高斯数据就是你所需要的一切吗？高维广义线性估计中普适性的范围与局限

Luca Pesce[1], Florent Krzakala[1], Bruno Loureiro[2], 和 Ludovic Stephan[1]

[1]洛桑联邦理工学院（EPFL），信息、学习与物理（IdePHICS）实验室，瑞士洛桑 CH-1015。[2]计算机科学系，巴黎高等师范学院 - PSL 与法国国家科学研究中心，45 rue d'Ulm，F-75230 巴黎 cedex 05，法国。

arXiv:2302.08923v1 [math.ST] 17 Feb 2023

摘要

本文研究标签由单指标模型给出的高斯混合数据上的广义线性估计问题。我们的首个结果给出了高维状态下测试误差和训练误差的精确渐近表达式。受近期关于广义线性估计中测试误差和训练误差的高斯普适性的一系列研究启发，我们提出这样一个问题：何时单个高斯分布足以刻画误差？我们的公式能够从正反两方面对此给出精确答案。具体而言，我们表明，高斯普适性成立（或不成立）的充分条件，关键取决于目标权重与混合簇的均值和协方差之间的对齐程度；我们对此进行了精确量化。对于最小二乘插值这一特殊情形，我们证明了训练误差具有强普适性，并表明其遵循一个简单的闭式表达式。最后，我们将结果应用于真实数据集，澄清了文献中关于这一情境下误差的高斯普适性的近期讨论。

1 引言

机器学习领域普遍认为，数据中的结构是成功学习的重要要素。如何量化这一观点，尤其是特征中的结构如何对训练误差和泛化误差产生最大影响，是深入理解现代机器学习这个黑箱这一广泛研究计划中的重要课题。在理论方面，广义线性估计领域近期在这一方向上取得了一些重要进展。例如，近期一系列关于基于高斯数据训练的线性回归的研究表明，即使在训练误差恰好为零的"过参数化状态"下，也能实现良好的泛化 [1–3]。这种良性过拟合性质关键取决于协方差结构：当目标的信号分量与数据的一个低维子空间对齐时，噪声便有空间分散到更高维的正交子空间中 [1]。在类似条件下，广义线性任务也有类似结论 [4–6]。事实上，这只是过去几年对高斯数据上的广义线性模型进行研究所获得的诸多出人意料的发现之一 [7–10]。尽管特征分布的假设看似有所限制，但近期一系列研究在不同设定下，为广义估计中训练误差和泛化误差的 高斯普适性提供了有力证据。这些研究包括针对非高斯设计的严格结果 [11, 12], 随机特征映射 [13–16], 神经切线特征 [17], 带随机标签的高斯混合 [18], , 以及针对其他特征映射的广泛数值证据 [15, 19], 甚至包括真实数据集 [19–21]。这些研究引出了一个问题："何时高斯特征是适用于学习的良好模型？"我们的目标是在广义线性估计的背景下，对这一问题给出明确答案。我们关注的是一种广受欢迎的多模态数据模型——高斯混合模型；众所周知，它能够近似任意分布。

1.1 主要结果

本工作的**主要贡献**如下：

• **广义线性模型的精确渐近行为： </style}我们给出了高维情况下凸损失广义线性模型训练误差和测试误差的精确渐近极限，其中数据来自具有单指标目标的高斯混合模型。这些渐近结果基于统计物理学中所谓的副本方法** ，并沿用了 [22], 的思路；该研究考察的则是学习混合标签这一任务。

• 训练误差和测试误差的普适性： </style}我们给出了关于目标权重 θ_0 的一组充分条件，使得高斯混合模型的渐近训练误差和测试误差均与簇均值无关。特别是，当目标方向在 $\mathbb{S}^{d-1}$上均匀分布时，这些条件便得到满足。对于岭回归，我们还证明了一个更强的结果：训练损失也与簇协方差无关，并且等同于具有单位协方差的单个高斯分布对应的训练损失。

• **同方差性的重要性**：在同方差高斯混合（具有相同协方差矩阵的高斯分布混合）这一特定情形下，我们进一步证明了普适性结果；这些结果在经过随机特征映射后，实际上可以在真实数据上观察到（例如参见图1和 2）。我们还揭示了线性可分性转变的普适行为。这一现象已针对纯高斯数据得到详细研究，见 [10] ；而对于同方差混合分布，它似乎也具有普适性。

• **打破普适性：**与上一段的结果相反，我们表明有两种方式可以打破高斯普适性。首先，强异方差性会破坏普适行为。其次，在同方差情形下，我们表明数据与任务之间的相关性至关重要：即使目标权重与簇均值之间只有微弱相关性，也足以打破普适性；具体而言，在高斯混合上训练的模型，其渐近误差会不同于在高斯数据上训练的模型。因此，起作用的似乎并非数据结构本身，而是这种结构与待学习任务之间的相关性。

1.2 相关工作

精确渐近行为：高斯数据的一个吸引人之处在于，在比例高维极限下，不同模型的渐近性能可以得到精确刻画，其中 $n,\ d \to \infty$ 样本复杂度固定为 $\alpha := n/d$。岭回归尤其如此，因为它与随机矩阵理论问题密切相关；例如参见 [2, 3, 23]。除二次情形外，也有许多严谨的渐近研究，例如， [24] 研究了M估计量， [25] 研究了LASSO估计量的性能，而 [19] 则针对具有任意协方差的凸损失和惩罚项给出了一般性结果。

对于高斯混合，相关研究大多致力于分类，即恢复簇标签，而非教师生成的标签。在二分类中，相关例子包括 [26–28], ；后者还揭示了分类与单指标模型之间的等价性。在多类别情形下， [29] 研究了岭回归分类器的性能；这一研究方向中最一般的结果是 [22], ，它考虑了任意凸损失（不一定可分）。

高斯普适性：值得注意的是，该模型还能刻画特定类别非高斯特征所产生的误差。 [30] 证明，对于采用随机特征 [13–16, 31], 神经切线特征 [17] 和核特征 [21, 32–35]的广义线性估计，这种高斯普适性性质（GEP）成立。 [36] 表明，在随机特征模型中，若权重训练的步数为常数阶，高斯普适性便得以保持；但若训练步数达到随规模增长的数量，则会失效。超越定理的范围， [19] 提供了数值证据，表明高斯普适性也适用于一类更广泛、来自训练好的神经网络的真实特征。与此密切相关的是， [18] 研究了纯随机二元标签下的高斯普适性，这为本研究提供了重要灵感；与此同时， [19–21, 34, 37, 38]通过数值研究表明，尤其对于岭回归，高斯公式能够刻画一些简单真实数据集的学习曲线。

2 设定与动机

令 $(\boldsymbol{x}^\nu,\ y^\nu) \in \mathbb{R}^d \times \mathcal{Y}$ 表示 $\nu = 1,\ \cdots,\ n$ 对独立采样的训练点。我们将研究广义线性估计 $\hat{}y(\boldsymbol{x}) = \hat{}f(\boldsymbol{\theta}^\top \boldsymbol{x})$ 的性质，其中权重 $\boldsymbol{\theta} \in \mathbb{R}^d$ 是通过最小化以下经验风险从训练数据中学习得到的：

$$\hat{\mathcal{R}}_n^\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^\nu, \frac{\boldsymbol{\theta}^\top \boldsymbol{x}}{\sqrt{d}}\right) + \lambda r(\boldsymbol{\theta}) \tag{1}$$

其中， $\ell\colon \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ 是一个凸损失函数， $r\colon \mathbb{R}^d \to \mathbb{R}$ 是一个凸惩罚项。例如，这包括岭回归这一特例，其中 $\ell(y,\ \hat{}y) = (y - \hat{}y)^2$ 且 $r(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_2^2$。接下来关注的关键量是

训练误差和泛化误差，其定义如下：

$$\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^{\nu}, \frac{\hat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}}\right) \tag{2}$$

$$\varepsilon_{\text{gen}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[g\left(y_{\text{new}}, \hat{f}\left(\frac{\hat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}_{\text{new}}}{\sqrt{d}}\right)\right)\right] \tag{3}$$

其中 $g\colon \mathcal{Y}\times\mathbb{R}\to\mathbb{R}_{+}$ 是统计学家选择的性能指标，不一定等于损失函数$\ell(y,\hat{y})$。例如，在采用 $\mathcal{Y}=\{-1,+1\}$ 的二分类问题中，我们可以令 $\ell(y,\hat{y}) = \log(1+e^{-y\hat{y}})$ 表示逻辑损失，而令 $g(y,\hat{y}) = \mathbb{I}[y\neq\hat{y}]$ 表示分类误差。请注意，在式 (3) 中，期望是针对一对新的数据 $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ 求取的，我们假设它是从与训练数据相同的分布中独立抽取的。

具体而言，我们将关注在标签由以下目标分布生成这一假设下对这些误差进行刻画：

$$y^{\nu} \sim P_0\left(\cdot \,\middle|\, \frac{\boldsymbol{\theta}_0^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}}\right) \tag{4}$$

其中，$\boldsymbol{\theta}_0\in\mathbb{R}^d$ 是某个固定向量，P_0是某个分布。对于加性高斯噪声 $\xi\sim\mathcal{N}(0,\Delta)$，$P_0$ 的常见选择是 $y^{\nu} = f_0\left(\frac{\boldsymbol{\theta}_0^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}} + \xi\right)$。这种设定有时称为教师-学生设定。我们有时会采用这一便于使用的术语，将目标分布 P_0 称为教师，将 $\boldsymbol{\theta}_0\in\mathbb{R}^d$ 称为教师权重。类似地，我们有时会将模型 $\hat{y}$ 称为学生，将 $\hat{\boldsymbol{\theta}}\in\mathbb{R}^d$ 称为学生权重。如前所述，我们将同时考虑回归 $\mathcal{Y}=\mathbb{R}$ 和二分类 $\mathcal{Y}=\{-1,+1\}$。

过去几年中，这一模型一直是高维统计学文献中大量研究工作的主题，尤其是在<style id='1'>高斯设计 假设下：

模型 2.1（高斯协变量模型）。 在高斯协变量模型（GCM）中，我们假设输入独立地采样自一个高斯分布：

$$\boldsymbol{x}^{\nu} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}}{\sqrt{d}}, \Sigma\right), \qquad \nu = 1,\cdots,n. \tag{5}$$

我们将 $\mathcal{G}_{\boldsymbol{\theta}_0,\ \Sigma,\ \boldsymbol{\mu}}$ 称为在该数据假设下的教师-学生问题。

这项工作的一个关键动机是这样一种普遍直觉：来自 MNIST 等标准分类任务的数据，与其说更接近单模态高斯分布，不如说更接近多模态分布。因此，在本文中，我们提出这样一个问题：“什么时候高斯数据就足够了？”。具体而言，我们关注一种用于建模多模态数据（以及密度的通用逼近器）的典型分布：K 簇高斯混合：

模型 2.2（高斯混合模型）。 在高斯混合模型（GMM）中，我们假设输入独立地采样自由 K 个高斯分布组成的混合分布：

$$\boldsymbol{x}^{\nu} \sim \sum_{c\in\mathcal{C}} p_c\,\mathcal{N}\left(\frac{\boldsymbol{\mu}_c}{\sqrt{d}}, \Sigma_c\right), \qquad \nu = 1,\cdots,n. \tag{6}$$

其中，$\mathcal{C}:=\{1,\cdots,K\}$ 是可能的簇集合，$p_c\in[0,1]$ 是属于簇 $c\in\mathcal{C}$的概率，其均值和协方差由（$\boldsymbol{\mu}_c$、Σ_c）给出。

我们注意到，尽管模型2.1 是模型 $K=1$ 2.2 的一个特例（当标签 y^{ν} 不依赖于输入簇时），但将 $K=1$和 $K>1$ 两种情况视为不同的模型会更有启发性。

3 主要理论结果

本节中，我们介绍关于高斯混合模型高维广义线性估计普适性的主要理论结果。我们的结果建立在一系列工作的基础上，这些工作在模型2的比例高维极限下，对经验风险最小化器给出了精确渐近刻画（2.1 [2, 19, 23, 24]）。特别是，与我们的推导更为接近的是 [19, 22]中的严格结果。我们证明，训练误差和泛化误差在高维中集中于一个确定性表达式，该表达式由一组自洽方程的解给出。随后我们

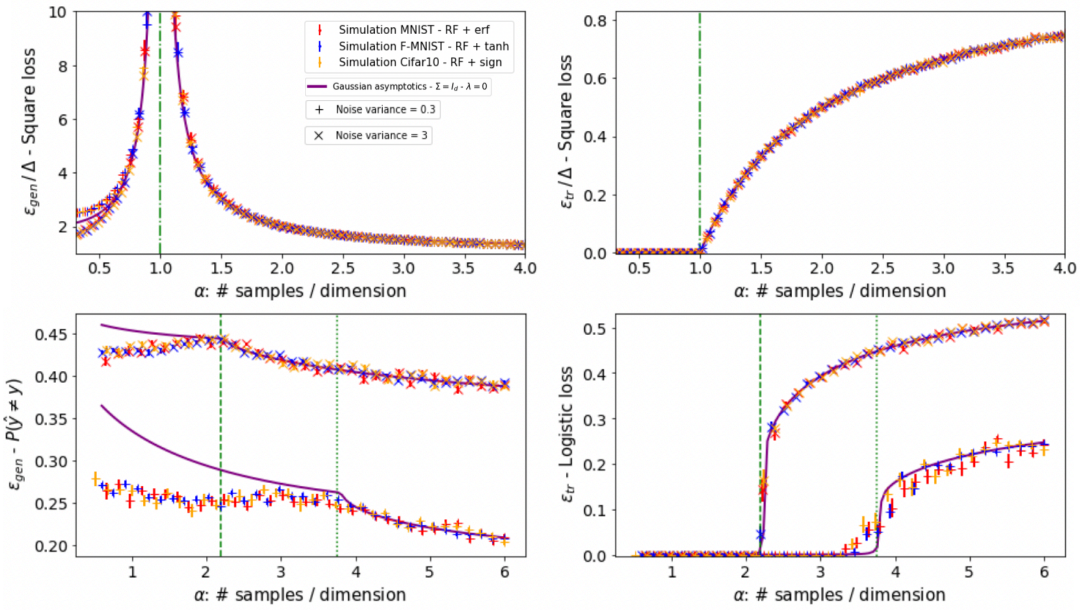

图 1：消失正则化条件下高斯普适性的示意图 $\lambda = 0^+$ 针对一组选定的数据集（MNIST、Fashion-MNIST、Cifar10），采用随机特征映射后使用随机教师函数：泛化误差（左）和训练误差（右）随每维样本数 $\alpha = n/d$ 的变化。岭回归（上排）和逻辑分类（下排）。紫色实线表示高斯协变量具有单位协方差时的精确渐近公式，绿色竖线表示阈值 $\alpha^\star(\Delta)$，在该阈值处，损失开始存在唯一极小值点。圆点表示使用随机特征映射在不同真实数据集上进行的数值模拟。所有数据的训练损失都符合高斯预测，这验证了定理3.4和3.6。特别是，可分性（或插值）阈值与高斯模型的阈值一致（见推论3.5）。每项学习任务均在两种不同的标签噪声方差下运行——叉号对应于 $\Delta = 3$，加号对应于 $\Delta = 0.3$。误差棒根据 30 次运行的标准差绘制。

