使用L_2-多项式回归进行 k-联合函数的无关 PAC 学习

Mohsen Heidari * 和 Wojciech Szpankowski †
* 印第安纳大学布卢明顿分校计算机科学系
† 普渡大学计算机科学系
*mheidar@iu.edu, †szpan@purdue.edu

arXiv:2303.04859v1 [cs.LG] 8 Mar 2023

摘要

许多传统学习算法为了提高计算效率和理论上的可处理性，会采用自然的 0-1 损失之外的损失函数。其中包括基于绝对损失的回归（$\mathcal{L}_1$ 回归）和基于平方损失的回归（$\mathcal{L}_2$ 回归）。已有证明表明，前者是若干重要概念类（如联合函数和半空间）的无关 PAC 学习器。另一方面，后者因计算效率高而更具优势，其计算复杂度关于样本量呈线性。然而，PAC 可学习性仍未可知，因为目前的保证仅在分布限制下得到证明。 $\mathcal{L}_2$ 回归是否是针对 0-1 损失的无关 PAC 学习器，这一问题自 1993 年以来一直悬而未决，至今仍未得到解答。

本文针对布尔立方体上的联合函数类解决了这一问题——证明了使用k-联合函数的无关 PAC 学习 $\mathcal{L}_2$ 多项式回归。此外，我们提出了一种基于布尔傅里叶展开的新 PAC 学习算法，其计算复杂度更低。基于傅里叶的算法（例如Linial 等人(1993）曾在均匀分布等分布限制下使用。我们证明，经过适当修改后，无需任何分布假设即可将这些算法应用于无关设置。我们通过将针对 0-1 损失的 PAC 学习与最小均方估计（MMSE）问题联系起来，证明了我们的结果。我们推导出了一个优美的上界，用 MMSE 误差表示 0-1 损失。基于此，我们证明，对于任何包含 MMSE 的概念类，MMSE 的符号都是一个 PAC 学习器。

一、引言

为提高计算效率或便于解析，许多传统学习方法（如支持向量机（SVM））依赖于自然的 0-1 损失以外的中间损失函数。绝对差（$\mathcal{L}_1$ 距离）就是一例。已知在 $\mathcal{L}_1$ 距离下进行多项式回归，可得到针对多种假设类的无分布假设的大概正确（PAC）学习器（Kalai 等人，2008），例如 k-联合函数、多项式近似的预测器和半空间。然而，计算 $\mathcal{L}_1$ 距离的运行时间与样本量呈二次关系，因此对于大型数据集而言难以承受。

另一方面，平方损失（$\mathcal{L}_2$ 距离）是一种计算复杂度随数据规模线性增长的替代方案。这促使人们采用低阶多项式算法（Linial 等人，1993）和最小二乘支持向量机（Suykens 和 Vandewalle，1999）等学习算法。从学习理论的角度来看，针对上述概念类，基于$\mathcal{L}_2$的方法进行 PAC 学习已有研究，但这些研究依赖于分布假设（Linial 等人，1993；Kalai 等人，2008；Jackson，2006）。

例如，在可实现性假设下，即可以实现零泛化损失的情形（$opt = 0$），$\mathcal{L}_2$-多项式回归是一种 PAC 学习器。除了可实现性假设外，在均匀输入分布下，低次算法也是一种 PAC 学习器（Mossel 等人，2004，2003年；Blais 等人，2010年）。在分布无关（无知者）设定下，对于各种概念类，目前已证明了形式为 c（opt）的 PAC 界，其中opt 是该类的最小损失，且 c 是一个最高可达 $c = 8$的常数（Kalai 等人，2008；

本研究部分得到 CCF-2006440、CCF-2007238、CCF-2 美国国家科学基金会信息科学中心（CSoI）资助，项目编号 CCF-0939370；此外还得到美国国家科学基金会资助，项目编号 211423，以及 Google 研究奖的支持。

Kearns 等人，1994；Jackson，2006）。因此，无知者PAC 可学习性 $\mathcal{L}_2$-方法仍是一个开放问题，尚未确定。

本文解决了在布尔立方体上学习 k-联合函数的问题，即输入为 d 个变量、输出至多取决于 $k < d$ 个变量的布尔函数，其中 k 通常是远小于 d 的常数。出于多种动机，联合函数学习已在文献中得到广泛研究，例如机器学习中的特征选择（Guyon 和 Elisseeff，2003年；Blais 等人，2010年；Heidari等，2021b；Kalai 等人，2008；Klivans 等人，2009年；Birnbaum 和 Shwartz，2012；Diakonikolas 等人，2019年）。我们证明，对于 k-联合函数，可以使用 $\mathcal{L}_2$ 次多项式回归实现无关PAC学习。此外，我们提出了一种更高效的变体，即使用布尔傅里叶展开进行 $\mathcal{L}_2$ 回归。我们表明，该算法也是一个针对 k-联合函数的无关 PAC 学习器。这一结果意味着，傅里叶算法（例如 Linial 等人（1993）提出的低次算法）即使最初是为均匀分布设计的，也适用于无关设置。

A. 贡献总结

使用最小二乘回归学习 k-联合函数：本文聚焦于在布尔输入上，使用 $\mathcal{L}_2$-回归和通常的 0-1 损失，对 k-联合函数类进行 PAC 学习。按照标准的 PAC 学习模型，训练集包含 n 个样本 $\{(\mathbf{x}(i)\ y(i))\}_{i=1}^{n}$，其中特征向量为 $\mathbf{x}(i) \in \{-1,1\}^d$，二元标签为 $y(i) \in \{-1,1\}$。$\mathcal{L}_2$-多项式回归的目标是最小化目标标签y 与次数不超过 k 的多项式 $p(\mathbf{x})$ 之间的经验平方损失。给定这样的多项式，只需将其符号作为 $g(\mathbf{x}) = \text{sign}[p(x)]$，即可得到一个预测器 g。

本文的首个主要结果表明，$\mathcal{L}_2$ 多项式回归能够对 k-联合函数进行无假设 PAC 学习。更准确地说，以至少 $(1-\delta)$ 的概率，预测器 g 的泛化损失与所有 k-联合函数中的最优损失相差不超过一个较小的量，即 $\mathbb{P}\{Y \neq g(\mathbf{X})\} \leq \text{opt} + \epsilon$，其中 opt 是 k-联合函数类中的最优损失。更正式地说，我们证明以下定理。

定理 1（简述）。给定 $k \leq d$，存在一种基于 $\mathcal{L}_2$-多项式回归、次数上限为 k 的算法（算法*1*），可无假设 PAC 学习 k-联合函数，样本复杂度为 n，最多相差 $O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta\epsilon^2})$，计算复杂度为 $O(nd^{\Theta(k)})$。

请注意，使用 $\mathcal{L}_1$-多项式回归学习 k-联合函数的计算复杂度为 $O(n^2 d^{(3+\omega)3k})$，当 n 较大时，复杂度会更高。

证明采用 $\mathcal{L}_1$ 或 $\mathcal{L}_2$ 回归得到 PAC 界时，主要的技术难点之一是分析 0-1 损失与平方损失或绝对损失之间的联系。关于 $\mathcal{L}_2$ 的传统结果依赖于不等式$\mathbb{1}\{y \neq \text{sign}[p(x)]\} \leq (y - p(x))^2$，该不等式对 $y \in \{-1,1\}$ 成立。基于这一界，可以推出 PAC 界 8opt （ Linial 等，1993 ）。因此，这引出了一个问题：在基于 $\mathcal{L}_2$ 的 PAC 学习中，取符号是否最优？当 $x \in \{-1,1\}^d$ 时，Blum 等人（Blum等，1994）和 Jackson（Jackson，2006年）提出了一种巧妙的随机舍入方法，以取代取符号。因此，他们将系数从 8opt 降至 2opt。在 II-A 节中，我们指出，至少对于二元输入而言，这些界并不紧。我们证明了连接 0-1 损失与平方损失的新界（二元输入对应引理 3，实值输入对应引理 6）。利用这些结果，我们证明，对于 k-联合函数类，取符号并无问题，并且可得到 opt，从而实现基于 $\mathcal{L}_2$ 的无关 PAC 可学习性。此外，我们还将更一般的、具有 $x \in \mathbb{R}^d$ 的类别的系数从 8opt 降至 2opt。

我们的方法依托一个框架：利用配备概率测度的向量空间作为代理来推导 PAC 学习界。除此之外，我们还考虑特征-标签集 $\mathcal{X} \times \mathcal{Y}$ 上函数的联合向量空间，并将样本-标签关系和底层联合分布 D 纳入其中。通过将 PAC学习模型与分析向量空间的强大工具联系起来，这一方法得以确立我们的结果。值得注意的是，我们证明了一个优雅的

基于 1-范数和 2-范数等适用量（参见第 1 和 3 节中的推论 III），得出的 0-1 损失上界。我们的方法有一个显著特点：所得表达式十分简洁且富有启发性。

表 I k-联合函数 PAC 学习算法比较。

算法	样本复杂度	计算复杂度	PAC 误差
暴力经验风险最小化	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$	$O(nd^k 2^{2^k})$	opt $+\epsilon$
$\mathcal{L}_1$-多项式回归 (Kalai 等人，2008)	$O(\frac{1}{\epsilon}k^{\Theta(k)}\log\frac{d}{\delta})$	$O(n^2 d^{(3+\omega)3k})$	opt $+\epsilon$
$\mathcal{L}_2$-多项式回归	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\epsilon^2\delta})$	$O(nd^{\Theta(k)})$	• $2opt+\epsilon$ (Jackson, 2006) • $opt+\epsilon$, [定理 I]
低次算法（均匀分布） (Linial 等人，1993)	$O(k2^k$ 对数$\frac{d}{\delta})$	$O(\frac{nkd^k}{(k-1)!}($	• $8opt+\epsilon$ (Linial 等人，1993) • $2opt+\epsilon$ (Jackson, 2006) • $\frac{1}{4}+opt(1-opt)+\epsilon$ (Kearns 等人，1994) • $opt+\epsilon$, [定理 2]
随机傅里叶 (算法 2)	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$	$O(\frac{nkd^k}{(k-1)!})$	$opt+\epsilon$, [定理 2]

使用傅里叶算法进行学习： 此外，我们还提出了另一种针对二值样本的、更高效的算法。该算法的运行时间关于 n 呈线性，并随 d^k 增长；渐近而言，这优于另外两种随 $d^{O(k)}$增长的方法。我们的结果基于针对均匀分布定义的布尔傅里叶展开（Wolf，2008；O'Donnell，2014）。我们通过证明一个反直觉的结果，表明均匀布尔傅里叶实际上适用于不可知分布无关设置。受 Linial 在均匀分布上的低次算法（Linial 等人，1993）启发，我们开发了一种傅里叶算法，能够更高效地进行 $\mathcal{L}_2$多项式回归，且不需要任何分布假设。随后我们证明，该算法也能对 k-junta 类进行无假设 PAC 学习。更正式地说，我们证明如下结论。

