arXiv:2305.02573v1 [stat.ML] 4 May 2023

拉普拉斯正则化分层模型中的联合图学习与模型拟合

Ziheng Cheng*† Junzi Zhang*‡ Akshay Agrawal § Stephen Boyd ¶

摘要

拉普拉斯正则化分层模型（LRSM）利用由类别特征定义的子问题的显式或隐式网络结构，这些类别特征称为层（例如年龄、地区、时间、预测范围等），并借助相邻层的数据来改进各子问题的参数学习。这类模型已广泛应用于机器学习和信号处理问题，包括但不限于时间序列预测、表示学习、图聚类、最大间隔分类和一般的少样本学习。然而，关于 LRSM 的现有研究要么假设图已知，要么局限于特定应用。本文首先展示图权重在 LRSM 中的重要性及其敏感性，并从理论上证明，当节点间的参数尺度和样本量严重不平衡时，这种敏感性可能变得任意大。随后，我们提出一种通用方法，通过求解单个优化问题，在拟合模型参数的同时联合学习图。我们从图连通性的角度和端到端贝叶斯视角阐释所提出的形式，并提出一种高效算法来求解该问题。尽管拉普拉斯正则化项缺乏现有文献通常要求的全局强光滑性，我们仍为所提出的优化算法提供了收敛保证，这一点本身也可能具有独立价值。最后，我们通过多个真实世界数值示例，展示了我们的方法相较于现有方法的效率。

1 引言

分层模型。 本文考虑为一组已确定的类别特征拟合分层模型，也就是在共享一个基础模型的情况下，同时为每个类别拟合多个（相互关联的）模型。这类模型在实践中无处不在，包括但不限于个性化点击率预测、多区域天气预报和多期限资产定价，其中类别分别对应用户类型、区域和预测期限。

*对本研究的贡献相同。

†北京大学数学科学学院。**电子邮箱：**alex-czh@stu.pku.edu.cn

‡Citadel Securities（相关工作完成于加入 Citadel Securities 之前）。**电子邮箱：**saslascroyale@gmail.com§Marimo Inc. **电子邮箱：**akshay@marimo.io

¶斯坦福大学电气工程系。**电子邮箱：**boyd@stanford.edu

更具体地说，给定形式为 $(z,\ x,\ y) \in \{1,\dots,K\} \times \mathcal{X} \times \mathcal{Y}$的数据记录，其中 z 是已确定的分类特征（我们称之为层，其取值有 K 种），我们据此进行分层； $x \in \mathcal{X}$ 是特征， $y \in \mathcal{Y}$ 是输出或结果； $\mathcal{X}$ 和$\mathcal{Y}$ 可以是数值型、类别型或任何其他数据类型。因此，分类和回归问题都包含在内。在某些情况下，我们没有 x ，因此数据记录的形式为 $(z,\ y)$。

拉普拉斯正则化分层模型。拉普拉斯正则化分层模型（LRSM）是 [TBB19] 的解。

$$\text{minimize}_{\Theta} \quad \sum_{k=1}^{K}(l_k(\theta_k) + r(\theta_k)) + \mathcal{L}(\theta_1,\dots,\theta_K), \tag{1.1}$$

其中优化变量为 $\Theta = [\theta_1 \cdots \theta_K] \in \mathbf{R}^{n\times K}$。这里， $l_k(\theta) = \sum_{i:z_i=k} l(\theta, x_i, y_i)$ 是第 k 个局部损失函数； $r : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ 是各个模型的正则化项（其中无穷值表示对允许的模型参数的约束）。

(1.1) 中的最后一项是拉普拉斯正则化，它鼓励在某个加权图下相邻的z具有相似的参数。它由 $W \in \mathbf{S}^K$ 表示，这是一个对称权重矩阵，对角线元素为零，非对角线元素为非负值。拉普拉斯正则化的形式为

$$\mathcal{L}(\Theta) = \mathcal{L}(\theta_1,\dots,\theta_K) = \frac{1}{2}\sum_{i=1}^{K}\sum_{i<j} W_{ij}\|\theta_i - \theta_j\|^2,$$

其中，当 θ_z 是向量时，范数为欧几里得范数或 ℓ_2 范数；当 θ_z 是矩阵时，范数为弗罗贝尼乌斯范数。我们将 W 视为定义了一个加权相似性图，其边对应于 W 的正元素，边权重为 W_{ij}。 W_{ij} 越大，就越鼓励 θ_i 和 θ_j 彼此接近。我们还可以将拉普拉斯正则化写成半正定二次型。

$$\mathcal{L}(\Theta) = (1/2)\mathbf{Tr}(\Theta L\Theta^T) = (1/2)\mathbf{Tr}(\Theta^T\Theta L),$$

且 $L = \mathcal{G}(W) \in \mathbf{R}^{K\times K}$ 是与该加权图相关联的（加权）拉普拉斯矩阵，其中 $\mathcal{G} : \mathbf{R}^{K\times K} \to \mathbf{R}^{K\times K}$ 是如下定义的线性映射：

$$\mathcal{G}(W)_{ij} = \begin{cases} -W_{ij} & i \neq j, \\ \sum_{k\neq i} W_{ik} & i = j. \end{cases}$$

在文献中，通常假设图边权重 W_{ij} 为非负且先验已知，数值越大表示相邻模型参数之间的接近程度越强。然而，这些权重通常由启发式方法确定，即使事先已知不同子模型之间相对准确的接近关系，这种做法也可能不准确且带有任意性（尤其是在尺度方面）。本文旨在拟合分层模型的同时，学习数据背后的图拓扑。我们将这一结合了图论与概率/统计建模思想的新方法称为联合拉普拉斯分层模型。

在本文中，我们始终假设 l_k 和 r 均为闭真函数。此外，假设 l_k 可微，而 r 具有闭式且易于计算的近端算子（例如， ℓ_1、岭正则化和弹性正则化）。

主要贡献。 我们的贡献主要有三点。首先，我们对零样本/少样本设置下 LRSM 中图权重的敏感性进行了理论分析和数值分析，结果表明，传统方法通过人工设定边权重的效率低下，尤其是在样本量和参数尺度严重不平衡时。其次，当图结构未知时，我们提出了一种新框架，通过对 Θ 和 W 求解联合优化问题，同时拟合 LRSM 并学习未知图。我们还从图论和概率模型的角度阐释了这一设计。此外，由于所得优化问题通常是非凸且非光滑的，我们采用加速近端梯度法求解，并在缺少全局强光滑性（这是现有文献通常要求拉普拉斯正则化项具备的性质）的情况下提供收敛保证。最后，我们还通过多个真实世界数据示例，展示了我们的方法相较于现有方法的有效性。

2 动机：图权重敏感性

在本节中，我们旨在说明，当节点间的参数尺度和样本量严重失衡时，即使已知图结构，LRSM 仍会受到边权重敏感性的影响。这里的敏感性是指，边权重的微小变化会导致较大的误差。鉴于此，与网格搜索和交叉验证等传统方法相比，基于梯度的权重端到端优化效率要高得多。

考虑这样一种情形：大多数层拥有大量数据样本，因此即使没有拉普拉斯正则化，也能准确估计。然而，少数层只能获取少量甚至没有数据样本，这在少样本学习或零样本学习中很常见。这样一来，要拟合这些节点上的模型参数，就需要设计良好的图拉普拉斯来传递数据信息。但如果参数的尺度在图上变化不平滑，那么边权重的轻微扰动就会导致估计参数发生剧烈变化。

粗略而言，当由 k_0 表示的节点只有很少或没有样本时，拉普拉斯正则化分层模型 (1.1) 的输出将是

$$\widehat{\theta}_{k_0} \approx \frac{\sum_{j\neq k_0} W_{jk_0}\widehat{\theta}_j}{\sum_{j\neq k_0} W_{jk_0}} \approx \frac{\sum_{j\neq k_0} W_{jk_0}\theta_j^*}{\sum_{j\neq k_0} W_{jk_0}}$$

$\widehat{}$前提是所有其他子模型都拟合良好。因此，对于任意 $i \neq k_0$， θ_{k_0} 关于 w_{ik_0} 的导数近似为

$$\frac{d\widehat{\theta}_{k_0}}{dW_{ik_0}} \approx \frac{\sum_{j\neq k_0, j} W_{jk_0}(\theta_i^* - \theta_j^*)}{(\sum_{j\neq k_0} W_{jk_0})^2}.$$

$\widehat{}$如果 $||\theta_i^*||$ 非常大，那么 $||\frac{d\theta_{k_0}}{dW_{ik_0}}||$ 也会很大，因此 $\widehat{\theta}_{k_0}$ 对边权重很敏感。我们在零样本设置下更严谨地阐述这一观点。证明见附录A。

定理 1。 考虑一个局部正则化为零的分层模型 (1.1)，其中一个记作 k_0 的层没有数据样本。将边权重矩阵记为 W。考虑一个较小的

对图拉普拉斯 $L = \mathcal{G}(W)$ 的扰动。具体而言，对于任意固定的 $i \neq k_0$、$\widetilde{\ }$，我们将 $W_{ik_0} > 0$ 替换为 $W_{ik_0} := W_{ik_0} + \epsilon$，其中 $\epsilon > 0$ 为某个值，而其他边保持不变
$\widetilde{\ }$，并将扰动后的拉普拉斯矩阵记为 L。假设在给定 L 和 $L^{\widetilde{\ }}$ 的情况下，(1.1) 的解分别为 Θ 和 Θ，且真实参数为 Θ^*。若对于某些 δ、$\delta' > 0$，两个解都是(δ、δ')-准确的，即最大值$_{j\neq k_0} \frac{||\theta_j - \theta_j^*||}{||\theta_j^*||} \leq \delta$、$\frac{||\theta_{k_0} - \theta_{k_0}^*||}{||\theta_{k_0}^*||} \leq \delta'$，最大值$_{j\neq k_0} \frac{||\widetilde{\theta}_j - \theta_j^*||}{||\theta_j^*||} \leq \delta$、$\frac{||\theta_{k_0} - \theta_{k_0}^*||}{||\theta_{k_0}^*||} \leq \delta'$，那么

$$\frac{\epsilon}{W_{ik_0}} \leq \frac{2(\delta + \delta' + \frac{\delta W_{ik_0}}{S} A_i)}{\left(1 - \frac{(1+2\delta)W_{ik_0}}{S} - \delta\right) A_i - 1 + \frac{W_{ik_0}}{S} - 2(\delta + \delta')} \frac{S}{W_{ik_0}}, \tag{2.1}$$

前提是分母为正。这里 $A_i := \frac{||\theta_i^*||}{\text{最大值}_{j\neq i} ||\theta_j^*||}$，$S := \sum_{j\neq k_0} W_{jk_0}$。

备注 1. 上述定理表明，当参数尺度与样本量极不平衡时，*LRSM* 对边权重十分敏感。为简化 (2.1)，令 $\frac{W_{ik_0}}{S}$为小于1 的正数常量，并令 δ 接近 *0*。那么，$\frac{\epsilon}{W_{ik_0}} \leq \frac{2\delta'(1 - \frac{W_{ik_0}}{S}) - 2\delta' W_{ik_0}}{(A_i - 1) - 2\delta'}$。如果存在某一层 i，使得 $||\theta_i^*||$相对较大，因而 A_i 也较大，那么$\frac{\epsilon}{W_{ik_0}} \lesssim \frac{\delta'}{A_i}$。换言之，在这种情况下，要获得稳健的解，我们需要高度精确的拉普拉斯矩阵 L，因此，网格搜索结合交叉验证等传统方法代价过高，因为必须使用非常密集的网格。

我们给出一个合成示例，以进一步阐释我们的想法。考虑最小二乘线性回归模型。该图包含 K 个节点，每个节点 v上有 n_k 个（可能为$n_k = 0$）样本（x_{ki}, y_{ki}）$\in \mathbf{R}^n \times \mathbf{R}$，其中 $i = 1, \cdots, n_k$。局部损失函数为平方误差$l_k(\theta) = \sum_{i=1}^{n_k} (y_{ki} - x_{ki}^T \theta)^2$，局部正则化采用平方和。

对于每个维度，我们从多元高斯分布 $\mathcal{N}(\mu, \Sigma)$ 中生成真实参数 Θ^*。输入 $x_{ki} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I_n)$ 和观测值均为 $y_{ki} = x_{ki}^T \theta_k^* + \varepsilon_{ki}$，其中 $\varepsilon_{ki} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2)$。

