arXiv:2303.01462v1 [cs.LG] 2 Mar 2023

基于间隔最大化的 KKT 条件研究线性分类器和 Leaky ReLU 网络中的良性过拟合

SpencerFrei* Gal Vardi*加州大学伯克利分校、TTI-Chicago 和
希伯来大学 frei@berkeley.edu galvardi@ttic.edu

Peter L. Bartlett 加州
大学伯克利分校和谷歌
peter@berkeley.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu

合作 深度学习理论基础deepfoundations.ai方面的合作

2023年3月3日

摘要

通过逻辑损失的梯度流训练的线性分类器和 Leaky ReLU 网络，具有趋向于满足间隔最大化的 Karush–Kuhn–Tucker（KKT）条件的解的隐式偏置。本文确立了若干种设定，在这些设定中，满足这些 KKT 条件意味着线性分类器和两层带泄漏 ReLU 网络会出现良性过拟合：估计量能够插值拟合含噪训练数据，同时对测试数据也有良好的泛化能力。这些设定包括先前工作中考察过的含噪类条件高斯分布的变体，也包括此前尚未观察到良性过拟合的新分布设定。我们的证明的关键在于以下观察：当训练数据近似正交时，满足各自间隔最大化问题的 KKT 条件的线性分类器和 Leaky ReLU 网络，其行为都类似于对训练样本近似均匀的平均。

1 引言

“良性过拟合”这一现象——指模型能够完美拟合含噪训练数据，同时仍能很好地泛化到未见数据的情形——近年来受到广泛关注。在 Zhang 等人最初的实验之后 [Zha+17]，研究者一直试图理解，尽管统计学习理论长期以来的直觉认为，对噪声过拟合会导致样本外预测性能不佳，这一现象为何仍会发生。

在这项工作中，我们针对线性分类器和两层带泄漏 ReLU 网络，在分类任务中的良性过拟合方面给出了若干新结果。我们研究采用指数尾损失函数（例如逻辑损失）时，经验风险上的梯度流。在对数据分布作出某些假设的条件下，我们证明梯度流会收敛到满足良性过拟合的解：这些预测器

*贡献相同。

能够插值含噪训练数据，同时对未见过的测试数据实现良好的泛化。我们的结果从两个方面拓展了现有研究：首先，我们证明了两层带泄漏 ReLU 网络中的良性过拟合，而现有结果并未涵盖这类模型。[1] 其次，我们在新的分布设定（即对数据分布的假设）下刻画了良性过拟合。

我们考虑的第一种分布设定是含噪次高斯分布 $(x, y) \sim \mathsf{Psg}$，其中标签由 x 的单个分量生成，并以概率 η翻转为相反符号。我们证明，如果该分量的方差相对于其他分量的方差足够大，且协方差矩阵的秩相对于样本数量足够高，那么通过梯度流训练的线性分类器和 Leaky ReLU 网络会表现出良性过拟合。在第二种分布设定中，我们考虑分布 $\mathsf{P}_{\mathsf{clust}}$，其中输入 x 从 k近乎正交的簇中均匀抽取，标签由簇决定，并以概率 η翻转为相反符号。我们证明，在对簇的尺度和相关性作出一些假设的条件下，在线性分类器和 Leaky ReLU 网络上运行梯度流会产生表现出良性过拟合的分类器。这一设定并未被此前关于良性过拟合的研究所涵盖，并且实质上推广了此前关于线性分类中的良性过拟合 [CL21；WT21] 以及具有平滑Leaky 激活函数的神经网络 [FCB22]的若干结果。

我们的证明通过分析梯度流的隐式偏置得出。Lyu and Li [LL20] 以及 Ji 和 Telgar- sky [JT 20] 证明了，在使用指数尾损失函数训练齐次神经网络时，梯度流会偏向于参数空间中间隔最大的解。也就是说，如果经验风险降至足够小的值，那么梯度流的方向会收敛到一个满足间隔最大化问题的 Karush–Kuhn–Tucker（KKT）条件的解。我们提出了新的证明方法，表明在上述分布设定中，任何满足这些 KKT 条件的解都会出现良性过拟合。具体而言，我们证明，在我们的设定中，每个 KKT 点都具有线性决策边界，即使是在 Leaky ReLU 网络的情况下也是如此。该线性决策边界可以表示为训练样本的加权和，且所有样本的权重大致均衡。利用这一均衡性，我们得以证明良性过拟合会发生。

相关工作

良性过拟合。良性过拟合现象近来引起了广泛关注，并在多种情形下得到研究，例如线性回归 [Has+20; BHX20; Bar+20; Mut+20; NDR20; CL20; Koe+21; WX20; TB20; Zho+22; WDY22; CLB21; BL21; Sha22]，核回归 [LR20; MM19; LRZ20; Mal+22; RZ19; BHM18]，以及分类 [CL21; WT21; CGB21; Mut+21; Mon+20; Sha22; FCB22; Cao+22; McR+22; LR21; TOS20; WMT21; Don+22]。下面我们将讨论几项与本文结果最相关的分类领域良性过拟合研究。

与线性回归不同，在线性分类中，梯度流已知会收敛到的解，即最大间隔预测器，并没有闭式表达式。因此，分析线性分类中的良性过拟合更具挑战性。Chatterji 和 Long [CL21] 以及 Wang 和 Thrampoulidis [WT21] 证明了高维次高斯混合模型中的线性分类存在良性过拟合。我们的结果将次高斯混合分布中的良性过拟合作为一个特例，其结论与他们的结果类似。Cao、Gu 和 Belkin [CGB21]也研究了次高斯混合模型中的良性过拟合，但没有考虑标签翻转噪声。Muthukumar 等人 [Mut+21] 研究了参数化最大间隔分类器的行为，

[1] Frei、Chatterji 和 Bartlett [FCB22] 证明了在具有 平滑 Leaky ReLU 激活函数的两层网络中存在良性过拟合，正如我们稍后将讨论的那样。

该分类器处于带有标签翻转噪声的判别式分类模型中，他们将最大间隔分类器的行为与普通最小二乘解联系起来。他们表明，在某些条件下，所有训练数据点都会成为最大间隔分 [HMX21]（另见 Hsu、Muthukumar 和 Xu [HMX+20] ）。Montanari 等人 [Mon22] 研究了输入服从高斯分布、标签根据逻辑斯蒂链接函数生成的设定。他们推导出了最大间隔线性分类器的渐近预测误差表达式，并假设维度与样本量之比收敛到某个固定的正极限。Shamir Sha也研究了线性分类，并在一种不同于上述研究及我们所采用设定的分布设定下证明了良性过拟合。

非线性神经网络中的良性过拟合甚至更难以理解。Frei、Chatterji 和 Bartlett[FCB22] 展示了在高维次高斯混合模型下，采用平滑的Leaky ReLU 激活函数的两层网络中的良性过拟合；我们在第 5 节末尾将我们的结果与他们的结果进行比较。Cao 等人 [Cao+22] 研究了使用逻辑损失训练两层卷积神经网络时的良性过拟合，但他们没有考虑我们所研究的标签翻转噪声。

隐式偏置。 近年来，关于神经网络隐式偏置的文献迅速增多（综述见 Vardi [Var22] ）。接下来，我们将讨论适用于使用梯度流进行线性分类，或适用于在分类设定下通过梯度流训练的非线性两层网络的研究结果。

Soudry 等人 [Sou+18]表明，对于具有指数尾损失（例如指数损失和逻辑损失）的线性可分二分类问题，梯度下降会收敛到最大ℓ_2-间隔方向。此分析后来扩展至其他损失函数、更紧的收敛速率、不可分数据，以及基于梯度的优化算法的变体 [Nac+19；JT18；Ji+20；Gun+18；Sha20；JT21；NSS19；JST21]。

Lyu and Li [LL20] 以及 Ji and Telgarsky [JT20] 证明，使用指数尾分类损失训练的齐次神经网络（特别是本文关注的两层带泄漏 ReLU 网络）在方向上收敛到最大间隔问题的一个 KKT 点。我们注意到，上述 KKT 点未必是最大间隔问题的全局最优解 [VSS21; Lyu+21]。最近，Kunin 等人 [Kun+22] 扩展了这一结果，证明在一类更广泛的、称为准齐次网络的网络中也存在偏向间隔最大化的趋势。Lyu 等人 [Lyu+21]、Sarussi、Brutzkus 和 Globerson [SBG21], 以及 Frei 等人 [Fre+23] 研究了线性可分数据下两层带泄漏 ReLU 网络的隐式偏置，并证明在一些额外假设下，梯度流会收敛到线性分类器。具体而言，Frei 等人 [Fre+23] 分析了使用近乎正交数据训练的 Leaky ReLU 网络中的隐式偏置，而我们对 Leaky ReLU 网络的分析正是建立在他们的结果之上（详见第 3 节）。此外，Vardi、Yehudai 和 Shamir [VYS22], 研究了 ReLU 网络在近乎正交数据下的隐式偏置，并证明其偏向于不具备对抗鲁健性的网络。其他研究梯度流在非线性两层网络中分类隐式偏置的工作包括 [CB20; PL20; SVL22; TVS22]。

2 预备知识

记号。我们用 $\|x\|$ 表示向量 x的欧几里得范数；对于矩阵 W，用 $\|W\|_F$表示其弗罗贝尼乌斯范数，用 $\|W\|_2$ 表示其谱范数。我们用 $\mathbb{I}(z)$ 表示指示函数，因此当 $z \geq 0$时，$\mathbb{I}(z) = 1$；否则为 0。我们将 sign(z) 作为函数：当 $z > 0$时取值为 1，否则取值为 -1。对于整数 $n \in \mathbb{N}$，我们使用 $[n] = \{1, \ldots, n\}$。均值为 a、方差为 σ^2的高斯分布记为 N(a, σ^2)，而均值为 μ、协方差矩阵为 Σ 的多元高斯分布记为

N(μ, Σ)。我们将两个数 a、b中的最小值记为 $a \wedge b$，最大值记为 $a \vee b$。对于向量$x \in \mathbb{R}^d$，我们用 $[x]_i \in \mathbb{R}$ 表示该向量的第 i个分量，并用 $[x]_{i:j} \in \mathbb{R}^{j-i+1}$ 表示分量为 $[x]_i$、$[x]_{i+1}$、……、$[x]_j$的向量。我们使用标准的大 O 记号 $O(\cdot)$、$\Omega(\cdot)$来省略通用常数，其中 $O(\cdot)$、$\Omega(\cdot)$还省略了对数因子。我们将与维度 d、样本数量 n、失败概率 δ 或网络中的神经元数量 m无关的量称为常数。

设定。我们考虑这样的分类任务：训练数据 $\{(x_i, y_i)\}_{i=1}^n$ 独立同分布地取自定义在 $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ 上的分布 P。本文研究两种不同的模型。第一种是最大间隔线性分类器 $x \mapsto \text{sign}(\langle w, x\rangle)$，它们是以下约束优化问题的解：

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{such that for all } i \in [n], \quad y_i \langle w, x_i\rangle \geq 1. \tag{1}$$

Soudry 等人 [Sou+18]指出，对逻辑损失等指数尾损失进行梯度下降，会隐式偏向这类解。我们将证明，在若干设定下，问题 (1)的任何解都会表现出良性过拟合。

第二种模型是带有 Leaky ReLU 激活函数的两层神经网络，其中第一层 $W \in \mathbb{R}^{m \times d}$ 经过训练，而第二层权重 $\{a_j\}_{j=1}^m$ 固定为随机初始化值：

$$f(x; W) := \sum_{j=1}^m a_j \phi(\langle w_j, x\rangle), \quad \phi(q) = \max(\gamma q, q), \quad \gamma \in (0, 1). \tag{2}$$

为简单起见，我们假设 m 是偶数，并且一半神经元满足 $a_j = 1/\sqrt{m}$，另一半神经元满足 $a_j = -1/\sqrt{m}$。我们考虑一个二分类任务，其训练数据为$S = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$。我们将神经网络$f(x; W)$在训练数据 S上的间隔最大化问题定义为

$$\min_{W \in \mathbb{R}^{m \times d}} \|W\|_F^2 \quad \text{such that for all } i \in [n], \quad y_i f(x_i; W) \geq 1. \tag{3}$$

回顾非光滑优化问题的 Karush–Kuhn–Tucker（KKT）条件定义（参见 Lyu and Li [LL20]和 Dutta 等人 [Dut+13]）。设 $h: \mathbb{R}^p \to \mathbb{R}$ 是一个局部 Lipschitz 函数。Clarke 次微分 [Cla+08] 在 $\theta \in \mathbb{R}^p$ 处是凸集

$$\partial^\circ h(\theta) := \text{conv}\left\{ \lim_{s \to \infty} \nabla h(\theta_s) \;\middle|\; \lim_{s \to \infty} \theta_s = \theta,\ h \text{ is differentiable at } \theta_s \right\}.$$

如果 h在 θ 处连续可微，则 $\partial^\circ h(\theta) = \{\nabla h(\theta)\}$。给定局部 Lipschitz 函数$h$、$g_1$、……、$g_n$：$\mathbb{R}^p \to \mathbb{R}$，我们称 $\theta \in \mathbb{R}^p$ 为该问题的可行点

$$\min h(\theta) \quad \text{s.t.} \quad \text{for all } n \in [N],\ g_n(\theta) \leq 0,$$

如果 θ 对所有 $n \in [N]$都满足 $g_n(\theta) \leq 0$。如果某个可行点 θ 是 KKT 点，则存在λ_1、……、$\lambda_N \geq 0$ 使得

1. $0 \in \partial^\circ h(\theta) + \sum_{n \in [N]} \lambda_n \partial^\circ g_n(\theta)$;
2. 对于所有 $n \in [N]$，我们都有 $\lambda_n g_n(\theta) = 0$。

我们将证明，在多种情形下，问题 (3) 的任何 KKT 点都能很好地泛化，即使有固定比例的训练标签是均匀随机标签也是如此。由于问题 (3)的任何可行点都能插值训练数据，这意味着该网络表现出良性过拟合。

问题 （3）的 KKT 点会自然地出现在神经网络的训练过程中。对于损失函数ℓ： $\mathbb{R} \to [0, \infty)$ 以及神经网络 $f(x;\ W)$的参数 W，定义在 ℓ 下的经验风险为

$$\widehat{L}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i f(x_i; W)\big).$$

目标函数 $L(W)$ 的梯度流是由初始点 $W(0)$ 定义的轨迹 $W(t)$，并且 $W(t)$ 满足微分方程 $\frac{\mathrm{d}}{\mathrm{d}t} W(t) \in -\partial^\circ \widehat{L}(W(t))$，初始点为 $W(0)$。由于网络 $f(x;\ \cdot)$ 是 1-齐次，Lyu and Li [LL20] 以及 Ji and Telgarsky [JT20] 的近期研究表明，若 ℓ 是指数损失 $\ell(q) = \exp(-q)$ 或逻辑损失 $\ell(q) = \log(1 + \exp(-q))$，那么只要存在某个时刻 t_0 使得 $L(W(0)) < \log(2)/n$，梯度流就会在方向上收敛到问题（3）的一个 KKT 点；也就是说，存在问题（3）的某个 KKT 点 W^*，使得 $\frac{W(t)}{\|W(t)\|} \to \frac{W^*}{\|W^*\|}$。因此，尽管存在许多能够正确分类训练数据的神经网络，但如果梯度流到达损失足够小的点，那么它只会产生在方向上收敛到满足问题（3）KKT 条件的网络。请注意，这并不意味着 $W(t)$ 会在方向上收敛到问题（3）的全局最优解 [VSS21; Lyu+21]。这与线性分类中的间隔最大化问题（见式（1））形成对比；该问题的约束和目标函数都是线性的，因此 KKT 条件是全局最优性的必要且充分条件。

3 近似正交数据的 KKT 点的性质

本节将证明，当训练数据近似正交（稍后将对此给出形式化定义）时，(i) 线性最大间隔问题 (1) 的 KKT 点和 (ii) 非线性Leaky ReLU 网络(3) 的 KKT 点，其决策边界都具有加权平均估计量的形式$w = \sum_{i=1}^{n} s_i y_i x_i$ ，其中$\{s_i\}_{i=1}^{n}$ 均严格为正且同阶；也就是说， w 是训练数据近似均匀的平均值。接下来的章节将利用这一性质，在某些分布假设下证明良性过拟合。我们先定义 p-正交和 τ-均匀分类器。

定义 3.1。记 $R_{min}^2 = \min_i \|x_i\|^2$, $R_{max}^2 = \max_i \|x_i\|^2$， 以及 $R^2 = R_{max}^2 / R_{min}^2$。若训练数据满足 $R_{min}^2 \geq pR^2 n \max_{i \neq j} |\langle x_i, x_j \rangle|$，则称其为 p-正交。

显然，如果训练数据精确正交，那么对于任意 $p > 0$，它都是 p-正交的。与精确正交不同， p-正交允许从一大类分布中独立同分布采样得到的训练数据是 p-正交的，正如我们稍后将看到的。

定义 3.2。我们称 $w \in \mathbb{R}^d$ 关于 $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \{-1, 1\}$ 是 τ-一致的，如果 $w = \sum_{i=1}^{n} s_i y_i x_i$, 其中系数 $\{s_i\}_{i=1}^{n}$ 均严格为正，且 $\frac{\text{最大值}_i\ s_i}{\text{最小值}_i\ s_i} \leq \tau$。

我们的第一个引理表明，如果训练数据在较大的 p 下是 p-正交的，且训练样本的范数都处于同一数量级，那么线性最大间隔分类器就是一个 τ-均匀向量。

命题 3.3. 假设训练数据对于 $p \geq 3$ 是 p-正交的。记 $R^2 = \max_{i,\ j} \|x_i\|^2/\|x_j\|^2$。令 $\widehat{w} = \operatorname{argmin}\{\|w\|^2 : y_i\langle w, x_i\rangle \geq 1\ \forall i\}$ 为最大间隔线性分类器。那么， $\widehat{w}$ 相对于 $\{(x_i, y_i)\}_{i=1}^n$ 是 τ-均匀的，其中

$$\tau = R^2\left(1 + \frac{2}{pR^2 - 2}\right).$$

该命题的证明基于对最大间隔问题的 KKT 条件的分析，详见附录A。注意，当 $p \to \infty$且 $R^2 \to 1$时，我们看到线性最大间隔分类器与 $\sum_{i=1}^n y_i x_i$成正比，即经典样本平均估计量。

接下来，我们证明，当训练数据对于足够大的 p是 p-正交的时，Leaky ReLU 网络间隔最大化问题(3) 的任意 KKT 点都具有与 τ-一致线性分类器相同的决策边界，尽管两层带泄漏 ReLU 网络通常是非线性的。该证明基于 Frei 等人的一项近期工作 [Fre+23]，详见附录B。

命题 3.4。令 $R^2 = \text{最大值}_{i,\ j} \|x_i\|^2/\|x_j\|^2$。令 f 表示 Leaky ReLU 网络(2) ，并令 W表示问题(3)的一个 KKT 点。假设训练数据对于 p是 $p \geq 3\gamma^{-3}$-正交的。那么，存在 $z \in \mathbb{R}^d$ ，使得对于任意 $x \in \mathbb{R}^d$,

$$\operatorname{sign}\left(f(x; W)\right) = \operatorname{sign}\left(\langle z, x\rangle\right),$$

并且 z相对于 $\{(x_i, y_i)\}_{i=1}^n$是 τ-一致的

$$\tau = \frac{R^2}{\gamma^2}\left(1 + \frac{2}{\gamma pR^2 - 2}\right).$$

此外，对于任意初始化 $W(0)$，在 logistic 损失或指数损失下的梯度流，其方向都会收敛到这样的 KKT 点。

命题 3.4 给出了在 非凸 情形下，通过梯度流训练的神经网络分类器极限行为的显式公式。值得强调的是，命题 3.4 并未对网络宽度或初始化作出任何假设，因此这一刻画适用于特征学习阶段的神经网络。最后，注意到当 $p \to \infty$ 且 $R^2 \to 1$时，问题 (3) 的 KKT 点与一个 τ-一致分类器具有相同的决策边界，其中 $\tau \to \gamma^{-2}$。特别地，如果进一步满足 Leaky 参数 $\gamma \to 1$，那么 Leaky ReLU 网络间隔最大化问题的 KKT 点会随着样本均值 $\sum_{i=1}^n y_i x_i$成比例变化，线性最大间隔预测器也是如此。

结合命题 3.3 和 3.4 可知，通过理解 τ-一致分类器 $x \mapsto$ 符号（$\langle \sum_{i=1}^n s_i y_i x_i,\ x\rangle$）的行为，我们便能把握线性最大间隔分类器估计量以及在近乎正交数据上通过梯度流训练的 Leaky ReLU 网络的行为。在接下来的章节中，我们将介绍两种分布设定，并证明该估计量可能出现良性过拟合：它在含噪数据集上实现零训练误差，同时测试误差接近噪声率。

4 次高斯边缘分布下的良性过拟合

本节考虑定义在 (x, y) 上的分布 P_{sg} ，其中 x 具有相互独立的次高斯分量，且只有一个高方差分量决定标签 y，其余分量

x 的方差都很小。设 P_x 是定义在 $\mathbb{R}^d$ 上的分布。我们假设协变量 $x \sim \mathsf{P}_x$ 的均值为零，协方差矩阵为 $\Sigma = \mathbb{E}_{x\sim\mathsf{P}_x}[xx^\top]$，满足 $\Sigma = \mathsf{diag}(\lambda_1, \ldots, \lambda_d)$，其中 $\lambda_1 \geq \cdots \geq \lambda_d$。我们假设 $z := \Sigma^{-1/2}x \sim \mathsf{P}_z$，其中 P_z 是一个分量相互独立、次高斯范数至多为 σ_z 的次高斯随机向量（关于次高斯分布的更多详情，参见 Vershynin [Ver18]）。给定 $x \sim \mathsf{P}_x$，标签按如下方式生成。对于某个标签噪声参数 $\eta \in (0, 1/2)$，以概率 $1-\eta$，我们有$y = \mathsf{sign}([x]_1)$；以概率 η，我们有 $y = -\mathsf{sign}([x]_1)$，其中 $[x]_1$ 表示 x 的第一分量。最后，我们假设对于某个绝对常数 $\beta > 0$，对所有 $t \geq 0$，都有 $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$。在下文中，我们将假设 σ_z、η 和 β 都是绝对常数；只要 d 和 n 相对于这些常数及其他通用常数足够大，我们的结果就成立。

读者或许会对要求 $\mathbb{P}_{z\sim\mathsf{P}_z}(|[z]_1| \leq t) \leq \beta t$ 感到好奇。这是一项技术性假设，确保模型中的“信号”足够强，因为它排除了 $[z]_1$ 的质量高度集中在零附近的可能性。此外，值得注意的是，根据各向同性对数凹分布的反集中性质 [LV07，定理 5.1 和定理 5.14]，如果 z 或 $[z]_1$ 的分布是（各向同性）对数凹的，那么该假设就成立。[2] 该假设还意味着 $\mathbb{E}[|[z]_1|] \geq 1/(4\beta)$，因为取 $t = 1/(2\beta)$ 后，$\mathbb{P}(|z| \geq 1/(2\beta)) \geq 1/2$。原则上，我们也可以采用更一般的条件，例如对某个 $p > 0$，$\mathbb{P}(|[z]_1| \leq t) \leq \beta t^p$；这是一种“软间隔”条件，先前关于学习含噪半空间的研究中已使用过这种条件 [FCG21a；FCG21b]。

我们假设可以获得 n 独立同分布训练样本 $\{(x_i, y_i)\} \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\mathsf{sg}}$。对于给定的失败概率$\delta \in (0, 1/2)$，当常数$C > 1$足够大时，我们对问题参数作出以下假设。

