arXiv:2303.02048v1 [stat.ML] 3 Mar 2023

高斯混合上的半监督多任务学习的渐近贝叶斯风险

Minh-Toan Nguyen
GIPSA-lab，格勒诺布尔阿尔卑斯大学

Romain Couillet
LIG-lab，格勒诺布尔阿尔卑斯大学

摘要

本文研究高斯混合模型（GMM）上的半监督多任务学习。我们利用统计物理学的方法，计算高维、大规模数据集情形下各任务的渐近贝叶斯风险，据此分析任务相似度在学习中的作用，并评估将多个任务联合学习而非分别学习所带来的性能提升。在监督情形下，我们推导出一种简单的算法，可达到贝叶斯最优性能。

1 引言

多任务学习（MTL）是一种同时学习多个任务的机器学习方法。它有助于在任务之间进行知识迁移，并能形成信息量更丰富的数据表示（Ruder，2017年）。尽管从相关任务中学习有助于将某个任务获得的有用信息传递给其他任务，但无关任务的存在也可能带来益处。若已知两个给定任务彼此无关，算法便可以学会忽略数据分布中的无关特征，从而获得更好的数据表示（Paredes 等人，2012年）。

在这项工作中，我们提出了一种基于高斯混合的简单多任务学习模型，重点在于捕捉任务之间的知识迁移，而不考虑数据表示方面。本文扩展了 Lelarge 和 Miolane（2019a）研究的半监督学习模型；该模型考察了无标签数据在单任务分类中的附加价值。我们转而研究多个分类任务，其中每个任务的数据都只有部分标注，且来自两个类别。得益于模型的简洁性，我们可以将两个任务之间的相关性定义为 $[-1, 1]$ 中的一个数值。我们关注的是性能增益：假设采用最佳算法，相关任务共同学习相较于分别学习能带来多少提升。这引出了贝叶斯风险这一概念，它定义为对一个不属于训练数据集的新数据点进行误分类的最小可行概率。尽管数据具有随机性，但在数据的数量和维度都很大且二者之比固定的极限情况下，贝叶斯风险会收敛到一个确定值。

尽管本研究的主要目标是计算最小分类错误率，但需要强调的是，给定观测数据后信号的后验分布是一个更为基础的对象，因为它是根据特定准则推导最优估计量的基础。在高维状态下，信号的后验分布是一个高维积分；尽管它十分复杂，其表现却类似于一个更简单的分布。这一性质使得本文能够进行精确计算。

贡献与相关工作。

作为首项贡献，我们基于一个简单的论证，推导出渐近贝叶斯风险的精确公式；这一论证与统计物理中的腔方法类似（Mezard and Montanari，2009）。尽管本文的论证并非完全严谨，但旨在清晰阐释高维状态下出现的渐近等价。这一概念是本文大多数公式的基础。本文力求通俗易懂，读者无需具备物理学的先验知识即可理解其内容。我们的研究与一系列探索各类高维统计模型基本极限的工作相契合，这些模型包括张量模型（Barbier 等人，2017年；Lesieur 等人，2017年）、广义线性模型（Barbier 等人，2019年）以及高斯混合模型（Lesieur 等人，2016年；Lelarge 和 Miolane，2019a）。

其次，我们分析任务相关性所起的作用，以及它如何与模型中的其他要素相互影响，例如每个任务中有标签数据所占的比例。众所周知，在单个任务上使用高斯混合数据进行无监督学习会导致相变，从而区分高噪声区间和低噪声区间。我们证明，这种相变在多任务情形下仍然存在，并研究其如何

26th人工智能与统计国际会议（AISTATS）论文集，2023年，西班牙瓦伦西亚。PMLR：第206卷。版权所有 © 2023年，作者所有。

受到任务相关性的影响。在源任务—目标任务的背景下，我们确定了源任务对目标任务最有利的条件。

最后，我们推导出一种简单的算法，在监督学习情形下能够达到最优性能。尽管在合成数据集上表现最优并不一定意味着在真实数据上也有良好表现，但该算法展示了在考虑相关性时，应如何调整针对各个独立任务的最优算法。这或许能为实践中设计多任务学习方法提供有益启示。

尽管我们的研究重点有所不同，但我们的工作与一些理论研究存在联系，这些研究考察了简单数据模型上的基于优化的推断。这些研究计算算法的精确渐近性能，并分析损失函数的选择、正则化以及模型参数数量等因素如何影响这种性能。另一方面，我们的工作关注统计问题的基本极限，而不依赖于任何特定算法。不过，有趣的是，在某些情况下，基于优化的方法几乎可以达到或确实能够达到最优性能（Mai 和 Couillet，2021；Thrampoulidis 等人，2020；Mignacco 等人，2020；Loureiro 等人，2021；Aubin 等人，2020）。对于高斯混合上的多任务学习，Tiomoko 等人（2021b）利用随机矩阵理论，获得了最小二乘支持向量机的精确渐近结果。

研究高斯混合模型有多方面的原因。除了便于进行理论分析之外，它还是能够涵盖我们所关注的多任务学习要素的最简单模型：任务相关性和信息传递。在应用方面，Lesieur 等人（2016）指出，以高斯混合模型作为先验分布进行贝叶斯统计，可以重新发现机器学习中的若干关键方法，例如 K均值或谱聚类算法。贝叶斯统计与算法之间存在密切关系。一方面，可以为主成分分析和支持向量机等知名算法建立贝叶斯解释（Tipping 和 Bishop，1999；Polson 和 Scott，2011），这些算法最终可视为相当简单的数据模型上的标准估计量。另一方面，从简单的数据模型出发，根据特定准则设计最优算法，可以改进现有方法或创造新方法（Bishop，1998；Krzakala 等人，2012）。

记号： 我们用符号 $\langle \cdot, \cdot \rangle$ 表示向量的标量积（或内积）。若 $\boldsymbol{X} = (X_{ij})$，则 $\boldsymbol{X}_{i\cdot} = (X_{ij})_j$ 且 $\boldsymbol{X}_{\cdot j} = (X_{ij})_i$。对于 $n \in \mathbb{N}$，我们用 $[n]$ 表示集合 $\{1, 2, \ldots, n\}$。记号 $\boldsymbol{D}_{\boldsymbol{x}}$ 表示对角元素由向量$\boldsymbol{x}$给出的对角矩阵。若给定诸如 $\boldsymbol{X}_i$ 这样的带索引对象，则 $\boldsymbol{X}$ 仅表示 $(\boldsymbol{X}_i)_i$。

本文中模拟的源代码可在以下地址获取：

https://github.com/Minh-Toan/Bayes-risk

2 模型

我们考虑 T 个分类任务，其中任务 t 包含N_t 个位于 $\mathbb{R}^D$中的数据点。任务 t中的第 i个数据点，记作 $\boldsymbol{Y}_{ti}$，定义为

$$\boldsymbol{Y}_{ti} = V_{ti}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{ti} \tag{1}$$

其中 $\sigma_t > 0$。随机变量 $\boldsymbol{V}$，$\boldsymbol{U}$，$\boldsymbol{Z}$相互独立，且

$$V_{ti} \overset{i.i.d}{\sim} \mathcal{U}(\{-1, 1\}),$$
$$Z_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0, I_D),$$

而 $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$在单位球面 $S^{D-1} = \{\boldsymbol{x} \in \mathbb{R}^D, \|\boldsymbol{x}\| = 1\}$ 上均匀随机选取，条件是事件

$$\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle = C_{tt'}, t \neq t'.$$

矩阵 $\boldsymbol{C} = (\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle)_{t,\, t'=1}^{T}$称为任务相关性矩阵。由定义可知，$\boldsymbol{C}$是一个正定矩阵，且其对角元均等于1。若对于任意两个任务 t 和 t'，都存在一个任务序列 t_1，.....，t_k，使得C_{tt_1}，$C_{t_1t_2} \ldots C_{t_kt'} \neq 0$，则称这些任务是连通的。

换言之，任务 t 中的数据来自两个类别，分别对应于以$\pm\boldsymbol{U}_t$为中心且具有相同协方差 $\sigma_t^2 I_D$的两个高斯分布。中心的位置未知，只能根据数据进行估计。数据点 $\boldsymbol{Y}_{ti}$ 的类别由V_{ti}表示，因此每个数据点属于每个类别的概率均为 $1/2$ 。若我们知道一个数据点属于哪个类别，则称该数据点为有标签的；否则，它是无标签的。与所有其他随机变量独立地，任务 t 中的每个数据点以概率 η_t被标记。$\eta_t = 1$ 和$\eta_t = 0$ 两种情形分别对应于监督学习和无监督学习。$C_{tt'}$ 度量任务 t 与 t' 之间的相关性。参数 $\lambda_t = 1/\sigma_t^2$ 称为信噪比（SNR）。随着信噪比增大，两个类别逐渐分离，分类也变得更容易。我们研究如下设定下的模型：维度以及每个任务中的数据量以固定比率 $\alpha_t = \lim_{D\to\infty} N_t/D$趋于无穷大，该比率称为采样比。请注意，Lelarge 和 Miolane（2019a）中研究的半监督学习模型对应于 $T = 1$的情形。