我们构造一个星形图边权重矩阵 W，其中中心节点记为 k_0，并令$\Sigma = \mathcal{G}(W)^\dagger + \delta I_n$。这里，$\mathcal{G}(W)^\dagger$是拉普拉斯矩阵 $\mathcal{G}(W)$ 的伪逆，且 $\delta > 0$ 足够小，使得 Σ 严格正定。中心层的样本量较小，而其余层可以获取大量样本。我们按照上述方式生成数据，但故意将真实参数的尺度设置得极不平衡。通过求解方程 $(\sum_{k\neq k_0} w_k)\theta_{k_0}^* = \sum_{k\neq k_0} w_k \theta_k^*$，构造最优星形图的边权重 W^*。随后，我们逐渐向 W^* 添加扰动，并使用扰动后的拉普拉斯矩阵 $\sim$ $\widetilde{\ }$ $W = W^* + \epsilon W_1$ 拟合模型，其中 W 仍然是星形图的边权重矩阵。具体而言，我们令 $K = 30$，$n = 20$。所有非中心层的样本量均为 100。W_1 是一个星形图的边权重矩阵，它是稀疏的，且其元素服从 Unif($[0, 1]$) 分布。随后，我们对 W_1 重新缩放，使其弗罗贝尼乌斯范数与 W^* 的相同。

我们考察了两组参数尺度失衡程度不同的情形，如图2所示。我们对两组分别进行了50次 W_1 随机试验。图1显示了中心层处估计参数与真实参数之间的归一化距离，$\widehat{\ }$即，$||\theta_{k_0} - \theta_{k_0}^*||/||\theta_{k_0}^*||$ 随扰动 ϵ的变化。实线表示均值，而

阴影区域表示90%置信区间。请注意， W_1 的尺度与 W^*相同，因此当 ϵ 足够小时，曲线近似为直线（直观而言，这是通过一阶泰勒展开线性化得到的）。尽管扰动幅度很小，我们仍观察到性能显著变化，尤其是在参数尺度失衡更严重时。比较图1和图2中的左、右子图还可以看出，参数尺度失衡程度越高，敏感性就越强。

此外，我们增大中心层的样本量，并在图3中展示不同样本量下归一化距离随扰动变化的斜率。这表明，我们的方法也适用于少样本学习。

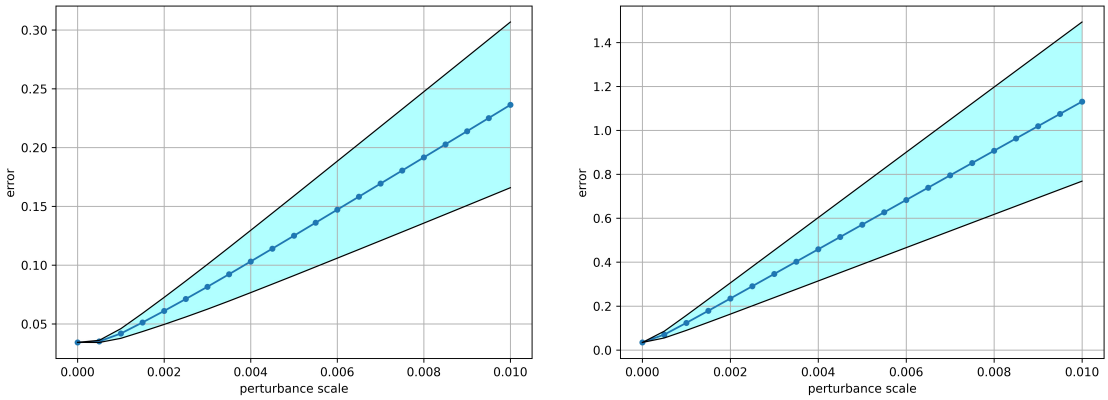

图1： 拉普拉斯矩阵的敏感性

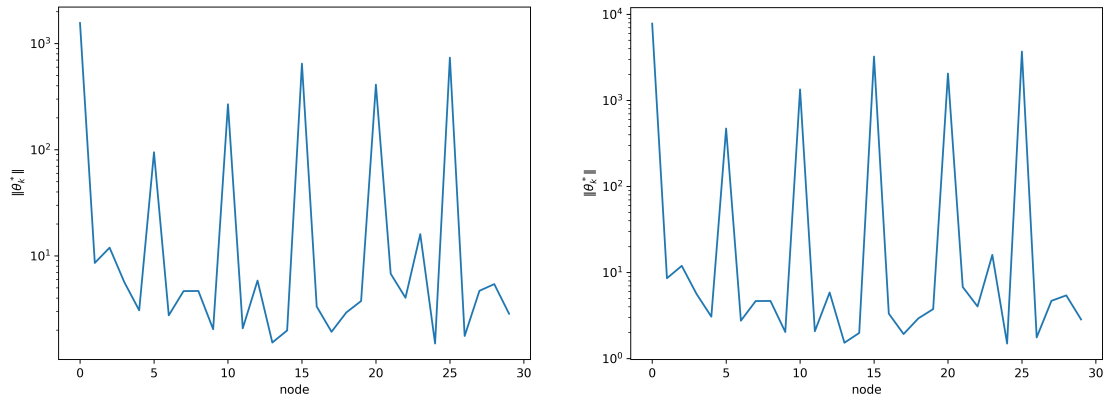

图2： 真实参数的范数（最后一个为中心层）

3 我们的方法

为了处理未知图和边权重，一个直接的解决方案是联合最小化 Θ 和 W，其中 $W \geq 0$ 取自 (1.1)。然而，直接这样做会导致无意义的解 $W = 0$。为了解决这个问题，我们需要对W 施加一些额外的准则，以抵消平凡最小化带来的这种影响。我们首先将提出的新

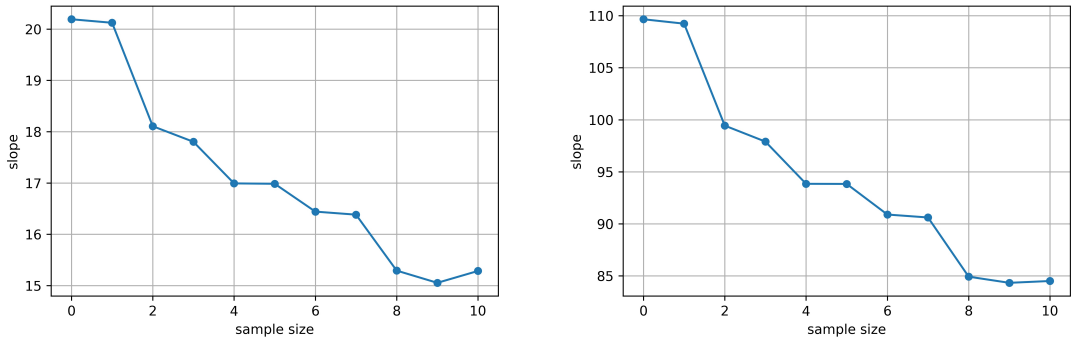

图3： 不同样本量下的斜率

框架作为一种启发式方法进行阐述，然后从图论和统计学两个角度对其进行验证。

3.1 联合拉普拉斯分层模型

在本文中，我们提出在 (1.1) 中添加两个额外项（其中 $\lambda_1,\ \lambda_2 \geq 0$）：

$$\begin{array}{ll} \text{minimize}_{W,\Theta} & \sum_{k=1}^{K}(l_k(\theta_k)+r(\theta_k))+\frac{1}{2}\sum_{i=1}^{K}\sum_{i<j}W_{ij}\|\theta_i-\theta_j\|_2^2+\lambda_1 R_1(W)+\lambda_2 R_2(W), \\ \text{subject to} & W\geq 0,\ \mathbf{diag}(W)=0,\ W\in\mathbf{S}^K. \end{array} \tag{3.1}$$

这里的第一项

$$R_1(W) = -\log\det(\mu I + \mathcal{G}(W)) \tag{3.2}$$

是正则化拉普拉斯矩阵 $\mathcal{G}(W)$ 的负对数行列式，它能够提升图的可访问性、连通性和鲁棒性。或者，也可以通过概率模型轻松推导得出。更多详情请参见 §3.2。此外，它还能防止 W 退化为零，从而增大其幅度。这里 $\mu > 0$ 是一个正数常量。

第二项（带有 $\eta \in [0,1]$）

$$R_2(W) = \frac{1-\eta}{2}\|W - W_0\|_F^2 + \eta\|W\|_1 \tag{3.3}$$

是作用于 W 的一种弹性网络式正则化项，由一个近端项和一个辅助惩罚项组成：前者用于使其接近某个先验猜测 $W_0 \in \mathbf{R}_+^{K\times K}$，后者则鼓励相似性图中的边稀疏。这里，$\|\cdot\|_1$ 是逐元素的 ℓ_1 范数。在 [ZLWN21] 中也采用了使用先验猜测 W_0 的思路。本文始终假设 W_0 的对角线元素为零。

示例选择 W_0。我们可以简单地将 W_0 设为零矩阵，也可以像 [TBB19],中那样，根据先验信息、人工设计和交叉验证的组合来设定；正如我们将在下文的数值实验中看到的，这两种做法在实践中都非常有效。此外，下文还将介绍三种稍微高级一些的方法，用于获得非平凡（且可能更优）的初始猜测 W_0。为简单起见，而且由于朴素方法已经效果良好，

正如我们在下文的实验中所示，这些更高级的 W_0 选择方法尚未在数值实验中进行测试；其优缺点有待今后研究。

第一种方法依赖于关于分层变量 z 的某些度量。例如，当 $z \in \{1, \dots, K\}$ 的每个取值都自然对应于度量空间 $\mathcal{M}$ 中的某种表示 $f_z \in \mathcal{M}$，且该度量空间配有度量 d 时，可以设定 $[W_0]_{ij} = w(f_i, f_j)$，其中 $w\colon \mathcal{M} \times \mathcal{M} \to \mathbf{R}_+$ 是某个“核”函数，也就是说，$w(f_i, f_j)$ 在 f_i 和 f_j 按度量d衡量时距离越近的情况下取值越大。以下是 w 的一些常见示例：

- 指数核：$w(f_i, f_j) = \exp(-\tau d(f_i, f_j))$，其中 $\tau > 0$ 为正数常量。
- k-近邻核 [ZDGA20]：$w(f_i, f_j) = 1$，若 f_j 属于 f_i 的 d 度量下的 k 近邻范围内；否则，$w(f_i, f_j) = 0$。

另请参阅 [TBB21]，了解如何在投资组合构建这一特定场景中设计 w。

第二种选择是先独立求解所有子模型，分别得到每个子模型独立学习得到的参数 $\theta_i = \operatorname{argmin}_\theta l_i(\theta) + r(\theta)$ $i = 1$、.....、K，然后定义 $[W_0]_{ij} = w(\theta_i, \theta_j)$，其中 $w\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}_+$ 仍是某种“核”函数。上文的 w 示例在此处同样适用，度量 $d(\theta_i, \theta_j) = \|\theta_i - \theta_j\|_2$。这种方法也可以看作一种通用方法，用于为分层变量 z 寻找表示 f_z（当这些表示无法直接获得时）。需要注意的是，这种方法的一个潜在局限是，当某个分层取值 z 对应的数据点很少甚至没有时，学习得到的参数 θ_z 可能毫无意义。

第三种方法基于自动调优的思想。在这种方法中，会留出一组数据点 $\{z_i^{\mathrm{val}}, x_i^{\mathrm{val}}, y_i^{\mathrm{val}}\}_{i=1}^{N_{\mathrm{val}}}$作为验证集，用于构造以下验证损失函数：

$$L^{\mathrm{val}}(\Theta) = \sum_{k=1}^{K} l_k^{\mathrm{val}}(\theta_k), \tag{3.4}$$

其中，$l_k^{\mathrm{val}}(\theta) = \sum_{i:\ z_i^{\mathrm{val}}=k} l(\theta, x_i^{\mathrm{val}}, y_i^{\mathrm{val}})$。令 $\mathcal{S}\colon \mathbf{R}^{K\times K} \to \mathbf{R}^{n\times K}$ 表示一个映射，将图权重 W 映射到训练问题（1.1）的解 Θ。然后，我们将 W_0 计算为以下问题的解：

$$\begin{array}{ll} \mathrm{minimize}_W & L^{\mathrm{val}}(\mathcal{S}(W)), \\ \mathrm{subject\ to} & W \geq 0,\ \mathbf{diag}(W) = 0,\ W \in \mathbf{S}^K, \end{array} \tag{3.5}$$

该问题在所有可能的非负图权重选择中最小化验证损失。问题（3.5）可以通过梯度下降方法进行最小化，其中梯度是通过对凸规划 [AAB+19]求导来计算的。请注意，[BB20]中已针对最小二乘估计提出了类似方法。直接将通过该方法获得的 W_0 用作图权重的最终选择似乎也是可行的。然而，对于一般的（非最小二乘）问题，除非问题规模较小，否则（1.1）的优化本身就需要大量计算，而且我们发现，如果不做进一步修改，直接将其用作最终权重，通常效果并不理想。

