arXiv:2302.03931v1 [stat.ML] 8 Feb 2023

# PILOT：快速线性模型树

Jakob Raymaekers，计量经济学系，马斯特里赫特大学，荷兰

Peter J. Rousseeuw，统计与数据科学系，鲁汶大学，比利时

Tim Verdonck
数学系，安特卫普大学 - imec，比利时

Ruicong Yao，统计与数据科学系，鲁汶大学，比利时

2023年2月7日

**摘要**

线性模型树是一种在叶节点中纳入线性模型的回归树。这既保留了决策树直观易懂的特点，又使其能够更好地捕捉线性关系，而标准决策树难以做到这一点。然而，现有的大多数线性模型树拟合方法都很耗时，因此无法扩展到大型数据集。此外，与标准回归树相比，这些方法更容易出现过拟合和外推问题。本文介绍 PILOT，这是一种用于线性模型树的新算法，具有快速、正则化、稳定且可解释的特点。PILOT 像经典回归树一样采用贪心方式进行训练，同时结合 $L^2$ 提升方法和一种模型选择规则，以拟合节点中的线性模型。PILOT 是 **PI**iecewise **L**inear **O**rganic **T**ree 的缩写，其中“organic”（有机）指的是不进行剪枝。与不剪枝的 CART 相比，PILOT 具有同样低的时间和空间复杂度。实证研究表明，在多种数据集上，PILOT 往往优于标准决策树和其他线性模型树。此外，我们证明了在弱假设下，PILOT 在加性模型设定中具有一致性。当数据由线性模型生成时，其收敛速率为多项式级别。

关键词：一致性、分段线性模型、回归树、可扩展算法。

# 1 引言

尽管历史悠久，CART（Breiman 等，1984）和 C4.5（Quinlan，1993）等决策树仍然是广受欢迎的机器学习工具。决策树训练速度快，需要调整的参数很少，而且最终模型易于解释和可视化，这在实践中是一项颇具吸引力的优势。因此，这些方法广泛应用于工程学（Shamshirband 等，2020）、生物信息学（Stiglic 等，2012）、农业（Boryan 等，2011）和商业分析（Delen 等，2013；Noulas 等，2012）等多个领域。除了单独使用之外，决策树也广泛应用于集成方法中，通常是现有最佳的“弱学习器”。典型的例子包括随机森林（Breiman，2001）以及梯度提升方法，例如 XGBoost（Chen 和 Guestrin，2016）和 LightGBM（Ke 等，2017）。在本研究中，我们假设目标变量是连续的，因此重点关注回归树，而非分类。

经典回归树的一个局限是，其分段常数的特性使其不适合捕捉连续关系。为了逼近线性函数，它们需要进行大量分裂，这并不可取。解决这一问题主要有两种方法。第一种是使用决策树集成，例如随机森林或梯度提升树。这类集成能够平滑预测，因此可以对预测变量与响应变量之间更连续的关系进行建模。此类集成的一个缺点是可解释性降低。组合多个回归树后，模型不再能通过简单的可视化来呈现，也无法清晰地展示从预测变量到预测结果的逐步路径。捕捉连续关系的第二种方法是使用模型树。模型树具有类似树的结构，但允许在树的叶节点中进行非恒定拟合。因此，模型树在保留经典回归树直观可解释性的同时，也更加灵活。可以说，最直观且最常见的模型树是线性模型树，它允许在叶节点中使用线性模型。

据我们所知，第一个线性模型树算法由 Friedman（1979年）提出。我们将其简称为 FRIED。该算法在每个节点使用单变量分段线性拟合，取代 CART 中的分段常数拟合。在某个节点上拟合模型后，其残差会传递给子节点，以进行进一步拟合。最终预测值由沿路径的线性模型之和给出。我们的实验表明，这种方法可能存在过拟合和外推问题。与其更复杂的后继方法 MARS（Friedman，1991年）相比，FRIED 算法受到的关注要少得多。

M5算法（Quinlan，1992年）无疑是最流行的线性模型树，至今仍广泛使用（Agrawal 和 Choudhary，2016年；Golbeck 等，2011年；Noulas 等，2012年；Quercia 等，2011年；Shamshirband 等，2020年）。它首先拟合 CART。构建好这棵树后，再在叶节点处引入线性回归。随后进行剪枝和平滑，以降低其泛化误差。对该算法的一种潜在质疑是：构建树结构时，完全没有考虑到叶节点将使用线性模型这一事实。

GUIDE算法（Loh，2002）在每个节点上对数值型预测变量拟合多元线性模型，然后进行 $\chi^2$ 检验，比较正残差与负残差，以决定在哪个预测变量上进行分裂。该算法还可用于检测预测变量之间的交互作用。然而，没有理论保证能够证明这一分裂过程的合理性。

还有一些算法在分裂规则中采用了聚类的思想。例如，SECRET（Dobra 和 Gehrke，2002）使用期望最大化算法将数据拟合为两个高斯簇，并根据数据与这些簇的接近程度，在局部将回归问题转化为分类问题。实验结果并不支持该方法优于 GUIDE，而且其计算成本很高。

SMOTI 方法（Malerba 等，2004年）使用两种类型的节点：回归节点和分裂节点。在每个叶节点中，最终模型是对其“活跃”变量进行多元回归拟合得到的。这是通过从响应变量和其他变量中同时“回归剔除”回归节点对应的变量来实现的。所得拟合效果相当出色，但该算法的复杂度为 $\mathcal{O}(n^2p^3)$，对于大型数据集而言可能难以承受。因此，该方法不太适合扩展到随机森林或梯度提升。

上述大多数算法都需要剪枝程序来确保其泛化能力，而剪枝通常耗时较长。一个例外是 LLRT（Vogel 等，2007年），它采用逐步回归，并通过 $k$ 折交叉验证评估每个节点中的模型。为降低计算成本，该算法使用预测变量的分位数作为潜在分裂点。它还保留用于在两个子节点上拟合线性模型的数据矩阵，以便对其进行更新。遗憾的是，LLRT 的时间复杂度相当高，为 $\mathcal{O}(knp^3 + np^5)$，其中 $k$ 是树的深度。

比较线性模型树集成与经典决策树集成也很有意思。最近，Shi 等人（2019年）将分段线性树与梯度提升相结合。对于每棵树，他们采用了 Friedman（1979年）提出的加性拟合，或采用半加性拟合，即将此前的预测乘以一个因子。为了控制树的复杂度，他们设置了一个调优参数，限制每条路径上可使用的预测变量的最大数量。实证结果表明，在多种数据集上，该方法在 XGBoost（Chen 和 Guestrin，2016年）和 LightGBM（Ke 等人，2017年）中的表现优于经典决策树，且所需的训练迭代次数更少。这表明线性模型树在集成中具有潜在优势。

总而言之，现有线性模型树的主要问题在于计算成本高。对叶节点和／或内部节点进行多元回归拟合的方法，会使其时间复杂度增加一个因子 $p^2$ 或 $p^3$ 。而采用简单线性拟合的方法（如 FRIED）仍需进行剪枝，这在大型数据集上同样代价高昂。此外，这些方法可能产生较大的外推误差（Loh 等人，2007年）。最后，据我们所知，文献中尚无线性模型树的理论依据。

针对这一系列棘手问题，我们提出了一种新颖的线性模型树算法。其缩写 PILOT 代表 **PI**ecewise **L**inear **O**rganic **T**ree（分段线性有机树），其中“有机”指的是该算法不进行剪枝。PILOT 的主要特点如下：

- **速度**：其时间复杂度与 CART 相同，且无需进行剪枝。
- **正则化**：在每个节点中，都会对潜在线性模型应用模型选择程序。此过程不会增加额外的计算复杂度。
- **可解释**：得益于叶节点中线性模型的简洁性，最终的树仍具有很强的可解释性。此外，还可以计算特征重要性。

- **稳定外推**：我们采用两种截断程序，以避免训练数据上出现极端拟合，以及测试数据上产生较大的外推误差；这些都是线性模型树中常见的问题。

- **有理论支持**：PILOT 已证明其在加性模型中的一致性。当数据由线性模型生成时，与 CART 相比，PILOT 可达到多项式收敛率。

本文结构安排如下。第2节介绍 PILOT 算法并讨论其性质，包括推导其时间复杂度和空间复杂度。第3节给出两项主要理论结果。首先，讨论并证明该算法对于一般加性模型的一致性。其次，当真实底层函数确实为线性函数时，展示其收敛速率有所提升。定理、命题和引理的证明见附录。第4节给出实证评估，其中将 PILOT 与若干替代方法在多个基准数据集上进行比较。在已知线性模型拟合效果良好的数据集上，PILOT 的表现优于其他基于树的方法；而在 CART 通常表现良好的数据集上，PILOT 的表现优于其他线性模型树。第5节总结全文。

# 2 方法

本节介绍 PILOT 学习算法的运作方式。我们首先说明如何构建其树，并阐述所作选择的理由，随后推导其计算成本。

我们将 $n \times p$ 设计矩阵记为 $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$，将 $n \times 1$ 响应向量记为 $Y = (y_1, \ldots, y_n)^\top$。我们考虑标准回归模型 $y = f(X) + \epsilon$，其中 $f : \mathbb{R}^p \to \mathbb{R}$ 是未知的回归函数，且 $\epsilon$ 的均值为零。

顺带一提，我们在此讨论的线性模型和树对于响应变量加上一个常数具有等变性，也就是说，预测模型的拟合参数保持不变，只有截距项例外。因此，在下文中，我们假设响应变量以零为中心，即 $Y_{min} = -Y_{max}$ 不失一般性。

## 2.1 PILOT的主要结构

典型的回归树包含四个要素：构造规则、对模型/分裂的评估、停止规则以及预测。大多数回归树算法采用贪心方式自上而下构建。也就是说，它们沿着一个预测变量对原始空间进行分裂，并在各个子集上重复这一过程。这种方法具有一些显著优势。第一是速度快。第二是它从全局视角开始研究数据，这有助于检测线性关系。

算法1概述了PILOT的总体流程。它首先对每个预测变量进行排序并存储其秩。在每个节点处，PILOT随后通过一个选择程序选出一个预测变量和一个单变量分段线性模型；我们将在第2.2节和第2.3节中介绍该选择程序。然后，它拟合该模型，并将残差传递给子节点，以便进一步拟合。这一递归

过程会持续进行，直到算法停止。停止条件有三个。具体而言，当满足以下情况时，节点中的递归会停止：

- 节点中的样本数量低于阈值$n_{\text{拟合}}$；
- 某个子节点中的样本数量将低于阈值$n_{\text{叶节点}}$；
- 候选模型中没有任何一个模型的表现显著优于常数预测，具体内容将在第 2.3 节中介绍。

当所有节点都达到停止触发条件后，算法即完成。最终预测值通过汇总从根节点到叶节点的线性模型预测值得到，如图 1 的示例所示。

**算法1：** PILOT 算法概要

输入：训练集（$\boldsymbol{X}, Y$）， $n_{\text{fit}}$：拟合模型所需的最小样本数； $n_{\text{leaf}}$：叶节点中的最小样本数； $K_{\max}$：最大树深度。输出： $T$，一个决策树对象。函数 PILOT($\boldsymbol{X}, Y, n_{fit}, n_{leaf}, K_{max}$):$I \leftarrow \{1, \ldots,$ length($Y$)$\}Y_0 \leftarrow Y$； $p \leftarrow$ 预测变量的数量 $B \leftarrow Y_{max}$对每个预测变量 $X^{(j)}$ 中的值 $X_i^{(j)}$ 排序，并将排序后样本的索引存储在一个 $n \times p$ 整数矩阵 $idx$中。使用（$\boldsymbol{X}, Y$） $n_{\text{fit}}, n_{\text{leaf}}, K_{\max}, I, Y_0, p, idx, B$和函数 Model Selection递归构建决策树。拟合选定模型，更新残差 $Y$以及每个子节点中样本的索引集 $I$ 。返回 Tend

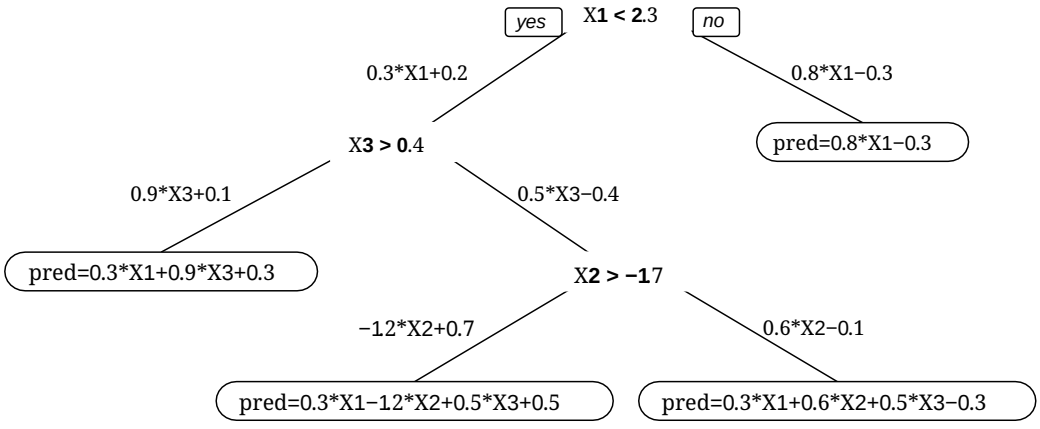



图 1：PILOT 树示例。

请注意，PILOT 保持了良好的可解释性。树结构将决策过程可视化，而每一步中的简单线性模型则揭示了所用的预测变量，以及预测值如何随该预测变量发生微小变化而改变。此外，可以基于每个节点的不纯度增益，定义一种类似于 CART 的特征重要性度量。这是因为 PILOT 在每个节点只选择一个预测变量，因此增益完全取决于该变量。这一点不同于 M5 、 GUIDE 和 LLRT 等方法；这些方法在每个节点使用多个预测变量，因此难以将增益公平地分配给多个预测变量，以得出每个预测变量总体重要性的度量。

## 2.2 节点中使用的模型

如算法1的高层概述所述，在每个节点，PILOT 会从多个线性模型和分段线性模型中选择一种拟合。具体而言，PILOT 会针对每个预测变量考虑以下回归模型：

- CON：**常数**拟合。拟合出一个 CON 模型后，我们会停止该节点中的递归 .
- LIN：简单**线性**回归。
- PCON：如 CART 中所用的**分段常数**拟合。
- BLIN：**分段线性**拟合：由两个线性片段构成的连续函数。
- PLIN：一种无需连续的双**Piece LINear** 拟合。

为避免不稳定拟合，仅当预测变量的唯一值数量至少为 5 时，才考虑 LIN 和 BLIN 模型。同样，只有当两个潜在子节点中的预测变量都至少有 5 个唯一值时，才考虑 PLIN 模型。

这些模型将 CART 的分段常数 PCON 拟合扩展为线性拟合。具体而言，PLIN 是 PCON 的直接扩展，而 BLIN 可视为 PLIN的正则化和平滑版本。对于类别型变量或只有少数几个唯一值的变量，只使用PCON 和 CON 。最后，如果在某个节点中选择了 CON ，该节点中的递归便会停止。

如果只选择 PCON 模型，PILOT 就会退化为 CART（不含其剪枝），因为子节点中的最小二乘常数拟合等于其平均响应。

请注意，选择 LIN 时节点不会分裂。因此，在节点中进行 LIN 拟合时，PILOT 不会增加其报告的深度。这会影响最终树的报告深度，而该深度是此方法的一个调优参数。

可能会在同一个节点中连续进行多次 LIN 拟合，这种情况确实出现在我们的实证研究中。在发生这种情况的节点中，PILOT 实际上执行的是$L^2$ 提升（B¨uhlmann，2006），即通过重复进行简单线性回归来拟合多元回归。研究表明， $L^2$ 提升适用于高维线性回归，且结果与套索回归相当。Freund 等人（2017）还证明，在某些假设下，其收敛速率可能相当快。对于 LIN 模型，PILOT 不会增加节点的深度值，以免中断这一提升过程。

## 2.3 模型选择规则

在每个节点中，PILOT 都会使用选择程序，从五种模型类型CON、 LIN、 PCON、 BLIN 和 PLIN中进行选择。根据残差平方和（RSS）降幅最大来选择模型并不恰当，因为这样总会选中 PLIN，因为该模型可以泛化其他所有模型。因此，我们需要某种正则化方法，以便在一路选择到 PLIN 所带来的额外 RSS 增益不够显著时，选择更简单的模型。

经过大量实验，我们采用了以下正则化方案。在每个节点中，PILOT 会选择使贝叶斯信息准则值最低的预测变量与回归模型组合

$$n \log\left(\frac{\mathrm{RSS}}{n}\right) + v \log(n) \tag{1}$$

该准则是残差平方和 RSS、节点中的样本数 $n$ 以及模型自由度 $v$ 的函数。借助这一选择规则，PILOT 在局部缓解过拟合；而在整棵树中应用该规则，则会产生全局正则化效果。

接下来要确定的是，对于五种模型中的每一种，式 (1) 中的 $v$ 应取何值。自由度由两部分相加确定：不包括分裂在内的模型参数数量，以及分裂的自由度。模型类型 CON、 LIN、 PCON、BLIN 和 PLIN 除分割点外，分别包含 1、2、2、3 和 4 个系数。根据经验和理论论据，Hall 等人（2017年）发现，一个不连续点应占 3 个自由度。对于 PCON 和 PLIN，我们采用这一建议，因此它们各增加 3 个自由度。 BLIN 也包含一个分割点，但其拟合函数是连续的。为体现这种中间复杂度，我们为 BLIN增加 2 个自由度。综上，对于模型类型 CON、 LIN、PCON、 BLIN 和 PLIN，我们得到的自由度依次为 $v = 1$、2、5、5 和 7。

式（1）中的 BIC 是我们本可以采用的若干模型选择准则之一。我们也尝试了其他指标，例如赤池信息准则（AIC）和修正赤池信息准则。结果发现，AIC 往往过于频繁地选择 PLIN，从而削弱了正则化效果。修正赤池信息准则需要采用剪枝程序，才能达到与 BIC准则相当的效果。由于剪枝会带来显著的额外计算成本，我们决定不采用修正赤池信息准则。