定理 2（简述）。给定 $k < d$，傅里叶算法（算法2）以 PAC 学习的方式学习 k-联合函数，样本复杂度为 $O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$，计算复杂度为 $O(\frac{nkd^k}{(k-1)!})$。

表 I 比较了各种 PAC 学习算法的样本复杂度、运行时间和 PAC 损失。与其他方法相比， $\mathcal{L}_2$ 和傅里叶算法具有更低的样本复杂度和计算复杂度。与使用 $\mathcal{L}_1$-多项式回归的 (Kalai 等人，2008) 相比，我们获得了更低的样本复杂度和计算复杂度。请注意， $\mathcal{L}_1$ 回归的运行时间随 $O(n^2 d^{O(k)})$ 增长，这对样本量 n 呈二次增长，因此在大型数据集中难以采用。 $\mathcal{L}_2$ 回归的运行时间为 $O(nd^{O(k)})$，相对于 n呈线性增长。最后，傅里叶算法的运行时间随 $O(nd^k)$ 增长，其指数优于 $\mathcal{L}_2$。总的来说，鉴于 k 通常是独立于 d的常数， $\mathcal{L}_2$ 回归和傅里叶算法适用于大型数据集。最后，我们给出了 k-联合函数类样本复杂度的下界。根据标准的 VC 维数论证，可得 $O(\frac{1}{\epsilon^2}(VC+\log(\frac{1}{\delta}))$。 k-联合函数类的 VC 维数的确切表达式尚不清楚，但介于 2^k 和 $2^k + k\log d$之间。

B. 相关工作

学习联合函数是机器学习中的一个经典问题。关于联合函数学习与测试，已有大量研究（Mossel 等，2004年；Bshouty 和 Costa，2016；Liu 等人，2019年；Arpe 和 Mossel，2008；Fischer 等人，2004；Servedio 等人，2015；De 等人，2019年；Vempala 和 Xiao，2011；Chen 等人，2021；Iyer 等人，2021）。联合函数在学习理论中备受关注，因为它们与其他基础问题相关，例如特征选择学习（Guyon 和 Elisseeff，2003年）、析取范式公式和决策树（Mossel 等，2004年）。特别是，特征选择学习可以表述为学习 k-联合函数（从 k

个 d 特征中）。此外，每个 k-联合函数都可以由规模为 2^k 的决策树或析取范式公式实现；反过来，任何规模为k的决策树也是一个 k-联合函数，而任何 k项析取范式都可以用 ϵ-近似为一个 $k\log(\frac{k}{\epsilon})$-联合函数。因此，为这些问题设计高效算法与学习联合函数密切相关（Mossel 等人，2004年）。关于 k-联合函数的 PAC 学习已采用多种方法进行研究。下面我们简要回顾学习这些概念类的方法，并在表I中对其进行总结。

朴素经验风险最小化（ERM）：这是通常的穷举搜索，通过遍历所有预测器来最小化经验损失。对于 k-联合函数，ERM 是一种无关 PAC 学习算法，其样本复杂度为 $O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$，计算复杂度为 $O(nd^k2^{2^k})$ (Shalev-Shwartz 和 Ben-David，2014）。由于其计算复杂度相对于 k呈双指数级增长，即使 k取值很小，ERM 也难以实行。

基于 $\mathcal{L}_1$ 回归的学习。Kalai 等人（Kalai 等人，2008）提出将多项式回归作为使用 $0-1$ 损失函数进行 PAC 学习的一种方法。他们证明，$\mathcal{L}_1$-多项式回归能够针对包含 k-联合函数的 $(k,\ \epsilon)$-集中假设类进行无假设 PAC 学习。将该算法应用于 k-联合函数需要 $O(\frac{1}{\epsilon}d^{\Theta(k)})$ 的样本复杂度。采用线性规划实现时，该算法的计算复杂度为 $O(n^2d^{(3+\omega)3k})$，其中 $\omega<2.4$ 是矩阵乘法指数。这种方法的计算复杂度呈二次增长，因此在样本量较大时成本高昂。这促使我们研究基于 $\mathcal{L}_2$ 的方法。

通过 $\mathcal{L}_2$ 多项式回归进行学习。这种方法与其 $\mathcal{L}_1$ 对应方法类似，只是将绝对误差替换为平方损失。已有研究探讨了样本量呈线性复杂度的快速 $\mathcal{L}_2$ 回归实现（Drineas等，2006年、2010年）。采用这种方法进行 PAC 学习已在（Kalai等，2008；Jackson，2006）中得到研究。在不可知设定下，研究表明这种方法是一个误差为 8opt 的弱学习器。使用（Blum等，1994年；Jackson，2006年）提出的非确定性舍入，可以将 PAC界 降至 2opt。本文表明，对于 k-联合函数，无需任何随机舍入即可达到 opt。对于其他非二元类别，我们在第四节证明了界为 2opt。

傅里叶算法。这种方法被视为 $\mathcal{L}_2$ 回归的一种特殊解法。Linial 等人（Linial 等，1993）从另一种视角研究了 PAC 学习，并提出了著名的“低次算法”。他们针对 $\{-1,1\}^d$ 样本的均匀且已知分布给出了理论保证。低次算法基于布尔立方体上的傅里叶展开。尽管该算法在计算上高效，但由于其分布限制，实际应用有限——在许多应用中，均匀（且已知）分布并不现实。Furst 等，1991）通过采用低次算法，在乘积概率分布下学习 AC^0 函数，从而放宽了这种分布限制。傅里叶展开已用于分析布尔函数（Wolf，2008；O’Donnell，2014），应用范围广泛，包括计算学习（Linial 等，1993；Mossel等，2004年）、噪声敏感性（O’Donnell，2014；Kalai，2005年；Li and M´edard，2018；Heidari等，2019年）、近似（Blais 等，2010年）、特征选择（Heidari等，2021b）以及其他信息论问题（Courtade和Kumar，2014年；Weinberger和Shayevitz，2017年、2018年；Heidari 等，2021a）。在本研究中，我们还将这种方法推广到无关PAC学习，从而消除了分布假设。

II. 形式化定义与主要结果

模型：我们采用无关 PAC 学习模型的标准表述（Valiant，1984；Kearns 等人，1994）。本文聚焦于采用 0-1 损失的二分类。如果对于任意 $\epsilon,\delta\in(0,1)$，给定从任意分布 D 中抽取的 $n>n(\epsilon,\delta)$ 个训练样本后，算法以概率 $(1-\delta)$ 输出预测器 g，且其期望损失至多为 opt $+\epsilon$（其中 opt 是 $\mathcal{H}$ 中的最小损失），则称该算法无假设 PAC 学习假设类$\mathcal{H}$。

记号：对于任意自然数 d，集合 $\{1,2,\cdots,d\}$ 记作 $[d]$。对于定义在 $\mathcal{X}$ 上的一对函数 f,g，记号 $f\equiv g$表示对于所有 $x\in\mathcal{X}$，都有 $f(x)=g(x)$。对于任意函数 $h\colon\mathcal{Z}\to\mathbb{R}$和输入分布 D,

算法 1： $\mathcal{L}_2$-算法

输入：训练样本 $\mathcal{S}_n = \{(\mathbf{x}(i)\ y(i))\}_{i=1}^n$，次数参数 k。1 对每个满足 $|\mathcal{J}| = k$ 的子集 $\mathcal{J} \subseteq [d]$，执行 2 求出一个次数不超过 k 的多项式 $\hat{}p_{\mathcal{J}}$，使 $\frac{1}{n}\sum_i \big(y(i) - p(\mathbf{x}(i))\big)^2$ 最小。3 选择平方损失最小的 $\hat{}p_{\mathcal{J}}$ 作为 $\hat{}p$ 。4 返回 $\hat{}g \equiv \text{sign}[\hat{}p]$。

1-范数和2-范数分别定义为 $\|h\|_{1,D} := \mathbb{E}_D[|h(Z)|]$ 和 $\|h\|_{2,D} := \sqrt{\mathbb{E}_D[h(Z)^2]}$, 。

A. 预备部分

我们首先指出，使用 $\mathcal{L}_1$ 或 $\mathcal{L}_2$ 回归证明 PAC 界时的一个主要难点。主要挑战在于分析对所得多项式取符号后的 0-1 损失。为此，需要研究 0-1 损失与平方损失或绝对损失之间的关系。具体来说，令 p 为使平方损失最小的多项式。那么，不难看出 $\mathbb{1}\{y \neq \text{符号}[p(x)]\} \leq (y - p(x))^2$, 其中 $y \in \{-1, 1\}$。因此，符号 $[p]$ 的 0-1 损失满足 $\mathbb{P}\{Y \neq \text{符号}[p(\mathbf{X})]\} \leq \mathbb{E}[(Y - p(\mathbf{X}))^2]$. 这是一个较松的界，由此得到的 PAC 界为 8opt。下面说明这一论证。令 f 为 0-1 损失为 opt 的最优预测器。此外，假设 f 可由多项式 $\tilde{p}$ 近似，且平方误差小于 ϵ^2。于是，我们可以写出

$$\begin{aligned}
\mathbb{P}\{Y \neq \text{sign}[p(\mathbf{X})]\} &\leq \mathbb{E}[(Y - p(\mathbf{X}))^2] \\
&\overset{(a)}{\leq} \mathbb{E}[(Y - \tilde{p}(\mathbf{X}))^2] \\
&\overset{(b)}{\leq} 2\mathbb{E}[(Y - f(\mathbf{X}))^2 + (f(\mathbf{X}) - \tilde{p}(\mathbf{X}))^2] \\
&\overset{(c)}{\leq} 8opt + 2\epsilon^2,
\end{aligned}$$

其中，(a) 是因为 p 是最优多项式，(b) 根据算术—几何平均不等式成立，而 (c) 是因为对于任意 f: $\mathcal{X} \to \{-1, 1\}$，$\mathbb{1}\{y \neq f(x)\} = \frac{1}{4}(y - f)^2$。

这些观测值引出了一个问题：取符号在 PAC 学习中是否最优。当 $x \in \{-1, 1\}^d$时，Blum 等人(Blum 等人, 1994) 和 Jackson (Jackson, 2006) 提出了一个巧妙的想法：用随机舍入代替取符号。因此，他们将系数从8opt 降至2opt。