3.2 图论与统计解释

在本小节中，我们从图论和统计学的角度阐释目标函数 (3.1)。首先，我们对与以下内容相关的附加项作出两种解释：

图权重 W 在 (3.1) 中；随后，我们将 (3.1) 作为一个贝叶斯模型的最大后验估计进行端到端阐释。

学习 W 作为生成森林总权重最大化。 我们首先解释加权图G_W 中与 W相关的对数行列式项（3.2）如何与生成森林的总权重相关。假设该图有 m 个连通分量，每个分量包含n_i($i = 1$、.....、 m)个节点。那么，图拉普拉斯矩阵 $\mathcal{G}(W)$的零特征值的重数等于 m。假设其余的 $K - m$ 个非零特征值为$0 < \lambda_1 \leq \cdots \leq \lambda_{K-m}$。那么，加权矩阵树/森林定理 [CK78、CS06b、Pen17]指出

$$\tau(G_W) = \frac{1}{\prod_{i=1}^{m} n_i} \lambda_1 \cdots \lambda_{K-m}, \tag{3.6}$$

其中， $\tau(G_W)$是图 G_W中生成森林的总权重，定义为

$$\tau(G_W) = \prod_{i=1}^{m} \sum_{T \in \mathcal{T}_i} w(T).$$

这里， $\mathcal{T}_i$ 表示分量 i中的所有生成树构成的集合，而 $w(T)$表示树 T的权重，定义为 $w(T) = \prod_{(i,\ j) \in T} W_{ij}$，即， T中所有边的权重之积。

因此我们可以看到，当 $\mu \to 0$ 时，有

$$\begin{aligned} \log\det(\mu I + \mathcal{G}(W)) &= \log\left(\mu^m \prod_{j=1}^{K-m} (\mu + \lambda_j)\right) \\ &= m\log\mu + \log\tau(G_W) + \sum_{i=1}^{m} \log n_i + O(\mu). \end{aligned}$$

这表明，对于较小的正则化参数 $\mu \in (0,1)$，最小化 $-\log\det(\mu I + \mathcal{G}(W))$ 大致可以同时实现三个目标：最大化生成森林的总权重（$\log\ \tau(G_W)$ 项）、最小化连通分量的总数（ $-m\log(1/\mu)$ 项），并使节点数在不同连通分量间的划分尽可能均衡（ $\sum_{i=1}^{m} \log n_i$ 项）。生成森林数量更多、连通分量更少以及节点划分更均衡，都有助于提高图的连通性和可达性 G_W[CS06a]。这能够更稳健、更全面地描述各子模型之间的关系，正是我们所希望的。

层次高斯马尔可夫随机场。 AGMRF [RH05]基于无向图定义如下

$$p(\Theta|W) \propto \exp\left(-E(\Theta)\right),$$

其能量函数为 $E(\Theta) := \frac{1}{2}\sum_{i<j} W_{ij}\|\theta_i - \theta_j\|^2 + \frac{\mu}{2}\|\Theta\|^2$。将其改写后可得到归一化密度

$$p(\Theta|W) = (2\pi)^{-\frac{nK}{2}} \det\left(\mu I + \mathcal{G}(W)\right)^{n/2} \exp\left(-\frac{1}{2} Tr(\Theta(\mu I + L)\Theta^T)\right).$$

而且，我们也可以在 W 上设定一个先验分布，使得

$$p(W) \propto \exp\left(-\frac{c_1}{2}\|W\|_F^2 - c_2\|W\|_1\right),$$

它是高斯分布与拉普拉斯分布的混合分布。此外，考虑一个具有似然函数 $p(y|x,z,\Theta) \propto \exp(-l(y,x,z;\Theta))$ 的分层贝叶斯模型。那么，整个模型的负对数似然为（相差一个加性常数）：

$$-\log p = l(y,x,z;\Theta) + \frac{\mu}{2}\|\Theta\|^2 + \mathcal{L}(\Theta) - \frac{n}{2}\log\det(\mu I + \mathcal{G}(W)) + \frac{c_1}{2}\|W\|_F^2 + c_2\|W\|_1.$$

通过设置 $\lambda_1 = \frac{n}{2}$， $r(\theta) = \frac{\mu}{2}\|\theta\|^2$， $\lambda_2 = c_1 + c_2$， $\eta = \frac{c_2}{c_1+c_2}$，我们发现 (Θ, W) 的最大后验估计恰好由 (3.1) 给出。注意，如果我们设置 $\mu = 0$ 并将伪行列式定义为 $\mathcal{G}(\mathcal{W})$ 的所有非零特征值之积，那么即可恢复 [KYdMCP19] 中的非正规高斯马尔可夫随机场。

4 与现有工作的比较

在继续求解我们提出的联合拉普拉斯分层模型之前，本节首先回顾并总结相关工作，并将我们的方法与其进行比较。

拉普拉斯正则化分层模型拟合。 已知图上的拉普拉斯正则化分层模型拟合已被研究多年 [AZ06, She08], ，近期人们越来越关注提升其可扩展性 [TB20a] ，以及其稳健性和自适应性[DW22, LWWZ22]。此外，近期还有许多研究将这些模型应用于半监督学习 [ST17, CPVBR20], 、联合协方差估计 [TB20b], 、少样本学习[ZDGA20], 、多种市场状况下的投资组合构建 [TBB21] 以及联邦学习 [DVT+21]，仅举数例。

多任务学习。 当将各层解释为任务时，分层模型也可以被视为多任务学习。然而，为了建立基于任务的联系，多任务学习领域的文献主要关注并非拉普拉斯正则化的正则化项 [JVB08, ZCY11, GVZB16, SCST17, EP04, ZY10], 这些正则化项通常不如图拉普拉斯/权重那样明确且易于解释，因此超出了本文的范围。

图学习。 从数据中进行图学习是一个典型问题，已受到文献界的广泛关注。大量研究探讨了如何在利用各种形式的先验知识的同时学习结构化图，包括谱约束、稀疏性、节点度和边模式 [KP17, KYdMCP19, KLZX21]。此外，我们还可以将节点方面的先验知识与观测方面的先验知识结合起来，如 [PCDS20]所示。许多现实世界的应用中经常可以获得结构化信息，包括基因网络分析、社区检测和聚类。然而，实践者应使用特定的算法来处理不同类型的约束。

当缺乏图结构方面的知识时，已有一些从数据（或信号）中学习图的通用方法。在对图结构了解有限的情况下，已有多种从数据或信号中进行图学习的通用方法。在文献中，这些方法大致可分为几类。

分为三类，它们都涉及最小化拉普拉斯项与 W 的正则化函数之和。首先， [DTFV14] 考虑求解

$$\begin{array}{ll}\text{minimize}_L & R_{\text{TC}}(L) := Tr(\Theta L \Theta^T) + \frac{C_2}{2}\|L\|^2 \\ \text{subject to} & Tr(L) = C_1, L_{ij} = L_{ji} \leq 0, L \cdot \mathbf{1} = 0.\end{array} \tag{4.1}$$

我们将（4.1）称为 Tr-约束法。它直接对边权重总和施加硬约束，以避免得到平凡解。但对拉普拉斯矩阵使用弗罗贝尼乌斯范数会带来问题： L 中的元素尺度各异，且线性相关。第二种形式是求解

$$\begin{array}{ll}\text{minimize}_W & R_{\text{LD}}(W) := Tr(\Theta \mathcal{G}(W) \Theta^T) - C_1 \mathbf{1}^T \log(W \cdot \mathbf{1}) + \frac{C_2}{2}\|W\|^2 \\ \text{subject to} & W_{ij} = W_{ji} \geq 0, W_{ii} = 0.\end{array} \tag{4.2}$$

我们将（4.2）称为对数对角法。作用于节点度向量 $W \cdot \mathbf{1}$ 的对数障碍项可以防止退化并提升图的整体连通性，同时不会损害稀疏性 [Kal16]。然而，对数障碍项会严格要求每个节点的度均为正，因此无法允许孤立节点。这一性质通常并不可取，因为某些节点很可能确实与其他节点不存在相关性，甚至呈负相关，因此应在图中将其孤立出来。对于基于熵的第三类方法，这个问题甚至更加严重：

$$\begin{array}{ll}\text{minimize}_W & R_{\text{Ent}}(W) := Tr(\Theta \mathcal{G}(W) \Theta^T) + \sigma^2 \sum_{i \neq j} W_{ij}(\log W_{ij} - 1) \\ \text{subject to} & W_{ij} = W_{ji} \geq 0, W_{ii} = 0,\end{array} \tag{4.3}$$

其闭式解为 $W_{ij} = \exp\left(-\frac{\|\theta_i - \theta_j\|^2}{2\sigma^2}\right)$。我们将（4.3） 称为熵方法。由此得到的图必须是全连通的，这会对性能产生负面影响。与上述另外两种方法相比，这种方法通常不是首选 [Kal16]，但可作为良好的初始猜测（见§3）。

文献中关于图学习的大多数最小假设方法只关注节点的度属性，而忽略了单个节点的边之间的相互作用。然而，如 §3 所示，在给定节点度数的情况下为每条边赋予权重，是一个至关重要却复杂的过程。

基于拉普拉斯正则化的联合图学习。联合学习模型和图并非学术界的新理念，近期已有一些针对特定应用的相关研究。大多数研究将具体任务与上述两种通用图学习方法（Tr-Constraint 和 Log-Diagonal）相结合，或采用针对任务特点进行调整的变体方法。 [VVBD20] 研究了用于基于图的半监督学习的表示学习问题。该方法联合学习表示和底层图，以供下游基于图的半监督学习任务使用。它采用 Log-Diagonal 方法学习图，其目标是使表示能够编码注入图中的标签信息，同时使图相对于变换后的数据提供平滑的拓扑结构。 [HTK22] 研究了金融领域的应用，并假设图具有 k-聚类结构。因此，该方法采用 Tr-Constraint 方法，并在拉普拉斯矩阵上增加了低秩约束，从而得到一个更为

关于图权重的困难优化问题。 [ZLWN21] 采用类似的方法，旨在进行无监督特征选择。它还利用由原始高维数据构建的现有相似性图矩阵作为先验。 [KA22, LZZ+16] 针对异构数据进行图聚类。他们使用 Tr-Constraint 方法，在学习给定数据的标签（社群）的同时学习图拓扑。

与现有方法的比较。 我们的方法最显著的区别在于采用了正则化对数行列式项，这不仅提供了更好的可解释性，还克服了现有方法所存在的局限性。值得注意的是，该方法能够对非连通图甚至孤立节点进行建模，从而提高模型的表达能力，并且与 Log-Diagonal 和 Entropy 方法相比，能够学习更广泛的图结构。例如，当某一层与其他层存在显著差异时，我们的方法能够更加灵活地将其建模为一个孤立节点，从而提高模型的可解释性和准确性。此外，对数行列式项考虑了每条边及其权重之间的相互作用，而不是像 Tr-Constraint、Log-Diagonal 和 Entropy 方法那样，仅利用单条边或单个节点度的信息。这有助于基于谱图论刻画图中几乎所有的关键信息。

5 优化：单调加速近端梯度法

请注意，当假设所有 $l_k(\cdot)$ 和 $r(\cdot)$ 均为凸函数时，训练目标关于 Θ 和 W 是双凸的，因此自然会考虑交替最小化方法。然而，给定 Θ 或 W 时，难以高效求解子问题。鉴于此，我们将 (1.1) 改写如下：

$$\min F(\Theta, W) := f(\Theta, W) + g(\Theta, W) \tag{5.1}$$

其中

$$\begin{aligned} g(\Theta, W) &= \sum_k r(\theta_k) + \lambda_2 \eta \|W\|_1 + \sum_{i<j} \mathcal{I}_{W_{ij}=W_{ji}\geq 0} + \mathcal{I}_{\mathbf{diag}(W)=0} \\ f(\Theta, W) &= \sum_k l_k(\theta_k) + \frac{1}{2}\sum_{i<j} W_{ij}\|\theta_i - \theta_j\|_2^2 + \frac{\lambda_2(1-\eta)}{2}\|W - W_0\|_F^2 \\ &\quad - \lambda_1 \log\det\left(\mu I + \mathcal{G}(W)\right) \end{aligned} \tag{5.2}$$

我们将 f 作为光滑项，并将 g 作为具有近端预言机的简单正则项。由于拉普拉斯正则化，整体问题是一个非凸、非光滑的复合优化问题。这里我们采用 MAPG算法，该算法最初发表于 [LL15]。

5.1 算法

令 $\boldsymbol{x} = (\Theta,\ W)$，具体过程如算法 1 所示。 $k = 1,\ \cdots,\ K$ 关于 l_k 的梯度可以并行计算，因此尽管拉普拉斯正则化项中的变量存在耦合，该算法仍然采用分布式方式。

