中国科学院
CHINESE ACADEMY OF SCIENCES

机器学习中的优化问题

从半监督学习到深度神经网络模型

答辩人：刘为

指导教师：刘歆 研究员

勤数 笃系 求天 真地
中国科学院数学与系统科学研究院
Academy of Mathematics and Systems Science
Chinese Academy of Sciences

计算数学与科学工程计算研究所
科学与工程计算国家重点实验室

中国科学院大学

博士学位论文答辩

2022 年 5 月 17 日，中国北京

目录

引言

机器学习

- 机器学习旨在设计**通用算法**以**自动地**从给定数据中得到这些数据所属的一个概率分布或泛函空间，简称提取特征

- 理论背景
  - 计算机科学 [Zhou 2016]
  - 应用数学 [Deisenroth-Faisal-Ong 2020]
  - 生物信息学 [Baldi-Brunak 2001]

- 应用场景
  - 搜索引擎 [Mahesh 2020]
  - 软件推荐系统 [Grandinetti 2021]
  - 人脸识别 [Zhao et al. 2003]
  - 人工智能 [Goodfellow et al. 2016]
  - 航天航空 [Maheshwari-Davendralingam-DeLaurentis 2018]
  - 美国总统竞选 [Rothwell-Diego 2016]

机器学习（续）

- 核心
  - 数据
  - 模型：利用统计方法构建模型
  - 学习：利用优化方法求解模型

- 按标签有无分类
  - 监督学习：如分类问题
  - 无监督学习：如聚类问题

- 无监督学习中的常见算法
  - k 均值聚类 [Steinhaus 1956; Lloyd 1982; ...]
  - 谱聚类 [Hagen-Kahng 1992; Liu et al. 2018; Zhang 2019; ...]

半监督学习

- **适用范围**
  - 对所有数据进行标签标注代价过高
  - 存在部分数据的标签已标注或容易标注

- 常用于分类问题、半监督聚类问题

	聚类	半监督聚类	分类
标签	无	**部分**	部分
训练目标	找出数据中的特征	找出数据中的特征	确认数据属于哪个类别
测试目标	根据特征进行聚类	根据特征进行聚类	将新数据分配到已知类别中
类的个数	未知	**有一定估计**	已知
类别	无监督学习	半监督学习	半监督学习，监督学习

- **标注标签分类**

强标注标签	数据与数据之间属于同一类 数据与数据之间不属于同一类
弱标注标签	数据与数据之间**大概率**不属于同一类 数据与数据之间**大概率**属于同一类

CHINESE ACADEMY OF SCIENCES

半监督学习中的优化方法

传统半监督学习方法

- 半监督中的分类问题 [Hastie-Tibshirani-Friedman 2009; ...]
- 基于 k 均值聚类的半监督学习[Wagstaff et al. 2001; ...]
- 基于谱聚类的半监督学习 [Zhou et al. 2003; Zhu 2005; ...]

基于深度神经网络的半监督学习 (包括分类与聚类)

[Hinton-Salakhutdinov 2006; van Engelen-Hoos 2020; ...]

- 更好的数值结果
- 更多的实际应用

中国科学院 CHINESE ACADEMY OF SCIENCES

基于谱聚类的半监督学习

谱聚类

$$\min_{H\in\mathbb{R}^{N\times k}} \operatorname{tr}\left(H^{\mathrm{T}}\mathcal{L}(S)H\right)$$
$$\text{s. t. } H^{\top}H = I_k$$

- 样本数：N，类的划分个数：k，数据矩阵：$X = (x_1, \ldots, x_N) \in \mathbb{R}^{N_0\times N}$
- 相似度矩阵：$S \in \mathbb{R}^{N\times N}$，$S_{ij} = \exp(-\frac{\|x_i-x_j\|^2}{0.1^2})$ [von Luxburg 2007]
- 拉普拉斯算子：$\mathcal{L} : \mathbb{R}^{N\times N} \mapsto \mathbb{R}^{N\times N}$
- 基于 Rayleigh-Ritz 定理 [Lutkepohl 1947]求解
- k 均值聚类后处理理解

半监督形式

$$\min_{H\in\mathbb{R}^{N\times k}} \operatorname{tr}\left(H^{\mathrm{T}}\mathcal{L}(S)H\right) + \lambda\|AH - H^*\|^2$$

- 指示矩阵：$A \in \mathbb{R}^{r\times N}$，标签数 r
- 预设的聚类矩阵（弱标注标签）：$H^* \in \mathbb{R}^{r\times k}$
- 参数：$\lambda > 0$
- 无正交约束

基于谱聚类的半监督学习

谱聚类

$$\min_{H\in\mathbb{R}^{N\times k}} \operatorname{tr}\left(H^{\mathrm{T}}\mathcal{L}(S)H\right)$$
$$\mathrm{s.t.}\ H^{\top}H = I_k$$

- 样本数：N，类的划分个数：k，数据矩阵：$X = (x_1, \ldots, x_N) \in \mathbb{R}^{N_0\times N}$
- 相似度矩阵：$S \in \mathbb{R}^{N\times N}$，$S_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{0.1^2})$ [von Luxburg 2007]
- 拉普拉斯算子：$\mathcal{L}: \mathbb{R}^{N\times N} \mapsto \mathbb{R}^{N\times N}$
- 基于 Rayleigh-Ritz 定理 [Lutkepohl 1947]求解
- k 均值聚类后处理解

半监督形式

$$\min_{H\in\mathbb{R}^{N\times k}} \operatorname{tr}\left(H^{\mathrm{T}}\mathcal{L}(S)H\right) + \lambda\|AH - H^*\|^2$$

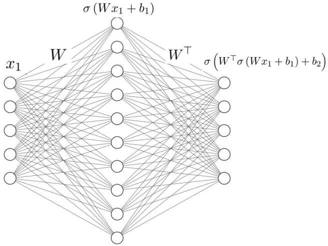
- 指示矩阵：$A \in \mathbb{R}^{r\times N}$，标签数 r
- 预设的聚类矩阵（弱标注标签）：$H^* \in \mathbb{R}^{r\times k}$
- 参数：$\lambda > 0$
- 无正交约束

基于非光滑自编码的半监督学习

自编码：一种特殊的两层神经网络[Goodfellow et al. 2016]

$$\min_{W,b} \frac{1}{N} \sum_{n=1}^{N} \|\sigma(W^\top \sigma(W x_n + b_1) + b_2) - x_n\|^2 \tag{AE}$$

- 样本数：N, 样本空间：$\mathbb{R}^{N_0}$, 隐藏层单元数：N_1
- 权重矩阵：$W \in \mathbb{R}^{N_1 \times N_0}$, 偏差向量：$b = (b_1^\top, b_2^\top)^\top \in \mathbb{R}^{N_1+N_0}$
- 激活函数：$\sigma(z) := \max\{z, \alpha z\}$, $0 \le \alpha < 1$
  表现最好[Jarrett et al. 2009; Agarap 2018]
  - ReLU：$\alpha = 0$
  - leaky ReLU：$0 < \alpha < 1$



基于非光滑自编码的半监督学习（续）

半监督形式：

$$\min_{W,b} \frac{1}{N}\sum_{n=1}^{N}\|\sigma(W^\top\sigma(Wx_n+b_1)+b_2)-x_n\|^2+\frac{\lambda_1}{r}\|A\sigma(WX+b_1e_N)-H^*\|^2$$
$$+\lambda_2\,\mathrm{tr}\left(\sigma(WX+b_1e_N)^\top\mathcal{L}(S)\sigma(WX+b_1e_N)\right)$$

- 参数：$\lambda_1, \lambda_2 > 0$
- 全为 1 的向量：$e_N \in \mathbb{R}^N$
- 对 $H = \sigma(WX+b_1e_N^\top)$**后处理**以得到聚类结果

训练非光滑深度神经网络的已有算法

- 随机梯度下降算法 (简称 SGD 类算法, 基于反向传播算法[1] [Werbos 1990])
  - Vanilla SGD [Cramir 1946]
  - RMSProp [Riedmiller-Braun 1993]
  - Adagrad, AdagradDecay [Duchi-Hazan-Singer 2011]
  - Adadelata [Zeiler 2012]
  - Adam, Adamax [Kingma-Ba 2014]
  - AMSGrad [Reddi-Kale-Kumar 2019]
  - ProxSGD [Yang et al. 2021]
- 随机次梯度下降算法[2] [Davis et al. 2020]
- 罚方法
  - l_2 罚方法[1] [Carreira Perpinan-Wang 2014; Lau et al. 2018; Zeng et al. 2019; Evens et al. 2021]
  - 增广拉格朗日法[1] [Taylor et al. 2016]
  - l_1 罚方法 MM+SN[2] [Cui-He-Pang 2021]: 利用变分分析知识

[1] 无法保证收敛到 Clarke 稳定点
[2] 有收敛性保证, 但算法效率低

变分分析[Rockafellar-Wets 2009]

对于 Lipschitz 连续函数 f

- 方向导数：$f'(z;d) := \lim_{t\downarrow 0} \frac{f(z+td)-f(z)}{t}$
- regular 次微分：$\widehat{\partial} f(z) := \left\{ v : \liminf_{z^k\to z} \frac{f(z^k)-f(z)-\langle v, z^k - z\rangle}{\|z^k - z\|} \geq 0 \right\}$
- limiting 次微分：$\partial f(z) := \left\{ v : \exists z^k \xrightarrow{f} z, v^k \to v \text{使得} v^k \in \widehat{\partial} f(z^k),\ \forall k \right\}$
- Clarke 次微分：
  $\partial^c f(z) := \left\{ v \in \mathbb{R}^n : \limsup_{z^k\to z, t\downarrow 0} \frac{f(z^k+tw)-f(z^k)-tv^\top w}{t} \geq 0, \quad \forall w \in \mathbb{R}^n \right\}$