分析高斯普适性，并严格刻画学习任务必须满足的一组充分条件，使得模型2.2的测试误差和训练误差在渐近意义上与模型2.1的相一致。

由于我们处理的是随机变量序列，因此需要对收敛给出严格定义。对于两个数列 (a_n)，(b_n)，我们记

$$a_n \simeq b_n \quad \text{iff} \quad \lim_{n\to\infty} a_n - b_n = 0. \tag{7}$$

相应地，对于两个随机变量序列 (X_n)，(Y_n)，我们将依概率接近定义为

$$X_n \overset{P}{\simeq} Y_n \quad \text{iff} \quad X_n - Y_n \overset{P}{\to} 0, \tag{8}$$

其中 $\overset{P}{\to}$表示依概率收敛。

3.1 精确渐近

我们的第一个结果是给出高斯混合模型 2 中极小值点 2.2 性能的闭式渐近刻画，从而推广了 [22]的严格结果：

命题 3.1. （精确渐近结果，非正式表述）考虑式 (1) 中提出的经验风险最小化问题，其中数据服从模型 *2.2*给出的高斯混合分布。对于任意伪利普希茨性能度量 g，训练误差和泛化误差 (2) 在 n、$d \to \infty$ 且比例 $\alpha = n/d$ 固定时的高维极限下，收敛到由一组自洽复制鞍点方程 (33)的解完全确定的确定性表达式。

$$\begin{aligned} \varepsilon_{tr}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \varepsilon_{tr}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \\ \varepsilon_{gen}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \varepsilon_{gen}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \end{aligned} \tag{9}$$

命题 3.1 与 [22]中的定理 1 的关键区别在于标签的分布。 [22]中的标签由 GMM 聚类索引给出，而在命题 3.1 中，我们考虑由目标函数（4）生成的标签。尽管我们没有给出这些结果的正式证明，但请注意， [22], 中将解映射到所谓的近似消息传递算法研究的通用证明框架 [39], 可以直接推广到我们的设定。实际上，在我们的情形中，近似消息传递方案并无不同；证明中唯一的区别是在其渐近分析中加入一个教师，这与 [40]中的做法类似。

一旦我们明确了 GMM，并在方程 (1) 中恰当地定义了经验风险最小化，就可以将方程 2、(3) 中的表达式表示为低维量的函数。这些量定义了渐近误差的充分统计量，也称为序参数。计算细节见附录 A。

3.2 不相关教师

深入比较模型2.1 &2.2的渐近误差表达式可知，它们的关键差异在于主导目标方向 $\boldsymbol{\theta}_0$ 与簇均值和协方差相关联的方式。因此，迈向普适性的第一步是明确渐近误差在什么条件下与均值无关。我们作出以下假设：

假设 1。教师 $\boldsymbol{\theta}_0$ 满足 $\forall(c,\ c') \in \mathcal{C} \times \mathcal{C}$:

$$\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0 \tag{10}$$

$$\lim_{n,d\to\infty} \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_{c'} \left(\lambda + \sum_{c\in\mathcal{C}} \hat{V}_c^\star \Sigma_c\right)^{-1} \boldsymbol{\mu}_c \to 0 \tag{11}$$

其中 $\{\hat{V}_c^\star\}_{c=1}^K$ 是描述中心化高斯混合模型问题的（复制）鞍点方程的不动点。

假设 2。 损失函数和教师分布均为对称的：

$$\ell(x,y) = \ell(-x,-y) \tag{12}$$

$$P_0(y|\tau) = P_0(-y|-\tau), \tag{13}$$

并且正则化为 ℓ^2 惩罚项 $\lambda/2\|\cdot\|_2^2$。

命题 3.2。 （均值普适性）
在假设 *1* 和 *2*下，簇均值 $\{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}$ 与高维经验风险最小化估计无关：

$$\varepsilon_{gen}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \varepsilon_{gen}^{\mathrm{GMM}}\left(\mathbf{0}, \{\Sigma_c\}_{c=1}^K\right) \tag{14}$$

$$\varepsilon_{tr}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \varepsilon_{tr}^{\mathrm{GMM}}\left(\mathbf{0}, \{\Sigma_c\}_{c=1}^K\right) \tag{15}$$

因此，直观地说，当标签生成过程与数据结构不相关时，可以实现均值普适性；详细讨论见附录 B 。对目标和损失函数的假设并不具有限制性，而且奇目标激活函数 f_0 以及如下形式的基于间隔的损失都很容易满足这些假设： $l(y,\ z) = \ell(yz)$。通过对数据结构作出简化假设，可以证明更强的普适性。事实上，如果混合分布是齐次的（即所有协方差均相同；统计学中通常称之为同方差性），则有高斯普适性：

定理 3.3。*(*高斯 同方差高斯混合模型的普适性）在*3.2*的假设下，考虑一个同方差高斯混合模型：

$$\Sigma_c = \Sigma \qquad \forall c \in \mathcal{C}$$

那么，对于所有 α 和 Δ，高斯混合模型的误差渐近等于高斯协方差模型的误差：

$$\varepsilon_{gen}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \Sigma\right) \simeq \varepsilon_{gen}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, \Sigma, \mathbf{0}) \tag{16}$$

$$\varepsilon_{tr}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \Sigma\right) \simeq \varepsilon_{tr}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, \Sigma, \mathbf{0}) \tag{17}$$

这些结果可由定理直接得出。 3.2。在可控假设下，我们将一个高斯混合模型问题映射到了一个更简单的高斯问题。

此外，如果考虑消失正则化，我们可以证明误差与协方差的形状无关，因此可以采用各向同性混合分布：

定理 3.4。（$\lambda = 0^+$的协方差普适性）在定理 3.3的假设下，假设经验风险 (1) 在零正则化下存在唯一极小值点。那么，同方差高斯混合模型 (2.2)估计问题中的测试误差和训练误差，与中心各向同性高斯模型 $\mathcal{G}_{(\boldsymbol{\theta}_0,\mathbf{I},\mathbf{0})}$ (2.1)渐近一致。

这个一般性结果的证明见附录B。

由于逻辑损失或合页损失等损失的训练误差刻画了 $\lambda = 0^+$处的可分性转变，定理3.4的一个有趣推论如下：

推论 3.5。（同方差高斯混合模型的线性可分性普适性）对于同方差高斯混合模型和高斯模型，导致数据不再线性可分的可分性／插值转变位置 α_c 相同。

鉴于文献 [10] 对随机教师权重和高斯数据下的可分性转变进行了详细研究，这种普适性尤其值得关注。文献 [41] 在线性估计的重构转变中也观察到了类似的普适性现象。

令人惊讶的是，如果进一步考虑平方损失最小化，那么满足均值普适性条件的任意高斯混合模型（无论是否同方差！）的估计问题都可以映射为一个简单的高斯问题：

定理 *3.6*。（$\lambda = 0^+$的平方损失强普适性）考虑具有一般均值和协方差、且满足定理 *3.2* 假设的高斯混合模型的欠参数化学习。在式（*1*）中设定：$l(y, \hat{y}) = (y - \hat{y})^2$，并考虑一个一般的带噪线性教师激活函数 $P_0(\tau; \Delta) = \mathcal{N}(\tau, \Delta)$，其中 $\xi \sim \mathcal{N}(0, 1)$。那么，高斯混合模型估计问题的训练误差为

$$\varepsilon_{tr}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \frac{(\alpha-1)\Delta}{\alpha} \tag{18}$$

请注意，强普适性结论并不适用于测试误差（附录C 讨论了一个反例，其中采用了强异方差情形）。然而，正如图1所示，令人惊讶的是，在参数不足区域，真实数据上的随机教师回归似乎遵循高斯渐近预测。此外，我们从图1 下图可以观察到，即使对于非二次损失，也能得出类似结论；我们将在下一节进一步探讨这一点。

最后，我们指出，在无限噪声和二元标签的极限下，我们的结果回归到 [18]中关于纯随机拉德马赫标签所观察到的结果。

3.3 相关教师

现在我们考虑一般情形：目标权重与数据中的结构相关，从而放宽定理中的假设。3.2。我们还能得出什么结论吗？我们的第一个结果表明答案是肯定的。我们聚焦于一个简单且可控的设定，在该设定下，对于任意教师向量，都可以将经验风险最小化性能表示为闭式：

定理 3.7。（各向同性协方差下的精确渐近结果）考虑一个具有双簇高斯混合模型的岭回归任务（2.2）。注意，不失一般性，我们可以取 $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \boldsymbol{\mu}$。假设协方差为各向同性：

$$\Sigma_+ = \Sigma_- = I_d, \tag{19}$$

并记为：

$$\rho = \lim_{d\to\infty} \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma \boldsymbol{\theta}_0, \qquad \gamma = \lim_{d\to\infty} \frac{1}{d}||\boldsymbol{\mu}||_2^2 \tag{20}$$

$$\pi = \lim_{d\to\infty} \frac{1}{d}\boldsymbol{\mu}^\top \boldsymbol{\theta}_0 \qquad P_0(\tau;\Delta) = \mathcal{N}(\tau,\Delta). \tag{21}$$

定义：

$$A(\eta) = \frac{(\eta-1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{\left(\Delta + (\eta-1)^2\rho\right)(1+\gamma\eta)^2} \quad \text{with} \tag{22}$$

$$\eta = 1 + \frac{1}{2}\left(\alpha - 1 + \lambda - \sqrt{4\lambda + (\alpha - 1 + \lambda)^2}\right)$$

渐近误差可用 η 表示为闭式表达式：

$$\varepsilon_{gen}^{\mathrm{GMM}} \simeq g(\alpha,\Delta,\rho,\eta)(1-\pi^2 A(\eta)) \tag{23}$$

$$\varepsilon_{tr}^{\mathrm{GMM}} \simeq t(\alpha,\Delta,\rho,\eta)(1-\pi^2 A(\eta)) \tag{24}$$

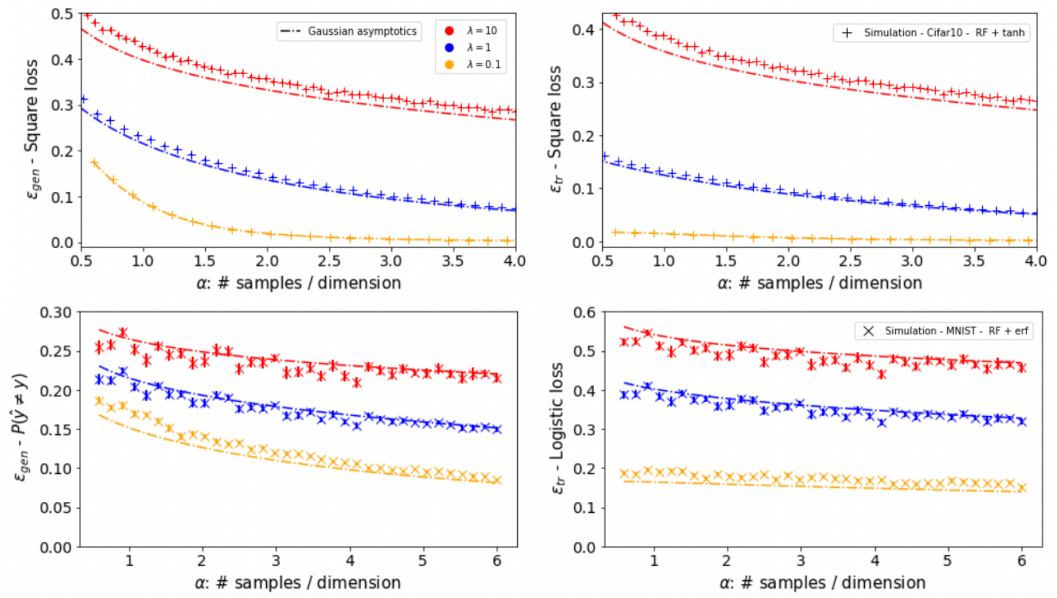

图 2：有限正则化下高斯普适性的示意图 λ：针对一组数据集（MNIST、Fashion-MNIST、Cifar10），使用随机教师函数经过随机特征映射后，泛化误差（左）和训练误差（右）随每维样本数变化的情况 $\alpha = n/d$。上图展示了采用随机特征和双曲正切函数激活函数、并经随机特征预处理的 Cifar10 上的岭回归；下图则展示了采用随机特征和误差函数激活函数的 MNIST 上的逻辑回归。虚线曲线是通过匹配协方差得到的高斯理论精确渐近预测。两者的良好吻合体现了定理3.3的性质。不同颜色分别代表不同的正则化强度$\lambda \in \{0.1, 1, 10\}$，依次为黄色、蓝色和红色。误差棒基于 30 次运行计算得出。

其中，g 和 t 分别是单位协方差、教师不相关的单个高斯模型的泛化 $\varepsilon_{gen}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, I, \mathbf{0})$ 误差和训练 $\varepsilon_{tr}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, I, \mathbf{0})$ 误差的渐近极限：

$$g(\alpha, \Delta, \rho, \eta) = \frac{\alpha\left(\Delta + (\eta - 1)^2\rho\right)}{(\alpha - \eta^2)} \tag{25}$$

$$t(\alpha, \Delta, \rho, \eta) = \frac{(\alpha - \eta)^2\left(\Delta + (\eta - 1)^2\rho\right)}{\alpha\,(\alpha - \eta^2)} \tag{26}$$

完整的闭式表达式以及对如下形式协方差的推广 $\Sigma = \sigma I_d$ 在附录中推导C。请注意，高斯性能的修正因子随 π^2变化，它衡量教师向量与数据结构之间的相关性。因此，我们可能会认为，与数据结构相关的目标，即 $\pi \neq 0$，总会破坏高斯普适性。尽管通常确实如此，但在消失正则化的极限下，我们可以给出定理3.7的一个有趣推论：