我们可以获取数据集 $\boldsymbol{Y} = (\boldsymbol{Y}_{ti})$、标签以及模型参数 (σ_t)，(η_t)，(α_t) 和 $\boldsymbol{C}$。[1]我们的任务

[1] $\boldsymbol{\sigma}$ 和 $\boldsymbol{C}$ 确实可以在$D \to \infty$时以趋于零的误差进行估计，前提是每个任务都有正比例的有标签数据可用，即对于所有 t，$\eta_t > 0$（附录 D）。

是利用这些可用信息，对任意给定任务中的一个新数据点 $\boldsymbol{Y}_{\text{新的}}$进行分类 t

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{2}$$

我们关注最小分类错误率，即贝叶斯风险

$$\inf_{\hat{V}} \mathbb{P}(\hat{V} \neq V_{\text{new}}) \tag{3}$$

其中，下确界是在 V_{new}的所有估计量中取得的。

3 结果

在介绍结果之前，我们需要先给出一些定义，以便清晰、简洁地阐述我们的发现。

定义 3.1。从数据 $\boldsymbol{Y}$中对 $\boldsymbol{X} \in \mathbb{R}^D$ 进行推断满足复制对称（RS）性质，其重叠度q 在极限 $D \to \infty$下，

$$\langle \boldsymbol{X}, \boldsymbol{X}^1 \rangle, \langle \boldsymbol{X}^1, \boldsymbol{X}^2 \rangle, \langle \boldsymbol{X}, \hat{\boldsymbol{X}} \rangle, \|\hat{\boldsymbol{X}}\|^2 \tag{4}$$

都收敛到同一个极限 q，其中 $\boldsymbol{X}^1$，$\boldsymbol{X}^2$ 是从给定 $\boldsymbol{Y}$ 时 $\boldsymbol{X}$ 的后验分布中独立采样得到的，而 $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$ 称为给定 $\boldsymbol{Y}$ 时 $\boldsymbol{X}$ 的最小均方误差估计器。[2] 在某些情境下，我们用 $\hat{\boldsymbol{X}}$ 表示 $\boldsymbol{X}$ 的一般估计量，而 $\boldsymbol{X}$ 的最小均方误差估计器记为 $\hat{\boldsymbol{X}}_{\text{MMSE}}$。

当信号由已知分布生成时，这一性质在广泛的推断问题中都成立。我们假设这一性质对我们的模型也成立：

假设。 $\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ 和 $N_t^{-1/2}\boldsymbol{V}_t|\boldsymbol{Y}$ 对所有 $t \in [T]$ 都满足 *RS* 性质，其重叠度分别记为 q_{ut} 和 q_{vt} 。

在 q_{ut} 的定义中加入归一化因子 σ_t 是为了方便。

在本文后文中，为了证明结果，我们需要以下定义：

定义 3.2。 考虑以下高斯信道

$$Y_i = \sqrt{\lambda_i} X_i + Z_i, \quad i = 1, \ldots, n \tag{5}$$

，其输入为 X_i，输出为 Y_i ，信噪比为 λ_i。令 $\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$。重叠函数 $F_{\boldsymbol{X},i}: \mathbb{R}^n \to \mathbb{R}$ 定义如下

$$F_{\boldsymbol{X},i}(\boldsymbol{\lambda}) = \mathbb{E}[\hat{X}_i X_i] = \mathbb{E}[\hat{X}_i^2] \tag{6}$$

$F_{\boldsymbol{X},\,i}$ 也被称为信号 X_i的重叠度。

[2]MMSE 是最小均方误差的缩写。

本文的主要结果如下。

结果。 *i)* 在该模型设定下，当 $D \to \infty$ 时，贝叶斯风险收敛至

$$1 - \Phi(\sqrt{q_{ut}}),$$

其中 $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2} dx$

ii) 重叠度 q_{ut}, q_{vt} 满足以下方程

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{7a}$$
$$q_{vt} = \eta_t + (1 - \eta_t) F(q_{ut}) \tag{7b}$$

with

$$\boldsymbol{M} = \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^{T}$$
$$\boldsymbol{D} = \text{diag}\{\alpha_t q_{vt}\}_{t=1}^{T}$$
$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1).$$

注释 3.1。 当 $q_{ut} = 0$ 时，任务 t 的贝叶斯风险等于 0。5，这对应于随机猜测的分类误差水平。在这种情况下，我们称任务 t 的分类是不可能的。另一方面，如果 q_{ut} 为正，则称任务 t 的分类是可行的。

注释 3.2。 不动点方程 (7a) 和 (7b) 可能无法唯一确定重叠度。具体而言，对于高信噪比下的无监督学习，存在两个解：零解不稳定，而非零解稳定，因此自然选择稳定解作为重叠度。其他情况下则只有一个解。

我们可以通过考察以下特殊情况来对结果进行合理性检验：如果任意两个任务之间的相似度为零，则结果表明，多任务学习的渐近贝叶斯风险与单独学习任务时相同。这显然是因为不同任务的数据彼此独立；而如果对于所有 t，$\sigma_t = \sigma$ 和 $C_{tt'} = 1$ 均成立，t'，即所有任务的数据分布都相同，则所有任务的渐近贝叶斯风险均等于参数为 $\alpha = \sum_t \alpha_t$ 和 $\alpha\eta = \sum_t \alpha_t \eta_t$ 的单个任务的渐近贝叶斯风险（附录B）。

4 推论

本节将介绍主要结果的一些推论。

4.1 监督学习。

对于仅有一个任务的监督学习，新数据点 $\boldsymbol{Y}_{\text{new}}$ 的最小分类错误率由估计量 $\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \bar{\boldsymbol{Y}} \rangle)$ 达到，其中 $\bar{\boldsymbol{Y}} = N^{-1}\sum_i V_i \boldsymbol{Y}_i$（Lelarge 和 Miolane, 2019a）。在多任务情形下，若 $\boldsymbol{Y}_{\text{new}}$ 是任务 t中的一个新数据点，则以下算法可达到最优性能：

1. 计算

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} V_{ti}\boldsymbol{Y}_{ti}$$

2. 计算

$$\tilde{\boldsymbol{Y}}_t = \sum_{s=1}^{T} a_{ts}\bar{\boldsymbol{Y}}_s$$

where $\boldsymbol{A} = (a_{ts})_{t,s=1}^{T} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$.

3. 渐近贝叶斯风险由以下方法达到

$$\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}, \tilde{\boldsymbol{Y}}_t \rangle). \tag{8}$$

我们可以看到，多个任务的最优估计量¯通过考虑任务之间的相关性，以及各自的难度水平和相对规模（分别由 $\boldsymbol{C}$,(σ_t) 和 (α_t) 衡量），对独立任务的最优估计量 $\boldsymbol{Y}_t$ 进行了调整。有趣的是，这一最优算法与 Tiomoko 等人（2021a）采用另一种方法提出的方法一致。

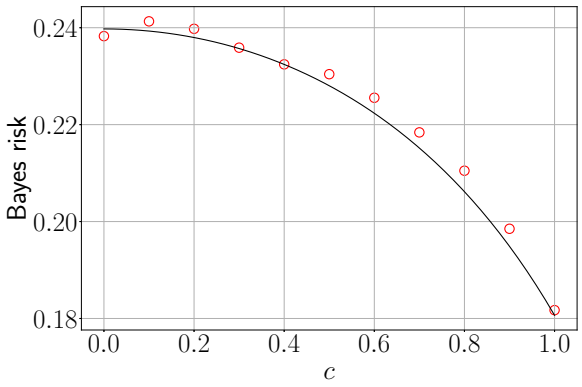


图 1：渐近最优算法的贝叶斯风险与性能。
$\alpha_1 = \alpha_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 0.5$,$D = 1000$。

4.2 无监督学习与相变

只有在无监督学习中才会出现的一种特别有趣的现象是相变。这一现象最著名的例子之一是*BBP* 相变 （Baik et al., 2005），它研究的是一个满足 $\lim_{D\to\infty} N/D = 1$ 的单一学习任务。当 $\lambda = 1/\sigma^2 \leq 1$ 时，没有任何估计量能够使分类错误率低于 0.5。换言之，由于两个类别在统计上完全相同，因此客观上不可能进行分类。另一方面，如果能够获得低于 0.5 的分类错误率，我们就称该任务是可行的 。事实证明，相变在多任务情形下依然存在。图

