arXiv:2303.01566v1 [stat.ML] 2 Mar 2023

无监督预训练的可证明优势

Jiawei Ge*† Shange Tang *† Jianqing Fan† Chi Jin‡

摘要

无监督预训练利用大量无标签数据学习有用的表示，以促进下游任务的学习，是现代大规模机器学习系统的重要组成部分。尽管无监督预训练在实践中取得了巨大成功，但对于其为何通常能发挥作用，严格的理论理解仍相当有限——现有结果大多局限于特定的无监督预训练方法或途径，并依赖专门的结构假设。本文研究一个通用框架，其中，无监督表征学习任务由一类抽象的潜变量模型 Φ 刻画，下游任务则由一类预测函数 Ψ刻画。我们考虑一种自然的方法：在无监督预训练中使用最大似然估计（MLE），在下游任务学习中使用经验风险最小化（ERM）。我们证明，在一个温和的"信息性"条件下，我们的算法在下游任务上的超额风险为 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_\Phi/m}+\sqrt{\mathcal{C}_\Psi/n})$，其中 $\mathcal{C}_\Phi$、$\mathcal{C}_\Psi$ 是函数类 Φ、Ψ的复杂度度量，m、n 分别是无标签数据和有标签数据的数量。与仅使用有标签数据进行监督学习所达到的 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_{\Phi\circ\Psi}/n})$基线相比，我们的结果严格证明了当$m \gg n$ 且 $\mathcal{C}_{\Phi\circ\Psi} > \mathcal{C}_\Psi$时，无监督预训练能够带来益处。本文进一步表明，我们的通用框架涵盖了多种无监督预训练方法，包括因子模型、高斯混合模型和对比学习。

1 引言

无监督预训练旨在高效利用大量无标签数据，学习有助于下游任务学习的有用表示。这项技术已广泛应用于现代机器学习系统，包括计算机视觉（Caron 等，2019年；Dai 等，2021年）、自然语言处理（Radford 等，2018年；Devlin 等，2018年；Song 等，2019年）和语音处理（Schneider 等，2019年；Baevski 等，2020年）。尽管在实证上取得了巨大成功，但人们仍不清楚，为何在未利用下游任务相关信息的情况下学习得到的预训练表征，往往有助于下游任务的学习。

近期已有多项研究尝试从理论角度理解无监督预训练的各种方法，包括语言模型Saunshi 等人（2020年）；Wei 等人（2021年）、对比学习Arora 等人（2019年）；Tosh 等人（2021年（b），a）；HaoChen 等人（2021年）；Saunshi 等人（2022年），以及基于重构的自监督学习Lee 等人（2021年）。

*共同贡献

†普林斯顿大学运筹学与金融工程系；{jg5300,shangetang,jqfan}@princeton.edu

‡普林斯顿大学电气与计算机工程系；chij@princeton.edu

尽管这类研究为在相应场景中采用无监督预训练提供了依据，但其中许多研究并未证明无监督学习在样本复杂度方面的优势，即使是与仅使用有标签数据进行监督学习这一朴素基线相比也是如此。此外，这些结果仅适用于论文中考察的特定无监督预训练方法，而且关键在于依赖专门的结构假设，这些假设无法推广到所研究场景之外。因此，我们提出以下问题：

我们能否构建一个通用框架，以可证明的方式解释无监督预训练的优势？

Thi本文对上述问题给出了肯定的回答。 ly.

我们考虑一个通用设定，其中数据 x 及其标签 y 通过一个未观测到的表示 z相联系。具体而言，我们假设 (x, z) 是从抽象类 Φ中的潜变量模型 ϕ^* 采样得到的，并且给定表示 z 时，标签 y 的分布取自类 Ψ中的分布 ψ^* 。我们考虑一种自然方法：在无监督预训练中使用最大似然估计（MLE），通过 m 个无标签数据近似学习潜变量模型 ϕ^* 。随后，我们结合表征学习和经验风险最小化（ERM）的结果，使用 n 个有标签数据学习下游预测器 ψ^* 。研究这一通用设定，使我们能够绕开先前研究的局限；那些研究仅限于特定的无监督预训练方法。

我们证明，在一个温和的“信息性”条件（假设 3.2）下，我们的算法在下游任务上的超额风险为 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_\Phi/m} + \sqrt{\mathcal{C}_\Psi/n})$，其中 $\mathcal{C}_\Phi, \mathcal{C}_\Psi$ 是函数类 Φ, Ψ 和 m, n 的复杂度度量，分别对应无标签数据和有标签数据的数量。与仅使用有标签数据进行监督学习所达到的基线 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_{\Phi\circ\Psi}/n}$相比，我们的结果严格证明了：当无标签数据 $m \gg n$ 充足，且复合类 $\mathcal{C}_{\Phi\circ\Psi}$ 的复杂度远高于单独下游任务的复杂度时，无监督预训练具有优势 $\mathcal{C}_\Psi$。

最后，本文证明，我们的通用框架（包括“信息性”条件）涵盖了广泛的无监督预训练设定，包括：(1) 以线性回归为下游任务的因子模型；(2) 以分类为下游任务的高斯混合模型；以及 (3) 以线性回归为下游任务的对比学习。

1.1 相关工作

无监督预训练的应用与方法。无监督预训练已在图像识别（Caron et al., 2019）、目标检测（Dai et al., 2021）、自然语言处理（Devlin et al., 2018; Radford et al., 2018; Song et al., 2019）和语音识别（Schneider et al., 2019; Baevski et al., 2020）领域取得巨大成功。最广泛使用的两种预训练方法是：(1) 基于特征的方法（Brown et al., 1992; Mikolov et al., 2013; Melamud et al., 2016; Peter et al., 2018），这类方法通过预训练模型来提取表示，并直接将预训练表示作为下游任务的输入；(2) 基于微调的方法（例如，参见 Devlin et al., 2018），这类方法根据下游任务，在预训练表示附近对所有模型参数进行微调。Erhan et al. (2010) 首次从经验上阐明了预训练的作用。他们认为，预训练是一种正则化形式，能够有效引导下游任务的学习。

预训练的多数应用场景都属于半监督学习（例如，Zhu，2005年），在这些场景中，训练过程中会观察到大量无标签数据和少量有标签数据。半监督学习方法旨在构建

更好的预测器，并有效利用无标签数据。一些传统方法包括：生成模型（例如，Ratsaby 和 Venkatesh，1995年）、低密度分离（Joachims 等人，1999；Lawrence 和 Jordan，2004；Szummer 和 Jaakkola，2002年）、基于图的方法（Belkin 等人，2006年）。尽管这一领域的大多数研究都提出了新方法，并展现出良好的实证性能，但并未对无监督预训练的优势提供严谨的理论解释。

无监督预训练的理论理解。近年来，理论结果大量涌现，为各种从无标签数据中提取表示的无监督预训练方法提供了解释。例如，(Saunshi et al., 2020; Wei et al., 2021) 研究了语言模型中的向量嵌入预训练，而 (Arora et al., 2019; Tosh et al., 2021b,a; HaoChen et al., 2021; Saunshi et al., 2022; Lee et al., 2021) 则研究了几种用于预训练的自监督学习（SSL）方法。在结果方面，Wei et al. (2021) 表明，在预训练语言模型之上使用线性预测器可以恢复其真实模型；Arora et al. (2019); Saunshi et al. (2020); Tosh et al. (2021b,a); Saunshi et al. (2022) 表明，下游任务的预测损失可以由无监督预训练任务的损失界定。这两类结果都未能证明，与仅使用有标签数据进行监督学习这一基线相比，无监督学习具有样本复杂度优势。

最相关的研究结果来自 Lee 等人 (2021年)；HaoChen 等人 (2021年) 明确展示了某些无监督预训练方法在样本复杂度方面的优势。然而，Lee 等人(2021年) 关注基于重构的自监督学习，并且关键依赖于这样一个条件独立性假设：给定标签后，特征与其重构相互独立；HaoChen 等人 (2021年) 研究对比学习，其结果依赖于确定性特征映射以及归一化邻接矩阵的谱条件。这两项研究结果都仅适用于各自论文中特定的无监督预训练设定和方法，通常无法推广到其他设定（例如第 4、5、6 节中的三个例子）。相反，本文仅借助抽象函数类，构建了一个适用于多种设定的无监督预训练通用框架。

表征学习的其他方法。近年来，另一类理论研究通过多任务学习来学习表示。Baxter (2000年) 在假设存在生成模型和任务间共享表征的前提下，给出了多任务迁移学习的泛化界。Maurer 等人 (2016) 从理论上分析了一种从多个任务中学习表示的通用方法，并在线性特征设定下展示了该方法。Tripuraneni 等人(2021年)；Du 等人 (2020年) 提出了能够解决多任务线性回归问题的样本高效算法。Tripuraneni 等人 (2020年) 进一步研究了通用的非线性特征表征，并为多种训练任务给出了样本复杂度保证。他们的研究与我们的工作不同之处在于，他们利用其他任务的有标签数据通过监督学习来学习表示，而我们的工作则利用无标签数据通过无监督学习来学习表示。

2 问题设定

记号。我们分别用 $\mathbb{P}(x)$ 和 $p(x)$表示定义在 $x \in \mathcal{X}$上的累积分布函数和概率密度函数。我们定义 $[n] = \{1, 2, \ldots, n\}$。集合的基数

$\mathcal{A}$ 记为 $|\mathcal{A}|$。令 $\|\cdot\|_2$ 表示向量的 ℓ_2 范数或矩阵的谱范数。我们用 $\|\cdot\|_{\mathrm{F}}$ 表示矩阵的弗罗贝尼乌斯范数。对于矩阵 $M \in \mathbb{R}^{m\times n}$，我们分别用 $\sigma_{\min}(M)$ 和 $\sigma_{\max}(M)$表示 M的最小奇异值和最大奇异值。对于两个概率分布 $\mathbb{P}_1$ 和 $\mathbb{P}_2$，我们分别用 $d_{\mathrm{TV}}(\mathbb{P}_1,\mathbb{P}_2)$ 和 $H(\mathbb{P}_1,\mathbb{P}_2)$表示这两个分布之间的全变差（TV）距离和海林格距离。

我们分别用 $x \in \mathcal{X}$ 和 $y \in \mathcal{Y}$ 表示下游任务的输入数据和目标。我们的目标是使用 x 预测 y。我们假设 x 通过未观测到的潜变量 $z \in \mathcal{Z}$ 与 y 相连（该潜变量也被视为 x 的一种表示）。给定潜变量 z 后，数据 x 与目标 y 相互独立。为了涵盖一大类现实世界应用，例如对比学习，我们考虑学习过程可能能够获取一些辅助信息 $s \in \mathcal{S}$ 的情形。我们假设 $(x,s,z) \sim \mathbb{P}_{\phi^*}(x,s,z)$ 且 $y|z \sim \mathbb{P}_{\psi^*}(y|z)$，其中 $\mathbb{P}_{\phi^*}$ 和 $\mathbb{P}_{\psi^*}$ 是分别由 $\phi^* \in \Phi$ 和 $\psi^* \in \Psi$ 索引的分布。因此有

$$\mathbb{P}_{\phi^*,\psi^*}(x,y) = \int \mathbb{P}_{\phi^*}(x,z)\mathbb{P}_{\psi^*}(y|z)\,dz,$$

这意味着 (x,y) 的概率分布同时取决于 ϕ^* 和 ψ^*。我们的设定包括这样一种特殊情形：$y = f^*(z) + \varepsilon$。函数 $f^* \in \mathcal{F}$ 是真实函数，而 $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ 是与 z 独立的高斯噪声。在这种情况下，条件随机变量$y|z \sim \mathcal{N}(f^*(z)\ \sigma^2)$ 的概率分布 $\mathbb{P}_{f^*}(y|z)$ 由 $f^* \in \mathcal{F}$ 参数化。

令 $\ell(\cdot,\cdot)$ 为损失函数。对于任意一对 $(\phi,\psi) \in \Phi\times\Psi$，最优预测器 $g_{\phi,\ \psi}$ 定义如下，

$$g_{\phi,\psi} \leftarrow \arg\min_{g} \mathbb{E}_{\mathbb{P}_{\phi,\psi}}\big[\ell\big(g(x),y\big)\big], \tag{1}$$

其中，最小值是在所有可能的函数上取的，且 $\mathbb{E}_{\mathbb{P}_{\phi,\ \psi}} := \mathbb{E}_{(x,y)\sim\mathbb{P}_{\phi,\ \psi}(x,\ y)}$。因此，我们的预测函数类定义为

$$\mathcal{G}_{\Phi,\Psi} := \big\{g_{\phi,\psi} \big| \phi \in \Phi, \psi \in \Psi\big\}.$$

特别地，如果 $\ell(\cdot,\cdot)$ 是平方损失函数，那么最优预测器存在闭式解 $g_{\phi,\ \psi}(x) = \mathbb{E}_{\mathbb{P}_{\phi,\ \psi}}[y|x]$ 而预测函数类 $\mathcal{G}_{\Phi,\ \Psi} = \{\mathbb{E}_{\mathbb{P}_{\phi,\psi}}[y|x] \,|\, \phi \in \Phi, \psi \in \Psi\}$。

给定一个估计器对 $(\hat{\phi},\hat{\psi})$，我们定义相对于损失 $\ell(\cdot,\cdot)$ 的超额风险 a s

$$\mathrm{Error}_{\ell}(\hat{\phi},\hat{\psi}) := \mathbb{E}_{\mathbb{P}_{\phi^*,\psi^*}}\big[\ell\big(g_{\hat{\phi},\hat{\psi}}(x),y\big)\big] - \mathbb{E}_{\mathbb{P}_{\phi^*,\psi^*}}\big[\ell\big(g_{\phi^*,\psi^*}(x),y\big)\big], \tag{2}$$

^ ^ 其中 ϕ^* 和 ψ^* 是真实参数。根据 $g_{\phi^*,\ \psi^*}$ 的定义，我们有 $\mathrm{Error}(\phi,\psi) \geq$^ ^ 0。我们的目标是从数据中学习一个估计器对 (ϕ,ψ)，使超额风险达到最小阶。

我们考虑无法观测到潜变量 z 的情形。具体而言，我们有大量无标签数据及其对应的辅助信息 $\{x_i,s_i\}_{i=1}^m$，这些数据独立同分布地采样自未知分布 $\mathbb{P}_{\phi^*}(x,s)$；此外，还有少量有标签数据 $\{x_j,y_j\}_{j=1}^n$，它们独立同分布地采样自未知分布 $\mathbb{P}_{\phi^*,\ \psi^*}(x,y)$。这里我们假设有标签数据 $\{x_j,y_j\}_{j=1}^n$与无标签数据 $\{x_i,s_i\}_{i=1}^m$ 相互独立，并以 $m \gg n$为前提。

算法 1 两阶段最大似然估计+经验风险最小化

1: 输入： $\{x_i, s_i\}_{i=1}^m$, $\{x_j, y_j\}_{j=1}^n$ ^2: 利用无标签数据及其对应的辅助信息 $\{x_i, s_i\}_{i=1}^m$ ， 通过最大似然估计学习 ϕ ：

$$\hat{\phi} \leftarrow \arg\max_{\phi \in \Phi} \sum_{i=1}^{m} \log p_\phi(x_i, s_i). \tag{3}$$

^ ^3: 固定 ϕ ，并使用有标签数据 $\{x_j, y_j\}_{j=1}^n$ 通过经验风险最小化学习 ψ ：

$$\hat{\psi} \leftarrow \arg\min_{\psi \in \Psi} \sum_{j=1}^{n} \ell\big(g_{\hat{\phi},\psi}(x_j), y_j\big). \tag{4}$$

^ ^4: 输出： ϕ 和 ψ。

学习算法。我们考虑一种由两个阶段组成的自然学习算法（算- 法 1）。在无监督预训练阶段，我们使用最大似然估计，基于无标签数据 ϕ^* 估计 $\{x_i, s_i\}_{i=1}^m$。在下游任务学习阶段，我们使用经验风险最小化，基于 ^预训练的 ψ^* 和有标签数据 ϕ 估计 $\{x_j, y_j\}_{j=1}^n$。详见算法 1 。

我们指出，在我们的设定中，另一种自然的学习算法是采用两阶段最大似然估计。具体而言，在无监督预训练阶段，我们使用最大似然估计，基于无标签数据 ϕ^* 按照式（ $\{x_i, s_i\}_{i=1}^m$ ）估计 3。在下游任务学习阶段，我们再次使用最大似然估计，基于预训练的 ψ^* 和有标签数据 ϕ ^估计 $\{x_j, y_j\}_{j=1}^n$。然而，我们可以证明，该两阶段最大似然估计方案在最坏情况下会失败。详见附录 E 。

复杂度度量。 算法 1 的样本复杂度保证将用三种复杂度度量来表述，即括号覆盖数、覆盖数和拉德马赫复杂度，定义如下。我们用 $\mathcal{P}_\mathcal{X}(\Phi)$ 表示定义在 $p_\phi(x)$ 上的一组参数化密度函数 $x \in \mathcal{X}$

$$\mathcal{P}_\mathcal{X}(\Phi) := \{p_\phi(x) \,|\, \phi \in \Phi\},$$

其中 $\phi \in \Phi$ 是参数。

定义 2.1（ϵ-括号与括号覆盖数）。令 $\epsilon > 0$。在 $\|\cdot\|_1$ 距离下，若函数集 $\mathcal{N}_{[\,]}(\mathcal{P}_\mathcal{X}(\Phi)\ \epsilon)$ 是 $\mathcal{P}_\mathcal{X}(\Phi)$ 的一个 ϵ-括号，则对于任意 $p_\phi(x) \in \mathcal{P}_\mathcal{X}(\Phi)$，都存在一个函数$\bar{p}_\phi(x) \in \mathcal{N}_{[\,]}(\mathcal{P}_\mathcal{X}(\Phi)\ \epsilon)$，使得以下两个性质成立：

- $\bar{p}_\phi(x) \geq p_\phi(x),\ \forall x \in \mathcal{X}$
- $\|\bar{p}_\phi(x) - p_\phi(x)\|_1 = \int |\bar{p}_\phi(x) - p_\phi(x)|\, dx \leq \epsilon$

注意， $\bar{p}_\phi(x)$ 不必属于 $\mathcal{P}_\mathcal{X}(\Phi)$。括号覆盖数 $N_{[\,]}(\mathcal{P}_\mathcal{X}(\Phi)\ \epsilon)$ 是覆盖 $\mathcal{P}_\mathcal{X}(\Phi)$ 所需的最小 ϵ-括号的基数。熵定义为括号覆盖数的对数。

为了衡量一个函数类的复杂度，我们考虑如下定义的覆盖数和 Rademacher 复杂度。

定义 2.2（ϵ-覆盖与覆盖数）。设 $\mathcal{F}$ 是一个函数类，且 $(\mathcal{F}, \|\cdot\|)$ 是一个度量空间。对于每个 $\epsilon > 0$，函数集 $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|)$ 称为 $\mathcal{F}$ 的一个 ϵ-覆盖，当且仅当对任意 $f \in \mathcal{F}$，都存在一个函数 $g \in \mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|)$，使得 $\|f - g\| \leq \epsilon$。覆盖数 $N(\mathcal{F}, \epsilon, \|\cdot\|)$ 定义为覆盖 $\mathcal{F}$ 所需的最小 ϵ-覆盖的基数。

定义 2.3（拉德马赫复杂度）。假设从定义在集合 $\mathcal{X}$ 上的概率分布 $\mathcal{D}$ 中独立同分布地抽取了 $x_1, \ldots, x_n$。设 $\mathcal{G}$ 是从 $\mathcal{X}$ 映射到 $\mathbb{R}$ 的函数类。$\mathcal{G}$ 的经验拉德马赫复杂度定义如下：

$$\hat{R}_n(\mathcal{G}) := \mathbb{E}_{\{\sigma_i\}_{i=1}^n \sim \mathrm{Unif}\{\pm 1\}} \left[\sup_{g \in \mathcal{G}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \right],$$

其中，$\{\sigma_i\}_{i=1}^n$ 是从拉德马赫分布中抽取的独立随机变量，期望是对 $\{\sigma_i\}_{i=1}^n$ 的随机性取的。$\mathcal{G}$ 的拉德马赫复杂度定义为

$$R_n(\mathcal{G}) := \mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{D}}[\hat{R}_n(\mathcal{G})].$$

3 主要结果

本节首先介绍无监督预训练的一个温和的“信息性”条件。我们证明，该“信息性”条件是预训练能够使下游任务受益的必要条件。接着，我们给出主要结果——算法 1 的无监督预训练和下游任务的统计保证。最后，在第 3.1 节中，我们将结果推广到一个技术性更强但较弱的“信息性”条件版本。事实证明，该版本有助于刻画对比学习的第三个示例（第 6 节）。

信息丰富的预训练任务。我们首先指出，在我们的通用设定下，如果不作进一步假设，那么在最坏情况下，无监督预训练可能完全无法使下游任务受益。

命题 3.1。存在第 2 节中的类 (Φ, Ψ)，使得无论使用何种无监督预训练算法，利用无标签数据进行预训练都无法为学习预测器 g_{ϕ^*, ψ^*} 提供额外信息。

> 考虑潜变量模型 $z = Ax$，其中 $x \sim \mathcal{N}(0, I_d)$，$A \in \Phi$ 是模型的参数。那么，无论我们拥有多少无标签 $\{x_i\}$，都无法从数据中获得关于 A 的任何信息！在这种情况下，无监督预训练对任何下游任务都没有帮助。

因此，提出一个能保证无监督预训练具有信息性的假设至关重要。作为一种思想实验，假设在预训练步骤中，我们找到了一个关于 x, s 边缘分布的精确^密度估计器 ϕ，即对于任意 x, s，$p\hat{}_\phi(x, s) = p_{\phi^*}(x, s)$ 均成立。我们理应预期，该估计量也能充分揭示 x 与 z 之间的关系，即对于任意 x, z，$p\hat{}_\phi(x, z) = p_{\phi^*}(x, z)$ 均成立。遗憾的是，在大多数实际设置中，这一条件并不成立，而且往往过于严格。例如，考虑高斯混合模型，其中 $z \in [K]$ 是数据点 $x \in \mathbb{R}^d$ 所属的簇。在这种情况下，我们无法确保 $p\hat{}_\phi(x, z) = p_{\phi^*}(x, z)$，因为对 z 进行排列不会改变 x 的边缘分布。不过请注意，在许多情况下，对类别标签进行排列并不会

影响下游任务学习。在这些情况下，允许对簇进行排列。受此观察启发，我们提出以下信息性假设，该假设允许下游任务所诱导的某些“变换”：

假设 3.2（κ^{-1}-信息性条件）。我们假设模型类 Φ 相对于变换群 $\mathcal{T}_\Phi$是 κ^{-1}-信息性的。也就是说，对于任意 $\phi \in \Phi$，都存在 $T_1 \in \mathcal{T}_\Phi$ 使得

$$d_{\mathrm{TV}}\big(\mathbb{P}_{T_1\circ\phi}(x,z), \mathbb{P}_{\phi^*}(x,z)\big) \le \kappa \cdot d_{\mathrm{TV}}\big(\mathbb{P}_{\phi}(x,s), \mathbb{P}_{\phi^*}(x,s)\big). \tag{5}$$

这里， ϕ^* 是真实参数。此外，我们假设 $\mathcal{T}_\Phi$ 由 Ψ上的变换群 $\mathcal{T}_\Psi$ 诱导，即对于任意 $T_1 \in \mathcal{T}_\Phi$，都存在 $T_2 \in \mathcal{T}_\Psi$，使得对于任意 $(\phi, \psi) \in \Phi \times \Psi$，

$$\mathbb{P}_{\phi,\psi}(x,y) = \mathbb{P}_{T_1\circ\phi, T_2\circ\psi}(x,y). \tag{6}$$

ˆ在假设 3.2下，如果预训练的 ϕ 能够高精度地估计 x的边缘分布， s那么它也能揭示 x 与表示 z 之间的正确关系（可允许某种变换 $\mathcal{T}_\Phi$，且该变换为下游任务所允许），从而可以使用更少的有标签数据来学习下游任务。

命题3.1 表明，信息性条件对于预训练带来优势是必要的，因为该命题中的反例恰好是 0-信息性的。我们还将证明，这一信息性条件足够宽泛，能够涵盖多种无监督预训练方法，详见第 4、5、6节，包括因子模型、高斯混合模型和对比学习模型。

无监督预训练的保证。 回顾 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi) := \{p_\phi(x, s) \mid \phi \in \Phi\}$。对于算法 1 的最大似然估计步骤（第 2 行），我们有如下保证。

ˆ定理 3.3。 设 ϕ 为 (3) 中定义的最大化者。则以至少 $1-\delta$ 的概率，有

$$d_{\mathrm{TV}}\big(\mathbb{P}_{\hat\phi}(x,s), \mathbb{P}_{\phi^*}(x,s)\big) \le 3\sqrt{\frac{1}{m}\log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \frac{1}{m})}{\delta}},$$

其中 $N_{[\,]}$ 是定义 *2.1* 中的括号覆盖数。

定理 3.3 声称，估计（x， s）联合分布时的 TV 误差会随着$\mathcal{O}(\mathcal{C}_\Phi/m)$ 而减小，其中 m 是无标签数据的数量，而 $\mathcal{C}_\Phi = \log N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ 1/m)$ 衡量学习潜变量模型 Φ 的复杂度。该结果主要来自最大似然估计的标准分析（Van deGeer， 2000年）。为完整起见，我们在附录 A.1 中给出证明。如果模型是 κ^{-1}-informative，则定理 3.3 进一步意味着，以至少 $1-\delta$ 的概率，

$$\min_{\psi} \mathbb{E}_{\mathbb{P}_{\phi^*,\psi^*}}\big[\ell\big(g_{\hat\phi,\psi}(x), y\big)\big] - \mathbb{E}_{\mathbb{P}_{\phi^*,\psi^*}}\big[\ell\big(g_{\phi^*,\psi^*}(x), y\big)\big] \le 12\kappa L\sqrt{\frac{1}{m}\log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}}.$$

详见引理 A.2 。该不等式表明，如果我们使用估计得到的表示学习出一个完美的下游预测器 ˆ ϕ，超额风险就很小。

下游任务学习的保证。 在实践中，我们只能使用少量有标签数据学习一个近似的下游预测器。我们给出算法1 的超额风险上界。

ˆ ˆ定理 3.4。 设 ϕ 和 ψ 分别为算法 *1*的输出。假设损失函数 $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 是 L-有界的，且我们的模型是 κ^{-1}-信息性的。那么，以至少 $1-\delta$的概率，算法 *1* 的超额风险满足以下界：

$$\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) \le 2 \max_{\phi \in \Phi} R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) + 12\kappa L \cdot \sqrt{\frac{1}{m} \log \frac{2N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}} + 2L \cdot \sqrt{\frac{2}{n} \log \frac{4}{\delta}}.$$

这里， $R_n(\cdot)$表示拉德马赫复杂度，并且

$$\ell \circ \mathcal{G}_{\phi,\Psi} := \big\{ \ell\big(g_{\phi,\psi}(x), y\big) : \mathcal{X} \times \mathcal{Y} \to [-L, L] \,\big|\, \psi \in \Psi \big\}.$$

注意，函数类的拉德马赫复杂度可以用其度量熵来界定。因此，我们得到以下推论。

推论 3.5。 在与定理 *3.4*相同的前提下，我们有：

$$\begin{aligned}\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) \le \tilde{c} \max_{\phi \in \Phi} L\sqrt{\frac{\log N(\ell \circ \mathcal{G}_{\phi,\Psi}, L/\sqrt{n}, \|\cdot\|_\infty)}{n}} + 2L\sqrt{\frac{2}{n}\log\frac{4}{\delta}} \\ + 12\kappa L\sqrt{\frac{1}{m}\log\frac{2N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}},\end{aligned}$$

其中， $\tilde{c}$是一个绝对常数， $N(\mathcal{F}, \delta, \|\cdot\|_\infty)$是函数类 $\mathcal{F}$相对于度量 $\|\cdot\|_\infty$的 δ–覆盖数。

根据推论 3.5，我们的算法 1 的超额风险约为 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_\Phi/m} + \sqrt{\mathcal{C}_\Psi/n})$，其中 $\mathcal{C}_\Phi$ 和 $\mathcal{C}_\Psi$ 分别约为类 Φ 的对数括号覆盖数和 Ψ的对数覆盖数。注意，对于仅使用有标签数据学习下游任务的基线算法，其超额风险为 $\tilde{\mathcal{O}}(\sqrt{\mathcal{C}_{\Phi\circ\Psi}/n})$，其中 $\mathcal{C}_{\Phi\circ\Psi}$ 是复合函数类 $\Phi \circ \Psi$的对数覆盖数。在许多实际场景中，例如在预训练深度神经网络上训练线性预测器，复杂度 $\mathcal{C}_{\Phi\circ\Psi}$ 远大于 $\mathcal{C}_\Psi$。我们通常也会拥有远多于有标签数据的无标签数据（$m \gg n$）。在这些场景中，我们的结果严格证明了，相较于不使用无标签数据而直接进行监督学习的基线算法，无监督预训练具有显著优势。

3.1 弱信息模型的保证

我们引入假设 3.2 的一个宽松版本，以便涵盖更广泛的示例。

假设 3.6（κ^{-1}-弱信息条件）。我们假设模型 (Φ, Ψ) 是 κ^{-1}-弱信息的，也就是说，对于任意 $\phi \in \Phi$，存在 $\psi \in \Psi$ 使得

$$d_{\mathrm{TV}}\big(\mathbb{P}_{\phi,\psi}(x,y), \mathbb{P}_{\phi^*,\psi^*}(x,y)\big) \le \kappa \cdot H\big(\mathbb{P}_\phi(x,s), \mathbb{P}_{\phi^*}(x,s)\big). \tag{7}$$

这里我们用 ϕ^*, ψ^* 表示真实参数。

假设 3.6 通过两项修改放宽了假设 3.2 ： (i) 将式（5）左端的量替换为（x, y）的联合分布之间的全变差距离；(ii) 将右端的全变差距离替换为海林格距离。关于这两个假设之间关系的更多内容，请参见附录 A.4.1。