推论 3.8。（恢复 $\lambda = 0^+$ 的普适性）考虑定理 *3.7* 中相同的设定，并进一步考虑消失正则化极限下的欠参数化学习。对于任意教师向量，高斯混合模型估计的测试误差和训练误差均与高斯情形相同（即 *23* 式和 *24* 式在 $\eta = 0$ 下成立）：

$$\varepsilon_{gen}^{\mathrm{GMM}} = \Delta\frac{\alpha}{\alpha - 1} \tag{27}$$

$$\varepsilon_{tr}^{\mathrm{GMM}} = \Delta\frac{(\alpha - 1)}{\alpha} \tag{28}$$

因此，我们以类似于定理3.4 所述的方式，在欠参数化区域恢复了相关教师的高斯普适性。该定理在均值普适性条件下，针对一般的凸损失和协方差给出了相应结论。相关教师的普适性对于具有单位协方差的更一般同方差混合分布同样成立。为简洁起见，我们请参阅附录C，其中我们将详细讨论这一有趣的扩展。

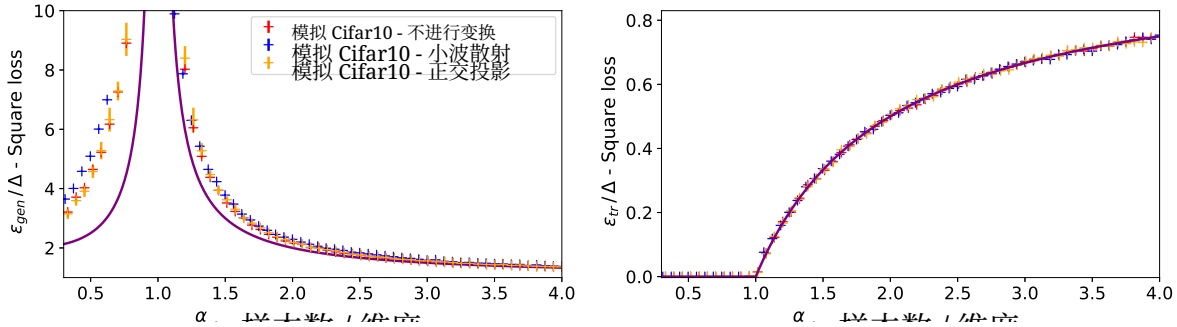

图 3：在正则化趋于零时，训练过程中的高斯普适性，以及泛化过程中缺乏这种普适性的示意图。不同于图 1 中使用随机特征映射，我们对灰度 Cifar10 数据进行小波散射预处理（蓝点），或使用正交哈达玛投影（黄点），也可以直接使用原始数据（红点）。再次使用随机教师函数后，发现训练误差与具有单位协方差的普适高斯渐近预测完全吻合，且 $\lambda = 0^+$，与定理 3.6 一致；但泛化结果则明显偏离。这凸显了数据中异方差性的作用。误差棒基于30 次运行绘制。

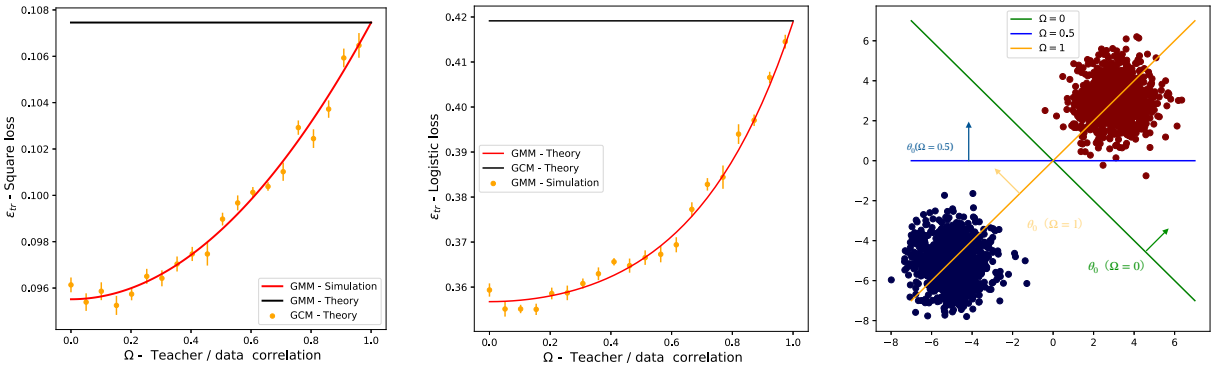

图 4：任务相关性与数据结构的重要性：**左图和中图**：使用 $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_-$ 和 $\Sigma_+ = \Sigma_- = I_d$进行双簇高斯混合模型估计时，经验风险最小化所得到的训练误差随相关性参数 Ω 变化的曲线：我们构造了一系列教师权重依赖于 Ω 的学习任务，即取 $\boldsymbol{\theta}_0(\Omega) = \Omega\boldsymbol{\mu}_\perp + \sqrt{1-\Omega^2}\boldsymbol{\mu}_+$，其中 $\boldsymbol{\mu}_+^\top \boldsymbol{\mu}_\perp = 0$。我们将维度固定为 $d = 500$，并令 $(\alpha, \lambda) = (1.2,\ 0.7)$。黑色实线表示高斯渐近理论预测。红色实线表示高斯混合模型性能的理论预测，橙色圆点则为数值模拟结果，符合预期，与理论预测一致。左图使用真实标签并进行岭回归；中图则使用二元标签并进行逻辑回归。误差棒根据 30 次运行的标准差绘制。**右图**：$d = 2$ 的几何直觉图。图中展示了三个超平面（二维中的直线），分别对应 $\Omega = \{0, 0.5, 1\}$，颜色依次为绿色、蓝色和橙色。随着 Ω 增大，标签与数据结构的相关性逐渐减弱。

4 合成数据集与真实数据集上的示例

本节将探讨主要理论结果的意义。首先，我们考虑真实数据的情形，说明我们的普适性定理适用于目标与数据结构不相关的情况。基于这些观察，我们进一步研究定理3.7 所描述的简单合成设置；针对这些设置，我们可以推导解析结果，并系统地检验高斯普适性。我们实验中使用的所有代码均可在GitHub 仓库中获取。

4.1 随机教师普适性

我们根据命题3.2的假设，分析随机教师函数条件下真实数据的高斯普适性。我们考虑三个标准数据集：MNIST、Fashion-MNIST 和灰度 CIFAR-10，以及合成数据。

随机特征映射的普适性——为了超越简单的线性拟合，我们用随机特征映射 [42] 对数据进行预处理，具体如下：取一幅图像 $\boldsymbol{\omega}_\nu \in \mathbb{R}^{d'}$，并将其映射为 $\boldsymbol{x}_\nu = \sigma(F\boldsymbol{\omega}_\nu)$，其中各元素 $F \in \mathbb{R}^{d\times d'}$独立同分布地取自 $\mathcal{N}(0,1)$，而 $\sigma(\cdot)$ 是一个激活函数。如 [42], 所示，对于 $d \to \infty$，这会收敛到一种核方法。对于每个数据集，我们采用不同的非线性函数（如图 1 所述，分别为误差函数、双曲正切函数和符号函数）。接着，将式 (4) 代入 $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, I_d)$，以此构造新的标签；对于回归任务，我们设 $f_0(\tau) = \tau + \Delta$，而对于分类任务，我们设 $f_0(\tau) =$ 符号函数$(\tau + \Delta)$。

首先，我们分析图 1 中的消失正则化情形：如定理 3.6 所述，经过预处理的真实数据集上的随机教师回归遵循训练损失的强普适性。有趣的是，对于泛化误差，我们还观察到高斯渐近预测与图 1 中的模拟结果完全吻合。这表明，经过随机特征映射后，数据已经充分接近同方差混合分布。此外，图 1 的下方面板显示，强普适性似乎不仅适用于平方损失最小化。

我们还分析了图 2 中的有限正则化情形：我们将真实数据上的模拟结果与精确高斯渐近的预测进行比较：匹配每个经过预处理的真实数据集的协方差，并借助确定性副本公式 3.1 计算经验风险最小化估计量的性能。高斯理论的预测与数值模拟结果相吻合。正如图 1 所讨论的，定理 3.3 中的同方差假设有时似乎可以放宽。

双重下降现象的普适性——现代机器学习中一项与经典统计理论认知相悖的发现，是学习曲线呈现双重下降行为 [13, 43——45], 见图1上图：随着参数数量增加，测试误差并不会恶化，而是在$\alpha = 1$处出现特征性的发散，这被称为插值峰。如图1 和图2所示，对于 $\alpha > 1$，同方差数据的泛化误差及其特征性的“双重下降”行为具有普适性，而训练误差似乎即使在异方差数据下也具有普适性（见图3）。

线性可分性的普适性——最近， [10] 研究了具有随机教师和噪声的高斯数据的线性可分性 Δ，并将 [46] 的经典结果推广到了单指标目标。 [10] 证明了存在一个临界相变 $\alpha_c(\Delta)$，它连续地从 $\alpha_c(\infty) = 2$（无限噪声时）变化到 $\alpha_c > 2$（有限 Δ 时（当 $\alpha_c \to \infty$ 在 $\Delta \to 0$时）。正如推论3.5所述，该相变对于同方差混合分布具有普适性。因此，这一相变的意义远不止于高斯数据；事实上，它与图1 和2中的真实数据实验相吻合。至于双重下降现象，我们认为，我们的定理带来了一个非常有趣的推论：上述理论研究的适用范围远远超出了简单的高斯假设。

泛化中的非普适行为——最后，我们在预处理步骤中实现了随机特征以外的不同变换：我们考虑了小波散射变换 [47], 正交随机投影 [48],，甚至完全不进行变换。由于没有对随机投影进行打乱，而且数据是复杂的，可能更符合异方差而非同方差的情况，我们预期普适性的形式会较弱。图3展示了由随机教师向量生成标签的岭回归结果。定理3.6中的强普适性结论不对数据作任何假设，因此仍然成立；我们也观察到训练数据完美重合。然而，正如预期，泛化误差相对于高斯行为呈现出明显偏差。推测这些变换对数据的齐次化程度不足，因而测试误差不满足高斯普适性。

4.2 相关教师

此前，我们证明，即使数据具有结构，采用随机教师模型的岭回归也可能产生普适行为。现在我们考虑其对偶任务：采用简单的齐次数据模型，并研究相关的目标权重。事实上，近期一些研究考察了高斯混合与高斯模型之间缺乏普适性的实例 [49, 50]；我们希望沿着这一方向，采用定理3.7中所述的设定：考虑一个双簇高斯混合模型，其簇均值方向相反，范数为 $\boldsymbol{\mu}$，且协方差相同 Σ。我们构造一系列学习任务，在固定(α, λ) 的情况下，按以下方式改变教师向量与簇均值之间的重叠程度：

$$\boldsymbol{\theta}_0(\Omega) = \Omega\boldsymbol{\mu}_\perp + \sqrt{1-\Omega^2}\boldsymbol{\mu} \tag{29}$$

$$\text{such that} \qquad \boldsymbol{\mu}^\top\boldsymbol{\mu}_\perp = 0 \tag{30}$$

结果见图 4，涵盖岭回归和逻辑回归。结果清楚地表明，目标权重与混合分布均值之间即使只有微弱的相关性，也会破坏高斯普适性。图 4给出了该结果的直观几何解释：随着我们减小Ω，生成的标签会越来越与数据分布相关，

因此，正如定理 3.7 所预测的那样，高斯预测中会出现修正因子。关于相关教师的更详细讨论，请参阅附录 C。我们由此得出结论：在标签与结构相关的任务中，普适性会被破坏。然而请注意，正如推论 3.8所证明的那样，GMM 预测与高斯预测之间的差异会趋于零，对任意 相关性度量 Ω ，当 $\lambda \to 0^{+}$时，普适性便会恢复，正如定理 3.6所示。

5 致谢

感谢 Yatin Dandi、Federica Gerace、Sebastian Goldt、Denny Wu 和 Lenka Zdeborová 提供的宝贵讨论。我们感谢瑞士国家科学基金会 SNFS OperaGOST 项目的资助， 200021_200390 以及法国国家科学研究中心（*CNRS*）“选择法国——*CNRS AI* 新秀人才”计划的支持。

参考文献

[1] Peter L. Bartlett、Philip M. Long、Gá bor Lugosi 和 Alexander Tsigler。《线性回归中的良性过拟合》。《美国国家科学院院刊》，117(48):30063–30070，2020年4月。[2] Trevor Hastie、Andrea Montanari、Saharon Rosset 和 Ryan J. Tibshirani。《高维无岭最小二乘插值中的意外发现》。《统计年刊》，50(2):949–986，2022年4月。[3] Denny Wu 和 Ji Xu。《过参数化线性回归中的最优加权 \ell_2正则化》。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，《神经信息处理系统进展》，第33卷，第10112–10123页。Curran Associates, Inc.，2020年。[4] Vidya Muthukumar、Adhyyan Narang、Vignesh Subramanian、Mikhail Belkin、Daniel Hsu 和 Anant Sahai。《过参数化情形下的分类与回归：损失函数是否重要？》《机器学习研究期刊》，22(222):1–69，2021年。[5] Ke Wang 和 Christos Thrampoulidis。《高斯混合分布二分类中的良性过拟合》。载于*ICASSP 2021 - 2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第4030–4034页，2021年。[6] Ohad Shamir。《良性过拟合的隐式偏差》。载于 Po-Ling Loh 和 Maxim Raginsky 编，《第三十五届学习理论会议论文集》，《机器学习研究论文集》第178卷，第448–478页。PMLR，2022年7月2–5日。[7] Elizabeth Gardner 和 Bernard Derrida。《神经网络模型的最优存储特性》。《物理学杂志*A*：数学与一般物理》，21(1):271，1988年。[8] Anders Krogh 和 John A Hertz。《简单的权重衰减能够改善泛化》。载于《神经信息处理系统进展》，第950–957页，1992年。[9] David Donoho 和 Jared Tanner。《高维几何中相变的观测普适性及其对现代数据分析与信号处理的启示》。《英国皇家学会哲学汇刊*A*辑：数学、物理与工程科学》，367(1906):4273–4293，2009年。[10] Emmanuel J Candès、Pragya Sur 等。《高维逻辑回归中最大似然估计存在性的相变》。统计年刊，48(1):27–42，2020年。[11] Andrea Montanari 和 Phan-Minh Nguyen。《弹性网络误差的普适性》。载于*2017*年*IEEE*信息论国际研讨会（*ISIT*），第2338–2342页，2017年。[12] Ashkan Panahi 和 Babak Hassibi。《大规模正则化最小二乘解的普适分析》。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，《神经信息处理系统进展》，第30卷。Curran Associates, Inc.，2017年。[13] Song Mei 和 Andrea Montanari。《随机特征回归的泛化误差：精确渐近行为与双下降曲线》，2019年。[14] Federica Gerace、Bruno Loureiro、Florent Krzakala、Marc Mezard 和 Lenka Zdeborova。《随机特征与隐流形模型学习中的泛化误差》。载于 Hal Daumé III 和 Aarti Singh 编，第*37*届国际机器学习会议论文集，《机器学习研究论文集》第119卷，第3452–3462页。PMLR，2020年7月13–18日。[15] Sebastian Goldt、Bruno Loureiro、Galen Reeves、Florent Krzakala、Marc Mézard 和 Lenka Zdeborová。《浅层神经网络学习中生成模型的高斯等价性》。载于*MSML*，2021年。[16] Hong Hu 和 Yue M. Lu。《随机特征高维学习的普适性定律》。*IEEE*信息论汇刊，第1–1页，2022年。[17] Andrea Montanari 和 Basil N. Saeed。《经验风险最小化的普适性》。载于 Po-Ling Loh 和 Maxim Raginsky 编，《第三十五届学习理论会议论文集》，《机器学习研究论文集》第178卷，第4310–4312页。PMLR，2022年7月2–5日。[18] Federica Gerace、Florent Krzakala、Bruno Loureiro、Ludovic Stephan 和 Lenka Zdeborová。《高维随机标签线性分类器的高斯普适性》，2022年。