相变。

2 展示了两个任务、 $N_1 = N_2 = D$ 且相关性为 $c = 0.7$ 时，任务 1 在不同信噪比下的表现。黑色曲线所界定的区域内无法进行分类。

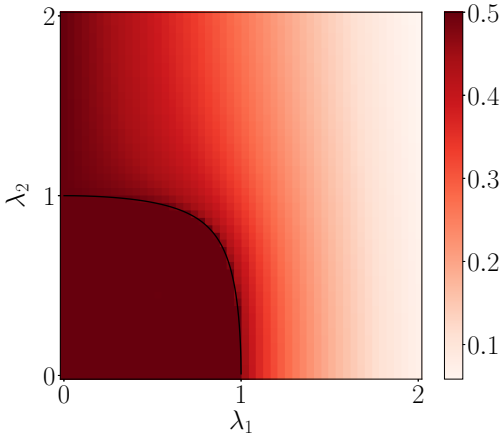


图3：任务1的贝叶斯风险与各任务信噪比的关系。两个任务均为无监督，且 $N_1 = N_2 = D$，相关性为 $c = 0.7$。黑色曲线所界定的区域内无法进行分类。两个任务的不可能区域相同。

模拟还表明，两个任务的不可能区域相同。换言之，两个相关任务要么都可行，要么都不可能。对于任意数量任务的一般情况，只要这些任务彼此连通，它们就会同时可行或同时不可能。

请注意，只要有正比例的标注数据可用，相变就会消失，因为仅基于标注数据进行监督学习已经能取得非平凡的性能。

对于具有 $N_1 = N_2 = D$的两个任务，不可能分类区域为

$$\left\{(\lambda_1, \lambda_2) \in [0,1]^2 : (1-\lambda_1^2)(1-\lambda_2^2) \geq c^4\lambda_1^2\lambda_2^2\right\} \tag{9}$$

如图3所示。随着任务相关性 c 从0增至1，该区域会从单位正方形 $[0,1]^2$缩小为四分之一圆盘。

另一种可以得到不可能区域显式公式的特殊情形是：存在 T 个任务，其 $N_1 = \cdots = N_T = D$，任意两个任务之间的相关性为 $c > 0$，并且对于所有 t，均有 $\lambda_t = \lambda$。可以证明，只要满足以下条件，分类就不可能实现：

$$\lambda \leq \frac{1}{\sqrt{1 + (T-1)c^2}}. \tag{10}$$

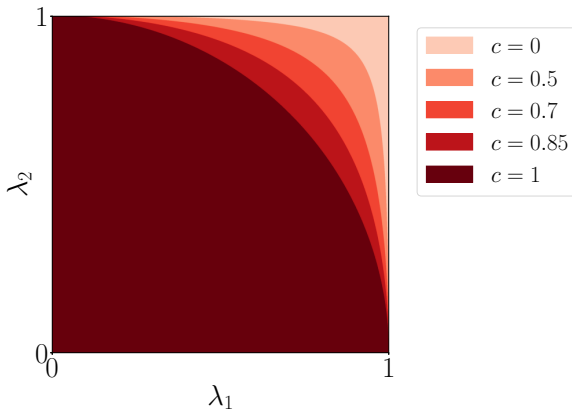


图3：随着任务相关性增强，不可能分类区域会缩小。当两个任务不相关（$c = 0$）时，不可能分类区域是整个正方形$[0, 1]^2$。随着 c 从 0 增加到 1，不可能区域会从单位正方形 $[0, 1]^2$ 缩小为四分之一圆盘。

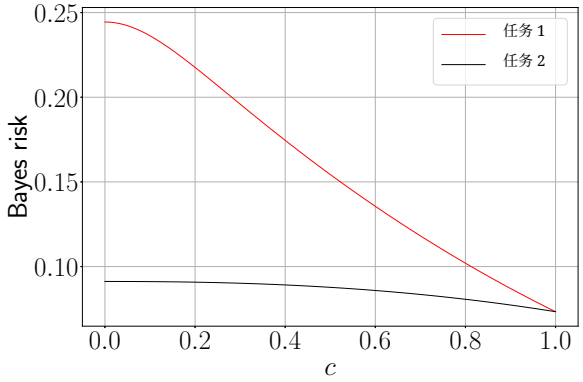


图4：双任务设定：贝叶斯风险随任务相关性 c、有标签数据比例$\eta_1 = 0$、 $\eta_2 = 1$、过采样比 $\alpha_1 = 0.1$、$\alpha_2 = 0.2$以及信噪比 $\lambda_1 = \lambda_2 = 4$的变化。当两个任务的相关性很高（$c \gtrsim 0.5$）时，任务1的性能会显著提升。

4.3 半监督学习。

为了减少模拟中的模型参数数量，我们在此聚焦于一种特定设定，其中包含一个源任务和一个目标任务。源任务相对容易：它可以拥有完整标签、较高的信噪比，或者更大的数据集。我们希望了解目标任务如何从源任务中获益。

图4展示了任务相关性的影响。任务相关性 c 的取值范围为 0 到 1。请注意，相关性c和 $-c$本质上相同，因为只需交换一个任务中的标签，就可以将其中一个转换为另一个。第一个任务（目标任务）由一个没有标签（$\eta_1 = 0$）的小型数据集（$\alpha_1 = 0.1$）组成，而第二个任务（源任务）则由一个完全标注的数据集（$\eta_2 = 1$）组成，其数据量是前者的两倍（$\alpha_2 = 0.2$）。如果两个任务高度相关（$c \gtrsim 0.5$），目标任务的性能就能显著提升。当 c 接近零时，贝叶斯风险的下降速度较慢，量级为 O（c^2）。请注意，两个任务具有相同的信噪比（$\lambda_1 = \lambda_2 = 4$），因此当 $c = 1$ 时，它们具有相同的数据分布，可以合并为一个任务，从而获得相同的性能。

在图5中，我们计算了从源任务迁移信息后，目标任务的误差降低率。我们发现，当目标任务的信噪比接近相变点且低于源任务的信噪比，同时标注数据的比例较低时，多任务学习最为有效。

直观而言，这有三个原因。首先，目标任务中的标注数据比源任务中的更有价值，即使在两个任务高度相关的情况下也是如此（$c = 0.8$）。因此，当目标任务中标注数据的比例较高时，收益会较低。其次，

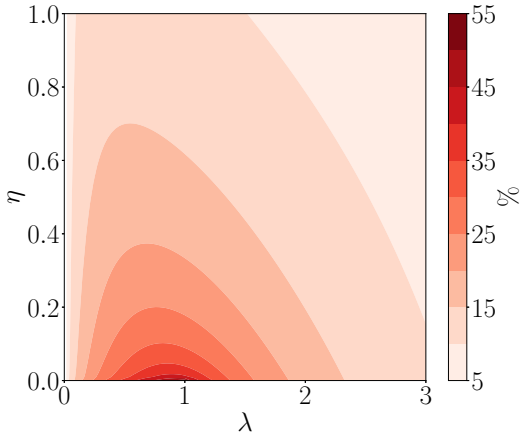


图5：目标任务的贝叶斯风险降低百分比随信噪比和有标签数据比例的变化，参数为 $c = 0.8$、 $N_1 = N_2 = D$、$\lambda_1 = 2$、 $0 \leq \lambda_2 \leq 3$、$\eta_1 = 1$、 $0 \leq \eta_2 \leq 1$。

如果源任务比目标任务更难，也就是说目标任务的信噪比更高，那么源任务的帮助就不大。最后，在目标任务面临困难的相变点附近，源任务中的标注数据可以提供宝贵的帮助。

5 腔体论证

本文得到的各种方程均以高维极限中出现的渐近等价现象为基础。在这一极限下，一个相当复杂的统计模型会解耦为彼此独立的分量，并且可以分别在每个分量中进行推

断。这种解耦现象是通过所谓的腔方法证明的。以下引理在本文提出的腔体论证中起着关键作用:

引理 5.1. 假设我们要根据数据 $\boldsymbol{Y}$估计信号 $X \in \mathbb{R}$，其先验分布为 P_X。数据 $\boldsymbol{Y}$可分为以下两部分。第一部分记为 $\boldsymbol{Y}^x$，由关于 X的以下观测组成，

$$\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}, \tag{11}$$

其中

- $\boldsymbol{U} \in R^D$ *is unknown with prior* P_U,
- $\boldsymbol{Z} \sim \mathcal{N}(0, I_D)$,
- X, $\boldsymbol{U}$ *and* $\boldsymbol{Z}$ *are independent.*

第二个数据集记为 $\boldsymbol{Y}^u$，且与 $\boldsymbol{X}$独立。假设分布 $\boldsymbol{U}|\boldsymbol{Y}^u$具有重叠度为 q的 *RS* 性质。那么在极限 $D \to \infty$下，

i) 给定 $\boldsymbol{Y}$时，$\boldsymbol{X}$ 的后验分布渐近等价于 $\bar{}$分布 P，其定义如下:

$$\frac{d\bar{P}(x|\boldsymbol{Y})}{dP_X(x)} \propto \exp\Big(x\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle - \frac{1}{2}qx^2\Big) \tag{12}$$

