

SVD and Regularization of Least-Squares Problems

Objectives

- Analyze impact of errors in least-squares problems using SVD
- Introduce truncated SVD regularization
- Analyze ridge regression using SVD

III-conditioned least-squares problems 2

$$\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 \Rightarrow \underline{w} = (\underline{A}^\top \underline{A})^{-1} \underline{A}^\top \underline{d}$$

have small singular values

$$\text{SVD: } \underline{A} = \underline{U} \Sigma \underline{V}^\top \Rightarrow \underline{w} = \underline{V} \Sigma^{-1} \underline{U}^\top \underline{d} = \sum_{i=1}^P \frac{1}{\sigma_i} \underline{v}_i (\underline{u}_i^\top \underline{d})$$

$N \times P$, rank P

$$\Rightarrow \|\underline{w}\|_2^2 = \sum_{i=1}^P \left(\frac{1}{\sigma_i}\right)^2 (\underline{u}_i^\top \underline{d})^2$$

Here we do not have \underline{v}_i because \underline{v}_i is orthonormal, $(\underline{v}_i)^\top = 1$

Small $\sigma_i \Rightarrow \underline{w}$ large $\|\underline{w}\|_2$

$$\text{Prediction with errors: } \tilde{\underline{y}} = (\tilde{\underline{x}} + \underline{\Sigma})^\top \underline{w} = \underline{x}^\top \underline{w} + \underbrace{\underline{\Sigma}^\top \underline{w}}_{\text{error}}$$

$$|\underline{\Sigma}^\top \underline{w}|^2 = \|\underline{w}\|_2^2 \|\underline{\Sigma}\|_2^2 \cos^2 \theta$$

large $\|\underline{w}\|_2^2 \Rightarrow$ sensitive

to errors

$\|\underline{w}\|_2^2$ is very large.

The solution can be disturbed by the error terms.
(we don't want this)

What if $\text{rank}(\underline{A}) < P$?

$\sigma_p = 0 \rightarrow b_1 \geq b_2 \geq \dots \geq b_p = 0$, we will encounter $\frac{1}{0}$, which is undefined

no unique solution

第1步：理想世界 (理论上的线性关系)

在最理想的情况下，我们假设数据是完美的，特征 x 和目标 y 之间存在一个精确的线性关系。我们可以用一个简单的方程来描述：

$$Xw = y$$

这里的 w 就是我们想求的权重向量。如果 X 是一个方阵并且可逆，我们可以直接解出 $w = X^{-1}y$ 。

第2步：现实世界 (充满噪声的数据)

现实中的数据总是有噪声的，而且通常样本数量（行数）远大于特征数量（列数），即 X 是一个“瘦高”的矩阵。在这种情况下， $Xw = y$ 这个方程几乎不可能有精确解。你不可能找到一条直线完美穿过所有的数据点。

因此，我们退而求其次，不再强求解，而是寻找一个“最优解”。

第3步：最小二乘法 (Least Squares) - 寻找最优解

最小二乘法的核心思想是：既然找不到能让 $Xw - y$ 等于零的 w ，那我们就找一个能让 $\|Xw - y\|_2^2$ （误差的平方和）最小的 w 。

这在几何上就相当于找到了一个离所有数据点“整体上最近”的线或超平面。

它的解析解就是我们熟悉的正规方程 (Normal Equation)：

$$w = (X^T X)^{-1} X^T y$$

这个方法在大多数情况下都非常好用。但是，它有两个“致命弱点”。

第4步：最小二乘法的“致命弱点”

这里的关键点在于 $(X^T X)^{-1}$ 这一项，也就是对 $X^T X$ 求逆。矩阵求逆是有前提条件的，那就是矩阵必须是可逆的。

1. 病态问题 (Ill-conditioned):

- 情况:** $X^T X$ 虽然可逆，但它“接近于不可逆”（术语叫“奇异”）。这通常发生在特征之间高度相关（但不完全线性相关）时。
- 后果:** 它的逆矩阵 $(X^T X)^{-1}$ 会变得非常大，导致最终求出的解 w 的范数 $\|w\|$ 也非常大。这样的模型对数据中的微小噪声极其敏感，预测结果会非常不稳定。

2. 秩亏缺问题 (Rank-deficient):

- 情况:** $X^T X$ 完全不可逆。这通常发生在特征之间存在完全线性相关（即多重共线性），例如特征3 = 2 * 特征1。此时， $\text{rank}(X) < P$ 。
- 后果:** $(X^T X)^{-1}$ 根本不存在！这意味着有无穷多个 w 向量都能让误差 $\|Xw - y\|_2^2$ 达到那个最小值。模型不知道该选哪一个解，因此最小二乘法失效了。

