arXiv:2602.13278v1 [cs.CC] 6 Feb 2026

p进线性回归的 NP 困难性

Greg Baker

澳大利亚国立大学，堪培拉，澳大利亚

摘要

p进线性回归是寻找系数 β 使其最小化$\sum_i |y_i - x_i^\top \beta|_p$的问题。我们利用正则化构件，通过从最大割进行多项式时间归约，证明了计算最优解是 NP 困难的。

关键词： p进分析、线性回归、计算复杂度、NP 困难性、最大割

1. 引言

p进数由 Hensel 于1897年引入，是相对于非阿基米德绝对值对有理数进行的一种替代性完备化。尽管它们在数论和代数几何中的应用由来已久，近期研究也探索了它们在机器学习中的用途，其中 p进空间的超度量结构为分层数据和树结构数据 [1, 2, 3, 4, 5, 6]提供了自然的表示方式。

这一领域的一项基本任务是 p进线性回归：给定数据点 $(x_i,\ y_i)$，其中 $x_i \in \mathbb{Q}^n$ 或 $\mathbb{Z}^n$ 且 $y_i \in \mathbb{Q}$ 或 $\mathbb{Z}$，求使下式最小化的系数 $\beta \in \mathbb{Q}^n$

$$L(\beta) = \sum_{i=1}^{m} |y_i - x_i^\top \beta|_p,$$

其中 $|\cdot|_p$ 表示 p进绝对值。

据我们所知，目前还没有已知的例子表明，p-进线性回归是作为预言机（黑箱）子程序问题出现的。我们所了解的 p-进方法在现实世界中的应用，反而是直接被表述为

电子邮箱地址： greg.baker@anu.edu.au (Greg Baker)

优化问题——例如，语法形态学 [6] 以及与本体相关的优化问题 [7]。因此，我们针对 p-进线性回归的优化版本给出了复杂性结果。

这也凸显了它与经典回归之间的鲜明对比：在欧几里得空间中，最小二乘回归有闭式解，而 ℓ_1回归可通过凸优化在多项式时间内求解。这里，仅将损失改为超度量 $|\cdot|_p$，就会得到一个 NP 困难的优化问题。

这个问题与经典最小二乘回归有着根本区别。超度量不等式 $|a+b|_p \le \max(|a|_p,\ |b|_p)$ 在损失景观中形成了离散的层次结构。先前的工作 [7] 证明，在非退化条件下，最优解至少经过 $n+1$ 个数据点，从而为固定维数情形带来了多项式时间算法。

本文通过证明当维数作为输入的一部分时，p-adic 线性回归是 NP 困难的，从而补全了复杂性图景（第 3 节）。

2. 预备知识

2.1. p-*adic* 绝对值

对于一个素数 p和一个非零有理数 $x = p^k \cdot (a/b)$，其中 $\gcd(ab,\ p) = 1$，p-*adic* 赋值为 $v_p(x) = k$，而 p-*adic* 绝对值为 $|x|_p = p^{-v_p(x)}$。按照惯例，$|0|_p = 0$。

p进绝对值满足超度量不等式：

$$|x+y|_p \le \max(|x|_p, |y|_p),$$

当 $|x|_p \neq |y|_p$时取等号。这种非阿基米德性质使p进分析区别于实分析，也是我们结果的核心。

2.2. 问题定义

定义 1（p进线性回归）。

给定数据点 (x_1, y_1).....，(x_m, y_m)，其中 $x_i \in \mathbb{Z}^n$（或 $\mathbb{Q}^n$）和 $y_i \in \mathbb{Z}$（或$\mathbb{Q}$），以及一个素数 p，求使其最小化的 $\beta \in \mathbb{Q}^n$

$$\sum_{i=1}^{m} |y_i - x_i^\top \beta|_p.$$

定义 2（最大割）。 给定一个无权图 $G = (V, E)$，将 V 划分为集合 S 和 T，使一个端点在 S、另一个端点在 T的边数最大。

最大割是 NP 困难 [8, 9]。将顶点着色为两种颜色之一，并将边描述为单色边或异色边，会比较方便。一种颜色的顶点归为一组，另一种颜色的顶点归为另一组。

3. p进回归的 NP 困难性

我们将最大割归约到 2进线性回归。

定理3。 p进线性回归是 *NP* 困难的（即使对于 $p = 2$ 也是如此）。

证明。 设 $G = (V, E)$ 是一个具有 $|V| = n$ 个顶点和$|E| = m$ 条边的无权图。我们构造一个含有 n 个系数β_1, ..., β_n 的回归实例（每个顶点对应一个系数）。

边点。 对于每条边 $(u, v) \in E$，添加一个点 $(e_u + e_v, 1)$，其中 e_u 是第 u 个标准基向量。这些点对残差$|1 - \beta_u - \beta_v|_2$作出贡献。

注意，如果 $\beta_u = \beta_v = 1$，那么残差为 $|-1|_2 = 1$；同样，如果 $\beta_u =\beta_v = 0$，那么残差为 $|1|_2 = 1$。因此，如果 β_u 且 $\beta_v \in \{0, 1\}$，残差就等于单色边的数量。