或者，也可以按照 B¨uhlmann（2006）的方法，计算聚合模型在每个节点中的自由度，然后计算修正赤池信息准则，以决定何时停止。但采用这种方法时，必须为每条路径维护帽子矩阵，这需要更多的计算时间和内存空间。此外，随着节点中的样本数量发生变化，帽子矩阵的评估也会变得更加复杂。基于这些原因，我们倾向于在每个节点采用选择规则（1）。

算法 2 展示了在给定节点中执行模型选择程序的伪代码。我们分别遍历每个预测变量。如果预测变量是数值型的，我们就在 5 个候选模型中找出贝叶斯信息准则最低的单变量模型。对于 CON 和 LIN，这一评估可以立即完成。对于另外三个模型，我们需要考虑所有可能的分裂点。我们按递增顺序依次考虑这些分裂点（请回想一下，每个预测变量的排序已在算法开始时完成），这样便可以在每一步中高效地维护和更新线性模型的 Gram 矩阵和矩矩阵，以便

保持并高效更新线性模型。另一方面，如果预测变量是类别型的，我们就采用 CART 的方法，如 Hastie等人（2009）的第9.2.4节所述。也就是说，我们首先计算响应变量在每个水平 $c$ 下的均值 $m_c$ ，然后按照 $m_c$对样本排序。接着，我们按常规方式拟合模型 PCON 。

分别确定每个预测变量的最佳模型类型后，算法 2 会返回 BIC准则最低的预测变量与模型组合。

## 2.4 预测截断

PILOT 依靠两种截断程序，避免训练数据和测试数据上的预测值不稳定。

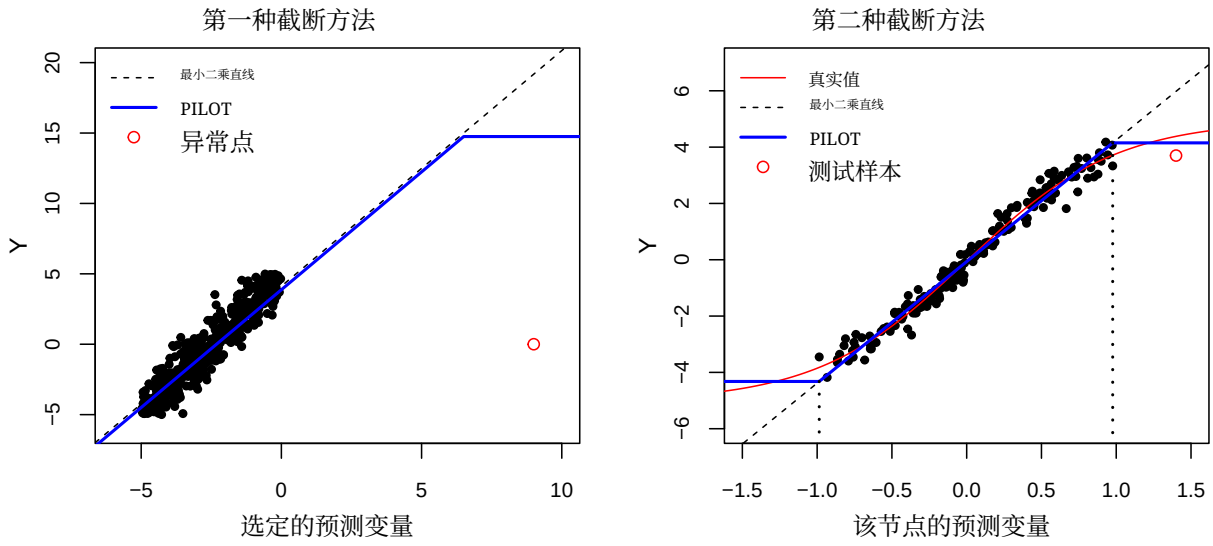



图2：左：树中一个节点采用第一种截断方法的示例。右：第二种截断方法的示例。

第一个截断过程的动机如下。图2左侧面板展示了 PILOT 中某个节点的数据。选定的预测变量位于横轴上。该图说明了这样一种情况：节点中的大多数数据点集中在预测变量范围的一部分，导致线性模型对这些点赋予较大的权重，而对其他点赋予较小的权重。这可能导致极端预测，例如右侧的红点。尽管更深层的节点或许能够校正这类不稳定的预测，但这可能会引入额外的分裂，使模型变得不必要地复杂。此外，如果这种情况发生在叶节点上，预测便无法得到校正。为避免这一问题，我们将截断预测函数。注意，如果给响应变量加上一个常数，除了截距项会按应有的方式变化外，PILOT 得到的估计参数都相同。因此，不失一般性，我们可以假设 $Y_{max} = -Y_{min}$。将该量记为$B$，原始响应变量的范围为 $[-B, B]$。然后，我们将预测值截断，使其落在 $[-3B, 3B]$内。接着，计算响应变量 $Y$ 减去截断后的预测值，以得到新的残差，并继续构建下一深度的树。整个过程

**算法 2：** 训练集节点中的模型选择与分裂查找

**全局变量：** $(\boldsymbol{X}, Y)$：当前节点中的预测变量和残差向量；$p$：预测变量的数量；$idx$：$p$ 预排序预测变量的索引矩阵；$n_{\text{leaf}}$：叶节点中的最小样本数。

**输入：** $I$，当前节点中的样本索引。

**输出：** $best\ pivot$；$best_model$；$best\ pred$；$_lm\ l, _lm\ r$，左侧和右侧的线性模型右子节点；$range$，当前节点中选定的预测变量的范围。

**函数** `Model Selection(`$I$`)`**：**

**初始化：** 最佳枢轴、最佳模型、最佳贝叶斯信息准则、最佳预测变量、线性模型左侧、线性模型右侧、范围

**对于** 每个预测变量 $X^{(j)}$ 及其 $j = 1, \ldots, p$ **执行**

根据 $X^{(j)}$，使用 $I$ 和 $idx$对样本排序并选取。

拟合 CON，并使用其结果初始化最佳模型的参数。

**如果** $X^{(j)}$ 是数值型预测变量**则**

计算 $(D^{\text{LIN}})^\top D^{\text{LIN}}$ 和 $(D^{\text{LIN}})^\top Y$，其中 $D^{\text{LIN}}$ 是线性模型的设计矩阵。对于 PCON，它是 $|I| \times 1$；否则为 $|I| \times 2$。

使用上述矩阵拟合 LIN。存储最佳模型的参数。

对于 PCON 和 PLIN，初始化格拉姆矩阵 $G^{\text{PLIN}}{}_{l,r}$ 以及矩阵 $M_l^{\text{PLIN}}{}_{,r}$ 左子节点和右子节点中的简单线性模型，使得 $G_l^{\text{PLIN}} = G_r^{\text{PLIN}} = \mathbf{0}$, $M_l^{\text{PLIN}} = (D^{\text{LIN}})^\top D^{\text{LIN}}$, $M_r^{\text{PLIN}} = (D^{\text{LIN}})^\top Y$。

对于 BLIN，初始化 Gram 矩阵 $G^{\text{BLIN}} = (D^{\text{BLIN}})^\top D^{\text{BLIN}}$ 和矩矩阵 $M^{\text{BLIN}} = (D^{\text{BLIN}})^\top Y$，其中 $D^{\text{BLIN}}$ 是 BLIN 的设计矩阵，节点等于 $\min X^{(j)}$。

**针对**每个可能的分裂枢轴**执行**

更新 $G^{\text{BLIN}}$，$M^{\text{BLIN}}$, $G_{l,r}^{\text{PLIN}}$, $M_{l,r}^{\text{PLIN}}$.

如果满足 $n_{\text{leaf}}$，使用 Gram 矩阵和矩矩阵拟合 $lm\ l$, $lm\ r$，得到左右模型 $lm\ l$ 和 $lm\ r$。

存储最佳模型的参数。

**end**

**else**

计算每个类别水平 $c$对应的响应变量均值 $m_c$。

根据其 $m_c$ 对样本重新排序。

贪心搜索最佳 $value$，使得 PCON 采用划分 $\{c|m_c \leqslant value\}$ 和 $\{c|m_c > value\}$ 的贝叶斯信息准则最低。

存储最佳模型的参数。枢轴是集合 $\{c|m_c \leqslant value\}$。

**end**

**返回：** 最佳枢轴、最佳模型、最佳贝叶斯信息准则、最佳预测、左线性模型、右线性模型、范围

**end**

总结于算法 3 中。在 $\pm 3B$ 处截断似乎可能过于宽松；如果需要，也可以在 $B$ 的较小倍数处截断。但我们的实证结果表明，在 $\pm 3B$ 处截断足以满足我们的目的。

第一种截断方法不仅应用于训练期间，也应用于预测

**算法 3：** 在训练数据的树构建过程中进行截断

**全局变量：** $p$; $idx$; $n_{\text{fit}}$：拟合模型所需的最小样本数； $n_{\text{leaf}}$：最小叶节点中的最小样本数； $K_{\max}$：最大深度； $Y_0$：原始响应变量； $B$：中心化响应变量的最大值。

**输入：** $(\boldsymbol{X}, Y)$：训练集； $K$：当前深度； $I$：样本索引。

**输出：** 一个树对象 $T$，其属性包括 $m$：模型； $l, r$：左侧和右侧子节点； $lm_l$, $lm_r$：左右子节点中的模型系数。

**函数** `Build Tree`$(\boldsymbol{X},\ Y)\ K,\ I)$**：**

- $T \leftarrow$ 一个空节点
- **如果** $K < K_{max}$ 且 *length(I)* $\geqslant n_{fit}$ **则**
  - 执行 `Model Selection`$(I)$ 并记录最佳模型的参数
  - $K \leftarrow K + 1$ 如果该 $best\ model$ 不是 LIN 或 CON
  - $T \leftarrow$ 记录了参数的树节点
  - **如果** 最佳模型为 CON**则**
    - **返回** $T$
  - **否则，如果** 最佳模型为 LIN**则**
    - $raw_pred[I] \leftarrow$ 在 $\boldsymbol{X}[I]$ 上的 $lm_l$ 预测
    - 在 $\pm 3B$ 处截断实际预测 $Y_0[I] - Y[I] + raw\ pred[I]$。更新使用这一新预测值时的残差。
    - $T.l \leftarrow$ `Build Tree`$(\boldsymbol{X},\ Y)\ K,\ I)$
  - **else**
    - $I_l,\ I_r \leftarrow$ 左子节点和右子节点中的样本索引
    - $raw_pred[I_l] \leftarrow \boldsymbol{X}[I_l]$上的 $lm_l$ 预测值
    - $raw_pred[I_r] \leftarrow \boldsymbol{X}[I_r]$上的 $lm_r$ 预测值
    - 截断实际预测值 $Y_0[I_l] - Y[I_l] + raw\ pred[I_l]$ 以及 $Y_0[I_r] - Y[I_r] + raw_pred[I_r]$ 在 $\pm 3B$ 处，并更新残差 $Y[I_l]$ 和 $Y[I_r]$ 在子节点中。
    - $T.l \leftarrow$ `Build Tree`$(\boldsymbol{X},\ Y)\ K,\ I_l)$
    - $T.r \leftarrow$ `Build Tree`$(\boldsymbol{X},\ Y)\ K,\ I_r)$
  - **end**
- **else**
  - **返回** $T$
- **end**
- **返回** $T$

新数据上。训练数据的响应变量范围会被保存；当对新数据进行预测时，如果预测结果超出指定范围 $[-3B, 3B]$，就会以相同的方式截断。此过程可作为预测的基本安全防护措施。然而，这还不够，因为可能会出现一个新问题：测试数据中的某个预测变量取值可能超出训练数据中同一预测变量的范围。图2右图展示了这一情况。垂直虚线表示训练数据中该预测变量的范围。然而，测试样本的预测变量取值（红点）远远超出该范围。使用训练期间拟合的直线来预测其响应变量将会

这会导致外推，结果可能不切实际。这是线性模型树的一个已知问题。例如，在第4节的实验中，FRIED 和 M5 在某些数据集上的预测值急剧增大，结果发现这是由这一问题造成的。

因此，我们增加了第二个截断过程，该过程仅应用于新数据（因为训练数据中的所有预测变量值都在其范围内）。因此，我们提出按照 Loh 等人（2007）的做法截断预测值。更确切地说，在训练期间，我们记录在此节点中选定的预测变量的范围 $[x_{\min}, x_{\max}]$，并存储相应预测值的范围 $[f(x_{\min})\ f(x_{\max})]$。对新数据进行预测时，我们截断预测值，使其保持在训练范围内。更确切地说，我们将此节点上的原始预测值 $f(x_{\text{test}})$ 替换为 $\max(\min(f(x_{\text{test}})\ f(x_{\max}))\ f(x_{\min}))$。

这两种截断方法相辅相成。例如，在图2右侧的情形中，第一种方法并不足够，因为新样本的线性预测值仍会落在 $[-3B, 3B]$ 内。此外，在图2左侧的情形中，第二种方法也不起作用，因为异常样本位于训练数据的范围内。我们的实证研究表明，结合这两种截断方法有助于消除极端预测，从而提高模型的稳定性。而且，这并未削弱模型整体的预测能力。算法 4 描述了 PILOT 对新数据进行预测时如何应用这两种截断方法。

**算法 4：** 对新数据进行预测时的截断

**输入：** $T$：一个决策树对象； $x$：一个新的观测值； $B$：中心化响应变量的最大值在训练数据中。

**输出：** $pred$， $x$ 中的预测值。

**函数** `Predict`$(T, X)$**：**

- $pred \leftarrow 0$
- **当*T* 不为 *None* 时执行**
  - **如果** $x$ 位于 $T$ 的左子节点中**那么**
    - $X_{min}$， $X_{max} \leftarrow$ $T.range$ 在选定的预测变量 $j$ 在训练数据中的
    - $X^* \leftarrow \min(\max(X^{(j)},\ X_{min})\ X_{max})$
    - $pred\text{-}add \leftarrow$ $T.lm\ l$ 在 $X^*$上的预测
    - $pred \leftarrow \max(\min(pred + pred\ add, 3B)\ -3B)$
    - $T \leftarrow T.l$
  - **else**
    - $X_{min}$， $X_{max} \leftarrow$ 选定的训练数据中的预测变量 $j$。
    - $X^* \leftarrow \max(\min(X^{(j)}, X_{min})\ X_{max})$
    - $pred\text{-}add \leftarrow$ 在 $X^*$ 上对 $T.lm\ r$ 的预测
    - $pred \leftarrow \max(\min(pred + pred\ add, 3B)\ -3B)$
    - $T \leftarrow T.r$
  - **end**
- **end**
- **返回** $pred$

**end**

## 2.5 停止规则与剪枝

众所周知，如果任由决策树生长得很深，就容易对数据过拟合。为避免这种情况，首先要规定一个节点在可以分裂之前必须包含的最小样本数，以及每个子节点必须包含的最小样本数。此外，如果贝叶斯信息准则模型选择准则选择 CON，PILOT 也会停止分裂该节点。

包括 CART 在内的若干决策树采用了不同的方法。它们先让树生长，之后再进行剪枝。这确实有助于降低泛化误差。然而，剪枝可能非常耗时。通常需要使用交叉验证程序来选择剪枝参数。PILOT 也可以像 CART 一样采用代价复杂度剪枝。（如果同一节点中出现一系列 LIN fits，PILOT 要么全部保留，要么一起移除。）然而，在我们的实证研究中，我们发现，在各种数据集上，将这种剪枝方法加入 PILOT 并未优于现有的 CON 停止规则。此外，我们将在下一节看到，泛化误差会渐近地消失。而且，不进行剪枝在计算上更加高效。基于这些原因，PILOT 不采用剪枝。这一选择所带来的计算增益，使得在集成方法中使用 PILOT 更加可行。

## 2.6 时间和空间复杂度

最后，我们证明，PILOT 的时间复杂度和空间复杂度与未进行剪枝的 CART 相同。证明的要点是，对于单个预测变量，五种模型中的每一种都可以通过对预测变量值进行一次遍历来求值。计算下一个分裂点只需要对 Gram 矩阵和矩矩阵进行一次秩一更新，而这可以在常数时间内完成。自然地，$\mathcal{O}(.)$复杂度前的隐含比例因子对于 PILOT 而言高于 CART，但该算法在实践中相当快。

**命题1**。 *PILOT* 的时间复杂度和空间复杂度与未进行剪枝的 *CART* 相同。

证明。 对于 PILOT 和 CART，我们都假设 $p$ 预测变量已经过预排序，这只需一次性耗费 $\mathcal{O}(np\log(n))$ 时间。

我们首先检查算法 2 中模型选择的复杂度。已知 CART 的分裂查找时间复杂度为 $\mathcal{O}(np)$。对于 PCON 模型，PILOT 使用与 CART 相同的算法。对于 CON，我们只需用 $\mathcal{O}(n)$ 的时间计算一个平均值。对于LIN，评估 $(D^{\text{LIN}})^\top D^{\text{LIN}}$ 和 $(D^{\text{LIN}})^\top Y$ 也需要 $\mathcal{O}(n)$。在评估PLIN 模型时，根据定义，Gram 矩阵和矩矩阵始终满足 $G_l^{\text{PLIN}}+G_r^{\text{PLIN}}=G^{\text{LIN}}=(D^{\text{LIN}})^\top D^{\text{LIN}}$和 $M_l^{\text{PLIN}}+M_r^{\text{PLIN}}=(D^{\text{LIN}})^\top Y$。这些矩阵最多只有 4 个元素，在评估下一个分裂点时，可在 $\mathcal{O}(1)$时间内对其进行增量更新。对于 BLIN，推理过程类似。在每个模型中，由于 Gram 矩阵的大小固定，对其求逆只需 $\mathcal{O}(1)$ 时间。因此，我们可以在一次遍历数据的过程中，评估一个预排序预测变量上的所有选项，这仍为 $\mathcal{O}(n)$，所以对所有预测变量而言，所需时间为 $\mathcal{O}(np)$，与 CART 相同。就空间复杂度而言，Gram 矩阵和矩矩阵只需要 $\mathcal{O}(1)$ 的