事实上，假设3.6 足以使我们获得与定理3.4相同的理论保证。

T定理 3.7。定理*3.4* 在 κ^{-1}-弱信息假设下仍然成立。

定理3.7 的证明需要比定理3.3更强的最大似然估计保证，以保证海林格距离意义下的接近程度。详细内容见附录A.4。

4 通过因子模型进行预训练

高维数据在现代统计学和机器学习中十分常见，而直接分析高维空间中的数据时，我们往往会遭遇维数灾难。为了解决这一问题，我们通常假设高维数据具有某些低维结构。在这一情境下，因子模型是广受研究的一类模型，它通过数据矩阵中的低秩加稀疏结构对高维测量进行建模，从而去相关协变量。学习这种潜在结构属于无监督统计学习的框架。本节将以因子模型为例，并将线性回归作为下游任务，具体阐述我们的理论框架。我们将严格证明，在这种情况下，无监督预训练如何帮助降低样本复杂度。

模型设定。因子模型（例如参见Lawley & Maxwell，1971年；Bai& Ng，2002年；Forni 等人，2005年；Fan 等人，2021年）广泛应用于金融、计算生物学和社会学领域，在这些领域中，高维测量之间存在很强的相关性。对于潜变量模型，我们考虑由标准高斯分量构成的因子模型，其定义如下。

定义 4.1（因子模型）。假设我们有一个 d维随机向量 x，其依赖关系由 r 个因子 z驱动。因子模型假设

$$x = B^* z + \mu,$$

其中 B^* 是一个 $d \times r$ 因子载荷矩阵。这里， $\mu \sim N(0, I_d)$ 是与共同因子 $z \sim N(0, I_r)$ 不相关的特质分量。我们假设真实参数$B^* \in \mathcal{B}$，其中 $\mathcal{B} := \{B \in \mathbb{R}^{d\times r} \mid \|B\|_2 \le D\}$ ，对于某个 $D > 0$。

对于下游任务，我们假设潜在因子 z 对响应变量 y 的影响方式与对 x 的影响方式相似，因此考虑以下线性回归问题

$$y = \beta^{*T} z + \nu,$$

其中， $\nu \sim N(0, \varepsilon^2)$ 是高斯噪声，与因子 z 和特质分量 μ 不相关。我们假设真实参数 $\beta^* \in \mathcal{C}$ 满足 $\mathcal{C} := \{\beta \in \mathbb{R}^r \mid \|\beta\|_2 \le D\}$ ，其中 $D > 0$。因此，潜变量模型（即 Φ）和预测类（即 Ψ）分别表示为 $\mathcal{B}$ 和 $\mathcal{C}$。在下文中，我们考虑没有辅助信息可用的情形，即我们只能获取独立同分布的无标签数据 $\{x_i\}_{i=1}^m$ 和独立同分布的有标签数据 $\{x_j, y_j\}_{j=1}^n$。

对于回归模型，自然会考虑平方损失函数 $\ell(x,y) := (y-x)^2$。那么，在分布 $\mathbb{P}_{B,\ \beta}$ 下，最优预测器 $g_{B,\ \beta}$ 具有以下闭式解，

$$g_{B,\beta}(x) = \mathbb{E}_{\mathbb{P}_{B,\beta}}[y \mid x] = \beta^T B^T (BB^T + I_d)^{-1} x.$$

现在，超额风险定义为

$$\mathrm{Error}_\ell(\hat{B},\hat{\beta}) := \mathbb{E}_{\mathbb{P}_{B^*,\beta^*}}\left[\left(y - g_{\hat{B},\hat{\beta}}(x)\right)^2\right] - \mathbb{E}_{\mathbb{P}_{B^*,\beta^*}}\left[\left(y - g_{B^*,\beta^*}(x)\right)^2\right].$$

信息性条件。我们首先证明，对于以线性回归为下游任务的因子模型，假设3.2 成立。因子模型的思路是学习一个低维表示 z，其中允许对 z 进行旋转，因为在下游任务中，我们也可以旋转β 以适配旋转后的 z。

引理 4.2. 以线性回归为下游任务的因子模型是 κ^{-1}-信息性的，其中

$$\kappa = \frac{c_1(\sigma^*_{\max}+1)^4}{(\sigma^*_{\min})^3}.$$

这里， c_1 是某些绝对常数， $\sigma^*_{\max}$ 和 $\sigma^*_{\min}$ 分别是B^*的最大和最小奇异值。

理论结果。回顾定理3.4，我们假设损失函数有 L界，以保证算法1的性能。因此，我们不直接将算法1 应用于平方损失函数，而是考虑使用截断平方损失的算法1 ，即

$$\tilde{\ell}(x,y) := (y-x)^2 \cdot \mathbb{1}_{\{(y-x)^2 \le L\}} + L \cdot \mathbb{1}_{\{(y-x)^2 > L\}}. \tag{8}$$

这里， L 是经过仔细选择的截断水平。具体而言，在第一阶段，我们仍然使用ˆ MLE来学习一个估计量 B ，与算法1第 2 行中的做法相同。在第二阶段，我们对截断平方损失应用 ERM ˆ，学习一个估计量 β，即

$$\hat{\beta} \leftarrow \arg\min_{\beta \in \mathcal{C}} \sum_{j=1}^{n} \tilde{\ell}\left(g_{\hat{B},\beta}(x_j), y_j\right).$$

接下来我们给出如下理论保证。

定理 4.3. 我们考虑采用截断平方损失的算法 *1* (8) 其中 $L = (D^2+1)^3 \log n$。 ˆ ˆ设 B, β 为算法 *1*的输出。则对于以下游任务为线性回归的因子模型，以至少 $1-\delta$的概率，超额风险可界定如下：

$$\mathrm{Error}_\ell(\hat{B},\hat{\beta}) \le \tilde{\mathcal{O}}\left(\kappa L\sqrt{\frac{dr}{m}} + L\sqrt{\frac{r}{n}}\right),$$

˜其中 D 定义于集合 $\mathcal{B}$ 和 $\mathcal{C}$中，且 κ 在引理 *4.2*中给出。此处 $\mathcal{O}(\cdot)$ 省略了 m、d、r、D、$1/\delta$ 中的绝对常数和多对数因子。

注意，我们在定理4.3 中得到的收敛率对于这一特定任务而言并非最优：由于平方损失的性质，如果我们考虑从 d–维到 x 维的 y线性回归，并使用 n 数据，我们 ˜通常可以达到快速收敛率，此时超额风险按 $\mathcal{O}(d/n)$ 的速度下降。为弥合这一差距，我们考虑算法 1 $\Phi = \mathbb{R}^{d\times r}$，其中 $\Psi = \mathbb{R}^r$，并记 D 为：= 最大值$\{\|B^*\|_2, \|\beta^*\|_2\}$。通过采用比一致集中技术更精细的分析（后者适用于一般问题，但对于这一特定任务并非最优），我们得到了以下具有更快风险收敛率的理论保证：

定理 4.4 (快速收敛率)。令 ˆB、ˆβ 为算法 1的输出。那么，若 $m \gtrsim (D^2+1)^2 d\log(1/\delta)$，$n \gtrsim (D^2+1)^2 r\log(1/\delta)$，对于以线性回归为下游任务的因子模型，以至少 $1-\delta$ 的概率，超额风险可界定如下：

$$\mathrm{Error}_\ell(\hat{B},\hat{\beta}) \le \mathcal{O}\left((D^2+1)^6(D^4+\sigma_{\min}^{*-4})\frac{d\log(1/\delta)}{m} + (D^2+1)^2\frac{r\log(4/\delta)}{n}\right).$$

这里， $\mathcal{O}(\cdot)$省略了一些绝对常数。

定理 4.4 展示了无监督预训练在以下意义上的优势。假设 D˜和 $\sigma^*_{\min}$ 均为常数。学习载荷矩阵所付出的代价为 $\mathcal{O}(d/m)$，当 m 非常大时，这一代价很小。注意，由于 x 是一个 d 维向量，通常的线性回归˜的风险为 $\mathcal{O}(d/n)$。在定理 4.4 给出的风险界中，与 n 相关的风险按 ˜缩放为 $\mathcal{O}(r/n)$。通常假设因子相对于输入是低维的（$d \gg r$）。˜ ˜那么当 $m \gg n$ 时，风险界 $\mathcal{O}(d/m + r/n)$远优于 $\mathcal{O}(d/n)$。

5 通过高斯混合模型进行预训练

在本节中，我们将在理论框架下展示，使用高斯混合模型（GMM）进行预训练如何有助于下游分类任务。

模型设置。对于潜变量模型，我们考虑一个 d 维、包含 K 个分量且权重相等的高斯混合模型。具体而言，表示簇的潜变量 z 从 $[K]$ 中均匀采样。在每个簇中，数据均从标准高斯分布中采样，即对于任意 $i \in [K]$， $x|z=i \sim \mathcal{N}(u_i^*, I_d)$。于是可得

$$x \sim \sum_{i=1}^{K} \frac{1}{K}\mathcal{N}(u_i^*, I_d).$$

我们用 $\mathcal{U}$ 表示参数空间，其中每个元素由 K 个中心（d 维向量）组成。

我们假设参数集合 $\mathcal{U}$ 满足归一化条件——存在 $D>0$，使得对于任意 $\mathbf{u} = \{u_i\}_{i=1}^K \in \mathcal{U}$，都有 $\|u_i\|_2 \le D\sqrt{d\log K}$，$\forall i \in [K]$。我们进一步假设真实中心 $\{u_i^*\}_{i=1}^K \in \mathcal{U}$ 满足以下分离条件。

假设5.1（分离条件）。真实参数 $\{u_i^*\}_{i=1}^K \in \mathcal{U}$ 满足

$$\|u_i^* - u_j^*\|_2 \ge 100\sqrt{d\log K},\ \forall i \ne j.$$

对于下游任务，我们考虑标签为 $y \in \{0,1\}$的二元分类问题。我们将 Ψ 记为满足以下条件的 2^K 分类器集合：对于每个 $\psi \in \Psi$及任意 $i \in [K]$，要么有$\mathbb{P}_\psi(y=1|z=i) = 1-\varepsilon$，要么有 $\mathbb{P}_\psi(y=0|z=i) = 1-\varepsilon$，其中 ε 表示噪声。于是，潜变量模型和预测类分别表示为 $\mathcal{U}$ 和 Ψ。下文中，我们考虑没有辅助信息可用的情形，即我们只能获得独立同分布的无标签数据 $\{x_i\}_{i=1}^m$ 和独立同分布的有标签数据 $\{x_j, y_j\}_{j=1}^n$。对于分类问题，自然可以考虑 $0-1$ 损失函数 $\ell(x,y) := \mathbb{1}_{\{x \neq y\}}$，其上界为1。

信息性条件。 我们证明，对于上述模型，假设 3.2 成立。我们有如下保证。

~引理 5.2。 设 $\mathcal{U} = \{\mathbf{u} \in \mathcal{U} \mid d_{\mathrm{TV}}(p_{\mathbf{u}}(x)\ p_{\mathbf{u}^*}(x)) \leq 1/(4K)\}$。在假设 5.1下，具有~参数的 $\mathcal{U}$ 是 $\mathcal{O}(1)$- 信息性的，这里的变换群由下游分类任务诱导。

理论结果 我们有如下理论保证。

^定理5.3。 设 ^$\mathbf{u}$ ψ 是算法 1的输出。假设 5.1 成立，并且$m = \tilde{\Omega}(dK^3)$。那么，对于以分类为下游任务的高斯混合模型，以至少 $1-\delta$的概率，超额风险可如下界定：

$$\mathrm{Error}_\ell(\hat{\mathbf{u}}, \hat{\psi}) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{dK}{m}} + \sqrt{\frac{K}{n}}\right),$$

~此处 $\mathcal{O}(\cdot)$省略了 m、 d、 K、 D、 $1/\delta$中的一些常数和多对数因子。

定理 5.3 在以下意义上展示了此设置下无监督预训练的优势：注意，高斯混合模型的参数数量为 dK，因此，如果不进行无监督预训练而直接进行分类，风险将按 $\tilde{\mathcal{O}}(\sqrt{dK/n})$ 的比例增长。当 d 较大且 $m \gg n$ 时，我们可以获得比仅使用有标签数据的监督学习更优的风险界。

6 通过对比学习进行预训练

对于人类而言，看到许多不同动物的图片时，即使对这些动物一无所知，我们也能推断出哪些图片展示的是同一种动物。在这一过程中，我们不知不觉地学会了每张图片的一种表示，用来捕捉不同图片之间的相似性。对比学习模拟了人类的学习方式。具体来说，对比学习基于正负样本对，将数据嵌入到某个空间中，使相似的样本对彼此接近，而不相似的样本对彼此远离。本节将在我们的理论框架下，说明预训练（学习嵌入函数）如何使下游线性回归任务受益。

模型设置。在对比学习的设定中，我们假设 x 和 x' 独立地从相同的分布 $\mathbb{P}(x)$ 中采样。 x 和 x' 之间的相似性由一个

表示函数 f_{θ^*}： $\mathcal{X} \to \mathbb{R}^r$ 以如下方式捕捉：

$$\mathbb{P}(t=1 \mid x, x') = \frac{1}{1+e^{-f_{\theta^*}(x)^T f_{\theta^*}(x')}},$$
$$\mathbb{P}(t=-1 \mid x, x') = \frac{1}{1+e^{f_{\theta^*}(x)^T f_{\theta^*}(x')}}.$$

这里， t 是一个随机变量，用于标记 x 与 x'之间的相似性。如果数据对（x, x'）相似，那么 t 趋于 1。如果数据对（x, x'）不相似（负样本），那么 t 趋于 -1。我们假设（x, x', t） $\sim \mathbb{P}_{f_{\theta^*}}$（$x, x', t$）。这里，（$x', t$）可以视为辅助信息。潜变量 z定义为 $z := f_{\theta^*}(x) + \mu$，其中 $\mu \sim \mathcal{N}(0, I_r)$是与 x不相关的高斯噪声。我们将（x, z）记为 $\sim \mathbb{P}_{f_{\theta^*}}$（$x, z$）。

对于下游任务，我们考虑以下线性回归问题

$$y = \beta^{*T} z + \nu,$$

其中， $\nu \sim \mathcal{N}(0, 1)$是高斯噪声。我们假设真实参数 $\theta^* \in \Theta$和 $\beta^* \in \mathcal{B}$满足标准归一化假设，即对于任意 $\theta \in \Theta$ 和 $x \in \mathcal{X}$， $\|f_\theta(x)\|_2 \le 1$ ，并且对于任意 $\beta \in \mathcal{B}$， $\|\beta\|_2 \le D$ 。我们可以获取独立同分布的无标签数据 $\{x_i, x_i', t_i\}_{i=1}^m$ 和独立同分布的有标签数据 $\{x_j, y_j\}_{j=1}^n$。这里， (x_i', t_i) 是与 x_i对应的辅助信息。

接下来，我们考虑平方损失函数 $\ell(x, y) := (y-x)^2$。我们采用与 (8) 中相同形式的截断平方损失。

弱信息条件。 我们首先证明上述模型满足假设 3.6：

引理 6.1. 以线性回归作为下游任务的对比学习是 κ^{-1}-弱信息的，其中

$$\kappa = c_3 \cdot \sqrt{\frac{1}{\sigma_{\min}(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T])}}.$$

这里， c_3 是一个绝对常数。

理论结果。我们定义一个密度函数集合 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta) := \{p_{f_\theta}(x, x', t) \mid \theta \in \Theta\}$。于是我们得到如下理论保证。

定理 6.2. 我们考虑算法 1 采用截断平方损失 (8) 其中 $L = 36(D^2+1)\log n$。 ˆ ˆ令 θ, β 为算法 1的输出。那么，对于以线性回归作为下游任务的对比学习，以至少 $1-\delta$ 的概率，超额风险可界定如下：

$$\mathrm{Error}_\ell(\hat{\theta}, \hat{\beta}) \le \tilde{\mathcal{O}}\left(\kappa L \sqrt{\frac{\log N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta), 1/m^2)}{m}} + L\sqrt{\frac{1}{n}}\right),$$

其中 $L = 36(D^2+1)\log n$ 且 κ 在引理 6.1中给出。这里 $\tilde{\mathcal{O}}(\cdot)$ 省略了 $1/\delta$中的一些常数和多对数因子。

请注意，直接使用有标签数据进行训练时，其超额风险很大程度上取决于函数类的复杂度 $\mathcal{F}_\theta$。在 $m \gg n$ 的情况下，定理 6.2 的超额风险按$\tilde{O}(\sqrt{1/n})$ 的规模变化；当 $\mathcal{F}_\theta$ 的复杂度相当高时，这优于纯监督学习。因此，无监督预训练对比学习的效用便显现出来。

7 结论

本文提出了一个通用的理论框架，用于解释无监督预训练在统计上的益处。我们研究了这样一种自然方案：使用最大似然估计进行无监督预训练，并使用经验风险最小化进行下游任务学习。我们提出了一个自然的“信息性”条件；在该条件下，我们的算法能够达到一个超额风险界，在典型的实际应用场景中显著优于纯监督学习所达到的基线。我们进一步以三种具体的无监督预训练方法实例化了这一理论框架，并给出了相应的保证。

参考文献

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe：低秩 mdps 的结构复杂度与表征学习。 神经信息处理系统进展, 33:20095–20107, 2020年。 Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. 对比无监督表示学习的理论分析。 arXiv 预印本 arXiv:1902.09229, 2019年。 Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0：用于自监督学习语音表示的框架。 神经信息处理系统进展, 33: 12449–12460, 2020年。 Bai, J. and Ng, S. 确定近似因子模型中的因子数量。 *Econometrica*, 70(1):191–221, 2002年。 Baxter, J. 归纳偏置学习模型。 人工智能研究杂志, 12: 149–198, 2000年3月。 doi: 10.1613/jair.731. URL https://doi.org/10.1613%2Fjair.731。 Belkin, M., Niyogi, P., and Sindhwani, V. 流形正则化：从有标签和无标签样本中学习的几何框架。 《机器学习研究期刊》, 7(11), 2006年。 Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., and Mercer, R. L. 自然语言的基于类别的 n 元语法模型。 计算语言学, 18(4):467–480, 1992年。 Caron, M., Bojanowski, P., Mairal, J., and Joulin, A. 在未经筛选的数据上对图像特征进行无监督预训练。载于 IEEE/CVF 国际计算机视觉会议, pp. 2959–2968, 2019年。 Chen, Y., Chi, Y., Fan, J., and Ma, C. 2021年。 Dai, Z., Cai, B., Lin, Y., and Chen, J. Up-DETR：利用 Transformer 进行目标检测的无监督预训练。载于IEEE/CVF 计算机视觉与模式识别会议, pp. 1601–1610, 2021年。 Davis, C. and Kahan, W. M. 扰动导致特征向量旋转。iii. SIAM 数值分析杂志, 7(1):1–46, 1970年。 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT：用于语言理解的深度双向 Transformer 预训练。 arXiv 预印本 arXiv:1810.04805, 2018年。 Devroye, L., Mehrabian, A., and Reddad, T. 高维高斯分布之间的全变差距离。 arXiv 预印本 arXiv:1810.08693, 2018年。 Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. 通过学习表示实现可证明的小样本学习。 arXiv 预印本 arXiv:2002.09434, 2020年。 Erhan, D., Courville, A., Bengio, Y., and Vincent, P. 为什么无监督预训练有助于深度学习？ 载于 第十三届人工智能与统计学国际会议论文集, pp. 201–208.JMLR Workshop and Conference Proceedings, 2010.

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. 用于统计机器学习的稳健高维因子模型。《统计科学：数理统计学会评论期刊》, 36(2):303, 2021年。 Forni, M., Hallin, M., Lippi, M., and Reichlin, L. 广义动态因子模型：单边估计与预测。《美国统计学会会刊》, 100(471):830–840, 2005年。 HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. 基于谱对比损失的自监督深度学习的可证明保证。神经信息处理系统进展 , 34: 5000–5011, 2021 年。 Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. 关于随机向量次高斯范数集中不等式的简短说明。 arXiv 预印本 arXiv:1902.03736, 2019年。 Joachims, T. et al. 使用支持向量机进行文本分类的传导推断。载于国际机器学习大会, volume 99, pp. 200–209, 1999. Lawley, D. N. and Maxwell, A. E. 因子分析作为一种统计方法。1971年。 Lawrence, N. and Jordan, M. 通过高斯过程进行半监督学习。 神经信息处理系统进展,17, 2004.Ledoux, M. and Talagrand,M. 《巴拿赫空间中的概率：等周理论与过程》。施普林格科学与商业媒体，2013年。 Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. 预测已知内容大有裨益：可证明的自监督学习。 神经信息处理系统进展, 34:309–323, 2021年。 Liu, Q., Chung, A., Szepesvari, C., and Jin, C. 部分可观测强化学习何时不再令人望而生畏？载于 Loh, P.-L. and Raginsky, M. (编), 第三十五届学习理论会议论文集, 《机器学习研究论文集》第178卷， 《机器学习研究论文集》, pp. 5175–5220. PMLR, 2022年7月02–05日。 Ma, C., Wang, K., Chi, Y., and Chen, Y. 非凸统计估计中的隐式正则化：梯度下降在线性收敛下实现相位恢复和矩阵补全。载于 国际机器学习大会, pp. 3345–3354. PMLR, 2018年。 Marshall, A. W., Olkin, I., and Arnold, B. C. 不等式：主序理论及其应用, 第143卷。施普林格，第二版，2011年。 doi: 10.1007/978-0-387-68276-1。 Maurer, A., Pontil, M., and Romera-Paredes, B. 多任务表征学习的益处。《机器学习研究期刊》, 17(81):1–32, 2016年。 URLhttp://jmlr.org/papers/v17/15-242.html。 Melamud, O., Goldberger, J., and Dagan, I. context2vec：利用双向长短期记忆网络学习通用上下文嵌入。载于 第20届 SIGNLL 计算自然语言学习会议论文集, pp. 51–61, 2016年。 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 词语和短语的分布式表示及其组合性。 神经信息处理系统进展, 26, 2013年。

Peter, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., 和 Zettlemoyer, L. 深度上下文化词表示。arXiv 预印本 arXiv:1802.05365，2018年。Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 等。通过生成式预训练提升语言理解。2018年。Ratsaby, J. 和 Venkatesh, S. S. 利用参数化侧信息从有标签和无标签样本混合数据中学习。载于第八届计算学习理论年度会议论文集，第 412–417 页，1995年。Saunshi, N., Malladi, S., 和 Arora, S. 对语言模型为何有助于解决下游任务的数学探索。arXiv 预印本 arXiv:2010.03648，2020年。Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., 和 Krishnamurthy, A. 理解对比学习需要纳入归纳偏置。arXiv 预印本 arXiv:2202.14037，2022年。Schmitt, B. A. 矩阵平方根和勾股和的扰动界。《线性代数及其应用》，174:215–227，1992年。Schneider, S., Baevski, A., Collobert, R., 和 Auli, M. wav2vec：用于语音识别的无监督预训练。arXiv 预印本 arXiv:1904.05862，2019年。Song, K., Tan, X., Qin, T., Lu, J., 和 Liu, T.-Y. MASS：用于语言生成的掩码序列到序列预训练。arXiv 预印本 arXiv:1905.02450，2019年。Szummer, M. 和 Jaakkola, T. 利用部分有标签数据进行信息正则化。《神经信息处理系统进展》，15，2002年。Tosh, C., Krishnamurthy, A., 和 Hsu, D. 对比学习、多视图冗余与线性模型。载于《算法学习理论》，第 1179–1206 页。PMLR，2021年（a）。Tosh, C., Krishnamurthy, A., 和 Hsu, D. 对比估计揭示了线性模型中的主题后验信息。《机器学习研究杂志》，22:281–1，2021年（b）。Tripuraneni, N., Jordan, M., 和 Jin, C. 迁移学习理论：任务多样性的重要性。《神经信息处理系统进展》，33:7852–7862，2020年。Tripuraneni, N., Jin, C., 和 Jordan, M. 线性表示的可证明元学习。载于《国际机器学习大会》，第 10434–10443 页。PMLR，2021年。Van de Geer, S. 《M估计中的经验过程》，第 6 卷。剑桥大学出版社，2000年。Vershynin, R. 《高维概率：数据科学应用导论》，第 47 卷。剑桥大学出版社，2018年。Wainwright, M. J. 《高维统计：非渐近视角》，第 48 卷。剑桥大学出版社，2019年。Wei, C., Xie, S. M., 和 Ma, T. 预训练语言模型为何有助于下游任务？对头部与提示调优的分析。《神经信息处理系统进展》，34: 16158–16170，2021年。

Zhang, T. 从 ε-熵到 KL-熵：最小信息复杂度密度估计分析。《统计学年鉴》，34(5)，2006年10月。

Zhu, X. J. 半监督学习文献综述。2005年。

A 第 3节的证明

在 A.1节中，我们证明定理 3.3，该定理为算法 1中的最大似然估计步骤给出了全变差距离保证。我们的证明受到 Van de Geer（2000年）；Zhang（2006年）的启发，并在很大程度上遵循 Agarwal等人（2020年）；Liu等人（2022年）的工作。在 A.2节中，我们证明定理 3.4，该定理通过给超额风险设定上界来保证算法 1的性能。证明依赖于这样一个事实：有标签数据 $\{x_j, y_j\}_{j=1}^n$ 与无标签数据 $\{x_i, s_i\}_{i=1}^m$ 相互独立。在 A.3节中，我们基于高斯复杂度分析证明推论 3.5。在 A.4节中，我们通过首先证明算法 1中的最大似然估计步骤实际上保证了海林格距离的一个上界，进而证明定理 3.7；这一保证比定理 3.3中提到的全变差距离保证更强。

A.1 定理 3.3的证明

接下来，我们证明定理 3.3。

定理 3.3 的证明。为简化记号，我们记作 $\boldsymbol{x} := (x, s)$。回顾一下，我们定义 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi) := \{p_\phi(x, s) \mid \phi \in \Phi\}$。令 $\mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)$ 为 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)$ 的最小 ϵ-括号。我们有 $|\mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)| = N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)$，其中 $N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)$ 是 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)$ 的括号覆盖数。根据马尔可夫不等式和布尔不等式，以至少 $1-\delta$ 的概率，对所有 $\bar{p}_\phi(\boldsymbol{x})$ 均有 $\in \mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)$

$$\frac{1}{2}\sum_{i=1}^{m}\log\frac{\bar{p}_\phi(\boldsymbol{x}_i)}{p_{\phi^*}(\boldsymbol{x}_i)} \le \log\mathbb{E}\Big[e^{\frac{1}{2}\sum_{i=1}^{m}\log\frac{\bar{p}_\phi(\boldsymbol{x}_i)}{p_{\phi^*}(\boldsymbol{x}_i)}}\Big] + \log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta}. \tag{9}$$

ˆ注意，ϕ 是似然函数的最大化者，即

$$\hat{\phi} \leftarrow \arg\max_{\phi\in\Phi}\sum_{i=1}^{m}\log p_\phi(\boldsymbol{x}_i),$$

这意味着

$$\frac{1}{2}\sum_{i=1}^{m}\log\frac{\bar{p}_{\hat{\phi}}(\boldsymbol{x}_i)}{p_{\phi^*}(\boldsymbol{x}_i)} \ge 0. \tag{10}$$

那么，我们以至少为 $1-\delta$ 的概率有

$$\begin{aligned}
0 &\le \log\mathbb{E}\Big[e^{\frac{1}{2}\sum_{i=1}^{m}\log\frac{\bar{p}_{\hat{\phi}}(\boldsymbol{x}_i)}{p_{\phi^*}(\boldsymbol{x}_i)}}\Big] + \log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta},\\
&= m\log\mathbb{E}\left[\sqrt{\frac{\bar{p}_{\hat{\phi}}(\boldsymbol{x})}{p_{\phi^*}(\boldsymbol{x})}}\right] + \log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta},\\
&= m\log\int\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})p_{\phi^*}(\boldsymbol{x})}\,d\boldsymbol{x} + \log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta},\\
&\le m\left(\int\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})p_{\phi^*}(\boldsymbol{x})}\,d\boldsymbol{x} - 1\right) + \log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta},
\end{aligned} \tag{11}$$

其中，最后一个不等式源于以下事实：$\log x \leq x - 1$。重新排列各项可得

$$1 - \int \sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} \leq \frac{1}{m} \log \frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon)}{\delta}. \tag{12}$$

根据括号的定义，我们得到

$$\int \bar{p}_{\hat{\phi}}(\boldsymbol{x}) d\boldsymbol{x} = \int (\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\hat{\phi}}(\boldsymbol{x})) d\boldsymbol{x} + \int p_{\hat{\phi}}(\boldsymbol{x}) d\boldsymbol{x} \leq \epsilon + 1,$$

这意味着

$$\int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \leq 2\left(1 - \int \sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})} d\boldsymbol{x}\right) + \epsilon \tag{13}$$

and

$$\int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} + \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \leq 2 \int \bar{p}_{\hat{\phi}}(\boldsymbol{x}) + p_{\phi^*}(\boldsymbol{x})\, d\boldsymbol{x} \leq 2\epsilon + 4. \tag{14}$$

结合（12）和（13），我们得到

$$\int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \leq \frac{2}{m} \log \frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon)}{\delta} + \epsilon. \tag{15}$$

根据柯西-施瓦茨不等式，可得

$$\begin{aligned}
\left(\int |\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right)^2 &\leq \int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} + \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \cdot \int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x}, \\
&\leq (2\epsilon + 4) \cdot \left(\frac{2}{m} \log \frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon)}{\delta} + \epsilon\right),
\end{aligned} \tag{16}$$