[19] Bruno Loureiro、Cedric Gerbelot、Hugo Cui、Sebastian Goldt、Florent Krzakala、Marc Mezard 和 Lenka Zdeborová。具有教师-学生模型的真实数据集上的通用特征映射学习曲线。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第34卷，第18137–18151页。Curran Associates, Inc.，2021年。[20] Arthur Jacot、Berfin Simsek、Francesco Spadaro、Clement Hongler 和 Franck Gabriel。核对齐风险估计器：从训练数据预测风险。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，神经信息处理系统进展，第33卷，第15568–15578页。Curran Associates, Inc.，2020年。[21] Blake Bordelon、Abdulkadir Canatar 和 Cengiz Pehlevan。核回归和宽神经网络中依赖谱的学习曲线。2020年。[22] Bruno Loureiro、Gabriele Sicuro、Cédric Gerbelot、Alessandro Pacco、Florent Krzakala 和 Lenka Zdeborová。使用广义线性模型学习高斯混合：高维下的精确渐近行为。2021年。[23] Edgar Dobriban 和 Stefan Wager。预测的高维渐近行为：岭回归与分类。统计年刊，46(1):247–279，2018年。[24] Christos Thrampoulidis、Ehsan Abbasi 和 Babak Hassibi。高维正则化 M-估计器的精确误差分析。*IEEE*信息论汇刊，64(8):5592–5628，2018年8月。[25] Michael Celentano、Andrea Montanari 和 Yuting Wei。具有一般高斯设计的套索回归及其在假设检验中的应用，2020年。[26] Xiaoyi Mai、Zhenyu Liao 和 Romain Couillet。逻辑回归的大规模分析：渐近性能与新见解。载于 *ICASSP 2019 - 2019 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第3357–3361页，2019年5月。ISSN：2379-190X。[27] Francesca Mignacco、Florent Krzakala、Yue M. Lu 和 Lenka Zdeborová。正则化在高维含噪高斯混合分类中的作用。2020年。[28] Zeyu Deng、Abla Kammoun 和 Christos Thrampoulidis。高维二元线性分类的双下降模型。信息与推断：*IMA*期刊，11(2):435–495，2022年6月。[29] Christos Thrampoulidis、Samet Oymak 和 Mahdi Soltanolkotabi。多类分类的理论见解：高维渐近视角。载于 第*34*届神经信息处理系统国际会议论文集，NIPS'20，第8907–8920页，美国纽约州雷德胡克，2020年12月。Curran Associates Inc.。[30] Sebastian Goldt、Marc Mézard、Florent Krzakala 和 Lenka Zdeborová。神经网络中数据结构对学习影响的建模：隐流形模型。《物理评论*X*》，10:041044，2020年12月。[31] Dominik Schröder、Hugo Cui、Daniil Dmitriev 和 Bruno Loureiro。深度随机特征学习的确定性等价与误差普适性，2023年。[32] Noureddine El Karoui。使用随机矩阵理论估计大维协方差矩阵的谱。统计年刊，第2757–2790页，2008年。[33] Song Mei、Theodor Misiakiewicz 和 Andrea Montanari。随机特征与核方法的泛化误差：超收缩性与核矩阵集中。《应用与计算调和分析》，59:3–84，2022年。调和分析与机器学习专题。[34] Hugo Cui、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborová。核回归中的泛化误差率：从无噪区域到含噪区域的交叉转变。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第34卷，第10131–10143页。Curran Associates, Inc.，2021年。[35] Hugo Cui、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborová。源条件与容量条件下核分类的误差率，2022年。[36] Jimmy Ba、Murat A. Erdogdu、Taiji Suzuki、Zhichao Wang、Denny Wu 和 Greg Yang。特征学习的高维渐近行为：一次梯度步如何改善表征，2022年。

[37] Stefano Spigler、Mario Geiger 和 Matthieu Wyart。核方法的渐近学习曲线：经验数据与教师–学生范式。《统计力学杂志：理论与实验》，2020(12):124001，2020年。[38] Alexander Wei、Wei Hu 和 Jacob Steinhardt。不仅仅是玩具：随机矩阵模型预测现实世界神经表征的泛化方式。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，第*39*届国际机器学习会议论文集，第162卷《style id='7'>机器学习研究论文集》，第23549–23588页。PMLR，2022年7月17–23日。[39] David L Donoho、Arian Maleki 和 Andrea Montanari。用于压缩感知的消息传递算法。《美国国家科学院院刊》，106(45):18914–18919，2009年。[40] Elisabetta Cornacchia、Francesca Mignacco、Rodrigo Veiga、Cédric Gerbelot、Bruno Loureiro 和 Lenka Zdeborová。多类教师-学生感知机的学习曲线。《机器学习：科学与技术》，2022年。[41] Ehsan Abbasi、Fariborz Salehi 和 Babak Hassibi。从线性测量中学习的普适性。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett 编，神经信息处理系统进展，第32卷。Curran Associates, Inc.，2019年。[42] Ali Rahimi 和 Benjamin Recht。大规模核机器的随机特征。载于 神经信息处理系统进展，第20卷。Curran Associates, Inc.，2007年。[43] Manfred Opper 和 Wolfgang Kinzel。泛化统计力学。载于 《神经网络模型 *III*》，第151–209页。施普林格，1996年。[44] Mikhail Belkin、Daniel Hsu、Siyuan Ma 和 Soumik Mandal。调和现代机器学习实践与经典偏差–方差权衡。《美国国家科学院院刊》，116(32):15849–15854，2019年7月。[45] S Spigler、M Geiger、S d'Ascoli、L Sagun、G Biroli 和 M Wyart。欠参数化到过参数化的堵塞转变影响深度学习中的泛化。《物理学杂志 *A*：数学与理论》，52(47):474001，2019年10月。[46] Thomas M Cover。线性不等式系统的几何与统计性质及其在模式识别中的应用。《*IEEE*电子计算机汇刊》，(3):326–334，1965年。[47] Mathieu Andreux、Tomás Angles、Georgios Exarchakis、Roberto Leonarduzzi、Gaspar Rochette、Louis Thiry、John Zarka、Stéphane Mallat、Joakim Andén、Eugene Belilovsky、Joan Bruna、Vincent Lostanlen、Muawiz Chaudhary、Matthew J. Hirn、Edouard Oyallon、Sixin Zhang、Carmine Cella 和 Michael Eickenberg。Kymatio：Python中的散射变换。《机器学习研究期刊》，21(60):1–6，2020年。[48] Krzysztof M Choromanski、Mark Rowland 和 Adrian Weller。结构化随机正交嵌入的非理性高效性。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。[49] Umberto M. Tomasini、Antonio Sclocchi 和 Matthieu Wyart。拉普拉斯核岭回归谱偏置预测的失效与成功：低维数据的情形。第21548–21583页。PMLR，2022年6月。[50] Alessandro Ingrosso 和 Sebastian Goldt。神经网络中卷积结构的数据驱动的涌现。《美国国家科学院院刊》，119(40):e2201854119，2022年10月。[51] Marc Mezard、Giorgio Parisi 和 Miguel Angel Virasoro。《自旋玻璃理论及其拓展：副本方法及其应用导论》。世界科学出版社，1987年11月。Google-Books-ID：DwY8DQAAQBAJ。[52] Michel Talagrand。什么是量子场论？剑桥大学出版社，剑桥，2022年。[53] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。Scikit-learn：Python中的机器学习。《机器学习研究期刊》，12:2825–2830，2011年。

A 复制计算

A.1 定理的正式表述

W我们首先给出命题3.1的完整表述。考虑如下形式的最小化问题

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{\boldsymbol{d}}}{\arg\min} \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^{\nu}, \boldsymbol{\theta}^{\top}\boldsymbol{x}^{\nu}\right) + r(\boldsymbol{\theta}), \tag{31}$$

其中，数据（$\boldsymbol{x}^{\nu}$， y^{ν}）根据以下高斯混合模型生成：

$$\boldsymbol{x}^{\nu} \sim \sum_{c\in\mathcal{C}} p_c \,\mathcal{N}\left(\frac{\boldsymbol{\mu}_c}{\sqrt{d}}, \Sigma_c\right) \quad \text{and} \quad y^{\nu} \sim P_0(\cdot \,|\, \boldsymbol{\theta}_0^{\top}\boldsymbol{x}^{\nu}) \tag{32}$$

并假设以下条件：

1. 函数 ℓ（y， $\cdot$）和 r 连续且强制性，且函数 ℓ（y， $\cdot$）$+r$（$\cdot$）强凸；2. 协方差矩阵 Σ_c 为正定，且其谱范数一致有界；3. 均值 $\boldsymbol{\mu_c}/\sqrt{d}$ 和教师向量 $\boldsymbol{\theta}_0$ 一致有界；4. 簇的数量 $|\mathcal{C}|$ 有限；5. 分布 P_0 为次高斯，且其范数一致有界。

那么，当 n， $d\to\infty$ 且 $n/d\to\alpha>0$ 时，我们有

$$\begin{aligned}
\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \sum_{c\in\mathcal{C}} p_c\, \mathbb{E}_{\omega_c^{(s)},\omega_c^{(t)},y}\left[\ell\left(y, \text{prox}_{V_c^{\star}\ell(y,\cdot)}(\omega_c^{(s)})\right)\right] &&=: \varepsilon_{\text{tr}}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \\
\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \sum_{c\in\mathcal{C}} p_c\, \mathbb{E}_{\omega_c^{(s)},\omega_c^{(t)},y}\left[\ell\left(y, \omega_c^{(s)}\right)\right] &&=: \varepsilon_{\text{tr}}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}})
\end{aligned} \tag{33}$$

其中

$$\begin{pmatrix}\omega_c^{(s)}\\ \omega_c^{(t)}\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\pi_c\\ h_c^{\star}\end{bmatrix}, \begin{bmatrix}\rho & m_c^{\star}\\ m_c^{\star} & q_c^{\star}\end{bmatrix}\right) \quad \text{and} \quad y\sim P_0\left(\cdot \,|\, \omega_c^{(t)}\right). \tag{34}$$

近端函数为

$$\text{prox}_f(x) = \min_{z\in\mathbb{R}}\left[\frac{1}{2}(z-x)^2 + f(z)\right] \tag{35}$$

方程中使用的重叠量定义如下：

$$\rho_c = \frac{1}{d}\boldsymbol{\theta}_0^{\top}\Sigma_c\boldsymbol{\theta}_0, \quad \pi_c = \frac{1}{d}\boldsymbol{\theta}_0^{\top}\boldsymbol{\mu}_c, \tag{36}$$

且 $(V_c^{\star},\ q_c^{\star},\ m_c^{\star},\ h_c^{\star})_{c\in\mathcal{C}}$是以下自洽复制鞍点方程组的唯一不动点：

$$\begin{cases}
\hat{V}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right)\partial_{\omega} f_{\ell}(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_{\ell}(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right]\\
\hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \partial_{\omega}\mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_{\ell}(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_{\ell}(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]
\end{cases} \tag{37}$$

$$\begin{cases}
V_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\hat{\boldsymbol{\eta}}^{\top}\hat{q}_c^{-1/2}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right]\\
q_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\hat{\boldsymbol{\eta}}^{\top}\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
m_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\boldsymbol{\theta}_0^{\top}\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
h_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\boldsymbol{\mu}_c^{\top}\hat{\boldsymbol{\eta}}\right]
\end{cases}$$

并且我们定义了以下辅助函数：

$$\mathcal{Z}_0(y,\omega,V)=\int\frac{\mathrm{d}\lambda}{\sqrt{2\pi V}}P_0(y|\lambda)\,e^{-\frac{(\lambda-\omega)^2}{2V}} \tag{38}$$

$$f_\ell(y,\omega,V)=-V^{-1}\partial_\omega\mathcal{M}_{V\ell(y,\cdot)}(\omega)=V^{-1}\left(\mathrm{prox}_{V\ell(y,\cdot)}(\omega)-\omega\right). \tag{39}$$

^ ^ 最后，辅助变量$\hat{\boldsymbol{\eta}}$是$\{\boldsymbol{\xi}_c,V_c,\hat{m}_c,\hat{q}_c,h_c\}$的函数：

$$\hat{\boldsymbol{\eta}}=\mathrm{prox}_{r(\hat{\Sigma}^{-1/2}\cdot)}\left(\hat{\Sigma}^{-1/2}\left(\sum_{c\in\mathcal{C}}\hat{\boldsymbol{\mu}}_c+\sqrt{\hat{q}_c}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right)\right) \tag{40}$$

with

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{c}}=\hat{h}_c\boldsymbol{\mu}_{\boldsymbol{c}}+\hat{m}_c\Sigma_c\boldsymbol{\theta}_0\quad\text{and}\quad\hat{\Sigma}=\sum_{c\in\mathcal{C}}\hat{V}_c\Sigma_c \tag{41}$$

我们不会证明这个定理，因为其证明实际上与 [22, 40]中的证明等价。接下来的章节将专门使用统计物理中的所谓副本方法 来推导这些方程 [51]。这些方程的较简单特例可见附录A.9。

A.2 吉布斯测度与自由能

起　　点是为我们的复制计算定义权重 $W\in\mathbb{R}^{K\times d}$上的吉布斯测度　　：

$$\mu_\beta(\boldsymbol{\theta})=\frac{1}{\mathcal{Z}_\beta}e^{-\beta\left[\sum_{\nu=1}^n\ell\left(y^\nu,\boldsymbol{\theta}^\top\boldsymbol{x}^\nu\right)+\lambda r(\boldsymbol{\theta})\right]}=\frac{1}{\mathcal{Z}_\beta}e^{-\beta r(\boldsymbol{\theta})}\prod_{\nu=1}^n e^{-\beta\ell(y^\nu,\boldsymbol{\theta}^\top\boldsymbol{x}^\nu)} \tag{42}$$

