Model Repair: Robust Recovery of Over-Parameterized Statistical Models

Chao Gao
Department of Statistics
University of Chicago

John Lafferty
Department of Statistics and Data Science
Yale University

March 19, 2020

Abstract: A new type of robust estimation problem is introduced where the goal is to recover a statistical model that has been corrupted after it has been estimated from data. Methods are proposed for "repairing" the model using only the design and not the response values used to fit the model in a supervised learning setting. Theory is developed which reveals that two important ingredients are necessary for model repair—the statistical model must be over-parameterized, and the estimator must incorporate redundancy. In particular, estimators based on stochastic gradient descent are seen to be well suited to model repair, but sparse estimators are not in general repairable. After formulating the problem and establishing a key technical lemma related to robust estimation, a series of results are presented for repair of over-parameterized linear models, random feature models, artificial neural networks, and certain families of Gaussian sequence models. A series of simulation studies are presented that corroborate and illustrate the theoretical findings.

1. Introduction

In this paper we introduce a new type of robust estimation problem—how to recover a statistical model that has been corrupted after estimation. Traditional robust estimation assumes that the data are corrupted, and studies methods of estimation that are immune to these corruptions or outliers in the data. In contrast, we explore the setting where the data are "clean" but a statistical model is corrupted after it has been estimated using the data. We study methods for recovering the model that do not require re-estimation from scratch, using only the design and not the original response values.

The problem of model repair is motivated from several different perspectives. First, it can be formulated as a well-defined statistical problem that is closely related to, but different from, traditional robust estimation, and that deserves study in its own right. From a more practical perspective, modern machine learning practice is increasingly working with very large statistical models. For example, artificial neural networks having several million parameters are now routinely estimated. It is anticipated that neural networks having trillions of parameters will be built in the coming years, and that large models will be increasingly embedded in systems, where they may be subject to errors and corruption of the parameter values. In this setting, the maintenance of models in a fault tolerant manner becomes a concern. A different perspective takes inspiration from plasticity in brain function, with the human brain in particular having a remarkable ability to repair itself after trauma. The framework for model repair that we introduce in this paper can be viewed as a

crude but mathematically rigorous formulation of this ability in neural networks.

At a high level, our findings reveal that two important ingredients are necessary for model repair. First, the statistical model must be over-parameterized, meaning that there should be many more parameters than observations. While over-parameterization leads to issues of identifiability from traditional perspectives, here it is seen as a necessary property of the model. Second, the estimator must incorporate redundancy in some form; for instance, sparse estimators of over-parameterized models will not in general be repairable. Notably, we show that estimators based on gradient descent and stochastic gradient descent are well suited to model repair.

At its core, our formulation and analysis of model repair rests upon representing an estimator in terms of the row space of functions of the data design matrix. This leads to a view of model repair as a form of robust estimation. The recovery algorithms that we propose are based on solving a linear program that is equivalent to median regression. Our key technical lemma, which may be of independent interest, gives sharp bounds on the probability that this linear program successfully recovers the model, which in turn determines the level of over-parameterization that is required. An interesting facet of this formulation is that the response vector is not required by the repair process. Because the model is over-parameterized, the estimator effectively encodes the response. This phenomenon can be viewed from the perspective of communication theory, where the corruption process is seen as a noisy channel, and the design matrix is seen as a linear error-correcting code for communication over this channel.

After formulating the problem and establishing the key technical lemma, we present a series of results for repair of over-parameterized linear models, random feature models, and artificial neural networks. These form the main technical contributions of this paper. We also explain how the concepts of over-paramaterization and redundancy apply to repair of nonparametric models, including Gaussian sequence models for Sobolev spaces and isotonic regression. A series of simulation experiments are presented that corroborate and illustrate our theoretical results.

In the following section we give a more detailed overview of our results, including the precise formulation of the model repair problem, its connection to robust estimation and error correcting codes, and an example of the repair algorithm in simulation. We then present the key lemma, followed by detailed analysis of model repair for specific model classes. Section 3 presents further simulation results that confirm the theory. In Section 4 we discuss directions for further research and potential implications of our findings for applications.