存储空间。由于 CART 也必须存储每个子节点中的平均响应变量，这同样需要 $\mathcal{O}(1)$ 的存储空间，因此两种方法的空间复杂度相同。

对于算法 3 中的树构建，在每个深度 $1 \leqslant k \leqslant K_{max}$，计算索引 $I_{l,\ r}$和残差的时间复杂度为 $\mathcal{O}(n)$。CART 也会计算索引 $I_{l,\ r}$，每一步所需的复杂度同样为 $\mathcal{O}(n)$。因此，两种方法的总耗时均为 $\mathcal{O}(K_{max}n)$。由于 PILOT 树的每个节点仅比 CART 多出 $\mathcal{O}(1)$ 个属性，因此 PILOT 的空间复杂度与 CART 相同。

初始化和预测部分的结果是显而易见的。 □

# 3 理论结果

## 3.1 普适一致性

本节证明 PILOT 的一致性。我们遵循 Klusowski（2021）的做法，并假设底层函数 $f \in \mathcal{F} \subset L^2([0,1]^p)$ 具有加性形式

$$f(X) := f_1(X^{(1)}) + \cdots + f_p(X^{(p)}) \tag{2}$$

其中，$f_i$ 具有有界变差，$X^{(j)}$ 是第 $j$ 个预测变量。我们将 $f \in \mathcal{F}$ 的全变差范数 $||f||_{TV}$ 定义为所有可能的 $f$ 表示中 $\sum_{i=1}^{p} ||f_i||_{TV}$ 的下确界，并假设式（2）中的表示达到该下确界。

对于所有 $f, g \in L^2([0,1]^p)$ 以及大小为 $n$ 的数据集 $X_1, \ldots, X_n$，我们将经验范数和经验内积定义为

$$||f||_n^2 := \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|^2 \quad \text{and} \quad \langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i). \tag{3}$$

对于响应向量 $Y = (Y_1, \ldots, Y_n)$，我们记

$$||Y - f||_n^2 := \frac{1}{n}\sum_{i=1}^{n} (Y_i - f(X_i))^2 \quad \text{and} \quad \langle Y, f\rangle_n := \frac{1}{n}\sum_{i=1}^{n} Y_i f(X_i). \tag{4}$$

向量和函数上的一般范数记作不带下标的 $||\cdot||$。

为了表示节点 $T$ 上含有 t 个观测值的函数的范数和内积，我们将下标 $n$ 替换为 $t$。我们用 $\mathcal{T}_k$ 表示深度为 $k$ 的树节点集合，以及深度小于 $k$ 的叶节点。特别地，$\mathcal{T}_K$ 包含一棵深度为 $K$ 的树的所有叶节点。

现在我们可以陈述主要定理：

**定理 2.** 设 $f \in \mathcal{F}$满足 $||f||_{TV} \leqslant A$，并记 $f(\mathcal{T}_K)$为深度为 $K$ 的 *PILOT* 树的预测。假设 $X \sim P$在 $[0,1]^{p_n}$ 上，响应变量几乎必然有界于 $[-B, B]$，且深度 $K_n$ 和预测变量数量 $p_n$ 满足 $K_n \to \infty$ 和 $2^{K_n} p_n \log(np_n)/n \to 0$。那么，*PILOT* 是一致的，即

$$\lim_{n\to\infty} \mathbb{E}[||f - \hat{f}(\mathcal{T}_{K_n})||^2] = 0. \tag{5}$$

**备注 3.** 注意，如果令 $K_n = \log_2(n)/r$，其中 $r > 1$ 且 $p_n = n^s$ 满足 $0 < s < 1 - 1/r$，则深度和维度的条件很容易满足。由此得到的收敛速率为 $\mathcal{O}(1/\log(n))$。这与 *Klusowski*（*2021*）在非常相似的假设下针对 *CART* 得到的速率相同。

定理 2 证明的关键在于，为每个深度 $k$ 建立训练误差 $R_k := ||Y - \hat{f}(\mathcal{T}_k)||_n^2 - ||Y - f||_n^2$ 的递归公式。然后，我们可以利用经验过程理论的结果（Gy¨orfi 等，2002），为 PILOT 推导出一个预言机不等式，并最终证明一致性。

为建立递归公式，我们首先注意到 $R_k = \sum_{T \in \mathcal{T}_K} w(T) R_k(T)$，其中 $w(T) := t/n$ 是节点 $T$ 的权重，且 $R_k(T) := ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - ||Y - f||_t^2$。于是我们立即得到 $R_{k+1} = R_k - \sum_{T \in \mathcal{T}_k} w(T)\Delta^{k+1}(T)$，其中

$$\Delta^{k+1}(T) := ||Y - \hat{f}(\mathcal{T}_k)||_t^2 \; - \; t_l||Y - \hat{f}(\mathcal{T}_{k+1})||_{t_l}^2/t \; - \; t_r||Y - \hat{f}(\mathcal{T}_{k+1})||_{t_r}^2/t$$

是模型在 $T$ 上的不纯度增益。这里，$t_l$ 和 $t_r$ 分别表示左子节点和右子节点中的样本数。如果没有发生分裂，其中一个数就设为零。对于 PILOT，我们需要从递归公式中移除那些不会产生不纯度增益的节点。为此，我们定义 $C_k^+ = \{T | T = \text{CON}, T \in \mathcal{T}_{k-1}, R_k(T) > 0\}$，即拟合了 CON 的“坏”节点集合；对于满足 $R_k(T) \leqslant 0$ 的节点，也类似地定义 $C_k^-$。于是可以定义 $R_{C_k^+} := \sum_{T \in C_k^+} w(T) R_k(T)$，即这些 $\sim$CON 节点中样本所对应的正训练误差。接下来，我们可以考察 $R_k := R_k - R_{C_k^+}$，并证明渐近地，$R_k$ 和 $R_{C_k^+}$ 都会变小。

现在，问题可归结为将选定模型的不纯度增益与此步骤之前的训练误差联系起来。最近，Klusowski（2021）在 CART 中为这类估计提出了一种新方法。然而，他的证明用到了这样一个事实：每一步的预测都是分段常数，而这里并非如此。因此，我们为 PILOT 推导了一个广义结果PCON （见附录 B.1 中的引理 12）。为了控制其他模型的增益，我们可以使用以下命题：

**命题 4.** 设 $\Delta_1$、$\Delta_2$ 和 $v_1$、$v_2$ 分别为某个包含 $t$ 个样本的节点 $T$ 上两个回归模型的不纯度增益和自由度。设 $R_0$ 为 $T$ 中的初始残差平方和。则

- 如果模型 *1* 的表现优于 CON，即 $BIC_{con} > BIC_1$，则对于某个依赖于 $v_1$ 和 $t$ 的正函数 $C$，有 $t\Delta_1/R_0 > C(v_1, t) > 0$。
- 如果 $BIC_{con} > BIC_1$ 且 $BIC_2 > BIC_1$ 满足 $v_2 \geqslant v_1$，则有 $\frac{\Delta_1}{\Delta_2} \geqslant \frac{v_1 - 1}{v_2 - 1}$。

此外，如果在某个节点选择了 CON，那么在后续模型中也会选择它 *s.*

因此我们知道，选定模型的增益始终与 PCON 的增益相当，至多相差一个常数因子。

该命题也从模型选择的角度为 CON 停止规则提供了依据：当为某个节点选择了 CON 时，该节点后续的模型仍将是 CON。

当其他回归模型带来的增益不足以弥补其增加的复杂度时，就会选择 CON 模型。由于 CON 只对均值进行回归，因此一旦遇到第一个 CON 拟合，就可以停止该节点中的算法。

我们可以给出以下递归公式：

**定理 5.** 在与定理 *2* 相同的假设下，对于任意 $K \geqslant 1$ 和任意$f \in \mathcal{F}$，我们有

$$||Y - \hat{f}(\mathcal{T}_K)||_n^2 \leqslant ||Y - f||_n^2 + \frac{(||f||_{TV} + 6B)^2}{6(K+3)} + R_{C_K^+} . \tag{6}$$

此外，若令 $K = 对数_2(n)/r$，其中 $r > 1$，则有 $R_{C_K^+} \sim \mathcal{O}(\sqrt{对数(n)/n^{(r-1)/r}})$。

BIC准则确保所有 CON 节点中的训练误差在 $n \to \infty$ 时趋于零。这使我们能够利用前述结果，以及 Gy¨orfi 等人（2002）的定理11.4和9.4及引理 13.1，证明主要定理。所有中间引理和证明均见附录B。

## 3.2 线性模型的收敛速率

定理 2 得出的收敛速率与 Klusowski（2021）推导出的 CART 收敛速率相同。这是因为我们没有对真实底层加性函数的性质作出具体假设。因此，这一收敛速率适用于各种各样的函数，包括表现不佳的函数。不过在实践中，如果函数具有某种良好的性质，收敛速度可能会快得多。特别是，鉴于 PILOT 在其节点中采用线性模型，当真实底层函数为线性函数时，PILOT 很可能表现更好，而线性函数是加性模型的一种特殊情况。我们将证明，在这种情况下，PILOT 确实具有更优的收敛速率。CART 不具备这一结果；据我们所知，其他任何线性模型树算法也尚未被证明具有这一性质。

为了证明 PILOT 在线性模型中的收敛速率，我们采用 Freund 等人（2017）的一项最新研究成果。对于 $L^2$ 提升算法的收敛速率，Freund 等人（2017）在定理 2.1 中证明，对于固定设计矩阵，均方误差（MSE）会随深度 $K$呈指数衰减。该证明基于 $L^2$ 提升与平方损失 $||Y - \boldsymbol{X}\beta||_n^2$下的二次规划（QP）之间的联系。事实上，假设每个预测变量都已中心化并标准化，平方损失的梯度的 $L^\infty$ 范数等于所有简单线性模型中最大的 不纯度增益。因此，第 $k$步的增益与第 $k-1$步的均方误差呈线性关系，从而带来更快的收敛速率。根据二次规划的性质，该速率取决于 $\boldsymbol{X}^\top \boldsymbol{X}$的最小特征值 $\lambda_{min}$。

在我们的框架中，我们可以在每个节点中考虑一个局部二次规划问题，以估计LIN的增益，从而得到选定模型改进幅度的下界（ CON被选中时除外）。为确保误差以恒定比例衰减，我们必须对预测变量局部相关矩阵的最小特征值作出一些假设。这个特征值可视为预测变量之间多重共线性的度量，而多重共线性是影响最小二乘回归收敛速率的关键因素。

具体而言，我们要求理论相关矩阵的最小特征值（其第 $jl$个元素为 $\mathbb{E}[(X-\bar{X})(X-\bar{X})^\top]_{jl}/(\sigma_j\sigma_l)$，其中 $\sigma_j$ 是第 $j$个预测变量的标准差）在任意足够小的立方区域内都以某个常数 $\lambda_0$ 为下界。随后，我们应用两个集中不等式，证明此类立方体内数据的协方差矩阵的最小特征值以高概率大于 $\lambda_0$ 。最后，我们可以证明误差的期望呈指数衰减，从而实现快速收敛。

我们的条件如下：

- 条件1： </style}PILOT 算法在以下情况下会停止分裂节点：
  – 节点中的样本数小于 $n_{\min}=n^\alpha, 0<\alpha<1$。– 某个预测变量的方差小于 $2\sigma_0^2$，其中 $0<\sigma_0<1$。– 节点的体积小于阈值 $\eta$, $0<\eta<1$。
- 条件2： </style}我们假设 $X\sim P$ 在 $[0,1]^p$ 上，且误差 $\epsilon$ 具有有限的四阶矩。此外，对于体积小于 $\eta$ 的任意立方体 $C$ ，我们假设 $\lambda_{min}(Cor(X|X\in C))\geqslant 2\lambda_0>0$，其中 $Cor(X)$ 是相关矩阵。

第一个条件对于基于树的方法而言很自然。对于第二个条件，一个直接的例子是所有预测变量相互独立的情形。另一个例子是这些变量服从多元高斯分布；当将其限制在 $[0,1]^p$ 或其他某个立方体 $C$内时，该分布就变成截断高斯分布。可以证明，在体积较小的立方体 $C$ 中，高斯分布的相关性相对较小；二维情形可参见 Muthen（1990）。这并不令人意外，因为在较小的立方体上，密度趋于常数。因此，这类立方体上的相关矩阵的最小特征值将有一个正的下界。

在这些条件下，我们得到以下结果，其证明见附录 C。

**定理 6.** 假设数据由线性模型 $Y\sim \boldsymbol{X}\beta+\epsilon$ 生成，其中有 $n$ 个样本。在条件 *1* 和 *2* 下，*PILOT* 在深度 $k$时的训练损失 $L_n^k$ 与最小二乘线性回归的训练损失 $L_n^*$ 之差满足

$$\mathbb{E}[L_n^k-L_n^*]\leqslant \gamma^k\sqrt{\mathbb{E}[(L_n^0-L_n^*)^2]}+\mathcal{O}(N_{leaves}\log(n)/n) \tag{7}$$

其中

$$\gamma:=1-\frac{\lambda_0}{4p}\,.$$

将这一结果与定理 2 证明中的技巧结合起来，我们可以证明我们的方法在线性函数上的收敛速率为多项式级。

**推论 7 (**线性模型的快速收敛)。 假设定理 *6* 的条件成立，且 $|Y|$ 几乎必然被某个常数 $B$ 所界定。令 $K_n=log_\gamma(n)$。那么，对于任意$0<\alpha<1$ 及其对应的 $n_{min}=n^\alpha$ ， 我们有

$$\mathbb{E}[||\hat{f}(\mathcal{T}_K)-\boldsymbol{X}\beta||^2]\leqslant\mathcal{O}\Big(\frac{\log(n)}{n^\alpha}\Big)\,. \tag{8}$$

因此，调优参数 $\alpha$ 的选择决定了收敛速率。当我们使用较低的 $\alpha$ 时，会得到较小的节点，因此节点数量更多，从而导致预言机不等式中的界以及 $\mathrm{CON}$ 节点误差的界更宽松，收敛也更慢。这是自然的，因为每个节点用于估计的样本更少。如果我们选择相对较高的 $\alpha$，则会获得更快的收敛速率；这也符合直觉，因为对于 $\alpha \to 1$，我们只有一个节点。

Cattaneo 等人（2022）针对更一般的函数类，证明了斜决策树具有多项式收敛率；这项工作或许可以推广到 PILOT。不过，我们注意到，他们的结果有一个很强的假设，即训练样本应近似均匀地分布在各个叶节点上。本文的条件则侧重于算法规则以及数据本身的分布。

# 4 实证评估

本节通过实证评估所提出的 PILOT 算法，并将其结果与一些常见竞争方法进行比较。我们首先介绍所比较的数据和方法。

## 4.1 数据集和方法

我们分析了 20 个基准数据集，样本数从 71 到 21263 不等，预测变量数从 6 到 4088 不等。我们从 UCI 仓库（Dua 和 Graff，2017年）使用了数据集 `Abalone`, `Airfoil`, `Bike`, `Communities`, `Concrete`,`Diabetes`, `Electricity`, `Energy`, `Residential`, `Skills`, `Superconductor`, `Temperature` and`Wine`。从 Kaggle 获取了 `Graduate Admission`, `Bodyfat`, `Boston Housing` and `Walmart`。`California Housing` 数据来自 StatLib 仓库（`http://lib.stat.cmu.edu/datasets/`）。`Ozone` 数据位于 `R`-package `hdi`（Dezeure 等人，2015年）中，而`Riboflavin` 来自 `R`-package `missMDA`（Josse 和 Husson，2016年）。表 1 按字母顺序列出了数据集及其规模。

在比较研究中，我们运行了以下基于树的方法：我们提出的 PILOT 算法（使用 Python 实现，并通过 NUMBA 加速分裂函数）、由我们实现的 FRIED（Friedman，1979年）、来自 `R` package Rweka（Hornik 等，2009年）的 M5（Quinlan，1992年），以及 CART。默认情况下，我们采用 $K_{\max} = 12$、$n_{\text{fit}} = 10$ 和 $n_{\text{leaf}} = 5$，这些是回归树的常见选择。我们还运行了 Scikit-Learn 软件包（Pedregosa 等，2011年）中的岭线性回归，以及套索回归（Tibshirani，1996年）。对于那些并非为处理分类变量而设计的方法，我们会先对这类变量进行独热编码。

我们将每个数据集随机划分为 5 折，并计算交叉验证均方误差，即每种方法在每一折上的测试均方误差的平均值。每种方法的最终得分为其均方误差除以该数据集上的最低均方误差。因此，得分为 1 表示该方法在该数据集上的表现最佳，其他方法的得分则反映其均方误差高出多少。

表 1：实证研究中使用的数据集。

| 来源 | 数据集 | $n$ | $p$ |
|---|---|---|---|
| UCI 仓库 | 鲍鱼 | 4177 | 8 |
| Kaggle | 研究生录取 | 400 | 9 |
| UCI 仓库 | 翼型 | 1503 | 6 |
| UCI 仓库 | Bike | 17389 | 16 |
| Kaggle | 体脂 | 252 | 14 |
| Kaggle | 波士顿房价 | 506 | 13 |
| StatLib 仓库 | 加利福尼亚房价 | 20640 | 8 |
| UCI 仓库 | 社区 | 1994 | 128 |
| UCI 仓库 | 混凝土 | 1030 | 9 |
| UCI 仓库 | 糖尿病 | 442 | 10 |
| UCI 仓库 | 电力 | 10000 | 14 |
| UCI 仓库 | 能源 | 768 | 8 |
| hdi | 臭氧 | 112 | 11 |
| UCI 仓库 | 住宅 | 372 | 105 |
| missMDA | 核黄素 | 71 | 4088 |
| UCI 仓库 | 技能 | 3395 | 20 |
| UCI 仓库 | 超导体 | 21263 | 81 |
| UCI 仓库 | 温度 | 7750 | 25 |
| Kaggle | 沃尔玛 | 6435 | 8 |
| UCI 仓库 | Wine | 4898 | 12 |

## 4.2 结果

结果汇总于表 2。蓝色条目对应得分不超过 1.05 的方法，也就是说，这些方法的均方误差不超过该数据集上最佳均方误差的 1.05 倍。PILOT 在 19 次比较中有 16 次优于 FRIED。它在 17 个数据集上的表现也优于 CART，并在 15 个数据集上胜过 M5。从表格底部两行可以看出，PILOT 的平均得分最佳，且标准差最低。

