Compressed and Privacy-Sensitive Sparse Regression

Shuheng Zhou, John Lafferty, *Fellow, IEEE*, and Larry Wasserman

***Abstract*—Recent research has studied the role of sparsity in high-dimensional regression and signal reconstruction, establishing theoretical limits for recovering sparse models. This line of work shows that ℓ_1-regularized least squares regression can accurately estimate a sparse linear model from noisy examples in high dimensions. 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 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." We also 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, establishing upper bounds on the mutual information between the compressed and uncompressed data that decay to zero.**

***Index Terms*—Capacity of multiple-antenna channels, compressed sensing, high-dimensional regression, lasso, ℓ_1 regularization, privacy, sparsity.**

I. 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 individuals 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.

Manuscript received June 04, 2007; revised April 04, 2008. Current version published February 04, 2009. This work was supported in part by the National Science Foundation under Grant CCF-0625879. The material in this paper was presented in part at The 21st Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, December 2007.

S. Zhou is with ETH Zürich, CH 8092 Zürich, Switzerland (e-mail: zhou@stat.math.ethz.ch).

J. Lafferty is with the Computer Science Department, Machine Learning Department, and Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: lafferty@cs.cmu.edu)

L. Wasserman is with the Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: larry@stat.cmu.edu).

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color versions of Figures 1 and 2 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.2009605

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. In fact, the original data need never 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

$$X \mapsto \tilde{X} \equiv \Phi X \tag{1}$$

where Φ is a random $m \times n$ matrix with $m \ll n$. It is also natural to consider a random affine transformation

$$X \mapsto \tilde{X} \equiv \Phi X + \Delta \tag{2}$$

where Δ is a random $m \times p$ matrix. Such transformations have been called "matrix masking" in the privacy literature [1]. The entries of Φ and Δ are taken to be independent Gaussian random variables, but other distributions are possible. We think of $\tilde{X}$ 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 $\tilde{X}$ requires solving a highly underdetermined linear system and comes with information-theoretic privacy guarantees, as we demonstrate.

In standard regression, a response $Y = X\beta + \epsilon \in \mathbb{R}^n$ 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 $\tilde{Y} \in \mathbb{R}^m$ given by

$$Y \mapsto \tilde{Y} \equiv \Phi Y \tag{3a}$$

$$= \Phi X\beta + \Phi\epsilon \tag{3b}$$

$$= \tilde{X}\beta + \tilde{\epsilon}. \tag{3c}$$

Note that under compression, the transformed noise $\tilde{\epsilon} = \Phi\epsilon$ is not independent across examples.

In the sparse setting, the parameter vector $\beta \in \mathbb{R}^p$ is sparse, with a relatively small number s of nonzero coefficients $\mathrm{supp}(\beta) = \{j : \beta_j \neq 0\}$. Two key tasks are to identify the relevant variables, and to predict the response $x^T\beta$ for a new input vector $x \in \mathbb{R}^p$. The method we focus on is ℓ_1-regularized least squares, also known as the lasso [2]. The main contributions of this paper are two technical results on the performance of this estimator, and an information-theoretic analysis of the privacy properties of the procedure. Our first result shows that the lasso is *sparsistent* under compression, meaning that the correct sparse set of relevant variables is identified asymptotically. Omitting details and technical assumptions for clarity, our result is the following.

Sparsistence (Theorem 3.4): If the number of compressed examples m satisfies

$$C_1 s^2 \log np s \leq m \leq \sqrt{\frac{C_2 n}{\log n}} \tag{4}$$

and the regularization parameter λ_m satisfies

$$\lambda_m \to 0 \quad \text{and} \quad \frac{m\lambda_m^2}{\log p} \to \infty \tag{5}$$

then the compressed lasso solution

$$\tilde{\beta}_m = \arg\min_{\beta} \frac{1}{2m}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_m \|\beta\|_1 \tag{6}$$

includes the correct variables, asymptotically

$$\mathbb{P}\left(\mathrm{supp}(\tilde{\beta}_m) = \mathrm{supp}(\beta)\right) \to 1. \tag{7}$$

Our second result shows that the lasso is *persistent* under compression. Roughly speaking, persistence [3] means that the procedure predicts well, as measured by the predictive risk

$$R(\beta) = \mathbb{E}\,(Y - X\beta)^2 \tag{8}$$

where now $X \in \mathbb{R}^p$ 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 $B_{n,m}$, the sequence of compressed lasso estimators

$$\tilde{\beta}_{n,m} = \arg\min_{\|\beta\|_1 \leq L_{n,m}} \|\tilde{Y} - \tilde{X}\beta\|_2^2 \tag{9}$$

is persistent with the oracle risk over uncompressed data with respect to $B_{n,m}$, meaning that

$$R(\tilde{\beta}_{n,m}) - \inf_{\|\beta\|_1 \leq L_{n,m}} R(\beta) \xrightarrow{P} 0, \quad \text{as} \quad n \to \infty \tag{10}$$

in case $\log^2(np) \leq m \leq n$ and the radius of the ℓ_1 ball satisfies $L_{n,m} = o\left(m/\log(np)\right)^{1/4}$.

