arXiv:2303.04145v2 [cs.LG] 4 Nov 2023

良性 双层 ReLU 卷积神经网络中的过拟合

Yiwen Kou [*1] **Zixiang Chen** [*1] **Yuanzhou Chen** [1] **Quanquan Gu** [1]

摘要

表达能力强大的现代深度学习模型经过训练后，可能会对训练数据过拟合，却仍能很好地泛化。这种现象被称为良性过拟合。近来，一些研究尝试从理论上理解神经网络中的良性过拟合。然而，这些研究要么仅限于采用平滑激活函数的神经网络，要么局限于神经切线核区域。ReLU 神经网络中良性过拟合何时以及如何发生，仍是一个尚未解决的问题。在本研究中，我们通过建立算法依赖的风险界，研究在标签翻转噪声下学习双层 ReLU 卷积神经网络的问题，以回答这一问题。我们证明，在温和的条件下，通过梯度下降训练的神经网络能够实现接近零的训练损失和贝叶斯最优测试风险。我们的结果还揭示，针对测试风险，数据分布条件不同时，良性过拟合与有害过拟合之间存在明显的转变。合成数据上的实验支持了我们的理论。

1. 引言

现代深度学习模型拥有大量参数，通常超过训练数据点的数量。尽管这些模型参数过多，并且对训练数据过拟合，但仍能对未见过的测试数据做出准确预测（Zhang 等人，2017年；Neyshabur 等人，2018年b）。这一现象通常称为良性过拟合（Bartlett 等人，2020），它彻底改变了传统统计学习理论，并引起统计学和机器学习领域的广泛关注（Belkin 等人，2018年；2019年；2020；Hastie 等人，2022年）。

[*]贡献相同 [1]加州大学洛杉矶分校计算机科学系。通信作者：Quanquan Gu<qgu@cs.ucla.edu>。

论文集： 40^{th} 国际机器学习大会，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023 作者。

近年来，已有一系列研究从理论角度探讨良性过拟合。尽管这些研究对良性过拟合现象作出了贡献并带来了新的认识，但大多数研究聚焦于线性模型（Belkin 等人，2020；Bartlett 等人，2020；Hastie 等人，2022年；Wu 和 Xu，2020；Chatterji 与 Long，2021年；Zou 等人，2021b；Cao 等人，2021年）或核模型/随机特征模型（Belkin 等人，2018年；Liang 与 Rakhlin，2020；Montanari 与 Zhong，2022年）。Adlam 与 Pennington（2020）以及 Li 等人（2021年）研究了神经网络模型中的良性过拟合，但他们的研究结果仅适用于神经切线核（NTK）区域（Jacot 等人，2018年），在该区域中，神经网络训练本质上等同于核回归。

由于问题具有非凸性，要理解超越 NTK 区域的神经网络中的良性过拟合，挑战要大得多。近期，Frei 等人(2022年研究了使用平滑化带泄漏ReLU激活函数的全连接两层神经网络，在标签翻转噪声下学习对数凹混合数据的问题。他们在特定正则性条件下证明了风险上界；当标签翻转噪声为零时，该上界与Cao 等人(2021年给出的下界相匹配。Cao 等人(2022年分析了使用多项式 ReLU 激活函数（ReLU^q，$q > 2$）的两层卷积神经网络（CNN）学习问题。他们的分析还指出了一个控制良性过拟合与有害过拟合之间相变的条件。Frei 等人(2022年）和Cao 等人(2021年的分析都高度依赖平滑激活函数，无法处理使用最广泛的 ReLU 激活函数。因此，仍有一个尚未解决的问题：

ReLU 神经网络中的良性过拟合何时发生，又是如何发生的？

本文旨在通过建立学习具有 ReLU 激活函数的双层卷积神经网络的风险界，回答上述问题。

1.1. 问题设定

我们考虑一种类似的Cao 等人(2022年所研究的数据分布。在这种特定的分布中，输入

数据由两类成分组成：标签相关信号和标签无关噪声。这种数据生成模型的灵感源自图像数据：输入由各种图像块构成，只有其中某些图像块与图像的类别标签相关。近期的研究也探讨了类似模型，包括Li 等人(2019年）；Allen-Zhu 和 Li(2020a；b）；Zou 等人(2021a）；Shen 和 Bubeck(2022年）。

定义 1.1. 令 $\boldsymbol{\mu} \in \mathbb{R}^d$ 为一个固定向量，表示每个数据点中包含的信号。每个数据点 $(\mathbf{x}\ y)$ 及其预测变量 $\mathbf{x} = [\mathbf{x}^{(1)\top}, \mathbf{x}^{(2)\top}]^\top \in \mathbb{R}^{2d}$, $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^d$ 和标签 $y \in \{-1, 1\}$ 均由分布 $\mathcal{D}$ 生成，具体定义如下：

1. 真实标签 $\widehat{y}$ 由 *Rademacher* 随机变量生成，即 $\mathbb{P}[\widehat{y} = 1] = \mathbb{P}[\widehat{y} = -1] = 1/2$。随后以概率 p 翻转 $\widehat{y}$，生成观测标签 y，其中 $p < 1/2$，即 $\mathbb{P}[y = \widehat{y}] = 1 - p$ 和$\mathbb{P}[y = -\widehat{y}] = p$。*2.* 从高斯分布 $\mathcal{N}(\mathbf{0}, \sigma_p^2\mathbf{I})$中生成噪声向量 $\boldsymbol{\xi}$。*3.* 随机选取 $\mathbf{x}^{(1)}$,$\mathbf{x}^{(2)}$ 中的一个，并将其指定为 $\widehat{y} \cdot \boldsymbol{\mu}$，表示信号；另一个则由 $\boldsymbol{\xi}$ 给出，表示噪声。

定义1.1 严格推广了Cao 等人(2022年）中的数据分布，即在真实标签 $\widehat{y}$中引入标签翻转噪声，并放宽信号向量 $\boldsymbol{\mu}$ 与噪声向量 $\boldsymbol{\xi}$ 之间的正交条件（参见Cao 等人（2022年）中的定义 3.1 以作比较）。

给定从某个未知联合分布 $\mathcal{D}$ 中抽取的训练数据集 $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$，其中分布定义在 $\mathbf{x} \times y$ 上，我们通过最小化以下采用逻辑损失衡量的经验风险，使用 ReLU 激活训练一个双层卷积神经网络

$$L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell[y_i \cdot f(\mathbf{W}, \mathbf{x}_i)], \tag{1.1}$$

其中 $\ell(z) = \log(1 + \exp(-z))$，且 $f(\mathbf{W}, \mathbf{x})$ 是双层卷积神经网络（详细定义见第 3 节）。我们将使用梯度下降来最小化训练损失$L_S(\mathbf{W})$，并且关注定义如下的测试误差（即真实误差）

$$L_{\mathcal{D}}^{0-1}(\mathbf{W}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \neq \operatorname{sign}\big(f(\mathbf{W}, \mathbf{x})\big)\big]. \tag{1.2}$$

1.2. 主要贡献

我们证明了以下主要结果，该结果刻画了经通过梯度下降训练。

定理 1.2（非正式）。对于任意 $\epsilon > 0$，在满足某些正则性条件的情况下，以至少 $1 - \delta$ 的概率，存在 $0 \le t \le T$ 使得：

1. 训练损失收敛至 ϵ，即 $L_S(\mathbf{W}^{(t)}) \le \epsilon$。*2.* 若 $n\|\boldsymbol{\mu}\|_2^4 \ge \Omega(\sigma_p^4 d)$，则有 $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \le p + \exp\Big(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\Big)$。*3.* 若 $n\|\boldsymbol{\mu}\|_2^4 \le O(\sigma_p^4 d)$，则有 $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \ge p + 0.1$。

定理 1.2 的重要性概括如下：

• 通过标准梯度下降法在逻辑损失上训练的 ReLU 卷积神经网络可以插值含噪训练数据，并实现近乎为零的训练损失。

• 在数据分布和训练样本量满足 $n\|\boldsymbol{\mu}\|_2^4 \ge \Omega(\sigma_p^4 d)$ 的条件下，学习得到的卷积神经网络可以实现近乎最优的测试误差（即贝叶斯风险 p）。

• 另一方面，如果 $n\|\boldsymbol{\mu}\|_2^4 \le O(\sigma_p^4 d)$，插值卷积神经网络模型的测试误差至少比贝叶斯风险高一个常数。这与正向结果共同揭示了良性过拟合与有害过拟合之间存在尖锐相变。

我们的分析依赖于若干新的证明技巧，它们显著推广了信号-噪声分解技术（Cao 等人，2022年）。具体而言，为了处理 ReLU 激活，我们直接利用激活模式和数据结构来刻画每个训练数据点的损失，而不是采用光滑性条件。为处理标签翻转噪声，我们证明，在整个训练过程中，每个训练数据点的损失都以大致相同的速率下降，从而确保即使存在标签噪声，也能学习到信号。

2. 相关工作

本节将详细讨论前文简要提及的一些相关工作。

线性模型的良性过拟合。 一项研究致力于从理论上理解线性模型中的良性过拟合现象。其中一些研究聚焦于线性回归问题。 Belkin 等人 (2020) 对高斯分布模型和傅里叶级数模型中最小二乘预测器的风险曲线形状进行了精确分析。 Hastie 等人 (2022年); Wu 和 Xu (2020) 研究了维度和样本量均增长、但二者之比固定的情形，并展示了相对于该比值的双下降风险

曲线。 Bartlett 等人 (2020) 为过参数化的最小范数插值器建立了匹配的风险上界和下界，并表明在数据协方差的谱满足某些条件时可能出现良性过拟合。 Zou 等人 (2021b) 研究了在过参数化情形下，采用迭代平均或尾部平均的常数步长随机梯度下降具有怎样的泛化能力。另有多项研究探讨了最大间隔线性分类器的良性过拟合。 Muthukum ar 等人 (2021年) 表明，在过参数化情形下，最大间隔预测器与最小二乘预测器相同；而使用 0-1 损失函数和平方损失函数时，它们的泛化表现不同。 Wang 与 Thrampoulidis (2021年); Cao 等人 (2021年) 分别研究了无标签噪声的高斯混合和次高斯混合数据模型，并刻画了良性过拟合可能发生的条件。Chatterji 与 Long (2021年) 表明，在噪声数据上训练的最大间隔算法经过充分过参数化后，可以达到近乎最优的风险。 Shamir (2022年) 研究了线性回归中的最小范数插值预测器以及分类中的最大间隔预测器，并讨论了良性过拟合可能发生或不会发生的条件。

神经网络的良性过拟合。近期一系列研究探讨了神经网络的良性过拟合。Liang 等人(2020) 表明，对于输入维度和样本量的多种缩放方式，核“无岭”回归可能会产生多重下降风险曲线。Adlam 与 Pennington(2020) 对核回归下的泛化进行了精确分析，并揭示了测试误差的非单调行为。Li 等人 (2021年) 研究了定义为双层神经网络的随机特征模型中的良性过拟合。Montanari 与 Zhong(2022年) 研究了 NTK 区域中的双层神经网络，重点考察了维度、样本量和神经元数量均过参数化且呈多项式关系时的泛化性质。Chatterji 与 Long (2022年) 对通过梯度流训练的插值深度线性网络的超额风险给出了界，并表明随机初始化的深度线性网络可以很好地逼近最小范数插值器的风险界。

3. 预备知识

本节将介绍符号约定、双层卷积神经网络模型以及基于梯度下降的训练算法。

记号。 我们用小写字母、小写粗体字母和大写粗体字母分别表示标量、向量和矩阵。对于向量$\mathbf{v} = (v_1, \cdots, v_d)^\top$，我们用 $\|\mathbf{v}\|_2 := \big(\sum_{j=1}^d v_j^2\big)^{1/2}$ 表示其 l_2 范数。对于两个序列 $\{a_k\}$ 和 $\{b_k\}$，若存在某个绝对常数 C使得 $|a_k| \le C|b_k|$，我们记作$a_k = O(b_k)$；若 $b_k = O(a_k)$，则记作 $a_k = \Omega(b_k)$；若 $a_k = O(b_k)$ 且 $a_k = \Omega(b_k)$，则记作 $a_k = \Theta(b_k)$。我们还将 $\widetilde{\ }$ $\widetilde{\ }$if $\lim |a_k/b_k| = 0$ 时的 $a_k = o(b_k)$ 记作。最后，我们用 $O(\cdot)$ 和 $\Omega(\cdot)$ 来省略记号中的对数项。

双层卷积神经网络。 我们考虑如下所述的双层卷积神经网络：其第一层包含 m 个正滤波器和 m 个负滤波器，每个滤波器分别作用于两个图像块 $\mathbf{x}^{(1)}$ 和 $\mathbf{x}^{(2)}$；第二层中正、负卷积滤波器对应的参数分别固定为 $+1/m$ 和 $-1/m$ 。因此，该网络可写为$f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$，其中，正、负滤波器对应的部分网络函数$F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$、 $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$定义如下：

$$
\begin{aligned}
F_j(\mathbf{W}_j, \mathbf{x}) &= \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(1)}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(2)}\rangle)\big] \\
&= \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{j,r}, \widehat{y}\cdot\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle)\big]
\end{aligned}
$$

对于 $j \in \{\pm 1\}$。这里 $\sigma(z) = \max\{0, z\}$ 是 ReLU 激活函数，$\mathbf{W}_j$ 表示与 F_j （正/负滤波器）相关联的模型权重集合，$\mathbf{w}_{j,r} \in \mathbb{R}^d$ 表示 $\mathbf{W}_j$ 中第 r 个滤波器/神经元的权重向量。我们用 $\mathbf{W}$ 表示所有模型权重的集合。注意，我们的卷积神经网络模型也可以看作采用平均全局池化的卷积神经网络（Lin 等，2013）。除了式（1.1）和（1.2）中定义的训练损失和测试误差外，我们还将真实损失（测试损失）定义为 $L_{\mathcal{D}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell[y\cdot f(\mathbf{W}, \mathbf{x})]$。

训练算法。我们使用梯度下降来优化（1.1）。卷积神经网络中滤波器的梯度下降更新可写为

$$
\begin{aligned}
\mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta\cdot\nabla_{\mathbf{w}_{j,r}} L_S(\mathbf{W}^{(t)}) \\
&= \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm}\sum_{i=1}^{n}\ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\cdot j y_i\boldsymbol{\xi}_i \\
&\quad - \frac{\eta}{nm}\sum_{i=1}^{n}\ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)\cdot \widehat{y}_i y_i j\boldsymbol{\mu}.
\end{aligned}
\tag{3.1}
$$

对于所有的 $j \in \{\pm 1\}$ 和 $r \in [m]$，其中我们引入简写记号 $\ell_i'^{(t)} = \ell'[y_i\cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)]$，并假设 ReLU 激活函数在 0 处的梯度为 $\sigma'(0) = 1$，这不失一般性。我们通过高斯初始化来初始化梯度下降，其中 $\mathbf{W}^{(0)}$ 的所有元素均从独立同分布的高斯分布 $\mathcal{N}(0, \sigma_0^2)$ 中采样，其中

σ_0^2 为方差。

4. 主要结果

本节介绍我们的主要理论结果。我们的结果基于以下条件：维度 d、样本量 n、神经网络宽度 m、初始化尺度 σ_0、信号范数 $\|\boldsymbol{\mu}\|_2$、噪声率 p以及学习率 η。本文考虑学习周期 $0 \leq t \leq T^*$，其中 $T^* = \eta^{-1}\text{poly}(\epsilon^{-1}, d, n, m)$ 为最大允许迭代次数。我们可以处理任意大于 $\widetilde{\Omega}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ 的多项式最大允许迭代次数 T^*。

条件 4.1。 假设存在一个充分大的常数 C，使得以下条件成立：

1. 维度 d 足够大：$d \geq C\max\{n\sigma_p^{-2}\|\boldsymbol{\mu}\|_2^2\log(T^*), n^2\log(nm/\delta)(\log(T^*))^2\}$。*2.* 训练样本量 n 和神经网络宽度满足 $m \geq C\log(n/\delta)$ $n \geq C\log(m/\delta)$。*3.* 信号范数满足 $\|\boldsymbol{\mu}\|_2^2 \geq C \cdot \sigma_p^2\log(n/\delta)$。*4.* 噪声率 p 满足 $p \leq 1/C$。*5.* 高斯分布初始化的标准差 σ_0 经过适当选择，使得 $\sigma_0 \leq \big(C\max\big\{\sigma_p d/\sqrt{n}, \sqrt{\log(m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\big\}\big)^{-1}$。

6. The learning rate η satisfies $\eta \leq \big(C\max\big\{\sigma_p^2 d^{3/2}/(n^2 m\sqrt{\log(n/\delta)}), \sigma_p^2 d/n\big\}\big)^{-1}$.

对 d、n、m 的条件旨在确保学习问题处于充分过参数化的设置中，Chatterji & Long (2021)、Cao 等人（2022）和 Frei 等人（2022）也提出了类似条件。对 σ_0 和 η 的条件旨在确保梯度下降能够有效地最小化训练损失。条件 4.1 与 Frei 等人（2022）中的假设（A1）–（A6）之间的区别在于，我们的设置对学习率 η 假设的条件较宽松，其阶为 $O(d^{-3/2})$，而非 $O(d^{-2})$（Frei 等人（2022）中的（A5））；对初始化 σ_0 假设的条件也较宽松，其阶为 $O(d^{-1}n^{1/2})$，而非 $O(d^{-5/2}m^{-1/2})$（Frei 等人（2022）中的（A6））。另一处区别是，Frei 等人（2022）允许神经网络具有任意宽度 m，而我们的条件要求作出一个较宽松的假设，即 m 不应超过维度 d 的指数阶。我们还要求另一项较宽松的条件：m 和 n 彼此不能超过对方的指数阶。此外，与 Cao 等人（2022）不同，我们的条件 4.1 放宽了对 m 和 d 的依赖，即我们不要求神经网络宽度 m满足任何多项式上界，而 Cao 等人 (2022年) 的条件4.2要求 m 以上界受限于 d 的某个分数阶。对 Cao 等人 (2022年) 的另一项改进是，我们在问题中加入了标签翻转噪声 p，但 Frei 等人也考虑了这一点。(2022年)。详细比较见表1 和表 2。

基于这些条件，我们在以下定理中给出主要结果。

定理 4.2。 对于任意 $\epsilon > 0$，在条件 *4.1*下，以至少 $1-\delta$ 的概率存在 $t = \widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$，使得：

1. 训练损失收敛至 ϵ，即 $L_S(\mathbf{W}^{(t)}) \leq \epsilon$。*2.* 当 $n\|\boldsymbol{\mu}\|_2^4 \geq C_1\sigma_p^4 d$ 时，训练后的卷积神经网络将以接近噪声率 p 的分类误差实现泛化：$L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \leq p + \exp\Big(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\Big)$。*3.* 当 $n\|\boldsymbol{\mu}\|_2^4 \leq C_3\sigma_p^4 d$ 时，测试误差 $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \geq p + 0.1$。

这里 C_1、C_2、C_3 是一些绝对常数。

备注4.3。 定理*4.2* 表明，训练损失会在 ϵ 内 $\widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ 次迭代后收敛。此外，当训练损失收敛时，如果信噪比较大，模型可以达到最优测试误差。然而，如果信噪比较小，模型的测试误差至少会比贝叶斯风险差一个常数。区分这两种情况的阈值由条件 $n\|\boldsymbol{\mu}\|_2^4 = \Theta(\sigma_p^4 d)$决定。除了定理 *4.2*中提到的结果外，还需强调，关于测试误差的第二和第三条结论，对于训练时间 t 大于 $\widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$、但小于最大允许迭代次数 $T^* = \eta^{-1}\text{poly}(\epsilon^{-1}, d, n, m)$时，同样成立。

与先前工作的比较。尽管定理 4.2 和 Frei 等人（2022）的定理 3.1 都表明，网络能够达到任意小的逻辑损失，同时使测试误差接近噪声率，但我们的结果与 Frei 等人（2022）不同，因为他们研究的是采用平滑化带泄漏 ReLU激活函数的神经网络，而我们研究的是采用非平滑 ReLU 激活的神经网络。此外，为了使训练误差小于 ϵ，Frei 等人（2022）需要 $O(\epsilon^{-2})$ 次迭代，而我们的结果仅需 $O(\epsilon^{-1})$ 次迭代。与研究采用 ReLU^q($q > 2$) 激活函数且不含标签噪声的卷积神经网络模型的 Cao 等人（2022）相比，我们的设定更具实际意义，因为我们采用 ReLU 激活，将标签翻转噪声纳入考量，同时还去除了信号块与噪声块之间的正交

假设。由于存在标签翻转噪声，通过将测试误差与贝叶斯最优分类器进行比较来评估泛化性能更为自然。这就是我们的定理 4.2 给出测试误差界，而Cao 等人 (2022年) 给出测试损失界的原因。尽管存在差异，我们的结果和他们的结果都给出了精确的相变条件。

5. 证明技巧概述

本节将讨论在我们的设定下研究良性过拟合所面临的主要挑战，并介绍我们在证明中采用的一些关键技巧，以克服这些挑战。基于这些技巧，我们的主定理4.2 的证明将自然而然地得出。所有结果的完整证明均见附录。

5.1. 关键技巧 1：时不变系数比

我们的第一个主要挑战是处理 ReLU 激活函数，即 $\sigma(z) = \max\{0, z\}$。这是最常见、应用最广泛的激活函数之一，但正如下文所述，它也很难分析。要建立良性过拟合保证，关键难点在于证明神经网络能够插值拟合数据。Frei 等人(2022年) 采用了 Frei & Gu(2021年) 提出的基于光滑性的收敛证明技巧。该技巧要求激活函数严格递增且光滑，因此无法应用于 ReLU 激活函数。Cao 等人(2022年) 对信号-噪声分解中的系数进行了迭代分析，相关定义如下。

定义 5.1。 设 $\mathbf{w}_{j,r}^{(t)}$，对于 $j \in \{\pm 1\}, r \in [m]$，它们是卷积神经网络在梯度下降第 t 次迭代时的卷积滤波器。则存在唯一的系数 $\gamma_{j,r}^{(t)}$，和 $\rho_{j,r,i}^{(t)}$，使得

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

进一步记 $\bar{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$，$\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$。则

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i. \tag{5.1}$$

(5.1) 称为 $\mathbf{w}_{j,r}^{(t)}$ 的信号-噪声分解，其中归一化因子 $\|\boldsymbol{\mu}\|_2^{-2}$，$\|\boldsymbol{\xi}_i\|_2^{-2}$用于确保$\gamma_{j,r}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu} \rangle$，$\rho_{j,r,i}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$。根据定义 5.1，通过在整个训练过程中仔细评估系数 $\gamma_{j,r}^{(t)}$，$\rho_{j,r,i}^{(t)}$，$\rho_{j,r,i}^{(t)}$，可以将对 CNN 学习过程的研究归结于此。该技术不依赖激活函数严格递增和光滑的性质，并将作为我们分析的基础。然而，Cao 等人 (2022年)只通过研究 $\max_r \gamma_{j,r}^{(t)}$，$\max_r \rho_{j,r,i}^{(t)}$刻画了主导神经元的行为。为了确保训练后主导神经元能够支配其他神经元，他们要求初始权重不同的神经元具有不同的更新速度，这得益于带有$q > 2$的 ReLU 激活函数^q。但 ReLU 函数是分段线性的，并且每个激活神经元都具有相同的学习速度 $\sigma'(x) = 1$。因此，处理 ReLU 需要新的技术。