其中 $\beta>0$ 是一个我们最终希望令其趋于 $\beta\to0^+$的参数，而 $\mathcal{Z}_\beta\in\mathbb{R}$ 是配分函数（吉布斯测度的归一化常数）。为方便起见，我们定义以下有用记号：

$$P_{\boldsymbol{\theta}}(\boldsymbol{\theta})\equiv e^{-\beta r(\boldsymbol{\theta})},\qquad\qquad P_\ell(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu)\equiv e^{-\beta\ell(y^\nu,\boldsymbol{\theta}^\top\boldsymbol{x}^\nu)} \tag{43}$$

这使我们可以将吉布斯测度写为：

$$\mu_\beta(\boldsymbol{\theta})=\frac{1}{\mathcal{Z}_\beta}P_{\boldsymbol{\theta}}(\boldsymbol{\theta})\prod_{\nu=1}^n P_\ell(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu) \tag{44}$$

当 $\beta\to\infty$时，吉布斯测度 μ_β 将集中在使经验风险最小的 $\boldsymbol{\theta}$ (1) 附近。因此，自由能密度定义为

$$-\beta f_\beta=\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\log\mathcal{Z}_\beta \tag{45}$$

其中，在 $n/d=\alpha$ 固定时取极限，期望是对数据分布求取的，它将集中在最小风险附近。为了显式地计算期望，我们使用复制技巧：

$$\log\mathcal{Z}_\beta=\lim_{s\to0^+}\frac{1}{s}\partial_s\mathcal{Z}_\beta^s|_{s=0} \tag{46}$$

因此，复制计算归结为计算平均复制配分函数：

$$\mathbb{E}\mathcal{Z}_\beta^s=\mathbb{E}\left[\int_{\mathbb{R}^d}\mathrm{d}\boldsymbol{\theta}\;P_{\boldsymbol{\theta}}(\boldsymbol{\theta})\prod_{\nu=1}^n P_\ell\left(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu\right)\right]^s \tag{47}$$

$$=\prod_{\nu=1}^n\mathbb{E}_{(\boldsymbol{x}^\nu,y^\nu)}\int\prod_{a=1}^s\mathrm{d}\boldsymbol{\theta}^a\;P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a)P_\ell\left(y^\nu|\boldsymbol{\theta}^{a\top}\boldsymbol{x}^\nu\right) \tag{48}$$

A.3 对数据取平均

与通常的二元高斯混合相比，这一计算的主要区别在于，标签由教师目标函数生成。换言之，数据的联合分布为：

$$P_{\boldsymbol{\theta}_0}(\boldsymbol{x}, y) = P_0(y|\tau) \sum_{c\in\mathcal{C}} \rho_c \, \mathcal{N}\left(\boldsymbol{x}; \frac{\boldsymbol{\mu}_c}{\sqrt{d}}, I_d\right) \tag{49}$$

其中 $P_0(y|\tau)$ 是由教师激活函数 $f_0(\tau)$诱导的概率分布。(48) 中的表达式于是可以写成：

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a) \left(\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{x} P_0(\boldsymbol{y}|\boldsymbol{\theta}_0^\top \boldsymbol{x}) P_{\boldsymbol{x}}(\boldsymbol{x}) \prod_{a=1}^s P_\ell\left(y^\nu | \boldsymbol{\theta}^{a\top} \boldsymbol{x}^\nu\right) \right)^n \tag{50}$$

我们通过引入局部场 $\tau = \boldsymbol{\theta}_0^\top \boldsymbol{x}$ 和 $\lambda^a = \boldsymbol{\theta}^{a\top} \boldsymbol{x}$进行变量替换。局部场的分布本身将是一个高斯混合分布，我们可以计算每个模态的矩：

$$\begin{aligned}
\pi_c &\equiv \mathbb{E}_{\boldsymbol{x}^\nu}\left[\tau_c\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c \\
\rho_c &\equiv \mathrm{Var}_{\boldsymbol{x}^\nu}\left[\tau_c\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}_0 \\
h_c^a &\equiv \mathbb{E}_{\boldsymbol{x}^\nu}\left[\lambda_c^a\right] = \frac{1}{d}\boldsymbol{\theta}^{a\top} \boldsymbol{\mu}_c \\
q_c^{ab} &\equiv \mathrm{Cov}_{\boldsymbol{x}^\nu}\left[\lambda_c^a, \lambda_c^b\right] = \frac{1}{d}\boldsymbol{\theta}^{a\top} \Sigma_c \boldsymbol{\theta}^b \\
m_c^a &\equiv \mathrm{Cov}_{\boldsymbol{x}^\nu}\left[\tau_c, \lambda_c^a\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a
\end{aligned} \tag{51}$$

等价地，局部场分布是以下低维高斯混合分布：

$$P(\tau, \boldsymbol{\lambda}) = \sum_{c\in\mathcal{C}} p_c \, \mathcal{N}\left(\begin{bmatrix} \pi_c \\ \boldsymbol{h}_c \end{bmatrix}, \begin{bmatrix} \rho_c & \boldsymbol{m}_c^\top \\ \boldsymbol{m}_c & Q_c \end{bmatrix} \right), \tag{52}$$

其中，我们定义了向量 $\boldsymbol{m}$，$\boldsymbol{h} \in \mathbb{R}^s$ 其分量为 m^a、h^a，以及矩阵 $Q \in \mathbb{R}^{s\times s}$，其分量为 q^{ab}。因此，只需对局部场分布进行积分，我们便可以将配分函数因式分解：

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a) \left[\int \mathrm{d}y \int \mathrm{d}\tau \, P_0(y|\tau) \int \left(\prod_{a=1}^s \mathrm{d}\lambda^a P_\ell(y|\lambda^a) \right) P(\tau, \boldsymbol{\lambda} | \boldsymbol{m}, \boldsymbol{h}, q) \right]^n \tag{53}$$

A.4 将其写成鞍点问题

注意，(π_c, ρ_c) 是问题中固定的输入。根据傅里叶变换的论证，我们可以写成

$$\delta\left(m_c^a - \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right) = \int_{i\mathbb{R}} \frac{\mathrm{d}\hat{m}_c^a}{2\pi} \, e^{\hat{m}_c^a \left(d m_c^a - \boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right)}, \tag{54}$$

其中积分路径是虚轴。对 q_c^{ab} 和 h_c^a 进行同样的论证，可得

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{c\in\mathcal{C}} \prod_{a=1}^s \frac{\mathrm{d}m_c^a \mathrm{d}\hat{m}_c^a}{2\pi} \frac{\mathrm{d}h_c^a \mathrm{d}\hat{h}_c^a}{2\pi} \prod_{1\le a\le b\le s} \frac{\mathrm{d}q_c^{ab} \mathrm{d}\hat{q}_c^{ab}}{2\pi} e^{d\Phi_{\mathrm{rs}}^{(s)}(\boldsymbol{m}, \hat{\boldsymbol{m}}, q, \hat{q})} \tag{55}$$

其中我们定义了自由能势：

$$\begin{aligned}
\Phi_{\mathrm{rs}}^{(s)}(\boldsymbol{m}, \hat{\boldsymbol{m}}, q, \hat{q}) &= \sum_{c\in\mathcal{C}} \left[\sum_{a=1}^s \hat{m}_c^a m_c^a + \sum_{a=1}^s \hat{h}_c^a h_c^a + \sum_{1\le a\le b\le s} \hat{q}_c^{ab} q_c^{ab} \right] - \alpha \Psi_g^{(s)}(\boldsymbol{m}, \boldsymbol{h}, q) - \Psi_\theta^{(s)}(\hat{m}, \hat{h}, \hat{q}) \\
\Psi_\ell^{(s)}(\boldsymbol{m}, \boldsymbol{h}, q) &\equiv \log \int \mathrm{d}y \int \mathrm{d}\tau \, P_0(y|\tau) \int \prod_{a=1}^s \mathrm{d}\lambda^a P_\ell(y|\lambda^a) P(\tau, \boldsymbol{\lambda} | \boldsymbol{m}, \boldsymbol{h}, q) \\
\Psi_\theta^{(s)}(\hat{\boldsymbol{m}}, \hat{\boldsymbol{h}}, \hat{q}) &\equiv \frac{1}{d} \log \int \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{c\in\mathcal{C}} e^{\sum_{a=1}^s \left[\hat{h}_c^a \boldsymbol{\theta}^{a\top} \boldsymbol{\mu}_c + \hat{m}^a \boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right] + \sum_{1\le a\le b\le s} \hat{q}_c^{ab} \boldsymbol{\theta}^{a\top} \Sigma_c \boldsymbol{\theta}^b}
\end{aligned} \tag{56}$$

因此，当我们取 $d \to \infty$极限时，可以应用鞍点法 [51] 来计算 f_β

$$-\beta f_\beta = \underset{\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{h},\hat{\boldsymbol{h}},q,\hat{q}}{\operatorname{extr}} \lim_{s\to 0^+} \frac{\Phi^{(s)}(\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{h},\hat{\boldsymbol{h}},q,\hat{q})}{s} \tag{57}$$

ˆ注：A.1。我们将参数 ˆq、ˆm、 h 引入为纯虚数，但优化问题将它们视为实数。这是因为函数 Φ 是全纯函数，因此积分与积分路径无关。更多细节请参见 [52]。

A.5 复制对称假设

为了推进 $s \to 0^+$ 极限，我们作出如下复制对称假设：

$$\begin{aligned}
m_c^a &= m_c & \hat{m}_c^a &= \hat{m}_c, & a &= 1,\cdots,s \\
h_c^a &= h_c & \hat{h}_c^a &= \hat{h}_c, & a &= 1,\cdots,s \\
q_c^{aa} &= r_c, & \hat{q}_c^{aa} &= -\frac{1}{2}\hat{r}_c, & a &= 1、\ \cdots、\ s \\
q_c^{ab} &= q_c, & \hat{q}_c^{aa} &= \hat{q}_c, & 1 &\le a < b \le s
\end{aligned} \tag{58}$$

对所有 $c \in \mathcal{C}$ 均成立。由于 ℓ 和 r 是凸函数，函数 $\Phi^{(s)}$ 有唯一的鞍点，因此它必定与复制对称假设一致。我们还定义

$$V_c = r_c - q_c, \qquad \text{and} \qquad \hat{V}_c = \hat{r}_c - \hat{q}_c \tag{59}$$

将此假设代入上式后，我们可以取 $s \to 0^+$ 极限。这需要用到一种经典但计算量很大的方法，称为 *Hubbard-Stratonovich*变换；详情见 [14], 附录 C。我们仅在此列出最终结果：

$$-\beta f_\beta = \underset{\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}}{\operatorname{extr}} \Phi_{\text{rs}}(\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}) \tag{60}$$

复制自由能 Φ_{rs} 由以下公式给出：

$$\Phi_{\text{rs}}(\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}) = \sum_{c\in\mathcal{C}} \left[\frac{1}{2}\left(\hat{V}_c q_c - \hat{q}_c V_c\right) - \frac{1}{2}\hat{V}_c V_c + \hat{m}_c m_c + \hat{h}_c h_c\right] \tag{61}$$

$$- \alpha \Psi_\ell(\{m_c,h_c,q_c,V_c\}) - \Psi_{\boldsymbol{\theta}}(\{\hat{m}_c,\hat{h}_c,\hat{q}_c,\hat{V}_c\}) \tag{62}$$

其中，我们将来自损失（Ψ_ℓ）和正则化（$\Psi_{\boldsymbol{\theta}}$）的贡献分解如下：

$$\Psi_\ell = \sum_{c\in\mathcal{C}} \mathbb{E}_{\xi_c \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)} \int \mathrm{d}y \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi + \pi_c, \rho - \frac{m_c^2}{q_c}\right) \log \mathcal{Z}_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c) \tag{63}$$

$$\Psi_{\boldsymbol{\theta}} = \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi}_c \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0},I_d)} \log \mathcal{Z}_{\boldsymbol{\theta}}\left(\sum_{c\in\mathcal{C}} \hat{V}_c\Sigma_c, \sum_{c\in\mathcal{C}} \hat{h}_c\boldsymbol{\mu}_c + \hat{m}_c\Sigma_c\boldsymbol{\theta}_0 + \sqrt{\hat{q}_c}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right), \tag{64}$$

并定义了以下辅助自由能：

$$\mathcal{Z}_{\ell/0}(y,\omega,V) = \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi V}} P_{\ell/0}(y|\lambda)\, e^{-\frac{(\lambda-\omega)^2}{2V}} \tag{65}$$

$$\mathcal{Z}_{\boldsymbol{\theta}}(A,\boldsymbol{b}) = \int \mathrm{d}\boldsymbol{\theta} e^{-\beta r(\boldsymbol{\theta})} e^{-\frac{1}{2}\boldsymbol{\theta}^\top A\boldsymbol{\theta} + \boldsymbol{b}^\top\boldsymbol{\theta}} \tag{66}$$

A.6 取零温极限

为了取 $\beta \to \infty$极限，我们进行如下缩放：

$$V_c \to \beta^{-1}V_c \qquad \hat{V}_c \to \beta\hat{V}_c \qquad \hat{q}_c \to \beta^2\hat{q}_c \qquad \hat{m}_c \to \beta\hat{m}_c. \tag{67}$$

很容易检验这种缩放对 $\beta^{-1}\Phi^{(\mathrm{rs})}$的影响，因此我们只需考虑 Ψ_ℓ 和 $\Psi_{\boldsymbol{\theta}}$。

我们从后者开始：令

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{2}\,\boldsymbol{\theta}^\top A\boldsymbol{\theta} - \boldsymbol{b}^\top\boldsymbol{\theta} + r(\boldsymbol{\theta})$$

根据拉普拉斯方法，对于任意 A，$\boldsymbol{b}$,

$$\lim_{\beta\to\infty}\frac{1}{\beta}\log Z_{\boldsymbol{\theta}}(\beta A, \beta\boldsymbol{b}) = -\inf_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\eta}) \tag{68}$$

其中

$$\boldsymbol{\eta} = \mathrm{prox}_{r(A^{1/2}\cdot)}(A^{1/2}\boldsymbol{b}). \tag{69}$$

因此， $\Psi_{\boldsymbol{\theta}}$ 中涉及的每个积分（以及之后的偏导数）都集中在其于$\hat{}\boldsymbol{\eta}$ （在40中定义）处的取值附近。对于项 Ψ_ℓ，项 $\mathcal{Z}_0$ 保持不变，并由与上文相同的推理可得

$$\lim_{\beta\to\infty}\frac{1}{\beta}\log\mathcal{Z}_\ell(y,\omega,\beta V) = -V^{-1}\mathcal{M}_{V\ell(y,\cdot)}(\omega). \tag{70}$$

A.7 鞍点方程

鞍点方程　　这些方程是通过对自由能势关于重叠　　ap
参数求导得到的。我们得到一组需要求解的自洽方程，以找到一个不动点：

$$\begin{cases}
\hat{V}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right)\partial_\omega f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right]\\
\hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \partial_\omega\mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\ \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]
\end{cases} \tag{71}$$

$$\begin{cases}
V_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\hat{\boldsymbol{\eta}}^\top\hat{q}_c^{-1/2}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right]\\
q_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\hat{\boldsymbol{\eta}}^\top\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
m_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\boldsymbol{\theta}_0^\top\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
h_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\mathbf{0},I_d)}\left[\boldsymbol{\mu}_c^\top\hat{\boldsymbol{\eta}}\right]
\end{cases} \tag{72}$$

其中所有相关量均已在附录 A.1.

