Technical Note: The Soft-Thresholding Operator

Paul F. Roysdon[†] Jay A. Farrell[‡]

***Abstract*— Due to space limitations in [1] and [2], this Technical Note is supplied to explain the formulation of the l_1-regularization for a nonlinear sliding window estimator.**

I. Introduction

Our work in outlier detection and exclusion, or accommodation, is motivated by recent advances in computer vision where sparse representation of candidate tracking sets [3] is applied to face recognition [4]. While it is common in the robotics community to solve state estimation problems by a formulation of the Maximum Likelihood Estimate (MLE), e.g. the Kalman filter, the MLE is sensitive to measurements which deviate from their stochastic noise model. The authors of [3] demonstrate that l_1-regularization can exploit the sparseness of outliers in a candidate dataset. However, success of the regularization depends on measurement redundancy.

II. Linear Problem Formulation

Consider the simple linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{H} \in \mathbb{R}^{m\times n}$ for $m > n$, state vector $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\boldsymbol{\sigma}^2)$ is the measurement noise. The maximum likelihood estimate for $\boldsymbol{x}$ is found by

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{arg\,min} \left\{ -\frac{1}{2}\|\mathbf{y} - \mathbf{H}\boldsymbol{x}\|_2^2 \right\}. \tag{2}$$

Given a dataset without outliers, the residual $\mathbf{r} \triangleq \mathbf{y} - \mathbf{H}\boldsymbol{x}$ will be dense with variance $\mathbf{I}\boldsymbol{\sigma}^2$. However, in the presence of outliers, $\mathbf{r}$ will contain both dense values from nominal measurements, and sparse values resulting from outliers. We can exploit the sparseness of the outliers by solving the problem in (1) as an l_1-regularized least squares problem, which is known to yield sparse solutions [3]. The Least Soft-thresholded Squares (LSS) [5] estimate for $\boldsymbol{x}$ is found by

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{arg\,min} \left\{ -\frac{1}{2}\|\mathbf{y} - \mathbf{H}\boldsymbol{x} - \mathbf{s}\|_2^2 + \lambda\|\mathbf{s}\|_1 \right\}, \tag{3}$$

where $\mathbf{s} \in \mathbb{R}^m$, and the regularizing or *soft-thresholding parameter* [6] is $\lambda \in \mathbb{R}$. The $\|.\|_1$ and $\|.\|_2$ denote the l_1 and l_2 norms respectively.

[†]Ph.D. graduate, [‡]Professor at the Dept. of Electrical & Computer Engineering, UC Riverside. {proysdon, farrell}@ece.ucr.edu.

A. Example 1: Necessity of Measurement Redundancy

Consider a simple 2D line-fit problem, $\mathbf{y} = \mathbf{H}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^2$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{H} \in \mathbb{R}^{m\times 2}$. The vertical shift along the y-axis is $\boldsymbol{x}(1)$, and slope is $\boldsymbol{x}(2)$. Suppose the true values are $\boldsymbol{x} = [0, 0]$, then true line lies on the x-axis of the x-y plane.

Assume $m = 2$. Given two measurements, $\tilde{\mathbf{y}} = [5, 0]$, the Least-Square (LS) estimate of the two unknowns is $\hat{\boldsymbol{x}} = (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\tilde{\mathbf{y}} = [5, -5]$, i.e. the estimated line is shifted up by 5 and has a slope of -5. Clearly, without measurement redundancy, it is impossible to reject, or accommodate, the bad measurement $\mathbf{y}(1) = 5$.

For the overdetermined problem where $m \geq 3$, there are $(m - 2)$ degrees-of-freedom with which to make a decision given any pair measurements. If a measurement is bad, an algorithm can be employed to remove or accommodate for the bad measurement, and the simple 2D line-fit problem can still be solved. While this is a trivial example, it motivates the necessity of measurement redundancy.

B. Example 2: Sparsity of L-1 Regularization

Here we extend the 2D line-fit problem of Section II-A, such that $m = 200$. Applying eqn. (3), Fig. 1 illustrates the residuals for two cases, with and without outliers. It is clear that the top plot of Fig. 1 (the case without outliers) contains residuals which are dense with zero mean. However, the bottom plot of Fig. 1 (the case *with* outliers) clearly shows that outliers are generally sparse, substantiating the claim of [3].

Applying equations (2) and (3) to the 2D line-fit problem, it is trivial to demonstrate the LS sensitivity to outliers. In this example, the LS residuals have a mean $\mu = 7.39$ and standard deviation $\sigma = 2.75$, whereas the LSS residuals have $\mu = 0.05$ and $\sigma = 0.99$.

The resulting model fit is shown in Fig. 2, where the true line lies on the x-axis, the LS fit is shifted up along the y-axis, and the LSS result nearly overlaps the true line. [1]

III. Soft-Thresholding Operator Proof

This section solves the optimization problem

$$f(r) = \underset{s}{arg\,min} \left\{ \frac{1}{2}\left(r - \frac{s}{\sigma}\right)^2 + \frac{1}{\nu}|s| \right\} = \underset{s}{arg\,min}\, g_r(s),$$

where $r, s \in \mathbb{R}$, $\sigma > 0$ and $\nu > 0$ are the parameters of the Normal and Laplacian distributions, and

$$g_r(s) \triangleq \frac{1}{2}\left(r - \frac{s}{\sigma}\right)^2 + \frac{1}{\nu}|s|. \tag{4}$$

[1]PFR: I think this paragraph and the Fig. 2 are unnecessary.