为克服这一困难，我们提出一种 时不变系数比分析方法，将 Cao 等人 (2022年) 的技术加以推广。关键引理如下，它刻画了任意时刻 $t \leq T^*$ 的系数阶，并有助于推导定理 4.2中关于测试误差上界和下界的第二和第三部分。

命题 5.2。 在条件 *4.1*下，以下界对于 $t \in [0, T^*]$成立：

- $\rho_{(t),j,r,i}$ 是一个递增序列。此外，$0 \leq \rho_{(t),j,r,i} \leq 4\log(T^*)$ 对所有 $j \in \{\pm 1\}$、$r \in [m]$ 和 $i \in [n]$均成立。
- $\rho_{(t),j,r,i}$ 是一个递减序列。此外，$-4\log(T^*) \leq -2\max_{i,j,r}\{|\langle w_{(0),j,r}, \mu\rangle|, |\langle w_{(0),j,r}, \xi_i\rangle|\} - 10n\sqrt{\log(6n^2/\delta)}/d \cdot 4\log(T^*) \leq \rho_{(t),j,r,i} \leq 0$ 对所有$j \in \{\pm 1\}$、$r \in [m]$ 和 $i \in [n]$均成立。
- $\gamma_{(t),j,r}$ 是一个严格递增序列。此外，$\gamma_{(t),j,r} = \Theta(\|\boldsymbol{\mu}\|_2^2/(d\sigma_p^2)) \sum_{i=1}^{n} \rho_{(t),j,r,i}$ 对所有 $j \in \{\pm 1\}$ 和 $r \in [m]$均成立。

在定义 5.1 中，$\gamma_{j,r}^{(t)}$ 刻画了学习信号向量 $\boldsymbol{\mu}$的进展，而 $\rho_{j,r,i}^{(t)}$ 刻画了滤波器记忆噪声的程度。命题 5.2 的第一和第二条告诉我们，对于任意迭代 t，噪声记忆程度 $\rho_{j,r,i}^{(t)}$，$\rho_{j,r,i}^{(t)}$ 均受总训练轮数 T^*的对数阶上界约束。特别地，当 $T^* = \eta^{-1}\text{poly}(\epsilon^{-1}, d, n, m)$ 时，$\rho_{j,r,i}^{(t)}$，$\rho_{j,r,i}^{(t)} = O(1)$。命题 5.2 的第三条是我们的方法相较于 Cao 等人（2022）取得的主要改进。它表明 $\gamma_{j,r}^{(t)}$ 严格递增，这意味着尽管存在标签翻转噪声，卷积神经网络仍会学习信号 $\boldsymbol{\mu}$。此外，系数比 $\gamma_{j,r}^{(t)}/(\sum_{i=1}^{n} \rho_{j,r,i}^{(t)})$ 的阶是时不变的。当信号强度 $\|\boldsymbol{\mu}\|_2$ 相较于噪声方差 $\sqrt{d}\sigma_p$较大时，神经元倾向于学习信号。

当 $\|\boldsymbol{\mu}\|_2$ 相较于 $\sqrt{d}\sigma_p$较小时，神经元倾向于学习噪声。借助时不变系数比技术，我们可以刻画训练过程中所有神经元的行为，从而处理 ReLU 激活函数。

为了证明第三条，我们需要刻画 $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$ 的激活模式。注意到，对于任意权重 $\mathbf{w}_{j,r}^{(t)}$，$\sum_i \rho_{i,j}^{(t)}$，r 的增量会按 $\sum_i \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$ 缩放，因此我们考虑集合序列 $\{S_{j,r}^{(t)}\}_{t=0}^{T^*}$，其中 $S_{j,r}^{(t)}$，定义为 $\{i|y_i = j, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}$。我们证明在整个训练过程中，这是一个递增的集合序列，从而得到 $|S_{j,r}^{(t)}| = \Theta(n)$。直观而言，这意味着对于给定样本，一旦某个神经元被噪声块激活，即使梯度下降更新了该神经元的权重，它在整个训练过程中仍会保持激活。将这一发现应用于 (5.3) 和 (5.4)，可直接推出 $\gamma_{j,r}^{(t)}$，与 $\sum_i \rho_{j,r}^{(t)}$，$_i$ 的增量比始终保持为 $\Theta(\|\boldsymbol{\mu}\|_2^2/(\sigma_p^2 d))$。

5.2. 关键技术 2：系数更新的自动平衡

我们的第二个主要挑战是应对标签翻转噪声。实证研究发现，过参数化神经网络在使用带有标签噪声的数据进行训练时，仍能实现良好的泛化（Belkin 等人，2019年；Zhang 等人，2021年），这与统计学习领域长期以来的理论相悖。为了拟合包含信号 $-y\boldsymbol{\mu}^\top$ 和噪声 $\boldsymbol{\xi}$的受污染数据，神经网络权重必须捕捉随机噪声$\boldsymbol{\xi}$，这会损害泛化。更糟的是，标签翻转噪声可能诱使学习器捕捉对抗信号 $-\boldsymbol{\mu}$，而非 $\boldsymbol{\mu}$。让我们考察系数 $\gamma_{j,r}$,$\rho_{j,r,i}$, $\rho_{j,r,i}$的更新规则。

引理 5.3。 系数 $\gamma_{j,r}^{(t)}$、、 $\rho_{j,r}^{(t)}$、、$_i$、 $\rho_{j,r}^{(t)}$、、$_i$在定义 *5.1*中给出，满足以下迭代方程：

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0, \tag{5.2}$$

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \Bigg[\sum_{i\in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) - \sum_{i\in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) \Bigg] \cdot \|\boldsymbol{\mu}\|_2^2, \tag{5.3}$$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = j), \tag{5.4}$$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = -j), \tag{5.5}$$

对于所有的 $r \in [m]$、 $j \in \{\pm 1\}$ 和 $i \in [n]$，其中 $S_+ := \{i \in [n] | y_i = \widehat{y}_i\}$ 和 $S_- := \{i \in [n] | y_i \neq \widehat{y}_i\}$。

当不存在标签翻转噪声时，我们可以得出结论：$S_- = \varnothing$，且信号系数 $\gamma_{j,r}^{(t)}$,严格递增，因为 $\ell_i'^{(t)}$ 严格为负。这一关键观察在Cao 等人 （2022年）的证明中发挥了重要作用。遗憾的是，噪声标签的存在会引入负项 $\sum_{i\in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle)$。因此，我们无法直接根据公式（5.2）判断$\gamma_{j,r}^{(t)}$,是否递增。如果（噪声）样本 $i \in S_-$ 对应的损失梯度 $\ell_i'^{(t)}$ 相较于（干净）样本 $i \in S_+$对应的损失梯度 $\ell_i'^{(t)}$ 特别大，那么（5.2）的确可能无法保证 $\gamma_{j,r}^{(t)}$,增加。为了说明神经网络在拟合噪声数据 S_-的同时仍能学习信号，我们需要对学习过程进行更深入、更审慎的刻画。

为克服处理标签翻转噪声的困难，我们采用一种称为系数更新的自动平衡的关键技术。如（5.3）所示，如果我们能够证明损失梯度 $\ell_i'^{(t)}$ 在所有样本间基本上是“平衡的”，即 $\ell_i'^{(t)}/\ell_k'^{(t)} \leq C$，$\forall i$，$j \in [n]$，那么只要噪声标签的比例不是太大，干净标签就会抵消噪声标签的影响，最终可以证明 $\gamma_{j,r}^{(t)}$，在增加。这为下一个引理提供了依据。

引理 5.4。 在条件*4.1*下，对于任意 $t \in [0, T^*]$:

$$y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_4, \tag{5.6}$$

$$\ell_i'^{(t)}/\ell_k'^{(t)} \leq C_5, \tag{5.7}$$

对于任意 i、 $k \in [n]$，其中 $C_4 = \Theta(1)$ 是一个正常数，$C_5 = \exp(C_4)$，且 $\ell_i'^{(t)} = \ell'(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i))$，$\ell_k'^{(t)} = \ell'(y_k f(\mathbf{W}^{(t)}, \mathbf{x}_k))$。

对 $\ell_i'^{(t)}/\ell_k'^{(t)}$ 进行界定的策略最早由 Chatterji 和 Long（2021年）在研究线性分类时提出，后来又被扩展到使用平滑泄漏 ReLU 激活函数的神经网络中（Frei 等人，2021年；2022年）。其主要思想是，根据 logit 函数 $\ell'(z) = -1/(1+\exp(z))$ 的性质，即 $\ell'(z_1)/\ell'(z_2) \approx \exp(z_2 - z_1)$，为了对 ℓ_i' 与 ℓ_k' 的比值进行上界估计，只需界定 $y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ 与 $y_k f(\mathbf{W}^{(t)}, \mathbf{x}_k)$ 之间的差异。为了进一步刻画这一差异，上述工作利用光滑性，将函数差异转化为梯度差 $\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ 与 $\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_k)$。然而，这种基于光滑性的技术无法直接应用于 ReLU 神经网络。

在本文中，我们应用信号-噪声分解，并近似 $y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)$，使得 $\sum_r \rho_{y_i,r,i}^{(t)}$ 对于任意 $i \in [n]$，且近似误差较小。因此，引理5.4 可以进一步简化为证明以下中间结果。

引理 5.5。在条件4.1下，以下界限对 $t \in [0, T^*]$ 成立：

$$\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)} \leq \kappa, \tag{5.8}$$

对于任意 $i, k \in [n]$，其中 $\kappa = \Theta(1)$ 是一个正常数。

注意，（5.8）比（5.6）容易处理得多，因为我们可以直接使用（5.4）的迭代分析，从而得到如下更新规则：

$$\begin{aligned}\sum_{r=1}^{m} [\overline{\rho}_{y_i,r,i}^{(t+1)} - \overline{\rho}_{y_k,r,k}^{(t+1)}] = \sum_{r=1}^{m} [\overline{\rho}_{y_i,r,i}^{(t)} - \overline{\rho}_{y_k,r,k}^{(t)}] \\ - \frac{\eta}{nm} \cdot \big(|S_i^{(t)}| \ell_i'^{(t)} \|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(t)}| \ell_k'^{(t)} \|\boldsymbol{\xi}_i\|_2^2\big),\end{aligned} \tag{5.9}$$

其中 $S_i^{(t)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \geq 0\}$，$i \in [n]$。现在，我们考虑两种情况：

- 如果 $\sum_{r=1}^{m} \rho_{(t) y_i, r, i} - \sum_{r=1}^{m} \rho_{(t) y_k, r, k}$ 相对较小，我们将证明，对于足够小的步长 η，$\sum_{r=1}^{m} \rho_{(t+1) y_i, r, i} - \sum_{r=1}^{m} \rho_{(t+1) y_k, r, k}$ 不会增长太多。
- 如果 $\sum_{r=1}^{m} \rho_{(t) y_i,r,i} - \sum_{r=1}^{m} \rho_{(t) y_k,r,k}$ 相对较大，则会导致 $\ell_{(t) i} / \ell_{(t) k}$ 收缩，因为 $\ell_{(t) i} / \ell_{(t) k}$ 可以近似为 $\exp(\sum_{r=1}^{m} \rho_{(t) y_k,r,k} - \sum_{r=1}^{m} \rho_{(t) y_i,r,i})$。此外，由于我们可以证明 $\|\boldsymbol{\xi}_i\|_2^2 \approx \|\boldsymbol{\xi}_k\|_2^2$ 和 $|S_{(t) i}| / |S_{(t) k}| = \Theta(1)$，因此根据（5.9），$\sum_{r=1}^{m} \rho_{(t+1) y_i,r,i} - \sum_{r=1}^{m} \rho_{(t+1) y_k,r,k}$ 将会减小。

结合这两种情况，$\sum_{r=1}^{m} \rho_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \rho_{y_k,r,k}^{(t)}$ <style id='16'>可以由一个常数作为上界，这就完成了引理 5.5 的证明，而引理 5.4 的证明可直接得出。

5.3. 关键技术 3：依赖算法的测试误差分析

通过选择 $\epsilon = 1/(4n)$，定理4.2 可得 $L_S(\mathbf{W}^{(t)}) \leq 1/(4n)$，这进一步意味着训练误差为0。另一方面，由于存在标签翻转噪声，我们知道贝叶斯最优测试误差至少为 p。因此，测试误差与训练误差之间的差距至少为 p，这使我们无法应用常用的标准基于一致收敛的界（Bartlett 等人，2017年；Neyshabur 等人，2018年a）或基于稳定性的界（Hardt 等人，2016；Mou 等人，2017年；Chen 等人，2018年）。本文将给出依赖于算法的测试误差分析。首先，我们可以将测试误差分解如下

$$\begin{aligned}&\mathbb{P}\big(y \neq \operatorname{sign}(f(\mathbf{W}^{(t)}, \mathbf{x}))\big) \\ &= p + (1-2p)\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big).\end{aligned} \tag{5.10}$$

根据（5.10），测试误差分析可归结为界定错误预测概率 $\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big)$。为此，我们需要界定训练损失收敛至 ϵ 时的系数量级。以下结果表明，在时刻 $T_1 < T^*$，$\rho_{y_i}^{(t)}, r, i$ 中有常数比例会达到常数量级。

引理 5.6。在条件4.1下，存在 $T_1 = \Theta(\eta^{-1} nm\sigma_p^{-2} d^{-1})$ 使得 $\rho_{y_i, r, i}^{(T_1)} \geq 2$ 对所有 $r \in S_i^{(0)} := \{r \in [m] : \langle \mathbf{w}_{y_i, r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$ 以及 $i \in [n]$ 均成立。

证明该引理的主要思路是，$\ell_i'^{(t)}$ 在 $\Theta(1)$ 在时刻 T_1 之前保持不变，而通过将5.4）中系数的 $\ell_i'^{(t)}$ 因子替换为其常数下界，可以大幅简化系数的动态变化。在时刻 T_1 之后，根据命题5.2中系数的单调性和阶数，我们可以用以下引理描述系数的阶数。

引理 5.7. 在条件 4.1下，对于 $t \in [T_1, T^*]$，以下系数阶成立：

- $\sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} = \Omega(n) = O(n \log(T^*))$ *for any* $j \in \{\pm 1\}$ *and* $r \in [m]$.
- $\sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} / \gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2})$ *for any* $j, j' \in \{\pm 1\}$ *and* $r, r' \in [m]$.
- $\max_{j,r,i} |\rho_{j,r,i}^{(t)}| = \max\big\{O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big), O\big(\sqrt{\log(n/\delta)} \log(T^*) \cdot n/\sqrt{d}\big)\big\}$.

利用引理 $\gamma_{j r}^{(t)}$，$\rho_{j r, i}^{(t)}$ 中给出的尺度5.7，以及利普希茨函数的高斯集中，可以沿用 Chatterji 与 Long（2021年）中定理 1 的证明以及 Frei 等人(2022年) 中引理 3 的类似思路，直接得到测试误差上界（定理 4.2的第二部分）。为证明测试误差下界（定理 4.2的第三部分），我们首先对错误预测概率项 $\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big)$ 给出下界：

$$0.5\mathbb{P}\Bigg(\underbrace{\Big|\sum_{j,r} j\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle)\Big| \geq C_6 \max_j \Big\{\sum_r \gamma_{j,r}^{(t)}\Big\}}_{\text{event } \Omega}\Bigg),$$

其中所有随机性都位于等式左侧，因此可以将其视为高斯随机向量 $\boldsymbol{\xi}$ 的函数。

接下来，我们给出关键引理，该引理可以通过对$\mathbf{w}_{j,r}^{(t)}$进行分解，并利用引理5.7中给出的分解系数的尺度来证明。

引理 5.8。对于 $t \in [T_1, T^*]$，记作 $g(\boldsymbol{\xi}) = \sum_{j,r} j\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle)$。存在一个固定向量 $\mathbf{v}$，满足$\|\mathbf{v}\|_2 \leq 0.06\sigma_p$，使得

$$\sum_{j' \in \{\pm 1\}} [g(j'\boldsymbol{\xi} + \mathbf{v}) - g(j'\boldsymbol{\xi})] \geq 4C_6 \max_{j \in \{\pm 1\}} \Big\{ \sum_r \gamma_{j,r}^{(t)} \Big\}, \tag{5.11}$$

对所有 $\boldsymbol{\xi} \in \mathbb{R}^d$。

根据引理5.8，由抽屉原理，$\boldsymbol{\xi}$、$\boldsymbol{\xi} + \mathbf{v}$、$-\boldsymbol{\xi}$、$-\boldsymbol{\xi} + \mathbf{v}$ 中必有一个属于 Ω，也就是说，$\Omega \cup (-\Omega) \cup (\Omega - \mathbf{v}) \cup (-\Omega - \mathbf{v}) = \mathbb{R}^d$。由并集界可得

$$\mathbb{P}(\Omega) + \mathbb{P}(-\Omega) + \mathbb{P}(\Omega - \mathbf{v}) + \mathbb{P}(-\Omega - \mathbf{v}) \geq 1. \tag{5.12}$$

由于噪声 $\boldsymbol{\xi}$ 服从对称分布，我们有 $\mathbb{P}(\Omega) = \mathbb{P}(-\Omega)$。我们可以使用基于全变差（TV）距离的一些技巧来证明

$$|\mathbb{P}(\Omega) - \mathbb{P}(\Omega - \mathbf{v})|, |\mathbb{P}(-\Omega) - \mathbb{P}(-\Omega - \mathbf{v})| \leq 0.03. \tag{5.13}$$

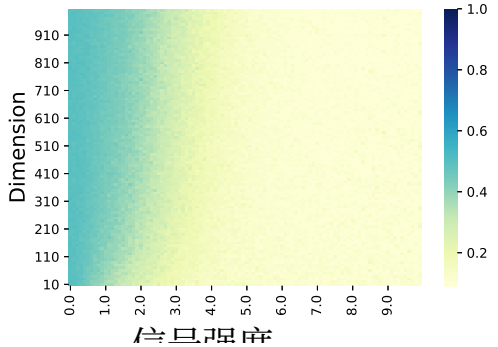
根据 (5.12) 和 (5.13)，我们已经证明 $\mathbb{P}(\Omega) \geq 0.22$。将 $\mathbb{P}(\Omega) \geq 0.22$ 代入 (5.10)，即可完成测试错误下界的证明。

6. 实验

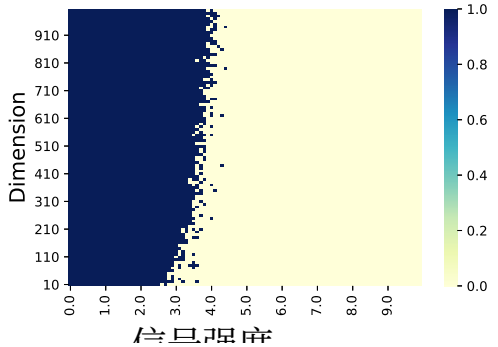
本节将展示合成数据的模拟，以佐证上一节的理论分析。我们的实验代码可在 GitHub [1] 上找到。

合成数据实验。我们在此严格按照定义 1.1 生成合成数据。具体而言，我们将训练数据规模设为 $n = 20$，并将标签翻转噪声设为 0。1。由于该学习问题具有旋转不变性，不失一般性，我们设定 $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|_2 \cdot [1, 0, \ldots, 0]^\top$。随后，我们从高斯分布 $\boldsymbol{\xi}$ 中生成噪声向量$\mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$，其标准差固定为 $\sigma_p = 1$。

我们使用 ReLU 激活函数训练第 3 节定义的双层卷积神经网络模型。滤波器数量设为$m = 10$。我们使用 PyTorch 的默认初始化方法来初始化卷积神经网络参数，并以 0.1的学习率进行全批量梯度下降，训练卷积神经网络 100 次迭代。我们考虑维度 d 从 100 到 1100不等的情形，以及信号强度 $\|\boldsymbol{\mu}\|_2$从 1 到11不等的情形。根据我们的结果，对于任意维度 d 和信号强度 μ 设置，我们考虑的训练方案都能保证训练损失小于 0.01。训练后，我们使用1000 个测试数据点估算每种情况下的测试误差。结果以参数 d 和 $\|\boldsymbol{\mu}\|_2$ 的热力图形式展示于图 1。



(a) 原始测试误差热力图



(b) 截断测试误差热力图

图 *1.* a) 展示了不同维度 d 和信号强度 $\boldsymbol{\mu}$下合成数据的测试误差热力图。较高的测试误差用蓝色标记，较低的测试误差用黄色标记。b) 展示了截断值热力图，其中小于 0.2 的值设为 0（黄色），大于 0.2 的值设为 1（蓝色）。

对于特定情况 $\|\boldsymbol{\mu}\|_2 = 5$ 和 $d = 100$，我们在图 2中绘制了训练过程中训练损失、测试损失和测试误差的变化。从图中可以看出，测试误差达到了贝叶斯最优误差 0.1，而训练损失则收敛至零。

在第 5 节中，我们直接利用激活模式和数据结构刻画每个样本的损失，并在引理 5.7 中证明，对于 $t \leq T^*$ 和任意 $i, k \in [n]$，$y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_4$。为展示这一点，我们针对 $\|\boldsymbol{\mu}\|_2 = 5, d =$100 的情况进行了另一项实验，并绘制了每次迭代的 $\max\ y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ 和 $\min\ y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$（间隔）。如图 3 所示，在训练过程中二者的差值始终不会变得太大（不超过 6）。

[1] https://github.com/uclaml/Benign ReLU 卷积神经网络

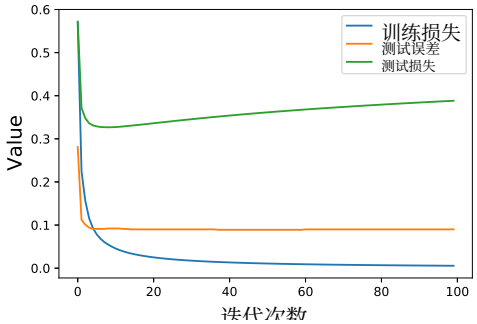


图 2。 在 $\|\boldsymbol{\mu}\|_2 = 5$ 和 $d = 100$ 条件下，经过 100 次迭代的训练损失、测试损失和测试误差。

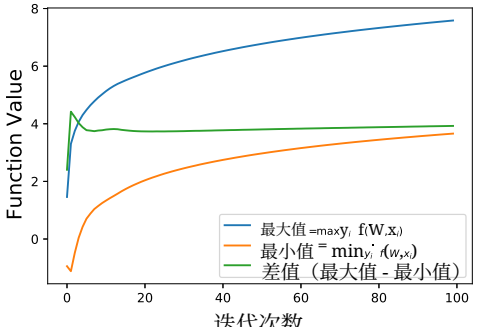


图 3。 在 $\|\boldsymbol{\mu}\|_2 = 5$ 和 $d = 100$ 条件下，经过 100 次迭代的最大值 $y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ 和最小值 $y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$（间隔）。