在引理3中，我们证明了 0-1 损失与平方损失之间更紧的界。利用该引理，我们在定理1 中证明，对于 k-junta 类，取符号是最优的，并且可达到opt （即无关 PAC 可学习性）。此外，我们还进行了更一般的分析，并证明给定观测值$\mathbf{X}$时， Y 的最小均方误差的符号可构成一个 PAC 学习器，见定理3。

B. 使用 $\mathcal{L}_2$-多项式回归进行学习

我们采用一种使用 $\mathcal{L}_2$-多项式回归的 PAC 学习算法。给定一个训练集，多项式回归的目标是在所有次数不超过 k 的多项式中最小化经验平方损失。这个过程可以通过随机梯度下降或求解线性方程组来实现。基于这种回归，可以研究各种概念类的 PAC 学习。本文考虑 k-联合函数。

k-联合函数类： k-联合函数是一个布尔函数 $h : \{-1, 1\}^d \to \{-1, 1\}$，它有 d 个输入变量，且其输出至多取决于 d 个输入变量中的 k 个。

对于 k-联合函数类，我们使用一种变体的 $\mathcal{L}_2$-多项式回归（见算法1），其计算复杂度与标准的 $\mathcal{L}_2$ 多项式回归相同。采用这种方法，我们得出以下定理。

定理 1。算法*1*，次数上限为 $k \le d$，可无假设 PAC 学习 k-联合函数。更准确地说，给定$\delta \in [0,1]$，以概率 $(1-\delta)$，其泛化损失不超过以下数值

$$\text{opt} + O\Big(\sqrt{\frac{2^k + k\log d}{n}\log\frac{n}{2^k + k\log d}}\Big) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

其中 n 是样本数。此外，所得计算复杂度为 $O(nd^{\Theta(k)})$。

简化上述表达式后，我们得到样本复杂度界 $n(\delta,\ \epsilon) = O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\epsilon^2\delta})$。证明见第 III-C节。

算法 1 中的多项式回归过程可通过在 $\mathbb{R}^D$ 中进行线性 L_2 回归来实现，其中$D = d^k$。因子 d^k 是因为次数最高为 k 的多项式是形如 $\prod_{j=1}^{k} X_{i_j}$的单项式的线性组合，其中 $i_j \in [d]$。线性回归可通过 Moore-Penrose（广义）逆实现。广义逆使用 O中的经典方法计算（$nD^2 + D^3$）。因此，鉴于 $n > D$，我们可以在 O（nD^2）$= O$（nd^{2k}）内进行多项式回归。不过，我们注意到，在特殊情况下（例如 $d^k = \lceil n^r \rceil$），回归可在 O（$nd^{k\omega(r)}$）内完成，其中 ω（r）是 Gall 和 Urrutia，2018 中给出的常数。因此，如表 I 所述，该算法的计算复杂度为 O（$nd^{\Theta(k)}$）。

C. 基于傅里叶的学习算法

我们提出另一种基于 $\mathcal{L}_2$ 的方法，其计算效率高于 $\mathcal{L}_2$ 多项式回归。$\mathcal{L}_2$ 回归的计算成本随 $O(nd^{\Theta(k)})$ 增长，这比其复杂度为 $O(n^2 d^{(3+\omega)3k})$ 的 $\mathcal{L}_1$ 变体效率更高。这引出了一个问题：因子 $d^{\Theta(k)}$ 能否进一步降低？我们通过对布尔立方体进行傅里叶分析来回答这个问题。具体而言，我们提出一种复杂度为 $O(\frac{nkd^k}{(k-1)!})$ 的算法。

我们的解决方案基于应用于布尔立方体上均匀分布的布尔傅里叶展开（O'Donnell，2014；Wolf，2008）。令人惊讶的是，我们计划将这种傅里叶方法用于不可知设定。下面简要介绍标准的布尔傅里叶展开。

事实 1（布尔傅里　叶）。任意（有界）函数 $f:\ \{-1,1\}^d \to \mathbb{R}$ 均可分解如下

$$f(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} \mathsf{f}_{\mathcal{S}}\ \chi_{\mathcal{S}}(\mathbf{x}), \quad \forall \mathbf{x} \in \{-1,1\}^d,$$

其中 $\chi_{\mathcal{S}}(\mathbf{x})$ 是与子集 $\mathcal{S} \subseteq [d]$ 对应的单项式，定义为 $\chi_{\mathcal{S}}(\mathbf{x}) = \prod_{j\in\mathcal{S}} x_j$。此外，系数 $\mathsf{f}_{\mathcal{S}} \in \mathbb{R}$ 称为 f 的傅里叶系数，计算方式如下：

$$\mathsf{f}_{\mathcal{S}} = \frac{1}{2^d}\sum_{\mathbf{x}} f(\mathbf{x})\chi_{\mathcal{S}}(\mathbf{x}), \quad \forall \mathcal{S} \in [d]$$

该展开依赖于一个限制条件：输入变量在布尔立方体上均匀分布。这使得基于傅里叶的算法（如 (Linial 等人，1993)）难以应用于 无知学习问题，尤其是在没有任何分布假设的情况下。通过 Gram-Schmidt 型正交化过程可以解决这一问题，该过程可得到广义布尔傅里叶展开（Heidari 等人，2021a）。

然而，本文采取了略有不同的思路，提出对标准布尔傅里叶展开作简单调整，使其适用于某些无知学习问题。因此，我们兼得 PAC 可学习性与计算效率。接下来，我们将介绍这一调整。

令 D_X 为 $\{-1,1\}^d$ 上的任意概率分布，f 为一个布尔函数。定义

$$f_{\mathcal{S}} := \frac{1}{2^d}\sum_{\mathbf{x}} f(\mathbf{x})D_X(\mathbf{x})\chi_{\mathcal{S}}(\mathbf{x}).$$

根据事实 1， f_S是实值函数 $f(\mathbf{x})D_X(\mathbf{x})$的傅里叶系数。注意，在均匀 D_X下，$f_S = \frac{1}{2^n}\mathrm{f}_S$，其中 f_S是 $f(\mathbf{x})$的傅里叶系数，如事实 1所述。在 D_X未知的无关设置中，f_S 无法直接获取。不过，我们可以通过经验估计它。

在解释估计方法之前，我们先介绍另一种扩展。在无关设置中，标签 y 不一定是特征 $\mathbf{x}$ 的函数。因此，为了使傅里叶展开适用于无关 PAC，我们将其从确定性函数扩展到随机映射：

考虑一个随机向量 $\mathbf{X}$ 和一个标记变量 Y。令 $(\mathbf{X}\ Y) \sim D$，其中 D 是 $\{-1,1\}^d \times \{-1,1\}$ 上的概率分布。那么，随机傅里叶系数定义为

$$a_S := \frac{1}{2^d}\mathbb{E}[Y\chi_S(\mathbf{X})], \tag{1}$$

对所有 $S \subseteq [d]$ 均成立。若 $Y = f(\mathbf{X})$，则 $a_S = f_S$。给定独立同分布（i.i.d.）样本$\{x(i)\ y(i)\}_{i=1}^n$， a_S 的经验估计为

$$\hat{a}_S := \frac{1}{2^d}\frac{1}{n}\sum_{i=1}^{n} y(i)\chi_S(\mathbf{x}(i)). \tag{2}$$

请注意，估计方　　法不依赖于底层分布 D，但我们将证明它收敛到 a_S　　.

引理 1。令 D 为 $\{-1,1\}^{d+1}$ 上的任意概率分布。令 $S_1, S_2, \cdots, S_m$ 为 $[d]$ 的 m 子集。给定$\delta \in (0,1)$ 以及从 D 中独立同分布抽取的 n 个样本，不等式

$$\sup_{1\le j\le m} |\hat{a}_{S_j} - a_{S_j}| \le \frac{1}{2^d}\sqrt{\frac{1}{2n}\log\frac{2m}{\delta}}$$

成立的概率至少为 p 至少以概率 $(1-\delta)$成立，其中 a_S和 $\hat{}a_S$ 分别由 (1) 和 (2)给出　　y.

证明：注意，对于任意 $S \subseteq [d]$

$$\begin{aligned}\mathbb{E}_{S_n\sim D^n}[\hat{a}_S] &= \frac{1}{2^d}\mathbb{E}[Y(1)\chi_S(\mathbf{X}(1))] \\ &= \frac{1}{2^d}\sum_{\mathbf{x},y} D(\mathbf{x},y)y\ \chi_S(\mathbf{x}) = a_S.\end{aligned}$$

在 a_S 和 $\hat{}a_S$ 的定义中取因子 $\frac{1}{2^d}$，可得

$$|\hat{a}_S - a_S| = \frac{1}{2^d}\Big|\frac{1}{n}\sum_{i=1}^{n} y(i)\chi_S(\mathbf{x}(i)) - \mathbb{E}[Y\chi_S(X)\Big|.$$

我们应用 McDiarmid不等式 来界定右侧项。不难验证

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} y(i)\chi_S(\mathbf{x}(i)) - \mathbb{E}[Y\chi_S(X)\Big| \ge \epsilon\Big\} \le 2e^{-\frac{n\epsilon^2}{2}}.$$

因此，考虑因子 $\frac{1}{2^d}$，利用并集界，并令右侧等于 δ，即可证明该引理。

■

采用这种方法，我们的傅里叶算法（见算法2）通过根据训练样本估计标签的傅里叶系数，在傅里叶域中进行多项式回归。在下面的定理中，我们给出使用这种方法学习 k-联合函数的 PAC 界。

定理 2。傅里叶算法以无关分布的方式学习 k-联合函数，对于 $k \le d/2$，误差小于

$$\mathrm{opt} + O\Big(\sqrt{\frac{2^k}{n}\log\frac{d^k}{(k-1)!\delta}}\Big),$$

算法 2：随机傅里叶

输入：训练样本 $\mathcal{S}_n = \{(\mathbf{x}(i)\ y(i))\}_{i=1}^n$，次数参数 k。输出：预测器 $\hat{}g1$ 对每个至多包含 k 个元素的 $\mathcal{S} \subseteq [d]$，计算经验傅里叶系数 $\hat{}a_{\mathcal{S}} = \frac{1}{2^d}\frac{1}{n}\sum_{i=1}^n y(i) \prod_{j\in\mathcal{S}} x_j$ (i)。2 对每个包含 k 个元素的 $\mathcal{J} \subseteq [d]$，构造函数 $\hat{}f^{\mathcal{J}}(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{}a_{\mathcal{S}}\chi_{\mathcal{S}}(\mathbf{x})$.
3 $\mathcal{J}$ $[f^{\mathcal{J}}]$ 。

找到使 sign $\hat{}$ 的经验损失最小的 $\hat{}$。4 返回 $\hat{}g \equiv \text{sign}[\hat{}f\hat{}^{\mathcal{J}}]$。

以至少$(1-\delta)$的概率，其中 n 是样本数。此外，所得计算复杂度为 $O(\ ^{nkd^k}\ (k-1)!)$。

在接下来的部分，我们将讨论主要思路。该定理的证明见第 III- E.