表 2 中有一些不寻常的条目，需要加以说明。FRIED 一列在 `Residential` 数据集上的测试误差极高。仔细检查后发现，这是由外推造成的。在某个节点，测试样本的预测变量值超出了训练样本的范围，因此 LIN 进行了外推，给出了不合理的预测值。最后一列中的 M5 结果经过了部分调整，因为其预测值不具备对预测变量缩放的不变性。因此，我们分别在对预测变量进行缩放和不进行缩放的情况下运行了 M5，并在此处列出最佳结果。基于缩放的结果以括号标出。遗憾的是，M5 的表现往往仍然不佳。在 `Superconductor` 数据上，尽管采用了用于交叉验证和剪枝的优化函数（来自 Scikit-Learn），FRIED 和 CART 的训练时间仍超过了预设的墙钟时间。

仔细观察结果可以发现，PILOT 在 `Airfoil`数据上的表现不佳。一个可能的解释是，`Airfoil` 数据的底层模型相当复杂，通常会使用梯度提升和随机森林等集成方法进行建模（Patri 和 Patnaik，2015）。包括 PILOT 在内的基于树的方法给出的预测值相对较差

表 2：表 1 中各数据集相对于最佳结果的均方误差比。

| | PILOT | 岭回归 | 套索回归 | FRIED | CART | M5 |
|---|---|---|---|---|---|---|
| 鲍鱼 | **1.00** | **1.02** | **1.01** | 1.11 | 1.12 | 1.56 |
| 入学 | 1.19 | **1.00** | **1.00** | 1.11 | 1.41 | (1.21) |
| 翼型 | 1.98 | 4.40 | 4.40 | **1.00** | 1.75 | 9.52 |
| Bike | **1.00** | 3.03 | 3.05 | 1.84 | **1.05** | 1.45 |
| 体脂 | 1.15 | **1.00** | **1.03** | 1.47 | 1.55 | (1.16) |
| 波士顿 | **1.02** | **1.00** | 1.06 | 1.17 | 1.16 | 3.99 |
| 加利福尼亚 | **1.00** | 1.95 | 1.95 | 1.08 | 1.42 | (11.74) |
| 社区 | 1.08 | **1.01** | **1.00** | 1.17 | 1.32 | **1.04** |
| 混凝土 | **1.00** | 2.61 | 2.62 | 3.42 | 1.38 | 2.48 |
| 糖尿病 | 1.07 | **1.00** | **1.00** | 1.24 | 1.31 | 1.12 |
| 电力 | **1.00** | 2.75 | 2.75 | 1.28 | 1.88 | 2.19 |
| 能源 | 1.17 | 3.60 | 3.57 | **1.00** | 1.29 | 1.12 |
| 臭氧 | 1.36 | **1.05** | **1.00** | 1.37 | 1.18 | 1.22 |
| 住宅 | 1.20 | 1.08 | **1.01** | 116.31 | 1.99 | **1.00** |
| 核黄素 | 2.21 | **1.00** | 1.14 | 3.43 | 2.89 | 2.44 |
| 技能 | **1.00** | **1.02** | **1.03** | 1.06 | 1.17 | (1.06) |
| 超导体 | **1.00** | 2.11 | 2.12 | ** | ** | (9.74) |
| 温度 | 1.07 | 1.48 | 1.54 | 1.36 | 2.19 | **1.00** |
| 沃尔玛 | **1.00** | 1.36 | 1.37 | 11.96 | 12.61 | 1.30 |
| Wine | **1.00** | 1.11 | 1.11 | 1.07 | 1.08 | 1.35 |
| Mean | **1.18** | 1.73 | 1.74 | 8.08 | 2.09 | (2.88) |
| Std | **0.33** | 1.03 | 1.02 | 26.33 | 2.59 | (3.32) |

每个条目表示该方法在该数据集上的均方误差除以该行中的最低均方误差。值不超过 1.05 的条目以 蓝色显示。**：超过预设的最长墙钟时间。（xxx）：M5 结果基于重新缩放后的预测变量。

`Riboflavin` 数据的预测值相对较差。这组数据具有高维特征，观测值仅有 78 个，但预测变量超过 7000 个，而且其响应变量与回归变量之间存在线性关系。这解释了为什么岭回归和套索回归在这些数据上表现如此出色。基于树的方法受训练集中样本数量相对较少的影响，导致测试误差较高。尽管如此，PILOT 在这些数据上的表现仍优于 CART、FRIED 和 M5，这表明它确实不易过拟合，也更善于捕捉线性关系。

还有若干其他数据集，岭回归和/或套索回归在这些数据集上表现很好，说明数据具有线性模式。结果表明，PILOT 在这些数据集上往往表现相当出色，并且优于 CART。这证实了 PILOT 的线性组成部分使其非常适合这类数据；而对于其他数据集，我们看到它通常至少与其他基于树的方法表现相当。

## 4.3 转换预测变量后的结果

已有观察发现，如果数值预测变量的偏斜程度不太严重，线性回归方法往往表现更好。为处理可能存在偏斜的预测变量，我们对所有数值预测变量进行 Yeo-Johnson（YJ）变换预处理后，重新开展了实证实验（Yeo 和 Johnson，2000）。我们在训练集上通过最大似然法拟合 YJ 变换的参数，随后将拟合得到的变换参数应用于测试集。这种变换可能有助于缓解外推问题，因为外推问题预计主要出现在偏斜预测变量长尾的一侧，而变换会使这一侧变短。

表 3：转换预测变量后相对于最佳结果的均方误差比。

| | PILOT | 岭回归 | 套索回归 | FRIED | CART | M5 |
|---|---|---|---|---|---|---|
| 鲍鱼 | 0.98 | 1.02 | 1.00 | 1.21 | 1.12 | 1.08 |
| 入学 | 1.19 | 1.00 | 1.00 | 1.12 | 1.39 | 1.20 |
| 翼型 | 1.66 | 4.70 | 4.71 | 1.11 | 1.78 | 9.98 |
| Bike | 1.01 | 3.04 | 3.05 | 1.29 | 1.05 | 2.20 |
| 体脂 | 1.13 | 0.96 | 0.94 | 1.49 | 1.58 | 1.16 |
| 波士顿 | 1.06 | 0.93 | 0.96 | 1.09 | 1.12 | 5.97 |
| 加利福尼亚 | 1.03 | 1.87 | 1.87 | 1.15 | 1.43 | 24.45 |
| 社区 | 1.05 | 1.04 | 1.03 | 1.16 | 1.30 | 1.01 |
| 混凝土 | 0.95 | 1.24 | 1.24 | 1.23 | 1.39 | 2.48 |
| 糖尿病 | 1.09 | 1.01 | 1.01 | 1.22 | 1.32 | 1.20 |
| 电力 | 1.00 | 2.75 | 2.75 | 1.29 | 1.87 | 2.71 |
| 能源 | 1.34 | 3.84 | 3.86 | 1.01 | 1.29 | 1.11 |
| 臭氧 | 1.25 | 1.07 | 1.00 | 1.16 | 1.01 | 1.22 |
| 住宅 | 0.78 | 0.98 | 0.81 | 116.31 | 2.01 | 0.81 |
| 核黄素 | 1.53 | 1.01 | 1.28 | 3.57 | 3.00 | 2.65 |
| 技能 | 0.96 | 0.93 | 0.93 | 1.03 | 1.17 | 1.06 |
| 超导体 | 1.05 | 2.04 | 2.05 | *** | *** | 9.74 |
| 温度 | 1.10 | 1.53 | 1.61 | 1.42 | 2.17 | 1.60 |
| 沃尔玛 | 1.00 | 1.36 | 1.38 | 11.89 | 11.82 | 1.30 |
| Wine | 1.01 | 1.09 | 1.09 | 1.07 | 1.08 | 1.35 |
| Mean | **1.11** | 1.67 | 1.68 | 7.94 | 2.05 | 3.71 |
| Std | **0.20** | 1.08 | 1.09 | 26.36 | 2.42 | 5.59 |

每个条目表示该方法在该数据集上的均方误差，除以表 2 同一行中的最低均方误差所得的值。**：超过预设的最大墙钟时间。

表 3 展示了对预测变量进行变换后的结果。均方误差比是相对于表 2 中每行的最低均方误差计算的，因此表 3 中的数值可以直接与表 2 中的数值进行比较。我们可以看到，这种变换通常提升了 PILOT 以及其他一些方法的表现。在高维 `Riboflavin` 数据集上，PILOT 的得分更接近岭回归和套索回归，而其他决策树仍然

表现不佳。此外，在其他具有线性结构的数据集上，PILOT 的表现也与两种线性方法相当。经过变换后，其得分的平均值和标准差也有所降低。PILOT 的整体表现仍然最佳。

# 5 结论

本文介绍了一种名为 PILOT 的新型线性模型树算法。该算法计算效率高，其时间和空间复杂度与 CART 相同。PILOT 通过模型选择规则进行正则化，以避免过拟合。这种正则化不会增加计算复杂度，而且树构建完成后无需剪枝。因此，PILOT 比大多数现有线性模型树更快。由于其树结构以及加性的分段线性特性，预测结果易于解释。为了指导变量选择，可以像 CART 中那样定义特征重要性度量。PILOT 对预测值采用两种截断方法，以避免其他线性模型树有时出现的极端外推误差。实证研究发现，在多个数据集上，PILOT 的表现往往优于 CART、M5 和 FRIED 决策树。应用于大致呈线性的数据时，与其他基于树的方法相比，PILOT 的表现更接近高维线性方法，这表明它发现线性结构的能力更强。我们证明了 PILOT 在一类一般函数上的一致性理论保证。对于由线性模型生成的数据，PILOT 的收敛速率为多项式级。据我们所知，这是首个具有已证理论保证的线性模型树。

我们认为，PILOT特别适合那些既要求性能又要求可解释性的领域，例如医疗保健（Ahmad 等，2018年）、商业分析（Bohanec 等，2017年；Delen 等，2013年）和公共政策（Brennan 和 Oliver，2013年），在这些领域中，PILOT可以支持决策。

未来的一个研究方向是将 PILOT 树集成为集成方法中的基学习器，因为众所周知，基学习器的准确性和多样性有助于提升集成模型的性能。一方面，我们已经看到，PILOT能够在多个数据集上给出准确的预测值。另一方面，每个节点上可用的模型种类十分丰富，这使基学习器具有更高的多样性。基于这些原因，以及 PILOT 所需的计算成本较低这一事实，我们预计它将成为随机森林和梯度提升中一种很有前景的基学习器。

# 参考文献

Ankit Agrawal 和 Alok Choudhary。展望：材料信息学与大数据：材料科学中“科学第四范式”的实现。APL Materials，4（5）：053208，2016年。

Muhammad Aurangzeb Ahmad、Carly Eckert 和 Ankur Teredesai。《医疗保健中的可解释机器学习》。载于《2018年 ACM 生物信息学、计算生物学与健康信息学国际会议论文集》，第559–560页，2018年。

ˇ Marko Bohanec、Mirjana Kljaji´c Borˇstnar 和 Marko Robnik-Sikonja。《解释用于销售预测的机器学习模型》。Expert Systems with Applications，71：416–428，2017年。Claire Boryan、Zhengwei Yang、Rick Mueller 和 Mike Craig。《监测美国农业：美国农业部国家农业统计局耕地数据层计划》。Geocarto International，26(5)：341–358，2011年。Leo Breiman。《随机森林》。Machine learning，45(1)：5–32，2001年。Leo Breiman、Jerome H Friedman、Richard A Olshen 和 Charles J Stone。《分类回归树》。Routledge，1984年。Tim Brennan 和 William L Oliver。《犯罪学中机器学习技术的兴起：数据复杂性与研究问题复杂性的影响》。Criminology & Public Policy，12：551，2013年。Peter B¨uhlmann。《高维线性模型的提升》。The Annals of Statistics，34 (2)：559–583，2006年。Matias D Cattaneo、Rajita Chandak 和 Jason M Klusowski。《灵活函数库中斜回归树的收敛速率》。arXiv 预印本 arXiv:2210.14429，2022 年。Tianqi Chen 和 Carlos Guestrin。《 XGBoost：一种可扩展的树提升系统》。载于第22届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第785–794页，2016年。Dursun Delen、Cemil Kuzey 和 Ali Uyar。《使用财务比率衡量企业绩效：一种决策树方法》。Expert Systems with Applications，40(10)：3970–3983，2013年。Ruben Dezeure、Peter B¨uhlmann、Lukas Meier 和 Nicolai Meinshausen。《高维推断：置信区间、p值与 R 软件 hdi》。Statistical Science，30(4)：533–558，2015年。Alin Dobra 和 Johannes Gehrke。《SECRET：一种可扩展的线性回归树算法》。载于第8届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第481–487页，2002年。Dheeru Dua 和 Casey Graff。UCI机器学习知识库，2017年。网址 http://archive.ics.uci.edu/ml。Robert M Freund、Paul Grigas 和 Rahul Mazumder。《通过次梯度优化及相关方法重新审视线性回归中的提升》。The Annals of Statistics，45 (6)：2328–2364，2017年。Jerome H Friedman。《非参数多元回归的树结构方法》。载于 Th. Gasser 和 M. Rosenblatt 编，《曲线估计平滑技术》，第5–22页。Springer，1979年。

Jerome H Friedman. 多元自适应回归样条. The Annals of Statistics, 19 (1):1–67, 1991. Jennifer Golbeck, Cristina Robles, and Karen Turner. 利用社交媒体预测人格. 载于 CHI'11《计算机系统中的人为因素》扩展摘要, 第253–262页. ACM数字图书馆, 2011. L´aszl´o Gy¨orfi, Michael Kohler, Adam Krzyzak, and Harro Walk. 无分布的非参数回归理论. Springer统计学丛书, 2002. Alastair R Hall, Denise R Osborn, and Nikolaos Sakkas. 含多个断点模型中残差平方和的渐近行为. 计量经济学评论, 36: 667–698, 2017. Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. 统计学习要素：数据挖掘、推断与预测. Springer, 2009. Kurt Hornik, Christian Buchta, and Achim Zeileis. 开源机器学习：R 与 Weka 相遇. 计算统计学, 24(2):225–232, 2009. Julie Josse and Fran。cois Husson. missMDA：用于处理多元数据分析中缺失值的软件包. 统计软件杂志, 70(1):1–31, 2016. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM：一种高效的梯度提升决策树. 神经信息处理系统进展, 30, 2017. Jason M Klusowski. 高维决策树的普适一致性. arXiv 预印本 arXiv:2104.13881, 2021. Wei-Yin Loh. 具有无偏变量选择和交互作用检测的回归树. 中华统计学志 , 第 361–386 页 , 2002. Wei-Yin Loh, Chien-Wei Chen, and Wei Zheng. 线性模型树中的外推误差. ACM数据知识发现汇刊 (TKDD), 1(2):1–17, 2007. Donato Malerba, Floriana Esposito, Michelangelo Ceci, and Annalisa Appice. 带有回归节点和分裂节点的模型树自顶向下归纳. IEEE模式分析与机器智能汇刊, 26(5):612–625, 2004. Andreas Maurer and Massimiliano Pontil. 经验伯恩斯坦界与样本方差惩罚. arXiv 预印本 arXiv:0907.3740, 2009. Bengt Muthen. 删失与截断二元正态分布的矩. 英国数学与统计心理学杂志, 43(1):131–143, 1990. Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 挖掘用户移动特征以预测基于位置的服务中的下一地点. 载于 2012年IEEE第12届国际数据挖掘会议, 第1038–1043页. IEEE, 2012.

Ashutosh Patri and Yugesh Patnaik. 基于随机森林和随机梯度树提升的翼型自噪声预测方法. Procedia Computer Science, 46: 109–121, 2015. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon- del, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn：Python 机器学习. Journal of Machine Learning Research, 12:2825–2830, 2011. Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 我们的 Twitter 个人资料就是我们自己：利用 Twitter 预测人格. 载于 2011年IEEE第三届隐私、安全、风险与信任国际会议和2011年IEEE第三届社会计算国际会议, 第180–185页. IEEE, 2011. J R Quinlan. 连续类别学习. 载于 第五届澳大利亚人工智能联合会议, 第92卷, 第343–348页. World Scientific, 1992. J R Quinlan. C 4.5：机器学习程序. Morgan Kaufmann机器学习系列, 1993. Erwan Scornet, G´erard Biau, and Jean-Philippe Vert. 随机森林的一致性. The Annals of Statistics, 43(4):1716–1741, 2015. Shahabbodin Shamshirband, Sajjad Hashemi, Hana Salimi, Saeed Samadianfard, Esmaeil Asadi, Sadra Shadkani, Katayoun Kargar, Amir Mosavi, Narjes Nabipour, and Kwok-Wing Chau. 使用机器学习模型预测水文干旱的标准化径流指数. Engineering Applications of Computational Fluid Mechanics, 14(1):339–350, 2020. Yu Shi, Jian Li, and Zhize Li. 基于分段线性回归树的梯度提升. 载于 第二十八届人工智能国际联合会议论文集, IJCAI'19, 第3432–3438页. AAAI Press, 2019. ISBN 9780999241141. Gregor Stiglic, Simon Kocbek, Igor Pernek, and Peter Kokol. 生物信息学中的综合决策树模型. PLOS ONE, 7(3):e33812, 2012. Robert Tibshirani. 通过套索回归进行回归收缩与变量选择. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996. Joel A Tropp 等人. 矩阵集中不等式导论. Foundations and Trends® in Machine Learning, 8(1-2):1–230, 2015. David S Vogel, Ognian Asparouhov, and Tobias Scheffer. 可扩展的前瞻线性回归树. 载于 第十三届ACM SIGKDD知识发现与数据挖掘国际会议论文集, 第757–764页, 2007. In-Kwon Yeo and Richard A Johnson. 一类用于改善正态性或对称性的新幂变换. Biometrika, 87(4):954–959, 2000.