其中最后一个不等式由（14）和（15）得出。注意

$$\begin{aligned}
&\left(\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right)^2 - \left(\int |\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right)^2 \\
&= \left(\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})| + |\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right) \cdot \left(\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})| - |\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right) \\
&\leq \left(\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})| + |\bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right) \cdot \int |p_{\hat{\phi}}(\boldsymbol{x}) - \bar{p}_{\hat{\phi}}(\boldsymbol{x})|\, d\boldsymbol{x} \\
&\leq (\epsilon + 4) \cdot \epsilon.
\end{aligned} \tag{17}$$

将（16）和（17）相加，可得

$$\left(\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})|\, d\boldsymbol{x}\right)^2 \leq (2\epsilon + 4) \cdot \left(\frac{2}{m} \log \frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon)}{\delta} + \epsilon\right) + (\epsilon + 4) \cdot \epsilon, \tag{18}$$

这意味着

$$d_{\mathrm{TV}}\left(\mathbb{P}_{\hat{\phi}}(\boldsymbol{x}), \mathbb{P}_{\phi^*}(\boldsymbol{x})\right) = \frac{1}{2}\int \left|p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})\right| d\boldsymbol{x}$$
$$\leq \frac{1}{2}\sqrt{(2\epsilon+4)\cdot\left(\frac{2}{m}\log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),\epsilon)}{\delta}+\epsilon\right)+(\epsilon+4)\cdot\epsilon}. \tag{19}$$

设定 $\epsilon = 1/m$，我们以至少 $1-\delta$ 的概率有

$$d_{\mathrm{TV}}\left(\mathbb{P}_{\hat{\phi}}(\boldsymbol{x}), \mathbb{P}_{\phi^*}(\boldsymbol{x})\right) \leq \frac{1}{2}\sqrt{\left(\frac{2}{m}+4\right)\cdot\left(\frac{2}{m}\log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m)}{\delta}+\frac{1}{m}\right)+\left(\frac{1}{m}+4\right)\cdot\frac{1}{m}}$$
$$\leq 3\cdot\sqrt{\frac{1}{m}\log\frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m)}{\delta}}. \tag{20}$$

因此，我们证明定理3.3。 □

A.2 定理 3.4的证明

在证明该定理之前，我们先给出一些将在定理3.4的证明中用到的有用结果。引理 A.1 通过应用有界差分不等式和标准的对称化论证，对经验损失与总体损失之差给出了上界。引理 A.2 将超额风险与概率分布之间的全变差距离联系起来。为简化记号，以下我们用 $\mathbb{E}_{(x,y)\sim\mathbb{P}_{\phi,\ \psi}(x,\ y)}$ 表示 $\mathbb{E}_{\phi,\ \psi}$。我们进一步用$\mathbb{E}$ 表示对真实参数取期望，即 $\mathbb{E} := \mathbb{E}_{(x,y)\sim\mathbb{P}_{\phi^*,\ \psi^*}(x,\ y)}$。

引理 A.1. 假设 $\ell(\cdot,\cdot)$ 是一个 L-有界损失函数。对于任意给定的 $\phi\in\Phi$，以至少 $1-\delta$ 的概率，

$$\sup_{\psi\in\Psi}\left|\mathbb{E}[\ell(g_{\phi,\psi}(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\phi,\psi}(x_j),y_j)\right| \leq R_n(\ell\circ\mathcal{G}_{\phi,\Psi}) + L\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{21}$$

其中 $R_n(\ell\circ\mathcal{G}_{\phi,\ \Psi})$ 是函数类 $\ell\circ\mathcal{G}_{\phi,\ \Psi}$ 在定理3.4中定义的拉德马赫复杂度。

引理 A.1 的证明。 首先注意，当一对 $(x_j,\ y_j)$ 发生变化时，由于 ℓ 是 L-有界的，随机变量

$$\sup_{\psi\in\Psi}\left(\mathbb{E}[\ell(g_{\phi,\psi}(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\phi,\psi}(x_j),y_j)\right) \tag{22}$$

变化幅度不超过 $2L/n$。麦克迪亚米德不等式表明，概率至少为$1-\delta/2$时，

$$\sup_{\psi\in\Psi}\left(\mathbb{E}[\ell(g_{\phi,\psi}(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\phi,\psi}(x_j),y_j)\right)$$
$$\leq \mathbb{E}\left[\sup_{\psi\in\Psi}\left(\mathbb{E}[\ell(g_{\phi,\psi}(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\phi,\psi}(x_j),y_j)\right)\right] + L\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{23}$$

令 $\{x'_j, y'_j\}_{j=1}^n$ 为 $\{x_j, y_j\}_{j=1}^n$ 的独立副本，令 $\{\sigma_j\}_{j=1}^n$ 为独立同分布的 Rademacher 随机变量。使用标准的对称化技巧，我们有

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\psi\in\Psi}\left(\mathbb{E}[\ell(g_{\phi,\psi}(x),y)]-\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x_j),y_j)\right)\right]\\
&=\mathbb{E}\left[\sup_{\psi\in\Psi}\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x'_j),y'_j)-\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x_j),y_j)\middle|\{x_j,y_j\}_{j=1}^n\right]\right]\\
&\le\mathbb{E}\left[\sup_{\psi\in\Psi}\left(\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x'_j),y'_j)-\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x_j),y_j)\right)\right]\\
&\le\mathbb{E}\left[\sup_{\psi\in\Psi}\frac{1}{n}\sum_{j=1}^n\sigma_j\left(\ell(g_{\phi,\psi}(x'_j),y'_j)-\ell(g_{\phi,\psi}(x_j),y_j)\right)\right]\\
&\le 2\mathbb{E}\left[\sup_{\psi\in\Psi}\frac{1}{n}\sum_{j=1}^n\sigma_j\ell(g_{\phi,\psi}(x_j),y_j)\right]\\
&=R_n(\ell\circ\mathcal{G}_{\phi,\Psi}).
\end{aligned}\tag{24}$$

因此，以至少 $1-\delta/2$ 的概率，

$$\sup_{\psi\in\Psi}\left(\mathbb{E}[\ell(g_{\phi,\psi}(x),y)]-\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x_j),y_j)\right)\le R_n(\ell\circ\mathcal{G}_{\phi,\Psi})+L\sqrt{\frac{2\log(2/\delta)}{n}}\tag{25}$$

同理，以至少 $1-\delta/2$ 的概率，

$$\sup_{\psi\in\Psi}\left(\frac{1}{n}\sum_{j=1}^n\ell(g_{\phi,\psi}(x_j),y_j)-\mathbb{E}[\ell(g_{\phi,\psi}(x),y)]\right)\le R_n(\ell\circ\mathcal{G}_{\phi,\Psi})+L\sqrt{\frac{2\log(2/\delta)}{n}}\tag{26}$$

结合以上结果，我们证明引理 A.1。 □

引理 A.2. 假设 $\ell(\cdot,\cdot)$ 是一个 L-有界损失函数。那么，对于任意 $\phi\in\Phi,\psi\in\Psi$，有

$$\mathbb{E}[\ell(g_{\phi,\psi}(x),y)]-\mathbb{E}[\ell(g_{\phi^*,\psi^*}(x),y)]\le 4L\cdot d_{\mathrm{TV}}(\mathbb{P}_{\phi,\psi}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y)).\tag{27}$$

引理 *A.2* 的证明。

$$\begin{aligned}
&\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x),y)]-\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)]\\
&=\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x),y)]-\mathbb{E}_{\phi,\psi}[\ell(g_{\phi,\psi}(x),y)]\\
&\quad+\mathbb{E}_{\phi,\psi}[\ell(g_{\phi,\psi}(x),y)]-\mathbb{E}_{\phi,\psi}[\ell(g_{\phi^*,\psi^*}(x),y)]\\
&\quad+\mathbb{E}_{\phi,\psi}[\ell(g_{\phi^*,\psi^*}(x),y)]-\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)].
\end{aligned}\tag{28}$$

首先注意到，根据 $g_{\phi,\ \psi}$ 的定义，

$$\mathbb{E}_{\phi,\psi}[\ell(g_{\phi,\psi}(x),y)]-\mathbb{E}_{\phi,\psi}[\ell(g_{\phi^*,\psi^*}(x),y)]\le 0.\tag{29}$$

对于另外两项，根据 ℓ 是 L-有界的这一事实，我们有

$$
\begin{aligned}
&|\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x),y)] - \mathbb{E}_{\phi,\psi}[\ell(g_{\phi,\psi}(x),y)]| \\
&= \left| \int \ell(g_{\phi,\psi}(x),y) p_{\phi_*,\psi_*}(x,y) \mathrm{d}x\mathrm{d}y - \int \ell(g_{\phi,\psi}(x),y) p_{\phi,\psi}(x,y) \mathrm{d}x\mathrm{d}y \right| \\
&= \left| \int \ell(g_{\phi,\psi}(x),y)(p_{\phi_*,\psi_*}(x,y) - p_{\phi,\psi}(x,y)) \mathrm{d}x\mathrm{d}y \right| \\
&\leq \int |\ell(g_{\phi,\psi}(x),y)||(p_{\phi_*,\psi_*}(x,y) - p_{\phi,\psi}(x,y))| \mathrm{d}x\mathrm{d}y \\
&\leq \int L|(p_{\phi_*,\psi_*}(x,y) - p_{\phi,\psi}(x,y))| \mathrm{d}x\mathrm{d}y \\
&= 2L \cdot d_{\mathrm{TV}}(P_{\phi,\psi}(x,y), P_{\phi^*,\psi^*}(x,y)).
\end{aligned} \tag{30}
$$

同理，有

$$
|\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)] - \mathbb{E}_{\phi,\psi}[\ell(g_{\phi^*,\psi^*}(x),y)]| \leq 2L \cdot d_{\mathrm{TV}}(P_{\phi,\psi}(x,y), P_{\phi^*,\psi^*}(x,y)). \tag{31}
$$

结合式（28）、（29）、（30）和（31），可得

$$
\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x),y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)] \leq 4L \cdot d_{\mathrm{TV}}(P_{\phi,\psi}(x,y), P_{\phi^*,\psi^*}(x,y)). \tag{32}
$$

□

利用引理 A.1 和引理 A.2，我们可以给出定理3.4 的证明如下。证明的主要思路是分解风险。一个关键观察是，$\hat{}$ 有标签数据 $\{x_j, y_j\}_{j=1}^n$ 与预训练 ϕ相互独立，后者是根据无标签数据 $\{x_i\}_{i=1}^m$ 学习得到的。

定理 *3.4* 的证明。 设

$$
\tilde{\psi} := \arg\min_{\psi \in \Psi} d_{\mathrm{TV}}(P_{\hat{\phi},\psi}(x,y), P_{\phi^*,\psi^*}(x,y)). \tag{33}
$$

对于任意 $\phi \in \Phi, \psi \in \Psi$，我们定义

$$
\Delta_{\phi,\psi} := \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x),y)] - \frac{1}{n}\sum_{j=1}^n \ell(g_{\phi,\psi}(x_j),y_j). \tag{34}
$$

回顾一下，超额风险定义于（2）中。因此有

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\phi},\hat{\psi}) &= \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\hat{\psi}}(x),y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)] \\
&= \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\hat{\psi}}(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\hat{\phi},\hat{\psi}}(x_j),y_j) \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\hat{\phi},\hat{\psi}}(x_j),y_j) - \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\hat{\phi},\tilde{\psi}}(x_j),y_j) \quad (\le 0, \text{ by ERM in Algorithm 1}) \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\ell(g_{\hat{\phi},\tilde{\psi}}(x_j),y_j) - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x),y)] \\
&\quad + \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x),y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)] \\
&\le \Delta_{\hat{\phi},\hat{\psi}} - \Delta_{\hat{\phi},\tilde{\psi}} + \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x),y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)]. \qquad (35)
\end{aligned}$$

由引理 A.2，我们得到

$$\begin{aligned}
&\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x),y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)] \\
&\le 4L\cdot d_{\mathrm{TV}}(P_{\hat{\phi},\tilde{\psi}}(x,y), P_{\phi^*,\psi^*}(x,y)) \\
&= 4L\cdot \min_{\psi\in\Psi} d_{\mathrm{TV}}(P_{\hat{\phi},\psi}(x,y), P_{\phi^*,\psi^*}(x,y)) \quad (\text{by definition of } \tilde{\psi}) \\
&\le 4\kappa L\cdot d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)). \qquad (36)
\end{aligned}$$

ˆ最后一行成立，因为根据假设 3.2，对于任意 $\phi\in\Phi$，我们选择 T_1 使其满足（5），并选择 T_2使其满足（6）。令 $\psi = T_2^{-1}\circ\psi^*$。于是有

$$\begin{aligned}
\min_{\psi\in\Psi} d_{\mathrm{TV}}(P_{\hat{\phi},\psi}(x,y), P_{\phi^*,\psi^*}(x,y)) &\le d_{\mathrm{TV}}\big(\mathbb{P}_{\hat{\phi},\psi}(x,y), \mathbb{P}_{\phi^*,\psi^*}(x,y)\big) \\
&= d_{\mathrm{TV}}\big(\mathbb{P}_{T_1\circ\hat{\phi},\psi^*}(x,y), \mathbb{P}_{\phi^*,\psi^*}(x,y)\big) \\
&\le d_{\mathrm{TV}}\big(\mathbb{P}_{T_1\circ\hat{\phi}}(x,z), \mathbb{P}_{\phi^*}(x,z)\big) \\
&\le \kappa\cdot d_{\mathrm{TV}}\big(\mathbb{P}_{\hat{\phi}}(x,s), \mathbb{P}_{\phi^*}(x,s)\big). \qquad (37)
\end{aligned}$$

结合（35）和（36），我们得到

$$\mathrm{Error}_\ell(\hat{\phi},\hat{\psi}) \le \Delta_{\hat{\phi},\hat{\psi}} - \Delta_{\hat{\phi},\tilde{\psi}} + 4\kappa L\cdot d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)). \qquad (38)$$

我们定义以下事件

$$D := \left\{ d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)) \le 3\sqrt{\frac{1}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m)}{\delta}} \right\} \qquad (39)$$

and

$$R := \left\{ \sup_{\psi\in\Psi} |\Delta_{\hat{\phi},\psi}| \le R_n(\ell\circ\mathcal{G}_{\hat{\phi},\Psi}) + L\sqrt{\frac{2\log(4/\delta)}{n}} \right\}. \qquad (40)$$

成立的是

$$\mathbb{P}(D \cap R) = \mathbb{E}[\mathbb{1}_{D\cap R}] = \mathbb{E}[\mathbb{E}[\mathbb{1}_D \mathbb{1}_R|\hat{\phi}]] = \mathbb{E}[\mathbb{1}_D \mathbb{E}[\mathbb{1}_R|\hat{\phi}]] = \mathbb{E}[\mathbb{1}_D \mathbb{P}(R|\hat{\phi})], \tag{41}$$

其中，第三个等式是由以下事实得出的： D 是 ˆ ϕ-可测的。注意， $\{x_j, y_j\}_{j=1}^n$ 是 ˆ ˆ与 ϕ相互独立。根据引理 A.1，对于任意给定的 ϕ，其概率至少为 $1 - \delta/2$,

$$\sup_{\psi\in\Psi} |\Delta_{\hat{\phi},\psi}| \le R_n(\ell \circ \mathcal{G}_{\hat{\phi},\Psi}) + L\sqrt{\frac{2\log(4/\delta)}{n}}, \tag{42}$$

即,

$$\mathbb{P}(R|\hat{\phi}) \ge 1 - \delta/2. \tag{43}$$

ˆ 根据引理 3.3，以至少 $1 - \delta/2$ 的概率，我们算法第一步的输出 ϕ 满足

$$d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)) \le 3\sqrt{\frac{1}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}} \tag{44}$$

即,

$$\mathbb{P}(D) \ge 1 - \delta/2. \tag{45}$$

根据（41）、（43）和（45），我们有

$$\mathbb{P}(D \cap R) \ge (1 - \delta/2)^2 \ge 1 - \delta. \tag{46}$$

那么，在事件 $D \cap R$ 下，根据我们的分解（38），我们有

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) &\le \Delta_{\hat{\phi},\hat{\psi}} - \Delta_{\hat{\phi},\tilde{\psi}} + 4\kappa L \cdot d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)) \\
&\le 2\sup_{\psi\in\Psi} |\Delta_{\hat{\phi},\psi}| + 4\kappa L \cdot d_{\mathrm{TV}}(P_{\hat{\phi}}(x,s), P_{\phi^*}(x,s)) \\
&\le 2R_n(\ell \circ \mathcal{G}_{\hat{\phi},\Psi}) + 2L\sqrt{\frac{2\log(4/\delta)}{n}} + 12\kappa L\sqrt{\frac{1}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}} \\
&\le 2\max_{\phi\in\Phi} R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) + 2L\sqrt{\frac{2\log(4/\delta)}{n}} + 12\kappa L\sqrt{\frac{1}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m)}{\delta}}.
\end{aligned} \tag{47}$$

因此，我们证明定理3.4。

□

A.3 推论的证明 3.5

下面给出推论3.5的证明，该证明基于对高斯复杂度的分析。

证明。 由定理3.4，可得

$$\mathrm{Error}_\ell(\hat{\phi},\hat{\psi}) \le 2\max_{\phi\in\Phi} R_n(\ell\circ\mathcal{G}_{\phi,\Psi}) + 2L\cdot\sqrt{\frac{2}{n}\log\frac{4}{\delta}} + 12\kappa L\cdot\sqrt{\frac{1}{m}\log\frac{2N_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m)}{\delta}}. \quad (48)$$

因此，接下来只需对拉德马赫复杂度项进行界定。根据 Ledoux & Talagrand(2013)，拉德马赫复杂度的上界为高斯复杂度，即

$$R_n(\mathcal{F}) \le c\cdot G_n(\mathcal{F}) = c\cdot\mathbb{E}\hat{G}_n(\mathcal{F}), \quad (49)$$

其中 c 是某些绝对常数。这里 $G_n(\mathcal{F})$ 是高斯复杂度，其经验版本定义为

$$\hat{G}_n(\mathcal{F}) := \mathbb{E}_{g_i}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^n g_i f(x_i)\right| \,\middle|\, x_1,\cdots,x_n\right] \quad (50)$$

其中 g_1，$\cdots$，g_n 是独立同分布的 $\mathcal{N}(0,\ 1)$ 随机变量。根据 Wainwright (2019年)，我们有 e

$$\begin{aligned}
\hat{G}_n(\ell\circ\mathcal{G}_{\phi,\Psi}) &\le \frac{1}{\sqrt{n}}\cdot\min_{\delta\in[0,L]}\left\{\delta\sqrt{n} + 2L\sqrt{\log N(\ell\circ\mathcal{G}_{\phi,\Psi},\delta,\|\cdot\|_\infty)}\right\}\\
&\le \frac{1}{\sqrt{n}}\left(L + 2L\sqrt{\log N(\ell\circ\mathcal{G}_{\phi,\Psi},L/\sqrt{n},\|\cdot\|_\infty)}\right) \quad (\text{Take } \delta = L/\sqrt{n})\\
&\le 3L\sqrt{\frac{\log N(\ell\circ\mathcal{G}_{\phi,\Psi},L/\sqrt{n},\|\cdot\|_\infty)}{n}}. \quad (51)
\end{aligned}$$

结合 (49) 和 (51)，可得

$$R_n(\ell\circ\mathcal{G}_{\phi,\Psi}) \le 3cL\sqrt{\frac{\log N(\ell\circ\mathcal{G}_{\phi,\Psi},L/\sqrt{n},\|\cdot\|_\infty)}{n}}. \quad (52)$$

由 (48) 和 (52)，证明完毕。 □

A.4 定理 3.7的证明

本节首先说明假设 3.2 与假设 3.6之间的关系。随后，我们说明算法 1 第 2 行的最大似然估计步骤保证了海林格距离的一个上界$H(\mathbb{P}_{\hat{\phi}}(x,s)\ \mathbb{P}_{\phi^*}(x,s))$。然后，运用与定理 3.4证明中相同的技巧，我们证明定理 3.7。

A.4.1 假设 3.2 与假设3.6之间的关系

假设3.6 实际上是假设3.2的一个放宽。为此，根据假设3.2，对于任意$\phi\in\Phi$，我们选取 T_1 使其满足（5），并选取 T_2 使其满足（6）。令 $\psi = T_2^{-1}\circ\psi^*$。于是有

$$\begin{aligned}
&d_{\mathrm{TV}}\big(\mathbb{P}_{\phi,\psi}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y)\big)\\
&= d_{\mathrm{TV}}\big(\mathbb{P}_{T_1\circ\phi,\psi^*}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y)\big)\\
&\le d_{\mathrm{TV}}\big(\mathbb{P}_{T_1\circ\phi}(x,z),\mathbb{P}_{\phi^*}(x,z)\big)\\
&\le \kappa\cdot d_{\mathrm{TV}}\big(\mathbb{P}_{\phi}(x,s),\mathbb{P}_{\phi^*}(x,s)\big).
\end{aligned}$$

注意，全变差距离可以用海林格距离作为上界。因此，假设3.2直接蕴含假设3.6。

A.4.2 海灵格距离保证

ˆ 假设 ϕ 是算法1中最大似然估计步骤的输出，且满足

$$\hat{\phi} \leftarrow \arg\max_{\phi \in \Phi} \sum_{i=1}^{m} \log p_{\phi}(x_i, s_i). \tag{53}$$

我们有 t 以下关于 $\mathbb{P}_{\hat{\phi}}(x, s)$ 与 $\mathbb{P}_{\phi^*}(x, s$ 之间海林格距离的理论保证).

ˆ引理 A.3。 令 ϕ 为算法 1的输出。则以至少 $1-\delta$的概率，有

$$H\big(\mathbb{P}_{\hat{\phi}}(x,s), \mathbb{P}_{\phi^*}(x,s)\big) \leq \sqrt{\frac{2}{m} \log \frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m^2\big)}{\delta}}, \tag{54}$$

其中，我们记 $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi) := \{p_{\phi}(x, s) \mid \phi \in \Phi\}$。

引理 *A.3.* 为简化记号，我们记为 $\boldsymbol{x} := (x, s)$。令 $\epsilon > 0$。与定理 3.3的证明类似，我们以至少 $1-\delta$证明

$$1 - \int \sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} \leq \frac{1}{m} \log \frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon\big)}{\delta}. \tag{55}$$

其中 $\bar{p}_{\hat{\phi}}(\boldsymbol{x}) \in \mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi)\ \epsilon)$ 满足 $\bar{p}_{\hat{\phi}}(\boldsymbol{x}) \geq p_{\phi^*}(\boldsymbol{x})$，对任意 $\boldsymbol{x}$ 以及

$$\int \bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})\, d\boldsymbol{x} \leq \epsilon. \tag{56}$$

注意

$$\begin{aligned}
&1 - \int \sqrt{p_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} - \left(1 - \int \sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x}\right) \\
&= \int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\hat{\phi}}(\boldsymbol{x})}\right) \sqrt{p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} \\
&\leq \sqrt{\int \left(\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\hat{\phi}}(\boldsymbol{x})}\right)^2 d\boldsymbol{x}} \\
&= \sqrt{\int \bar{p}_{\hat{\phi}}(\boldsymbol{x}) + p_{\hat{\phi}}(\boldsymbol{x}) - 2\sqrt{\bar{p}_{\hat{\phi}}(\boldsymbol{x}) p_{\hat{\phi}}(\boldsymbol{x})}\, d\boldsymbol{x}} \\
&\leq \sqrt{\int \bar{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\hat{\phi}}(\boldsymbol{x})\, d\boldsymbol{x}} \\
&\leq \sqrt{\epsilon}.
\end{aligned} \tag{57}$$

这里第一个不等式由柯西-施瓦茨不等式得出，第二个不等式则由以下事实得出：$\sqrt{\bar{p}_{\hat{\phi}}}(\boldsymbol{x})\, p_{\hat{\phi}}(\boldsymbol{x}) \geq p_{\hat{\phi}}(\boldsymbol{x})$。根据（55）和（57），我们有

$$1 - \int \sqrt{p_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} \leq \sqrt{\epsilon} + \frac{1}{m} \log \frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon\big)}{\delta}, \tag{58}$$

这意味着

$$H^2\big(\mathbb{P}_{\hat{\phi}}(\boldsymbol{x}), \mathbb{P}_{\phi^*}(\boldsymbol{x})\big) = 1 - \int \sqrt{p_{\hat{\phi}}(\boldsymbol{x}) p_{\phi^*}(\boldsymbol{x})}\, d\boldsymbol{x} \leq \sqrt{\epsilon} + \frac{1}{m} \log \frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), \epsilon\big)}{\delta}. \tag{59}$$

令 $\epsilon = 1/m^2$。我们有

$$H^2\big(\mathbb{P}_{\hat{\phi}}(x, s), \mathbb{P}_{\phi^*}(x, s)\big) \leq \frac{2}{m} \log \frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m^2\big)}{\delta}. \tag{60}$$

□

A.4.3 定理3.7 的证明

有了引理 A.3 之后，我们就可以证明定理3.7。

ˆ定理 *3.7*的证明。 设 ϕ 为算法 1中最大似然估计步骤的输出。并且对任意 $\phi \in \Phi$， $\psi \in \Psi$，我们定义

$$\Delta_{\phi,\psi} := \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi,\psi}(x), y)] - \frac{1}{n}\sum_{j=1}^{n} \ell(g_{\phi,\psi}(x_j), y_j). \tag{61}$$

按照定理 3.4证明中的相同论证，我们以至少 $1-\delta$的概率得到，

$$H\big(\mathbb{P}_{\hat{\phi}}(x, s), \mathbb{P}_{\phi^*}(x, s)\big) \leq \sqrt{\frac{2}{m} \log \frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi), 1/m^2)}{\delta}} \tag{62}$$

and

$$\sup_{\psi\in\Psi} |\Delta_{\hat{\phi},\psi}| \leq R_n(\ell \circ \mathcal{G}_{\hat{\phi},\Psi}) + L\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{63}$$

此外，正如（35）中所述，我们有

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) &\leq \Delta_{\hat{\phi},\hat{\psi}} - \Delta_{\hat{\phi},\tilde{\psi}} + \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x), y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x), y)] \\
&\leq 2R_n(\ell \circ \mathcal{G}_{\hat{\phi},\Psi}) + 2L\sqrt{\frac{2\log(4/\delta)}{n}} \\
&\quad + \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x), y)] - \mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x), y)],
\end{aligned} \tag{64}$$

˜其中 $\psi:=\arg\min_{\psi\in\Psi} d_{\mathrm{TV}}(\mathbb{P}_{\hat{\phi},\psi}(x,y)\,\mathbb{P}_{\phi^*,\psi^*}(x,y))$，第二个不等式由（63）得出。根据引理 A.2，我们有

$$\begin{aligned}
&\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\hat{\phi},\tilde{\psi}}(x),y)]-\mathbb{E}_{\phi^*,\psi^*}[\ell(g_{\phi^*,\psi^*}(x),y)]\\
&\le 4L\cdot d_{\mathrm{TV}}(\mathbb{P}_{\hat{\phi},\tilde{\psi}}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y))\\
&= 4L\cdot\min_{\psi\in\Psi} d_{\mathrm{TV}}(\mathbb{P}_{\hat{\phi},\psi}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y)) \quad \text{(by definition of } \tilde{\psi})\\
&\le_{1)} 4\kappa L\cdot H(\mathbb{P}_{\hat{\phi}}(x,s),\mathbb{P}_{\phi^*}(x,s))\\
&\le_{2)} 4\kappa L\sqrt{\frac{2}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m^2)}{\delta}},
\end{aligned}\tag{65}$$

其中，1）由假设 3.6 得出，2）由（62）得出。结合（64）和（65），可得

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\phi},\hat{\psi}) &\le 2R_n(\ell\circ\mathcal{G}_{\hat{\phi},\Psi})+2L\sqrt{\frac{2\log(4/\delta)}{n}}+4\kappa L\sqrt{\frac{2}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m^2)}{\delta}}\\
&\le 2\max_{\phi\in\Phi} R_n(\ell\circ\mathcal{G}_{\phi,\Psi})+2L\sqrt{\frac{2\log(4/\delta)}{n}}+4\kappa L\sqrt{\frac{2}{m}\log\frac{2N(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi),1/m^2)}{\delta}}.
\end{aligned}\tag{66}$$

□

B 第 4 节的证明

在第 B.1 节中，我们通过分析两个高维高斯分布之间的全变差距离并应用 Davis-Kahan 定理，证明了以线性回归为下游任务的因子模型具有 κ 可迁移性（引理 4.2），其中 κ 取决于真实参数 B^* 的最大和最小奇异值。在第 B.2 节和第 B.3 节中，我们证明了两个将在定理 4.3 的证明中用到的引理。具体而言，在第 B.2 节中，我们利用 ϵ 离散化来给集合 $\mathcal{P}(\mathcal{B})$ 的括号覆盖数提供上界（引理 B.5）。在第 B.3 节中，我们证明引理 B.6，该引理将用于给函数类 $\ell\circ\mathcal{G}_{B,\mathcal{C}}$ 的拉德马赫复杂度提供上界。在第 B.4 节中，我们证明定理 4.3。最后，在第 B.5 节中，我们给出精细分析，以证明定理 4.4。

B.1 引理证明 4.2

首先，我们给出一些有用的引理，这些引理将在引理 4.2的证明中用到。给定两个高维高斯分布，我们可以如下界定它们之间的全变差距离。

引理 B.1 （Devroye 等人 （2018年）中的定理 1.2 和命题 2.1）。 假设 $d>1$。令$\mu_1\neq\mu_2\in\mathbb{R}^d$。那么，我们有

$$\frac{1}{200}\le\frac{d_{\mathrm{TV}}\big(\mathcal{N}(\mu_1,I_d),\mathcal{N}(\mu_2,I_d)\big)}{\min\{1,\|\mu_1-\mu_2\|_2\}}\le 1.$$