(SG1) 样本数量满足 $n \geq C\log(6/\delta)$。

(SG2) 协方差矩阵满足 $\mathsf{StableRank}(\Sigma_{2:d}) \geq C\log(6n/\delta)$，其中 $\Sigma_{2:d}$ 表示矩阵 $\mathsf{diag}(\lambda_2, \ldots, \lambda_d)$。

(SG3) 协方差矩阵满足 $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)} \geq C\log(6n^2/\delta)$。

提醒读者，矩阵 $M \in \mathbb{R}^{m\times d}$ 的稳定秩为 $\mathsf{StableRank}(M) := \|M\|_F^2/\|M\|_2^2$。我们注意到，$\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)}$ 这一量在 (SG3) 中曾出现在此前关于良性过拟合的研究中：它是 Bartlett 等人 $R_0(\Sigma)$ 提出的“有效秩” [Bar+20] 的平方根。事实上，如果$\sqrt{\mathsf{StableRank}(\Sigma^{1/2})}$ 较大，则该量也较大，因为

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} \geq \frac{\mathrm{tr}(\Sigma)}{\sqrt{\|\Sigma\|_2 \mathrm{tr}(\Sigma)}} = \frac{\sqrt{\mathrm{tr}(\Sigma)}}{\|\Sigma^{1/2}\|_2} = \frac{\|\Sigma^{1/2}\|_F}{\|\Sigma^{1/2}\|_2} = \sqrt{\mathsf{StableRank}(\Sigma^{1/2})}.$$

因此，假设（SG3）大致可以理解为要求矩阵 $\Sigma^{1/2}$ 具有足够大的秩。此外，我们注意到，$\mathsf{StableRank}(\Sigma_{2:d}) = \Theta(1)$ 与 $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)} = \Theta(\sqrt{d})$ 可以同时成立（取 $\Sigma = \mathsf{diag}(\sqrt{d}, \sqrt{d}, 1, \ldots, 1)$）；而 $\mathsf{StableRank}(\Sigma_{2:d}) = \Theta(\sqrt{d})$ 与 $\mathrm{tr}(\Sigma)/\sqrt{\mathrm{tr}(\Sigma^2)} = \Theta(1)$ 也可以同时成立（取 $\Sigma = \mathsf{diag}(d, d^{1/4}, 1, \ldots, 1)$）。因此，假设（SG2）和（SG3）是相互独立的。

我们的第一个引理指出，随着前述假设中的常数 C 增大，训练数据会变得更加正交。

[2] 对于 $z \sim \mathsf{P}_z$，其中 P_z 是对数凹且各向同性的，Lovász 和 Vempala [LV07，定理 5.1] 意味着一维边缘分布 $[z]_1$ 是各向同性且对数凹的。同一参考文献中的定理 5.14 表明，（一维）$[z]_1$ 的密度函数有一个常数上界，由此可得 $\mathbb{P}(|[z]_1| \leq t) \leq \beta t$，其中 $\beta > 0$ 为绝对常数。

引理 4.1。 存在一个绝对常数 $C_1 > 0$（仅取决于 σ_z），使得对于所有足够大的常数 $C > 0$，对于任意 $\delta \in (0, 1/2)$，在假设 *(SG1)* 至 *(SG3)*（针对这些 C 和 δ 定义）下，以至少 $1 - 2\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{sg}}^n$），训练数据是 C/C_1-正交的，并且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \le (1 + C_1/\sqrt{C})^4$。

引理 4.1 的证明，以及本节的所有证明，均见附录 C。回顾命题 3.3 和 3.4，对于 p-正交训练数据，当 $R^2 = \max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \to 1$ 和 $p \to \infty$ 时，线性最大间隔问题（1）的解变为 τ-一致。类似地，带泄漏 ReLU 最大间隔问题的 KKT 点表现得像 τ-一致线性分类器，其中 $\tau \to \gamma^{-2}$ as $p \to \infty$and $R \to 1$。在本节的主定理中，我们证明 τ-一致线性分类器表现出良性过拟合。提醒读者，我们将与维度 d、样本数量 n、失败概率 δ 或网络中的神经元数量 m 无关的量称为常数。

定理 4.2。令 $\tau \ge 1$ 为一个常数，并假设 $\eta \le \frac{1}{2\tau} - \Delta$ 对某些绝对常数 η、$\Delta > 0$ 成立。存在常数 C、$C' > 0$（仅取决于 η、σ_z、β、τ 和 Δ），使得对于任意 $\delta \in (0, 1/7)$，在假设（SG1）至（SG3）（针对这些 C 和 δ 定义）下，以至少 $1 - 7\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{sg}}^n$），若 $u \in \mathbb{R}^d$ 是关于 $\{(x_i, y_i)\}_{i=1}^n$ 的 τ-一致的，则

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle u, x_k\rangle\big), \qquad \textit{while simultaneously,}$$
$$\eta \le \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \ne \operatorname{sign}(\langle u, x\rangle)\big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

特别地，如果 $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么线性分类器 $x \mapsto \operatorname{sign}(\langle u, x\rangle)$ 表现出良性过拟合。

定理 4.2 表明，任何 τ-一致估计量都会表现出良性过拟合，其所能容忍的噪声水平由数量 τ决定。此外，通过考虑 1-一致估计量 $\sum_{i=1}^n y_i x_i$，我们看到存在一种能够容忍接近 $1/2$的噪声水平的估计量。

利用引理4.1 和命题 3.3，我们可以使用定理 4.2 刻画线性最大间隔预测器。

推论 4.3。 假设 $0 < \eta \le 0.49$。存在常数 C、$C' > 0$，使得对于任意 $\delta \in (0, 1/9)$，在假设 *(SG1)* 至 *(SG3)*（针对这些 C和 δ定义）下，以至少 $1 - 9\delta$的概率 $\mathsf{P}_{\mathsf{sg}}^n$，最大间隔线性分类器 $w = \operatorname{argmin}\{\|w\|^2 : y_i\langle w,\ x_i\rangle \ge 1\ \forall i\}$ 满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle w, x_k\rangle\big), \qquad \textit{while simultaneously,}$$
$$\eta \le \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \ne \operatorname{sign}(\langle w, x\rangle)\big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

特别地，如果 $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么 w 表现出良性过拟合。

推论的证明 4.3 可通过简单计算得出（为完整起见，计算过程见附录 C）：根据定理 4.2，当 η接近 $1\,2$时，只要 τ接近 1，我们就能容忍较高的噪声率。根据引理 4.1，随着 C增大，训练数据会变得更加正交，样本范数之比也会更接近 1。根据命题 3.3，这意味着 $\tau \to 1$ 会随着 C 增大。

我们同样可以使用引理 4.1 和命题 3.4 来证明，来自问题 (3) 的 Leaky ReLU 网络的最大间隔问题的 KKT 点同样表现出良性过拟合。

推论 4.4。 假设 $0 < \eta \le \frac{49\gamma^2}{100}$。存在常数 C、$C' > 0$，使得对于任意 $\delta \in (0,\ 1/9)$，在假设 *(SG1)* 至 *(SG3)* （针对这些 C 和 δ定义）下，以至少 $1 - 9\delta$ 的概率$\mathsf{P}_{\mathsf{sg}}^n$，问题 （3）的任意 KKT 点 W 满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big), \quad \textit{while simultaneously,}$$

$$\eta \le \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\Big(y \ne \operatorname{sign}\big(f(x;W)\big)\Big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

特别地，如果 $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么神经网络 $f(x; W)$ 会表现出良性过拟合。此外，对于任意初始化 $W(0)$，梯度流都会在方向上收敛到一个满足上述条件的网络。

推论 4.4 的证明同样需要进行一项简单计算，我们在附录 C中给出。值得注意的是，带泄漏 ReLU 的最大间隔问题 （3）与线性最大间隔问题 （1）的 KKT 点表现出的唯一差异，是其所能容忍的噪声水平：在带泄漏 ReLU 的情形下，泄漏参数越小 γ，所能容忍的噪声就越少；而当 $\gamma \to 1$时，我们便恢复了推论 4.3中的线性最大间隔预测器的表现。此外，推论 4.4中的泛化界与网络中的神经元数量 m 无关。

从上述结果可以看出，要使线性最大间隔分类器或通过梯度流训练的两层带泄漏 ReLU 网络实现良性过拟合，数据必须同时满足两个条件：（1）协方差矩阵的秩足够高，即满足假设（SG2）和（SG3）；（2）第一个坐标的方差必须相对于最后 $d-1$ 个坐标的方差足够大。这里存在一种制约关系，考虑协方差矩阵 $\Sigma = (\xi, 1, \ldots, 1)$（其中 $\xi \ge 1$）即可看出：当 $\xi \to \infty$ 时，$\operatorname{tr}(\Sigma)/\sqrt{\operatorname{tr}(\Sigma^2)} \to 1$，因此，随着信噪比 $\lambda_1^2/\operatorname{tr}(\Sigma_{2:d}^2)$ 增大，满足假设（SG3）就变得更加困难。不过，确实可以同时满足条件（1）和（2）。考虑分布 $\mathsf{P}_{\mathsf{gaus}}$，其定义在 (x, y) 上，$\in \mathbb{R}^d \times \{\pm 1\}$，其中 $x \sim \mathsf{N}(0, \Sigma)$，协方差矩阵为 $\Sigma = \operatorname{diag}(d^\rho, 1, \ldots, 1)$，其中某个 $\rho > 0$；并且 $y =$ 以概率 $1-\eta$取 $\operatorname{sign}([x]_1)$，以概率 η 取 $\operatorname{sign}([x]_1)$，其中某个常数为 $\eta > 0$。在下面的推论中，我们证明，若$\rho \in (1/2, 1)$，则条件（1）和（2）均成立，因此，带泄漏 ReLU 最大间隔问题（3）的 KKT 点会表现出良性过拟合（线性最大间隔分类器也有类似结果）。

推论 4.5。 假设 $0 < \eta \le \frac{49\gamma^2}{100}$。那么，对于分布 $\mathsf{P}_{\mathsf{gaus}}$，对于任意 $\delta \in (0, 1/9)$，若 $\rho \in (1/2, 1)$，$d = \tilde{\Omega}(n^{1/(1-\rho)})$，且 $n = \tilde{\Omega}(1)$，则假设 *(SG1)* 至 *(SG3)* 均成立。此外，以至少 $1 - 9\delta$ 的概率，在 $\mathsf{P}_{\mathsf{gaus}}^n$下，问题 (3) 的 KKT 点表现出良性过拟合：

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big),$$

$$\textit{while simultaneously,} \quad \eta \le \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{gaus}}}\Big(y \ne \operatorname{sign}\big(f(x_k;W)\big)\Big) \le \eta + \tilde{O}\left(d^{\frac{1}{2}(1-2\rho)}\right) = \eta + o_d(1).$$

此外，对于任意初始化 $W(0)$，梯度流在方向上收敛到一个满足上述条件的网络。

Muthukumar 等人已证明，对于来自 $\mathsf{P}_{\mathsf{gaus}}$ 的数据，线性最大间隔分类器也存在类似的良性过拟合结果 [Mut+21]，但采用了截然不同的证明方法。

5 聚类数据的良性过拟合

本节考虑这样一种分布：数据来自多个簇，每个簇中的数据最初共享同一标签，但随后以某个常数概率 η 被翻转。具体而言，我们

考虑定义在 $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ 上的分布 $\mathsf{P}_{\mathsf{clust}}$，其定义如下。令 $k \geq 2$ 和 $Q := [k]$。给定簇均值 $\mu^{(1)}$、...、$\mu^{(k)}$ 及其簇标签 $\tilde{y}^{(1)}$、...、$\tilde{y}^{(k)} \in \{\pm 1\}$。簇索引从$q \sim \mathsf{Unif}(Q)$中采样，随后 $x|q \sim \mu^{(q)} + z$，其中 $z \sim \mathsf{P}'_z$ 满足：z 的各个分量均为均值为零、相互独立的次高斯随机变量，且次高斯范数至多为 1；以及 $\mathbb{E}[\|z\|^2] = d$。[3]最后，x 的（干净）标签为 $\tilde{y} = y^{(q)}$，而观测标签以 $1-\eta$ 的概率为 $y = \tilde{y}$，以 $1-\eta$ 的概率为 $y = -\tilde{y}$。

对于给定的 $\delta \in (0, 1/2)$，当常数 $C > 1$ 足够大时，我们对参数作出如下假设：

(CL1) 样本数量 $n \geq Ck^2 \log(k/\delta)$。

(CL2) 维度 $d \geq C \max\{n \max_q \|\mu^{(q)}\|^2, n^2 \log(n/\delta)\}$。

(CL3) 簇均值满足：$\min_q \|\mu^{(q)}\| \geq Ck\sqrt{\log(2nk/\delta)}$。

(CL4) 簇均值满足以下意义下的近似正交性：$\min_q \|\mu^{(q)}\|^2 \geq Ck \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle|$。

下面我们将证明，在这些假设下，训练数据以很高的概率是线性可分的。

我们的第一个引理表明，在上述假设下，随着 C 增加，训练数据会变得更加正交，样本范数之比趋于 1。

引理 5.1. 存在一个绝对常数 $C_2 > 0$，使得对于每个足够大的常数 $C > 0$，对于任意$\delta \in (0, 1/7)$，在假设 *(CL1)* 至*(CL4)*（针对这些 C 和 δ定义）下，以至少$1 - 7\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{clust}}$），训练数据是 C/C_2-正交的，并且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$。

上述引理的证明以及本节所有证明均见附录D。与之前一样，引理5.1 使我们能够利用命题 3.3 和3.4，证明线性最大间隔问题(1) 以及 Leaky ReLU 网络最大间隔问题(2) 的 KKT 点都具有以下形式：$\sum_{i=1}^n s_i y_i x_i$。以下定理刻画了该预测器的性能。

定理 5.2。设 $\tau \geq 1$ 为一个常数，并假设对于某些绝对常数 η、$\Delta > 0$，$\eta \leq \frac{1}{1+\tau} - \Delta$ 成立。存在常数 C、$C' > 0$（仅取决于 η、τ 和 Δ），使得对于任意 $\delta \in (0, 1/14)$，在假设 (CL1) 至 (CL4)（针对这些 C 和 δ定义）下，以至少 $1 - 14\delta$ 的概率（关于 $\mathsf{P}^n_{\mathsf{clust}}$），如果 $u \in \mathbb{R}^d$ 是关于 $\{(x_i, y_i)\}_{i=1}^n$的 τ-均匀的，那么

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle u, x_k \rangle\big),$$
$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle u, x \rangle)\big) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，如果 $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，那么线性分类器 $x \mapsto \operatorname{sign}(\langle u, x \rangle)$表现出良性过拟合。

要实现良性过拟合，上述定理要求满足假设 (CL1)至 (CL4)，同时满足 $\min_q \|\mu^{(q)}\|^4 = \omega(k^2 d/n)$。以下是几种可以满足这些条件的情形：

[3]我们可以轻松地处理良态簇，例如对于某个绝对常数 $\kappa > 0$，$\kappa d \leq \mathbb{E}\|z\|^2 \leq d$ 成立，不过此时可容忍的噪声率将取决于 κ（较小的 κ 需要较小的 η）。为使论述简洁，我们不作此处理。

(*i*) 正交簇，其中每个 $q \in Q$ 都满足 $\|\mu^{(q)}\| = \Theta(d^{\beta})$，且 $\beta \in (1/4, 1/2)$、$k = O(1)$、$n = \tilde{\Omega}(1)$ 和 $d = \tilde{\Omega}(n^{1/(1-2\beta)})$。在这种情形下，测试误差至多为 $\eta + \exp\left(-\tilde{\Omega}(nd^{4\beta-1})\right)$。

(*ii*) 非正交簇，其中 $\max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle| = O(d^{\frac{3}{5}})$ 且 $\|\mu^{(q)}\| = \Theta(d^{\frac{1}{5}})$ 对每个 q，$n = \Theta$ $k = \Theta$ $(d^{\frac{1}{5}})$ 成立，并且 $(d^{0.05})$。在此设定下，测试误差至多为 $\eta + \exp\left(-\Omega(d^{0.43})\right)$。

尽管上述两种设定都未被定理 4.2 明确涵盖，设定 (i) 与该定理的思路相似，因为标签由数量固定的高方差方向决定。相比之下，设定 (ii) 则截然不同，因为它允许（干净的）标签由线性分类器在 $k = \Theta(d^{0.05})$ 个分量上的输出决定，即 符号 $\left(\langle \sum_{q=1}^{k} y^{(q)} \mu^{(q)}, \ x \rangle\right)$。

与定理 4.2的情形类似，定理 5.2有若干推论。第一个推论是命题 3.3的直接结果。

推论 5.3。 假设 $0 < \eta \leq 0.49$。存在常数 $C, C' > 0$ 使得对于任意 $\delta \in (0, 1/21)$，在假设 *(CL1)* 至 *(CL4)*（针对这些 C 和 δ定义）下，以至少 $1 - 21\delta$ 的概率，在$\mathsf{P}_{\mathsf{clust}}^{n}$上，最大间隔线性分类器 $w = \operatorname{argmin}\{\|w\|^2 : \ y_i \langle w, x_i \rangle \geq 1 \ \forall i\}$ 满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle w, x_k \rangle\right),$$

$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y) \sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}(\langle w, x \rangle)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，如果 $n\min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，那么 w 表现出良性过拟合。

类似地，我们可以证明，问题 (3) 中的 Leaky ReLU 网络的最大间隔问题的 KKT 点也表现出良性过拟合。

推论 5.4. 假设 $0 < \eta \leq \frac{49\gamma^2}{100}$。存在常数 C， $C' > 0$ 使得对于任意 $\delta \in (0, \ 1/21)$，在假设 *(CL1)* 至 *(CL4)*（针对这些 C 和 δ定义）下，以至少 $1 - 21\delta$ 相对于$\mathsf{P}_{\mathsf{clust}}^{n}$的概率，任意一个 W 问题 (3) 的 KKT 点均满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\left(f(x_k; W)\right),$$

$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y) \sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}\left(f(x; W)\right)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，若 $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，则神经网络 $f(x; W)$ 呈现良性过拟合。此外，对于任意初始化 $W(0)$，梯度流都会在方向上收敛到一个满足上述条件的网络。

我们想指出，定理 5.2 （以及后续推论）并未明确涵盖数据由两个相反簇组成的情况，即 $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$ 且 $x|\tilde{y} \sim \tilde{y}\mu + z$ 对于某个随机向量 z，并且标签以 $\tilde{y} \mapsto -\tilde{y}$ 的概率翻转 $\eta < 1/2$的情形。近期的一些研究表明，线性最大间隔分类器 [CL21; WT21] 以及采用平滑Leaky ReLU 激活函数的两层神经网络 [FCB22] 在这种分布设定下表现出良性过拟合。然而，只需对我们用于定理 (3) 5.2的证明稍作修改，就可以轻松扩展我们的分析，证明该分布下线性最大间隔分类器以及 Leaky ReLU 网络最大间隔问题的 KKT 点也会表现出良性过拟合。我们还希望强调，Frei、Chatterji 和 Bartlett 的结果

[FCB22] 仅适用于采用平滑泄漏 ReLU 激活函数的网络。这类网络并非齐次网络，且对于它们，梯度流是否具有促使其满足间隔最大化 KKT 条件的已知隐式偏置尚无定论。具体而言，他们的分析基于追踪神经网络在整个训练轨迹中的泛化误差，而我们的分析则依赖于齐次网络中间隔最大化 KKT 条件所施加的结构。我们的工作与他们的另一处不同在于标签噪声模型。我们推导出了所能容忍噪声水平的显式上界，而他们的结果适用于低于某个未指明常数的噪声水平。我们的分析适用于标签以固定概率翻转的情形，而他们的研究允许对抗性标签噪声。

6 证明思路

本节将介绍一些直觉，说明最大间隔线性分类器如何能够实现良性过拟合。我们考虑一个分布 $\mathrm{P}_{\mathrm{opp}}$，它由均值向量 $\mu \in \mathbb{R}^d$ 和标签噪声参数$\eta \in (0, 1/2)$定义，其中样本 $(x, y) \sim \mathrm{P}_{\mathrm{opp}}$ 按如下方式抽取：

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad z \sim \mathsf{N}(0, I_d), \quad x|\tilde{y} \sim \tilde{y}\mu + z, \quad \begin{cases} y = \tilde{y}, & \text{w.p. } 1-\eta, \\ y = -\tilde{y}, & \text{w.p. } \eta. \end{cases} \tag{4}$$

正如上一节所述，该分布并未被定理 5.2 明确涵盖，但其直觉和证明本质上相同。我们的出发点是命题 3.3，它表明当训练数据近乎正交时，最大间隔线性分类器在训练数据上是 τ-均匀的。为简单起见，我们考虑这种形式中最简单的估计量，即 1-均匀向量 $\widehat{\mu} = \sum_{i=1}^n y_i x_i$。我们将满足 $y_i = \tilde{y}_i$ 的训练样本 (x_i, y_i) 称为干净样本，并将对应的索引记为 $\mathcal{C} \subset [n]$；将满足 $y_i = -\tilde{y}_i$ 的样本称为噪声样本，其索引记为 $\mathcal{N} \subset [n]$（因此 $\mathcal{C} \cup \mathcal{N} = [n]$）。对于分布 (4) 和训练数据 $\{(x_i, y_i)\}_{i=1}^n$，估计量 $\widehat{\mu}$ 因而具有如下形式

$$\widehat{\mu} = \sum_{i=1}^n y_i x_i = \sum_{i \in \mathcal{C}} (\mu + y_i z_i) + \sum_{i \in \mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i \propto \mu + \frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i,$$

其中 $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$。现在，只要 n 足够大，且噪声率小于例如$1/4$，那么以高概率有 $9n/10 \geq |\mathcal{C}| - |\mathcal{N}| \geq n/10$。特别地，估计量 $\widehat{\mu}$ 与以下两个分量之和成正比：μ，它是能在该分布上达到最优准确率的线性分类器；以及 $(|\mathcal{C}| - |\mathcal{N}|)^{-1} \sum_{i=1}^n y_i z_i$，它只包含噪声。后一个分量对新测试样本的预测毫无用处，但对实现较小的训练误差却相当有用。事实上，对于训练样本 $(x_k, y_k) = (\tilde{y}_k \mu + z_k, y_k)$，

$$\begin{aligned} \left\langle y_k x_k, \sum_{i=1}^n y_i z_i \right\rangle &= \left\langle y_k \tilde{y}_k \mu + y_k z_k, y_k z_k + \sum_{i \neq k} y_i z_i \right\rangle \\ &= \|z_k\|^2 + y_k \tilde{y}_k \left\langle \mu, \sum_{i=1}^n y_i z_i \right\rangle + \left\langle y_k z_k, \sum_{i \neq k} y_i z_i \right\rangle. \end{aligned}$$

由于 $z_k \sim \mathsf{N}(0, I_d)$ 且 $\sum_{i=1}^n y_i z_i \sim \mathsf{N}(0, n I_d)$，标准集中界表明 $\|z_k\|^2 \gtrsim d$，而 $|\langle \mu, \sum_{i=1}^n y_i z_i \rangle| \lesssim \sqrt{n}\|\mu\|$ 和 $|\langle y_k z_k, \sum_{i \neq k} y_i z_i \rangle| \lesssim \sqrt{nd}$（为简洁起见，忽略对数因子）。因此，

