Vol. 39 (2019)
No. 6

数 学 杂 志
J. of Math. (PRC)

稀疏正则非凸优化问题之全局收敛分析

储 敏
(武汉大学数学与统计学院, 湖北 武汉 430072)

摘 要: 本文研究了一类稀疏正则化的非凸优化问题. 利用近端梯度法, 获得了其全局收敛的结果, 推广了算法模型在神经网络训练中的应用.

关键词: 非凸组合优化; 稀疏正则化; 近端梯度
MR(2010) **主题分类号**: 49J30 **中图分类号**: O224
文献标识码: A **文章编号**: 0255-7797(2019)06-0852-07

1 引言

近年来, 随着数据量的加大和计算机性能的急速提升, 极大地促进了以机器学习为主导的人工智能技术研究. 然而, 当前应用数学家所关心的是如何把实际的问题进行数学上的刻画, 并且求出其显示解或者数值解. 以目前最流行的深度神经网络为例, 在训练集上, 我们可以把它归纳为一个非凸非光滑的优化问题 [1]. 同样地, 在矩阵分解以及张量填充中, 其目标函数也是非凸的. 另一方面, 由于大数据的高维特性 (观测样本量个数小于人们关心的属性的维数), 使得很多传统的数学工具、统计方法不再有效, 对所观测到的大数据本身作更好的先验假设, 则是有效处理大数据的关键. 幸运的是, 大多数的实际问题中造成某种结果的影响因素有可能有很多, 但是真正有显著影响的因素实际上很少, 只需要很少的某些属性就能较好的满足于表征我们所关心的这些事物, 反映到数学思想方法上, 稀疏性这个合理的先验假设给处理大数据问题打开了一扇窗. 例如, 在图像处理领域, 近些年的发展很大程度上得益于提出: "自然图像可以在某些变换下稀疏表示"这样一个合理的假设 [2]. 又例如, 在日常生活中, 一个人的健康指标通常只采用由少数的生物指标来反映. 由此, 寻求稀疏解不仅符合问题本身的需求同时也有益于节省存储成本.

考虑如下非凸组合优化问题

$$\begin{cases} \text{minmize } F(x) := f(x) + r(x), \\ \text{subject to } \ x \in X, \end{cases} \tag{1.1}$$

其中 X 是欧式空间 R^d 上的凸的紧集, $f : X \to R$ 是一个光滑非凸函数, $r : X \to R$ 是一个凸的但非光滑的正则化项. 若: 稀疏 l_1 正则化 [3], 问题 (1.1) 涵盖了一系列非凸组合优化问题.

*收稿日期: 2018-11-11 接收日期: 2019-03-25
基金项目: 国家自然科学基金委员会 (61179039); 国家重点基础研究发展规划项目 (973 计划项目) (2011CB707100).
作者简介: 储敏 (1993–), 女, 安徽六安, 硕士, 主要研究方向: 图像处理与模式识别.

例 1 给定一个 n 维序列 $(a_1,b_1),\cdots,(a_n,b_n)$, 其中 $a_i\in R^d, b_i\in R$, 若令 $f(x)=\sum\limits_{i=1}^{n}(\varphi(a_i^Tx-c)-b_i)$, $r(\cdot)\equiv 0$, 其中 c 是偏差, φ 是 Sigmoid 函数, 即 $\varphi(t)=\frac{1}{1+\exp(-t)}$, 那么问题 (1.1) 即化为感知机问题（非凸）; 若令 $f(x)=\sum\limits_{i=1}^{n}(\langle a_i,x\rangle-b_i)^2, r(\cdot)=\lambda\|\cdot\|_1,\lambda>0$ 在函数 f 和正则化项 r 间起到了平衡的作用, 这时问题 (1.1) 即为著名的 Lasso[3].

2 预备知识

对于实值函数 $f:X\to R\cup\{+\infty\}$, f 的定义域 dom $f:=x\to X:f(x)<+\infty$; f 为正常函数, 即 dom$f\neq\emptyset$ 且 $f\neq-\infty$; f 为闭函数, 即 f 是下半连续的.