算法 1 单调 加速 近端梯度

初始化 $\boldsymbol{z}_1 = \boldsymbol{x}_1 = \boldsymbol{x}_0$, $t_1 = 1$, $t_0 = 0$, α_x, $\alpha_y > 0$。**对于** $k = 1, \cdots$ **执行**

$$\begin{aligned}
\boldsymbol{y}_k &= \boldsymbol{x}_k + \frac{t_{k-1}}{t_k}(\boldsymbol{z}_k - \boldsymbol{x}_k) + \frac{t_{k-1}-1}{t_k}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}),\\
\boldsymbol{z}_{k+1} &= \mathbf{prox}_{\alpha_y g}(\boldsymbol{y}_k - \alpha_y \nabla f(\boldsymbol{y}_k)),\\
\boldsymbol{v}_{k+1} &= \mathbf{prox}_{\alpha_x g}(\boldsymbol{x}_k - \alpha_x \nabla f(\boldsymbol{x}_k)),\\
t_{k+1} &= \frac{\sqrt{4(t_k)^2+1}+1}{2},\\
\boldsymbol{x}_{k+1} &= \begin{cases} \boldsymbol{z}_{k+1}, & \text{if } F(\boldsymbol{z}_{k+1}) \le F(\boldsymbol{v}_{k+1}),\\ \boldsymbol{v}_{k+1}, & otherwise \end{cases}
\end{aligned}$$

结束循环

计算 g **的近端算子**。注意，$g(\Theta, W) = \sum_{k=1}^{K} r(\theta_k) + g_2(W)$，其中

$$g_2(W) = \lambda_2 \eta \|W\|_1 + \sum_{i<j} \mathcal{I}_{W_{ij}=W_{ji}\ge 0} + \mathcal{I}_{\mathbf{diag}(W)=0}.$$

根据可分离和函数 [PB+14 的近端算子的基本性质，§2.1]，对任意 $\alpha \ge 0$，我们有：

$$\mathbf{prox}_{\alpha g}(\boldsymbol{x}) = (\mathbf{prox}_{\alpha r}(\theta_1), \ldots, \mathbf{prox}_{\alpha r}(\theta_K), \mathbf{prox}_{\alpha g_2}(W)).$$

如本文开篇所假设，$r(\theta_k)$ 的近端算子具有闭式解，而 $g_2(W)$ 的近端算子为

$$\begin{aligned}
\mathbf{prox}_{\alpha g_2}(U) &= \arg\min_W \lambda_2 \eta \|W\|_1 + \sum_{i,j} \mathcal{I}_{W_{ij}=W_{ji}\ge 0} + \mathcal{I}_{\mathbf{diag}(W)=0} + \frac{1}{2\alpha}\|W-U\|_F^2\\
&= \max\left(0, (U+U^\top)/2 - \lambda_2 \alpha \eta\right),
\end{aligned}$$

其中，max 是逐元素取值，且 U 的对角线元素为零。推导 **近端**$_{\alpha g_2}$ 的更多细节请参见附录C。

复杂度。 $g(\boldsymbol{x})$的近端算子通常具有闭式解，并且可以并行计算。计算复杂度主要由计算 $f(\boldsymbol{x})$的梯度决定，其中包括计算拉普拉斯矩阵 $\mathcal{G}(W) \in \mathbf{S}^K$的逆矩阵，以及在单调性检查步骤中计算对数行列式。实践中，K 通常不会太大，最好为几百。

初始化。 由于该问题是非凸的，不同的初始化可能会得到不同的解。这里我们启发式地提出将 W 初始化为 W_0，这一方法在我们的实验中效果良好。至于 Θ的初始点，我们将其设为零向量。

步长。为简单起见，在所有数值实验中，我们通过网格搜索选择固定步长 α_x，α_y。[LL15] 提出了一种采用 Barzilai-Borwein 初始化的线搜索方法，用于自适应调整步长。请注意，对于我们的模型，由于其中包含对数行列式项，计算值 $F(\boldsymbol{x})$ 的代价非常高，其复杂度为 $O(K^3)$。线搜索方法可能需要在单次迭代中多次计算函数值，因此这里不采用这种方法。

停止准则。如果我们有一个良好的先验 W_0，那么 W 与 W_0 的差异就不会太大，我们的算法也容易收敛。在这种情况下，我们的停止准则是$\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\| < \epsilon_{tol}$。

然而，当我们的知识有限或 W_0 与最优的 W 存在显著差异时，继续训练直至收敛既耗时又没有必要。因此，我们采用提前停止的思路。为了获得更好的泛化性能，我们在每次迭代时都使用验证集进行测试，并在验证分数连续若干次迭代未提升时停止训练。

请注意，对于用于拟合 [TBB19], 中 LRSM 的基于 ADMM 的算法，他们采用了基于 KKT 条件的停止准则，这与我们的方法略有不同。为确保公平，在第 6 节的所有实验中，我们都让他们的算法运行足够多次迭代，以确保算法已经收敛且性能不再提升。

5.2 收敛性分析

由于拉普拉斯正则化项并非全局强光滑，而全局强光滑是 [LL15], 中原始证明的关键假设，因此我们需要新的技术来确保迭代序列有界。具体而言，我们作出以下假设。

假设 5.1。对于任意 $k \leq K$，l_k 在任意有界集合上都是适当且强光滑的。

假设 5.2。r 是适当的、下半连续的、强制的且有下界的（不要求凸）。

假设 5.3。对于任意 $k \leq K$，$l_k(\theta) + r(\theta)$ 是强制的，即 $l_k(\theta) + r(\theta) \to +\infty$ 当 $\|\theta\| \to \infty$。

假设 5.1 可确保光滑项 f 在任意有界集合上局部强光滑。与文献中的大多数假设相比，该假设通常更弱，因为我们不要求全局强光滑。假设 5.3 仅出于技术原因提出，在 [LL15] 中也有要求。l_k 的典型例子包括回归和分类模型，如 [TBB19] 中所述。r 的例子包括有界约束、平方和、带有 $p > 0$ 的 ℓ_p、弹性网络、OSCAR，以及在无穷远处被违反的等式或不等式约束（例如 斯蒂费尔流形、球面、多面体）。

定理 2. 在假设 *5.1-5.3* 下，算法 *1* 中存在 $\alpha_x, \alpha_y > 0$，使得

(i) 序列 $\{\boldsymbol{x}_k\}$ 有界。

(ii) $\{\boldsymbol{x}_k\}$ 的极限点集合非空。如果 $\boldsymbol{x}^*$ 是 $\{\boldsymbol{x}_k\}$的极限点，那么 $\boldsymbol{x}^*$ 是 F的临界点。

(iii) 如果 $F(\boldsymbol{x})$ 满足 *KL* 性质，且去奇异化函数具有如下形式： $\phi(t) = \overset{C}{_s}t^s$，其中 $C \succ 0$, $s \in (0, 1],$，那么 $\{F(\boldsymbol{x}_k)\}$ 将收敛到某个 $F(\boldsymbol{x}^*)$，其收敛速率为线性或次线性（取决于 s）。

证明思路。我们首先用归纳法证明序列 $\{\boldsymbol{x}_k\}$ 和 $\{\boldsymbol{v}_k\}$ 有界。只要步长足够小，我们算法的函数值就具有单调性，因此这一点可以保证。然后，根据假设 5.1，迭代序列满足强光滑性。其余部分与 [LL15]相同。完整证明见附录 B。 □

6 数值结果

我们以 LRSM 为基线，展示联合图学习方法的优势。本节分为两部分：一部分展示相较于良好先验 W_0 的改进，另一部分则说明在没有先验（即，$W_0 = 0$）时的灵活性。我们还将共同模型（不依赖于 z 的模型）和分离模型（$W_{ij} \equiv 0$）作为两个极端情况。实现所提方法以及基线方法、用于复现以下实验的代码可在线获取：
https://github.com/cvxgrp/joint-lrsm。

6.1 带先验的分层模型

在本小节中，我们考虑 [TBB19] 中的两个示例，并将其中人工设计的图作为先验 W_0。我们的目标是通过联合拉普拉斯分层模型，在给定先验信息 W_0 的情况下学习更好的 W。结果表明，我们的方法能够显著提升性能。

6.1.1 间皮瘤分类

我们研究一个预测患者是否患有间皮瘤（一种癌症）的问题，依据是患者的性别、年龄及其他医疗特征；这些特征是在一系列患者就诊和实验室检查期间收集的。对于每个类别，我们使用逻辑回归，根据特征对患间皮瘤的条件概率进行建模。数据总计有 324 个样本，其中 10% 被划分为测试集。

数据预处理、模型设置和基线算法均与 [TBB19], 相同，不同之处在于，我们通过交叉验证重新调整局部正则化权重，以提升标准 LRSM 的性能（使其成为更强的基线）。此外，在联合模型中，我们将先验 W_0 设为 [TBB19]中原始正则化图的边权重矩阵。进一步地，我们通过 5 折交叉验证选择 $\lambda_1 = 100,\ \lambda_2 = 12,\ \eta = 0.1,\ \mu = 20.0$。

结果。表 1 展示了两个模型的测试预测误差、平均负对数似然、AUC 和 F1 分数。可以看到，联合分层模型始终优于标准模型。图 4 展示了权重差异的热图 $W - W_0$。请注意，W_0 的边权重为 10 或 500，因此差异的幅度相对较小。这进一步表明，拉普拉斯正则化的性能对

（高维）权重十分敏感，因此难以手动调优。因此，在这种情况下，我们的联合学习方法可以作为一种便捷的调优方法。

	误差（%）	ANLL	AUC	F1 分数
共同	33.3	0.728	0.529	0.593
分离	39.4	0.695	0.402	0.439
LRSM	24.2	0.532	0.582	0.592
联合拉普拉斯分层	21。221。221。2	0.4790.4790.479	0.5830.5830.583	0.6180.6180.618

表 1：间皮瘤结果。

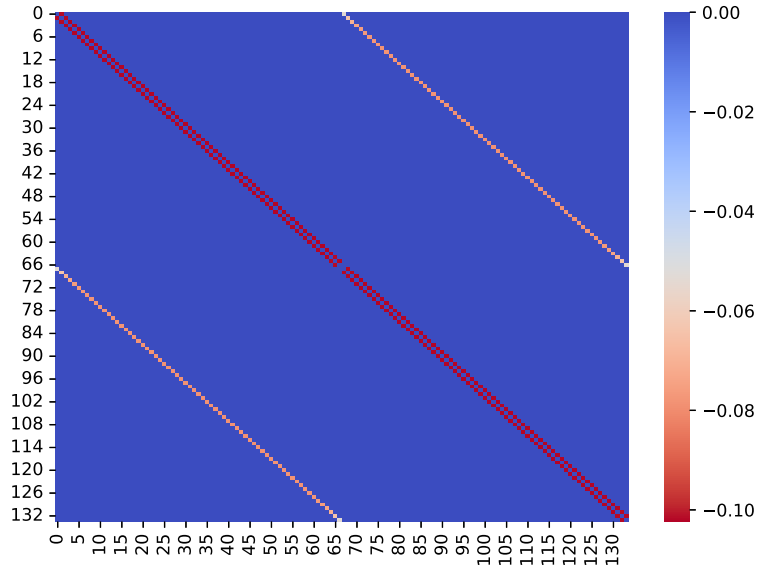

图 4：权重差异热力图 $W - W_0$

6.1.2 参议院选举

我们对某个州在某个选举年份举行的美国参议院选举由民主党获胜的概率进行建模。对于每个州和选举年份，我们都有一个伯努利参数，可将其解释为该州选出民主党候选人的概率。我们的损失函数是负对数似然。训练记录有 639 条，测试记录有 68 条。

模型设置和基线算法与 [TBB19] 相同。此外，我们将联合模型中的先验 W_0 设为 [TBB19] 中原始正则化图的边权重矩阵。我们通过执行 5 折交叉验证来设置 $\lambda_1 = 0.2$, $\lambda_2 = 2.0, \eta = 1 \times 10^{-3}, \mu = 0.1$ 。

结果。 表 2 展示了两个模型的测试准确率、平均负对数似然、AUC 和 F1 分数。可以看到，联合拉普拉斯分层模型的表现显著优于标准模型。图 5 和图 6 展示了两个模型的伯努利参数。较高的模型

	准确率（%）	ANLL	AUC	F1 分数
共同	35.3	0.701	0.500	0.522
分离	35.3	0.998	0.500	0.522
LRSM	72.1	0.608	0.756	0.689
联合拉普拉斯分层	80.980.980.9	0。5460。5460。546	0.8140.8140.814	0.7550.7550.755

表 2： 选举预测结果。

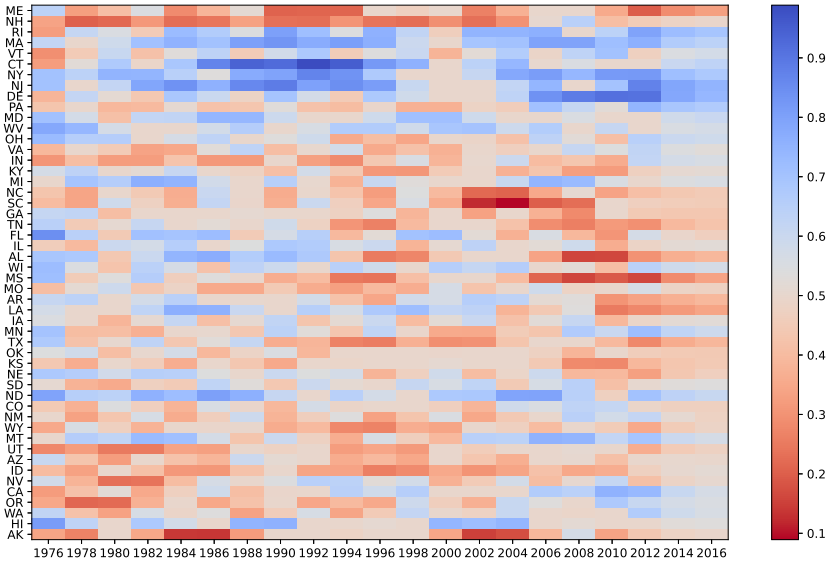

图 5： LRSM 中不同选举年份和州的伯努利参数 .