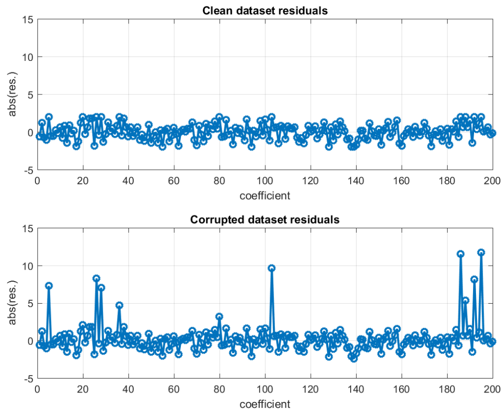

Fig. 1. Top: Clean dataset residuals without outliers. Bottom: Corrupted dataset residuals with 5% outliers.

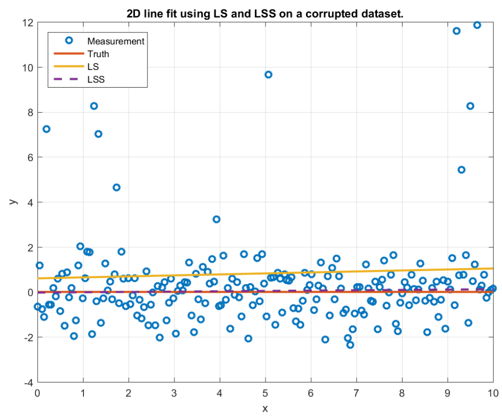

Fig. 2. 2D line fit with a corrupted dataset containing 5% outliers.

Note first that $g_r(s)\Big|_{s=0} = \frac{1}{2}r^2$.

Because $g_r(s)$ is not differentiable in s, three cases can be considered ($s < 0$, $s = 0$, and $s > 0$), with the final answer $f(r)$ being the value of s over the three cases that gives the lowest cost. For $s \neq 0$:

$$\frac{\partial}{\partial s} g_r(s) = -\frac{r}{\sigma} + \frac{s}{\sigma^2} + \frac{1}{\nu}\ \mathrm{sgn}(s).$$

For $s > 0$, $\frac{\partial}{\partial s} g_r(s) = 0$ yields the critical value $s_+^* = \sigma(r - \mu)$, where $\mu \triangleq \frac{\sigma}{\nu}$. Because, in this case $s_+^* > 0$, it must be that $r > \mu$. The cost at s_+^* is:

$$g_r(s)\Big|_{s=s_+^*} = g_r(\sigma(r-\mu)) = \mu r - \frac{1}{2}\mu^2.$$

Note that:

$$\frac{1}{2}(r-\mu)^2 \geq 0 \quad \forall\ r, \mu;$$

therefore,

$$\frac{1}{2}r^2 \geq r\mu - \frac{1}{2}\mu^2 \quad \forall\ r, \mu.$$

This ensures that in this case (i.e., $s > 0$), for any value of r, it is true that $g_r(s_+^*) \leq g_r(0)$.

For $s < 0$, $\frac{\partial}{\partial s} g_r(s) = 0$ yields the critical value $s_-^* = \sigma(r + \mu)$. Because, in this case $s_-^* < 0$, it must be that $r < -\mu$. The cost at s_-^* is:

$$g_r(s)\Big|_{s=s_-^*} = g_r(\sigma(r+\mu)) = -\mu r - \frac{1}{2}\mu^2.$$

Note that:

$$\frac{1}{2}(r+\mu)^2 \geq 0 \quad \forall\ r, \mu;$$

therefore,

$$\frac{1}{2}r^2 \geq -r\mu - \frac{1}{2}\mu^2 \quad \forall\ r, \mu.$$

This ensures that in this case (i.e., $s < 0$), for any value of r, it is true that $g_r(s_+^*) \leq g_r(0)$.

When $|r| < \mu$, it is straightforward to show that any non-zero value of s will increase the second term of $g_r(s)$ more than it decreases the first term; therefore, in this case $s^* = 0$.

Given the analysis above, the unique optimal solution for s as a function of r and $\mu > 0$ is:

$$s = \begin{cases} \sigma(r+\mu), & \text{if } r < -\mu, \\ \sigma(r-\mu), & \text{if } r > \mu, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Eqn. (5) can be more compactly stated as

$$S_{\sigma,\nu}(r) = \sigma\ \mathrm{sgn}(r)\ \max\left(|r| - \frac{\sigma}{\nu}, 0\right).$$

References

[1] P. F. Roysdon and J. A. Farrell, "GPS-INS Outlier Detection and Elimination using a Sliding Window Filter," *American Control Conference, In Presc.*, 2017.

[2] ——, "Robust GPS-INS Outlier Accomodation using a Sliding Window Filter," *22th IFAC World Congress*, 2017.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, 2009.

[4] X. Mei and H. Ling, "Robust Visual Tracking using L-1 Minimization," *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 2009.

[5] D. Wang, H. Lu, and M. Yang, "Robust Visual Tracking via Least Soft-threshold Squares," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[6] P. Huber, *Robust Statistics*. New York: John Wiley and Sons Inc., 1986.