定义 2.1 [4] 给定一个正常函数 $f:X\to R\cup\{+\infty\}$, 对每个 $x\in\mathrm{dom}(f)$, f 在 x 处的 Fréchet 次微分记为 $\hat{\partial}f(x)$, 其定义为

$$\hat{\partial}f(x):=\{x^*\in X^*|\liminf_{\substack{z\to x\\ z\neq x}}\frac{1}{||x-z||}[f(z)-f(x)-\langle x^*,z-x\rangle]\geq 0\}. \tag{2.1}$$

定义 2.2 [5] 给定一个正常函数 $f:X\to R\cup\{+\infty\}$, 对每个 $x\in\mathrm{dom}(f)$, f 在 x 处的次微分记为 $\partial f(x)$, 其定义为

$$\partial f(x):=\{x^*\in X^*|\exists x_k\to x, f(x_k)\to f(x), x_k^*\in\hat{\partial}f(x_k), x_k^*\to x^*\}. \tag{2.2}$$

定理 2.1 [5] 令 $J(x,z):=H(x,z)+f_1(x)+f_2(x)$, 其中 $f_1:X\to R\cup\{+\infty\}$ 是一个正常的下半连续的凸函数, $f_2:X\to R\cup\{+\infty\}$ 是一个正常的连续可微函数, H 也是连续可微函数. 那么 $\forall(x,z)\in X\times X$, 有

$$\partial J(x,z)=(\bigtriangledown_x H(x,z)+\partial f_1(x),\bigtriangledown_z H(x,z)+\bigtriangledown f_2(z))=(\partial_x J(x,z),\bigtriangledown_z J(x,z)). \tag{2.3}$$

定义 2.3 [5] f 的临界点 $\{x|0\in\partial f(x)\}$, 满足 $\min\limits_x f$ 的必要非充分条件.

定义 2.4 [6] (KL 函数) (a) 设 $x^{'}\in$ dom$\partial f,\zeta\in[0,+\infty)$, 若存在 $x^{'}$ 的某个邻域 U, 连续凹函数 $\varphi:[0,\zeta)\to R_+$ 满足

(i) $\varphi(0)=0$;

(ii) φ 在 $(0,\zeta)$ 上是一阶连续可导的;

(iii) 任意 $z\in(0,\zeta)$, $\varphi^{'}(z)>0$;

(iv) 任意 $x\in U\cap[f(x)<f(x^{'})<f(x)+\zeta]$, 有 Kurdyka–Lojasiewicz 不等式成立

$$\varphi^{'}(f(x^{'})-f(x))\,\mathrm{dist}\,(0,\partial f(x^{'}))\geq 1, \tag{2.4}$$

则称 f: $R^n\to R\cup+\infty$ 在 x^* 满足 Kurdyka-Lojasiewicz 性质 [6].

(b) 在 dom∂f 内每个点都满足 Kurdyka-Lojasiewicz 不等式的正常下半连续函数, 称为 KL 函数.

3 模型及收敛性分析

首先, 纵观全文对函数 f 和 g 做如下假设.

(i) f 是 Lipschitz 连续可微函数, Lipschitz 常数 $L>0$, 即 $\forall x,y\in X$ 都有 $||\bigtriangledown f(x)-\bigtriangledown f(y)||\leq L||x-y||$;

(ii) f 和 g 是非负、正常、强制、半代数函数.

基于以上假设, 给出如下近端梯度算法 [7]

表 1: 近端梯度算法

算法 1 近端梯度算法
1. 输入: 序列的步长为 $\{\eta_k\}$;
2. 初始化: $x_1=0$;
3. 当 $k=1,2,\ldots$ 令
$x_{k+1}=\arg\min\limits_{x\in X}\{\langle x,\bigtriangledown f(x_k)\rangle+\frac{1}{2\eta_k}||x-x_k||^2+r(x)\}$;
4. 结果

在这个部分, 分析算法 1 的收敛性. 有必要先对算法 1 中的序列 $\{x_k\}$ 的特性进行分析.