引理 B.2 （定理 1.1，见 Devroye 等人 （2018年））。 假设 $d>1$。令 $\mu\in\mathbb{R}^d$ 和 $\Sigma_1\neq\Sigma_2$为正定 $d\times d$ 矩阵。则有

$$\frac{1}{100}\le\frac{d_{\mathrm{TV}}\big(\mathcal{N}(\mu,\Sigma_1),\mathcal{N}(\mu,\Sigma_2)\big)}{\min\{1,\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}-I_d\|_{\mathrm{F}}\}}\le\frac{3}{2}.$$

回顾一下，我们定义 $\mathcal{B} := \{B \in \mathbb{R}^{d\times r} \,|\, \|B\|_2 \le D\}$。令 $B \in \mathcal{B}$ 和 B^* 为真实参数。我们分别用 $\sigma^*_{\max}$ 和 $\sigma^*_{\min}$ 表示 B^*的最大和最小奇异值。此外，我们分别用 $B = U\Sigma V$和 $B^* = U^*\Sigma^*V^*$表示 B 和 B^* 的奇异值分解。这里， Σ、$\Sigma^* \in \mathbb{R}^{r\times r}$是对角矩阵， U、$U^* \in \mathbb{R}^{d\times r}$、$V$、$V^* \in \mathbb{R}^{r\times d}$ 是具有正交列的矩阵。令

$$M := BB^T = U\Lambda U^T, \quad M^* := B^*B^{*T} = U^*\Lambda^*U^{*T}, \tag{67}$$

其中 $\Lambda := \Sigma\Sigma^T$ 且 $\Lambda^* := \Sigma^*\Sigma^{*T}$。我们定义

$$O := \arg\min_{O\in\mathcal{O}^{r\times r}} \|UO - U^*\|_{\mathrm{F}}. \tag{68}$$

于是，我们有以下引理。

引理 B.3. 对于 M， M^* 定义于 (67) 和 O 定义于 (68)，存在一些绝对常数$c > 1$ 使得

$$\|UO - U^*\|_{\mathrm{F}} \le \frac{c}{(\sigma^*_{\min})^2}\|M - M^*\|_{\mathrm{F}}.$$

这里， $\sigma^*_{\min}$ 是真实参数 B^*的最小奇异值。

证明。 应用 Davis-Kahan 定理（Davis 和 Kahan， 1970）。 □

引理 B.4. 对于 M， M^* 定义于 (67) 和 O 定义于 (68)，存在一些绝对常数c 使得

$$\|\Lambda^{1/2}O - O\Lambda^{*1/2}\|_{\mathrm{F}} \le \frac{4c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3}\|M - M^*\|_{\mathrm{F}}.$$

这里， $\sigma^*_{\min}$ 是真实参数 B^*的最小奇异值。

引理*B.4*的证明。我们的证明受到Ma 等人(2018年)的启发。根据Schmitt (1992年)中的引理 2.1，我们有

$$\|\Lambda^{1/2}O - O\Lambda^{*1/2}\|_{\mathrm{F}} \le \frac{1}{\sqrt{\sigma_{\min}(M^*)}}\|O^T\Lambda O - \Lambda^*\|_{\mathrm{F}} = \frac{1}{\sigma^*_{\min}}\|O^T\Lambda O - \Lambda^*\|_{\mathrm{F}}. \tag{69}$$

注意， $\Lambda = U^TMU$和 $\Lambda^* = U^{*T}M^*U^*$。因此，我们有

$$\begin{aligned}
\|O^T\Lambda O - \Lambda^*\|_{\mathrm{F}} &= \|O^TU^TMUO - U^{*T}M^*U^*\|_{\mathrm{F}} \\
&\le \|O^TU^TMUO - O^TU^TM^*UO\|_{\mathrm{F}} + \|O^TU^TM^*UO - U^{*T}M^*UO\|_{\mathrm{F}} \\
&\quad + \|U^{*T}M^*UO - U^{*T}M^*U^*\|_{\mathrm{F}} \\
&\le \|M - M^*\|_{\mathrm{F}} + 2\|M^*\|_2\|UO - U^*\|_{\mathrm{F}} \\
&\le \|M - M^*\|_{\mathrm{F}} + 2c\left(\frac{\sigma^*_{\max}}{\sigma^*_{\min}}\right)^2\|M - M^*\|_{\mathrm{F}} \\
&\le 4c\left(\frac{\sigma^*_{\max}}{\sigma^*_{\min}}\right)^2\|M - M^*\|_{\mathrm{F}},
\end{aligned} \tag{70}$$

其中第三个不等式由引理B.3推出。结合（69）和（70），我们有

$$\|\Lambda^{1/2}O - O\Lambda^{*1/2}\|_{\mathrm{F}} \le \frac{4c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3}\|M - M^*\|_{\mathrm{F}}.$$

□

现在我们准备好证明引理 4.2。

引理 *4.2* 的证明。令 $\mathcal{O}^{r\times r} := \{O \in \mathbb{R}^{r\times r} \mid OO^T = O^TO = I_r\}$。首先，我们证明对于任意 $(B,\beta,O) \in \mathcal{B}\times\mathcal{C}\times\mathcal{O}$，都有 $\mathbb{P}_{B,\beta}(x,y) = \mathbb{P}_{BO,O^T\beta}(x,y)$。这一点可由以下观察轻易得出，

$$\mathbb{P}_{BO,O^T\beta} \sim \mathcal{N}\left(0, \begin{bmatrix} BO(BO)^T & BOO^T\beta \\ \beta^TOO^TB^T & (O^T\beta)^TO^T\beta \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} BB^T & B\beta \\ \beta^TB^T & \beta^T\beta \end{bmatrix}\right) \sim \mathbb{P}_{B,\beta}.$$

根据引理 B.3，存在某个常数 $c > 1$ 使得

$$\|UO - U^*\|_{\mathrm{F}} \le \frac{c}{(\sigma^*_{\min})^2}\|BB^T - B^*B^{*T}\|_{\mathrm{F}}. \tag{71}$$

根据引理 B.4，存在某个常数 $c > 1$ 使得

$$\|\Sigma O - O\Sigma^*\|_{\mathrm{F}} \le \frac{4c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3}\|BB^T - B^*B^{*T}\|_{\mathrm{F}}. \tag{72}$$

令 $\hat{}O := V^{-1}OV^* \in \mathcal{O}^{r\times r}$。根据 (71) 和 (72)，我们有

$$\begin{aligned}
\|B\hat{O} - B^*\|_{\mathrm{F}} &= \|U\Sigma OV^* - U^*\Sigma^*V^*\|_{\mathrm{F}} \\
&\le \|U\Sigma O - U^*\Sigma^*\|_{\mathrm{F}} \\
&\le \|U\Sigma O - UO\Sigma^*\|_{\mathrm{F}} + \|UO\Sigma^* - U^*\Sigma^*\|_{\mathrm{F}} \\
&\le \|\Sigma O - O\Sigma^*\|_{\mathrm{F}} + \|UO - U^*\|_{\mathrm{F}}\|\Sigma^*\|_2 \\
&\le c\cdot\left(\frac{4(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3} + \frac{\sigma^*_{\max}}{(\sigma^*_{\min})^2}\right)\cdot\|BB^T - B^*B^{*T}\|_{\mathrm{F}} \\
&\le \frac{5c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3}\cdot\|BB^T - B^*B^{*T}\|_{\mathrm{F}}.
\end{aligned} \tag{73}$$

注意，

$$\begin{aligned}
d_{\mathrm{TV}}\big(\mathbb{P}_{B\hat{O}}(x,z), \mathbb{P}_{B^*}(x,z)\big) &= \int |p_{B\hat{O}}(x\,|\,z) - p_{B^*}(x\,|\,z)|p(z)\,dxdz \\
&= \int d_{\mathrm{TV}}\big(\mathcal{N}(B\hat{O}z, I_d), \mathcal{N}(B^*z, I_d)\big)p(z)\,dz \\
&\le \int \min\{1, \|B\hat{O}z - B^*z\|_2\}p(z)\,dz \\
&\le \min\big\{1, \mathbb{E}[\|B\hat{O}z - B^*z\|_2]\big\},
\end{aligned} \tag{74}$$

其中第一个不等式由引理 B.1 得出。我们可以证明,

$$\begin{aligned}\mathbb{E}[\|B\hat{O}z - B^*z\|_2] &\le \Big(\mathbb{E}\big[\|B\hat{O}z - B^*z\|_2^2\big]\Big)^{1/2} \\ &= \Big(\mathbb{E}\big[z^T(B\hat{O} - B^*)^T(B\hat{O} - B^*)z\big]\Big)^{1/2} \\ &= \Big(\mathbb{E}\big[\mathrm{Tr}\big((B\hat{O} - B^*)^T(B\hat{O} - B^*)zz^T\big)\big]\Big)^{1/2} \\ &= \Big(\mathrm{Tr}\big((B\hat{O} - B^*)^T(B\hat{O} - B^*)\big)\Big)^{1/2} \\ &= \|B\hat{O} - B^*\|_{\mathrm{F}}. \end{aligned} \tag{75}$$

根据（73）、（74）和（75，可得

$$\begin{aligned}&d_{\mathrm{TV}}\big(\mathbb{P}_{B\hat{O}}(x,z), \mathbb{P}_{B^*}(x,z)\big) \\ &\quad\le \min\big\{1, \|B\hat{O} - B^*\|_{\mathrm{F}}\big\} \\ &\quad\le \min\left\{1, \frac{5c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3} \cdot \|BB^T - B^*B^{*T}\|_{\mathrm{F}}\right\} \\ &\quad\le \frac{5c(\sigma^*_{\max})^2}{(\sigma^*_{\min})^3} \cdot \big((\sigma^*_{\max})^2 + 1\big) \cdot \min\left\{1, \frac{\|BB^T - B^*B^{*T}\|_{\mathrm{F}}}{(\sigma^*_{\max})^2 + 1}\right\}, \end{aligned} \tag{76}$$

其中，最后一个不等式由 $c > 1$ and 得出

$$\frac{(\sigma^*_{\max})^2 + 1}{\sigma^*_{\min}} \ge \frac{2\sigma^*_{\max}}{\sigma^*_{\min}} > 1.$$

根据引理 B.2，可得

$$\begin{aligned}&d_{\mathrm{TV}}(p_B(x), p_{B^*}(x)) \\ &\quad\ge \frac{1}{100}\min\big\{1, \|(B^*B^{*T} + I_d)^{-1/2}(BB^T - B^*B^{*T})(B^*B^{*T} + I_d)^{-1/2}\|_{\mathrm{F}}\big\}. \end{aligned} \tag{77}$$

注意到

$$\begin{aligned}&\|(B^*B^{*T} + I_d)^{-1/2}(BB^T - B^*B^{*T})(B^*B^{*T} + I_d)^{-1/2}\|_{\mathrm{F}} \\ &\quad\ge \frac{\|BB^T - B^*B^{*T}\|_{\mathrm{F}}}{\|B^*B^{*T} + I_d\|_2} \ge \frac{\|BB^T - B^*B^{*T}\|_{\mathrm{F}}}{(\sigma^*_{\max})^2 + 1}. \end{aligned} \tag{78}$$

因此，根据（77）和（78），可得

$$d_{\mathrm{TV}}(p_B(x), p_{B^*}(x)) \ge \frac{1}{100}\min\left\{1, \frac{\|BB^T - B^*B^{*T}\|_{\mathrm{F}}}{(\sigma^*_{\max})^2 + 1}\right\} \tag{79}$$

最后，由（76）和（79，我们得到

$$\begin{aligned}d_{\mathrm{TV}}\big(\mathbb{P}_{B\hat{O}}(x,z), \mathbb{P}_{B^*}(x,z)\big) &\le \frac{500c(\sigma^*_{\max})^2\big((\sigma^*_{\max})^2 + 1\big)}{(\sigma^*_{\min})^3} \cdot d_{\mathrm{TV}}(p_B(x), p_{B^*}(x)) \\ &\le \frac{500c(\sigma^*_{\max} + 1)^4}{(\sigma^*_{\min})^3} \cdot d_{\mathrm{TV}}(p_B(x), p_{B^*}(x)). \end{aligned}$$

□

B.2 括号覆盖数

通过应用 ϵ-离散化技术，我们得到 $\mathcal{P}(\mathcal{B})$ 的括号覆盖数上界如下。

引理 B.5. 设 $\mathcal{P}_{\mathcal{X}}(\mathcal{B}) := \{\mathcal{N}(0, BB^T + I_d) \mid B \in \mathcal{B}\}$，其中 $\mathcal{B} = \{B \in \mathbb{R}^{d\times r} \mid \|B\|_2 \le D\}$ 对某个 $D > 0$ 成立。则熵可由下式界定，

$$\log N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{B}), 1/m) \le 4dr \log\big(24mdr(D^2+1)\big).$$

引理的证明 *B.5.* 考虑一个高斯分布集合

$$\mathcal{P}_{\mathcal{X}}(\mathcal{B}) := \left\{ p_\Sigma(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}x^T\Sigma^{-1}x} \,\middle|\, \Sigma = BB^T + I_d, B \in \mathcal{B} \right\},$$

其中 $\mathcal{B} = \{B \in \mathbb{R}^{d\times r} \mid \|B\|_2 \le D\}$。注意到

$$\lambda_{\max}(\Sigma^{-1}) = \big(\lambda_{\min}(\Sigma)\big)^{-1} = 1,\ \lambda_{\min}(\Sigma^{-1}) = \big(\lambda_{\max}(\Sigma)\big)^{-1} \ge \frac{1}{D^2+1}. \tag{80}$$

这里，我们分别用 $\lambda_{\max}(\Sigma^{-1})$ 和 $\lambda_{\min}(\Sigma^{-1})$ 表示 Σ^{-1} 的最大特征值和最小特征值。我们的目标是为 $\mathcal{P}_{\mathcal{X}}(\mathcal{B})$ 找到一个 $1/m$-括号 $\mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{B})\ 1/m)$。换言之，对于任意 $p_\Sigma(x) \in \mathcal{P}_{\mathcal{X}}(\mathcal{B})$，我们需要定义 $\bar{p}_\Sigma(x) \in \mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{B})\ 1/m)$，使得

- $\bar{p}_\Sigma(x) \ge p_\Sigma(x) \quad \forall x \in \mathbb{R}^d$
- $\int |\bar{p}_\Sigma(x) - p_\Sigma(x)|\, dx \le 1/m$。

注意，$\mathrm{rank}(BB^T) = r < d$ 和 $\Sigma = BB^T + I_d$。因此，Σ^{-1} 的特征分解形式如下

$$\Sigma^{-1} = V \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} V^T = U \begin{bmatrix} \lambda_1 - 1 & & \\ & \ddots & \\ & & \lambda_r - 1 \end{bmatrix} U^T + I_d, \tag{81}$$

其中，$VV^T = V^TV = I_d$ 和 $U \in \mathbb{R}^{d\times r}$ 是 V 的前 r 列。为简化符号，我们记作

$$\Lambda := \begin{bmatrix} \lambda_1 - 1 & & \\ & \ddots & \\ & & \lambda_r - 1 \end{bmatrix}.$$

因此，我们有 $\Sigma^{-1} = U\Lambda U^T + I_d$。对于某个固定的 $0 < \epsilon \le (D^2+1)^{-1}/2$（我们稍后将选定它），如果对于某个 $k \in \mathbb{Z}$，有 $\lambda_i \in [k\epsilon, (k+1)\epsilon)$，我们定义 $\bar{\lambda}_i := (k-1)\epsilon$。注意，$\lambda_i \ge \lambda_{\min}(\Sigma^{-1}) \ge (D^2+1)^{-1}$。因此，有 $k \ge 2$ 和 $\lambda_i = (k-1)\epsilon \ge \epsilon > 0$。此外，我们有 $\epsilon \le \lambda_i - \bar{\lambda}_i \le 2\epsilon$。我们定义

$$\bar{\Lambda} := \begin{bmatrix} \bar{\lambda}_1 - 1 & & \\ & \ddots & \\ & & \bar{\lambda}_r - 1 \end{bmatrix}.$$

对于矩阵 $U = (u_{i,\ j}) \in \mathbb{R}^{d\times r}$，如果对于某个 $k \in \mathbb{Z}$，有 $u_{i,\ j} \in [\frac{k\epsilon}{3\sqrt{dr}}, \frac{(k+1)\epsilon}{3\sqrt{dr}})$，我们定义 $\bar{u}_{i,\ j} := \frac{k\epsilon}{3\sqrt{dr}}$和 $\bar{U} := (\bar{u}_{i,\ j}) \in \mathbb{R}^{d\times r}$。于是有

$$\|U - \bar{U}\|_2 \le \|U - \bar{U}\|_F = \sqrt{\sum_{i,j} |u_{i,j} - \bar{u}_{i,j}|^2} \le \sqrt{dr} \cdot \frac{\epsilon}{3\sqrt{dr}} = \frac{\epsilon}{3}. \tag{82}$$

我们定义

$$\overline{\Sigma^{-1}} := \bar{U}\bar{\Lambda}\bar{U}^T + I_d. \tag{83}$$

注意，$(D^2+1)^{-1} \le \lambda_i \le 1$ 和 $|u_{i,j}| \le 1$。因此，我们总共有

$$\left(\frac{1-(D^2+1)^{-1}}{\epsilon}\right)^r \cdot \left(\frac{6\sqrt{dr}}{\epsilon}\right)^{dr} = \left(\frac{D^2}{(D^2+1)\epsilon}\right)^r \cdot \left(\frac{6\sqrt{dr}}{\epsilon}\right)^{dr} \tag{84}$$

许多 $\bar{\Sigma}^{-1}$。注意，对于任意 $\|x\|_2 = 1$，我们有

$$\begin{aligned}
x^T(\Sigma^{-1} - \overline{\Sigma^{-1}})x &= x^T(U^T\Lambda U - \bar{U}\bar{\Lambda}\bar{U}^T)x \\
&= x^T U^T(\Lambda - \bar{\Lambda})Ux + x^T(U-\bar{U})^T\bar{\Lambda}(U+\bar{U})x \\
&\ge \lambda_{\min}(\Lambda - \bar{\Lambda}) - \|(U-\bar{U})^T\bar{\Lambda}(U+\bar{U})\|_2 \\
&\ge \lambda_{\min}(\Lambda - \bar{\Lambda}) - \|U-\bar{U}\|_2 \cdot \|\bar{\Lambda}(U+\bar{U})\|_2 \\
&\ge \epsilon - 3\left(2\epsilon + \frac{D^2}{D^2+1}\right)\|U-\bar{U}\|_2 \\
&\ge \epsilon - 3\left(2\epsilon + \frac{D^2}{D^2+1}\right)\cdot\frac{\epsilon}{3} \ge 0,
\end{aligned}$$

其中第三个不等式由

$$\|\bar{\Lambda}(U+\bar{U})\|_2 \le \|\bar{\Lambda}\|_2\|U+\bar{U}\|_2 \le \left(2\epsilon + 1 - \frac{1}{D^2+1}\right)\cdot\left(2+\frac{\epsilon}{3}\right) \le 3\left(2\epsilon + \frac{D^2}{D^2+1}\right).$$

最后一个不等式由我们的假设 $\epsilon \le (D^2+1)^{-1}/2$ 得出。因此，对于任意 $x \in \mathbb{R}^d$，有

$$x^T(\Sigma^{-1} - \overline{\Sigma^{-1}})x \ge 0. \tag{85}$$

我们考虑如下形式的 $\bar{p}_\Sigma(x)$

$$\bar{p}_\Sigma(x) = c\sqrt{\frac{|\overline{\Sigma^{-1}}|}{(2\pi)^d}}e^{-\frac{1}{2}x^T\overline{\Sigma^{-1}}x}.$$

根据（85），我们有：$\bar{p}_\Sigma(x) \ge p_\Sigma(x)$ 对任意 $x \in \mathbb{R}^d$ 当且仅当

$$c \ge \sqrt{\frac{|\Sigma^{-1}|}{|\overline{\Sigma^{-1}}|}} = \sqrt{\frac{\lambda_1\dots\lambda_r}{\bar{\lambda}_1\dots\bar{\lambda}_r}}.$$

注意，

$$\frac{\lambda_i}{\bar{\lambda}_i} \le \frac{(k+1)\epsilon}{(k-1)\epsilon} = 1 + \frac{2}{k-1} \le 1 + \frac{4}{k} \le 1 + 4(D^2+1)\epsilon,$$

其中，第二个不等式由 $k \ge 2$ 得出，最后一个不等式由 $k\epsilon \ge (D^2+\sigma^2)^{-1}$ 得出。于是我们得到

$$\sqrt{\frac{\lambda_1 \dots \lambda_r}{\bar{\lambda}_1 \dots \bar{\lambda}_r}} \le \big(1 + 4(D^2+1)\epsilon\big)^{r/2}.$$

令 $c = (1 + 4(D^2+1)\epsilon)^{r/2}$。于是有

$$c \ge \sqrt{\frac{\lambda_1 \dots \lambda_r}{\bar{\lambda}_1 \dots \bar{\lambda}_r}},$$

这意味着对于任意 $x \in \mathbb{R}^d$， $\bar{p}_\Sigma(x) \ge p_\Sigma(x)$ 均成立。注意到

$$\int |\bar{p}_\Sigma(x) - p_\Sigma(x)|\, dx = c - 1 = (1 + 4(D^2+1)\epsilon)^{r/2} - 1 \le 4(D^2+1)\epsilon r,$$

其中，最后一个不等式由 $(1+x)^{r/2} \le 1 + rx$ 对 $x \le r^{-1}$ 成立得出。令

$$\epsilon = \frac{1}{4(D^2+1)mr}. \tag{86}$$

我们有

$$\int |\bar{p}_\Sigma(x) - p_\Sigma(x)|\, dx \le 4(D^2+1)\epsilon r = \frac{1}{m}.$$

由（84）和（86），我们证明

$$N_{[\,]}(\mathcal{P}_\mathcal{X}(\mathcal{B}), 1/m) \le (4rmD^2)^r \cdot \big(24rm(D^2+1)\sqrt{dr}\big)^{dr},$$

这意味着

$$\log N_{[\,]}(\mathcal{P}_\mathcal{X}(\mathcal{B}), 1/m) \le 4dr \log\big(24mdr(D^2+1)\big).$$

□

B.3 拉德马赫复杂度

注意，对于固定的 B，预测函数类

$$\mathcal{G}_{B,\mathcal{C}} := \big\{ g_{B,\beta}(x) = \beta^T B^T (BB^T + \sigma^2 I_d)^{-1} x \,\big|\, \beta \in \mathcal{C} \big\}$$

属于线性假设类。对于线性假设类 $\mathcal{H}$，其经验拉德马赫复杂度可界定如下。

引理 B.6. 对于线性假设类 $\mathcal{H} = \{h_\beta(x) = \beta^T x \,|\, \beta \in \mathbb{R}^r, \|\beta\|_2 \le D\}$，其中 $x \in \mathbb{R}^r$且 $\|x\|_2 \le X$，经验拉德马赫复杂度可界定如下，

$$\hat{R}_n(\mathcal{H}) \le \frac{2DX}{\sqrt{n}}.$$

引理的证明 *B.6*。 注意

$$\begin{aligned}\hat{R}_n(\mathcal{H}) &= \frac{2}{n}\mathbb{E}_{\sigma_i}\left[\sup_{\|\beta\|_2\le D}\sum_{i=1}^n \sigma_i \cdot \beta^T x_i\right] = \frac{2}{n}\mathbb{E}_{\sigma_i}\left[\sup_{\|\beta\|_2\le D}\beta^T\left(\sum_{i=1}^n \sigma_i x_i\right)\right] \\ &\le \frac{2}{n}\mathbb{E}_{\sigma_i}\left[\sup_{\|\beta\|_2\le D}\|\beta\|_2\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2\right] \le \frac{2D}{n}\mathbb{E}_{\sigma_i}\left[\sqrt{\sum_{i,j}\sigma_i\sigma_j x_i^T x_j}\right].\end{aligned}$$

根据詹森不等式，我们可得

$$\hat{R}_n(\mathcal{H}) \le \frac{2D}{n}\mathbb{E}_{\sigma_i}\left[\sqrt{\sum_{i,j}\sigma_i\sigma_j x_i^T x_j}\right] \le \frac{2D}{n}\sqrt{E_{\sigma_i}\left[\sum_{i,j}\sigma_i\sigma_j x_i^T x_j\right]} = \frac{2D}{n}\sqrt{\sum_{i=1}^n \|x_i\|^2} \le \frac{2DX}{\sqrt{n}}.$$

□

B.4 定理 4.3的证明

本节通过证明定理 1 来验证算法的效用 4.3。回顾一下，截断平方损失定义为

$$\tilde{\ell}(x, y) := (y - x)^2 \mathbb{I}_{\{(y-x)^2 \le L\}} + L \cdot \mathbb{I}_{\{(y-x)^2 > L\}}, \tag{87}$$

它关于第一个参数是 L–有界且 $2\sqrt{L}$–利普希茨的。在证明定理 4.3 之前，我们需要陈述一些核心引理。回顾 $g_{B,\beta}(x)$ 的定义：

$$g_{B,\beta}(x) := \arg\min_g \mathbb{E}_{B,\beta}[\ell(g(x), y)]. \tag{88}$$

由于 ℓ 是平方损失，显然

$$g_{B,\beta}(x) := \arg\min_g \mathbb{E}_{B,\beta}[\ell(g(x), y)] = \mathbb{E}_{\mathbb{P}_{B,\beta}(x,y)}[y \mid x] = \beta^T B^T (BB^T + I_d)^{-1} x. \tag{89}$$

下一个引理表明，在平方损失 ℓ 和截断˜平方损失 ℓ 下，最优预测器保持不变。

引理 B.7. 我们将截断平方损失下的最优预测器记为 $\tilde{g}_{B,\ \beta}$ 即：

$$\tilde{g}_{B,\beta} \leftarrow \arg\min_g \mathbb{E}_{B,\beta}[\tilde{\ell}(g(x), y)]. \tag{90}$$

于是有

$$\tilde{g}_{B,\beta}(x) = \mathbb{E}_{\mathbb{P}_{B,\beta}(x,y)}[y \mid x] = g_{B,\beta}(x). \tag{91}$$

引理 B.7的证明。注意，在参数 B,β 下，给定 x 时 y 的分布是一个均值为 $\mu=\mathbb{E}_{P_{B,\beta}(x,y)}[y\mid x]$ 、方差为 v^2 （无关紧要）的高斯分布。我们将函数 f 定义为

$$\begin{aligned}f(a)&:=\mathbb{E}_{B,\beta}[\tilde{\ell}(a,y)\mid x]\\&=\int_{a-\sqrt{L}}^{a+\sqrt{L}}(y-a)^2\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y+\int_{a+\sqrt{L}}^{+\infty}L\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y\\&\quad+\int_{-\infty}^{a-\sqrt{L}}L\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y.\end{aligned}\tag{92}$$

于是，有

$$\begin{aligned}f'(a)&=\frac{L}{v\sqrt{2\pi}}e^{-\frac{(a-\mu+\sqrt{L})^2}{2v^2}}-\frac{L}{v\sqrt{2\pi}}e^{-\frac{(a-\mu-\sqrt{L})^2}{2v^2}}+\int_{a-\sqrt{L}}^{a+\sqrt{L}}2(a-y)\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y\\&\quad-\frac{L}{v\sqrt{2\pi}}e^{-\frac{(a-\mu+\sqrt{L})^2}{2v^2}}+\frac{L}{v\sqrt{2\pi}}e^{-\frac{(a-\mu-\sqrt{L})^2}{2v^2}}\\&=\int_{a-\sqrt{L}}^{a+\sqrt{L}}2(a-y)\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y\\&=\int_{a-\sqrt{L}}^{a}2(a-y)\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y+\int_{a}^{a+\sqrt{L}}2(a-y)\frac{1}{v\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2v^2}}\mathrm{d}y\\&=\int_{0}^{\sqrt{L}}2z\frac{1}{v\sqrt{2\pi}}e^{-\frac{(a-z-\mu)^2}{2v^2}}\mathrm{d}z-\int_{0}^{\sqrt{L}}2z\frac{1}{v\sqrt{2\pi}}e^{-\frac{(a+z-\mu)^2}{2v^2}}\mathrm{d}z\\&=\int_{0}^{\sqrt{L}}\frac{2z}{v\sqrt{2\pi}}(e^{-\frac{(a-z-\mu)^2}{2v^2}}-e^{-\frac{(a+z-\mu)^2}{2v^2}})\mathrm{d}z.\end{aligned}\tag{93}$$

注意，对于 $z\in[0,\sqrt{L}]$,

$$e^{-\frac{(a-z-\mu)^2}{2v^2}}-e^{-\frac{(a+z-\mu)^2}{2v^2}}>0\text{ when }a>\mu,\tag{94}$$

$$e^{-\frac{(a-z-\mu)^2}{2v^2}}-e^{-\frac{(a+z-\mu)^2}{2v^2}}<0\text{ when }a<\mu.\tag{95}$$

因此，当 $a<\mu$ 时，有 $f'(a)<0$；当 $a>\mu$ 时，有 $f'(a)>0$，这意味着 $a=\mu$ 是 $f(a)$ 的唯一极小值点，即

$$\tilde{g}_{B,\beta}(x)=\mathbb{E}_{\mathbb{P}_{B,\beta}(x,y)}[y\mid x]=g_{B,\beta}(x).\tag{96}$$

□

下面的引理表明，截断对超额风险没有显著影响。

引理 B.8. 存在 $c_2=(D^2+1)^3$, 使得

$$\mathrm{Error}_\ell(\hat{B},\hat{\beta})\le\mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)]-\mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)]+\sqrt{\frac{2Lc_2}{\pi}}e^{-\frac{L}{2c_2}}.\tag{97}$$

引理的证明*B.8*。

$$
\begin{aligned}
\text{Error}_\ell(\hat{B},\hat{\beta}) &= \mathbb{E}_{B^*,\beta^*}[\ell(g_{\hat{B},\hat{\beta}}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\ell(g_{B^*,\beta^*}(x),y)] \\
&= \mathbb{E}_{B^*,\beta^*}[\ell(g_{\hat{B},\hat{\beta}}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)] \\
&\quad + \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)] \\
&\quad + \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\ell(g_{B^*,\beta^*}(x),y)] \quad (\leq 0 \text{ since } \tilde{\ell} \leq \ell) \\
&\leq \sup_{B,\beta}\{\mathbb{E}_{B^*,\beta^*}[\ell(g_{B,\beta}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B,\beta}(x),y)]\} \\
&\quad + \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)] \qquad (98)
\end{aligned}
$$