正如你所说，我们陷入了困境：因为有噪声，我们用了最小二乘法；但因为特征相关，最小二乘法又可能失效。

第5步：岭回归 (Ridge Regression) - 解决方案

为了解决这个困境，岭回归应运而生。它对最小二乘法的目标函数做了一个小小的修改：

$$\min \{ \|Xw - y\|_2^2 + \alpha \|w\|_2^2 \}$$

我们来分析这个新目标：

- 第一部分** $\|Xw - y\|_2^2$ ：和最小二乘法一样，我们仍然希望模型能很好地拟合数据（误差小）。
- 第二部分** $\alpha \|w\|_2^2$ ：这是新加入的惩罚项 (Penalty Term) 或 正则化项 (Regularization Term)。
 - $\|w\|_2^2$ 是权重向量 w 的 L2 范数的平方。它要求 w 里的每个元素都尽可能小。
 - α 是一个超参数，由我们自己设定。它用来平衡“拟合数据”和“保持权重小”这两个目标。

岭回归如何解决问题的？

岭回归的解析解变成了：

$$w = (X^T X + \alpha I)^{-1} X^T y$$

对比最小二乘法的解，我们发现变化就在于求逆的那一项从 $X^T X$ 变成了 $(X^T X + \alpha I)$ (I 是单位矩阵)。

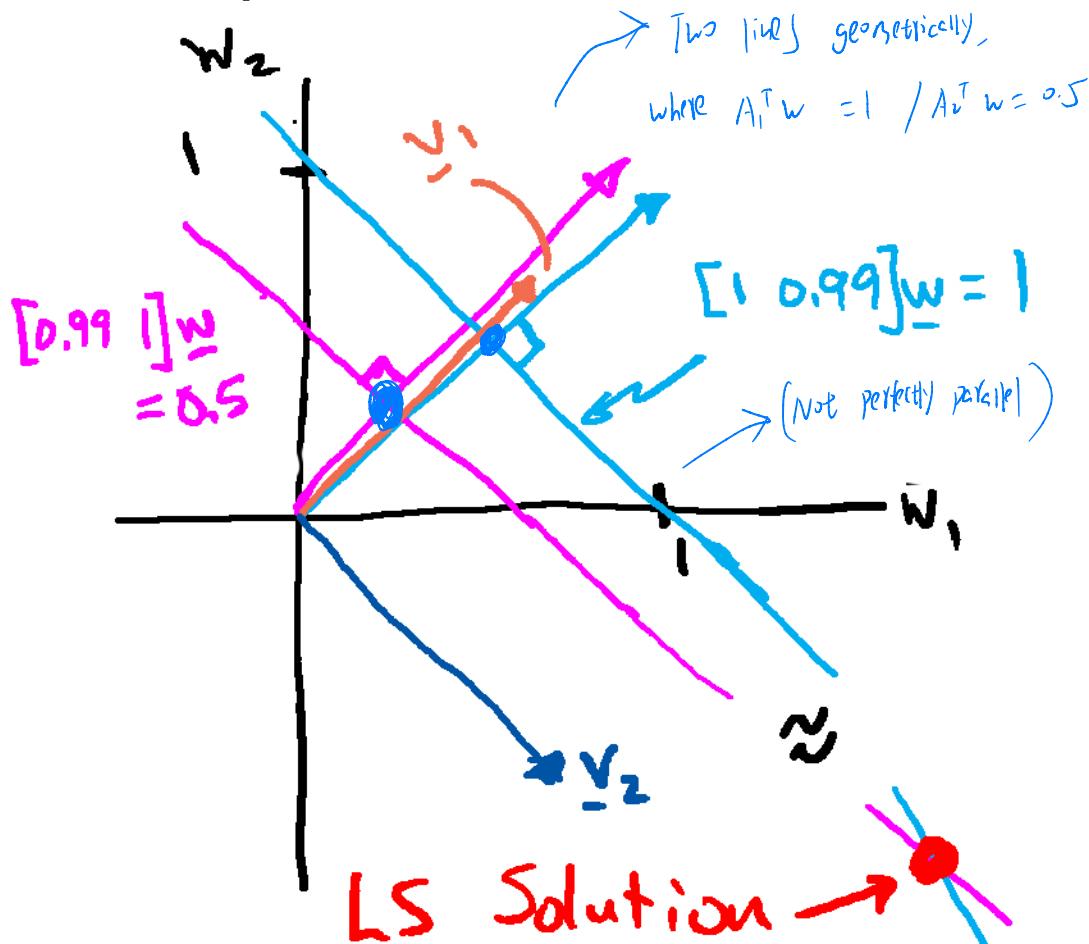
- 这个 $+ \alpha I$ 的操作非常巧妙。** 它相当于给 $X^T X$ 矩阵的对角线上的每个元素都加上了一个小的正数 α 。
- 在数学上可以证明，即使 $X^T X$ 是不可逆的（秩亏缺）， $(X^T X + \alpha I)$ 也一定是可逆的（只要 $\alpha > 0$ ）。