引理 3.1 假设 (i) 成立且 $\eta_k<\frac{1}{L}$, $\forall k\geq 1$, 算法 1 产生的序列 $\{x_k\}$ 满足

(i) 存在两个常量, 即 $0<\mu\leq\frac{1}{\eta_k}-\frac{L}{2}$, 使得

$$F(x_{k+1})+\mu||x_{k+1}-x_k||\leq F(x_k). \tag{3.1}$$

(ii)

$$\lim_{k\to\infty}||x_{k+1}-x_k||^2=0. \tag{3.2}$$

(iii) 令 $x_k^*\in\partial F(x_k)$, 则对 $\{x_k\}$ 的所有有界子序列 $\{x_{k_i}\}$, 都有当 $i\to\infty$, 有 $x_{k_i}^*\to 0$, 即

$$\mathrm{dist}(\partial F(x_{k_i}),0)\to 0, \text{当 } i\to\infty. \tag{3.3}$$

证 (i) 首先定义如下函数

$$\Phi_k(x):=f(x_k)+\langle\bigtriangledown f(x_k),x-x_k\rangle+\frac{1}{2\eta_k}||x-x_k||^2+r(x). \tag{3.4}$$

则算法 1 中的步骤 4 可以表示为 $x_{k+1}=\arg\min\limits_{x\in X}\Phi_k(x)$, 由 Φ_k 的 $\frac{1}{\eta_k}$ 强凸性, 得到

$$\Phi_k(x_{k+1})-\Phi_k(x_k)\leq-\frac{1}{2\eta_k}||x_{k+1}-x_k||^2. \tag{3.5}$$

进行化简后可得到

$$f(x_k)+\langle\bigtriangledown f(x_k),x_{k+1}-x_k\rangle+r(x_{k+1}\leq F(x_k)-\frac{1}{\eta_k}||x_{k+1}-x_k)||^2. \tag{3.6}$$

利用 f 的 Lipschitz 连续可微性, 得到

$$\begin{aligned}&f(x_{k+1})-\frac{L}{2}||x_{k+1}-x_k||^2+r(x_{k+1})\\&\leq f(x_k)+\langle\bigtriangledown f(x_k),x_{k+1}-x_k\rangle+r(x_{k+1})\\&\leq F(x_k)-\frac{1}{\eta_k}||x_{k+1}-x_k||^2.\end{aligned}$$

从而 (i) 得证.

对于 (ii), 将上 (3.1) 式两边同时进行求和, 得到

$$\sum_{k=1}^{\infty}\mu||x_{k+1}-x_k||^2 \leq F(x_1) < +\infty. \tag{3.7}$$

从而 (ii) 得证.

对于 (iii), 由 $\partial F(x)$ 的定义, 令

$$x_{k+1}^* := \bigtriangledown f(x_{k+1}) + \partial r(x_{k+1}). \tag{3.8}$$

另一方面, 由算法 1 的一阶优化条件, 得到

$$0 \in \bigtriangledown f(x_k) + \frac{1}{\eta_k}(x_{k+1}-x_k) + \partial r(x_{k+1}), \tag{3.9}$$

化简得到

$$x_{k+1}^* = \bigtriangledown f(x_{k+1}) - \bigtriangledown f(x_k) - \frac{1}{\eta_k}(x_{k+1}-x_k). \tag{3.10}$$

由 (ii) 中的不等式 (3.2), (iii) 得证.

为了证明算法 1 的收敛性, 还需要如下定理.

定理 3.1 [8−10] 假设 (i) 成立且 $\eta_k < \frac{2}{L}$, $\{x_k\}$是算法 1 产生的序列, 则 $\{x_k\}$收敛到 F 的临界点 $\mathring{x}$.

证 为了证明算法 1 的收敛性, 首先要证明以下三个条件.