若 $d \gg \sqrt{nd}$ 且 $d \gg \sqrt{n}\|\mu\|$，则估计量 $\xi := (|\mathcal{C}| - |\mathcal{N}|)^{-1} \sum_{i=1}^{n} y_i z_i$ 满足

$$\langle \xi, y_k x_k \rangle \gtrsim d/n. \tag{5}$$

另一方面，ξ 对一个独立测试样本（x, y）的影响为，

$$|\langle yx, \xi \rangle| = \left| \frac{1}{|\mathcal{C}| - |\mathcal{N}|} \left\langle y\tilde{y}(z+\mu), \sum_{i=1}^{n} y_i z_i \right\rangle \right| \lesssim \frac{1}{n} \left(\sqrt{n}\|\mu\| + \sqrt{dn} \right) = \frac{\|\mu\| + \sqrt{d}}{\sqrt{n}}. \tag{6}$$

将 (5) 和 (6) 结合起来，我们可以看到，只要 ξ 对训练数据的影响就会显著大于其对测试数据表现的影响 $d \gg \sqrt{n}\|\mu\| + \sqrt{dn}$。假设 $\|\mu\| < \sqrt{d}$ 当 $d \gg n^2$时成立。特别是，至少当信号 $\|\mu\|$不太大时，在高维情况下，更容易出现能够实现良性过拟合的分量。

以上示意表明，当 $\|\mu\| < \sqrt{d}$ 和 $d \gg n^2$时，过拟合分量 ξ 有助于插值（含噪）训练数据。然而，估计量 $\widehat{\mu} \propto \mu + \xi$ 还包含分量 μ，它会使含噪训练数据更容易被错误拟合。因此，为了证明 $\mu + \xi$ 表现出良性过拟合，我们需要证明：（i）来自 μ 的信号强度不至于强到阻止对含噪训练数据的过拟合；但（ii）来自 μ的信号强度足够大，能够实现基于测试数据的良好泛化。对于（i），标准集中界意味着

$$|\langle \mu, y_k x_k \rangle| = |y_k \tilde{y}_k \|\mu\|^2 + \langle y_k z_k, \mu \rangle| \lesssim \|\mu\|^2 + \|\mu\|. \tag{7}$$

根据 (5)，估计量 $\mu + \xi$ 仍会对训练数据进行插值，前提是 $d/n \gg \max(\|\mu\|, \|\mu\|^2)$。对于部分 (ii)，给定一个 干净测试样本 $(x, \tilde{y})$，

$$\langle \tilde{y}x, \mu \rangle = \langle \mu + \tilde{y}z, \mu \rangle \gtrsim \|\mu\|^2 - C\|\mu\|. \tag{8}$$

因此，若满足 $\|\mu\| \gg C$ 和 $\|\mu\|^2 \gg n^{-1/2}(\|\mu\| + \sqrt{d})$，我们便可确保 $\mu + \xi$ 也能通过结合 (8)和 (6)，正确分类干净测试样本。

总而言之，我们已经确定了在何种条件下可以保证估计量 $\widehat{\mu} \propto \mu + \xi$ 对于分布 Popp会发生良性过拟合。首先，训练数据的维度必须足够高，以确保过拟合分量 ξ 对训练数据有显著影响，但对未来测试数据的影响很小。其次，$\widehat{\mu}$ 的潜在信号（其强度由 $\|\mu\|$衡量）不能强到足以阻止模型对噪声标签过拟合，但也必须足够强，才能确保准确预测未来的测试数据。

7 讨论

我们刻画了若干新的情形，在这些情形下，线性分类器和双层神经网络能够表现出良性过拟合。我们展示了梯度流的隐式偏置如何为通过这种方法训练的线性分类器和神经网络施加重要的结构约束，以及如何利用这种结构来理解存在噪声标签时插值模型的泛化。

未来研究有许多方向。例如，使用逻辑损失训练的更广泛的齐次神经网络，同样具有趋向于满足间隔最大化的 KKT 条件的隐式偏置。能否利用这种隐式偏置，证明深度为 $L \geq 2$ 的 ReLU 激活神经网络存在良性过拟合？此外，尽管两层带泄漏 ReLU 网络通常是非线性的，但我们的证明适用于其决策边界为线性的情形。理解学习到的决策边界为非线性时神经网络中的良性过拟合，将会很有意义。

致谢

SF、GV、PB 和 NS 感谢美国国家科学基金会和西蒙斯基金会通过 DMS-2031883 和 #814639 资助的深度学习理论基础合作项目所提供的支持。

目录

1 引言 **12** 预备知识 **33** 近似正交数据的 KKT 点性质 **54** 次高斯边缘分布的良性过拟合 **65** 聚类数据的良性过拟合 **96** 证明思路 **127** 讨论 **13**A 命题证明 **3.3 14**B 命题证明 **3.4 16**C 次高斯边缘分布的证明 **22**C.1 初步集中不等式.22C.2 引理证明4.1 .26C.3 定理证明 4.2 .27C.4 推论证明 4.3，推论 4.4，以及推论 4.5 .32

D 聚类数据的证明 **34**

D.1 初步集中不等式.35D.2 引理的证明5.1 .40D.3 定理的证明5.2 .42D.4 推论的证明5.3和推论 5.4 .48

命题3.3的证明

由于 $\widehat{w}$ 满足最大间隔问题的 KKT 条件，因此我们有 $\widehat{w}=\sum_{i=1}^{n}\lambda_i y_i x_i$，其中对于所有 $i\in[n]$，都有 $\lambda_i\geq 0$，且当 $y_i\widehat{w}^\top x_i\neq 1$ 时，$\lambda_i=0$。我们记 $R_{\min}=\min_i\|x_i\|$、$R_{\max}=\max_i\|x_i\|$ 和 $R=R_{\max}/R_{\min}$。

在下面的引理中，我们为 λ_i 得到一个上界：

引理 A.1. 假设训练数据是 p-正交的。那么对于所有 $i\in[n]$，我们有 $\lambda_i\leq\frac{1}{R_{min}^2\left(1-\frac{1}{pR^2}\right)}$。

证明. 令 $j \in \operatorname{argmax}_{i\in[n]} \lambda_i$。我们有

$$y_j \widehat{w}^\top x_j = \sum_{i=1}^{n} \lambda_i y_j y_i x_i^\top x_j = \lambda_j \|x_j\|^2 + \sum_{i\neq j} \lambda_i y_j y_i x_i^\top x_j \geq \lambda_j R_{\min}^2 - n \left(\max_{i\in[n]} \lambda_i\right) \left(\max_{i\neq j} |\langle x_i, x_j\rangle|\right) . \tag{9}$$

由 p-正交性假设，我们还有

$$n \max_{i\neq j} |\langle x_i, x_j\rangle| \leq \frac{R_{\min}^2}{pR^2} . \tag{10}$$

假设

$$\lambda_j = \max_{i\in[n]} \lambda_i > \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} . \tag{11}$$

结合(9)、(10) 和(11)，我们得到

$$y_j \widehat{w}^\top x_j \geq \lambda_j R_{\min}^2 - \lambda_j \cdot \frac{R_{\min}^2}{pR^2} = \lambda_j R_{\min}^2 \left(1 - \frac{1}{pR^2}\right) > 1 .$$

根据 KKT 条件，若 $y_j \widehat{w}^\top x_j > 1$，则必有 $\lambda_j = 0$，从而导致矛盾。 □

接下来，我们给出 λ_i 的一个下界：

引理A.2。假设训练数据是 p-正交的。那么，对所有 $i \in [n]$，我们有 $\lambda_i \geq \frac{1}{R_{max}^2} \left(1 - \frac{1}{pR^2-1}\right)$。

证明。令 $j \in [n]$。根据 $\widehat{w}$ 的定义，我们有

$$\begin{aligned} 1 &\leq y_j \widehat{w}^\top x_j \\ &= \sum_{i=1}^{n} \lambda_i y_j y_i x_i^\top x_j \\ &= \lambda_j \|x_j\|^2 + \sum_{i\neq j} \lambda_i y_j y_i x_i^\top x_j \\ &\leq \lambda_j R_{\max}^2 + n \left(\max_{i\in[n]} \lambda_i\right) \left(\max_{i\neq j} |\langle x_i, x_j\rangle|\right) . \end{aligned} \tag{12}$$

根据 p-正交假设，我们有

$$n \max_{i\neq j} |\langle x_i, x_j\rangle| \leq \frac{R_{\min}^2}{pR^2} , \tag{13}$$

并且根据引理A.1 我们有

$$\max_{i\in[n]} \lambda_i \leq \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} . \tag{14}$$

结合(12)、(13) 和(14)，可得

$$1 \leq \lambda_j R_{\max}^2 + \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} \cdot \frac{R_{\min}^2}{pR^2} = \lambda_j R_{\max}^2 + \frac{1}{pR^2 - 1} .$$

因此，

$$\lambda_j \geq \left(1 - \frac{1}{pR^2 - 1}\right) \frac{1}{R_{\max}^2} .$$

□

结合引理A.1 和A.2，我们得出 $\widehat{w} = \sum_{i=1}^{n} \lambda_i y_i x_i$ 其中

$$\begin{aligned}
\frac{\max_i \lambda_i}{\min_i \lambda_i} &\leq \frac{1}{R_{\min}^2 \left(1 - \frac{1}{pR^2}\right)} \cdot R_{\max}^2 \left(1 - \frac{1}{pR^2 - 1}\right)^{-1} \\
&= \frac{pR^2}{R_{\min}^2 (pR^2 - 1)} \cdot R_{\max}^2 \cdot \frac{pR^2 - 1}{pR^2 - 2} \\
&= \frac{R_{\max}^2}{R_{\min}^2} \cdot \frac{pR^2}{pR^2 - 2} \\
&= R^2 \left(1 + \frac{2}{pR^2 - 2}\right) .
\end{aligned}$$

B 命题3.4的证明

我们先介绍一些记号。为方便起见，我们将对正神经元（即满足 $a_j = 1/\sqrt{m}$ 的神经元）和负神经元（即满足 $a_j = -1/\sqrt{m}$ 的神经元）使用不同的记号。具体而言，

$$f(x; W) = \sum_{j=1}^{m} a_j \phi(w_j^\top x) = \sum_{j=1}^{m/2} \frac{1}{\sqrt{m}} \phi(v_j^\top x) - \sum_{j=1}^{m/2} \frac{1}{\sqrt{m}} \phi(u_j^\top x) .$$

记 $\zeta =$最大值$_{i \neq j} |\langle x_i, \ x_j \rangle|$。因此，我们的近似正交假设可以写作 $n\zeta \leq \frac{R_{\min}^2}{pR^2}$。由于 W满足问题(3)的 KKT 条件，因此存在 λ_1，...，λ_n（称为KKT 乘子），使得对每个 $j \in [m/2]$ 都有

$$v_j = \sum_{i \in [n]} \lambda_i \nabla_{v_j} \left(y_i f(x_i; W)\right) = \frac{1}{\sqrt{m}} \sum_{i \in [n]} \lambda_i y_i \phi'_{i, v_j} x_i , \tag{15}$$

其中 $\phi'_{i,\ v_j}$ 是 ϕ 在 $v_j^\top x_i$处的次梯度，即，若 $v_j^\top x_i > 0$，则 $\phi'_{i,\ v_j} = 1$；若 $v_j^\top x_i < 0$，则 $\phi'_{i,\ v_j} = \gamma$；否则 $\phi'_{i,\ v_j}$是 $[\gamma, 1]$中的某个值。此外，对所有 i，我们有 $\lambda_i \geq 0$；并且当 $y_i f(x_i; W) \neq 1$时，$\lambda_i = 0$。同样，对所有 $j \in [m/2]$，我们有

$$u_j = \sum_{i \in [n]} \lambda_i \nabla_{u_j} \left(y_i f(x_i; W)\right) = \frac{1}{\sqrt{m}} \sum_{i \in [n]} \lambda_i (-y_i) \phi'_{i, u_j} x_i , \tag{16}$$

其中 ϕ'_{i,u_j} 的定义与 $\phi'_{i,v}$ 类似

$_j$.

我们的证明基于以下引理，该引理由 Frei 等人[Fre+23]推出（请注意，在他们的记号中，在我们的设定下有 $m_1 = m_2 = m/2$）：

引理 B.1（Frei 等人 [Fre+23]，定理 3.2 和推论 3.5）。记 $R_{min}^2 = \min_i \|x_i\|^2$ 和 $R_{max}^2 = \max_i \|x_i\|^2$。令 f 表示 Leaky ReLU 网络（2），并令 W 表示问题（3）的一个 KKT 点。令λ_1、...、$\lambda_n \geq 0$ 表示相应的 KKT 乘子。假设对于 $p \geq 3\gamma^{-3}$，训练数据是 p-正交的。那么，对于所有 $i \in [n]$，我们有 $\lambda_i \in \left(\frac{1}{2R_{max}^2}, \frac{3}{2\gamma^2 R_{min}^2}\right)$；并且对于任意 $x \in \mathbb{R}^d$，有 $\text{sign}\left(f(x; W)\right) = \text{sign}\left(\langle z, x\rangle\right)$，其中 $z = \frac{\sqrt{m}}{2}v - \frac{\sqrt{m}}{2}u$ for

$$v = \frac{1}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i \,,$$

and

$$u = \frac{1}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i \,.$$

此外，对于任意初始化 $W(0)$，逻辑损失或指数损失下的梯度流在方向上收敛到这样的 KKT 点。

注意，上述引理意味着 $\text{sign}\left(f(x; W)\right) = \text{sign}\left(\langle z, x\rangle\right)$ for

$$\begin{aligned} z &= \frac{\sqrt{m}}{2}\left(\frac{1}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i\right) - \frac{\sqrt{m}}{2}\left(\frac{1}{\sqrt{m}} \sum_{i:y_i=-1} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i:y_i=1} \lambda_i x_i\right) \\ &= \frac{1+\gamma}{2} \sum_{i:y_i=1} \lambda_i x_i - \frac{1+\gamma}{2} \sum_{i:y_i=-1} \lambda_i x_i \\ &= \frac{1+\gamma}{2} \sum_{i=1}^{n} y_i \lambda_i x_i \,. \end{aligned} \tag{17}$$

该引理还意味着 $\lambda_i \in \left(\frac{1}{2R_{\max}^2}, \frac{3}{2\gamma^2 R_{\min}^2}\right)$ 对所有 i均成立。然而，这些界对我们来说还不够精确。在接下来的引理中，我们将得到一些界，它们明确给出了 p 对于 p-正交数据的依赖关系。这些引理的证明与 Frei 等人的证明思路相似，但需要做一些修改。 [Fre+23]。

引理 B.2。 记 $R_{min}^2 = \text{最小值}_i \|x_i\|^2$， $R_{max}^2 = \text{最大值}_i \|x_i\|^2$，并记 $R^2 = R_{max}^2/R_{min}^2$。 令 f 表示 Leaky ReLU 网络 (2)，并令 W 表示问题 (3)的一个 KKT 点。假设训练数据为p-正交的，其中 $p \geq 3\gamma^{-3}$。采用式 (15) 和(16)中的记号，对于所有 $i \in [n]$，有

$$\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \leq \frac{m}{R_{min}^2\left(\gamma - \frac{1}{pR^2}\right)} \,,$$

and

$$\lambda_i \leq \frac{1}{R_{min}^2 \gamma\left(\gamma - \frac{1}{pR^2}\right)} \,.$$

证明。令 $\xi = \max_{q\in[n]}\left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j}\right)$，并假设 $\xi > \frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)}$。令$r = \text{argmax}_{q\in[n]}\left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j}\right)$。由于根据我们的假设，$p \geq 3\gamma^{-3} \geq \frac{3}{\gamma}$ 且 $R \geq 1$，因此 $\xi > 0$，从而 $\lambda_r > 0$。因此，根据 KKT 条件，我们必有 $y_r f(x_r; W) = 1$。

我们考虑两种情况：

情况 1：假设 $y_r = -1$。利用 (15) 和 (16)，可得

$$\begin{aligned}
\sqrt{m} f(x_r; W) &= \sum_{j\in[m/2]} \phi(v_j^\top x_r) - \sum_{j\in[m/2]} \phi(u_j^\top x_r) \\
&= \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) \\
&= \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r y_r \phi'_{r,v_j} x_r^\top x_r + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) \\
&\quad - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r (-y_r) \phi'_{r,u_j} x_r^\top x_r + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) \\
&\le \sum_{j\in[m/2]} \phi\left(-\frac{1}{\sqrt{m}} \lambda_r \phi'_{r,v_j} R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) \\
&\quad - \sum_{j\in[m/2]} \phi\left(\frac{1}{\sqrt{m}} \lambda_r \phi'_{r,u_j} R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) .
\end{aligned}$$

由于 ϕ 的导数以下界为 γ，我们知道对于所有 z_1、$z_2 \in \mathbb{R}$，$\phi(z_1) - \phi(z_2) \ge \gamma(z_1 - z_2)$。利用这一点以及 ξ 的定义，上式至多为

$$\begin{aligned}
&\sum_{j\in[m/2]} \left[\phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right) - \frac{1}{\sqrt{m}} \gamma \lambda_r \phi'_{r,v_j} R_{\min}^2\right] \\
&\quad - \sum_{j\in[m/2]} \left[\phi\left(\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right) + \frac{1}{\sqrt{m}} \gamma \lambda_r \phi'_{r,u_j} R_{\min}^2\right] \\
&\le -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \sum_{j\in[m/2]} \left|\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right| + \sum_{j\in[m/2]} \left|\frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{r\}} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right| \\
&\le -\frac{1}{\sqrt{m}} \gamma \xi R_{\min}^2 + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \left|\lambda_q y_q \phi'_{q,v_j} x_q^\top x_r\right| + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \sum_{q\in[n]\setminus\{r\}} \left|\lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_r\right| .
\end{aligned}$$

在 $q \neq r$ 中使用 $|x_q^\top x_r| \leq \zeta$，上式至多为

$$
\begin{aligned}
&-\frac{1}{\sqrt{m}}\gamma\xi R_{\min}^2 + \frac{1}{\sqrt{m}}\sum_{j\in[m/2]}\sum_{q\in[n]\setminus\{r\}}\lambda_q\phi'_{q,v_j}\zeta + \frac{1}{\sqrt{m}}\sum_{j\in[m/2]}\sum_{q\in[n]\setminus\{r\}}\lambda_q\phi'_{q,u_j}\zeta \\
&= -\frac{1}{\sqrt{m}}\gamma\xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}}\sum_{q\in[n]\setminus\{r\}}\left(\sum_{j\in[m/2]}\lambda_q\phi'_{q,v_j} + \sum_{j\in[m/2]}\lambda_q\phi'_{q,u_j}\right) \\
&\leq -\frac{1}{\sqrt{m}}\gamma\xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}}\cdot n\cdot\max_{q\in[n]}\left(\sum_{j\in[m/2]}\lambda_q\phi'_{q,v_j} + \sum_{j\in[m/2]}\lambda_q\phi'_{q,u_j}\right) \\
&= -\frac{1}{\sqrt{m}}\gamma\xi R_{\min}^2 + \frac{\zeta}{\sqrt{m}}n\xi \\
&= -\frac{\xi}{\sqrt{m}}(\gamma R_{\min}^2 - n\zeta)\,.
\end{aligned}
$$

根据我们的 p-正交假设，上述表达式至多为

$$
-\frac{\xi}{\sqrt{m}}\left(\gamma R_{\min}^2 - \frac{R_{\min}^2}{pR^2}\right) = -\frac{\xi R_{\min}^2}{\sqrt{m}}\left(\gamma - \frac{1}{pR^2}\right) < -\frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)}\cdot\frac{R_{\min}^2}{\sqrt{m}}\left(\gamma - \frac{1}{pR^2}\right) = -\sqrt{m}\,,
$$

其中，在该不等式中，我们使用了关于 ξ 的假设、假设 $p \geq 3\gamma^{-3} \geq \frac{3}{\gamma}$ 以及 $R \geq 1$。因此，我们得到 $f(x_r; W) < -1$，这与 $y_r f(x_r; W) = 1$ 矛盾。

情况 2：假设 $y_r = 1$。类似于情况 1 中的计算（为简洁起见，此处不再重复）可得 $f(x_r; W) > 1$，这与 $y_r f(x_r; W) = 1$ 矛盾。由此完成 $\xi \leq \frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)}$ 的证明

最后，由于 $\xi \leq \frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)}$，且 ϕ 的导数以下界 γ 为界，因此对于所有 $i \in [n]$，我们有 have

$$
\frac{m}{R_{\min}^2\left(\gamma - \frac{1}{pR^2}\right)} \geq \sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j} + \sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j} \geq m\lambda_i\gamma\,,
$$

因此 $\lambda_i \leq \frac{1}{R_{\min}^2\gamma\left(\gamma - \frac{1}{pR^2}\right)}$. □

引理 B.3. 记 $R_{min}^2 =$ 最小值$_i \|x_i\|^2$，$R_{max}^2 =$ 最大值$_i \|x_i\|^2$，并且 $R^2 = R_{max}^2/R_{min}^2$。令 f 表示 Leaky ReLU 网络 (2)，并令 W 表示问题 (3)的一个 KKT 点。假设训练数据为p-正交的（对 $p \geq 3\gamma^{-3}$）。使用公式 (15) 和 (16)中的记号，对于所有 $i \in [n]$，有

$$
\sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j} + \sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j} \geq \frac{m(\gamma pR^2 - 2)}{R_{max}^2(\gamma pR^2 - 1)}\,,
$$

and

$$
\lambda_i \geq \frac{\gamma pR^2 - 2}{R_{max}^2(\gamma pR^2 - 1)} = \frac{1}{R_{max}^2}\left(1 - \frac{1}{\gamma pR^2 - 1}\right)\,.
$$

证明。 假设存在 $i \in [n]$ 使得 $\sum_{j\in[m/2]} \lambda_i \phi'_{i,\ v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,\ u_j} < {}^{m}_{(\gamma p R^2 - 2)\ \underline{R^2_{\max}(\gamma p R^2 - 1)}}$。 根据 (15) 和 (16)，可得

$$
\begin{aligned}
\sqrt{m} \le \left|\sqrt{m} f(x_i; W)\right| &= \left| \sum_{j\in[m/2]} \phi(v_j^\top x_i) - \sum_{j\in[m/2]} \phi(u_j^\top x_i) \right| \le \sum_{j\in[m/2]} \left| v_j^\top x_i \right| + \sum_{j\in[m/2]} \left| u_j^\top x_i \right| \\
&= \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q y_q \phi'_{q,v_j} x_q^\top x_i \right| + \sum_{j\in[m/2]} \left| \frac{1}{\sqrt{m}} \sum_{q\in[n]} \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_i \right| \\
&\le \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\left| \lambda_i y_i \phi'_{i,v_j} x_i^\top x_i \right| + \sum_{q\in[n]\setminus\{i\}} \left| \lambda_q y_q \phi'_{q,v_j} x_q^\top x_i \right| \right) \\
&\quad + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\left| \lambda_i (-y_i) \phi'_{i,u_j} x_i^\top x_i \right| + \sum_{q\in[n]\setminus\{i\}} \left| \lambda_q (-y_q) \phi'_{q,u_j} x_q^\top x_i \right| \right) .
\end{aligned}
$$