7. 结论与未来工作

本文研究了存在标签翻转噪声时双层 ReLU 卷积神经网络的良性过拟合。我们将Cao 等人(2022年首次提出的信号—噪声分解技术加以推广，并提出三项关键技术：时间不变的系数比、系数更新的自动平衡和依赖算法的测试误差分析。借助这些技术，我们证明了训练损失的收敛性，给出了卷积神经网络的测试误差接近噪声率的精确条件，并揭示了良性过拟合与有害过拟合之间清晰的相变。我们的结果从理论上说明了良性过拟合在 ReLU 神经网络中如何以及何时发生。未来工作的一个重要方向是将我们的分析推广到学习其他数据模型的深层 ReLU 神经网络。

致谢

我们感谢匿名评审提出的宝贵意见。YK、ZC、YC 和 QG 的部分研究经费来自美国国家科学基金会 CAREER 奖 1906169 和 IIS-2008981 项目，以及斯隆研究奖学金。本文所述观点和结论属于作者，不应被解读为代表任何资助机构的立场。

参考文献

Adlam, B. 和 Pennington, J. 高维下的神经切线核：三重下降与多尺度泛化理论。载于国际机器学习大会，第 74–84 页。PMLR，2020。

Allen-Zhu, Z. 和 Li, Y. 特征净化：对抗训练如何实现鲁棒深度学习。*arXiv* 预印本 *arXiv:2005.10190*，2020a。

Allen-Zhu, Z. 和 Li, Y. 致力于理解深度学习中的集成、知识蒸馏和自蒸馏。*arXiv* 预印本 *arXiv:2012.09816*，2020b。

Bartlett, P. L., Foster, D. J. 和 Telgarsky, M. J. 神经网络的谱归一化间隔界。载于 神经信息处理系统进展，第 6240–6249 页，2017年。

Bartlett, P. L., Long, P. M., Lugosi, G. 和 Tsigler, A. 线性回归中的良性过拟合。美国国家科学院院刊，2020。

Belkin, M., Ma, S. 和 Mandal, S. 要理解深度学习，我们需要理解核学习。载于 国际机器学习会议，第 540–548 页，2018年。

Belkin, M., Hsu, D., Ma, S. 和 Mandal, S. 调和现代机器学习实践与经典的偏差–方差权衡。美国国家科学院院刊，116(32):15849–15854，2019年。

Belkin, M.、Hsu, D. 和 Xu, J. 弱特征的双重下降的两种模型。*SIAM* 数据科学数学期刊，2(4)：1167–1180，2020。

Cao, Y.、Gu, Q. 和 Belkin, M. 子高斯混合上的过参数化最大间隔分类的风险界。神经信息处理系统进展，34，2021年。

Cao, Y.、Chen, Z.、Belkin, M. 和 Gu, Q. 两层卷积神经网络中的良性过拟合。*arXiv* 预印本 *arXiv:2202.06526*，2022年。

Chatterji, N. S. 和 Long, P. M. 过参数化情形下插值线性分类器的有限样本分析。机器学习研究杂志，22:129– 1，2021年。

Chatterji, N. S. 和 Long, P. M. 当浅层网络能够良性过拟合时，深度线性网络也可以良性过拟合。*arXiv* 预印本 *arXiv:2209.09315*，2022年。

Chen, Y.、Jin, C. 和 Yu, B. 迭代优化算法的稳定性与收敛性权衡。*arXiv* 预印本 *arXiv:1804.01619*，2018年。

Devroye, L.、Mehrabian, A. 和 Reddad, T. 高维高斯分布之间的全变差距离。*arXiv* 预印本 *arXiv:1810.08693*，2018年。

Frei, S. 和 Gu, Q. 代理凸性：分析经梯度下降训练的神经网络的统一框架。神经信息处理系统进展，34:7937–7949，2021年。

Frei, S.、Cao, Y. 和 Gu, Q. 在存在对抗性标签噪声的情况下，对任意宽度的经随机梯度下降训练的神经网络进行可证明的泛化分析。载于国际机器学习大会，第 3427–3438 页。PMLR，2021年。

Frei, S., Chatterji, N. S., and Bartlett, P. 无线性条件下的良性过拟合：针对含噪线性数据经梯度下降训练的神经网络分类器。载于学习理论会议，第 2668–2703 页。PMLR，2022年。

Hardt, M., Recht, B., and Singer, Y. 训练更快，泛化更好：随机梯度下降的稳定性。载于第 *33* 届国际机器学习会议论文集（第 *48* 卷），第 1225–1234 页。JMLR. org，2016。

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. 高维无正则最小二乘插值中的意外发现。统计年刊，50(2):949–986，2022年。

Jacot, A., Gabriel, F., and Hongler, C. 神经切线核：神经网络中的收敛与泛化。载于神经信息处理系统进展，第 8571–8580 页，2018年。

Li, Y.、Wei, C. 和 Ma, T. 试图解释初始大学习率在神经网络训练中的正则化效应。载于神经信息处理系统进展，第 11669–11680 页，2019年。

Li, Z.、Zhou, Z.-H. 和 Gretton, A. 试图理解神经网络中的良性过拟合。*arXiv* 预印本 *arXiv:2106.03212*，2021年。

Liang, T. 和 Rakhlin, A. 只需插值：核“无岭”回归也能实现泛化。统计年刊，48(3):1329–1347，2020。

Liang, T.、Rakhlin, A. 和 Zhai, X. 关于最小范数插值解的多重下降现象与核的受限下等距性。载于学习理论会议，第 2683–2711 页。PMLR，2020。Lin, M.、Chen, Q. 和 Yan, S. 网络中的网络。*arXiv* 预印本 *arXiv:1312.4400*，2013年。Montanari, A. 和 Zhong, Y. 神经网络中的插值相变：惰性训练下的记忆与泛化。统计年刊，50(5):2816–2847，2022年。

Mou, W.、Wang, L.、Zhai, X. 和 Zheng, K. 非凸学习中随机梯度朗之万动力学的泛化界：两种理论视角。*arXiv* 预印本 *arXiv:1707.05947*，2017年。

Muthukumar, V.、Narang, A.、Subramanian, V.、Belkin, M.、Hsu, D. 和 Sahai, A. 过参数化区域中的分类与回归：损失函数是否重要？机器学习研究杂志，22(1)：10104–10172，2021年。

Neyshabur, B., Bhojanapalli, S., McAllester, D. 和 Srebro, N. 神经网络谱归一化间隔界的 PAC-贝叶斯方法。载于国际学习表征会议，2018年a。

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. 和 Srebro, N. 探讨过度参数化在神经网络泛化中的作用。*arXiv*预印本 *arXiv:1805.12076*，2018年b。

Shamir, O. 良性过拟合的隐式偏差。*arXiv* 预印本 *arXiv:2201.11489*，2022年。

Shen, R. 和 Bubeck, S. 数据增强作为特征操控：沙漠奶牛与草地奶牛的故事。*ArXiv.org*，2022年。

Vershynin, R. 高维概率：数据科学应用导论。剑桥统计与概率数学丛书。剑桥大学出版社，2018年。doi：10.1017/9781108231596。

Wang, K. 和 Thrampoulidis, C. 高斯混合的二元分类中的良性过拟合。收录于*ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and*信号*Processing (ICASSP)*，第 4030–4034 页。IEEE，2021年。

Wu, D. 和 Xu, J. 过参数化线性回归中最优加权 ℓ_2 正则化。神经信息处理系统进展，33，2020。

Zhang, C.、Bengio, S.、Hardt, M.、Recht, B. 和 Vinyals, O. 理解深度学习需要重新思考泛化。载于国际学习表征会议，2017年。

Zhang, C.、Bengio, S.、Hardt, M.、Recht, B. 和 Vinyals, O.《理解深度学习（仍然）需要重新思考泛化》。《*ACM* 通讯》，64(3)：107–115，2021年。

Zou, D.、Cao, Y.、Li, Y. 和 Gu, Q.《通过适当正则化理解 Adam 在神经网络学习中的泛化》。*arXiv* 预印本 *arXiv:2108.11371*，2021a。Zou, D.、Wu, J.、Braverman, V.、Gu, Q. 和 Kakade, S.《线性回归中恒定步长随机梯度下降的良性过拟合》。载于学习理论会议，第 4633–4635 页。PMLR，2021b。

A. 相关研究提出的条件比较

本节将通过以下两张表（表 4.1 和相关工作对参数所作条件之间的差异进行介绍（Frei 等人，2022年；Cao 等人，2022年）（表 1 和 2）。

样本数量	Frei 等人 (2022年)	$n \geq C\log(1/\delta)$
	Ours	$n \geq C\log(m/\delta)$
神经网络宽度	Frei 等人 (2022年)	-
	Ours	$m \geq C\log(n/\delta)$
维度	Frei 等人 (2022年)	$d \geq$ 最大值$\{n\|\boldsymbol{\mu}\|_2^2, n^2$ 对数$(n/\delta)\}$
	Ours	$d \geq C$最大值$\{n\sigma_p^{-2}\|\boldsymbol{\mu}\|_2^2$对数$(T^*)\ n^2$对数$(nm/\delta)$(对数$(T^*))^2\}$
信号范数	Frei 等人 (2022年)	$\|\boldsymbol{\mu}\|_2^2 \geq C\cdot\log(n/\delta)$
	Ours	$\|\boldsymbol{\mu}\|_2^2 \geq C\cdot\sigma_p^2\log(n/\delta)$
噪声率	Frei 等人 (2022年)	$p \leq 1/C$
	Ours	$p \leq 1/C$
学习率	Frei 等人 (2022年)	$\eta \leq (C$最大值$\{1, H/\sqrt{m}\}d^2)^{-1}$
	Ours	$\eta \leq \left(C\text{最大值}\left\{\sigma_p^2 d/n, \sigma_p^2 d^{3/2}/\left(n^2 m\cdot\sqrt{\text{对数}(n/\delta)}\right)\right\}\right)^{-1}$
初始化方差	Frei 等人（2022年）	$\sigma_0 \leq \eta/\sqrt{md}$
	Ours	$\sigma_0 \leq \left(C\text{最大值}\left\{\sigma_p d/\sqrt{n}, \sqrt{\text{对数}(m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\right\}\right)^{-1}$

表 *1.* 与 Frei 等人（2022年）的条件比较。 H 是 Frei 等人（2022年）所设定的 Leaky ReLU 激活函数的光滑性。在本文中， σ_p 是可视为常数的噪声尺度。

样本数量	Cao 等人（2022年）	$n = \Omega(\text{polylog}(d))$
	Ours	$n \geq C\log(m/\delta)$
神经网络宽度	Cao 等人 (2022年)	$m = \Omega(\text{polylog}(d))$
	Ours	$m \geq C\log(n/\delta)$
维度	Cao 等人 (2022年)	$d = \Omega(m^{2\vee[4/(q-2)]}n^{4\vee[(2q-2)/(q-2)]})$
	Ours	$d \geq C$最大值$\{n\sigma_p^{-2}\|\boldsymbol{\mu}\|_2^2$对数$(T^*)\ n^2$对数$(nm/\delta)$(对数$(T^*))^2\}$
信号范数	Cao 等人 (2022年)	-
	Ours	$\|\boldsymbol{\mu}\|_2^2 \geq C\cdot\sigma_p^2\log(n/\delta)$
噪声率	Cao 等人 （2022年）	$p = 0$
	Ours	$p \leq 1/C$
学习率	Cao 等人 （2022年）	$\eta \leq \widetilde{O}$（最小值$\{\|\boldsymbol{\mu}\|_2^{-2}, \sigma_p^{-2}d^{-1}\}$）
	Ours	$\eta \leq \left(C\text{最大值}\left\{\sigma_p^2 d/n, \sigma_p^2 d^{3/2}/\left(n^2 m\cdot\sqrt{\text{对数}(n/\delta)}\right)\right\}\right)^{-1}$
初始化方差	Cao 等人 (2022年)	$\sigma_0 \leq \widetilde{O}(m^{-2/(q-2)}n^{-[1/(q-2)\vee 1]})\cdot$分钟$\{(\sigma_p\sqrt{d})^{-1}, \|\boldsymbol{\mu}\|_2^{-1}\}$, $\sigma_0 \geq \widetilde{O}(nd^{-1/2})\cdot$分钟$\{(\sigma_p\sqrt{d})^{-1}, \ \|\boldsymbol{\mu}\|_2^{-1}\}$
	Ours	$\sigma_0 \leq \left(C\text{最大值}\left\{\sigma_p d/\sqrt{n}, \sqrt{\text{对数}(m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\right\}\right)^{-1}$

表*2*。 与Cao 等人 (2022年)的条件比较。 q 是在 Cao 等人 (2022年)的设定下，多项式 ReLU 激活函数的阶数。

B. 预备引理

本节给出一些关键引理，以说明数据及其随机初始化时神经网络参数的一些重要性质。

我们首先给出关于数据集 S的一些集中引理。令 $S_+ = \{i|y_i = \widehat{y}_i\}$ 和 $S_- = \{i|y_i \neq \widehat{y}_i\}$ 分别表示真实标签和翻转标签对应的数据点索引集。首先有以下引理。

引理 B.1. 给定 $\delta > 0$，以至少 $1-\delta$ 的概率，

$$\left||S_+| - (1-p)n\right| \leq \sqrt{\frac{n}{2}\log\left(\frac{4}{\delta}\right)},\ \left||S_-| - pn\right| \leq \sqrt{\frac{n}{2}\log\left(\frac{4}{\delta}\right)}.$$

引理 *B.1* 的证明。 由于 $|S_+| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i = y_i]$, $|S_-| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i \neq y_i]$, 根据霍夫丁不等式, 对于任意 $t > 0$ 有

$$\mathbb{P}\big(\big||S_+| - \mathbb{E}[|S_+|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \mathbb{P}\big(\big||S_-| - \mathbb{E}[|S_-|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big).$$

根据定义 $\mathcal{D}$中定义的数据分布 1.1, 我们有 $\mathbb{E}[S_+] = (1-p)n$, $\mathbb{E}[S_-] = pn$。令 $t = \sqrt{(n/2)\log(4/\delta)}$并应用并集界可知, 以至少 $1-\delta$ 的概率,

$$\big||S_+| - (1-p)n\big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}, \big||S_-| - pn\big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)},$$

由此完成证明。 □

接下来, 令 $S_1 = \{i|y_i = 1\}$ 和 $S_{-1} = \{i|y_i = -1\}$。我们有以下引理来刻画它们的大小 s.

引理 B.2. 假设 $\delta > 0$ 和 $n \geq 8\log(4/\delta)$。那么, 以至少 $1-\delta$ 的概率,

$$|S_1|, |S_{-1}| \in [n/4, 3n/4].$$

引理*B.2*的证明。 根据数据分布 $\mathcal{D}$ （定义 1.1中给出）, 对于 $(\mathbf{x}\ y) \sim \mathcal{D}$, 我们有

$$\begin{aligned}\mathbb{P}(y=1) &= \mathbb{P}(\widehat{y}=1)\times\mathbb{P}(y=\widehat{y}) + \mathbb{P}(\widehat{y}=-1)\times\mathbb{P}(y=-\widehat{y})\\ &= \frac{1}{2}(1-p) + \frac{1}{2}p\\ &= \frac{1}{2},\end{aligned}$$

因此 $\mathbb{P}(y=-1) = 1/2$ 也成立。由于 $|S_1| = \sum_{i=1}^n \mathbb{1}[y_i = 1]$, $|S_{-1}| = \sum_{i=1}^n \mathbb{1}[y_i = -1]$, 我们有 $\mathbb{E}[|S_1|] = \mathbb{E}[|S_{-1}|] = n/2$。根据霍夫丁不等式, 对于任意 $t > 0$, 以下结论成立:

$$\begin{aligned}\mathbb{P}\big(\big||S_1| - \mathbb{E}[|S_1|]\big| \geq t\big) &\leq 2\exp\Big(-\frac{2t^2}{n}\Big),\\ \mathbb{P}\big(\big||S_{-1}| - \mathbb{E}[|S_{-1}|]\big| \geq t\big) &\leq 2\exp\Big(-\frac{2t^2}{n}\Big).\end{aligned}$$

令 $t = \sqrt{(n/2)\log(4/\delta)}$, 并取并集界, 可得以至少 $1-\delta$ 的概率,

$$\Big||S_1| - \frac{n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}, \Big||S_{-1}| - \frac{n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}.$$

因此, 只要 $n \geq 8\log(4/\delta)$, 我们就有 $\sqrt{n\log(4/\delta)/2} \leq n/4$, 从而 $3n/4 \geq |S_1|$, $|S_{-1}| \geq n/4$。

□

引理 B.3。 对于 $|S_+ \cap S_y|$ 和 $|S_- \cap S_y|$, 其中 $y \in \{\pm 1\}$, 以至少 $1-\delta$ （$\delta > 0$） 的概率有

$$\Big||S_+ \cap S_y| - \frac{(1-p)n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \Big||S_- \cap S_y| - \frac{pn}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \forall\, y \in \{\pm 1\}.$$

证明。 由于 $|S_+ \cap S_y| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i = y_i = y]$, $|S_- \cap S_y| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i \neq y_i, y_i = y]$, 根据霍夫丁不等式, 我们有

$$\mathbb{P}\big(\big||S_+ \cap S_y| - \mathbb{E}[|S_+ \cap S_y|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \forall\, y \in \{\pm 1\},$$

$$\mathbb{P}\big(\big||S_- \cap S_y| - \mathbb{E}[|S_- \cap S_y|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \forall\, y \in \{\pm 1\}.$$

根据定义 $\mathcal{D}$（定义 1.1），我们有 $\mathbb{E}[|S_+ \cap S_y|] = (1-p)n/2$，$\mathbb{E}[|S_- \cap S_y|] = pn/2$。以至少 $1-\delta$ 的概率可得

$$\Big||S_+ \cap S_y| - \frac{(1-p)n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \Big||S_- \cap S_y| - \frac{pn}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \forall\, y \in \{\pm 1\},$$

由此完成证明。 □

下面的引理估计噪声向量 $\boldsymbol{\xi}_i$、 $i \in [n]$的范数，并给出它们彼此之间以及与信号向量 $\boldsymbol{\mu}$的内积上界。

引理 B.4. 假设 $\delta > 0$ 和 $d = \Omega(\log(6n/\delta))$。那么，以至少 $1-$ 的概率 δ,

$$\begin{aligned}
&\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2, \\
&|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(6n^2/\delta)}, \\
&|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu} \rangle| \leq \|\boldsymbol{\mu}\|_2 \sigma_p \cdot \sqrt{2\log(6n/\delta)}
\end{aligned}$$

对所有 i、 $i' \in [n]$均成立。

引理 *B.4* 的证明。 根据伯恩斯坦不等式，以至少 $1-\delta/(3n)$ 的概率，我们有

$$\big|\|\boldsymbol{\xi}_i\|_2^2 - \sigma_p^2 d\big| = O(\sigma_p^2 \cdot \sqrt{d\log(6n/\delta)}).$$

因此，若我们适当设定 $d = \Omega(\log(6n/\delta))$，即可得到

$$\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2.$$

此外，显然 $\langle \boldsymbol{\xi}_i,\ \boldsymbol{\xi}_{i'} \rangle$ 的均值为零。对于任意 i, i'，若满足 $i \neq i'$，根据伯恩斯坦不等式，以至少$1-\delta/(3n^2)$ 的概率，我们有

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(6n^2/\delta)}.$$

最后，注意 $\langle \boldsymbol{\xi}_i,\ \boldsymbol{\mu} \rangle \sim \mathcal{N}(0,\ \|\boldsymbol{\mu}\|_2^2\sigma_p^2)$。根据高斯尾界，以至少 $1-\delta/3n$ 的概率，我们有

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu} \rangle| \leq \|\boldsymbol{\mu}\|_2 \sigma_p \cdot \sqrt{2\log(6n/\delta)}.$$

应用并集界即可完成证明。 □

现在转向网络初始化，以下引理研究随机初始化的 CNN 卷积滤波器 $\mathbf{w}_{j\,r}^{(0)}$ ，($j \in \{\pm 1\}$ 和 $r \in [m]$) 与训练数据中的信号/噪声向量之间的内积。这些计算刻画了初始化时的神经网络如何随机捕获信号与噪声信息。

引理 B.5. 假设 $d = \Omega(\log(mn/\delta))$, $m = \Omega(\log(1/\delta))$。那么，以至少1的概率 $-\delta$,

$$\begin{aligned}
&\sigma_0^2 d/2 \leq \|\mathbf{w}_{j,r}^{(0)}\|_2^2 \leq 3\sigma_0^2 d/2, \\
&|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2, \\
&|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \leq 2\sqrt{\log(12mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}
\end{aligned}$$

对所有 $r \in [m]$、 $j \in \{\pm 1\}$ 和 $i \in [n]$均成立。此外，

$$\sigma_0 \|\boldsymbol{\mu}\|_2/2 \leq \max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2,$$

$$\sigma_0\sigma_p\sqrt{d}/4 \le \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \le 2\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}$$

对所有 $j\in\{\pm 1\}$ 和 $i\in[n]$均成立。

引理 *B.5* 的证明。 首先，初始权重 $\mathbf{w}_{j,r}^{(0)}\sim\mathcal{N}(\mathbf{0},\sigma_0\mathbf{I})$。根据伯恩斯坦不等式，以至少$1-\delta/(6m)$的概率，有

$$\big|\|\mathbf{w}_{j,r}^{(0)}\|_2^2-\sigma_0^2 d\big| = O(\sigma_0^2\cdot\sqrt{d\log(12m/\delta)}).$$

因此，若适当设置 $d=\Omega(\log(mn/\delta))$，则以至少 $1-\delta/3$ 的概率，对所有 $j\in\{\pm1\}$ 和$r\in[m]$,

$$\sigma_0^2 d/2 \le \|\mathbf{w}_{j,r}^{(0)}\|_2^2 \le 3\sigma_0^2 d/2.$$

接下来，显然，对于每个 $r\in[m]$， $j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle$ 都是均值为零、方差为 $\sigma_0^2\|\boldsymbol{\mu}\|_2^2$ 的高斯随机变量。因此，根据高斯尾界和并集界，以至少 $1-\delta/6$ 的概率，对所有 $j\in\{\pm1\}$ 和 $r\in[m]$,

$$j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle \le |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \le \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2.$$