其中 $\hat{} \boldsymbol{U} = \mathbb{E}[\boldsymbol{U}|\boldsymbol{Y}]$。因此，统计量 $S = \hat{}\langle \boldsymbol{Y}^x, \boldsymbol{U}\rangle$ 对于根据 $\boldsymbol{Y}$估计 $\boldsymbol{X}$是渐近充分的。

ii) $S/\sqrt{q}$ 依分布收敛于 $\sqrt{q}X \bar{} + \xi$，其中 ξ 服从标准正态分布，且与 X独立。因此，根据 $\boldsymbol{Y}$估计 X，在渐近意义上等价于根据信噪比为 q 的高斯信道输出估计 X。

证明。由于 X 与 $\boldsymbol{U}$和 $\boldsymbol{Y}^u$独立，因此有

$$\begin{aligned}\frac{dP(x|\boldsymbol{Y})}{dP_X(x)} &= \int dP(\boldsymbol{u}|\boldsymbol{Y}^u)P(x|\boldsymbol{u}, \boldsymbol{Y}^x) \\ &\propto \int dP(\boldsymbol{u}|\boldsymbol{Y}^u)\exp\Big(x\langle \boldsymbol{Y}^x, \boldsymbol{u}\rangle - \frac{1}{2}x^2\|\boldsymbol{u}\|^2\Big) \\ &:= \mathcal{A}\end{aligned}$$

定义

$$\mathcal{B} = \exp\Big(x\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle - \frac{1}{2}qx^2\Big) \tag{13}$$

为了证明 (i)，我们将证明在高维极限下，对于 x的任意取值，$\mathbb{E}[(\mathcal{A} - \mathcal{B})^2] \to 0$ $D \to \infty$。为此，只需证明利用 $\boldsymbol{U}|\boldsymbol{Y}^u$ 的 RS 性质，$\mathbb{E}[\mathcal{A}^2]$、$\mathbb{E}[\mathcal{B}^2]$ 和 $\mathbb{E}[\mathcal{AB}]$ 收敛到同一个极限。事实上，$\mathbb{E}[\mathcal{A}^2]$可以写成

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} x\langle \boldsymbol{Y}^x, \boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

其中，$\boldsymbol{U}^1$，$\boldsymbol{U}^2$是从 $\boldsymbol{U}|\boldsymbol{Y}^u$ 中独立采样的。将 $\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}$ 代入前一个表达式，我们得到

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle \boldsymbol{U}, \boldsymbol{U}^a\rangle + x\langle \boldsymbol{Z}, \boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

对 $\boldsymbol{Z}$ 取期望，并利用$\mathbb{E}[e^{\langle \boldsymbol{a}, \boldsymbol{Z}\rangle}] = e^{\frac{1}{2}\|\boldsymbol{a}\|^2}$这一事实，可得

$$\mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle \boldsymbol{U}, \boldsymbol{U}^a\rangle + x^2\langle \boldsymbol{U}^1, \boldsymbol{U}^2\rangle\Big)$$

由 $\boldsymbol{U}|\boldsymbol{Y}^u$ 的 RS 性质可得

$$\lim_{D\to\infty} \mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\big(2qXx + qx^2\big) \tag{14}$$

为计算 $\mathbb{E}[\mathcal{AB}]$ 和 $\mathbb{E}[\mathcal{B}^2]$的极限，我们完全按照相同的步骤进行：代入 $\boldsymbol{Y}^x$的定义，对 $\boldsymbol{Z}$取期望，并利用 RS 性质。这将得到与 (14) 相同的极限，从而证明 (i)。

由 $P(x|\boldsymbol{Y})$ 与 $P(x|\boldsymbol{Y})$ 的渐近等价可立即得出，统计量 $\langle \boldsymbol{Y}^x, \boldsymbol{U}\rangle$ 对于根据 $\boldsymbol{Y}$ 估计 X 是渐近充分的。这意味着，可以从 $\langle \boldsymbol{Y}^x, \boldsymbol{U}\rangle$ 而非 $\boldsymbol{Y}$ 中提取关于 X 的所有相关信息，且在高维极限下不会损失任何信息。

现在我们有

$$\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle = \langle X\boldsymbol{U} + \boldsymbol{Z}, \hat{\boldsymbol{U}}\rangle = X\langle \boldsymbol{U}, \hat{\boldsymbol{U}}\rangle + \langle \boldsymbol{Z}, \hat{\boldsymbol{U}}\rangle.$$

鉴于 $\langle \boldsymbol{Z}, \hat{}\boldsymbol{U}\rangle \sim \mathcal{N}(0, \|\hat{}\boldsymbol{U}\|^2)$ 且 $\boldsymbol{Z}$ 与 X 独立，在极限 $D \to \infty$ 下，这个内积依分布收敛于 $\sqrt{q}\xi$，其中 ξ 是一个与 $\boldsymbol{X}$ 独立的标准正态随机变量。因此

$$\frac{\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle}{\sqrt{q}} \xrightarrow{d} \sqrt{q}X + \xi, \quad D \to \infty,$$

这证明了 (ii)，因为最后一个表达式左侧也是在给定 $\boldsymbol{Y}$ 的条件下关于 X 的充分统计量。 □

为了展示引理 5.1 的一个应用，并让读者在深入了解本文主要结果的证明之前熟悉腔体论证，我们将分析 Miolane（2017年）研究的以下张量模型。我们的目标是根据以下观测估计信号 $\boldsymbol{U}$ 和 $\boldsymbol{V}$:

$$Y_{ij} = \sqrt{\frac{\lambda}{N}}U_iV_j + Z_{ij}, i \in [N_u], j \in [N_v] \tag{15}$$

这里，我们假设 $U_i \overset{i.i.d}{\sim} P_U$，$V_j \overset{i.i.d}{\sim} P_V$且噪声$Z_{ij}$服从相互独立的标准高斯分布，适用于

所有 i， j。我们研究 N， N_u， N_v 以固定比率 $N_u/N \to \alpha_u$ 和 $N_v/N \to \alpha_v$趋于无穷时的模型极限。此外，我们假设 $\boldsymbol{U}$， $\boldsymbol{V}$和 $\boldsymbol{Z} = (Z_{ij})$ 相互独立。可以证明， $N_u^{-1/2}\boldsymbol{U}|\boldsymbol{Y}$和$N_v^{-1/2}\boldsymbol{V}|\boldsymbol{Y}$都满足副本对称性，其重叠度分别为 q_u 和 q_v。我们将使用引理 5.1 推导 q_u， q_v所满足的不动点方程。

设 $i \in [N_u]$ 为固定值。腔方法涉及将数据 $\boldsymbol{Y}$ 分成两部分。第一部分记为 $\boldsymbol{Y}^1$，其中包含与 U_i 相关的观测，表示为

$$\boldsymbol{Y}_{i\cdot} = \sqrt{\frac{\lambda}{N}} U_i \boldsymbol{V} + \boldsymbol{Z}_{i\cdot} \tag{16}$$

其余数据记为 $\boldsymbol{Y}^2$。由于数据集 $\boldsymbol{Y}^1$ 中仅包含与 $\boldsymbol{V}$ 相关的极少量信息（比较 $\boldsymbol{Y}^1$ 和 $\boldsymbol{Y}^2$ 的大小即可看出），因此，从 $\boldsymbol{Y}$ 中估计 $\boldsymbol{V}$，与从 $\boldsymbol{Y}^2$ 中估计 $\boldsymbol{V}$ 基本等同。因此，
$N_v^{-1/2}\boldsymbol{V}|Y^2$ 也满足重叠度为 q_v 的 RS 性质。容易验证，引理 5.1 适用于该模型，其中 U_i 和 $\sqrt{\lambda/N}\boldsymbol{V}$ 分别对应引理中的 X 和 $\boldsymbol{U}$。因此，从$\boldsymbol{Y}$ 中估计 U_i，渐近等价于从信噪比为 $\lambda\alpha_v q_v$ 的高斯信道输出中估计信号 U_i。

对于不同的 i， $k \in N_u$，由于 $\boldsymbol{Z}_{i\cdot}$ 和 $\boldsymbol{Z}_{k\cdot}$ 相互独立，根据引理 5.1-ii 的证明可知，与 U_i， U_k相关联的等效高斯信道的噪声 ξ_i 和 ξ_k 相互独立。因此，对于所有 i，取决于 ξ_i 和 U_i的 $\boldsymbol{U}_i$在渐近意义下相互独立。根据大数定律

$$q_u = \lim_{N_u \to \infty} \frac{1}{N_u} \sum_{i=1}^{N_u} \hat{U}_i^2 = F_U(\lambda q_v) \tag{17}$$

其中， F_U 是信号为 U的高斯信道的重叠函数。对 V_j 和 $j \in N_v$重复相同的论证，我们得到 q_u， q_v的不动点方程：

$$q_u = F_U(\lambda\alpha_v q_v)$$
$$q_v = F_V(\lambda\alpha_u q_u)$$

注意，不动点方程可能无法唯一确定重叠，因为它们可能存在多个解。然而，严格的方法（Barbier 和 Macris（2019））表明，重叠可以被唯一确定为某个函数的极小极大点。

6 证明

6.1 不动点方程

重新表述为张量模型。令 $\tilde{\boldsymbol{U}}_t = \sqrt{D}\boldsymbol{U}_t$，可以证明，在极限情况下，˜和 $D \to \infty$在附录 F 中 U_{tj} 是渐近高斯，其协方差为

$$\mathbb{E}[\tilde{U}_{tj}\tilde{U}_{t'j'}] = C_{tt'}\delta_{jj'} \tag{18}$$

设 $\boldsymbol{W}_t = \sqrt{D}\overline{U}_t/\sigma_t$，原始模型可以写成一组一维高斯信道

$$Y_{ijt} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij} \tag{19}$$

对于 $1 \le t \le T$， $1 \le i \le N_t$， $1 \le j \le D$。当 $D \to \infty$时，随机变量 W_{tj} 渐近高斯，其协方差为

$$\mathbb{E}[W_{tj}W_{t'j'}] = M_{tt'}\delta_{jj'} \tag{20}$$

where $M_{tt'} = C_{tt'}/(\sigma_t\sigma_{t'})$.