(H1) (充分下降条件) $\forall k > 0$, 存在 $a > 0$, $F(x_k) - F(x_{k+1}) \geq a||x_{k+1}-x_k||^2$;

(H2) (相对误差条件) $\forall k > 0$, 存在 $b > 0$, 存在 $x_{k+1}^* \in \partial F(x_{k+1})$ 使得

$$||x_{k+1}^*|| \leq b||x_{k+1}-x_k||;$$

(H3) (连续条件) 存在子列 x_{k_i} 和聚点 $\mathring{x}$ 使得 当$i \to +\infty$,有 $x_{k_i} \to \mathring{x}$ 且 $F(x_{k_i}) \to F(\mathring{x})$.

事实上, 令 $a = \mu$, (H1) 很容易由引理 3.1 得出. 令 $b = \frac{1}{\eta_k} + \frac{2}{L}$, (H2) 易由引理 3.1 得出. 下面证明 (H3).

由 $F(x)$ 的强制性, 知道 $\{x_k\}$ 包含在水平集 $\{x_k \in X : F(x_k) \leq F(x_1)\}$ 中, 利用 Bolzano-Weierstrass 定理, 得出存在子集记为 x_{k_i} 收敛到某个聚点 $\mathring{x}$. 由 x_{k+1} 的定义有

$$F(x_{k_{i+1}}) + (\frac{1}{\eta_k} - \frac{L}{2})||x_{k_{i+1}}-x_{k_i}||^2 \leq \Phi_{k_i}(\mathring{x}).$$

又由 Φ_k 的定义, 有 $\lim\limits_{i\to\infty}\Phi_{k_i}(\mathring{x}) = F(\mathring{x})$, 由上可得 $F(x_{k_{i+1}}) \leq F(\mathring{x})$.

一方面, 由 F 的连续性, 得到 $\lim\limits_{i\to+\infty}\sup F(x_{k_i}) \leq F(\mathring{x})$, 其中 $\{x_{k_i}\}$ 是收敛到 $\mathring{x}$ 的序列, 由引理 3.1, 得到 $\{x_{k_{i+1}}\}$ 也收敛到 $\mathring{x}$. 另一方面, 由 $F(\cdot)$ 的下半连续性, 得到 $\lim\limits_{i\to+\infty}\inf F(x_{k_i}) \geq F(\mathring{x})$, 于是可以得到: 存在一个子列 $\{x_{k_i}\}$ 收敛到 $\mathring{x}$, 且当 $i \to +\infty$, $F(x_{k_i}) \to F(\mathring{x})$. (H3) 得证.

回到算法 1 的收敛性证明, 知道 $F(x)$ 是半代数的, 且是一个 KL 函数, 由 KL 函数的性质 (见定义 2.4), 存在 $\zeta > 0$, $\mathring{x}$ 的邻域 $\mathcal{V}$ 和一个连续凹函 $\varphi : [0, \zeta) \to R_+$, 对所有的 $x \in \mathcal{V}$, 有

$$F^* < F(x) < F^* + \zeta,$$

其中 $F^* := F(\mathring{x})$.

取 $r > 0$, 则 $\mathcal{B}_r(\mathring{x}) \subseteq \mathcal{V}$. 已知存在子列 $\{x_{k_i}\}$ 收敛到 $\mathring{x}$, 则意味着存在一个 x_{k_N}, 使得

(a) $x_{k_N} \in \mathcal{V}$;

(b) $F^* < F(x_{k_N}) < F^* + \zeta$;

(c) $||x_{x_N} - \mathring{x}|| + 2\sqrt{\frac{\bigtriangledown F^*}{b}} + \frac{a}{b}(\bigtriangledown J^*) < r$, 其中 $\bigtriangledown J^* := F(x_{k_N}) - F^*$.

通过 (H1), (H2), (H3), (a), (b) 和 (c), 利用文献 [10] 的定理 2.9, 可以得到 $\{x_k\}$ 收敛到 $\mathring{x}$.