**附录**

# A 初步结果

本节给出第3节定理所需的记号和初步结果。

## A.1 记号

我们沿用第2节和第3节中的记号。 $n$ 响应值构成向量 $Y$，其均值为 $\bar{Y}$。一个样本中各预测变量的取值合并为 $X \in \mathbb{R}^p$。所有 $n$ 个样本的设计矩阵记为 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$。给定一个包含 $t$ 个样本的树节点 $T$，我们将受限数据矩阵记为 $\boldsymbol{X}_T := (X_{T_1}、\ .....、\ X_{T_t})^\top \in \mathbb{R}^{t\times p}$，其第 $j$ 列记为 $X_T^{(j)}$。第 $j$ 列的方差记为 $\hat{\sigma}^2_{j,\ T} := \sum_{k=1}^{t}(X_{T_k}^{(j)} - X_T^{(j)})^2/t$。我们还将分母为 $t-1$ 的经典无偏方差估计记为 $(\hat{\sigma}^U_{j,\ T})^2$。当 $T$ 为根节点或不存在歧义时，这些符号中省略 $T$。

我们用 $\mathcal{T}_k$ 表示深度为 $k$ 的节点集合，以及深度小于 $k$ 的叶节点。记这些节点上的 PILOT 预测为 $\hat{}\ f(\mathcal{T}_k) \in \mathbb{R}^n$，而 $\hat{}\ f_T^k$ 表示选定模型（例如 PLIN）在某些 $T \in \mathcal{T}_k$ 上的预测，该预测通过拟合上一层的残差 $\hat{}Y - f(\mathcal{T}_{k-1})$ 得到。具有 $t$ 个样本的节点 $T$ 上模型的不纯度增益可写为

$$\widehat{\Delta}^{k+1}(T) = ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - P(t_l)||Y - \hat{f}(\mathcal{T}_k) - \hat{f}_{T_l}^{k+1}||_{t_l}^2 - P(t_r)||Y - \hat{f}(\mathcal{T}_k) - \hat{f}_{T_r}^{k+1}||_{t_r}^2,$$

其中，$T_l$ 和 $T_r$ 是左、右子节点，分别包含 $t_l$ 和 $t_r$ 个样本，且 $P(t_l) = t_l/t$、$P(t_r) = t_r/t$。

由于 $f(\mathcal{T}_k)$ 是加性的，并且我们假设所有 $f \in \mathcal{F}$ 也都是加性的，因此可以将第 $j$ 个预测变量上的函数 $\hat{}$ 写为 $f_j(\mathcal{T}_k)$ 和 $f_j$。函数 $f$ 在 $T$ 上的全变差记为 $||f||_{TV(T)}$。对于依赖于 $n \to \infty$ 的非零值 $A_n$ 和 $B_n$，当且仅当 $A_n/B_n \le \mathcal{O}(1)$ 时，我们记为 $A_n \precsim B_n$；$\succsim$ 的定义类似。若 $A_n/B_n = \mathcal{O}(1)$，则记为 $A_n \asymp B_n$。集合 $C_1$ 与 $C_2$ 的差记为 $C_1 \backslash C_2$。

## A.2 增益与残差的表示

本节证明， LIN 和 PCON 的不纯度增益可以写成一个内积的平方。我们还将说明， PLIN 可以看作这两个模型的组合。这些公式构成了证明 PILOT 的一致性及其在线性模型上的收敛速率的基础。

**命题 8.** 设 $\beta$ 是一个线性模型的最小二乘估计量，该模型以设计矩阵 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ 为自变量拟合 $Y \in \mathbb{R}^p$。则有

$$\langle Y, \boldsymbol{X}\hat{\beta}\rangle_n = ||\boldsymbol{X}\hat{\beta}||_n^2 . \tag{9}$$

证明。 这由 $\hat{}\ \beta = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top Y$ 以及 $\langle Y, \boldsymbol{X}\hat{}\beta\rangle_n = Y^\top\boldsymbol{X}\hat{}\beta/n$ 得出。 □

**引理 9 (LIN 的表示)。** 设 $T$ 为一个包含 $t$ 个样本的节点。对于变量 $X^{(j)}$ 上的 LIN 模型，有：

$$\widehat{\Delta}_{lin}^{k+1} = \left| \left\langle Y - \hat{f}(\mathcal{T}_k), \frac{X_T^{(j)} - \overline{X_T^{(j)}}}{\hat{\sigma}_{j,T}} \right\rangle_t \right|^2 + \left( \overline{Y} - \overline{\hat{f}(\mathcal{T}_k)} \right)^2 . \tag{10}$$

此外，如果我们对最后一个表达式中的第二项进行归一化，并记为

$$\widetilde{f}_{lin}^{T} := \frac{X^{(j)} - \overline{X_T^{(j)}}}{\sqrt{w(t)}\hat{\sigma}_{j,T}} \mathbb{1}\{X \in T\}$$

其中 $w(t) = t/n$，我们有

$$\hat{f}_T^{k+1} = \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{lin}^{T} \rangle_n \, \widetilde{f}_{lin}^{T} + \overline{Y - \hat{f}(\mathcal{T}_k)} \, .$$

证明。 根据命题 8，我们有

$$\begin{aligned} \widehat{\Delta}_{lin}^{k+1} &= ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - ||(Y - \hat{f}(\mathcal{T}_k)) - (\hat{\alpha} + \hat{\beta} X_T^{(j)})||_t^2 \\ &= 2\langle Y - \hat{f}(\mathcal{T}_k), \hat{\alpha} + \hat{\beta} X_T^{(j)} \rangle_t - ||\hat{\alpha} + \hat{\beta} X_T^{(j)}||_t^2 \\ &= \langle Y - \hat{f}(\mathcal{T}_k), \hat{\alpha} + \hat{\beta} X_T^{(j)} \rangle_t \, . \end{aligned}$$

ˆ对于 $\beta$，我们无需考虑残差的均值，因此

$$\hat{\beta} = \frac{(X_T^{(j)} - \overline{X_T^{(j)}})^\top (Y - \hat{f}(\mathcal{T}_k))}{(X_T^{(j)} - \overline{X_T^{(j)}})^\top (X_T^{(j)} - \overline{X_T^{(j)}})} = \frac{\langle Y - \hat{f}(\mathcal{T}_k), X_T^{(j)} - \overline{X_T^{(j)}} \rangle_t}{\hat{\sigma}_{j,T}^2} \, .$$

另一方面，ˆ$\alpha = Y -$ ˆ$f(\overline{\mathcal{T}_k)} -$ ˆ$\beta \overline{X_T^{(j)}}$，因此

$$\begin{aligned} \widehat{\Delta}_{lin}^{k+1} &= \hat{\beta} \langle Y - \hat{f}(\mathcal{T}_k), X^{(j)} - \overline{X^{(j)}} \rangle_t + \left( \overline{Y} - \overline{\hat{f}(\mathcal{T}_k)} \right)^2 \\ &= \left| \left\langle Y - \hat{f}(\mathcal{T}_k), \frac{X^{(j)} - \overline{X^{(j)}}}{\hat{\sigma}_{j,T}} \right\rangle_t \right|^2 + \left( \overline{Y} - \overline{\hat{f}(\mathcal{T}_k)} \right)^2 \, . \end{aligned}$$

Th ˆ第二个结果由ˆ$\alpha, \beta$ 的定义及上述公式得出。 □

ˆ 接下来，我们将针对 PCON 给出一个类似的结果。请注意，这里我们的 $f(\mathcal{T}_k)$ 在每个节点上并非常数，因此我们需要将 Klusowski（2021）中关于 CART 的结果调整到我们的情形。

**引理 10 (PCON 的表示)。** 设 $T$ 是一个包含 $t$ 个样本的节点，$T_l$ 和 $T_r$ 分别是它的左子节点和右子节点。于是我们有

$$\widehat{\Delta}_{pcon}^{k+1} = \left| \left\langle Y - \hat{f}(\mathcal{T}_k), \frac{\mathbb{1}\{X_T \in T_l\} t_r - \mathbb{1}\{X_T \in T_r\} t_l}{\sqrt{t_r t_l}} \right\rangle_t \right|^2 + \left( \overline{Y} - \overline{\hat{f}(\mathcal{T}_k)} \right)^2 .$$

此外，如果我们对内积中的第二项进行归一化，并记 $e$

$$\widetilde{f}^T_{pcon} := \frac{\mathbb{1}\{X_T \in T_l\}P(t_r) - \mathbb{1}\{X_T \in T_r\}P(t_l)}{\sqrt{w(t)P(t_r)P(t_l)}},$$

其中 $w(t) = t/n$，我们有

$$\hat{f}^{k+1}_T = \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}^T_{pcon}\rangle_n \, \widetilde{f}^T_{pcon} + \overline{Y - \hat{f}(\mathcal{T}_k)} \, .$$

证明。 根据 PCON 的定义，我们有

$$\hat{f}^{k+1} = (\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l)\mathbb{1}(X_T \in T_l) + (\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)\mathbb{1}(X_T \in T_r)$$

^^其中 $Y_l$ 是 $T_l$ 中 $Y$ 的均值，$\overline{f(\mathcal{T}_k)}_l$ 是 $T_l$ 中 $f(\mathcal{T}_k)$ 的均值，右子节点中的样本也同理。下文中，我们还将左右节点的常数预测分别记为 $\hat{f}^{k+1}_l := \hat{f}^{k+1}\mathbb{1}(X_T \in T_l)$ 和 $\hat{f}^{k+1}_r := \hat{f}^{k+1}\mathbb{1}(X_T \in T_r)$。因此我们有

$$\begin{aligned}
\widehat{\Delta}^{k+1}_{pcon} &= \frac{1}{t}\Big[2\sum_{X_T\in T_l}(Y - \hat{f}(\mathcal{T}_k))\hat{f}^{k+1} - \sum_{X_T\in T_l}(\hat{f}^{k+1})^2 \\
&\quad + 2\sum_{X_T\in T_r}(Y - \hat{f}(\mathcal{T}_k))\hat{f}^{k+1} - \sum_{X_T\in T_r}(\hat{f}^{k+1})^2\Big] \\
&= \frac{1}{t}[2t_l(\hat{f}^{k+1}_l)^2 - t_l(\hat{f}^{k+1}_l)^2 + 2t_r(\hat{f}^{k+1}_r)^2 - t_r(\hat{f}^{k+1}_r)^2] \\
&= \frac{1}{t}[t_l\left(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l\right)^2 + t_r\left(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r\right)^2] \\
&= \frac{t_l t_r}{t^2}\Big[\left(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l\right)^2 + \left(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r\right)^2 + \frac{t_l}{t_r}\left(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l\right)^2 + \frac{t_r}{t_l}\left(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r\right)^2\Big] \\
&\overset{(i)}{=} \frac{t_l t_r}{t^2}\Big[\left(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l\right)^2 + \left(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r\right)^2 - 2(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l)(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)\Big] \\
&\quad + \left(\overline{Y} - \overline{\hat{f}(\mathcal{T}_k)}\right)^2 \\
&= \frac{t_l t_r}{t^2}\Big[(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l) - (\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)\Big]^2 + \left(\overline{Y} - \overline{\hat{f}(\mathcal{T}_k)}\right)^2
\end{aligned}$$

其中，(i) 可由以下事实得出：

$$\begin{aligned}
&\frac{t_l}{t_r}\left(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l\right)^2 + \frac{t_r}{t_l}\left(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r\right)^2 + 2(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l)(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r) \\
&\qquad = \frac{1}{t_l t_r}\Big(\sum_{X_T\in T_l}(Y - \hat{f}(\mathcal{T}_k)) + \sum_{X_T\in T_r}(Y - \hat{f}(\mathcal{T}_k))\Big)^2 \\
&\qquad = \frac{t^2}{t_l t_r}\left(\overline{Y} - \overline{\hat{f}(\mathcal{T}_k)}\right)^2 .
\end{aligned}$$

因此，

$$\begin{aligned}\widehat{\Delta}_{pcon}^{k+1} &= \left|\frac{1}{t}\frac{t_l t_r(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l) - t_l t_r(\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)}{\sqrt{t_l t_r}}\right|^2 + \left(\overline{Y} - \overline{\hat{f}(\mathcal{T}_k)}\right)^2 \\ &= \left|\left\langle Y - \hat{f}(\mathcal{T}_k), \frac{\mathbb{1}\{X_T \in T_l\}t_r - \mathbb{1}\{X_T \in T_r\}t_l}{\sqrt{t_r t_l}}\right\rangle_t\right|^2 + \left(\overline{Y} - \overline{\hat{f}(\mathcal{T}_k)}\right)^2\end{aligned}$$

最后，我们得到

$$\begin{aligned}&\langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{pcon}^T\rangle_n \widetilde{f}_{pcon}^T + \overline{Y - \hat{f}(\mathcal{T}_k)} \\ &= \frac{1}{t}[(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l) - (\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)](\mathbb{1}\{X_T \in T_l\}t_r - \mathbb{1}\{X_T \in T_r\}t_l) + \overline{Y - \hat{f}(\mathcal{T}_k)} \\ &= \frac{t_l + t_r}{t}[(\overline{Y}_l - \overline{\hat{f}(\mathcal{T}_k)}_l)\mathbb{1}\{X_T \in T_l\} + (\overline{Y}_r - \overline{\hat{f}(\mathcal{T}_k)}_r)\mathbb{1}\{X_T \in T_r\}] \\ &\quad - \overline{Y - \hat{f}(\mathcal{T}_k)}(\mathbb{1}\{X_T \in T_l\} + \mathbb{1}\{X_T \in T_r\}) + \overline{Y - \hat{f}(\mathcal{T}_k)} \\ &= \hat{f}_T^k\end{aligned}$$

□

注意， PLIN 等价于前两个模型的组合（PCON 后接两个 LIN 模型）。因此，我们可以类似地展开其预测。

**命题 11 (PLIN 的表示)。** 对于 PLIN，存在一个依赖于 $X$、 $T$、 $T_l$ 和 $T_r$ 的函数 $\tilde{f}_{plin}^T$，使得

$$\hat{f}_T^{k+1} = \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{plin}^T\rangle_n \, \widetilde{f}_{plin}^T + \overline{Y - \hat{f}(\mathcal{T}_k)}$$

证明。 由于 PLIN 可视为一个 PCON 后接两个基于同一预测变量进行拟合的 LIN，因此我们有

$$\begin{aligned}\hat{f}_T^{k+1} &= \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{pcon}^T\rangle_n \, \widetilde{f}_{pcon}^T + \langle Y - \hat{f}(\mathcal{T}_k) - \hat{f}_{pcon}, \widetilde{f}_{lin}^{T_l}\rangle_n \, \widetilde{f}_{lin}^{T_l} + \\ &\quad + \langle Y - \hat{f}(\mathcal{T}_k) - \hat{f}_{pcon}, \widetilde{f}_{lin}^{T_r}\rangle_n \, \widetilde{f}_{lin}^{T_r} + \overline{Y - \hat{f}(\mathcal{T}_k)} \\ &= \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{pcon}^T\rangle_n \, \widetilde{f}_{pcon}^T + \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{lin}^{T_l}\rangle_n \, \widetilde{f}_{lin}^{T_l} + \langle Y - \hat{f}(\mathcal{T}_k), \widetilde{f}_{lin}^{T_r}\rangle_n \, \widetilde{f}_{lin}^{T_r} \\ &\quad + \overline{Y - \hat{f}(\mathcal{T}_k)}\end{aligned}$$

~其中，最后一个等式由 $\langle 1, f_{lin}\rangle_n = 0$ 以及 PCON的预测在两个子节点上为常数这一事实得出。此外，由于 $\widetilde{f}_{lin}^{T_r}$、 $\widetilde{f}_{lin}^{T_l}$、 $\widetilde{f}_{pcon}^T$彼此正交，我们可以推得

$$\begin{aligned}\hat{f}^{k+1} = \langle Y - \hat{f}(\mathcal{T}_k), A\widetilde{f}_{pcon}^T + B\widetilde{f}_{lin}^{T_l} + C\widetilde{f}_{lin}^{T_r}\rangle_n \, (A\widetilde{f}_{pcon}^T + B\widetilde{f}_{lin}^{T_l} + C\widetilde{f}_{lin}^{T_r}) \\ + \overline{Y - \hat{f}(\mathcal{T}_k)}\end{aligned}$$

对于某个 $A$， $B$，以及依赖于 $X$、节点 $T$、 $T_l$ 和 $T_r$ 的 $C$。 □

# B PILOT 普适一致性的证明

本节将证明 PILOT 对一般加性模型的一致性。如第3节所述，我们首先估计 PILOT 的不纯度增益。然后推导出一个递推公式，据此建立预言机不等式，并在最后证明一致性。对于所有中间结果，我们假设定理 2 中的条件成立。

## B.1 $\Delta_{pcon}$的下界

一致性证明的关键在于将不纯度增益 $\Delta^{k+1}$ 与训练误差$R_k := ||Y - \hat{f}(\mathcal{T}_k)||_n^2 - ||Y - f||_n^2$ 联系起来，从而可以推导出关于 $R_k$ 的递推公式。最近，Klusowski（2021）为 CART 推导了这样的递推公式。他在引理7.1中证明，对于 CART， $\Delta^{k+1}$ 有一个下界 $R_k^2$，相差若干因子。然而，该证明用到了 $\langle Y - Y_t, Y_t \rangle = 0$这一事实；当 $Y_t$ 被 $f(\mathcal{T}_k)$替代时，这个公式便不再成立。原因在于，由于较早的节点，我们无法假设 PILOT 树叶节点中的预测值等于该节点中的响应变量均值。为了解决这一问题，我们将在下文给出关于 PCON 的推广结果。

**引理 12.** 假设 $R_k(T) > 0$ 位于某个节点 $T$中，则 PCON 在该节点上的不纯度增益满足：

$$\hat{\Delta}_{pcon}^{k+1}(\hat{s}, \hat{j}, T) \geqslant \frac{R_k^2(T)}{(||f||_{TV} + 6B)^2}$$

其中，$\hat{s}, \hat{j}$ 是最优分裂点和预测。

证明。 在整个证明过程中，计算均针对单个节点 $T$，因此为方便记号，有时省略证明中的 $T$。我们首先考虑节点中先前残差的均值为零的情形。我们将证明

$$\hat{\Delta}_{pcon}^{k+1}(\hat{s}, \hat{j}, t) \geqslant \frac{|\langle Y - \hat{f}(\mathcal{T}_k), f - \hat{f}(\mathcal{T}_k) \rangle_t|^2}{4||f - \hat{f}(\mathcal{T}_k)||_{TV}^2} .$$

注意到 $\hat{f}(\mathcal{T}_k)$ 是由定义在 $p_n$ 个预测变量上的 $p_n$ 个线性函数之和。设 $\beta_j$ 为第 $j$个预测变量对应的线性函数的斜率。接着，我们在空间 $\mathbb{R} \times \{1, \ldots, p_n\}$ 上定义概率测度 $\Pi(s, j)$，其拉东–尼科迪姆导数如下：

$$\frac{d\Pi(s,j)}{d(s,j)} = \frac{|f_j'(s) - a_j| \sqrt{P(t_l)P(t_r)}}{\sum_{j=1}^{p_n} \int |f_j'(s') - a_j| \sqrt{P(t_l')P(t_r')} ds'} .$$

这里，$(t_l,\ t_r)$ 和 $(t_l',\ t_r')$ 是由分裂点 $s$ 和预测变量 $j$ 产生的子节点，而 $a_j$ 和 $f_j'(x)$ 定义为

$$a_j = \hat{\beta}_j, \quad f_j'(s) = \begin{cases} \dfrac{f_j(X_{(i)}^{(j)}) - f_j(X_{(i-1)}^{(j)})}{X_{(i)}^{(j)} - X_{(i-1)}^{(j)}} & X_{(i-1)}^{(j)} < s < X_{(i)}^{(j)} \\ 0 & \text{otherwise} \end{cases}$$

其中， $X^{(j)}{}_{(1)} \dots X^{(j)}{}_{(t)}$ 是沿第 $j$ 个预测变量排序的数据点 .