接下来，标签所传递的信息可以并入 $\boldsymbol{V}$ 的先验分布中。具体而言，若 V_{ti} 的值未知，则其先验分布在 $\{-1, 1\}$ 上仍为均匀分布。否则，若已知 $V_{ti} = 1$，则 V_{ti} 的先验分布为 $\delta(v-1)$。请注意，在这种情况下，后验分布与先验分布一致。

$\sigma_t^{-2}\boldsymbol{U}_t|\boldsymbol{Y}$ 的 RS 性质意味着 $D^{-1/2}\boldsymbol{W}_t|\boldsymbol{Y}$ 也具有 RS 性质，其重叠度为 $\boldsymbol{q}_{ut}$。

总之，该问题可以表述为一个张量模型，其目标是根据关于向量 $\boldsymbol{V}_t$ 和 $\boldsymbol{W}_t$ 的先验信息以及张量积 $\boldsymbol{V}_t \otimes \boldsymbol{W}_t$ 的含噪观测来估计信号 $\boldsymbol{V}_t$ 和 $\boldsymbol{W}_t$。

腔体论证。分析中的关键步骤是证明，在高维极限下，给定 $\boldsymbol{Y}$ 估计 $\boldsymbol{V}_t$ 和 $\boldsymbol{W}_t$，渐近等价于通过具有相互独立噪声的高斯信道来估计这些向量的各个坐标。因此，原始模型等价于一个更加解耦的模型，并且可以分别在每个信道上进行推断。

为得到不动点方程，我们采用与第 5 节所示例子相同的方法。我们假设在任意任务中，无标签数据的比例均为正。将这些比例取极限至零，即可得到监督情形下的结果。固定 $t \in [T]$ 和 $i \in [N_t]$，使得 V_{ti} 未知。我们将数据 $\boldsymbol{Y}$ 分成两部分： $\boldsymbol{Y}^1$ 包含关于 V_{ti} 的观测，即

$$\boldsymbol{Y}_{ti} = \frac{1}{\sqrt{D}} V_{ti} \boldsymbol{W}_t + \boldsymbol{Z}_{ti}$$

以及剩余的数据 $\boldsymbol{Y}^2$。由于数据集 $\boldsymbol{Y}^1$ 中与$\boldsymbol{W}_t$相关的信息量微不足道，因此根据 $\boldsymbol{Y}$ 估计 $\boldsymbol{W}_t$，本质上等同于根据 $\boldsymbol{Y}^2$估计 $\boldsymbol{W}_t$。因此， $D^{-1/2}\boldsymbol{W}_t|\boldsymbol{Y}^2$ 也满足重叠度为 q_u的 RS 性质。容易验证，引理 5.1 适用，其中 V_{ti} 和 $D^{-1/2}\boldsymbol{W}_t$分别扮演引理中 X 和 $\boldsymbol{U}$ 的角色。

因此，根据 $\boldsymbol{Y}$估计 V_{ti}，渐近等价于从信噪比为 q_{ut}的高斯信道输出中估计信号 V_{ti}。对于不同的 i, $k \in [N_t]$，由于 $\boldsymbol{Z}_{ti}$和 $\boldsymbol{Z}_{tk}$ 相互独立，从引理 5.1-ii 的证明可知，与 V_{ti}，V_{tk} 对应的等效高斯信道中的噪声 ξ_i 和 ξ_k 也相互独立。因此，对于所有满足 V_{ti} 未标注的 i，依赖于 ξ_i 和 V_{ti}的 V_{ti}渐近独立。根据大数定律,

$$r_{vt} := \lim_{N_t\to\infty} \frac{1}{(1-\eta_t)N_t} \sum_i \hat{V}_{ti}^2 = F_v(q_{ut}) \tag{21}$$

其中，求和遍及所有满足 $i \in [N_t]$ 为无标签且 V_{ti} 为具有拉ademacher信号的高斯信道的重叠函数的 F_v。由附录 E.1 可知,

$$F_v(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1). \tag{22}$$

另一方面，根据 r_{vt}的定义，我们有

$$q_{vt} = \eta_t + (1-\eta_t) r_{vt} \tag{23}$$

不动点方程 (7b) 由 (21)、(22) 和 (23) 推得。

沿用完全相同的腔体论证，在给定 $\boldsymbol{Y}$ 的条件下对 W_{tj} 的估计，渐近等价于根据具有信噪比 $\alpha_t q_{vt}$ 的高斯信道输出对信号 W_{tj} 的估计。此外，当 $(t,\ j) \neq (t',\ j')$ 时，与信号 W_{tj} 和 $W_{t'j'}$ 对应的噪声渐近独立。当 $j \neq j'$ 时，信号 W_{tj} 和 $W_{t'j'}$ 独立。因此，可在由 T 个标量高斯信道组成的组 $(W_{tj})_{t=1}^T$ 上独立地进行等效高斯信道的推断。根据大数定律,

$$q_{ut} = \lim_{D\to\infty} \frac{1}{D} \sum_{j=1}^{D} \hat{W}_{tj}^2 = F_{w,t}(\{\alpha_t q_{vt}\}_{t=1}^T) \tag{24}$$

其中，$F_{w,\ t}$ 是信号为 $\mathcal{N}$（0，$\boldsymbol{M}$）的高斯信道的重叠函数。$F_{w,\ t}$的显式公式见附录 E.2，由此得到不动点方程 (7a)。

6.2 贝叶斯风险与最优算法

假设我们想要对任务 t 中的一个新数据点 $\boldsymbol{Y}_{\text{new}}$ 进行分类

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}} \boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{25}$$

不难验证，引理 5.1 可应用于这个问题，其中 $V_{\text{new}}, \boldsymbol{U}_t$ 在引理中扮演 $X, \boldsymbol{U}$的角色，因为后验分布 $\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ 满足 RS 性质，重叠度为 q_{ut}。因此，在高维极限下，给定 $\boldsymbol{Y}, \boldsymbol{Y}_{\text{new}}$ 来估计 V_{new}，本质上等同于根据信噪比为 q_{ut} 的高斯信道输出估计信号 V_{new}。这意味着，V_{new} 的最小分类错误率等于信噪比为 q_{ut}、信号为拉ademacher信号的高斯信道的分类错误率，即（附录 E.1）

$$1 - \Phi(\sqrt{q_{ut}}),$$

根据引理 5.1，$S = \langle \boldsymbol{Y}_{\text{新}},\ \hat{\boldsymbol{U}}_t \rangle / \sqrt{q_{ut}}$ 足以用于估计 V_{new}。此外，S 依分布收敛到信号为 V_{new}、信噪比为 q_{ut} 的高斯信道的输出。该信道中使贝叶斯风险最小的估计量就是 sgn(S)，由此得到 V_{new} 的最优ˆ估计量 sgn($\langle \boldsymbol{Y}_{\text{新}}\ \boldsymbol{U}_t \rangle$)。下一步是确定 $\boldsymbol{U}_t$ 的值。我们将利用向量 $\boldsymbol{U}_t$ 渐近高斯这一事实，因此后续论证将基于模型的重述形式 (18)。我们需要以下结果

引理 6.1。 以下这组高斯信道

$$Y_i = c_i X_i + Z_i, \quad i = 1, \ldots, n \tag{26}$$

输入为 X_i、输出为 Y_i、信噪比为 c_i^2 且具有独立标准高斯噪声 Z_i，等价于一个信号为 X、输出为 $\langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ 且信噪比为$\sum_{i=1}^n c_i^2$的单一高斯信道。此外，

证明。很容易验证，统计量 $S := \langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ 对于根据 $\boldsymbol{Y}$估计 X 是充分的。此外，$S = \|\boldsymbol{c}\| X + \xi$，其中 $\xi = \|\boldsymbol{c}\|^{-1} \langle \boldsymbol{c}, \boldsymbol{Z} \rangle$ 服从标准高斯分布，且独立于 X。这就证明了该引理的结论。 □

注记 6.1。从引理 6.1 的证明中还可以看出，简化信道的噪声 ξ 来自原始信道的噪声。

引理 6.1 表明，对于每个固定的 ($t,\ j$)，以下高斯信道

$$Y_{tij} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij}, \quad i = 1, \ldots, N_t$$