对于第一项，我们有

$$
\begin{aligned}
&\sup_{B,\beta}\{\mathbb{E}_{B^*,\beta^*}[\ell(g_{B,\beta}(x),y)] - \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B,\beta}(x),y)]\} \\
&= \sup_{B,\beta}\{\mathbb{E}_{B^*,\beta^*}((g_{B,\beta}(x)-y)^2 - L)\mathbb{1}_{\{(g_{B,\beta}(x)-y)^2 \geq L\}}\}. \qquad (99)
\end{aligned}
$$

注意到

$$
g_{B,\beta}(x) - y = \beta^T B^T (BB^T + I_d)^{-1} x - y \sim \mathcal{N}(0,\lambda^2), \qquad (100)
$$

其中

$$
\begin{aligned}
\lambda^2 &= Var_{B^*,\beta^*}[g_{B,\beta}(x) - y] \\
&= \mathbb{E}_{B^*,\beta^*}(\beta^T B^T (BB^T + I_d)^{-1} x - y)^2 \\
&= \epsilon^2 + \beta^T B^T (BB^T + I_d)^{-1}(B^* B^{*T} + I_d)(BB^T + I_d)^{-1} B\beta \\
&\quad + \beta^{*T}\beta^* - 2\beta^T B^T (BB^T + I_d)^{-1} B^* \beta^* \\
&\leq \epsilon^2 + \beta^{*T}\beta^* + \|(BB^T + I_d)^{-1}\|_2^2 \cdot \|B^* B^{*T} + I_d\|_2 \cdot \|B\beta\|_2^2 \\
&\quad + 2\|(BB^T + I_d)^{-1}\|_2 \cdot \|B^*\beta^*\|_2 \cdot \|B\beta\|_2 \\
&\leq \epsilon^2 + \beta^{*T}\beta^* + D^4\|B^* B^{*T} + I_d\|_2 + 2D^2\|B^*\beta^*\|_2 \\
&\leq 1 + D^2 + D^4(D^2+1) + 2D^4 \\
&\leq c_2. \qquad (101)
\end{aligned}
$$

因此

$$
\begin{aligned}
&\sup_{B,\beta}\{\mathbb{E}_{B^*,\beta^*}((g_{B,\beta}(x)-y)^2-L)\mathbb{1}_{\{(g_{B,\beta}(x)-y)^2\geq L\}}\}\\
&=\sup_{\lambda}2\int_{\sqrt{L}}^{+\infty}\frac{1}{\lambda\sqrt{2\pi}}(x^2-L)e^{-\frac{x^2}{2\lambda^2}}\mathrm{d}x\\
&=2\sup_{\lambda}\left\{-\frac{\lambda}{\sqrt{2\pi}}xe^{-\frac{x^2}{2\lambda^2}}\Big|_{\sqrt{L}}^{+\infty}+(\lambda^2-L)\int_{\sqrt{L}}^{+\infty}\frac{1}{\lambda\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}\mathrm{d}x\right\}\\
&=2\sup_{\lambda}\left\{\sqrt{\frac{L}{2\pi}}\lambda e^{-\frac{L}{2\lambda^2}}+(\lambda^2-L)\int_{\sqrt{L}}^{+\infty}\frac{1}{\lambda\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}\mathrm{d}x\right\}\\
&\leq 2\sup_{\lambda}\left\{\sqrt{\frac{L}{2\pi}}\lambda e^{-\frac{L}{2\lambda^2}}\right\}\quad(\text{since } L\geq c_2\geq\lambda^2)\\
&=\sqrt{\frac{2Lc_2}{\pi}}e^{-\frac{L}{2c_2}}.
\end{aligned}
\tag{102}
$$

最后一个等式成立，是因为 $\lambda e^{-\frac{L}{2\lambda^2}}$ 关于 λ单调递增，且 $\lambda\leq\sqrt{c_1}$。结合式（98）、（99）和（102），证明得证。 □

现在我们可以证明定理4.3了。

~定理4.3的证明。注意到 l 是 L–有界的。根据引理B.7，我们可以应用定理3.4 于 l，得到

$$
\begin{aligned}
&\mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)]-\mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)]\\
&\leq 2\max_{B\in\mathcal{B}}R_n\left(\tilde{\ell}\circ\mathcal{G}_{B,\mathcal{C}}\right)+L\cdot\sqrt{\frac{2}{n}\log\frac{4}{\delta}}+12\kappa L\cdot\sqrt{\frac{1}{m}\log\frac{2N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{B}),1/m)}{\delta}}.
\end{aligned}
\tag{103}
$$

这里 $\kappa=c_1(\sigma^*_{max}+1)^4/\sigma^{*3}_{min}$ 是引理4.2中定义的可迁移性。

根据引理B.5，我们有

$$
\log N_{[\,]}(\mathcal{P}(\mathcal{B}),1/m)\leq 4dr\log(24mdr(D^2+1)).
\tag{104}
$$

由于 $\tilde{l}$ 关于第一个参数是 $2\sqrt{L}$–利普希茨的，压缩原理（Ledoux & Talagrand（2013）中的定理4.12）给出

$$
R_n\left(\tilde{\ell}\circ\mathcal{G}_{B,\mathcal{C}}\right)\leq 2\sqrt{L}R_n\left(\mathcal{G}_{B,\mathcal{C}}\right).
\tag{105}
$$

因此，只需对 $R_n\ (\mathcal{G}_{B,\ \mathcal{C}})$ 进行界定。 根据引理 B.6，对于固定的 B,

$$\begin{aligned}
R_n\left(\mathcal{G}_{B,\mathcal{C}}\right) &= \mathbb{E}_{\{x_j\}_{j=1}^n}\mathbb{E}_{\{\sigma_j\}_{j=1}^n}[\sup_{\beta}\frac{2}{n}\sum_{j=1}^{n}\sigma_j g_{B,\beta}(x_j)] \\
&= \mathbb{E}_{\{x_j\}_{j=1}^n}\mathbb{E}_{\{\sigma_j\}_{j=1}^n}[\sup_{\beta}\frac{2}{n}\sum_{j=1}^{n}\sigma_j \beta^T B^T(BB^T+I_d)^{-1}x_j] \\
&\leq \mathbb{E}_{\{x_j\}_{j=1}^n}[\frac{2D}{\sqrt{n}}\sup_j\|B^T(BB^T+I_d)^{-1}x_j\|_2] \quad \text{(By Lemma B.6, since } \|\beta\|_2\leq D) \\
&= \frac{2D}{\sqrt{n}}\mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|B^T(BB^T+I_d)^{-1}x_j\|_2]. \qquad (106)
\end{aligned}$$

注意到 $x_j\sim\mathcal{N}(0,B^*B^{*T}+I_d)$。因此 $B^T(BB^T+I_d)^{-1}x_j\sim\mathcal{N}(0,\Sigma)$，其中

$$\Sigma := B^T(BB^T+I_d)^{-1}(B^*B^{*T}+I_d)(BB^T+I_d)^{-1}B. \qquad (107)$$

因此，我们有

$$\Sigma^{-\frac{1}{2}}B^T(BB^T+I_d)^{-1}x_j\sim\mathcal{N}(0,I_r). \qquad (108)$$

令 $u_j := \Sigma^{-\frac{1}{2}}B^T(BB^T+I_d)^{-1}x_j$，则

$$\begin{aligned}
&\mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|B^T(BB^T+I_d)^{-1}x_j\|_2] \\
&= \mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|\Sigma^{\frac{1}{2}}u_j\|_2] \\
&\leq \mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|\Sigma^{\frac{1}{2}}\|_2\|u_j\|_2] \\
&\leq \sup\|\Sigma^{\frac{1}{2}}\|_2\mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|u_j\|_2]. \qquad (109)
\end{aligned}$$

根据 Vershynin （2018年）中的定理 3.1.1， $\|u_j\|-\sqrt{r}$ 对于某个绝对常数 c_4–是次高斯的c_4。因此，对于任意 $t>0$,

$$\begin{aligned}
e^{\mathbb{E}[t\sup_j\|u_j\|_2]} &\leq \mathbb{E}[e^{t\sup_j\|u_j\|_2}] \quad \text{(by Jensen's inequality)} \\
&\leq \sum_{j=1}^{n}\mathbb{E}[e^{t\|u_j\|_2}] \\
&= \sum_{j=1}^{n}\mathbb{E}[e^{t\|u_j\|_2-\sqrt{r}}]e^{t\sqrt{r}} \\
&\leq \sum_{j=1}^{n}e^{\frac{t^2}{2}c_4}e^{t\sqrt{r}} \\
&= ne^{t\sqrt{r}+\frac{t^2}{2}c_4}. \qquad (110)
\end{aligned}$$

两边取对数，得到

$$\mathbb{E}[\sup_j\|u_j\|_2]\leq\frac{\log n}{t}+\sqrt{r}+\frac{t}{2}c_4, \qquad (111)$$

这对任意 $t>0$ 都成立。取 $t=\sqrt{2\log\frac{n}{c_4 r}}$ 得到

$$\mathbb{E}[\sup_j \|u_j\|_2] \le \sqrt{2c_4\log n}+\sqrt{r}. \tag{112}$$

注意到

$$\begin{aligned}\|\Sigma\|_2 &= \|B^T(BB^T+I_d)^{-1}(B^*B^{*T}+I_d)(BB^T+I_d)^{-1}B\|_2\\ &\le \|B\|_2^2\cdot\|(BB^T+I_d)^{-1}\|_2^2\cdot\|B^*B^{*T}+I_d\|\\ &\le (D^2+1)^2,\end{aligned} \tag{113}$$

即，$\sup\|\Sigma^{\frac{1}{2}}\|_2\le(D^2+1)$。结合（106）、（109）、（112）和（113），可得

$$\begin{aligned}R_n\left(\mathcal{G}_{\phi,\Psi}\right) &\le \frac{2D}{\sqrt{n}}\mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|B^T(BB^T+I_d)^{-1}x_j\|_2]\\ &\le \frac{2D}{\sqrt{n}}\sup\|\Sigma^{\frac{1}{2}}\|_2\mathbb{E}_{\{x_j\}_{j=1}^n}[\sup_j\|u_j\|_2]\\ &\le \frac{2D}{\sqrt{n}}(D^2+1)(\sqrt{2c_4\log n}+\sqrt{r}),\end{aligned} \tag{114}$$

这意味着

$$\max_{\phi\in\Phi}R_n\left(\tilde{\ell}\circ\mathcal{G}_{\phi,\Psi}\right)\le 2\sqrt{L}\max_{\phi in\Phi}R_n\left(\mathcal{G}_{\phi,\Psi}\right)\le 2\sqrt{L}\frac{2D}{\sqrt{n}}(D^2+1)(\sqrt{2c_4\log n}+\sqrt{r}) \tag{115}$$

现在我们可以对超额风险进行界定。根据引理 B.8，可得

$$\begin{aligned}\mathrm{Error}_\ell(\hat{B},\hat{\beta}) &\le \mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{\hat{B},\hat{\beta}}(x),y)]-\mathbb{E}_{B^*,\beta^*}[\tilde{\ell}(g_{B^*,\beta^*}(x),y)]+\sqrt{\frac{2Lc_2}{\pi}}e^{-\frac{L}{2c_2}}\\ &\le 2\max_{\phi\in\Phi}R_n\left(\tilde{\ell}\circ\mathcal{G}_{\phi,\Psi}\right)+L\cdot\sqrt{\frac{2}{n}\log\frac{4}{\delta}}\\ &\quad+12\kappa L\cdot\sqrt{\frac{1}{m}\log\frac{2N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{B}),1/m)}{\delta}}+\sqrt{\frac{2Lc_2}{\pi}}e^{-\frac{L}{2c_2}}\\ &\le 4\sqrt{L}\frac{2D}{\sqrt{n}}(D^2+1)(\sqrt{2c_4\log n}+\sqrt{r})+L\cdot\sqrt{\frac{2}{n}\log\frac{4}{\delta}}\\ &\quad+12\kappa L\sqrt{\frac{1}{m}(4dr\log(24mdr(D^2+1))+\log(2/\delta))}+\sqrt{\frac{2Lc_2}{\pi}}e^{-\frac{L}{2c_2}},\end{aligned} \tag{116}$$

其中第二个不等式由（103）得出，最后一个不等式由（104）、（115）得出。这里 c_4 是一个绝对常数。注意 $c_2=(D^2+1)^3$ 以及 $L=c_2\log n$。因此，我们有

$$\begin{aligned}\mathrm{Error}_\ell(\hat{B},\hat{\beta}) &\le 8\sqrt{2c_4}L\sqrt{\frac{1}{n}}+8L\sqrt{\frac{r}{n}}+L\cdot\sqrt{\frac{2}{n}\log\frac{4}{\delta}}\\ &\quad+12\kappa L\sqrt{\frac{1}{m}(4dr\log(24mdr(D^2+1))+\log(2/\delta))}+L\sqrt{\frac{2}{\pi n}}\\ &\le \tilde{\mathcal{O}}\left(\kappa L\sqrt{\frac{dr}{m}}+L\sqrt{\frac{r}{n}}\right),\end{aligned} \tag{117}$$

w这里 $L = (D^2+1)^3 \log n$，并且 $\kappa = c_1(\sigma^*_{max}+1)^4/\sigma^{*3}_{min}$，其中 c_1 为某些绝对常数。

□

B.5 定理 4.4的证明

本节将对证明定理 4.4进行精细分析。首先注意到，我们可以将我们的模型（不含 z）重写为

$$y = \beta^{*T} C^* x + w, \tag{118}$$

其中 $\beta^* \in \mathbb{R}^{r\times 1}$、$C^* = B^{*T}(B^*B^{*T}+I_d)^{-1} \in \mathbb{R}^{r\times d}$、$x \sim N(0, B^*B^{*T}+I_d)$、$w \sim N(0, \epsilon^2 + \|\beta^*\|_2^2 - \beta^{*T}B^{*T}(B^*B^{*T}+I_d)^{-1}B^*\beta^*)$。这里 w 和 x 是独立的。因此，我们可以将数据写为

$$Y = XC^{*T}\beta^* + W, \tag{119}$$

其中 $Y = (y_1, \cdots, y_n)^T \in \mathbb{R}^{n\times 1}$、$X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d}$、$W = (w_1, \cdots, w_n)^T \in \mathbb{R}^{n\times 1}$。

在第一步（MLE）中，我们得到 C^T 的估计量 $\hat{C}^T = (\hat{B}^T, I_d)^{-1}$。然后，第二步（经验风险最小化）中的估计量 $\hat{\beta}$ 为

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} \|Y - X\hat{C}^T\beta\|_2^2 \\
&= ((X\hat{C}^T)^T(X\hat{C}^T))^{-1}(X\hat{C}^T)^T Y \\
&= (\hat{C}X^TX\hat{C}^T)^{-1}\hat{C}X^TY.
\end{aligned} \tag{120}$$

于是，我们的风险为

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{B}, \hat{\beta}) &= \mathbb{E}_{\mathbb{P}_{B^*,\beta^*}(x,y)}\left[\left(y - g_{\hat{B},\hat{\beta}}(x)\right)^2\right] - \mathbb{E}_{\mathbb{P}_{B^*,\beta^*}(x,y)}\left[\left(y - g_{B^*,\beta^*}(x)\right)^2\right] \\
&= \mathbb{E}[(\beta^{*T}C^*x + w - \hat{\beta}^T\hat{B}^T(\hat{B}\hat{B}^T + I_d)^{-1}x)^2] - \mathbb{E}[w^2] \\
&= \mathbb{E}[(\beta^{*T}C^*x - \hat{\beta}^T\hat{C}x)^2] \\
&= (\beta^{*T}C^* - \hat{\beta}^T\hat{C})(B^*B^{*T} + I_d)(\beta^{*T}C^* - \hat{\beta}^T\hat{C})^T \\
&\le \|B^*B^{*T} + I_d\|_2 \|\hat{C}^T\hat{\beta} - C^{*T}\beta^*\|_2^2
\end{aligned} \tag{121}$$

我们的目标是界定 $\|\hat{C}^T\hat{\beta} - C^{*T}\beta^*\|_2^2$。考虑 C^{*T} 和 $\hat{C}^T$ 的奇异值分解（SVD），即 $C^{*T} = U^*\Lambda^*V^{*T}$、$\hat{C}^T = \hat{U}\hat{\Lambda}\hat{V}^T$。于是，我们有

$$\begin{aligned}
&\hat{C}^T\hat{\beta} - C^{*T}\beta^* \\
&= \hat{C}^T(\hat{C}X^TX\hat{C}^T)^{-1}\hat{C}X^TY - C^{*T}\beta^* \\
&= \hat{C}^T(\hat{C}X^TX\hat{C}^T)^{-1}\hat{C}X^T(XC^{*T}\beta^* + W) - C^{*T}\beta^* \\
&= (\hat{C}^T(\hat{C}X^TX\hat{C}^T)^{-1}\hat{C}X^TXC^{*T} - C^{*T})\beta^* + \hat{C}^T(\hat{C}X^TX\hat{C}^T)^{-1}\hat{C}X^TW \\
&= (\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TXU^* - U^*)\Lambda^*V^{*T}\beta^* + \hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW.
\end{aligned} \tag{122}$$

因此

$$\begin{aligned}
\|\hat{C}^T\hat{\beta} - C^{*T}\beta^*\|_2^2 \le\ & 2\|(\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TXU^* - U^*)\|_2^2\|\Lambda^*\|_2^2\|\beta^*\|_2^2 \\
& + 2\|\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW\|_2^2
\end{aligned} \tag{123}$$

我们给出用于界定相关项的两个引理。第一个引理考虑偏差项 ：

引理 B.9. 设 $\Sigma := B^* B^{*T} + I_d$。若 $n \gtrsim \|\Sigma\|^2 r\log(1/\delta)$，则以至少 $1-\delta$ 的概率，

$$\|(\hat{U}(\hat{U}^T X^T X \hat{U})^{-1}\hat{U}^T X^T X U^* - U^*)\|_2^2 \le \mathcal{O}(\|\Sigma\|^2\Delta^2), \tag{124}$$

其中 $\Delta = dist(\hat{U}, U^*) := \|\hat{U}\hat{U}^T - U^* U^{*T}\|$。

第二个引理考虑方差项：

引理 B.10. 设 $\Sigma := B^* B^{*T} + I_d$。如果 $n \gtrsim \|\Sigma\|^2 r\log(1/\delta)$，则以至少 $1-\delta$ 的概率，

$$\|\hat{U}(\hat{U}^T X^T X \hat{U})^{-1}\hat{U}^T X^T W\|_2^2 \le \mathcal{O}\left(\frac{\sigma^2 r\log(4/\delta)}{n}\right), \tag{125}$$

其中 $\sigma^2 := \mathbb{E}(w^2) = \epsilon^2 + \|\beta^*\|_2^2 - \beta^{*T} B^{*T}(B^* B^{*T} + I_d)^{-1} B^* \beta^*$ 是 w 的方差。

结合这两个引理以及分解（123），我们有

$$\|\hat{C}^T\hat{\beta} - C^{*T}\beta^*\|_2^2 \le \mathcal{O}\left(\|\beta^*\|^2\|\Lambda^*\|^2\|\Sigma\|^2\Delta^2 + \frac{\sigma^2 r\log(4/\delta)}{n}\right). \tag{126}$$

现在还需控制 Δ，它与第一步（最大似然估计）的估计误差有关。以下引理给出了 Δ 的一个上界。

引理 B.11。如果 $m \gtrsim \|\Sigma\|^2 d\log(1/\delta)$，则以至少 $1-\delta$ 的概率，

$$\Delta^2 \le \mathcal{O}\left(\|\Sigma\|^2\frac{d\log(1/\delta)}{m}\lambda_r^{-2}(C^{*T}C^*)\right), \tag{127}$$

其中，$\lambda_r(C^{*T}C^*)$ 是 $C^{*T}C^*$ 的第 r 个（最小的）非零特征值。

定理 *4.4* 的证明。根据引理 B.9、B.10、B.11，我们有

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{B},\hat{\beta}) &\le \|\Sigma\|\|\hat{C}^T\hat{\beta} - C^{*T}\beta^*\|_2^2 \\
&\le \mathcal{O}(\|\beta^*\|^2\|\Lambda^*\|^2\|\Sigma\|^3\Delta^2 + \|\Sigma\|\frac{\sigma^2 r\log(4/\delta)}{n}). \\
&\le \mathcal{O}(\|\beta^*\|^2\|\Lambda^*\|^2\|\Sigma\|^5\lambda_r^{-2}(C^{*T}C^*)\frac{d\log(1/\delta)}{m} + \|\Sigma\|\frac{\sigma^2 r\log(4/\delta)}{n}).
\end{aligned} \tag{128}$$

利用 $\|\beta^*\| \le D$ 和 $\|B^*\| \le D$ 这两个假设，我们可以将这些项界定为 D 以及与真实值相关的量。首先注意到，Σ 的特征值为 $\sigma_1^{*2}+1 \ge \sigma_2^{*2}+1 \ge \cdots \ge \sigma_r^{*2}+1 \ge 1 = \cdots = 1$，其中 σ_i^* 是 B^* 的奇异值，因此 $\|\Sigma\| \le D^2+1$。此外，由于

$$\begin{aligned}
C^{*T}C^* &= (B^* B^{*T} + I_d)^{-1} B^* B^{*T}(B^* B^{*T} + I_d)^{-1} \\
&= (B^* B^{*T} + I_d)^{-1} - (B^* B^{*T} + I_d)^{-2} \\
&= \Sigma^{-1} - \Sigma^{-2},
\end{aligned} \tag{129}$$

我们知道，$C^{*T}C^*$ 有 r 个非零特征值 $\{(\sigma_i^* + \sigma_i^{*-1})^{-2}\}_{i=1}^r$。因此 $\|\Lambda^*\|^2 = \|C^{*T}C^*\| \le 1/4$，

$$\begin{aligned}
\lambda_r^{-2}(C^{*T}C^*) &\le \max((\sigma_1^* + \sigma_1^{*-1})^4, (\sigma_r^* + \sigma_r^{*-1})^4) \\
&\le \mathcal{O}(D^4 + \sigma_r^{*-4}).
\end{aligned} \tag{130}$$

对于 σ^2，我们有

$$\begin{aligned}\sigma^2 &= \epsilon^2 + \|\beta^*\|_2^2 - \beta^{*T}B^{*T}(B^*B^{*T} + I_d)^{-1}B^*\beta^* \\ &\leq 1 + \|\beta^*\|^2\|I_r - B^{*T}(B^*B^{*T} + I_d)^{-1}B^*\| \\ &\leq 1 + D^2. \end{aligned} \tag{131}$$

综合以上所有界，我们得到

$$\begin{aligned}\text{Error}_\ell(\hat{B}, \hat{\beta}) &\leq \mathcal{O}(\|\beta^*\|^2\|\Lambda^*\|^2\|\Sigma\|^5\lambda_r^{-2}(C^{*T}C^*)\frac{d\log(1/\delta)}{m} + \|\Sigma\|\frac{\sigma^2 r\log(4/\delta)}{n}). \\ &\leq \mathcal{O}((D^2+1)^6(D^4 + \sigma_{min}^{*-4})\frac{d\log(1/\delta)}{m} + (D^2+1)^2\frac{r\log(4/\delta)}{n}). \end{aligned} \tag{132}$$

□

接下来，我们给出引理 B.9、B.10 和 B.11的证明。我们首先证明一些额外的技术引理。以下引理是 Tripuraneni 等人 (2021年) 引理 20 的一个简单推论，它展示了经验协方差矩阵的集中性质。

引理 B.12. 设 $\Sigma \in \mathbb{R}^d$ 为一个正定矩阵。设 $\{x_i\}_{i=1}^n$ 为从 $N(0, \Sigma)$ 独立同分布抽取的 $d-$ 维高斯随机向量样本， $X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d}$。那么，对于任意 $A, B \in \mathbb{R}^{d\times r}$，以至少 $1-\delta$ 的概率，我们有

$$\|A^T(\frac{X^TX}{n})B - A^T\Sigma B\|_2 \leq \mathcal{O}(\|A\|\|B\|\|\Sigma\|(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}). \tag{133}$$

证明. 我们写出 A 和 B 的奇异值分解： $A = U_1\Lambda_1V_1^T$, $B = U_2\Lambda_2V_2^T$，其中 $U_1, U_2 \in \mathbb{R}^{d\times r}$,$\Lambda_1$, $\Lambda_2, V_1, V_2 \in \mathbb{R}^{r\times r}$。于是

$$\begin{aligned}\|A^T(\frac{X^TX}{n})B - A^T\Sigma B\|_2 &= \|V_1\Lambda_1U_1^T(\frac{X^TX}{n})U_2\Lambda_2V_2^T - V_1\Lambda_1U_1^T\Sigma U_2\Lambda_2V_2^T\|_2 \\ &\leq \|V_1\Lambda_1\|\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\|\|\Lambda_2V_2^T\| \\ &\leq \|A\|\|B\|\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\|. \end{aligned} \tag{134}$$

现在，由于 U_1、$U_2 \in \mathbb{R}^{d\times r}$ 是投影矩阵，我们可以应用 Tripuraneni 等人 （2021年）的引理 20，因此

$$\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\| \leq \mathcal{O}(\|\Sigma\|(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})) \tag{135}$$

这就得到了我们想要的结果。 □

下面的引理是一个基本的矩阵扰动结果（参见Tripuraneni 等人（2021年）的引理 25）。

引理 B.13. 设 A 为一个正定矩阵， E 为另一个满足 $\|EA^{-1}\| \leq_{14}$的矩阵，则 $F := (A+E)^{-1} - A^{-1}$满足 $\|F\| \leq \frac{4}{3}\|A^{-1}\|\|EA^{-1}\|$。

利用这两个技术引理，我们能够证明引理 B.9、B.10。

引理 B.9 的证明。我们考虑 $\hat{U} \in \mathbb{R}^{d\times r}$ 和 $\hat{U}_\perp^T \in \mathbb{R}^{d\times(d-r)}$ 为正交归一投影矩阵，分别张成秩为 r 和秩为 $d-r$ 的正交子空间，因此 range($\hat{U}$) $\oplus$range($\hat{U}_\perp$) $= \mathbb{R}^d$.。于是 $\Delta = dist(\hat{U}, U^*) = \|\hat{U}_\perp^T U^*\|_2$（见 Chen 等人 (2021) 引理 2.5）。注意到 $I_d = \hat{U}\hat{U}^T + \hat{U}_\perp\hat{U}_\perp^T$，我们有

$$\begin{aligned}
&\hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T XU^* - U^* \\
&= \hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X(\hat{U}\hat{U}^T + \hat{U}_\perp\hat{U}_\perp^T)U^* - U^* \\
&= \hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}\hat{U}^T U^* + \hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}_\perp\hat{U}_\perp^T U^* - U^* \\
&= \hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}_\perp\hat{U}_\perp^T U^* + \hat{U}\hat{U}^T U^* - U^* \\
&= \hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}_\perp\hat{U}_\perp^T U^* - \hat{U}_\perp\hat{U}_\perp^T U^*
\end{aligned} \tag{136}$$

因此

$$\|\hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T XU^* - U^*\|_2^2 \le 2\|\hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}_\perp\hat{U}_\perp^T U^*\|_2^2 + 2\|\hat{U}_\perp\hat{U}_\perp^T U^*\|_2^2. \tag{137}$$

对于第二项,

$$\|\hat{U}_\perp\hat{U}_\perp^T U^*\|_2^2 \le \|\hat{U}_\perp|^2\|\hat{U}_\perp^T U^*\|^2 \le \Delta^2. \tag{138}$$

对于第一项,

$$\begin{aligned}
&\|\hat{U}(\hat{U}^T X^T X\hat{U})^{-1}\hat{U}^T X^T X\hat{U}_\perp\hat{U}_\perp^T U^*\| \\
&= \|\hat{U}(\hat{U}^T\frac{X^T X}{n}\hat{U})^{-1}\hat{U}^T\frac{X^T X}{n}\hat{U}_\perp\hat{U}_\perp^T U^*\| \\
&= \|\hat{U}((\hat{U}^T\Sigma\hat{U})^{-1} + F)(\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^T U^* + E_1)\| \\
&\le \|(\hat{U}^T\Sigma\hat{U})^{-1}(\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^T U^*)\| + \|(\hat{U}^T\Sigma\hat{U})^{-1}E_1\| + \|F\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^T U^*\| + \|FE_1\|,
\end{aligned} \tag{139}$$

其中 $E_1 = \hat{U}^T\frac{X^T X}{n}\hat{U}_\perp\hat{U}_\perp^T U^* - \hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^T U^*$, $F = (\hat{U}^T\frac{X^T X}{n}\hat{U})^{-1} - (\hat{U}^T\Sigma\hat{U})^{-1}$。为了界定 $\|F\|$，令 $E = \hat{U}^T\frac{X^T X}{n}\hat{U} - \hat{U}^T\Sigma\hat{U}$，则根据引理 B.12，以至少 $1-\delta$ 的概率,

$$\|E\| \le \mathcal{O}(\|\Sigma\|(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})). \tag{140}$$

因此，由于 $\lambda_{min}(\Sigma) = 1$,

$$\begin{aligned}
\|E(\hat{U}^T\Sigma\hat{U})^{-1}\| &\le \|E\|\|(\hat{U}^T\Sigma\hat{U})^{-1}\| \\
&\le \|E\|\lambda_{min}(\Sigma)^{-1} \\
&\le \mathcal{O}(\|\Sigma\|(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}))
\end{aligned} \tag{141}$$

