arXiv:2305.02304v1 [stat.ML] 3 May 2023

插值与支持向量机之间的新等价关系：核与结构化特征

Chiraag Kaushik* Andrew D.McRae†、Mark A. Davenport†
Vidya Muthukumar†‡

2023年5月4日

摘要

支持向量机（SVM）是一种监督学习算法，它通过寻找最大间隔线性分类器来完成学习；通常会先借助核技巧将数据映射到高维特征空间。近期研究表明，在某些过参数化程度足够高的情形下，支持向量机决策函数与最小范数标签插值函数完全一致。这种支持向量增殖（SVP）现象尤其值得关注，因为它使我们能够利用近期关于线性模型和核模型中无害插值的分析来理解支持向量机的性能。然而，先前关于SVP 的研究对数据/特征分布和谱作出了较为严格的假设。本文提出一种全新且灵活的分析框架，用于证明任意再生核希尔伯特空间中的 SVP，并允许标签采用灵活的一类生成模型。我们针对一般有界正交系统（例如傅里叶特征）和独立次高斯特征这两类特征，给出了 SVP 的成立条件。在这两种情形下，我们证明 SVP 会出现在许多先前研究未涵盖的重要情形中，并利用这些结果证明了关于核支持向量机分类的新型泛化结果。

1 引言

监督式机器学习领域近期的实证和理论研究发现了现代过参数化区域中出现的各种令人惊讶的现象（即模型中的自由参数数量远多于训练样本数量 [13, 6]）。例如，在观察到深度神经网络能够完美拟合含噪训练数据，同时仍能很好地泛化到新数据之后（参见例如 [35, 43]），一些理论研究表明，这种“无害插值”现象实际上甚至可以出现在更简单的线性回归和核回归设置中 [8, 7, 5]。在这一过参数化区域中，另一个同样令人惊讶的观察是，优化不同损失函数的训练过程仍可能产生相似的测试性能。例如， [36, 22,26, 16] 的实证研究表明，使用平方损失训练的核机器和深度神经网络——平方损失传统上用于带有连续标签的回归问题——所获得的分类性能可以与使用更受欢迎的交叉熵损失训练的模型相当。

受这些观察结果的启发，近期研究开始从线性模型入手，力求加深对损失函数在过参数化分类任务中所产生影响的理论认识。硬间隔支持向量机（SVM） [12] 是一种常用的分类算法。给定 n 线性可分的训练样本（x_i, y_i） $\in \mathbf{R}^d \times \{-1, 1\}$，其目标是通过求解以下凸规划，找到决策函数 $\hat{} \ f(x) = \langle \hat{}\theta, x \rangle$:

$$\hat{\theta} = \arg\min_{\theta \in \mathbf{R}^d} \|\theta\|_2 \quad \text{s.t. } y_i \theta^\top x_i \geq 1 \text{ for } i = 1, \ldots, n.$$

*佐治亚理工学院电气与计算机工程学院，美国佐治亚州亚特兰大（ckaushik7@gatech.edu, mdav@gatech.edu, vmuthukumar8@gatech.edu）。
†瑞士洛桑 EPFL 数学研究所（andrew.mcrae@epfl.ch）
‡佐治亚理工学院工业与系统工程学院，佐治亚州亚特兰大。

ˆ 等价地，该解使超平面 $\{x\colon\ f(x) = 0\}$ 到任意训练样本的最小距离最大化。 [33] 首次证明，在过参数化情形下，支持向量机决策函数 ˆ 可以与训练数据的最小范数插值器完全一致（即，学习得到的 $\theta =$ ˆ $\arg\min_{\theta\in\mathbf{R}^d}\|\theta\|_2$ 对所有 $i = 1,\dots,n$ 均产生 $f(x_i) = y_i$）。这一现象被称为支持向量增殖（SVP），具有若干重要意义。首先，由于已知这两种解分别可以通过对平方损失和逻辑损失进行梯度下降获得，因此，这为一项经验观察提供了理论依据：在过参数化情形下，不同的损失函数可能产生相似的行为 [36, 22]。此外，已有多项研究成果（例如， [33, 42, 41]）利用这一等价关系，通过理解离散标签的最小范数插值器来研究支持向量机的泛化性能，因为借助一种便于分析的闭式解，可以更轻松地对其进行分析。事实上，这种方法还可用于证明支持向量机在经典泛化界无法覆盖的情形下所具有的分类一致性结果，正如 [33, 21] 中所讨论的那样。

尽管此后我们对 SVP [21, 1, 42, 41, 10], 的理解取得了相当大的进展，但当前的刻画仍然关键性地依赖于关于数据或特征的强分布假设。具体而言，关于 SVP 何时出现的条件，仅在特征具有高斯条目或独立次高斯条目的情形下得到证明。因此，这些条件本质上仅适用于直接在输入数据上学习线性模型的情况，而此类输入数据本身具有高维特性。相比之下，在更一般的核支持向量机设置中，训练数据 $x \in \mathrm{R}^d$ 通常首先通过某种特征映射 $\phi\colon \mathrm{R}^d \to \mathrm{R}^D$进行变换（显式变换，或通过核技巧隐式变换），随后上述支持向量机解学习一个具有如下形式的函数 $f(x) = \langle\theta, \phi(x)\rangle$。对于几乎任意核或非线性特征映射的选择， $\phi(x_i)$的条目都不是独立的。事实上，在许多常见的学习场景中， $D \gg d$; 也就是说，低维或固定维度的训练数据（或有效维度较低的数据，例如来自低维流形的数据）会被映射到某些更高维甚至无限维的特征空间中，而这些特征彼此之间存在复杂的依赖关系。作为一个案例研究，SVP 最初是在 [33], 的实验中通过实证观察到的，该实验使用了一维数据和一种傅里叶特征映射，其条目具有显著的结构。现有的理论结果并不适用于这类情形。

1.1 贡献

本文研究核方法（即一般特征映射类别）中的 SVP，并给出其发生的精确非渐近条件。我们考虑如下用于分类的生成式标签模型： $\eta^*(x) = 2\mathrm{P}(y = 1 \mid x) - 1$，其中 η^* 是任意再生核希尔伯特空间中的函数，表示分类条件概率的一种自然仿射变换。我们的结果简要总结如下。

一般有界特征下的 SVP 在上述标签模型下，我们给出了核积分算子的谱和训练样本的充分条件；在这些条件下，训练数据的最小再生核希尔伯特空间范数插值函数与核支持向量机学到的决策函数相同。我们在第 3.1 节正式给出结果成立所需的主要假设。从较高层次来看，我们只要求主导核特征函数构成一个有界正交系统（BOS），并利用近期文献中关于噪声无害插值的条件，确保最小范数插值函数接近 [1] of η^*的一个衰减版本。随后，我们在第 3.2 节刻画这些假设下 SVP 发生的情形。与以往工作相比，我们对特征映射和核谱的要求显著更弱，并且表明 SVP 可在一般再生核希尔伯特空间情形下发生，而不要求特征相互独立或服从次高斯分布。我们的证明技术与假设标签为最坏情形 [33, 21, 42, 41, 10, 1], 的既有文献显著不同，并在 SVP 与插值方法精确误差分析工具之间建立了新的概念联系。

独立次高斯特征的精细界 在第3.3节中，我们针对特征相互独立且服从次高斯分布的情形，对结果进行了具体化和强化。值得注意的是，与以往研究成果不同，我们采用概率生成式标签模型和较弱的谱假设，因此能够证明，即使最小范数插值器在回归任务中具有良好的泛化性能，SVP 仍可能发生。据我们所知，这是首次在此类情形下确立 SVP。

[1]具体而言，我们定义一个算子 $\overline{\mathcal{S}}$，并要求$\hat{\eta} \approx \overline{\mathcal{S}}\eta^*$，且 $\overline{\mathcal{S}}$ 是一个 L_∞ 压缩算子。

核支持向量机在新情形下的泛化 与以往关于 SVP 的研究一样，我们的结果可以与对最小范数插值器的分析相结合，从而分析支持向量机的泛化性能。在第3.4节中，我们证明，我们的分析可用于证明核支持向量机在一些新情形下具有渐近分类一致性；在这些情形下，维度（有效意义上）过高，经典泛化界无法提供有用信息。与以往研究输入数据上线性模型的 SVP 研究成果相比，我们的结果表明，对于一系列有趣且对应于结构更丰富的特征映射的核，核支持向量机能够实现良好泛化。

1.2 相关工作

我们从两个方面梳理相关工作的讨论。

支持向量增殖 经验观察表明，在现代深度学习和核模型中，平方损失与交叉熵损失会产生相当的分类性能 [36, 22, 26, 16] ，这促使人们对支持向量增殖 [33, 21, 1, 10, 42, 41],这一现象以及更普遍的由各种损失函数 [27]获得的解之间的等价性产生了越来越浓厚的理论兴趣。特别是，由于支持向量机和最小范数插值器可以通过对交叉熵（或类似的指数尾部） [37, 24, 25, 18, 23] 损失与平方损失执行梯度下降获得，SVP 为理解可能出现这种损失函数等价性的过参数化情形提供了一条途径。

先前的研究给出了在何种条件下（通常以数据协方差和/或信噪比来刻画），对于任意（即最坏情况）标签选择，SVP现象能够以高概率发生。然而，这些结果都对数据分布和维度提出了较强的假设。最初的研究是在高斯设计下证明了SVP现象 [33], ，随后针对独立次高斯数据和Haar设计给出了充分条件 [21] ，并为高斯情形提供了一个逆向结果。此外， [1] 针对各向异性次高斯数据给出了SVP的逆向结果，并证明在各向同性情形下，SVP要求数据维度 $d = \Theta(n\log\ n)$。我们在第3.2.1节中进一步详细阐述了我们的上界与 [1] 的逆向结果之间的联系。 [42, 10, 41] 的研究在高斯混合模型和次高斯混合模型下，针对二元、多类别以及一对多支持向量机设置，推导了类似的结果。在我们的研究中，我们聚焦于一种针对二元标签的特定生成模型，并且在主要结果中并未明确对协变量分布施加次高斯尾部假设。与现有文献的第二个关键区别在于，由于我们对标签采用了概率生成模型，因此能够在更弱的随机矩阵条件下获得结果（进而只需对核积分算子的谱施加相应条件）。这些谱条件与最小范数插值器分析中采用的条件相似（例如， [5, 34, 39, 32,31]），并揭示了SVP与噪声的无害插值之间一种新颖的联系。

过参数化情形下分类器的泛化分析支持向量机泛化误差的经典界限依赖于样本压缩（例如， [17, 15]）或拉德马赫复杂度（例如，[4, 30]），这两种方法通常分别根据支持向量占训练样本的比例，或学习得到的超平面在训练集上实现的间隔，给出测试误差的上界。此外，研究表明，在一系列训练点数量与数据维度 [14, 9, 29], 成正比或远大于数据维度 [14, 9, 29], 的情形下，支持向量机中的支持向量比例为 $o(1)$，这意味着模型具有良好的泛化能力。然而，当数据维度或特征维度远大于训练点数量时，这些结果便不再适用。事实上，近期研究表明，在高度过参数化情形下，支持向量机决策函数能够插值拟合训练标签（即每个点都会成为支持向量），因此，传统界限会变得无效，也就是说，以高概率而言，这些界限并不优于一个常数 [33, 21]。

近期对线性和核回归模型中良性过拟合的分析，为研究过参数化分类器的行为和泛化提供了另一条途径。这些研究成果已被用于在带符号高斯模型 [33], 、高维线性判别分析 [11], 以及高斯和次高斯混合模型 [42, 41, 10] 中给出误差界。大多数这类研究成果（ [11], 除外，该研究还要求数据维度极高）都明确使用 SVP 来关联性能。

[2]我们注意到，关于渐近误差分析以及一般 ℓ_p-正则化器的研究非常广泛，因此这里将讨论范围限定在最相关的情形，即采用 ℓ_2-正则化的支持向量机的非渐近误差分析。