其思路是，最优不纯度增益 $\hat{}\,\Delta_{pcon}(\hat{}s, \hat{}j, t)$ 始终大于相对于该概率测度的不纯度增益平均值:

$$\begin{aligned}\hat{\Delta}_{pcon}(\hat{s},\hat{j},t) &\geqslant \int \hat{\Delta}(s,j,t)d\Pi(s,j)\\ &= \int |\langle Y-\hat{f}(\mathcal{T}_k), \sqrt{w(t)}\widetilde{f}^T_{pcon}\rangle_t|^2 d\Pi(s,j)\\ &\geqslant \left(\int |\langle Y-\hat{f}(\mathcal{T}_k), \hat{f}_t\rangle_t| d\Pi(s,j)\right)^2\end{aligned}$$

其中，最后一个不等式由詹森不等式得出，并使用了引理 10 中的表示。现在我们聚焦于方括号内的项，并得到 Klusowski（2021）中式 (29) 的一个类似结果:

$$\int |\langle Y-\hat{f}(\mathcal{T}_k), \hat{f}_t\rangle_t| d\Pi(s,j) \geqslant \frac{|\langle Y-\hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n}\int (f_j'(s)-a_j)\mathbb{1}_{\{X^{(j)}>s\}}ds\rangle_t|}{\sum_{j=1}^{p}\int |f_j'(s')-a_j|\sqrt{P_{t_l'}P_{t_r'}}ds'}.$$

对于分子，我们有

$$\begin{aligned}&\left\langle Y-\hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n}\int (f_j'(s)-a_j)\mathbb{1}_{\{X^{(j)}>s\}}ds\right\rangle_t\\ \overset{(i)}{=}&\left\langle Y-\hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n}\int (f_j'(s)-a_j)\mathbb{1}_{\{X^{(j)}>s\}}ds + \sum_{j=1}^{p_n}(f_j(X^{(j)}_{(1)}) - \hat{f}_j(\mathcal{T}_k)|_{x=X^{(j)}_{(1)}})\right\rangle_t\\ =&\left\langle Y-\hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n}\left(f_j(X^{(j)}_{(1)}) + \sum_{X^{(j)}_{(2)}\leqslant X^{(j)}_{(i)}\leqslant X^{(j)}} \frac{f_j(X^{(j)}_{(i)}) - f_j(X^{(j)}_{(i-1)})}{X^{(j)}_{(i)} - X^{(j)}_{(i-1)}}(X^{(j)}_{(i)} - X^{(j)}_{(i-1)})\right.\right.\\ &\left.\left. - \hat{f}_j(\mathcal{T}_k)|_{x=X^{(j)}_{(1)}} - a_j(X^{(j)} - X^{(j)}_{(1)})\right)\right\rangle_t\\ =&\left\langle Y-\hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n}(f_j - \hat{f}_j(\mathcal{T}_k))\right\rangle_t\\ =&\left\langle Y-\hat{f}(\mathcal{T}_k), f-\hat{f}(\mathcal{T}_k)\right\rangle_t\end{aligned}$$

$\hat{}$ 其中 (i) 再次使用了 $\langle 1, Y-f(\mathcal{T}_k)\rangle_t = 0$。

对于仅限一个预测变量的分母，我们有

$$\begin{aligned}&\int |f_j'(s')-a_j|\sqrt{P_{t_l'}P_{t_r'}}ds'\\ =&\sum_{i=0}^{N(t)}\int_{N(t)P'(t_l')=i} |f_j'(s')-a_j|\sqrt{(i/N(t))(1-i/N(t))}ds'\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N(t)-1} \int_{X_{(i)}^{(j)}}^{X_{(i+1)}^{(j)}} |f_j'(s') - a_j| ds' \sqrt{(i/N(t))(1-i/N(t))} \\
&= \sum_{i=1}^{N(t)-1} |f_j(X_{(i+1)}^{(j)}) - a_j X_{(i+1)}^{(j)} - f_j(X_{(i)}^{(j)}) + a_j X_{(i)}^{(j)}| \sqrt{(i/N(t))(1-i/N(t))} \\
&\leqslant \frac{1}{2} \sum_{i=1}^{N(t)-1} |f_j(X_{(i+1)}^{(j)}) - a_j X_{(i+1)}^{(j)} - f_j(X_{(i)}^{(j)}) + a_j X_{(i)}^{(j)}| \\
&= \frac{1}{2} \sum_{i=1}^{N(t)-1} |(f_j(X_{(i+1)}^{(j)}) - \hat{f}_j(\mathcal{T}_k)|_{x=X_{(i+1)}^{(j)}}) - (f_j(X_{(i)}^{(j)}) - \hat{f}_j(\mathcal{T}_k)|_{x=X_{(i)}^{(j)}})| \\
&= \frac{1}{2} ||f_j - \hat{f}(\mathcal{T}_k)_j||_{TV}(T) \,.
\end{aligned}
$$

如果我们对所有 $j$ 预测变量求和，就有

$$
\sum_{j'=1}^{p_n} \int |f_j'(s') - a_j| \sqrt{P_{t_L'} P_{t_R'}} ds' \leqslant \frac{1}{2} ||f - \hat{f}(\mathcal{T}_k)||_{TV(T)} \,.
$$

将分子和分母的界结合起来，我们得到：

$$
\begin{aligned}
\int |\langle Y - \hat{f}(\mathcal{T}_k), \hat{f}_t \rangle_t| d\Pi(s,j) \geqslant & \frac{|\langle Y - \hat{f}(\mathcal{T}_k), \sum_{j=1}^{p_n} \int (f_j'(s) - a_j) \mathbb{1}_{\{X^{(j)} > s\}} ds \rangle_t|}{\sum_{j=1}^{p} \int |f_j'(s') - a_j| \sqrt{P_{t_l'} P_{t_r'}} ds'} \\
\geqslant & \frac{\left\langle Y - \hat{f}(\mathcal{T}_k), f - \hat{f}(\mathcal{T}_k) \right\rangle_t}{\frac{1}{2} ||f - \hat{f}(\mathcal{T}_k)||_{TV(T)}} \,.
\end{aligned}
$$

现在注意到

$$
\begin{aligned}
|\langle Y - \hat{f}(\mathcal{T}_k), f - \hat{f}(\mathcal{T}_k) \rangle_t| = & |\langle Y - \hat{f}(\mathcal{T}_k), Y - \hat{f}(\mathcal{T}_k) \rangle_t + \langle Y - \hat{f}(\mathcal{T}_k), f - Y \rangle_t| \\
\overset{(i)}{\geqslant} & ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - |\langle Y - \hat{f}(\mathcal{T}_k), f - Y \rangle_t| \\
\geqslant & ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - ||Y - f||_t ||Y - \hat{f}(\mathcal{T}_k)||_t \\
\geqslant & \frac{1}{2} ||Y - \hat{f}(\mathcal{T}_k)||_t^2 - \frac{1}{2} ||Y - f||_t^2 \\
= & \frac{1}{2} R_k(t) \overset{(ii)}{>} 0
\end{aligned}
$$

其中，(i) 是三角不等式，(ii) 则源于假设 $R_k(T)$ 严格为正。

最后，我们处理 $\|f\|_{TV(T)}$。我们有 $\|f\|_{TV(T)} \leqslant \|f - \hat{f}(\mathcal{T}_k)\|_{TV(T)} + \|\hat{f}(\mathcal{T}_k)\|_{TV(T)}$。因此，接下来需要界定节点 $T$ 上 $\hat{f}(\mathcal{T}_k)$ 的全变差。现在我们证明，沿所有预测变量的全变差之和有界于 $(\max \hat{f}(\mathcal{T}_k) - \min \hat{f}(\mathcal{T}_k)) \leqslant 6B$。由于 $\hat{f}(\mathcal{T}_k)$ 是线性的，最大值

最小值总是在节点（或立方体）的顶点处取得。因此，不失一般性，我们可以假设立方体为 $[0,1]^{p_n}$ 且 $\hat{}\ f(\mathcal{T}_k) = \sum_{j=1}^{p_n} \hat{}\beta_j X^{(j)}$，使得 $\hat{}\ \hat{}\ \hat{}\forall j < j'$, $\beta_j < \beta_{j'}$ 和 $\beta_j < 0 \Leftrightarrow j < p_0$。现在我们从顶点 $v_0 \in [0,1]^{p_n}$ 开始，使得当且仅当 $j < p_0$ 时，其第 $j$ 个变量为 1（否则为 0）。然后，我们沿着边移动到另一个顶点 $v_1$，该顶点除了第一项外与 $v_0$ 相同，即其第一项为 0。此外，$\hat{}\ \hat{}\ \hat{}\ ||f_1(\mathcal{T}_k)||_{TV} = f(\mathcal{T}_k)|_{X=v_1} - f(\mathcal{T}_k)|_{X=v_0}$。类似地，对于 $\forall j$，我们令顶点$v_{j+1}$ 除第 $j$ 个变量外与 $v_j$ 相同（将 1 改为 0，或将 0 改为 1），并且有$||\hat{}f_j(\mathcal{T}_k)||_{TV} = \hat{}f(\mathcal{T}_k)|_{X=v_j} - \hat{}f(\mathcal{T}_k)|_{X=v_{j-1}}$。此外，有 $\sum_{j=1}^{p_n} ||\hat{}f(\mathcal{T}_k)_j||_{TV} = \max \hat{}f(\mathcal{T}_k) - \min \hat{}f(\mathcal{T}_k) \leqslant 6B$。因此，我们得到 $\left(\frac{1}{2}||f - \hat{}f(\mathcal{T}_k)||_{TV(T)}\right)^2 \leqslant \frac{1}{4}(||f||_{TV} + 6B)^2$。代入

g everything tog以太 y域

$$\hat{\Delta}_{pcon}^{k+1}(\hat{s}, \hat{j}, t) \geqslant \frac{|\langle Y - \hat{f}(\mathcal{T}_k), f - \hat{f}(\mathcal{T}_k)\rangle_t|^2}{\frac{1}{4}||f - \hat{f}(\mathcal{T}_k)||_{TV}^2} \geqslant \frac{R_k^2(T)}{(||f||_{TV} + 6B)^2}.$$

最后，我们讨论残差在节点中的均值不一定为零的一般情形（这可能发生在拟合 BLIN 模型之后）。注意，$R_k(T) = \{||Y - \hat{}f(\mathcal{T}_k) - (Y - \hat{}f(\mathcal{T}_k))||_t^2 - ||Y - f||_t^2\} + (Y - \hat{}f(\mathcal{T}_k))^2 := R_{k(1)} + R_{k(2)}$，其中 $R_{k(1)}$ 是回归掉均值后的平方训练误差，$R_{k(2)}$ 是平方均值。由于第一个截断过程确保 $R_k(T) \leqslant 16B^2$，我们有 $(||f||_{TV} + 6B)^2 \geqslant 36B^2 \geqslant 2R_k(T) = 2(R_{k(1)} + R_{k(2)})$，因此

$$\frac{R_{k(1)}^2}{(||f||_{TV} + 6B)^2} + R_{k(2)} \geqslant \frac{(R_{k(1)} + R_{k(2)})^2}{(||f||_{TV} + 6B)^2}.$$

因此，根据前述结果以及 PCON 使残差的均值为零这一事实，我们有

$$\Delta_{pcon}^{k+1}(\hat{s}, \hat{j}, t) \geq \frac{R_{k(1)}^2}{(||f||_{TV} + 6B)^2} + R_{k(2)} \geqslant \frac{R_k(T)^2}{(||f||_{TV} + 6B)^2}.$$

□

## B.2 命题4的证明

由于 PLIN 泛化了 PCON，我们总有 $\hat{}\ \Delta_{plin}^k \geqslant \hat{}\Delta_{pcon}^k$。然而，对于其他模型，无法像引理 12 中那样，通过构造精心设计的概率测度来建立下界。这是因为我们没有一个指示函数，在与积分中的 $f'$ 相关联时，会自然地产生 $f$（参见分子的估计）。幸运的是，我们可以依靠 BIC 准则来获得不同模型之间相对不纯度增益的界。更准确地说，命题 4 表明，如果选择了 BLIN或 LIN，其增益必须与 PCON的增益相当（即相差至多一个常数因子）。这确保了不会发生欠拟合。此外，如果在某个节点选择了 CON，那么该节点上所有后续模型也都会是 CON，这证明了使用 CON停止规则的合理性。证明如下。

证明。 不失一般性，我们可以假设没有模型能完美拟合。先来看第一个结论。根据 $BIC_1 > BIC_2$ 这一假设，对于任意 $v_1$，$v_2$，我们有

$$
\begin{aligned}
BIC_2 &< BIC_1 \\
t\log\left(\frac{R_0 - t\Delta_2}{t}\right) + v_2\log(t) &< t\log\left(\frac{R_0 - t\Delta_1}{t}\right) + v_1\log(t) \\
\log\left(1 - \frac{t\Delta_2}{R_0}\right) - \log\left(1 - \frac{t\Delta_1}{R_0}\right) &< (v_1 - v_2)\frac{\log t}{t} \\
\left(1 - \frac{t\Delta_2}{R_0}\right)\Big/\left(1 - \frac{t\Delta_1}{R_0}\right) &< \exp\left((v_1 - v_2)\frac{\log t}{t}\right) \\
\frac{t\Delta_2}{R_0} &> 1 + \exp\left((v_1 - v_2)\frac{\log t}{t}\right)\left(\frac{t\Delta_1}{R_0} - 1\right) \qquad (11)
\end{aligned}
$$

因此，对于任何通过 BIC准则优于 CON 的模型 $i$，我们有

$$
\frac{t\Delta_i}{R_0} > 1 - \exp\left((1 - v_i)\frac{\log t}{t}\right) \qquad (12)
$$

由于 $t \geqslant 2$，$\log(t)/t$ 始终为正，因此前面公式中的下界也为正。

现在我们继续证明第二个结论。若模型 1 满足 (12)，则根据 (11)，我们有

$$
\begin{aligned}
\frac{\Delta_2}{\Delta_1} &> \left(1 - \exp\left((v_1 - v_2)\frac{\log t}{t}\right)\right)\frac{R_0}{t\Delta_1} + \exp\left((v_1 - v_2)\frac{\log t}{t}\right) \\
&> \left(1 - \exp\left((v_1 - v_2)\frac{\log t}{t}\right)\right)\Big/\left(1 - \exp\left((1 - v_1)\frac{\log t}{t}\right)\right) + \exp\left((v_1 - v_2)\frac{\log t}{t}\right) \\
&= \frac{1 - \exp((1 - v_2)\log t/t)}{1 - \exp((1 - v_1)\log t/t)} \\
&:= C^*\left(\frac{\log t}{t}\right) > 0
\end{aligned}
$$

另一方面，若模型 1 不满足 (12)，则仍可通过对模型 2 使用 (11)，对模型 1 使用 CON 及其逆不等式 CON，以及函数 $f(x) = (1 + ax)/(1 + bx)$ 对 $x \in [-1, 0]$ 和 $1 > a > b > 0$ 单调递增这一事实，得到 $\Delta_2 > C^*\Delta_1$。

接下来求 $C^*$ 的最小值。我们对 $C^*$ 的分子和分母关于 $\log t/t$ 求导，得到 $(v_2 - 1)\exp((1 - v_2)\log t/t)$ 和 $(v_1 - 1)\exp((1 - v_1)\log t/t)$。因此，对于任意 $\log t/t := s \in (0, 1/2]$，我们有

$$
\begin{aligned}
C^*(s) &= \frac{\int_0^s (v_2 - 1)\exp((1 - v_2)r)dr}{\int_0^s (v_1 - 1)\exp((1 - v_1)r)dr} \\
&\geqslant \frac{(v_2 - 1)\int_0^s \exp((1 - v_1)r)dr}{(v_1 - 1)\int_0^s \exp((1 - v_1)r)dr} = \frac{v_2 - 1}{v_1 - 1}
\end{aligned}
$$

由于 $v_1 \geqslant v_2$。因此，我们得出结论：对于任意 $t \geqslant 2$，$C^* \geqslant (v_2 - 1)/(v_1 - 1)$。

根据前面的引理，我们知道模型的不纯度增益可以分为两部分。第一部分来自回归掉均值，第二部分来自模型假设，与响应变量的均值无关。因此，当模型 1 为 CON 时，我们可以在（11）中令 $\Delta_2 = \Delta_1 + \Delta_2'$。这里，$\Delta_2'$ 是回归掉均值后的增益。现在，如果 CON 更好，根据（11）的逆不等式，我们有

$$\begin{aligned}
&\frac{t\Delta_2'}{R_0} + \frac{t\Delta_1}{R_0} < 1 - \exp\left((1-v_2)\frac{\log t}{t}\right)\left(1 - \frac{t\Delta_1}{R_0}\right) \\
\iff &\frac{t\Delta_2'}{R_0} < \left(1 - \exp\left((1-v_2)\frac{\log t}{t}\right)\right)\left(1 - \frac{t\Delta_1}{R_0}\right) \\
&< 1 - \exp\left((1-v_2)\frac{\log t}{t}\right), \qquad (13)
\end{aligned}$$