注意到 $n \gtrsim \|\Sigma\|^2 r\log(1/\delta)$ 意味着 $\sqrt{\frac{r}{n}}+\frac{r}{n}+\sqrt{\frac{\log(1/\delta)\,n}{}}+\frac{\log(1/\delta)}{n} \lesssim \|\Sigma\|^{-1}$。因此，我们证明，当 n 足够大时，有 $\|E(\hat{U}^T\Sigma\hat{U})^{-1}\| \le \frac{1}{4}$。因此，我们可以应用引理 B.13，得到

$$\begin{aligned}\|F\| &\le \frac{4}{3}\|E(\hat{U}^T\Sigma\hat{U})^{-1}\|\|(\hat{U}^T\Sigma\hat{U})^{-1}\| \\ &\le \frac{4}{3}\times\frac{1}{4}\|(\hat{U}^T\Sigma\hat{U})^{-1}\| \\ &\le \frac{1}{3}.\end{aligned} \tag{142}$$

对于 $\|E_1\|$，直接应用引理 B.12，使用 $n \gtrsim \|\Sigma\|^2 r\log(1/\delta)$，可得

$$\begin{aligned}\|E_1\| &\le \mathcal{O}(\|\Sigma\|\|\hat{U}_\perp\hat{U}_\perp^T U^*\|(\sqrt{\frac{r}{n}}+\frac{r}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n})) \\ &\le \mathcal{O}(\|\Sigma\|\Delta\|\Sigma\|^{-1}) \\ &\le \mathcal{O}(\Delta)\end{aligned} \tag{143}$$

结合（139）、（142）和（143），可得

$$\begin{aligned}&\|\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TX\hat{U}_\perp\hat{U}_\perp^TU^*\| \le \|(\hat{U}^T\Sigma\hat{U})^{-1}(\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^TU^*)\| + \|(\hat{U}^T\Sigma\hat{U})^{-1}E_1\| + \|F\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^TU^*\| + \|FE_1\| \le \|(\hat{U}^T\Sigma\hat{U})^{-1}\|\|(\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^TU^*)\| + \|(\hat{U}^T\Sigma\hat{U})^{-1}\|\|E_1\| + \|F\|\|\hat{U}^T\Sigma\hat{U}_\perp\hat{U}_\perp^TU^*\| + \|F\|\|E_1\| \le \lambda_{min}(\Sigma)^{-1}\|\Sigma\|\|\hat{U}_\perp^TU^*\| + \lambda_{min}(\Sigma)^{-1}\|E_1\| + \|F\|\|\Sigma\|\|\hat{U}_\perp^TU^*\| + \|F\|\|E_1\| \\ &\le \lambda_{min}(\Sigma)^{-1}\|\Sigma\|\Delta + \lambda_{min}(\Sigma)^{-1}\mathcal{O}(\lambda_{min}(\Sigma)\Delta) + \frac{1}{3}\lambda_{min}(\Sigma)^{-1}\|\Sigma\|\Delta + \frac{1}{3}\lambda_{min}(\Sigma)^{-1}\mathcal{O}(\lambda_{min}(\Sigma)\Delta) \\ &\le \mathcal{O}(\|\Sigma\|\Delta)\end{aligned} \tag{144}$$

最后，结合（137）、（138）和（144），即可得到

$$\|(\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TXU^*-U^*)\|_2^2 \le \mathcal{O}(\|\Sigma\|^2\Delta^2), \tag{145}$$

以至少 $1-\delta$ 的概率，这正是我们想要的结果。 □

引理的证明*B.10*。

$$\begin{aligned}\|\hat{U}(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW\|_2^2 &\le \|(\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW\|_2^2 \\ &= ((\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW)^T((\hat{U}^TX^TX\hat{U})^{-1}\hat{U}^TX^TW) \\ &= W^T(\frac{1}{n}\frac{X\hat{U}}{\sqrt{n}}(\hat{U}^T\frac{X^TX}{n}\hat{U})^{-2}\frac{\hat{U}^TX^T}{\sqrt{n}})W.\end{aligned} \tag{146}$$

令 $A=\frac{1}{n}\frac{X\hat{U}}{\sqrt{n}}(\hat{U}^T\frac{X^TX}{n}\hat{U})^{-2}\frac{\hat{U}^TX^T}{\sqrt{n}}$, $W=\sigma V$, 则 $V\sim N(0,I_n)$。根据 Hanson-Wright 不等式（见 Vershynin (2018年) 定理 6.2.1），

$$\mathbb{P}(|V^TAV-\mathbb{E}[V^TAV]|\ge t) \le 2\exp(-c\min(\frac{t^2}{\|A\|_F^2},\frac{t}{\|A\|_2})). \tag{147}$$

因此，以至少 $1-\delta$ 的概率，

$$V^TAV \le \mathbb{E}[V^TAV] + \mathcal{O}(\|A\|_F\sqrt{\log\frac{2}{\delta}}) + \mathcal{O}(\|A\|_2\log\frac{2}{\delta}). \tag{148}$$

注意到 $\mathbb{E}[V^TAV] = \mathbf{Tr}(A)$，因此只需对 $\mathbf{Tr}(A)$、$\|A\|_F$ 和 $\|A\|_2$ 进行界定。若定义 $B = \frac{XU}{\sqrt{n}} \in \mathbb{R}^{n\times r}$，则 $A = \frac{1}{n}B(B^TB)^{-2}B^T$。因此

$$\begin{aligned}\mathrm{Tr}(A) &= \mathrm{Tr}(\frac{1}{n}B(B^TB)^{-2}B^T)\\ &= \frac{1}{n}\mathrm{Tr}((B^TB)^{-2}B^TB)\\ &= \frac{1}{n}\mathrm{Tr}((B^TB)^{-1})\\ &\le \frac{r}{n}\|(B^TB)^{-1}\|_2\end{aligned} \tag{149}$$

设 B 的奇异值分解为 $B = PMQ^T$，其中 $P\in\mathbb{R}^{n\times r}$、$M$、$Q\in\mathbb{R}^{r\times r}$，则

$$\begin{aligned}\|A\|_2 &= \frac{1}{n}\|B(B^TB)^{-2}B^T\|_2\\ &= \frac{1}{n}\|PMQ^T(QM^2Q^T)^{-2}QMP^T\|_2\\ &= \frac{1}{n}\|PM^{-2}P^T\|_2\\ &\le \frac{1}{n}\|M^{-2}\|_2\\ &= \frac{1}{n}\|(B^TB)^{-1}\|_2\end{aligned} \tag{150}$$

还注意到 A 的秩为 r，因此 $\|A\|_F \le \sqrt{r}\|A\|_2$。所以只需对 $\|(B^TB)^{-1}\|_2 = \|(\hat{U}^T\frac{X^TX}{n}\hat{U})^{-1}\|_2$ 进行界定。令 $F = (\hat{U}^T\frac{X^TX}{n}\hat{U})^{-1} - (\hat{U}^T\Sigma\hat{U})^{-1}$。回顾 (142)，其指出以至少 $1-\delta$ 的概率，有 $\|F\| \le \frac{1}{3}\lambda_{min}(\Sigma)^{-1}$。因此

$$\begin{aligned}\|(\hat{U}^T\frac{X^TX}{n}\hat{U})^{-1}\| &= \|(\hat{U}^T\Sigma\hat{U})^{-1} + F\|\\ &\le \|(\hat{U}^T\Sigma\hat{U})^{-1}\| + \|F\|\\ &\le \mathcal{O}(\lambda_{min}(\Sigma)^{-1}).\end{aligned} \tag{151}$$

因此 $\|A\| \le \mathcal{O}(\frac{1}{n}\lambda_{min}(\Sigma)^{-1})$，$\|A\|_F \le \mathcal{O}(\frac{\sqrt{r}}{n}\lambda_{min}(\Sigma)^{-1})$, $\mathbf{Tr}(A) \le \mathcal{O}(\frac{r}{n}\lambda_{min}(\Sigma)^{-1})$。因此，以至少 $1-2\delta$ 的概率，

$$\begin{aligned}V^TAV &\le \mathbb{E}[V^TAV] + \mathcal{O}(\|A\|_F\sqrt{\log\frac{2}{\delta}}) + \mathcal{O}(\|A\|_2\log\frac{2}{\delta})\\ &\le \mathcal{O}(\frac{r}{n}\lambda_{min}(\Sigma)^{-1}) + \mathcal{O}(\frac{\sqrt{r}}{n}\lambda_{min}(\Sigma)^{-1}\sqrt{\log\frac{2}{\delta}}) + \mathcal{O}(\frac{1}{n}\lambda_{min}(\Sigma)^{-1}\log\frac{2}{\delta})\\ &\le \mathcal{O}(\frac{r}{n}\lambda_{min}(\Sigma)^{-1}\log\frac{2}{\delta})\\ &= \mathcal{O}(\frac{r}{n}\log\frac{2}{\delta}).\end{aligned} \tag{152}$$

最后一行成立是因为 $\lambda_{min}(\Sigma) = 1$。回想

$$\|\hat{U}(\hat{U}^T X^T X \hat{U})^{-1}\hat{U}^T X^T W\|_2^2 = W^T A W = \sigma^2 V^T A V, \tag{153}$$

将此结果与上述针对 $V^T AV$的界结合，即可得到我们所需的结果。 □

最后，我们证明以下B.11 引理。

引理*B.11*的证明。第一步中，我们有 m 无标签数据 $\{x_i\}_{i=1}^m$ 来自$N(0,\Sigma)$ 的独立同分布样本。令 $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m x_i x_i^T$为经验协方差矩阵。根据引理B.12，以至少 $1-\delta$的概率，

$$\|\Sigma - \hat{\Sigma}\| \le \mathcal{O}(\|\Sigma\|(\sqrt{\frac{d}{m}} + \frac{d}{m} + \sqrt{\frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m})) \tag{154}$$

我们断言

$$\|\hat{B}\hat{B}^T - (\hat{\Sigma} - I_d)\|_2 \le \|\hat{\Sigma} - \Sigma\|, \tag{155}$$

该断言的证明将在本节末尾给出。有了这个断言，

$$\begin{aligned}\|\hat{B}\hat{B}^T - B^* B^{*T}\| &= \|\hat{B}\hat{B}^T - (\hat{\Sigma} - I_d) + (\hat{\Sigma} - I_d) - (\Sigma - I_d)\| \\ &\le \|\hat{B}\hat{B}^T - (\hat{\Sigma} - I_d)\| + \|\Sigma - \hat{\Sigma}\| \\ &\le 2\|\Sigma - \hat{\Sigma}\|. \end{aligned} \tag{156}$$

注意到

$$\begin{aligned} C^{*T} C^* &= (B^* B^{*T} + I_d)^{-1} B^* B^{*T} (B^* B^{*T} + I_d)^{-1} \\ &= (B^* B^{*T} + I_d)^{-1} - (B^* B^{*T} + I_d)^{-2} \end{aligned} \tag{157}$$

类似地

$$\hat{C}^T \hat{C} = (\hat{B}\hat{B}^T + I_d)^{-1} - (\hat{B}\hat{B}^T + I_d)^{-2}. \tag{158}$$

令 $E_2 = (\hat{B}\hat{B}^T + I_d) - (B^* B^{*T} + I_d)$，$F_2 = (\hat{B}\hat{B}^T + I_d)^{-1} - (B^* B^{*T} + I_d)^{-1}$。则

$$\|E_2\| \le 2\|\Sigma - \hat{\Sigma}\| \le \mathcal{O}(\|\Sigma\|(\sqrt{\frac{d}{m}} + \frac{d}{m} + \sqrt{\frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m})). \tag{159}$$

因此，当 $m \gtrsim \|\Sigma\|^2 d\log(1/\delta)$，$\|E_2\| \le \mathcal{O}(\|\Sigma\|\sqrt{d\log(1/\delta)/m})$，$\|E_2\Sigma^{-1}\| \le \|E_2\|\|\Sigma^{-1}\| \le 1/4$。

接下来，我们可以应用引理 B.13，得到

$$\begin{aligned}\|F_2\| &\le \frac{4}{3}\|\Sigma^{-1}\|\|E_2\Sigma^{-1}\| \\ &\le \frac{4}{3}\|\Sigma^{-1}\|^2\|E_2\| \\ &\le \mathcal{O}(\lambda_{min}^{-2}(\Sigma)\|\Sigma\|\sqrt{\frac{d\log(1/\delta)}{m}}) \\ &= \mathcal{O}(\|\Sigma\|\sqrt{\frac{d\log(1/\delta)}{m}}). \end{aligned} \tag{160}$$

最后一行成立，因为 $\lambda_{min}(\Sigma) = 1$。因此

$$\begin{aligned}\|C^{*T}C^* - \hat{C}^T\hat{C}\| &= \|(\Sigma^{-1} + F_2) - (\Sigma^{-1} + F_2)^2 - (\Sigma^{-1} - \Sigma^{-2})\| \\ &= \|F_2 - \Sigma^{-1}F_2 - F_2\Sigma^{-1} - F_2^2\| \\ &\leq \|F_2\| + 2\|\Sigma^{-1}\|\|F_2\| + \|F_2\|^2 \\ &\leq \mathcal{O}(\|\Sigma\|\sqrt{\frac{d\log(1/\delta)}{m}}).\end{aligned} \tag{161}$$

因此，根据 Davis-Kahan 定理，

$$\Delta = dist(U^*, \hat{U}) \leq \mathcal{O}(\lambda_r^{-1}(C^{*T}C^*)\|C^{*T}C^* - \hat{C}^T\hat{C}\|). \tag{162}$$

结合以上三个不等式，我们得到

$$\Delta^2 \leq \mathcal{O}(\|\Sigma\|^2\frac{d\log(1/\delta)}{m}\lambda_r^{-2}(C^{*T}C^*)). \tag{163}$$

最后，我们需要证明这一结论（155）。注意，最大似然估计量 $\hat{B}$ 由下式给出

$$\begin{aligned}\hat{B} &= \arg\max_{B\in\mathbb{R}^{d\times r}} \sum_{i=1}^m p_B(x_i) \\ &= \arg\max_{B\in\mathbb{R}^{d\times r}} (-\log\det(BB^T + I_d) - \mathrm{Tr}(\hat{\Sigma}(BB^T + I_d)^{-1})) \\ &= \arg\min_{B\in\mathbb{R}^{d\times r}} (\log\det(BB^T + I_d) + \mathrm{Tr}(\hat{\Sigma}(BB^T + I_d)^{-1}))\end{aligned} \tag{164}$$

令 $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$和 $(BB^T + I_d) = U\Lambda U^T$，其中 $\hat{U}$和 U为正交矩阵，$\hat{\Lambda} =$diag$(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$, $\Lambda =$diag$(\lambda_1, \cdots, \lambda_d)$ 且 $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_d$, $\lambda_1 \geq \ldots \geq \lambda_d$。由于 rank$(BB^T) \leq r$，我们有 $\lambda_{r+1} = \ldots\lambda_d = 1$。根据 Ruhe迹不等式（参见 Marshall 等人（2011）第341页），我们有

$$\mathrm{Tr}(\hat{\Sigma}(BB^T + I_d)^{-1})) \geq \sum_{j=1}^d \lambda_j^{-1}\hat{\lambda}_j, \tag{165}$$

且等号成立当且仅当这两个矩阵具有同时有序的谱分解，即 $U = U$。因此

$$\begin{aligned}&\min_{B\in\mathbb{R}^{d\times r}} (\log\det(BB^T + I_d) + \mathrm{Tr}(\hat{\Sigma}(BB^T + I_d)^{-1})) \\ =& \min_{\{\lambda_j\}_{j=1}^d} \sum_{j=1}^d (\log\lambda_j + \lambda_j^{-1}\hat{\lambda}_j) \quad \text{subject to } \lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} = \cdots = \lambda_d = 1\end{aligned} \tag{166}$$

$\hat{}$ and the minimum is achieved when $\lambda_j = \lambda_j$, for $j = 1, \cdots, r$. Therefore the 最大似然估计量 $\hat{}$ B satisfies ($\hat{}_B\hat{}B^T + I_d) = \hat{}U\Lambda\hat{}U^T$where $\Lambda =$diag$(\hat{}\lambda_1, \cdots, \hat{}\lambda_r, 1, \cdots, 1)$. Thus, we have $\hat{}_B\hat{}B^T =$

$\hat{U}(\Lambda - I_d)\hat{U}^T$，这意味着

$$\begin{aligned}
&\|\hat{B}\hat{B}^T - (\hat{\Sigma} - I_d)\|_2 \\
&= \|\hat{U}(\Lambda - I_d)\hat{U}^T - \hat{U}(\hat{\Lambda} - I_d)\hat{U}^T\| \\
&\leq \|\Lambda - \hat{\Lambda}\| \\
&= \max_{j=r+1,\cdots,d} |\hat{\lambda}_j - 1| \\
&\leq \max_{j=1,\cdots,d} |\hat{\lambda}_j - \lambda_j(\Sigma)| \\
&\leq \|\hat{\Sigma} - \Sigma\|. \qquad (167)
\end{aligned}$$

这里最后一个不等式由韦尔定理得出。因此，我们证明了断言（155）。 □

C 第 5 节的证明5

在第 C.1 节中，我们证明高斯混合模型以分类为下游任务时，对于某些绝对常数 c_2，具有 -可迁移性（引理 5.2）。在第 C.2 节和第 C.3 节中，我们证明两个将在定理 5.3 的证明中用到的引理。具体而言，在第 C.2 节中，我们利用 $\mathcal{P}(\mathcal{U})$ 离散化，对集合 ϵ 的括号覆盖数给出上界（引理 C.5）。在第 C.3 节中，我们证明引理 C.6，该引理将用于给函数类$\ell \circ \mathcal{G}_{\mathbf{u}}$的拉德马赫复杂度$\Psi$ 给出上界。最后，在第 C.4 节中，我们证明定理 5.3。

C.1 引理 5.2 的证明

在开始证明该定理之前，我们先给出一些基本定义和有用的引理。我们将以每个 u_i^* 和 u_i 为中心、半径为 $8\sqrt{d}\log K$ 的球定义为

$$\Omega_i^* := \left\{x \in \mathbb{R}^d \mid \|x - u_i^*\| \leq 8\sqrt{d\log K}\right\} \qquad (168)$$

$$\Omega_i := \left\{x \in \mathbb{R}^d \mid \|x - u_i\| \leq 8\sqrt{d\log K}\right\} \qquad (169)$$

我们分别用 P_i 和 P_i^* 表示 $\mathcal{N}(u_i, I_d)$ 和 $\mathcal{N}(u_i^*, I_d)$ 的概率密度函数。

引理 C.1. 若

$$d_{\mathrm{TV}}\left(p_{\mathbf{u}}(x), p_{\mathbf{u}^*}(x)\right) \leq \frac{1}{4K}, \qquad (170)$$

则存在 $\boldsymbol{u}$ 的一个排列，使得对每个 $1 \leq i \leq K$， $\|u_i^* - u_i\| \leq \overline{16\sqrt{d}\log K}$ 均成立。

Bef在证明引理 C.1之前，我们先陈述一个关于高斯范数集中的有用结果。

引理 C.2. 令 $X \sim \mathcal{N}(0, I_d)$，则

$$\mathbb{P}(\|X\| \geq t) \leq 2\exp(-\frac{t^2}{16d}). \qquad (171)$$

证明. 这是 Jin 等人 (2019年) 引理 1.3 的一个简单应用。注意到 X 是 1-次高斯，因此在 $\sigma=\sqrt{d}$ 中采用 Jin 等人 (2019年) 引理 1.3，即可得到所需结论。 □

引理的证明 C.1。我们用反证法证明。如果结论不成立，由于分离度满足 $100\sqrt{d}\log K \geq 2\cdot 16\sqrt{d}\log K$，则必然存在一个 u_i^*（不失一般性，记为 u_1^*），使得$\|u_1^*-u_j\|>16\sqrt{d}\log K$对任意 $1\leq j\leq K$均成立。于是

$$\begin{aligned}
2d_{\mathrm{TV}}\left(p_{\mathbf{u}}(x),p_{\mathbf{u}^*}(x)\right) &= \int_{\mathbb{R}^d}\left|\frac{1}{K}\sum_{i=1}^{K}P_i^*-\frac{1}{K}\sum_{i=1}^{K}P_i\right|\mathrm{d}x \\
&\geq \int_{\Omega_1^*}\left|\frac{1}{K}\sum_{i=1}^{K}P_i^*-\frac{1}{K}\sum_{i=1}^{K}P_i\right|\mathrm{d}x \\
&\geq \int_{\Omega_1^*}\frac{1}{K}\sum_{i=1}^{K}P_i^*\mathrm{d}x-\int_{\Omega_1^*}\frac{1}{K}\sum_{i=1}^{K}P_i\mathrm{d}x \\
&\geq \int_{\Omega_1^*}\frac{1}{K}P_1^*\mathrm{d}x-\frac{1}{K}\sum_{i=1}^{K}\int_{\Omega_1^*}P_i\mathrm{d}x \\
&= \frac{1}{K}\mathbb{P}(\mathcal{N}\left(u_1^*,I_d\right)\in\Omega_1^*)-\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}(\mathcal{N}\left(u_i,I_d\right)\in\Omega_1^*) \qquad (172)
\end{aligned}$$

由于$\|u_1^*-u_i\|>16\sqrt{d\log K}$，因此 $\Omega_1^*\cap\Omega_i=\emptyset$，这意味着（由引理 C.2）

$$\mathbb{P}(\mathcal{N}\left(u_i,I_d\right)\in\Omega_1^*)\leq\mathbb{P}(\mathcal{N}\left(u_i,I_d\right)\in\Omega_i^{\mathbf{C}})\leq 2\exp(-\frac{(8\sqrt{d\log K})^2}{16d})=2e^{-4\log K} \qquad (173)$$

此外，由引理 C.2，

$$\mathbb{P}(\mathcal{N}\left(u_1^*,I_d\right)\in\Omega_1^*)\geq 1-2\exp(-\frac{(8\sqrt{d\log K})^2}{16d})=1-2e^{-4\log K} \qquad (174)$$

因此，

$$\begin{aligned}
2d_{\mathrm{TV}}\left(p_{\mathbf{u}}(x),p_{\mathbf{u}^*}(x)\right) &\geq \frac{1}{K}\mathbb{P}(\mathcal{N}\left(u_1^*,I_d\right)\in\Omega_1^*)-\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}(\mathcal{N}\left(u_i,I_d\right)\in\Omega_1^*) \\
&\geq \frac{1}{K}(1-2e^{-4\log K})-\frac{1}{K}\sum_{i=1}^{K}2e^{-4\log K} \\
&= \frac{1}{K}-(2+\frac{2}{K})e^{-4\log K} \\
&\geq \frac{1}{K}-3e^{-4\log K} \\
&= \frac{1}{K}-3(\frac{1}{K})^4 \\
&= \frac{1}{2K} \qquad (175)
\end{aligned}$$

这就构成了矛盾。 □

接下来，我们陈述证明引理 5.2所需的核心引理。

引理 C.3. 若对任意 i，$\|u_i - u_i^*\| \leq 16\sqrt{d\log K}$，则对于 Ω_1^*（对于每个 Ω_i^*，相应的结果也成立），

$$\int_{\Omega_1^*} |P_1^* - P_1| \mathrm{d}x \geq c_1 \min\left\{\|u_1^* - u_1\|, 1\right\}, \tag{176}$$

其中 $c_1 = \frac{1}{200}$。

引理 C.4. 若对于任意 i，$\|u_i - u_i^*\| \leq 16\sqrt{d\log K}$，则对于 Ω_1^*（对于每个 Ω_i^*，相应结论均成立），对于每个 $j \neq 1$,

$$\int_{\Omega_1^*} |P_j^* - P_j| \mathrm{d}x \leq \frac{c_2}{K} \min\left\{\|u_j^* - u_j\|, 1\right\}, \tag{177}$$

其中 $c_2 = 2688\left(\frac{1}{2}\right)^{69}$.

利用这些引理，我们现在可以证明引理 5.2。

引理 5.2的证明。 根据引理 C.1，存在 $\boldsymbol{u}$ 的一个排列，使得 $\|u_i^* - u_i\| \leq 16\sqrt{d\log K}$ 对每个 $1 \leq i \leq K$均成立。因此可以应用引理 C.3、C.4 。注意到

$$\begin{aligned}
\int_{\Omega_1^*} |p_{\mathbf{u}}(x) - p_{\mathbf{u}^*}(x)| \mathrm{d}x &= \int_{\Omega_1^*} \left|\frac{1}{K}\sum_{i=1}^K P_i^* - \frac{1}{K}\sum_{i=1}^K P_i\right| \mathrm{d}x \\
&\geq \int_{\Omega_1^*} \left|\frac{1}{K}P_1^* - \frac{1}{K}P_i\right| \mathrm{d}x - \int_{\Omega_1^*} \left|\frac{1}{K}\sum_{i=2}^K P_i^* - \frac{1}{K}\sum_{i=2}^K P_i\right| \mathrm{d}x \\
&\geq \frac{1}{K}\int_{\Omega_1^*} |P_1^* - P_i| \mathrm{d}x - \frac{1}{K}\sum_{i=2}^K \int_{\Omega_1^*} |P_i^* - P_i| \mathrm{d}x \\
&\geq \frac{c_1}{K} \min\left\{\|u_1^* - u_1\|, 1\right\} - \frac{c_2}{K^2}\sum_{i=2}^K \min\left\{\|u_i^* - u_i\|, 1\right\},
\end{aligned} \tag{178}$$

其中最后一行由引理 C.3、C.4得出。

对相应的 $1 \leq i \leq K$ 将上述所有方程相加，由于 $\{\Omega_i^*\}_{i=1}^K$ 不相交，我们

have

$$
\begin{aligned}
d_{\mathrm{TV}}\left(p_{\mathbf{u}}(x), p_{\mathbf{u}^*}(x)\right) &= \frac{1}{2}\int_{\mathbb{R}^d} |p_{\mathbf{u}}(x) - p_{\mathbf{u}^*}(x)| \mathrm{d}x \\
&\geq \frac{1}{2}\sum_{i=1}^{K}\int_{\Omega_i^*} |p_{\mathbf{u}}(x) - p_{\mathbf{u}^*}(x)| \mathrm{d}x \\
&\geq \frac{1}{2}\left(\frac{c_1}{K} - \frac{(K-1)c_2}{K^2}\right)\sum_{i=1}^{K}\min\left\{\|u_i^* - u_i\|, 1\right\} \\
&\geq \frac{1}{2}\left(c_1 - c_2\right)\cdot\frac{1}{K}\sum_{i=1}^{K}\min\left\{\|u_i^* - u_i\|, 1\right\} \\
&= \frac{1}{2}\left(\frac{1}{200} - 2688\left(\frac{1}{2}\right)^{69}\right)\cdot\frac{1}{K}\sum_{i=1}^{K}\min\left\{\|u_i^* - u_i\|, 1\right\} \\
&\geq \frac{1}{500}\cdot\frac{1}{K}\sum_{i=1}^{K}\min\left\{\|u_i^* - u_i\|, 1\right\}. \qquad (179)
\end{aligned}
$$

最后，我们引用引理 B.1，其表述如下

$$
d_{\mathrm{TV}}(\mathcal{N}(u_i^*, I_d), \mathcal{N}(u_i, I_d)) \leq \min(\|u_i^* - u_i\|, 1). \qquad (180)
$$

取 $\sigma(\boldsymbol{u}) = \{u_i\}_{i=1}^K$,

$$
\begin{aligned}
d_{\mathrm{TV}}\left(p_{\sigma(\mathbf{u})}(x,z), p_{\mathbf{u}^*}(x,z)\right) &= \sum_{i=1}^{K}\mathbb{P}(z=i) d_{\mathrm{TV}}(\mathcal{N}(u_i^*, I_d), \mathcal{N}(u_i, I_d)) \\
&\leq \sum_{i=1}^{K}\frac{1}{K}\min(\|u_i^* - u_i\|, 1) \\
&\leq 500 d_{\mathrm{TV}}\left(p_{\mathbf{u}}(x), p_{\mathbf{u}^*}(x)\right). \qquad (181)
\end{aligned}
$$

□

最后，我们给出引理C.3 和C.4的证明。

引理*C.3*的证明。 不失一般性，令 $u_1^* = 0$， $\Delta := \|u_1\| \leq 16\sqrt{d}$对数 K，以及 $u_1 = (-\Delta, 0, 0, \cdots, 0)$。密度为

$$
P_1^*(x) = (\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}\|x\|^2} \qquad (182)
$$

$$
P_1(x) = (\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}\|x-u_1\|^2} \qquad (183)
$$

我们考虑一个面积为 $S \subset \Omega_1^*$的区域:

$$
S := \left\{x = (x_1, \cdots, x_d) \middle| x \in \Omega_1^*, x_1 \geq \frac{1}{10}\right\} \qquad (184)
$$

那么对于任意 $x \in S$, $\|x\|^2 \le \|x-u_1\|^2$，这意味着 $P_1^*(x) \ge P_1(x)$。因此

$$\begin{aligned}
\int_{\Omega_1^*} |P_1^* - P_1| \mathrm{d}x &\ge \int_S |P_1^* - P_1| \mathrm{d}x \\
&= \int_S (\frac{1}{\sqrt{2\pi}})^d \left(e^{-\frac{1}{2}\|x\|^2} - e^{-\frac{1}{2}\|x-u_1\|^2} \right) \mathrm{d}x \\
&= \int_S (\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}\|x\|^2} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \mathrm{d}x \\
&\ge \min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \int_S (\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}\|x\|^2} \mathrm{d}x \\
&= \min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \mathbb{P}(\mathcal{N}(0, I_d) \in S)
\end{aligned} \tag{185}$$

对于 $\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right)$，注意到对于任意 $x = (x_1, \cdots, x_d) \in S$,

$$\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2 = -x_1\Delta - \frac{1}{2}\Delta^2 \le -\frac{1}{10}\Delta \tag{186}$$

Thus

$$\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \ge 1 - e^{-\frac{1}{10}\Delta} \tag{187}$$

取 $c_3 = \frac{1}{20}$。我们声称

$$\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \ge c_3 \min\{\Delta, 1\}. \tag{188}$$