Our third result analyzes the privacy properties of compressed regression. We consider the problem of recovering the uncompressed data X from the compressed data $\tilde{X} = \Phi X + \Delta$. To preserve privacy, the random matrices Φ and Δ should remain private. However, even in the case where $\Delta = 0$ and Φ is known, if $m \ll \min(n, p)$ the linear system $\tilde{X} = \Phi X$ is highly underdetermined. We evaluate privacy in information-theoretic terms by bounding the average mutual information $I(\tilde{X}; X)/np$ 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 [4], [5].

Information Resistance (Propositions 5.1 and 5.2): The rate at which information about X is revealed by the compressed data $\tilde{X}$ satisfies

$$r_{n,m} = \sup \frac{I(X; \tilde{X})}{np} = O\left(\frac{m}{n}\right) \to 0 \tag{11}$$

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. This basic technique has connections in the privacy literature with matrix masking and other methods, yet most of the existing work in this direction has been heuristic and without theoretical guarantees; connections with this literature are briefly reviewed in Section II-C. Compressed regression builds on the ideas underlying compressed sensing and sparse inference in high-dimensional data, topics which have attracted a great deal of recent interest in the statistics and signal processing communities; the connections with this literature are reviewed in Sections II-B and II-A.

The remainder of the paper is organized as follows. In Section II, we review relevant work from high-dimensional statistical inference, compressed sensing, and privacy. Section III presents our analysis of the sparsistency properties of the compressed lasso. Our approach follows the methods introduced by [6] in the uncompressed case. Section IV proves that compressed regression is persistent. Section V derives upper bounds on the mutual information between the compressed data $\tilde{X}$ and the uncompressed data X, after identifying a correspondence with the problem of computing channel capacity for a certain model of a multiple-antenna mobile communication channel. Section VI includes the results of experimental simulations, showing that the empirical performance of the compressed lasso is consistent with our theoretical analysis. We evaluate the ability of the procedure to recover the relevant variables (sparsistency) and to predict well (persistence). The technical details of the proof of sparsistency are collected at the end of the paper, in Section VII-B. The paper concludes with a discussion of the results and directions for future work in Section VIII.

II. Background and Related Work

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

A. Sparse Regression

We adopt standard notation where a data matrix X has p variables and n records; in a linear model, the response $Y =$

$X\beta + \epsilon \in \mathbb{R}^n$ is thus an n-vector, and the noise ϵ_i is independent and mean zero, $\mathbb{E}(\epsilon) = 0$. The usual estimator of β is the least squares estimator

$$\widehat{\beta} = (X^T X)^{-1} X^T Y. \tag{12}$$

However, this estimator has very large variance when p is large, and is not even defined when $p > n$. An estimator that has received much attention in the recent literature is the *lasso* $\widehat{\beta_n}$ [2], defined as

$$\widehat{\beta_n} = \arg\min \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \tag{13a}$$

$$= \arg\min \frac{1}{2n} ||Y - X\beta||_2^2 + \lambda_n ||\beta||_1 \tag{13b}$$

where λ_n is a regularization parameter. The practical success and importance of the lasso can be attributed to the fact that in many cases β is sparse, that is, it has few large components. For example, data are often collected with many variables in the hope that at least a few will be useful for prediction. The result is that many covariates contribute little to the prediction of Y, although it is not known in advance which variables are important. Recent work has greatly clarified the properties of the lasso estimator in the high-dimensional setting.