最后, 由引理 3.1 的 (iii), 可知 $\mathring{x}$ 是 JF 的一个临界点. 算法 1 的收敛性得证.

4 数值试验

考虑 (1.1) 优化问题, 我们通过设计加 $L1, L2$ 正则化项的神经网络做分类试验, 来验证算法的有效性.

神经网络 [11] 的模型如图 1 所示, 一个神经元对输入信号 $X = [x_1, x_2, \ldots, x_n]$ 的输出为 $y = f(u + b)$, 其中 $u = \sum_{i=1}^{n} w_i x_i$, 公式中各字符含义如图 1 所示. 神经网络的训练通常用误差函数 (也称目标函数)E 来衡量, 当误差函数小于设定的值时即停止神经网络的训练. 误差函数为衡量实际输出向量 Y_k 与期望向量 T_k 误差大小的函数, 常采用二乘误差函数来定义为 $E = \frac{1}{2}\sum_{k=1}^{n}[Y_k - T_k]^2$, $k = 1, 2, \ldots, n$ 为训练样本个数. 在模型训练时, 如果参数过多, 模型过于复杂, 容易造成过拟合 (overfit), 即模型在训练样本数据上表现得很好, 但在实际测试样本上表现得较差, 不具备良好的泛化能力. 为了避免过拟合, 最常用的一种方法是使用使用正则化, 例如 $L1$ 和 $L2$ 正则化, 其中 $L1$ 正则化产生更加稀疏的权值. 在误差函数的基础上加正则化项后的损失函数为 $F = E + \lambda\|\cdot\|_i, \lambda > 0, i = 0, 1, 2\cdots$.

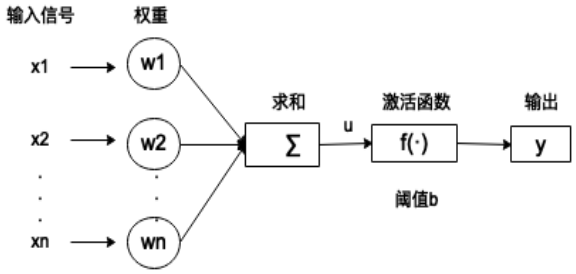

图 1: 人工神经元模型

本试验中, 选取 Sigmoid 函数作为激活函数, 即 $\varphi(t) = \frac{1}{1+\exp(-t)}$, 分别加 $L1$ 和 $L2$ 正则化进行对比试验, 加 $L1$ 正则化的损失函数即为本文讨论的非凸组合优化问题, 我们采用近端梯度下降算法对模型进行训练; 加 $L2$ 正则化的损失函数我们采用梯度下降算法对模型进行训练. 对两组组试验数据集上, 得到试验结果如下

表 2: $L1+$ 近端梯度下降算法与 $L2+$ 梯度下降算法分类错误率

模型	BTEAST_CANCER	DIGITS
$L1$+PG	0.119	0.059
$L2$+GD	0.225	0.075

在不同数据集上, 采用 $L1$+PG 和 $L2$+GD 的模型进行神经网络训练, 可以看到 $L1$+PG 模型的分类错误率均低于 $L2$+GD 模型, 通过损失曲线的对比, 可以看出 $L1$+PG 比 $L2$+GD 模型的训练更加快速达到收敛.