对于闭集 $\mathcal{Z}$ 上点 z

- 示性函数：$\delta_{\mathcal{Z}}(z)$
- limiting 法锥：$\mathcal{N}_{\mathcal{Z}}(z) := \partial \delta_{\mathcal{Z}}(z)$
- Clarke 法锥：$\mathcal{N}^c_{\mathcal{Z}}(z) := \mathrm{clco}\mathcal{N}_{\mathcal{Z}}(z)$

稳定点

对于问题 $\min_{z\in\mathcal{Z}} f(z)$, 称 z 为

- D(irectional)-稳定点, 若

$$f'(z;d) \geq 0,\ \ \forall d \in \mathcal{T}_{\mathcal{Z}}(z)$$

- 强 l(imiting)-稳定点, 若

$$0 \in \partial\left(f(z) + \delta_{\mathcal{Z}}(z)\right)$$

- **l-稳定点**, 若

$$0 \in \partial f(z) + \mathcal{N}_{\mathcal{Z}}(z)$$

- **C(larke)-稳定点**, 若

$$0 \in \partial^c f(z) + \mathcal{N}^c_{\mathcal{Z}}(z)$$

$$\boxed{\text{D-稳定点} \Rightarrow \text{强 l-稳定点} \Rightarrow \text{l-稳定点} \Rightarrow \text{C-稳定点}.}$$

若 $\mathcal{Z}$ 为闭凸集且 f 为适当的凸函数, 则上述定义等价

工作一：半监督谱聚类新模型

研究动机：传统的的半监督谱聚类方法

- 需要额外用某种聚类方法后处理问题的解
- 需要预先知道精确的类的划分个数，且对类的划分个数敏感
- 无法同时处理强标注标签和弱标注标签

解决方案：提出**新方法**

- 新模型的解**显式**地给出聚类结果
- 新模型**不需要**精确的类的划分个数
- 新模型**同时支持**强标注标签和弱标注标签
- 设计快速有效算法

工作二、工作三：非光滑网络的优化算法

研究动机：现有的训练深度神经网络的算法

- 基于链式法则求导数，而其在非光滑网络上不严格成立
- 缺乏收敛性证明

工作二解决方案：提出训练非光滑自编码的新模型与算法

- 保证**收敛性**，严格求得其某一稳定点
- 保证**数值有效性**

自编码 $\Rightarrow$ 深度神经网络：变量更多、问题结构不再保持！

工作三解决方案：提出训练非光滑深度神经网络的新模型与新算法

- 保证**收敛性**，严格求得其某一稳定点
- 保证**数值有效性**

主要工作与贡献

- 设计了一个**半监督谱聚类连续优化模型**, 提出了一种有限步收敛的块坐标下降算法
  - 无需后处理
  - 无需类的划分个数的精确值
  - 同时处理强标注标签和弱标注标签

- 为 (稀疏) 自编码码问题 (R) 提出一类 l_1 罚模型 (LRP), 以及一类收敛到 C-稳定点的**光滑化临近点算法**
  - 解集有界性
  - 精确罚性
  - 数值有效性

- 为带组稀疏正则的非光滑网络模型 (P) 提出一类 l_1 罚模型 (PP), 以及一类收敛到 l-稳定点的**非精确增广拉格朗日算法**
  - 解集有界性
  - 精确罚性
  - 数值有效性

工作一：一类新的半监督谱聚类的优化模型与算法

模型设计思路

- 变量：结果图 $\mathcal{G}(X, S \circ Z)$ 的无权重邻接矩阵 $Z \in \mathcal{S}_S^N \cap \{0,1\}^{N\times N}$
  - $\mathcal{S}_S^n = \{Z \in \mathcal{S}^n : \operatorname{supp}(Z) \subset \operatorname{supp}(S)\}$, 对称矩阵集：$\mathcal{S}^n$, 支撑集：supp
  - $$Z_{ij} = \begin{cases} 1 & \text{如果保留图}\mathcal{G}(X,S)\text{ 的边}(i,j), \\ 0 & \text{如果未保留图}\mathcal{G}(X,S)\text{ 的边}(i,j). \end{cases}$$
- 整数规划模型

$$\min_{Z} \ \operatorname{rank}(\mathcal{L}(S \circ Z)) - \beta \operatorname{tr}(SZ)$$
$$\text{s.t. } Z \in \mathcal{S}_S^N \cap \{0,1\}^{N\times N}$$

  - 使 $\mathcal{G}(X, S \circ Z)$ 边的**总权重 $\frac{1}{2}\operatorname{tr}(SZ)$ 尽可能大**
  - 其**连通分支尽可能多**但不多于 d：$N - \operatorname{rank}(\mathcal{L}(S \circ Z)) < d$ [Mohar et al. 1991]
  - 衡量参数 β

模型设计思路

松弛+

$$N-d=\min_{Z\in\mathcal{S}_S^N\cap\{0,1\}^{N\times N}}\operatorname{rank}(\mathcal{L}(S\circ Z))\Leftrightarrow 0=\min_{\substack{H^\top H=I_d\\ Z\in\mathcal{S}_S^N\cap\{0,1\}^{N\times N}}}\operatorname{tr}\left[H^{\mathrm{T}}\mathcal{L}(S\circ Z)H\right]$$

$$\Downarrow$$

$$\begin{aligned}\min_{Z,H}\ & f(Z,H)=\operatorname{tr}\left[H^\top\mathcal{L}(S\circ Z)H\right]-\beta\operatorname{tr}(SZ)\\ \text{s.t. }\ & Z\in\mathcal{S}_S^N\cap[0,1]^{N\times N}\\ & H^\top H=I_d\end{aligned}\tag{SC}$$

根据半监督信息构造相似度矩阵 S, 主要考虑前三种

- x_i 与 x_j 大概率属于同一类：按概率大小同比例扩大 S_{ij} 和 S_{ji} 的值
- x_i 与 x_j 大概率不属于同一类：按概率大小同比例缩小 S_{ij} 和 S_{ji} 的值
- x_i 与 x_j 属于同一类：令 $S_{ij}=S_{ji}=\infty$
- x_i 与 x_j 不属于同一类：我们令 $S_{ij}=S_{ji}=0$, 且添加 x_i,x_j 不属于同一连通分支的约束

模型设计思路

松弛+

$$N - d = \min_{Z \in \mathcal{S}_S^N \cap \{0,1\}^{N\times N}} \operatorname{rank}(\mathcal{L}(S \circ Z)) \Leftrightarrow 0 = \min_{\substack{H^\top H = I_d \\ Z \in \mathcal{S}_S^N \cap \{0,1\}^{N\times N}}} \operatorname{tr}\left[H^\mathrm{T} \mathcal{L}(S \circ Z) H\right]$$

$$\Downarrow$$

$$\begin{aligned} \min_{Z,H}\ & f(Z,H) = \operatorname{tr}\left[H^\top \mathcal{L}(S \circ Z) H\right] - \beta \operatorname{tr}(SZ) \\ \text{s.t.}\ & Z \in \mathcal{S}_S^N \cap [0,1]^{N\times N} \\ & H^\top H = I_d \end{aligned} \tag{SC}$$

根据半监督信息构造相似度矩阵 S, 主要考虑前三种

- x_i 与 x_j 大概率属于同一类：按概率大小同比例扩大 S_{ij} 和 S_{ji} 的值
- x_i 与 x_j 大概率不属于同一类：按概率大小同比例缩小 S_{ij} 和 S_{ji} 的值
- x_i 与 x_j 属于同一类：令 $S_{ij} = S_{ji} = \infty$
- x_i 与 x_j 不属于同一类：我们令 $S_{ij} = S_{ji} = 0$, 且添加 x_i, x_j 不属于同一连通分支的约束

块坐标下降算法设计思路

步一： $Z^{(k)} \to Z^{(k+1)}$

$$\begin{aligned} Z^{(k+1)} \in \arg\min_{Z} \; & f(Z, H^{(k)}) \\ \text{s.t. } & Z \in \mathcal{S}_S^N \cap [0,1]^{N\times N} \end{aligned} \Rightarrow Z_{ij}^{(k+1)} = \begin{cases} 0, & \text{if } g(H^{(k)})_{ij} > 0 \\ Z_{ij}^{(k)}, & \text{if } g(H^{(k)})_{ij} = 0 \\ 1, & \text{if } g(H^{(k)})_{ij} < 0 \end{cases} \tag{1}$$

其中 $g(H) := \nabla_Z f(Z,H) + \nabla_Z f(Z,H)^\top$

步二： $H^{(k)} \to H^{(k+1)}$

$$H^{(k+1)} \in \arg\min_{H^\top H = I_d} \operatorname{tr}\left[H^\top \mathcal{L}(S \circ Z^{(k+1)}) H\right] \tag{2}$$

算法框架

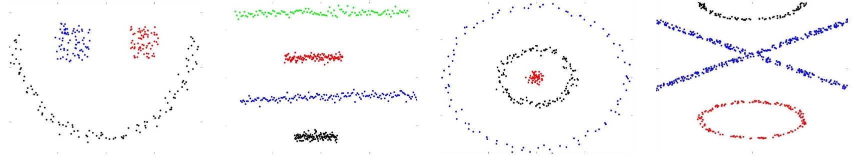
求解问题 (SC) 的块坐标下降算法 (CO-BCD)

1. 输入：相似度矩阵 S 和 $\beta > 0$. 令 $Z^0 \in \mathcal{S}_S^N \cap \{0,1\}^{N\times N}$, $k := 0$
2. 通过 (1) 计算 $Z^{(k+1)}$
3. 基于 Rayleigh-Ritz 定理求解 (2) 并得到 $H^{(k+1)}$
4. 令 $k := k+1$; 若终止条件未满足，回到第 2 步
5. 输出：在 $Z^{(k)}$ 上利用广度优先搜索[Cormen et al. 2009]求得 $\mathcal{G}(X, S \circ Z)$ 的连通分支及聚类结果

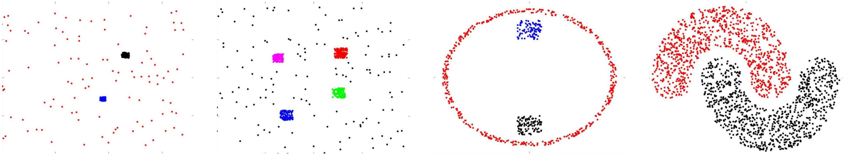
定理 1

令 $\{(Z^{(k)}, H^{(k)})\}$ 是由算法 CO-BCD 生成的序列，则序列 $\{Z^{(k)}\}$ 有限步收敛，且序列 $\{(Z^{(k)}, H^{(k)})\}$ 的任一聚点 (Z^*, H^*) 是问题 (SC) 的一个 KKT 点.