共享相同信号 W_{tj}，可简化为单个高斯信道，其输出为 $\sqrt{N_t}\bar{Y}_{tj}$，信噪比为 $N_t/D \simeq \alpha_t$，其中 Y_{tj} 是算法中向量 $\boldsymbol{Y}_t$ 的第 j 个坐标。

对于 $(t,\ j) \neq (t',\ j')$，根据注 6.1，与 W_{tj} 和 $W_{t'j'}$ 相关的简化高斯信道的噪声相互独立。此外，若 $j \neq j'$，则信号W_{tj} 和 $W_{t'j'}$ 相互独立。因此，可以对每组包含 T 个信道的简化高斯信道独立进行推断，其

信号为 $(W_{tj})_{t=1}^T$。这些组中每一组的最小均方误差估计器可显式计算如下

$$(\hat{W}_{tj})_{t=1}^T = \boldsymbol{B}(\sqrt{N_t}\bar{Y}_{tj})_{t=1}^T$$

其中

$$\boldsymbol{B} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2}(\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2}\boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2})^{-1}$$

（附录 E.2）。等价地，

$$\hat{\boldsymbol{W}}_t = \sum_s B_{ts}\sqrt{N_s}\bar{\boldsymbol{Y}}_s$$

两边同时除以 $\sqrt{D}$ 并使用 $N_t/D \simeq \alpha_t$，可得

$$\tilde{\boldsymbol{Y}}_t := \sigma_t^{-1}\hat{\boldsymbol{U}}_t \simeq \sum_s A_{ts}\bar{\boldsymbol{Y}}_s \tag{27}$$

其中 $A_{ts} = B_{ts}\sqrt{\alpha_s}$。因此，

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$$

如最优算法所示。对于 V_{new}，最优估计量 $\hat{}$ 为 $\text{sgn}(\langle \boldsymbol{Y}_{\text{新}}\ \boldsymbol{U}_t\rangle) = \text{sgn}(\langle \boldsymbol{Y}_{\text{新}}\ \boldsymbol{Y}_t\rangle)$。

7 结论

本文提出了一种多任务学习的高斯混合模型，其中每个任务都是一个半监督分类问题。我们推导出了贝叶斯风险的显式公式，并通过各种数值模拟研究了该模型的行为。

本文中的模型涉及高斯随机变量和拉ademacher随机变量。不过，我们的方法也适用于具有有限二阶矩随机变量的更一般的张量模型。

致谢。我们感谢审稿人提供宝贵的反馈和富有见地的评论，这些意见对改进本文起到了重要作用。我们还要感谢 Malik Tiomoko 就模型算法方面进行的深入讨论，并感谢 Hugues Souchard de Lavoreille 提供实习报告；第一作者在撰写本文期间多次参考了该报告。本研究得到 MIAI 的支持。

参考文献

B. Aubin、F. Krzakala、Y. Lu 和 L. Zdeborov´a。《高维感知机中的泛化误差：通过凸优化逼近贝叶斯误差》。神经信息处理系统进展，33：12199–12210，2020。

J. Baik、G. B. Arous 和 S. P´ech´e. 非零复样本协方差矩阵最大特征值的相变。《概率年刊》，33(5):1643–1697，2005年。

J. Barbier 和 N. Macris. 自适应插值法：一种证明贝叶斯推断中副本公式的简单方案。《概率理论及相关领域》，174:1133–1185，2019年。J. Barbier、N. Macris 和 L. Miolane. 张量估计的分层结构及其互信息。载于*2017*年通信、控制与计算第*55*届年度奥尔顿会议（*Allerton*），第1056–1063页。IEEE，2017年。

J.Barbier, F.Krzakala, N.Macris, L. Miolane, 和 L.Zde- borov´a. 高维广义线性模型中的最优误差和相变。《美国国家科学院院刊》, 116(12):5451–5460, 2019.

C. Bishop. 贝叶斯主成分分析。神经信息处理系统进展, 11, 1998.

F. Krzakala, M. M´ezard, F. Sausset, Y. Sun, 和 L. Zdeborov´a. 压缩感知中的基于统计物理的重建。《物理评论 *X*》, 2(2):021005, 2012.

M. Lelarge 和 L. Miolane. 半监督设定下高斯混合的渐近贝叶斯风险。发表于*2019 IEEE* 第八届多传感器自适应处理计算进展国际研讨会（*CAMSAP*）, 第639–643页。IEEE, 2019a.

M. Lelarge 和 L. Miolane。对称低秩矩阵估计的基本极限。《概率理论及相关领域》，173(3):859–929，2019年b。

T. Lesieur、C. De Bacco、J. Banks、F. Krzakala、C. Moore 和 L. Zdeborov´a。高维高斯混合聚类中的相变与最优算法。载于*2016*年第*54*届*Allerton*通信、控制与计算年会（*Allerton*），第601–608页。IEEE，2016年。

T. Lesieur、L. Miolane、M. Lelarge、F. Krzakala 和 L. Zdeborov´a。尖峰张量估计中的统计与计算相变。载于*2017*年*IEEE*国际信息论研讨会（*ISIT*），第511–515页。IEEE，2017年。

B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala 和 L. Zdeborov´a。使用广义线性模型学习高斯混合：高维情形下的精确渐近分析。神经信息处理系统进展，34：10144–10157，2021。

X. Mai 和 R. Couillet。适用于高维数据的一致半监督图正则化。机器学习研究杂志，22：94–1，2021。

M. Mezard 和 A. Montanari。信息、物理与计算。牛津大学出版社，2009。

F. Mignacco、F. Krzakala、Y. Lu、P. Urbani 和 L. Zdeborova。正则化在高维含噪高斯混合分类中的作用。

载于国际机器学习会议，第 6874–6883 页。PMLR，2020年。L. Miolane。低秩矩阵估计的基本极限：非对称情形。*arXiv* 预印本 *arXiv:1702.00473*，2017年。B. R. Paredes、A. Argyriou、N. Berthouze 和 M. Pontil。利用多任务学习中的无关任务。载于人工智能与统计，第 951–959 页。PMLR，2012年。N. G. Polson 和 S. L. Scott。支持向量机的数据增强。贝叶斯分析，6(1):1–23，2011年。S. Ruder。深度神经网络中多任务学习综述。*arXiv* 预印本 *arXiv:1706.05098*，2017年。C. Thrampoulidis、S. Oymak 和 M. Soltanolkotabi。多类别分类的理论洞见：高维渐近视角。神经信息处理系统进展，33:8907–8920，2020年。M. Tiomoko、R. Couillet 和 F. Pascal。基于主成分分析的多任务学习：一种随机矩阵方法。*arXiv* 预印本 *arXiv:2111.00924*，2021a年。M. Tiomoko、H. Tiomoko 和 R. Couillet。解读并优化多任务学习：一种随机矩阵方法。载于*ICLR 2021*第九届国际学习表征会议，2021b年。M. E. Tipping 和 C. M. Bishop。概率主成分分析。皇家统计学会杂志：*B*辑（统计方法论），61(3):611–622，1999年。

设定与主要结果

本文总结了总体设定和主要结果。我们考虑 T 个任务，其中任务 t 是对 $\mathbb{R}^D$ 中的 N_t 个数据点进行分类，这些数据点属于两个具有相同协方差 $\sigma_t^2 I_D$的不同高斯簇。每个任务的数据集都只有部分标注。我们在高维设定 $D \to \infty$ 下研究该模型，并假设已知以下参数:

- $\boldsymbol{C} = (C_{tt'})_{t,\ t'=1}^{T}$: 任务相关性，其中对所有 t， $C_{tt} = 1$。
- $\alpha_t = \lim_{D\to\infty} N_t/D$: 过采样比
- $\lambda_t = 1/\sigma_t^2$: 信噪比（SNR）
- η_t: 任务 t中有标签数据的比例

我们关注于 任务 t中新数据点的最小误分类概率，即任务 t的贝叶斯风险 .
结果。 在该模型设定下，当 $D \to \infty$时，任务 t 的贝叶斯风险收敛至

$$1 - \Phi(\sqrt{q_{ut}}),$$

其中 $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2} dx$ 以及 $(q_{ut}, q_{vt})_{t=1}^{T}$ 是方程组的稳定解

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{28a}$$
$$q_{vt} = \eta_t + (1 - \eta_t) F(q_{ut}) \tag{28b}$$

with

$$\begin{aligned}
\boldsymbol{M} &= \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^{T} \\
\boldsymbol{D} &= \operatorname{diag}\{\alpha_t q_{vt}\}_{t=1}^{T} \\
F(q) &= \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1).
\end{aligned}$$