对 $|x_q^\top x_i| \le \zeta$、 $q \ne i$ 和 $x_i^\top x_i \le R^2_{\max}$ 使用上述结果，上式至多为

$$
\begin{aligned}
&\frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\lambda_i \phi'_{i,v_j} R^2_{\max} + \sum_{q\in[n]\setminus\{i\}} \lambda_q \phi'_{q,v_j} \zeta \right) + \frac{1}{\sqrt{m}} \sum_{j\in[m/2]} \left(\lambda_i \phi'_{i,u_j} R^2_{\max} + \sum_{q\in[n]\setminus\{i\}} \lambda_q \phi'_{q,u_j} \zeta \right) \\
&= \frac{1}{\sqrt{m}} \left(\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} R^2_{\max} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} R^2_{\max} \right) + \\
&\quad \frac{1}{\sqrt{m}} \sum_{q\in[n]\setminus\{i\}} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} \zeta + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \zeta \right) \\
&= \frac{R^2_{\max}}{\sqrt{m}} \left(\sum_{j\in[m/2]} \lambda_i \phi'_{i,v_j} + \sum_{j\in[m/2]} \lambda_i \phi'_{i,u_j} \right) + \frac{\zeta}{\sqrt{m}} \sum_{q\in[n]\setminus\{i\}} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) \\
&< \frac{R^2_{\max}}{\sqrt{m}} \cdot \frac{m(\gamma p R^2 - 2)}{R^2_{\max}(\gamma p R^2 - 1)} + \frac{\zeta}{\sqrt{m}} \cdot n \cdot \max_{q\in[n]} \left(\sum_{j\in[m/2]} \lambda_q \phi'_{q,v_j} + \sum_{j\in[m/2]} \lambda_q \phi'_{q,u_j} \right) ,
\end{aligned}
$$

其中在最后的　　等式中，我们使用了关于 i的假设。结合上述结果与我们的 p-正交

假设 $n\zeta \le \frac{R_{\min}^2}{pR^2}$，可得

$$\begin{aligned}\max_{q\in[n]}\left(\sum_{j\in[m/2]}\lambda_q\phi'_{q,v_j}+\sum_{j\in[m/2]}\lambda_q\phi'_{q,u_j}\right) &> m\left(1-\frac{\gamma pR^2-2}{\gamma pR^2-1}\right)\cdot\frac{1}{n\zeta}\\ &\ge m\left(1-\frac{\gamma pR^2-2}{\gamma pR^2-1}\right)\cdot\frac{pR^2}{R_{\min}^2}\\ &= m\left(\frac{1}{\gamma pR^2-1}\right)\cdot\frac{pR^2}{R_{\min}^2}\\ &= \frac{m}{R_{\min}^2}\left(\frac{1}{\gamma-\frac{1}{pR^2}}\right),\end{aligned}$$

与引理 B.2矛盾。因此完成了对 $\sum_{j\in[m/2]}\lambda_i\phi'_{i,\ v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,\ u_j}\ge\frac{m}{(\gamma pR^2-2)\,R_{\max}^2(\gamma pR^2-1)}$ 的证明。

最后，由于 $\sum_{j\in[m/2]}\lambda_i\phi'_{i,\ v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,\ u_j}\ge\frac{m}{(\gamma pR^2-2)\,R_{\max}^2(\gamma pR^2-1)}$，且 ϕ 的导数上界为 1，因此对所有 $i\in[n]$，我们有

$$\frac{m(\gamma pR^2-2)}{R_{\max}^2(\gamma pR^2-1)}\le\sum_{j\in[m/2]}\lambda_i\phi'_{i,v_j}+\sum_{j\in[m/2]}\lambda_i\phi'_{i,u_j}\le m\lambda_i\,,$$

因此

$$\lambda_i\ge\frac{\gamma pR^2-2}{R_{\max}^2(\gamma pR^2-1)}=\frac{1}{R_{\max}^2}\left(1-\frac{1}{\gamma pR^2-1}\right).$$

□

结合公式 (17) 与引理 B.2和B.3，并令对所有 $i\in[n]$，$s_i=\frac{(1+\gamma)}{2}^{\lambda_i}$，我们得到 sign $(f(x;W))$ sign $(\langle z,x\rangle)$ 对于 $z=\sum_{i=1}^n s_iy_ix_i$，其中对所有 $i\in[n]$，我们有

$$s_i\in\left[\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\max}^2}\left(1-\frac{1}{\gamma pR^2-1}\right),\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\min}^2\gamma^2\left(1-\frac{1}{\gamma pR^2}\right)}\right].$$

因此，z 是 τ-一致的，且

$$\begin{aligned}\tau&=\frac{(1+\gamma)}{2}\cdot\frac{1}{R_{\min}^2\gamma^2\left(1-\frac{1}{\gamma pR^2}\right)}\cdot\frac{2}{(1+\gamma)}\cdot R_{\max}^2\left(1-\frac{1}{\gamma pR^2-1}\right)^{-1}\\ &=\frac{\gamma pR^2}{R_{\min}^2\gamma^2(\gamma pR^2-1)}\cdot R_{\max}^2\frac{\gamma pR^2-1}{\gamma pR^2-2}\\ &=\frac{R_{\max}^2}{R_{\min}^2\gamma^2}\cdot\frac{\gamma pR^2}{\gamma pR^2-2}\\ &=\frac{R^2}{\gamma^2}\left(1+\frac{2}{\gamma pR^2-2}\right).\end{aligned}$$

C 次高斯边缘分布的证明

本节将证明引理4.1 和定理 4.2 以及推论 4.3、推论 4.4和推论 4.5。我们的证明策略大致如下。

1. 首先，我们在引理 C.1 中证明，为了使线性分类器 $x \mapsto \operatorname{sign}(\langle w, x\rangle)$ 达到接近噪声率的测试误差，只需 $\|[\Sigma^{1/2}w]_{2:d}\|/\sqrt{\lambda_1}[w]_1$ 足够小。

2. 接下来，我们在引理 C.2 中证明，训练数据 $\{(x_i, y_i)\}_{i=1}^n$ 在假设（SG1）至（SG3）下以高概率满足若干性质。引理 4.1 将作为该引理的确定性推论成立，因此当 p 较大时，训练数据是 p-正交的（回顾定义 3.1），且所有样本的范数彼此接近。这使我们能够应用命题 3.3 和命题 3.4；这两个命题表明，线性最大间隔问题（1）和 Leaky ReLU 最大间隔问题（3）的 KKT 点相对于 $\{(x_i, y_i)\}_{i=1}^n$ 是 τ-一致的。

3. 根据第一步，要证明定理 4.2，只需证明某个 τ-一致 $w \in \mathbb{R}^d$ 满足以下条件：

   (i) 范数 $\|[\Sigma^{1/2}w]_{2:d}\|$较小，且 (ii) 第一分量 $[w]_1$ 较大且为正。

   回顾一下， τ-一致 w具有如下形式：$\sum_{i=1}^n s_i y_i x_i$其中 最大值$_{i,\ j}\, s_i/s_j \le \tau$。对于 (i)，引理 C.2给出了训练样本 $\|[\Sigma^{1/2}x_i]_{2:d}\|$的界，这是本部分的基本构件。对于$(ii)$，注意，对于干净样本 $i \in \mathcal{C} \subset [n]$ （其中 $y_i = \tilde{y}_i$），$[s_i y_i x_i]_1 = s_i|[x_i]_1|$；而对于噪声样本 $i \in \mathcal{N} \subset [n]$ （其中 $y_i = -\tilde{y}_i$），$[s_i y_i x_i]_1 = -s_i|[x_i]_1|$。因此，只需刻画以下内容：

$$\begin{aligned}\left[\sum_{i=1}^n s_i y_i x_i\right]_1 &= \sum_{i\in\mathcal{C}} s_i|[x_i]_1| - \sum_{i\in\mathcal{N}} s_i|[x_i]_1| \\ &= \sum_{i=1}^n s_i|[x_i]_1| - 2\sum_{i\in\mathcal{N}} s_i|[x_i]_1| \\ &\ge \min_i s_i \sum_{i=1}^n |[x_i]_1| - 2\max_i s_i \sum_{i\in\mathcal{N}} |[x_i]_1|.\end{aligned}$$

   引理C.4 直接界定上述各项。定理 4.2 的证明随后基于(i) 和 (ii) 进行直接计算即可。

4. 推论 4.3 和 4.4 可通过基于引理 4.1 和定理 4.2的直接计算得到。推论 4.5 可通过直接计算验证 $\mathsf{P}_{\mathsf{gaus}}$ 满足所需性质得到。

C.1 初步集中不等式

我们首先证明，对于满足 $[w]_1 > 0$ 的任意线性分类器 $w \in \mathbb{R}^d$，只要$\|[\Sigma^{1/2}w]_{2:d}\|/\sqrt{\lambda_1}[w]_1$ 足够小，其测试误差就接近噪声率。

引理 C.1. 存在一个绝对常数 $c_1 \ge 2$，使得当 $w \in \mathbb{R}^d$ 满足 $[w]_1 > 0$ 时，以下结论成立。若 $[\Sigma^{1/2}w]_{2:d} = 0$，则 $\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y \ne \text{符号}(\langle w, x\rangle)) \le \eta$。否则，

$$\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y \ne \operatorname{sign}(\langle w, x\rangle)) \le \eta + \frac{c_1\|[\Sigma^{1/2}w]_{2:d}\|}{\sqrt{\lambda_1}[w]_1} \cdot \left(1 + \sqrt{0 \vee \log\left(\frac{\sqrt{\lambda_1}[w]_1}{\|[\Sigma^{1/2}w]_{2:d}\|}\right)}\right).$$

证明. 根据 P_{sg} 的定义，我们有，

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\big(y\neq\operatorname{sign}(\langle w,x\rangle)\big) &= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0)\\ &= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=-\operatorname{sign}([x]_1))\\ &\quad+\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\operatorname{sign}([x]_1))\\ &\le \eta+\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\operatorname{sign}([x]_1)).\end{aligned}\tag{18}$$

将向量 $u\in\mathbb{R}^d$ 的最后 $d-1$ 个分量记为 $[u]_{2:d}\in\mathbb{R}^{d-1}$，则有

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\operatorname{sign}([x]_1)) &= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(|[x]_1|[w]_1<-\operatorname{sign}([x]_1)\langle[w]_{2:d},[x]_{2:d}\rangle)\\ &\overset{(i)}{=} \mathbb{P}(\sqrt{\lambda_1}[w]_1|[z]_1|<-\operatorname{sign}([z]_1)\langle[w]_{2:d},[\Sigma^{1/2}z]_{2:d}\rangle)\\ &\overset{(ii)}{=} \mathbb{P}\left(|[z]_1|<-\frac{\operatorname{sign}([z]_1)\langle[\Sigma^{1/2}w]_{2:d},[z]_{2:d}\rangle}{\sqrt{\lambda_1}[w]_1}\right).\end{aligned}\tag{19}$$

等式（i）利用了 $x=\Sigma^{1/2}z_\circ$ 等式（ii）利用了 $[w]_1>0$ 这一假设。由此可见，如果 $[\Sigma^{1/2}w]_{2:d}=0$，那么我们有 $\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\text{符号}([x]_1))=\mathbb{P}(|[z]_1|<0)=0$，根据（18）可知 $\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\big(y\neq\text{符号}(\langle w,x\rangle)\big)\le\eta_\circ$ 因此，在证明的其余部分中，我们将假设 $[\Sigma^{1/2}w]_{2:d}\neq 0_\circ$

我们定义以下项

$$\rho:=\frac{\operatorname{sign}([z]_1)\langle[\Sigma^{1/2}w]_{2:d},[z]_{2:d}\rangle}{\sqrt{\lambda_1}[w]_1}.$$

当 $\sqrt{\lambda_1}[w]_1\gg\|[\Sigma^{1/2}w]_{2:d}\|$ 时，该项的绝对值很小。特别地，由于 z 是一个次高斯随机向量，且其次高斯范数至多为 σ_z，根据霍夫丁不等式，对于某个 $c>0$ 以及任意 $t\ge0$，我们有

$$\mathbb{P}(|\rho|\ge t)=\mathbb{P}\left(\frac{|\langle[z]_{2:d},[\Sigma^{1/2}w]_{2:d}\rangle|}{\sqrt{\lambda_1}[w]_1}\ge t\right)\le 2\exp\left(-\frac{c\lambda_1[w]_1^2t^2}{\sigma_z^2\|[\Sigma^{1/2}w]_{2:d}\|^2}\right).\tag{20}$$

由 (19)继续推得，对于任意 $t\ge0$,

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\operatorname{sign}([x]_1)) &= \mathbb{P}(|[z]_1|<-\rho)\\ &= \mathbb{P}(|[z]_1|<-\rho,\ |\rho|\ge t)+\mathbb{P}(|[z]_1|<-\rho,|\rho|<t)\\ &\le \mathbb{P}(|\rho|\ge t)+\mathbb{P}(|[z]_1|<t)\\ &\overset{(i)}{\le} 2\exp\left(-\frac{c\lambda_1[w]_1^2t^2}{\sigma_z^2\|[\Sigma^{1/2}w]_{2:d}\|^2}\right)+\beta t.\end{aligned}$$

在不等式 (i) 中，我们使用了 (20)，以及假设 $\mathbb{P}(|[z]_1|\le t)\le\beta t$ 对任意 $t\ge0$均成立。特别地，如果令

$$\xi:=\frac{\|[\Sigma^{1/2}w]_{2:d}\|}{\sqrt{\lambda_1}[w]_1},\quad t:=c^{-1/2}\sigma_z\xi\sqrt{0\vee\log(1/\xi)},$$

于是我们有，

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}(y\langle w,x\rangle<0,\ y=\operatorname{sign}([x]_1)) &\le 2\exp\left(-\frac{ct^2}{\sigma_z^2\xi^2}\right)+\beta t\\ &= 2(1\wedge\xi)+\xi\beta c^{-1/2}\sigma_z\sqrt{0\vee\log(1/\xi)})\\ &\le \xi\cdot\max(2,\beta c^{-1/2}\sigma_z)(1+\sqrt{0\vee\log(1/\xi)}).\end{aligned}$$

证明方法是令 $c_1 = \max(2, \beta c^{-1/2}\sigma_z)$。 □

T以下引理刻画了训练数据的若干有用性质 .

引理 C.2. 存在一个绝对常数 $C_0 > 1$，使得对于所有足够大的 $C > 1$ (其中 C，C_0仅依赖于 σ_z) 以及任意 $\delta \in (0, 1/2)$，在假设 *(SG1)* 至 *(SG3)* (针对这些C和 δ定义）下，以至少 $1 - 2\delta$ 的概率，对 $\mathsf{P}_{\mathsf{sg}}^n$，以下结论成立：

1. 样本的范数满足：

$$\left| \frac{\|x_i\|}{\sqrt{\mathrm{tr}(\Sigma)}} - 1 \right| \le C_0 \sqrt{\frac{\|\Sigma\|_2 \log(6n/\delta)}{\mathrm{tr}(\Sigma)}}, \quad \textit{for all } i \in [n],$$

and

$$\left| \frac{\|[\Sigma^{1/2} x_i]_{2:d}\|}{\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}} - 1 \right| \le C_0 \sqrt{\frac{\|\Sigma_{2:d}^2\|_2 \log(6n/\delta)}{\mathrm{tr}(\Sigma_{2:d}^2)}}.$$

2. 不同样本之间的相关性满足：

$$|\langle x_i, x_j \rangle| \le C_0 \sqrt{\mathrm{tr}(\Sigma^2)} \log(6n^2/\delta), \quad \textit{for all } i \ne j.$$

3. 样本满足：

$$\min_i \|x_i\|^2 \ge \frac{\mathrm{tr}(\Sigma)}{C_0 \sqrt{\mathrm{tr}(\Sigma^2)} \log(6n^2/\delta)} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \max_{i \ne j} |\langle x_i, x_j \rangle|.$$

证明。我们分部分证明该引理。

第 1 部分：样本的范数。我们首先证明范数的集中性。令 $x \in \{x_1, \ldots, x_n\}$。回忆一下$x = \Sigma^{1/2} z$ 其中，z 的各分量相互独立、均值为零，且 z 的次高斯范数至多为σ_z，其中 $\mathbb{E}[zz^\top] = I_d$。因此，我们可以应用 Hanson-Wright 不等式[Ver18，定理 6.3。2], 于是存在一个绝对常数 $c > 0$，使得对任意 $t \ge 0$,

$$\mathbb{P}\left(\left| \|x\| - \sqrt{\mathrm{tr}(\Sigma)} \right| > t \right) \le 2 \exp\left(-\frac{ct^2}{\sigma_z^4 \|\Sigma\|_2} \right),$$

这里我们用到了 $\|\Sigma^{1/2}\|_F^2 = \sum_{i=1}^d \lambda_i = \mathrm{tr}(\Sigma)$。取 $t = c^{-1/2}\sigma_z^2 \sqrt{\|\Sigma\|_2 \log(6n/\delta)}$，并对 $x \in \{x_1, \ldots, x_n\}$ 使用并集界，可得，

$$\mathbb{P}\left(\exists i \in [n] : \left| \|x_i\| - \sqrt{\mathrm{tr}(\Sigma)} \right| > c^{-1/2}\sigma_z^2 \sqrt{\|\Sigma\|_2 \log(6n/\delta)} \right) < \delta/3. \tag{21}$$

现在我们给出 $\|\Sigma^{1/2} x_i\|^2 = z_i^\top \Sigma^2 z_i$ 的一个上界。再次固定 $x \in \{x_1, \ldots, x_n\}$。我们可以采用与上述几乎相同的论证：由于 $\|\Sigma\|_F^2 = \mathrm{tr}(\Sigma^2)$，根据 Hanson-Wright 不等式，对于任意 $t \ge 0$,

$$\mathbb{P}\left(\left| \|\Sigma^{1/2} x\| - \sqrt{\mathrm{tr}(\Sigma^2)} \right| > t \right) \le 2 \exp\left(-\frac{ct^2}{\sigma_z^4 \|\Sigma^2\|_2} \right), \tag{22}$$

取 $t = c^{-1/2}\sigma_z^2\sqrt{\|\Sigma^2\|_2 \log(6n/\delta)}$，并注意到 $\mathrm{tr}(\Sigma^2) \geq \|\Sigma^2\|_2$，可得

$$\sqrt{\mathrm{tr}(\Sigma^2)} + c^{-1/2}\sigma_z^2\sqrt{\|\Sigma^2\|_2 \log(6n/\delta)} \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\mathrm{tr}(\Sigma^2)\log(6n/\delta)},$$

由并集界可得

$$\mathbb{P}\left(\exists i \in [n] : \|\Sigma^{1/2}x_i\| > (1 + c^{-1/2}\sigma_z^2)\sqrt{\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right) \leq \delta/3. \tag{23}$$

使用推导(21)时完全相同的论证，我们也有

$$\mathbb{P}\left(\exists i \in [n] : \left|\|[\Sigma^{1/2}x_i]_{2:d}\| - \sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}\right| > c^{-1/2}\sigma_z^2\sqrt{\|\Sigma_{2:d}^2\|\log(6n/\delta)}\right) \leq \delta/3. \tag{24}$$

第 2 部分：样本的相关性。现在我们对不同样本之间的相关性进行界定。固定$j \in [n]$ 并考虑 $i \in [n] \setminus \{j\}$。于是存在一个次高斯随机向量 z_i，其次高斯范数至多为 σ_z，使得 $\langle x_i, x_j\rangle = z_i^\top \Sigma^{1/2}x_j$。特别地，$\langle x_i, x_j\rangle = z_i^\top \xi \cdot \|\Sigma^{1/2}x_j\|$，其中 ξ是单位范数向量。由于 z_i 是次高斯随机向量，这意味着对于某个 $c > 0$以及任意 $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(|\langle x_i, x_j\rangle| > t \middle| \|\Sigma^{1/2}x_j\| \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\quad\leq 2\exp\left(-c \cdot \frac{t^2}{\sigma_z^2(1 + c^{-1/2}\sigma_z^2)^2\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right).
\end{aligned} \tag{25}$$

令$c' = c/[\sigma_z^2\ (1 + c^{-1/2}\sigma_z^2)^{\ 2}]$,，因此我们可以得到如下界：

$$\begin{aligned}
&\mathbb{P}\left(\exists i \neq j :\ |\langle x_i, x_j\rangle| > t\right) \\
&\quad\leq \mathbb{P}\left(\exists i \neq j : |\langle x_i, x_j\rangle| > t \middle| \|\Sigma^{1/2}x_j\| \leq (1 + c^{-1/2}\sigma_z^2)\sqrt{\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\qquad + \mathbb{P}\left(\exists j : \|\Sigma^{1/2}x_j\| > (1 + c^{-1/2}\sigma_z^2)\sqrt{\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right) \\
&\quad\overset{(i)}{\leq} 2n^2\exp\left(-\frac{c't^2}{\mathrm{tr}(\Sigma^2)\log(6n/\delta)}\right) + \frac{\delta}{3},
\end{aligned} \tag{26}$$

其中 (i) 使用 (23)。取 $t = (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)\log(6n^2/\delta)}$ 并利用 (25)，可得

$$\mathbb{P}\left(\exists i \neq j :\ |\langle x_i, x_j\rangle| > (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta)\right) \leq \frac{2\delta}{3}. \tag{27}$$

将上述表达式与 (21) 和 (24) 结合起来，并使用并集界，可得其成立的概率至少为 $1 - 2\delta$,

$$\begin{cases}
\left|\frac{\|x_i\|}{\sqrt{\mathrm{tr}(\Sigma)}} - 1\right| \leq c^{-1}\sigma_z^2\sqrt{\frac{\|\Sigma\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma)}}, & \text{for all } i \in [n], \\
\left|\frac{\|[\Sigma^{1/2}x_i]_{2:d}\|}{\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}} - 1\right| \leq c^{-1}\sigma_z^2\sqrt{\frac{\|\Sigma_{2:d}^2\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma_{2:d}^2)}}, & \text{for all } i \in [n], \\
|\langle x_i, x_j\rangle| \leq (c')^{-1/2}\sqrt{\mathrm{tr}(\Sigma^2)}\log(6n^2/\delta), & \text{for all } i \neq j.
\end{cases} \tag{28}$$

引理的前两部分至此完成。

第 3 部分：样本的近正交性。现在我们给出 $R = \max_{i,j} \|x_i\|/\|x_j\|$ 的上界。注意，以至少 $1 - 2\delta$ 的概率，(28) 成立，我们将证明这意味着存在一个绝对常数 $C_0 > 0$ 使得：

$$\frac{\min_i \|x_i\|^2}{R^2 \max_{i\neq j} |\langle x_i, x_j\rangle|} \geq \frac{\operatorname{tr}(\Sigma)}{C_0 \sigma_z \sqrt{\operatorname{tr}(\Sigma^2)\log(6n^2/\delta)}}.$$

根据假设(SG3)，我们有

$$\xi := \frac{\|\Sigma\|_2 \log(6n/\delta)}{\operatorname{tr}(\Sigma)} \leq \frac{\|\Sigma\|_F \log(6n/\delta)}{\operatorname{tr}(\Sigma)} = \frac{\sqrt{\operatorname{tr}(\Sigma^2)}\log(6n/\delta)}{\operatorname{tr}(\Sigma)} \leq \frac{1}{C}. \tag{29}$$

因此，由(28)可知，量 $R = \max_{i,\ j} \|x_i\|/\|x_j\|$ 满足

$$\begin{aligned} R &= \max_{i,j} \frac{\|x_i\|}{\|x_j\|} \\ &\leq \frac{1 + c^{-1}\sigma_z^2\sqrt{\xi}}{1 - c^{-1}\sigma_z^2\sqrt{\xi}} \\ &\overset{(i)}{\leq} \frac{1 + c^{-1}\sigma_z^2/\sqrt{C}}{1 - c^{-1}\sigma_z^2/\sqrt{C}} \\ &\overset{(ii)}{\leq} \left(1 + \frac{2c^{-1}\sigma_z^2}{\sqrt{C}}\right)^2. \end{aligned} \tag{30}$$

不等式(i) 由(29)得出，而(ii) 用到了不等式 $1/(1-x) \leq 1 + 2x$，该不等式适用于 $[0, 1/2]$，且当$C > 1$ 足够大时成立。特别地，通过取更大的 C，我们可以确保 R 更接近 1。