此外， $\mathbb{P}(\sigma_0\|\boldsymbol{\mu}\|_2/2 > j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle)$是一个绝对常数，因此在条件 $m=\Omega(\log(1/\delta))$ 下，我们有

$$\begin{aligned}\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 \le \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle\big) &= 1-\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 > \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle\big)\\ &= 1-\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 > j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle\big)^{2m}\\ &\ge 1-\delta/6,\end{aligned}$$

因此，以至少 $1-\delta/3$ 的概率，有 $\sigma_0\|\boldsymbol{\mu}\|_2/2 \le \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j\,r\,\cdot\,,}^{(0)}\ ,\ \boldsymbol{\mu}\rangle \le \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2$。

最后，根据引理 B.4 的结果，对于所有 $i\in[n]$，我们有$\sigma_p\sqrt{d}/\sqrt{2} \le \|\boldsymbol{\xi}_i\|_2 \le \sqrt{3}/2\cdot\sigma_p\sqrt{d}$。因此，按照与 $j\cdot\langle \mathbf{w}_{j\,r\,\cdot\,,}^{(0)}\ ,\ \boldsymbol{\mu}\rangle$ 相同的证明思路，以至少 $1-\delta/3$ 的概率，我们可以得到 $\langle \mathbf{w}_{j\,r\,\cdot\,,}^{(0)}\ ,\ \boldsymbol{\xi}_i\rangle$ 的结果。 □

接下来，我们将 $S_i^{(0)}$ 记为 $\{r\in[m]:\ \langle \mathbf{w}_{y_i\,r\,\cdot\,,}^{(0)}\ r,\ \boldsymbol{\xi}_i\rangle > 0\}$，并将 $S_{j\,r\,\cdot\,,}^{(t)}$ 记为 $\{i\in[n]:\ y_i=j,\ \langle \mathbf{w}_{j\,r\,\cdot\,,}^{(t)}\ ,\ \boldsymbol{\xi}_i\rangle > 0\}, j\in\{\pm1\}, r\in[m]$。我们将在下面的两个引理中给出 $|S_i^{(0)}|$ 和 $|S_{j\,\cdot\,,\,r}^{(0)}|$ 的下界。

引理 B.6. 假设 $\delta>0$ 和 $m\ge 50\log(2n/\delta)$。那么，以至少 $1-\delta$ 的概率，

$$|S_i^{(0)}| \ge 0.4m,\ \forall i\in[n].$$

引理 *B.6*的证明。 注意到 $|S_i^{(0)}| = \sum_{r=1}^m \mathbb{1}[\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle > 0]$ 和 $P(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle > 0) = 1/2$，那么根据霍夫丁不等式，以至少 $1-\delta/n$ 的概率，我们有

$$\left|\frac{|S_i^{(0)}|}{m}-\frac{1}{2}\right| \le \sqrt{\frac{\log(2n/\delta)}{2m}}.$$

因此，只要 $m\ge 50\log(2n/\delta)$，应用并集界可得，以至少 $1-\delta$ 的概率，我们有

$$|S_i^{(0)}| \ge 0.4m,\ \forall i\in[n].$$

□

引理 B.7. 假设 $\delta > 0$ 且 $n \geq 32\log(4m/\delta)$。则以至少 $1-$ 的概率 δ,

$$|S_{j,r}^{(0)}| \geq n/8, \forall j \in \{\pm 1\}, r \in [m].$$

引理的证明 *B.7*。 注意， $|S_{j}^{(0)}{}_{,r}| = \sum_{i=1}^{n} \mathbb{1}[y_i = j]\,\mathbb{1}[\langle \mathbf{w}_{j\,r}^{(0)}{}_{,}, \boldsymbol{\xi}_i\rangle > 0]$ 以及 $\mathbb{P}(y_i = j,\ \langle \mathbf{w}_{j\,r}^{(0)}{}_{,}, \boldsymbol{\xi}_i\rangle > 0) = 1/4$，那么根据霍夫丁不等式，以至少 $1 - \delta/2m$ 的概率，我们有

$$\big||S_{j,r}^{(0)}|/n - 1/4\big| \leq \sqrt{\frac{\log(4m/\delta)}{2n}}.$$

因此， 只要 $n \geq 32\log(4m/\delta)$，通过应用并集界，我们就以至少 $1-$ 的概率有 δ,

$$|S_{j,r}^{(0)}| \geq n/8, \forall j \in \{\pm 1\}, r \in [m].$$

□

C. 信号-噪声分解系数分析

在本节中，我们将建立一系列关于信号-噪声分解的结果。这些结果基于附录 B中的结论，这些结论以高概率成立。记 $\mathcal{E}_{\text{prelim}}$ 为附录 B 中的所有结果均成立这一事件（对于给定的 δ，我们通过 $\mathbb{P}(\mathcal{E}_{\text{prelim}}) \geq 1 - 7\delta$ 并集界可知）。为简单明了起见，在本节及后续各节中，我们陈述的所有结果均以 $\mathcal{E}_{\text{prelim}}$为条件。

C.1. 分解系数的迭代表达式

我们首先分析定义 5.1 中信号-噪声分解的系数。第一个引理给出了系数的迭代表达式。

引理 C.1. （引理 *5.3* 的重述）定义 $\gamma_{j\,r}^{(t)}$、、 $\rho_{j\,r}^{(t)}$、、$_{i}$、 $\underline{\rho}_{j\,r}^{(t)}$、$\underline{\ }_{i}$ 中定义的系数满足以下迭代方程：*5.1*

$$\begin{aligned}
&\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0,\\
&\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \Bigg[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) \Bigg] \cdot \|\boldsymbol{\mu}\|_2^2,\\
&\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = j),\\
&\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = -j),
\end{aligned}$$

对所有 $r \in [m]$、 $j \in \{\pm 1\}$ 和 $i \in [n]$ 均成立。

引理 *C.1* 的证明。首先，我们将梯度下降更新规则（3.1） t 迭代次，得到

$$\begin{aligned}
\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(0)} &- \frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i=1}^{n} \ell_i'^{(s)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i\rangle) \cdot j y_i \boldsymbol{\xi}_i\\
&- \frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i=1}^{n} \ell_i'^{(s)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(s)}, \widehat{y}_i \boldsymbol{\mu}\rangle) \cdot \widehat{y}_i y_i j \boldsymbol{\mu}.
\end{aligned}$$

根据 $\gamma_{j\,r}^{(t)}$、和 $\rho_{j\,r}^{(t)}{}_{,,i}$的定义，

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

注意，$\boldsymbol{\xi}_i$ 和 $\boldsymbol{\mu}$ 是线　以概率 1 线性无关，在此条件下我们得到唯一表示

$$\gamma_{j,r}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i=1}^{n}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2\cdot\widehat{y}_i y_i,$$

$$\rho_{j,r,i}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot j y_i.$$

回顾 $S_+ = \{i|y_i = \widehat{y}_i\}$、$S_- = \{i|y_i \neq \widehat{y}_i\}$，我们可以进一步写成

$$\gamma_{j,r}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i\in S_+}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2 + \frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i\in S_-}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2. \tag{C.1}$$

现在，利用记号 $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)}\geq 0)$、$\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)}\leq 0)$ 以及事实 $\ell_i'^{(s)} < 0$，我们得到

$$\overline{\rho}_{j,r,i}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = j), \tag{C.2}$$

$$\underline{\rho}_{j,r,i}^{(t)} = \frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = -j). \tag{C.3}$$

写出（C.1）、（C.2）和（C.3）的迭代形式即可完成证明。 □

C.2. 分解系数的尺度

本节余下部分将用于证明以下命题 C.2，该命题表明，信号-噪声分解中的系数将在相当长的一段时间内保持在合理范围内。考虑训练周期 $0\leq t\leq T^*$，其中 $T^* = \eta^{-1}\mathrm{poly}(\epsilon^{-1},\ d,\ n,\ m)$是定理 4.2中定义的最大允许迭代次数。现在记

$$\alpha := 4\log(T^*), \tag{C.4}$$

$$\beta := 2\max_{i,j,r}\{|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle|, |\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle|\}, \tag{C.5}$$

$$\mathrm{SNR} := \|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d}). \tag{C.6}$$

根据引理 B.5，β 可以由 $4\sigma_0\cdot\max\{\sqrt{\log(12mn/\delta)}\cdot\sigma_p\sqrt{d},\ \sqrt{\log(12m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\}$界定。然后，根据条件 4.1，通过选取足够大的常数 C，可以直接验证以下不等式:

$$\max\left\{\beta, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha\right\}\leq\frac{1}{12}. \tag{C.7}$$

命题 C.2。（命题 5.2的部分重述）在条件 4.1下，对于 $0\leq t\leq T^*$，我们有

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0 \tag{C.8}$$

$$0\leq\overline{\rho}_{j,r,i}^{(t)}\leq\alpha, \tag{C.9}$$

$$0\geq\underline{\rho}_{j,r,i}^{(t)}\geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha\geq -\alpha, \tag{C.10}$$

并且存在一个正常数 C'，使得

$$0\leq\gamma_{j,r}^{(t)}\leq C'\widehat{\gamma}\alpha, \tag{C.11}$$

对于所有 $r \in [m]$， $j \in \{\pm 1\}$ 和 $i \in [n]$，其中 $\widehat{\gamma} := n \cdot$ 信噪比^2。此外， $\gamma_{j,r}^{(t)}$ 对于 $0 \le t \le T^*$是非递减的 .

我们将使用归纳法证明命题 C.2。首先，我们介绍若干技术引理（引理C.3、C.4 和C.5），这些引理将用于命题C.2的归纳证明。

引理 C.3。 在条件 *4.1*下，假设 (C.9)、(C.10) 和 (C.11) 在迭代 t处成立。那么，对于所有 $r \in [m]$， $j \in \{\pm 1\}$和 $i \in [n]$,

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle - j \cdot \gamma_{j,r}^{(t)}\big| \le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, \tag{C.12}$$

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle - \underline{\rho}_{j,r,i}^{(t)}\big| \le 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j \ne y_i, \tag{C.13}$$

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle - \overline{\rho}_{j,r,i}^{(t)}\big| \le 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j = y_i. \tag{C.14}$$

引理 *C.3*的证明。 首先，对于任意时刻 $t \ge 0$，根据信号-噪声分解（5.1）可得

$$\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle = j \cdot \gamma_{j,r}^{(t)} + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle$$

根据引理B.4，我们有

$$\begin{aligned}
&\bigg|\sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle\bigg| \\
&\le \sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}| \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu} \rangle| \\
&\le \frac{2\|\boldsymbol{\mu}\|_2\sqrt{2\log(6n/\delta)}}{\sigma_p d}\bigg(\sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&= \mathrm{SNR}\sqrt{\frac{8\log(6n/\delta)}{d}}\bigg(\sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha,
\end{aligned}$$

其中，第一个不等式由三角不等式得到，第二个不等式由引理B.4得到，等式由信噪比 $= \|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$的定义得到，最后一个不等式由（C.9）、（C.10）得到。因此

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle - j \cdot \gamma_{j,r}^{(t)}\big| \le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha.$$

其次，对于 $j \ne y_i$ 和任意 $t \ge 0$，我们有 $\rho_{j,r,i}^{(t)} = 0$，因此

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle &= j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \underline{\rho}_{j,r,i}^{(t)} + j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle + \sum_{i' \ne i} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle + \sum_{i' \ne i} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle.
\end{aligned}$$

现在我们来看

$$\bigg|j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle + \sum_{i' \ne i} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle + \sum_{i' \ne i} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle\bigg|$$

$$
\begin{aligned}
&\leq \gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot|\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle| + \sum_{i'\neq i}(|\underline{\rho}_{j,r,i'}^{(t)}| + |\overline{\rho}_{j,r,i'}^{(t)}|)\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot|\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle| \\
&\leq \gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-1}\sigma_p\sqrt{2\log(6n/\delta)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&= \mathrm{SNR}^{-1}\sqrt{\frac{2\log(6n/\delta)}{d}}\gamma_{j,r}^{(t)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\leq \mathrm{SNR}\sqrt{\frac{8C'^2\log(6n/\delta)}{d}}n\alpha + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}
$$

其中，第一个不等式由三角不等式和 $\gamma_{j,r}^{(t)} \geq 0$得出；第二个不等式由引理 B.4得出；等式根据 信噪比 $= \|\boldsymbol{\mu}\|_2/\sigma_p\sqrt{d}$的定义成立；倒数第二个不等式由(C.10) 和 (C.11) 得出；最后一个不等式由信噪比 $\leq 1/\sqrt{8C'^2}$ 得出。因此，对于 $j \neq y_i$

$$
\big|\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle - \underline{\rho}_{j,r,i}^{(t)}\big| \leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha.
$$

类似地，对于 $y_i = j$，我们有 $\underline{\rho}_{j,r,i}^{(t)} = 0$且

$$
\begin{aligned}
\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle &= j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle \\
&= \overline{\rho}_{j,r,i}^{(t)} + j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle,
\end{aligned}
$$

并且

$$
\begin{aligned}
&\bigg|j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle\bigg| \\
&\leq \mathrm{SNR}^{-1}\sqrt{\frac{2\log(6n/\delta)}{d}}\gamma_{j,r}^{(t)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\leq \mathrm{SNR}\sqrt{\frac{8C'^2\log(6n/\delta)}{d}}n\alpha + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}
$$

其中倒数第二个不等式由（C.9）、（C.11得出；最后一个不等式由信噪比 $\leq 1/\sqrt{8C'^2}$得出。因此，对于 $j = y_i$

$$
\big|\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle - \overline{\rho}_{j,r,i}^{(t)}\big| \leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
$$

由此完成证明。 □

引理 C.4。 在条件 *4.1*下，假设 (C.9)、(C.10) 和 (C.11) 在第 t次迭代时成立。那么，对于所有 $j \neq y_i$, $j \in \{\pm 1\}$以及 $i \in [n]$, $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 0.5$。

引理的证明 *C.4*。 根据引理 C.3，我们有

$$\begin{aligned}
F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)] \\
&\le 5\max\left\{|\langle \mathbf{w}_{j,r}^{(0)}, \widehat{y}_i\boldsymbol{\mu}\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha, C'\widehat{\gamma}\alpha\right\} \\
&\le 5\max\left\{\beta, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha, C'\widehat{\gamma}\alpha\right\} \\
&< 0.5,
\end{aligned}$$

其中，第一个不等式由（C.12）、（C.13）和（C.14）得出；第二个不等式根据 β 的定义得出；第三个不等式由（C.7）得出。 □

引理 C.5。 在条件 *4.1* 下，假设 (C.9)、(C.10) 和 (C.11) 在第 t 次迭代时成立。则有

$$\begin{aligned}
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \ge -0.25, \\
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \le \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \le \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25,
\end{aligned}$$

对任意 $i \in [n]$，均成立。

引理 *C.5* 的证明。根据引理 C.14）C.3 中的（，可得

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle &\ge \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{y_i,r,i}^{(t)} - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha \\
&\ge -\beta - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha \\
&\ge -0.25,
\end{aligned}$$

其中，第二个不等式由 $\rho_{y_i,r,i}^{(t)} \ge 0$ 得出，第三个不等式由 $\beta < 1/8$ 和 $5n\sqrt{\log(4n^2/\delta)/d}\cdot\alpha < 1/8$ 得出。对于第二个不等式，左侧自然成立，因为 $z \le \sigma(z)$。对于右侧，若 $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \le 0$，则

$$\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) = 0 \le \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25.$$

若 $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0$，则

$$\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) = \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle < \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25.$$

□

接下来，我们给出一个重要的引理 C.7，它保证不同 $\ell_i'^{(t)}$ 对应的 $i \in [n]$ 的 logit 是平衡的。正如我们稍后将看到的，这保证了系数 $\gamma_{j,r}^{(t)}$ 尽管存在标签翻转噪声，仍关于 t 单调递增，这对命题C.2的证明至关重要。为此，我们先给出一个补充引理。

引理 C.6。 令 $g(z) = \ell'(z) = -1/(1+\exp(z))$，则对于所有满足 $c \ge 0$ 的 $z_2 - c \ge z_1 \ge -1$，我们有 *t*

$$\frac{\exp(c)}{4} \le \frac{g(z_1)}{g(z_2)} \le \exp(c).$$

引理的证明 *C.6*。 一方面，我们有

$$\frac{1+\exp(z_2)}{1+\exp(z_1)} \le \max\{1, \exp(z_2 - z_1)\} = \exp(c),$$

另一方面，我们有

$$\frac{1+\exp(z_2)}{1+\exp(z_1)} = \frac{\exp(-z_1)+\exp(z_2-z_1)}{\exp(-z_1)+1} \geq \frac{\exp(-z_1)+\exp(c)}{\exp(-z_1)+1} \geq \frac{\exp(1)+\exp(c)}{\exp(1)+1} \geq \frac{\exp(c)}{4}.$$

□

引理 C.7。 在条件 *4.1*下，假设 （C.9）、 （C.10）和 （C.11）对任意迭代 $t' \leq t$均成立。那么，对于任意迭代 $t' \leq t$，以下条件成立：

1. $\sum_{r=1}^{m}\big[\rho_{y_i,r,i}^{(t')} - \rho_{y_k,r,k}^{(t')}\big] \leq \kappa$ 对所有 i、$k \in [n]$均成立。
2. $y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) \leq C_1$对所有 i、$k \in [n]$均成立，
3. $\ell'^{(t')}_i / \ell'^{(t')}_k \leq C_2 = \exp(C_1)$ 对所有 i、$k \in [n]$均成立。
4. $S_i^{(0)} \subseteq S_i^{(t')}$, *where* $S_i^{(t')} := \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle > 0\}$, *and hence* $|S_i^{(t')}| \geq 0.4m$ *for all* $i \in [n]$.
5. $S_{j,r}^{(0)} \subseteq S_{j,r}^{(t')}$, *where* $S_{j,r}^{(t')} := \{i \in [n] : y_i = j, \langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_i \rangle > 0\}$, *and hence* $|S_{j,r}^{(t')}| \geq n/8$ *for all* $j \in \{\pm 1\}, r \in [m]$.

这里我们分别取 κ 和 C_1 为 3。25 和 5 。

引理的证明 *C.7*。 我们通过归纳法证明该引理。当 $t' = 0$ 时，第四和第五个条件自然成立，因此我们只需验证前三个假设。根据（C.8），我们有 $\rho_{j,r,i}^{(0)} = 0$，对任意 j, r, i 均成立，因此$\sum_{r=1}^{m}\big[\rho_{y_i,r,i}^{(0)} - \rho_{y_k,r,k}^{(0)}\big] = 0$ 对所有 i, $k \in [n]$都成立，故第一个条件对 $t' = 0$ 成立。对于第二个条件，对任意 $i, k \in [n]$

$$\begin{aligned}
&y_i \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_k) \\
&= F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^{(0)}, \mathbf{x}_i) + F_{-y_k}(\mathbf{W}_{-y_k}^{(0)}, \mathbf{x}_i) - F_{y_k}(\mathbf{W}_{y_k}^{(0)}, \mathbf{x}_i) \\
&\leq F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) + F_{-y_k}(\mathbf{W}_{-y_k}^{(0)}, \mathbf{x}_i) \\
&= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \widehat{y}_i \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle)] + \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{-y_k,r}^{(t)}, \widehat{y}_k \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{-y_k,r}^{(t)}, \boldsymbol{\xi}_i \rangle)] \\
&\leq 2\beta \leq 1/3 \leq C_1,
\end{aligned}$$

其中，第一个不等式是因为 $F_j(\mathbf{W}_j^{(0)}, \mathbf{x}_i) > 0$ 对所有 $i \in [n], j \in [m]$ 都成立；第二个不等式根据 β 在（C.5）中的定义得出，而第三个不等式则由（C.7）得出。最后，利用第二个条件，第三个条件可由以下推得

$$\frac{\ell_i'^{(0)}}{\ell_k'^{(0)}} \leq \exp\big(y_k \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_i)\big) \leq \exp(C_1),$$

根据引理C.6。

现在假设存在 $t \leq t$，使得对于任意 $0 \leq t' \leq t-1$这五个条件都成立。我们旨在证明这些条件对于 $t' = t$同样成立。

我们首先证明，对于任意 $0 \leq t' \leq t$， $y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k)$可以由 $\sum_{r=1}^{m}\big[\rho_{y_i,r,i}^{(t')} - \rho_{y_k,r,k}^{(t')}\big]$

其近似误差为一个较小的常数。我们首先将其展开为

$$
\begin{aligned}
& y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) \\
&= y_i \sum_{j\in\{\pm 1\}} j \cdot F_j(\mathbf{W}_j^{(t')}, \mathbf{x}_i) - y_k \sum_{j\in\{\pm 1\}} j \cdot F_j(\mathbf{W}_j^{(t')}, \mathbf{x}_k) \\
&= F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) + F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_k}(\mathbf{W}_{y_k}^{(t')}, \mathbf{x}_k) \\
&= F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) + \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle)] \\
&\quad - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle)] \\
&= \underbrace{F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i)}_{I_1} + \underbrace{\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu}\rangle)]}_{I_2} \\
&\quad + \underbrace{\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle)]}_{I_3},
\end{aligned}
\tag{C.15}
$$

其中所有等式均由网络定义得出。接下来，我们逐一估计 I_1、 I_2 和 I_3。对于 $|I_1|$，根据引理C.4，我们有以下上界：

$$
|I_1| \le |F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k)| + |F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i)| = F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) + F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) \le 1. \tag{C.16}
$$

对于 $|I_2|$，我们有以下上界：

$$
\begin{aligned}
|I_2| &\le \max\left\{\frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle), \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu}\rangle)\right\} \\
&\le 3\max\left\{|\langle \mathbf{w}_{y_i,r}^{(0)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle|, |\langle \mathbf{w}_{y_k,r}^{(0)}, \widehat{y}_k \cdot \boldsymbol{\mu}\rangle|, \gamma_{j,r}^{(t')}, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha\right\} \\
&\le 3\max\left\{\beta, C'\widehat{\gamma}\alpha, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha\right\} \\
&\le 0.25,
\end{aligned}
\tag{C.17}
$$

其中第二个不等式由（C.12）得出；第二个不等式由 β 的定义和（C.11）得出；最后一个不等式由条件4.1 和（C.7）得出。

对于 I_3，我们有以下上界

$$
\begin{aligned}
I_3 &= \frac{1}{m}\sum_{r=1}^{m}\left[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle)\right] \\
&\le \frac{1}{m}\sum_{r=1}^{m}\left[\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_i\rangle - \langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle\right] + 0.25 \\
&\le \frac{1}{m}\sum_{r=1}^{m}\left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')} + 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha + 0.25\right] \\
&\le \frac{1}{m}\sum_{r=1}^{m}\left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\right] + 0.5,
\end{aligned}
\tag{C.18}
$$

其中，第一个不等式根据引理 C.5，由 $\sigma(\langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle) \le \langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle + 0.25$ 和 $\sigma(\langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle) \ge \langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle$ 得出；第二个不等式由引理 C.3 中的 (C.14) 得出；最后一个不等式根据条件 4.1，由 $5\sqrt{\log(6n^2/\delta)/dn\alpha} \le 1/8$ 得出。类似地，我们有以下下界

$$
\begin{aligned}
I_3 &= \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle)\big] \\
&\ge \frac{1}{m}\sum_{r=1}^{m}\big[\langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle - \langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle\big] - 0.25 \\
&\ge \frac{1}{m}\sum_{r=1}^{m}\bigg[\overline{\rho}^{(t')}_{y_i,r,i} - \overline{\rho}^{(t')}_{y_k,r,k} - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha - 0.25\bigg] \\
&\ge \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}^{(t')}_{y_i,r,i} - \overline{\rho}^{(t')}_{y_k,r,k}\big] - 0.5,
\end{aligned}
\tag{C.19}
$$

其中，第一个不等式根据引理 C.5 得出，因为 $\sigma(\langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle) \ge \langle \mathbf{w}^{(t')}_{y_i,r}, \boldsymbol{\xi}_i\rangle$ 和 $\sigma(\langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle) \le \langle \mathbf{w}^{(t')}_{y_k,r}, \boldsymbol{\xi}_k\rangle + 0.25$；第二个不等式由引理 C.3 中的 (C.14) 得出；最后一个不等式根据条件 4.1 得出，因为 $5\sqrt{\log(6n^2/\delta)/dn\alpha} \le 1/8$。现在，将 (C.16)–(C.18) 代入 (C.15)，得到

$$
\begin{aligned}
y_i\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) &\le |I_1| + |I_2| + I_3 \le \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}^{(t')}_{y_i,r,i} - \overline{\rho}^{(t')}_{y_k,r,k}\big] + 1.75 \\
y_i\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) &\ge -|I_1| - |I_2| + I_3 \ge \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}^{(t')}_{y_i,r,i} - \overline{\rho}^{(t')}_{y_k,r,k}\big] - 1.75,
\end{aligned}
$$

这等价于

$$
\bigg|y_i\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k\cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) - \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}^{(t')}_{y_i,r,i} - \overline{\rho}^{(t')}_{y_k,r,k}\big]\bigg| \le 1.75. \tag{C.20}
$$

由此，　我们可以看到，当第一个条件对 t' 成立时，第二个条件立即对　　t'.