将分类器的任务转化为更易于分析的插值回归任务。因此，它们仅适用于已证明 SVP 成立的限制性情形。

相比之下，我们的分析与 [31], 的研究最为接近，他们在一般的再生核希尔伯特空间框架下，通过最小化平方损失，给出了标签插值函数的分类损失界限。 [31] 所提供的分析工具，又在一定程度上借鉴了 [20, 44] 的研究方法，后者考察了随机设计下的显式正则化核回归，并对特征仅作出最小限度的假设。值得注意的是，我们针对有界正交系统的结果，同样是以核积分算子上的谱条件来表述的，而非依赖于数据分布的特定性质。由此得到的误差表达式不如通过高斯假设或次高斯假设所获得的结果精确，但适用于更一般的核和特征映射选择。

2 背景与设置

我们的主要研究结果建立在 [31] 中关于核回归和分类的分析框架之上。我们在此概述该框架的主要组成部分。具体而言，我们将介绍与再生核希尔伯特空间的谱性质相关的重要概念，并建立监督学习框架，以便在该框架下研究支持向量机（SVM）分类问题与最小范数插值（MNI）分类问题之间的等价性。

2.1 符号

对于概率空间 (X, μ) 和 $f\colon X \to \mathrm{R}$，我们定义范数 $\|f\|_{L_p} := (\mathrm{E}_{x\sim\mu}|f(x)|^p)^{1/p}$ 和 $\|f\|_{L_\infty} := \operatorname{ess\,sup}_x |f(x)|$。对于 $u \in \mathrm{R}^n$，$\|u\|_{\ell_2}$ 或 $\|u\|_2$ 表示欧几里得范数。更一般地，对于希尔伯特空间 $\mathcal{H}$ 中的函数 f，$\|f\|_{\mathcal{H}}$ 表示希尔伯特范数。我们将平方可积函数 $f\colon X \to \mathrm{R}$ 构成的希尔伯特空间记为 $L_2(X, \mu)$，其内积为 $\langle f, g\rangle_{L_2} = \mathrm{E}_{x\sim\mu} f(x)g(x)$。我们分别用 $\langle\cdot,\cdot\rangle_{\ell_2}$ 和 $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ 表示 ℓ_2 和 $\mathcal{H}$ 内积。对于算子 $T\colon \mathcal{H}_1 \to \mathcal{H}_2$（其中 $\mathcal{H}_1$、$\mathcal{H}_2$ 是不同的希尔伯特空间），我们将其关于 $\mathcal{H}_1$ 和 $\mathcal{H}_2$ 范数的算子范数记为 $\|T\|_{\mathcal{H}_1\to\mathcal{H}_2} := \sup_{\|f\|_{\mathcal{H}_1}\le 1}\|Tf\|_{\mathcal{H}_2}$；当$T\colon \mathcal{H}_1 \to \mathcal{H}_1$时，则简记为 $\|T\|_{\mathcal{H}_1}$。对于希尔伯特空间 $\mathcal{H}$ 的子空间 A，我们将到 A 上的正交投影算子记为 $\mathcal{P}_A$；对于 $\mathcal{H}$ 上的算子 T，记 $T_A := T\mathcal{P}_A$。令 $\mathrm{diag}^{\perp}(Z)$ 表示将矩阵 Z 投影到对角线为零的矩阵集合上。恒等算子记为 $\mathcal{I}$（若为普通矩阵，则记为 I）。记号 e_i 表示 R^n 中的第 i^{th} 个标准基向量。最后，记号 $a \lesssim b$ 表示 $a \le Cb$，其中 $C > 0$ 是某个通用常数；$a \asymp b$ 表示 $a \lesssim b$ 和 $b \lesssim a$；而$a \ll b$表示 $a < cb$，其中 $c > 0$ 是某个（足够小的）常数。

2.2 再生核希尔伯特空间

设 $\mathcal{H}$ 是定义在概率空间 (X, μ) 上的实值函数再生核希尔伯特空间（RKHS）。我们将 $\mathcal{H}$ 的再生核记为 $k\colon X \times X \to \mathrm{R}$。根据定义，$k$ 满足如下性质：对于任意 $x \in X$ 和 $f \in \mathcal{H}$，$f(x) = \langle k_x, f\rangle_{\mathcal{H}}$，且 $k_x := k(x,\cdot)$ 是 $\mathcal{H}$ 中具有此性质的唯一元素。

我们的结果关键地依赖于学习过程中所用核的谱性质。为此，我们首先假设样本是按照概率测度 μ 从 X 中抽取的。于是，核积分算子 $\mathcal{T}\colon L_2(X, \mu) \to L_2(X, \mu)$ 定义为

$$\mathcal{T}f(\cdot) = \int_X k(\cdot, y)f(y)d\mu(y).$$

假设 k 连续，则 Mercer 定理给出

$$k(x, y) = \sum_{\ell=1}^{\infty} \lambda_\ell v_\ell(x)v_\ell(y),$$

其中，$\{\lambda_\ell\}_{\ell=1}^{\infty}$ 和 $\{v_\ell\}_{\ell=1}^{\infty}$ 分别是 $\mathcal{T}$的特征值和特征函数（关于这种分解成立所需的相对较弱条件，可参见 [38]）。此外，特征函数 $\{v_\ell\}_{\ell=1}^{\infty}$ 构成了 $L_2(X, \mu)$的一个正交归一基。不失一般性，我们假设

特征值已排序，使得 $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$。对于某个索引 p（我们可以在分析中选定该索引），令 $G := \mathrm{span}(v_1, \ldots, v_p)$为 $L_2(X, \mu)$中对应于最大 p 个特征值的子空间，并令 $\mathcal{T}_G$为核积分算子在 G上的限制。

2.3 分类设定

假设我们观测到一个训练集 $\{x_i, y_i\}_{i=1}^n$，其中包含 (x, y) 的独立同分布副本，且 $x \sim \mu$和 $y \in \{-1, 1\}$。我们假设 x 与 y 之间的关系由回归函数决定

$$\eta^*(x) := \mathbf{E}[y \mid x] = 2\mathbf{P}(y = 1 \mid x) - 1 \tag{1}$$

其中某个 $\eta^* \in G$ 满足条件。[3] 换言之，给定 x， y 以概率 $\frac{\eta^*(x)}{2} + 1$ 取值为 1，否则以概率 -1 取值。公式 (1) 构成了一种自然的分类生成模型：通过对分类条件概率（的仿射变换）进行回归来实现分类；插件分类器相关文献中也出现了类似模型（例如 [2]），不过对 η^*的正则性假设略有不同。因此，我们考虑能够给出 η^* 的估计量 $\hat{\eta}$ 的分类方法，并通过 $\hat{y} = \mathrm{sign}(\hat{\eta}(x))$ 来预测新点 $x \in X$的标签。将二元观测视为 $y_i = \eta^*(x_i) + \xi_i$ 的形式通常很方便，其中 $\xi_i := y_i - \eta^*(x_i)$ 被解释为观测噪声，并满足 $\mathrm{E}[\xi | x] = 0$。我们将观测y_i 和噪声 ξ_i 分别收集到向量 y 和 ξ中；根据上下文可以清楚判断这些符号指的是标量随机变量还是观测值向量。

对于上述形式的数据集，我们令 $\mathcal{A}$: $\mathcal{H} \to \mathrm{R}^n$ 表示 $\mathcal{H}$ 上的采样算子，即满足 $(\mathcal{A}f)_i = f(x_i)$ 的算子。采样算子关于 $\mathcal{H}$ 和 ℓ_2 内积的伴随算子 $\mathcal{A}^*$: $\mathrm{R}^n \to \mathcal{H}$ 的特征为：对于任意 $\beta \in \mathrm{R}^n$， $\mathcal{A}^*\beta = \sum_{i=1}^n \beta_i k_{x_i}$。最后，我们将与该数据集对应的 Gram 矩阵记为 $K := \mathcal{A}\mathcal{A}^* \in \mathrm{R}^{n \times n}$。该矩阵的元素由 $K_{ij} = k(x_i, x_j)$ 给出。

在我们的分析中，我们还将采样算子拆分为分别作用于 G 和 $G^\perp$ 的分量，记为 $\mathcal{A}_G := \mathcal{C}$ 和 $\mathcal{R} := \mathcal{A}_{G^\perp}$。我们将 $\mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$ 定义为 $\mathcal{A}_G$ 关于内积 $\langle \cdot, \cdot \rangle_{L_2}$的伴随算子。最后，估计量 $\hat{\eta}$ 的超额分类风险定义为

$$\mathcal{E}(\hat{\eta}) := \mathbf{P}(\hat{y} \neq y) - \mathbf{P}(y \neq \mathrm{sign}(\eta^*)).$$

2.4 无岭核回归

无岭核 回归，或称最小范数插值（MNI），过程求解以下问题 :

$$\begin{aligned} \hat{\eta} &= \arg\min_{\eta \in \mathcal{H}} \|\eta\|_{\mathcal{H}} \\ &\text{s.t. } \eta(x_i) = y_i \text{ for all } i = 1, \ldots, n. \end{aligned} \tag{2}$$

该问题存在闭式解 $\hat{\eta}(x) = \sum_{i=1}^n \hat{\beta}_i k(x, x_i)$，其中 $\hat{\beta} = K^{-1}y$。近来已有多项研究成果广泛研究了此类估计量的性质（例如， [28, 5, 19, 32, 31]）。总体而言，这些研究成果大多将学习到的估计分解为

$$\hat{\eta} = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}y = \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*}_{=:\hat{\eta}_0} + \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi}_{=:\epsilon} \tag{3}$$

然后，在各种假设下，我们刻画以下情形何时成立：(i) 第一项近似于 η^* ；(ii) 第二项可以忽略。对我们的分析而言最为核心的是， [31] 证明了，在核满足某些谱条件时，该估计满足 $\hat{\eta} \approx \overline{\mathcal{S}}\eta^*$，其中 $\overline{\mathcal{S}} = n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ 是理想化生存算子，[4]它表示通过 MNI 恢复真实函数的效果。我们将在下文看到，SVP 可以出现在各种各样的情形中，此时 $\overline{\mathcal{S}}$ 会在 L_∞ 范数下均匀衰减 η^* 。

[3]假设 η^* 属于子空间 G 是为了方便起见；若放宽这一假设，我们的误差表达式中会多出一项。[4]我们将在假设 1 中正式定义 $\bar{\alpha}$ 。

2.5 核支持向量机

硬间隔核支持向量机（支持向量机）过程旨在寻找一个线性分类器，使其在空间 $\mathcal{H}$ 中最大化训练数据间隔。具体而言，该问题的原始形式和对偶形式可以写成

$$\begin{array}{c}\hat{\eta}=\underset{\eta\in\mathcal{H}}{\text{求最小值}}\|\eta\|_{\mathcal{H}}\\ \text{约束条件：}\ y_i\eta(x_i)\geq 1\ \text{对于}\ i=1,\dots,n.\end{array}\quad\Longleftrightarrow\quad\begin{array}{c}\hat{\beta}=\underset{\beta\in\mathbb{R}^n}{\text{求最大值}}\ y^\top\beta-\frac{1}{2}\beta^TK\beta\\ \text{约束条件：}\ y_i\beta_i\geq 0\ \text{对于}\ i=1,\dots,n,\end{array}\tag{4}$$

其中 $\hat{\eta}=\sum_{i=1}^n\hat{\beta}_ik_{x_i}$。观察对偶形式，我们注意到，由 MNI 和支持向量机学习过程得到的估计量 $\hat{\eta}$ 都可以表示为输入特征 k_{x_i} 的线性组合（系数由 $\hat{}$向量 β 给出），其中 $i=1,\dots,n$。正如在关于 [33,21], 的讨论中最初指出的那样，很容易验证，选择 $\hat{}\ \beta=K^{-1}y$ （其刻画了（2）的解）也会使（4）的对偶目标函数达到最大值；然而，这一选择可能不满足支持向量机的所有附加约束。下一节的主要结果旨在刻画：通过求解这两个问题得到的估计量 $\hat{\eta}$ 何时会完全一致。由于支持向量是满足 $\hat{\eta}(x_i)=\pm1$ 的训练点（因此原始问题中的约束在这些点上是有效的），这种等价性意味着，在学习得到的支持向量机中，每个训练点都是支持向量。