参数对应蓝色（民主党），较低的模型参数对应红色（共和党及其他政党）。请注意，在联合分层模型中，颜色通常更深，表明参数更接近 0 或 1。

6.2 无先验的分层模型

在本小节中，我们将联合拉普拉斯分层模型中的 W_0 设为零矩阵；对于标准 LRSM，我们则按照 [TBB19] 中采用的类似模式手动设计图，并通过交叉验证调整边权重。由于 W_0 为零，我们没有任何先验信息，权重矩阵 W 会在训练过程中自动学习。我们还考虑了之前提到的联合学习基线方法，即 Log-Diagonal 和 Tr-Constraint。我们的目标是考察本方法能否与标准方法相媲美，甚至表现得更好。

由于我们对MAPG算法的理论分析也可以轻松推广到上述两种设置，因此我们将其应用于这两个基线方法。我们还考虑了使用交替方法来求解优化问题，这与文献中的做法一致。所示的Log-Diagonal和Tr-Constraint结果是两种优化算法中的最佳结果。

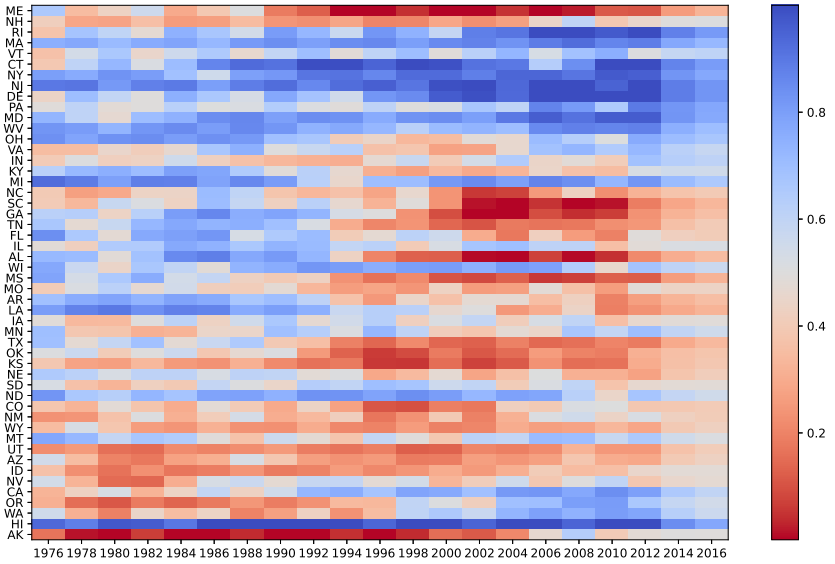

图 6： 联合拉普拉斯分层模型中，不同选举年份和州对应的伯努利参数。

6.2.1 葡萄酒质量分类

我们的目标是根据红葡萄酒的理化特征和感官特征预测其红葡萄酒品质。

数据集。 我们获取了一个描述葡萄牙“绿酒”红葡萄酒品质的数据集。该数据集包含 11 个特征，包括固定酸度、挥发性酸度、柠檬酸、残糖、氯化物、游离二氧化硫、总二氧化硫、密度、pH值、硫酸盐和酒精。输出是一个表示葡萄酒质量的二元变量。

数据记录。 该数据集包含 1599 个实例。分层特征为密度和硫酸盐。由于酒精与质量高度相关，我们移除了这一特征。因此，特征向量为 $x \in \mathbf{R}^9$，并包含截距。我们将数据集的 20% 作为测试集，其余部分作为训练集。我们将训练数据随机分为五折，并对特征进行标准化，使其具有零均值和单位方差。

数据模型。 我们的模型采用逻辑回归和局部 ℓ_1-正则化，权重为 γ_{local}。

标准 LRSM 的正则化图。 我们取两个正则化图的笛卡尔积：

- 密度。我们将密度分为 10 个大小相等的分箱。正则化图是连接密度分箱的路径图，边权重为 γ_{den}。

- 硫酸盐。我们将硫酸盐特征分为 10 个大小相等的分箱。正则化图是连接各分箱的路径图，边权重为 γ_{sul}。

因此，我们有 $K = 10 \times 10 = 100$ 个分层模型。请注意，上述正则化图结构仅用于标准拉普拉斯分层模型；对于我们的方法，除图由这 100 个分层节点组成外，我们不对图作任何假设。

结果。 在 LRSM 中，我们通过 5 折交叉验证设置 $\gamma_{den} = 20$，$\gamma_{sul} = 10$。在联合拉普拉斯分层模型中，我们通过 5 折交叉验证设置 $\lambda_1 = 5.0$，$\lambda_2 = 2 \times 10^{-3}$，$\eta = 0.1$，$\mu = 0.2$。局部正则化权重为 $\gamma_{local} = 0.01$。

	误差（%）	ANLL	AUC	F1 分数
共同	31.2	0.585	0.761	0.679
分离	28.1	0.827	0.689	0.715
LRSM	27.8	0.5370.5370.537	0.8020.8020.802	0.717
对数-对角分层	28.4	0.621	0.625	0.712
Tr 约束分层	27.5	0.781	0.692	0.721
联合拉普拉斯分层	26。626。626。6	0.549	0.789	0.7290.7290.729

表3： 葡萄酒结果。

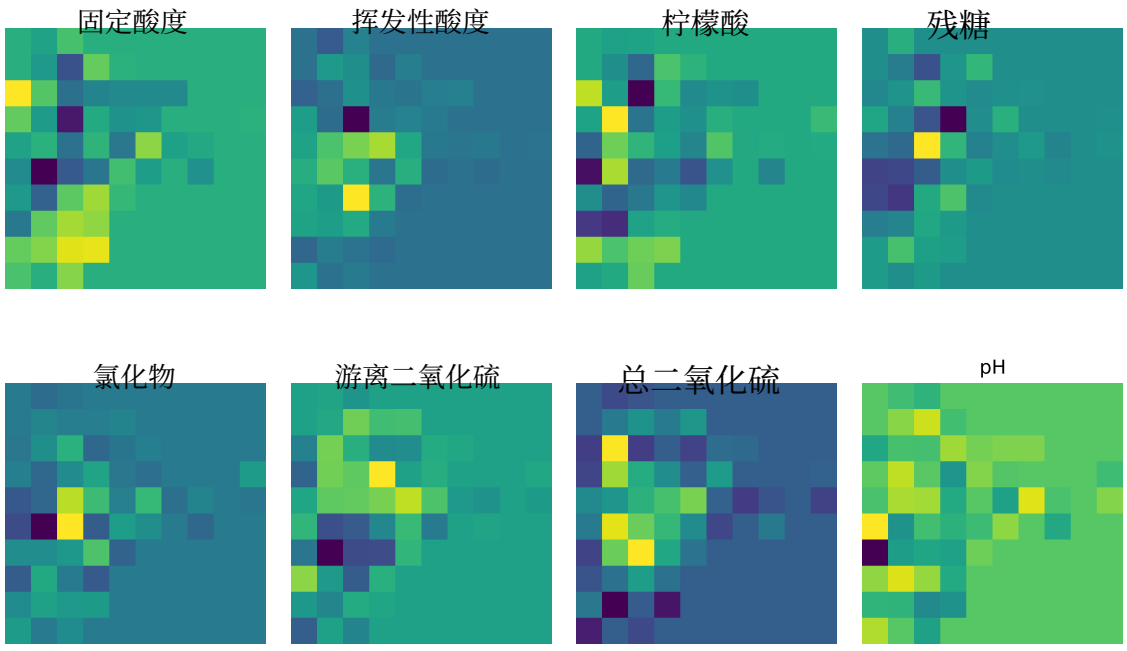

图7： 联合拉普拉斯分层模型中的葡萄酒质量系数。

表3展示了测试集上的四项评估指标，结果表明，联合拉普拉斯分层模型的表现与标准模型相当。我们的联合优化方法

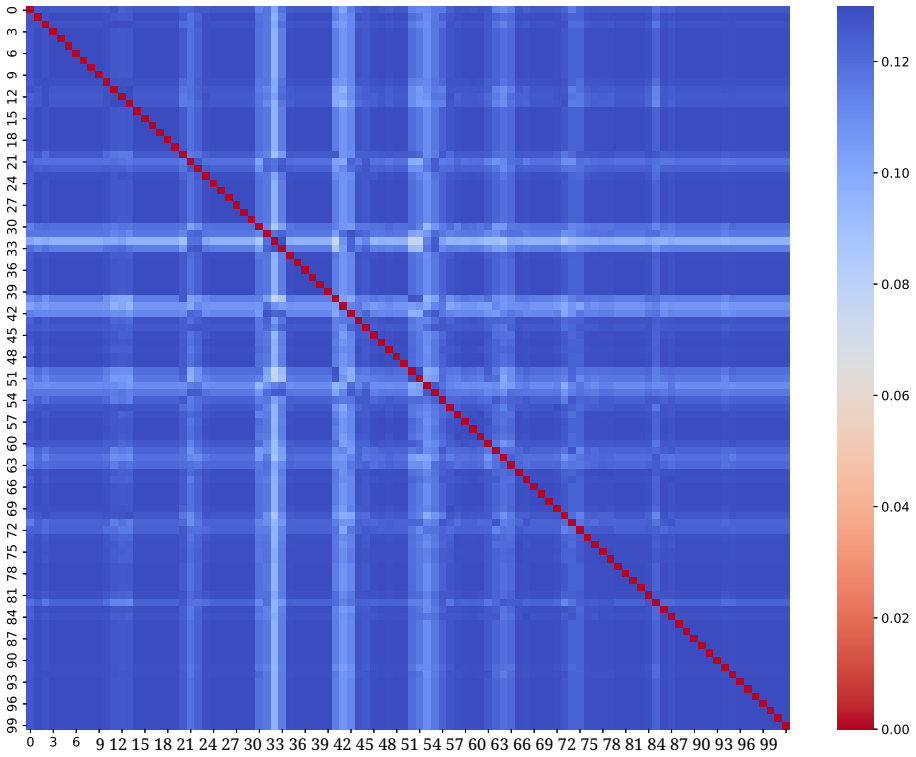

图8： 通过联合拉普拉斯分层模型学习的葡萄酒质量边权重矩阵。

即使没有先验知识，也能准确调整边权重。图7展示了联合拉普拉斯分层模型中每个分箱对应的8个系数。图8可视化了我们的方法学习到的边权重。

6.2.2 混凝土强度预测

我们对给定年龄和粉煤灰、水泥等原料的混凝土抗压强度条件分位数进行建模。

数据集。 我们获取了描述混凝土抗压强度的数据。该混凝土由7种组分构成，包括水泥、高炉矿渣、粉煤灰、水、高效减水剂、粗骨料和细骨料。此外，该数据集还包含混凝土年龄的信息。

数据记录。 该数据集中共有 1030 个实例。分层特征为年龄和粉煤灰组分。因此，我们得到特征向量 $x \in \mathbf{R}^7$，其中包含截距项。我们将数据集的 25% 作为测试集，其余部分作为训练集。我们将训练数据随机划分为五折，并对特征进行标准化处理，使其具有零均值和单位方差。

数据模型。 我们的模型是统计学中的分位数回归 [KBJ78]。不同于普通最小二乘法中的条件均值，分位数回归的目标是给定 $\boldsymbol{X}$ 时 y 的 τ 阶条件分位数，其定义为 $Q_{Y|\boldsymbol{X}}(\tau) := \inf\{y : \mathbf{F}^{Y|\boldsymbol{X}}(y) \geq \tau\}$，其中 $\mathbf{F}^{Y|\boldsymbol{X}}$ 是条件 CDF。

局部损失函数由弹球损失给出：

$$l_k(\theta) = \sum_{i=1}^{n_k} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}),$$

其中，$\rho_\tau(z) = (1-\tau)\max\{-z, 0\} + \tau \max\{z, 0\}$。对于联合分层模型，由于损失函数在零点处不可微，因此我们改用 Huber 损失来计算梯度。对于 LRSM，通过调用 CVXPY 来估计 l_k 的近端算子。此外，我们还使用正则化权重为 γ_{local} 的平方和局部正则化函数。