事实上，当 $0 \le \Delta \le 1$ 时,

$$\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \ge 1 - e^{-\frac{1}{10}\Delta} \ge \frac{1}{20}\Delta. \tag{189}$$

最后一个不等式成立，因为若令 $f(x) = e^{-\frac{1}{10}x} + \frac{1}{20}x - 1$，则 $f(0) = 0$,

$$f'(x) = -\frac{1}{10}e^{-\frac{1}{10}x} + \frac{1}{20} \le 0 \tag{190}$$

对于任意 $x \in [0, 10\log 2]$.，因此对于任意 $\Delta \in [0, 1]$,

$$e^{-\frac{1}{10}\Delta} + \frac{1}{20}\Delta - 1 = f(\Delta) \le f(0) = 0. \tag{191}$$

当 $1 \le \Delta \le 16\sqrt{d\log K}$ 时,

$$\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \ge 1 - e^{-\frac{1}{10}\Delta} \ge 1 - e^{-\frac{1}{10}} \ge \frac{1}{20} \cdot 1 \tag{192}$$

因此，我们已经证明

$$\min_{x\in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x-u_1\|^2} \right) \ge c_3 \min\{\Delta, 1\}. \tag{193}$$

其中 $c_3 = \frac{1}{20}$。

对于 $\mathbb{P}(\mathcal{N}(0, I_d) \in S)$，取

$$S' := \left\{ x = (x_1, \cdots, x_d) \middle| 2\sqrt{d \log 2} \geq x_1 \geq \frac{1}{10}, x_2^2 + \cdots + x_d^2 \leq 60 d \log K \right\}. \tag{194}$$

于是 $S' \subset S$。因此

$$\begin{aligned}
\mathbb{P}(\mathcal{N}(0, I_d) \in S) &\geq \mathbb{P}(\mathcal{N}(0, I_d) \in S') \\
&= \mathbb{P}(2\sqrt{d \log 2} \geq x_1 \geq \frac{1}{10}, x_2^2 + \cdots + x_d^2 \leq 60 d \log K, x \sim \mathcal{N}(0, I_d)) \\
&= \mathbb{P}\left(2\sqrt{d \log 2} \geq \mathcal{N}(0,1) \geq \frac{1}{10}\right) \mathbb{P}\left(\|\mathcal{N}(0, I_{d-1})\|^2 \leq 60 d \log K\right) \\
&\geq \mathbb{P}\left(2\sqrt{\log 2} \geq \mathcal{N}(0,1) \geq \frac{1}{10}\right) \mathbb{P}\left(\|\mathcal{N}(0, I_{d-1})\|^2 \leq 60(d-1) \log 2\right) \\
&> \mathbb{P}\left(2\sqrt{\log 2} \geq \mathcal{N}(0,1) \geq \frac{1}{10}\right) \cdot (1 - 2e^{-2}) \quad \text{(by Lemma C.2)} \\
&> \frac{1}{4} \cdot (1 - 2e^{-2}) \\
&> \frac{1}{10}
\end{aligned} \tag{195}$$

综合以上所有结果，我们得到

$$\begin{aligned}
\int_{\Omega_1^*} |P_1^* - P_1| \mathrm{d}x &\geq \min_{x \in S} \left(1 - e^{\frac{1}{2}\|x\|^2 - \frac{1}{2}\|x - u_1\|^2}\right) \mathbb{P}(\mathcal{N}(0, I_d) \in S) \\
&\geq c_3 \min\{\Delta, 1\} \cdot \frac{1}{10} \\
&= \frac{1}{200} \min\{\|u_1^* - u_1\|, 1\}
\end{aligned} \tag{196}$$

□

引理的证明 *C.4*。 对于任意 $i \neq 1$,

$$\int_{\Omega_1^*} |P_i^* - P_i| \mathrm{d}x = \int_{\Omega_1^*} (\frac{1}{\sqrt{2\pi}})^d |e^{-\frac{1}{2}\|x - u_i^*\|^2} - e^{-\frac{1}{2}\|x - u_i\|^2}| \mathrm{d}x. \tag{197}$$

注意，若我们记 $a(x) := \|x - u_i^*\|$, $\delta(x) := \|x - u_i^*\| - \|x - u_i\|$,$\Delta := \|u_i - u_i^*\|$, 则$|\delta(x)| \leq \Delta \leq 16\sqrt{d \log K}$，且对任意 $x \in \Omega_1^*$, $a(x) \geq 92\sqrt{d \log K}$（由分离条件可知）。

因此

$$
\begin{aligned}
&\max_{x\in\Omega_1^*}\left|e^{-\frac{1}{2}\|x-u_i^*\|^2}-e^{-\frac{1}{2}\|x-u_i\|^2}\right| \\
&=\max_{x\in\Omega_1^*}\left|e^{-\frac{1}{2}a(x)^2}-e^{-\frac{1}{2}(a(x)-\delta(x))^2}\right| \\
&\le \max\left\{\left|e^{-\frac{1}{2}a(x)^2}-e^{-\frac{1}{2}(a(x)-\delta(x))^2}\right|\middle| a(x)\ge 92\sqrt{d\log K}, |\delta(x)|\le\Delta\right\} \\
&\le \max_{a\ge 92\sqrt{d\log K}}\{\max(|e^{-\frac{a^2}{2}}-e^{-\frac{(a-\Delta)^2}{2}}|, |e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}}|)\} \\
&= \max_{a\ge 92\sqrt{d\log K}}\{\max(e^{-\frac{(a-\Delta)^2}{2}}-e^{-\frac{a^2}{2}}, e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}})\} \\
&\le \max(\max_{a\ge 92\sqrt{d\log K}} e^{-\frac{(a-\Delta)^2}{2}}-e^{-\frac{a^2}{2}}, \max_{a\ge 92\sqrt{d\log K}} e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}})) \\
&\le \max(\max_{a\ge 76\sqrt{d\log K}} e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}}, \max_{a\ge 92\sqrt{d\log K}} e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}})). \qquad (198)
\end{aligned}
$$

最后一个不等式成立，因为 $\Delta\le 16\sqrt{d\log K}$ 固定 Δ，令 $f(a)=e^{-\frac{a^2}{2}}-e^{-\frac{(a+\Delta)^2}{2}}$ 于是

$$
f'(a)=-ae^{-\frac{a^2}{2}}+(a+\Delta)e^{-\frac{(a+\Delta)^2}{2}} \qquad (199)
$$

我们首先证明，对于任意 $a\ge 76\sqrt{d\log K}$，都有 $f'(a)\le 0$。注意到

$$
\begin{aligned}
f'(a)=-ae^{-\frac{a^2}{2}}+(a+\Delta)e^{-\frac{(a+\Delta)^2}{2}}&\le 0 \\
\iff (a+\Delta)e^{-\frac{(a+\Delta)^2}{2}}&\le ae^{-\frac{a^2}{2}} \\
\iff 1+\frac{\Delta}{a}&\le e^{a\Delta+\frac{1}{2}\Delta^2} \qquad (200)
\end{aligned}
$$

最后一个结论成立，因为

$$
e^{a\Delta+\frac{1}{2}\Delta^2}\ge 1+a\Delta+\frac{1}{2}\Delta^2\ge 1+\frac{\Delta}{a} \qquad (201)
$$

当 $a\ge 76\sqrt{d\log K}>1$. 时
由于对于任意 $a\ge 76\sqrt{d\log K}$，$f'(a)\le 0$，因此我们有

$$
\begin{aligned}
f(a)&\le f(76\sqrt{d\log K}) \\
&=\exp(-\frac{1}{2}(76\sqrt{d\log K})^2)-\exp(-\frac{1}{2}(76\sqrt{d\log K}+\Delta)^2) \\
&=e^{-\frac{1}{2}(76\sqrt{d\log K})^2}(1-e^{-76\sqrt{d\log K}\Delta-\frac{1}{2}\Delta^2}) \\
&\le e^{-\frac{1}{2}(76\sqrt{d\log K})^2}(76\sqrt{d\log K}\Delta+\frac{1}{2}\Delta^2) \\
&\le e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 84\sqrt{d\log K}\Delta \quad (\text{since } \Delta\le 16\sqrt{d\log K}). \qquad (202)
\end{aligned}
$$

这表明

$$\max_{a\geq 76\sqrt{d\log K}} e^{-\frac{a^2}{2}} - e^{-\frac{(a+\Delta)^2}{2}} \leq e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 84\sqrt{d\log K}\Delta \tag{203}$$

类似地

$$\max_{a\geq 92\sqrt{d\log K}} e^{-\frac{a^2}{2}} - e^{-\frac{(a+\Delta)^2}{2}} \leq e^{-\frac{1}{2}(92\sqrt{d\log K})^2}\cdot 100\sqrt{d\log K}\Delta \tag{204}$$

因此

$$\max_{x\in\Omega_1^*}\left|e^{-\frac{1}{2}\|x-u_i^*\|^2} - e^{-\frac{1}{2}\|x-u_i\|^2}\right| \leq e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 84\sqrt{d\log K}\Delta \leq c_4\min\{\|u_i^*-u_i\|,1\} \tag{205}$$

其中 $c_4 = e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 1344 d\log K$（由于 $\Delta\leq 16\sqrt{d\log K}\min\{\Delta,1\}$）。注意，

$$\begin{aligned}
c_4 &= e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 1344 d\log K\\
&\leq e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 1344 k^d K\\
&\leq e^{-\frac{1}{2}(76\sqrt{d\log K})^2}\cdot 1344 k^{2d}\\
&= 1344 e^{-2886 d\log K}\\
&\leq 1344 e^{-\frac{1}{2}(70\sqrt{d\log K})^2}
\end{aligned} \tag{206}$$

不失一般性，令 $u_1^*=0$，并定义 $u' = (50\sqrt{d\log K}, 0, \cdots, 0)$，则

$$\begin{aligned}
\int_{\Omega_1^*}|P_i^*-P_i|\mathrm{d}x &= \int_{\Omega_1^*}(\frac{1}{\sqrt{2\pi}})^d|e^{-\frac{1}{2}\|x-u_i^*\|^2} - e^{-\frac{1}{2}\|x-u_i\|^2}|\mathrm{d}x\\
&\leq \int_{\Omega_1^*}(\frac{1}{\sqrt{2\pi}})^d \max_{x\in\Omega_1^*}|e^{-\frac{1}{2}\|x-u_i^*\|^2} - e^{-\frac{1}{2}\|x-u_i\|^2}|\mathrm{d}x\\
&\leq \int_{\Omega_1^*}(\frac{1}{\sqrt{2\pi}})^d 1344 e^{-\frac{1}{2}(70\sqrt{d\log K})^2}\min\{\|u_i^*-u_i\|,1\}\,\mathrm{d}x\\
&= \min\{\|u_i^*-u_i\|,1\}\int_{\Omega_1^*}(\frac{1}{\sqrt{2\pi}})^d 1344 e^{-\frac{1}{2}(70\sqrt{d\log K})^2}\mathrm{d}x\\
&\leq 1344\min\{\|u_i^*-u_i\|,1\}\int_{\Omega_1^*}(\frac{1}{\sqrt{2\pi}})^d e^{-\frac{1}{2}\|x-u'\|^2}\mathrm{d}x\\
&\leq 1344\min\{\|u_i^*-u_i\|,1\}\,\mathbb{P}(\mathcal{N}(u',I_d)\in\Omega_1^*)\\
&\leq 1344\min\{\|u_i^*-u_i\|,1\}\,\mathbb{P}(\|\mathcal{N}(u',I_d)-u'\|\geq 34\sqrt{d\log K})\\
&\leq 1344\min\{\|u_i^*-u_i\|,1\}\cdot 2\exp(-\frac{(34\sqrt{d\log K})^2}{16d}) \quad \text{(by Lemma C.2)}\\
&\leq 1344\min\{\|u_i^*-u_i\|,1\}\cdot 2\exp(-70\log K)\\
&= 2688\min\{\|u_i^*-u_i\|,1\}\left(\frac{1}{K}\right)^{70}\\
&\leq 2688\left(\frac{1}{2}\right)^{69}\left(\frac{1}{K}\right)\min\{\|u_i^*-u_i\|,1\}
\end{aligned} \tag{207}$$

□

C.2 括号覆盖数

我们将 $\mathcal{P}_{\mathcal{X}}(\mathcal{U})$ 的括号覆盖数上界估计如下。

引理 C.5. 令

$$\mathcal{P}_{\mathcal{X}}(\mathcal{U}) := \left\{ \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(u_i, I_d) \,\middle|\, \mathbf{u} = \{u_i\}_{i=1}^{K} \in \mathcal{U} \right\}.$$

我们假设存在 $D > 0$，使得对于任意 $\mathbf{u} = \{u_i\}_{i=1}^{K} \in \mathcal{U}$，都有

$$\|u_i\|_2 \le D\sqrt{d \log K}, \ \forall i \in [K].$$

那么，熵可以界定如下，

$$\log N\big(\mathcal{P}_{\mathcal{X}}(\mathcal{U}), 1/m\big) \le 2dK \log(6mdKD).$$

引理 *C.5* 的证明。 首先，我们考虑一个标准高斯分布集合

$$\mathcal{P}_{\mathcal{X}}(\mathcal{A}) := \left\{ p_a(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x-a\|_2^2}{2}} \,\middle|\, a \in \mathcal{A} \right\},$$

其中 $\mathcal{A} = \{a \in \mathbb{R}^d \,|\, \|a\|_2 \le D\sqrt{d \log K}\}$。我们的目标是找到一个 $1/m$-括号 $\mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{A}) \ 1/m)$，使其覆盖$\mathcal{P}_{\mathcal{X}}(\mathcal{A})$。换言之，对于任意 $p_a(x) \in \mathcal{P}_{\mathcal{X}}(\mathcal{A})$，我们需要定义 $\bar{p}_a(x) \in \mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{A}) \ 1/m)$，使得

- $\bar{p}_a(x) \ge p_a(x) \quad \forall x \in \mathbb{R}^d$
- $\int |\bar{p}_a(x) - p_a(x)|\, dx \le 1/m$。

我们考虑具有以下形式的 $\bar{p}_a(x)$

$$\bar{p}_a(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{c_1\|x-\bar{a}\|_2^2}{2} + c_2}.$$

接下来，我们指定 $\bar{a} \in \mathbb{R}^d$、 $c_1 \in \mathbb{R}$ 和 $c_2 \in \mathbb{R}$。令 $a = (a_1, \ldots, a_d)$，并令 $\epsilon > 0$ 为稍后选取的参数。若对于某个 $k \in \mathbb{Z}$，有 $a_i \in [k\epsilon, (k+1)\epsilon)$，则定义 $\bar{a}_i := k\epsilon$ 和 $\bar{a} := (\bar{a}_1, \ldots, \bar{a}_d)$，由此可得

$$\|a - \bar{a}\|_2^2 \le d\epsilon^2. \tag{208}$$

注意，对于任意 $x \in \mathbb{R}^d$， $\bar{p}_a(x) \ge p_a(x)$ 成立，当且仅当

$$(c_1 - 1)\left\| x + \frac{a - c_1\bar{a}}{c_1 - 1} \right\|_2^2 + \frac{c_1}{1 - c_1}\|a - \bar{a}\|_2^2 \le 2c_2, \ \forall x \in \mathbb{R}^d.$$

令 $c_1 = 1 - \epsilon$。那么，当且仅当 $\bar{p}_a(x) \ge p_a(x)$

$$-\epsilon\left\| x + \frac{a - c_1\bar{a}}{c_1 - 1} \right\|_2^2 + \frac{1 - \epsilon}{\epsilon}\|a - \bar{a}\|_2^2 \le 2c_2, \ \forall x \in \mathbb{R}^d.$$

注意，

$$-\epsilon\left\|x+\frac{a-c_1\bar{a}}{c_1-1}\right\|_2^2+\frac{1-\epsilon}{\epsilon}\|a-\bar{a}\|_2^2\le\frac{1-\epsilon}{\epsilon}\|a-\bar{a}\|_2^2\le d(1-\epsilon)\epsilon,$$

其中，最后一个不等式由（208）得出。因此，选择 $c_2=d(1-\epsilon)\epsilon/2$，我们得到$\bar{p}_a(x)\ge p_a(x)$，对任意 $x\in\mathbb{R}^d$ 均成立。注意到

$$\int|\bar{p}_a(x)-p_a(x)|\,dx=\frac{1}{\sqrt{c_1}}\cdot e^{c_2}-1=\frac{e^{\frac{d(1-\epsilon)\epsilon}{2}}}{\sqrt{1-\epsilon}}-1\le\big(1+d(1-\epsilon)\epsilon\big)\cdot(1+\epsilon)-1\le(1+2d)\epsilon.$$

这里，第一个不等式是由以下事实得出的：对于任意 $0<x<1/2$，$e^x\le1+2x$且 $\frac{1}{\sqrt{1-x}}\le1+x$。令 $(1+2d)\epsilon=m^{-1}$。于是有

$$\int|\bar{p}_a(x)-p_a(x)|\,dx\le(1+2d)\epsilon=\frac{1}{m}.$$

回忆一下，对于任意 $a\in\mathcal{A}$，都有 $\|a\|_2\le D\sqrt{d}$因此，我们有

$$N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{A}),1/m)\le\left(\frac{2D\sqrt{d\log K}}{\epsilon}\right)^d=\left(2mD(1+2d)\sqrt{d\log K}\right)^d.$$

接下来，我们考虑一个高斯混合模型集合

$$\mathcal{P}_{\mathcal{X}}(\mathcal{U}):=\left\{\sum_{i=1}^K\frac{1}{K}\mathcal{N}(u_i,I_d)\,\middle|\,\mathbf{u}=\{u_i\}_{i=1}^K\in\mathcal{U}\right\},$$

其中 $\mathcal{U}=\{\{u_i\}_{i=1}^K\,|\,\|u_i\|_2\le D\sqrt{d}$，$\forall i\in[K]\}$。我们的目标是找到 $\mathcal{P}_{\mathcal{X}}(\mathcal{U})$ 的一个 $1/m$-括号 $\mathcal{N}(\mathcal{P}_{\mathcal{X}}(\mathcal{U})\ 1/m)$。对于任意 $p_{\mathbf{u}}(x)\in\mathcal{P}_{\mathcal{X}}(\mathcal{U})$，都有

$$p_{\mathbf{u}}(x)=\sum_{i=1}^K\frac{1}{K}p_{u_i}(x),$$

其中 $p_{u_i}(x)\in\mathcal{P}_{\mathcal{X}}(\mathcal{A})$。注意，对于任意 $i\in[K]$，都存在 $\bar{p}_{u_i}(x)\in\mathcal{N}_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{A})\ 1/m)$，使得

- $\bar{p}_{u_i}(x)\ge p_{u_i}(x)\quad\forall x\in\mathbb{R}^d$
- $\int|\bar{p}_{u_i}(x)-p_{u_i}(x)|\,dx\le1/m$。

我们定义

$$\bar{p}_{\mathbf{u}}(x)=\sum_{i=1}^K\frac{1}{K}\bar{p}_{u_i}(x).$$

于是有

$$\bar{p}_{\mathbf{u}}(x)=\sum_{i=1}^K\frac{1}{K}\bar{p}_{u_i}(x)\ge\sum_{i=1}^K\frac{1}{K}p_{u_i}(x)=p_{\mathbf{u}}(x),\ \forall x\in\mathbb{R}^d$$

and

$$\int |\bar{p}_{\mathbf{u}}(x) - p_{\mathbf{u}}(x)|\, dx \le \sum_{i=1}^{K} \frac{1}{K} \int |\bar{p}_{u_i}(x) - p_{u_i}(x)|\, dx \le \sum_{i=1}^{K} \frac{1}{mK} = \frac{1}{m}.$$

因此，我们得到

$$N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{U}), 1/m) \le \Big(N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{A}), 1/m)\Big)^{K} \le \Big(2mD(1+2d)\sqrt{d \log K}\Big)^{dK},$$

这意味着

$$\log N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\mathcal{U}), 1/m) \le dK \log\Big(2mD(1+2d)\sqrt{d \log K}\Big) \le 2dK \log(6mdKD).$$

□

C.3 拉德马赫复杂度

给定有标签数据 $\{x_j, y_j\}_{j=1}^{n}$ 和预训练的 $\hat{\mathbf{u}}$，函数类

$$\big\{(\mathbb{1}_{g_{\hat{\mathbf{u}},\psi}(x_1)\neq y_1}, \ldots, \mathbb{1}_{g_{\hat{\mathbf{u}},\psi}(x_n)\neq y_n}) \,\big|\, \psi \in \Psi\big\}$$

是一个有限 函数类，其拉德马赫复杂度可由以下引理界定 .

引理 C.6. 设 $A = \{a^1, \ldots, a^N\}$ 是 $\mathbb{R}^n$中的一个有限向量集。那么，拉德马赫复杂度可由下式界定,

$$R_n(A) \le \max_{a\in A} \|a\|_2 \cdot \frac{2\sqrt{2\log N}}{n}.$$

证明。注意，对于任意 $\lambda > 0$

$$\begin{aligned} R_n(A) &= \mathbb{E}\Big[\sup_{a\in A} \frac{2}{n} \sum_{i=1}^{n} \sigma_i a_i\Big] \le \frac{1}{\lambda} \log \mathbb{E}\Big[e^{\sup_{a\in A} \frac{2\lambda}{n} \sum_{i=1}^{n} \sigma_i a_i}\Big] \\ &\le \frac{1}{\lambda} \log \sum_{a\in A} \mathbb{E}\Big[e^{\frac{2\lambda}{n} \sum_{i=1}^{n} \sigma_i a_i}\Big] = \frac{1}{\lambda} \log \sum_{a\in A} \prod_{i=1}^{n} \mathbb{E}\Big[e^{\frac{2\lambda}{n} \sigma_i a_i}\Big], \end{aligned} \tag{209}$$

其中，第一个不等式由詹森不等式推出。回顾一下， σ_i 是一个拉德马赫随机变量。因此，我们有

$$\mathbb{E}\Big[e^{\frac{2\lambda}{n}\sigma_i a_i}\Big] = \frac{1}{2} e^{\frac{2\lambda}{n} a_i} + \frac{1}{2} e^{-\frac{2\lambda}{n} a_i} \le e^{\frac{2\lambda^2 a_i^2}{n^2}}, \tag{210}$$

其中，最后一个不等式由以下事实推出：$(e^x + e^{-x})/2 \le e^{x^2/2}$。根据 (209) 和 (210)，我们有

$$R_n(A) \le \frac{1}{\lambda} \log \sum_{a\in A} e^{\frac{2\lambda^2 \|a\|^2}{n^2}} \le \frac{1}{\lambda} \log |A| e^{\frac{2\lambda^2}{n^2} \cdot \max_{a\in A} \|a\|^2} = \frac{1}{\lambda} \log N + \frac{2\lambda}{n^2} \cdot \max_{a\in A} \|a\|^2. \tag{211}$$

令 $\lambda = \sqrt{n \log N / 2 \max_{a\in A} \|a\|^2}$。我们得到

$$R_n(A) \le \max_{a\in A} \|a\| \cdot \frac{2\sqrt{2\log N}}{n}.$$

□

C.4 定理 5.3的证明

接下来，我们证明定理 5.3。

证明。 令 $\Phi = \mathcal{U}$ 和 Ψ 为 2^K 分类的集合。回顾一下，损失函数定义为$\ell(x, y) = \mathbb{1}_{\{x \neq y\}}$，其上界为 1。令 $m = \tilde{\Omega}(dK^3)$。根据定理 3.3 和引理 C.5，有

$$d_{\mathrm{TV}}\left(\mathbb{P}_{\hat{\phi}}(x), \mathbb{P}_{\phi^*}(x)\right) \lesssim \sqrt{\frac{1}{m} \log \frac{N_{[\,]}(\mathcal{P}_{\mathcal{X}}(\Phi), 1/m)}{\delta}} \lesssim \sqrt{\frac{dK}{m} \log \frac{mdKD}{\delta}} \lesssim \frac{1}{K}.$$

然后，根据引理 5.2，假设 3.2 对高斯混合模型成立。根据定理3.4，以至少 $1 - \delta$ 的概率，我们得到以下超额风险界：

$$\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) \leq 2 \max_{\phi \in \Phi} R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) + \sqrt{\frac{2}{n} \log \frac{4}{\delta}} + 12\kappa \cdot \sqrt{\frac{1}{m} \log \frac{2N(\mathcal{P}_{\mathcal{X}}(\Phi), 1/m)}{\delta}},$$

其中 $\kappa = c_2$ 是某个表示模型可迁移性的绝对常数。根据引理C.5，进一步可得

$$\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) \leq 2 \max_{\phi \in \Phi} R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) + \sqrt{\frac{2}{n} \log \frac{4}{\delta}} + 12\kappa \cdot \sqrt{\frac{2dK}{m} \log \frac{12mdKD}{\delta}}. \tag{212}$$

对于任意 $\phi \in \Phi$，有

$$R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) = \mathbb{E}\left[\sup_{\psi \in \Psi} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{g_{\phi,\psi}(x_i) \neq y_i\}}\right]. \tag{213}$$

注意 $|\Psi| = 2^K$。根据引理 C.6，对于任意 $\phi \in \Phi$，都有

$$R_n(\ell \circ \mathcal{G}_{\phi,\Psi}) \leq \sqrt{n} \cdot \frac{2\sqrt{2 \log 2^K}}{n} = 2\sqrt{\frac{2K \log 2}{n}}. \tag{214}$$

根据（212）和（214），我们有

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\phi}, \hat{\psi}) &\leq 4\sqrt{\frac{2K \log 2}{n}} + \sqrt{\frac{2}{n} \log \frac{4}{\delta}} + 12\kappa \cdot \sqrt{\frac{2dK}{m} \log \frac{12mdKD}{\delta}} \\
&= \mathcal{O}\left(\sqrt{\frac{K \log \frac{1}{\delta}}{n}} + \kappa\sqrt{\frac{dK \log \frac{mdKD}{\delta}}{m}}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\frac{K}{n}} + \kappa\sqrt{\frac{dK}{m}}\right),
\end{aligned}$$

其中， $\kappa = c_2$ 是表示模型可迁移性的某些绝对常数。 □

因此，我们证明了定理 5.3。

D 第 6 节的证明

在第D.1节中，我们通过证明引理κ^{-1}6.1，说明以线性回归为下游任务的对比学习是弱信息的。在第D.2节中，我们证明定理6.2。

D.1 引理的证明 6.1

回想一下，在对比学习的设定中，我们假设 x 和 x' 从同一分布 $\mathbb{P}(x)$中独立采样。并且我们假设，刻画x与 x' 之间相似性的标签 t 满足

$$\mathbb{P}(t=1\,|\,x,x') = \frac{1}{1+e^{-f_{\theta^*}(x)^T f_{\theta^*}(x')}},$$
$$\mathbb{P}(t=-1\,|\,x,x') = \frac{1}{1+e^{f_{\theta^*}(x)^T f_{\theta^*}(x')}}.$$

引理 6.1 可直接由以下引理推出。

引理 D.1. 存在 $O\in\mathbb{R}^{r\times r}$, $O^TO=OO^T=I_r$ 使得

$$d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta}(x,z),\mathbb{P}_{f_{\theta^*}}(x,z)\big)\le c\cdot\sqrt{\frac{1}{\sigma_{\min}(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T])}}\cdot H\big(\mathbb{P}_{f_\theta}(x,x',t),\mathbb{P}_{f_{\theta^*}}(x,x',t)\big).$$

这里的 c是一些绝对常数。

我们先证明以下引理，这是引理 D.1的核心。

引理 D.2. 假设 $\mathbb{E}[f_\theta(x)f_{\theta^*}(x)^T]=\mathbb{E}[f_{\theta^*}(x)f_\theta(x)^T]$ 是半正定矩阵。那么我们有

$$\mathbb{E}\big[\big(f_\theta(x)^Tf_\theta(x')-f_{\theta^*}(x)^Tf_{\theta^*}(x')\big)^2\big]\ge(2\sqrt{2}-2)\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)\cdot\mathbb{E}[\|f_{\theta^*}(x)-f_\theta(x)\|_2^2]$$

引理 D.2 的证明。 为简化记号，我们记 $\Delta(x):=f_{\theta^*}(x)-f_\theta(x)$。于是有

$$\begin{aligned}
&\mathbb{E}\big[\big(f_\theta(x)^Tf_\theta(x')-f_{\theta^*}(x)^Tf_{\theta^*}(x')\big)^2\big]\\
&=\mathbb{E}\big[\big(f_{\theta^*}(x)^T\Delta(x')+\Delta(x)^Tf_{\theta^*}(x')-\Delta(x)^T\Delta(x')\big)^2\big]\\
&=\mathbb{E}\big[\big(\Delta(x)^T\Delta(x')\big)^2-2\sqrt{2}\Delta(x)^T\Delta(x')f_{\theta^*}(x')^T\Delta(x)+2f_{\theta^*}(x)^T\Delta(x')f_{\theta^*}(x')^T\Delta(x)\big]\\
&\quad+(4-2\sqrt{2})\mathbb{E}[f_\theta(x')^T\Delta(x)\Delta(x)^Tf_{\theta^*}(x')]+(2\sqrt{2}-2)\mathbb{E}[f_{\theta^*}(x')^T\Delta(x)\Delta(x)^Tf_{\theta^*}(x')].
\end{aligned}\tag{215}$$

对于式（215）的第一项，有

$$\begin{aligned}
&\mathbb{E}\big[\big(\Delta(x)^T\Delta(x')\big)^2-2\sqrt{2}\Delta(x)^T\Delta(x')f_{\theta^*}(x')^T\Delta(x)+2f_{\theta^*}(x)^T\Delta(x')f_{\theta^*}(x')^T\Delta(x)\big]\\
&=\mathrm{Tr}\Big(\mathbb{E}[\Delta(x')\Delta(x')^T\Delta(x)\Delta(x)^T-2\sqrt{2}\Delta(x')f_{\theta^*}(x')^T\Delta(x)\Delta(x)^T+2\Delta(x')f_{\theta^*}(x')^T\Delta(x)f_{\theta^*}(x)^T]\Big)\\
&=\mathrm{Tr}\Big(\big(\mathbb{E}[\Delta(x)\Delta(x)^T]\big)^2-2\sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]\cdot\mathbb{E}[\Delta(x)\Delta(x)^T]+2\big(\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]\big)^2\Big)\\
&=\mathrm{Tr}\Big(\big(\mathbb{E}[\Delta(x)\Delta(x)^T]-\sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]\big)^2\Big),
\end{aligned}\tag{216}$$