遗憾的是，目前还没有任何约束能保证 β_u 为 $\{0, 1\}$。为此，我们加入正则化。

设 $M := m + 1$。我们创建以下数据点：

强制点。 对于每个顶点 $j \in \{1, \ldots, n\}$，添加 M 个 $(e_j, 0)$ 的副本，以及 M 个 $(e_j, 1)$ 的副本。这些分别贡献残差 $|\beta_j|_2$ 和 $|\beta_j - 1|_2$。

这是欧几里得线性回归问题中正则化的 p-进制对应形式：将系数值的某个倍数纳入损失中。

那么，我们模型的总损失为：

$$L(\beta) = M \sum_{j=1}^{n} \big(|\beta_j|_2 + |\beta_j - 1|_2\big) + \sum_{(u,v)\in E} |\beta_u + \beta_v - 1|_2.$$

我们将证明，使该损失最小化的 n 最优系数（β）对应于最大割问题中 n 顶点的着色。

为此，我们需要证明 $\beta_j \in \{0,1\}$ 对所有 j 都成立（因此它们对应于有效的着色），并且最小损失也是最大割的最大化解。

我们证明四个引理。

引理4. 对于上述给出的 *p-adic* 线性回归问题，任何导致损失等于或大于 $M(n+1)$ 的解都不可能是最优解。

证明. 先考虑一个朴素的解： β^0 (其中 $\beta_j = 0$ 对所有 j都成立)，则可以计算出

$$L(\beta^0) = M\sum_{j=1}^{n}(|0|_2 + |-1|_2) + \sum_{(u,v)\in E}|-1|_2 = Mn + m$$

因此，任何导致损失大于 $Mn+m$的解都不可能是最优解。由于 $m < M$，我们还可以进一步断言，任何损失等于或大于 $M(n+1)$ 的解都不可能是最优解。

□

引理5. 对于任意 $a \in \mathbb{Q}$，都有 $|a|_2 + |a-1|_2 \geq 1$，当且仅当 $a \in \{0,1\}$ 时等号成立。

证明. 由于 $|1|_2 = |a-(a-1)|_2 \leq \max(|a|_2, |a-1|_2)$，至少有一个项为 ≥ 1。要使求和等于 1，另一项必须为 0，因此可得 $a = 0$ 或 $a = 1$。

□

引理6. 若对于某些 β, k，有 $|\beta_k|_2 > 1$ 或 $|\beta_k - 1|_2 > 1$，则 β 不是最优解。

证明. 查看 $|\cdot|_2$ 的公式可知，不可能得到负的测度；也不存在任何满足 $1 < |a|_2 < 2$ 的 a。因此$|\beta_k|_2 \geq 2$。

代入这些不等式并使用引理5，我们得到

$$\begin{aligned} L(\beta) &\geq M\sum_{j\neq k}\big(|\beta_j|_2 + |\beta_j - 1|_2\big) + M\big(|\beta_k|_2 + |\beta_k - 1|_2\big) \\ &\geq M(n-1) + 2M \;=\; M(n+1), \end{aligned}$$

这里我们还利用了边项和非负这一点。

因此，根据引理4，这不会是一个最优解。 □

由于在最优解中，对于所有 j，我们知道 β_j 的最简形式的分母不能被 2 整除。

因此，根据 Gouvêa [10], 中的引理 4.3.2，我们可以将任意 β_j 表示为递增的 2 的幂的求和，其中每个 $a_t \in \{0,1\}$。

$$\beta_j = a_0 + 2a_1 + 4a_2 + \dots$$

注意，在 2-adically 意义下，后面的项会逐渐变小。如果我们需要用无限多项来表示 β_j （例如当它不是整数时），那么极限是良好定义的，我们可以在有限项之后截断，从而对其进行近似。

现在，我们将模 2 定义为取该级数中的第一项（a_0）。这与整数中模 2 的通常含义完全相同，但对非整数也同样有良好定义。

引理 7。 对于任意 $\beta \in \mathbb{Z}_2^n$，令 $s_j = \beta_j$ 模 $2 \in \{0,1\}$。则 $L(s) \le L(\beta)$。

证明。 根据引理5，强制项满足 $|s_j|_2 + |s_j - 1|_2 = 1 \le |\beta_j|_2 + |\beta_j - 1|_2$。

对于边 $(u,\ v)$：若 $s_u = s_v$，则 $\beta_u + \beta_v - 1$ 为奇数，因此 $|\beta_u + \beta_v - 1|_2 = 1 = |s_u + s_v - 1|_2$。若 $s_u \neq s_v$，则 $\beta_u + \beta_v - 1 \equiv 0$（mod 2），因此 $|\beta_u + \beta_v - 1|_2 \le 1/2$，而 $|s_u + s_v - 1|_2 = |0|_2 = 0$。 □