图像聚类结果



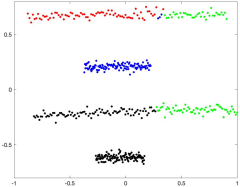
(a) $d = 3$ 到 10 (b) $d = 4$ 到 10 (c) $d = 3$ 到 10 (d) $d = 3$ 到 10



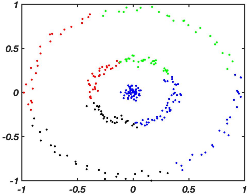
(e) $d = 3$ 到 5 (f) $d = 5$ 到 8 (g) $d = 3$ 到 10 (h) $d = 2$ 到 10

- $\beta = d/N$
- 弱标注标签：对于数据 x_i 和 x_j, 若数据 x_i 不为 x_j 最近的 k 个数据或 x_j 不为 x_i 最近的 k 个数据，令 $S_{ij} = S_{ji} = 0$, $k = \lfloor \log(N) \rfloor + 1$ [Brito et al. 1997]

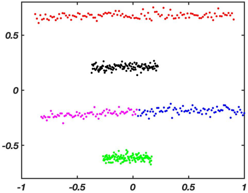
其它算法的图像聚类结果



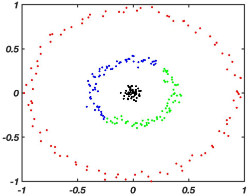
(a) $d = 4$, k 均值



(b) $d = 4$, k 均值



(c) $d = 5$, 谱聚类



(d) $d = 4$, 谱聚类

只知道类的数目的上界估计, 我们的方法仍然有效, 但其它方法不一定

工作二：自编码的一类带线性约束的非光滑优化模型及算法

非光滑网络中链式法则不严格成立

链式法则不成立

$$\min_{z\in\mathbb{R}} f(z) = \frac{1}{2}\left(-(zx_1)_+ + (zx_2)_+ + 1\right)^2 + \frac{1}{2}\left(-(zy_1)_+ + (zy_2)_+ + 1\right)^2$$

其中 $x = (-1, 1), y = (-2, 0), z_+ := \max\{0, z\}$.
SGD 类算法将

$$\begin{aligned} g(z) := &\left(-(zx_1)_+ + (zx_2)_+ + 1\right)\left(-x_1 h(zx_1) + x_2 h(zx_2)\right) \\ &+ \left(-(zy_1)_+ + (zy_2)_+ + 1\right)\left(-y_1 h(zy_1) + x_2 h(zy_2)\right) \end{aligned}$$

视为 f 在 z 点处的导数, 其中 $h(z) := \operatorname{sign}(z_+)$.
然而 $g(0) = 0$ 并不属于 f 在 z 点处的 Clarke 次微分.
0 不是一个稳定点!

更多的例子: [April-July-Kummer 2011]

$\alpha = 0$ 时 (AE) 解集无界[3]

解集无界性

令 (W^*, b^*) 为 (AE) 的一个全局极小值点, $z_+ = \max\{z, 0\}$

$$X = (x_1, x_2) = \begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

$$W = [w_1, w_2] \in \mathbb{R}^{1\times 2}, b_1 \in \mathbb{R}, b_2 = \begin{bmatrix} b_{2,1} \\ b_{2,2} \end{bmatrix} \in \mathbb{R}^2$$

令 $\hat{b}_1 = b_1^*, \hat{b}_{2,2} = b_{2,2}^*$,
$\hat{b}_{2,1} \leq \min\left\{-w_1^*\left(w_2^* + b_1^*\right)_+, -w_1^*\left(2w_2^* + b_1^*\right)_+\right\}$.
则 $\left(W^*, \hat{b}\right)$ 也是 (AE) 的一个全局极小值点.
考虑到 $\hat{b}$ 的定义 $\Rightarrow$(AE) 的解集无界.

[3] 第 9 页

模型设计思路

- ReLU **激活函数**, $\alpha = 0$ + （**稀疏**）**自编码**

$$\underbrace{\min_{W,b} \frac{1}{N}\sum_{n=1}^{N} \|(W^\top (Wx_n + b_1)_+ + b_2)_+ - x_n\|_2^2}_{\text{拟合项}} + \underbrace{\lambda_1 \sum_{n=1}^{N} e_{N_1}^\top (Wx_n + b_1)_+ + \lambda_2 \|W\|_F^2}_{\text{正则项[Ng et al. 2011]}}$$

  - **正则项系数**： $\lambda_1, \lambda_2 > 0$

- 链式法则不成立 $\Rightarrow$ 考虑非凸非光滑分析

$$\min_{z} \underbrace{\frac{1}{N}\sum_{n=1}^{N} \|(W^\top v_n + b_2)_+ - x_n\|_2^2}_{\mathcal{F}(z)} + \underbrace{\lambda_1 \sum_{n=1}^{N} e_{N_1}^\top v_n + \lambda_2 \|W\|_F^2}_{\mathcal{R}(z)} \tag{R}$$

$$\text{s. t. } v_n = (Wx_n + b_1)_+, \ n \in [N] := \{1, 2, \ldots, N\}$$

  - $z = (\mathrm{vec}(W)^\top, b^\top, \mathrm{vec}(V)^\top)^\top \in \mathbb{R}^{N_2}$

中国科学院 CHINESE ACADEMY OF SCIENCES

模型设计思路

- ReLU 激活函数, $\alpha = 0$ + (稀疏) 自编码

$$\min_{W,b} \underbrace{\frac{1}{N}\sum_{n=1}^{N} \|(W^\top (Wx_n + b_1)_+ + b_2)_+ - x_n\|_2^2}_{\text{拟合项}} + \underbrace{\lambda_1 \sum_{n=1}^{N} e_{N_1}^\top (Wx_n + b_1)_+ + \lambda_2 \|W\|_F^2}_{\text{正则项[Ng et al. 2011]}}$$

  - 正则项系数: $\lambda_1, \lambda_2 > 0$
- 链式法则不成立 $\Rightarrow$ 考虑非凸非光滑分析

$$\min_{z} \underbrace{\frac{1}{N}\sum_{n=1}^{N} \|(W^\top v_n + b_2)_+ - x_n\|_2^2}_{\mathcal{F}(z)} + \underbrace{\lambda_1 \sum_{n=1}^{N} e_{N_1}^\top v_n + \lambda_2 \|W\|_F^2}_{\mathcal{R}(z)} \tag{R}$$

$$\text{s.t. } v_n = (Wx_n + b_1)_+,\ n \in [N] := \{1, 2, \ldots, N\}$$

  - $z = (\mathrm{vec}(W)^\top, b^\top, \mathrm{vec}(V)^\top)^\top \in \mathbb{R}^{N_2}$

中国科学院 CHINESE ACADEMY OF SCIENCES

模型设计思路（续）

- 链式法则不成立 ⇒ 考虑非凸非光滑分析 ⇒ **构建罚模型**

$$\min_z \mathcal{O}(z) := \mathcal{F}(z) + \mathcal{R}(z) + \underbrace{\beta \sum_{n=1}^{N} e_{N_1}^\top (v_n - (Wx_n + b_1)_+)}_{\text{惩罚项}\mathcal{P}(z),\ \beta>0} \tag{RP}$$

$$\text{s.t. } v_n \geq (Wx_n + b_1)_+,\ n \in [N]$$

  - $\sum_{n=1}^N \|v_n - (Wx_n + b_1)_+\|_1 = \sum_{n=1}^N e_{N_1}^\top (v_n - (Wx_n + b_1)_+)$ $\Longrightarrow$ **$\partial^c \mathcal{O}(z)$ 有显式表达**

- 解集无界性 ⇒ 构造“有界”集合，使 (RP) 在该集合内有全局解

$$\min_z \mathcal{O}(z) \tag{LRP}$$

$$\text{s.t. } \|b\|_\infty \leq a,\ v_n \geq (Wx_n + b_1)_+,\ n \in [N]$$

  - 正常数 $a := \frac{\|X\|_F^2}{\lambda_1 N} + \sqrt{\frac{N_1 N_0}{\lambda_2 N}} \|X\|_F \|X\|_1$：可计算！

模型设计思路（续）

- 链式法则不成立 ⇒ 考虑非凸非光滑分析 ⇒ 构建罚模型

$$\min_{z} \mathcal{O}(z) := \mathcal{F}(z) + \mathcal{R}(z) + \underbrace{\beta \sum_{n=1}^{N} e_{N_1}^{\top} (v_n - (Wx_n + b_1)_+)}_{\text{惩罚项}\mathcal{P}(z),\ \beta>0} \qquad \text{(RP)}$$

$$\text{s.t. } v_n \geq (Wx_n + b_1)_+,\ n \in [N]$$

  - $\sum_{n=1}^{N} \|v_n - (Wx_n + b_1)_+\|_1 = \sum_{n=1}^{N} e_{N_1}^{\top} (v_n - (Wx_n + b_1)_+)$
    $\Longrightarrow \partial^c \mathcal{O}(z)$ 有显式表达
- 解集无界性 ⇒ 构造“有界”集合，使 (RP) 在该集合内有全局解

$$\min_{z} \mathcal{O}(z) \qquad \text{(LRP)}$$

$$\text{s.t. } \|b\|_\infty \leq a, v_n \geq (Wx_n + b_1)_+,\ n \in [N]$$

  - 正常数 $a := \frac{\|X\|_F^2}{\lambda_1 N} + \sqrt{\frac{N_1 N_0}{\lambda_2 N}} \|X\|_F \|X\|_1$：可计算！

中国科学院 CHINESE ACADEMY OF SCIENCES

模型分析

令 $\theta > \frac{1}{N}\|X\|_F^2$，水平集 $\Omega_\theta := \{z : \mathcal{O}(z) \le \theta, v_n \ge (Wx_n + b_1)_+, n \in [N]\}$

定理 2（解集有界性）

对于 $z \in \Omega_\theta$，我们有：

(a) $\|W\|_F^2 \le \frac{\theta}{\lambda_2}, \|V\|_1 \le \frac{\theta}{\lambda_1}$ 和 $\|b_+\|_\infty \le a$.

