Sparse Representation of The Signal: A Brief Introduction and Some Discussion

inlmouse
Glasssix Research
March 20, 2020
inlmouse@glasssix.com

Abstract

使用少量基本信号的线性组合表示一目标信号，称为信号的稀疏表示。信号的稀疏表示是过去近 25 年来信号处理界一个非常引人关注的研究领域（现在凉透了），众多研究论文和专题研讨会表明了该领域曾经的的蓬勃发展。信号稀疏表示的目的就是在给定的超完备字典中用尽可能少的原子来表示信号，可以获得信号更为简洁的表示方式，从而使我们更容易地获取信号中所蕴含的信息，更方便进一步对信号进行加工处理，如压缩、编码等 [24]。

1 稀疏向量与稀疏表示 (Sparse Vector and Sparse Representation)

1.1 Preliminary Knowledge

对矩阵 $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ 有以下是常用的 7 种范数[1]：

1. m_1 范数：$||\boldsymbol{A}||_{m_1} = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$
2. F 范数[2] ：$||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{tr(\boldsymbol{A}^H\boldsymbol{A})}$
3. M 范数/最大范数：$||\boldsymbol{A}||_M = \max\{m,n\}\max_{i,j}|a_{ij}|$
4. G 范数/几何平均范数：$||\boldsymbol{A}||_G = \sqrt{mn}\max_{i,j}|a_{ij}|$
5. 1 范数/列和范数[3]：$||\boldsymbol{A}||_1 = \max_j\sum_{i=1}^{m}|a_{ij}|$
6. 2 范数/谱范数：$||\boldsymbol{A}||_2 = \sqrt{\boldsymbol{A}^H\boldsymbol{A}\text{的最大特征值}}$
7. ∞ 范数/行和范数：$||\boldsymbol{A}||_\infty = \max_i\sum_{j=1}^{n}|a_{ij}|$

1.2 Sparse Representation

一个含有大多数零元素的向量或者矩阵成为稀疏向量 (sparse vector) 或者稀疏矩阵 (sparse matrix)。也即是：给定 $K \in \mathbb{N}^+, \|\boldsymbol{x}\|_0 \le K$。

[1] 矩阵的范数定义除开满足非负性，齐次性和三角不等式外，还需满足**相容性**:

$$\forall \boldsymbol{A},\boldsymbol{B} \in \mathbb{C}^{n\times n}, \exists ||\boldsymbol{AB}|| \le ||\boldsymbol{A}||\cdot||\boldsymbol{B}||$$

[2] Here, $\boldsymbol{A}^H$ means **Hermitian Matrix**(埃尔米特/厄米/自伴随矩阵): $\boldsymbol{A}^H = (\bar{a}_{ji})_{n\times n}$.

[3] 5∼7 的范数并不是按照这里给出的公式定义的，而是从属于某向量范数 $||\cdot||_v$ 导出的矩阵范数:$||\boldsymbol{A}|| = \max_{\mathbf{x}\neq\mathbf{0}}\frac{||\boldsymbol{Ax}||_v}{||\boldsymbol{x}||_v}$，简称**导出范数/从属范数**，且满足：$||\boldsymbol{E}|| = 1$.

给定一个向量 $\boldsymbol{x} \in \mathbb{R}^n$，可以定义如下稀疏测度 (sparse measure)：

$$\text{sparseness}(\boldsymbol{x}) = \frac{\sqrt{n} - \|\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_2}{\sqrt{n}-1} \tag{1}$$

信号向量 $\boldsymbol{y} \in \mathbb{R}^m$ 最多可以分解为 m 个正交基 $\boldsymbol{g}_k \in \mathbb{R}^m$，这些正交基的集合成为完备正交基 (complete orthogonal basis)。此时，信号分解

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{c} = \sum_{i=1}^{m} c_i \boldsymbol{g}_i \tag{2}$$

中的系数向量 $\boldsymbol{c}$ 一定是非稀疏的。

若将信号向量 $\boldsymbol{y} \in \mathbb{R}^m$ 分解为 n 个 m 维向量 $\boldsymbol{a}_i \in \mathbb{R}$（其中 $n > m$）的线性组合：