2. Problem formulation and overview of results

In this section we formulate the problem of model repair, and give an overview of our results. Suppose that $\widehat{\theta} \in \mathbb{R}^p$ is a model with p parameters estimated on n data points $\{(x_i, y_i)\}_{i=1}^n$ as a classification or regression model. The model $\widehat{\theta}$ is then corrupted by noise. The primary noise model we study in this paper is

$$\eta = \widehat{\theta} + z \tag{2.1}$$

where $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q$ and Q is an arbitrary distribution. In other words, each component $\widehat{\theta}_j$ of $\widehat{\theta}$ is corrupted by additive noise from an arbitrary distribution Q with probability ε, where

$0 \leq \varepsilon \leq 1$, and is uncorrupted with probability $1 - \varepsilon$. We discuss alternative error models later in the paper. The goal is to recover $\widehat{\theta}$ from η, without reestimating the model using the response values $\{y_i\}$.

Overparameterized linear models. To explain the main ideas, let us first consider the setting of under-determined linear regression. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix and $y \in \mathbb{R}^n$ a vector of response values, and suppose that we wish to minimize the squared error $\|y - X\theta\|_2^2$. If $n > p$ then this is an under-determined optimization problem. Among all solutions to the linear system $y = X\theta$, the solution of minimal norm $\|\theta\|_2$ is given by

$$\widehat{\theta} = X^T(XX^T)^{-1}y \tag{2.2}$$

assuming that X has full rank n (Boyd and Vandenberghe, 2004). Thus, $\widehat{\theta}$ lies in the row space of the $n \times p$ design matrix X.

Now suppose that $\eta = \widehat{\theta} + z$ where $z_j \sim (1 - \varepsilon)\delta_0 + \varepsilon Q$. The method we propose to recover $\widehat{\theta}$ from η is to let $\widetilde{u} \in \mathbb{R}^n$ be the solution to the optimization

$$\widetilde{u} = \arg\min_u \|\eta - X^T u\|_1 \tag{2.3}$$

and define the repaired model as $\widetilde{\theta} = X^T\widetilde{u}$. The linear program defined in (2.3) can be thought of as performing median regression of η onto the rows of X. Our analysis shows that, under appropriate assumptions, the model is repaired with high probability, so that $\widetilde{\theta} = \widehat{\theta}$, as long as $n/p \leq c(1-\varepsilon)^2$ for some sufficiently small constant c.

Figure 1 shows the performance of the repair algorithm in simulation. The design is sampled as $X_{ij} \sim N(0, 1)$ and the corruption distribution is $Q = N(1, 1)$. With the sample size fixed at $n = 50$, the dimension p is varied according to $p_j/n = 200/j^2$ with j ranging from 1 to 6. The plots show the empirical probability of exact repair $\widetilde{\theta} = \widehat{\theta}$ as a function of ε. The roughly equal spacing of the curves agrees with our theory, which indicates that $\sqrt{n/p}/(1 - \varepsilon)$ should be sufficiently small for successful repair. The theory indicates that the repair probability for dimension p_j as a function of the adjusted value $\varepsilon_j = \varepsilon + c' \cdot j - \frac{1}{2}$ should exhibit a threshold at $\varepsilon_j = 1/2$ for the constant $c' = \frac{\sqrt{2}}{20c}$; this is seen in the right plot of Figure 1.

Robust regression. This procedure can be viewed in terms of robust regression. Specifically, η can be viewed as a corrupted response vector, and $A = X^T \in \mathbb{R}^{p \times n}$ can be viewed as design matrix that is *not corrupted*. Our result makes precise conditions under which this robust regression problem can be successfully carried out. In particular, we show that model repair is possible even if $\varepsilon \to 1$, so that the proportion of corrupted model components approaches one. This is in stark contrast to the traditional Huber model where the design is corrupted (Huber, 1964), under which consistent estimation is only possible if $\varepsilon \to 0$ (Chen et al., 2016; Gao, 2020).

