programming*, **155**, 57-79. https://doi.org/10.1007/s10107-014-0826-5

[13] Lin, T., Ma, S. and Zhang, S. (2017) Global Convergence of Unmodified 3-Block ADMM for a Class of Convex Minimization Problems. *Journal of Scientific Computing*, **76**, 69-88. https://doi.org/10.1007/s10915-017-0612-7

[14] Wang, Y., Yin, W. and Zeng, J. (2018) Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. *Journal of Scientific Computing*, **78**, 29-63. https://doi.org/10.1007/s10915-018-0757-z

[15] Chao, M.T., Zhang, Y. and Jian, J.B. (2020) An Inertial Proximal Alternating Direction Method of Multipliers for Nonconvex Optimization. *International Journal of Computer Mathematics*, **98**, 1199-1217. https://doi.org/10.1080/00207160.2020.1812585

[16] Liavas, A.P. and Sidiropoulos, N.D. (2015) Parallel Algorithms for Constrained Tensor Factorization via Alternating Direction Method of Multipliers. *IEEE Transactions on Signal Processing*, **63**, 5450-5463. https://doi.org/10.1109/tsp.2015.2454476

[17] Shen, Y., Wen, Z. and Zhang, Y. (2012) Augmented Lagrangian Alternating Direction Method for Matrix Separation Based on Low-Rank Factorization. *Optimization Methods and Software*, **29**, 239-263.

https://doi.org/10.1080/10556788.2012.700713

[18] Mai, V. and Johansson, M. (2020) Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization. *Proceedings of the* 37*th International Conference on Machine Learning*, Online, 13-18 July 2020, 6630-6639.

[19] Emmert-Streib, F. and Dehmer, M. (2019) High-Dimensional Lasso-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Machine Learning and Knowledge Extraction*, **1**, 359-383. https://doi.org/10.3390/make1010021

[20] Avron, H., Clarkson, K.L. and Woodruff, D.P. (2017) Faster Kernel Ridge Regression Using Sketching and Preconditioning. *SIAM Journal on Matrix Analysis and Applications*, **38**, 1116-1138. https://doi.org/10.1137/16m1105396

[21] Zhong, W. and Kwok, J. (2014) Gradient Descent with Proximal Average for Nonconvex and Composite Regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**, 2206-2212. https://doi.org/10.1609/aaai.v28i1.8994

[22] Li, X., Ding, S. and Li, Y. (2017) Outlier Suppression via Non-Convex Robust PCA for Efficient Localization in Wireless Sensor Networks. *IEEE Sensors Journal*, **17**, 7053-7063. https://doi.org/10.1109/jsen.2017.2754502