(b) $\bar{z} = \mathrm{Proj}_{z:\|b\|_\infty \le a}(z)$ 为 (LRP) 的可行点，且 $\mathcal{O}(\bar{z}) = \mathcal{O}(z)$.

(LRP) 的解集非空有界，且属于 (RP) 的解集.

精确罚性

R :	全局极小值点	局部极小值点	d-稳定点	C-稳定点
	$\Updownarrow$	$\Downarrow$	$\Uparrow$	$\Uparrow$
RP :	全局极小值点	局部极小值点	d-稳定点	C-稳定点
	$\|\bar{b}\|_\infty \le a \Downarrow\Uparrow$	$\|\bar{b}\|_\infty \le a \Downarrow\Uparrow$	$\Uparrow \mathcal{O}(\bar{z}) < \theta$	$\Uparrow \|\bar{b}\|_\infty < a, \mathcal{O}(\bar{z}) < \theta$
LRP :	全局极小值点	局部极小值点	d-稳定点	C-稳定点

- 相应的点为 $\bar{z}$，其函数值小于 θ
- β 大于 L_{FR}，其为 $\mathcal{F} + \mathcal{R}$ 在 Ω_θ 上的 Lipschitz 常数

模型分析

令 $\theta > \frac{1}{N}\|X\|_F^2$, 水平集 $\Omega_\theta := \{z : \mathcal{O}(z) \le \theta, v_n \ge (Wx_n + b_1)_+, n \in [N]\}$

定理 2（解集有界性）

对于 $z \in \Omega_\theta$, 我们有:

(a) $\|W\|_F^2 \le \frac{\theta}{\lambda_2}, \|V\|_1 \le \frac{\theta}{\lambda_1}$ 和 $\|b_+\|_\infty \le a$.

(b) $\bar{z} = \mathrm{Proj}_{z:\|b\|_\infty \le a}(z)$ 为 (LRP) 的可行点, 且 $\mathcal{O}(\bar{z}) = \mathcal{O}(z)$.

(LRP) 的解集非空有界, 且属于 (RP) 的解集.

精确罚性

R :	全局极小值点	局部极小值点	d-稳定点	C-稳定点
	$\Updownarrow$	$\Downarrow$	$\Uparrow$	$\Uparrow$
RP :	全局极小值点	局部极小值点	d-稳定点	C-稳定点
	$\|\bar{b}\|_\infty \le a \Downarrow\Uparrow$	$\|\bar{b}\|_\infty \le a \Downarrow\Uparrow$	$\Uparrow\ \mathcal{O}(\bar{z}) < \theta$	$\Uparrow\ \|\bar{b}\|_\infty < a, \mathcal{O}(\bar{z}) < \theta$
LRP :	全局极小值点	局部极小值点	d-稳定点	C-稳定点

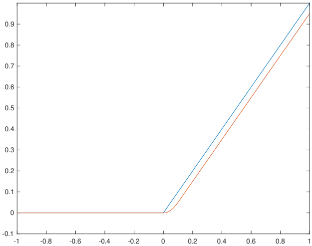
- 相应的点为 $\bar{z}$, 其函数值小于 θ
- β 大于 L_{FR}, 其为 $\mathcal{F} + \mathcal{R}$ 在 Ω_θ 上的 Lipschitz 常数

光滑化函数

$\mathcal{O}$ 非光滑性！⇒ **光滑化技术！**

$\sigma(z)$ 的光滑化函数

$$\widetilde{\sigma}_i(z,\mu)=\begin{cases}0 & \text{若} y_i<0\\ \dfrac{z_i^2}{2\mu} & \text{若} 0\le z_i\le\mu\\ z_i-\dfrac{\mu}{2} & \text{若} z_i>\mu\end{cases}$$



⇒ $\mathcal{O}(z)$ 的光滑化函数

$$\widetilde{\mathcal{O}}(z,\mu):=\widetilde{\mathcal{F}}(z,\mu)+\widetilde{P}(z,\mu)+\mathcal{R}(z)$$

- $\mu>0$
- $\widetilde{\mathcal{F}}(z,\mu)=\frac{1}{N}\sum_{n=1}^{N}\|(W^\top v_n+b_2)_+\|^2+\frac{1}{N}\|X\|_F^2-\frac{2}{N}\sum_{n=1}^{N}x_n^\top\widetilde{\sigma}(W^\top v_n+b_2,\mu)$
- $\widetilde{P}(z,\mu)=\beta\sum_{n=1}^{N}e_{N_1}^\top(v_n-\widetilde{\sigma}(Wx_n+b_1,\mu))$

中国科学院 CHINESE ACADEMY OF SCIENCES

算法框架

求解问题 (LRP) 的一类光滑化临近梯度算法 (SPG)

1. 输入： 选取 $z^{(0)}$ 为 (LRP) 的可行点, $0 < \mu^{(0)} < 1$, $0 < \tau_1 < 1$, $\tau_2 > 0$, $\tau_3 \geq 1$, $L^{(0)} \geq 1$, 取 $k := 0$
2. 令 $z^{(k+1)}$ 为下面这个带线性约束的强凸规划的唯一解[4]

$$\begin{aligned} \min_z \ & \left\langle \nabla_z(\widetilde{\mathcal{F}} + \widetilde{\mathcal{P}})(z^{(k)}, \mu^{(k)}), z - z^{(k)} \right\rangle + \mathcal{R}(z) + \frac{L^{(k)}}{2}\|z - z^{(k)}\|_2^2 \\ \text{s. t. } & \|b\|_\infty \leq a, v_n \geq (Wx_n + b_1)_+, \ n \in [N] \end{aligned}$$

3. 令

$$\begin{cases} (\mu^{(k+1)}, L^{(k+1)}) := (\mu^{(k)}, L^{(k)}), & \text{若 } \widetilde{O}(z^{(k+1)}, \mu^{(k)}) - \widetilde{O}(z^{(k)}, \mu^{(k)}) < -\tau_2 \dfrac{\mu^{(k)}}{L^{(k)}} \\ (\mu^{(k+1)}, L^{(k+1)}) := (\tau_1 \mu^{(k)}, \tau_3 L^{(k)}), & \text{否则} \end{cases}$$

4. 令 $k := k + 1$, 若终止条件未满足, 回到第 2 步

[4] 可被 ‘quadprog’[Weingessel 2007]和 ‘CVX’ 求解[Grant-Boyd 2014]

中国科学院 CHINESE ACADEMY OF SCIENCES

收敛性分析

定理 3

令 $\{z^{(k)}\}$ 和 $\{\mu^{(k)}\}$ 为算法 SPG 生成的序列. 假设 $\mathcal{O}(z^{(0)}) < \theta$, $\tau_1\tau_3 \geq 1$, $\mu^{(0)}L^{(0)}$ 满足

$$\mu^{(0)}L^{(0)} \geq \max\left\{6\lambda_2 N_1 N_0 + \frac{2}{\eta}(N_2 L_{FR} + \lambda_1 N_1 N),\ 8\lambda_2 + L_{FR}\right\}.$$

则下述命题成立

(a) 序列 $\left\{\widetilde{\mathcal{O}}(z^{(k)}, \mu^{(k)})\right\}$ 非增, 且 $\{z^{(k)}\} \subset \Omega_\theta$;

(b) $\lim_{k\to\infty} \mu^{(k)} = 0$, 且 $\{\mathcal{O}(z^{(k)})\}$ 与 $\{\widetilde{\mathcal{O}}(z^{(k)}, \mu^{(k)})\}$ 收敛;

(c) 令 $\mathcal{K} = \{k : \mu^{(k+1)} = \tau_1\mu^{(k)}, k \geq 0\}$. 则 $\{z^{(k)} : k \in \mathcal{K}\}$ 有界, 且序列 $\{z^{(k)} : k \in \mathcal{K}\}$ 的任一聚点 z^* 是 (LRP) 的一个 C-稳定点.