One of the most basic desirable properties of an estimator is consistency; an estimator $\widehat{\beta_n}$ is *consistent* in case

$$||\widehat{\beta_n} - \beta||_2 \xrightarrow{P} 0. \tag{14}$$

The authors of [7] have recently shown that the lasso is consistent in the high-dimensional setting. If the underlying model is sparse, a natural yet more demanding criterion is to ask that the estimator correctly identify the relevant variables. This may be useful for interpretation, dimension reduction, and prediction. For example, if an effective procedure for high-dimensional data can be used to identify the relevant variables in the model, then these variables can be isolated and their coefficients estimated by a separate procedure that works well for low-dimensional data. An estimator is *sparsistent*[1] if

$$\mathbb{P}\left(\operatorname{supp}(\widehat{\beta}_n) = \operatorname{supp}(\beta)\right) \to 1 \tag{15}$$

where $\operatorname{supp}(\beta) = \{j : j \neq 0\}$. Asymptotically, a sparsistent estimator has nonzero coefficients only for the true relevant variables. Sparsistency proofs for high-dimensional problems have appeared recently in a number of settings. In [8], the authors consider the problem of estimating the graph underlying a sparse Gaussian graphical model by showing sparsistency of the lasso with exponential rates of convergence on the probability of error. Zhou and Yu [9] show sparsistency of the lasso under more general noise distributions. Wainwright [6] characterizes the sparsistency properties of the lasso by showing that there is a threshold sample size $n(p, s)$ above which the relevant variables are identified, and below which the relevant variables fail to be identified, where $s = ||\beta||_0$ is the number of relevant variables. More precisely, [6] shows that when X comes from a Gaussian ensemble, there exist fixed constants $0 < \theta_\ell \leq 1$ and $1 \leq \theta_u < +\infty$, where $\theta_\ell = \theta_u = 1$ when each row of X is chosen as an independent Gaussian random vector $\sim N(0, I_{p\times p})$, then for any $\nu > 0$, if

$$n > 2(\theta_u + \nu) s \log(p - s) + s + 1 \tag{16}$$

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

$$n < 2(\theta_\ell - \nu) s \log(p - s) + s + 1 \tag{17}$$

then the probability of recovering the true variables using the lasso approaches zero. These results require certain *incoherence* assumptions on the data X; intuitively, it is required that an irrelevant variable cannot be too strongly correlated with the set of relevant variables. Wainwright's method [6] of analysis is particularly relevant to the current paper; the details will be described in the following section. In particular, we refer to this result as the Gaussian Ensemble result. However, it is important to point out that under compression, the noise $\tilde{\epsilon} = \Phi\epsilon$ is not independent. This prevents one from simply applying the Gaussian Ensemble results to the compressed case.

An alternative goal is accurate prediction. In high dimensions, it is essential to regularize the model in some fashion in order to control the variance of the estimator and attain good predictive risk. Persistence for the lasso was first defined and studied in [3]. Given a sequence of sets of estimators B_n, the sequence of estimators $\widehat{\beta}_n \in B_n$ is called *persistent* in case

$$R(\widehat{\beta}_n) - \inf_{\beta \in B_n} R(\beta) \xrightarrow{P} 0 \tag{18}$$

where $R(\beta) = \mathbb{E}(Y - X^T\beta)^2$ is the prediction risk of a new pair (X, Y). Thus, a sequence of estimators is persistent if it asymptotically predicts as well as the oracle within the class, which minimizes the population risk; it can be achieved under weaker assumptions than are required for sparsistence. In particular, persistence does not assume the true model is linear, and it does not require strong incoherence assumptions on the data. The results of the current paper show that sparsistence and persistence are preserved under compression.

B. Compressed Sensing

Our work has connections to compressed sensing [10]–[13]. However, in a sense, our motivation here is the opposite to that 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 V.

Several authors have considered variations on compressed sensing for statistical signal processing tasks [14]–[17]. The focus of this work is to consider 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

$$\tilde{H}_i : \; y = \Phi(s_i + \epsilon) \tag{19}$$

[1]This terminology is due to Pradeep Ravikumar.

where $\epsilon \in \mathbb{R}^n$ is additive Gaussian noise. Importantly, the setup exploits the "universality" of the matrix Φ, which is not selected with knowledge of s_i. The proof techniques use concentration properties of random projection, which underlie the celebrated Johnson–Lindenstrauss Lemma [18]. 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. Moreover, a key motivation for compressed regression is privacy; if privacy is not a concern, simple subsampling of the data matrix could be an effective compression procedure.

C. Privacy

A typical type of scenario we envision our framework being applied to is where one wishes to perform a regression analysis of medical data without revealing detailed information about individual members of the population. For example, it may be of interest to analyze which genes are relevant to a particular disease, using a database of gene expression profiles collected from microarrays, but it is desirable not to make public the full gene profiles of individuals in the database. Under compression, only a small number of random averages of the individual gene profiles is revealed.