根据引理 7，$\{0,1\}^n$ 中存在一个最优解。对于 $s \in \{0,1\}^n$，边 $(u,\ v)$ 的贡献为：

- 0 计入损失，当 $s_u \neq s_v$（边跨越割）时，因为 $s_u + s_v = 1$。
- 1 计入损失，当 $s_u = s_v$（边未跨越割）时，因为 $s_u + s_v \in \{0,2\}$ 和 $|\pm 1|_2 = 1$。

Thus

$L(s) = Mn +$(不跨割边的数量) $= Mn + m -$割大小(s)。

由于 $Mn + m$ 项与 s 无关，最小化 $L(s)$ 等价于最大化 cutsize(s)。特别是，根据最优回归值 $L^\star$，我们可以恢复最大割大小 cutsize$^\star = Mn + m - L^\star$。等价地，对于任意整数 K，G 存在大小至少为 K 的割，当且仅当所构造的回归实例存在满足 $L(\beta) \le Mn + m - K$ 的 β。因此，任何最大割问题都可以转化为 p-adic 线性回归问题；如果能在多项式时间内求解 p-adic 线性回归问题，那么 $P = NP$。

□

NP 困难性是一种最坏情况分类：它并不排除针对结构化实例的快速算法，同时也促使人们研究近似算法并识别可处理情形。

4. 讨论：可处理情形与开放问题

定理3说明了当维数作为输入的一部分时的最坏情况复杂度。它补充了已知的固定维数多项式时间方法，并指出了若干影响计算行为的自然参数。

4.1. 固定维数

当维数固定时，Baker等人证明（在非退化条件下），最优解会经过至少 $n+1$ 个数据点 [7]。枚举所有此类候选超平面，可得到一个 $O(m^{n+1})$ 算法。

4.2. 对 p 和数据集的依赖

对 p 的依赖体现在残差的整除性上。对于整数残差 r，若 $r=0$，则有 $|r|_p=0$；而当 $p\nmid r$ 时，$|r|_p=1$。因此，如果数据是整数，且某个候选拟合产生的整数残差的（普通）绝对值都为 $<p$，那么每个非零残差的贡献都是 1，损失就只是非零残差的个数。在这种情况下，最小化 p-进损失等价于精确拟合：使精确插值的数据点数量最大化。在二维中，这就是共线点最大化问题；已有近二次算法 [11]，而 [7] 的算法在数据点数量上将是三次时间复杂度。

4.3. 其他聚合方式

Martins [4] 研究了采用最大值聚合而非求和的 p-adic 学习。目前，采用这种及其他聚合方式的 p-adic 回归问题的复杂度尚未解决。

4.4. 正则化逐渐减弱时会发生什么？

在第 3 节中，我们需要正则化项来约束 $\beta\in\{0,1\}^n$。该归约为每个强制点使用 $M=m+1$ 个副本，但如果没有任何强制点，构造出的实例就是平凡的：对所有 j 设置 $\beta_j=\frac{1}{2}$，即可使每条边的残差都为零。随着正则化强度减弱，计算复杂度如何变化仍是一个开放问题。

5. 结论

我们证明了，当维数作为输入的一部分时， p-进线性回归是 NP 困难的；结合已知的固定维数情形下的多项式时间算法，这就完成了对其复杂度的分类。从最大割进行的归约从根本上利用了超度量不等式：正是由于非阿基米德性质，正则化项才将 β 限制为二进制值。

尚待研究的问题包括 p-进线性回归的近似算法、复杂度如何随 p 的选择和正则化强度而变化，以及采用其他损失聚合方式（最大值与求和）时的复杂度。

参考文献

[1] F. Murtagh，论超度量性、数据编码与计算，《分类学杂志》21（2004年）167–184。[2]P.E. Bradley，芒福德树状图，《计算机杂志》51(5)（2008年）561–571。[3] A. Khrennikov、S. Kozyrev、W.A. Zúñiga-Galindo，《超度量伪微分方程及其应用》，《数学及其应用百科全书》，剑桥大学出版社，2018年。[4] A.F.T. Martins，基于 p进数的学习，arXiv:2512.22692，2025年。[5]A.P. Zubarev，以 p进多项式回归替代神经网络来逼近多变量 p进函数，arXiv:2503.23488，2025年。[6] G. Baker、D. Mollà-Aliod，数论与语言学的交汇：使用2 进度量对1497种语言的名词形态学进行建模，AACL-IJCNLP 2022会议论文集，台北，台湾，2022年。[7] G. Baker、S. McCallum、D. Pattinson， p进度量空间中的线性回归，《p 进数、超度量分析及其应用》17(4)（2025年）。

[8] R.M. Karp，组合问题之间的可约性，载于：R.E. Miller、J.W. Thatcher（编），《计算机计算的复杂性》，全会出版社，1972年，第85–103页。[9] M.R. Garey、D.S. Johnson，《计算机与难解性：NP完全性理论指南》，W.H. Freeman，1979年。[10] F.Q. Gouvêa，《p进数：导论》，第3版，Universitext，施普林格，2020年。[11] J. Chen、Q. Huang、I. Kanj、G. Xia，点线拟合问题的近似最优算法，计算几何学报13(1)（2022年）226–243。