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i \tag{3}$$

则 $\boldsymbol{a}_i \in \mathbb{R}$ 不可能是正交基的集合。为区别于基，这些列向量通常称为**原子** (atom) 或框架。由于原子数大于向量空间的维数，所以称这些原子的集合是过完备的 (overcompete)。过完备的原子组成的矩阵 $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n]$ 称为字典或者库 (dictionary)。

对于字典 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$，可以做如下假设：

1. $n > m$;
2. $rank(\boldsymbol{A}) = m$;
3. $\|\boldsymbol{a}_i\|_2 = 1, i = 1, \cdots, n$;

信号过完备分解式3为欠定方程，存在无穷多组解向量 $\boldsymbol{x}$。求解这种欠定方程有两种常用方法：

1. 古典方法（求最小 L_2 范数解），即是：

$$\min \|\boldsymbol{x}\|_2, s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{4}$$

   这种方式的优点是：有唯一解，其物理意义为最小能量解。然而由于这种解的每个元素通常为非零值，故不符合很多实际应用的稀疏表示要求。

2. 现代方法（求最小 L_0 范数解），即是：

$$\min \|\boldsymbol{x}\|_0, s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{5}$$

   这种方式的优点是：很多实际应用只选择一个稀疏解向量。然而在计算上难以处理。

假定观测向量存在加性误差或者噪声，最小 L_0 范数解为：

$$\min \|\boldsymbol{x}\|_0, s.t. \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \varepsilon \tag{6}$$

其中 ε 为很小的误差或者扰动。

当系数向量 $\boldsymbol{x}$ 是稀疏向量时，信号分解 $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ 称为（信号的）稀疏分解 (sparse decomposition)。其中字典矩阵 $\boldsymbol{A}$ 的列常称为解释变量 (explanatory variables)；向量 $\boldsymbol{y}$ 称为相应变量 (response variable) 或目标信号；$\boldsymbol{A}\boldsymbol{x}$ 称为相应的线性预测；$\boldsymbol{x}$ 可视为目标信号相对于字典 $\boldsymbol{A}$ 的一种表示。

因此，称式5是目标信号 $\boldsymbol{y}$ 相对于字典 $\boldsymbol{A}$ 的**稀疏表示** (sparse representation)，而式6称为目标信号的**稀疏逼近** (sparse approximation)。

稀疏表示属于线性求逆问题 (linera inverse problem)。在通信和信息论中，$\boldsymbol{A} \in \mathbb{R}^{m \times N}$ 和 $\boldsymbol{x} \in \mathbb{R}^N$ 分别代表编码矩阵和待发送的明文，观测向量 $\boldsymbol{y} \in \mathbb{R}^m$ 则称密文。线性求逆问题便成了解码问题：即如何从密文恢复明文。

2 人脸识别的稀疏表示 (Sparse Representation in Face Recognition)

我们考虑 close-set 的人脸识别应用：假定共有 c 类目标，每一目标的脸部图像已经被向量化编码（可以是直接矩阵拉直，也可以是通过 CNN 进行特征提取），表示为了 $m \times 1$ 的归一化列向量（通常我们这里的 m 为 512 或者 128）。于是第 i 类目标的 N_i 张训练图像即可表示成 $\boldsymbol{D}_i = [\boldsymbol{d}_{i,1}, \cdots, \boldsymbol{d}_{i,N_i}] \in \mathbb{R}^{m \times N}$。给定一个足够丰富的训练集 $\boldsymbol{D}_i$，则第 i 类目标的非训练集新图片 $\boldsymbol{y}$ 可以被表示为已知训练图像的一线性组合 $\boldsymbol{y} \approx \boldsymbol{D}_i \boldsymbol{\alpha}_i$，其中 $\boldsymbol{\alpha}_i$ 为系数向量。问题是：在实际应用中往往不知道新图像分属哪一类，而需要识别：判断该样本的属性。

于是我们已这 c 类目标的所有训练样本构造一个字典：

$$\boldsymbol{D} = [\boldsymbol{D}_1, \cdots, \boldsymbol{D}_c] = [\boldsymbol{d}_{1,1}, \cdots, \boldsymbol{d}_{1,N_1}, \cdots, \boldsymbol{d}_{c,1}, \cdots, \boldsymbol{d}_{c,N_c}] \in \mathbb{R}^{m \times N} \tag{7}$$

