Compressed Regression

Shuheng Zhou[*] **John Lafferty**[*†] **Larry Wasserman**[‡†]

[*]Computer Science Department
[‡]Department of Statistics
[†]Machine Learning Department

Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Recent research has studied the role of sparsity in high dimensional regression and signal reconstruction, establishing theoretical limits for recovering sparse models from sparse data. In this paper we study a variant of this problem where the original n input variables are compressed by a random linear transformation to $m \ll n$ examples in p dimensions, and establish conditions under which a sparse linear model can be successfully recovered from the compressed data. A primary motivation for this compression procedure is to anonymize the data and preserve privacy by revealing little information about the original data. We characterize the number of random projections that are required for ℓ_1-regularized compressed regression to identify the nonzero coefficients in the true model with probability approaching one, a property called "sparsistence." In addition, we show that ℓ_1-regularized compressed regression asymptotically predicts as well as an oracle linear model, a property called "persistence." Finally, we characterize the privacy properties of the compression procedure in information-theoretic terms, establishing upper bounds on the rate of information communicated between the compressed and uncompressed data that decay to zero.

1 Introduction

Two issues facing the use of statistical learning methods in applications are *scale* and *privacy*. Scale is an issue in storing, manipulating and analyzing extremely large, high dimensional data. Privacy is, increasingly, a concern whenever large amounts of confidential data are manipulated within an organization. It is often important to allow researchers to analyze data without compromising the privacy of customers or leaking confidential information outside the organization. In this paper we show that sparse regression for high dimensional data can be carried out directly on a compressed form of the data, in a manner that can be shown to guard privacy in an information theoretic sense.

The approach we develop here compresses the data by a random linear or affine transformation, reducing the number of data records exponentially, while preserving the number of original input variables. These compressed data can then be made available for statistical analyses; we focus on the problem of sparse linear regression for high dimensional data. Informally, our theory ensures that the relevant predictors can be learned from the compressed data as well as they could be from the original uncompressed data. Moreover, the actual predictions based on new examples are as accurate as they would be had the original data been made available. However, the original data are not recoverable from the compressed data, and the compressed data effectively reveal no more information than would be revealed by a completely new sample. At the same time, the inference algorithms run faster and require fewer resources than the much larger uncompressed data would require. The original data need not be stored; they can be transformed "on the fly" as they come in.

In more detail, the data are represented as a $n \times p$ matrix X. Each of the p columns is an attribute, and each of the n rows is the vector of attributes for an individual record. The data are compressed by a random linear transformation $X \mapsto \widetilde{X} \equiv \Phi X$, where Φ is a random $m \times n$ matrix with $m \ll n$. It is also natural to consider a random affine transformation $X \mapsto \widetilde{X} \equiv \Phi X + \Delta$, where Δ is a random $m \times p$ matrix. Such transformations have been called "matrix masking" in the privacy literature [6]. The entries of Φ and Δ are taken to be independent Gaussian random variables, but other distributions are possible. We think of $\widetilde{X}$ as "public," while Φ and Δ are private and only needed at the time of compression. However, even with $\Delta = 0$ and Φ known, recovering X from $\widetilde{X}$ requires solving a highly under-determined linear system and comes with information theoretic privacy guarantees, as we demonstrate.

In standard regression, a response variable $Y = X\beta + \epsilon \in \mathbb{R}^n$ is associated with the input variables, where ϵ_i are independent, mean zero additive noise variables. In compressed regression, we assume that the response is also compressed, resulting in the transformed response $\widetilde{Y} \in \mathbb{R}^m$ given by $Y \mapsto \widetilde{Y} \equiv \Phi Y = \Phi X\beta + \Phi\epsilon = \widetilde{X}\beta + \widetilde{\epsilon}$. Note that under compression, $\widetilde{\epsilon}_i, i \in \{1, \ldots, m\}$, in the transformed noise $\widetilde{\epsilon} = \Phi\epsilon$ are no longer independent. In the sparse setting, the parameter $\beta \in \mathbb{R}^p$ is sparse, with a relatively small number $s = \|\beta\|_0$ of nonzero coefficients in β. The method we focus on is ℓ_1-regularized least squares, also known as the lasso [17]. We study the ability of the compressed lasso estimator to identify the correct sparse set of relevant variables and to predict well.

We omit details and technical assumptions in the following theorems for clarity. Our first result shows that the lasso is *sparsistent* under compression, meaning that the correct sparse set of relevant variables is identified asymptotically.