3 主要结果

本节给出我们关于上述再生核希尔伯特空间设定下 SVP 的主要结果。首先，我们介绍主要假设并定义结果中涉及的关键量；随后，在第 3.2 节中，我们针对一般的有界特征族陈述一个定理，并给出一个呈现 SVP 的双层特征设计示例。接着，在第 3.3 节中，我们运用分析框架，针对独立次高斯特征的情形进一步强化结果，从而表明，在我们的生成式标签模型下，SVP 能够出现在一个严格大于 [21] 主要结果所涵盖的范围内。最后，我们证明，我们的结果立即意味着核支持向量机在新的过参数化情形下具有良好的泛化能力（即具备统计一致性）。

3.1 特征模型和采样条件

我们考虑满足以下两种性质之一的核积分算子 $v_\ell(x)$ 的特征函数：

1. G 上的有界正交系统（BOS）： $\mathcal{T}_G$ 的特征函数满足 $\sum_{\ell=1}^p v_\ell^2(x)\leq Cp$，且该条件对 x 以概率 1 成立，其中 $C\geq1$ 为某个常数。
2. 独立次高斯（SG）：对于 $x\sim\mu$， $\{v_\ell(x)\}_{\ell\geq1}$ 是相互独立、零均值的次高斯随机变量。

G 上的 BOS 条件可由若干常见的特征族满足，例如傅里叶特征、某些正交归一多项式族、离散正交归一系统等。值得注意的是，在其中许多情形下，特征之间远非独立。

我们的分析考虑这样一种设定：训练集中的采样位置从分布 μ 中独立抽取，并且以高概率满足某些性质（这些性质可通过附录 B 中给出的集中界另行验证）。前两个性质等价于 [31] 的定理 1 中的性质。最后一个条件是我们分析中的一项性质，对于多种特征和 η^* 的选择都成立。下面列出我们的主要假设，随后简要解释并讨论该假设在何种情况下成立。

假设 1. 对于某个 $0<\delta<1/2$，训练样本 $x_1,\dots,x_n\overset{i.i.d.}{\sim}\mu$ 以至少 $1-\delta$ 的概率满足以下所有条件：

1.（残差 Gram 矩阵集中）存在常数 $\alpha_L \asymp \alpha_U > 0$，使得残差 Gram 矩阵满足

$$\alpha_L I_n \preccurlyeq \mathcal{R}\mathcal{R}^* \preccurlyeq \alpha_U I_n.$$

2.（关于 G 的特征协方差集中）设 $\mathcal{C}_{\setminus i}$ 是针对留一数据集 $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ 在 G 上的采样算子，其 L_2-伴随算子为 $\mathcal{C}^*_{\setminus i}$。那么，对于所有 $i \in \{1, \ldots, n\}$，以下结论对某个 $c \le \frac{1}{2}$ 成立：

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*_{\setminus i}\mathcal{C}_{\setminus i} - n\mathcal{I}_G\|_{L_2} \le c.$$

3.（L_∞ 对期望恢复信号的控制）设 $\bar{\alpha}$ 是 α_L 与 α_U 的调和平均数。设 $\overline{\mathcal{S}} := n\mathcal{T}_G(\bar{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ 为理想化生存算子，$\overline{\mathcal{B}} := \mathcal{I}_G - \overline{\mathcal{S}}$ 为理想化偏差算子。那么

$$b := \|\overline{\mathcal{B}}\eta^*\|_\infty \lesssim 1 - \|\overline{\mathcal{S}}\eta^*\|_\infty.$$

残差 Gram 矩阵集中的第一个条件是确保 MNI 中噪声的无害插值所必需的，并且已证明在许多情况下该条件都以高概率成立，例如，对于次高斯特征 [5, 39]，以及主特征子空间满足超收缩性条件 [32] 的情形。我们在附录 B 中给出了相关集中界，可用于验证 BOS 和 SG 两种情形下的这一条件。请注意，与 SVP 的先前结果不同，后者与标签无关，并且关键依赖于整个 Gram 矩阵（包括 [33, 21]）集中到单位矩阵的某个倍数；我们只要求残差分量满足这一条件。

第二个条件要求在 G 上满足协方差集中。该假设的全样本版本在回归中是标准的。对于 BOS 特征，当$n \gtrsim p \log n$ 时（例如，参见引理 7），以及对于 SG 特征，当 $n \gtrsim p$ 时（引理 8），我们的留一法版本以高概率成立。

第三个条件要求理想化生存算子 $\overline{\mathcal{S}}$（近似地）一致收缩 η^*。例如，在以下任一情况下，该条件都成立：

- 主特征值相同：$\mathcal{T}$的主特征值均等于某个 $\lambda > 0$。于是，$\overline{\mathcal{S}} = \frac{n\lambda}{\bar{\alpha}+n\lambda}\mathcal{I}_G$ 和 $\overline{\mathcal{B}} = \frac{\bar{\alpha}}{\bar{\alpha}+n\lambda}\mathcal{I}_G$，而且 $\|\eta^*\|_\infty \le 1$ 这一事实意味着该假设成立（其中.中的常数恰好为 1）。我们将在第3.2.1节探讨这种情形。

- 1-稀疏目标函数：如果 η^* 位于单个主特征向量的张成空间中，即对于某个 $1 \le \ell \le p$ 和常数 γ，$\eta^* = \gamma v_\ell$，则即使主特征值不相同，我们也能得到相同的结果。
- 极端收缩：一般来说，如果 $\bar{\alpha} \gg n\lambda_1$，那么 $\overline{\mathcal{S}}\eta^*$ 将收缩至 0，而 $\overline{\mathcal{B}}\eta^*$ 趋近于η^*（同时，有界的 v_ℓ 与充分的收缩相结合，可得到所需的 L_∞ 控制）。请注意，这与 [21]中的 SVP 条件 $\sum_{i=1}^n \lambda_i \gg n\lambda_1$ 相似。

3.2 支持向量机等价的一般条件

我们的主要结果在假设 1 下给出了充分条件，使每个训练点都是支持向量，从而使 MNI 和支持向量机得到相同的解。该结果的证明见第 4 节。

定理 1（有界正交系统中的 SVP）。考虑第 2 节中所述且采用 BOS 特征的分类设定。假设对于 $\eta^* \in G$，假设 1 得到满足。那么，以至少 $1 - \delta - \frac{2}{n}$ 的概率，MNI 程序 (2) 和 SVM 程序 (4) 的解相同，如果

$$\frac{1}{b}\sqrt{\frac{p\log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\left(\frac{\sqrt{\log n}}{b} + \sqrt{n}\right) + c^2\sqrt{p} \ll 1. \tag{5}$$

我们的结果以一种微妙的方式取决于理想化偏差项 b、样本数量 n、 G的维度 p 、残差 Gram 矩阵的集中性 $\mathcal{R}\mathcal{R}^*$，以及（通过常数 c）对 G的样本协方差集中性。该结果表明，在核设定中，即使不存在次高斯或独立特征，也可以更普遍地出现 SVP（包括 [33] 中的傅里叶特征设定，其中

这一现象最初被观察到）。请注意，与 [33, 21], 的结果不同，我们的分析假设标签遵循特定模型，不允许进行对抗性标签选择。不过，正如我们将在下一节中讨论的那样，我们的结果适用于对特征所作的假设弱得多的情形。

3.2.1 双层集成的启示

作为一个具体示例，我们现在介绍 [33, 31] 中定义的过参数化双层集成。此示例有助于理解 SVP 可能出现的各种情形，其参数由元组 (n, β, q, r) 表示，其中 $\beta > 1$、$0 < r < 1$ 和 $q \leq \beta - r$。这里，$d = n^\beta$ 是用于训练的特征数量，前 $p = n^r$ 个特征对应较大的特征值，并且

$$\lambda_\ell = \begin{cases} 1, & \text{for } 1 \leq \ell \leq p \\ n^{-\beta+r+q}, & \text{for } p < \ell \leq d. \end{cases} \tag{6}$$

请注意，$q \leq \beta - r$（等号成立时表示各向同性情形）是正确排列特征值所必需的，而 $r < 1$ 确保 $p < n$。这里，q 可理解为控制“有效过参数化”，即领先特征值与残余特征值之间的差距。这种特征化方法受到尖峰协方差模型 [33] 的启发，并具有一项实用性质：在其允许的参数范围内，它允许在线性回归中对噪声进行无害插值。鉴于我们的分析与 MNI 的无害插值分析密切相关，这种集成是探索 SVP 现象的自然情形。下面，我们将定理 1 的结果具体化到双层集成下特征有界的情形。证明见附录 C。

推论 1（BOS 中的 SVP：双层集成）。考虑参数为 (n, β, q, r)、具有有界特征 $\{v_\ell(x)\}_{\ell=1}^d$ 的双层集成。假设 $\eta^* \in G$ 满足第 2 节中的分类设定，且 $\sup_x \left|\sum_{\ell>p}(|v_\ell(x)|^2 - 1)\right| \lesssim n^{\frac{\beta}{2}+1}$。那么，当 $n \to \infty$ 时，若满足 $\beta > 3$、$q > \frac{1-r}{2}$ 和 $r < \frac{2}{3}$，SVP 发生的概率为 $\to 1$。

该推论表明，只要有效过参数化程度（由 β 和 q 控制）足够高，这种设定下就会发生 SVP。值得注意的是，这允许特征具有高度结构化的情形。事实上，我们只假设 v_ℓ 有界并满足薄壳型条件 $\sup_x \left|\sum_{\ell>p}(|v_\ell(x)|^2 - 1)\right| \ll n^{\frac{\beta}{2}+1}$，这确保了 $k(x,x)$ 的变化不会太大，也是残差 Gram 矩阵集中性的必要条件（相关研究在考察这种集中现象时也提出了类似要求，例如 [32], [3]）。对于许多结构化特征族，包括傅里叶特征，都可以很容易地验证这一条件成立；对于后者，$|v_\ell(x)|^2 = 1$ 对所有 x 都一致成立。我们进一步指出，尽管为清晰起见，我们在双层模型下陈述渐近结果，但定理 1（以及附录 C 中的证明）使我们能够在具有非均匀特征值的广泛设定下，获得以高概率成立的有限样本保证。

3.3 独立次高斯情形下的改进

现在我们考虑特征 $\{v_\ell(x)\}_\ell$ 是相互独立的次高斯随机变量的情形。这是早期关于支持向量增殖 [21, 33] 的研究成果所考虑的设定。这些研究成果通常不假设特定的标签模型，而是用整个特征协方差的某些“有效秩”概念来表述条件。为了与这类研究成果进行比较，我们也给出了独立次高斯情形下的结果。我们的结果适用于一种更具体的标签模型；与先前工作相比，该模型隐含地施加了额外的分布假设，因为在我们的模型中，$\eta^*(x)$（它是 $\{v_i(x)\}_{i=1}^p$ 的线性组合）必须一致有界。不过，我们的分析框架既能复现许多现有结果，也表明 SVP 可以在更广泛的情形下发生。该定理的证明见第 4.5 节。

定理 2（独立次高斯特征的 SVP）。考虑第 2 节所述的分类设定。假设对于 $\eta^* \in G$，假设 1 得到满足，并且特征 $\{v_\ell(x)\}_{\ell>1}$ 是零均值、独立且次高斯的。那么，以至少 $1 - \delta - ne^{-p} - \frac{3}{n}$ 的概率，当且仅当满足以下条件时，MNI 程序 (2) 和 SVM 程序 (4) 的解一致：

$$\frac{1}{b}\sqrt{\frac{p\log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\frac{\sqrt{\log n}}{b} + \sqrt{\frac{\lambda_{p+1} n \log n}{\alpha_L}} + c\sqrt{\log n} \ll 1. \tag{7}$$