其中 $N = \sum_{i=1}^{c} N_i$。于是，待识别的人脸图像编码 $\boldsymbol{y}$ 可以表示为线性组合：

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\alpha}_0 = [\boldsymbol{d}_{1,1}, \cdots, \boldsymbol{d}_{1,N_1}, \cdots, \boldsymbol{d}_{c,1}, \cdots, \boldsymbol{d}_{c,N_c}] \begin{bmatrix} \boldsymbol{0}_{N_1} \\ \vdots \\ \boldsymbol{0}_{N_{i-1}} \\ \boldsymbol{\alpha}_i \\ \boldsymbol{0}_{N_{i+1}} \\ \vdots \\ \boldsymbol{0}_{N_c} \end{bmatrix} \tag{8}$$

现在，人脸识别变成了一个矩阵方程求解的问题或者线性求你问题：已知数据向量 $\boldsymbol{y}$ 和数据矩阵 $\boldsymbol{D}$，求矩阵方程 $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\alpha}_0$ 的解向量 $\boldsymbol{\alpha}_0$。需要注意的是，通常 $m < N$，因为方程欠定，有无穷多解，其中**最稀疏的解才是我们感兴趣的**。鉴于此，问题划归为式5的问题。

3 稀疏矩阵方程求解的优化理论 (Optimization Theory for Solving Sparse Matrix Equations)

3.1 L_1 范数最小化 (L_1 Norm Minimization)

L_1 范数最小化也称为 L_1 线性规划或者 L_1 范数正则化最小二乘。

直接求解优化问题 P0，必须筛选出系数向量 $\boldsymbol{x}$ 中所有可能的非零元素。这个方法是**不可跟踪的** (untractable) 或者 NP hard[4]的，因为搜索空间过于庞大。

向量 $\boldsymbol{x}$ 的非零元素指标集称为**支撑集**，记为 $\text{supp}(\boldsymbol{x}) = \{i : x_i \neq 0\}$，支撑集的长度即是 L_0 拟范数[5]：

$$\|\boldsymbol{x}\|_0 = |supp(\boldsymbol{x})| \tag{9}$$

K-稀疏向量的集合记为 $\Sigma_K = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_0 \leq K\}$。若 $\hat{\boldsymbol{x}} \in \Sigma_K$，则称向量 $\hat{\boldsymbol{x}}$ 是 $\boldsymbol{x}$ 的 K-项逼近或者 K-稀疏逼近。

一般地，称向量 $\hat{\boldsymbol{x}}$ 是 $\boldsymbol{x}$ 在 L_p 范数[6]下的 K-稀疏逼近，若：

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_p = \inf_{\boldsymbol{z} \in \Sigma_K} \|\boldsymbol{x} - \boldsymbol{z}\|_p \tag{10}$$

显然 L_0 是 L_p 范数范数的特殊形式：$\|\boldsymbol{x}\|_0 = \lim_{p \to 0} \|\boldsymbol{x}\|_p^p$。由于当且仅当 $p \geq 1$ 时 $\|\boldsymbol{x}\|_p$ 为凸函数，所以 L_1 范数时最接近于 L_0 拟范数的凸目标函数。于是从最优化角度讲，称 L_1 范数是 L_0 拟范数的凸松弛 [22]。因此，L_0 **拟范数最小化问题便可以转变为凸松弛的 L_1 范数最小化问题**：

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, s.t. \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{11}$$

---

[4]指所有 NP 问题都能在多项式时间复杂度内归约到的问题.

[5]L_0 范数不满足范数公理中的齐次性：$\|c\boldsymbol{x}\|_0 = |c|\|\boldsymbol{x}\|_0$，故严格来讲它是一种虚拟的范数，也称拟范数。

[6]$\forall p \in \mathbb{R}^+, \|\boldsymbol{x}\|_p = \left(\sum_{i \in supp(\boldsymbol{x})} |x_i|^p\right)^{1/p}$.