III. 主要技术结果

本文的主要结果依赖于平方损失与 0-1 损失之间的一个基本联系，该联系以推论 1 和 3 的形式在第 III-B节中给出。本节将阐述这一联系，并介绍证明定理1 和 2的步骤。

A. 向量空间表示

我们引入一种融入特征-标签分布的向量表示。这种表示是一种代理，可用于运用为向量空间开发的强大代数工具。接下来，我们将描述这种表示。

令 $\mathcal{X}$ 表示输入集，$\mathcal{Y} = \{-1, 1\}$ 表示标签集，D 表示 $\mathcal{X} \times \mathcal{Y}$ 上的底层分布。考虑由所有满足 $\mathbb{E}_D[h(\mathbf{X}\ Y)^2]$ 有限[1] 的函数 $h\colon \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ 构成的向量空间。自然地，两个函数 h_1 和 h_2 之间的内积定义为

$$\langle h_1, h_2\rangle_D \triangleq \mathbb{E}_D[h_1(\mathbf{X}, Y)h_2(\mathbf{X}, Y)].$$

采用这种表述后，真实标记就是函数 $(\mathbf{x}\ y) \mapsto y$。请注意，这符合不可知设定，即标签不一定是 $\mathbf{x}$ 的函数。此外，学习模型中的预测器 g 可视为映射 $(\mathbf{x}\ y) \mapsto g(\mathbf{x})$。由于 $\mathcal{Y} = \{-1, 1\}$，任何预测器 g 的 0-1 损失都可以写为

$$\mathcal{L}_D(g) = \frac{1}{2} - \frac{1}{2}\langle Y, g\rangle_D = \frac{1}{4}\|Y - g\|_{2,D}^2, \tag{3}$$

其中，稍微滥用一下记号，Y 和 g 分别表示映射$(\mathbf{x}\ y) \mapsto y$ 和$(\mathbf{x}\ y) \mapsto g(\mathbf{x})$。式（3）中的第一个等号是由恒等式$\mathbb{1}\{a \neq b\} = \frac{1}{2}(1 - ab)$ 对任意 $a, b \in \{-1, 1\}$均成立这一事实得出的。第二个等号则来自2-范数的定义以及$\|Y\|_{2,D} = \|g\|_{2,D} = 1$这一事实。

这种表示的一个优点是，在已知分布 D 下得到的理论结果可以轻松地$\hat{}$推广到不可知设定。只需将 D 替换为经验分布 $\hat{D}$ 即可；该经验分布在训练集上均匀分布，在训练集之外为零。例如，g 的经验损失立即满足与式（3）相同类型的关系：

$$\frac{1}{n}\sum_i \mathbb{1}\{y_i \neq g(\mathbf{x}_i)\} = \frac{1}{2} - \frac{1}{2}\langle Y, g\rangle_{\hat{D}} = \frac{1}{4}\|Y - g\|_{2,\hat{D}}^2.$$

[1]该空间中的零函数是指一个函数，它将 $(\mathbf{x}\ y) \mapsto 0$ 映射为所有 $\mathbf{x}\ y$，但零概率子集除外。

B. PAC 与最小均方误差

接下来，我们将推导期望损失和经验损失的界，并证明主要定理。主要结果（定理 1 和 2）的证明关键在于建立最小均方误差与 PAC 学习损失之间的联系。

考虑一个一般问题，其中 Z 为观测值，目标是预测 Y。这里 Z 取值于一个一般集合 $\mathcal{Z}$，而 Y 取值于 $\{-1,1\}$。令 Y_{MMSE} 表示给定 Z 时 Y 的最小均方误差。已知 $Y_{MMSE}=\mathbb{E}[Y|Z]$。在以下引理中，我们建立最小均方误差与 PAC 之间的联系。证明见附录A 和B。

引理 2。假设 $(Y,Z)\sim D$ 是一对随机变量，其中 Y 取值于 $\{-1,1\}$，而 Z 取值于某个集合 $\mathcal{Z}$。假设 $g:\mathcal{Z}\to\{-1,1\}$ 是根据 Z 预测 Y 的任意预测器。那么，

$$\mathbb{P}\Big\{Y\neq g(Z)\Big\}=\frac{1}{2}-\frac{1}{2}\langle Y_{MMSE},g\rangle.$$

此外，令 opt_Z 为在给定 Z 时所有 Y 的预测器中的最小 0-1 损失。那么，

$$opt_Z=\frac{1}{2}-\frac{1}{2}\mathbb{E}\Big[\big|\mathbb{E}[Y|Z]\big|\Big]. \tag{4}$$

最后，$g^*\equiv\text{sign}[Y_{MMSE}]$ 是最优预测器。

引理 3。令 $\mathcal{Z}$ 为任意集合，且 $h:\mathcal{Z}\to\mathbb{R}$ 为任意有界函数。假设 $(Y,Z)\sim D$ 是一对随机变量，其中 Y 取值于 $\{-1,1\}$，而 Z 取值于 $\mathcal{Z}$。那么，

$$\mathbb{P}\Big\{Y\neq\text{sign}[h(Z)]\Big\}\leq opt_Z+U\big(\mathbb{E}\Big[\big(Y_{MMSE}-h(Z)\big)^2\Big]\big),$$

其中 U 是一个定义为 $U(x)=x^3+\frac{3}{2}x^2+\frac{3}{2}x$ 的多项式。

与学习 k-联合函数的联系：根据上述结果，我们可以推导出学习许多类别（例如 k-联合函数）时的误差界。令 $\mathcal{J}$ 为 $[d]$ 的一个子集，其中包含 k 个元素。将 $Z=X^{\mathcal{J}}$ 设为我们的观测变量。考虑所有定义在 $\mathcal{J}$ 坐标上的多项式，并将其作为输入变量。使平方损失最小的多项式，定义为将 Y 投影到子集 $\mathcal{J}$ 上所得的结果。该多项式的形式化定义如下：

$$\Pi_Y^{\mathcal{J}}:=\arg\min_{p\in\mathcal{P}_k}\|Y-p(X^{\mathcal{J}})\|_{2,D} \tag{5}$$

其中 opt 是次数至多为 k 的多项式集合。注意，$\Pi_Y^{\mathcal{J}}$ 是根据观测 $Z=X^{\mathcal{J}}$ 对 Y 的最小均方误差。于是，我们立即可以从引理 2 得到以下结果。

推论 1。令 opt 为给定 $k\leq d$ 时所有 k-联合函数中的最小 0-1 误差。那么，

$$\text{opt}=\frac{1}{2}-\frac{1}{2}\max_{\mathcal{J}\subseteq[d],|\mathcal{J}|=k}\|\Pi_Y^{\mathcal{J}}\|_{1,D}. \tag{6}$$

基于这些结果，我们可以证明关于使用 $\mathcal{L}_2$ 回归进行PAC 学习 k-联合函数的定理1 。

C. 定理的证明*1*

对于任意 $\mathcal{J}$，令 $\hat{}p_{\mathcal{J}}$ 为经验多项式回归的输出，即 $\hat{}p_{\mathcal{J}}=\arg\min_{p\in\mathcal{P}_k}\|Y-p_{\mathcal{J}}\|_{2,\hat{}D}$，其中 $\hat{}D$ 为经验分布。注意，所选预测器的形式为 $\text{sign}[\hat{}p_{\mathcal{J}}]$，与算法 1 中相同。因此，根据推论 3，将 D 替换为 $\hat{}D$，并将 $Z=$ 替换为 $X^{\mathcal{J}}$，可得 $\text{sign}[\hat{}p_{\mathcal{J}}]$ 的经验损失满足 $\mathcal{L}_{\hat{}D}(\text{sign}[\hat{}p_{\mathcal{J}}])\leq\frac{1}{2}-\frac{1}{2}\|\hat{}p_{\mathcal{J}}\|_{1,\hat{}D}$，其中引理 3 中的 $U(\cdot)$ 项为零，因为 $\hat{}p_{\mathcal{J}}$ 是 $\hat{}\hat{}D$ 下 Y 的最小均方误差。接下来，我们对所有 k 元子集 $\mathcal{J}$ 两边取最小值。根据推论 1，将 D 替换为 D 后，上述不等式右侧关于 $\mathcal{J}$ 取最小值即为最小经验损失 opt。这意味着

$\min_{\mathcal{J}:|\mathcal{J}|=k} \hat{\mathcal{L}}_D(\text{sign}[\hat{p}_{\mathcal{J}}]) = \widehat{opt}$. 因此，我们证明了最小经验损失可由 $\mathcal{L}_2$ 多项式回归取得。自然，下一步是将这一结果推广到泛化损失。这部分可由 VC 理论中的标准论证得出（参见 Mohri 等人，2018 中的推论 3.19），并利用 k-junta 类的 VC 维度小于 $2^k + O(k \log d)$ 这一事实。具体而言，给定 $\delta \in (0,1)$，以概率 $(1-\delta)$，泛化损失小于 opt $+ O\left(\sqrt{\frac{2^k+k}{n}\log d \log \frac{2^k+k}{n}}\right) + \sqrt{\frac{\log(1/\delta)}{2n}}$，其中 n 为样本数。由此不等式即可证明该定理。

D. 傅里叶域中的 PAC 学习

接下来，我们分析傅里叶算法并证明定理 2。我们研究傅里叶域中的 PAC 学习问题。为此，我们先给出以下引理，以建立预测损失与傅里叶系数之间的联系。

引理 4. 设 $(\mathbf{X}\ Y) \sim D$，其中 D 是定义在 $\{-1,1\}^{d+1}$ 上的分布。那么，任意 $g(\mathbf{x})$ 的预测损失等于

$$\mathcal{L}_D(g) = \frac{1}{2} - 2^{d-1} \sum_{\mathcal{S}\subseteq[d]} a_{\mathcal{S}} g_{\mathcal{S}},$$

其中 $g_{\mathcal{S}}$ 是 g 对应于 $\mathcal{S}$的（均匀）傅里叶系数，如事实 *1*中所述；而 $a_{\mathcal{S}}$ 是 Y 的随机傅里叶系数，如 (1) 中所述。