接下来，我们证明第一个条件对 $t' = \widetilde{t}$ 成立。首先，我们为 $\sum_{r=1}^{m}[\rho^{(\widetilde{t})}_{y_i,r,i} - \rho^{(\widetilde{t})}_{y_k,r,k}]$ 写出一个迭代更新规则。回顾引理 5.3，可知

$$
\overline{\rho}^{(t+1)}_{j,r,i} = \overline{\rho}^{(t)}_{j,r,i} - \frac{\eta}{nm}\cdot \ell_i'^{(t)}\cdot \sigma'(\langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\xi}_i\rangle)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2
$$

对所有 $j \in \{\pm 1\}$，$r \in [m]$，$i \in [n]$，$t \in [0, T^*]$ 均成立。此外，回顾 $S_i^{(t)} = \{r \in [m]$ 的定义：$\langle \mathbf{w}^{(t)}_{y_i,r}, \boldsymbol{\xi}_i\rangle > 0\}$，可得

$$
\sum_{r=1}^{m}\big[\overline{\rho}^{(t+1)}_{y_i,r,i} - \overline{\rho}^{(t+1)}_{y_k,r,k}\big] = \sum_{r=1}^{m}\big[\overline{\rho}^{(t)}_{y_i,r,i} - \overline{\rho}^{(t)}_{y_k,r,k}\big] - \frac{\eta}{nm}\cdot\big(|S_i^{(t)}|\ell_i'^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(t)}|\ell_k'^{(t)}\cdot\|\boldsymbol{\xi}_k\|_2^2\big),
$$

对所有 i，$k \in [n]$ 和 $0 \le t \le T^*$ 均成立。现在我们考虑两种不同的情况：$\sum_{r=1}^{m}\big[\rho^{(\widetilde{t}-1)}_{y_i,r,i} - \rho^{(\widetilde{t}-1)}_{y_k,r,k}\big] \le 0.9\kappa$ 和 $\sum_{r=1}^{m}\big[\rho^{(\widetilde{t}-1)}_{y_i,r,i} - \rho^{(\widetilde{t}-1)}_{y_k,r,k}\big] > 0.9\kappa$。

当 $\sum_{r=1}^{m}\big[\rho^{(\widetilde{t}-1)}_{y_i,r,i} - \rho^{(\widetilde{t}-1)}_{y_k,r,k}\big] \le 0.9\kappa$ 时，我们有

$$
\begin{aligned}
\sum_{r=1}^{m}\big[\overline{\rho}^{(\widetilde{t})}_{y_i,r,i} - \overline{\rho}^{(\widetilde{t})}_{y_k,r,k}\big] &= \sum_{r=1}^{m}\big[\overline{\rho}^{(\widetilde{t}-1)}_{y_i,r,i} - \overline{\rho}^{(\widetilde{t}-1)}_{y_k,r,k}\big] - \frac{\eta}{nm}\cdot\Big(|S_i^{(\widetilde{t}-1)}|\ell_i'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(\widetilde{t}-1)}|\ell_k'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_k\|_2^2\Big) \\
&\le \sum_{r=1}^{m}\big[\overline{\rho}^{(\widetilde{t}-1)}_{y_i,r,i} - \overline{\rho}^{(\widetilde{t}-1)}_{y_k,r,k}\big] - \frac{\eta}{nm}\cdot|S_i^{(\widetilde{t}-1)}|\ell_i'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_i\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] + \frac{\eta}{n} \cdot \|\boldsymbol{\xi}_i\|_2^2 \\
&\leq 0.9\kappa + 0.1\kappa \\
&= \kappa,
\end{aligned}
$$

其中，第一个不等式由 $\ell_i'^{(\widetilde{t}-1)} < 0$ 得出；第二个不等式由 $|S_i^{(\widetilde{t}-1)}| \leq m$ 和 $-\ell_i'^{(\widetilde{t}-1)} < 1$ 得出；第三个不等式由本情形中的假设以及 $\eta \leq C^{-1} \cdot n\sigma_p^{-2} d^{-1}$ 条件 4.1 得出。
另一方面，当 $\sum_{r=1}^{m} \big[\rho_{y_i,r,i}^{(\widetilde{t}-1)} - \rho_{y_k,r,k}^{(\widetilde{t}-1)}\big] > 0.9\kappa$ 时，由（C.20）可得

$$
\begin{aligned}
y_i \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_k) &\geq \frac{1}{m} \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] - 1.75 \\
&\geq 0.9\kappa - 0.54\kappa \\
&= 0.36\kappa,
\end{aligned} \tag{C.21}
$$

其中，第二个不等式由 $\kappa = 3.25$给出。因此，根据引理 C.6，我们有

$$
\frac{\ell_i'^{(\widetilde{t}-1)}}{\ell_k'^{(\widetilde{t}-1)}} \leq \exp\big(y_k \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_i)\big) \leq \exp(-0.36\kappa).
$$

Since 我们有 $|S_i^{(\widetilde{t}-1)}| \leq m$ 且 $|S_k^{(\widetilde{t}-1)}| \geq 0.4m$ 根据第四个条件，可得

$$
\frac{|S_i^{(\widetilde{t}-1)}| \ell_i'^{(\widetilde{t}-1)}}{|S_k^{(\widetilde{t}-1)}| \ell_k'^{(\widetilde{t}-1)}} \leq 2.5 \exp(-0.36\kappa) < 0.8.
$$

根据引理 B.4，在事件 $\mathcal{E}_{\text{prelim}}$下，我们有

$$
\big|\|\boldsymbol{\xi}_i\|_2^2 - d \cdot \sigma_p^2\big| = O\big(\sigma_p^2 \cdot \sqrt{d \log(6n/\delta)}\big), \forall i \in [n].
$$

注意到 $d = \Omega(\log(6n/\delta))$ 由条件 4.1，可得

$$
|S_i^{(\widetilde{t}-1)}|(-\ell_i'^{(\widetilde{t}-1)}) \cdot \|\boldsymbol{\xi}_i\|_2^2 < |S_k^{(\widetilde{t}-1)}|(-\ell_k'^{(\widetilde{t}-1)}) \cdot \|\boldsymbol{\xi}_k\|_2^2.
$$

于是我们得到

$$
\sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big] \leq \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] \leq \kappa,
$$

这就完成了第 $t' = t$ 次迭代时第一个假设的证明。接下来，通过应用式（C.20）中的近似，我们可以验证第 t次迭代时的第二个假设。事实上，我们有

$$
y_i \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_k) \leq \frac{1}{m} \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big] + 1.75 \leq C_1,
$$

其中，第一个不等式由(C.20)得到；最后一个不等式由归纳假设得出，并取 κ 为 3.25，取 C_1 为 5。注意到

$$
\frac{\ell_i'^{(\widetilde{t})}}{\ell_k'^{(\widetilde{t})}} \leq \exp\big(y_k \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_i)\big) \leq \exp(C_1) = C_2.
$$

为验证第四个假设，根据梯度下降规则，我们有

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle \\
&\quad - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle \\
&\quad - \frac{\eta}{nm} \cdot \ell_i^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 - \frac{\eta}{nm} \cdot \sum_{i' \neq i} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \underbrace{\ell_i^{(\widetilde{t}-1)} \cdot \|\boldsymbol{\xi}_i\|_2^2}_{I_4} - \frac{\eta}{nm} \cdot \underbrace{\sum_{i' \neq i} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}_{I_5} \\
&\quad - \frac{\eta}{nm} \cdot \underbrace{\sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle}_{I_6},
\end{aligned}$$

对于任意 $r \in S_i^{(\widetilde{t}-1)}$，其中最后一个等号是由 $\langle \mathbf{w}_{y_i\, r}^{(\widetilde{t}-1)}$得到的. ， $\boldsymbol{\xi}_i \rangle > 0$。然后，我们分别估计 I_4， I_5， I_6。对于 I_4，根据引理 B.4，我们有

$$-I_4 \geq |\ell_i^{(\widetilde{t}-1)}| \cdot \sigma_p^2 d/2.$$

对于 I_5，我们有以下上界

$$\begin{aligned}
|I_5| &\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot 2\sigma_p^2 \cdot \sqrt{d \log(6n^2/\delta)} \\
&\leq nC_2 |\ell_i^{(\widetilde{t}-1)}| \cdot 2\sigma_p^2 \cdot \sqrt{d \log(6n^2/\delta)},
\end{aligned}$$

其中第一个不等式由三角不等式得出；第二个不等式由 $\sigma'(z) \in \{0,1\}$ 得出；第三个不等式~由引理 B.4 得出；最后一个不等式由迭代 $t-1$ 时的第三个假设得出。

对于 I_6，我们有以下上界

$$\begin{aligned}
|I_6| &\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot |\langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot |\langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \|\boldsymbol{\mu}\|_2 \sigma_p \sqrt{2 \log(6n/\delta)} \\
&\leq nC_2 |\ell_i^{(\widetilde{t}-1)}| \cdot \|\boldsymbol{\mu}\|_2 \sigma_p \sqrt{2 \log(6n/\delta)},
\end{aligned}$$

其中，第一个不等式由三角不等式得出；第二个不等式是由于 $\sigma'(z) \in \{0,1\}$；第三个不等式由引理 B.4 得出；最后一个不等式是由于迭代 $\widetilde{t}-1$ 时的第三个假设。由于 $d \geq$ 最大值$\{32C_2^2 n^2 \cdot$

$\log(6n^2/\delta)$ $4C_2n\|\boldsymbol{\mu}\|\sigma_p^{-1}\sqrt{2\log(6n/\delta)}$，我们有 $-I_4 \geq$最大值$\{|I_5|/2, |I_6|/2\}$，因此 $-I_4 \geq |I_5| + |I_6|$。由此可得

$$\langle \mathbf{w}_{y_i,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i \rangle \geq \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle > 0,$$

对于任意 $r \in S_i^{(\widetilde{t}-1)}$。因此， $S_i^{(0)} \subseteq S_i^{(\widetilde{t}-1)} \subseteq S_i^{(\widetilde{t})}$。根据引理 B.6，可直接推出 $|S_i^{(\widetilde{t})}| \geq 0.4m, \quad \forall i \in [n]$， ~这意味着第四个假设对 $t' = t$成立。对于第五个假设，类似于第四个假设的证明，我们也有

$$\langle \mathbf{w}_{j,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}_{j,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \ell_i^{(\widetilde{t}-1)} \cdot \|\boldsymbol{\xi}_i\|_2^2 - \frac{\eta}{nm} \cdot \sum_{i' \neq i} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle$$

$$- \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle$$

对于任意 $i \in S_{j\,r}^{(\widetilde{t}-1)}$，，其中等号成立是因为 $\langle \mathbf{w}_{j\,r}^{(\widetilde{t}-1)}$，，$\boldsymbol{\xi}_i \rangle > 0$ 和 $y_i = j$。应用第四个假设证明中使用的相同技巧，可得

$$\langle \mathbf{w}_{j,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i \rangle \geq \langle \mathbf{w}_{j,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle > 0,$$

对于任意 $i \in S_{j\,r}^{(\widetilde{t}-1)}$，。因此，我们有 $S_j^{(0)}$，$_r \subseteq S_j^{(\widetilde{t}-1)}$，$_r \subseteq S_{j\,r}^{(\widetilde{t})}$。由引理 B.7 可直接推出 $|S_j^{(\widetilde{t})}$，$_r| \geq n/8$，这意味着第四个假设对 $t' = \widetilde{t}$ 成立。因此，五个假设对 $t' = \widetilde{t}$ 均成立，从而完成归纳法。

□

现在我们可以证明命题C.2。

命题 *C.2* 的证明。我们的证明基于归纳法。当 $t = 0$ 时，结论显然成立，因为所有系数均为零。

~假设存在 $T \leq T^*$，使得命题 C.2 中的结论对所有时刻 $0 \leq t \leq T-1$均成立。我们要证明

~它们对 $t = T$也成立。注意，根据引理 C.7，对于任意 $0 \leq t \leq T-1$，我们也有

1. $\sum_{r=1}^{m} [\rho_{y_i,r,i}^{(t)} - \rho_{y_k,r,k}^{(t)}] \leq \kappa$ 对于所有 $i, k \in [n]$。 2. $y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_1$对于所有 $i, k \in [n]$， 3. $\ell_i'^{(t)}/\ell_k'^{(t)} \leq C_2 = \exp(C_1)$ 对于所有 $i, k \in [n]$。 4. $S_i^{(0)} \subseteq S_i^{(t)}$ 对于所有 $i \in [n]$，其中 $S_i^{(t)}$: $= \{r \in [m] : \langle \mathbf{w}_{y_i}^{(t)}, r, \boldsymbol{\xi}_i \rangle > 0\}$，因此 $|S_i^{(t)}| \geq 0.4m$，对于所有 $i \in [n]$。 5. $S_j^{(0)}$，$_r \subseteq S_{j\,r}^{(t)}$，，其中 $S_{j\,r}^{(t)}$，: $= \{i \in [n] : y_i = j, \langle \mathbf{w}_{j\,r}^{(t)}$，，$\boldsymbol{\xi}_i \rangle > 0\}$，因此 $|S_j^{(t)}$，$_r| \geq n/8$ 对于所有 $j \in \{\pm 1\}, r \in [m]$。

我们首先证明，对于 $t = \widetilde{T}$，即 $\rho_{j\,r}^{(t)}$，，$_i \geq -\beta - 10\sqrt{\log(6n^2/\delta)/d \cdot n\alpha}$，对 $t = \widetilde{T}$ 以及任意 $r \in [m]$、$j \in \{\pm 1\}$ 和 $i \in [n]$，(C.10) 成立。注意到 $\rho_{j\,r}^{(t)-}$，，$_i = 0$ 对于 $j = y_i$ 成立，因此我们只需考虑 $j \neq y_i$ 的情况。当 ~$\rho_{j\,r}^{(T-1)}$，，$_t < -0.5\beta - 5\sqrt{\log(6n^2/\delta)/d \cdot n\alpha}$ 时，根据引理 C.3 中的 (C.14)，我们有

$$\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i \rangle \leq \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + 5\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha < 0,$$

从而

$$\begin{aligned} \underline{\rho}_{j,r,i}^{(\widetilde{T})} &= \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \frac{\eta}{nm} \cdot \ell_i'^{(\widetilde{T}-1)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i \rangle \geq 0) \cdot \mathbb{1}(y_i = -j) \|\boldsymbol{\xi}_i\|_2^2 \\ &= \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} \end{aligned}$$

$$\geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,$$

其中最后一个不等式由归纳假设得到。当 $\rho_{j,r,i}^{(T-1)} \geq -0.5\beta - 5\sqrt{\log(6n^2/\delta)/d}\cdot n\alpha$ 时，我们有

$$\begin{aligned}
\rho_{j,r,i}^{(\widetilde{T})} &= \rho_{j,r,i}^{(\widetilde{T}-1)} + \frac{\eta}{nm}\cdot \ell_i^{\prime(\widetilde{T}-1)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = -j)\|\boldsymbol{\xi}_i\|_2^2 \\
&\geq -0.5\beta - 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha - \frac{3\eta\sigma_p^2 d}{2nm} \\
&\geq -0.5\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}$$

其中，第一个等号由 $\ell_i^{\prime(T-1)} \in (-1, 0)$ 以及 $\|\boldsymbol{\xi}_i\|_2^2 \leq (3/2)\sigma_p^2 d$ 根据引理 B.4得出；第二个不等式则由条件 $5\sqrt{\log(6n^2/\delta)/d}\cdot n\alpha \geq 3\eta\sigma_p^2 d/2nm$ 中关于 η 的条件得出，见条件 4.1。

接下来，我们证明（C.9）对 $t = T$成立。考虑

$$\begin{aligned}
|\ell_i^{\prime(t)}| &= \frac{1}{1+\exp\{y_i\cdot[F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i)]\}} \\
&\leq \exp\{-y_i\cdot[F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i)]\} \\
&\leq \exp\{-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) + 0.5\},
\end{aligned} \tag{C.22}$$

其中，最后一个不等式根据引理 C.4，由 $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 0.5$ 关于 $j \neq y_i$ 得出。现在回顾 $\rho_{j,r,i}^{(t)}$的迭代更新规则：

$$\rho_{j,r,i}^{(t+1)} = \rho_{j,r,i}^{(t)} - \frac{\eta}{nm}\cdot \ell_i^{\prime(t)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2.$$

令 $t_{j,r,i}$ 为最后一次时间 $t < T^*$ 使得 $\rho_{j,r,i}^{(t)} \leq 0.5\alpha$。然后，从 $t = t_{j,r,i}$ 到 $t = \widetilde{T} - 1$迭代应用更新规则，可得

$$\begin{aligned}
\rho_{j,r,i}^{(\widetilde{T})} = \rho_{j,r,i}^{(t_{j,r,i})} &\underbrace{- \frac{\eta}{nm}\cdot \ell_i^{\prime(t_{j,r,i})} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t_{j,r,i})}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2}_{I_7} \\
&\underbrace{- \sum_{t_{j,r,i}<t<\widetilde{T}} \frac{\eta}{nm}\cdot \ell_i^{\prime(t)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2}_{I_8}.
\end{aligned} \tag{C.23}$$

我们首先如下界定 I_7：

$$|I_7| \leq (\eta/nm)\cdot\|\boldsymbol{\xi}_i\|_2^2 \leq (\eta/nm)\cdot 3\sigma_p^2 d/2 \leq 1 \leq 0.25\alpha,$$

其中，第一个不等式由 $\ell^{\prime(t_{j,r,i})}{}_i \in (-1,0)$得出；第二个不等式由引理 B.4得出；第三个不等式由 $\eta \leq C^{-1}\cdot n/(\sigma_p^2 d)$ 根据条件 4.1得出；最后一个不等式由我们对 $\alpha = 4\log(T^*)$ 和 $T^* \geq e$的选择得出。

其次，我们界定 d I_8。对于 $t_{j,r,i} < t < \widetilde{T}$ 和 $y_i = j$，我们可以对内积 $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ 给出如下下界 s

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle &\geq \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \overline{\rho}_{j,r,i}^{(t)} - 5\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \\ &\geq -0.5\beta + 0.5\alpha - 5\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \\ &\geq 0.25\alpha,\end{aligned} \tag{C.24}$$

其中，第一个不等式由引理 C.3中的（C.13）得出；第二个不等式由 $\rho_{j,r,,i}^{(t)} \gtrsim 0.5\alpha$ 和 $\langle \mathbf{w}_{j,r}^{(0)}, , \boldsymbol{\xi}_i \rangle \geq -0.5\beta$根据 $t_{j,r,i}$ 和 β的定义得出；最后一个不等式由条件 4.1 中的 $\beta \leq 1/8 \leq 0.1\alpha$ 和 $5\sqrt{\log(6n^2/\delta)/d \cdot n\alpha} \leq 0.2\alpha$ by $d \geq C \cdot n^2 \log(nm/\delta)(\log\ T^*)^2$ 得出。因此，将 $\langle \mathbf{w}_{j,r}^{(t)}, , \boldsymbol{\xi}_i \rangle$ 的下界代入 I_8 可得

$$\begin{aligned}|I_8| &\leq \sum_{t_{j,r,i}<t<\widetilde{T}} \frac{\eta}{nm} \cdot \exp(-\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) + 0.5) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2 \\ &\leq \frac{2\eta(\widetilde{T} - t_{j,r,i} - 1)}{nm} \cdot \exp(-0.25\alpha) \cdot \frac{3\sigma_p^2 d}{2} \\ &\leq \frac{2\eta T^*}{nm} \cdot \exp(-\log(T^*)) \cdot \frac{3\sigma_p^2 d}{2} \\ &= \frac{2\eta}{nm} \cdot \frac{3\sigma_p^2 d}{2} \leq 1 \leq 0.25\alpha,\end{aligned}$$

其中，第一个不等式由 (C.22) 得出；第二个不等式由 (C.24) 得出；第三个不等式由 $\alpha = 4\log(T^*)$ 得出；根据条件 4.1 中关于 η和 d 的条件，第四个不等式由 $\eta \leq C^{-1} n^2 m \sqrt{\log(n/\delta)} \sigma_p^{-2} d^{-3/2} \leq nm/(3\sigma_p^2 d)$ 得出；最后一个不等式由 $\log(T^*) \geq 1$ 和 $\alpha = 4\log(T^*)$ 得出。将 I_7, I_8的界代入 (C.23)，即完成对 ρ的证明。 –