由于 $\|\boldsymbol{x}\|_1$ 是凸函数，并且约束等式 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$ 为一个仿射变换，因此这是一个凸优化问题。

存在观测噪声的情况下，等式约束可以松弛为不等式约束的最优化问题（L_1 最小化）：

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, s.t. \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\| \leq \varepsilon \tag{12}$$

L_1 范数下的最优化问题又称为基追踪 (base pursuit, BP)。这是一个二次约束线性规划问题 (quadratically constrained linear problem, QCLP)。

若 $\boldsymbol{x}_1$ 是 L_1 的解，$\boldsymbol{x}_0$ 是 L_0 的解，则有 [11]：

$$\|\boldsymbol{x}_1\|_1 \leq \|\boldsymbol{x}_0\|_1 \tag{13}$$

因为 $\boldsymbol{x}_1$ 是可行解，$\boldsymbol{x}_0$ 是最优解。同时 $A\boldsymbol{x}_1 = A\boldsymbol{x}_0$。

同样的，式12也有两种变形：

1. （L_1 惩罚最小化）利用 $\boldsymbol{x}$ 是 K 稀疏向量的约束，将 L_1 不等式范数最小化变成 L_2：

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2, s.t. \|\boldsymbol{x}\| \leq q \tag{14}$$

   划归为一类二次规划 (quadratic program, QP) 问题。

2. 利用 Lagrangian 乘子法，将 L_1 不等式范数最小化变成：

$$\min_{\lambda,\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{15}$$

   划归为一类基追踪去噪 (basis pursuit denoising, BPDN) 问题 [7]。

在基于小波变换的图像/信号重构和恢复（deconv）中，也经常会遇到基追踪去噪问题。

参数稀疏的好处主要有以下两点：

1. 特征选择 (Feature Selection)
2. 可解释性 (Interpretability)

3.2 约束等距性条件 (Restricted Isometry Property Condition)

前文讨论了 L_1 范数最小化问题是 L_0 范数最小化某种程度的凸松弛。接下来考察两种问题的解之间的关系。

定义 1 (约束等距性 (Restricted Isometry Property, RIP) 条件 [5, 12])**.** *若存在矩阵 $\boldsymbol{A}$ 和 K-稀疏向量 $\|\boldsymbol{x}\|_0 \leq K$，*

$$(1-\delta_K)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}_K\boldsymbol{x}\|_2^2 \leq (1+\delta_K)\|\boldsymbol{x}\|_2^2 \tag{16}$$

其中 $0 \leq \delta_K < 1$ 是一个与稀疏度 K 有关的常数 (约束等距常数, restricted isometry constants, RIC)，$\boldsymbol{A}_K$ 是字典矩阵 $\boldsymbol{A}$ 的任意 K 列组成的子矩阵。

则称矩阵 $\boldsymbol{A}$ 满足 K 阶 RIP 条件。

当 RIP 条件满足时，非凸的 L_0 范数最小化问题与 L_1 范数最小化问题等价。也即是：

$$\min ||\boldsymbol{x}||_0 \ \ s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \overset{\text{概率为 1 的}}{\Longleftrightarrow} \min ||\boldsymbol{x}||_1 \ \ s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

带参数 δ_K 的 K 阶 RIP 条件简记为 $RIP(K,\delta_K)$，定义为所有使 $RIP(K,\delta_K)$ 成立的参数 δ 的下确界：

$$\delta_K = \inf\left\{\delta : (1-\delta)\|\boldsymbol{z}\|_2^2 \leq \|\boldsymbol{A}_{supp(z)}\boldsymbol{z}\|_2^2 \leq (1+\delta)\|\boldsymbol{z}\|_2^2, \forall |supp(z)| \leq K, \forall \boldsymbol{z} \in \mathbb{R}^{|supp(\boldsymbol{z})|}\right\} \tag{17}$$

显然若 $\boldsymbol{A}_K$ 正交，则 $\delta_K = 0$。于是，RIC 的非零值实际上可以评价该矩阵的非正交程度。此外，由于 $\boldsymbol{A}_K$ 的任意性，要求 $\boldsymbol{A}$ 在每一列的能量分布投影尽可能均匀。