Sparsistence (Theorem 3.3): *If the number of compressed examples m satisfies $C_1 s^2 \log nps \leq m \leq \sqrt{C_2 n/\log n}$, and the regularization parameter λ_m satisfies $\lambda_m \to 0$ and $m\lambda_m^2/\log p \to \infty$, then the compressed lasso estimator $\widetilde{\beta}_m = \arg\min_\beta \frac{1}{2m}\|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda_m\|\beta\|_1$ is sparsistent: $\mathbb{P}\left(supp(\widetilde{\beta}_m) = supp(\beta)\right) \to 1$ as $m \to \infty$, where* $\mathrm{supp}(\beta) = \{j : j \neq 0\}$.

Our second result shows that the lasso is *persistent* under compression. Roughly speaking, persistence [10] means that the procedure predicts well, as measured by the predictive risk $R(\beta) = \mathbb{E}\left(Y - \beta^T X\right)^2$, where $X \in \mathbb{R}^p$ is a new input vector and Y is the associated response. Persistence is a weaker condition than sparsistency, and in particular does not assume that the true model is linear.

Persistence (Theorem 4.1): *Given a sequence of sets of estimators $B_{n,m} \subset \mathbb{R}^p$ such that $B_{n,m} = \{\beta : \|\beta\|_1 \leq L_{n,m}\}$ with* $\log^2(np) \leq m \leq n$*, the sequence of compressed lasso estimators $\widetilde{\beta}_{n,m} = \arg\min_{\|\beta\|_1 \leq L_{n,m}} \|\widetilde{Y} - \widetilde{X}\beta\|_2^2$ is persistent with the predictive risk $R(\beta) = \mathbb{E}\left(Y - \beta^T X\right)^2$ over uncompressed data with respect to $B_{n,m}$, meaning that $R(\widetilde{\beta}_{n,m}) - \inf_{\|\beta\|_1 \leq L_{n,m}} R(\beta) \xrightarrow{P} 0$, as $n \to \infty$, in case $L_{n,m} = o\left(m/\log(np)\right)^{1/4}$.*

Our third result analyzes the privacy properties of compressed regression. We evaluate privacy in information theoretic terms by bounding the average mutual information $I(\widetilde{X}; X)/np$ per matrix entry in the original data matrix X, which can be viewed as a communication rate. Bounding this mutual information is intimately connected with the problem of computing the channel capacity of certain multiple-antenna wireless communication systems [13].

Information Resistence (Propositions 5.1 and 5.2): *The rate at which information about X is revealed by the compressed data $\widetilde{X}$ satisfies* $r_{n,m} = \sup \frac{I(X;\widetilde{X})}{np} = O\left(\frac{m}{n}\right) \to 0$*, where the supremum is over distributions on the original data X.*

As summarized by these results, compressed regression is a practical procedure for sparse learning in high dimensional data that has provably good properties. Connections with related literature are briefly reviewed in Section 2. Analyses of sparsistence, persistence and privacy properties appear in Section 3–5. Simulations for sparsistence and persistence of the compressed lasso are presented in Section 6. The proofs are included in the full version of the paper, available at `http://arxiv.org/abs/0706.0534`.

2 Background and Related Work

In this section we briefly review related work in high dimensional statistical inference, compressed sensing, and privacy, to place our work in context.

Sparse Regression. An estimator that has received much attention in the recent literature is the *lasso* $\widehat{\beta}_n$ [17], defined as $\widehat{\beta}_n = \arg\min \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n\|\beta\|_1$, where λ_n is a regularization parameter. In [14] it was shown that the lasso is consistent in the high dimensional setting under certain assumptions. Sparsistency proofs for high dimensional problems have appeared recently in [20] and [19]. The results and method of analysis of Wainwright [19], where X comes from a Gaussian ensemble and ϵ_i is i.i.d. Gaussian, are particularly relevant to the current paper. We describe this Gaussian Ensemble result, and compare our results to it in Sections 3, 6.Given that under compression, the noise $\widetilde{\epsilon} = \Phi\epsilon$ is not i.i.d, one cannot simply apply this result to the compressed case. Persistence for the lasso was first defined and studied by Greenshtein and Ritov in [10]; we review their result in Section 4.