A.8 训练误差和泛化误差

It 现在剩下的任务是根据自由能计算训练误差和泛化误差。回想一下，　　t

$$\varepsilon_{\mathrm{tr}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{\nu=1}^{n}\ell\left(y^\nu, \hat{\boldsymbol{\theta}}^\top\boldsymbol{x}^\nu\right),$$

根据自由能的定义，我们可以写成（45）：

$$\lim_{n\to\infty}\varepsilon_{\mathrm{tr}} = \lim_{\beta\to\infty}\partial_\beta f_\beta - r(\hat{\boldsymbol{\theta}}).$$

显式计算导数，并再次对数据取平均，得到

$$\lim_{n\to\infty}\varepsilon_{\mathrm{tr}} = -\lim_{\beta\to\infty}\partial_\beta\Psi_\ell,$$

w这里 Ψ_ℓ 是式（63）中定义的损失所贡献的自由能。将该导数明确写出　　，

$$-\partial_\beta\Psi_\ell = \sum_{c\in\mathcal{C}}\mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\int_{\mathbb{R}}\mathrm{d}y\frac{\mathcal{Z}_0\left(y,\omega_c^{(0)},\rho-\frac{m_c^2}{q_c}\right)}{\mathcal{Z}_\ell(y,\omega_c^{(\ell)},V_c)}\underbrace{\int_{\mathbb{R}}\frac{\mathrm{d}\lambda}{\sqrt{2\pi V_c}}e^{-(\lambda-\omega_c^{(\ell)})V_c^{-1}(\lambda-\omega_c^{(\ell)},V_c))-\beta\ell(y,\lambda)}\ell(y,\lambda)}_{\tilde{\mathcal{Z}}_\ell(y,\omega_c^{(\ell)},V_c)} \tag{73}$$

其中

$$\omega_c^{(0)} = \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c \quad \text{and} \quad \omega_c^{(\ell)} = h_c + \sqrt{q_c}\xi_c.$$

现在，我们可以对（67）作变量替换，并再次使用拉普拉斯近似：若

$$\eta_c = \mathrm{prox}_{V\ell(y,\cdot)}(\omega_c^{(\ell)}) \quad \text{and} \quad \mathcal{L}(\lambda) = -(\lambda - \omega_c^{(\ell)})V_c^{-1}(\lambda - \omega_c^{(\ell)}) - \beta\ell(y,\lambda),$$

then

$$\mathcal{Z}_\ell(y, \omega_c^{(\ell)}, V_c) \sim e^{-\beta\mathcal{L}(\eta_c)} \quad \text{and} \quad \tilde{\mathcal{Z}}_\ell(y, \omega_c^{(\ell)}, V_c) \sim e^{-\beta\mathcal{L}(\eta_c)}\ell(y,\eta_c),$$

所有重叠量都会集中在（71）、（72）的解附近。最后，包含 $\mathcal{Z}_0$ 的项是根据 $P_0(y \mid \tau_c)$ 对 y 求得的期望，其中

$$\tau_c = \omega_c^{(0)} + \sqrt{\rho - \frac{m_c^2}{q_c}}\xi_c^{(0)}$$

a且 $\xi_c^{(0)}$是一个与其他所有变量独立的高斯变量。综合以上各项，可得

$$\lim_{n\to\infty} \varepsilon_{\mathrm{tr}}(\hat{\boldsymbol{\theta}}) = \sum_{c\in\mathcal{C}} p_c \, \mathbb{E}_{\nu_c,\tau_c,y}\left[\ell\left(y, \mathrm{prox}_{V_c^\star \ell(y,\cdot)}(\nu_c)\right)\right] \tag{74}$$

with

$$\begin{pmatrix}\nu_c \\ \tau_c\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\pi_c \\ h_c\end{bmatrix}, \begin{bmatrix}\rho_c & m_c^\star \\ m_c^\star & q_c^\star\end{bmatrix}\right) \quad \text{and} \quad y \sim P_0(\cdot \mid \tau_c). \tag{75}$$

ˆ泛化误差更容易求得：由于 x_{new}， y_{new} 与估计器 $\boldsymbol{\theta}$ 相互独立，因此我们有

$$\lim_{n\to\infty} \varepsilon_{\mathrm{gen}}(\hat{\boldsymbol{\theta}}) = \sum_{c\in\mathcal{C}} p_c \, \mathbb{E}_{\nu_c,\tau_c,y}\left[\ell\left(y,\nu_c\right)\right] \tag{76}$$

其中 ν_c， τ_c， y 的分布与上文相同。

A.9 示例

岭惩罚 考虑上述一般方程所述的一个特例：采用岭惩罚的情形

$$r(\boldsymbol{\theta}) = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2.$$

于是有：

$$\mathrm{prox}_r(\boldsymbol{x}) = (1+\lambda)^{-1}\boldsymbol{x} \tag{77}$$

这大大简化了之前的方程。事实上，现在我们可以计算72中的每一个期望，从而得到以下不动点方程：

$$\begin{cases}
\hat{V}_c &= -\alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right)\partial_\omega f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\
\hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right] \\
\hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \partial_\omega\mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\
\hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]
\end{cases} \tag{78}$$

$$\begin{cases}
V_c &= \frac{1}{d}\,\mathrm{tr}\left[\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right] \\
q_c &= \frac{1}{d}\,\mathrm{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{q}_{c'}\Sigma_{c'} + \sum_{c',c''\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\hat{\boldsymbol{\mu}}_{c''}^\top\right)\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-2}\right] \\
m_c &= \frac{1}{d}\,\mathrm{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\boldsymbol{\theta}_0^\top\right)\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right] \\
h_c &= \frac{1}{d}\,\mathrm{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\boldsymbol{\mu}_c^\top\right)\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right]
\end{cases} \tag{79}$$

高斯协变量模型 高斯协变量模型的方程见 [19]；它们也相当于在上述方程中取 $|\mathcal{C}| = 1$。为完整起见，我们在此重述如下：

$$\begin{cases} \hat{V} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) \partial_\omega f_\ell(y, h + \sqrt{q}\xi, V) \right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V)^2 \right] \\ \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \partial_\omega \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V) \right] \\ \hat{h} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V) \right] \end{cases} \tag{80}$$

$$\begin{cases} V &= \frac{1}{d} \operatorname{tr} \left[\Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ q &= \frac{1}{d} \operatorname{tr} \left[\left(\hat{q}\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \right) \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-2} \right] \\ m &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ h &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \end{cases} \tag{81}$$

此时

$$\hat{\boldsymbol{\mu}} = \hat{h}\boldsymbol{\mu} + \hat{m}\Sigma\boldsymbol{\theta}_0 \qquad \text{and} \qquad \hat{\Sigma} = \hat{V}\Sigma. \tag{82}$$

误差为

$$\lim_{n\to\infty} \varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\nu,\tau,y} \left[\ell \left(y, \operatorname{prox}_{V^\star \ell(y,\cdot)}(\nu) \right) \right] \tag{83}$$

$$\lim_{n\to\infty} \varepsilon_{\text{gen}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\nu,\tau,y} \left[\ell \left(y, \nu \right) \right] \tag{84}$$

with

$$\begin{pmatrix} \nu \\ \tau \end{pmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \pi \\ h \end{bmatrix}, \begin{bmatrix} \rho & m^\star \\ m^\star & q^\star \end{bmatrix} \right) \quad \text{and} \quad y \sim P_0(\cdot \,|\, \tau). \tag{85}$$

岭回归 我们考虑岭回归的情形，此时

$$P_0(\tau|\Delta) = \mathcal{N}(\tau, \Delta) \qquad\qquad l(y, \hat{y}) = (y - \hat{y})^2. \tag{86}$$

在这种情况下，式（79）中的方程可进一步简化，得到

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\rho_c + \Delta + q_c - 2m_c + (h_c - \pi_c)^2) \\ \hat{m}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{h}_c &= \frac{\alpha p_c(\pi_c - h_c)}{1+V_c} \end{cases} \tag{87}$$

$$\begin{cases} V &= \frac{1}{d} \operatorname{tr} \left[\Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ q &= \frac{1}{d} \operatorname{tr} \left[\left(\hat{q}\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \right) \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-2} \right] \\ m &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ h &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \end{cases} \tag{88}$$

训练误差和泛化误差也能通过一个非常简单的表达式来表示

$$\varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star}{\alpha} \qquad\qquad \varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star (1 + V_c^\star)^2}{\alpha}. \tag{89}$$

B 高斯混合模型的普适性

在本节中，我们将证明第3 节所展示的关于高斯混合模型普适性的主要结果。

B.1 均值普适性：命题3.2的证明

我们在此重述假设。设 $\{\hat{V}_c^\star\}_{c=1}^K$ 为描述中心化高斯混合问题的（副本）鞍点方程的不动点，即方程（78）、（79）的解。我们假设教师向量和数据结构满足以下条件：

$$\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0 \qquad \forall c \in \mathcal{C} \tag{90}$$

$$\lim_{n,d\to\infty} \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_{c'} \left(\lambda + \sum_{c\in\mathcal{C}} \hat{V}_c^\star \Sigma_c\right)^{-1} \boldsymbol{\mu}_c \to 0 \qquad \forall (c,c') \in \mathcal{C}\times\mathcal{C} \tag{91}$$

并且损失和教师均为对称的：

$$\ell(x,y) = \ell(-x,-y) \tag{92}$$

$$P_0(y|\tau) = P_0(-y|-\tau). \tag{93}$$

在鞍点方程（78）、（79）中，均值向量总是与重叠量 $\{h_c, h_c\}_{c\in\mathcal{C}}$耦合。因此，要证明命题3.2，只需证明以下结果：

引理 B.1. 若$(a+b+c)$ 成立，则 $\{h_c = h_c = 0\}_{c\in\mathcal{C}}$ 是该问题的一个不动点。

首先考虑 $\{h_c\}_{c\in\mathcal{C}}$ 的更新方程：

$$h_c = \frac{1}{d}\sum_l \mathrm{tr}\left[\left(\hat{h}_l \mu_l \mu_c^\top + \hat{m}_l \Sigma_l \boldsymbol{\theta}_0 \mu_c^\top\right)\left(\lambda I_d + \sum_{l'} \hat{V}_{l'}\Sigma_{l'}\right)^{-1}\right] \tag{94}$$

$$\underset{a+c}{\simeq} \quad \frac{1}{d}\sum_l \mathrm{tr}\left[\left(\hat{h}_l \mu_l \mu_c^\top\right)\left(\lambda I_d + \sum_{l'} \hat{V}_{l'}\Sigma_{l'}\right)^{-1}\right] \tag{95}$$

根据鞍点方程的连续性，如果在不动点 $\{h_c^\star \to 0\}_{c\in\mathcal{C}}$ 处成立，我们也很容易得到 $\{h_c^\star \to 0\}_{c\in\mathcal{C}}$。现在假设在不动点处有 $\{h_c^\star = 0\}_{c\in\mathcal{C}}$。我们利用更新函数中固有的对称性论证，证明在对教师函数和损失函数施加较弱假设的情况下，$\{h_c^\star = 0\}_{c\in\mathcal{C}}$。根据假设 c)，将更新方程写为：

$$\hat{h}_c = \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi + \pi_c, \rho_c - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi, V)\right] \tag{96}$$

$$\underset{h_c\to 0}{\simeq} \quad \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi, \rho_c - \frac{m_c^2}{q_c}\right) f_\ell(y, \sqrt{q_c}\xi, V_c)\right] \tag{97}$$

回顾教师测度项的定义：

$$Z_0(y,\omega,V) = \mathbb{E}_{\lambda\sim\mathcal{N}(\omega,V)}\left[P_0(y|\lambda)\right] \qquad\qquad \lambda \equiv \frac{\boldsymbol{x}^\top\boldsymbol{\theta}_0}{\sqrt{d}} \tag{98}$$

我们看到，对称性条件对标签生成施加了基本限制：

$$f) \quad \mathbb{E}_{\lambda\sim\mathcal{N}(\omega,V)}\left[P_0(y|\lambda)\right] = \mathbb{E}_{\lambda\sim\mathcal{N}(-\omega,V)}\left[P_0(-y|\lambda)\right] \tag{99}$$

因此， $\mathcal{Z}_0$ 关于其第二个自变量是偶函数。此外，我们有

$$f_\ell(y,\omega,V) = \frac{1}{V}(\mathrm{prox}_{\,V l(y,\cdot)}(\omega) - \omega) \qquad \mathrm{prox}_{\,V\ell(y,\cdot)}(w) = \arg\min_z \left[\frac{1}{2V}(z-w)^2 + \ell(y,z)\right] \tag{100}$$

关于 ℓ 的对称性条件随后意味着

$$\mathrm{prox}_{\,V\ell(-y,\cdot)}(-\omega) = -\,\mathrm{prox}_{\,V\ell(y,\cdot)}(\omega), \tag{101}$$

因此 f_ℓ 关于其第二个自变量是奇函数。需要注意的只剩下（97）是一个奇函数的高斯积分，因此等于 0。

B.2 高斯普适性：命题3.3的证明

ˆ我们证明了在命题 $\{h_c^\star, h_c^\star = 0, 0\}_{c\in\mathcal{C}}$ 的假设下，不动点满足 3.2。现在假设混合分布是齐次的：

$$\Sigma_c = \Sigma \qquad \forall c \in \mathcal{C} \tag{102}$$

于是我们有

$$\varepsilon_{gen}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma\}_{c=1}^K\right) \simeq \varepsilon_{gen}^{\mathrm{GMM}}\left(\mathbf{0}, \{\Sigma\}_{c=1}^K\right) \tag{103}$$

$$\varepsilon_{tr}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma\}_{c=1}^K\right) \simeq \varepsilon_{tr}^{\mathrm{GMM}}\left(\mathbf{0}, \{\Sigma\}_{c=1}^K\right) \tag{104}$$

但这些方程右侧对应于如下形式的分布

$$\boldsymbol{x} \sim \sum_{c\in\mathcal{C}} \mathcal{N}(\mathbf{0}, \Sigma) = \mathcal{N}(\mathbf{0}, \Sigma)!$$