RIC 有三个重要性质：

1. 系数信号精确重构的充分条件 [4]：若字典矩阵 $\boldsymbol{A}$ 分别满足 $\delta_K, \delta_{2K}, \delta_{3K}$ 的 RIP 条件，并且：
$$\delta_K + \delta_{2K} + \delta_{3K} < 1$$
则 L_1 范数最小化可以精确重构所有 K 稀疏信号。也即是，在此条件下，若无噪声存在，则 K 稀疏信号可以确保由 L_1 范数最小化精确恢复；并且在有噪声的情况下可以稳定估计。
2. RIC 与特征值的关系 [9]：若字典矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 满足 $RIP(K, \delta_K)$，则：
$$1 - \delta_K \leq \lambda_{min}(\boldsymbol{A}_K^T \boldsymbol{A}_K) \leq \lambda_{max}(\boldsymbol{A}_K^T \boldsymbol{A}_K) \leq 1 + \delta_K$$
3. 单调性 [4]：若 $K \leq K'$，则 $\delta_K \leq \delta_{K'}$

3.3 Tikhonov 正则化最小二乘 (Tikhonov Regularization LSM)

类似式15，作为最小二乘法代价函数 $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ 的改进，Tikhonov 在 1963 年提出了 L_2 正则化代价函数 [21]：
$$J(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2 \tag{18}$$
其中 $\lambda \geq 0$ 称正则化参数。

同样地，考察其共轭梯度：
$$\frac{\partial J(\boldsymbol{x})}{\partial \boldsymbol{x}^T} = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^T\boldsymbol{b} + \lambda\boldsymbol{x} = \boldsymbol{0} \tag{19}$$
立即得到：
$$\hat{\boldsymbol{x}}_{Tik} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{E})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{20}$$
这种使用 $(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{E})^{-1}$ 代替协方差矩阵 $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ 直接求逆的方式称之为 Tikhonov 正则化 (Tikhonov Regularization)[7]，或者简称正则化方法 (regularized method)[8]，在信号处理中称为松弛法 (relaxation method)。

Tikhonov 正则化本质是：通过对秩亏缺的协方差矩阵 $\boldsymbol{A}^T\boldsymbol{A}$ 的每一个对角线元素加一个很小的扰动 λ，使其变为非奇异阵，从而大大改善求逆的数值稳定性[9][10] [11]。

[7]值得注意的是 Weight Decay 并非 L_2 正则，而是在标准的随机梯度下降中 (SGD)，权重衰减正则化和 L_2 正则化的效果相同。而在选取 decay 值上，目前尚没有比较普适的公式 [16]。

[8]所有损害优化的方法都可以理解为正则化。主要分为增加优化约束和干扰优化过程两种思路。

[9]condition number 是一个矩阵（或者它所描述的线性系统）的稳定性或者敏感度的度量，如果一个矩阵的 condition number 在 1 附近，那么它就是 well-conditioned 的，如果远大于 1，那么它就是 ill-conditioned 的，如果一个系统是 ill-conditioned 的，它的输出结果就不具有相当的置信度。

[10]矩阵 $\boldsymbol{A}^T\boldsymbol{A}$ 半正定，所以 $\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{E}$ 的特征值位于区间 $[\lambda, \lambda+\|\boldsymbol{A}\|_F^2]$，使得条件数 $cond(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{E}) \leq \frac{\lambda+\|\boldsymbol{A}\|_F^2}{\lambda}$，相较之于 $cond(\boldsymbol{A}^T\boldsymbol{A}) < \infty$ 有极大改善。