接下来，由该引理的第 1 部分和第 2 部分可得，

$$\frac{\min_i \|x_i\|^2}{\max_{i\neq j} |\langle x_i, x_j\rangle|} \geq \frac{\operatorname{tr}(\Sigma)\left(1 - c^{-1}\sigma_z^2\sqrt{\xi}\right)}{(c')^{-1/2}\sqrt{\operatorname{tr}(\Sigma^2)\log(6n^2/\delta)}} \geq \frac{\operatorname{tr}(\Sigma)\left(1 - c^{-1}\sigma_z^2/\sqrt{C}\right)}{(c')^{-1/2}\sqrt{\operatorname{tr}(\Sigma^2)\log(6n^2/\delta)}}.$$

对于 $C > 100c^{-2}\sigma_z^4$，由(30)可得 $1 - c^{-1}\sigma_z^2/\sqrt{C} \geq 0.9$ 以及 $R^{-2} \geq 0.68$。因此可知，当 C足够大时，

$$\frac{\min_i \|x_i\|^2}{R^2 \max_{i\neq j} |\langle x_i, x_j\rangle|} \geq \frac{\operatorname{tr}(\Sigma)}{2(c')^{-1/2}\sqrt{\operatorname{tr}(\Sigma^2)\log(6n^2/\delta)}}.$$

□

C.2 引理4.1的证明

接下来我们证明引理4.1：随着 C 增大，在假设 (SG1) 至 (SG3)下，当 p 足够大且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \to 1$ 时，训练数据变得p-正交。

引理 4.1. 存在一个绝对常数 $C_1 > 0$（仅取决于 σ_z），使得对于任意足够大的常数 $C > 0$，对于任意 $\delta \in (0, 1/2)$，在假设 *(SG1)* 至 *(SG3)*（针对这些 C和 δ 定义）的条件下，以至少 $1 - 2\delta$ 的概率，在 $\mathrm{P}_{\mathrm{sg}}^n$ 上，训练数据是 C/C_1-正交的，并且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_1/\sqrt{C})^4$。

证明。 首先注意，引理 C.2 中的所有结果都以至少 $1-2\delta$ 的概率成立。我们将证明，训练数据是 C/C_1-正交的，且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2$ 是这一高概率事件的确定性推论。根据引理C.2，

$$\min_i \|x_i\|^2 \geq \frac{\operatorname{tr}(\Sigma)}{C_0\sqrt{\operatorname{tr}(\Sigma^2)}\log(6n^2/\delta)} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \max_{i\neq j} |\langle x_i, x_j\rangle|.$$

根据假设 (SG3)，这意味着

$$\min_i \|x_i\|^2 \geq \frac{C}{C_0} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot n \max_{i\neq j} |\langle x_i, x_j\rangle|.$$

特别地，训练数据是 C/C_0-正交的（见定义 3.1）。

对于比值 $R = \text{最大值}_{i,\ j} \|x_i\|/\|x_j\|$，若令 $\xi := \frac{\|\Sigma\|_2}{\log(6n/\delta)\operatorname{tr}(\Sigma)}$ 则根据引理 C.2 的第 1 部分，我们有

$$\sqrt{\operatorname{tr}(\Sigma)}(1 - C_0\sqrt{\xi}) \leq \|x_i\| \leq \sqrt{\operatorname{tr}(\Sigma)}(1 + C_0\sqrt{\xi}).$$

根据假设 (SG3) 我们知道 $\xi \leq 1/C$（见 (29)）。因此，当 $C > 1$ 足够大时，利用 $1/(1-x) \leq 1 + 2x$ 对于 $x \in [0, 1/2]$,

$$R = \max_{i,j} \frac{\|x_i\|}{\|x_j\|} \leq \frac{1 + C_0\sqrt{\xi}}{1 - C_0\sqrt{\xi}} \leq \frac{1 + C_0/\sqrt{C}}{1 - C_0/\sqrt{C}} \leq \left(1 + \frac{2C_0}{\sqrt{C}}\right)^2.$$

这就得到了所声称的关于 R 的上界。 □

C.3 定理的证明4.2

现在我们开始证明，如果 w 是 τ-一致的，也就是说，存在严格为正的 s_i，$i = 1, \ldots\ldots, n$ 使得$w = \sum_{i=1}^n s_i y_i x_i$ 且 $\text{最大值}_{i,\ j}\, s_i/s_j \leq \tau$，那么 w 的第一分量较大且为正。根据引理 C.1，这是证明该线性预测器的测试误差接近噪声率的一步。

开始，注意到对于 $i \in \mathcal{C}$，$y_i = \text{sign}([x_i]_1)$，而对于 $i \in \mathcal{N}$，$y_i = -\text{sign}([x_i]_1)$，因此我们有，

$$\left[\sum_{i=1}^n s_i y_i x_i\right]_1 = \sum_{i\in\mathcal{C}} s_i|[x_i]_1| - \sum_{i\in\mathcal{N}} s_i|[x_i]_1| = \sum_{i=1}^n s_i|[x_i]_1| - 2\sum_{i\in\mathcal{N}} s_i|[x_i]_1|.$$

因此，为了证明这个量较大且为正，我们希望证明第一项较大且为正，同时第二项不会太负。我们在下面的引理中证明这一点。

引理 C.3. 存在一个通用常数 $C_1' > 1$（仅取决于 η 和 σ_z），使得对于任意 $\delta \in (0, 1/3)$，若 $n \geq C_1' \log(2/\delta)$，则在训练数据 $\{(x_i, y_i)\}_{i=1}^n \sim \mathrm{P}_{\mathrm{sg}}^n$ 上，以至少 $1 - 3\delta$ 的概率，以下结论成立：

$$\sum_{i=1}^n |[x_i]_1| \geq n\sqrt{\lambda_1}\,\mathbb{E}[|[z]_1|]\left(1 - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right), \quad \textit{and}$$

$$\sum_{i\in\mathcal{N}} |[x_i]_1| \leq n\sqrt{\lambda_1}\,\mathbb{E}[|[z]_1|]\left(\eta + C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right).$$

证明。根据定义，存在独立同分布的 $z_i \sim \mathrm{P}_z$，使得 $x_i = \Sigma^{1/2} z_i$。特别地，$[x_i]_1 = \sqrt{\lambda_1}[z_i]_1$，因此只需从下方界定总和 $\sum_{i=1}^n |[z_i]_1| = \lambda_1^{-1/2} \sum_{i=1}^n |[x_i]_1|$，并从上方界定总和 $\sum_{i\in\mathcal{N}} |[z_i]_1|$。

令 义 $\alpha := \mathbb{E}\,|[z_i]_1|$。注意，由于 $\mathbb{P}(|[z]_1| \le t) \le \beta t$，取 $t = 1/(2\beta)$ 后可知 t

$$\alpha = \mathbb{E}\,|[z_i]_1| \ge \frac{1}{4\beta}. \tag{31}$$

量 $|[z_i]_1| - \alpha$ 的次高斯范数至多为 $c_1\sigma_z$，其中某个绝对常数为 $c_1 > 0$ [Ver18，引理 2.6。8], 并且在指标 $i \in [n]$上独立同分布。因此，根据霍夫丁不等式，这意味着存在某个绝对常数 $c > 0$ 以及任意 $t \ge 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (|[z_i]_1| - \alpha)\right| \ge t\right) \le 2\exp\left(-\frac{cnt^2}{\sigma_z^2}\right).$$

取 $t = c^{-1/2}\sigma_z\sqrt{\log(2/\delta)/n}$ 并使用 (31) 可得，其发生概率至少为 $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n (|[z_i]_1| - \alpha)\right| \le c^{-1/2}\sigma_z\sqrt{\frac{\log(2/\delta)}{n}} \implies \sum_{i=1}^n |[z_i]_1| \ge n\alpha\left(1 - 4c^{-1/2}\sigma_z\beta\sqrt{\frac{\log(2/\delta)}{n}}\right). \tag{32}$$

使用相同的论证（并不失一般性地假设 $|\mathcal{N}| > 0$，因为否则我们可以完全忽略这一项），可得其发生概率至少为 $1-\delta$,

$$\left|\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} (|[z_i]_1| - \alpha)\right| \le c^{-1/2}\sigma_z\sqrt{\frac{\log(2/\delta)}{|\mathcal{N}|}}. \tag{33}$$

由此可见，必须控制噪声点的数量。噪声点的数量 $|\mathcal{N}|$是 n个均值为 η的独立同分布随机变量之和。因此，根据霍夫丁不等式，对于任意 $u \ge 0$,

$$\mathbb{P}\left(||\mathcal{N}| - n\eta| \ge u\right) \le 2\exp\left(-\frac{2u^2}{n}\right).$$

I特别地，取 $u = \sqrt{n\log(2/\delta)/2}$，我们可知其概率至少为 $1-\delta$,

$$\left|\frac{|\mathcal{N}|}{n} - \eta\right| \le \sqrt{\frac{\log(2/\delta)}{n}}.$$

重新整理可得

$$\eta n - \sqrt{n\log(2/\delta)} \le |\mathcal{N}| \le \eta n + \sqrt{n\log(2/\delta)}.$$

由于 η 是绝对常数，利用引理中关于 $n \ge C_1'\log(2/\delta)$ 的假设，我们得到当 C_1' 相对于 η^{-2} 足够大时，

$$\eta n - \sqrt{n\log(2/\delta)} = \eta n\left(1 - \sqrt{\frac{\eta^{-2}\log(2/\delta)}{n}}\right) \ge \frac{1}{2}\eta n,$$

因此

$$\frac{1}{2}\eta n \le |\mathcal{N}| \le \eta n + \sqrt{n\log(2/\delta)} \tag{34}$$

将前两个公式代入(33) 可得

$$\begin{aligned}\sum_{i\in\mathcal{N}} |[z_i]_1| &\le |\mathcal{N}|\alpha\left(1 + c^{-1/2}\sigma_z\alpha^{-1}\sqrt{\frac{\log(2/\delta)}{|\mathcal{N}|}}\right)\\ &\le n\alpha\left(\eta + \sqrt{\frac{\log(2/\delta)}{n}}\right)\cdot\left(1 + 4c^{-1/2}\sigma_z\beta\sqrt{\frac{2\eta^{-1}\log(2/\delta)}{n}}\right)\\ &\le n\alpha\left(\eta + 12c^{-1/2}\sigma_z\beta\sqrt{\frac{2\eta^{-1}\log(2/\delta)}{n}}\right).\end{aligned}$$

第二个不等式用到了(31)。最后一个不等式用到了引理的假设，即 $n \ge C_1'\log(2/\delta)$ 对于足够大的 C_1' 成立，并且 η^{-1} 是一个绝对常数。对这三个事件应用并集界，并取足够大的 C_1' 即可完成证明，因为 σ_z 和 η 都是绝对常数。 □

现在我们证明，以 τ为均匀分布的分类器 u 以高概率具有较大且为正的第一分量，而 $\|[\Sigma^{1/2}u]_{2:d}\|$较小。根据引理D.1，这足以证明泛化误差接近噪声率。

引理 C.4。设 $\tau \ge 1$ 为一个常数，并假设对于某些绝对常数 η、$\Delta > 0$，$\eta \le \frac{1}{2\tau} - \Delta$ 成立。存在一个绝对常数 $C > 1$（仅依赖于 η、σ_z、β、τ和 Δ），使得对于任意 $\delta \in (0, 1/5)$，在假设*(SG1)* 至*(SG3)*（针对这些 C和 δ定义）下，若 $u = \sum_{i=1}^n s_i y_i x_i \in \mathbb{R}^d$相对于 $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{P_{sg}}$是 τ-均匀的，则关于 $\mathrm{P}_{\mathrm{sg}}^n$，以至少 $1 - 5\delta$的概率，有

$$[u]_1 \ge \frac{\tau\Delta n\alpha\sqrt{\lambda_1}}{8\beta}\left(\min_i s_i\right), \quad \textit{and} \quad \|[\Sigma^{1/2}u]_{2:d}\| \le \frac{3}{2}n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}.$$

特别地，如果 $[\Sigma^{1/2}u]_{2:d} \ne 0$，那么

$$\frac{[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|} \ge \frac{\Delta}{12\beta}\cdot\sqrt{\frac{\lambda_1}{\mathrm{tr}(\Sigma_{2:d}^2)}}.$$

证明。 首先，根据并集界，对于 C 充分大的情况，引理 C.3 和引理 C.2以至少 $1 - 5\delta$的概率成立。在证明的其余部分，我们将在这个高概率事件下进行，并将证明该引理是由此得到的确定性推论。

根据定义，当 $i \in \mathcal{C}$ 时，$y_i =$符号$([x_i]_1)$；当 $i \in \mathcal{N}$ 时，$y_i = -$符号$([x_i]_1)$。由于 u 是 τ-均匀的，因此存在严格为正的数 s_i，使得 $u = \sum_{i=1}^n s_i y_i x_i$，其中 $\max_{i,\ j}\frac{s_i}{s_j} \le \tau$。因此我们可以写成，

$$\begin{aligned}[u]_1 = \left[\sum_{i=1}^n s_i y_i x_i\right]_1 &= \sum_{i\in\mathcal{C}} s_i|[x_i]_1| - \sum_{i\in\mathcal{N}} s_i|[x_i]_1|\\ &= \sum_{i=1}^n s_i|[x_i]_1| - 2\sum_{i\in\mathcal{N}} s_i|[x_i]_1|\\ &\ge \min_i s_i\sum_{i=1}^n |[x_i]_1| - 2\max_i s_i\sum_{i\in\mathcal{N}}|[x_i]_1|.\end{aligned} \tag{35}$$

记 $\alpha := \mathbb{E}|[z]_1|$。由 (31) 可知，对 $[z]_1$ 的反集中假设 $\alpha \geq 1/(4\beta)$蕴含 $\alpha \geq 1/(4\beta)$。现在利用引理 C.3，我们有，

$$\begin{aligned}
[u]_1 &\geq n\alpha\sqrt{\lambda_1}\left(\min_i s_i\right)\left(1 - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right) \\
&\quad - n\alpha\sqrt{\lambda_1}\left(\max_i s_i\right)\left(2\eta + 2C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right) \\
&\geq n\alpha\sqrt{\lambda_1}\left(\min_i s_i\right)\left[1 - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}} - \tau\left(2\eta + 2C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right)\right] \\
&\overset{(i)}{\geq} n\alpha\sqrt{\lambda_1}\left(\min_i s_i\right)\left[1 - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}} - \tau\left(\frac{1}{\tau} - 2\Delta + 2C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right)\right].
\end{aligned}$$

不等式 (i) 使用了引理关于 $\eta \leq 1/(2\tau) - \Delta$ 的假设。重新整理上式，我们得到 e

$$\begin{aligned}
[u]_1 &\geq n\alpha\sqrt{\lambda_1}\left(\min_i s_i\right)\left[2\tau\Delta - C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}} - 2\tau C_1'\beta\sqrt{\frac{\log(2/\delta)}{n}}\right] \\
&\geq \frac{1}{2}\tau\Delta n\alpha\sqrt{\lambda_1}\left(\min_i s_i\right) > 0. \qquad (36)
\end{aligned}$$

最后一个不等式利用了 τ、Δ、β 是绝对常数这一事实，并通过取足够大的 C，使得 $n \geq C\log(2/\delta)$，从而推出该不等式。

接下来，我们希望给 $\|[\Sigma^{1/2}u]_{2:d}\|$ 一个上界。为此，我们将使用引理 C.2的第一部分。我们有，

$$\begin{aligned}
\left\|\left[\Sigma^{1/2}u\right]_{2:d}\right\| &= \left\|\left[\textstyle\sum_{i=1}^n s_i y_i \Sigma^{1/2}x_i\right]_{2:d}\right\| \\
&\leq n\left(\max_i s_i\right)\max_i\left\|\left[\Sigma^{1/2}x_i\right]_{2:d}\right\| \\
&\leq n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}\left(1 + C_0\sqrt{\frac{\|\Sigma_{2:d}^2\|_2\log(6n/\delta)}{\mathrm{tr}(\Sigma_{2:d}^2)}}\right) \\
&= n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}\left(1 + C_0\sqrt{\frac{\log(6n/\delta)}{\mathsf{StableRank}(\Sigma_{2:d})}}\right) \\
&\leq \frac{3}{2}n\left(\max_i s_i\right)\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)}. \qquad (37)
\end{aligned}$$

最后一个不等式利用了假设 (SG2)，从而有 稳定秩$(\Sigma_{2:d}) > C\log(6n/\delta)$，并通过取$C$ 足够大得到。结合 (36) 及以上结果，若 $[\Sigma^{1/2}u]_{2:d} \neq 0$，则有

$$\frac{[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|} \geq \frac{1}{3}\tau\Delta\alpha\cdot\frac{\min_i s_i}{\max_i s_i}\cdot\sqrt{\frac{\lambda_1}{\mathrm{tr}(\Sigma_{2:d}^2)}} \geq \frac{\Delta\alpha}{3}\cdot\sqrt{\frac{\lambda_1}{\mathrm{tr}(\Sigma_{2:d}^2)}}.$$

由于根据 (31) 我们有 $\alpha \geq 1/(4\beta)$，证明至此完成。 □

现在我们可以证明定理 4.2。为方便读者，现将其重新陈述如下。

定理 4.2。设 $\tau \geq 1$ 为一个常数，并假设 $\eta \leq \frac{1}{2\tau} - \Delta$，其中 η、$\Delta > 0$为某些绝对常数。存在常数 C、$C' > 0$（仅取决于 η、σ_z、β、τ 和 Δ），使得对于任意 $\delta \in (0, 1/7)$，在假设（SG1）至（SG3）（针对这些 C 和 δ定义）下，以至少 $1 - 7\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{sg}}^n$），若$u \in \mathbb{R}^d$是关于 $\{(x_i, y_i)\}_{i=1}^n$的 τ-一致的，则

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle u, x_k\rangle\big), \qquad \textit{while simultaneously,}$$
$$\eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x\rangle)\big) \leq \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

特别地，如果 $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么线性分类器 $x \mapsto \operatorname{sign}(\langle u, x\rangle)$ 表现出良性过拟合。

证明。根据并集界，以至少 $1 - 7\delta$ 的概率，引理 C.4 和引理 C.2的结论均成立；此前我们已经证明，引理 4.1是引理 C.2 以及假设 (SG1) 至 (SG3)的确定性推论。在证明的余下部分，我们将在该事件上进行讨论，并证明定理是这些引理和假设 (SG1) 至 (SG3)的推论。

由于 u 是 τ-均匀的，存在严格为正的常数 s_i 使得 $u = \sum_{i=1}^n s_i y_i x_i$。我们首先证明 u 对训练数据进行插值：对于任意 $k \in [n]$，有

$$\begin{aligned}
\langle u, y_k x_k\rangle &= s_k\|x_k\|^2 + \sum_{i\neq k}\langle s_i y_i x_i, y_k x_k\rangle \\
&\geq s_k\|x_k\|^2 - n\max_i s_i \cdot \max_{i\neq j}|\langle x_i, x_j\rangle| \\
&= s_k\|x_k\|^2\left(1 - \frac{n\max_i s_i \cdot \max_{i\neq j}|\langle x_i, x_j\rangle|}{s_k\|x_k\|^2}\right) \\
&\geq s_k\|x_k\|^2\left(1 - \frac{n\tau\max_{i\neq j}|\langle x_i, x_j\rangle|}{\|x_k\|^2}\right) \\
&\overset{(i)}{\geq} s_k\|x_k\|^2\left(1 - \frac{C_1\tau}{C}\right) \\
&\overset{(ii)}{\geq} \frac{1}{2}s_k\|x_k\|^2.
\end{aligned} \tag{38}$$

不等式 (i) 利用了训练数据根据引理 C/C_1-正交这一事实4.1，而 (ii) 则由取$C \geq 2C_1\tau$得到。根据引理 C.2，最后这个量严格为正。因此，u 对训练数据进行插值。

现在我们证明泛化误差接近噪声率。根据引理 C.1，如果 $[\Sigma^{1/2}u]_{2:d} = 0$，那么由于 $[u]_1 > 0$，根据引理 C.4，我们有 $\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \text{符号}(\langle u,\ x\rangle)\big) \leq \eta$，证明完成。

因此考虑 $[\Sigma^{1/2}u]_{2:d} \neq 0$ 的情形。令 $c := \Delta/(12\beta)$，其中 $c < 1$ 是一个绝对常数（不妨设 $\beta \geq 1$），因为 Δ、β 根据假设均为绝对常数。由引理 C.4 可得：

$$\frac{[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|} \geq \frac{\Delta}{12\beta}\sqrt{\frac{\lambda_1}{\operatorname{tr}(\Sigma_{2:d}^2)}} = c\sqrt{\frac{\lambda_1}{\operatorname{tr}(\Sigma_{2:d}^2)}}. \tag{39}$$

应用引理 C.1，存在 $c_1 \geq 2$ 使得

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \operatorname{sign}(\langle u, x\rangle)\big) \leq \eta + \frac{c_1\|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1}\left(1 + \sqrt{0 \vee \log\left(\frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right). \tag{40}$$

我们现在考虑两种情况。

情况 1: $c^{-1}\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2} \le 1/2$。由于函数 $\xi \mapsto \xi(1+\sqrt{\log(1/\xi)})$ 在区间 $(0,1/2]$上单调递增，对于任意 $\xi, \xi' \in [0,1/2]$ 满足 $\xi \le \xi'$，我们有 $\xi(1+\sqrt{\log(1/\xi)}) \le \xi'(1+\sqrt{\log(1/\xi')})$。根据 (39) 和情况假设，我们得到

$$\frac{\|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1} \le \sqrt{\frac{c^{-2}\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}} \le \frac{1}{2}. \tag{41}$$

因此，接着由 (40)，

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \mathrm{sign}(\langle u, x\rangle)\big) &\le \eta + \frac{c_1\|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1}\left(1+\sqrt{\log\left(0 \vee \frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right) \\
&\overset{(i)}{=} \eta + \frac{c_1\|[\Sigma^{1/2}u]_{2:d}\|}{\sqrt{\lambda_1}[u]_1}\left(1+\sqrt{\log\left(\frac{\sqrt{\lambda_1}[u]_1}{\|[\Sigma^{1/2}u]_{2:d}\|}\right)}\right) \\
&\overset{(ii)}{\le} \eta + c_1\sqrt{\frac{c^{-2}\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1+\sqrt{\log\left(\sqrt{\frac{\lambda_1^2}{c^{-2}\mathrm{tr}(\Sigma_{2:d}^2)}}\right)}\right) \\
&\overset{(iii)}{\le} \eta + c_1c^{-1}\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1+\sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right).
\end{aligned}$$

等式 (i) 利用了 $\log(x) \ge 0$对 $x \ge 1$成立。不等式 (ii) 利用了 (41)。最后的不等式 (iii) 利用了$c < 1$ 以及 $a \le a \vee b$ 对任意 a, $b \in \mathbb{R}$。

情况 2: $c^{-1}\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2} > 1/2$。在这种情况下，显然有

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \mathrm{sign}(\langle u, x\rangle)\big) \le \eta + c_1c^{-1}\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1+\sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right),$$

由于 $c_1 \ge 2$ 且 $c^{-1}\sqrt{\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2} > 1/2$，右侧至少为 1。由此可见，取 $C' = c_1c^{-1}$ 即可得出定理。

□

C.4 推论4.3、推论 4.4和推论 4.5

本节给出了推论 4.3、推论 4.4和推论 4.5的证明。

推论 4.3. 假设 $0 < \eta \le 0.49$。存在常数 $C, C' > 0$ 使得对于任意 $\delta \in (0, 1/9)$，在假设 *(SG1)* 至 *(SG3)*（针对这些 C和 δ定义）下，以至少 $1 - 9\delta$ 的概率，在 $\mathsf{P}_{\mathsf{sg}}^n$上，最大间隔线性分类器 $w = \mathrm{argmin}\{\|w\|^2 \colon y_i\langle w, x_i\rangle \ge 1\, \forall i\}$ 满足

$$\textit{for all } k \in [n], \quad y_k = \mathrm{sign}\big(\langle w, x_k\rangle\big), \qquad \textit{while simultaneously,}$$

$$\eta \le \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{sg}}}\big(y \neq \mathrm{sign}(\langle w, x\rangle)\big) \le \eta + C'\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1+\sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