Error-correcting codes. Model repair can also be viewed in terms of error-correcting codes. Specifically, viewing the response vector $y \in \mathbb{R}^n$ as a "message" to be communicated over a noisy channel, the minimum norm model $\widehat{\theta} = X^T u = X^T(XX^T)^{-1}y$ redundantly encodes y since $p > n$ (see Figure 2). The decoding algorithm $\widetilde{u} = \arg\min_u \|\eta - X^T u\|$ then recovers the data y according

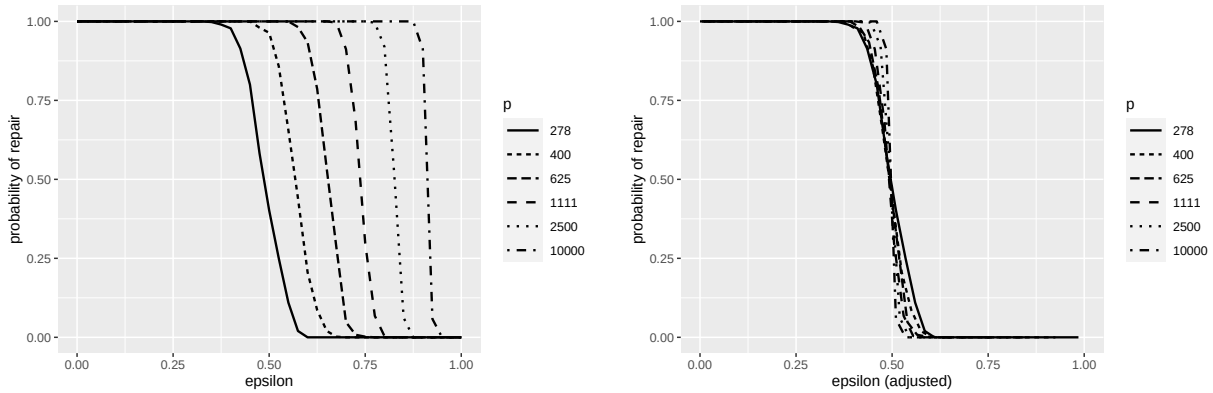

FIG 1. *Left: Empirical probability of exact repair as a function of ε. The sample size is $n = 50$ and the model dimension p varies as $p_j/n = 200/j^2$, for $j = 1, \ldots, 6$; each point is an average over 500 trials. The plot on the right shows the repair probability as a function of the adjusted value $\varepsilon_j = \varepsilon + c' \cdot j - \frac{1}{2}$ for dimension p_j, where the constant is $c' = \frac{\sqrt{2}}{20c} = 0.085$.*

to $y = (XX^T)\widetilde{u}$. The inequality $n/p < c(1-\varepsilon)^2$ gives a condition on the rate of the code, that is, the level of redundancy that is sufficient for this decoding procedure to recover the message with high probability. When X is a random Gaussian matrix, the mapping $u \to X^T u = \sum_{i=1}^n u_i X_i^T$ can be viewed as a superposition of random codewords in $\mathbb{R}^p$ (Joseph and Barron, 2012; Rush et al., 2017). The fundamental difference with channel coding is that in our regression setting, the design matrix X is fixed, and is not chosen for optimal channel coding. Indeed, the noise model $w \to w + z$ that we consider corresponds to a channel having infinite capacity, and a simple repetition code would suffice for identifying components that are uncorrupted.

Estimators based on gradient descent. The observations made above carry over to estimators of linear models based on gradient descent. Consider objective functions of the form

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(y_i, x_i^T\theta) \tag{2.4}$$

where $\mathcal{L}(y, f)$ is a general loss function; this includes a broad range of estimators for problems such as linear least squares and logistic regression, robust regression, support vector machines, and others. The gradient descent update rule is

$$\theta^{(t+1)} = \theta^{(t)} - \gamma_t \frac{1}{n}\sum_{i=1}^n \nabla_\theta \mathcal{L}(y_i, x_i^T\theta^{(t-1)}) \tag{2.5}$$

$$= \theta^{(t)} - \gamma_t \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial f}\mathcal{L}(y_i, x_i^T\theta^{(t-1)})x_i \tag{2.6}$$

$$= \theta^{(t)} - \sum_{i=1}^n w_i^{(t)} x_i \tag{2.7}$$

where γ_t is a step size parameter. If the model is initialized at $\theta^{(0)} = 0 \in \mathbb{R}^p$ then the estimate at time t can thus be written as