Research on privacy in statistical data analysis has a long history, going back at least to [19]; we refer to [1] for discussion and further pointers into this literature. The compression method we employ has been called *matrix masking* in the privacy literature. In the general method, the $n \times p$ data matrix X is transformed by premultiplication, postmultiplication, and addition into a new $m \times q$ matrix

$$\tilde{X} = AXB + C. \tag{20}$$

The transformation A operates on data records for fixed covariates, and the transformation B operates on covariates for a fixed record. The method encapsulated in this transformation is quite general, and allows the possibility of deleting records, suppressing subsets of variables, data swapping, and including simulated data. In our use of matrix masking, we transform the data by replacing each variable with a relatively small number of random averages of the instances of that variable in the data. In other work, the authors in [20] consider the problem of privacy-preserving regression analysis in distributed data, where different variables appear in different databases but it is of interest to integrate data across databases. The recent work of [21] considers random orthogonal mappings $X \mapsto RX = \tilde{X}$ where R is a random rotation (rank n), designed to preserve the sufficient statistics of a multivariate Gaussian and therefore allow regression estimation, for instance. This use of matrix masking does not share the information-theoretic guarantees we present in Section V. We are not aware of previous work that analyzes the asymptotic properties of a statistical estimator under matrix masking in the high-dimensional setting.

Our setting differs from the classical information-theoretic scenarios for private communication. Shannon [22] formalized the notion of communication with perfect security in information-theoretic terms. If Alice sends a k-bit message M to Bob across a channel via an encoded n-bit message X, then the transmission is secure if the mutual information satisfies $I(X; M) = 0$. Thus, Alice and Bob need to share a k-bit key. Wyner [23] introduced the wiretap channel, where Bob receives a noisy version Y of the message X, but an eavesdropper Eve also receives a noisy version Z of message X through a different channel. The communication is considered to be secure as long as $\frac{1}{n} I(M; Z) \to 0$, and reliable as long as $\mathbb{P}(\hat{X}(Y) \,|\, Y) \to 1$ for some decoder $\hat{X}(Y)$. The capacity of such a channel is the largest value of the rate k/n for which these competing goals are possible. In our setting the goal is to estimate the vector β where $Y = X^T\beta + \epsilon$, from noisy observations $\tilde{Y}$ and $\tilde{X}$, in such a way that the mutual information $\frac{1}{n}\, I(\tilde{X}; X) \to 0$, is asymptotically vanishing if $p = O(1)$.

The work in [24] is closely related to the current paper 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 in [24] note the Johnson–Lindenstrauss lemma, which implies that ℓ_2 norms are approximately preserved under random projection, and argue heuristically that data mining procedures that exploit correlations or pairwise distances in the data, such as principal components analysis and clustering, are just as effective under random projection. The privacy analysis is restricted to observing that recovering X from $\tilde{X}$ requires solving an underdetermined linear system, and arguing that this prevents the exact values from being recovered. In our work, we identify privacy with the rate of information communicated about X through $\tilde{X}$ under matrix masking, maximizing over all distributions on X. We furthermore identify this with the problem of computing, or bounding, the Shannon capacity of a multiple-antenna wireless communication channel, as modeled by [5] and [4]. A related information-theoretic quantification of privacy was formulated by [25].

Finally, we mention the currently active line of work on cryptographic approaches to privacy, which have come mainly from the theoretical computer science community. Dwork [26] revisits the notion of privacy formulated by Dalenius [27], which intuitively demands that nothing can be learned about an individual record in a database that cannot be learned without access to the database. An impossibility result is given which shows that, appropriately formalized, this strong notion of privacy cannot be achieved. An alternative notion of *differential privacy* is proposed, which allows the probability of a disclosure of private information to change by only a small multiplicative factor, depending on whether or not an individual participates in the database. This line of work has recently been built upon in [28], with connections to compressed sensing, showing that any method that gives accurate answers to a large fraction of randomly generated subset sum queries must violate privacy.

III. Compressed Regression is Sparsistent

In the standard setting, X is an $n \times p$ 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 quadratic program:

$$(P_1) \text{ minimize } \|Y - X\beta\|_2^2 \text{ such that } \|\beta\|_1 \le L. \tag{21}$$