[11] 首先假定只存在 $\boldsymbol{b}$ 的扰动 $\delta\boldsymbol{b}$，$\boldsymbol{A}$ 稳定，考察 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ 的解析解向量 $\boldsymbol{x} + \delta\boldsymbol{x}$，即：
$$\boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b}$$
求解得：
$$\delta\boldsymbol{x} = \boldsymbol{A}^{-1}\delta\boldsymbol{b}$$
由向量范数性质：
$$||\delta\boldsymbol{x}||_2 \leq ||\boldsymbol{A}^{-1}||_2 \cdot ||\delta\boldsymbol{b}||_2$$
同理显然：
$$||\boldsymbol{b}||_2 \leq ||\boldsymbol{A}||_2 \cdot ||\boldsymbol{x}||_2$$
于是：
$$\frac{||\delta\boldsymbol{x}||_2}{||\boldsymbol{x}||_2} \leq ||\boldsymbol{A}||_2 \cdot ||\boldsymbol{A}^{-1}||_2 \frac{||\delta\boldsymbol{b}||_2}{||\boldsymbol{b}||_2}$$
然后考察 $\delta\boldsymbol{A}$ 的影响：
$$(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b}$$
同理显然：
$$\frac{||\delta\boldsymbol{x}||_2}{||\boldsymbol{x}||_2 + ||\delta\boldsymbol{x}||_2} \leq ||\boldsymbol{A}||_2 \cdot ||\boldsymbol{A}^{-1}||_2 \frac{||\delta\boldsymbol{A}||_2}{||\boldsymbol{A}||_2}$$

如果使用迭代优化的算法，condition number 太大仍然会导致问题：它会拖慢迭代的收敛速度，而规则项从优化的角度来看，实际上是将目标函数变成 λ-strongly convex[12]（λ 强凸）的，也即可以在迭代中避免大平原的情况。

式15即是 L_1 正则化最小二乘，它总是有解的，但是不一定是唯一解。

同时 L_1 范数最小化解向量与 L_2 范数最小化解向量之间有如下关系 [4]：

$$0 \leq \|\boldsymbol{x}\|_2 - \frac{\|\boldsymbol{x}\|_1}{\sqrt{n}} \leq \frac{\sqrt{n}}{4}\left(\max_{1\leq i\leq n} x_i - \min_{1\leq i\leq n} x_i\right) \tag{21}$$

4 稀疏矩阵方程求解的优化算法 (Optimization Algorithm for Solving Sparse Matrix Equations)

上一节讨论了优化理论，本节讨论求解的具体算法。本质上来讲欠定的稀疏矩阵方程的求解都是将方程变换为超定的非稀疏矩阵方程求解。

方式众多，大致的思路有三种：

1. 贪婪启发式；
2. 统计拟合式；
3. 拓扑同伦式；

这里我们仅仅介绍一种贪婪启发式的**正交匹配追踪算法** [10, 19]：

Algorithm 1: 正交匹配追踪算法

Input: 观测数据向量 $\boldsymbol{y} \in \mathbb{R}^m$ 和字典矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$
Output: 稀疏的系数（权重）向量 $\boldsymbol{x} \in \mathbb{R}^n$

初始化：令标签集 $\Omega_0 = \emptyset$，初始化残差向量 $\boldsymbol{r}_0 = \boldsymbol{y}$，$k = 1$;
while 达成停止判据 * **do**
  辨识：求矩阵 $\boldsymbol{A}$ 中与残差 $\boldsymbol{r}_{k-1}$ 最相关的列;

$$j_k \in \arg\min_j |< \boldsymbol{r}_{k-1}, A_j >|, \Omega_k = \Omega_{k-1} \cup \{j_k\}$$

  估计：求解最小化问题 $\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}_{\boldsymbol{\Omega_k}}\boldsymbol{x}\|$ 的解;

$$\boldsymbol{x}_k = (\boldsymbol{A}_{\Omega_k}^T \boldsymbol{A}_{\Omega_k})^{-1} \boldsymbol{A}_{\Omega_k}^T \boldsymbol{y}$$

  更新残差：$\boldsymbol{r}_k = \boldsymbol{y} - \boldsymbol{A}_{\Omega_k}\boldsymbol{x}_k$;
  $k = k + 1$;
end
return $\boldsymbol{x}(i) = \begin{cases} \boldsymbol{x}_k(i), & k \in \Omega_k \\ 0, & else \end{cases}$;

下面是三种常用的停止判据 [23]：

1. 运行到某个固定的迭代步数；
2. 残差能量小于某个预先给定的固定值：$\|\boldsymbol{r}_k\|_2 \leq \varepsilon$;

于是可知解向量的相对误差应正比于 $||\boldsymbol{A}||_2 \cdot ||\boldsymbol{A}^{-1}||_2 \equiv \text{cond}(\boldsymbol{A}) \equiv \kappa(\boldsymbol{A})$，称 condition number.