具体而言，如果 $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么 w 表现出良性过拟合。

证明。 根据并集界，定理 4.2 和引理 4.1 以至少 $1-9\delta$ 的概率τ-一致线性分类器均会以定理所述的意义表现出良性过拟合，其噪声容忍度由 τ决定。因此，我们只需验证：在这一高概率事件成立且满足这些假设的情况下，线性最大间隔解是 τ-一致的，并且 τ 很小。

根据引理4.1，训练数据是 C/C_1-正交的，并且我们有以下关于 R^2的上界：

$$R^2 = \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \le \left(1+\frac{C_1}{\sqrt{C}}\right)^4 \le \frac{100}{99}. \tag{42}$$

最后一个不等式是通过取 C 为足够大的绝对常数得到的。因此，命题 3.3保证线性最大间隔分类器 w是 τ-一致的，其中 $\tau = R^2\left(1+\frac{2}{pR^2-2}\right)$。特别地，

$$\tau \le R^2\left(1+\frac{2}{CR^2/C_1-2}\right) \le \frac{201}{198}.$$

最后一个不等式用到了 $R^2 \le 100/99$，并且需要取 $C>1$ 足够大。因此，最大间隔线性分类器是 τ-一致的，其中$\tau \le \frac{201}{198}$。由于 $\frac{1}{2\tau} \ge \frac{198}{402} \ge 0.492$，若 $\eta \le 0.49 = 0.492 - 0.002$，我们可以应用定理 4.2。 □

推论 4.4。 假设 $0<\eta\le\frac{49\gamma^2}{100}$。存在常数 C，$C'>0$，使得对于任意 $\delta\in(0,1/9)$，在假设 *(SG1)* 至 *(SG3)*（针对这些 C和 δ定义）下，以至少 $1-9\delta$ 的概率，在$\mathsf{P}_{\mathsf{sg}}^n$上，问题 W 的任意 KKT 点 (3) 均满足

$$\textit{for all } k\in[n],\quad y_k = \operatorname{sign}\big(f(x_k;W)\big),\quad \textit{while simultaneously,}$$

$$\eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{sg}}}\Big(y\neq\operatorname{sign}\big(f(x;W)\big)\Big) \le \eta + C'\sqrt{\frac{\operatorname{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1+\sqrt{0\vee\frac{1}{2}\log\left(\frac{\lambda_1^2}{\operatorname{tr}(\Sigma_{2:d}^2)}\right)}\right).$$

特别是，如果 $\operatorname{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o(1)$，那么神经网络 $f(x;W)$ 会表现出良性过拟合。此外，对于任意初始化 $W(0)$，梯度流在方向上收敛到一个满足上述条件的网络。

证明。与前一推论的证明相同，通过并集界，以至少 $1-9\delta$ 的概率，定理 4.2 和引理 4.1 均成立，并且任意 τ- 一致线性分类器都以由 τ 决定的噪声容忍度表现出良性过拟合。由引理 4.1 可知，训练数据是 C/C_1- 正交的，因此对于 $C>3C_1\gamma^{-3}$，我们可以应用命题 3.4，从而有 $\operatorname{sign}(f(x;W)) = \operatorname{sign}(\langle z,x\rangle)$，其中 z 关于 $\{(x_i,y_i)\}_{i=1}^n$ 对 $\tau = R^2\gamma^{-2}\left(1+\frac{2}{\gamma CR^2/C_1-2}\right)$ 是 τ-一致的。引理 4.1 还意味着，对于足够大的 C，$R^2 \le (1+C_1/\sqrt{C})^4 \le \frac{100}{99}$。因此，对于足够大的 C，$\tau \le \frac{201}{198}\gamma^{-2}$。由于 $\frac{1}{2\tau} \ge \frac{198\gamma^2}{402} > 0.492\gamma^2$，如果 $\eta \le 0.49\gamma^2 = 0.492\gamma^2 - 0.002\gamma^2$，我们可以应用定理 4.2，因为 γ 是一个绝对常数。 □

推论 4.5. 假设 $0<\eta\le\frac{49\gamma^2}{100}$。那么，对于分布 $\mathsf{P}_{\mathsf{gaus}}$，对于任意 $\delta\in(0,1/9)$，若 $\rho\in(1/2,1)$，$d = \tilde{\Omega}(n^{1/(1-\rho)})$，且 $n = \tilde{\Omega}(1)$，则假设 *(SG1)* 至 *(SG3)* 均成立。此外，以至少 $1-9\delta$ 的概率，在 $\mathsf{P}_{\mathsf{gaus}}^n$下，问题 (3) 的 KKT 点表现出良性过拟合：

$$\textit{for all } k\in[n],\quad y_k = \operatorname{sign}\big(f(x_k;W)\big),$$

$$\textit{while simultaneously,}\quad \eta \le \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{gaus}}}\Big(y\neq\operatorname{sign}\big(f(x_k;W)\big)\Big) \le \eta + \tilde{O}\left(d^{\frac{1}{2}(1-2\rho)}\right) = \eta + o_d(1).$$

此外，对于任意初始化 $W(0)$，梯度流都收敛到某个满足上述条件的网络方向。

证明。 首先，显然 P_{gaus} 是 P_{sg}的一个实例，因为 $\Sigma^{-1/2}x$是一个各向同性高斯分布，显然满足反集中性质 $\mathbb{P}(|[z]_1| \le t) \le \beta t$ 对于 $\beta = 1/\sqrt{2\pi}$。因此，我们只需验证假设 (SG1) 至 (SG3) 均成立，且 $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2$ 较小。显然，稳定秩$(\Sigma_{2:d})$ $=d-1$ 并且

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} = \frac{d^\rho + d - 1}{\sqrt{d^{2\rho} + d - 1}}.$$

根据假设， $\rho \in (1/2, 1)$，因此 $d^\rho + d - 1 = \Theta(d)$，而 $d^{2\rho} + d - 1 = \Theta(d^{2\rho})$。所以，

$$\frac{\mathrm{tr}(\Sigma)}{\sqrt{\mathrm{tr}(\Sigma^2)}} = \Theta(d^{1-\rho}).$$

因此，我们看到，如果 $n = \tilde{\Omega}(1)$ 且 $d = \tilde{\Omega}(n^{1/(1-\rho)})$，那么假设 (SG1) 至 (SG3) 均成立，因此在给定的噪声率假设 η下，定理 4.2 和推论 4.4均适用。另一方面，

$$\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2} = \frac{d-1}{d^{2\rho}} = \Theta(d^{1-2\rho}).$$

由于 $\rho > 1/2$，我们可知 $\mathrm{tr}(\Sigma_{2:d}^2)/\lambda_1^2 = o_d(1)$，因此问题 (3) 的 KKT 点的测试误差至多为

$$\eta + C'\sqrt{\frac{\mathrm{tr}(\Sigma_{2:d}^2)}{\lambda_1^2}}\left(1 + \sqrt{0 \vee \frac{1}{2}\log\left(\frac{\lambda_1^2}{\mathrm{tr}(\Sigma_{2:d}^2)}\right)}\right) = \eta + \tilde{O}(d^{\frac{1}{2}(1-2\rho)}) = \eta + o_d(1).$$

□

D 聚类数据的证明

本节给出第 5 节的证明。我们的证明策略与附录 4.2 中定理 C 的证明所采用的策略类似，可概括如下：

1. 我们首先证明，对于线性分类器 $x \mapsto \mathrm{sign}(\langle w, x\rangle)$，若要获得较小的测试误差，只需对每个 $q \in Q$， $\langle w、\ y^{(q)}\mu^{(q)}\rangle$ 均较大且为正。

2. 命题 3.3 和 3.4 表明，当训练数据为 p-正交时，线性分类器和 Leaky ReLU 网络的最大间隔解对应于 τ-均匀分类器。为了利用这一结果，我们需要刻画样本的范数和两两相关性。此外请注意，如果 $w \in \mathbb{R}^d$ 相对于 $\{(x_i, y_i)\}_{i=1}^n$ 是 τ-均匀的，那么对于某个 $s_i > 0$，$w = \sum_{i=1}^n s_i y_i x_i$。因此，根据上述第一步，我们可以看出，刻画样本 $i \in [n]$ 和簇 $q \in Q$ 的 $\langle y_i x_i、\ y^{(q)}\mu^{(q)}\rangle$ 将会很有帮助。引理 D.2 给出了一些初步界限，有助于实现这些目标；引理 D.3 则汇总了我们将用到的训练数据的所有重要性质。特别地，引理 5.1 将由引理 D.3 推出，而定理 5.2 中针对 τ-均匀分类器的测试误差上界也将关键地依赖于该引理。

3. 然后，我们利用上述性质证明定理5.2。

4. 推论5.3和5.4的证明可通过直接计算得到。

D.1 初步集中不等式

我们的第一个引理为 $\mathsf{P}_{\mathsf{clust}}$ 上的任意线性分类器给出了一个泛化界。

引理 D.1. 存在一个绝对常数 $c > 0$，使得若 $w \in \mathbb{R}^d$ 满足 $\langle w, y^{(q)}\mu^{(q)}\rangle \geq 0$ 对每个 $q \in [k]$ 均成立，则

$$\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta + \frac{1}{k}\sum_{q=1}^{k} \exp\left(-c\frac{\langle w, \mu^{(q)}\rangle^2}{\|w\|^2}\right).$$

证明。 我们采用与引理 C.1 相同的证明。根据 $\mathsf{P}_{\mathsf{clust}}$ 的定义，可得 $y = \tilde{y}$ （“干净”标签）的概率为 $1-\eta$，而 $y = -\tilde{y}$ 的概率为 η。因此我们可以计算：

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) &= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0) \\
&= \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = -\tilde{y}) \\
&\quad + \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}) \\
&\leq \eta + \mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}). \qquad (43)
\end{aligned}$$

上式中的第二项可按如下方式估计：

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathsf{P}_{\mathsf{clust}}}(y\langle w, x\rangle < 0,\ y = \tilde{y}) &= \frac{1}{k}\sum_{q=1}^{k} \mathbb{P}_{z\sim\mathsf{P}'_z}(\langle w, y^{(q)}\mu^{(q)} + y^{(q)}z\rangle < 0) \\
&= \frac{1}{k}\sum_{q=1}^{k} \mathbb{P}_{z\sim\mathsf{P}'_z}(\langle w, y^{(q)}z\rangle < -y^{(q)}\langle w, \mu^{(q)}\rangle) \\
&\leq \frac{1}{k}\sum_{q=1}^{k} \exp\left(-c\frac{\langle w, \mu^{(q)}\rangle^2}{\|w\|^2}\right).
\end{aligned}$$

在最后一个不等式中，我们用到了 $y^{(q)}\langle w, \mu^{(q)}\rangle \geq 0$，以及 $y^{(q)}z$ 是次高斯的这一事实（其次高斯范数至多为绝对常数 σ_z），并应用了霍夫丁不等式。将上述结果代入(43) 即可完成证明。 □

根据命题 3.3 和 3.4，我们关注由 $w \in \mathbb{R}^d$ 定义且相对于训练数据是 τ-一致的分类器的行为。这类分类器的形式为 $\sum_{i=1}^{n} s_i y_i x_i$，其中 $s_i > 0$。根据引理 D.1，为了证明 $x \mapsto \operatorname{sign}(\langle w, x\rangle)$ 具有较小的泛化误差，因此有必要刻画$\langle y_i x_i, \mu^{(q)}\rangle$ 对不同簇 q的情况。我们从以下引理着手。

引理 D.2. 设 P'_z 为一个分布，使得 $z \sim \mathsf{P}'_z$ 的各分量均为零均值、相互独立的次高斯随机变量，且次高斯范数至多为一；并且对某个绝对常数 $\kappa > 0$，
$\kappa d \leq \mathbb{E}[\|z\|^2] \leq d$ $\delta \in (0,1)$ $\{z_i\}_{i=1}^{n}$ z' $v_1 \ldots v_k$ 成立。设 ( )。假设 $\overset{\text{i.i.d.}}{\sim}\mathsf{P}$，并设 , , 为 $\mathbb{R}^d$中的任意一组向量。存在绝对常数 C、$C_1 > 1$，使得当 $d \geq C\log(n/\delta)$ 时，以下结论以至少 $1 - 4\delta$ 的概率成立。

(i) 对所有 i，

$$\kappa d\left(1 - C_1\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) \leq \|z_i\|^2 \leq d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right).$$

(ii) 对所有 $i \neq j, |\langle z_i, z_j\rangle| \leq C_1\sqrt{d\log(2n/\delta)}$。

(iii) 对所有 $i = 1, \ldots, k$ 和 $j = 1, \ldots, n, |\langle v_i, z_j\rangle| \leq C_1\|v_i\|\sqrt{\log(2nk/\delta)}$。

证明。 我们分部分证明该引理。我们采用与 Chatterji 和 Long [CL21，引理16]。

对于第一部分，固定 $i \in [n]$。量 $\|z_i\|^2$ 是 d 个独立随机变量之和，这些变量是范数至多为一的次高斯随机变量的平方，因此根据 Vershynin [Ver18，引理 2.7。6], 它是 d 个次指数随机变量之和，且次指数范数至多为一。因此根据伯恩斯坦不等式（见 [Ver18，定理 2.8。1]），存在某个绝对常数 $c > 0$使得对任意 $t \geq 0$,

$$\mathbb{P}(|\|z_i\|^2 - \mathbb{E}\|z_i\|^2| \geq t) \leq 2\exp\left(-c\left(t \wedge \frac{t^2}{d}\right)\right).$$

取 $t = c^{-1}\sqrt{d\log(2n/\delta)}$，我们得到

$$d \geq c^{-2}\log(2n/\delta) \implies t \wedge t^2/d = c^{-2}\log(2n/\delta).$$

因此，我们有

$$\mathbb{P}\left(\exists i : \left|\|z_i\|^2 - \mathbb{E}[\|z_i\|^2]\right| \geq c^{-1}\sqrt{d\log(2n/\delta)}\right) \leq \delta.$$

根据假设， $\kappa d \leq \mathbb{E}[\|z_i\|^2] \leq d$。利用

$$\kappa d\left(1 - c^{-1}\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) = \kappa d - c^{-1}\sqrt{d\log(2n/\delta)},$$

$$d + c^{-1}\sqrt{d\log(2n/\delta)} = d\left(1 + c^{-1}\sqrt{\frac{\log(2n/\delta)}{d}}\right),$$

因此我们有

$$\mathbb{P}\left(\exists i : \kappa d\left(1 - c^{-1}\sqrt{\frac{\kappa^{-2}\log(2n/\delta)}{d}}\right) > \|z_i\|^2 \text{ or } \|z_i\|^2 > d\left(1 + c^{-1}\sqrt{\frac{\log(2n/\delta)}{d}}\right)\right) \leq \delta. \quad (44)$$

接下来，注意到对于任意 i、 $j \in [n]$ 以及任意 $t \geq 0$,

$$\mathbb{P}(|\langle z_i, z_j\rangle| \geq t) \leq \mathbb{P}\left(|\langle z_i, z_j\rangle| \geq t \middle| \|z_j\| \leq \sqrt{2d}\right) + \mathbb{P}(\|z_j\| > \sqrt{2d}).$$

对于 $i \neq j$，在给定z_j的条件下，由于 z_i 具有相互独立的次高斯分量，且次高斯范数至多为 1，因此随机变量 $\langle z_i, z_j\rangle$是零均值次高斯变量，其次高斯范数至多为 $c_1\|z_j\|$，其中 $c_1 > 0$ [Ver18，命题 2.6。1]。因此，根据霍夫丁不等式 [Ver18，定理 2.6。3]，对于某个绝对常数 $c_2 > 0$,

$$\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq t \middle| \|z_j\| \leq \sqrt{2d}\right) \leq 2\exp\left(-c_2 \cdot \frac{t^2}{2d}\right).$$

令 $t = c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}$，我们得到

$$\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}\,\middle|\,\|z_j\| \leq \sqrt{2d}\right) \leq \frac{\delta}{n^2}.$$

利用这一点和(44)，

$$\begin{aligned}
&\mathbb{P}(\text{ for some } i \neq j, |\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}) \\
&\quad \leq n^2\mathbb{P}\left(|\langle z_i, z_j\rangle| \geq c_2^{-1/2}\sqrt{2d\log(2n^2/\delta)}\,\middle|\,\|z_j\| \leq \sqrt{2d}\right) + \mathbb{P}(\text{for some } j \in [n],\ \|z_j\| > \sqrt{2d}) \\
&\quad \leq 2\delta. \qquad (45)
\end{aligned}$$

在最后一个不等式中，我们利用了引理的假设，即 $d \geq 4c^{-2}\log(2n/\delta)$，从而有 $\{\|z_i\|^2 > \sqrt{2}d\} \subset \{\|z_i\|^2 > d(1 + c^{-1}\sqrt{\log(2n/\delta)/d})\}$。

最后，对于 $v \in \{v_1, \ldots, v_k\}$ 和固定的 j，由于 z_j 的分量相互独立且为次高斯分量，我们知道 $\langle z_j, v\rangle$ 是次高斯随机变量，其次高斯范数至多为 $c_1\|v\|$。因此，根据霍夫丁不等式，对于某个常数 $c_3 > 0$，

$$\mathbb{P}\left(|\langle z_j, v\rangle| \geq t\right) \leq 2\exp\left(-c_3 \cdot \frac{t^2}{\|v\|^2}\right).$$

令 $t = c_3^{-1}\|v\|\sqrt{\log(2nk/\delta)}$，并对 $j \in [n]$ 以及 v 的 k 种可能选项应用并集界，我们得到

$$\mathbb{P}\left(\exists j \in [n], v \in \{v_1, \ldots, v_k\} \text{ s.t. } |\langle z_j, v\rangle| \geq c_3^{-1}\|v\|\sqrt{\log(2nk/\delta)}\right) \leq \delta.$$

利用并集界以及(44)、(45) 和上述结果，可得总失败概率为 4δ，证明完毕。

□

接下来，我们说明如何利用上述结果对训练数据进行一些论述。回顾一下，我们观测到样本 $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbf{P}_{\mathsf{clust}}$，它们是 $\{(x_i, \tilde{y}_i)\}_{i=1}^n$ 的含噪版本。我们将 $\mathcal{C} \subset [n]$ 记为干净样本，将 $\mathcal{N} \subset [n]$ 记为噪声样本，因此 $\mathcal{C} \cup \mathcal{N} = [n] = I$。特别地，对于 $i \in \mathcal{N}$，$y_i = -\tilde{y}_i$；而对于 $i \in \mathcal{C}$，$y_i = \tilde{y}_i$。我们还使用记号 $\text{cluster}(i) = q_i$ 和 $I^{(q)} = \{i \in I : \text{cluster}(i) = q\}$and

$$I_{\mathcal{C}}^{(q)} := \{i \in I \cap \mathcal{C} : \text{cluster}(i) = q\}, \quad I_{\mathcal{N}}^{(q)} := \{i \in I \cap \mathcal{N} : \text{cluster}(i) = q\},$$

因此 $I^{(q)} = I_{\mathcal{C}}^{(q)} \cup I_{\mathcal{N}}^{(q)}$。

引理 D.3。 存在一个绝对常数 $C_1' > 1$，使得以下结论成立。对于 $C > 1$在假设 *(CL1)* 至 *(CL4)*下充分大的情形，概率至少为 $1 - 7\delta$时，引理 *D.2* 的 (i) 至 (iii) 项成立（其中 $v_i = \mu^{(i)}$ 对于 $i = 1, \ldots, k$），并且有以下结论。

(i) 对于所有 i,

$$d\left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right) \leq \|x_i\|^2 \leq d\left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right).$$

(ii) 对于每个 $q \in Q$ 和 $i \in I^{(q)}$,

$$\left|\langle \mu^{(q)}, x_i\rangle - \|\mu^{(q)}\|^2\right| \leq C_1'\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)}.$$

(iii) 对于每个 $q \in Q$，若 i、$j \in I^{(q)}$ 和 $i \neq j$，则

$$\left|\langle x_i, x_j\rangle - \|\mu^{(q)}\|^2\right| \leq C_1'\sqrt{d\log(2n/\delta)}.$$

(iv) 对于每个满足 $q \neq r$ 的 q、$r \in Q$，如果 $i \in I^{(q)}$ 且 $j \in I^{(r)}$，那么

$$|\langle x_i, x_j\rangle| \leq \max_{q\neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle| + C_1'\sqrt{d\log(2n/\delta)}.$$

(v) 对所有 $q \in Q$,

$$\left|\frac{|I^{(q)}|}{n} - \frac{1}{k}\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}},$$

and

$$\left|\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \eta\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}, \quad \left|\frac{|I_{\mathcal{C}}^{(q)}|}{|I^{(q)}|} - (1-\eta)\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

证明。根据定义,

$$\|x_i\|^2 = \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle.$$

我们首先注意到，由于 $d \geq Cn^2\log(n/\delta)$ 根据假设 (CL2)，以至少 $1-4\delta$ 的概率，引理 D.2 中的所有结论都成立，其中 v_1、.....、v_k 取为簇均值， $v_i = \mu^{(i)}$。在证明的其余部分，我们都在这一高概率事件上进行。

根据定义，由于我们假设了 $\mathbb{E}[\|z\|^2] = d$,

$$\begin{aligned}
\|x_i\|^2 &= \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle \\
&\geq d\left(1 - C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + \|\mu^{(q_i)}\|^2 - 2C_1\|\mu^{(q_i)}\|\sqrt{\log(2nk/\delta)} \\
&\geq d\left(1 - C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right).
\end{aligned}$$

其中最后一个不等式使用了假设 (CL3) （当 $C > 1$ 足够大时）。另一方面，根据引理D.2，我们还有,

$$\begin{aligned}
\|x_i\|^2 &= \|z_i\|^2 + \|\mu^{(q_i)}\|^2 + 2\langle z_i, \mu^{(q_i)}\rangle \\
&\leq d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + \|\mu^{(q_i)}\|^2 + 2C_1\|\mu^{(q_i)}\|\sqrt{\log(2nk/\delta)} \\
&\overset{(i)}{\leq} d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}}\right) + 2\|\mu^{(q_i)}\|^2 \\
&\overset{(ii)}{\leq} d\left(1 + C_1\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right).
\end{aligned}$$

不等式 (i) 使用假设 (CL3)，而不等式 (ii) 使用假设 (CL2)。

对于引理的第二部分，注意，对于 $i \in I^{(q)}$， $\langle \mu^{(q)}, x_i \rangle - \|\mu^{(q)}\|^2 = \langle z_i, \mu^{(q)} \rangle$，引理 D.2因此给出了该量绝对值的上界。

对于引理的第三部分，考虑属于同一簇的那些 $i \neq j$。对于这些对象，我们有$\mu^{(q_i)} = \mu^{(q_j)}$，因此

$$\begin{aligned}\langle x_i, x_j \rangle &= \langle \mu^{(q_i)} + z_i, \mu^{(q_j)} + z_j \rangle \\ &= \|\mu^{(q_i)}\|^2 + \langle z_i, \mu^{(q_j)} \rangle + \langle \mu^{(q_j)}, z_i \rangle + \langle z_i, z_j \rangle.\end{aligned}$$