接下来，我们证明（C.11）对 $t = \widetilde{T}$ 成立。回顾 $\gamma_{j,r}^{(t)}$ 的迭代更新规则，我们有

$$\gamma_{j,r}^{(\widetilde{T})} = \gamma_{j,r}^{(\widetilde{T}-1)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \bigg] \cdot \|\boldsymbol{\mu}\|_2^2$$

~ ~ ~我们首先证明，系数 $\gamma_j^{(T)}$、$_r \geq \gamma_{j\,r}^{(T-1)}$，从而 $\gamma^{(T)}{}_{j,\ r} \geq \gamma_j^{(0)}$、$_r = 0$ 对任意 $j \in \{\pm 1\}$、$r \in [m]$均成立。回顾 $S_+ = \{i | y_i = \widehat{y}_i\}$、$S_- = \{i | y_i \neq \widehat{y}_i\}$、$S_1 = \{i | \widehat{y}_i = 1\}$ 和 $S_{-1} = \{i | \widehat{y}_i = -1\}$的定义。我们将分别考虑

~ ~ ~以下两种情形：$\langle \mathbf{w}_{j\,r}^{(T-1)}, , \boldsymbol{\mu} \rangle \geq 0$以及 $\langle \mathbf{w}^{(T-1)}{}_{j,\ r}, \boldsymbol{\mu} \rangle < 0$。如果 $\langle \mathbf{w}^{(T-1)}{}_{j,\ r}, \boldsymbol{\mu} \rangle \geq 0$，那么

$$\begin{aligned}&- \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \\ &= - \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0) \\ &= \sum_{i \in S_+ \cap S_1} |\ell_i'^{(\widetilde{T}-1)}| - \sum_{i \in S_- \cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| \\ &\geq |S_+ \cap S_1| \cdot \min_{i \in S_+ \cap S_1} |\ell_i'^{(\widetilde{T}-1)}| - |S_- \cap S_{-1}| \cdot \max_{i \in S_- \cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|,\end{aligned}$$

其中 sec ~ ~等式成立是由于 $\ell_i'^{(T-1)} < 0$。如果 $\langle \mathbf{w}^{(T-1}{}_{)j,\ r}, \boldsymbol{\mu} \rangle < 0$，那么通过类似的推理，我们有

$$- \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle)$$

$$
\begin{aligned}
&= -\sum_{i\in S_+} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle \geq 0) + \sum_{i\in S_-} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle \geq 0) \\
&= \sum_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| - \sum_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}| \\
&\geq |S_+\cap S_{-1}| \cdot \min_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| - |S_-\cap S_1| \cdot \max_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|.
\end{aligned}
$$

根据引理 B.3 以及引理 C.7 中的第三个结论，$\ell_i'^{(T-1)}/\ell_k'^{(T-1)} \leq C_2$，$\forall i,\ k\in[n]$，在事件$\mathcal{E}_{\mathrm{prelim}}$下，我们有

$$
\begin{aligned}
\frac{|S_+\cap S_1|\cdot \min_{i\in S_+\cap S_1}|\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_{-1}|\cdot \max_{i\in S_-\cap S_{-1}}|\ell_i'^{(\widetilde{T}-1)}|} &\geq \frac{|S_+\cap S_1|}{C_2|S_-\cap S_{-1}|} \geq \frac{(1-p)n-\sqrt{2n\log(8/\delta)}}{C_2\cdot(pn+\sqrt{2n\log(8/\delta)})}, \\
\frac{|S_+\cap S_{-1}|\cdot \min_{i\in S_+\cap S_{-1}}|\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_1|\cdot \max_{i\in S_-\cap S_1}|\ell_i'^{(\widetilde{T}-1)}|} &\geq \frac{|S_+\cap S_{-1}|}{C_2|S_-\cap S_1|} \geq \frac{(1-p)n-\sqrt{2n\log(8/\delta)}}{C_2\cdot(pn+\sqrt{2n\log(8/\delta)})}.
\end{aligned}
$$

只要 $p<1/[2(1+C_2)]$ 和 $n\geq 8(C_2+1)^2\log(8/\delta)$，我们就有

$$
\frac{|S_+\cap S_1|\cdot \min_{i\in S_+\cap S_1}|\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_{-1}|\cdot \max_{i\in S_-\cap S_{-1}}|\ell_i'^{(\widetilde{T}-1)}|} \geq 1,\ \frac{|S_+\cap S_{-1}|\cdot \min_{i\in S_+\cap S_{-1}}|\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_1|\cdot \max_{i\in S_-\cap S_1}|\ell_i'^{(\widetilde{T}-1)}|} \geq 1.
$$

对于两种情况，均可推出 $\langle \mathbf{w}_{j,r}^{(T-1)}, \boldsymbol{\mu}\rangle \geq 0$ 以及 $\langle \mathbf{w}^{(T-1)}{}_{j,r}, \boldsymbol{\mu}\rangle < 0$ 满足

$$
\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(\widetilde{T}-1)}, \tag{C.25}
$$

因此

$$
\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(0)} = 0.
$$

对于（C.11）的另一部分，我们证明一个加强的假设：存在一个 $i^*\in[n]$，使得 $y_{i^*}=j$ 对于$1\leq t\leq T^*$，我们有

$$
\gamma_{j,r}^{(t)}/\overline{\rho}_{j,r,i^*}^{(t)} \leq C' n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d,
$$

并且 i^* 可以取集合 $S_{j,r}^{(0)}$ 中的任意样本，而 C' 可以取为 $2C_2$。

回忆 $\gamma_{j,r}^{(t)}$ 和 $\rho_{j,r,i}^{(t)}$ 的更新规则，我们有

$$
\begin{aligned}
\gamma_{j,r}^{(\widetilde{T})} &= \gamma_{j,r}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\Bigg[\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\,\mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0) - \sum_{i\in S_-}\ell_i'^{(\widetilde{T}-1)}\,\mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0)\Bigg]\cdot\|\boldsymbol{\mu}\|_2^2, \\
\overline{\rho}_{j,r,i}^{(\widetilde{T})} &= \overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\ell_i'^{(\widetilde{T}-1)}\cdot\mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i\rangle\geq 0)\cdot\mathbb{1}(y_i=j)\|\boldsymbol{\xi}_i\|_2^2.
\end{aligned}
$$

根据引理 C.7 的第五个陈述，对于任意 $i^*\in S_{j,r}^{(0)}$，均有 $j=y_{i^*}$ 以及 $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i^*}\rangle \geq 0$ 对于任意$0\leq t\leq T-1$。因此，我们有

$$
\overline{\rho}_{j,r,i^*}^{(\widetilde{T})} = \overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\ell_{i^*}'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\xi}_{i^*}\|_2^2 \geq \overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\ell_{i^*}'^{(\widetilde{T}-1)}\cdot\sigma_p^2 d/2.
$$

~ $\gamma_{j\,r}^{(T)}$ 的更新规则，为

$$\left|\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\mathbb{1}(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0)-\sum_{i\in S_-}\ell_i'^{(\widetilde{T}-1)}\mathbb{1}(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0)\right|\leq\sum_{i=1}^{n}|\ell_i'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2$$
$$\leq C_2 n\cdot|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2,$$

其中，第一个不等式由三角不等式得出；第二个不等式由引理C.7的第三个结论得出， C_2 是一个正常数。于是，我们有

$$\frac{\gamma_{j,r}^{(\widetilde{T})}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T})}}\leq\max\left\{\frac{\gamma_{j,r}^{(\widetilde{T}-1)}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)}},\frac{C_2 n\cdot|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2}{|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\sigma_p^2 d/2}\right\}=\max\left\{\frac{\gamma_{j,r}^{(\widetilde{T}-1)}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)}},\frac{2C_2 n\|\boldsymbol{\mu}\|_2^2}{\sigma_p^2 d}\right\}\leq\frac{2C_2 n\|\boldsymbol{\mu}\|_2^2}{\sigma_p^2 d},$$

~ ~其中，最后一个不等式由 $\gamma_{j}^{(T-1)}{}_{,\,r}\quad/\rho_{j\,r}^{(T-1)}{}_{,\,,\,i^*}\leq C'\widehat{\gamma}=C'n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d$ 得出，且 C' 可取为 $2C_2$，由此完成归纳。 □

至此，我们已经证明了命题 C.2。然后，根据引理 C.7，下一个命题可直接得出。

命题 C.8。在条件 *4.1*下，对于 $0\leq t\leq T^*$，我们有

1. $\sum_{r=1}^{m}\left[\overline{\rho}_{y_i,r,i}^{(t)}-\overline{\rho}_{y_k,r,k}^{(t)}\right]\leq\kappa$ *for all* $i,k\in[n]$.
2. $y_i\cdot f(\mathbf{W}^{(t)},\mathbf{x}_i)-y_k\cdot f(\mathbf{W}^{(t)},\mathbf{x}_k)\leq C_1$ *for all* $i,k\in[n]$,
3. $\ell_i'^{(t)}/\ell_k'^{(t)}\leq C_2=\exp(C_1)$ *for all* $i,k\in[n]$.
4. $S_i^{(0)}\subseteq S_i^{(t)}$, *where* $S_i^{(t)}:=\{r\in[m]:\langle\mathbf{w}_{y_i,r}^{(t)},\boldsymbol{\xi}_i\rangle>0\}$, *and hence* $|S_i^{(t)}|\geq 0.4m$ *for all* $i\in[n]$.
5. $S_{j,r}^{(0)}\subseteq S_{j,r}^{(t)}$, *where* $S_{j,r}^{(t)}:=\{i\in[n]:y_i=j,\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle>0\}$, *and hence* $|S_{j,r}^{(t)}|\geq n/8$ *for all* $j\in\{\pm1\},r\in[m]$.

这里， κ 和 C_1 分别可取为 *3.25* 和 *5*。

D. 使用两阶段分析进行解耦

我们采用两阶段分析来解耦系数 $\gamma_{j\,r}^{(t)}{}_{,\,,}\rho_{j\,r}^{(t)}{}_{,\,,i}$ 和 $\rho_{j\,r}^{(t)}{}_{,\,,i}$ 之间复杂的关系。直观而言，初始神经网络权重足够小，因此神经网络在初始化时，对所有训练数据的交叉熵损失导数都处于常数水平： $\ell_i'^{(0)}=\ell'[y_i\cdot f(\bar{\mathbf{W}}^{(0)},\mathbf{x}_i)]=\Theta(1)$，对所有 $i\in[n]$ 均成立。受此启发，我们可以考察训练过程的第一阶段，其中 $\ell_i'^{(0)}=\Theta(1)$；在这种情况下，我们可以证明 $\gamma_{j\,r}^{(t)}{}_{,\,,}\rho_{j\,r}^{(t)}{}_{,\,,i}$ 与 $\rho_{j\,r}^{(t)}{}_{,\,,i}$ 之间存在显著的尺度差异。基于第一阶段的结果，我们接着考察训练过程的第二阶段，此时损失导数不再处于常数水平；我们证明，训练损失可以优化到任意小，同时第一学习阶段所呈现的尺度差异在整个训练过程中保持不变。回顾一下，我们记 $\alpha=4\log(T^*)$、 $\beta=2\max_{i,j,r}\{|\langle\mathbf{w}_{j\,r}^{(0)}{}_{,\,,}\boldsymbol{\mu}\rangle|,|\langle\mathbf{w}_{j\,r}^{(0)}{}_{,\,,}\boldsymbol{\xi}_i\rangle|\}$，以及信噪比 $=\|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$。提醒读者，本节中的证明基于 C 节的结果，这些结果以高概率成立。

D.1. 第一阶段

引理 D.1. 若我们记

$$n\cdot\mathrm{SNR}^2=\widehat{\gamma},$$

则存在

$$T_1=C_3\eta^{-1}nm\sigma_p^{-2}d^{-1},T_2=C_4\eta^{-1}nm\sigma_p^{-2}d^{-1}$$

其中， $C_3=\Theta(1)$是一个较大的常数，而 $C_4=\Theta(1)$ 是一个较小的常数，使得

- $\rho_{(T_1)\,j,\,r^*,\,i} \geq 2$ 对于任意 $r^* \in S_{(0)\,i} = \{r \in [m] : \langle \mathbf{w}_{(0)y_i,\,r}, \boldsymbol{\xi}_i \rangle > 0\}$、$j \in \{\pm 1\}$ 以及 $i \in [n]$，且 $y_i = j$。

- 最大值$_{j,\,r}\, \gamma_{(t)\,j,\,r} = O(\widehat{\gamma})$ 对所有 $0 \leq t \leq T_1$ 均成立。

- 最大值$_{j,\,r,\,i} |\rho_{(t)\,j,\,r,\,i}| =$ 最大值$\{O(\sqrt{\log(mn}/\delta) \cdot \sigma_0 \sigma_p \sqrt{d}),\ O(n\sqrt{\log(n/\delta)} \log(T^*)/\sqrt{d})\}$ 对所有 $0 \leq t \leq T_1$。

- 最小值$_{j,\,r}\, \gamma_{(t)\,j,\,r} = \Omega(\widehat{\gamma})$ 对所有 $t \geq T_2$ 均成立。

- 最大值$_{j,\,r}\, \rho_{(T_1)\,j,\,r,\,i} = O(1)$ 对所有 $i \in [n]$ 均成立。

引理 *D.1* 的证明。 根据命题 C.2，我们有 $\rho_{j\,r\,,,i}^{(t)} \geq -\beta - 10n\sqrt{\log(6n^2/\delta)\, d}\ \alpha$ 对所有 $j \in \{\pm 1\}$，$r \in [m]$，$i \in [n]$以及 $0 \leq t \leq T^*$。根据引理 B.5，对于 β 我们有

$$\begin{aligned}
\beta &= 2 \max_{i,j,r} \{ |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \} \\
&\leq 2 \max\{ \sqrt{2\log(12m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2, 2\sqrt{\log(12mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d} \} \\
&= O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d} \big)
\end{aligned}$$

其中，最后一个等式由条件4.1的第一个条件得出。由于$\rho_{j\,r\,,,i}^{(t)} \leq 0$，我们有

$$\begin{aligned}
\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| &= \max_{j,r,i} -\underline{\rho}_{j,r,i}^{(t)} \\
&\leq \beta + 10\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \\
&= \max \left\{ O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d} \big), O\big(\sqrt{\log(n/\delta)} \log(T^*) \cdot n/\sqrt{d} \big) \right\}.
\end{aligned}$$

接下来，对于 $\gamma_{j\,r\,,}^{(t)}$的增长，我们有以下上界

$$\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \left[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \right] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \|\boldsymbol{\mu}\|_2^2 \\
&\leq \gamma_{j,r}^{(t)} + \frac{\eta}{m} \cdot \|\boldsymbol{\mu}\|_2^2,
\end{aligned}$$

其中，不等式由 $|\ell'| \leq 1$得出。注意到 $\gamma_{j\,,r}^{(0)} = 0$，并递归地使用不等式 t 次，我们得到

$$\gamma_{j,r}^{(t)} \leq \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2. \tag{D.1}$$

由于 $n \cdot$ 信噪比$^2 = n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d = \widehat{\gamma}$，我们有

$$T_1 = C_3 \eta^{-1} nm \sigma_p^{-2} d^{-1} = C_3 \eta^{-1} m \|\boldsymbol{\mu}\|_2^{-2} \widehat{\gamma}.$$

由此可得

$$\gamma_{j,r}^{(t)} \leq \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2 \leq \frac{\eta T_1}{m} \cdot \|\boldsymbol{\mu}\|_2^2 \leq C_3 \widehat{\gamma},$$

对所有 $0 \leq t \leq T_1$。

对于 $\rho_{j,r,i}^{(t)}$，回顾(5.4)可知

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2.$$

根据命题 C.8，对于任意 $r^* \in S_i^{(0)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$，我们有 $\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle > 0$ 对所有 $0 \le t \le T^*$ 均成立，因此

$$\overline{\rho}_{y_i,r^*,i}^{(t+1)} = \overline{\rho}_{y_i,r^*,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^2.$$

注意到 $\rho_{y_i,r^*,i}^{(0)} = 0$并递归地使用该方程 t 次，我们有

$$\overline{\rho}_{y_i,r^*,i}^{(t)} = -\frac{\eta}{nm} \cdot \sum_{s=0}^{t-1} \ell_i'^{(s)} \cdot \|\boldsymbol{\xi}_i\|_2^2.$$

对于每个 i，记 $T_1^{(i)}$ 为周期 $[0, T_1]$ 内满足 $\max_{j,r} |\rho_{j,r,i}^{(t)}| \le 2$ 的最后时刻。则对于 $0 \le t \le T_1^{(i)}$，$\max_{j,r}\{|\rho_{j,r,i}^{(t)}|, |\rho_{j,r,i}^{(t)}|\} = O(1)$ 且 $\max_{j,r} \gamma_{j,r}^{(t)} = O(1)$。因此，我们知道 $F_{-1}(\mathbf{W}^{(t)}, \mathbf{x}_i)$ $F_{+1}(\mathbf{W}^{(t)}, \mathbf{x}_i) = O(1)$。因此，存在正常数 C，使得 $-\ell_i'^{(t)} \ge C$ 对于 $0 \le t \le T_1^{(i)}$ 成立。于是我们有

$$\overline{\rho}_{y_i,r^*,i}^{(t)} \ge \frac{C\eta\sigma_p^2 dt}{2nm}.$$

因此，$\rho_{y_i,r^*,i}^{(t)}$ 将在

$$T_1 = C_3 \eta^{-1} nm \sigma_p^{-2} d^{-1}$$

次迭代内达到 2，对任意 $r^* \in S_i^{(0)}$ 均成立，其中 C_3 可取为 $4/C$。

接下来，我们将讨论 $\gamma_{j,r}^{(t)}$ 增长的下界。对于 $\rho_{j,r,i}^{(t)}$，有

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2 \le \overline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \|\boldsymbol{\xi}_i\|_2^2 \le \overline{\rho}_{j,r,i}^{(t)} + \frac{3\eta\sigma_p^2 d}{2nm},$$

其中，第一个不等式由 $-\ell_i' \in (0, 1)$ 和 $\sigma' \in \{0, 1\}$ 得出；第二个不等式由引理 B.4 得出。根据（D.1）和 $\rho_{j,r,i}^{(0)} = 0$，可得

$$\overline{\rho}_{j,r,i}^{(t)} \le \frac{3\eta\sigma_p^2 dt}{2nm}, \gamma_{j,r}^{(t)} \le \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2. \tag{D.2}$$

因此，$\max_{j,r,i} \rho_{j,r,i}^{(t)}$ 将在 $\gamma_{j,r}^{(t)}$ 小于 $\Theta(n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d) = \Theta$（信噪比^2） $= \Theta(\widehat{\gamma}) = O(1)$ 的

$$T_2 = C_4 \eta^{-1} nm \sigma_p^{-2} d^{-1}$$

次迭代内小于 1，其中 C_4 可取为 $2/3$。因此，我们知道，在 $[0, T_2]$ 中，$F_{-1}(\mathbf{W}^{(t)}, \mathbf{x}_i)$ $F_{+1}(\mathbf{W}^{(t)}, \mathbf{x}_i) = O(1)$。因此，存在正常数 C，使得 $-\ell_i'^{(t)} \ge C$ 对 $0 \le t \le T_2$成立。

R回顾一下，我们将 $\{i \in [n] | y_i = y\}$ 记为 S_y。对于 $\gamma_{j,r}^{(t)}$ 的增长，如果 $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \ge 0$，我们有

$$\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+ \cap S_1} \ell_i'^{(t)} - \sum_{i \in S_- \cap S_{-1}} \ell_i'^{(t)} \bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&\ge \gamma_{j,r}^{(t)} + \frac{\eta}{nm} \cdot (C|S_+ \cap S_1| - |S_- \cap S_{-1}|) \cdot \|\boldsymbol{\mu}\|_2^2.
\end{aligned} \tag{D.3}$$

如果 $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0$，我们有

$$\begin{aligned}\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+ \cap S_{-1}} \ell_i'^{(t)} - \sum_{i \in S_- \cap S_1} \ell_i'^{(t)} \bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\ &\geq \gamma_{j,r}^{(t)} + \frac{\eta}{nm} \cdot (C|S_+ \cap S_{-1}| - |S_- \cap S_1|) \cdot \|\boldsymbol{\mu}\|_2^2.\end{aligned} \tag{D.4}$$

根据引理 B.3，在事件 $\mathcal{E}_{\mathrm{prelim}}$ 下，我们有

$$\begin{aligned}&\frac{|S_+ \cap S_1|}{|S_- \cap S_{-1}|}, \frac{|S_+ \cap S_{-1}|}{|S_- \cap S_1|} \geq \frac{(1-p)n - \sqrt{2n\log(8/\delta)}}{pn + \sqrt{2n\log(8/\delta)}}, \\ &|S_+ \cap S_1|, |S_+ \cap S_{-1}| \geq (1-p)n - \sqrt{2n\log(8/\delta)}.\end{aligned} \tag{D.5}$$

只要 $p < C/6$ 和 $n \geq 72C^{-2}\log(8/\delta)$，就有

$$\begin{aligned}&\frac{|S_+ \cap S_1|}{|S_- \cap S_{-1}|}, \frac{|S_+ \cap S_{-1}|}{|S_- \cap S_1|} \geq 2/C, \\ &|S_+ \cap S_1|, |S_+ \cap S_{-1}| \geq n/4.\end{aligned}$$

因此，我们有

$$\begin{aligned}\gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} + \frac{C\eta}{2nm} \cdot |S_+ \cap S_{-1}| \cdot \|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j,r}^{(t)} + \frac{C\eta}{8m} \cdot \|\boldsymbol{\mu}\|_2^2, \text{ if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \geq 0, \\ \gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} + \frac{C\eta}{2nm} \cdot |S_+ \cap S_1| \cdot \|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j,r}^{(t)} + \frac{C\eta}{8m} \cdot \|\boldsymbol{\mu}\|_2^2, \text{ if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0.\end{aligned} \tag{D.6}$$

注意到 $\gamma_{j,r}^{(0)} = 0$，可得

$$\gamma_{j,r}^{(t)} \geq \frac{C\|\boldsymbol{\mu}\|_2^2 \eta t}{8m}, \gamma_{j,r}^{(T_2)} \geq \frac{CC_4 n\|\boldsymbol{\mu}\|_2^2}{8\sigma_p^2 d} = \Theta(n \cdot \mathrm{SNR}^2) = \Theta(\widehat{\gamma}).$$

Note 我们已在引理 C.2 中证明了（C.25），即 $\gamma_{j,r}^{(t)}$ 对 $0 \leq t \leq T^*$ 递增，因此我们有 e

$$\gamma_{j,r}^{(t)} = \Omega(\widehat{\gamma})$$

对于 $T_2 \leq t \leq T^*$。并且由（D.2）可直接得到

$$\overline{\rho}_{j,r,i}^{(T_1)} \leq \frac{3\eta\sigma_p^2 d T_1}{2nm} = \frac{3C_3}{2}, \overline{\rho}_{j,r,i}^{(T_1)} = O(1),$$

由此完成证明。 □

D.2. 第二阶段

根据信号-噪声分解，在第一阶段结束时，我们有

$$\mathbf{w}_{j,r}^{(T_1)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(T_1)} \cdot \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2^2} + \sum_{i=1}^n \overline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^n \underline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}$$

对于 $j \in [\pm 1]$ 和 $r \in [m]$。根据我们在第一阶段得到的结果，我们知道，在这一阶段开始时，以下性质成立:

- $\underline{\rho}_{j,r^*}^{(T_1)}{}_i \geq 2$ 对于任意 $r^* \in S_i^{(0)} = \{r \in [m]: \langle \mathbf{w}_{y_i,\, r,}^{(0)} \boldsymbol{\xi}_i \rangle > 0\}$、$j \in \{\pm 1\}$ 以及满足 $y_i = j$ 的 $i \in [n]$。
- $\max_{j,\, r,\, i} |\underline{\rho}_{j,\, r,\, i}^{(T_1)}| = \max\{O(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}),\ O(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d})\}$。
- $\gamma_{j,\, r}^{(T_1)} = \Theta(\widehat{\gamma})$ 对于任意 $j \in \{\pm 1\}$，$r \in [m]$。

其中 $\widehat{\gamma} = n \cdot$信噪比^2。现在我们选择 $\mathbf{W}^*$ 如下

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 5\log(2/\epsilon)\Big[\sum_{i=1}^n \mathbb{1}(j = y_i) \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}\Big].$$

引理 D.2。在此条件下 与定理 *4.2*具有相同条件时，我们有 $\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F \leq \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2})$.