我们可以看到，最后两项比 BOS 情形中的相应项锐利得多（例如，是 $c^2\sqrt{p}$ 的 $c\sqrt{\log n}$，而不是 $c^2\sqrt{p}$）。这一改进后的界，结合算子 $\mathcal{R}\mathcal{R}^*$ 和 $\mathcal{C}^*\mathcal{C}$ 针对次高斯特征更快的集中速率，可以证明会弱化双层集成下 SVP 的条件。我们在下面的推论中陈述这一条件，该推论已在附录 C 中证明。

推论 2（独立次高斯情形下的 SVP：双层集成）。考虑参数为 (n, β, q, r) 且具有独立、零均值、次高斯特征 $\{v_\ell(x)\}_{\ell=1}^d$ 的双层集成。假设 $\eta^* \in G$ 满足第 2 节中的分类设定。那么，当 $n \to \infty$ 时，如果 $\beta > \max(1, 3-2r-2q)$ 且 $q > \frac{1-r}{2}$，则 SVP 以概率 $\to 1$ 发生。

该推论与 [21], 的结果一致，该结果表明，对于$q > 1-r$ 和 $\beta > 1$，在双层模型中，对于任意标签 y_i，会出现 SVP。我们在标签模型下恢复了这一结果，同时还发现，只要 β 足够大（即只要存在足够的过参数化），SVP 就可以在范围 $\frac{1-r}{2} < q < 1-r$ 内发生。不同于以往已证明 SVP 会发生的情形，这是一种 MNI 程序最终表现出回归一致性的情形，也就是说，它可以在 $n \to \infty$ [33][5]的极限下准确恢复 η^*。

3.4 核支持向量机在新情形下的泛化

在发生支持向量增殖的情形下，MNI 与核支持向量机之间的等价性为刻画支持向量机分类器的泛化误差提供了一种途径。具体而言，我们可以利用具有闭式解且近期文献已对其进行了广泛分析的 MNI，来理解核支持向量机的泛化。事实上，正如 [21], 所指出的，传统的支持向量机泛化界通常基于支持向量所占比例，因此在发生 SVP 的过参数化情形中会变得毫无信息量（此时该比例趋近于 1）。尽管如此，利用与 MNI 的等价性可以证明，即使在每个训练点都是支持向量的高度过参数化情形下，核支持向量机确实仍能实现良好的泛化（就渐近一致性而言）。此前，这类结果已针对高斯数据和独立次高斯数据上的线性模型得到证明 [33,21, 42, 41, 10]。现在，我们将证明，在双层集成下，一般核方法也能得到这类结果，从而部分回答了 [33] 提出的一个未解问题。回顾一下，在这种集成下，可以利用经典间隔界证明核支持向量机对于 $q < 1-r$ 是分类一致的。通过运用 [31], 中关于 MNI 的泛化结果，我们将在下文证明，在某些 $q > 1-r$ 的情形下，即过参数化程度很高时，SVP 也可用于证明核支持向量机具有一致性。

推论 3（双层集成中核支持向量机的泛化）。考虑参数为 (n, β, q, r) 的双层集成。假设 $\eta^* \in G$ 满足第 2 节中的分类设定。那么，

1. 若 $\{v_\ell(x)\}_{\ell=1}^d$ 满足推论 1 中的 BOS 条件和薄壳条件，则当 $n \to \infty$ 时，核支持向量机的超额分类风险以概率 $\to 1$ 趋于 0，条件是

$$\beta > 3, r < \frac{2}{3}, \ and \ 0 \le q < \frac{3}{2}(1-r).$$

2. 若 $\{v_\ell(x)\}_{\ell=1}^d$ 是零均值、独立且次高斯的，则当 $n \to \infty$ 时，核支持向量机的超额分类风险以概率 $\to 1$ 趋于 0，条件是

$$0 \le q < \min\left(1 - r + \frac{\beta - 1}{2}, \frac{3}{2}(1-r)\right).$$

我们注意到，推论 3 的第二部分与研究成果 [33, 21],所隐含的条件非常相似，但对 q的条件要求略强。不过，这些先前结果假设真实函数

[5]我们在此指出，[1] 证明了在 $d_2 := \frac{(\sum_{\ell>1}\lambda_j)^2}{\sum_{\ell>1}\lambda_j^2} \lesssim n\log n$ 且 $d_\infty^2 := \frac{(\sum_{j\ge 1}\lambda_j)}{\lambda_1^2}^2 \gtrsim d_2 n$ 的情形下，SVP 很可能发生（对于次高斯数据）。然而，这种情形无法由双层集成的任何参数化涵盖，因为我们始终有 $d_\infty^2 \ll d_2 n$。我们的结果适用于另一种情形，在该情形下我们允许 $d_\infty \le n$，并且仍能证明 SVP 可以成立。

是 1-稀疏且线性的；相比之下，我们考虑一种灵活分类设定，其中目标函数 η^*被假设属于 $\mathcal{H}$的子空间 G，且 y_i 遵循第 2 节所述的标签模型。此外，推论 3 的第一部分适用于广泛的非独立特征族，这些特征族的分布可能具有更重的尾部。

4 主要定理的证明

在本节中，我们将证明主要结果。首先介绍整体证明方法，然后分别考察分析中的主要组成部分，并将一些有用技术引理的证明推迟至附录 B。

4.1 分析概述

我们的结果基于 [21] 中给出的支持向量增殖的确定性条件。为确保内容自洽，我们首先使用第 2 节中引入的符号，重新陈述这一针对核设定的结果：

命题 1（ [21] 中的引理 1）。对于对应于再生核希尔伯特空间 $\mathcal{H}$ 的核 $k(x,y)$，最小范数标签插值函数 $\hat{}\eta$ 由以下公式给出。

$$\hat{\eta}(x) = \sum_{i=1}^{n} z_i k(x, x_i),$$

其中， $z = K^{-1}y$ 和 $K = [k(x_i, x_j)]_{ij}$ 是核 Gram 矩阵，假设其非奇异。此外，设 $\hat{}\eta_{\setminus i}$ 为使用由点 $\{(x_j, y_j)\}_{j=1, j\neq i}^{n}$ 组成的留一数据集构造的最小范数标签插值函数。

那么，以下条件等价：

1. 每个数据点（x_i, y_i）都是最大间隔支持向量机的支持向量，即支持向量机决策边界与二元标签插值函数完全相同。
2. 对于每个 $i \in \{1, \dots, n\}$, $z_i y_i > 0$，即 $z_i \neq 0$ 和 $\operatorname{sign}(z_i) = y_i$。
3. 对于每个 $i \in \{1, \dots, n\}$， $y_i \hat{}\eta_{\setminus i}(x_i) < 1$。

我们的分析旨在给出使得 $|\hat{}\eta_{\setminus i}(x_i)| < 1$ 对所有 i 都成立的条件。当这一情况发生时，命题 1 立即意味着所有训练点都是支持向量。回顾分解 (3)，我们可以写出

$$|\hat{\eta}_{\setminus i}(x_i)| \le |\epsilon_{\setminus i}(x_i)| + |\mathcal{P}_{G^\perp}\hat{\eta}_{\setminus i,0}(x_i)| + |\mathcal{P}_G\hat{\eta}_{\setminus i,0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)| + |\overline{\mathcal{S}}\eta^*(x_i)|.$$

根据假设 1， $1 - |\overline{\mathcal{S}}\eta^*(x_i)| \gtrsim b$，因此当.....时，上述界将严格小于 1。

$$\underbrace{\frac{|\epsilon_{\setminus i}(x_i)|}{b}}_{A} + \underbrace{\frac{|\mathcal{P}_{G^\perp}\hat{\eta}_{\setminus i,0}(x_i)|}{b}}_{B} + \underbrace{\frac{|\mathcal{P}_G\hat{\eta}_{\setminus i,0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)|}{b}}_{C} \lesssim 1 \tag{8}$$

对所有 i。项 A 表示标签噪声的插值。项 B 和 C 表示无噪声估计 $\hat{}\eta_{\setminus i,\,0}$ 在 x_i 处由理想化生存函数 $\overline{\mathcal{S}}\eta^*$ 近似的程度： B 表示与 G正交的分量，而 C 是 G中的分量。在接下来的三个小节中，我们将重点讨论 BOS 特征情形，分别对这三项进行界定，从而得出满足条件 (8) 的一般条件（定理 1）。随后，在第 4.5 节中，我们考虑独立次高斯特征的情形，并对项 B 和 C 得出更精确的界（从而得到定理 2）。

4.2 噪声误差

为界定 (8) 中的项 A，首先注意到，对于所有 $i \le n$，噪声变量 ξ 均有界，且 $|\xi_i| \le 2$。因此，向量 ξ 具有如下意义上的次高斯性质：对于任意固定向量 $z \in \mathbf{R}^n$，$\langle z, \xi \rangle$ 是一个次高斯随机变量，其次高斯范数在常数 $\|z\|_2$ 以内。这意味着，对于任意固定的 $z \in \mathbf{R}^n$，相对于 ξ，以至少 $1-\delta'$ 的概率，$|\langle z, \xi \rangle| \lesssim \|z\|_2 \sqrt{\log\, \delta'^{-1}}$。因此，由于

$$|\epsilon(x)| = |\langle k_x, \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi \rangle_{\mathcal{H}}| = |\langle (\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x, \xi \rangle|,$$

it s只需对 $\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2$ 进行界定。我们将在以下引理中利用这一观察。 .

引理 1（噪声项的上界）。设 $x \in X$ 与训练数据独立，并假设假设 1 成立。那么，以至少 $1 - \frac{1}{n^2}$ 的概率，

$$|\epsilon(x)| \lesssim \sqrt{\frac{p \log n}{n}} + \frac{1}{\tilde{\alpha}} \|\mathcal{R} k_x^R\|_2 \sqrt{\log n},$$

其中 $\tilde{\alpha} = \frac{\alpha_L + \alpha_U}{2}$ 和 $k_x^R = \mathcal{P}_{G^\perp} k_x$。

引理 1 的证明。设 $x \in X$ 与数据独立。我们有

$$\epsilon(x) = \langle k_x, \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi \rangle_{\mathcal{H}} = \langle (\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x, \xi \rangle.$$

关于 ξ 中的随机性，这是一个次高斯随机变量，其次高斯范数与 $\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2$ 相差一个常数因子。根据 [31], 的引理 8，我们有

$$\begin{aligned}
\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2 &= \|(\mathcal{A}\mathcal{A}^*)^{-1}(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}k_x\|_2 \\
&\le \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right)\|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}k_x\|_2 \\
&\lesssim \|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G k_x^G\|_2 + \|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}k_x^R\|_2,
\end{aligned}$$

其中，$\tilde{\alpha} = \frac{\alpha_L + \alpha_U}{2}$、$k_x^G$ 和 k_x^R 分别表示 $\mathcal{P}_G k_x$ 和 $\mathcal{P}_{G^\perp} k_x$。对于第一项，我们有

$$\begin{aligned}
\|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G k_x^G\|_2 &= \|\mathcal{C}(\tilde{\alpha}\mathcal{I}_G + \mathcal{T}_G\mathcal{C}^*\mathcal{C})^{-1}k_x^G\|_2 \\
&= \|\mathcal{C}(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\mathcal{T}_G^{-1}k_x^G\|_2 \\
&\le \|\mathcal{C}\|_{L_2 \to \ell_2}\|(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\|_{L_2 \to L_2}\|\mathcal{T}_G^{-1}k_x^G\|_{L_2} \\
&\lesssim \sqrt{n}\left(\frac{1}{n}\right)\sqrt{p} = \sqrt{\frac{p}{n}}.
\end{aligned}$$

这里，最后一个不等式由关于 $\mathcal{C}^*\mathcal{C}$ $\|\mathcal{T}_G^{-1}k_x^G\|_{L_2}^2 =$ 集中性的假设，$\|\sum_{\ell=1}^p v_\ell\ x\ v_\ell\|_{L_2}^2 = \sum_{\ell=1}^p v_\ell^2\ x\ \lesssim p$ 以及根据帕塞瓦尔恒等式和 BOS 假设可知 ( ) ( ) 成立。