证明： 回忆式（3）可知， $\mathcal{L}_D(g) = \frac{1}{2} - \frac{1}{2}\mathbb{E}[Yg(\mathbf{X})]$。接着，根据式（ $a_{\mathcal{S}}$ 中 1）的定义，我们有

$$\begin{aligned}
\mathbb{E}[Yg(\mathbf{X})] &= \sum_{y,\mathbf{x}} D(x,y) y g(\mathbf{x}) \\
&= \sum_{y,\mathbf{x}} yD(x,y)\Big(\sum_{\mathcal{S}} g_{\mathcal{S}}\chi_{\mathcal{S}}(\mathbf{x})\Big) \\
&= \sum_{\mathcal{S}} g_{\mathcal{S}} \sum_{y,\mathbf{x}} yD(x,y)\chi_{\mathcal{S}}(\mathbf{x}) \\
&= \sum_{\mathcal{S}\subseteq[d]} g_{\mathcal{S}}(2^d a_{\mathcal{S}}),
\end{aligned}$$

正如所需。 ■

有趣的是，借助这个引理，在任意分布 D 下的预测损失都可以用 $g_{\mathcal{S}}$的傅里叶系数表示，这些系数对应于 g 在均匀分布下的傅里叶系数。我们利用这一思路，在附录 C 中证明以下引理。

引理 5. 设 $(\mathbf{X}\ Y) \sim D$，其中 D是 $\{-1,1\}^{d+1}$上的一个分布。给定任意子集坐标 $\mathcal{J}$，令 $f^{\mathcal{J}}(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} a_{\mathcal{S}}\chi_{\mathcal{S}}(\mathbf{x})$，其中 $a_{\mathcal{S}}$是 Y的随机傅里叶系数。令 $h_{\mathcal{J}}$为定义在坐标 $\mathcal{J}$上的任意实值函数，则 $g \equiv \text{sign}[h_{\mathcal{J}}]$ 的预测损失有如下界：

$$\mathcal{L}_D(g) \le \frac{1}{2}(1 - \|f^{\mathcal{J}}\|_{1,\text{unif}}) + U(\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,\text{unif}}),$$

其中，范数是在均匀分布下计算的，且 $U(x) = x^3 + \frac{3}{2}x^2 + \frac{3}{2}x$。

这个引理与引理 3 的不同之处在于， $f^{\mathcal{J}}$ 并非 Y 的最小均方误差估计，因为它是基于均匀傅里叶展开定义的。然而，它给出了最优损失 *opt*的一种不同刻画。

推论 2。 *ptim*在任意分布下， k-联合函数的 al 损失满足以下等式

$$\text{opt} = \frac{1}{2} - \frac{1}{2}\max_{\mathcal{J}\subseteq[d],|\mathcal{J}|=k} \|f^{\mathcal{J}}\|_{1,\text{unif}}.$$

基于 根据这些结果，我们证明关于傅里叶算法的 PAC 学习的定理 2 。 .

E. 定理2的证明

我们通过证明 $\hat{g}$ 在算法 2 中达到了 k-联合函数的 opt 来证明该定理。回顾一下，$\hat{g} \equiv$ 符号$[\hat{f}^{\hat{\mathcal{J}}}]$，其中 $\hat{f}^{\hat{\mathcal{J}}}$ 是为所选子集 $\hat{\mathcal{J}}$ 构造的。因此，根据引理 5，$\hat{g}$ 的预测损失有如下界：

$$\mathcal{L}_D(\hat{g}) \leq \frac{1}{2}(1 - \|f^{\hat{\mathcal{J}}}\|_{1,unif}) + U(\|f^{\hat{\mathcal{J}}} - \hat{f}^{\hat{\mathcal{J}}}\|_{2,unif}).$$

接下来，我们对右侧的第二项进行界定。注意，$\hat{f}^{\hat{\mathcal{J}}} \equiv \sum_{\mathcal{S}\subseteq\hat{\mathcal{J}}} \hat{a}_{\mathcal{S}}\chi_{\mathcal{S}}$。由帕塞瓦尔恒等式可得

$$\|f^{\hat{\mathcal{J}}} - \hat{f}^{\hat{\mathcal{J}}}\|^2_{2,unif} = \sum_{\mathcal{S}\subseteq\hat{\mathcal{J}}} (a_{\mathcal{S}} - \hat{a}_{\mathcal{S}})^2. \tag{7}$$

考虑所有 $\mathcal{S} \subseteq [d]$ 个元素数至多为 k 的 K。令 $|\mathcal{J}| = k$为此类子集的数量。那么，当 $|\mathcal{J}| = k$时，利用引理1 可将上述求和界定为：

$$\|f^{\mathcal{J}} - \hat{f}^{\mathcal{J}}\|^2_{2,unif} \leq 2^k \sup_{\mathcal{S}:|\mathcal{S}|\leq k} (a_{\mathcal{S}} - \hat{a}_{\mathcal{S}})^2 \leq \frac{2^k}{2n}\log\frac{2K}{\delta},$$

其中第二个不等式 等式成立的概率至少为$(1-\delta)$。因此，预测损失满足 fies

$$\mathcal{L}_D(\hat{g}) \leq \frac{1}{2}(1 - \|f^{\hat{\mathcal{J}}}\|_{1,unif}) + O\left(\sqrt{\frac{2^k}{n}\log\frac{K}{\delta}}\right),$$

其中我们用到了以下事实：$U(x) \leq 4x$ 对于 $x \in [0,1]$成立。接下来，我们通过考察所有 $\mathcal{J}$个 k元素的坐标 $\mathcal{J}$，对右侧进行最小化，以选择 $\mathcal{J}^*$。令 $\mathcal{J}^*$ 为最优集合。于是，由推论2可得：

$$\mathcal{L}_D(\text{sign}[\hat{f}^{\mathcal{J}^*}]) \leq opt + O\left(\sqrt{\frac{2^k}{n}\log\frac{K}{\delta}}\right).$$

注意，$\mathcal{J}^*$ 不一定与算法所选择的 $\mathcal{J}$ 相同。然而，由于 $\mathcal{J}$ 是使经验损失最小的 k 元坐标，因此 $\mathcal{L}_D(\text{sign}[\hat{f}^{\hat{\mathcal{J}}}]) \leq \mathcal{L}_D(\text{sign}[\hat{f}^{\mathcal{J}^*}])$。因此，根据麦克迪米德不等式，以概率 $(1-\delta)$ 可得

$$\mathcal{L}_{\hat{D}}(\text{sign}[\hat{f}^{\mathcal{J}^*}]) \leq \mathcal{L}_D(\text{sign}[\hat{f}^{\mathcal{J}^*}]) + \sqrt{\frac{k}{2n}\log\frac{2}{\delta}},$$

这里我们利用了如下事实：坐标 $\mathcal{J}^*$ 上至多有 2^k 个布尔函数。总而言之，我们证明了

$$\mathcal{L}_{\hat{D}}(\text{sign}[\hat{f}^{\hat{\mathcal{J}}}]) \leq opt + O\left(\sqrt{\frac{2^k}{n}\log\frac{K}{\delta}}\right).$$

假设 $k \leq d/2$，我们将 K 上界估计为

$$K \leq \sum_{\ell=0}^{k}\binom{d}{\ell} \leq 1 + k\binom{d}{k} = 1 + \frac{d^k}{(k-1)!}.$$

其余部分 论证可由 VC 理论推出，只需在左侧将 D 替换为 $\hat{D}$ 即可 e.

算法 2 的计算复杂度主要由估计所有 K 傅里叶系数的过程决定。每次估计需要 $O(nk)$。因此，该算法的总体计算复杂度为 $O(nkK) = O(nk\ \frac{d^k}{(k-1)!})$，如表 I所示。

算法 3：使用 $\mathcal{L}_2$-多项式回归进行学习

输入：训练样本 $\mathcal{S}_n = \{(\mathbf{x}(i)\ y(i))\}_{i=1}^n$，次数参数 k。1 找到一个次数不超过 k 的多项式 $\hat{}p$，使 $\frac{1}{n}\sum_i \big(y(i) - p(\mathbf{x}(i))\big)^2$ 最小。2 找到 $\theta \in [-1,1]$，使 $\mathrm{sign}[\hat{}p(\mathbf{x}) - \theta]$ 的经验误差最小。3 返回 $\hat{}g \equiv \mathrm{sign}[\hat{}p - \theta]$。

F. 使用最小均方误差进行 PAC 学习

最后，我们将讨论一个更一般的结论，以说明最小均方误差的 PAC 可学习性。

$\hat{}$ $\hat{}$定理 3。假设 $\mathcal{A}$ 是一个输出 符号$[Y_{MMSE}]$ 的算法，其中 Y_{MMSE} 是给定观测样本 $\mathcal{S}_n$ 后 $Y\hat{}$ 的经验最小均方误差。那么，$\mathcal{A}$ 无假设 PAC 学习任意包含 符号$[Y_{MMSE}]$ 的概念类 $\mathcal{C}$，其误差至多为

$$opt_{\mathcal{C}} + O\Big(\sqrt{\frac{VC}{n}\log(\frac{n}{\delta VC}}\Big),$$

其中 VC 是 $\mathcal{C}$ 的 VC 维度。

这一结果可由引理 3 应用于经验损失后，再结合 VC 理论得出。

IV. 学习其他假设类

本节中，我们使用基础版 $\mathcal{L}_2$ 多项式回归（见算法 3）来研究更一般的概念类。一个重要的概念类是由可用固定次数多项式近似的预测器构成的集合，Kalai 等人（2008；Blais 等人，2010）对此进行了研究。

(ϵ, k)-近似概念类：给定 $\epsilon \in [0,1]$、$k \in \mathbb{N}$ 以及定义在 $\mathcal{X}$ 上的任意概率分布 $D_{\mathbf{X}}$，若由函数 $c: \mathbb{R}^d \mapsto \{-1,1\}$ 构成的概念类 $\mathcal{C}$ 满足以下条件，则称其为 (ϵ,k)-近似的：

$$\sup_{c\in\mathcal{C}} \inf_{p\in\mathcal{P}_k} \mathbb{E}\big[\big(c(\mathbf{X}) - p(\mathbf{X})\big)^2\big] \le \epsilon^2,$$

其中 $\mathcal{P}_k$ 是所有次数不超过 k 的多项式构成的集合。

我们在附录 D 中证明，$\mathcal{L}_2$ 多项式回归能够学习近似概念类，且误差至多为 $2opt + \epsilon$。与 Linial 等人（1993）中的最佳已知界 $8opt$ 相比，这是一个改进。

定理 4。给定 $\epsilon > 0$ 和 $k \in \mathbb{N}$，次数为 k $\mathcal{L}_2$ 的多项式回归（算法 3）可以学习任意 (ϵ,k)-近似概念类，成功概率大于 $(1-\delta)$，且误差至多为

$$2\mathrm{opt} + 3\epsilon + O\Big(\sqrt{\frac{2\ d^{k+1}}{n}\log\frac{en}{d^{k+1}}}\Big) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