这意味着后续节点仍然倾向于选择 CON，即使其增益为 0。 □

例如，对于我们选择的自由度，若选择 LIN，则有 $\Delta_{lin} \geqslant \Delta_{pcon}/4$。类似地，若 BLIN 是优选模型，由于它们的自由度相同，必有 $\Delta_{blin} \geqslant \Delta_{pcon}$。

## B.3 定理 5 的证明

剩下的问题是，随着深度增加，CON 节点中的残差平方和（RSS）并不会改善，因为后续节点会再次选择 CON。因此，我们首先构造一个递推公式，排除与 CON 节点中的 RSS 对应的项。接着我们将证明，这些 CON 节点中的训练误差渐近趋于零，从而证明 CON 停止规则是合理的。

回顾 $C_k^+ = \{T|T = \text{CON}, T \in \mathcal{T}_{k-1}, R_{k-1}(T) > 0\}$ 的定义，即在第 $k$ 步之前拟合 CON 的节点集合；以及对于满足 $R_k(T) \leqslant 0$ 的节点，$C_k^-$ 的定义。进一步定义 $C_k^* := C_k^+ \backslash C_{k-1}^+$ 和 $C_k^\# := C_k^- \backslash C_{k-1}^-$。最后，令 $A_k^+ := \{T|T \in \mathcal{T}_k, T \neq \text{CON}, R_k(T) > 0\}$，以及 $A_k^- := \{T|T \in \mathcal{T}_k, T \neq \text{CON}, R_k(T) \leqslant 0\}$。注意，利用这些互不相交的节点集合，我们现在有

$$\mathcal{T}_{k-1} = C_{k-1}^+ \cup C_{k-1}^- \cup A_{k-1}^+ \cup A_{k-1}^- \cup C_k^* \cup C_k^\# .$$

接下来，我们可以计算这些节点集合中的误差。令 $R_{C_k^+} := \sum_{T \in C_k^+} w(T)R(T)$，并以类似方式定义 $R_{C_k^-}$, $R_{C_k^*}$, $R_{C_k^\#}$, $R_{A_{k-1}^+}$, $R_{A_{k-1}^-}$。由此得到

$$R_{k-1} = R_{C_{k-1}^+} + R_{C_{k-1}^-} + R_{A_{k-1}^+} + R_{A_{k-1}^-} + R_{C_k^\#} + R_{C_k^*} .$$

最后，令 $R_k := R_k - R_{C_k^+}$。

在整个证明过程中，我们不妨假设，在深度 k 拟合的 CON 节点的增益已经包含在 $R_k$ 中，因此 $R_{C_k^+}$ 可以被视为在回归掉深度 $k$ 之前的 CON 节点均值后的剩余误差。

证明。 不妨假设 $R_{C_k^+} \leqslant R_k$ ，从而 $R_k > 0$。否则，该结论可立即得出。我们有

$$\begin{aligned}
\widetilde{R}_k &= R_k - R_{C_k^+} \\
&\leqslant R_{k-1} - R_{C_k^+} - \sum_{T \in A_{k-1}^+} w(T)\Delta(T) \\
&\leqslant \widetilde{R}_{k-1} - R_{C_k^*} - \sum_{T \in A_{k-1}^+} \frac{w(T)}{F} R_{k-1}^2(T) \\
&\overset{(i)}{\leqslant} \widetilde{R}_{k-1} - R_{C_k^*} - \frac{1}{F}\Big( \sum_{T \in A_{k-1}^+} w(T) R_{k-1}(T) \Big)^2 \\
&= \widetilde{R}_{k-1} - R_{C_k^*} - \frac{1}{F}\Big( R_{k-1} - R_{C_{k-1}^+} - R_{C_{k-1}^-} - \sum_{T \in A_{k-1}^-} w(T) R_{k-1}(T) - R_{C_k^\#} - R_{C_k^*} \Big)^2 \\
&=: \widetilde{R}_{k-1} - R_{C_k^*} - \frac{1}{F}\Big( R_{k-1}^* - R_{C_k^*} \Big)^2
\end{aligned}$$

其中，第二个不等式由前面的引理和命题得出，而 (i) 是詹森不等式。由于 $R_{k-1} - R_{C_{k-1}^+}$ 为正，且 $R_{k-1}^*$ 中的另外三项均为~负数，我们可以将右侧的不等式中的 $R_{k-1}^*$ 替换为 $R_{k-1} - R_{C_{k-1}^+} = R_{k-1} > 0$ 。此外，根据定义，我们有：

$$-R_{C_k^*} - (\widetilde{R}_{k-1} - R_{C_k^*})^2/F \leqslant \begin{cases} -\widetilde{R}_{k-1} + F/4 < -\widetilde{R}_{k-1}/2 & \text{if } \widetilde{R}_{k-1} > F/2 \\ -\widetilde{R}_{k-1}^2/F & \text{otherwise.} \end{cases} \tag{14}$$

~事实上，假设第一个回归模型不是 CON，那么根据 (14) 中任一不等式，对于任意初始 $R_0$ ，我们已经有 $R_1 \leqslant F/4$。因此，后续所有估计都遵循第二个不等式，并且对于任意 $k \geqslant 2$，我们有：

$$\widetilde{R}_k \leqslant \widetilde{R}_{k-1} - \frac{1}{F}\widetilde{R}_{k-1}^2.$$

通过类似于 Klusowski（2021年）引理 4.1 的归纳论证，将 $R_{C_K^+}$ 移至右侧后，我们得到$R_K$ 的估计。

接下来需要控制 CON 节点中的误差。我们可以假设 CON 模型已对每个 $C_k^+$ 中的均值进行了回归处理。第一步是使用命题 4 和引理 12，为任意 CON 节点 $T$得到以下上界：

$$\frac{R(T)^2}{FR_0} \leqslant \frac{\Delta_{\text{pcon}}}{R_0} \leqslant \frac{1}{t}\Big(1 - \exp\Big((1 - v_{pcon})\frac{\log t}{t}\Big)\Big) \tag{15}$$

其中，第一个不等式用到了引理 12，第二个不等式用到了 (13)。由于我们假设响应变量几乎必然被 $\pm B$ 有界，因此有

$$R(T) \leqslant \sqrt{B^2 F}\sqrt{1 - \exp\Big((1 - v_{pcon})\frac{\log t}{t}\Big)} \precsim \sqrt{\frac{(v_{pcon} - 1)\log t}{t}}$$

因此，$C_K^+$ 上 $R(T)$ 的加权和渐近地

$$\sum_{T\in C_K^+} w(T)R(T) \precsim \frac{1}{n}\sum_{T\in C_K^+}\sqrt{t\log t}\,.$$

根据柯西-施瓦茨不等式，我们得到

$$\sum_{T\in C_K^+}\sqrt{t\log t} \leqslant \sqrt{2^K}\Big(\sum_{T\in C_K^+} t\log t\Big)^{1/2} \leqslant \sqrt{2^K}\Big(\log n\sum_{T\in C_K^+} t\Big)^{1/2} = \sqrt{2^K n\log n}\,.$$

因此，若令 $K = \log_2(n)/r$，其中 $r>1$，则 $R_{C_K^+}$ 的阶为 $\mathcal{O}(\sqrt{\log(n)/n^{(r-1)/r}})$，当样本数趋于无穷时，该值趋于零。 □

## B.4 定理 2 的证明

现在，我们可以应用定理 5，为我们的方法推导出一个预言机不等式，并最终证明其一致性。我们将使用 Gy¨orfi 等人（2002）的定理 11.4、定理 9.4 和引理 13.1，因为这些结果不要求估计量是经验风险最小化器。

证明。令 $f$ 为 $\mathcal{F}$中的真实底层函数，令 $\mathcal{F}_n$ 为线性模型树类。我们首先将 $L^2$误差写为 $||f - \hat{}f(\mathcal{T}_K)||^2 = E_1 + E_2$，以便应用定理 11.4，其中

$$E_1 := ||f-\hat{f}(\mathcal{T}_K)||^2 - 2(||Y-\hat{f}(\mathcal{T}_K)||_n^2 - ||Y-f||_n^2) - \alpha - \beta$$

and

$$E_2 := 2(||Y-\hat{f}(\mathcal{T}_K)||_n^2 - ||Y-f||_n^2) + \alpha + \beta.$$

根据我们的假设，我们可以定义 $B_0 = \max\{3B, 1\}$，使得函数类和 $Y$ 几乎处处有界于 $[-B_0, B_0]$，这满足定理 11.4 的条件。因此，根据定理 11.4，并在其记号中取 $\epsilon = 1/2$，可得

$$P(\{\exists \hat{f}(\mathcal{T}_K)\in\mathcal{F}_n \text{ s.t. } E_1>0\}) \leqslant 14\sup_{\mathbf{x}^n}\mathcal{N}\Big(\frac{\beta}{40B_0},\mathcal{F}_n, L_1(\nu_{\mathbf{x}^n})\Big)\exp\Big(-\frac{\alpha n}{2568B_0^4}\Big)$$

其中 $\alpha,\ \beta\to 0$ 作为 $n\to\infty$。

定理 9.4 给出了 $\mathcal{G}_n$ 的覆盖数估计，其中 $\mathcal{G}_n$ 表示叶节点中的函数类。我们注意到，如果 $\beta\to 0$ 定义良好，那么当 $n$ 足够大时，定理的条件 $\beta/40B_0 < B_0/2$ 自动满足。

结合定理 9.4 和引理 13.1，我们得到 $\mathcal{F}_n$的覆盖数估计。具体而言，

$$\mathcal{N}\Big(\frac{\beta}{40B_0},\mathcal{F}_n, L_1(\nu_{\boldsymbol{x}^n})\Big) \leqslant \Gamma_n(\Lambda_n)\Big[3\Big(4eB_0\frac{40B_0}{\beta}\log\Big(6eB_0\frac{40B_0}{\beta}\Big)\Big)^{V(\mathcal{G}_n)}\Big]^{2^K}$$

$$\leqslant \Gamma_n(\Lambda_n)\Big[3\Big(\frac{160eB_0^2}{\beta}\log\frac{240eB_0^2}{\beta}\Big)^{V(\mathcal{G}_n)}\Big]^{2^K}$$
$$\leqslant (np)^{2^K}\Big[3\Big(\frac{160eB_0^2}{\beta}\log\frac{240eB_0^2}{\beta}\Big)^{p+1}\Big]^{2^K}.$$

这里， $\Lambda_n$ 是训练集上的所有可能二叉树的集合，该训练集包含 $n$ 个样本； $\Gamma_n(\Lambda_n)$是二叉树在该集合上所能诱导出的不同划分数量的上界。 $\Gamma_n(\Lambda_n)$的上界源自 Klusowski（2021）和 Scornet 等人（2015）。指数 $V(\mathcal{G}_n)$是 $\mathcal{G}_n$的 VC 维数；由于我们采用多变量线性预测， $V(\mathcal{G}_n) = p+1$。

如果进一步令 $\beta \asymp \frac{B_0^2}{n}$and $\alpha \asymp 3B_0^4\log(np\log n)2^{K+\log(p+1)}/n$，我们有

$$\mathcal{N}\Big(\frac{\beta}{40B_0}, \mathcal{F}_n, L_1(\nu_{\boldsymbol{x}^n})\Big) \precsim (np)^{2^K}(n\log n)^{2^{K+\log(p+1)}}$$

and

$$\begin{aligned}P(E_1>0) &\leqslant P(\{\exists \hat{f}(\mathcal{T}_K)\in\mathcal{F}_n \text{ s.t. } E_1>0\})\\ &\precsim (np)^{2^K}(n\log n)^{2^{K+\log(p+1)}}\frac{1}{(np\log n)^{3\times 2^{K+\log(p+1)}}}\\ &\precsim \frac{1}{(np\log n)^{2^{K+\log(p+1)}}}\\ &\leqslant \mathcal{O}\Big(\frac{1}{n}\Big).\end{aligned}$$

此外，由于 $||f||_{TV} \leqslant A$，且 $Y$ 和 $f(\mathcal{T}_k)$ 在 $[-B_0, B_0]$ 中对任意 $k$ 都有界，因此 $E_1 \leqslant C_1$ 对某个常数 $C_1$ 成立。于是，对某个 $C_1>0$，有 $\mathbb{E}[E_1]\leqslant C_1/n$。根据定理 5，我们还可得，对任意 $f\in\mathcal{F}$，

$$\mathbb{E}[E_2] \leqslant \frac{2F}{K+3} + \frac{C_2\sqrt{2^K n\log n}}{n} + \alpha + \beta,$$

因此，汇总以上所有结果可得，

$$\mathbb{E}[||f-\hat{f}(\mathcal{T}_k)||^2] \leqslant \frac{2F}{K+3} + \frac{C_3\sqrt{2^K n\log n}}{n} + \frac{C_4\log(np\log n)2^{K+\log(p+1)}}{n} \tag{16}$$

其中， $C_3$ 和 $C_4$ 仅取决于 $A$ 和 $B$。

根据我们的假设，期望误差趋于零。此外，如果我们选取 $K_n = \log(n)/r$，其中 $r>1$ 为某个常数，且 $p_n = n^s$ 满足 $0<s<1-1/r$，则 (16) 中的第一项为 $\mathcal{O}(1/\log(n))$，第二项和第三项则为 $o(1/\log(n))$。因此，总体收敛速率为 $\mathcal{O}(1/\log(n))$。 □

# C 线性模型数据的收敛速率

## C.1 预备知识与思路

上一节中，我们推导了当底层函数为一般加性函数时 PILOT 的一致性。在这种情况下得到的收敛速率为 $\mathcal{O}(1/\log(n))$，与

CART 相同，但相当缓慢。当然，这仅仅是一个上界；如果真实底层函数具有某种良好的性质，我们很可能能够获得快得多的收敛速度。

在本节中，我们考虑这一情形。具体而言，我们旨在展示 PILOT 对线性函数具有更快的收敛速率。为解决这一问题，我们不能采用一般一致性结果中的相同方法。具体来说，关于 LIN 的增益的引理 12 的对应结论，仍会导致左侧出现平方误差，从而限制收敛速率。因此，我们采用不同的方法，转而使用 Freund 等人（2017）在提升模型背景下提出的一种近期工具。在我们的情形中，我们在每个节点 $T$ 上对 $Y_T$ 关于 $\boldsymbol{X}_T$ 进行线性回归，并估计相关二次规划（QP）问题的梯度，从而估计不纯度增益。

在整个证明过程中，我们使用与 Freund 等人（2017）相同的记号，以表明与 Klusowski（2021）相比超额误差中的一个细微差异。首先，我们定义与损失函数相关的量的记号。令 $L_n^*(T) := \min_{\hat{\beta}} ||Y - \boldsymbol{X}_T \hat{}\beta||_T^2$ 表示节点 $T$ 上的最小二乘损失，令 $L_n^* := \min_{\hat{\beta}} ||Y - \boldsymbol{X} \hat{}\beta||_n^2$ 表示完整数据上的最小二乘损失。我们进一步将深度为 $k$ 的 PILOT 的损失记为 $L_n^k := ||Y - \hat{}f(\mathcal{T}_k)||_n^2$。最后，$L_n^k(T) := ||Y_T - \boldsymbol{X}_T \beta_T||_t^2$ 表示其在节点 $T$ 中的损失，其中 $\beta_T$ 为某个值（我们可以这样写损失，因为在 $T$ 上预测函数是线性的）。当深度 $k$ 不重要时，我们省略上标。

以下我们使用记号 $\tilde{\boldsymbol{X}}_T$ 表示标准化后的预测变量矩阵，该矩阵是将 $\boldsymbol{X}$ 中每个（非截距）列 $j$ 除以 $\sqrt{n}\hat{}\sigma_j^U$ 得到的，其中 $\sigma_j^U\hat{}$ 是列 $j$ 的标准差的常用估计值。然后我们考虑 $L_n(T) := ||Y_T - \tilde{\boldsymbol{X}}_T \tilde{\beta}_T||_t^2$，并将其梯度记为 $\nabla L_n(T)|_{\tilde{\beta}=\tilde{\beta}_T}$。

Freund 等人（2017）的定理 2.1 使用了一个固定设计矩阵，因此可以对其施加全局特征值条件。由于我们是在局部拟合线性模型，因此需要一个更灵活的结果。为此，我们将施加一个分布条件，其中会用到以下矩阵集中不等式：

**定理 13**（Tropp 等人（2015）的定理 1.6.2）。设 $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$ 为相互独立、中心化且维度相同的随机矩阵，并假设 $\mathbb{E}[\boldsymbol{S}_i] = 0$ 和 $||\boldsymbol{S}_i|| \leqslant L$。设 $\boldsymbol{Z} = \sum_{k=1}^n \boldsymbol{S}_i$ 并定义

$$v(\boldsymbol{Z}) = \max\{||\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]||, ||\mathbb{E}[\boldsymbol{Z}^\top\boldsymbol{Z}]||\}.$$

*Then*

$$P[||Z|| \geqslant t] \leqslant (d_1 + d_2)\exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3}\right) \quad \textit{for all } t \geqslant 0.$$

## C.2 $\Delta_{lin}$的概率下界

我们的第一步是基于上一节中的思路，证明 $\Delta_{lin}$ 的不纯度增益的概率界。

**引理14**。设 $T$ 为一个包含 $n$ 个样本的节点。我们将 $L_n(T)$ 定义为 *PILOT* 的训练损失，将 $L_n^*(T)$ 定义为最小二乘回归的训练损失。除条件 *1* 和 *2* 外，我们还假设

响应变量在 $T$ 上的残差均值为零。那么存在一个常数 $C$，使得 LIN 模型满足

$$P\left[\Delta_{lin} \geqslant \frac{\lambda_0(L_n(T) - L_n^*(T))}{4p}\right] \geqslant 1 - \exp(C_{\lambda_0,\sigma_0,p} n).$$

证明。根据梯度的定义，我们有：

$$\begin{aligned}
\frac{n||\nabla L_n(T)|_{\tilde{\beta}=\tilde{\beta}_T}||_\infty}{2} &= ||\tilde{\boldsymbol{X}}^\top r||_\infty = \max_{j\in 1\ldots p}\{|r^\top \tilde{X}^{(j)}|\} \\
&= \max_{j\in 1\ldots p} \frac{|r^\top X^{(j)}|}{\sqrt{n}\hat{\sigma}_j^U} \\
&\leqslant \max_{j\in 1\ldots p} \frac{|r^\top (X^{(j)} - \overline{X}^{(j)})|}{\sqrt{n}\hat{\sigma}_j} \\
&= \sqrt{n\Delta_{lin}}
\end{aligned} \tag{17}$$