数值实验

默认设置

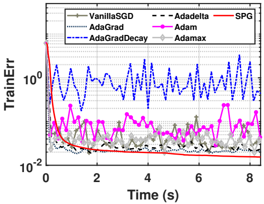
- 最大迭代数： 4000
- 初始化： $W^{(0)} = \mathrm{randn}(N_1, N_0)/N$, $b^{(0)} = 0$, $v_n^{(0)} = (W^{(0)}x_n)_+$, $\forall n \in [N]$
- 停机准则: $\mu^{(k)} \le 10^{-7}$
- 测试集样本数： $N_{\text{test}} = \lceil N/5 \rceil$
- SGD 类算法直接求解问题 (AE), 每个迭代小样本选取为 $\lceil \sqrt{N} \rceil$ [Kasai 2018]

衡量标准

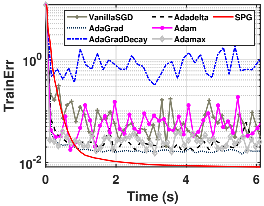
- TrainErr： $\frac{1}{N}\sum_{n=1}^{N} \|(W^\top(Wx_n + b_1)_+ + b_2)_+ - x_n\|^2$
- TestErr： $\frac{1}{N_{\text{test}}}\sum_{n=N_{\text{test}}+1}^{N+N_{\text{test}}} \left\|(W^\top v_n + b_2)_+ - x_n\right\|^2$
- Time (s)

中国科学院 CHINESE ACADEMY OF SCIENCES

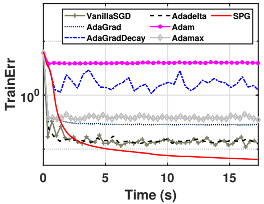
在高斯分布数据集上的数值变化比较



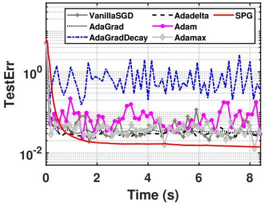
(a) $N = 75,\ N_1 = 10,\ N_0 = 5$



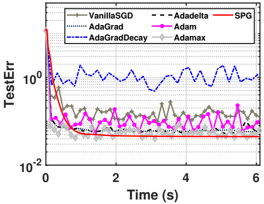
(b) $N = 100,\ N_1 = 10,\ N_0 = 5$



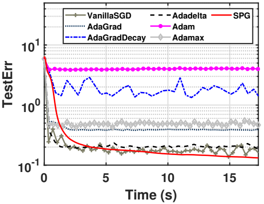
(c) $N = 150,\ N_1 = 20,\ N_0 = 10$



(d) $N = 75,\ N_1 = 10,\ N_0 = 5$



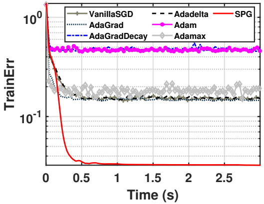
(e) $N = 100,\ N_1 = 10,\ N_0 = 5$



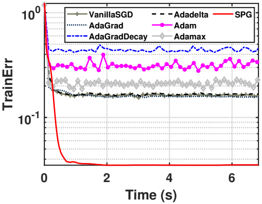
(f) $N = 150,\ N_1 = 20,\ N_0 = 10$

中国科学院 CHINESE ACADEMY OF SCIENCES

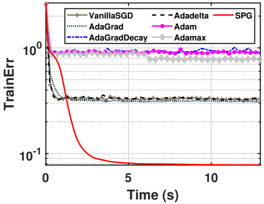
在均匀分布数据集上的数值变化比较



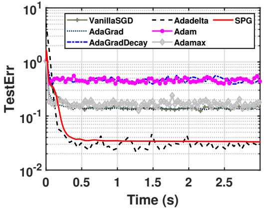
(a) $N = 75,\ N_1 = 10,\ N_0 = 5$



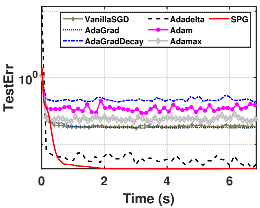
(b) $N = 100,\ N_1 = 10,\ N_0 = 5$



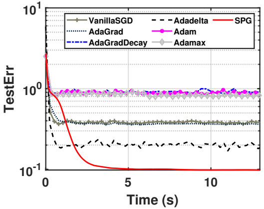
(c) $N = 150,\ N_1 = 20,\ N_0 = 10$



(d) $N = 75,\ N_1 = 10,\ N_0 = 5$



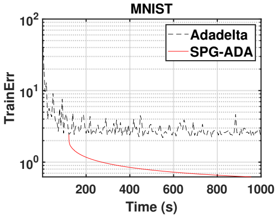
(e) $N = 100,\ N_1 = 10,\ N_0 = 5$



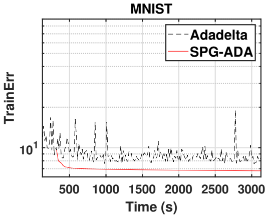
(f) $N = 150,\ N_1 = 20,\ N_0 = 10$

在 MNIST[LeCun et al. 1998]上的数值变化比较

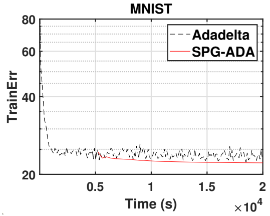
SPG-ADA: Adadelta (**预训练** 1000 **迭代步**) + SPG



(a) $N = 100,\ N_1 = 500$



(b) $N = 1000,\ N_1 = 1000$

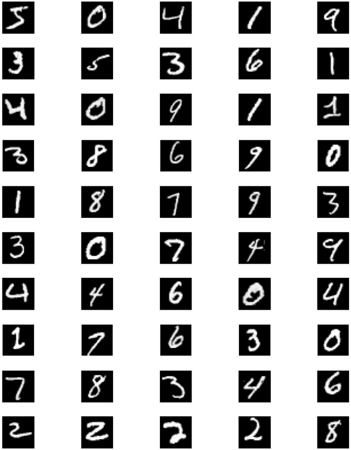


(c) $N = 10000,\ N_1 = 2000$

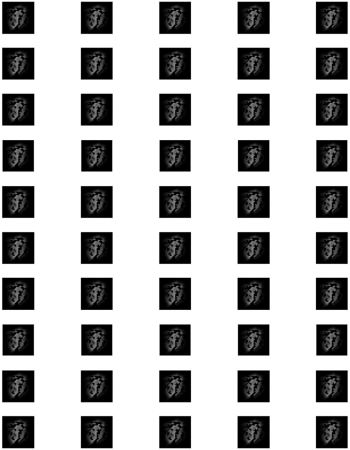
- SPG-ADA **略优于** Adadelta

中国科学院
CHINESE ACADEMY OF SCIENCES

在 MNIST 上的重构结果



(a) SPG



(b) Adam

工作三：一类训练稀疏 leaky ReLU 网络的非精确增广拉格朗日算法

深度神经网络模型

$$\min_{w,b} \frac{1}{N}\sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n+b_1)+\cdots)+b_L)-y_n\|^2+\mathcal{R}(w)$$

- 权重参数： $w=(\mathrm{vec}(W_1)^\top,\ldots,\mathrm{vec}(W_L)^\top)^\top\in\mathbb{R}^{N_{L+1}}$, 维数 $N_{L+1}=\sum_{\ell=1}^{L}N_\ell N_{\ell-1}$
- 偏差参数： $b=(b_1^\top,\ldots,b_L^\top)^\top\in\mathbb{R}^{N_{L+2}}$, 维数 $N_{L+2}=\sum_{\ell=1}^{L}N_\ell$
- leaky ReLU： $\sigma=\max\{z,\alpha z\}$, $0<\alpha<1$
- 数据矩阵： $X\in\mathbb{R}^{N_0\times N}$, 标签矩阵： $Y=(y_1,\ldots,y_N)\in\mathbb{R}^{N_L\times N}$
- 层数 L, 隐藏层单元数： $N_1,\ldots,N_{L-1}$
- $l_{2,1}$ 正则项： $\mathcal{R}(w):=\lambda_w\sum_{\ell=1}^{L}\|W_\ell\|_{2,1}=\lambda_w\sum_{\ell=1}^{L}\sum_{j=1}^{N_{\ell-1}}\|(W_\ell)_{\cdot,j}\|$, $\lambda_w>0$

为什么用 leaky ReLU? 良好的理论性质, 较好的数值结果
[Maas-Hannun-Ng 2013; Pedamonti 2018; ...]

为什么用 $l_{2,1}$ 正则项? 相较于 l_2 正则与 l_1 正则数值表现更好; SGD 训练时间更短[Zhou-Jin-Hoi 2010; Scardapane et al. 2017; Yoon-Hwang 2017; Hoefler et al. 2021; ...]

深度神经网络模型

$$\min_{w,b} \frac{1}{N}\sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n+b_1)+\cdots)+b_L)-y_n\|^2+\mathcal{R}(w)$$

- 权重参数： $w=(\text{vec}(W_1)^\top,\ldots,\text{vec}(W_L)^\top)^\top\in\mathbb{R}^{N_{L+1}}$，维数 $N_{L+1}=\sum_{\ell=1}^L N_\ell N_{\ell-1}$
- 偏差参数： $b=(b_1^\top,\ldots,b_L^\top)^\top\in\mathbb{R}^{N_{L+2}}$，维数 $N_{L+2}=\sum_{\ell=1}^L N_\ell$
- leaky ReLU： $\sigma=\max\{z,\alpha z\}$，$0<\alpha<1$
- 数据矩阵： $X\in\mathbb{R}^{N_0\times N}$，标签矩阵： $Y=(y_1,\ldots,y_N)\in\mathbb{R}^{N_L\times N}$
- 层数 L，隐藏层单元数： $N_1,\ldots,N_{L-1}$
- $l_{2,1}$ 正则项： $\mathcal{R}(w):=\lambda_w\sum_{\ell=1}^L\|W_\ell\|_{2,1}=\lambda_w\sum_{\ell=1}^L\sum_{j=1}^{N_{\ell-1}}\|(W_\ell)_{\cdot,j}\|$，$\lambda_w>0$

为什么用 leaky ReLU? **良好的理论性质，较好的数值结果**
[Maas-Hannun-Ng 2013; Pedamonti 2018; ...]

为什么用 $l_{2,1}$ 正则项? **相较于 l_2 正则与 l_1 正则数值表现更好；SGD 训练时间更短**[Zhou-Jin-Hoi 2010; Scardapane et al. 2017; Yoon-Hwang 2017; Hoefler et al. 2021; ...]