其中，d 是输入维度，n 是样本量。

注意，将输入从二元改为非二元时，$\mathcal{L}_2$ 多项式回归不一定是无关 PAC 学习器，因为标量会增至 2opt。

该结果通过使用附录D-A中证明的以下引理推导得出，从而无需进行随机舍入。

引理 6. 假设 θ 是一个随机变量，其概率密度函数为 $f_\theta(t) = 1 - |t|$，其中 $t \in [-1,1]$。那么，对于任意多项式 p，以下界成立

$$\mathbb{E}_\theta\Big[\mathcal{L}_{\hat{D}}(\mathrm{sign}[p(\mathbf{X}) - \theta])\Big] \le \frac{1}{2}\|Y - p\|_{2,\hat{D}}^2.$$

结论

本文研究基于 $\mathcal{L}_2$ 多项式回归的算法进行 PAC 学习。主要地，我们证明，基于$\mathcal{L}_2$ 的算法是针对 k-junta 类的 PAC 学习器。此外，我们提出了一种基于（均匀）布尔傅里叶展开的效率更高的 PAC 学习算法。我们的方法依赖于两个框架：一个将最小均方误差与 PAC 联系起来，另一个将 PAC 与布尔傅里叶展开联系起来。借助这种方法以及分析向量空间的强大工具，我们推导出 0-1 损失与平方损失之间更紧的界。

参考文献

J. Arpe 和 E. Mossel。通过随机游走无关地学习联合函数。2008年6月。A. Birnbaum 和 S. S. Shwartz。使用零一损失学习半空间：时间-准确率权衡。载于《神经信息处理系统进展》，第 926–934 页，2012年。E. Blais、R. O'Donnell 和 K. Wimmer。任意乘积分布下的多项式回归。《机器学习》，80(2-3):273–294，2010年。A. Blum、M. Furst、J. Jackson、M. Kearns、Y. Mansour 和 S. Rudich。利用傅里叶分析弱学习析取范式并刻画统计查询学习。载于第二十六届 ACM计算理论研讨会论文集——STOC 94。ACM 出版社，1994。doi: 10.1145/195058.195147。N. H. Bshouty 和 A. Costa。通过成员查询精确学习联合函数。载于算法学习理论（ALT），第 115–129 页。施普林格国际出版，2016。doi: 10.1007/978-3-319-46379-7 8。X. Chen、R. Jayaram、A. Levi 和 E. Waingarten。通过子立方体条件化学习和测试 junta 分布。载于 M. Belkin 和 S. Kpotufe（编），第三十四届学习理论会议论文集，机器学习研究论文集第 134 卷，第 1060–1113 页。PMLR，2021年8月15日至19日。T. A. Courtade 和 G. R. Kumar。哪些布尔函数能使噪声输入上的互信息最大化？IEEE 信息论汇刊，60(8):4515–4525，2014。A. De、E. Mossel 和 J. Neeman。junta 相关性可测试。载于 2019年 IEEE 第60届计算机科学基础年会（FOCS）。IEEE，2019年11月。I. Diakonikolas、T. Gouleakis 和 C. Tzamos。具有 Massart 噪声的分布无关 PAC 学习半空间。载于《神经信息处理系统进展》，第 4749–4760 页，2019年。P. Drineas、M. W. Mahoney 和 S. Muthukrishnan。L2 回归及其应用的采样算法。载于第十七届 ACM-SIAM 离散算法年会论文集（SODA '06），第 1127–1136 页，美国，2006年。工业与应用数学学会。ISBN 0898716055。P. Drineas、M. W. Mahoney、S. Muthukrishnan 和 T. Sarlos。更快的最小二乘近似。《数值数学》，117(2):219–249，2010年10月。doi: 10.1007/s00211-010-0331-6。E. Fischer、G. Kindler、D. Ron、S. Safra 和 A. Samorodnitsky。junta 测试。《计算机与系统科学杂志》，68:753–787，2004年。M. L. Furst、J. C. Jackson 和 S. W. Smith。改进对 AC^0 函数的学习。载于 COLT，第 91 卷，第 317–325 页，1991年。F. L. Gall 和 F. Urrutia。利用 Coppersmith-Winograd 张量的幂改进矩形矩阵乘法。载于第二十九届 ACM-SIAM 离散算法年会论文集，第 1029–1046 页。工业与应用数学学会，2018年1月。doi: 10.1137/1.9781611975031.67。I. Guyon 和 A. Elisseeff。变量与特征选择简介。《机器学习研究期刊》，3(Mar):1157–1182，2003年。M. Heidari、S. S. Pradhan 和 R. Venkataramanan。具有有偏输入的布尔函数：近似与

噪声敏感性。载于 IEEE 国际信息论研讨会（ISIT）论文集，第 1192–1196 页，2019年7月。doi: 10.1109/ISIT.2019.8849233。M. Heidari、J. Sreedharan、G. I. Shamir 和 W. Szpankowski。通过傅里叶展开实现信息充分性。载于 IEEE 国际信息论研讨会（ISIT）论文集，2021a年7月。M. Heidari、J. Sreedharan、G. I. Shamir 和 W. Szpankowski。通过离散傅里叶展开寻找相关信息。载于 M. Meila 和 T. Zhang（编），第38届国际机器学习会议论文集，机器学习研究论文集第 139 卷，第 4181–4191 页。PMLR，2021b年7月18–24日。V. Iyer、A. Tal 和 M. Whitmeyer。使用亚指数查询近似 junta 距离。2021年6月。J. C. Jackson。注意力焦点受限时含噪线性阈值函数的均匀分布可学习性。载于 G. Lugosi 和 H. U. Simon（编），学习理论，第 304–318 页，柏林、海德堡，2006年。Springer Berlin Heidelberg。ISBN 978-3-540-35296-9。A. T. Kalai、A. R. Klivans、Y. Mansour 和 R. A. Servedio。无关地学习半空间。SIAM计算杂志，37(6):1777–1805，2008年1月。doi: 10.1137/060649057。G. Kalai。社会选择理论中的噪声敏感性与混沌。技术报告，希伯来大学，2005年。M. J. Kearns、R. E. Schapire 和 L. M. Sellie。迈向高效的无关学习。《机器学习》，17(2-3):115–141，1994年。doi: 10.1007/bf00993468。A. R. Klivans、P. M. Long 和 R. A. Servedio。在恶意噪声下学习半空间。《机器学习研究期刊》，10(12)，2009年。J. Li 和 M. M´edard。布尔函数：噪声稳定性、非交互相关性与互信息。载于 IEEE 信息论研讨会论文集，2018年。N. Linial、Y. Mansour 和 N. Nisan。常深度电路、傅里叶变换与可学习性。《ACM期刊》，40(3):607–620，1993年。Z. Liu、X. Chen、R. A. Servedio、Y. Sheng 和 J. Xie。分布无关junta测试。《ACM算法汇刊》，15(1):1–23，2019年1月。doi: 10.1145/3264434。M. N. Y. U. Mohri、A.（Google）、I. Rostamizadeh、A.（加利福尼亚大学）和 B. Talwalkar。《机器学习基础》。麻省理工学院出版社，2018年。ISBN 0262039400。E. Mossel、R. O’Donnell 和 R. P. Servedio。学习联合函数。载于 ACM计算理论研讨会论文集，第 206–212 页，2003年。E. Mossel、R. O’Donnell 和 R. A. Servedio。学习由 k 个相关变量构成的函数。《计算机与系统科学杂志》，69 (3):421–434，2004年。R. O’Donnell。《布尔函数分析》。剑桥大学出版社，2014年。R. A. Servedio、L.-Y. Tan 和 J. Wright。自适应性有助于测试 junta。载于第30届计算复杂性会议论文集（CCC ’15），第 264–279 页，德国达格斯图尔，2015年。Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik。ISBN 9783939897811。S. Shalev-Shwartz 和 S. Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，美国纽约州纽约市，2014年。ISBN 1107057132, 9781107057135。J. Suykens 和 J. Vandewalle。《神经处理快报》，9(3):293–300，1999年。doi: 10.1023/a:1018628609742。L. G. Valiant。一种可学习理论。《ACM通讯》，27(11):1134–1142，1984年11月。doi: 10.1145/1968.1972。S. S. Vempala 和 Y. Xiao。从局部最优解中提取结构：通过高阶主成分分析学习子空间junta。2011年8月。N. Weinberger 和 O. Shayevitz。关于二次损失下预测的最优布尔函数。IEEE 信息论汇刊。

信息论，63(7):4202–4217，2017。N. Weinberger 和 O. Shayevitz。《自预测布尔函数》。载于 IEEE 信息论研讨会论文集，2018。R. d. Wolf。《布尔立方体上的傅里叶分析简明导论》。研究生综述第 1 号。计算理论图书馆，2008。doi: 10.4086/toc.gs.2008.001。

附录 A 引理2的引理

证明

证明：根据正文中的（3），g的泛化误差可以写成$\frac{1}{2}-\frac{1}{2}\langle Y,\ g\rangle$。该内积等于以下表达式

$$\begin{aligned}\langle Y,g\rangle &= \mathbb{E}\big[Y g(Z)\big] = \mathbb{E}_Z\Big[\mathbb{E}_{Y|Z}\big[Y g(Z)\big]\mid Z\big]\Big]\\ &= \mathbb{E}_Z\Big[\mathbb{E}_{Y|Z}\big[Y|Z\big]g(Z)\Big].\end{aligned}$$

令 $Y_{MMSE}=\mathbb{E}[Y|Z]$。因此，我们得到

$$\mathbb{P}\Big\{Y\neq g(Z)\Big\}=\frac{1}{2}-\frac{1}{2}\langle Y_{MMSE},g\rangle \tag{8}$$

注意

$$\mathbb{P}\Big\{Y\neq g(\mathbf{X})\Big\}=\frac{1}{2}-\frac{1}{2}\langle Y_{MMSE},g\rangle_D\geq\frac{1}{2}-\frac{1}{2}\langle|Y_{MMSE}|,|g|\rangle_D\geq\frac{1}{2}-\frac{1}{2}\|Y_{MMSE}\|_{1,D},$$

其中，最后一个不等式由$|g(Z)|=1$得出。因此，我们得到界$opt_Z\geq\frac{1}{2}-\frac{1}{2}\|Y_{MMSE}\|_{1,\ D}$。由此，我们建立了关于 opt_Z 的一个下界。接下来，我们证明该界可以达到。为此，构造一个预测器 $g^*=$符号$[Y_{MMSE}]$。那么，根据上述论证，这样的 g 的泛化误差等于

$$\mathbb{P}\Big\{Y\neq \operatorname{sign}[Y_{MMSE}]\Big\}=\frac{1}{2}-\frac{1}{2}\langle Y_{MMSE},\operatorname{sign}[Y_{MMSE}]\rangle_D=\frac{1}{2}-\frac{1}{2}\|Y_{MMSE}\|_{1,D},$$