其中，第二个等式是由我们的假设 x、x' 独立同分布得出的。注意，$\mathbb{E}[f_\theta(x)f_{\theta^*}(x)^T]=\mathbb{E}[f_{\theta^*}(x)f_\theta(x)^T]$。因此，我们得到

$$\begin{aligned}
\Big(\mathbb{E}[\Delta(x)\Delta(x)^T]-\sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]\Big)^T&=\mathbb{E}[\Delta(x)\Delta(x)^T]-\sqrt{2}\mathbb{E}[f_{\theta^*}(x)\Delta(x)^T]\\
&=\mathbb{E}[\Delta(x)\Delta(x)^T]-\sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T],
\end{aligned}\tag{217}$$

w这意味着 $\mathbb{E}[\Delta(x)\Delta(x)^T] - \sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]$ 是对称的。因此有

$$\begin{aligned}&\mathbb{E}\big[\big(\Delta(x)^T\Delta(x')\big)^2 - 2\sqrt{2}\Delta(x)^T\Delta(x')f_{\theta^*}(x')\Delta(x) + 2f_{\theta^*}(x)^T\Delta(x')f_{\theta^*}(x')^T\Delta(x)\big]\\&= \mathrm{Tr}\Big(\big(\mathbb{E}[\Delta(x)\Delta(x)^T] - \sqrt{2}\mathbb{E}[\Delta(x)f_{\theta^*}(x)^T]\big)^2\Big) \geq 0.\end{aligned} \tag{218}$$

(215）的第二项为

$$\mathbb{E}[f_\theta(x')^T\Delta(x)\Delta(x)^T f_{\theta^*}(x')] = \mathrm{Tr}\Big(\mathbb{E}[f_{\theta^*}(x')f_\theta(x')^T]\cdot\mathbb{E}[\Delta(x)\Delta(x)^T]\Big) \geq 0, \tag{219}$$

其中，不等式是由以下事实得出的： $\mathbb{E}[f_{\theta^*}(x')f_\theta(x')^T] \succcurlyeq 0$ 以及 $\mathbb{E}[\Delta(x)\Delta(x)^T] \succcurlyeq 0$。
(215）的第三项为

$$\begin{aligned}\mathbb{E}[f_{\theta^*}(x')^T\Delta(x)\Delta(x)^T f_{\theta^*}(x')] &= \mathrm{Tr}\Big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\cdot\mathbb{E}[\Delta(x)\Delta(x)^T]\Big)\\&\geq \sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)\mathrm{Tr}\Big(\mathbb{E}[\Delta(x)\Delta(x)^T]\Big)\\&= \sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)\mathbb{E}[\|\Delta(x)\|_2^2].\end{aligned} \tag{220}$$

结合式（215）、（218）、（219）和（220），可得

$$\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x') - f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big] \geq (2\sqrt{2}-2)\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)\mathbb{E}[\|\Delta(x)\|_2^2] \tag{221}$$

□

利用引理 D.2，我们证明如下引理 D.1 。

引理 D.1 的证明。我们考虑 $\mathbb{E}[f_\theta(x)f_{\theta^*}(x)^T] = U_1\Sigma_1V_1^T$ 和 $\mathbb{E}[f_{\theta^*}(x)f_\theta(x)^T] = (\mathbb{E}[f_\theta(x)f_{\theta^*}(x)^T])^T = V_1\Sigma_1U_1^T$ 的奇异值分解（SVD）。定义 $O := V_1U_1^T \in \mathbb{R}^{r\times r}$，使其满足 $O^TO = OO^T = I_r$。于是有

$$\mathbb{E}[Of_\theta(x)f_{\theta^*}(x)^T] = \mathbb{E}\big[f_{\theta^*}(x)\big(Of_\theta(x)\big)^T\big] = V_1\Sigma_1V_1^T, \tag{222}$$

它们是半正定矩阵。根据引理 D.2，我们有

$$\begin{aligned}&\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x') - f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big]\\&\quad\geq (2\sqrt{2}-2)\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)\cdot\mathbb{E}[\|f_{\theta^*}(x) - Of_\theta(x)\|_2^2].\end{aligned} \tag{223}$$

对于海林格距离，我们有

$$\begin{aligned}&2H^2\big(\mathbb{P}_{f_\theta}(x,x',t), \mathbb{P}_{f_{\theta^*}}(x,x',t)\big)\\&= \int\Big(\sqrt{p_{f_\theta}(x,x',t)} - \sqrt{p_{f_{\theta^*}}(x,x',t)}\Big)^2 dtdxdx'\\&= \int\Big(\sqrt{p_{f_\theta}(t=1\,|\,x,x')} - \sqrt{p_{f_{\theta^*}}(t=1\,|\,x,x')}\Big)^2 p(x,x')\,dxdx'\\&\quad+ \int\Big(\sqrt{p_{f_\theta}(t=0\,|\,x,x')} - \sqrt{p_{f_{\theta^*}}(t=0\,|\,x,x')}\Big)^2 p(x,x')\,dxdx'\end{aligned} \tag{224}$$

对于（224）中的第一项，我们有

$$\int \left(\sqrt{p_{f_\theta}(t=1\,|\,x,x')}-\sqrt{p_{f_{\theta^*}}(t=1\,|\,x,x')}\right)^2 p(x,x')\,dxdx' \\ =\int \left(\sqrt{h\big(f_\theta(x)^T f_\theta(x')\big)}-\sqrt{h\big(f_{\theta^*}(x)^T f_{\theta^*}(x')\big)}\right)^2 p(x,x')\,dxdx', \tag{225}$$

其中

$$h(a):=\frac{1}{1+e^{-a}}. \tag{226}$$

根据柯西-施瓦茨不等式，我们有 $|f_\theta(x)^T f_\theta(x')|\le \|f_\theta(x)\|_2\|f_\theta(x')\|_2\le 1$。注意，对于任意 a、$b\in[-1,1]$，我们有

$$\left(\sqrt{h(a)}-\sqrt{h(b)}\right)^2 \\ =\frac{\big(h(a)-h(b)\big)^2}{\left(\sqrt{h(a)}+\sqrt{h(b)}\right)^2}\ge\frac{1}{4}\big(h(a)-h(b)\big)^2=\frac{1}{4}h'(\xi)^2(a-b)^2\ge\frac{1}{2+e+e^{-1}}(a-b)^2. \tag{227}$$

因此，有

$$\int \left(\sqrt{p_{f_\theta}(t=1\,|\,x,x')}-\sqrt{p_{f_{\theta^*}}(t=1\,|\,x,x')}\right)^2 p(x,x')\,dxdx' \\ \ge\frac{1}{2+e+e^{-1}}\int \big(f_\theta(x)^T f_\theta(x')-f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2 p(x,x')\,dxdx' \\ =\frac{1}{2+e+e^{-1}}\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x')-f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big]. \tag{228}$$

同样地，对于（224）的第二项，我们有

$$\int \left(\sqrt{p_{f_\theta}(t=0\,|\,x,x')}-\sqrt{p_{f_{\theta^*}}(t=0\,|\,x,x')}\right)^2 p(x,x')\,dxdx' \\ \ge\frac{1}{2+e+e^{-1}}\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x')-f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big] \tag{229}$$

结合（224）、（228）和（229），我们有

$$H^2\big(\mathbb{P}_{f_\theta}(x,x',t),\mathbb{P}_{f_{\theta^*}}(x,x',t)\big)\ge\frac{1}{2+e+e^{-1}}\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x')-f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big]. \tag{230}$$

我们选择 $O\in\mathbb{R}^{r\times r}$，其满足（223）。对于全变差距离，我们有

$$d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta}(x,z),\mathbb{P}_{f_{\theta^*}}(x,z)\big)=\frac{1}{2}\int |p_{Of_\theta}(z\,|\,x)-p_{f_{\theta^*}}(z\,|\,x)|p(x)\,dx. \tag{231}$$

注意， $z \mid x \sim \mathcal{N}$（$f_\theta$（$x$） I_r）。根据引理B.1，我们有

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta}(x,z), \mathbb{P}_{f_{\theta^*}}(x,z)\big) &= \frac{1}{2}\int |p_{Of_\theta}(z \mid x) - p_{f_{\theta^*}}(z \mid x)| p(x)\, dx \\ &\le \frac{1}{2}\int \min\{1, \|Of_\theta(x) - f_{\theta^*}(x)\|_2\} p(x)\, dx \\ &\le \frac{1}{2}\min\left\{1, \int \|Of_\theta(x) - f_{\theta^*}(x)\|_2 p(x)\, dx\right\} \\ &= \frac{1}{2}\min\left\{1, \mathbb{E}[\|Of_\theta(x) - f_{\theta^*}(x)\|_2]\right\}. \end{aligned} \tag{232}$$

结合（223）、（230）和（232），我们证明

$$\begin{aligned} & d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta}(x,z), \mathbb{P}_{f_{\theta^*}}(x,z)\big) \\ & \le \frac{1}{2}\mathbb{E}[\|Of_\theta(x) - f_{\theta^*}(x)\|_2] \\ & \le \frac{1}{2}\sqrt{\mathbb{E}[\|Of_\theta(x) - f_{\theta^*}(x)\|_2^2]} \\ & \le \frac{1}{2}\sqrt{\frac{1}{(2\sqrt{2}-2)\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T]\big)}\mathbb{E}\big[\big(f_\theta(x)^T f_\theta(x') - f_{\theta^*}(x)^T f_{\theta^*}(x')\big)^2\big]} \\ & \le \frac{1}{2}\sqrt{\frac{2+e+e^{-1}}{(2\sqrt{2}-2)\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T]\big)} H\big(\mathbb{P}_{f_\theta}(x,x',t), \mathbb{P}_{f_{\theta^*}}(x,x',t)\big)}. \end{aligned} \tag{233}$$

因此，我们证明引理 D.1。 □

引理 D.1 直接推出引理 6.1。

引理 *6.1*的证明。 对于任意 $\theta \in \Theta$，我们选择满足引理 D.1的 $O \in \mathbb{R}^{r\times r}$。于是有

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathbb{P}_{f_\theta, O^T\beta^*}(x,y), \mathbb{P}_{f_{\theta^*},\beta^*}(x,y)\big) &= d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta,\beta^*}(x,y), \mathbb{P}_{f_{\theta^*},\beta^*}(x,y)\big) \\ &\le d_{\mathrm{TV}}\big(\mathbb{P}_{Of_\theta}(x,z), \mathbb{P}_{f_{\theta^*}}(x,z)\big) \\ &\le c\cdot\sqrt{\frac{1}{\sigma_{\min}(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T])}}\cdot H\big(\mathbb{P}_{f_\theta}(x,x',t), \mathbb{P}_{f_{\theta^*}}(x,x',t)\big). \end{aligned}$$

因此，我们证明模型是 κ^{-1}-弱信息的，其中

$$\kappa = c\cdot\sqrt{\frac{1}{\sigma_{\min}(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T])}}. \tag{234}$$

这里 c 是一些绝对常数。

□

D.2 定理 6.2 的证明

ˆˆ 本节中，我们证明定理 6.2。设 θ， β 是算法 1的输出。令 ℓ 为˜平方损失，令 ℓ 为其截断形式，截断水平为 L。式（1）中定义的最优预测器具有如下闭式解

$$g_{\theta,\beta}(x) = \mathbb{E}_{\theta,\beta}[y \mid x] = \beta^T f_\theta(x). \tag{235}$$

我们有如下保证。

引理 D.3. 设截断水平为 $L = 36(D^2+1)\log n$。则有

$$\sup_{\theta,\beta} \left\{ \mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta,\beta}(x),y\big)\right] - \mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta,\beta}(x),y\big)\right] \right\} \le \sqrt{\frac{18(D^2+1)\log n}{\pi n}}. \tag{236}$$

引理 *D.3* 的证明。 注意

$$\big(g_{\theta,\beta}(x)-y\big)\big|x = \big(\beta^T f_\theta(x)-y\big)\big|x \sim \mathcal{N}\big(\beta^T f_\theta(x) - \beta^{*T} f_{\theta^*}(x), 1\big) \tag{237}$$

我们将 $c(x) := \beta^T f_\theta(x) - \beta^{*T} f_{\theta^*}(x)$ 记为。由此可得 $|c(x)| \le 2D$。因此，对于任意 θ, β，都有

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta,\beta}(x),y\big) - \tilde{\ell}\big(g_{\theta,\beta}(x),y\big) \mid x\right] \\
&= \mathbb{E}_{\theta^*,\beta^*}\left[\Big(\big(g_{\theta,\beta}(x)-y\big)^2 - L\Big)\mathbb{1}_{\{(g_{\theta,\beta}(x)-y)^2 > L\}} \mid x\right] \\
&= \int_{\sqrt{L}}^{+\infty} (u^2-L)\cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(u-c(x)\right)^2}{2}}\, du \\
&= \int_{\sqrt{L}-c(x)}^{+\infty} \big((u+c(x))^2 - L\big)\cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du \\
&= \frac{\sqrt{L}+c(x)}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{L}-c(x)\right)^2}{2}} + \frac{1+c(x)^2-L}{\sqrt{2\pi}} \int_{\sqrt{L}-c(x)}^{+\infty} e^{-\frac{u^2}{2}}\, du \\
&\le \frac{\sqrt{L}+c(x)}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{L}-c(x)\right)^2}{2}} \quad (L \ge 4D^2+1 \ge c(x)^2+1) \\
&\le \frac{2(\sqrt{L}-c(x))}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{L}-c(x)\right)^2}{2}} \quad (L \ge 36D^2 \ge (3c(x))^2) \\
&\le \frac{2(\sqrt{L}-2D)}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{L}-2D\right)^2}{2}} \\
&\le \frac{\sqrt{L}}{\sqrt{2\pi}} e^{-\frac{L}{8}} \quad (\sqrt{L}-2D \ge \frac{\sqrt{L}}{2})
\end{aligned} \tag{238}$$

因此，我们证明了

$$\begin{aligned}
&\sup_{\theta,\beta}\left\{\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta,\beta}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta,\beta}(x),y\big)\right]\right\}\\
&\le \mathbb{E}_{\theta^*,\beta^*}\left[\sup_{\theta,\beta}\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta,\beta}(x),y\big)-\tilde{\ell}\big(g_{\theta,\beta}(x),y\big)\mid x\right]\right]\\
&\le \frac{\sqrt{L}}{\sqrt{2\pi}}e^{-\frac{L}{8}}\\
&\le \sqrt{\frac{18(D^2+1)\log n}{\pi n}}. \quad (L=36(D^2+1)\log n)
\end{aligned} \tag{239}$$

□

^^ ~引理 D.4. 假设 θ,β 是算法 *1*的输出。设 ℓ 为截断水平为 L的截断平方损失。则存在绝对常数 c 使得，以至少$1-\delta$的概率，满足

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta^*,\beta^*}(x),y\big)\right]\\
&\le c\kappa L\cdot\sqrt{\frac{1}{m}\log\frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta),1/m^2\big)}{\delta}}+cL\sqrt{\frac{\log 1/\delta}{n}}+c\sqrt{L}\sup_{\theta\in\Theta}R_n(\mathcal{G}_{\theta,\mathcal{B}}),
\end{aligned} \tag{240}$$

其中

$$\kappa=c_3\sqrt{\frac{1}{\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x)f_{\theta^*}(x)^T]\big)}}$$

*fo*其中 c_3 是某些绝对常数。这里， $R_n(\mathcal{G}_{\theta,\mathcal{B}})$ 是定义如下的拉德马赫复杂度：

$$R_n(\mathcal{G}_{\theta,\mathcal{B}})=\mathbb{E}\left[\sup_{\beta\in\mathcal{B}}\frac{2}{n}\sum_{i=1}^{n}\sigma_i g_{\theta,\beta}(x_i)\right], \tag{241}$$

其中 σ_i 是拉德马赫随机变量。

引理 *D.4* 的证明。 根据引理 B.7 和引理 6.1 ， 引理 D.4 可直接由定理3.7 以及以下事实推出： $\tilde{\ell}$ 是 $2\sqrt{L}$-利普希茨。

□

根据引理 D.3 和引理 D.4 ， 我们现在可以证明定理6.2。

定理的证明 *6.2*。 注意

$$\begin{aligned}
\mathrm{Error}_\ell(\hat{\theta},\hat{\beta})&=\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta^*,\beta^*}(x),y\big)\right]\\
&=\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]\\
&\quad+\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta^*,\beta^*}(x),y\big)\right]\\
&\quad+\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta^*,\beta^*}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta^*,\beta^*}(x),y\big)\right]\\
&\le\sup_{\theta,\beta}\left\{\mathbb{E}_{\theta^*,\beta^*}\left[\ell\big(g_{\theta,\beta}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta,\beta}(x),y\big)\right]\right\}\\
&\quad+\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\hat{\theta},\hat{\beta}}(x),y\big)\right]-\mathbb{E}_{\theta^*,\beta^*}\left[\tilde{\ell}\big(g_{\theta^*,\beta^*}(x),y\big)\right].
\end{aligned} \tag{242}$$

令 t 截断水平为 $L = 36(D^2+1)\log n$。根据引理 D.3 和引理 D.4，我们有

$$\begin{aligned}
&\mathrm{Error}(\hat{\theta}, \hat{\beta}) \\
&\le c\kappa L \cdot \sqrt{\frac{1}{m}\log\frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta), 1/m^2\big)}{\delta}} + cL\sqrt{\frac{\log 1/\delta}{n}} + c\sqrt{L}\sup_{\theta\in\Theta} R_n(\mathcal{G}_{\theta,\mathcal{B}}) \\
&\quad + \sqrt{\frac{18(D^2+1)\log n}{\pi n}}.
\end{aligned} \tag{243}$$

对于拉德马赫复杂度，我们有

$$\begin{aligned}
R_n(\mathcal{G}_{\theta,\mathcal{B}}) &= \mathbb{E}\left[\sup_{\beta\in\mathcal{B}}\frac{2}{n}\sum_{i=1}^n \sigma_i g_{\theta,\beta}(x_i)\right] \\
&= \mathbb{E}\left[\sup_{\beta\in\mathcal{B}}\frac{2}{n}\sum_{i=1}^n \sigma_i \beta^T f_\theta(x_i)\right] \\
&\le \frac{2D}{\sqrt{n}},
\end{aligned} \tag{244}$$

其中，最后一个不等式由引理 B.6 得出。结合式（243）和（244），我们有

$$\begin{aligned}
&\mathrm{Error}(\hat{\theta}, \hat{\beta}) \\
&\le c\kappa L \cdot \sqrt{\frac{1}{m}\log\frac{N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta), 1/m^2\big)}{\delta}} + cL\sqrt{\frac{\log 1/\delta}{n}} + 2cD\sqrt{\frac{L}{n}} \\
&\quad + \sqrt{\frac{18(D^2+1)\log n}{\pi n}} \\
&= \tilde{\mathcal{O}}\left(\kappa L\sqrt{\frac{\log N_{[\,]}\big(\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\mathcal{F}_\theta), 1/m^2\big)}{m}} + L\sqrt{\frac{1}{n}}\right),
\end{aligned} \tag{245}$$

其中 $L = 36(D^2+1)\log n$ 且

$$\kappa = c_3\sqrt{\frac{1}{\sigma_{\min}\big(\mathbb{E}[f_{\theta^*}(x) f_{\theta^*}(x)^T]\big)}}$$

其中某些绝对常数为 c_3。 □

E 两阶段最大似然估计的失败

为简单起见，以下我们考虑没有辅助信息可用的情形，即我们可以访问无标签数据 $\{x_i\}_{i=1}^m$ 和有标签数据 $\{x_j, y_j\}_{j=1}^n$。另一种自然的方案是使用两阶段最大似然估计（算法 2）。具体来说，在第一阶段，我们使用最大似然估计，基于无标签数据 $\{x_i\}_{i=1}^m$来估计 ϕ^*。在第二阶段，我们再次使用最大似然估计，基于预训练的 ϕ 和有标签数据 $\{x_j, y_j\}_{j=1}^n$来估计 ψ^*。

算法 2 两阶段最大似然估计

1: 输入： $\{x_i\}_{i=1}^m$, $\{(x_j, y_j)\}_{j=1}^n$ ^2: 使用无标签数据 $\{x_i\}_{i=1}^m$ 通过最大似然估计学习 ϕ：

$$\hat{\phi} \leftarrow \arg\max_{\phi \in \Phi} \sum_{i=1}^{m} \log p_\phi(x_i).$$

^ ^3: 固定 ϕ，并使用有标签数据 $\{(x_j, y_j)\}_{j=1}^n$，通过最大似然估计学习 ψ：

$$\hat{\psi} \leftarrow \arg\max_{\psi \in \Psi} \sum_{j=1}^{n} \log p_{\hat{\phi},\psi}(x_j, y_j).$$

^ ^4: 输出： ϕ 和 ψ。

注意，两阶段最大似然估计并未将学习过程直接与损失函数联系起来。因此，评估超额风险的唯一方法是研究 $\mathbb{P}_{\hat{\phi},\hat{\psi}}(x, y)$ 与 $\mathbb{P}_{\phi^*,\psi^*}(x, y)$ 之间的全变差距离。在预训练阶段，最大似然估计保证估计量 $\mathbb{P}_{\hat{\phi}}$ 与 $\mathbb{P}_{\phi^*}$ 之间的全变差距离很小（定理 3.3）。然而，对于某些 x，$p_{\phi,\psi^*}(x, y) = -\infty$，$\mathbb{P}_{\hat{\phi}}(x) = 0$，仍有可能出现 $\mathbb{P}_{\phi^*}(x) \neq 0$。在下游任务的学习中，对于某些有标签数据，这种现象可能导致 $\log p_{\hat{\phi},\psi}(x, y)$，从而极大地影响最大似然估计在估计 ψ^* 时的表现，并最终导致第二阶段失败。受此想法启发，我们给出以下定理。

定理 E.1. 存在 $\Phi, \Psi, \phi^* \in \Phi, \psi^* \in \Psi$，使得对于任意常数 $c > 0$，都存在 $m, n \geq c$，使得以至少 $\frac{1}{2}(1 - e^{-1})e^{-1}$ 的概率，我们有

$$d_{\mathrm{TV}}\left(\mathbb{P}_{\hat{\phi},\hat{\psi}}(x, y), \mathbb{P}_{\phi^*,\psi^*}(x, y)\right) \geq \frac{1}{8},$$

其中 $\hat{\phi}$ 和 $\hat{\psi}$ 是算法 2 的输出。

定理 E.1 的证明。我们按如下方式构造反例。令 $(x, y, z) \in \mathbb{N}_+ \times \mathbb{N}_+ \times \mathbb{N}_+$。我们假设真实参数 $(\phi^*, \psi^*) = (\phi_1, \psi_1)$，其满足

$$\mathbb{P}_{\phi_1}(x = k, z = k) = \frac{1}{2^k} \ \ \forall k \in \mathbb{N}_+, \quad \mathbb{P}_{\phi_1}(x = m, z = n) = 0 \ \ \forall m \neq n,$$
$$\mathbb{P}_{\psi_1}(y = k | z = k) = 1, \ \forall k \in \mathbb{N}_+.$$

对于 $i \geq 2$，我们定义 $\mathbb{P}_{\phi_i}$ 如下，

$$\mathbb{P}_{\phi_i}(x = 1, z = 1) = \frac{1}{2} + \frac{1}{2^i}, \quad \mathbb{P}_{\phi_i}(x = k, z = k) = \frac{1}{2^k} \ \forall k \notin \{1, i\}$$
$$\mathbb{P}_{\phi_i}(x = m, z = n) = 0 \ \forall m \neq n \text{ or } m = n = i.$$

对于任意 $k \in \mathbb{N}_+$，我们定义 $\mathbb{P}_{\psi_2}$ 如下：

$$\mathbb{P}_{\psi_2}(y = 1 | z = k) = \frac{1}{4}, \quad \mathbb{P}_{\psi_2}(y = 2 | z = k) = \frac{1}{2}$$
$$\mathbb{P}_{\psi_2}(y = j | z = k) = \frac{1}{2^j} \ \forall j \notin \{1, 2\}.$$

我们记作 $\Phi:=\{\phi_i \,|\, i\in\mathbb{N}_+\}$ 和 $\Psi:=\{\psi_1,\psi_2\}$。接下来，我们将证明算法 2 在此情形下失效。回顾一下，我们分别用 $\{x_i\}_{i=1}^m$ 和 $\{x_j,y_j\}_{j=1}^n$ 表示无标签数据和有标签数据。我们有如下观察：

- 我们定义 $i:=\min\{k\neq 1\,|\,k\notin\{x_i\}_{i=1}^m\}$。若有 $1\in\{x_i\}_{i=1}^m$，则似然函数 ϕ 的最大化解满足 $\phi=\phi_i$。

  假设对于某个 $i\neq 1$ 和 $i\in\{y_j\}_{j=1}^n$，有 $\phi=\phi_i$。于是我们有 $\psi=\psi_2$。

  我们定义事件 $\mathcal{E}:=\{\exists i\neq 1,,$ 使得 $\phi=\phi_i$ 且 $i\in\{y_j\}_{j=1}^n\}$。在事件 $\mathcal{E}$ 下，对于某个 $i\neq 1$ 和 $\psi=\psi_2$，有 $\phi=\phi_i$，这意味着

$$\begin{aligned}
d_{\mathrm{TV}}\big(\mathbb{P}_{\hat{\phi},\hat{\psi}}(x,y),\mathbb{P}_{\phi^*,\psi^*}(x,y)\big) &= \frac{1}{2}\int\int |p_{\phi_i,\psi_2}(x,y)-p_{\phi_1,\psi_1}(x,y)|\,dxdy\\
&\geq \frac{1}{2}\int\left|\int p_{\phi_i,\psi_2}(x,y)-p_{\phi_1,\psi_1}(x,y)\,dx\right|dy\\
&= \frac{1}{2}\int |p_{\phi_i,\psi_2}(y)-p_{\phi_1,\psi_1}(y)|\,dy\\
&\geq \frac{1}{2}|\mathbb{P}_{\phi_i,\psi_2}(y=2)-\mathbb{P}_{\phi_1,\psi_1}(y=2)| = \frac{1}{8}
\end{aligned}\tag{246}$$

接下来，我们只需给事件 $\mathcal{E}$ 的概率提供一个下界。注意到

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}\big(\cup_{i=2}^{\infty}\{\hat{\phi}=\phi, i\in\{y_j\}_{j=1}^n\}\big)\\
&= \sum_{i=2}^{\infty}\mathbb{P}\big(\hat{\phi}=\phi_i, i\in\{y_j\}_{j=1}^n\big)\\
&= \sum_{i=2}^{\infty}\mathbb{P}(\hat{\phi}=\phi_i)\cdot\mathbb{P}\big(i\in\{y_j\}_{j=1}^n\big)\\
&= \sum_{i=2}^{\infty}\left(1-\left(1-\frac{1}{2^i}\right)^n\right)\cdot\mathbb{P}(\hat{\phi}=\phi_i).
\end{aligned}\tag{247}$$

因此，对于任意 $L\geq 2$，都有

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\geq \sum_{i=2}^{L}\left(1-\left(1-\frac{1}{2^i}\right)^n\right)\cdot\mathbb{P}(\hat{\phi}=\phi_i)\\
&\geq \left(1-\left(1-\frac{1}{2^L}\right)^n\right)\cdot\mathbb{P}\big(\exists 2\leq i\leq L,\hat{\phi}=\phi_i\big).
\end{aligned}\tag{248}$$

请注意

$$
\begin{aligned}
&\mathbb{P}\big(\exists 2 \le i \le L, \hat{\phi} = \phi_i\big) \\
&= \mathbb{P}\Big(\big\{1 \in \{x_i\}_{i=1}^m\big\} \cap \big\{\exists 2 \le i \le L, i \notin \{x_i\}_{i=1}^m\big\}\Big) \\
&\ge \mathbb{P}\Big(\big\{1 \in \{x_i\}_{i=1}^m\big\} \cap \big\{L \notin \{x_i\}_{i=1}^m\big\}\Big) \\
&\ge \mathbb{P}\big(1 \in \{x_i\}_{i=1}^m\big) + \mathbb{P}\big(L \notin \{x_i\}_{i=1}^m\big) - 1 \\
&= \mathbb{P}\big(L \notin \{x_i\}_{i=1}^m\big) - \mathbb{P}\big(1 \notin \{x_i\}_{i=1}^m\big) \\
&= \left(1 - \frac{1}{2^L}\right)^m - \frac{1}{2^m}.
\end{aligned} \tag{249}
$$

结合（248）和（249），对于任意 $L \ge 2$

$$
\mathbb{P}(\mathcal{E}) \ge \left(1 - \left(1 - \frac{1}{2^L}\right)^n\right) \cdot \left(\left(1 - \frac{1}{2^L}\right)^m - \frac{1}{2^m}\right). \tag{250}
$$

令 $2^L = m = n$，可得

$$
\mathbb{P}(\mathcal{E}) \ge \left(1 - \left(1 - \frac{1}{m}\right)^m\right) \cdot \left(\left(1 - \frac{1}{m}\right)^m - \frac{1}{2^m}\right) \to (1 - e^{-1}) \cdot e^{-1}, \text{ as } m \to \infty. \tag{251}
$$

因此，对于任意 $c > 0$，存在 m、$n \ge c$，使得

$$
\mathbb{P}(\mathcal{E}) \ge \frac{1}{2}(1 - e^{-1}) \cdot e^{-1}.
$$

□