对于第二项，我们有

$$\|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}k_x^R\|_2 \le \|(\tilde{\alpha} I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|\|\mathcal{R}k_x^R\|_2 \le \frac{1}{\tilde{\alpha}}\|\mathcal{R}k_x^R\|_2$$

应用次高斯尾界即可得到该结果。

我们还指出，对于独立的次高斯特征，上述论证同样适用，只是我们利用了以下事实：

$$\|\mathcal{T}_G^{-1}k_x^G\|_{L_2} = \left(\sum_{\ell=1}^p v_\ell^2(x)\right)^{1/2} \lesssim \sqrt{p}$$

概率至少为 $1 - e^{-p}$（例如参见 [40, 定理 3.1.1]）（注意，在 BOS 情形下，该不等式确定性成立）。因此，最终结果成立的概率至少为 $1 - e^{-p} - \frac{1}{n^2}$，而非 $1 - \frac{1}{n^2}$。我们将在第 4.5 节中用到这一事实。

□

现在，我们将此引理应用于索引 i 对应的留一训练集，并令 $x = x_i$。在这种情况下， $\|\mathcal{R}_{\setminus i} k^R_{x_i}\|_2 = \|\operatorname{diag}^{\perp}(\mathcal{R}\mathcal{R}^*)e_i\|_2 \le \|(\mathcal{R}\mathcal{R}^* - \alpha_L I_n)e_i\|_2 \le \alpha_U - \alpha_L$。回想 $\tilde{\alpha}$ 的定义，可得

$$|\epsilon_{\setminus i}(x_i)| \lesssim \sqrt{\frac{p\log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{\log n}$$

以至少 $1 - \frac{1}{n^2}$ 的概率成立。随后，我们可以对 n 个留一法集合取并集界，并除以 b，从而得到关于项 A 的最终界：以至少 $1 - \frac{1}{n}$ 的概率，对所有 $i = 1$、$\cdots$、n，

$$A \lesssim \frac{1}{b}\cdot\sqrt{\frac{p\log n}{n}} + \frac{1}{b}\cdot\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{\log n}. \tag{9}$$

4.3 由 $G^{\perp}$ 导致的偏差误差

对于（8）中的项 B，

$$|\mathcal{P}_{G^{\perp}}\hat{\eta}_{\setminus i,0}(x_i)| = \left|\sum_{j=1, j\neq i}^{n}(K^{-1}_{\setminus i}\mathcal{A}_{\setminus i}\eta^*)_j k^R(x_j, x_i)\right| \le \left\|K^{-1}_{\setminus i}\mathcal{A}_{\setminus i}\eta^*\right\|_2 \left\|\operatorname{diag}^{\perp}(\mathcal{R}\mathcal{R}^*)e_i\right\|_2$$

根据柯西-施瓦茨不等式。下面的引理给出了这些项中第一项的界。

引理 2（关于 $G^{\perp}$ 项的有用界）。在假设 1 以及第 2 节所述的分类模型下， $\|K^{-1}\mathcal{A}\eta^*\|_2 \lesssim \frac{b\sqrt{n}}{\alpha_L}$.

引理 2 的证明。回顾分解 $K = \mathcal{A}_G\mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*$。然后，根据推移恒等式，

$$\begin{aligned}
\|(\mathcal{R}\mathcal{R}^* + \mathcal{A}_G\mathcal{A}_G)^{-1}\mathcal{A}\eta^*\|_2 &= \|(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}\eta^*\|_2 \\
&= \|(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}\eta^*\|_2 \\
&\le \|(\mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\ell_2}\|\mathcal{A}_G\|_{L_2\to\ell_2}\|\mathcal{B}\eta^*\|_{L_2} \\
&\lesssim \frac{1}{\alpha_L}\sqrt{n}b\left(1 + \frac{c}{1-c}\right) \lesssim \frac{b\sqrt{n}}{\alpha_L}
\end{aligned}$$

在最后一行，我们使用了假设 1 的第二部分得出的事实 $\|\mathcal{A}_G\|_{L_2\to\ell_2} = \sqrt{\|\mathcal{C}^*\mathcal{C}\|_{L_2}} \asymp \sqrt{n}$。此外，我们将 $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ 记作，并利用引理 9 将这个“偏差算子”上界为假设 1 中定义的“理想化偏差算子”： $\|\mathcal{B}\eta^*\|_{L_2} \le (1 + \frac{c}{1-c})\|\overline{\mathcal{B}}\eta^*\|_{L_2} \le (1 + \frac{c}{1-c})b$。 □

利用此引理，并注意到（如上一节所示） $\left\|\operatorname{diag}^{\perp}(\mathcal{R}\mathcal{R}^*)e_i\right\|_2 \le \alpha_U - \alpha_L$，再除以 b，我们得出项 B 在相差一个常数因子的意义下被 $\left(\frac{\alpha_U}{\alpha_L} - 1\right)\sqrt{n} \asymp \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{n}$ 所界。

4.4 G 中的偏差误差

对于项 C，我们首先注意到， $|\mathcal{P}_G\hat{\eta}_{\setminus i,\ 0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)|$ 对应于 G-投影估计量与其“理想化”版本（由 $\overline{\mathcal{S}}\eta^*$ 给出）的预测值之间的绝对误差。在引理 3 中，我们证明了 $|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)|$ 的高概率界，其中 x 独立于所有训练点。

引理 3（ G-投影估计量的误差）。设 $x \in X$ 与训练数据独立，并假设假设 1 成立。那么，以至少 $1 - \frac{1}{n^2}$ 的概率，

$$\frac{1}{b}|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| \lesssim \sqrt{\frac{p\log n}{n}} + \sqrt{p}\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + c^2\sqrt{p}.$$

引理 3 的证明。首先注意，我们可以写成 $|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| = |((\mathcal{B} - \overline{\mathcal{B}})\eta^*)(x)|$，其中 $\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1} =: (\mathcal{I}_G + \mathcal{M})^{-1}$，且 $\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1} =: (\mathcal{I}_G + \overline{\mathcal{M}})^{-1}$.。接着注意到

$$\mathcal{B} - \overline{\mathcal{B}} = \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\mathcal{B} = \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}} + \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}}).$$

因此，我们得到

$$\begin{aligned}|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| &\le |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}}\eta^*)(x)| + |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*)(x)| \\ &\le |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}}\eta^*)(x)| + \|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\|_{L_\infty}\end{aligned}$$

然后我们使用分解

$$\overline{\mathcal{M}} - \mathcal{M} = \frac{n}{\bar{\alpha}}\mathcal{T}_G - \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G = \frac{n}{\bar{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right) + \mathcal{A}_G^*\left(\frac{1}{\bar{\alpha}}I_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G,$$

将其应用于第一项可得

$$\begin{aligned}|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| &\le \left|\left(\overline{\mathcal{B}}\frac{n}{\bar{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}}\eta^*\right)(x)\right| \\ &\quad + \left|\left(\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\bar{\alpha}}I_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right)(x)\right| + \left\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\right\|_{L_\infty} \\ &\le \left|\left(\overline{\mathcal{B}}\frac{n}{\bar{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}}\eta^*\right)(x)\right| \\ &\quad + \left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\bar{\alpha}}I_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right\|_{L_\infty} + \left\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\right\|_{L_\infty}\end{aligned} \tag{10}$$

我们通过分别对式（10）中的三个项进行界定，得到所需的界。对于第一项，注意到

$$\begin{aligned}\overline{\mathcal{B}} \cdot \frac{n}{\bar{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}} &= \frac{n}{\bar{\alpha}}\left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}} \\ &= \left(\frac{\bar{\alpha}}{n}\mathcal{T}_G^{-1} + \mathcal{I}_G\right)^{-1}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)\overline{\mathcal{B}}.\end{aligned}$$

我们不再对 L_∞ 范数进行界定，而是直接分析该算子对 η^* 的作用。令 $\mathcal{D} := \left(\frac{\bar{\alpha}}{n}\mathcal{T}_G^{-1} + \mathcal{I}_G\right)^{-1}$；注意， $\mathcal{D}$ 是 G 上的压缩映射（即，对 $f \in G$，有 $\|Df\|_{L_2} \le \|f\|_{L_2}$）。此外，令 $\delta_{x_i}^G = \mathcal{T}_G^{-1}k_{x_i}^G$。于是，对于固定的 $f \in G$ 和 $x \in X$，我们有

$$\left|\left(\mathcal{D}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)f\right)(x)\right| = \left|\frac{1}{n}\sum_{i=1}^n (f(x_i)\mathcal{D}\delta_{x_i}^G(x) - (\mathcal{D}f)(x))\right|,$$

它是独立零均值随机变量之和。具体而言，每个变量 $f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)$ 都是零均值的，且方差为

$$\mathbf{E}_{x_i}|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)|^2 \le \mathbf{E}_{x_i}|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x)|^2 \le \|f\|_\infty^2\|\mathcal{D}\delta_x^G\|_{L_2}^2 \lesssim p\|f\|_\infty^2,$$

其中，最后一个不等式用到了 D 是压缩映射这一事实以及 BOS 假设。此外，每一项都有如下界： $|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)| \lesssim \|f\|_\infty p$.因此，根据伯恩斯坦不等式（例如， [40,第 2 节。8]），以至少 $1 - \frac{1}{n^2}$ 的概率可得:

$$\left|\left(\mathcal{D}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)f\right)(x)\right| \lesssim \sqrt{\frac{p\log n}{n}}\|f\|_\infty + \frac{p\log n}{n}\|f\|_\infty \lesssim \sqrt{\frac{p\log n}{n}}\|f\|_\infty.$$

我们利用 $f = \overline{\mathcal{B}}\eta^*$ 并回忆 $\|\overline{\mathcal{B}}\eta^*\|_\infty = b$，对第一项进行界定。 .

对于　对于式 (10) 的第二项，[31] 的引理 7（我们将其表述为引理 9）证明了

$$\left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\bar{\alpha}}I_n-(\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\right\|_{L_2}\leq\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L}\left(1+\frac{1}{n}\|\mathcal{C}^*\mathcal{C}-n\mathcal{I}_G\|_{L_2}\right)\asymp\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L},$$

其中，最后一个表达式由假设 1 的第二部分推出。回顾一下，根据 BOS 假设，对所有 $f\in G$, $\|f\|_{L_\infty}\lesssim\sqrt{p}\|f\|_{L_2}$ 均成立。因此，将其应用于 $f=\overline{\mathcal{B}}\eta^*$，并注意到 $\|\overline{\mathcal{B}}\eta^*\|_{L_2}\leq b$，我们得到

$$\left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\bar{\alpha}}I_n-(\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right\|_{L_\infty}\lesssim\sqrt{p}\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L}b.$$

为了界定第三项，注意到 $\overline{\mathcal{B}}(\overline{\mathcal{M}}-\mathcal{M})$ 正是引理 9 中的算子 $\mathcal{Q}$。因此，$\|\overline{\mathcal{B}}(\overline{\mathcal{M}}-\mathcal{M})\|_{L_2}\leq c$. 此外，我们还有（再次利用引理 9 的结果）

$$\|(\overline{\mathcal{B}}-\mathcal{B})\eta^*\|_{L_2}\leq\frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_2}\leq\frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_\infty}\leq\frac{bc}{1-c}\tag{11}$$

现在，利用 BOS 假设，可以将式 (10) 中的第三项界定为：

$$\|(\overline{\mathcal{B}}(\overline{\mathcal{M}}-\mathcal{M})(\mathcal{B}-\overline{\mathcal{B}})\eta^*\|_{L_\infty}\leq\sqrt{p}\cdot\|\overline{\mathcal{B}}(\overline{\mathcal{M}}-\mathcal{M})\|_{L_2}\|(\overline{\mathcal{B}}-\mathcal{B})\eta^*\|_{L_2}\leq\frac{bc^2\sqrt{p}}{1-c}\lesssim bc^2\sqrt{p}$$

综合以上所有界，我们最终得到

$$|\mathcal{P}_G\hat{\eta}_0(x)-\overline{\mathcal{S}}\eta^*(x)|\lesssim\sqrt{\frac{p\log n}{n}}\cdot b+\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L}\cdot b\sqrt{p}+bc^2\sqrt{p}$$