其中，最后一个等式是根据恒等式 $\langle h,$ 符号$[h]\rangle=\|h\|_1$ 得出的，该恒等式对任意函数 h 均成立。因此，我们证明了下界可以达到，这意味着$opt_Z=\frac{1}{2}-\frac{1}{2}\|Y_{MMSE}\|_{1,\ D}$，且 $g^*=$符号$[Y_{MMSE}]$ 是最优预测器。■

Appendix B
Proof of Lemma 3

证明：为简便起见，令 $f(z)=\mathbb{E}[Y|z]$对任意 $z\in\mathcal{Z}$ 成立。因此，$f(Z)=Y_{MMSE}$。由引理2的证明中的（8）可知，$[h]$ 的符号泛化误差可写为.....因此，我们得到

$$\mathbb{P}\Big\{Y\neq \operatorname{sign}[h(Z)]\Big\}=\frac{1}{2}-\frac{1}{2}\langle f,\operatorname{sign}[h]\rangle$$

回顾一下，$\|f\|_{2,\ D}:=\sqrt{\mathbb{E}_D[f\ (X)\ ^2]}$。因此，$\|a-b\|^2_{2,\ D}=\|a\|^2_{2,\ D}+\|b\|^2_{2,\ D}-2\langle a,\ b\rangle$。因此，

$$\begin{aligned}\langle f,\operatorname{sign}[h]\rangle &= \frac{1}{2}\big(\|f\|^2_{2,D}+\|\operatorname{sign}[h]\|^2_{2,D}-\|f-\operatorname{sign}[h]\|^2_{2,D}\big)\\ &= \frac{1}{2}\big(\|f\|^2_{2,D}+1-\|f-\operatorname{sign}[h]\|^2_{2,D}\big),\end{aligned}$$

这里用到了这样一个事实：$|$符号$[h]|=1$。因此，

$$\mathbb{P}\Big\{Y\neq \operatorname{sign}[h(Z)]\Big\}=\frac{1}{4}\big(1-\|f\|^2_{2,D}+\|f-\operatorname{sign}[h_{\mathcal{J}}]\|^2_{2,D}\big). \tag{9}$$

接下来，我们对 $\|f-$符号$[h_{\mathcal{J}}]\|^2_{2,\ D}$ 进行界定。通过加上再减去 h，可得 t

$$\begin{aligned}\|f-\operatorname{sign}[h]\|^2_{2,D} &\overset{(a)}{\leq}\Big(\|f-h\|_{2,D}+\|h-\operatorname{sign}[h]\|_{2,D}\Big)^2\\ &= \Big(\|f-h\|^2_{2,D}+\underbrace{\|h-\operatorname{sign}[h]\|^2_{2,D}}_{\text{(I)}}+2\|f-h\|_{2,D}\underbrace{\|h-\operatorname{sign}[h]\|_{2,D}}_{\text{(II)}}\Big),\end{aligned} \tag{10}$$

其中(a) 由闵可夫斯基不等式对2-范数的应用可得。接下来，我们分别给出项 (I) 和 (II) 的上界：

对 (I) 进行界定：注意到 $|h - \text{符号}[h]| = |1 - |h||$. 因此，

$$\begin{aligned}\text{(I)} = \|h - \operatorname{sign}[h]\|_{2,D}^2 &= \mathbb{E}\left[(1 - |h(Z)|)^2\right] \\ &= 1 + \|h\|_{2,D}^2 - 2\|h\|_{1,D}. \end{aligned} \tag{11}$$

界定（II）：由（11），可得

$$\begin{aligned}\|h - \operatorname{sign}[h]\|_{2,D}^2 &= 1 + \|h\|_{2,D}^2 - 2\|h\|_{1,D} \\ &\overset{(a)}{\le} 1 + 2(\|f\|_{2,D}^2 + \|f - h\|_{2,D}^2) - 2\|h\|_{1,D} \\ &\overset{(b)}{=} 1 + 2(\|f\|_{2,D}^2 + \|f - h\|_{2,D}^2) - 2\big(\|f\|_{1,D} + (\|h\|_{1,D} - \|f\|_{1,D})\big) \\ &= 1 + 2(\|f\|_{2,D}^2 - \|f\|_{1,D}) + 2\|f - h\|_{2,D}^2 - 2\big(\|h\|_{1,D} - \|f\|_{1,D}\big) \\ &\overset{(c)}{\le} 1 + 2\|f - h\|_{2,D}^2 - 2\big(\|h\|_{1,D} - \|f\|_{1,D}\big) \\ &\overset{(d)}{\le} 1 + 2\|f - h\|_{2,D}^2 + 2\|f - h\|_{2,D}, \end{aligned} \tag{12}$$

其中，(a) 由 2-范数的闵可夫斯基不等式和不等式 $(x+y)^2 \le 2(x^2+y^2)$得到。等式(b)通过加上再减去 $\|f\|_1$ 得到。不等式 (c)成立，因为 $|f(x)| \le 1$意味着 $\|f\|_2^2 \le \|f\|_1$。最后，(d) 成立，这是因为以下不等式链

$$\Big|\|f\|_{1,D} - \|h\|_{1,D}\Big| \le \|f - h\|_{1,D} \le \|f - h\|_{2,D}, \tag{13}$$

其中，第一 项来自 1-范数的闵可夫斯基不等式，第二项则来自 Hölder 不等式 s.

接下来，我们证明，不带平方的量 $\|h - \text{符号}[h_{\mathcal{J}}]\|_{2,\ D}$ 上界不超过（12）右侧的同一项。也就是说

$$\text{(II)} = \big\|h - \operatorname{sign}[h_{\mathcal{J}}]\big\|_{2,D} \le \lambda_1 \triangleq 1 + 2\|f - h\|_{2,D}^2 + 2\|f - h\|_{2,D}. \tag{14}$$

论证如下：若 $\|h - \text{符号}[h_{\mathcal{J}}]\|_{2,\ D}$ 小于 1，则上界显然成立，因为 $\lambda_1 \ge 1$；否则，该量小于其平方，因此上界成立。

现在结合（14）、（11）和（10）可得

$$\|f - \operatorname{sign}[h]\|_{2,D}^2 \le \|f - h\|_{2,D}^2 + 1 + \|h\|_{2,D}^2 - 2\|h\|_1 + 2\lambda_1\|f - h\|_{2,D}. \tag{15}$$

由此界和（9），误差概率满足：

$$4\mathbb{P}\Big\{Y \ne \operatorname{sign}[h(Z)]\Big\} \le 2 - 2\|h\|_{1,D} + \underbrace{\|h\|_{2,D}^2 - \|f\|_{2,D}^2}_{\text{(III)}} + \|f - h\|_{2,D}^2 + 2\lambda_1\|f - h\|_{2,D}. \tag{16}$$

接下来，我们对记为 (III) 的项进行界定。

对 (III) 进行界定：由 2-范数的闵可夫斯基不等式可得

$$\begin{aligned}\|h\|_{2,D}^2 &\le \Big(\|f\|_{2,D} + \|h - f\|_{2,D}\Big)^2 \\ &= \|f\|_{2,D}^2 + \|h - f\|_{2,D}^2 + 2\|f\|_{2,D}\|h - f\|_{2,D} \\ &\le \|f\|_{2,D}^2 + \|h - f\|_{2,D}^2 + 2\|h - f\|_{2,D} \end{aligned}$$

其中，第二个不等式由贝塞尔不等式得出，这意味着 $\|f\|_{2,\ D} \le 1$。因此，(16) 中的项 (III) 可上界为

$$\text{(III)} \le \lambda_2 \triangleq \|h - f\|_{2,D}^2 + 2\|h - f\|_{2,D}. \tag{17}$$

根据 (16) 和 (17) 中的界，我们得到

$$\begin{aligned}4\mathbb{P}\Big\{Y \neq \operatorname{sign}[h(Z)]\Big\} &\leq 2 - 2\|h\|_{1,D} + \lambda_2 + \|f-h\|_{2,D}^2 + 2\lambda_1\|f-h\|_{2,D} \\ &= 2 - 2\|f\|_{1,D} + 2\Big(\|f\|_{1,D} - \|h\|_{1,D}\Big) + \lambda_2 + \|f-h\|_{2,D}^2 + 2\lambda_1\|f-h\|_{2,D} \\ &\leq 2 - 2\|f\|_{1,D} + 2\|f-h\|_{2,D} + \lambda_2 + \|f-h\|_{2,D}^2 + 2\lambda_1\|f-h\|_{2,D},\end{aligned}$$

其中，最后一个不等式由（13）得出。因此，根据 λ_1 和 λ_2的定义，以及引理陈述中的函数 U，我们得到

$$4\mathbb{P}\Big\{Y \neq \operatorname{sign}[h(Z)]\Big\} \leq 2 - 2\|f\|_{1,D} + 4U(\|f-h\|_{2,D}).$$

回顾 $f(z) = \mathbb{E}[Y|z]$，以及根据引理 2，$opt_Z = \frac{1}{2} - \frac{1}{2}\|f\|_{1,\ D}$，即可完成证明。 ■

Appendix C
Proof of Lemma 5

证明：根据正文中的引理 4，$g = \operatorname{sign}[h_{\mathcal{J}}]$ 的泛化误差可写为

$$\mathcal{L}_D(g) = \frac{1}{2} - 2^{d-1}\sum_{\mathcal{S}\subseteq[d]} a_{\mathcal{S}} g_{\mathcal{S}},$$

注意，由于 g 仅取决于坐标 $\mathcal{J}$，因此对于任意 $\mathcal{S} \nsubseteq \mathcal{J}$，都有 $g_{\mathcal{S}} = 0$。所以，上式可简化为

$$\mathcal{L}_D(g) = \frac{1}{2} - 2^{d-1}\sum_{\mathcal{S}\subseteq\mathcal{J}} a_{\mathcal{S}} g_{\mathcal{S}}.$$

注意，对于不同的 $\mathcal{S}$，$\chi_{\mathcal{S}}$彼此正交，并且$\sum_{\mathbf{x}} \chi_{\mathcal{S}}(\mathbf{x})^2 = 2^d$。因此，

$$\begin{aligned}\mathcal{L}_D(g) &= \frac{1}{2} - \frac{1}{2}\sum_{\mathbf{x}}\Big(\sum_{\mathcal{S}\subseteq\mathcal{J}} a_{\mathcal{S}}\chi_{\mathcal{S}}(\mathbf{x})\Big)\Big(\sum_{\mathcal{S}\subseteq\mathcal{J}} g_{\mathcal{S}}\chi_{\mathcal{S}}(\mathbf{x})\Big) \\ &= \frac{1}{2} - \frac{1}{2}\sum_{\mathbf{x}} f^{\mathcal{J}}(\mathbf{x})g(\mathbf{x}),\end{aligned}$$