标准 LRSM 的正则化图。 我们取两个正则化图的笛卡尔积：

- 年龄。我们将年龄分成 10 个大小相等的分箱。正则化图是年龄分箱之间的路径图，边权重为 γ_{age}。
- 粉煤灰。我们将粉煤灰特征分成 10 个大小相等的分箱。正则化图是分箱之间的路径图，边权重为 γ_{ash}。

因此，我们有 $K = 10 \times 10 = 100$ 个分层模型。再次注意，对于我们的联合拉普拉斯分层模型，除了相同的 100 个分层节点外，我们不对图作任何假设。

结果。 我们的目标是找到第 $\tau = 0.9$ 个分位数。在 LRSM 中，我们设置 $\gamma_{age} = \gamma_{ash} = 0.5$。在联合拉普拉斯分层模型中，我们设置 $\lambda_1 = 5.0$，$\lambda_2 = 10.0$，$\eta = 1 \times 10^{-3}$，$\mu = 2 \times 10^{-3}$。局部正则化权重为 $\gamma_{local} = 0.01$。

	训练损失	测试损失
共同	2.06	2.27
分离	1.231.231.23	2.52
LRSM	1.56	1.74
对数-对角分层	1.59	1.99
Tr 约束分层	1.37	1.85
联合拉普拉斯分层	1.49	1.581.581.58

表4： 混凝土结果。

表4展示了训练集和测试集的弹球损失结果。联合拉普拉斯分层模型的表现显著优于标准模型。值得注意的是，我们的联合优化方法无需先验知识即可高效调整边权重。图9展示了联合拉普拉斯分层模型中水泥和水参数的条形图。需要注意的是，图中的参数并不十分平滑。仅对节点度进行正则化不足以控制图连接结构。因此，Log-Diagonal 和 Tr-Constraint 的表现不如我们的方法。图10

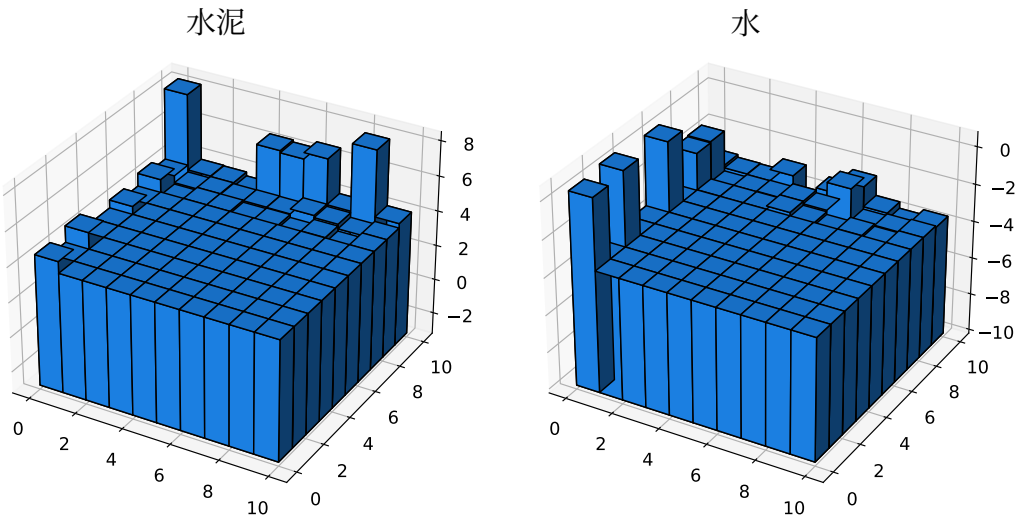

图9： 各子模型中的水泥和水参数。

展示了联合拉普拉斯分层模型学习得到的边权重矩阵的可视化结果，而这种结果很难通过手动调整获得。

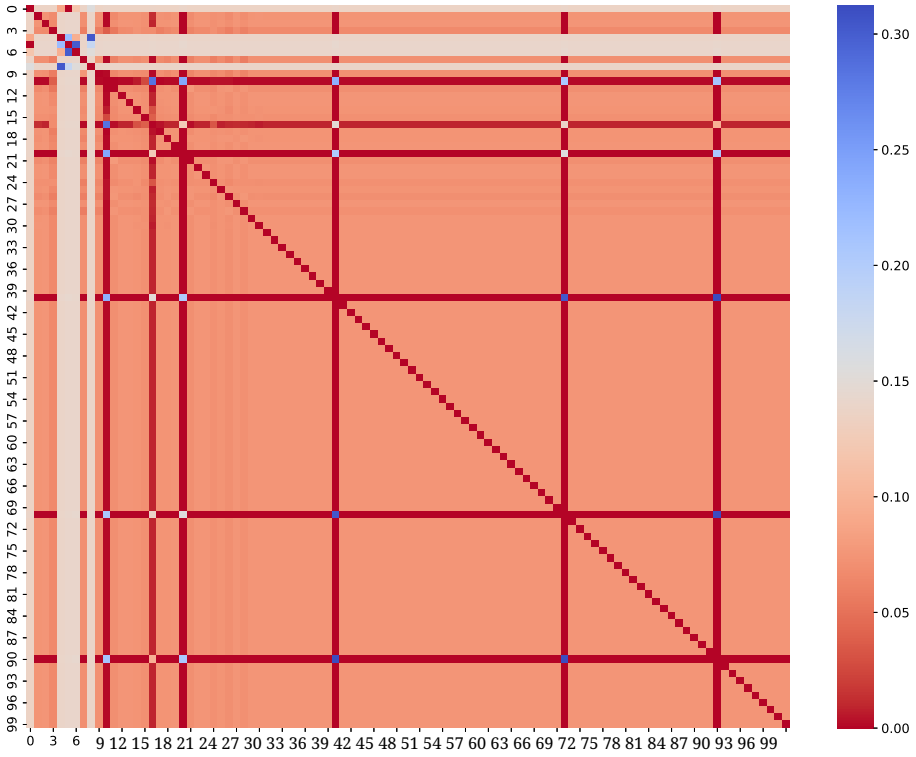

图10： 联合拉普拉斯分层模型学习到的混凝土边权重。

致谢

作者谨向 Jonathan Tuck 和 Anran Hu 致以诚挚谢意，感谢他们为本研究作出的宝贵贡献。他们富有见地的讨论、反馈与支持对本项目的成功至关重要。

参考文献

[AAB+19] Akshay Agrawal、Brandon Amos、Shane Barratt、Stephen Boyd、Steven Diamond 和 Zico Kolter。可微凸优化层。*arXiv* 预印本 *arXiv:1910.12430*，2019 年。[AZ06] Rie Ando 和 Tong Zhang。基于拉普拉斯正则化的图学习。神经信息处理系统进展，19，2006 年。[BB20] Shane T Barratt 和 Stephen P Boyd。最小二乘自动调优。《工程优化》，第 1–22 页，2020 年。[BST14] Jérôme Bolte、Shoham Sabach 和 Marc Teboulle。用于非凸和非光滑问题的近端交替线性化最小化。《数学规划》，146(1):459–494，2014 年。[CK78] Seth Chaiken 和 Daniel J Kleitman。矩阵树定理。《组合理论杂志》，*A*辑，24(3):377–381，1978 年。[CPVBR20] Vivien A. Cabannes、Loucas Pillaud-Vivien、Francis R. Bach 和 Alessandro Rudi。通过半监督学习中的拉普拉斯正则化克服维数灾难。载于神经信息处理系统，2020 年。[CS06a] Pavel Chebotarev 和 Elena Shamis。矩阵森林定理及小型社会群体中关系的度量。*arXiv* 预印本 *math/0602070*，2006 年。[CS06b] Pavel Chebotarev 和 Elena Shamis。矩阵森林定理。*arXiv* 预印本 *math/0602575*，2006 年。[DTFV14] Xiaowen Dong、Dorina Thanou、Pascal Frossard 和 Pierre Vandergheynst。平滑图信号表示中的拉普拉斯矩阵学习。《*IEEE*信号处理汇刊》，64:6160–6173，2014 年。[DVT+21] Canh T. Dinh、Thanh Tung Vu、Nguyen H. Tran、Minh N. Dao 和 Hongyu Zhang。重新审视拉普拉斯正则化联邦多任务学习及其收敛速率。《*IEEE*神经网络与学习系统汇刊》，2021 年。[DW22] Yaqi Duan 和 Kaizheng Wang。自适应鲁棒多任务学习。*ArXiv*，abs/2202.05250，2022 年。

[EP04] Theodoros Evgeniou 和 Massimiliano Pontil。正则化多任务学习。第十届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，2004 年。[GVZB16] André R Gonçalves、Fernando J Von Zuben 和 Arindam Banerjee。基于高斯Copula模型的多任务稀疏结构学习。机器学习研究杂志，17(1):1205–1234，2016 年。[HTK22] Jiayu He、Nguyen H. Tran 和 Matloob Khushi。基于图拉普拉斯多任务学习的股票预测器。载于概念结构国际会议，2022 年。[JVB08] Laurent Jacob、Jean-philippe Vert 和 Francis Bach。聚类多任务学习：凸公式化。神经信息处理系统进展，21，2008 年。[KA22] Abdullah Karaaslanli 和 Selin Aviyente。图信号同步聚类与图学习。载于国际机器学习大会，2022 年。[Kal16] Vassilis Kalofolias。如何从平滑信号中学习图。载于人工智能与统计国际会议，2016 年。[KBJ78] Roger Koenker 和 Gilbert Bassett Jr.。回归分位数。计量经济学：计量经济学会杂志，第 33–50 页，1978 年。[KLZX21] Zhao Kang、Zhiping Lin、Xiaofeng Zhu 和 Wenbo Xu。从单视图到多视图：面向可扩展子空间聚类的结构化图学习。*IEEE*控制论汇刊，52:8976–8986，2021 年。[KP17] Vassilis Kalofolias 和 Nathanael Perraudin。从平滑信号中进行大规模图学习。*ArXiv*，abs/1710.05654，2017 年。[KYdMCP19]Sandeep Kumar、Jiaxi Ying、José Vinícius de Miranda Cardoso 和 Daniel Pérez Palomar。通过谱约束进行结构化图学习的统一框架。《机器学习研究杂志》，21:22:1–22:60，2019 年。[LL15] Huan Li 和 Zhouchen Lin。非凸规划的加速近端梯度方法。载于 C. Cortes、N. Lawrence、D. Lee、M. Sugiyama 和 R. Garnett（编），神经信息处理系统进展，第 28 卷。Curran Associates, Inc.，2015 年。[LWWZ22] Henry Lam、Kaizheng Wang、Yuhang Wu 和 Yichen Zhang。用于多任务非光滑优化的自适应数据融合。*ArXiv*，abs/2210.12334，2022 年。[LZZ+16] Xi Li、Xueyi Zhao、Zhongfei Zhang、Fei Wu、Yueting Zhuang、Jingdong Wang 和 Xuelong Li。结合社区感知标签图学习的联合多标签分类。*IEEE*图像处理汇刊，25:484–493，2016 年。

[PB+14] Neal Parikh、Stephen Boyd 等。近端算法。《优化基础与趋势》，1(3):127–239，2014。[PCDS20] Xingyue Pu、Siu Lun Chau、Xiaowen Dong 和 D. Sej dinovic。基于平滑信号的核图学习：函数视角。《IEEE 网络信号与信息处理汇刊》，7:192–207，2020。[Pen17] Richard Peng。CS7540 第2讲，矩阵树定理。https://www.cc. gatech.edu/~rpeng/CS7540_S17/Jan12MatrixTree.pdf，2017。[RH05] Håvard Rue 和 Leonhard Held。高斯马尔可夫随机场：理论与应用。2005。[SCST17] Virginia Smith、Chao-Kai Chiang、Maziar Sanjabi 和 Ameet S. Talwalkar。联邦多任务学习。发表于 NIPS，2017。[She08] Daniel Sheldon。图形多任务学习。2008。[ST17] Dejan Slepčev 和 Matthew Thorpe。半监督学习中 p-拉普拉斯正则化的分析。ArXiv，abs/1707.06213，2017。[TB20a] Jonathan Tuck 和 Stephen Boyd。特征分层模型。arXiv 预印本 arXiv:2001.10389，2020。[TB20b] Jonathan Tuck 和 Stephen Boyd。拟合拉普拉斯正则化分层高斯模型。arXiv 预印本 arXiv:2005.01752，2020。[TBB19] Jonathan Tuck、Shane Barratt 和 Stephen Boyd。用于拟合拉普拉斯正则化分层模型的分布式方法。arXiv 预印本 arXiv:1904.12017，2019。[TBB21] Jonathan Tuck、Shane Barratt 和 Stephen Boyd。使用分层模型构建投资组合。arXiv 预印本 arXiv:2101.04113，2021。[VVBD20] Mariana Vargas-Vieyra、Aurélien Bellet 和 Pascal Denis。联合学习图与数据表示以实现基于图的半监督学习。发表于《自然语言处理的图方法》（TextGraphs）论文集，2020。[ZCY11] Jiayu Zhou、Jianhui Chen 和 Jieping Ye。通过交替结构优化实现聚类多任务学习。《神经信息处理系统进展》，24，2011。[ZDGA20] Imtiaz Ziko、Jose Dolz、Eric Granger 和 Ismail Ben Ayed。拉普拉斯正则化少样本学习。发表于《国际机器学习大会》，第11660–11670页。PMLR，2020。