以至少 $1-\frac{1}{n^2}$ 的概率。

□

为了界定式 (8) 中的项 C，我们可以将此引理应用于 n 留一训练集，并取并集界，从而以至少 $1-\frac{1}{n}$ 的概率得到最终界。结合前面三个小节中对项 A、B 和 C 的界，即可得到定理 1。

4.5 次高斯特征的证明

假设特征 $v_\ell(x)$ 不满足 BOS 条件，而是次高斯且相互独立。在这种情况下，我们首先注意到，对于所有 $f\in\mathcal{H}$，$f(x)=\langle k_x,f\rangle_{\mathcal{H}}$ 都是次高斯随机变量，其范数至多为 $\|f\|_{L_2}$（忽略常数因子）。利用这一观察，我们可以为式 (8) 中的每一项得到紧确界。

对于项 A，我们注意到，引理 1 的陈述在独立次高斯情形下也成立，且其成立概率至少为 $1-\frac{1}{n^2}-e^{-p}$（这一事实可在引理的证明中验证）。因此，将引理 1 应用于每个 n 留一法集合后，我们以概率 $1-ne^{-p}-\frac{1}{n}$ 得到项 A 的上界，该上界与 BOS 情形下相同。

对于项 B，我们注意到项 $|\mathcal{P}_{G^\perp}\hat{\eta}_{\setminus i,0}(x)|$ 是次高斯的，其范数至多为 $\left\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\right\|_{L_2}$. 我们通过以下引理对其进行界定。

引理 4（次高斯特征的 $G^\perp$ 误差项）。假设次高斯特征相互独立，且假设 1 成立。那么，

$$\left\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\right\|_{L_2}\lesssim\sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}b.$$

引理 4 的证明。我们采用与引理 2 的证明类似的方法。

$$\begin{aligned}\left\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\right\|_{L_2}&\leq\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\|_{\mathcal{H}}\\&=\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^*+\mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}\eta^*\|_{\mathcal{H}}\\&=\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}\eta^*\|_{\mathcal{H}}\\&\leq\|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\mathcal{H}}\|\mathcal{A}_G\|_{L_2\to\ell_2}\|\mathcal{B}\eta^*\|_{L_2}\\&\lesssim\sqrt{\lambda_{p+1}}\frac{1}{\sqrt{\alpha_L}}\sqrt{n}b=\sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}b.\end{aligned}$$

在倒数第二行中，我们利用引理 9 的结果来界定 $\|\mathcal{B}\eta^*\|_{L_2}$。 □

利用次高斯尾界，项 B 因此以至少 $1-\frac{1}{n^2}$ 的概率被界定为 $\sqrt{\lambda_{p+1} n \log n / a_L}$。随后，我们将其应用于每一个 n 留一法集合，从而以概率 $1-\frac{1}{n}$ 得到项 B 的上界。对于项 C，我们注意到，与引理 3 的证明中一样， that

$$|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| = |((\mathcal{B}-\overline{\mathcal{B}})\eta^*)(x)|,$$

其中 $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$（这可由 $\hat{}\eta$ 的定义以及推移恒等式推出）。因此，该项的次高斯范数至多为

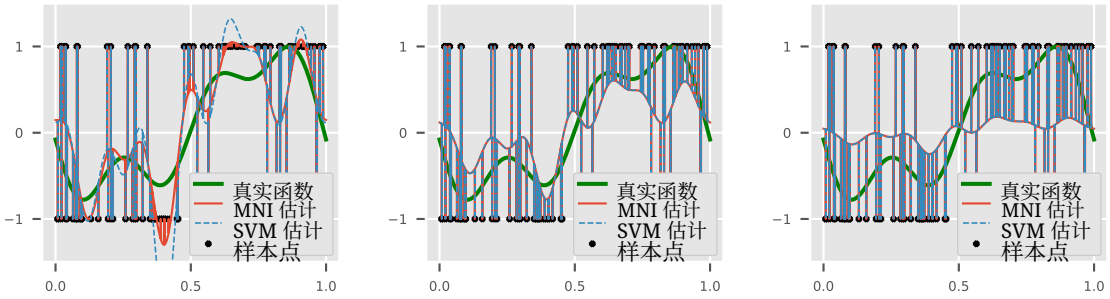
$$\|(\mathcal{B}-\overline{\mathcal{B}})\eta^*\|_{L_2} \le \frac{bc}{1-c} \lesssim bc,$$

其中我们用到了附录中的引理 9 的结果。因此，利用尾界可得，

$$|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| \lesssim bc\sqrt{\log n}$$

以至少 $1-\frac{1}{n^2}$ 的概率（因此，这对所有留一法集合以 $1-\frac{1}{n}$ 的概率成立）。结合上述界限可得定理 2。

5 模拟与讨论



(a) $q = -0.4$　　(b) $q = 0.4$　　(c) $q = 0.8$

图1：在双层集成中使用傅里叶特征时，不同有效过参数化程度下的 SVM 和 MNI 解。

为验证我们的理论发现，我们现在在双层模型下对 MNI 和 SVM 程序进行模拟。我们考虑输入空间 X 是区间 $[0,1]$，且采用均匀测度的情形。此外，我们考虑由特征函数[6] $v_\ell(x) = e^{j2\pi\ell x}$（$\ell = -d,\dots,d$）指定的傅里叶特征化。使用这些特征进行 MNI 或 SVM，并对特征进行双层加权，对应的核函数为

$$k(x,y) = \sum_{\ell=-d}^{d} \lambda_\ell v_\ell(x)\overline{v_\ell(y)} = \sum_{\ell=-d}^{d} \lambda_\ell e^{j2\pi\ell(x-y)}$$

其中，λ_ℓ 在 $\ell > 0$的公式 6 中给出，且 $\lambda_{-\ell} = \lambda_\ell$。对于所有实验，我们使用$n = 100$ 的训练样本量（请注意，n 的选择会影响核特征值）。

[6]为简单起见，我们使用复指数特征函数，但始终采用共轭对称系数，从而得到实函数。我们也可以使用纯实的正弦和余弦作为特征函数，但这样记号会稍微复杂一些。

我们首先生成一个随机真实函数 η^*，它是 $\{v_\ell\}_{\ell=-3}^{3}$ 的线性组合，系数服从标准正态分布，并经过缩放，使得 $\eta^*(x) \in [-1,1]$。对于训练集中的每个 x_i，我们根据规则 $\mathrm{P}(y_i = 1) = {}^{1+\eta^*}{}_{(x_i)\ 2}$ 生成 y_i。随后，我们针对三组双层~~参数计算~~并绘制 MNI 和支持向量机估计值：$(\beta, r, q) =$ (3.2, 0.4, -0.4)、(3.2, 0.4, 0.4) 和 (3.2, 0.4, 0.8)，它们对应不同程度的有效过参数化。结果如图1所示。根据推论 1，当 $q > \frac{1-0.4}{2} = 0.3$ 时，这两种估计以高概率一致。事实上，从图1b和图1c可以看出，第二组和第三组参数确实如此。我们的结果无法对 $q < 0.3$ 范围内是否会出现 SVP 得出结论，但从这个例子可以看出，有效过参数化程度很低时不会出现 SVP，如图1a所示。

为衡量双层集成参数的影响，我们还绘制了不同 r（优选特征的数量）和 q（有效过参数化）取值下 SVP 发生情况的热力图。在这里，我们固定 $n = 100$、$\beta = 3.2$ 以及真实函数 η^*（按照图 1 中的方法生成），并绘制对于每个（r，q），SVP 发生的试验比例（共 25 次）。图 2 给出了这些结果的热力图，并叠加显示了推论 1 预测 SVP 的参数区域。我们注意到，至少对于傅里叶特征而言，存在一些（r，q）取值，使得 SVP 会发生，但我们的理论结果并未预测到这些情况；理解这些区域中的 SVP（例如，$\frac{2}{3} < r < 1$）（以及更一般地，给出精确的下界）是未来研究中一个有趣且重要的方向。

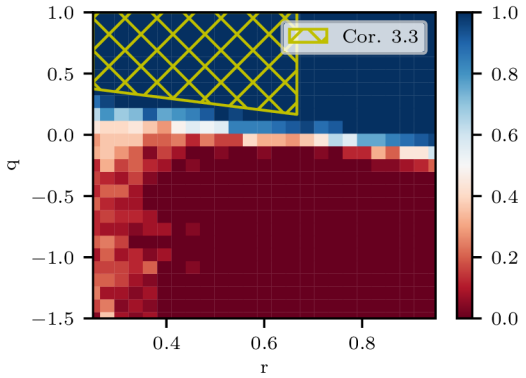


图 2：热力图展示了在傅里叶双层集成中，对不同参数选择进行 25 次试验时出现 SVP 的比例，并叠加了我们的结果所预测的 SVP 区域。

致谢

我们衷心感谢美国国家科学基金会、Adobe Research、Amazon Research 和 Google Research 对本研究的支持。CK 获得了美国国家科学基金会研究生研究奖学金 DGE-2039655 的资助。AM 和 MD 获得了美国国家科学基金会资助 DMS-2134037 和 CCF-2107455 的资助。VM 获得了美国国家科学基金会职业奖 CCF-239151、美国国家科学基金会奖项 IIS-2212182、Adobe数据科学研究奖、Amazon研究奖以及 Google Research Colabs奖的资助。

A 符号说明

表 1 回顾了我们在结果和证明中使用的主要符号。

B 集中界与技术引理

本节将陈述一些有用的集中界，可用于验证假设 1 中的算子集中性条件。这些不等式及其证明见于 [31]。类似的集中性

表 1：符号说明

符号(s)	定义(s)	描述
k_x	$k_x = k(\cdot, x)$	以 x 为中心的核函数
$\mathcal{T}$	$\mathcal{T}(f) = \int f(x)k_x \, d\mu(x)$	核 k 的积分算子
$\{(\lambda_\ell, v_\ell)\}_{\ell=1}^{\infty}$	$\mathcal{T}(f) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle f, v_\ell \rangle_{L_2} v_\ell, \ \lambda_1 \geq \lambda_2 \geq \cdots$	特征值分解 of $\mathcal{T}$
$\mathcal{A}$	$\mathcal{A}(f) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$	从 $\mathcal{H}$ 到的采样算子 $\mathbf{R}^n$
$\mathcal{A}^*$	$\mathcal{A}^*(z) = \sum_{i=1}^{n} z_i k_{x_i}$	$\mathcal{A}$ 关于 $\mathcal{H}$ 和 ℓ_2 的伴随算子内积
$G, G^\perp$	$G =$ 张成$\{v_1, \ldots, v_p\}$	前 p 个特征函数的张成空间的 $\mathcal{T}$（及其补空间）
$\mathcal{I}$ $(\mathcal{I}_G)$		恒等算子（限制于至 G）
$\mathcal{T}_G$, $\mathcal{T}_{G^\perp}$	$\mathcal{T}_G = \mathcal{T}\mathcal{P}_G, \quad \mathcal{T}_{G^\perp} = \mathcal{T}\mathcal{P}_{G^\perp}$	$\mathcal{T}$限制于 G 和 $G^\perp$
$\mathcal{A}_G$, $\mathcal{R}$	$\mathcal{A}_G = \mathcal{A}\mathcal{P}_G, \mathcal{R} = \mathcal{A}\mathcal{P}_{G^\perp}$	采样算子的限制-算子至 $G, G^\perp$
$\mathcal{C}, \mathcal{C}^*$	$\mathcal{C} = \mathcal{A}_G, \mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$	采样算子及其相关内容-关于联合 G 关于。L_2 内积 on G
α_L, α_U	$\alpha_L I_n \preceq \alpha I_n + \mathcal{R}\mathcal{R}^* \preceq \alpha_U I_n$	的下界和上界 残差 Gram 矩阵
c	$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}_{\setminus i}^* \mathcal{C}_{\setminus i} - n\mathcal{I}_G\|_{L_2} \leq c, \forall i \in [n]$	特征集中度上界于 G
b	$b = \|\overline{\mathcal{B}}\eta^*\|_\infty$	理想化偏差的无穷范数应用于 η^*
$\bar{\alpha}, \tilde{\alpha}$	$\bar{\alpha} = \frac{2\alpha_U \alpha_L}{\alpha_U + \alpha_L}, \tilde{\alpha} = \frac{\alpha_U + \alpha_L}{2}$	调和 and 算术 α_U、α_L 的均值
$\mathcal{B}$	$\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\alpha + \mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$	G 上的偏差算子
$\mathcal{S}$	$\mathcal{S} = \mathcal{I}_G - \mathcal{B}$	核回归算子
$\overline{\mathcal{B}}$	$\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	（“生存”）在 G 理想化 近似于偏置 $\mathcal{B}$
$\overline{\mathcal{S}}$	$\overline{\mathcal{S}} = \mathcal{I}_G - \overline{\mathcal{B}} = \frac{n}{\bar{\alpha}}\mathcal{T}_G\left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$	理想化近似至生存 $\mathcal{S}$