Compressed Sensing. Compressed regression has close connections to, and draws motivation from compressed sensing [4, 2]. However, in a sense, our motivation is the opposite of compressed sensing. While compressed sensing of X allows a sparse X to be reconstructed from a small number of random measurements, our goal is to reconstruct a sparse function of X. Indeed, from the point of view of privacy, approximately reconstructing X, which compressed sensing shows is possible if X is sparse, should be viewed as undesirable; we return to this point in Section **??**. Several authors have considered variations on compressed sensing for statistical signal processing tasks [5, 11]. They focus on certain hypothesis testing problems under sparse random measurements, and a generalization to classification of a signal into two or more classes. Here one observes $y = \Phi x$, where $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ and Φ is a known random measurement matrix. The problem is to select between the hypotheses $\widetilde{H}_i : \; y = \Phi(s_i + \epsilon)$. The proofs use concentration properties of random projection, which underlie the celebrated Johnson-Lindenstrauss lemma. The compressed regression problem we introduce can be considered as a more challenging statistical inference task, where the problem is to select from an exponentially large set of linear models, each with a certain set of relevant variables with unknown parameters, or to predict as well as the best linear model in some class.

Privacy. Research on privacy in statistical data analysis has a long history, going back at least to [3]. We refer to [6] for discussion and further pointers into this literature; recent work includes [16]. The work of [12] is closely related to our work at a high level, in that it considers low rank random linear transformations of either the row space or column space of the data X. The authors note the Johnson-Lindenstrauss lemma, and argue heuristically that data mining procedures that exploit correlations or pairwise distances in the data are just as effective under random projection. The privacy analysis is restricted to observing that recovering X from $\widetilde{X}$ requires solving an under-determined linear system. We are not aware of previous work that analyzes the asymptotic properties of a statistical estimator under random projection in the high dimensional setting, giving information-theoretic guarantees, although an information-theoretic quantification of privacy was proposed in [1]. We cast privacy in terms of the rate of information communicated about X through $\widetilde{X}$, maximizing over all distributions on X, and identify this with the problem of bounding the Shannon capacity of a multi-antenna wireless channel, as modeled in [13]. Finally, it is important to mention the active area of cryptographic approaches to privacy from the theoretical computer science community, for instance [9, 7]; however, this line of work is quite different from our approach.

3 Compressed Regression is Sparsistent

In the standard setting, X is a $n \times p$ matrix, $Y = X\beta + \epsilon$ is a vector of noisy observations under a linear model, and p is considered to be a constant. In the high-dimensional setting we allow p to grow with n. The lasso refers to the following: $(P_1) \quad \min \quad \|Y - X\beta\|_2^2 \;$ such that $\; \|\beta\|_1 \leq L$. In Lagrangian form, this becomes: $(P_2) \quad \min \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n\|\beta\|_1$. For an appropriate choice of the regularization parameter $\lambda = \lambda(Y, L)$, the solutions of these two problems coincide.

In compressed regression we project each column $X_j \in \mathbb{R}^n$ of X to a subspace of m dimensions, using an $m \times n$ random projection matrix Φ. Let $\widetilde{X} = \Phi X$ be the compressed design matrix, and let $\widetilde{Y} = \Phi Y$ be the compressed response. Thus, the transformed noise $\widetilde{\epsilon}$ is no longer i.i.d.. The compressed lasso is the following optimization problem, for $\widetilde{Y} = \Phi X\beta + \Phi\epsilon = \Phi\widetilde{X} + \widetilde{\epsilon}$, with $\widetilde{\Omega}_m$ being the set of optimal solutions:

$$(a) \ \ (\widetilde{P}_2) \min \ \ \frac{1}{2m}\|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda_m\|\beta\|_1, (b) \ \ \widetilde{\Omega}_m = \arg\min_{\beta\in\mathbb{R}^p} \ \frac{1}{2m}\|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda_m\|\beta\|_1. \tag{1}$$

Although sparsistency is the primary goal in selecting the correct variables, our analysis establishes conditions for the stronger property of sign consistency:

Definition 3.1. (Sign Consistency) *A set of estimators Ω_n is sign consistent with the true β if $\mathbb{P}\left(\exists\widehat{\beta}_n \in \Omega_n \text{ s.t. } sgn(\widehat{\beta}_n) = sgn(\beta)\right) \to 1$ as $n \to \infty$, where $sgn(\cdot)$ is given by $sgn(x) = 1, 0,$ and -1 for $x >, =,$ or < 0 respectively. As a shorthand, denote the event that a sign consistent solution exists with $\mathcal{E}\left(sgn(\widehat{\beta}_n) = sgn(\beta^*)\right) := \left\{\exists\widehat{\beta} \in \Omega_n \text{ such that } sgn(\widehat{\beta}) = sgn(\beta^*)\right\}$.*

Clearly, if a set of estimators is sign consistent then it is sparsistent.