对于超定方程（$\boldsymbol{A} \in \mathbb{C}^{m\times n}(m > n)$），必有最小二乘解 $\boldsymbol{x} = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\boldsymbol{b}$。容易证明（你信么?）$\text{cond}(\boldsymbol{A}^H\boldsymbol{A}) = \text{cond}^2(\boldsymbol{A})$。条件数是平方关系增大的，同时稳定性反比于条件数。

[12] A differentiable function f is called **λ-strongly convex** with parameter $\lambda > 0$ if the following inequality holds for all points x, y in its domain:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \lambda \|x - y\|_2^2$$

or, more generally,

$$\langle \nabla f(x) - \nabla f(y), (x - y) \rangle \geq \lambda \|x - y\|^2$$

where $\|\cdot\|$ is any norm.

3. 字典矩阵任何一列都没有残差向量的明显能量时：$\|\boldsymbol{A}^T\boldsymbol{r}_k\|_\infty \leq \varepsilon$；

5 卷积神经网络的中的稀疏问题 (The Sparse in CNN)

5.1 Generalization and Sparse

泛化能力可以看成模型的稀疏性。正如奥卡姆剃刀[13]指出的，面对不同的解释时，最简单的解释是最好的解释。在机器学习中，具有泛化能力的模型中应该有很多参数是接近 0 的。而在深度学习中，则是待优化的矩阵应该对稀疏性有偏好性。

泛化能力还可以看成模型的信息压缩能力。这里涉及到解释为什么深度学习有效的一种假说：信息瓶颈（information bottleneck）[20, 3, 2, 1]，说的是一个模型对特征进行压缩（降维）的能力越强，其就越更大的可能性做出准确的分类。

理解泛化能力的最后一种角度是风险最小化。这是从博弈论的角度来看，泛化能力强的模型能尽可能降低自己在真实环境中遇到意外的风险，因此会在内部产生对未知特征的预警机制，并提前做好应对预案。这是一种很抽象的也不那么精确的解释，但随着技术的进步，人们会找出在该解释下进行模型泛化能力的量化评价方法。

5.2 The Trade-off Between Architecture and Weights in CNN

在卷积神经网络中，稀疏带来最直观的结果就是剪枝式的权重稀疏化。首先介绍两个业内目前相对较新的工作。

Lottery Ticket Hypothesis(LTH)[13] 大意是按照这种迭代非结构化剪枝（剪枝将模型参数减少到 10%-20%）的方式可以找到一个小网络，用大网络同样的初始化方式可以在性能不会下降的情况下更快的训练这个小网络，但是随机初始化不能将这个网络训练到同样的性能。

而另一工作 [18] 中，认为对于结构化剪枝来说，剪枝完之后的模型是可以从随机初始化开始训练而达到和一般 fine-tuning 相比同样（甚至有些情况更好）的精度的，所以对于结构化剪枝来说更重要的是剪枝完得到的网络架构而非权重。

以上工作是同一年 ICLR2019 的工作，而且 LTH 还是当年的 best paper。然而两篇文章居然有矛盾的地方：那么在非结构剪枝中重头训练小模型时到底需不需要使用已经训练好的权重 (winning ticket)？使用 winning ticket 到底相比随机初始化有没有提升？

目前这个问题自提出之初 [17, 15] 至今还没有定论。在其他工作中，winning ticket 不一定有帮助的实验请参考 [14], train from scratch 不需要使用 winning tickets 的实验可以参考 [8]。

5.3 Do We really Need Sparse Convolution in CNN?

一般认为，稀疏卷积有三个问题：

1. 没有现成且成熟的稀疏卷积库，英伟达不推稀疏卷积优化，就没有合适的软硬件可以用；
2. 目前卷积的 GPU 实现是受带宽限制的，减少浮点计算潜在的提升不会很大；
3. LASSO 之类的稀疏训练多了个超参需要调节，而 LASSO 训练本身是无法加速的；