[ZLWN21] Haifeng Zhao、Qi Li、Z. Wang 和 Feiping Nie。面向无监督特征选择的联合自适应图学习与判别分析。认知计算，14:1211–1221，2021。

[ZY10] Yu Zhang 和 D. Y. Yeung。《多任务学习中任务关系学习的凸公式化》。载于人工智能中的不确定性会议，2010 年。

A 定理 1 的证明

证明。令 $S=\sum_{j\neq k_0}W_{jk_0}$。由于层 k_0 没有数据样本且局部正则化为零，因此（1.1）表明

$$\theta_{k_0}=\frac{\sum_{j\neq k_0}W_{jk_0}\theta_j}{S},\ \widetilde{\theta}_{k_0}=\frac{\sum_{j\neq k_0,i}W_{jk_0}\widetilde{\theta}_j+\widetilde{W}_{ik_0}\widetilde{\theta}_i}{S+\epsilon}. \tag{A.1}$$

~注意，根据假设，Θ,Θ 均为 δ-稳健。将上面两个方程相减并重新整理，

$$\begin{aligned}
||\frac{W_{ik_0}}{S}\theta_i-\frac{W_{ik_0}+\epsilon}{S+\epsilon}\widetilde{\theta}_i|| &\leq ||\theta_{k_0}-\widetilde{\theta}_{k_0}||+\sum_{j\neq k_0,i}||\frac{W_{jk_0}\theta_j}{S}-\frac{W_{jk_0}\widetilde{\theta}_j}{S+\epsilon}|| \\
&\leq 2\delta'||\theta^*_{k_0}||+\sum_{j\neq k_0,i}\left[\frac{W_{jk_0}}{S}||\theta_j-\theta^*_j||+\frac{W_{jk_0}}{S+\epsilon}||\widetilde{\theta}_j-\theta^*_j||+\frac{\epsilon W_{jk_0}}{S(S+\epsilon)}||\theta^*_j||\right] \\
&\leq 2\delta'||\theta^*_{k_0}||+\sum_{j\neq k_0,i}\left(\frac{\delta W_{jk_0}}{S}+\frac{\delta W_{jk_0}}{S+\epsilon}+\frac{\epsilon W_{jk_0}}{S(S+\epsilon)}\right)||\theta^*_j|| \\
&\leq \max_{j\neq i}||\theta^*_j||\left(2\delta'+2\delta+\frac{\epsilon}{S+\epsilon}(1-\frac{W_{ik_0}}{S})\right).
\end{aligned}$$

此外，我们有

$$\begin{aligned}
||\frac{W_{ik_0}}{S}\theta_i-\frac{W_{ik_0}+\epsilon}{S+\epsilon}\widetilde{\theta}_i|| &\geq (-\frac{W_{ik_0}}{S}+\frac{W_{ik_0}+\epsilon}{S+\epsilon})||\theta^*_i||-\frac{W_{ik_0}}{S}||\theta_i-\theta^*_i||-\frac{W_{ik_0}+\epsilon}{S+\epsilon}||\widetilde{\theta}_i-\theta^*_i|| \\
&\geq \left(\frac{\epsilon}{S+\epsilon}(1-\frac{W_{ik_0}}{S})-\frac{W_{ik_0}}{S}\delta-\frac{W_{ik_0}+\epsilon}{S+\epsilon}\delta\right)||\theta^*_i|| \\
&\geq \left(\frac{\epsilon}{S+\epsilon}(1-\frac{W_{ik_0}}{S}-\delta)-2\frac{W_{ik_0}}{S}\delta\right)||\theta^*_i||.
\end{aligned}$$

结合上述两个不等式，并令 $A_i:=\frac{||\theta^*_i||}{\text{最大值}_{j\neq i}||\theta^*_j||}$，可得（前提是分母为正），

$$\frac{\epsilon}{W_{ik_0}}\leq\frac{2(\delta+\delta'+\frac{\delta W_{ik_0}}{S}A_i)}{\left(1-\frac{(1+2\delta)W_{ik_0}}{S}-\delta\right)A_i-1+\frac{W_{ik_0}}{S}-2(\delta+\delta')}\frac{S}{W_{ik_0}}.$$

□

B 定理 2 的证明

我们先陈述证明定理 2 时用到的一个重要引理。

引理 B.1 ([BST14])。设Ω 是一个紧集，且 $F:\mathbf{R}^n\to(-\infty,+\infty]$ 是一个真且下半连续的函数。假设 F 在Ω 上为常数，并且在Ω 中的每一点都满足 ***KL***

性质。那么存在 $\epsilon > 0$、$\eta > 0$ 和 $\varphi \in \Phi_\eta$，使得对于$\mathbf{u} \in \Omega$ 以及以下交集中的所有$\mathbf{u}$

$$\{\mathbf{u} \in \mathbf{R}^n : \operatorname{dist}(\mathbf{u}, \Omega) < \epsilon\} \bigcap \{\mathbf{u} \in \mathbf{R}^n : f(\overline{\mathbf{u}}) < f(\mathbf{u}) < f(\overline{\mathbf{u}}) + \eta\},$$

以下不等式成立

$$\varphi'(f(\mathbf{u}) - f(\overline{\mathbf{u}})) \operatorname{dist}(0, \partial f(\mathbf{u})) > 1.$$

下面我们针对一般的 f, g 重新陈述定理 2，并给出完整的证明。注意，假设 (5.1)–(5.3) 可以确保满足以下假设，因此定理 2 是定理 B.2 的一个特殊情形。

定理 B.2. 考虑一般问题

$$\min_{\boldsymbol{x} \in \mathbf{R}^n} F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{B.1}$$

满足以下假设

A.1 f（$\boldsymbol{x}$）在任何有界集合上都是适当且强光滑的。

A.2 $g(\boldsymbol{x})$ 是适当的、下半连续的，且下界为 $g > -\infty$。

A.3 $F(\boldsymbol{x})$ 具有强制性，即，当 $F(\boldsymbol{x}) \to +\infty$ 时，$\|\boldsymbol{x}\| \to +\infty$。

那么，存在 α_x、$\alpha_y > 0$，使其满足算法 *1*，即

(i) 序列 $\{\boldsymbol{x}_k\}$ 和 $\{\boldsymbol{v}_k\}$ 有界，并且极限$\lim_{k \to \infty} \|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\| = 0$。

(ii) $\{\boldsymbol{x}_k\}$ 的极限点集合非空。若 $\boldsymbol{x}^*$ 是 $\{\boldsymbol{x}_k\}$的一个极限点，则 $\boldsymbol{x}^*$ 是 F的一个临界点。

(iii) 若 $F(\boldsymbol{x})$ 满足 *KL* 性质，且去奇异化函数具有 $\varphi(t) = \frac{C}{s} t^s$ 的形式，其中 $C > 0$、$s \in (0, 1]$，则

1. 若 $s = 1$，则存在 k_1，使得对所有 $k > k_1$ 均有 $F(\boldsymbol{x}_k) = F^*$，且算法在有限步内终止。

2. 若 $s \in \left[\frac{1}{2}, 1\right)$，则存在 k_2，使得对所有 $k > k_2$，

$$F(\boldsymbol{x}_k) - F^* \leq \left(\frac{d_1 C^2}{1 + d_1 C^2}\right)^{k - k_2} r_{k_2}.$$

3. 若 $s \in \left(0, \frac{1}{2}\right)$，则存在 k_3，使得对所有 $k > k_3$，

$$F(\boldsymbol{x}_k) - F^* \leq \left(\frac{C}{(k - k_3)\, d_2 (1 - 2s)}\right)^{\frac{1}{1-2s}}$$

其中 F^* 在 $\{\boldsymbol{x}_k\}$、$r_k =$的所有聚点处具有相同的函数值
$F(\boldsymbol{v}_k) - F^*$, $d_1 = \left(\frac{1}{\alpha_x} + L\right)^2 / \left(\frac{1}{2\alpha_x} - \frac{L}{2}\right)$, $d_2 = \min\left\{\frac{1}{2 d_1 C}, \frac{C}{1-2s}\left(2^{\frac{2s-1}{2s-2}} - 1\right) r_0^{2s-1}\right\}$.

证明。令次水平集 $B_0 := \{\boldsymbol{x}: F(\boldsymbol{x}) \le F(\boldsymbol{x}_0)\}$，根据强制性假设，该集合是有界集。因此，存在 G_f, $G_g < +\infty$，使得对于任意 $\boldsymbol{x} \in B_0$ 和 $g(\boldsymbol{x}) - g \le G_g$，均有 $\|\nabla f(\boldsymbol{x})\| \le G_f$，且对于任意 $\boldsymbol{x} \in B_0 \cap$**定义域** g。令 $\alpha_0 > 0$ 为任意正数常量，并令 $B_1 :=$**凸包**$\{\boldsymbol{x}: \mathrm{dist}(\boldsymbol{x}, B_0) \le \sqrt{2\alpha_0 G_g + 2\alpha_0 G_f}\}$。那么 B_1 也是一个有界集，因此 f 在 B_1 上具有强光滑系数 L。我们证明 $\alpha_x = \alpha_y < \min\{\alpha_0, \frac{1}{L}\}$ 已足够。

首先，我们通过归纳法证明，对于 $\delta = \frac{1}{2\alpha_x} - \frac{L}{2} > 0$, $F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k-1}) \le -\delta\|\boldsymbol{v}_k - \boldsymbol{x}_{k-1}\|^2$，因此对于所有 k, $\boldsymbol{x}_k$ 属于 B_0。

$k = 0$ 的情形显而易见。

假设 $\boldsymbol{x}_k$ 属于 B_0。那么根据 $\boldsymbol{v}_{k+1}$ 的构造，我们有

$$0 \in \partial g(\boldsymbol{v}_{k+1}) + \frac{\boldsymbol{v}_{k+1} - (\boldsymbol{x}_k - \alpha_x \nabla f(\boldsymbol{x}_k))}{\alpha_x}, \tag{B.2}$$

其中，∂g 表示极限 g次微分。

这意味着

$$g(\boldsymbol{v}_{k+1}) + \frac{1}{2\alpha_x}\|\boldsymbol{v}_{k+1} - (\boldsymbol{x}_k - \alpha_x \nabla f(\boldsymbol{x}_k))\|^2 \le g(\boldsymbol{x}_k) + \frac{\alpha_x}{2}\|\nabla f(\boldsymbol{x}_k)\|^2,$$

$$\begin{aligned} \|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\| &\le \sqrt{2\alpha_x(g(\boldsymbol{x}_k) - \underline{g}) + \alpha_x^2\|\nabla f(\boldsymbol{x}_k)\|^2} + \alpha_x\|\nabla f(\boldsymbol{x}_k)\| \\ &\le \sqrt{2\alpha_0 G_g} + 2\alpha_0 G_f. \end{aligned} \tag{B.3}$$

因此，$\boldsymbol{v}_{k+1}$ 属于 B_1。对于凸集 B_1 上的利普希茨可微 f，我们有

$$\begin{aligned} F(\boldsymbol{v}_{k+1}) &\le g(\boldsymbol{v}_{k+1}) + f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_{k+1} - \boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\|^2 \\ &\le F(\boldsymbol{x}_k) - (\frac{1}{2\alpha_x} - \frac{L}{2})\|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\|^2. \end{aligned} \tag{B.4}$$

因此

$$F(\boldsymbol{x}_{k+1}) \le F(\boldsymbol{v}_{k+1}) \le F(\boldsymbol{x}_k) - \delta\|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\|^2. \tag{B.5}$$

根据归纳法假设，我们有 $\boldsymbol{x}_{k+1}$, $\boldsymbol{v}_{k+1}$ 属于 B_0。

以下过程与 [LL15] 中的相同。

(i) 上述结论证明了 $\{\boldsymbol{x}_k\}$ 和 $\{\boldsymbol{v}_k\}$ 的有界性。将 B.5 从 1 加到k，得到

$$\delta\sum_{i=1}^{k}\|\boldsymbol{v}_{i+1} - \boldsymbol{x}_i\|^2 \le F(\boldsymbol{x}_1) - F(\boldsymbol{x}_{k+1}) \le F(\boldsymbol{x}_1) - F^*, \tag{B.6}$$

因此 $\sum_{i=1}^{\infty}\|\boldsymbol{v}_{i+1} - \boldsymbol{x}_i\|^2 < +\infty$，特别是 $\lim_{k\to\infty}\|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\| = 0$。