In Lagrangian form, this becomes the optimization problem

$$(P_2)\ \text{minimize}\ \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n\|\beta\|_1 \tag{22}$$

where the scaling factor $1/2n$ is chosen by convention and convenience. 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. We shall assume that the entries of Φ are independent Gaussian random variables

$$\frac{\Phi_{ij} \sim N(0,1}{n).} \tag{23}$$

Let $\tilde{X} = \Phi X$ be the compressed matrix of covariates, and let $\tilde{Y} = \Phi Y$ be the compressed response. Our objective is to estimate β in order to determine the relevant variables, or to predict well. The compressed lasso is the optimization problem, for $\tilde{Y} = \Phi X\beta + \Phi\epsilon = \Phi\tilde{X} + \tilde{\epsilon}$

$$(\tilde{P}_2)\ \text{minimize}\ \frac{1}{2m}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_m\|\beta\|_1, \tag{24}$$

with $\tilde{\Omega}_m$ being the set of optimal solutions

$$\tilde{\Omega}_m = \arg\min_{\beta\in\mathbb{R}^p}\ \frac{1}{2m}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_m\|\beta\|_1. \tag{25}$$

Thus, the transformed noise $\tilde{\epsilon}$ is no longer independent and identically distributed (i.i.d.), a fact that complicates the analysis. It is convenient to formalize the model selection problem using the following definitions.

Definition 3.1 (Sign Consistency): A set of estimators Ω_n is sign consistent with the true β if

$$\mathbb{P}\left(\exists\widehat{\beta}_n \in \Omega_n \text{ s.t. } \operatorname{sgn}(\beta_n) = \operatorname{sgn}(\widehat{\beta})\right) \to 1,\ \text{as } n \to \infty \tag{26}$$

where $\operatorname{sgn}(\cdot)$ is given by

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0. \end{cases} \tag{27}$$

As a shorthand, we use

$$\mathcal{E}\left(\operatorname{sgn}(\widehat{\beta}_n) = \operatorname{sgn}(\beta^*)\right) := \left\{\exists\widehat{\beta} \in \Omega_n \text{ s.t. } \operatorname{sgn}(\widehat{\beta}) = \operatorname{sgn}(\beta^*)\right\}$$

to denote the event that a sign consistent solution exists.

The lasso objective function is convex in β, and strictly convex for $p \le n$. Therefore, the set of solutions to the lasso and compressed lasso (24) is convex: if $\widehat{\beta}$ and $\widehat{\beta}'$ are two solutions, then by convexity $\widehat{\beta} + \rho(\widehat{\beta}' - \widehat{\beta})$ is also a solution for any $\rho \in [0, 1]$.

Definition 3.2 (Sparsistency): A set of estimators Ω_n is sparsistent with the true β if

$$\mathbb{P}\left(\exists\widehat{\beta}_n \in \Omega_n \text{ s.t. } \operatorname{supp}(\hat{\beta}_n) = \operatorname{supp}(\beta)\right) \to 1,\ \text{as } n \to \infty. \tag{28}$$

Clearly, if a set of estimators is sign consistent then it is sparsistent. Although sparsistency is the primary goal in selecting the correct variables, our analysis establishes conditions for the slightly stronger property of sign consistency.

All recent work establishing results on sparsity recovery assumes some form of *incoherence condition* on the data matrix X. Such a condition ensures that the irrelevant variables are not too strongly correlated with the relevant variables. Intuitively, without such a condition the lasso may be subject to false positives and negatives, where a relevant variable is replaced by a highly correlated relevant variable. 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 [29] and [30] in a deterministic setting.

Definition 3.3 (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 if for some $\eta \in (0, 1]$

$$\left\|\frac{1}{n}X_{S^c}^T X_S\right\|_\infty + \left\|\frac{1}{n}X_S^T X_S - I_{|S|}\right\|_\infty \le 1 - \eta \tag{29}$$

where $\|A\|_\infty = \max_i \sum_{j=1}^p |A_{ij}|$ denotes the matrix ∞-norm.

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

Theorem 3.4: Suppose that, before compression, we have $Y = X\beta^* + \epsilon$, where each column of X is normalized to have ℓ_2-norm n, and $\epsilon \sim N(0, \sigma^2 I_n)$. Assume that X is S-incoherent, where $S = \operatorname{supp}(\beta^*)$, and define $s = |S|$ and $\rho_m = \min_{i\in S}|\beta_i^*|$. We observe, after compression, that

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\epsilon} \tag{30}$$

where $\tilde{Y} = \Phi Y$, $\tilde{X} = \Phi X$, and $\tilde{\epsilon} = \Phi\epsilon$, where $\Phi_{ij} \sim N(0, 1\frac{}{n})$. Suppose

$$\left(\frac{16C_1 s^2}{\eta^2} + \frac{4C_2 s}{\eta}\right)(\ln p + \log 2n^2(s+1)) \le m \le \sqrt{\frac{n}{16\log n}} \tag{31}$$