根据引理D.2，我们因此有

$$\begin{aligned}|\langle x_i, x_j \rangle - \|\mu^{(q_i)}\|^2| &\leq |\langle z_i, \mu^{(q_j)} \rangle| + |\langle \mu^{(q_i)}, z_j \rangle| + |\langle z_i, z_j \rangle| \\ &\leq 2C_1 \max_q \|\mu^{(q)}\| \sqrt{\log(2nk/\delta)} + C_1 \sqrt{d \log(2n/\delta)} \\ &\overset{(i)}{\leq} 2C_1 \sqrt{\frac{d \log(2nk/\delta)}{Cn}} + C_1 \sqrt{d \log(2n/\delta)} \\ &\overset{(ii)}{\leq} 2C_1 \sqrt{d \log(2n/\delta)}. \end{aligned} \tag{46}$$

不等式 (i) 使用了假设 (CL2)。不等式 (ii) 成立，因为假设 (CL1) 意味着，对于$C > 1$ 足够大时，我们有 $n > 10C_1^2 k$ 因此

$$\log(2nk/\delta) < \log(2n^2/\delta) < 2 \log(2n/\delta).$$

对于引理的第四部分，如果 $i \in I^{(q)}$ 且 $j \in I^{(r)}$，对于 $q \neq r$，

$$\begin{aligned}|\langle x_i, x_j \rangle| &= |\langle \mu^{(q_i)} + z_i, \mu^{(q_j)} + z_j \rangle| \\ &\leq |\langle \mu^{(q_i)}, \mu^{(q_j)} \rangle| + |\langle z_i, \mu^{(q_j)} \rangle| + |\langle \mu^{(q_i)}, z_j \rangle| + |\langle z_i, z_j \rangle| \\ &\leq \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle| + 2C_1 \max_q \|\mu^{(q)}\| \sqrt{\log(2nk/\delta)} + C_1 \sqrt{d \log(2n/\delta)} \\ &\leq \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)} \rangle| + 2C_1 \sqrt{d \log(2n/\delta)}.\end{aligned}$$

倒数第二个不等式使用了引理D.2，最后一个不等式则运用了与 (46) 相同的论证。

对于引理的最后一部分，如果 $q \in Q$，那么该量

$$|I^{(q)}| = \sum_{i=1}^{n} \mathbb{I}(\text{cluster}(i) = q)$$

是.....之和 个均值为 $1/k$ 的 n独立同分布随机变量。根据霍夫丁不等式，对于任意 $u \geq 0$,

$$\mathbb{P}\left(\left| |I^{(q)}| - \frac{n}{k} \right| \geq u \right) \leq 2 \exp\left(-\frac{2u^2}{n} \right).$$

特别地，取 $u = \sqrt{n \log(2k/\delta)}$，并对 k 个簇应用并集界，我们可以看到，概率至少为 $1 - \delta$ 时，对所有 $q \in Q$,

$$\left| \frac{|I^{(q)}|}{n} - \frac{1}{k} \right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

最后，令 N_q 表示簇 q 中噪声样本的数量，

$$|I_{\mathcal{N}}^{(q)}| = N_q = \sum_{i \in I^{(q)}} \mathbb{I}(i \in \mathcal{N}).$$

在给定 $|I^{(q)}|$ 的值的条件下，由于我们考虑的是随机分类噪声， N_q 是 $|I^{(q)}|$ 个均值为 的独立同分布随机变量之和

$$m_q := \mathbb{P}(i \in \mathcal{N}) = \eta.$$

根据霍夫丁不等式，对于任意 $u \geq 0$,

$$\mathbb{P}\left(\left|N_q - |I^{(q)}| m_q\right| \geq u\right) \leq 2\exp\left(-\frac{2u^2}{|I^{(q)}|}\right).$$

特别地，取 $u = \sqrt{|I^{(q)}| \log(2k/\delta)}$，并对 k 个簇应用并集界，我们得到：以至少 $1-\delta$ 的概率，对所有 $q \in Q$,

$$\left|\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \eta\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.$$

由于样本是“干净的”，且位于簇 q 中的概率为 $1-\eta$，完全相同的论证给出了 $|I_{\mathcal{C}}^{(q)}|$ 的界。对引理 D.2中的事件与上述三个事件应用并集界，可得总失败概率为 7δ。

□

D.2 引理 5.1 的证明

现在我们证明，在关于问题参数的假设下，当 p 足够大时，训练数据是 p-正交的，并且每个样本的范数彼此非常接近。

引理 5.1。存在一个绝对常数 $C_2 > 0$，使得对于每个足够大的常数 $C > 0$，对于任意$\delta \in (0, 1/7)$，在假设 *(CL1)* 至 *(CL4)* （针对这些 C 和 δ定义）下，以至少$1 - 7\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{clust}}$），训练数据是 C/C_2-正交的，并且 $\max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$。

证明。 引理 D.3 的所有结论都以至少 $1 - 7\delta$ 的概率成立。我们将证明，该引理是这一高概率事件的确定性推论。

首先，若 $i,\ j \in I^{(q)}$ 且 $i \neq j$，则根据引理 D.3,

$$|\langle x_i, x_j\rangle| \leq \max_q \|\mu^{(q)}\|^2 + C_1'\sqrt{d\log(2n/\delta)} \leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)}\right).$$

另一方面，若 $i \in I^{(q)}$ 且 $j \in I^{(r)}$ 满足 $q \neq r$，则

$$\begin{aligned}
|\langle x_i, x_j\rangle| &\leq \max_{q \neq r} |\langle \mu^{(q)}, \mu^{(r)}\rangle| + C_1'\sqrt{d\log(2n/\delta)} \\
&\overset{(i)}{\leq} \min_q \|\mu^{(q)}\|^2 + C_1'\sqrt{d\log(2n/\delta)} \\
&\leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)}\right),
\end{aligned}$$

其中，在 (i) 中，我们使用假设 (CL4)。因此，对于任意 $i \neq j$，我们有

$$|\langle x_i, x_j \rangle| \leq 2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)}\right). \tag{47}$$

另一方面，根据引理D.3 我们还可以得到

$$d\left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right) \leq \min_i \|x_i\|^2 \leq \max_i \|x_i\|^2 \leq d\left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right). \tag{48}$$

因此我们可以给出上界

$$\begin{aligned}
\frac{\min_i \|x_i\|^4}{\max_i \|x_i\|^2} &\geq d \cdot \frac{\left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right)^2}{\left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right)} \\
&\overset{(i)}{\geq} d \cdot \left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right)^2 \cdot \left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}} - \frac{2}{Cn}\right) \\
&\overset{(ii)}{\geq} \frac{1}{2}d.
\end{aligned} \tag{49}$$

在不等式 (i) 中，我们用到了 $1/(1+x) \geq 1-x$ 对 $x > 0$成立；在不等式 (ii) 中，我们在假设 $C > 1$中取值足够大(CL2)。因此，我们有

$$\frac{\min_i \|x_i\|^4}{\max_i \|x_i\|^2 \max_{i\neq j} |\langle x_i, x_j \rangle|} \geq \frac{d}{2\max_{i\neq j} |\langle x_i, x_j \rangle|} \geq \frac{d}{2C_1' \max\left(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)}\right)}.$$

重新整理并利用假设 (CL2)，可得

$$\begin{aligned}
\min_i \|x_i\|^2 &\geq \frac{1}{2C_1'} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot \frac{d}{\max(\max_q \|\mu^{(q)}\|^2, \sqrt{d\log(2n/\delta)})} \cdot \max_{i\neq j} |\langle x_i, x_j \rangle| \\
&\geq \frac{C}{2C_1'} \cdot \frac{\max_i \|x_i\|^2}{\min_i \|x_i\|^2} \cdot n \max_{i\neq j} |\langle x_i, x_j \rangle|.
\end{aligned}$$

特别地，对于 C/C_2，训练数据是 C_2 -正交的：$= 2C_1'$ (参见定义 3.1)。此外，根据 (48)我们有

$$\begin{aligned}
R^2 &\leq \left(1 + C_1'\sqrt{\frac{\log(2n/\delta)}{d}} + \frac{2}{Cn}\right) \cdot \left(1 - C_1'\sqrt{\frac{\log(2n/\delta)}{d}}\right)^{-1} \\
&\overset{(i)}{\leq} \left(1 + C_1'/\sqrt{C} + \frac{2}{Cn}\right) \cdot \left(1 - C_1'/\sqrt{C}\right)^{-1} \\
&\leq \left(1 + 2C_1'/\sqrt{C}\right)^2.
\end{aligned}$$

不等式 (i) 使用了假设 (CL2)。最后一个不等式利用了：对于足够大的 $\frac{C_1'}{\sqrt{C}} + \frac{2}{Cn} \leq \frac{2C_1'}{\sqrt{C}}$，$C$，并且 $1/(1-x) \leq 1 + 2x$ 对于 $x \in (0, 1/2)$。 □

定理 **5.2** 的证明

现在我们证明，投影到任意 τ-一致线性分类器上的任意方向 $y^{(q)}\mu^{(q)}$ 都具有较大的正值。根据引理 D.1，这将是推导测试误差上界的关键要素。

引理 D.4. 设 $u \in \mathbb{R}^d$ 是 τ-一致的（关于 $\{(x_i, y_i)\}_{i=1}^n$，其中某个绝对常数为 $\tau \geq 1$。设 $\Delta > 0$ 为一个绝对常数，并假设 $\eta \leq \frac{1}{1+\tau} - \Delta$。那么，在假设 *(CL1)* 至 *(CL4)*成立的条件下，如果$C > 1$ 是足够大的绝对常数（仅依赖于 η、 τ和 Δ），那么对于每个 $q \in Q$，在关于 $\mathsf{P}_{\mathsf{clust}}^n$的概率至少为$1 - 7\delta$

$$\frac{\langle u, y^{(q)}\mu^{(q)}\rangle}{\|u\|} \geq \frac{\sqrt{3}(1+\tau)\Delta}{4\sqrt{10}\tau} \cdot \frac{\sqrt{n}\|\mu^{(q)}\|^2}{k\sqrt{d}}.$$

证明。 首先注意到，以至少 $1 - 7\delta$的概率，引理 D.3 和引理 D.2 (其中向量为 $v_i = \mu^{(i)}$) 中的结论均成立。我们还证明了，引理5.1 是这些引理的确定性推论。在证明余下部分，我们将在此高概率事件上进行论证，并说明在假设D.4 成立时，引理 D.2、D.3和 5.1 在假设 (CL1)至 (CL4)下的确定性推论。

由于 u 是 τ-均匀的，存在严格为正的数 s_i，使得 $u = \sum_{i=1}^n s_i y_i x_i$ 且 $\max_{i,j} s_i/s_j = \tau$。我们的证明分为两部分：首先，我们要证明对于每个 q，量

$$\begin{aligned}
\langle u, y^{(q)}\mu^{(q)}\rangle &= \left\langle \sum_{i=1}^n s_i y_i x_i, y^{(q)}\mu^{(q)} \right\rangle \\
&= \sum_{r=1}^k \sum_{i \in I^{(r)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \\
&= \sum_{i \in I^{(q)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle + \sum_{r \neq q} \sum_{i \in I^{(r)}} s_i \langle y_i x_i, y^{(q)}\mu^{(q)}\rangle
\end{aligned}$$

是很大的。为此，我们将考察上面的两项。直观地说，当 $i \in I^{(q)}$ 时，第一项中的加数 $\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle$ 对于干净点 $i \in I_{\mathcal{C}}^{(q)}$ 会是较大的正数，而对于噪声点 $i \in I_{\mathcal{N}}^{(q)}$ 则是负数。因此，只要干净点的数量多于噪声点，第一项就会是较大的正数。对于上面的第二项，由于各簇近似正交，该项的绝对值不会太大。在证明上述结论成立后，我们还要给出 $\|u\|^2$ 的一个上界。

我们将分别考察上述分解中的两项，以证明量 $\langle u, y^{(q)}\mu^{(q)}\rangle$ 是较大的正数。

第一项： $i \in I^{(q)}$。在这种情况下，我们有 $\mu^{(q_i)} = \mu^{(q)}$。如果 $i \in I_{\mathcal{C}}^{(q)}$，则 $y_i = y^{(q)}$；而如果 $i \in I_{\mathcal{N}}^{(q)}$，则 $y_i = -y^{(q)}$。因此，我们将证明干净点的正下界，以及噪声点绝对值的上界。

我们首先给出 干净样本的一个下界 $i \in \mathcal{C}$。对于此类样本， $x_i = \mu^{(q_i)} + z_i$ 并且

$y_i = \tilde{y}^{(q_i)} = \tilde{y}$ 因此

$$\begin{aligned}
\langle s_i y_i x_i, y^{(q)}\mu^{(q)}\rangle &= s_i\langle \mu^{(q)} + z_i, \mu^{(q)}\rangle \\
&\geq s_i\left[\|\mu^{(q)}\|^2 - |\langle z_i, \mu^{(q)}\rangle|\right] \\
&= s_i\|\mu^{(q)}\|^2\left(1 - \frac{|\langle z_i, \mu^{(q)}\rangle|}{\|\mu^{(q)}\|^2}\right) \\
&\overset{(i)}{\geq} s_i\|\mu^{(q)}\|^2\left(1 - \frac{C_1\sqrt{\log(2nk/\delta)}}{\|\mu^{(q)}\|}\right).
\end{aligned}$$

不等式 (i) 使用引理D.2。通过一系列相同的计算，我们可以为 $|\langle s_i y_i x_i,\ \tilde{y}(\mu^{(q)} + z)\rangle|$ 推导出类似的上界。对于噪声样本，我们有 $i \in I_{\mathcal{N}}^{(q)}$，

$$\begin{aligned}
|\langle s_i y_i x_i, \tilde{y}\mu^{(q)}\rangle| &= s_i|\langle \mu^{(q)} + z_i, \mu^{(q)}\rangle| \\
&\leq s_i\left[\|\mu^{(q)}\|^2 + |\langle z_i, \mu^{(q)}\rangle|\right] \\
&\leq s_i\|\mu^{(q)}\|^2\left(1 + \frac{C_1\sqrt{\log(2nk/\delta)}}{\|\mu^{(q)}\|}\right).
\end{aligned}$$

将前面两个公式合并，可得：

$$\begin{cases}
s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \geq s_i\|\mu^{(q)}\|^2 \cdot \left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right), & i \in I_{\mathcal{C}}^{(q)}, \\
|s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle| \leq s_i\|\mu^{(q)}\|^2 \cdot \left(1 + C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right), & i \in I_{\mathcal{N}}^{(q)}.
\end{cases} \tag{50}$$

第二项： $i \in I^{(r)}, r \neq q$。由于 $\mu^{(q_i)} \neq \mu^{(q)}$，对于噪声样本和干净样本，我们都有，

$$\begin{aligned}
|s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle| &= s_i|\langle \mu^{(r)} + z_i, \mu^{(q)}\rangle| \\
&\leq s_i\left(|\langle z_i, \mu^{(q)}\rangle| + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right) \\
&\overset{(i)}{\leq} s_i\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle \mu^{(q)}, \mu^{(r)}\rangle|\right)
\end{aligned} \tag{51}$$

不等式 (i) 使用引理 D.2。综合以上结果，我们得到，

$$
\begin{aligned}
&\langle u, y^{(q)}\mu^{(q)}\rangle \\
&= \sum_{i\in I^{(q)}} s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle + \sum_{r\neq q}\sum_{i\in I^{(r)}} s_i\langle y_i x_i, y^{(q)}\mu^{(q)}\rangle \\
&\geq \sum_{i\in I_{\mathcal{C}}^{(q)}} s_i\|\mu^{(q)}\|^2\left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) - \sum_{i\in I_{\mathcal{N}}^{(q)}} s_i\|\mu^{(q)}\|^2\left(1 + C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) \\
&\quad - \sum_{r\neq q}\sum_{i\in I^{(r)}} s_i\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|\right) \\
&\geq \left(\min_i s_i\right)|I_{\mathcal{C}}^{(q)}|\|\mu^{(q)}\|^2\left(1 - C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) - \left(\max_i s_i\right)|I_{\mathcal{N}}^{(q)}|\|\mu^{(q)}\|^2\left(1 + \sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}}\right) \\
&\quad - \left(\max_i s_i\right)\cdot\left(n - |I^{(q)}|\right)\cdot\left(C_1\|\mu^{(q)}\|\sqrt{\log(2nk/\delta)} + \max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|\right). \qquad (52)
\end{aligned}
$$

为简化记号，定义

$$
\nu := C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \ll 1, \qquad (53)
$$

其中 ν 较小，这由假设 (CL3) 得出。由于 $\tau :=$ 最大值$_i\, s_i$/最小值$_i\, s_i$ 且 $|I^{(q)}| = |I_{\mathcal{C}}^{(q)}| + |I_{\mathcal{N}}^{(q)}|$，因此我们可以将上述不等式写为

$$
\begin{aligned}
&\langle u, y^{(q)}\mu^{(q)}\rangle \\
&\geq \left(\min_i s_i\right)\cdot\left(|I^{(q)}| - |I_{\mathcal{N}}^{(q)}|\right)\cdot\|\mu^{(q)}\|^2(1-\nu) - \left(\max_i s_i\right)\cdot|I_{\mathcal{N}}^{(q)}|\cdot\|\mu^{(q)}\|^2(1+\nu) \\
&\quad - \left(\max_i s_i\right)\cdot(n - |I^{(q)}|)\cdot\left(\|\mu^{(q)}\|^2\nu + \max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|\right) \\
&= \left(\min_i s_i\right)|I^{(q)}|\|\mu^{(q)}\|^2\Bigg[1 - \nu - (1-\nu)\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - (1+\nu)\tau\cdot\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} \\
&\quad - \tau\left(\frac{n}{|I^{(q)}|} - 1\right)\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg] \\
&= \left(\min_i s_i\right)|I^{(q)}|\|\mu^{(q)}\|^2\Bigg[1 - (1+\tau)\cdot\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \left(1 - (1-\tau)\cdot\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|}\right)\nu \\
&\quad - \tau\left(\frac{n}{|I^{(q)}|} - 1\right)\cdot\left(\nu + \frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg]. \qquad (54)
\end{aligned}
$$

由此可见，我们需要控制 $|I^{(q)}|$ 和 $|I_{\mathcal{N}}^{(q)}|$。利用引理D.3，我们得到

$$
\left|\frac{|I^{(q)}|}{n} - \frac{1}{k}\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}, \qquad \left|\frac{|I_{\mathcal{N}}^{(q)}|}{|I^{(q)}|} - \eta\right| \leq \sqrt{\frac{\log(2k/\delta)}{n}}.
$$

特别地，我们有

$$|I^{(q)}| \geq \frac{n}{k} - \sqrt{n\log(2k/\delta)} = \frac{n}{k}\left(1 - \sqrt{\frac{k^2\log(2k/\delta)}{n}}\right) \overset{(i)}{\geq} \frac{n}{2k}, \tag{55}$$

wh不等式 (i) 使用假设 (CL1)，因此 $n \geq 4k^2\log(2k/\delta)$。因此我们有 e

$$\frac{n}{|I^{(q)}|} - 1 \leq 2k.$$

将这些不等式代入 (54) 并利用 $(1-(1-\tau)|I^{(q)}_{\mathcal{N}}|/|I^{(q)}|) \leq \tau$，可得：

$$\begin{aligned}\langle u, y^{(q)}\mu^{(q)}\rangle \geq |I^{(q)}|\|\mu^{(q)}\|^2\left(\min_i s_i\right)\cdot\Bigg[1-(1+\tau)\cdot\left(\eta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu \\ -2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\Bigg].\end{aligned} \tag{56}$$

通过代数计算完成对 $\langle u,\ y^{(q)}\mu^{(q)}\rangle$ 的界定。现在我们要证明 (56) 中括号内的量为正。根据假设，$\eta \leq \frac{1}{1+\tau} - \Delta$ 其中某个绝对常数为$\Delta > 0$,

$$\begin{aligned}&1-(1+\tau)\cdot\left(\eta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu-2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\\ &\geq 1-(1+\tau)\cdot\left(\frac{1}{1+\tau}-\Delta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu-2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\\ &= (1+\tau)\cdot\left[\Delta-\sqrt{\frac{\log(2k/\delta)}{n}}-\frac{\tau\nu}{1+\tau}-\frac{2k\tau}{1+\tau}\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right]\\ &= (1+\tau)\Delta\cdot\left[1-\sqrt{\frac{\Delta^{-2}\log(2k/\delta)}{n}}-\frac{\tau\nu\Delta^{-1}}{1+\tau}-\frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right].\end{aligned} \tag{57}$$

对于括号中的第二项，假设 (CL1) 意味着

$$\sqrt{\frac{\Delta^{-2}\log(2k/\delta)}{n}} \leq \frac{1}{8}.$$

对于接下来的两项，注意到 $\frac{\tau}{1+\tau} \leq 1$ 因为 $\tau \geq 1$。由于 $\nu = C_1\sqrt{\log(2nk/\delta)/\|\mu^{(q)}\|^2}$，当 $C > 1$ 根据假设取足够大的值时，第二项可以趋于零 (CL3) (也就是说，$\min_q \|\mu^{(q)}\| \geq Ck\sqrt{\log(2nk/\delta)}$)：

$$\frac{\tau\nu\Delta^{-1}}{1+\tau} \leq \Delta^{-1}\cdot C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \leq \frac{1}{8}.$$

同样利用假设 (CL3)，当 $C > 1$ 足够大时，我们有：

$$\frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\nu \le 2k\Delta^{-1}\cdot C_1\sqrt{\frac{\log(2nk/\delta)}{\|\mu^{(q)}\|^2}} \le \frac{1}{32}.$$

最后，假设 (CL4) 意味着当 $C > 1$足够大时，

$$\frac{2k\tau\Delta^{-1}}{1+\tau}\cdot\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\left\|\mu^{(q)}\right\|^2} \le \frac{1}{32}.$$

将以上结果代入 (57)，得到

$$\begin{aligned}
&1-(1+\tau)\cdot\left(\eta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu-2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\\
&\quad\ge(1+\tau)\Delta\left(1-\frac{1}{8}-\frac{1}{8}-\frac{1}{32}-\frac{1}{32}\right)>\frac{(1+\tau)\Delta}{2}.
\end{aligned}$$

将此式代入 (56)，得到

$$\begin{aligned}
\langle u, y^{(q)}\mu^{(q)}\rangle &\ge |I^{(q)}|\|\mu^{(q)}\|^2\left(\min_i s_i\right)\cdot\left[1-(1+\tau)\cdot\left(\eta+\sqrt{\frac{\log(2k/\delta)}{n}}\right)-\tau\nu\right.\\
&\qquad\left.-2k\tau\cdot\left(\nu+\frac{\max_{q\neq r}|\langle\mu^{(q)},\mu^{(r)}\rangle|}{\|\mu^{(q)}\|^2}\right)\right]\\
&\ge\frac{1}{2}(1+\tau)\Delta|I^{(q)}|\|\mu^{(q)}\|^2\left(\min_i s_i\right)\\
&\overset{(i)}{\ge}\frac{(1+\tau)n\|\mu^{(q)}\|^2\Delta(\min_i s_i)}{4k}.
\end{aligned}\tag{58}$$

不等式 (i) 使用了 (55)。这为 $\langle u,\ y^{(q)}\mu^{(q)}\rangle$ 提供了所需的下界。

$\|u\|$的上界。这里我们利用样本近似正交这一事实：我们有，

$$\begin{aligned}
\left\|\sum_{i=1}^n s_i y_i x_i\right\|^2 &\le \sum_{i=1}^n s_i^2\|x_i\|^2+\sum_{i\neq j}s_i s_j|\langle x_i,x_j\rangle|\\
&\le n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2\right)+n^2\left(\max_i s_i^2\right)\max_{i\neq j}|\langle x_i,x_j\rangle|\\
&= n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2+n\max_{i\neq j}|\langle x_i,x_j\rangle|\right)\\
&\overset{(i)}{\le}\frac{5}{4}n\left(\max_i s_i^2\right)\left(\max_i\|x_i\|^2\right)\\
&\overset{(ii)}{\le}\frac{5}{4}n\left(\max_i s_i^2\right)\cdot d\left(1+C_1\sqrt{\frac{\log(2n/\delta)}{d}}+\frac{2}{Cn}\right)\\
&\overset{(iii)}{\le}\frac{10}{3}nd\max_i s_i^2.
\end{aligned}\tag{59}$$