B 特殊情形

我们通过以下特殊情况来检验主要结果。

B.1 不相关任务

这里考虑对所有 $t \neq t'$， $C_{tt'} = 0$ 的情况。矩阵 $\boldsymbol{M}$ 为对角矩阵，并且我们得到每个 t 对应的以下方程

$$\begin{aligned}
q_{ut} &= \frac{1}{\sigma_1^2} \frac{\alpha_t q_{vt}}{\sigma_t^2 + \alpha q_{vt}} \\
q_{vt} &= \eta_t + (1 - \eta_t) F(q_{ut})
\end{aligned}$$

这与分别学习各任务时的不动点方程相同。

B.2 每个任务的数据都服从相同的分布

这里考虑对所有 t， $C_{tt'} = 1$ 和 $\sigma_t = \sigma$ 的情况， $t' = 1$， $\ldots$， T。我们有

$$\boldsymbol{M} = \frac{1}{\sigma^2} \mathbb{1}\mathbb{1}^T, \quad \boldsymbol{D}\boldsymbol{M} = \frac{\boldsymbol{u}\mathbb{1}^T}{\sigma^2}$$

其中 $\boldsymbol{u} = (\alpha_t q_{vt})_{t=1}^{\top}$ 以及 $\mathbb{1} = (\underbrace{1, \ldots\ldots, 1}_{T\ 1s})^{\top}$ 应用公式

$$(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{v}^T}{1 + \boldsymbol{u}^T\boldsymbol{v}}$$

对于 $\boldsymbol{v} = \mathbb{1}/\sigma^2$，我们得到

$$(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1} = \boldsymbol{I} - \frac{\boldsymbol{u}\mathbb{1}^T}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}}$$

so

$$\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1} = \frac{1}{\sigma^2}\frac{\boldsymbol{u}^T\mathbb{1}}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}}\mathbb{1}\mathbb{1}^T \tag{29}$$

由方程 (28a) 可知，对于所有 t,

$$q_{ut} = \frac{1}{\sigma^2}\frac{\boldsymbol{u}^T\mathbb{1}}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}} := q_u \tag{30}$$

定义 α，η 为

$$\alpha = \sum_t \alpha_t, \quad \alpha\eta = \sum_t \alpha_t\eta_t \tag{31}$$

我们有

$$\begin{aligned}\boldsymbol{u}^T\mathbb{1} &= \sum_t \alpha_t q_{ut} \\ &= \sum_t \alpha_t(\eta_t + (1-\eta_t)F(q_u)) \\ &= \alpha\eta + \alpha(1-\eta)F(q_u) \\ &= \alpha q_v\end{aligned} \tag{32}$$

其中 q_v 定义为

$$q_v = \eta + (1-\eta)F(q_u) \tag{33}$$

则由 (30) 和 (32) 可得

$$q_u = \frac{1}{\sigma^2}\frac{\alpha q_v}{\sigma^2 + \alpha q_v} \tag{34}$$

由于（33）和（34）恰好是参数为 α 和 η 的单任务学习情形下的方程，因此多任务学习问题可化简为一个参数为 α, η（由（31）给出）的单任务问题。

C 无监督学习与相变

C.1 不可能恢复区域

在无监督情形下，不动点方程为

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{35a}$$

$$q_{vt} = F(q_{ut}) \tag{35b}$$

它总是以 $(\boldsymbol{q}_u, \boldsymbol{q}_v) = (\boldsymbol{0}, \boldsymbol{0})$ 为解。当且仅当该解稳定时，分类才是不可能的。为了分析（35）在零点附近的稳定性，令 $q_{ut}, q_{vt} = O(h)$，其中 $h \to 0$。对于维度相同的向量 A 和 B，若 $|A - B| \simeq O(h^2)$，我们记作 $A \simeq B$，其中 $|.|$ 表示欧几里得范数。根据

$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \tag{36}$$

（附录 E.1），利用 $\tanh(x) = x - x^3/3 + o(x^3)$ 的泰勒展开，可得

$$q_{vt} = F(q_{ut}) \simeq q_{ut}$$

另一方面,

$$\begin{aligned} q_{ut} &= [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \\ &\simeq [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{D}\boldsymbol{M})]_{tt} \\ &= [\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}]_{tt} \\ &= \sum_{s=1}^{T} M_{ts}^2 \alpha_s q_{vs} \end{aligned}$$

Let

$$\boldsymbol{P} = (M_{ts}^2 \alpha_s)_{s,t=1}^T = \Big(\frac{C_{ts}^2}{\sigma_t^2 \sigma_s^2}\alpha_s\Big)_{s,t=1}^T = (\lambda_s \lambda_t C_{st}^2 \alpha_s)_{s,t=1}^T \tag{37}$$

当 n 较小时在（$\mathbf{0}$，$\mathbf{0}$）的邻域内，该方程组可近似到误差为 $O(h^2)$ 的程度:

$$\boldsymbol{q}_v = \boldsymbol{q}_u \tag{38}$$

$$\boldsymbol{q}_u = \boldsymbol{P}\boldsymbol{q}_v \tag{39}$$

因此，当且仅当 $\boldsymbol{P}$ 的每个特征值的模不大于 1 时，不动点（$\mathbf{0}$,$\mathbf{0}$）稳定。利用一般方阵 A, B 的 AB 和 BA 具有相同特征值这一性质，矩阵 $\boldsymbol{P}$ 与下列对称矩阵具有相同的特征值

$$\boldsymbol{R} = (\sqrt{\alpha_s \alpha_t}\lambda_s \lambda_t C_{st}^2)_{s,t=1}^T \tag{40}$$

注意， $\boldsymbol{R}$ 是一个半正定（p.s.d）矩阵，因为它可以表示为半正定矩阵的 Hadamard乘积。因此，当且仅当 $\boldsymbol{R}$ 的所有特征值均不大于 1 时，分类是不可能的。

当所有 $t \neq t'$ 和 $\lambda_t = \lambda, \alpha_t = 1$ 均满足 $C_{tt'} = c$，且所有 t 均满足 $\lambda_t = \lambda$ 时，我们有

$$\boldsymbol{R} = \lambda^2(c^2 \mathbb{1}\mathbb{1}^T + (1 - c^2)I) \tag{41}$$

注意，矩阵 $\mathbb{1}\mathbb{1}^T$的特征值为 0，.....，0， T，因此 $\boldsymbol{R}$的最大特征值为 $\lambda^2(1 + (T-1)c^2)$，由此可得不可能分类的条件

$$\lambda^2(1 + (T-1)c^2) \leq 1 \tag{42}$$

在特殊情形 $T = 1$下，该条件变为 $\lambda \leq 1$ 。

当 $T = 2$，任务相关性为 c 且 $\alpha_1 = \alpha_2 = 1$时，我们有

$$\boldsymbol{R} = \begin{pmatrix} \lambda_1^2 & c^2\lambda_1\lambda_2 \\ c^2\lambda_1\lambda_2 & \lambda_2^2 \end{pmatrix} \tag{43}$$

显然，不可能分类的（λ_1， λ_2）域是 $[0,1]^2$,的子集；否则，至少有一个任务是可行的。当且仅当 $\mathrm{Tr}(\boldsymbol{I} - \boldsymbol{R}) \geq 0$ 且 $\det(\boldsymbol{I} - \boldsymbol{R}) \geq 0$时， $\boldsymbol{R}$ 的所有特征值都小于 1。对于（λ_1， λ_2） $\in [0,1]^2$，第一个条件已满足，而第二个条件等价于

$$(1 - \lambda_1^2)(1 - \lambda_2^2) \leq c^4 \lambda_1^2 \lambda_2^2 \tag{44}$$

C.2 连通任务要么全部可行，要么全部不可能

在无监督情形中，如果任意两个任务通过其他任务直接或间接相关联，则称这些任务是连通的。我们将证明，如果任务是连通的，那么要么所有任务都可行，要么所有任务都不可能。提醒一下，对于任意任务 t， q_{ut} 的值始终非负。如果 $q_{ut} = 0$, 则任务 t 不可能；否则，该任务可行。

考虑具有输出 $(Y_t)_{t=1}^T$、信号 $(X_t)_{t=1}^T$ 的 T高斯信道，其联合分布为 $\mathcal{N}(\mathbf{0}, \boldsymbol{M})$，且噪声为相互独立的标准高斯噪声。每个信道的信噪比分别为 $(\alpha_t q_{vt})_{t=1}^T$。于是，(35a) 右侧对应于信号 X_t 与其最小均方误差估计器之间的重叠度（附录 E.2）。

反设这些任务可以划分为两个非空集合 S 和 S'，使得对于所有 $t \in S$， $q_{ut} = 0$ 成立，而对于所有 $t \in S'$，$q_{ut} > 0$ 成立。由于这些任务是连通的，存在相关任务 t, t'，使得 $t \in S, t' \in S'$。因此，存在 t, t'，使得 $q_{ut} = 0$,$q_{ut'} > 0$ 且 $C_{tt'} \neq 0$。