这就证明了命题3.3

B.3 协方差普适性

令人惊讶的是，在正则化消失的极限 $\lambda \to 0^+$ 下，协方差对高维学习问题并不重要。我们证明，当损失存在唯一极小值点，且可以在鞍点方程中安全地取极限 $\lambda \to 0^+$ 时，协方差 Σ 会从重叠表达式中完全消失。事实上，假设ˆ极小值点 $\boldsymbol{\theta}$ 是唯一的，我们可以安全地简化如下类型的表达式：

$$\lim_{\lambda\to 0^+} \Sigma^n(\lambda I_d + \hat{V}\Sigma)^{n'} = \frac{1}{\hat{V}}\Sigma^{n-n'} \tag{105}$$

然后代入 $\lambda \to 0^+$ 方程（80）、（81）中的化简结果，我们便得到其与协方差矩阵完全无关 Σ:

$$\begin{cases} \hat{V} &= -\alpha\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi, \rho_c - \frac{m_c^2}{q_c}\right)\partial_\omega f_\ell(y, \sqrt{q_c}\xi, V_c)\right] \\ \hat{q} &= \alpha\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) f_\ell(y, \sqrt{q}\xi, V)^2\right] \\ \hat{m} &= \alpha\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \partial_\omega\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) f_\ell(y, \sqrt{q}\xi, V)\right] \end{cases} \tag{106}$$

$$\begin{cases} V &= 1 \\ q &= \hat{q} + \rho\hat{m}^2 \\ m &= \hat{m}\rho \end{cases} \tag{107}$$

至此，证明完成。

B.4 强普适性

到目前为止，除了假设2中的一般性对称性条件外，我们并未对损失作出任何具体假设。本节将证明，对于满足均值普适性条件（命题 3.2）的任意高斯混合模型 估计问题，平方损失回归具有强普适性。我们假设考虑欠参数化区域中的岭回归问题 $\alpha > 1$:

$$P_0(\tau|\Delta) = \mathcal{N}(\tau, \Delta) \qquad \ell(y, \hat{y}) = (y - \hat{y})^2 \tag{108}$$

th这些副本随后对应于方程（87）、（88）。在 $\lambda \to 0^+$ 极限下，该方程大幅简化 ly:

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\rho_c + \Delta + q_c - 2m_c) \\ \hat{m}_c &= \frac{\alpha p_c}{1+V_c} \end{cases} \tag{109}$$

$$\begin{cases} V_c &= \frac{1}{d}\,\mathrm{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right] \\ q_c &= \frac{1}{d}\sum_{l\in\mathcal{C}}^K \mathrm{tr}\left[\hat{q}_l\Sigma_l\Sigma_c(\sum_{l'\in\mathcal{C}}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \frac{1}{d}\sum_{(l,l')\in\mathcal{C}\times\mathcal{C}}\mathrm{tr}\left[\hat{m}_l\hat{m}_{l'}\Sigma_l\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Sigma_{l'}\Sigma_c(\sum_{l''\in\mathcal{C}}\hat{V}_{l''}\Sigma_{l''})^{-2}\right] \\ m_c &= \frac{1}{d}\sum_{l=1}^K \mathrm{tr}\left[\hat{m}_l\Sigma_l\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-1}\right] \end{cases} \tag{110}$$

考虑重叠量的方程 $\{q_c\}_{c\in\mathcal{C}}$:

$$q_c = \frac{1}{d}\sum_{l=1}^{K}\operatorname{tr}\left[\hat{q}_l\Sigma_l\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \frac{1}{d}\sum_{l,k=1}^{K}\operatorname{tr}\left[\hat{m}_l\hat{m}_k\Sigma_l\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\Sigma_k\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] \tag{111}$$

$$\equiv R_c + G_c \tag{112}$$

误差度量可以分解为:

$$\varepsilon_{\text{tr}} = \sum_{c\in\mathcal{C}}\frac{p_c}{(1+V_c)^2}(R_c+\Delta) + \sum_{c\in\mathcal{C}}\frac{p_c}{(1+V_c)^2}(\rho_c + G_c - 2m_c) \tag{113}$$

$$\varepsilon_{\text{gen}} = \sum_{c\in\mathcal{C}}p_c(R_c+\Delta) + \sum_{c\in\mathcal{C}}p_c(\rho_c + G_c - 2m_c) \tag{114}$$

我们重点考察第二项，并证明它等于零。观察方程 (110),(109)，注意到:

$$\hat{m}_c = \hat{V}_c \qquad \forall c\in\mathcal{C} \tag{115}$$

利用方程 (115)，我们可以简化关于 $\{\hat{m}_c, G_c\}$ 的方程:

$$G_c = \rho_c, \qquad m_c = \rho_c \qquad \forall c\in\mathcal{C} \tag{116}$$

将这些关系代入鞍点方程，我们得到:

$$\begin{cases}\hat{V}_c &= \frac{\alpha p_c}{1+V_c} = \hat{m}_c\\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\Delta + R_c)\end{cases} \tag{117}$$

$$\begin{cases}V_c &= \frac{1}{d}\operatorname{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right]\\ q_c &= \frac{1}{d}\sum_{l=1}^{K}\operatorname{tr}\left[\hat{q}_l\Sigma_l\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \rho_c \equiv R_c + \rho_c\\ m_c &= \rho_c\end{cases} \tag{118}$$

这些方程的不动点　　不再依赖于 $\{\rho_c\}_{c\in\mathcal{C}}$，因此只剩下关于 $\{\hat{V}_c, \hat{q}_c, V_c, R_c\}_{c\in\mathcal{C}}$ 的方程　　:

$$\begin{cases}\hat{V}_c &= \frac{\alpha p_c}{1+V_c}\\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\Delta + R_c)\end{cases} \tag{119}$$

$$\begin{cases}V_c &= \frac{1}{d}\operatorname{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right]\\ R_c &= \frac{1}{d}\sum_{l=1}^{K}\operatorname{tr}\left[\hat{q}_l\Sigma_l\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-2}\right]\end{cases} \tag{120}$$

通过一系列代数运算，可以得到误差的闭式表达。通过代数运算可以验证以下关系:

$$\sum_c V_c\hat{V}_c = \sum_c\frac{1}{d}\operatorname{tr}\left[\hat{V}_c\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right] = 1 \tag{121}$$

$$\sum_c R_c\hat{V}_c - V_c\hat{q}_c = \frac{1}{d}\sum_{l=1}^{K}\operatorname{tr}\left[\hat{q}_l\Sigma_l(\sum_c\hat{V}_c\Sigma_c)(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] - \frac{1}{d}\operatorname{tr}\left[\sum_c\hat{q}_c\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right] = 0 \tag{122}$$

现在，我们将式（121）中的所有量都用未加帽的重叠量表示，得到:

$$\star)\ \sum_c\frac{\alpha p_c}{1+V_c}V_c = 1 \tag{123}$$

$$\#)\ \sum_c\frac{\alpha p_c}{(1+V_c)^2}R_c = \sum_c\frac{\alpha p_c}{(1+V_c)^2}V_c\Delta \tag{124}$$

值得注意的是，训练误差和泛化误差可以写为:

$$\varepsilon_{\text{tr}} = \sum_c\frac{\hat{q}_c}{\alpha} \qquad\qquad \varepsilon_{\text{tr}} = \sum_c\frac{\hat{q}_c^\star(1+V_c^\star)^2}{\alpha} \tag{125}$$

因此，如果我们计算量 $\sum_c \hat{}q_c$，便可得出结论。将 (#) 代入上面的表达式，我们得到：

$$\sum_c \hat{q}_c = \sum_c \frac{\alpha p_c}{(1+V_c)^2}(R_c + \Delta) \tag{126}$$

$$\underset{(\#)}{=} \Delta \sum_c \frac{\alpha p_c}{(1+V_c)^2} + \Delta \sum_c \frac{\alpha p_c}{(1+V_c)^2} V_c \tag{127}$$

$$= \Delta \sum_c \frac{\alpha p_c}{1+V_c} \tag{128}$$

最后，利用关系式（⋆），我们证明该定理：

$$\varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star}{\alpha} \tag{129}$$

$$= \sum_c \frac{\hat{q}_c^\star}{\alpha} \pm \frac{\Delta}{\alpha} \tag{130}$$

$$\underset{(\star)}{=} \Delta(1 - \frac{1}{\alpha}) \tag{131}$$

另一方面，正如我们将在下一节中分析的那样，泛化误差并不符合强普适性结论（见图 5）。

C 高斯混合模型的非普适性

在上一节中，我们列举了一系列揭示高斯混合模型普适性的结果。现在我们要研究一个相反的问题：高斯混合模型何时不具有普适性？打破普适性有两种方式：a）允许数据结构具有强异质性；b）考虑与数据结构强相关的标签。本节将更详细地回顾这两种导致普适性破缺的情形。

C.1 强异质混合分布

我们在定理 3.6 中证明了岭回归训练损失的一个强普适性结论。然而，本节要说明，该定理在超出其对一般混合分布的假设范围时并不成立。我们通过分析一个反例来证明这一点：一个强异质的双簇高斯混合：

$$\Sigma_+ = \text{diag}(0.1, \dots, 0.1, 1.9, \dots, 1.9) \qquad p_+ = 0.8 \tag{132}$$

$$\Sigma_- = \text{diag}(1.9, \dots, 1.9, 0.1, \dots, 0.1) \qquad p_- = 0.2 \tag{133}$$

我们在图中展示5 异质高斯混合模型性能与高斯理论的对比：尽管训练误差如定理所预测的那样一致。3.6，泛化误差却有所不同。不过，我们注意到，预处理后的真实数据似乎足够齐次，因此其结果与精确高斯渐近行为吻合良好，正如图中所示。1。

C.2 相关教师

本节将更深入地研究定理所述的受控设定。3.7。我们考虑的估计问题不一定满足均值普适性的假设（命题。3.2），实际上，我们会精确分析放宽这些条件所带来的后果。我们对均值相反、协方差矩阵相同且正比于 I_d的双成分高斯混合模型进行岭回归。在这种情形下，经验风险最小化估计量性能的精确渐近行为具有闭式表达式。这里重述理论分析中的主要参数：

$$\rho = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma \boldsymbol{\theta}_0 = \frac{1}{d}\|\boldsymbol{\theta}_0\|_2^2 \qquad \gamma = \frac{1}{d}||\boldsymbol{\mu}||_2^2 \qquad \Delta = \mathbb{E}_\xi \xi^2 \qquad \hat{\boldsymbol{\mu}}_\pm = \hat{h}_\pm \boldsymbol{\mu}_\pm + \hat{m}_\pm \Sigma \boldsymbol{\theta}_0 \tag{134}$$

在这一简单设定下，我们无需对教师作任何假设，即可简化方程（80）、（81）中的一般混合方程。我们首先将该方程映射为一个单高斯问题 $\mathcal{G}_{\boldsymbol{\theta},\ I_d,\ \boldsymbol{\mu}}$。我们需要对证明稍作修改

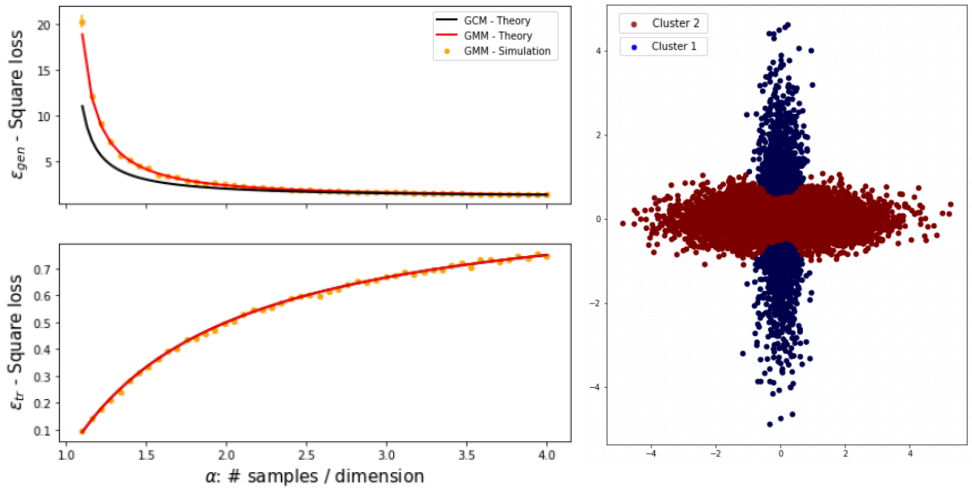

图5：高斯混合模型上的岭回归训练误差与泛化误差，对应的高斯混合模型定义见公式（132）。**左图**：我们比较高斯渐近结果（黑色实线）与高斯混合模型结果（红色实线）的性能；橙色圆点表示模拟结果，与定理所预测的理论预测一致。3.1。**右图**：来自异质高斯混合的10000 个样本的二维玩具图。

命题 3.3 见第 B.2节，实际上，我们不能盲目地假设均值重叠量 $\{h_+, h_-\}$ 为零，但可以证明，在不动点处，重叠量满足：

$$\hat{h}_+^\star = -\hat{h}_-^\star, \qquad \hat{m}_+^\star = \hat{m}_-^\star \qquad \to \qquad \hat{\boldsymbol{\mu}}_+^\star = \hat{\boldsymbol{\mu}}_-^\star \tag{135}$$

$$\hat{V} = \hat{V}_+ + \hat{V}_- \qquad \hat{q} = \hat{q}_+ + \hat{q}_- \qquad \hat{m} = \hat{m}_+ + \hat{m}_- \qquad \hat{h} = \hat{h}_+ - \hat{h}_- \tag{136}$$

$$V = V_+ = V_- \qquad q = q_+ = q_- \qquad m = m_+ = m_- \qquad h = h_+ = -h_- \tag{137}$$

上述混合分布的复制子方程不动点，可映射为单高斯问题中关于重叠量 $\{V, m, q, h, \hat{V}, \hat{m}, \hat{q}, \hat{h}\}$ 的不动点。现在，我们可以进一步简化这些方程：代入关于协方差的假设后，所有迹项都可化简，得到：

$$\begin{cases} V &= \frac{1}{\lambda+\hat{V}} \\ q &= V^2\left(\hat{q} + \rho\hat{m}^2 + 2\pi\hat{m}\hat{h} + \gamma\hat{h}^2\right) \\ m &= V\left(\rho\hat{m} + \pi\hat{h}\right) \\ h &= V\left(\gamma\hat{h} + \pi\hat{m}\right) \end{cases} \tag{138}$$