(ii) 由于 $\{\boldsymbol{x}_k\}$ 有界，其极限点集必定非空。假设存在一个收敛到 $\boldsymbol{x}^*$ 的子序列 $\{\boldsymbol{x}_{k_j}\}$。那么由 **(i)** 可得 $\|\boldsymbol{v}_{k_j+1} - \boldsymbol{x}^*\| \to 0$。注意，根据 $\boldsymbol{v}_k$ 的构造，

$$g(\boldsymbol{v}_{k_j+1}) + \frac{1}{2\alpha_x}\|\boldsymbol{v}_{k_j+1} - (\boldsymbol{x}_{k_j} - \alpha_x\nabla f(\boldsymbol{x}_{k_j}))\|^2 \le g(\boldsymbol{x}^*) + \frac{1}{2\alpha_x}\|\boldsymbol{x}^* - (\boldsymbol{x}_{k_j} - \alpha_x\nabla f(\boldsymbol{x}_{k_j}))\|^2.$$

So

$$\limsup_{j\to\infty} g(\boldsymbol{v}_{k_j+1}) \le g(\boldsymbol{x}^*).$$

因为 g 是下半连续的，因此上面的不等式实际上是等式，即

$$\lim_{j\to\infty} g(\boldsymbol{v}_{k_j+1}) = g(\boldsymbol{x}^*). \tag{B.7}$$

因此 $F(\boldsymbol{v}_{k_j+1}) \to F(\boldsymbol{x}^*)$，根据 (B.2)，我们有

$$0 \in \partial F(\boldsymbol{x}^*). \tag{B.8}$$

(iii) 由 (B.2)，我们有

$$\operatorname{dist}(0, \partial F(\boldsymbol{v}_{k+1})) \le \|\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{v}_{k+1}) + \frac{1}{\alpha_x}(\boldsymbol{v}_{k+1} - \boldsymbol{x}_k)\| \le (\frac{1}{\alpha_x} + L)\|\boldsymbol{v}_{k+1} - \boldsymbol{x}_k\|. \tag{B.9}$$

若存在 $\bar{k}$ 使得 $F\left(\boldsymbol{v}^{\bar{k}}\right) = F^*$，则 $F\left(\boldsymbol{v}^{\bar{k}}\right) = F\left(\boldsymbol{v}^{\bar{k}+1}\right) = \cdots = F^*$。因此 $\|\boldsymbol{v}^{\bar{k}+1} - \boldsymbol{x}^{\bar{k}}\| = \|\boldsymbol{v}^{\bar{k}+2} - \boldsymbol{x}^{\bar{k}+1}\| = \cdots = 0$。结论成立。若对于所有 k，都有 $F(\boldsymbol{v}_k) > F^*$，则由 $F(\boldsymbol{v}_k) \to F^*$ 可知，存在 k_1 使得当 $k > k_1$ 时，$F(\boldsymbol{v}_k) < F^* + \eta$。另一方面，由于 $\operatorname{dist}(\boldsymbol{v}_k, \Omega) \to 0$，存在 k_2，使得当 $k > \hat{k}_2$ 时，$\operatorname{dist}(\boldsymbol{v}_k, \Omega) < \varepsilon$。令 $k > k_0 = \max\left\{\hat{k}_1, \hat{k}_2\right\}$，我们有

$$\boldsymbol{v}_k \in \{\boldsymbol{v} : \operatorname{dist}(\boldsymbol{v}, \Omega) \le \varepsilon\} \bigcap \{F^* < F(\boldsymbol{v}) < F^* + \eta\}. \tag{B.10}$$

根据引理 B.1 中的一致 Kurdyka–Łojasiewicz 性质，存在一个凹函数 φ，使得

$$\varphi'\left(F\left(\boldsymbol{v}_k\right) - F^*\right) \operatorname{dist}\left(0, \partial F\left(\boldsymbol{v}_k\right)\right) \ge 1. \tag{B.11}$$

定义 $r_k = F(\boldsymbol{v}_k) - F^*$。不失一般性，我们假设对所有充分大的 k，$r_k > 0$。否则，$F(\boldsymbol{v}_k) = F(\boldsymbol{v}_{k+1}) = \cdots = F^*$，且算法会在有限步内终止。由 (B.9)、(B.11) 和 (B.5)，我们有

$$\begin{aligned}
1 &\le \left[\varphi'\left(F\left(\boldsymbol{v}_k\right) - F^*\right) \operatorname{dist}\left(0, \partial F\left(\boldsymbol{v}_k\right)\right)\right]^2 \\
&\le \left[\varphi'\left(r_k\right)\right]^2 \left(\frac{1}{\alpha_x} + L\right)^2 \left\|\boldsymbol{v}_k - \mathbf{x}_{k-1}\right\|^2 \\
&\le \left[\varphi'\left(r_k\right)\right]^2 \left(\frac{1}{\alpha_x} + L\right)^2 \frac{F\left(\boldsymbol{v}_{k-1}\right) - F\left(\boldsymbol{v}_k\right)}{\left(\frac{1}{2\alpha_x} - \frac{L}{2}\right)} \\
&= d_1 \left[\varphi'\left(r_k\right)\right]^2 \left(r_{k-1} - r_k\right),
\end{aligned}$$

对所有 $k > k_0$，其中 $d_1 = \left(\frac{1}{\alpha_x} + L\right)^2 / \left(\frac{1}{2\alpha_x} - \frac{L}{2}\right)$。由于 φ 的形式为 $\varphi(t) = {}^{C}_{s}t^s$，我们有 $\varphi'(t) = Ct^{s-1}$。因此得到

$$1 \le d_1 C^2 r_k^{2s-2} \left(r_{k-1} - r_k\right). \tag{B.12}$$

1. 情形 $s=1$. 在这种情形下，(B.12) 变为

$$1 \leq d_1 C^2 \left(r_k - r_{k+1}\right). \tag{B.13}$$

因为 $r_k \to 0$ 和 $d_1 > 0, C > 0$，这就构成了矛盾。因此，存在 k_1，使得$r_k = 0$ 对所有 $k > k_1$ 都成立。该算法在有限步内终止。

2. 情形 $s \in \left[\frac{1}{2}, 1\right)$. 在此情形下，$0 < 2 - 2s \leq 1$。由于 $r_k \to 0$，存在 $\hat{} k_3$，使得 $r_k^{2-2s} \geq r_k$ 对所有 $k > \hat{} k_3$成立。(B.12)变为

$$r_k \leq d_1 C^2 \left(r_{k-1} - r_k\right).$$

因此，对于所有 $k_2 > \max\left\{k_0, \hat{} k_3\right\}$

$$r_k \leq \left(\frac{d_1 C^2}{1 + d_1 C^2}\right)^{k-k_2} r_{k_2}.$$

and

$$F\left(\mathrm{x}_k\right) - F^* \leq F\left(\mathbf{v}_k\right) - F^* = r_k \leq \left(\frac{d_1 C^2}{1 + d_1 C^2}\right)^{k-k_2} r_{k_2}. \tag{B.14}$$

3. 情形 $s \in \left(0, \frac{1}{2}\right)$. 在这种情形下，$2s - 2 \in (-2, -1)$ $2s - 1 \in (-1, 0)$。由于 $r_{k-1} > r_k$，我们有 $r_{k-1}^{2s-2} < r_k^{2s-2}$和 $r_0^{2s-1} < \cdots < r_{k-1}^{2s-1} < r_k^{2s-1}$。定义 $\phi(t) = \frac{C}{1-2s} t^{2s-1}$，则$\phi'(t) = -Ct^{2s-2}$。

如果 $r_k^{2s-2} \leq 2r_{k-1}^{2s-2}$，那么

$$\begin{aligned}
\phi\left(r_k\right) - \phi\left(r_{k-1}\right) &= C \int_{r_k}^{r_{k-1}} t^{2s-2} dt \\
&\geq C\left(r_{k-1} - r_k\right) r_{k-1}^{2s-2} \\
&\geq \frac{C}{2}\left(r_{k-1} - r_k\right) r_k^{2s-2},
\end{aligned}$$

对所有 $k > k_0$。
如果 $r_k^{2s-2} \geq 2r_{k-1}^{2s-2}$，那么 $r_k^{2s-1} \geq 2^{\frac{2s-1}{2s-2}} r_{k-1}^{2s-1}$，并且

$$\begin{aligned}
\phi\left(r_k\right) - \phi\left(r_{k-1}\right) &= \frac{C}{1-2s}\left(r_k^{2s-1} - r_{k-1}^{2s-1}\right) \\
&\geq \frac{C}{1-2s}\left(2^{\frac{2s-1}{2s-2}} - 1\right) r_{k-1}^{2s-1} \\
&= q r_{k-1}^{2s-1} \\
&\geq q r_0^{2s-1},
\end{aligned}$$

其中 $q = \frac{C}{1-2s}\left(2^{\frac{2s-1}{2s-2}} - 1\right)$。

令 $d_2 = \min\left\{\frac{1}{2d_1C},\ qr_0^{2s-1}\right\}$，我们有

$$\phi(r_k) - \phi(r_{k-1}) \geq d_2,$$

对于所有 $k > k_0$ 和

$$\phi(r_k) \geq \phi(r_k) - \phi(r_{k_0}) \geq \sum_{i=k_0+1}^{k} \phi(r_i) - \phi(r_{i-1}) \geq (k - k_0)\, d_2.$$

因此,

$$r_k \leq \left(\frac{C}{(k-k_0)\, d_2(1-2s)}\right)^{\frac{1}{1-2s}},$$

and

$$F(\mathbf{x}_k) - F^* \leq F(\mathbf{v}_k) - F^* = r_k \leq \left(\frac{C}{(k-k_3)\, d_2(1-2s)}\right)^{\frac{1}{1-2s}}, \tag{B.15}$$

这就完成了证明。

□

C 技术细节

C.1 梯度的推导

考虑 $f(\theta,\ W) = \sum_k l_k(\theta_k) + \frac{1}{2}\sum_{i<j} W_{ij}\|\theta_i - \theta_j\|_2^2 + \frac{\lambda_2\ (1-\eta)}{2}\ \|W - W_0\|_F^2 - \lambda_1 \log\det(\mu I + \mathcal{G}(W))$。我们要计算关于 W 的梯度。尽管在算法中 W 必须是对称且对角线元素为零的，但在推导其梯度时，我们认为它有 $K(K-1)$ 个变量。

将第一项改写为 $f_1 := \frac{1}{2}\sum_{i<j} W_{ij}\|\theta_i - \theta_j\|_2^2 = \frac{1}{4}\sum_{i\neq j} W_{ij}\|\theta_i - \theta_j\|_2^2$。它的梯度为 $\nabla f_1 = (\|\theta_i - \theta_j\|_2^2/4)_{ij}$。

第二项 $f_2 := \frac{\lambda_2(1-\eta)}{2}\ \|W - W_0\|_F^2$ 的梯度为 $\nabla f_2 = \lambda_2(1-\eta)(W - W_0)$。注意，这成立是因为对角元素全为零。

经过繁琐的微积分运算，我们可以推导出最后一项 $f_3 := -\lambda_1 \log\det(\mu I + \mathcal{G}(W))$ 的梯度为 $\nabla f_3 = -\lambda_1 P((\mu I + \mathcal{G}(W))^{-1})$，其中 $[P(L)]_{ij} = L_{ii} - L_{ij}$。同样，对角元素全为零。

最终得到 $\nabla_W f(\theta,\ W) = \nabla f_1 + \nabla f_2 + \nabla f_3$，它通常不是对称的，但对角线元素为零。

C.2 近端算子的推导

令 $g(W) = \lambda_2\eta\|W\|_1 + \sum_{i,\ j} \mathcal{I}_{W_{ij}=W_{ji}\geq 0} + \mathcal{I}_{\mathbf{diag}(W)=0}$。我们希望找到 $\textbf{近端}_{\alpha g}(U) := \arg\min_W\ {}_g(W) \pm \frac{1}{2\alpha}\|W - U\|_F^2$。由于 $W_{ij} = W_{ji}$，我们将它们视为单个变量。KKT 条件是

$$0 \in \lambda_2\eta\alpha\partial\|W_{ij}\|_1 + \partial\mathcal{I}_{W_{ij}\geq 0} + (W_{ij} - \frac{U_{ij} + U_{ji}}{2}),\ \text{ for } i < j.$$

因此，对于 $i < j$ 和 $W_{ii} = 0$， ~~$W_{ij} =$~~ $W_{ji} = \max(0, \frac{U_{ij}+U_{ji}}{2} - \lambda_2\alpha\eta)$。一个关键观察是，在我们的 MAPG算法 中， U 具有 $W - \alpha\nabla f(W)$ 的形式。只要 $\widetilde{W}$ 是对角线元素为零的， U 就是对角线元素为零的（见 C.1）。因此，只要 W_0 和初始输入 $W^{(0)}$~~都~~是对角线元素为零的，表达式 $\mathbf{prox}^{\alpha g}(K) = \max(0, \frac{U+U^T}{2} - \lambda_2\alpha\eta)$ 就是有效的。