$$\theta^{(t)} = X^T u^{(t)} \tag{2.8}$$

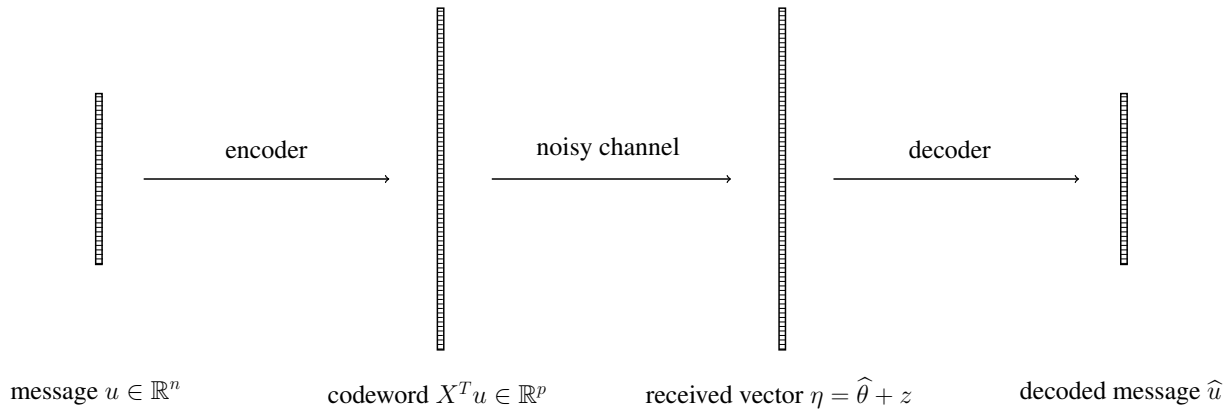

FIG 2. *Model repair viewed in terms of error-correcting codes. The model $\widehat{\theta} = X^T u \in \mathbb{R}^p$ is in the row-space of the design matrix, which gives a redundant representation of the "message" $u \in \mathbb{R}^n$ for $n < p$. The model is received as a noisy version η with each entry corrupted with probability ε. The received vector is decoded by solving a linear program.*

for some $u^{(t)} \in \mathbb{R}^n$. After contamination, we have $\eta = X^T u^{(t)} + z$. Therefore, we can recover the model by computing $\widetilde{\theta} = X^T \widetilde{u}$ where $\widetilde{u}$ is the solution to (2.3).

The same conclusion holds for estimators based on stochastic gradient descent. Let B_t be the set of samples used in the mini-batch of the tth iteration. Then, we can write

$$\theta^{(t)} = \sum_{i \in B_1 \cup \cdots \cup B_t} u_i^{(t)} x_i. \tag{2.9}$$

We can recover $\theta^{(t)}$ from a corrupted model η by computing $\widetilde{\theta} = X^T_{B_1 \cup \cdots B_t} \widetilde{u}$ with

$$\widetilde{u} = \arg\min_u \|\eta - X^T_{B_1 \cup \cdots B_t} u\|_1 \tag{2.10}$$

where the submatrix $X_{B_1 \cup \cdots B_t}$ only takes rows of X for indices that were visited during some stochastic gradient descent step. Our theory then establishes that the model is recovered with high probability in case

$$\frac{\sqrt{|B_1 \cup \cdots B_t|/p}}{1 - \varepsilon} < c. \tag{2.11}$$

Typically the training takes place in "epochs" where all n data points are visited in each epoch.

Random features and neural networks. Our theory extends to random features models (Rahimi and Recht, 2008), where the covariates are $\widetilde{X} = \psi(XW) \in \mathbb{R}^{n \times p}$ where $X \in \mathbb{R}^{n \times d}$, the matrix $W \in \mathbb{R}^{d \times p}$ is a random Gaussian matrix that is not trained, and ψ is a threshold function such as the hyperbolic tangent function or rectified linear unit. In particular, when the model is trained using gradient descent, the parameters $\widehat{\theta}$ lie in the row space of the matrix $\widetilde{X}$. We also show how the ideas can be extended to neural networks, where the weights W are trained. This requires modifications to the training and recovery algorithms that we detail below.

3. Simulation studies

4. Discussion

References

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Chen, M., Gao, C., and Ren, Z. (2016). A general decision theory for Huber's ε-contamination model. *Electron. J. Statist.*, 10(2):3752–3774.

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

Joseph, A. and Barron, A. R. (2012). Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Transactions on Information Theory*, 58(5):2541–2557.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rush, C., Greig, A., and Venkataramanan, R. (2017). Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Info. Theory*, 63(3):1476–1500.