其中 $f^{\mathcal{J}} \equiv \sum_{\mathcal{S}\subseteq\mathcal{J}} a_{\mathcal{S}}\chi_{\mathcal{S}}$。将 2^d乘入又除以，上述求和等于均匀分布上的内积，即

$$\sum_{\mathbf{x}} f^{\mathcal{J}}(\mathbf{x})g(\mathbf{x}) = 2^d\langle f^{\mathcal{J}}, g\rangle_{unif}.$$

因此，根据 g 的定义，我们得到

$$\mathcal{L}_D(g) = \frac{1}{2} - \frac{2^d}{2}\langle f^{\mathcal{J}}, \operatorname{sign}[h_{\mathcal{J}}]\rangle$$

采用推导（9）时的类似论证，我们可以证明

$$\mathcal{L}_D(g) = \frac{1}{2} - \frac{2^d}{4}\big(1 + \|f^{\mathcal{J}}\|_{2,unif}^2 - \|f^{\mathcal{J}} - \operatorname{sign}[h_{\mathcal{J}}]\|_{2,unif}^2\big). \tag{18}$$

注意，由于因子 2^d，且范数相关量是相对于均匀分布取的，这个方程与（9）不同。接下来我们对 2-范数相关量进行界定。注意，我们可以直接使用推导（15）时所用的相同论证，因为该论证对任意底层分布都成立。上述 2-范数相关量的上界如下

$$\|f^{\mathcal{J}} - \operatorname{sign}[h_{\mathcal{J}}]\|_{2,unif}^2 \leq \|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}^2 + 1 + \|h_{\mathcal{J}}\|_{2,unif}^2 - 2\|h_{\mathcal{J}}\|_{1,unif} + 2\lambda_1\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif},$$

其中 $\lambda_1 = 1 + 2\|f^{\mathcal{J}} - h\|_{2,\ unif}^2 + 2\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,\ unif}$。因此， g 的损失满足

$$\mathcal{L}_D(g) = \frac{1}{2} - \frac{2^d}{4}\Big(2\|h_{\mathcal{J}}\|_{1,unif} + \underbrace{\big(\|f^{\mathcal{J}}\|_{2,unif}^2 - \|h_{\mathcal{J}}\|_{2,unif}^2\big)}_{\text{(I)}} - \|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}^2 - 2\lambda_1\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}\Big).$$

注意，(I) $\geq -\lambda_2$满足 $\lambda_2 \triangleq \|h_{\mathcal{J}} - f^{\mathcal{J}}\|_{2,\ unif}^2 + 2\|h_{\mathcal{J}} - f^{\mathcal{J}}\|_{2,\ unif}$如（17）所示。接下来，通过加上并减去 $2\|f^{\mathcal{J}}\|_{1,\ unif}$，可得

$$\begin{aligned}\mathcal{L}_D(g) &\leq \frac{1}{2} - \frac{2^d}{4}\Big(2\|f^{\mathcal{J}}\|_{1,unif} + 2\underbrace{\big(\|h_{\mathcal{J}}\|_{1,unif} - \|f^{\mathcal{J}}\|_{1,unif}\big)} - \lambda_2 - \|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}^2 - 2\lambda_1\|f - h\|_{2,unif}\Big)\\ &\leq \frac{1}{2} - \frac{2^d}{4}\Big(2\|f^{\mathcal{J}}\|_{1,unif} - 2\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif} - \lambda_2 - \|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}^2 - 2\lambda_1\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}\Big),\end{aligned}$$

其中，我们利用（13）推导出该不等式。因此，根据 λ_1， λ_2和 U（x） $= x^3 + \frac{3}{2}x^2 + \frac{3}{2}x$的定义，可得

$$\mathcal{L}_D(g) \leq \frac{1}{2} - \frac{2^d}{2}\|f^{\mathcal{J}}\|_{1,unif} + 2^d U(\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}).$$

最 后，我们进一步对该表达式进行界定。注意，$\|\cdot\|_1 = 2^d\|\cdot\|_{1,\ unif}$。于是可得 t

$$\mathcal{L}_D(g) = \frac{1}{2} - \frac{1}{2}\|f^{\mathcal{J}}\|_1 + 2^d U(\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}) \leq \frac{1}{2} - \frac{1}{2}\|f^{\mathcal{J}}\|_1 + U(2^d\|f^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,unif}),$$

其中，最后的不 等式通过将 2^d 移入 U（$\cdot$）内而成立。引理证毕。 ■

Appendix D
Proof of Theorem 4

为了推导 $\hat{}g$ 的经验误差上界，我们首先考虑该算法的一个较弱版本。其思路是随机选择 θ，而不是像算法中那样对其进行优化。为此，我们使用第 6 节 IV中的引理。因此，根据该引理，并利用算法中选择 θ 以最小化经验误差这一事实，可得

$$\mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} \leq \frac{1}{2}\|Y - \hat{p}\|_{2,\hat{D}}^2, \tag{19}$$

其中 $\hat{}p$ 是 $\mathcal{L}_2$-多项式回归的输出，而 $\hat{}g \equiv \text{sign}[\hat{}p - \theta]$与算法 1中相同。令 c^* 为 (ϵ, k)-近似概念类中泛化误差最小的预测器。令 p 为一个次数为 k 的多项式，使得 $\|c^* - p\|_2 \leq \epsilon$。由于 $\hat{}p$ 使经验 2-范数最小，因此 (19) 的右侧满足

$$\frac{1}{2}\|Y - \hat{p}\|_{2,\hat{D}}^2 \leq \frac{1}{2}\|Y - p^*\|_{2,\hat{D}}^2. \tag{20}$$

接下来，我们对随机训练样本下经验误差的期望误差进行分析。由（19）和（20），可得以下不等式

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\}\Big] &\leq \frac{1}{2}\mathbb{E}\Big[\|Y - p^*\|_{2,\hat{D}}^2\Big] = \frac{1}{2}\|Y - p^*\|_{2,D}^2\\ &\overset{(a)}{\leq} \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \|p^* - c^*\|_{2,D}\Big)^2\\ &\leq \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \epsilon\Big)^2\\ &\overset{(b)}{\leq} \frac{1}{2}\Big(\|Y - c^*\|_{2,D}^2 + 4\epsilon + \epsilon^2\Big)\\ &\overset{(c)}{\leq} 2opt + \frac{5}{2}\epsilon,\end{aligned} \tag{21}$$

其中，（a）由 2-范数的闵可夫斯基不等式得出；（b）成立是因为 $\|Y - c^*\|_{2,\ D} \le 2$；（c）则是由（3）中的第二个等式以及 opt $= \mathbb{P}\{Y \neq c^*(\mathbf{X})\}$ 得出。

接下来，我们将 $\hat{}g$ 的经验误差与其泛化误差联系起来。注意，对于某个次数不超过 k 的多项式 d^{k+1}，所有形如 $\mathrm{sign}[p]$ 的函数，其 Vapnik–Chervonenkis（VC）维数不超过 d^{k+1}。因此，根据 VC 理论（参见（Mohri 等人，2018）中的推论 3.19），对于任意 δ，以至少 $(1-\delta)$ 的概率，以下不等式成立

$$\mathbb{P}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} \le \mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} + \sqrt{\frac{2\ d^{k+1}}{n} \log \frac{en}{d^{k+1}}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{22}$$

因此，证明 取期望并将其与（2）中的最后一个界结合起来，即可完成证明 1).

A. 引理 6 的证明

注意，如果 θ 介于 y 和 $p(\mathbf{x})$ 之间，则 $y \neq \mathrm{sign}(p(\mathbf{x}) - \theta)$。因此，相对于随机 θ，$\mathrm{sign}[p(\mathbf{X}) - \theta]$ 的期望经验误差等于

$$\begin{aligned}
&\mathbb{E}_\theta\Big[\mathbb{P}_{\hat{D}}\Big\{Y \neq \mathrm{sign}[p(\mathbf{X}) - \theta]\Big\}\Big] \\
&\quad = \frac{1}{n}\sum_i \mathbb{E}_\theta\Big[\mathbb{1}\big\{y_i \neq \mathrm{sign}(p(\mathbf{x}_i) - \theta)\big\}\Big] \\
&\quad = \frac{1}{n}\sum_i \underbrace{\mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), y_i] \bigcup [y_i, p(\mathbf{x}_i)]\Big\}}_{\mathbb{P}_i}.
\end{aligned} \tag{23}$$

接下来，我们证明，对于所有 $(\mathbf{x}_i, y_i)$，都有 $\mathbb{P}_i \le \frac{1}{2}(y_i - p(\mathbf{x}_i))^2$。设 $y_i = 1$。如果 $p(\mathbf{x}_i) > 1$，那么由于 $\theta \le 1$，$\mathbb{P}_i = 0$。如果 $p(\mathbf{x}_i) \in [0,1]$，那么

$$\begin{aligned}
\mathbb{P}_i = \mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), 1]\Big\} &= \int_{p(\mathbf{x}_i)}^{1} (1-t)dt \\
&= \frac{1}{2}\big(1 - p(\mathbf{x}_i)\big)^2 = \frac{1}{2}\big(y_i - p(\mathbf{x}_i)\big)^2.
\end{aligned}$$

若 $p(\mathbf{x}_i) \in [-1, 0]$，则

$$\begin{aligned}
\mathbb{P}_i = \mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), 1]\Big\} &= \int_{p(\mathbf{x}_i)}^{1} 1 - |t| dt \\
&= \frac{1}{2} + \int_{p(\mathbf{x}_i)}^{0} (1+t)dt \\
&= \frac{1}{2} - p(\mathbf{x}_i) - \frac{1}{2}(p(\mathbf{x}_i))^2 \\
&\le \frac{1}{2}(1 + |p(\mathbf{x}_i)|)^2 = \frac{1}{2}(y_i - p(\mathbf{x}_i))^2.
\end{aligned}$$

最后，若 $p(\mathbf{x}_i) < -1$，则 $\mathbb{P}_i = 1$，因为 $\theta \ge -1$。在这种情况下，同样有 $\mathbb{P}_i \le \frac{1}{2}(y_i - p(\mathbf{x}_i))^2$。由对称性，$y_i = -1$ 的情形也成立。因此，我们得到以下不等式

$$\mathbb{E}_\theta\Big[\mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\}\Big] \le \frac{1}{n}\sum_i \frac{1}{2}\big(y_i - p(\mathbf{x}_i)\big)^2.$$

注意到右侧等于 $\frac{1}{2}\|Y - p\|_{2,\hat{}D}^2$，即可完成证明

.