CPSC 490: Extensions of Model Repair for Robust Recovery of Over-Parametrized Models

Seungkyoon Bong
Advisor: John Lafferty

Introduction

Statistical models, such as regression models, are typically designed with underlying assumptions about the distribution and soundness of the data. When these idealized assumptions are violated, some methods, like ordinary least squares which is highly sensitive to outliers, can become compromised. This has led to a class of estimation techniques known as robust estimation which is less sensitive to departures from assumptions of the algorithm. Although traditional robust estimation assumes that the data is inherently corrupted, a recent paper by Gao and Lafferty also studies the case in which the data is sound but the statistical model becomes corrupted after fitting the data. The goal of this thesis is to continue and expand upon the ideas and simulations explored in the paper.

Background

As statistical models grow larger and more prevalent, questions about the validity and integrity of these models become more pertinent. While there is existing literature on model repair for corrupt data, the literature is less rich for corrupt models. Lafferty and Gao attempt to formulate and address this problem in a paper written in 2020.

They formulate the general problem of model repair with a corrupted estimator as follows:

- $\hat{\theta} \in R^p$ is a model with p parameters estimated on n data points $\{x_i, y_i\}_{i=1}^n$
- The model $\hat{\theta}$ is corrupted as $\eta = \hat{\theta} + z$ where
    - $z_j \sim (1-\epsilon)\delta_0 + \epsilon Q_j$ and Q_j is some distribution
    - z is independent of the design $\{x_i\}_{i=1}^n$
    - i.e. each $\hat{\theta}_j$ is corrupted by additive noise from Q_j with probability ϵ
- The goal is to recover $\hat{\theta}$ from η without re-estimating the model from scratch

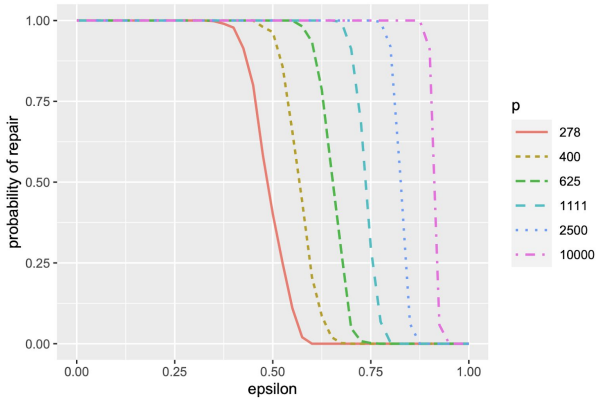
Using this general formulation, they then delve into multiple applications of robust recovery of statistical models, two of which are of foundational interest in this proposal: over-parameterized linear models and feedforward networks.

Repair of over-parametrized linear models largely follows the framework above with:

- Design matrix $X \in \mathbb{R}^{n \times p}$ and response vector $y \in \mathbb{R}^n$

- Minimization of the squared error $||y - X\theta||_2^2$
- The ordinary least squares (OLS) solution $\hat{\theta} = X^T(XX^T)^{-1}y$
- However, we only have a corrupted version $\eta = \hat{\theta} + z$, where z is the same as above

It turns out that by finding $\widetilde{\theta} = X^T\widetilde{u}$, where $\widetilde{u} = \text{argmin}_u||\eta - X^T u||_1$ we can recover the original model $\hat{\theta}$ with high probability. A plot (from the paper) of the probability of exact repair as a function of the corruption probability ϵ are to the right.



For feed-forward neural networks, the paper considers a feedforward network with one hidden layer:

- $$f(x) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p}\beta_j\phi(W_j^T x)$$
  - ϕ is an activation function (ReLU and hyperbolic tangent were studied here)
- Squared error loss: $$L(\beta, W) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \frac{1}{\sqrt{p}}\sum_{j=1}^{p}\beta_j\phi(W_j^T x)\right)^2$$
- Trained using gradient descent algorithm
  - Either normally or by training the first layer, freezing it, and retraining the second
- The model repair is performed in reverse layer order similar to that of the linear model. Further details of the repair can be found in the paper.

In general, the paper finds that to repair a corrupted model, the statistical model must (1) be over-parameterized and (2) incorporate redundancy. In fact, because of these two key properties, the response variable is not needed to recover the statistical models analyzed in this paper; the design matrix functionally encodes the response variable because of the over-parameterization of the model.

<u>Sub-projects</u>

Theoretical Proof: Minimax Lower Bound for Robust Regression

An interesting observation from the paper is that the model repair problem for a corrupted linear model can be viewed as a robust regression problem. If we consider η as the corrupted response

vector and $A = X^T \in \mathbb{R}^{p \times n}$ as the uncorrupted design matrix, we can solve the analogous robust regression problem.

As a robust regression problem, we can search broader related literature to determine the lower-bound on the run-time complexity. If we make a simplifying assumption that all $(1-\epsilon)n$ uncorrupted response variables exist on the same hyperplane (i.e. no noise), this is akin to the degenerate point subset problem (c-DPS) where $c = \epsilon$. In a publication from 2006, Thorsten Bernholt proves that c-DPS is NP-hard via a reduction from the NP-hard vertex cover problem. While this proof relies on the portion of points not on the hyperplane $c = \epsilon \leq 0.5$, we would like to see if the work could extend to the case of $c > 0.5$.

Simulation 1: Model Repair for a Large, Real-Life Dataset

In the paper, the datasets are primarily small toy datasets of known, generated distributions. It would thus be worthwhile to use a real, larger dataset and ensure that the results of the paper are truly robust and can be applied in other domains. In particular, the work in this paper has implications for privacy because the linear program for repairing a model can be solved distributively as opposed to re-estimating the model from scratch. Therefore, ensuring versatility would be important to take advantage of this potential privacy benefit in multiple domains.

Simulation 2: Application to Convolutional Networks

The applications to neural networks in the paper were restricted to feed-forward neural networks which are the simplest networks to analyze. However, as convolutional networks become more popular, especially in the computer vision domain, we would like to see if the same principles of model repair apply to convolutional networks. This would involve creating a 2-layer convolutional network and training it. Specifically, I will attempt to both train each layer of the convolutional network jointly and train the first layer followed by the second (after freezing the first) to parallel the experimentation of the feedforward network. Similar to the feedforward example, the latter approach can be interpreted as applying a second convolutional layer to the features extracted from the input by the first convolutional layer. The work will be attempted on a toy dataset, and potentially applied to a larger real-life dataset if the results are encouraging.

Deliverables

- Proof of the robust regression minimax lower bound complexity
- Code for both simulations (including related plots and analyses of simulations)

References

Bernholt, T. (2006). *Robust estimators are hard to compute* (No. 2005, 52). Technical report.

Gao, C., & Lafferty, J. (2020). Model Repair: Robust Recovery of Over-Parameterized Statistical Models. *arXiv preprint arXiv:2005.09912*.