引理 *D.2*的证明。 我们有

$$\begin{aligned}
\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F &\leq \|\mathbf{W}^{(T_1)} - \mathbf{W}^{(0)}\|_F + \|\mathbf{W}^* - \mathbf{W}^{(0)}\|_F \\
&\leq O(\sqrt{m}) \max_{j,r} \gamma_{j,r}^{(T_1)} \|\boldsymbol{\mu}\|_2^{-1} + O(\sqrt{m}) \max_{j,r} \Big\| \sum_{i=1}^n \overline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^n \underline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} \Big\|_2 \\
&\quad + O(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= O(m^{1/2}\widehat{\gamma}\|\boldsymbol{\mu}\|_2^{-1}) + \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2}) + O(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= O(m^{1/2}n \cdot \mathrm{SNR} \cdot \sigma_p^{-1}d^{-1/2}) + \widetilde{O}(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2}),
\end{aligned}$$

其中，第一个不等式由三角不等式得到，第二个不等式和第一个等式由我们对$\mathbf{W}^{(T_1)}$、$\mathbf{W}^*$ 的分解以及引理 B.4 得到；第二个等式由 $n \cdot$信噪比$^2 = \Theta(\widehat{\gamma})$ 和 信噪比 $= \|\boldsymbol{\mu}\|/\sigma_p d^{1/2}$得到；第三个等式由$n^{1/2} \cdot$信噪比 $= O(1)$得到。□

引理 D.3。 在与定理 *4.2*相同的条件下，我们有

$$y_i\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \geq \log(2/\epsilon)$$

对所有 $T_1 \leq t \leq T^*$ 均成立。

引理的证明 *D.3*. 回忆一下， $f(\mathbf{W}^{(t)}) = (1/m)\sum_{j,r} j \cdot [\sigma(\langle \mathbf{w}_{j,r}, y_i \cdot \boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle)]$, 因此我们有

$$\begin{aligned}
&y_i\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \\
&= \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)\langle \boldsymbol{\mu}, j\mathbf{w}_{j,r}^*\rangle + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\langle y_i\boldsymbol{\xi}_i, j\mathbf{w}_{j,r}^*\rangle \\
&= \frac{1}{m}\sum_{j,r}\sum_{i'=1}^n \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon)\, \mathbb{1}(j = y_{i'}) \cdot \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \\
&\quad + \frac{1}{m}\sum_{j,r}\sum_{i'=1}^n \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle) 5\log(2/\epsilon)\, \mathbb{1}(j = y_{i'}) \cdot \frac{\langle \boldsymbol{\mu}, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \\
&\quad + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)\langle \boldsymbol{\mu}, j\mathbf{w}_{j,r}^{(0)}\rangle + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\langle y_i\boldsymbol{\xi}_i, j\mathbf{w}_{j,r}^{(0)}\rangle \\
&\geq \frac{1}{m}\sum_{j=y_i,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon) - \frac{1}{m}\sum_{j,r}\sum_{i' \neq i} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon) \cdot \frac{|\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle|}{\|\boldsymbol{\xi}_{i'}\|_2^2}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{m}\sum_{j,r}\sum_{i'=1}^{n}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)5\log(2/\epsilon)\cdot\frac{|\langle\boldsymbol{\mu},\boldsymbol{\xi}_{i'}\rangle|}{\|\boldsymbol{\xi}_{i'}\|_2^2}\\
&- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big) - \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big)\\
\geq\ & \underbrace{\frac{1}{m}\sum_{j=y_i,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)5\log(2/\epsilon)}_{I_9} - \underbrace{\frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)5\log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}/\sqrt{d}\big)}_{I_{10}}\\
&\underbrace{- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)5\log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}\cdot\mathrm{SNR}\cdot d^{-1/2}\big)}_{I_{11}}\\
&\underbrace{- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\boldsymbol{\mu}\rangle)O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big)}_{I_{12}} \underbrace{- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big)}_{I_{14}},
\end{aligned}
\tag{D.7}
$$

其中，第一个不等式由引理 B.5 得出，最后一个不等式由引理 B.4得出。接下来，我们将分别界定式（D.7）中的内积项。对于 $I_{10}, I_{11}, I_{12}, I_{14}$，注意到 $\sigma' \in \{0,1\}$ 我们有

$$
\begin{aligned}
&|I_{10}| \leq \log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}/\sqrt{d}\big),\ |I_{11}| \leq \log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}\cdot\mathrm{SNR}\cdot d^{-1/2}\big),\\
&|I_{12}| \leq O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big),\ |I_{14}| \leq O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big).
\end{aligned}
\tag{D.8}
$$

对于 $j = y_i$ 和 $r \in S_i^{(0)}$，根据引理 C.3，我们有

$$
\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle &\geq \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{j,r,i}^{(t)} - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha\\
&\geq 2 - \beta - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha\\
&\geq 1
\end{aligned}
$$

其中，第一个不等式由引理 C.3得出；最后一个不等式由 $\beta \leq 0.5$ 和 $5n\sqrt{\frac{\log(4n^2/\delta)}{d}} \leq 0.5$得出。因此，对于 I_9，根据命题 C.8的第四个结论，我们有

$$
I_9 \geq \frac{1}{m}|S_i^{(t)}|5\log(2/\epsilon) \geq 2\log(2/\epsilon). \tag{D.9}
$$

将（D.8）和（D.9）代入（D.7），并根据三角不等式可得

$$
y_i\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle \geq I_9 - |I_{10}| - |I_{11}| - |I_{12}| - |I_{14}| \geq \log(2/\epsilon),
$$

由此完成证明。 □

引理 D.4。 在条件 *4.1*下，对于 $0 \leq t \leq T^*$，以下结论成立。

$$
\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \leq O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\})L_S(\mathbf{W}^{(t)}).
$$

引理的证明 *D.4*。 我们首先证明

$$
\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F = O(\max\{\|\boldsymbol{\mu}\|_2, \sigma_p\sqrt{d}\}). \tag{D.10}
$$

不失一般性，我们假设 $\widehat{y}_i = 1$ 且 $\mathbf{x}_i = [\boldsymbol{\mu}^\top, \boldsymbol{\xi}_i]$。于是我们有

$$\begin{aligned}\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F &\le \frac{1}{m}\sum_{j,r}\Big\|\big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}\rangle)\boldsymbol{\mu} + \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i\big]\Big\|_2 \\ &\le \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}\rangle)\|\boldsymbol{\mu}\|_2 + \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2 \\ &\le 4\max\{\|\boldsymbol{\mu}\|_2, 2\sigma_p\sqrt{d}\},\end{aligned}$$

其中，第一个和第二个不等式由三角不等式得到，第三个不等式由引理 B.4和 $\sigma' \le 1$得到。现在我们可以将梯度范数 $\|\nabla L_S(\mathbf{W}^{(t)})\|_F$上界估计如下:

$$\begin{aligned}\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 &\le \bigg[\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big)\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F\bigg]^2 \\ &\le \bigg[\frac{1}{n}\sum_{i=1}^n O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) - \ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big)\bigg]^2 \\ &\le O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) \cdot \frac{1}{n}\sum_{i=1}^n -\ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) \\ &\le O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) L_S(\mathbf{W}^{(t)}),\end{aligned}$$

其中，第一个不等式由三角不等式得到，第二个不等式由（D.10）得到，第三个不等式由柯西-施瓦茨不等式得到，最后一个不等式则由交叉熵损失的性质得到 $-\ell' \le \ell$。 □

引理 D.5。在与定理*4.2*相同的条件下，我们有

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \ge \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon$$

对所有 $T_1 \le t \le T^*$ 均成立。

引理*D.5*的证明。我们有

$$\begin{aligned}&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\ &= 2\eta\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^*\rangle - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\ &= \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^{(t)}[y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\ &\ge \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^{(t)}[y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \log(2/\epsilon)] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\ &\ge \frac{2\eta}{n}\sum_{i=1}^n [\ell\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) - \epsilon/2] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\ &\ge \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon,\end{aligned}$$

其中，第一个不等式由引理 D.3 得出；第二个不等式源于交叉熵函数的凸性；最后一个不等式由引理D.4 得出。 □

引理 D.6. 在与定理 *4.2*相同的条件下，对于所有 $T_1 \le t \le T^*$，有 最大值$_{j,r,i}|\rho_{j\,r\,,,i}^{(t)}| =$ _

最大值 $\{O(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}),\ O(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d})\}$。此外，

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t}L_S(\mathbf{W}^{(s)})\le\frac{\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\|_F^2}{\eta(t-T_1+1)}+\epsilon$$

对于所有 $T_1\le t\le T^*$。因此，我们可以在 $T=T_1+\Big\lfloor\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\|_F^2/(\eta\epsilon)\Big\rfloor=T_1+\widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ 次迭代内找到一个训练损失小于 2ϵ 的迭代点。

引理的证明 *D.6*。 注意， 最大值$_{j,\ r,\ i}|\rho_{j\ r,\ \ ,\ i}^{(t)}|=$最大值$\{O(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}),\ O(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d})\}$可用与引理相同的方法证明 D.1；此处省略证明细节。对于任意 $t\in[T_1,T]$，将引理 D.5 中的不等式求和，并将等式两边同时除以 $(t-T_1+1)$，可得

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t}L_S(\mathbf{W}^{(s)})\le\frac{\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\|_F^2}{\eta(t-T_1+1)}+\epsilon$$

对所有 $T_1\le t\le T$。根据 T 的定义，我们有

$$\frac{1}{T-T_1+1}\sum_{s=T_1}^{T}L_S(\mathbf{W}^{(s)})\le 2\epsilon.$$

那么，存在某次迭代 $T_1\le t\le T$，使得训练损失小于 ϵ。 □

此外，当训练损失小于 ϵ 时，关于 $\rho_{j\ r\ ,,i}^{(t)},\gamma_{j\ r\ ,}^{(t)}$的比值顺序，我们有如下引理。该引理将帮助我们证明关于测试误差的定理。

引理 D.7。 在与定理 *4.2*相同的条件下，我们有

$$\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}/\gamma_{j',r'}^{(t)}=\Theta(\mathrm{SNR}^{-2})\tag{D.11}$$

对所有 j、$j'\in\{\pm1\}$、r、$r'\in[m]$ 和 $T_1\le t\le T^*$。

引理 D.7 的证明。我们将使用归纳法证明该引理。首先验证 (D.11) 对 $t=T_1$ 成立。根据引理 D.1，我们有 $\gamma^{(T_1)}{}_{j',}$ $_{r'}=\Theta(\widetilde{\gamma})=\Theta(n\cdot$信噪比$^2)$ 和 $\sum_{i=1}^{n}\rho_{j\ r}^{(T_1)}$，，$_{\ i}=\Theta(n)$，因此 (D.11) 直接成立。现在假设存在 $\widetilde{T}\in[T_1,T^*]$，使得对于所有 $t\in[T_1,\widetilde{T}-1]$，均有 $\sum_{i=1}^{n}\overline{\rho}_{j\ r}^{(t)}$，，$_{i}/\gamma_{j'}^{(t)}{}_{,r'}=\Theta($信噪比$^2)$。那么对于 $\rho_{j\ r}^{(t)}$，，$_{\ i}$，根据引理 C.1，我们有

$$\begin{aligned}\overline{\rho}_{j,r,i}^{(t+1)}&=\overline{\rho}_{j,r,i}^{(t)}-\frac{\eta}{nm}\cdot\ell_i'^{(t)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle)\cdot\mathbb{1}(y_i=j)\|\boldsymbol{\xi}_i\|_2^2.\\ \gamma_{j',r'}^{(t+1)}&=\gamma_{j',r'}^{(t)}-\frac{\eta}{nm}\cdot\Bigg[\sum_{i\in S_+}\ell_i'^{(t)}\sigma'(\langle\mathbf{w}_{j',r'}^{(t)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)-\sum_{i\in S_-}\ell_i'^{(t)}\sigma'(\langle\mathbf{w}_{j',r'}^{(t)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\Bigg]\cdot\|\boldsymbol{\mu}\|_2^2\end{aligned}$$

由此可得

$$\begin{aligned}\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})}=\sum_{i:y_i=j}\overline{\rho}_{j,r,i}^{(\widetilde{T})}&=\sum_{i:y_i=j}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}-\frac{\eta}{nm}\cdot\sum_{i:y_i=j}\ell_i'^{(\widetilde{T}-1)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\\&=\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}-\frac{\eta}{nm}\cdot\sum_{i\in S_{j,r}^{(\widetilde{T}-1)}}\ell_i'^{(\widetilde{T}-1)}\|\boldsymbol{\xi}_i\|_2^2\\&\geq\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}+\frac{\eta\sigma_p^2 d}{16m}\cdot\min_{i\in S_{j,r}^{(\widetilde{T}-1)}}|\ell_i'^{(\widetilde{T}-1)}|,\end{aligned}\tag{D.12}$$

~其中最后一个等号由 $S_{j\,r}^{(T-1)}$、当 $\{i\in[n]:\ \ y_i=j$、$\langle\mathbf{w}_{j\,r}^{(T-1)}$、、$\boldsymbol{\xi}_i\rangle>0\}$; 最后一个不等式由引理 B.4和命题C.8的第五条陈述得到。并且

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})}&\leq\gamma_{j',r'}^{(\widetilde{T}-1)}-\frac{\eta}{nm}\cdot\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2\\&\leq\gamma_{j',r'}^{(\widetilde{T}-1)}+\frac{\eta\|\boldsymbol{\mu}\|_2^2}{m}\cdot\max_{i\in S_+}|\ell_i'^{(\widetilde{T}-1)}|.\end{aligned}\tag{D.13}$$

~根据命题 C.8的第三条陈述，我们有 最大值$_{i\in S_+}|\ell_i'^{(T-1)}|\leq C_2$最小值$_{i\in S^{(\widetilde{T}-1)}\,j,}$ $_r$ $|\widetilde{\ell}_i'^{(T-1)}|$。然后结合 (D.12) 和 (D.13)，可得

$$\frac{\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}}\geq\min\left\{\frac{\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}}{\gamma_{j',r'}^{(\widetilde{T}-1)}},\frac{\sigma_p^2 d}{16C_2\|\boldsymbol{\mu}\|_2^2}\right\}=\Theta(\mathrm{SNR}^{-2}).\tag{D.14}$$

另一方面，根据 (D.12) 并由引理B.4，可得

$$\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})}\leq\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}+\frac{9\eta\sigma_p^2 d}{8m}\cdot\max_{i\in S_{j,r}^{(\widetilde{T}-1)}}|\ell_i'^{(\widetilde{T}-1)}|,\tag{D.15}$$

~其中该不等式由 $|S_j^{(T-1)}$、$_r$ $|\leq|S_j|\leq 3n/4$得到。并且按照与 (D.3)、(D.4)、(D.5) 和 (D.6) 类似的推理方式，我们可以得到，只要 $q<C_2/6$ 且 $n\geq 72C_2^{-2}\log(8/\delta)$，就有

$$\sum_{i\in S_+}|\ell_i'^{(\widetilde{T}-1)}|\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\geq 2\sum_{i\in S_-}|\ell_i'^{(\widetilde{T}-1)}|\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)$$

因此

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})}&=\gamma_{j',r'}^{(\widetilde{T}-1)}-\frac{\eta}{nm}\cdot\left[\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)-\sum_{i\in S_-}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\right]\cdot\|\boldsymbol{\mu}\|_2^2\\&\geq\gamma_{j',r'}^{(\widetilde{T}-1)}-\frac{\eta}{2nm}\cdot\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2.\end{aligned}$$

接下来我们有

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})}&\geq\gamma_{j',r'}^{(\widetilde{T}-1)}-\frac{\eta}{2nm}\sum_{i\in S_+\cap S_1}\ell_i'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\mu}\|_2^2\geq\gamma_{j',r'}^{(\widetilde{T}-1)}+\frac{\eta\|\boldsymbol{\mu}\|_2^2}{8m}\min_{i\in S_+\cap S_1}\ell_i'^{(\widetilde{T}-1)},\ \text{if}\ \langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\boldsymbol{\mu}\rangle\geq 0,\\\gamma_{j',r'}^{(\widetilde{T})}&\geq\gamma_{j',r'}^{(\widetilde{T}-1)}-\frac{\eta}{2nm}\sum_{i\in S_+\cap S_{-1}}\ell_i'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\mu}\|_2^2\geq\gamma_{j',r'}^{(\widetilde{T}-1)}+\frac{\eta\|\boldsymbol{\mu}\|_2^2}{8m}\min_{i\in S_+\cap S_{-1}}\ell_i'^{(\widetilde{T}-1)},\ \text{if}\ \langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\boldsymbol{\mu}\rangle<0,\end{aligned}\tag{D.16}$$

其中第二个不等式由引理B.3得出。根据命题 C.8的第四条陈述，我们有

$$\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}} \le \max\left\{\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}}{\gamma_{j',r'}^{(\widetilde{T}-1)}}, \frac{9C_2\sigma_p^2 d}{\|\boldsymbol{\mu}\|_2^2}\right\} = \Theta(\mathrm{SNR}^{-2}). \tag{D.17}$$

根据（D.14）和（D.17），我们有

$$\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}} = \Theta(\mathrm{SNR}^{-2}),$$

这就完成了归纳。

□

实际上，引理 D.7 中的结果对于 $0 \le t \le T^*$ 也成立，即

引理 D.8。 在与定理 *4.2* 相同的条件下，我们有

$$\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(t)} / \gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2}) \tag{D.18}$$

对所有 j、$j' \in \{\pm 1\}$、r、$r' \in [m]$ 和 $0 \le t \le T^*$ 均成立。

证明思路与引理 D.7几乎相同，而后续论证中只需使用引理 D.7 ，因此此处省略证明细节。

E. 测试误差分析

E.1. 测试误差上界

接下来，我们给出定理 t 中定义的第 t 次迭代时的测试误差上界，其中定理 4.2中的训练损失收敛至 ϵ。首先，注意到 $T_1 \le t \le T^*$ 根据引理 D.6，我们可以将先前的结果归纳如下：

- $\sum_{i=1}^n \overline{\rho}_{(t)\,j,\,r,\,i} / \gamma_{(t)\,j',\,r'} = \Theta($信噪比$^{-2})$ (见引理 D.7),
- $\sum_{i=1}^n \rho_{(t)\,j,\,r,\,i} = \Omega(n) = O(n\log(T^*)) = \widetilde{\Theta}(n)$ (见命题 C.2 和引理 D.1)
- 最大值$_{j,\,r,\,i} |\rho_{(t)}^{t}$ max $\{O(\sqrt{\log(mn/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}), O(\sqrt{\log(n/\delta)}\log(T^*)\cdot n/\sqrt{d})\}$ (from Lemma D.).