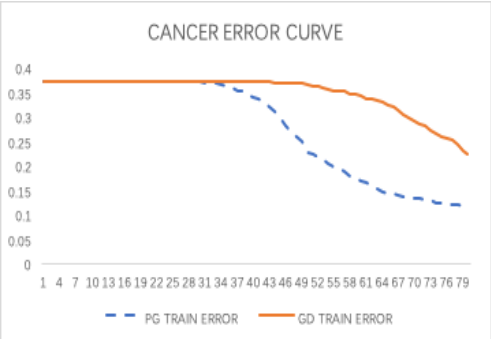

图 2: CANCER 数据集上的错误率曲线

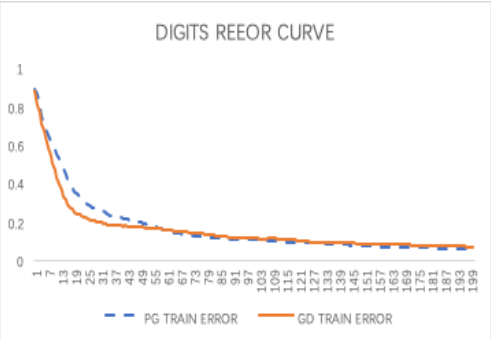

图 3: DIGITS 数据集上的错误率曲线

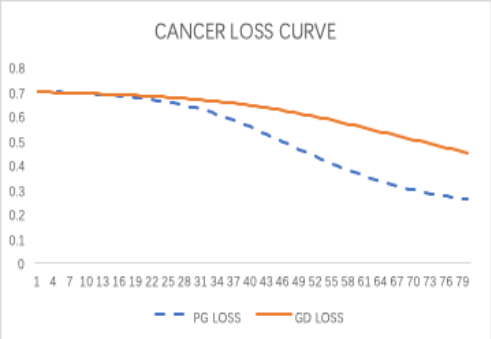

图 4: CANCER 数据集上的损失曲线

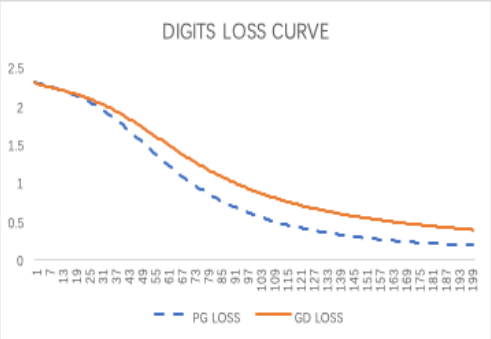

图 5: DIGITS 数据集上的损失曲线

参 考 文 献

[1] Cui Zhuoxu, Fan Qibin. A “nonconvex+ nonconvex” approach for image restoration with impulse noise removal [J]. Applied Mathematical Modelling, 2018, 62: 254–271.

[2] Cui Zhuoxu, Fan Qibin. A nonconvex nonsmooth regularization method for compressed sensing and low rank matrix completion [J]. Digital Signal Processing, 2017, 62: 101–111.

[3] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. Journal of the Royal Statistical Society Series B, 1996, 58: 267–288.

[4] 肖瑾. 非凸函数在图像复原中的应用 [D]. 长沙: 湖南大学, 2015.

[5] Boyd S, Vandenberghe L. Convex optimization[M]. New York: Cambridage Univ. Press, 2004.

[6] Bolte J, Daniilidis A. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems [J]. SIAM J. Optim., 2007, 17: 1205–1223.

[7] Aleksandr D, Vladimir T. Theory of extremal problems, studies in mathematics and its applications[M]. New York: Oxford, 2006.

[8] Attouch H, Bolte J, Svaiter B. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods [J]. Math. Program, 2013, 137: 91–129.

[9] Kruger A. Fréchet subdifferential calculus and optimality conditions in nondifferentiable programming [J]. Journal of Mathematical Sciences, 2003, 116: 3325–3358.

[10] Chieu N. The Fréchet and limiting subdifferentials of integral functionals on the spaces $L1(\Omega, E)$ [J]. J. Math. Anal. Appli., 2009, 360: 704–710.

[11] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.

GLOBAL CONVERGENCE ANALYSIS OF SPARSE REGULAR NONCONVEX OPTIMIZATION PROBLEMS

CHU Min

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: In this paper, we consider a class of sparse regularization nonconvex optimization problems. By using the proximal gradient method, we obtain the global convergence results, which generalize application of algorithm models in neural network training.

Keywords: nonconvex composite optimization; sparse regularization; proximal gradient

2010 MR Subject Classification: 49J30