次高斯/独立特征的界也可见 [5, 39]。

引理 5（一般残差 Gram 矩阵集中）。设 $k^R(x,y)$ 是核 k 在$G^\perp$上的限制。那么，

$$\mathbf{E}\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\|^2 \lesssim n^2\operatorname{tr}(\mathcal{T}_{G^\perp}^2) + \|k^R(\cdot,\cdot) - \operatorname{tr}\mathcal{T}_{G^\perp}\|_\infty^2,$$

其中 $\|k^R(\cdot,\cdot) - \operatorname{tr}\ \mathcal{T}_{G^\perp}\|_\infty = \sup_x\{|k^R(x,x) - \operatorname{tr}\ \mathcal{T}_{G^\perp}|\}$。

引理 6（独立次高斯特征的残差 Gram 矩阵集中）。假设特征 $\{v_\ell(x)\}_{\ell\geq 1}$ 是零均值、独立且次高斯的。那么，对于 $t>0$，以至少$1-e^{-t}$的概率，

$$\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^\perp})I_n\| \lesssim \sqrt{(n+t)\operatorname{tr}(\mathcal{T}_{G^\perp}^2)} + (n+t)\lambda_{p+1}.$$

引理 7（BOS 中 G 上的采样算子集中）。假设核特征 $\{v_\ell(x)\}_{\ell\geq 1}$ 满足常数为 $C\geq 1$ 的 BOS 条件。那么，对于 $t>0$，以至少 $1-e^{-t}$ 的概率，

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{Cp(t+\log p)}{n}} + \frac{Cp(t+\log p)}{n}.$$

引理 8（独立次高斯特征的 G 上的采样算子集中）。假设特征 $\{v_\ell(x)\}_{\ell\geq 1}$ 是零均值、独立且次高斯的。那么，以至少 $1-e^{-t}$ 的概率，

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{p+t}{n}} + \frac{p+t}{n}.$$

接下来我们陈述以下近似结果，该结果在 [31], 中得到证明，我们经常利用它来证明各种技术引理。

引理 9（ [31] 中的引理 7）。设 $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ 且 $\overline{\mathcal{B}} = \left(\mathcal{I}_G + \frac{n}{\bar{\alpha}}\mathcal{T}_G\right)^{-1}$。那么，在假设 1 的条件下，

$$\mathcal{B} - \overline{\mathcal{B}} = \left(\sum_{k=1}^{\infty}(-\mathcal{Q})^k\right)\overline{\mathcal{B}},$$

对于满足 $\|\mathcal{Q}\|_{L_2}\leq c$ 的 $\mathcal{Q} := \left(\mathcal{T}_G^{-1} + \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)^{-1}\left(\mathcal{C}^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\bar{\alpha}}\mathcal{I}_G\right)$。这里， $c<1$ 是量 $\frac{\alpha_U-\alpha_L}{\alpha_U+\alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}$ 的上界。

注意，应用无穷几何级数求和公式可知，上述结果意味着 $\|(\mathcal{B}-\overline{\mathcal{B}})\eta^*\|_{L_2}\leq\frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_2}$ 且 $\|\mathcal{B}\eta^*\|_{L_2}\leq\left(1+\frac{c}{1-c}\right)\|\overline{\mathcal{B}}\eta^*\|_{L_2}$。

C 推论 1 和 2 的证明

推论 1 的证明。我们按照 [31, 推论 1]中的方法，对双层集成的每一项进行界定。首先，注意到

$$n\sqrt{\operatorname{tr}(\mathcal{T}_{G^\perp}^2)} \asymp n\cdot((n^\beta - n^r)\cdot n^{-2\beta+2r+2q})^{1/2} \asymp n^{1-\frac{\beta}{2}+r+q}.$$

根据薄壳假设，$\sup_x\left|\sum_{\ell>p}(v_\ell(x)^2-1)\right| \lesssim n^{\beta/2+1}$，

$$\begin{aligned}\sup_x|k^R(x,x) - \operatorname{tr}\mathcal{T}_{G^\perp}| &= \sup_x\left|\sum_{\ell>p}\lambda_\ell(v_\ell(x)^2-1)\right| = n^{-\beta+r+q}\sup_x\left|\sum_{\ell>p}(v_\ell(x)^2-1)\right| \\ &\lesssim n^{-\beta+r+q}n^{\beta/2+1} = n^{1-\beta/2+r+q},\end{aligned}$$

因此，引理 5 意味着，当 $n\to\infty$ 趋于极限时，只要 $1-\frac{\beta}{2}+r+q<r+q$ 成立，我们就可以选择 $\alpha_L = \lambda_{min}(\mathcal{R}\mathcal{R}^*)$ 和 $\alpha_U = \lambda_{max}(\mathcal{R}\mathcal{R}^*)$，使得 $\alpha_U,\alpha_L,\bar{\alpha}\asymp n^{r+q}$，而根据我们的假设， $\beta>2$ 确实成立。更准确地说，引理 5 给出了一个期望界，我们随后可以利用马尔可夫不等式得到一个高概率界。对于任意小的.....，我们都可以在相差 n^ϵ 倍的范围内得到所述界限。

$\epsilon > 0$，且随着 $n \to \infty$，这些界限成立的概率趋于 1。现在，我们可以利用假设 1 对项 c 进行界定。引理 5 表明，

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} \lesssim \frac{1}{\bar{\alpha}} n \sqrt{\mathrm{tr}(\mathcal{T}_{G^\perp}^2)} \asymp n^{1-\frac{\beta}{2}}.$$

然后，对每个 n 留一法集合应用引理 7，即可对 c 进行如下界定。

$$c \lesssim \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n} \|\mathcal{C}_{\setminus i}^* \mathcal{C}_{\setminus i} - n\mathcal{I}_G\|_{L_2} \lesssim n^{1-\frac{\beta}{2}} + n^{\frac{r-1}{2}} \sqrt{\log n}.$$

现在请注意，对于双层集成，我们有 $b = \left\| \frac{\bar{\alpha}}{\bar{\alpha}+n} \mathcal{I}_G \eta^* \right\|_{L_\infty} = \frac{\bar{\alpha}}{\bar{\alpha}+n} \|\eta^*\|_{L_\infty} \asymp \frac{\bar{\alpha}}{\bar{\alpha}+n} \asymp \min\{1, n^{r+q-1}\}$。我们考虑两种情况（忽略对数因子，因为当 $n \to \infty$ 时，与 n 的幂次相比，它们可以忽略不计）：

1. 若满足 $q > 1 - r$、 $b \asymp 1$，则定理 1 中 SVP 的条件可写为

$$n^{(r-1)/2} + n^{1-\beta/2} n^{1/2} + (n^{2-\beta} + n^{r-1}) n^{r/2} \ll 1$$

很容易验证，当 $r < 2/3$ 和 $\beta > 3$ 时，这对 $n \to \infty$ 成立。

2. 如果 $q \leq 1 - r$，那么 $b \asymp n^{r+q-1}$，因此该条件是

$$n^{1-r-q} n^{(r-1)/2} + n^{1-\beta/2}(n^{1-r-q} + n^{1/2}) + (n^{2-\beta} + n^{r-1}) n^{r/2} \ll 1$$

这对 $r < 2/3$、 $q > \frac{1-r}{2}$ 和 $\beta > 3$ 成立，其中 $n \to \infty$

结合上述两种情况即可得到结论。

□

推论 2 的证明。该陈述的证明与推论 1 类似，但使用了关于独立次高斯特征的引理 6 和引理 8 所给出的更精确的集中界。同样地，我们首先对量 $\alpha_U - \alpha_L$ 和 c 进行界定。引理 6 意味着，以至少 $1 - n^{-1}$ 的概率，

$$\|\mathcal{R}\mathcal{R}^* - (\mathrm{tr}\, \mathcal{T}_{G^\perp}) I_n\| \lesssim n^{1/2-\beta/2+r+q} + n^{1-\beta+r+q}.$$

由于 $\beta > 1$，我们再次可以选择带有 α_U、 α_L、 $\bar{\alpha} \asymp n^{r+q}$ 的 $\alpha_L = \lambda_{min}(\mathcal{R}\mathcal{R}^*)$ 和 $\alpha_U = \lambda_{max}(\mathcal{R}\mathcal{R}^*)$。此外，

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} \lesssim \frac{1}{n^{r+q}} n^{(1-\beta)/2+r+q} \asymp n^{(1-\beta)/2},$$

而引理 8（再次对 n 留一法集合应用并集界）意味着我们可以取
$c \asymp n^{(1-\beta)/2} + n^{\frac{r-1}{2}} \gtrsim \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n} \|\mathcal{C}_{\setminus i}^* \mathcal{C}_{\setminus i} - n\mathcal{I}_G\|_{L_2}$。我们再次考虑前一个推论中的两种情况（同样省略对数因子）：

1. 如果 $q > 1 - r, b \asymp 1$，定理 2 中 SVP 的条件可以写成

$$n^{(r-1)/2} + n^{(1-\beta)/2} + \sqrt{\frac{n^{-\beta+r+q+1}}{n^{r+q}}} + (n^{(1-\beta)/2} + n^{\frac{r-1}{2}}) \ll 1$$

这里，左侧对于 $\beta > 1$ 和 $r < 1$ 而言总是随 n 递减。

2. 如果 $q \leq 1 - r$，那么 $b \asymp n^{r+q-1}$，因此条件为

$$n^{\frac{r-1}{2}+1-r-q} + n^{1-r-q} \cdot n^{(1-\beta)/2} + \sqrt{\frac{n^{-\beta+r+q+1}}{n^{r+q}}} + (n^{(1-\beta)/2} + n^{\frac{r-1}{2}}) \ll 1$$

在这种情况下，左侧对于 $q > \frac{1-r}{2}$ 和 $\beta > 3 - 2r - 2q$ 是递减的。

综合上述两种情形，即可得到所需结果。 □

D 推论 3 的证明

为证明这一推论，我们首先刻画参数 (n, β, q, r) 的取值范围；在该范围内，当 $n \to \infty$ 时，MNI解的分类超额风险趋于 0。在 BOS 情形中， [31] 的推论 1 已证明，这种情况会在以下两种设定下发生：

1. $q < 1 - r$、 $\beta > 2$ （回归也具有一致性） 2. $1 - r < q < \frac{3}{2}$ （$1 - r$）和 $\beta > 2$（$r + q$）（分类具有一致性，但回归不具有一致性）。

结合推论 1 中有界正交系统满足 SVP 的条件，我们可以立即得出结论：在两个条件都成立的情况下（即 SVP 发生且 MNI 一致时），支持向量机解是分类一致的。推论 3 的第一部分得证。

现在，为了证明该推论的第二部分，我们首先通过使用针对独立次高斯特征的更精确的集中界，修改 [31] 中推论 1 的证明。我们首先回顾 [31] 中超额风险的以下表达式： $\mathcal{E} \leq \frac{\|\hat{\eta}_r\|_{L_2}}{s}$，其中 $s > 0$ 是一个我们可以选择的参数，并且 $\hat{\eta} = s\eta^* + \hat{\eta}_r$。