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

$$\text{(a)}\quad \frac{m\eta^2\lambda_m^2}{\log(p-s)} \to \infty \text{ and}$$
$$\text{(b)}\quad \frac{1}{\rho_m}\left\{\sqrt{\frac{\log s}{m}} + \lambda_m\left\|\left(\frac{1}{n}X_S^T X_S\right)^{-1}\right\|_\infty\right\} \to 0. \tag{32}$$

Then the compressed lasso is sparsistent

$$\mathbb{P}\left(\exists\tilde{\beta}_m \in \tilde{\Omega}_m \text{ s.t. } \operatorname{supp}(\tilde{\beta}_m) = \operatorname{supp}(\beta)\right) \to 1 \text{ as } m \to \infty \tag{33}$$

where $\tilde{\beta}_m$ is an optimal solution to (24).

Note that an appropriate choice of the regularization parameter λ_m is

$$\lambda_m = c\sqrt{\frac{\log(p-s)\log m}{m}}. \tag{34}$$

We remark that the upper bound on m is required to bound the norm of the deviation $\Phi\Phi^T - I$. The lower bound on the compressed sample size m depends on s^2, the square of the number of relevant variables, rather than on s, as in the uncompressed case; however, our assumptions are significantly different. A detailed discussion of the relationship with the Gaussian ensemble results of [6] is given in Section VII-A.

A. Outline of Proof for Theorem 3.4

Our overall approach is to follow a deterministic analysis, in the sense that we analyze ΦX as a realization from the distribution of Φ from a Gaussian ensemble. Assuming that X satisfies the S-incoherence condition, we show that with high probability ΦX also satisfies the S-incoherence condition, and hence the incoherence conditions (106a) and (106b) used in [6]. In addition, we make use of a large-deviation result that shows $\Phi\Phi^T$ is concentrated around its mean $I_{m\times m}$, which is crucial for the recovery of the true sparsity pattern. It is important to note that the compressed noise $\tilde{\epsilon}$ is not i.i.d., even when conditioned on Φ.

In more detail, we first show that with high probability $1-n^{-c}$ for some $c \geq 2$, the projected data ΦX satisfies the following properties.

1) Each column of $\tilde{X} = \Phi X$ has ℓ_2-norm at most $m(1 + \eta/4s)$.
2) $\tilde{X}$ is S-incoherent, and also satisfies the incoherence conditions (106a) and (106b).

In addition, the projections satisfy the following properties.

1) Each entry of $\Phi\Phi^T - I$ is at most $\sqrt{b \log n/n}$ for some constant b, with high probability.
2) $\mathbb{P}\left(|\frac{n}{m}\langle \Phi x, \Phi y\rangle - \langle x, y\rangle| \geq \tau\right) \leq 2\exp\left(-\frac{m\tau^2}{C_1+C_2\tau}\right)$ for any $x, y \in \mathbb{R}^n$ with $\|x\|_2, \|y\|_2 \leq 1$.

These facts allow us to condition on a "good" Φ and incoherent ΦX, and to proceed as in the deterministic setting with Gaussian noise. Our analysis then follows that of [6]. Recall that S is the set of relevant variables in β and $S^c = \{1, \ldots, p\} \setminus S$ is the set of irrelevant variables. To explain the basic approach, first observe that the Karush–Kuhn–Tucker (KKT) conditions imply that $\tilde{\beta} \in \mathbb{R}^p$ is an optimal solution to (24) , i.e., $\tilde{\beta} \in \tilde{\Omega}_m$, if and only if there exists a subgradient

$$\tilde{z} \in \partial\|\tilde{\beta}\|_1 = \left\{z \in \mathbb{R}^p \,|\, z_i = \operatorname{sgn}(\tilde{\beta}_i), \text{ for } \tilde{\beta}_i \neq 0 \text{ and } |\tilde{z}_j| \leq 1 \text{ otherwise}\right\}$$

such that

$$\frac{1}{m}\tilde{X}^T\tilde{X}\tilde{\beta} - \frac{1}{m}\tilde{X}^T\tilde{Y} + \lambda\tilde{z} = 0. \tag{35}$$