All recent work establishing results on sparsity recovery assumes some form of *incoherence condition* on the data matrix X. To formulate such a condition, it is convenient to introduce an additional piece of notation. Let $S = \{j : \beta_j \neq 0\}$ be the set of relevant variables and let $S^c = \{1, \ldots, p\} \setminus S$ be the set of irrelevant variables. Then X_S and X_{S^c} denote the corresponding sets of columns of the matrix X. We will impose the following incoherence condition; related conditions are used by [18] in a deterministic setting. Let $\|A\|_\infty = \max_i \sum_{j=1}^p |A_{ij}|$ denote the matrix ∞-norm.

Definition 3.2. (S-Incoherence) *Let X be an $n \times p$ matrix and let $S \subset \{1, \ldots, p\}$ be nonempty. We say that X is S-incoherent in case*

$$\left\|\tfrac{1}{n}X_{S^c}^T X_S\right\|_\infty + \left\|\tfrac{1}{n}X_S^T X_S - I_{|S|}\right\|_\infty \leq 1 - \eta, \quad \textit{for some } \eta \in (0, 1]. \tag{2}$$

Although not explicitly required, we only apply this definition to X such that columns of X satisfy $\left\|X_j\right\|_2^2 = \Theta(n), \forall j \in \{1, \ldots, p\}$. We can now state our main result on sparsistency.

Theorem 3.3. *Suppose that, before compression, $Y = X\beta^* + \epsilon$, where each column of X is normalized to have ℓ_2-norm n, and $\varepsilon \sim N(0, \sigma^2 I_n)$. Assume that X is S-incoherent, where $S = supp(\beta^*)$, and define $s = |S|$ and $\rho_m = \min_{i\in S}|\beta_i^*|$. We observe, after compression, $\widetilde{Y} = \widetilde{X}\beta^* + \widetilde{\epsilon}$, where $\widetilde{Y} = \Phi Y$, $\widetilde{X} = \Phi X$, and $\widetilde{\epsilon} = \Phi\epsilon$, where $\Phi_{ij} \sim N(0, \frac{1}{n})$. Let $\widetilde{\beta}_m \in \widetilde{\Omega}_m$ as in (1b). If*

$$\left(\frac{16C_1 s^2}{\eta^2} + \frac{4C_2 s}{\eta}\right)(\ln p + 2\log n + \log 2(s+1)) \leq m \leq \sqrt{\frac{n}{16\log n}} \tag{3}$$

with $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ and $C_2 = \sqrt{8e} \approx 7.6885$, and $\lambda_m \to 0$ satisfies

$$(a) \ \ \frac{m\eta^2\lambda_m^2}{\log(p-s)} \to \infty, \quad \textit{and} \ \ (b) \ \ \frac{1}{\rho_m}\left\{\sqrt{\frac{\log s}{m}} + \lambda_m\left\|(\tfrac{1}{n}X_S^T X_S)^{-1}\right\|_\infty\right\} \to 0. \tag{4}$$

Then the compressed lasso is sparsistent: $\mathbb{P}\left(supp(\widetilde{\beta}_m) = supp(\beta)\right) \to 1$ as $m \to \infty$.

4 Compressed Regression is Persistent

Persistence (Greenshtein and Ritov [10]) is a weaker condition than sparsistency. In particular, the assumption that $\mathbb{E}(Y|X) = \beta^T X$ is dropped. Roughly speaking, persistence implies that a procedure predicts well. We review the arguments in [10] first; we then adapt it to the compressed case.

Uncompressed Persistence. Consider a new pair (X, Y) and suppose we want to predict Y from X. The predictive risk using predictor $\beta^T X$ is $R(\beta) = \mathbb{E}(Y - \beta^T X)^2$. Note that this is a well-defined quantity even though we do not assume that $\mathbb{E}(Y|X) = \beta^T X$. It is convenient to rewrite the risk in the following way: define $Q = (Y, X_1, \ldots, X_p)$ and $\gamma = (-1, \beta_1, \ldots, \beta_p)^T$, then

$$R(\beta) = \gamma^T\Sigma\gamma, \ \text{ where } \Sigma = \mathbb{E}(QQ^T). \tag{5}$$

Let $\mathbb{Q} = (Q_1^\dagger \, Q_2^\dagger \, \cdots Q_n^\dagger)^T$, where $Q_i^\dagger = (Y_i, X_{1i}, \ldots, X_{pi})^T \sim Q, \forall i = 1, \ldots, n$ are i.i.d. random vectors and the training error is

$$\widehat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 = \gamma^T \widehat{\Sigma}^n \gamma, \text{ where } \widehat{\Sigma}^n = \frac{1}{n}\mathbb{Q}^T\mathbb{Q}. \tag{6}$$

Given $B_n = \{\beta : \|\beta\|_1 \leq L_n\}$ for $L_n = o\left((n/\log n)^{1/4}\right)$, we define the oracle predictor $\beta_{*,n} = \arg\min_{\|\beta\|_1 \leq L_n} R(\beta)$, and the uncompressed lasso estimator $\widehat{\beta}_n = \arg\min_{\|\beta\|_1 \leq L_n} \widehat{R}_n(\beta)$.