注意，双层集成中的理想化生存算子由 $\overline{\mathcal{S}} = \frac{n}{n+\bar{\alpha}}\mathcal{I}_G \asymp \frac{n}{n+n^{r+q}}\mathcal{I}_g \asymp \min(1, n^{1-r-q})\mathcal{I}_G$ 给出。因此，我们将选择 $s = \min(1, n^{1-r-q})$，并应用上述关于超额分类风险的界。然后，根据定义，$\hat{\eta}_r = \epsilon + \hat{\eta}_0 - \overline{\mathcal{S}}\eta^*$，因此

$$\|\hat{\eta}_r\|_{L_2} \leq \|\epsilon\|_{L_2} + \|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2},$$

第一项可以由 [31], 的定理 2 界定，从而得到

$$\|\epsilon\|_{L_2} \lesssim \sqrt{\frac{p}{n}} + \left(\frac{n\sum_{\ell>p}\lambda_\ell^2}{\left(\sum_{\ell>p}\lambda_\ell\right)^2}\right)^{1/2} \asymp n^{(r-1)/2} + \left(\frac{nn^{\beta}n^{-2\beta+2r+2q}}{n^{2\beta}n^{-2\beta+2r+2q}}\right)^{1/2} = n^{(r-1)/2} + n^{(1-\beta)/2},$$

其中，我们使用了在推论 2 的证明中推导出的渐近尺度。

此外，根据 [31] 的引理 6，我们可以对 $\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2}$ 进行界定：

$$\begin{aligned}
\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2} &\lesssim \left(c + \sqrt{\frac{n\lambda_{p+1}}{\bar{\alpha}}}\right)\min\left(\lambda_1, \frac{\bar{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\bar{\alpha}}{n}}\right)\|\eta^*\|_{\mathcal{H}} \\
&\lesssim \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}} + n^{(1-\beta)/2}\right)\min\left(1, n^{r+q-1}, n^{(r+q-1)/2}\right) \\
&\lesssim \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}}\right)\min\left(1, n^{r+q-1}\right)
\end{aligned}$$

因此，我们可以得到风险上界。

$$\mathcal{E} \lesssim \max(1, n^{r+q-1})\left(n^{(r-1)/2} + n^{(1-\beta)/2} + \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}}\right)\min\left(1, n^{r+q-1}\right)\right)$$

由此我们得到如下结论：

1. If $q < 1 - r$, then $\mathcal{E} \to 0$ (and, in fact $s \to 1$).
2. if $q > 1 - r$, then $\mathcal{E} \to 0$ if $q < \frac{3}{2}(1 - r)$ and $q < 1 + \frac{\beta-1}{2} - r$.

将这一允许参数范围与推论 2 中 SVP 的条件结合起来，即可得到该推论的第二部分。同样，$q \leq \frac{1-r}{2} < 1 - r$ 的情形已由先前结果涵盖，详见 [21]。

参考文献

[1] Navid Ardeshir、Clayton Sanford 和 Daniel J Hsu。《支持向量机与线性回归在超高维特征下相一致》。载于：神经信息处理系统进展 34（2021），第 4907–4918 页。

[2] Jean-Yves Audibert 和 Alexandre B. Tsybakov。《插件分类器的快速学习率》。载于：《统计年刊》35.2（2007），第 608–633 页。 DOI: 10.1214/009053606000001217.[3] Daniel Bartl 和 Shahar Mendelson。《具有近似高斯畸变的随机嵌入》。载于：《数学进展》400（2022）。 DOI: 10.1016/j.aim.2022.108261.[4] Peter Bartlett 和 John Shawe-Taylor。《支持向量机及其他模式分类器的泛化性能》。载于：《核方法——支持向量学习》（1999），第 43–54 页。[5] Peter Bartlett 等。《线性回归中的良性过拟合》。载于：《美国国家科学院院刊》117.48（2020），第 30063–30070 页。 DOI: 10.1073/pnas.1907378117.[6] Mikhail Belkin。《无惧拟合：透过插值的棱镜看深度学习中引人注目的数学现象》。载于：《数值数学学报》30（2021），第 203–248 页。 DOI: 10.1017/s0962492921000039.[7] Mikhail Belkin、Daniel Hsu 和 Ji Xu。《弱特征的双下降模型》。载于：《SIAM数学与数据科学杂志》2.4（2020），第 1167–1180 页。 DOI: 10.1137/20M1336072.[8] Mikhail Belkin、Siyuan Ma 和 Soumik Mandal。《要理解深度学习，我们需要理解核学习》。载于：国际机器学习会议（ICML）论文集。第 35 卷。瑞典斯德哥尔摩，2018。[9] Arnaud Buhot 和 Mirta B Gordon。《支持向量机的稳健学习与泛化》。载于：《物理学A：数学与一般物理》34.21（2001），第 4377 页。[10] Yuan Cao、Quanquan Gu 和 Mikhail Belkin。《次高斯混合分布上过参数化最大间隔分类的风险界》。载于：神经信息处理系统大会（NeurIPS）论文集。线上会议，2021 年 12 月。 URL: https://proceedings.neurips.cc/paper/2021/hash/46e0eae7d5217c79c3ef6b4c212b

[11] Niladri S. Chatterji 和 Philip M. Long。《过参数化区域中插值线性分类器的有限样本分析》。载于：《机器学习研究杂志》22（2021）。[12] Corinna Cortes 和 Vladimir Vapnik。《支持向量网络》。载于：《机器学习》20（1995），第 273–297 页。[13] Yehuda Dar、Vidya Muthukumar 和 Richard G. Baraniuk。《告别偏差-方差权衡？过参数化机器学习理论概述》。载于：（2021）。arXiv：2109。02355 [stat.ML]

[14] Rainer Dietrich、Manfred Opper 和 Haim Sompolinsky。“支持向量网络的统计力学”。载于：《Physical Review Letters》82.14（1999），第 2975 页。[15] Pascal Germain 等。“一种针对核方法的 PAC-Bayes 样本压缩方法 ”。载于：ICML。2011。[16] Pavel Golik 、 Patrick Doetsch 和 Hermann Ney。“交叉熵与平方误差训练：理论与实验比较。”载于：Interspeech。第 13 卷。2013，第 1756–1760 页。[17] Thore Graepel、Ralf Herbrich 和 John Shawe-Taylor。“用于分类的学习算法预测误差的 PAC-Bayesian 压缩界 ”。载于：《 Machine Learning 》 59.1 （ 2005 ），第 55–76 页。[18] Suriya Gunasekar 等。“梯度下降在线性卷积网络中的隐式偏置”。载于：神经信息处理系统进展 31（2018）。[19] Trevor Hastie 等。“高维无岭最小二乘插值中的意外发现”。载于：《统计年刊》50.2（2022）。 DOI: 10.1214/21-AOS2133。[20] Daniel Hsu、Sham M. Kakade 和 Tong Zhang。“岭回归的随机设计分析”。载于：《Foundations of Computational Mathematics》14（2014），第 569–600 页。 DOI: 10.1007/s10208-014-9192-1。[21] Daniel Hsu、Vidya Muthukumar 和 Ji Xu。“高维中支持向量的激增”。载于：国际人工智能与统计学会议论文集（AISTATS）。线上会议，2021。 URL: http://proceedings.mlr.press/v130/hsu21a.html。[22] Like Hui 和 Mikhail Belkin。“分类任务中采用平方损失与交叉熵训练的神经网络架构评估”。载于：国际学习表征会议论文集（ICLR）。线上会议，2021。 URL: https://openreview.net/forum?id=hsFN92eQEla。[23] Ziwei Ji 和 Matus Telgarsky。“深度学习中的方向收敛与对齐”。载于：神经信息处理系统进展 33（2020），第 17176–17186 页。

[24] Ziwei Ji 和 Matus Telgarsky。“梯度下降在不可分数据上的隐式偏置”。载于：学习理论会议。P
MLR。2019，第 1772–1798 页。[25] Ziwei Ji 等。“梯度下降遵循一般损失函数的正则化路径”。载于：学习理论会议。PMLR。2020，第 2109–2136 页。[26] Douglas M Kline 和 Victor L Berardi。“重新审视用于训练神经网络分类器的平方误差函数与交叉熵函数”。载于：《Neural Computing & Applications》14（2005），第 310–318 页。[27] Kuo-Wei Lai 和 Vidya Muthukumar。“一般损失函数在高维中导致（近似）插值”。载于：arXiv 预印本 arXiv:2303.07475（2023）。[28] Tengyuan Liang 和 Alexander Rakhlin。“直接插值：核无岭回归可以实现泛化”。载于：《统计年刊》48.3（2020），第 1329–1347 页。 DOI：10.1214/19-aos1849。[29] D¨orthe Malzahn 和 Manfred Opper。“一种分析机器学习算法在真实数据上表现的统计物理学方法”。载于：《统计力学杂志：理论与实验》2005.11（2005），P11001。[30] David McAllester。“简化的 PAC-Bayesian 间隔界”。载于：《学习理论与核机器》。Springer，2003，第 203–215 页。[31] Andrew D. McRae 等。“结构化特征下回归与分类中的无害插值”。载于：国际人工智能与统计学会议论文集（AISTATS）。线上会议，2022。 URL：https://pro

ceedings.mlr.p

[32] Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。“随机特征与核方法的泛化误差：超收缩性与核矩阵集中性”。载于：《应用计算调和分析》59（2022），第 3–84 页。 DOI: 10.1016/j.acha.2021.12.003.[33] Vidya Muthukumar 等。“过参数化区域中的分类与回归：损失函数重要吗？”载于：《机器学习研究杂志》（2021）。arXiv：2005.08054。即将发表。[34] Vidya Muthukumar 等。“回归中噪声数据的无害插值”。载于：IEEE《精选信息理论领域期刊》1.1（2020），第 67–83 页。 DOI: 10.1109/jsait.2020.2984716.[35] Behnam Neyshabur 、 Ryota Tomioka 和 Nathan Srebro。“探寻真正的归纳偏置：论隐式正则化在深度学习中的作用”。载于：arXiv 预印本 arXiv:1412.6614（2014）。[36] Ryan Michael Rifkin。《一切旧事皆焕然一新：重新审视机器学习中的历史方法》。博士论文。麻省理工学院，2002 年。[37] Daniel Soudry 等。“梯度下降在可分数据上的隐式偏置”。载于：《机器学习研究杂志》19.1（2018），第 2822–2878 页。[38] Ingo Steinwart 和 Clint Scovel。“一般定义域上的 Mercer 定理：测度、核与再生核希尔伯特空间之间的相互作用”。载于：《构造逼近》35（2012），第 363–417 页。 DOI: 10.1007/s00365-012-

9153-3。

[39] Alexander Tsigler 和 Peter L Bartlett。《岭回归中的良性过拟合》。发表于：（2020）。arXiv：2009.14286 [math.ST].[40] Roman Vershynin。《高维概率：数据科学应用导论》。剑桥，2018 年。296 页。 ISBN：1108415199。[41] Ke Wang、Vidya Muthukumar 和 Christos Thrampoulidis。《多类别分类中的良性过拟合：条条大路通向插值》。载于《神经信息处理系统进展》第 34 卷（2021），第 24164–24179 页。[42] Ke Wang 和 Christos Thrampoulidis。《高斯混合模型二分类中的良性过拟合》。载于 IEEE 声学、语音与信号处理国际会议论文集（ICASSP）。加拿大多伦多，2021 年。 DOI: 10.1109/icassp39728.2021.9413946。[43] Chiyuan Zhang 等。《理解深度学习需要重新思考泛化》。载于国际学习表征会议论文集（ICLR）。法国土伦，2017 年。 URL：https://openreview.net/pdf?id=Sy

8gdB9xx.

[44] Tong Zhang。《利用有效数据维数进行核回归的学习界》。载于：Neural Comput. 17 (2005)，第 2077–2098 页。 DOI: 10.1162/0899766054323008。