Hence, the $\mathcal{E}\left(\operatorname{sgn}(\tilde{\beta}) = \operatorname{sgn}(\beta^*)\right)$ can be shown to be equivalent to requiring the existence of a solution $\tilde{\beta} \in \mathbb{R}^p$ such that $\operatorname{sgn}(\tilde{\beta}) = \operatorname{sgn}(\beta^*)$, and a subgradient $\tilde{z} \in \partial\|\tilde{\beta}\|_1$, such that the following equations hold:

$$\frac{1}{m}\tilde{X}_{S^c}^T\tilde{X}_S(\tilde{\beta}_S - \beta_S^*) - \frac{1}{m}\tilde{X}_{S^c}^T\tilde{\epsilon} = -\lambda\tilde{z}_{S^c} \tag{36a}$$

$$\frac{1}{m}\tilde{X}_S^T\tilde{X}_S(\tilde{\beta}_S - \beta_S^*) - \frac{1}{m}\tilde{X}_S^T\tilde{\epsilon} = -\lambda\tilde{z}_S \tag{36b}$$

where $\tilde{z}_S = \operatorname{sgn}(\beta_S^*)$ and $|\tilde{z}_{S^c}| \leq 1$ by definition of $\tilde{z}$. The existence of solutions to (36a) and (36b) can be characterized in terms of two events $\mathcal{E}(V)$ and $\mathcal{E}(U)$. The proof proceeds by showing that $\mathbb{P}(\mathcal{E}(V)) \longrightarrow 1$ and $\mathbb{P}(\mathcal{E}(U)) \longrightarrow 1$ as $m \longrightarrow \infty$.

In the remainder of this section, we present the main steps of the proof, relegating the technical details to Section VII-B. To avoid unnecessary clutter in notation, we will use Z to denote the compressed data $\tilde{X} = \Phi X$ and W to denote the compressed response $\tilde{Y} = \Phi Y$, and $\omega = \tilde{\epsilon}$ to denote the compressed noise.

B. Incoherence and Concentration Under Random Projection

In order for the estimated $\tilde{\beta}_m$ to be close to the solution of the uncompressed lasso, we require the stability of inner products of columns of X under multiplication with the random matrix Φ, in the sense that

$$\langle \Phi X_i, \Phi X_j\rangle \approx \langle X_i, X_j\rangle. \tag{37}$$

Toward this end, we have the following, adapted from [13], where for each entry in Φ, the variance is $\frac{1}{m}$ instead of $\frac{1}{n}$.

Lemma 3.5 (Adapted From [13]): Let $x, y \in \mathbb{R}^n$ with $\|x\|_2, \|y\|_2 \leq 1$. Assume that Φ is an $m \times n$ random matrix with independent $N(0, n^{-1})$ entries (independent of x, y). Then for all $\tau > 0$

$$\mathbb{P}\left(\left|\frac{n}{m}\langle \Phi x, \Phi y\rangle - \langle x, y\rangle\right| \geq \tau\right) \leq 2\exp\left(\frac{-m\tau^2}{C_1 + C_2\tau}\right) \tag{38}$$

with $C_1 = 4e/\sqrt{6\pi} \approx 2.5044$ and $C_2 = \sqrt{8e} \approx 7.6885$.

The proof follows the same reasoning as in [13], and is omitted. We next summarize the properties of ΦX that we require. The following result implies that, with high probability, incoherence is preserved under random projection.

Proposition 3.6: Let X be a (deterministic) design matrix that is S-incoherent with ℓ_2-norm n, and let Φ be an $m \times n$ random matrix with independent $N(0, n^{-1})$ entries. Suppose that

$$m \geq \left(\frac{16C_1s^2}{\eta^2} + \frac{4C_2s}{\eta}\right)(\ln p + c\ln n + \ln 2(s+1)) \tag{39}$$

for some $c \geq 2$, where C_1, C_2 are defined in Lemma 3.5. Then with probability at least $1 - \frac{1}{n^c}$, the following properties hold for $Z = \Phi X$:

1) Z is S-incoherent; in particular

$$\left|\left\|\frac{1}{m}Z_S^TZ_S - I_s\right\|_\infty - \left\|\frac{1}{n}X_S^TX_S - I_s\right\|_\infty\right| \leq \frac{\eta}{4} \tag{40a}$$

$$\left\|\frac{1}{m}Z_S^TZ_S - I_s\right\|_\infty + \left\|\frac{1}{m}Z_{S^c}^TZ_S\right\|_\infty \leq 1 - \frac{\eta}{2}. \tag{40b}$$

2) $Z = \Phi X$ is incoherent in the sense of (106a) and (106b)

$$\|Z_{S^c}^TZ_S\left(Z_S^TZ_S\right)^{-1}\|_\infty \leq 1 - \eta/2 \tag{41a}$$

$$\Lambda_{\min}\left(\frac{1}{m}Z_S^TZ_S\right) \geq \frac{3\eta}{4}. \tag{41b}$$

3) The ℓ_2 norm of each column is approximately preserved, for all j

$$|\|\Phi X_j\|_2^2 - m| \leq \frac{m\eta}{4s}. \tag{42}$$

Finally, we have the following large deviation result for the projection matrix Φ, which guarantees that $R = \Phi\Phi^T - I_{m\times m}$ is

small entry-wise. The proof of this result uses inequalities from [31] and [32] on χ^2 random variables and the sum of products of normals.