不等式 (i) 使用了引理5.1。不等式 (ii) 使用了引理 D.3，而不等式 (iii) 则可通过根据假设 $C > 1$ 取足够大而得出 (CL1) 和 (CL2)。结合 (59) 和 (58)，我们得到，

$$\frac{\left\langle u, y^{(q)}\mu^{(q)}\right\rangle^2}{\|u\|^2} \geq \frac{(1+\tau)^2 n^2 \|\mu^{(q)}\|^4 \Delta^2 \min_i s_i^2}{16k^2 \cdot \frac{10}{3} nd \max_i s_i^2} = \frac{3(1+\tau)^2\Delta^2}{160\tau^2} \cdot \frac{n\|\mu^{(q)}\|^4}{k^2 d}.$$

对上式开平方即可完成证明。 □

结合引理 D.4和引理 D.1，我们可以推导出线性分 $\sum_{i=1}^n s_i y_i x_i$类器的泛化界。

定理 5.2。设 $\tau \geq 1$ 为一个常数，并假设对于某些绝对常数 η、$\Delta > 0$，有 $\eta \leq \frac{1}{1+\tau} - \Delta$。存在常数 C、$C' > 0$（仅依赖于 η、τ 和 Δ），使得对于任意 $\delta \in (0, 1/14)$，在假设 (CL1) 至 (CL4)（针对这些 C 和 δ 定义）下，以至少 $1 - 14\delta$ 的概率（关于 $\mathsf{P}_{\mathsf{clust}}^n$），若 $u \in \mathbb{R}^d$ 关于 $\{(x_i, y_i)\}_{i=1}^n$ 是 τ-一致的，则

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\left(\langle u, x_k\rangle\right),$$
$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\left(y \neq \operatorname{sign}(\langle u, x\rangle)\right) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，若 $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，则线性分类器 $x \mapsto \operatorname{sign}(\langle u, x\rangle)$ 呈现良性过拟合。

证明。 根据并集界，以至少 $1 - 14\delta$ 的概率，引理 D.4 和引理 5.1 的结论成立。在证明的其余部分中，我们将在该高概率事件上进行讨论，并证明该定理是由该事件和假设 (CL1)至 (CL4)确定性推出的。

由于 u 是 τ-一致的，因此存在严格为正的数 s_i 使得 $u = \sum_{i=1}^n s_i y_i x_i$。我们首先证明该估计量能够拟合训练数据。采用与 (38) 中相同的计算可得

$$\begin{aligned}
\langle u, y_k x_k\rangle &= s_k \|x_k\|^2 + \sum_{i\neq k} \langle s_i y_i x_i, y_k x_k\rangle \\
&\geq s_k \|x_k\|^2 \left(1 - \frac{n\tau \max_{i\neq j} |\langle x_i, x_j\rangle|}{\|x_k\|^2}\right) \\
&\overset{(i)}{\geq} s_k \|x_k\|^2 \left(1 - \frac{C_2 \tau}{C}\right) \\
&\geq \frac{1}{2} s_k \|x_k\|^2 > 0.
\end{aligned}$$

不等式 (i) 利用了训练数据根据引理 C/C_2-正交这一事实 5.1，并且我们取 C 相对于绝对常数 C_2、τ 足够大。

现在我们证明，泛化误差接近噪声率。由于 $\eta \leq \frac{1}{1+\tau} - \Delta$，根据引理 D.4，我们知道，对于每个 q，都有：

$$\frac{\left\langle \sum_{i=1}^n s_i y_i x_i, y^{(q)}\mu^{(q)}\right\rangle}{\|\sum_{i=1}^n s_i y_i x_i\|} \geq \frac{\sqrt{3}(1+\tau)\Delta}{4\sqrt{10}\tau} \cdot \frac{\sqrt{n}\|\mu^{(q)}\|^2}{k\sqrt{d}}.$$

现在利用引理D.1，可得

$$\mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle \widehat{\mu}, x\rangle)\big) \leq \eta + \frac{1}{k}\sum_{q=1}^{k} \exp\left(-\frac{3c(1+\tau)^2 n\Delta^2 \|\mu^{(q)}\|^4}{160\tau^2 k^2 d}\right)$$
$$\leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right),$$

其中， C' 是一个与 d 和 n无关的绝对常数。 □

D.4 推论5.3 及推论5.4

本节将说明如何利用定理 5.2 和引理5.1 证明推论 5.3 及推论 5.4。

推论 5.3。 假设 $0 < \eta \leq 0.49$。存在常数 C， $C' > 0$ 使得对于任意 $\delta \in (0, 1/21)$，在假设 *(CL1)* 至 *(CL4)* （针对这些 C 和 δ定义）下，以至少 $1 - 21\delta$ 的概率，$\mathsf{P}_{\mathsf{clust}}^n$，最大间隔线性分类器 $w = \operatorname{argmin}\{\|w\|^2 \colon y_i\langle w, x_i\rangle \geq 1\, \forall i\}$ 满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(\langle w, x_k\rangle\big),$$
$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}(\langle w, x\rangle)\big) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，如果 $n\min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，那么 w 会出现良性过拟合。

证明。 计算过程与推论 4.3 的证明基本相同。根据并集界，定理 5.2 和引理 5.1 的结论以至少 $1 - 21\delta$ 的概率成立，并且任意 τ-一致线性分类器都表现出良性过拟合，其噪声容忍度由 τ 决定。因此，我们验证线性最大间隔分类器是 τ-一致的，且 τ 很小。

根据引理 5.1，训练数据是 C/C_2-正交的，并且 $R^2 = \max_{i,\ j} \|x_i\|^2/\|x_j\|^2 \leq (1 + C_2/\sqrt{C})^2$。由于当 C 足够大时，有 $C/C_2 \geq 3$，根据命题 3.3 ，这意味着线性最大间隔分类器 w 是τ-一致的，且 $\tau \leq R^2\left(1 + \frac{2}{CC_2^{-1}R^2 - 2}\right)$。特别地，我们有

$$\tau \leq \left(1 + \frac{C_2}{\sqrt{C}}\right)^2 \cdot \left(1 + \frac{2}{CC_2^{-1}R^2 - 2}\right)^2 \leq \frac{100}{99} \cdot \frac{201}{200} = \frac{201}{198}.$$

最后一个不等式是通过取 $C > 1$足够大的绝对常数得到的。因此，最大间隔线性分类器是 τ-一致的，其中$\tau \leq \frac{201}{198}$。由于 $\frac{1}{1+\tau} \geq \frac{198}{399} \geq 0.496$，取 $\eta \leq 0.49 = 0.496 - 0.006$ 后，我们可以应用定理5.2。

□

最后，我们证明推论5.4，并再次为方便起见重述该推论。

推论 5.4. 假设 $0 < \eta \leq \frac{49\gamma^2}{100}$。存在常数 C， $C' > 0$ 使得对于任意 $\delta \in (0, 1/21)$，在假设 *(CL1)* 至 *(CL4)* （针对这些 C 和 δ定义）下，以至少 $1 - 21\delta$ 的概率，关于$\mathsf{P}_{\mathsf{clust}}^n$，问题 问题 (3) 的任意 KKT 点 W 满足

$$\textit{for all } k \in [n], \quad y_k = \operatorname{sign}\big(f(x_k; W)\big),$$
$$\textit{while simultaneously,} \quad \eta \leq \mathbb{P}_{(x,y)\sim \mathsf{P}_{\mathsf{clust}}}\big(y \neq \operatorname{sign}\big(f(x; W)\big) \leq \eta + \exp\left(-\frac{n \min_q \|\mu^{(q)}\|^4}{C' k^2 d}\right).$$

特别地，如果 $n \min_q \|\mu^{(q)}\|^4 = \omega(k^2 d)$，那么神经网络 $f(x; W)$ 表现出良性过拟合。此外，对于任意初始化 $W(0)$，梯度流在方向上收敛到一个满足上述条件的网络。

证明。与前一推论一样，以至少 $1 - 21\delta$ 的概率，定理 5.2 和引理 5.1 的结论成立，且任意 τ-一致线性分类器都会表现出良性过拟合，其噪声容忍度由 τ 确定。根据引理 5.1，训练数据是 C/C_2-正交的，因此对 $C > 3C_2\gamma^{-3}$，我们可以应用命题 3.4，从而得到 $\text{sign}(f(x; W)) = \text{sign}(\langle z, x \rangle)$，其中 z 对于 $\tau = R^2\gamma^{-2}\left(1 + \frac{2}{\gamma C R^2/C_2 - 2}\right)$ 而言相对于训练数据是 τ- 一致的。引理 5.1 表明，当 C 足够大时，$R^2 \leq \frac{100}{99}$ 成立，因此当 C 足够大时，$\tau \leq \frac{201}{198}\gamma^{-2}$ 成立。注意，$\frac{1}{1+201\gamma^{-2}/198} \geq 0.496\gamma^2$。因此，我们可以应用定理 4.2，其中$\eta \leq 0.49\gamma^2 = 0.496\gamma^2 - 0.006\gamma^2$，因为 γ 是一个绝对常数。

□

参考文献

[BL21] Peter L Bartlett 和 Philip M Long。“最小范数插值中依赖模型的泛化界失效”。载于：《机器学习研究期刊》22.1（2021），第9297–9311页（引自第2页）。[Bar+20] Peter L. Bartlett、Philip M. Long、Gabor Lugosi 和 Alexander Tsigler。“线性回归中的良性过拟合”。载于：《美国国家科学院院刊》117.48（2020年），第30063–30070页（引自第2、7页）。[BHX20] Mikhail Belkin、Daniel Hsu 和 Ji Xu。“弱特征的双重下降两种模型”。载于：SIAM数据科学数学期刊 2.4（2020年），第1167–1180页（引自第2页）。[BHM18] Mikhail Belkin、Daniel J Hsu 和 Partha Mitra。“过拟合还是完美拟合？对插值分类和回归规则的风险界”。载于：神经信息处理系统进展（NeurIPS）。2018年（引自第2页）。[Cao+22] Yuan Cao、Zixiang Chen、Mikhail Belkin 和 Quanquan Gu。“双层卷积神经网络中的良性过拟合”。载于：arXiv预印本 arXiv:2202.06526（2022年）（引自第2、3页）。[CGB21] Yuan Cao、Quanquan Gu 和 Mikhail Belkin。“次高斯混合分布上过参数化最大间隔分类的风险界”。载于：神经信息处理系统进展（NeurIPS）。2021年（引自第2页）。[CLB21] Niladri S Chatterji、Philip M Long 和 Peter L Bartlett。“双层线性网络中隐式偏置与良性过拟合的相互作用”。载于：arXiv预印本 arXiv:2108.11489（2021年）（引自第2页）。[CL21] Niladri S. Chatterji 和 Philip M. Long。“过参数化情形下插值线性分类器的有限样本分析”。载于：机器学习研究期刊 22.129（2021年），第1–30页（引自第2、11、36页）。[CL20] Geoffrey Chinot 和 Matthieu Lerasle。“关于最小 ℓ_2 插值器的鲁棒性”。载于：arXiv预印本 arXiv:2003.05838（2020年）（引自第2页）。[CB20] Lenaic Chizat 和 Francis Bach。“使用逻辑损失训练的宽双层神经网络中梯度下降的隐式偏置”。载于：学习理论会议（COLT）。2020年（引自第3页）。

[Cla+08] Francis H Clarke、Yuri S Ledyaev、Ronald J Stern 和 Peter R Wolenski。《非光滑分析与控制理论》。第178卷。Springer Science & Business Media，2008年（引自第4页）。[Don+22] Konstantin Donhauser、Nicolo Ruggeri、Stefan Stojanovic 和 Fanny Yang。“噪声插值的快速收敛率要求重新思考归纳偏置的影响”。载于：国际机器学习会议（ICML）。2022年（引自第2页）。[Dut+13] Joydeep Dutta、Kalyanmoy Deb、Rupesh Tulshyan 和 Ramnik Arora。“近似 KKT 点及用于终止的邻近度量”。载于：《全球优化期刊》56.4（2013年），第1463–1499页（引自第4页）。[FCG21a] Spencer Frei、Yuan Cao 和 Quanquan Gu。“通过软间隔和梯度下降进行半空间的无先验学习”。载于：国际机器学习会议（ICML）。2021年（引自第7页）。[FCG21b] Spencer Frei、Yuan Cao 和 Quanquan Gu。“存在对抗性标签噪声时，任意宽度的 SGD 训练神经网络的可证明泛化性”。载于：国际机器学习会议（ICML）。2021年（引自第7页）。[FCB22] Spencer Frei、Niladri S. Chatterji 和 Peter L. Bartlett。“无需线性结构的良性过拟合：通过梯度下降训练、用于含噪线性数据的神经网络分类器”。载于：学习理论会议（COLT）。2022年（引自第2、3、11页）。[Fre+23] Spencer Frei、Gal Vardi、Peter L. Bartlett、Nathan Srebro 和 Wei Hu。“高维数据上训练的 Leaky ReLU 网络中的隐式偏置”。载于：国际学习表征会议。2023年（引自第3、6、16、17页）。[Gun+18] Suriya Gunasekar、Jason Lee、Daniel Soudry 和 Nathan Srebro。“从优化几何角度刻画隐式偏置”。载于：预印本，arXiv:1802.08246（2018年）（引自第3页）。[Has+20] Trevor Hastie、Andrea Montanari、Saharon Rosset 和 Ryan J. Tibshirani。“高维无岭最小二乘插值中的意外发现”。载于：预印本，arXiv:1903.08560（2020年）（引自第2页）。[HMX21] Daniel Hsu、Vidya Muthukumar 和 Ji Xu。“高维中支持向量的激增”。载于：人工智能与统计国际会议（AISTATS）。2021年，第91–99页（引自第3页）。[Ji+20] Ziwei Ji、Miroslav Dudik、Robert E Schapire 和 Matus Telgarsky。“梯度下降沿一般损失函数的正则化路径前进”。载于：学习理论会议。PMLR。2020年，第2109–2136页（引自第3页）。[JST21] Ziwei Ji、Nathan Srebro 和 Matus Telgarsky。“通过对偶加速快速实现间隔最大化”。载于：国际机器学习会议。PMLR。2021年，第4860–4869页（引自第3页）。[JT18] Ziwei Ji 和 Matus Telgarsky。“逻辑回归的风险与参数收敛”。载于：预印本，arXiv:1803.07300（2018年）（引自第3页）。[JT20] Ziwei Ji 和 Matus Telgarsky。“深度学习中的方向收敛与对齐”。载于：神经信息处理系统进展（NeurIPS）。2020年（引自第2、3、5页）。[JT21] Ziwei Ji 和 Matus Telgarsky。“通过原始-对偶分析刻画隐式偏置”。载于：算法学习理论。PMLR。2021年，第772–804页（引自第3页）。

[Koe+21] Frederic Koehler、Lijia Zhou、Danica J Sutherland 和 Nathan Srebro。《插值器的一致收敛：高斯宽度、范数界与良性过拟合》。载于：arXiv 预印本 arXiv:2106.09276（2021年）（引用见第2页）。[Kun+22] Daniel Kunin、Atsushi Yamamura、Chao Ma 和 Surya Ganguli。《拟齐次神经网络的非对称最大间隔偏置》。载于：arXiv 预印本 arXiv:2210.03820（2022年）（引用见第3页）。[LR20] Tengyuan Liang 和 Alexander Rakhlin。《直接插值：核“无岭”回归可以实现泛化》。载于：《统计学年鉴》48.3（2020年），第1329–1347页（引用见第2页）。[LRZ20] Tengyuan Liang、Alexander Rakhlin 和 Xiyu Zhai。《关于最小范数插值器的多重下降与核的受限下等距性》。载于：学习理论会议（COLT）。2020年（引用见第2页）。[LR21] Tengyuan Liang 和 Benjamin Recht。《插值分类器很少犯错》。载于：arXiv 预印本 arXiv:2101.11815（2021年）（引用见第2页）。[LV07] L´aszl´o Lov´asz 和 Santosh Vempala。《对数凹函数的几何与采样算法》。载于：Random Struct. Algorithms 30.3（2007年），第307–358页。ISSN：1042-9832（引用见第7页）。[LL20] Kaifeng Lyu 和 Jian Li。《梯度下降使齐次神经网络的间隔最大化》。载于：国际学习表征会议（ICLR）。2020年（引用见第2–5页）。[Lyu+21] Kaifeng Lyu、Zhiyuan Li、Runzhe Wang 和 Sanjeev Arora。《双层网络上的梯度下降：间隔最大化与简洁性偏置》。载于：神经信息处理系统进展34（2021年），第12978–12991页（引用见第3、5页）。[Mal+22] Neil Mallinar、James B Simon、Amirhesam Abedsoltan、Parthe Pandit、Mikhail Belkin 和 Preetum Nakkiran。《良性、温和还是灾难性：过拟合分类法》。载于：arXiv 预印本 arXiv:2207.06569（2022年）（引用见第2页）。[McR+22] Andrew D McRae、Santhosh Karnik、Mark Davenport 和 Vidya K Muthukumar。《结构化特征下回归与分类中的无害插值》。载于：人工智能与统计国际会议。PMLR。2022年，第5853–5875页（引用见第2页）。[MM19] Song Mei 和 Andrea Montanari。《随机特征回归的泛化误差：精确渐近性与双重下降曲线》。载于：纯粹与应用数学通讯（2019年）（引用见第2页）。[Mon+20] Andrea Montanari、Feng Ruan、Youngtak Sohn 和 Jun Yan。《最大间隔线性分类器的泛化误差：过参数化体制下的高维渐近性》。载于：预印本，arXiv:1911.01544（2020年）（引用见第2、3页）。[Mut+21] Vidya Muthukumar、Adhyyan Narang、Vignesh Subramanian、Mikhail Belkin、Daniel Hsu 和 Anant Sahai。《过参数化体制中的分类与回归：损失函数是否重要？》。载于：机器学习研究期刊22.222（2021年），第1–69页（引用见第2、9页）。[Mut+20] Vidya Muthukumar、Kailas Vodrahalli、Vignesh Subramanian 和 Anant Sahai。《回归中对含噪数据的无害插值》。载于：IEEE信息理论精选领域期刊（2020年）（引用见第2页）。

[Nac+19] Mor Shpigel Nacson、Jason Lee、Suriya Gunasekar、Pedro Henrique Pamplona Savarese、Nathan Srebro 和 Daniel Soudry。《梯度下降在可分数据上的收敛性》。载于：第22届人工智能与统计国际会议。PMLR。2019年，第3420–3428页（引用见第3页）。[NSS19] Mor Shpigel Nacson、Nathan Srebro 和 Daniel Soudry。《随机梯度下降在可分数据上的表现：固定学习率下的精确收敛》。载于：第22届人工智能与统计国际会议。PMLR。2019年，第3051–3059页（引用见第3页）。[NDR20] Jeffrey Negrea、Gintare Karolina Dziugaite 和 Daniel Roy。《为一致收敛辩护：通过去随机化实现泛化及其在插值预测器中的应用》。载于：国际机器学习会议。2020年，第7263–7272页（引用见第2页）。[PL20] Mary Phuong 和 Christoph H Lampert。《ReLU 网络在正交可分数据上的归纳偏置》。载于：国际学习表征会议（ICLR）。2020年（引用见第3页）。[RZ19] Alexander Rakhlin 和 Xiyu Zhai。《拉普拉斯核插值的一致性是一种高维现象》。载于：学习理论会议。PMLR。2019年，第2595–2623页（引用见第2页）。[SVL22] Itay Safran、Gal Vardi 和 Jason D Lee。《浅层单变量 ReLU 网络中线性区域的有效数量：收敛保证与隐式偏置》。载于：预印本，arXiv:2205.09072（2022年）（引用见第3页）。[SBG21] Roei Sarussi、Alon Brutzkus 和 Amir Globerson。《迈向理解线性教师下的神经网络学习》。载于：国际机器学习会议（ICML）。2021年（引用见第3页）。[Sha20] Ohad Shamir。《梯度方法在可分数据上永不过拟合》。载于：预印本，arXiv:2007.00028（2020年）（引用见第3页）。[Sha22] Ohad Shamir。《良性过拟合的隐式偏置》。载于：学习理论会议。PMLR。2022年，第448–478页（引用见第2、3页）。[Sou+18] Daniel Soudry、Elad Hoffer、Mor Shpigel Nacson、Suriya Gunasekar 和 Nathan Srebro。《梯度下降在可分数据上的隐式偏置》。载于：机器学习研究期刊（JMLR）19.70（2018年），第1–57页（引用见第3、4页）。[TOS20] Christos Thrampoulidis、Samet Oymak 和 Mahdi Soltanolkotabi。《多类别分类的理论见解：高维渐近视角》。载于：神经信息处理系统进展33（2020年），第8907–8920页（引用见第2页）。[TVS22] Nadav Timor、Gal Vardi 和 OhadShamir。《ReLU 网络中隐式正则化趋向秩最小化》。载于：arXiv 预印本 arXiv:2201.12760（2022年）（引用见第3页）。[TB20] A. Tsigler 和 P. L. Bartlett。《岭回归中的良性过拟合》。载于：预印本，arXiv:2009.14286（2020年）（引用见第2页）。[Var22] Gal Vardi。《深度学习算法中的隐式偏置》。载于：预印本，arXiv:2208.12591（2022年）（引用见第3页）。[VSS21] Gal Vardi、Ohad Shamir 和 Nathan Srebro。《线性网络与 ReLU 网络中的间隔最大化》。载于：arXiv 预印本 arXiv:2110.02732（2021年）（引用见第3、5页）。

[VYS22] Gal Vardi、Gilad Yehudai 和 Ohad Shamir。《梯度方法可证明收敛到非鲁棒网络》。载于：预印本，arXiv:2202.04347（2022年）（引自第 3 页）。[Ver18] Roman Vershynin。《高维概率：数据科学中的导论与应用》。剑桥大学出版社，2018年（引自第 7、24、28、36 页）。[WDY22] Guillaume Wang、Konstantin Donhauser 和 Fanny Yang。《含噪数据的最小 l1 范数插值的紧致界》。载于：人工智能与统计国际会议（AISTATS），2022年（引自第 2 页）。[WMT21] Ke Wang、Vidya Muthukumar 和 Christos Thrampoulidis。《多类别分类中的良性过拟合：殊途同归于插值》。载于：神经信息处理系统进展（NeurIPS），2021年（引自第 2 页）。[WT21] Ke Wang 和 Christos Thrampoulidis。《高斯混合的二元分类：支持向量的丰沛性、良性过拟合与正则化》。载于：预印本，arXiv:2011.09148（2021年）（引自第 2、11 页）。[WX20] Denny Wu 和 Ji Xu。《过参数化线性回归中的最优加权 ℓ_2 正则化》。载于：神经信息处理系统进展 33（2020年），第 10112–10123 页（引自第 2 页）。[Zha+17] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。《理解深度学习需要重新思考泛化》。载于：国际学习表征会议（ICLR），2017年（引自第 1 页）。[Zho+22] Lijia Zhou、Frederic Koehler、Pragya Sur、Danica J Sutherland 和 Nathan Srebro。《高维广义线性模型的非渐近 Moreau 包络理论》。载于：arXiv 预印本 arXiv:2210.12082（2022年）（引自第 2 页）。