Assumption 1. Suppose that, for each j and k, $\mathbb{E}\left(|Z|^q\right) \leq q! M^{q-2} s/2$, for every $q \geq 2$ and some constants M and s, where $Z = Q_j Q_k - \mathbb{E}(Q_j Q_k)$, where Q_j, Q_k denote elements of Q.

Following arguments in [10], it can be shown that under Assumption 1 and given a sequence of sets of estimators $B_n = \{\beta : \|\beta\|_1 \leq L_n\}$ for $L_n = o\left((n/\log n)^{1/4}\right)$, the sequence of uncompressed lasso estimators $\widehat{\beta}_n = \arg\min_{\beta \in B_n} \widehat{R}_n(\beta)$ is persistent, i.e., $R(\widehat{\beta}_n) - R(\beta_{*,n}) \xrightarrow{P} 0$.

Compressed Persistence. For the compressed case, again we want to predict (X, Y), but now the estimator $\widehat{\beta}_{n,m}$ is based on the lasso from the compressed data of size m_n. Let $\gamma = (-1, \beta_1, \ldots, \beta_p)^T$ as before and we replace $\widehat{R}_n$ with

$$\widehat{R}_{n,m}(\beta) = \gamma^T \widehat{\Sigma}^{n,m} \gamma, \text{ where } \widehat{\Sigma}^{n,m} = \frac{1}{m_n}\mathbb{Q}^T \Phi^T \Phi \mathbb{Q}. \tag{7}$$

Given compressed sample size m_n, let $B_{n,m} = \{\beta : \|\beta\|_1 \leq L_{n,m}\}$, where $L_{n,m} = o\left(\frac{m_n}{\log(np_n)}\right)^{1/4}$. We define the compressed oracle predictor $\beta_{*,n,m} = \arg\min_{\beta : \|\beta\|_1 \leq L_{n,m}} R(\beta)$ and the compressed lasso estimator $\widehat{\beta}_{n,m} = \arg\min_{\beta : \|\beta\|_1 \leq L_{n,m}} \widehat{R}_{n,m}(\beta)$.

Theorem 4.1. *Under Assumption* 1*, we further assume that there exists a constant* $M_1 > 0$ *such that* $\mathbb{E}(Q_j^2) < M_1, \forall j$*, where* Q_j *denotes the* j^{th} *element of* Q*. For any sequence* $B_{n,m} \subset \mathbb{R}^p$ *with* $\log^2(np_n) \leq m_n \leq n$*, where* $B_{n,m}$ *consists of all coefficient vectors* β *such that* $\|\beta\|_1 \leq L_{n,m} = o\left((m_n/\log(np_n))^{1/4}\right)$*, the sequence of compressed lasso procedures* $\widehat{\beta}_{n,m} = \arg\min_{\beta \in B_{n,m}} \widehat{R}_{n,m}(\beta)$ *is persistent:* $R(\widehat{\beta}_{n,m}) - R(\beta_{*,n,m}) \xrightarrow{P} 0$*, when* $p_n = O\left(e^{n^c}\right)$ *for* $c < 1/2$*.*

The main difference between the sequence of compressed lasso estimators and the original uncompressed sequence is that n and m_n together define the sequence of estimators for the compressed data. Here m_n is allowed to grow from $\Omega(\log^2(np))$ to n; hence for each fixed n, $\left\{\widehat{\beta}_{n,m}, \forall m_n \text{ such that } \log^2(np) < m_n \leq n\right\}$ defines a subsequence of estimators. In Section 6 we illustrate the compressed lasso persistency via simulations to compare the empirical risks with the oracle risks on such a subsequence for a fixed n.