所以，岭回归：

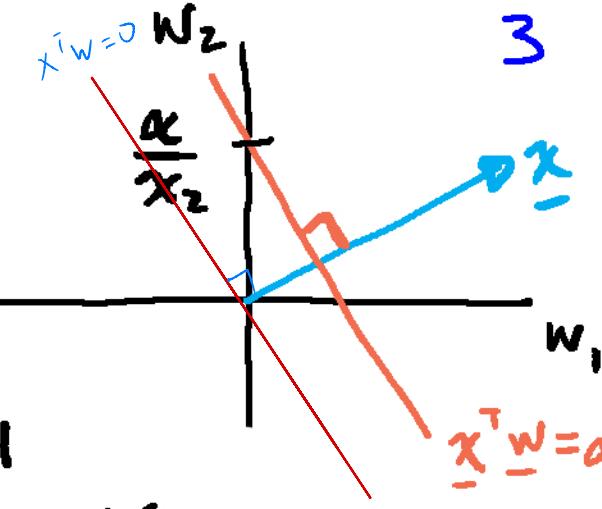
- 对于秩亏缺问题：**通过 $+ \alpha I$ 使得矩阵变得可逆，从而保证了解 w 的存在性和唯一性。在无穷多个解中，它会选择那个 L2 范数最小的解。
- 对于病态问题：** $+ \alpha I$ 操作同样稳定了求逆的过程，有效抑制了 w 的范数变得过大，从而提高了模型的稳定性和泛化能力。

Example: ill-conditioned A

$$\underline{A} = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}, \underline{d} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$



Geometry $\underline{x}^T \underline{w} = \alpha$



$$\sigma_1 = 1.99 \quad \sigma_2 = 0.01$$

$$1/\sigma_1 \approx 0.5, \quad 1/\sigma_2 = 100$$

$$\underline{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \end{bmatrix}^T \quad \underline{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix}^T$$

$$\underline{u}_1^T \underline{d} = \frac{1.5}{\sqrt{2}}, \quad \underline{u}_2^T \underline{d} = \frac{0.5}{\sqrt{2}}$$

$$\underline{w} = \underline{v}_1 \frac{\underline{u}_1^T \underline{d}}{\sigma_1} + \underline{v}_2 \frac{\underline{u}_2^T \underline{d}}{\sigma_2} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \frac{3}{\sqrt{2}} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \frac{25}{\sqrt{2}}$$

(Large norm is problematic because it leads to tremendous amplification of noise)

$\|\underline{w}\|_2 \uparrow$ as $\sigma_2 \downarrow (\rightarrow)$

Regularized LS via truncated SVD

4

Replace $\sum_{i=1}^p \frac{1}{\sigma_i} \underline{v}_i (\underline{u}_i^\top \underline{d})$ with $\sum_{i=1}^r \frac{1}{\sigma_i} \underline{v}_i (\underline{u}_i^\top \underline{d})$
where $r < p$.



(Solution to the Ill-conditioned LS problem)

- Avoid inverting small/zero singular values
- Equivalent to replacing $\underline{A} = \sum_{i=1}^p \sigma_i \underline{u}_i \underline{v}_i^\top$ with the rank- r approximation $\underline{A}_r = \sum_{i=1}^r \sigma_i \underline{u}_i \underline{v}_i^\top$
- Increases $\min_w \|\underline{A}\underline{w} - \underline{d}\|_2^2$
- Can choose r using intuition or cross-validation

Regularized LS via ridge regression

$$\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda \|\underline{w}\|_2^2 \Rightarrow \underline{w} = (\underline{A}^T \underline{A} + \lambda \underline{I})^{-1} \underline{A}^T \underline{d}$$

controls norm!

Use SVD: $\underline{A}^T \underline{A} = \underline{V} \Sigma^2 \underline{V}^T$, $\lambda \underline{I} = \underline{V} \lambda \underline{I} \underline{V}^T$

$\underline{w} = (\underline{V} (\Sigma^2 + \lambda \underline{I}) \underline{V}^T)^{-1} \underline{V} \Sigma \underline{U}^T \underline{d} = \underline{V} (\Sigma^2 + \lambda \underline{I})^{-1} \Sigma \underline{U}^T \underline{d}$

Because $\underline{V} \underline{V}^T$ is \underline{I} , so no difference

$$\underline{D} = \begin{bmatrix} \frac{1}{\sigma_1^2 + \lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_p^2 + \lambda} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{bmatrix} = \begin{bmatrix} \sigma_1 / (\sigma_1^2 + \lambda) & & 0 \\ & \ddots & \\ 0 & & \sigma_p / (\sigma_p^2 + \lambda) \end{bmatrix}$$

Controlled!

$\underline{w} = \sum_{i=1}^p \frac{\sigma_i}{\sigma_i^2 + \lambda} \underline{v}_i (\underline{u}_i^T \underline{d})$

↓ SVD version of Ridge Regression

- as $\sigma_i \rightarrow 0$, $\frac{\sigma_i}{\sigma_i^2 + \lambda} \rightarrow \sigma_i / \lambda$
- increased value $\|\underline{A}\underline{w} - \underline{d}\|_2^2$

**Copyright 2019
Barry Van Veen**