所以在谋求稀疏程度、加速比、精度损失平衡的时候，需要反反复复的搜索超参训练，考虑到 CNN 的训练速度，其潜力不如低精度方案。

然而在最近的来自 Intel 的一个工作中提出了：SLIDE(Sub-LInearDeep learning Engine)[6]。其工作流程如图5.3所示。号称提出的第一个针对现代 CPU 基于 OpenMP 并行性的可靠算法，可以胜过用于训练大型深度学习架构的最佳可用硬件 NVIDIA-V100。其精心定制的随机散列算法与允许异步并行的正确数据结构的组合。在训练时间上，我们在流行的极端分类数据集上的显示精度高达 TF-GPU 的 3.5 倍，而 TF-CPU 的 10 倍。接下来的步骤是将 SLIDE 扩展到包括卷积层。SLIDE 在随机存储器访问和并行性方面具有独特的优势。预计 SLIDE 的分布式实现将非常有吸引力，因为由于稀疏的渐变，通信成本最小。

[13]Occam's Razor: 这个原理称为“如无必要，勿增实体”，即“简单有效原理”。

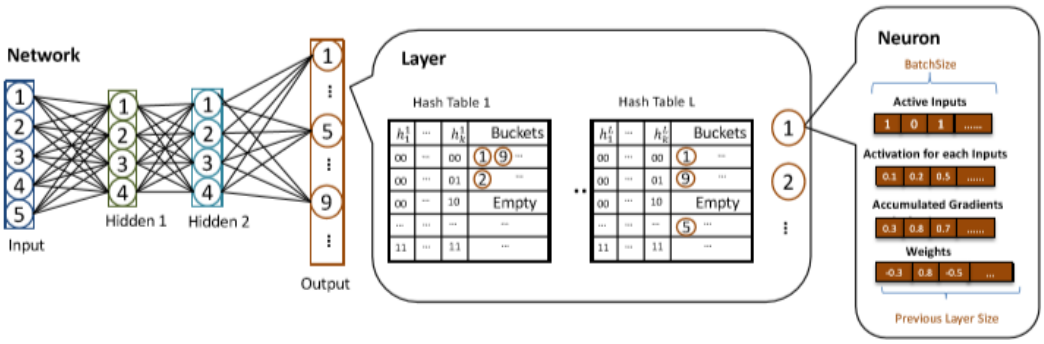

Figure 1: Architecture: The central module of SLIDE is Network. The network is composed of few-layer modules. Each layer module is composed of neurons and a few hash tables into which the neuron ids are hashed. Each neuron module has multiple arrays of batch size length: 1) a binary array suggesting whether this neuron is active for each input in the batch 2) activation for each input in the batch 3) accumulated gradients for each input in the batch. 4) The connection weights to the previous layer. The last array has a length equal to the number of neurons in the previous layer.

References

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*, 2017.

[2] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[3] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

[4] T Tony Cai, Lie Wang, and Guangwu Xu. New bounds for restricted isometry constants. *IEEE Transactions on Information Theory*, 56(9):4388–4394, 2010.

[5] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[6] Beidi Chen, Tharun Medini, James Farwell, Sameh Gobriel, Charlie Tai, and Anshumali Shrivastava. Slide: In defense of smart algorithms over hardware acceleration for large-scale deep learning systems. 2020.

[7] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[8] Elliot J Crowley, Jack Turner, Amos Storkey, and Michael O'Boyle. Pruning neural networks: is it time to nip it in the bud? 2018.

[9] Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE transactions on Information Theory*, 55(5):2230–2249, 2009.

[10] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[11] David L Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(6):797–829, 2006.

[12] Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via q-minimization for 0< q 1. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[13] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[14] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

[15] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[16] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[17] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[18] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[19] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

[20] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[21] Andrei N Tikhonov, A Tikhonov, and AN TIKHONOV. Solution of incorrectly formulated problems and the regularization method. 1963.

[22] Joel A Tropp. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006.

[23] Joel A Tropp and Stephen J Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[24] 郭金库, 刘光斌, 余志勇, and 吴瑾颖. 信号稀疏表示理论及其应用. 科学出版社, 2013.