5 Information Theoretic Analysis of Privacy

Next we derive bounds on the rate at which the compressed data $\widetilde{X}$ reveal information about the uncompressed data X. Our general approach is to consider the mapping $X \mapsto \Phi X + \Delta$ as a noisy communication channel, where the channel is characterized by multiplicative noise Φ and additive noise Δ. Since the number of symbols in X is np we normalize by this effective block length to define the information rate $r_{n,m}$ per symbol as $r_{n,m} = \sup_{p(X)} \frac{I(X;\widetilde{X})}{np}$. Thus, we seek bounds on the capacity of this channel. A privacy guarantee is given in terms of bounds on the rate $r_{n,m} \to 0$ decaying to zero. Intuitively, if the mutual information satisfies $I(X; \widetilde{X}) = H(X) - H(X \mid \widetilde{X}) \approx 0$, then the compressed data $\widetilde{X}$ reveal, on average, no more information about the original data X than could be obtained from an independent sample.

The underlying channel is equivalent to the multiple antenna model for wireless communication [13], where there are n transmitter and m receiver antennas in a Raleigh flat-fading environment. The propagation coefficients between pairs of transmitter and receiver antennas are modeled by the matrix entries Φ_{ij}; they remain constant for a coherence interval of p time periods. Computing the channel capacity over multiple intervals requires optimization of the joint density of pn transmitted signals, the problem studied in [13]. Formally, the channel is modeled as $Z = \Phi X + \gamma\,\Delta$, where $\gamma > 0$, $\Delta_{ij} \sim N(0,1)$, $\Phi_{ij} \sim N(0, 1/n)$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_{ij}^2] \leq P$, where the latter is a power constraint.

Theorem 5.1. *Suppose that $E[X_j^2] \leq P$ and the compressed data are formed by $Z = \Phi X + \gamma\,\Delta$, where Φ is $m \times n$ with independent entries $\Phi_{ij} \sim N(0, 1/n)$ and Δ is $m \times p$ with independent entries $\Delta_{ij} \sim N(0,1)$. Then the information rate $r_{n,m}$ satisfies $r_{n,m} = \sup_{p(X)} \frac{I(X;Z)}{np} \leq \frac{m}{n} \log\left(1 + \frac{P}{\gamma^2}\right)$.*

This result is implicitly contained in [13]. When $\Delta = 0$, or equivalently $\gamma = 0$, which is the case assumed in our sparsistence and persistence results, the above analysis yields the trivial bound $r_{n,m} \leq \infty$. We thus derive a separate bound for this case; however, the resulting asymptotic order of the information rate is the same.

Theorem 5.2. *Suppose that $E[X_j^2] \leq P$ and the compressed data are formed by $Z = \Phi X$, where Φ is $m \times n$ with independent entries $\Phi_{ij} \sim N(0, 1/n)$. Then the information rate $r_{n,m}$ satisfies $r_{n,m} = \sup_{p(X)} \frac{I(X;Z)}{np} \leq \frac{m}{2n} \log\left(2\pi e P\right)$.*

Under our sparsistency lower bound on m, the above upper bounds are $r_{n,m} = O(\log(np)/n)$. We note that these bounds may not be the best possible since they are obtained assuming knowledge of the compression matrix Φ, when in fact the privacy protocol requires that Φ and Δ are not public.

6 Experiments

In this section, we report results of simulations designed to validate the theoretical analysis presented in previous sections. We first present results that show the compressed lasso is comparable to the uncompressed lasso in recovering the sparsity pattern of the true linear model. We then show results on persistence that are in close agreement with the theoretical results of Section 4. We only include Figures 1–2 here; additional plots are included in the full version.

Sparsistency. Here we run simulations to compare the compressed lasso with the uncompressed lasso in terms of the probability of success in recovering the sparsity pattern of β^*. We use random matrices for both X and Φ, and reproduce the experimental conditions of [19]. A design parameter is the *compression factor* $f = \frac{n}{m}$, which indicates how much the original data are compressed. The results show that when the compression factor f is large enough, the thresholding behaviors as specified in (8) and (9) for the uncompressed lasso carry over to the compressed lasso, when X is drawn from a Gaussian ensemble. In general, the compression factor f is well below the requirement that we have in Theorem 3.3 in case X is deterministic. In more detail, we consider the Gaussian ensemble for the projection matrix Φ, where $\Phi_{i,j} \sim N(0, 1/n)$ are independent. The noise is $\epsilon \sim N(0, \sigma^2)$, where $\sigma^2 = 1$. We consider Gaussian ensembles for the design matrix X with both diagonal and Toeplitz covariance. In the Toeplitz case, the covariance is given by $T(\rho)_{i,j} = \rho^{|i-j|}$; we use $\rho = 0.1$. [19] shows that when X comes from a Gaussian ensemble under these conditions, there exist fixed constants θ_ℓ and θ_u such that for any $\nu > 0$ and $s = \text{supp}(\beta)$, if