中国科学院 CHINESE ACADEMY OF SCIENCES

l_1 罚模型

辅助模型

$$\min_{w,b,v,u} \bar{O}(w,v) := \frac{1}{N}\sum_{n=1}^{N}\|v_{n,L}-y_n\|^2+\mathcal{R}(w)+\lambda_v\|v\|^2 \tag{P}$$
$$\text{s.t. } \sigma(u_{n,\ell})-v_{n,\ell}=0,\ u_{n,\ell}-(W_L v_{n,L-1}+b_L)=0,\ n\in[N],\ \ell\in[L]$$

- $v := (v_{1,1}^\top, v_{2,1}^\top, \ldots, v_{1,L}^\top, v_{2,L}^\top, \ldots, v_{N,L}^\top)^\top \in \mathbb{R}^m$
- $u = (u_{1,1}^\top, u_{2,1}^\top, \ldots, u_{1,L}^\top, u_{2,L}^\top, \ldots, u_{N,L}^\top)^\top \in \mathbb{R}^m$
- 辅助变量维数 $m = NN_{L+2}$
- 正则项：$\lambda_v\|v\|^2, \lambda_v > 0$：控制 $\|v\|$ 的大小
- (P) 的可行集可表示为 $v-\sigma(u)=0, u=\Psi(v)w+Ab$
  线性算子：$\Psi(v):\mathbb{R}^m\mapsto\mathbb{R}^{m\times N_{L+1}}$，矩阵：$A\in\mathbb{R}^{m\times N_{L+2}}$

l_1 罚模型：$\beta=(\beta_1 e_{NN_1}^\top,\ldots,\beta_L e_{NN_L}^\top)^\top\in\mathbb{R}_+^m$

$$\min_{w,b,v,u} O(w,v,u)=\bar{O}(w,v)+\beta^\top(v-\sigma(u)) \tag{PP}$$
$$\text{s.t. } v-u\ge 0, v-\alpha u\ge 0,\ u=\Psi(v)w+Ab$$

l_1 罚模型

辅助模型

$$\min_{w,b,v,u} \bar{O}(w,v) := \frac{1}{N}\sum_{n=1}^{N}\|v_{n,L}-y_n\|^2 + \mathcal{R}(w) + \lambda_v\|v\|^2 \tag{P}$$

$$\text{s.t.}\ \sigma(u_{n,\ell}) - v_{n,\ell} = 0,\ u_{n,\ell} - (W_L v_{n,L-1} + b_L) = 0,\ n \in [N],\ \ell \in [L]$$

- $v := (v_{1,1}^\top, v_{2,1}^\top, \ldots, v_{1,L}^\top, v_{2,L}^\top, \ldots, v_{N,L}^\top)^\top \in \mathbb{R}^m$
- $u = (u_{1,1}^\top, u_{2,1}^\top, \ldots, u_{1,L}^\top, u_{2,L}^\top, \ldots, u_{N,L}^\top)^\top \in \mathbb{R}^m$
- 辅助变量维数 $m = NN_{L+2}$
- 正则项：$\lambda_v\|v\|^2, \lambda_v > 0$：控制 $\|v\|$ 的大小
- (P) 的可行集可表示为 $v - \sigma(u) = 0, u = \Psi(v)w + Ab$
  线性算子：$\Psi(v) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times N_{L+1}}$，矩阵：$A \in \mathbb{R}^{m\times N_{L+2}}$

l_1 罚模型：$\beta = (\beta_1 e_{NN_1}^\top, \ldots, \beta_L e_{NN_L}^\top)^\top \in \mathbb{R}_+^m$

$$\min_{w,b,v,u} O(w,v,u) = \bar{O}(w,v) + \beta^\top(v - \sigma(u)) \tag{PP}$$

$$\text{s.t.}\ \ v - u \geq 0, v - \alpha u \geq 0,\ u = \Psi(v)w + Ab$$

为什么不用 SPG 算法求解 (PP)？

- 变量维数更高 + **非光滑正则项**
- 约束非线性 ⇒ **约束规范性条件不一定成立**
- ***C*-稳定点局限性**

C-稳定点局限性

考虑
$$\min_{w_1\in\mathbb{R},w_2\in\mathbb{R},b_1\in\mathbb{R},b_2\in\mathbb{R}} f(w_1,w_2,b_1,b_2) := \left((w_2\sigma(w_1+b_1)+b_2)+1\right)^2+\left((w_2\sigma(2w_1+b_1)+b_2)-1\right)^2. \tag{3}$$

令 $w_2^*=1,\ b_1^*=0,\ w_1^*=0,\ b_2^*=0$, 有

$$\partial^c f(w_1^*,w_2^*,b_1^*,b_2^*)=\{(t,0,s,0)^\top : t\in[2\alpha-4,2-4\alpha], s\in[-2+2\alpha,2-2\alpha]\},$$
$$\partial\left(f(w_1^*,w_2^*,b_1^*,b_2^*)\right)$$
$$=\{(-2\alpha,0,0,0)^\top,(2\alpha-4,0,2\alpha-2,0)^\top,(2-4\alpha,0,2-2\alpha,0)^\top,(-2,0,0,0)^\top\},$$
$$f(w_1^*+\epsilon,w_2^*,b_1^*,b_2^*)=5\epsilon^2-2\epsilon+2<2=f(w_1^*,w_2^*,b_1^*,b_2^*),\ \epsilon \text{ 是一个小正数}.$$

对于 $0<\alpha<\frac{1}{2}$, $(w_1^*,w_2^*,b_1^*,b_2^*)$ 是 C-稳定点, 而不是 l-稳定点、局部极小值点.

中国科学院 CHINESE ACADEMY OF SCIENCES

为什么不用 SPG 算法求解 (PP)？

- 变量维数更高 + **非光滑正则项**
- 约束非线性 ⇒ **约束规范性条件不一定成立**
- **C-稳定点局限性**

C-稳定点局限性

考虑

$$\min_{w_1\in\mathbb{R},w_2\in\mathbb{R},b_1\in\mathbb{R},b_2\in\mathbb{R}} f(w_1,w_2,b_1,b_2) := \\ ((w_2\sigma(w_1+b_1)+b_2)+1)^2+((w_2\sigma(2w_1+b_1)+b_2)-1)^2. \tag{3}$$

令 $w_2^*=1$, $b_1^*=0$, $w_1^*=0$, $b_2^*=0$, 有

$$\partial^c f(w_1^*,w_2^*,b_1^*,b_2^*)=\{(t,0,s,0)^\top : t\in[2\alpha-4,2-4\alpha], s\in[-2+2\alpha,2-2\alpha]\},$$

$$\partial(f(w_1^*,w_2^*,b_1^*,b_2^*))$$

$$=\{(-2\alpha,0,0,0)^\top,(2\alpha-4,0,2\alpha-2,0)^\top,(2-4\alpha,0,2-2\alpha,0)^\top,(-2,0,0,0)^\top\},$$

$$f(w_1^*+\epsilon,w_2^*,b_1^*,b_2^*)=5\epsilon^2-2\epsilon+2<2=f(w_1^*,w_2^*,b_1^*,b_2^*),\ \epsilon \text{ 是一个小正数}.$$

对于 $0<\alpha<\frac{1}{2}$, $(w_1^*,w_2^*,b_1^*,b_2^*)$ 是 C-稳定点，而不是 l-稳定点、局部极小值点.

(P) 的一阶稳定点

由于 $v-\sigma(u)=0$ **可被表示为互补约束** $v-u\geq 0, (v-u)(v-\alpha u)=0, v-\alpha u\geq 0$,
我们称问题(P)的可行点 (w^*, b^*, v^*, u^*) 是其MPCC W-**稳定点**[Scheel-Scholtes 2000; Guo-Chen 2021], 若存在 $\mu^1\in\mathbb{R}^m$, $\mu^2\in\mathbb{R}^m$ 和 $\xi\in\mathbb{R}^m$ 使得下式成立

$$
\begin{aligned}
&0=\nabla_w\bar{O}(w^*,v^*)+\Psi(v^*)^\top\xi,\quad 0=A^\top\xi\\
&0=\nabla_v\bar{O}(w^*,v^*)-\mu^1-\mu^2+\nabla_v\xi^\top(u^*-\Psi(v^*)w^*)\\
&0=\mu^1+\alpha\mu^2+\xi\\
&(\mu^1)^\top(v^*-u^*)=0,\ (\mu^2)^\top(v^*-\alpha u^*)=0
\end{aligned}
$$

MPCC W-稳定点 + $\mu_i^1\mu_i^2\geq 0, \forall i: u_i^*=v_i^*=0 \Rightarrow$ MPCC C-稳定点

引理 4

无非零异常乘子规范性条件[a]对问题(P)的约束函数成立.

[a] [Ye-Zhang 2013]

- $\Rightarrow$ 问题(P)的任一局部极小值点是其 l-稳定点 $\Rightarrow$ MPCC C-**稳定点**

(P) 的一阶稳定点

由于 $v-\sigma(u)=0$ **可被表示为互补约束** $v-u\geq 0, (v-u)(v-\alpha u)=0, v-\alpha u\geq 0$, 我们称问题(P)的可行点 (w^*,b^*,v^*,u^*) 是其MPCC W-**稳定点**[Scheel-Scholtes 2000; Guo-Chen 2021], 若存在 $\mu^1\in\mathbb{R}^m$, $\mu^2\in\mathbb{R}^m$ 和 $\xi\in\mathbb{R}^m$ 使得下式成立

$$
\begin{aligned}
&0=\nabla_w\bar{O}(w^*,v^*)+\Psi(v^*)^\top\xi, \quad 0=A^\top\xi\\
&0=\nabla_v\bar{O}(w^*,v^*)-\mu^1-\mu^2+\nabla_v\xi^\top(u^*-\Psi(v^*)w^*)\\
&0=\mu^1+\alpha\mu^2+\xi\\
&(\mu^1)^\top(v^*-u^*)=0,\ (\mu^2)^\top(v^*-\alpha u^*)=0
\end{aligned}
$$

MPCC W-稳定点 $+\ \mu_i^1\mu_i^2\geq 0, \forall i: u_i^*=v_i^*=0 \Rightarrow$ MPCC C-稳定点

引理 4

无非零异常乘子规范性条件[a]对问题(P)的约束函数成立.