这些方程实际上是可解的！定义

$$\eta := V\hat{V} = \frac{\alpha V}{1+V};$$

现在，我们知道泛化误差满足：

$$V^2\hat{q} = \frac{V^2\alpha}{(1+V)^2}\varepsilon_{\text{gen}} = \frac{\eta^2}{\alpha}(\varepsilon_{\text{gen}}).$$

我们可以将其代入关于 q 的方程，得到

$$\varepsilon_{\text{gen}} = \rho + \underbrace{\left(\frac{\eta^2}{\alpha}(\varepsilon_{\text{gen}}) + \rho(V\hat{m})^2 + 2\pi(V\hat{m})(V\hat{h}) + \gamma(V\hat{h})^2\right)}_{q} - 2\underbrace{\left(\rho V\hat{m} + \pi V\hat{h}\right)}_{m} + (\pi - h)^2 \tag{139}$$

$$\varepsilon_{\text{tr}} = \frac{\varepsilon_{\text{gen}}}{(1+V)^2} \tag{140}$$

根据以下关系：

$$V = \frac{\eta}{\alpha - \eta}, \qquad V\hat{m} = \eta, \quad V\hat{h} = \eta(\pi - h)$$

以及将 $\pi - h$ 表示为 η的函数：

$$\pi - h = \pi - V(\gamma\hat{h} + \pi\hat{m}) = \pi - \gamma\eta(\pi - h) - \pi\eta \iff \pi - h = \pi\frac{1-\eta}{1+\gamma\eta} \tag{141}$$

我们仅用 η 表示所有内容，得到：

$$\varepsilon_{\text{gen}} = \frac{\alpha\left(\Delta + (\eta-1)^2\rho\right)}{(\alpha - \eta^2)}\left(1 - \pi^2\frac{(\eta-1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{\left(\Delta + (\eta-1)^2\rho\right)(1+\gamma\eta)^2}\right) \tag{142}$$

$$\varepsilon_{\text{tr}} = \frac{(\alpha - \eta)^2\left(\Delta + (\eta-1)^2\rho\right)}{\alpha\,(\alpha - \eta^2)}\left(1 - \pi^2\frac{(\eta-1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{\left(\Delta + (\eta-1)^2\rho\right)(1+\gamma\eta)^2}\right) \tag{143}$$

ˆ 剩下的工作就是利用关于 V和 V的方程求解 η，得到

$$\eta = 1 + \frac{1}{2}\left(\alpha - 1 + \lambda - \sqrt{4\lambda + (\alpha - 1 + \lambda)^2}\right). \tag{144}$$

消失正则化 我们可以将 $\lambda = 0$代入（144），即使 $\alpha < 1$，并得到

$$\eta(\lambda = 0) = \min(\alpha, 1) \tag{145}$$

最后得到

$$\varepsilon_{gen} = \begin{cases} \frac{\alpha\Delta}{\alpha-1} & \alpha \geq 1 \\ \frac{\pi^2(\alpha-1)\left(\alpha^2\gamma+2\alpha-1\right)}{(\alpha\gamma+1)^2} - \frac{(\alpha-1)^2\rho+\Delta}{\alpha-1} & \text{else} \end{cases} \tag{146}$$

$$\varepsilon_{\text{train}} = \begin{cases} \frac{(\alpha-1)\Delta}{\alpha} & \alpha \geq 1 \\ 0 & \text{else} \end{cases} \tag{147}$$

我们得到推论 3.8的结果：即使存在相关教师，我们也证明了在欠参数化区域内具有高斯普适性。

推论的推广。**3.8：**对于消失正则化下相关教师的高斯普适性结果，也可以推广到具有同方差协方差的一般平衡混合分布，即

$$p_c = \frac{1}{|\mathcal{C}|}, \quad \Sigma_c = \Sigma, \quad \lambda = 0.$$

为此，我们考虑式（87）、（88）中的岭回归一般鞍点方程，并代入以下假设：

$$\begin{cases} \hat{V} &= \frac{\alpha}{|\mathcal{C}|(1+V)} \\ \hat{q}_c &= \frac{\alpha}{|\mathcal{C}|(1+V)^2}(\rho + \Delta + q_c - 2m + (h_c - \pi_c)^2) \\ \hat{m} &= \frac{\alpha}{|\mathcal{C}|(1+V)} \\ \hat{h}_c &= \frac{\alpha(\pi_c - h_c)}{|\mathcal{C}|(1+V)} \end{cases} \tag{148}$$

$$\begin{cases} V &= \frac{1}{d}\operatorname{tr}\left[\Sigma\hat{\Sigma}^{-1}\right] \\ q_c &= \frac{1}{d}\operatorname{tr}\left[\left(\hat{q}_c\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right)\Sigma\hat{\Sigma}^{-2}\right] \\ m &= \frac{1}{d}\operatorname{tr}\left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top\Sigma\hat{\Sigma}^{-1}\right] \\ h_c &= \frac{1}{d}\operatorname{tr}\left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}_c^\top\hat{\Sigma}^{-1}\right] \end{cases} \tag{149}$$

ˆ 特别是，这些假设意味着 V_c，V_c，m_c，ˆm_c不依赖于类别标签 c，因此

$$\hat{\Sigma} = \hat{V}\Sigma.$$

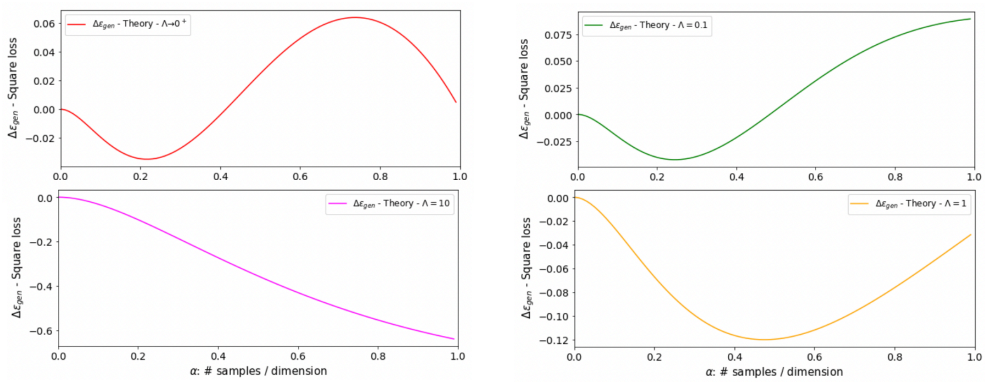

图 6：根据公式 (158) 得出的理论预测：高斯理论计算的泛化误差与高斯混合模型计算的泛化误差之差，随算法中拟合所用样本数量的变化情况。 C.3. 我们采用不同的正则化强度 $\Lambda \in \{1e-10, 0.1, 1, 10\}$，图中按顺时针方向递增。

现在，考虑如下赋值

$$\hat{h}_c = 0 \quad \text{and} \quad h_c = \pi_c; \tag{150}$$

容易验证它们是上述鞍点方程的一个不动点，因为此时

$$h_c = \frac{\hat{m}\pi_c}{\hat{V}} = \pi_c$$

将这一关系代入 $\{\hat{q}_c\}_{c\in\mathcal{C}}$ 的更新表达式中，我们得到：

$$\hat{q}_c^\star = \frac{\alpha p_c}{(1+V_c^\star)^2}(\rho_c + \Delta + q_c^\star - 2m_c^\star + (h_c^\star - \pi_c)^2) = \frac{\alpha p_c}{(1+V_c^\star)^2}(\rho_c + \Delta + q_c^\star - 2m_c^\star) \tag{151}$$

因此，通过考察式（125）中岭回归的泛化误差和训练误差表达式，我们可以得出，它们不会依赖于均值重叠量 $\{h_c^\star\}_{c\in\mathcal{C}}$，并且可以按照证明定理的相同方式证明高斯普适性。3.4。

C.3 插值教师

不同研究在不同数据集上观察到，如果我们找到一个插值教师向量，从而能够保留真实标签来研究经验风险最小化性能，那么来自高斯渐近的理论预测将与模拟结果一致， [19] 等等。这一观察似乎与定理 3.7 的结果相矛盾；在该定理中，我们证明了相关标签（正如我们预期真实标签那样）会破坏高斯普适性。我们尝试在与定理 3.7相同的设定下对此加以解释：

定理 C.1。 （真实标签普适性）考虑与定理相同的设定。 3.7。现在将教师向量固定为 $\tilde{}$相关性最大的那个 $\boldsymbol{\theta} = \boldsymbol{\mu}$，并寻找一个插值教师 $\boldsymbol{\theta}$，使用 $\tilde{n}$ 来自高斯混合模型的样本。如果我们比较 $\tilde{}$在正则化 Λ 固定时 $\alpha = \frac{n}{d}$ 高斯模型和高斯混合模型的精确渐近，会得到差异：

$$\Delta\varepsilon_{gen} = \frac{4(E-1)\alpha^2\left(E^2\gamma + 2e - 1\right)}{(\alpha+1)^2(E\gamma+1)^2} \tag{152}$$

其中 $E(\alpha,\Lambda)$ 是以下方程的解：

$$E = 1 + \frac{1}{2}\left(\alpha - 1 + \Lambda - \sqrt{4\Lambda + (\alpha - 1 + \Lambda)^2}\right). \tag{153}$$

上述结果解决了表面上的矛盾：如果我们对拟合模型进行强烈的过参数化，插值教师可能与数据结构不相关。

这些结果的证明高度依赖于这样一个事实：由于该情形与上一节相同，我们可以对其性能进行解析分析。回顾式（145），在极限 $\lambda \to 0$ 下，我们处于过参数化

算法 1 在保留真实标签的情况下获取理论学习曲线

数据生成： 根据C.1中描述的双簇高斯混合模型生成数据，并设置：

$$\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \sqrt{\gamma}(1, \dots, 1)^\top, \qquad \Sigma_+ = \Sigma_- = I_d, \qquad \boldsymbol{\theta}_0 = \boldsymbol{\mu}_+ \tag{154}$$

拟合插值教师 对 $\lambda \to 0$ 执行岭回归。取 $\alpha = \frac{n}{d} < 1$，使经验风险最小化估计量能够插值数据。

运行精确渐近分析 根据估计的 θ，针对相同的 α 和固定的正则化参数 Λ，计算相应双高斯混合模型和等价高斯协方差模型的理论性能。

设定 $\eta(\alpha, \lambda = 0) = \alpha$，我们可以写成：

$$h_{erm} = \frac{\boldsymbol{\mu}^\top \hat{\boldsymbol{\theta}}}{d} = \gamma\alpha\frac{\gamma+1}{1+\gamma\alpha} \tag{155}$$

$$q_{erm} = \frac{\hat{\boldsymbol{\theta}}^\top \Sigma \hat{\boldsymbol{\theta}}}{d} = \alpha\left(-\frac{\Delta}{\alpha-1} - \frac{(\alpha-1)\pi^2}{\alpha\gamma+1} + \rho\right) \tag{156}$$

我们在算法 C.3 中总结了该算法流程。为了根据式（142）计算泛化误差，我们需要以下量：$\{\rho, \pi\}$。注意，在算法 C.3 中，我们根据估计的$\hat{}$ $\boldsymbol{\theta}$计算高斯混合模型和高斯协方差模型的理论曲线。这意味着，前一步数值模拟中估计的重叠量等价于：

$$\pi = h_{erm} \qquad\qquad \rho = q_{erm} \tag{157}$$

我们将理论预测中的正则化固定为 Λ，为简便起见，使用相同的样本复杂度参数 α。现在可以比较高斯模型与高斯混合模型的理论预测；更确切地说，我们分析这两种数据模型的泛化误差之差。回顾式（142）中的泛化误差表达式，只需将估计的重叠量代入式（157）即可得到：

$$\Delta\varepsilon_{\text{gen}} = \frac{4\alpha^2(E-1)\left(E^2\gamma+2e-1\right)}{(\alpha+1)^2(E\gamma+1)^2} \tag{158}$$

其中 $E(\alpha, \Lambda)$ 是方程（144）的解，得到了定理 C.1中的结果。为清晰起见，我们在简单的设定下陈述了$\hat{}$定理。实际上，在估计出 $\boldsymbol{\theta}$ 后，我们本可以决定更改样本复杂度，因为现在我们想要运行理论曲线。不过，为了简化方程，我们假设理论预测采用与经验风险最小化拟合相同的 α 值。我们在图 6中展示了 $\gamma = 1$的结果：尽管真实标签与数据结构最大程度相关，我们仍看到，高斯理论预测 $\hat{}$与插值教师 $\boldsymbol{\theta}$ 的结果非常接近高斯混合模型。我们指出，图 1 左上图很好地刻画了定理（3.8）：当我们进入欠参数化区域时，即使对于相关教师，$\Lambda \to 0^+$ 也能恢复高斯普适性。

D 真实数据集模拟的详细信息

本节介绍在真实数据上创建随机回归任务的过程，详见算法。D。本文所述不同数值模拟的实现可在GitHub 仓库中找到。

算法 2 真实数据上的随机教师回归

数据处理： 加载数据，并使用变换矩阵 $F \in \mathbb{R}^{d \times d'}$ 执行预处理步骤，其中 d'为所选数据集中图像的维度。对于本文中的所有图，我们均选择 $d = 2000$。**匹配协方差** 计算 $\hat{\Sigma} = \frac{1}{n} X X^\top$

创建新标签忽略与该数据集关联的真实标签，并根据以下规则创建新标签：

$$y_\nu = \begin{cases} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}_\nu}{\sqrt{d}} + \sqrt{\Delta}\xi & \text{Ridge regression} \\ \operatorname{sign}\left(\frac{\boldsymbol{\theta}_0^\top \mathrm{x}_\nu}{\sqrt{\mathrm{d}}} + \sqrt{\Delta}\xi\right) & \text{Logistic regression} \end{cases} \tag{159}$$

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, I_d) \qquad \xi \sim \mathcal{N}(0, 1) \tag{160}$$

运行学习曲线 固定正则化参数 λ。

对于 α，在给定范围内**执行**

模拟 使用 sklearn 包求解方程 (1) 中的经验风险最小化问题 逻辑回归 [53] 或使用 Moore-Penrose 伪逆求解岭估计量：

$$\hat{\boldsymbol{\theta}} = \begin{cases} X^\top (X X^\top + \lambda I_d)^{-1} Y & n < d \\ (X^\top X + \lambda I_d)^{-1} X^\top y & n > d \end{cases} \tag{161}$$

高斯理论求解方程 (80),(81中定义模型高斯渐近行为的鞍点方程。 2.1。**计算误差**使用以下指标计算训练损失和泛化误差：

$$g(y, \hat{y}) = \begin{cases} (y - \hat{y})^2 & \text{Ridge regression} \\ \mathbb{P}(y \neq \hat{y}) & \text{Logistic regression} \end{cases} \tag{162}$$

结束于