$$n > 2(\theta_u + \nu)s\log(p-s) + s + 1, \tag{8}$$

then the lasso identifies true variables with probability approaching one. Conversely, if

$$n < 2(\theta_\ell - \nu)s\log(p-s) + s + 1, \tag{9}$$

then the probability of recovering the true variables using the lasso approaches zero. In the following simulations, we carry out the lasso using procedure $\texttt{lars}(Y, X)$ that implements the LARS algorithm of [8] to calculate the full regularization path. For the uncompressed case, we run $\texttt{lars}(Y, X)$ such that $Y = X\beta^* + \epsilon$, and for the compressed case we run $\texttt{lars}(\Phi Y, \Phi X)$ such that $\Phi Y = \Phi X\beta^* + \Phi\epsilon$. The regularization parameter is $\lambda_m = c\sqrt{(\log(p-s)\log s)/m}$. The results show that the behavior under compression is close to the uncompressed case.

Persistence. Here we solve the following ℓ_1-constrained optimization problem $\widetilde{\beta} = \arg\min_{\|\beta\|_1 \leq L} \|Y - X\beta\|_2$ directly, based on algorithms described by [15]. We constrain the solution to lie in the ball $B_n = \{\|\beta\|_1 \leq L_n\}$, where $L_n = n^{1/4}/\sqrt{\log n}$. By [10], the uncompressed lasso

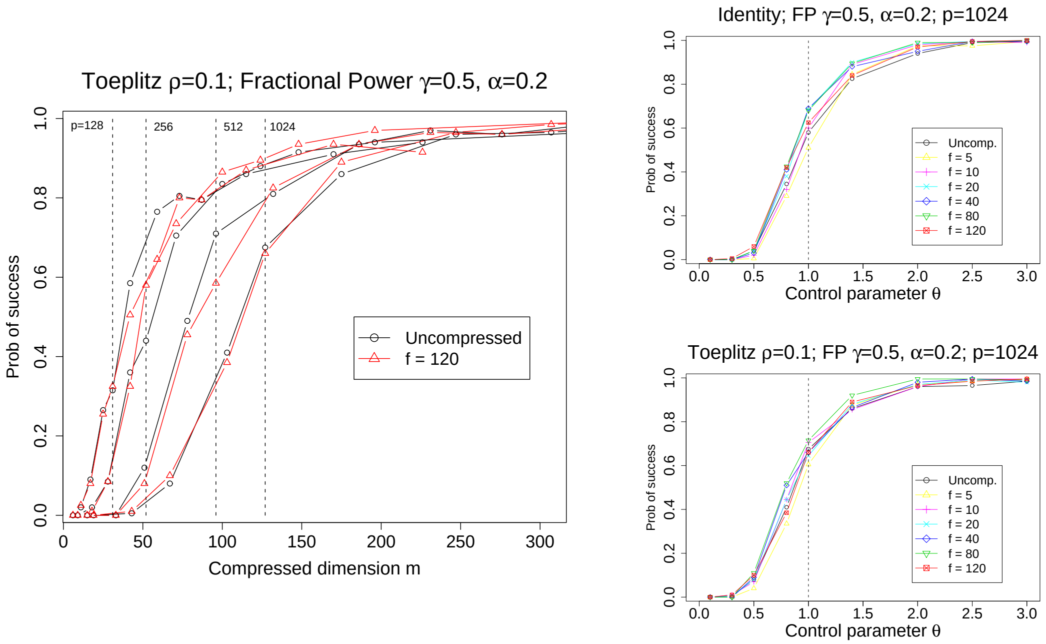

Figure 1: Plots of the number of samples versus the probability of success for recovering sgn(β^*). Each point on a curve for a particular θ or m, where $m = 2\theta\sigma^2 s \log(p-s) + s + 1$, is an average over 200 trials; for each trial, we randomly draw $X_{n\times p}$, $\Phi_{m\times n}$, and $\epsilon \in \mathbb{R}^n$. The covariance $\Sigma = \frac{1}{n}\mathbb{E}\left(X^T X\right)$ and model β^* are fixed across all curves in the plot. The sparsity level is $s(p) = 0.2p^{1/2}$. The four sets of curves in the left plot are for $p = 128, 256, 512$ and 1024, with dashed lines marking m for $\theta = 1$ and $s = 2, 3, 5$ and 6 respectively. In the plots on the right, each curve has a compression factor $f \in \{5, 10, 20, 40, 80, 120\}$ for the compressed lasso, thus $n = fm$; dashed lines mark $\theta = 1$. For $\Sigma = I$, $\theta_u = \theta_\ell = 1$, while for $\Sigma = T(0.1)$, $\theta_u \approx 1.84$ and $\theta_\ell \approx 0.46$ [19], for the uncompressed lasso in (8) and in (9).