[a] [Ye-Zhang 2013]

- $\Rightarrow$ 问题(P)的任一局部极小值点是其 l-稳定点 $\Rightarrow$ MPCC C-**稳定点**

(PP) 的一阶稳定点

我们称问题(PP)的可行点 (w^*, b^*, v^*, u^*) 是其 KKT 点, 若存在 $\mu^1 \in \mathbb{R}^m_+$, $\mu^2 \in \mathbb{R}^m_+$ 和 $\xi \in \mathbb{R}^m$ 使得下式成立

$$0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^\top \xi, \quad 0 = A^\top \xi$$

$$0 = \nabla_v \bar{O}(w^*, v^*) + \beta - \mu^1 - \mu^2 + \nabla_v \xi^\top (u^* - \Psi(v^*) w^*)$$

$$0 \in \partial_u(-\beta^\top \sigma(u^*)) + \mu^1 + \alpha\mu^2 + \xi$$

$$(\mu^1)^\top (v^* - u^*) = 0, \ (\mu^2)^\top (v^* - \alpha u^*) = 0$$

引理 5

Mangasarian-Fromovitz 约束规范性条件[a]对问题(PP)的约束函数成立.

[a] [Mangasarian 1994]

- $\Rightarrow$ (w^*, b^*, v^*, u^*) 是问题(PP)的一个 l-稳定点, 当且仅当 (w^*, b^*, v^*, u^*) 是问题(PP)的一个 KKT 点
- $\Rightarrow$ 问题(PP)的任一局部极小值点是其 l-稳定点
- $\Rightarrow$ 问题(PP)的任一 l-稳定点 $\Rightarrow$ 问题(P)的可行点 $\Rightarrow$ MPCC W-稳定点

中國科學院 CHINESE ACADEMY OF SCIENCES

(PP) 的一阶稳定点

我们称问题(PP)的可行点 (w^*, b^*, v^*, u^*) 是其 KKT 点, 若存在 $\mu^1 \in \mathbb{R}^m_+$, $\mu^2 \in \mathbb{R}^m_+$ 和 $\xi \in \mathbb{R}^m$ 使得下式成立

$$
\begin{aligned}
&0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^\top \xi, \quad 0 = A^\top \xi \\
&0 = \nabla_v \bar{O}(w^*, v^*) + \beta - \mu^1 - \mu^2 + \nabla_v \xi^\top (u^* - \Psi(v^*) w^*) \\
&0 \in \partial_u(-\beta^\top \sigma(u^*)) + \mu^1 + \alpha\mu^2 + \xi \\
&(\mu^1)^\top (v^* - u^*) = 0, \;\; (\mu^2)^\top (v^* - \alpha u^*) = 0
\end{aligned}
$$

引理 5

Mangasarian-Fromovitz 约束规范性条件[a]对问题(PP)的约束函数成立.

[a] [Mangasarian 1994]

- $\Rightarrow$ (w^*, b^*, v^*, u^*) 是问题(PP)的一个 l-稳定点, 当且仅当 (w^*, b^*, v^*, u^*) 是问题(PP)的一个 KKT 点
- $\Rightarrow$ 问题(PP)的任一局部极小值点是其 l-稳定点
- $\Rightarrow$ 问题(PP)的任一 l-稳定点 $\Rightarrow$ 问题(P)的可行点 $\Rightarrow$ MPCC W-稳定点

中国科学院 CHINESE ACADEMY OF SCIENCES

模型分析

令 $\theta > \frac{1}{N}\|Y\|_F^2$

$$\Omega_\theta = \{(w,b,v,u) : v-u \geq 0, v-\alpha u \geq 0,\ u = \Psi(v)w + Ab, \mathcal{O}(w,v,u) \leq \theta\}$$

定理 6 (解集有界性)

集合 Ω_θ 是有界的. 进一步地, 问题(PP) 的解集是非空且有界的.

精确罚性

(PP):	全局 (局部) 极小值点	d-稳定点		l-稳定点 $\Leftrightarrow$ KKT 点
	$\Downarrow\Uparrow$	$\Downarrow$	一些条件 $\Downarrow$	$\Downarrow$
(P):	全局 (局部) 极小值点	d-稳定点	MPCC C-稳定点	MPCC W-稳定点

- $\beta_\ell > LL_{\bar{O}} \max\{\theta_w, 1\}^L + 2\sum_{j=\ell+1}^{L} \beta_j \theta_w \max\{\theta_w, 1\}^{j-\ell-1}$, $L_{\bar{O}}$ 为 $\bar{O}$ 在 Ω_θ 上的 Lipschitz 常数
- 相应点的函数值小于 θ

推广到 ReLU 网络

当 $\alpha = 0$ 时，问题(PP) 的解集可能无界

$\Downarrow$

构建“有界”闭集，求解

$$\begin{aligned} \min_{w,b,v,u} \ & \mathcal{O}(w,v,u) \\ \text{s.t.} \ & v-u \geq 0, v-\alpha u \geq 0,\ u = \Psi(v)w + Ab \\ & b \geq -e_{N_{L+2}} N_{L+2}(\theta_w + \theta_v) \end{aligned} \tag{PP$_\mathrm{b}$}$$

- $(\mathrm{PP_b})$ 的解集非空且有界
- $(\mathrm{PP_b})$ 的任一 l-稳定点是(P)的 MPCC W-稳定点
- $(\mathrm{PP_b})$ 的任一全局（局部）极小值点是(P)的全局（局部）极小值点
- 后续算法也可被用于求解 ReLU 网络

ReLU 网络下，问题 $(\mathrm{PP_b})$ 没有解集有界性！

非精确增广拉格朗日函数法[Lu-Zhang 2012; Chen et al. 2017]

增广拉格朗日函数

$$\mathcal{L}_\rho(w,b,v,u;\xi) := \mathcal{O}(w,v,u) + \langle \xi, u - \Psi(v)w - Ab \rangle + \frac{\rho}{2}\|u - \Psi(v)w - Ab\|^2$$

- 增广拉格朗日罚参数：$\rho \in \mathbb{R}_+$
- 增广拉格朗日乘子：$\xi \in \mathbb{R}^m$

非精确增广拉格朗日函数法子问题

$$\min_{(w,b,v,u):v \geq u,\, v \geq \alpha u} \mathcal{L}_\rho(w,b,v,u;\xi) \tag{4}$$

- IALAM：IALM 框架 + 交替下降算法求解子问题

算法框架

一类求解问题(PP)的非精确增广拉格朗日算法 (IALM 框架)

1. 输入：初始点 $(w^{(0)}, b^{(0)}, v^{(0)}, u^{(0)}) \in \Omega_\theta$, 参数 $\rho^{(0)} > 0$, $\eta_1 \in (0,1)$, $\eta_2, \eta_3 > 1$, $\xi^{(0)} \in \mathbb{R}^m$, $\gamma \in \mathbb{N}_+$. 令 $k := 1$
2. 令 $(\xi, \rho) = (\xi^{(k-1)}, \rho^{(k-1)})$, 非精确求解子问题(4)
3. 通过 $\xi^{(k)} := \xi^{(k-1)} + \rho^{(k-1)}(u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)})$ 更新乘子
4. 若 $k \le \gamma$, 取 $\rho^{(k)} = \rho^{(k-1)}$. 若 $k > \gamma$ 以及
$$\left\|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\right\| \le \eta_1 \max_{t=k-\gamma,\dots,k-1} \left\|u^{(t)} - \Psi(v^{(t)})w^{(t)} - Ab^{(t)}\right\|$$
则令 $\rho^{(k)} = \rho^{(k-1)}$. 否则, 令
$$\rho^{(k)} = \max\left\{\rho^{(k-1)}/\eta_2, \|\xi^{(k)}\|^{1+\eta_3}\right\}$$
5. 令 $k := k+1$, 若终止条件未满足, 回到第 2 步
6. 输出：$(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})$

- 可行误差非单调下降

交替下降算法设计思路

步一： $(w^{(J)}, b^{(J)}) \to (w^{(J+1)}, b^{(J+1)})$

$$(w^{(J+1)}, b^{(J+1)}) := \arg\min_{w,b} \mathcal{L}_\rho(w, b, v^{(J)}, u^{(J)}; \xi)^5 \tag{5}$$

步二： $(v^{(J)}, u^{(J)}) \to (v^{(J+1)}, u^{(J+1)})$

$$(v^{(J+1)}, u^{(J+1)}) := \arg \min_{(v,u): v \geq u, v \geq \alpha u} \mathcal{L}_\rho(w^{(J+1)}, b^{(J+1)}, v, u; \xi) + \mathcal{P}(u, v; u^{(J)}, v^{(J)}, \tau^{(J)}) \tag{6}$$

$$\mathcal{P}(u, v; u^{(J)}, v^{(J)}, \tau^{(J)}) := \frac{1}{2} \sum_{n=1}^{N} \sum_{\ell=2}^{L} \left\| \begin{pmatrix} v_{n,\ell-1} \\ u_{n,\ell} \end{pmatrix} - \begin{pmatrix} v_{n,\ell-1}^{(J)} \\ u_{n,\ell}^{(J)} \end{pmatrix} \right\|_{S_\ell^{(J)}}^2 + \frac{\tau_1}{2} \sum_{n=1}^{N} \|u_{n,1} - u_{n,1}^{(J)}\|^2$$

- $\tau_1 > 0,\ \tau^{(J)} := (\tau_2^{(J)}, \ldots, \tau_L^{(J)})^\top \in \mathbb{R}^{L-1}$
- $S_\ell^{(J)} := \tau_\ell^{(J)} I_{N_\ell + N_{\ell-1}} - \rho \begin{bmatrix} -W_\ell^{(J+1)} & I_{N_\ell} \end{bmatrix}^\top \begin{bmatrix} -W_\ell^{(J+1)} & I_{N_\ell} \end{bmatrix} \succeq \tau_1 I_{N_\ell + N_{\ell-1}}$
- $\tau_\ell^{(J)} := \rho \left\| \begin{bmatrix} -W_\ell^{(J+1)} & I_{N_\ell} \end{bmatrix} \right\|^2 + \tau_1$

---

[5] 可被投影梯度法求解[Dai-Fletcher 2005]

算法框架

求解问题 (4)的一类交替下降算法

1. 输入：矩阵 A, 向量 ξ 以及参数 $\rho > 0$, 初始点 $(w^{(0)}, b^{(0)}, v^{(0)}, u^{(0)})$. 令 $\tau_1 > 0$, $\jmath = 0$
2. 通过求解问题 (5) 得到 $(w^{(\jmath+1)}, b^{(\jmath+1)})$
3. 通过求解问题 (6)得到 $(u^{(\jmath+1)}, v^{(\jmath+1)})$
4. 令 $\jmath := \jmath + 1$, 若终止条件未满足, 回到第 2 步
5. 输出：$(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})$

定理 7

令 $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ 为交替下降算法生成的序列, 则其任一聚点 (w^*, b^*, v^*, u^*) 是问题 (4)的一个 KKT 点.