此外，回顾前两个结论中的定义 $\widehat{\gamma} = n \cdot$ 信噪比^2,，我们有 $\gamma_{j\,,\,r}^{(t)} = \widetilde{\Theta}(\widehat{\gamma})$，对所有 j，r 均成立。还要注意，根据第三个结论，由于 $\sigma_0\sigma_p\sqrt{d} = \widetilde{O}(\sqrt{n}/\sqrt{d}) = o(1)$，且根据条件 4.1，$\sqrt{\log(n/\delta)}\log(T^*)\cdot n/\sqrt{d} = O(1)$，我们有 $\max_{j,\,r,\,i} |\rho_{j\,r\,,\,\,,\,i}^{(t)}| = O(1)$，因此 $\sum_{i=1}^n |\rho_{\widehat{y}\,r\,,\,\,,\,i}^{(t)}| = O(\sum_{i=1}^n \rho_{\widehat{y}\,r,-,\,i}^{(t)})$，所以当 ρ 的求和与 ρ 的求和同时出现时，可以忽略前者。现在我们可以在下述定理中分析测试误差。

定理 E.1 （定理 4.2的第二部分）。 在与定理 *4.2*相同的条件下，存在一个较大的常数C_1 使得当 $n\|\boldsymbol{\mu}\|_2^2 \ge C_1\sigma_p^4 d$ 时，对于 t 引理 *D.6*中定义的时间，测试误差为

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \ne \mathrm{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big) \le p + \exp\left(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\right),$$

其中 $C_2 = O(1)$。

证明。为方便起见，我们用$(\mathbf{x}\ \widehat{y}, y)\ \sim \mathcal{D}$表示如下：数据点$(\mathbf{x}\ y)$服从定义 $\mathcal{D}$中定义的分布1.1，并且 $\widehat{y}$ 是其真实标签。我们可以将测试误差写为

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \operatorname{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big)\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)},\mathbf{x}) \leq 0\big)\\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)},\mathbf{x}) \leq 0, y \neq \widehat{y}\big) + \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)},\mathbf{x}) \leq 0, y = \widehat{y}\big)\\
&= p\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x}) \geq 0\big) + (1-p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x}) \leq 0\big)\\
&\leq p + \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x}) \leq 0\big),
\end{aligned} \tag{E.1}$$

其中，在第二个等式中我们使用了定义 $\mathcal{D}$ 1.1。因此，只需给出 $\mathbb{P}_{(\mathbf{x},\widehat{y})\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\ \leq 0\big)$的上界。为此，我们将$\mathbf{x}\ = (\widehat{y}\boldsymbol{\mu}, \boldsymbol{\xi})$写成如下形式，并得到

$$\begin{aligned}
\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x}) &= \frac{1}{m}\sum_{j,r}\widehat{y}j[\sigma(\langle\mathbf{w}_{j,r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)]\\
&= \frac{1}{m}\sum_{r}[\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle)] - \frac{1}{m}\sum_{r}[\sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle)]
\end{aligned} \tag{E.2}$$

现在先考虑表达式 $\langle\mathbf{w}_{j\ r}^{(t)},,\widehat{y}\boldsymbol{\mu}\rangle$ （其中 $j = \pm\widehat{y}$）。回顾式（5.1）中的$\mathbf{w}_{j\ r}^{(t)}$,的信号-噪声分解:

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j\cdot\gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-2}\cdot\boldsymbol{\mu} + \sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\boldsymbol{\xi}_i + \sum_{i=1}^{n}\underline{\rho}_{j,r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\boldsymbol{\xi}_i,$$

因此，与 $j = \widehat{y}$ 的内积可按如下方式给出上界

$$\begin{aligned}
\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle &= \langle\mathbf{w}_{\widehat{y},r}^{(0)},\widehat{y}\boldsymbol{\mu}\rangle + \gamma_{\widehat{y},r}^{(t)} + \sum_{i=1}^{n}\overline{\rho}_{\widehat{y},r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\langle\boldsymbol{\xi}_i,\widehat{y}\boldsymbol{\mu}\rangle + \sum_{i=1}^{n}\underline{\rho}_{\widehat{y},r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\langle\boldsymbol{\xi}_i,\widehat{y}\boldsymbol{\mu}\rangle\\
&\geq \gamma_{\widehat{y},r}^{(t)} - \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\\
&\quad - \sqrt{2\log(6n/\delta)}\cdot\sigma_p\|\boldsymbol{\mu}\|_2\cdot(\sigma_p^2 d/2)^{-1}\bigg[\sum_{i=1}^{n}\overline{\rho}_{\widehat{y},r,i}^{(t)} + \sum_{i=1}^{n}|\underline{\rho}_{\widehat{y},r,i}^{(t)}|\bigg]\\
&= \gamma_{\widehat{y},r}^{(t)} - \Theta(\sqrt{\log(m/\delta)}\sigma_0\|\boldsymbol{\mu}\|_2) - \Theta\big(\sqrt{\log(n/\delta)}\cdot(\sigma_p d)^{-1}\|\boldsymbol{\mu}\|_2\big)\cdot\Theta(\mathrm{SNR}^{-2})\cdot\gamma_{\widehat{y},r}^{(t)}\\
&= \big[1 - \Theta\big(\sqrt{\log(n/\delta)}\cdot\sigma_p/\|\boldsymbol{\mu}\|_2\big)\big]\gamma_{\widehat{y},r}^{(t)} - \Theta\big(\sqrt{\log(m/\delta)}(\sigma_p d)^{-1}\sqrt{n}\|\boldsymbol{\mu}\|_2\big)\\
&= \Theta(\gamma_{\widehat{y},r}^{(t)}),
\end{aligned} \tag{E.3}$$

其中，不等式由引理 B.4 和引理 B.5 得出；第二个等式是将本节开头总结的系数阶代入后得到的；第三个等式由条件 4.1 中的条件 $\sigma_0 \leq C^{-1}(\sigma_p d)^{-1}\sqrt{n}$ 和信噪比 $= \|\boldsymbol{\mu}\|_2/\sigma_p\sqrt{d}$ 得出；对于第四个等式，注意到 $\gamma_j^{(t)}$,$\overline{r} = \Omega(\widehat{\gamma})$，此外，由条件 4.1 中的 $\|\boldsymbol{\mu}\|_2^2 \geq C\cdot\sigma_p^2\log(n/\delta)$ 和 $n \geq C\log(m/\delta)$ 可知，$\sqrt{\log(n/\delta)}\cdot\sigma_p/\|\boldsymbol{\mu}\|_2 \leq 1/\sqrt{C}$ and$\sqrt{\log(m/\delta)}(\sigma_p d)^{-1}\sqrt{n}\|\boldsymbol{\mu}\|_2/\widehat{\gamma} = \sqrt{\log(m/\delta)}\sigma_p/(\sqrt{n}\|\boldsymbol{\mu}\|_2) \leq \sqrt{\log}(m/\delta)/n\cdot 1/\big(\sqrt{C\log(n/\delta)}\big) \leq 1/(C\sqrt{\log(n/\delta)})$ 成立，因此对于充分大的常数 C，等式

成立。此外，我们可以用类似的方法推出

$$\begin{aligned}\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle &= \langle \mathbf{w}_{-\widehat{y},r}^{(0)}, \widehat{y}\boldsymbol{\mu}\rangle - \gamma_{-\widehat{y},r}^{(t)} + \sum_{i=1}^{n} \overline{\rho}_{-\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, -\widehat{y}\boldsymbol{\mu}\rangle + \sum_{i=1}^{n} \underline{\rho}_{-\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, \widehat{y}\boldsymbol{\mu}\rangle \\ &\le -\gamma_{-\widehat{y},r}^{(t)} + \sqrt{2\log(8m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2 \\ &\quad + \sqrt{2\log(6n/\delta)} \cdot \sigma_p \|\boldsymbol{\mu}\|_2 \cdot (\sigma_p^2 d/2)^{-1} \Big[\sum_{i=1}^{n} \overline{\rho}_{-\widehat{y},r,i}^{(t)} + \sum_{i=1}^{n} |\underline{\rho}_{-\widehat{y},r,i}^{(t)}| \Big] \\ &= -\Theta(\gamma_{-\widehat{y},r}^{(t)}) < 0,\end{aligned} \tag{E.4}$$

其中第二个等号是基于与（E.3）中类似的分析得出的。
记 $g(\boldsymbol{\xi})$ 为 $\sum_r \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle)$。根据 Vershynin (2018年中的定理 5.2.2，我们知道，对任意 $x \ge 0$，都有

$$\mathbb{P}(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \ge x) \le \exp\Big(-\frac{cx^2}{\sigma_p^2 \|g\|_{\mathrm{Lip}}^2}\Big), \tag{E.5}$$

其中 c是一个常数。为计算 Lipschitz 范数，我们有

$$\begin{aligned}|g(\boldsymbol{\xi}) - g(\boldsymbol{\xi}')| &= \Big| \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) - \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}'\rangle) \Big| \\ &\le \sum_{r=1}^{m} \big| \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) - \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}'\rangle) \big| \\ &\le \sum_{r=1}^{m} |\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi} - \boldsymbol{\xi}'\rangle| \\ &\le \sum_{r=1}^{m} \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2 \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2,\end{aligned}$$

其中，第一个不等式由三角不等式得出；第二个不等式由 ReLU 的性质得出；最后一个不等式由柯西-施瓦茨不等式得出。因此，我们有

$$\|g\|_{\mathrm{Lip}} \le \sum_{r=1}^{m} \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2, \tag{E.6}$$

并且由于 $\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle \sim \mathcal{N}(0, \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2^2 \sigma_p^2)$，我们可以得到

$$\mathbb{E}g(\boldsymbol{\xi}) = \sum_{r=1}^{m} \mathbb{E}\sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) = \sum_{r=1}^{m} \frac{\|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2 \sigma_p}{\sqrt{2\pi}} = \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^{m} \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2.$$

接下来，我们要对$\mathbf{w}_{j\,r}^{(t)}$,的 2-范数给出上界。首先，我们处理分解中的噪声部分，即:

$$\begin{aligned}&\Big\| \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i \Big\|_2^2 \\ &= \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 \cdot \|\boldsymbol{\xi}_i\|_2^{-2} + 2 \sum_{1 \le i_1 < i_2 \le n} \rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)} \cdot \|\boldsymbol{\xi}_{i_1}\|_2^{-2} \cdot \|\boldsymbol{\xi}_{i_2}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i_1}, \boldsymbol{\xi}_{i_2}\rangle \\ &\le 4\sigma_p^{-2} d^{-1} \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 + 2 \sum_{1 \le i_1 < i_2 \le n} |\rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)}| \cdot (16\sigma_p^{-4} d^{-2}) \cdot (2\sigma_p^2 \sqrt{d\log(6n^2/\delta)})\end{aligned}$$

$$
\begin{aligned}
&= 4\sigma_p^{-2}d^{-1}\sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2 + 32\sigma_p^{-2}d^{-3/2}\sqrt{\log(6n^2/\delta)}\bigg[\bigg(\sum_{i=1}^{n}|\rho_{j,r,i}^{(t)}|\bigg)^2 - \sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2\bigg] \\
&= \Theta(\sigma_p^{-2}d^{-1})\sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2 + \widetilde{\Theta}(\sigma_p^{-2}d^{-3/2})\bigg(\sum_{i=1}^{n}|\rho_{j,r,i}^{(t)}|\bigg)^2 \\
&\le \big[\Theta(\sigma_p^{-2}d^{-1}n^{-1}) + \widetilde{\Theta}(\sigma_p^{-2}d^{-3/2})\big]\bigg(\sum_{i=1}^{n}|\overline{\rho}_{j,r,i}^{(t)}| + \sum_{i=1}^{n}|\underline{\rho}_{j,r,i}^{(t)}|\bigg)^2 \\
&\le \Theta(\sigma_p^{-2}d^{-1}n^{-1})\bigg(\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}\bigg)^2
\end{aligned}
$$

其中，第一个不等式用到了引理 B.4；第二个不等式用到了 $\rho, \underline{\rho}$ 的定义；倒数第二个等式中代入了系数阶。因此，我们可以对$\mathbf{w}_{j,r}^{(t)}$,的范数给出上界：

$$
\begin{aligned}
\|\mathbf{w}_{j,r}^{(t)}\|_2 &\le \|\mathbf{w}_{j,r}^{(0)}\|_2 + \gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1} + \bigg\|\sum_{i=1}^{n}\rho_{j,r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\boldsymbol{\xi}_i\bigg\|_2 \\
&\le \|\mathbf{w}_{j,r}^{(0)}\|_2 + \gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1} + \Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)} \\
&= \Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}
\end{aligned} \tag{E.7}
$$

其中，第一个不等式由三角不等式得出，而等式则基于以下比较：

$$
\frac{\gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1}}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = \Theta(\sigma_p d^{1/2}n^{1/2}\|\boldsymbol{\mu}\|_2^{-1}\mathrm{SNR}^2) = \Theta(\sigma_p^{-1}d^{-1/2}n^{1/2}\|\boldsymbol{\mu}\|_2) = O(1)
$$

根据系数顺序 $\sum_{i=1}^{n}\rho_{j,r,,i}^{(t)}/\gamma_{j,r,}^{(t)} = \Theta($信噪比$^{-2})$、定义 信噪比 $=\|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$以及条件$d$在条件 4.1中的要求；此外

$$
\frac{\|\mathbf{w}_{j,r}^{(0)}\|_2}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = \frac{\Theta(\sigma_0\sqrt{d})}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = O(\sigma_0\sigma_p d n^{-1/2}) = O(1)
$$

根据引理 B.5、系数顺序 $\sum_{i=1}^{n}\rho_{j,r,,i}^{(t)} = \Omega(n)$以及条件 σ_0 在条件 4.1中的要求。结合这些结果和（E.3），我们对以下关键组成部分进行分析，

$$
\frac{\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^{m}\big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2} \ge \frac{\Theta\big(\sum_r\gamma_{\widehat{y},r}^{(t)}\big)}{\Theta(d^{-1/2}n^{-1/2})\cdot\sum_{r,i}\overline{\rho}_{-\widehat{y},r,i}^{(t)}} = \Theta(d^{1/2}n^{1/2}\mathrm{SNR}^2) = \Theta(n^{1/2}\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d^{1/2}) \tag{E.8}
$$

由（E.8）以及 $n\|\boldsymbol{\mu}\|_2^4 \ge C_1\sigma_p^4 d$ 其中 C_1 是一个充分大的常数，可直接推出

$$
\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^{m}\|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2 > 0. \tag{E.9}
$$

现在利用（E.5）中的方法，结合上述结果，将（E.4）代入（E.2），再代入（E.1，得到

$$
\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\le 0\big) \le \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\sum_r\sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle) \ge \sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)\bigg)
$$

$$
\begin{aligned}
&= \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \geq \sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2\bigg) \\
&\leq \exp\bigg[-\frac{c\Big(\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) - (\sigma_p/\sqrt{2\pi})\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2\Big)^2}{\sigma_p^2\Big(\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2\Big)^2}\bigg] \\
&= \exp\bigg[-c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2} - 1/\sqrt{2\pi}\Big)^2\bigg] \\
&\leq \exp(c/2\pi)\exp\bigg(-0.5c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2}\Big)^2\bigg) \qquad \text{(E.10)}
\end{aligned}
$$

其中，第二个不等式由（E.9）并将（E.6）代入（E.5）得到；第三个不等式则是由于以下事实：$(s-t)^2 \geq s^2/2 - t^2,\ \forall s,\ t \geq 0$。

我们可以由（E.8）和（E.10）得到

$$
\begin{aligned}
\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x}) \leq 0\big) &\leq \exp(c/2\pi)\exp\bigg(-0.5c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2}\Big)^2\bigg) \\
&= \exp\Big(\frac{c}{2\pi} - \frac{n\|\boldsymbol{\mu}\|_2^4}{C\sigma_p^4 d}\Big) \\
&\leq \exp\Big(-\frac{n\|\boldsymbol{\mu}\|_2^4}{2C\sigma_p^4 d}\Big) \\
&= \exp\Big(-\frac{n\|\boldsymbol{\mu}\|_2^4}{C_2\sigma_p^4 d}\Big),
\end{aligned}
$$

其中 $C = O(1)$；the 个不等式在选择 $C_1 \geq cC/\pi$ 时成立；最后一个等式在选择 C_2 为 $2C$ 时成立 .

□

E.2. 测试误差下界

本节将给出训练损失收敛到 ϵ时，定理 4.2 中定义的第 t 次迭代的测试误差下界4.2，这与定理 5.8。由此可知存在尖锐相变。首先，我们给出关键引理

引理的证明 5.8。 不失一般性，令 $\max\Big\{\sum_r \gamma_{1\,r}^{(t)},, \sum_r \gamma_{-1,r}^{(t)}\Big\} = \sum_r \gamma_{1\,r}^{(t)},$。 记 $\mathbf{v} = \lambda\cdot\sum_i \mathbb{1}(y_i = 1)\boldsymbol{\xi}_i$，其中 $\lambda = C_7\|\boldsymbol{\mu}\|_2^2/(d\sigma_p^2)$，且 C_7 是一个足够大的常数。于是我们只需证明

$$
\underbrace{g(\boldsymbol{\xi}+\mathbf{v}) - g(\boldsymbol{\xi}) + g(-\boldsymbol{\xi}+\mathbf{v}) - g(-\boldsymbol{\xi})}_{I} \geq 4C_6\sum_r \gamma_{1,r}^{(t)}. \qquad \text{(E.11)}
$$

由于 ReLU 是凸激活函数，我们有

$$
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) \geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle)\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \qquad \text{(E.12)}
$$

$$
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle)\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle. \qquad \text{(E.13)}
$$

将（E.12）和（E.13）相加可得，对于所有 $\boldsymbol{\xi}$，几乎必然有

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \\
&\geq \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \\
&\geq \lambda\bigg[\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} - 2n\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} - 5n^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg],
\end{aligned} \tag{E.14}$$

wher最后一个不等式由（C.14）和引理B.5得出。由于 ReLU 是利普希茨函数，我们还可以得到

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \\
&\leq 2|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}\rangle| \\
&\leq 2\lambda\bigg[\sum_{y_i=1} \underline{\rho}_{-1,r,i}^{(t)} + 2n\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} + 5n^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg],
\end{aligned} \tag{E.15}$$

最后一个不等式由（C.13）和引理 B.5得出。因此，将（E.14）和（E.15）代入左侧 I 的（E.11），我们得到

$$\begin{aligned}
&g(\boldsymbol{\xi}+\mathbf{v}) - g(\boldsymbol{\xi}) + g(-\boldsymbol{\xi}+\mathbf{v}) - g(-\boldsymbol{\xi}) \\
&\geq \lambda\bigg[\sum_{r}\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} - 6nm\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} - 15mn^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg] \\
&\geq (\lambda/2)\cdot\sum_{r}\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} \\
&\geq \lambda/2\cdot\Theta(\mathrm{SNR}^{-2})\sum_{r}\gamma_{1,r}^{(t)} \\
&\geq 4C_6\sum_{r}\gamma_{1,r}^{(t)},
\end{aligned}$$

第二个不等式由引理 D.1 和条件 4.1得出；第三个不等式由引理 D.7得出。最后值得注意的是，范数

$$\|\mathbf{v}\|_2 = \|\lambda\cdot\sum_i \mathbb{1}(y_i=1)\boldsymbol{\xi}_i\|_2 = \Theta\bigg(\sqrt{\frac{n\|\boldsymbol{\mu}\|_2^4}{\sigma_p^4 d}}\bigg) \leq 0.06\sigma_p,$$

最后一个不等式由条件 $n\|\boldsymbol{\mu}\|_2^4 \leq C_3\sigma_p^4 d$ 以及定理 C_3 中充分大的 4.2得出，至此证明完成。

□

接下来，我们给出一个重要引理，它界定了具有相同协方差矩阵的两个高斯分布之间的全变差（TV）距离。

引理 E.2（命题 2.1，出自 Devroye 等人（2018年））。 和 $\mathcal{N}(0,\sigma_p^2\mathbf{I}_d)$ 之间的全变差距离 $\mathcal{N}(\mathbf{v}\ \sigma_p^2\mathbf{I}_d)$ 小于 $\|\mathbf{v}\|_2/2\sigma_p$。

最后，我们可以证明定理4.2的第三部分：根据引理 E.2 和引理 5.8。

定理 E.3（定理 4.2的第三部分）。 假设 $n\|\boldsymbol{\mu}\|_2^4 \leq C_3 d\sigma_p^4$，则有 $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \geq p+0.1$，其中 C_3 是一个足够大的绝对常数。

证明。 为方便起见，我们用 $(\mathbf{x}\ \widehat{y}, y) \sim \mathcal{D}$ 表示如下情况：数据点 $(\mathbf{x}\ y)$ 服从分布 $\mathcal{D}$

定义于定义 1.1中，且 $\widehat{y}$ 是其真实标签。根据 (E.1)，我们有

$$\begin{aligned}&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y\neq\operatorname{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big)\\&=p\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\geq 0\big)+(1-p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big)\\&=p+(1-2p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big).\end{aligned}\tag{E.16}$$

因此，只需给出 $\mathbb{P}_{(\mathbf{x},\widehat{y})\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\ \mathbf{x})\ \leq 0\big)$ 的下界。为此，我们有

$$\begin{aligned}&\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\leq 0\big)\\&=\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\sum_r\sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle)-\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle)\geq\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)-\sum_r\sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)\bigg)\\&\geq 0.5\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\bigg|\sum_r\sigma(\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\rangle)-\sum_r\sigma(\langle\mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\rangle)\bigg|\geq C_6\max\Big\{\sum_r\gamma_{1,r}^{(t)},\sum_r\gamma_{-1,r}^{(t)}\Big\}\bigg)\end{aligned}\tag{E.17}$$

其中 C_6 是常数。该不等式成立是因为，如果 $\left|\sum_r\sigma(\langle\mathbf{w}_{1\ r}^{(t)},\ \boldsymbol{\xi}\rangle)-\sum_r\sigma(\langle\mathbf{w}_{-1,\ r}^{(t)},\ \boldsymbol{\xi}\rangle)\right|$ 过大，那么给定 $\boldsymbol{\xi}$，我们总能选取相应的 $\widehat{y}$ 来进行错误预测。令 $g(\boldsymbol{\xi})=\sum_r\sigma(\langle\mathbf{w}_{1\ r}^{(t)},\ \boldsymbol{\xi}\rangle)-\sum_r\sigma(\langle\mathbf{w}_{-1,\ r}^{(t)},\ \boldsymbol{\xi}\rangle)$。记该集合为

$$\Omega:=\bigg\{\boldsymbol{\xi}\bigg||g(\boldsymbol{\xi})|\geq C_6\max\Big\{\sum_r\gamma_{1,r}^{(t)},\sum_r\gamma_{-1,r}^{(t)}\Big\}\bigg\}.$$

将Ω 的定义代入（E.17），可得

$$\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\leq 0\big)\geq 0.5\mathbb{P}(\Omega)\tag{E.18}$$

接下来，我们将给出 $\mathbb{P}(\Omega)$ 的一个下界。根据引理 5.8，有 $\sum_j[g(j\boldsymbol{\xi}+\mathbf{v})\ -g(j\boldsymbol{\xi})]\geq 4C_6\max_j\Big\{\sum_r\gamma_j^{(t)},_r\Big\}$因此，根据鸽巢原理，必定存在 $\boldsymbol{\xi},\boldsymbol{\xi}+\mathrm{v}$、$-\boldsymbol{\xi},-\boldsymbol{\xi}+\mathrm{v}$ 中的一个属于 Ω。由此我们证明了 $\Omega\cup-\Omega\cup\Omega\ -\{\mathrm{v}\}\cup-\Omega\ -\{\mathrm{v}\}=\mathbb{R}^d$。因此，$\mathbb{P}(\Omega)$ $\mathbb{P}(-\Omega)$ $\mathbb{P}(\Omega\ -\{\mathrm{v}\})$ $\mathbb{P}(\Omega\ -\{\mathrm{v}\})$ $\mathbb{P}(-\Omega\ -\{\mathrm{v}\})$ 中至少有一个大于 0.25。注意，$\mathbb{P}(-\Omega)\ =\mathbb{P}(\Omega)$，并且

$$\begin{aligned}|\mathbb{P}(\Omega)-\mathbb{P}(\Omega-\mathbf{v})|&=|\mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma_p^2\mathbf{I}_d)}(\boldsymbol{\xi}\in\Omega)-\mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{v},\sigma_p^2\mathbf{I}_d)}(\boldsymbol{\xi}\in\Omega)|\\&\leq\mathrm{TV}(\mathcal{N}(0,\sigma_p^2\mathbf{I}_d),\mathcal{N}(\mathbf{v},\sigma_p^2\mathbf{I}_d))\\&\leq\frac{\|\mathbf{v}\|_2}{2\sigma_p}\\&\leq 0.03,\end{aligned}$$

其中，第一个不等式依据全变差（TV）距离的定义，第二个不等式依据引理 E.2。

因此，我们已经证明 $\mathbb{P}(\Omega)\ \geq 0.22$，将其代入式（E.16）和式（E.18），得到 t

$$\begin{aligned}&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y\neq\operatorname{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big)\\&=p+(1-2p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big)\\&\geq p+(0.5-p)\cdot\mathbb{P}(\Omega)\\&\geq 0.78p+0.11\\&\geq p+0.1,\end{aligned}$$

其中，最后一个不等式由条件 $p<1/C$ 4.1 得出，并且通过选取 $C>22$ 一个足够大的常数即可得到，这就完成了证明。 □