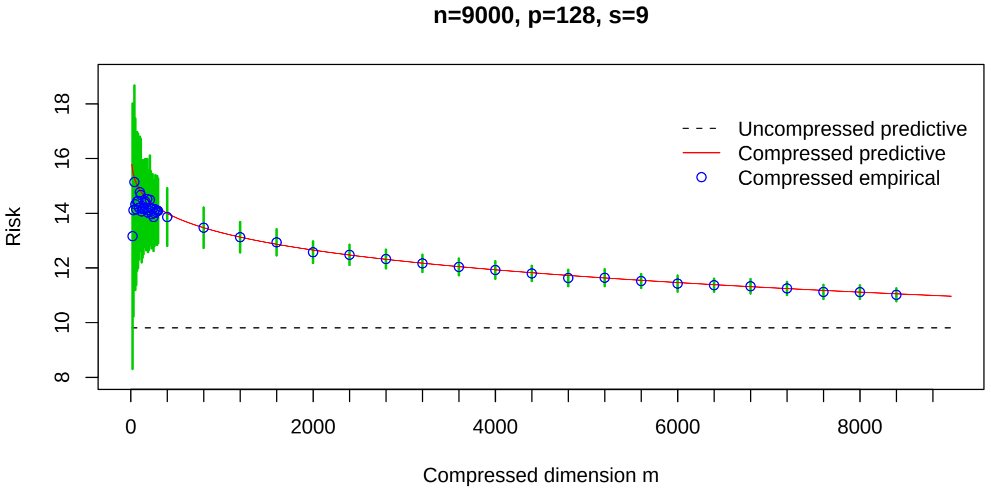

Figure 2: Risk versus compressed dimension. We fix $n = 9000$ and $p = 128$, and set $s(p) = 3$ and $L_n = 2.6874$. The model is $\beta^* = (-0.9, -1.7, 1.1, 1.3, -0.5, 2, -1.7, -1.3, -0.9, 0, \ldots, 0)^T$ so that $\left\|\beta_b^*\right\|_1 > L_n$ and $\beta_b^* \notin B_n$, and the uncompressed oracle predictive risk is $R = 9.81$. For each value of m, a data point corresponds to the mean empirical risk, which is defined in (7), over 100 trials, and each vertical bar shows one standard deviation. For each trial, we randomly draw $X_{n\times p}$ with i.i.d. row vectors $x_i \sim N(0, T(0.1))$, and $Y = X\beta^* + \epsilon$.

estimator $\widehat{\beta}_n$ is persistent over B_n. For the compressed lasso, given n and p_n, and a varying compressed sample size m, we take the ball $B_{n,m} = \{\beta : \|\beta\|_1 \leq L_{n,m}\}$ where $L_{n,m} = m^{1/4}/\sqrt{\log(np_n)}$. The compressed lasso estimator $\widehat{\beta}_{n,m}$ for $\log^2(np_n) \leq m \leq n$, is persistent over $B_{n,m}$ by Theorem 4.1. The simulations confirm this behavior.

7 Acknowlegments

This work was supported in part by NSF grant CCF-0625879.

References

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *In Proceedings of the 20th Symposium on Principles of Database Systems*, May 2001.

[2] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, 59(8):1207–1223, August 2006.

[3] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[4] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[5] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk. Sparse signal detection from incoherent projections. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2006.

[6] G. Duncan and R. Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, August 1991.

[7] C. Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming–ICALP 2006*, pages 1–12, 2006.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[9] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. J. Strauss, and R. N. Wright. Secure multiparty computation of approximations. *ACM Trans. Algorithms*, 2(3):435–472, 2006.

[10] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli*, 10:971–988, 2004.

[11] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, October 2006.

[12] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. on Knowl. and Data Engin.*, 18(1), Jan. 2006.

[13] T. L. Marzetta and B. M. Hochwald. Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Trans. Info. Theory*, 45(1):139–157, January 1999.

[14] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Technical Report 720, Department of Statistics, UC Berkeley, 2006.

[15] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *J. Comp. and Graph. Stat.*, 9(2):319–337, 2000.

[16] A. P. Sanil, A. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *Proceedings of Tenth ACM SIGKDD*, 2004.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[18] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[19] M. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical Report 709, Department of Statistics, UC Berkeley, May 2006.

[20] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Research*, 7:2541–2567, 2007.