收敛性分析

定理 8

令 $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ 为算法 IALAM 生成的序列, 则下述命题成立.

(a) $\liminf_{k\to\infty} \|u^{(k+1)} - \Psi(v^{(k+1)})w^{(k+1)} - Ab^{(k+1)}\| = 0$. 进一步地, 序列 $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ 至少有一个聚点, 且 $\liminf_{k\to\infty} \mathrm{dist}((w^*, b^*, v^*, u^*), \mathcal{Z}^*) = 0$, 其中 $\mathcal{Z}^*$ 是问题(PP)的 KKT 点集.

(b) 若 $\gamma = 1$ 或 $\lim_{k\to\infty} \rho^{(k)}$ 存在, 则有 $\lim_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$. 进一步地, 序列 $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ 至少有一个聚点, 且其任一聚点 (w^*, b^*, v^*, u^*) 是问题(PP)的一个 KKT 点.

- **理论贡献**: 与已有非精确增广拉格朗日算法不同[Lu-Zhang 2012; Chen et al. 2017], 我们证明了**聚点存在**
- **算法可拓展性**: IALAM 算法可求解自编码问题以及带 l_1, l_2 和 $l_{2,1}$ 正则的 (leaky) ReLU 网络, 区别仅在交替下降算法的第 2 步

中国科学院 CHINESE ACADEMY OF SCIENCES

默认设置

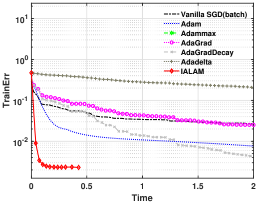
- 停机准则：$\epsilon_k < 10^{-6}$ 或 $\rho^{(k)} > 10^3 \rho^{(0)}$
- 初始化：令 $W_\ell^{(0)} = \mathrm{randn}(N_\ell, N_\ell - 1)/N$, $b^{(0)} = 0$, $u_{n,\ell}^{(0)} = W_\ell^{(0)} v_{n,\ell-1}^{(0)}$, $v_{n,\ell}^{(0)} = \sigma(u_{n,\ell}^{(0)})$
- 测试问题：人工合成数据集上的函数拟合问题、MNIST 数据集上的分类问题
- $N_{\text{test}} = \lceil N/5 \rceil$

衡量标准

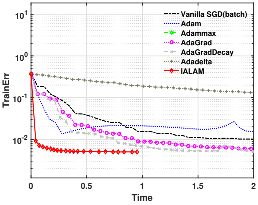
- TrainErr $= \frac{1}{N} \sum_{n=1}^{N} \|\sigma(W_L \sigma(\cdots \sigma(W_1 x_n + b_1) + b_2 \cdots) + b_L) - y_n\|^2$
- TestErr $= \frac{1}{N} \sum_{n=N+1}^{N+N_{\text{test}}} \|\sigma(W_L \sigma(\cdots \sigma(W_1 x_n + b_1) + b_2 \cdots) + b_L) - y_n\|^2$
- FeasVi $= \frac{1}{N} \sum_{n=1}^{N} \sum_{\ell=1}^{L} \|v_{n,\ell} - \sigma(u_{n,\ell})\|^2 + \frac{1}{N} \sum_{n=1}^{N} \sum_{\ell=1}^{L} \|u_{n,\ell} - (W_\ell v_{n,\ell-1} + b_\ell)\|^2$
- Column Sparse Ratio: 在所有的 W_ℓ 共 $\sum_{\ell=0}^{L-1} N_\ell$ 列中, l_2 模的值低于容忍度 ϵ 的列的比例
- 训练集精确度：Accuracy
- Time (s)

中國科學院 CHINESE ACADEMY OF SCIENCES

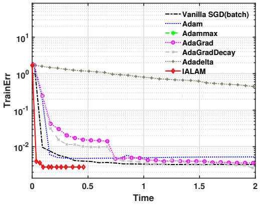
与 SGD 类算法在人工合成数据集上的数值变化比较



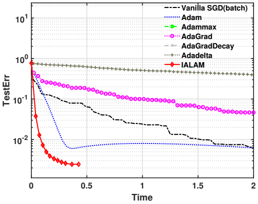
(a) $L = 2,\ N_1 = 10$



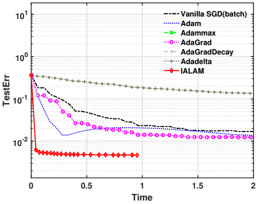
(b) $L = 3,\ N_1 = N_2 = 5$



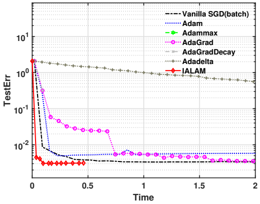
(c) $L = 4,\ N_1 = 4, N_2 = N_3 = 3$



(d) $L = 2,\ N_1 = 10$

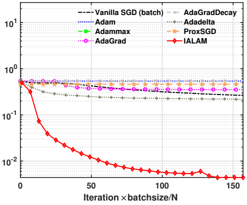


(e) $L = 3,\ N_1 = N_2 = 5$

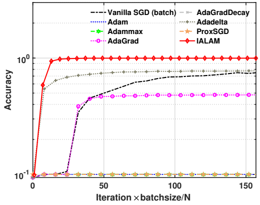


(f) $L = 4,\ N_1 = 4, N_2 = N_3 = 3$

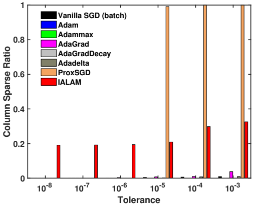
与 SGD 类算法在 MNIST 上的数值变化比较



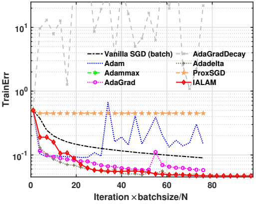
(a) TrainErr



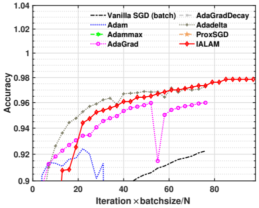
(b) Accuracy



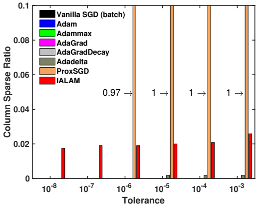
(c) Column Sparse Ratio



(d) TrainErr



(e) Accuracy

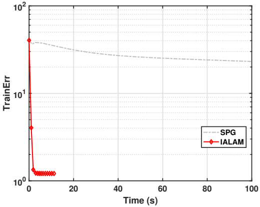


(f) Column Sparse Ratio

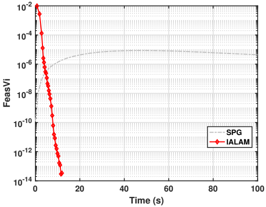
(a)–(c)： $N = 1000,\ N_1 = 100,\ N_2 = 50,\ L = 3$

(d)–(f)： $N = 60000,\ N_1 = 200,\ N_2 = 100,\ L = 3$

与 SPG 算法在 (AE) 与 MNIST 上的数值变化比较



(a) TrainErr



(b) FeasVi

- $N = 1000$

工作二与工作三的区别

- 数值：IALAM 数值效果优于 SPG，且能求解多层神经网络
- 理论：
  - 在求解两层神经网络（自编码）问题时，SPG 能找到原问题的 C-稳定点
  - IALAM 能找到原问题的 MPCC C-稳定点

总结与展望

总结

半监督聚类问题

- 新模型
  - **无需精确类的数目**
  - **无需后处理**
  - 同时处理强标签标注和弱标签标注**两类半监督信息**
- 有限收敛算法 CO-BCD

非光滑自编码问题 (R)

- 新模型 (LRP) 有有界解集
- **精确罚**：C-稳定点
- 快速算法 SPG：稳定收敛到 (LRP) 和 (R) 的**C-稳定点**
- 文章已被 SIAM Journal on Optimization 接收

总结（续）

非光滑深度神经网络 (P)

- 有界解新模型 (PP)
- **精确罚**：
  - 局部极小值点集
  - (PP) 的任一 KKT 点 $\Rightarrow$ 问题 (P) 的一个 MPCC C-稳定点
- 非精确IALAM **算法**：稳定收敛到 (PP) 问题的一个KKT **点**
  - 可行误差非单调下降
  - 无需假设聚点存在

展望

- 利用与半监督谱聚类连续模型类似的构造方式，设计分式模型 (无 β)

$$\begin{aligned} \min_{Z} \ & \frac{\operatorname{tr}\left[H^\top \mathcal{L}(S \circ Z) H\right] + d}{\operatorname{tr}(SZ)} \\ \text{s. t. } & Z \in \mathcal{S}_S^N \cap \{0,1\}^{N \times N} \\ & H^\top H = I_d \end{aligned}$$

  - 研究理论性质
  - 设计快速算法

- 利用 SPG 和 IALAM 求解半监督聚类

- 开发 IALAM 算法包并求解实际问题

中国科学院
CHINESE ACADEMY OF SCIENCES

谢谢各位专家！

liuwei175@lsec.cc.ac.cn