由于 $\mathbb{E}[X_t X_{t'}] = M_{tt'} = C_{tt'}/(\sigma_t \sigma_{t'}) \neq 0$， X_t 与 $X_{t'}$相关。此外，由于 $q_{vt'} = F(q_{ut'})$ 且 $q_{ut'} > 0$，我们有 $q_{vt'} > 0$。因此， X_t 不独立于 $\boldsymbol{Y} = \left\{\sqrt{\alpha_s} q_{vs} X_s + Z_s\right\}_{s=1}^{T}$，从而导致 $q_{ut} = \mathbb{E}[X_t \mathbb{E}[X_t | \boldsymbol{Y}]] > 0$，矛盾。

D 从数据中估计模型参数

尽管假设模型参数 $\boldsymbol{C}$ 和（σ_t）可用于分析，但我们在此表明，在每个任务中都有正比例的标注数据可用，即对所有 t均有 $\eta_t > 0$，当 $D \to \infty$ 时，它们确实可以被估计，且误差趋于零。首先考虑监督学习的情形。令

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Y}_{ti} \tag{45}$$

于是我们有

$$\bar{\boldsymbol{Y}}_t = \boldsymbol{U}_t + \sqrt{\frac{\sigma_t^2}{N_t}} \bar{\boldsymbol{Z}}_t \tag{46}$$

其中

$$\bar{\boldsymbol{Z}}_t = \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Z}_{ti} \tag{47}$$

显然 然，对于 $t = 1, \ldots, T$，有 $\bar{\boldsymbol{Z}}_t \overset{i.i.d}{\sim} \mathcal{N}(0, I_D)$。我们考虑如下的 $C_{tt'}$ 估计量，适用于 $t \neq t$ ′：

$$\hat{C}_{tt'} = \langle \bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Y}}_{t'} \rangle \tag{48}$$

将（46）代入 $\hat{} C_{tt'}$ 的定义，并利用 $\langle \bar{\boldsymbol{Z}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(\sqrt{D})$、 $\langle \bar{\boldsymbol{U}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(\overline{1})$（这两个结果直接由中心极限定理推出），可得 $\hat{} C_{tt'} = C_{tt'} + O(D^{-1/2})$。此外

$$\|\bar{\boldsymbol{Y}}_t\|^2 = 1 + \frac{\sigma_t^2}{\alpha_t} + O(D^{-1/2}), \tag{49}$$

由此也可以估计 σ_t。

在所有任务中有标签数据的比例均为正的情况下，我们可以将上述估计量限制在有标签数据上，并获得 $\boldsymbol{C}$ 和（σ_t）的近似值；当 $D \to \infty$ 时，误差趋于零。

E 简单高斯信道

E.1 拉ademacher信号。

考虑如下高斯信道

$$Y = \sqrt{\lambda} X + Z, \tag{50}$$

其中，拉ademacher信号 X 以相等的概率取值为 1 和 -1，且标准高斯噪声 Z 与 X相互独立。我们得到

$$\begin{aligned} P(x|Y) &= \frac{P(x) P(Y|x)}{P(Y)} \\ &\propto e^{-(Y - \sqrt{\lambda} x)^2/2} \\ &\propto e^{\sqrt{\lambda} Y x}, \end{aligned} \tag{51}$$

由此可得后验分布：

$$P(x|Y) = \frac{e^{\sqrt{\lambda}Yx}}{2\cosh(\sqrt{\lambda}Y)} \tag{52}$$

以及最小均方误差估计器 $\hat{X}_{\text{MMSE}} = \mathbb{E}[X|Y]$ as

$$\hat{X} = \sum_{x=\pm 1} xP(x|Y) = \tanh(\sqrt{\lambda}Y). \tag{53}$$

因此，最小均方误差估计器与信号之间的重叠度为

$$\begin{aligned}
\mathbb{E}[X\hat{X}_{\text{MMSE}}] &= \mathbb{E}[X\tanh(\sqrt{\lambda}(\sqrt{\lambda}X + Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda + \sqrt{\lambda}Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda - \sqrt{\lambda}Z)] \\
&= \mathbb{E}[\tanh(\sqrt{\lambda}Z + \lambda)]
\end{aligned} \tag{54}$$

接下来，对于任意估计量 $\hat{X}$ of X，误差 $\mathbb{P}(\hat{X} \neq X)$ 由最大似然估计量最小化：

$$\begin{aligned}
\hat{X}_{\text{ML}} &= \operatorname*{argmax}_{x=\pm 1} P(x, Y) \\
&= \operatorname*{argmax}_{x=\pm 1} e^{\sqrt{\lambda}Yx}
\end{aligned} \tag{55}$$

由此可得最大似然估计量：

$$\hat{X}_{\text{ML}} = \text{sgn}(Y). \tag{56}$$

因此，贝叶斯风险为

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}_{ML}) &= \frac{1}{2}\mathbb{P}(X = 1, \hat{X}_{ML} = -1) + \frac{1}{2}\mathbb{P}(X = -1, \hat{X}_{ML} = 1) \\
&= \frac{1}{2}\mathbb{P}(X = 1, Y < 0) + \frac{1}{2}\mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(Z > \sqrt{\lambda})
\end{aligned}$$

E.2 相关高斯信号

考虑 T高斯信道，其中信号 X_1、.....、 X_T的联合分布为 $\mathcal{N}(\mathbf{0},\ \boldsymbol{M})$，且与高斯噪声 Z_1、.....、 Z_T相互独立；这些噪声彼此独立，且服从 $\mathcal{N}(0,\ 1)$ 分布。具体而言，我们有：

$$Y_t = \sqrt{\lambda_t}X_t + Z_t, \quad t = 1, \ldots, T.$$

令 $\hat{X}_t = \mathbb{E}[X|\boldsymbol{Y}]$ 为 X_t的最小均方误差估计器。由于 $(X_t, Y_1, \ldots, Y_T)$ 是一个高斯向量，$\hat{X}_t$ 是 $Y_1, \ldots, Y_T$的线性组合。因此

$$\begin{aligned}
\text{MMSE}_t &:= \mathbb{E}[(X_t - \hat{X}_t)^2] \\
&= \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \mathbb{E}\big[(X_t - \langle \boldsymbol{\beta}_t, \boldsymbol{Y}\rangle)^2\big].
\end{aligned}$$

这可以写成一个二次优化问题

$$\text{MMSE}_t = \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \left\{ M_{tt} - 2\boldsymbol{a}_t^T \boldsymbol{\beta}_t + \boldsymbol{\beta}_t^T \boldsymbol{A} \boldsymbol{\beta}_t \right\}$$

with

$$\begin{aligned}
\boldsymbol{a}_t &= (\mathbb{E}[X_t Y_s])_{s=1}^T = \left(\sqrt{\lambda_t} M_{ts} \right)_{s=1}^T = \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{e}_t \\
\boldsymbol{A} &= (\mathbb{E}[Y_s Y_{s'}])_{s,s'=1}^T = \left(\sqrt{\lambda_s \lambda_{s'}} M_{ss'} + \delta_{ss'} \right)_{s,s'=1}^T = \boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2}.
\end{aligned}$$

此优化问题存在唯一最小化解 $\boldsymbol{\beta}_t = \boldsymbol{A}^{-1} \boldsymbol{a}_t$，由此我们得到

$$\hat{\boldsymbol{X}} = \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2})^{-1} \boldsymbol{Y} \tag{57}$$

$$\text{MMSE}_t = [\boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt} \tag{58}$$

$$\mathbb{E}[X_t \hat{X}_t] = [\boldsymbol{M} - \boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt}. \tag{59}$$

F 均匀先验渐近地服从高斯分布

为了根据模型中指定的先验分布生成 ($\boldsymbol{U}_1$、……、$\boldsymbol{U}_T$)，我们按以下步骤进行：

1. 生成 $\boldsymbol{Z}_1$、……、$\boldsymbol{Z}_T \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}$、$I_D)$。
2. 使用格拉姆–施密特过程将 $\boldsymbol{Z}_1$、……、$\boldsymbol{Z}_T$ 正交归一化，得到正交归一向量 $\boldsymbol{S}_1$、……、$\boldsymbol{S}_T$
3. $(\boldsymbol{U}_1$，……，$\boldsymbol{U}_T) = (\boldsymbol{S}_1$，……，$\boldsymbol{S}_T)\boldsymbol{C}^{1/2}$，其中（$\boldsymbol{U}_1$，……，$\boldsymbol{U}_T$）表示列为 $\boldsymbol{U}_1$，……，$\boldsymbol{U}_T$ 的 $D \times T$矩阵。

在高维极限下，向量 $\boldsymbol{Z}_1$，……，$\boldsymbol{Z}_T$ 渐近正交，因此正交归一化步骤产生近似的 $n^{-1/2}(\boldsymbol{Z}_1$，……，$\boldsymbol{Z}_T)$，这意味着如果 $\boldsymbol{W}_t = \sqrt{D}\boldsymbol{U}_t$，那么 $\boldsymbol{W}_t$ 服从渐近高斯分布，其协方差为

$$\mathbb{E}[W_{ti} W_{t'j}] = \delta_{ij} C_{tt'} \tag{60}$$

值得注意的是，这直接源于统计物理学中正则系综与微正则系综之间的等价性。