其中，$r$ 表示残差 $Y - \boldsymbol{X}\beta_T$，最后一个等式由（10）以及响应变量在 $T$ 上的残差均值为零这一假设得出。

令 $h(\cdot) = L_n(\cdot)$，根据 Freund 等人（2017）的命题 A.1 和（17），我们有（假设损失函数之差为正）：

$$\begin{aligned}
P\left[\Delta_{\text{lin}} \geqslant \frac{\lambda_0(L_n(T) - L_n^*)}{4p}\right] &\geqslant P\left[\frac{n||\nabla L_n|_{\tilde{\beta}=\tilde{\beta}_T}||_\infty^2}{4} \geqslant \frac{\lambda_0(L_n(T) - L_n^*)}{4p}\right] \\
&\geqslant P\left[\frac{n}{p}||\nabla L_n|_{\tilde{\beta}=\tilde{\beta}_T}||_2^2 \geqslant \frac{\lambda_0(L_n(\tilde{\beta}_T) - L_n^*)}{p}\right] \\
&\geqslant P[\lambda_{min}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) > \lambda_0]\,.
\end{aligned}$$

剩下要证明的仅是：

$$P[\lambda_{min}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) > \lambda_0] \geqslant 1 - \exp(-C_{\lambda_0,\sigma_0,p} n).$$

为此，我们计算 $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ 的最小特征值。我们将向量 $S_i \in \mathbb{R}^p$ 定义为其坐标 $S_i^{(j)} = X_i^{(j)}/\sigma_j$，其中 $\sigma_j$ 是第 $j$ 个预测变量的真实标准差。根据特征值和奇异值的性质，我们可以将估计分为三部分

$$\begin{aligned}
\lambda_{min}(\mathbb{E}[SS^\top]) &= \sigma_{min}(\mathbb{E}[SS^\top]) \\
&= \sigma_{min}\Big[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \underbrace{\frac{1}{n}\sum_{i=1}^n S_i S_i^\top - \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}}_{\tilde{A}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}[SS^\top] - S_i S_i^\top\Big)}_{\tilde{B}_n}\Big] \\
&\leqslant \lambda_{min}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) + \sigma_{max}(\tilde{A}) + \sigma_{max}(\tilde{B})
\end{aligned}$$

因此，我们可以利用另外 3 个矩阵的奇异值来给矩阵 $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ 的特征值构造下界。根据我们的假设，左侧已经由 $2\lambda_0$ 界定。对于 $\sigma_{max}$（$A_n$），我们注意到，这两个矩阵内部的差异仅在于标准差的估计。如果能够证明各项之间的差异均由某个常数统一界定，那么我们也可以给其最大奇异值构造上界。为此，我们使用 Maurer 和 Pontil（2009）中的定理10，该定理指出，对于 $[0,1]$ 和 $\delta > 0$ 中的独立同分布随机变量 $X_1$、..... $X_n$，经典无偏方差估计（$\hat{}\sigma^U$）$^2$ 与真实方差 $\sigma^2$ 满足

$$P\left[|\sigma - \hat{\sigma}^U| > \sqrt{\frac{2\log(1/\delta)}{n-1}}\right] < 2\delta\ .$$

由于

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} S_i S_i^\top - \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} &= \frac{1}{n}\sum_{i=1}^{n}[S_i S_i^\top - \tilde{X}_i \tilde{X}_i{}^\top] \\ &= \left[\frac{\sum_{i=1}^{n} X_i^{(k)} X_i^{(l)}}{n}\Big(\frac{1}{\sigma_k \sigma_l} - \frac{1}{\hat{\sigma}_k^U \hat{\sigma}_l^U}\Big)\right]_{kl}\end{aligned}$$

我们只需控制方差项中的差异，因为每个 $X_i^{(k)}$ 都有界于 1。令 $\sqrt{2\log(1/\delta)/(n-1)}$ $= \lambda_0\sigma_0^4/2p$，并设 $E_k$ 为事件 $|\sigma_k - \hat{}\sigma_k^U| \leqslant {}^{\lambda_0\sigma_0^4}_{2p}$，则有 ——

$$P[E_k] \geqslant 1 - \exp(-C_{\lambda_0,\sigma_0,p} n) \quad \text{and} \quad \sigma_k > \sigma_0 \text{ on } E_k$$

由于 $\sigma_0 < 1$， $\lambda_0 < 2\lambda_0 \leqslant tr(\mathbb{E}[(S-\overline{S})(S-\overline{S})^\top]) \leqslant p$，并且根据我们的假设，估计方差应有下界 $2\sigma_0^2$。在 $E := \bigcup_k E_k$ 上，我们有

$$\max_{k,l}(|\tilde{A}_{kl}|) \leqslant \max_{k,l} \frac{|\sigma_k\sigma_l - \hat{\sigma}_k^U\hat{\sigma}_l^U|}{\sigma_k\sigma_l\hat{\sigma}_k^U\hat{\sigma}_l^U} \leqslant \max_{k,l} \frac{\sigma_k(|\sigma_l - \hat{\sigma}_l^U|) + \hat{\sigma}_l^U(|\sigma_k - \hat{\sigma}_k^U|)}{2\sigma_0^4} \leqslant \frac{\lambda_0}{2p}$$

其中，最后一个不等式由 $\sigma_k\ \sigma_l^U \leqslant 1$ $P[\ _{k\,l}\ |\tilde{A}_{kl}| \leqslant \lambda_0/\ 2p\ ] \geqslant 1 -$，得出。因 $p^2$ $-C_{\lambda_0\,\sigma_0\,p}n$ $_{||x||_2=1}\,||\tilde{A}x||_2 \leqslant$ $_{k\,l}\ |\tilde{A}_{kl}|\ p$ 此，根据并集界，max. ( ) ( ) exp( . . )。由于 max max. ( )，我们有

$$\begin{aligned}P[\sigma_{max}(\tilde{A}) \leqslant \lambda_0/2] &= P[\max_{||x||_2=1} ||\tilde{A}x||_2 \leqslant \lambda_0/2] \\ &\geqslant P[\max_{k,l}(|\tilde{A}_{kl}|) \leqslant \lambda_0/2p] \geqslant 1 - p^2\exp(-C_{\lambda_0,\sigma_0,p}n)\end{aligned}$$

˜对于 $\lambda_{max}(B_n)$，我们可以应用定理 13，得到它以高概率有下界 $\lambda_0/2$。事实上，在 $E$ 上，$S_i$ 的平方 $L^2$ 范数有界于 $p/\sigma_0^2$，因此$||S_iS_i^\top|| = \lambda_{max}(S_iS_i^\top) \leqslant p/\sigma_0^2$，从而根据詹森不等式有 $||\mathbb{E}[SS^\top]|| \leqslant \mathbb{E}[||S_i||^2] \leqslant p/\sigma_0^2$。此外，对于 $\boldsymbol{S}_i := \mathbb{E}[SS^\top] - S_iS_i^\top$，我们有 $\mathbb{E}[\boldsymbol{S}_i] = 0$，并且

$$||\boldsymbol{S}_i|| \leqslant \frac{1}{n}(||S_iS_i^\top|| + ||\mathbb{E}[SS^\top]||) \leqslant \frac{2p}{n\sigma_0^2}\ .$$

使用 Tropp 等人（2015）第 1.6.3 节中的相同论证，我们得到 $v(\tilde{B}_n) \leqslant p^2/n\sigma_0^4$，因此在 $E$ 上，

$$\begin{aligned} P(||\tilde{B}_n|| > \frac{\lambda_0}{2}) &\leqslant P\left(||\tilde{B}_n|| > \frac{\lambda_0}{2} \text{ on } E\right) + P(E^c) \\ &\leqslant 2p \exp\left(\frac{-\lambda_0^2\sigma_0^4 n/8}{p^2 + p\lambda_0\sigma_0^2/3}\right) + p^2 \exp(-C_{\lambda_0,\sigma_0,p} n). \end{aligned}$$

最后我们指出

$$2\lambda_0 \leqslant \lambda_{min}(Cor(X)) = \lambda_{min}(\mathbb{E}[(S-\overline{S})(S-\overline{S})^\top]) \leqslant \lambda_{min}(\mathbb{E}[SS^\top]).$$

汇总所有估计后，由并集界可知，存在一个常数$C_{\lambda_0,\ \sigma_0,\ p} > 0$ 使得 $P[\lambda_{min}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) > \lambda_0] \geqslant 1 - \exp(-C_{\lambda_0,\ \sigma_0,\ p} n)$。 □

## C.3 定理 6 的证明

现在我们可以证明超额误差期望的递推公式。

证明。 我们先处理不包含 CON 的情况。我们知道，在 $T$ 上选取任何非 CON 模型，其不纯度增益始终大于 LIN 的不纯度增益（因为 LIN 的自由度最小）。此外，第一次截断过程不会降低模型的增益，因为它只会剔除错误预测。因此，当给定正数 $L_n^k - L_n^*$ 时，根据引理14，以至少 $\sum_{T\in\mathcal{T}_k} \exp(-C_{\lambda_0,\ \sigma_0,\ p} t)$ 的概率，有

$$\begin{aligned} L_n^{k+1} - L_n^* &= L_n^{k+1} - L_n^k + L_n^k - L_n^* = L_n^k - L_n^* - (L_n^{k+1} - L_n^k) \\ &\leqslant L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} w(T)\Delta_{lin}(T) \\ &= L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} w(T)(L_{lin,n}^{k+1}(T) - L_n^k(T)) \\ &\overset{(i)}{=} L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} w(T)\Big((\overline{Y} - \overline{X_T\beta_T})^2 + \Big\langle Y - X_T\beta_T, \frac{X^{(j)} - \overline{X^{(j)}}}{\sigma_j}\Big\rangle_t^2\Big) \\ &\overset{(ii)}{=} L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} w(T)(\bar{r}_T^k)^2 - \sum_{T\in\mathcal{T}_K} w(T)\tilde{\Delta}_{lin}(T) \\ &\overset{(iii)}{\leqslant} L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} w(T)(\bar{r}_T^k)^2 \\ &\quad - \sum_{T\in\mathcal{T}_K} \frac{\lambda_0}{4p} w(T)(||r_T^k - \bar{r}_T^k||_t^2 - L_n^*(T)) \\ &\overset{(iv)}{\leqslant} L_n^k - L_n^* - \sum_{T\in\mathcal{T}_K} \frac{\lambda_0}{4p} w(T)(L_n^k(T) - L_n^*(T)) \end{aligned}$$

$$\overset{(v)}{\leqslant} L_n^k - L_n^* - \frac{\lambda_0}{4p}(L_n^k - L_n^*)$$
$$= (L_n^k - L_n^*)\Big(1 - \frac{\lambda_0}{4p}\Big)$$

其中 $r_T^k := Y - X_T\beta_T$ 是 $T$ 上的残差。这里，(i) 根据方程（10）得出。(ii) 中的 $\tilde{}\Delta_{lin}(T)$ 表示回归掉均值后 LIN 的增益。(iii) 中的最后一个等式根据引理14得出。(iv) 是由 $||r_T^k - \bar{r}_T^k||_t^2 + ||\bar{r}_T^k||_t^2 = ||r_T^k||_t^2$ 以及系数 $\frac{\lambda_0}{4p}$ 小于 1 这一事实得到的。(v) 根据以下事实得出：每个叶节点 $\sum_{T\in\mathcal{T}_k} w(T)L_n^*(T)$ 中最优损失的加权和小于每个 $k$ 对应的 $L_n^*$。如前一个引理中所述，因子 $1 - \frac{\lambda_0}{4p}$ 始终为正。

接下来需要控制“坏”概率，即 $\lambda_{min}(T) < \lambda_0$ 的情况。我们有

$$P(\exists T \in \mathcal{T}_K \text{ s.t. } \lambda_{min}(T) < \lambda_0) \leqslant \sum_{T\in\cup_{k=1}^K \mathcal{T}_k} \exp(-C_{\lambda_0,\sigma_0,p}t) \leqslant K \sum_{T\in\mathcal{T}_K} \exp(-C_{\lambda_0,\sigma_0,p}t).$$

根据条件1，可得

$$K \sum_{T\in\mathcal{T}_K} \exp(-C_{\lambda_0,\sigma_0,p}t) \leqslant Kn^{1-\alpha}\exp(-C_{\lambda_0,\sigma_0,p}n^\alpha)$$
$$= \exp(-C_{\lambda_0,\sigma_0,p}n^\alpha + \log K + (1-\alpha)\log n)$$
$$\precsim \exp(-C'_{\lambda_0,\sigma_0,p}n^\alpha)$$

只要 $\log K \precsim n^\alpha$，这一条件可通过令 $K$ 的阶为 $\log n$ 来满足。总而言之，若令 $G := \{\forall T \in \mathcal{T}_K, \lambda_{pmin}(T) > \lambda_0\}$ 并假设 $\mathbb{E}[\epsilon^4] < \infty$，则对充分大的 $n$，有

$$\mathbb{E}[L_n^k - L_n^*] = \mathbb{E}_X[\mathbb{1}_G\mathbb{E}_{\epsilon|X}[L_n^k - L_n^*] + \mathbb{1}_{G^c}\mathbb{E}_{\epsilon|X}[L_n^k - L_n^*]]$$
$$\leqslant \sqrt{P(X\in G)}\sqrt{\mathbb{E}_X[(\mathbb{E}_{\epsilon|X\in G}[L_n^k - L_n^*])^2]}$$
$$+ \sqrt{P(X\in G^c)}\sqrt{\mathbb{E}_X[(\mathbb{E}_{\epsilon|X\in G^c}[L_n^0 - L_n^*])^2]}$$
$$\leqslant (1 - \exp(-C'_{\lambda_0,\sigma_0,p}n^\alpha))^{1/2}\Big(1 - \frac{\lambda_0}{4p}\Big)^k\sqrt{\mathbb{E}[(L_n^0 - L_n^*)^2]}$$
$$+ \exp(-C''_{\lambda_0,\sigma_0,p}n^\alpha)\sqrt{\mathbb{E}[(L_n^0 - L_n^*)^2]}$$
$$\leqslant \Big(1 - \frac{\lambda_0}{4p}\Big)^k\sqrt{\mathbb{E}[(L_n^0 - L_n^*)^2]} + \mathcal{O}(1/n).$$

由于概率呈指数衰减，若令 $K \asymp \log n$，误差可被界定为 $\mathcal{O}(1/n)$。

当包含 CON 时，我们可以利用引理14和命题 4 推导出 (15) 的类似结果，从而对所有满足 CON的 $T$ 有：

$$P_{X,\epsilon}\left[\lambda_0\frac{L_n(T) - L_n^*(T)}{4pR_0} < \frac{1}{t}\Big(1 - \exp\Big((1-v_{lin})\frac{\log t}{t}\Big)\Big)\right] \geqslant 1 - \exp(C_{\lambda_0,\sigma_0,p}n).$$

此外，由于误差项具有有限的四阶矩，且真实的底层 $f$ 是线性的，因此以至少 $1-\exp(C_{\lambda_0,\ \sigma_0,\ p}n)$ 的概率，有：

$$\mathbb{E}_{\epsilon|X}[L_n(T)-L_n^*(T)] \precsim \mathbb{E}_{\epsilon|X}[Y^2]\Big(1-\exp\Big((1-v_{lin})\frac{\log t}{t}\Big)\Big) \precsim \frac{(v_{lin}-1)\log t}{t}.$$

我们用 $\mathcal{T}_K^{\text{CON}}$ 表示树中深度不超过 $K$ 的所有 CON 。那么，这些节点上误差的加权和的期望满足

$$\begin{aligned}\mathbb{E}_{X,\epsilon}\Big[\sum_{T\in\mathcal{T}_K^{\text{CON}}} w(T)(L_n(T)-L_n^*(T))\Big] &\leqslant \mathbb{E}_X\Big[\mathbb{1}_{G^c}\sum_{T\in\mathcal{T}_K^{\text{CON}}}\mathbb{E}_{\epsilon|X}[w(T)(L_n(T)-L_n^*(T))]\Big]\\ &\quad+\mathbb{E}_X\Big[\mathbb{1}_G\sum_{T\in\mathcal{T}_K^{\text{CON}}}\frac{(v_{lin}-1)\log(t)}{n}\Big]\\ &\precsim \mathcal{O}\Big(\frac{1}{n}\Big)+\mathcal{O}\Big(\frac{N_{leaves}\log(n)}{n}\Big).\end{aligned}$$

注意到，在我们此前对非 CON 节点进行的估计中，当 $L_n^k(T)-L_n^*(T)>0$ 时，由于 $\lambda_0\leqslant p$ （见引理14的证明），对 $L_n^{k+1}(T)-L_n^*(T)$ 的估计始终为正，因此我们可以先忽略 CON 节点以得到递推公式，最后再加回所有 CON 节点上的误差，从而得到最终结果。□

## C.4 推论 7 的证明

最后，通过应用前述结果以及一致性定理的证明，我们得到 PILOT 在线性模型数据上的收敛速率。

证明。我们知道，对于真实参数 $\beta$，有 $\mathbb{E}[L_n^*-||\boldsymbol{X}\beta||_n^2]=\mathcal{O}(1/n)$。因此，如果我们选择 $K_n=log_\gamma(n)$，根据前述定理，我们有 $\mathbb{E}[L_n^{K_n}-||\boldsymbol{X}\beta||_n^2]=\mathcal{O}(N_{leaves}\log(n)/n)$。最后，我们可以沿用定理 2 的证明，将其中的 $2^K$ 替换为 $N_{leaves}$，并利用定理 6 推导出与（16）类似的预言机不等式：

$$\begin{aligned}\mathbb{E}[||\boldsymbol{X}\beta-\hat{f}(\mathcal{T}_K)||^2] &\precsim \mathcal{O}\Big(\frac{N_{leaves}\log(n)}{n}\Big)+2\mathbb{E}[|L_n^{K_n}-||\boldsymbol{X}\beta||_n^2|]\\ &\precsim \mathcal{O}\Big(\frac{\log(n)}{n^\alpha}\Big).\end{aligned}$$

□

顺便一提，在某些预测变量彼此相关的情况下，实践中 PLIN 和 PCON 的收敛速度甚至可能比纯粹的 $L^2$ 提升更快。