Theorem 3.7: If Φ is an $m \times n$ random matrix with independent entries $\Phi_{ij} \sim N(0, \frac{1}{n})$, then $R = \Phi\Phi^T - I$ satisfies

$$\mathbb{P}\left(\left\{\max_i |R_{ii}| \geq \sqrt{16 \log n/n}\right\} \cup \left\{\max_{i \neq j} |R_{ij}| \geq \sqrt{2 \log n/n}\right\}\right) \leq \frac{m^2}{n^3}.$$

C. Proof of Theorem 3.4

We first state necessary and sufficient conditions on the event $\mathcal{E}(\mathrm{sgn}(\tilde{\beta}_m) = \mathrm{sgn}(\beta^*))$. Note that this is essentially equivalent to Lemma 1 in [6]; a proof of this lemma is included in Section VII-D for completeness.

Lemma 3.8: Assume that the matrix $Z_S^T Z_S$ is invertible. Then for any given $\lambda_m > 0$ and noise vector $\omega \in \mathbb{R}^m$, $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m) = \mathrm{sgn}(\beta^*)\right)$ holds if and only if the following two conditions hold:

$$\left| Z_{S^c}^T Z_S (Z_S^T Z_S)^{-1} \left[\frac{1}{m} Z_S^T \omega - \lambda_m \mathrm{sgn}(\beta_S^*)\right] - \frac{1}{m} Z_{S^c}^T \omega \right| \leq \lambda_m \tag{43}$$

$$\mathrm{sgn}\left(\beta_S^* + \left(\frac{1}{m} Z_S^T Z_S\right)^{-1} \left[\frac{1}{m} Z_S^T \omega - \lambda_m \mathrm{sgn}(\beta_S^*)\right]\right) = \mathrm{sgn}(\beta_S^*). \tag{44}$$

Let $\vec{b} := \mathrm{sgn}(\beta_S^*)$ and $e_i \in \mathbb{R}^s$ be the vector with 1 in the ith position, and zeros elsewhere; hence , $\|e_i\|_2 = 1$. Our proof of Theorem 3.4 follows that of [6]. We first define a set of random variables that are relevant to (43) and (44)

$$\begin{aligned} \forall j \in S^c, \quad V_j &:= Z_j^T Z_S (Z_S^T Z_S)^{-1} \lambda_m \vec{b} \\ &\quad + Z_j^T \left\{I_{m \times m} - Z_S (Z_S^T \tilde{X}_S)^{-1} Z_S^T\right\} \frac{\omega}{m} \\ \forall i \in S, \quad U_i &:= e_i^T \left(\frac{1}{m} \tilde{X}_S^T Z_S\right)^{-1} \left[\frac{1}{m} Z_S^T \omega - \lambda_m \vec{b}\right]. \end{aligned}$$

Condition (43) holds if and if only the event

$$\mathcal{E}(V) := \left\{\max_{j \in S^c} |V_j| \leq \lambda_m\right\} \tag{45}$$

holds. Condition (44) holds if the event

$$\mathcal{E}(U) := \left\{\max_{i \in S} |U_i| \leq \rho_m\right\} \tag{46}$$

holds, where $\rho_m := \min_{i \in S} |\beta_i^*|$ is sufficient to guarantee that condition (44) holds.

Now, in the proof of Theorem 3.4, we assume that Φ has been fixed, and $Z = \Phi X$ and $\Phi\Phi^T$ behave nicely, in accordance with the results of Section III-B. Let $R = \Phi\Phi^T - I_{m \times m}$ as defined in Theorem 3.7. From here on, we use $(|r_{i,j}|)$ to denote a fixed symmetric matrix with diagonal entries that are $\sqrt{16 \log n/n}$ and off-diagonal entries that are $\sqrt{2 \log n/n}$. We now prove that $\mathbb{P}(\mathcal{E}(V))$ and $\mathbb{P}(\mathcal{E}(U))$ both converge to one. We begin by stating two technical lemmas that will be required.

Lemma 3.9: Suppose that $\|\frac{1}{n} X_S^T X_S - I_s\|_\infty$ is bounded away from 1 and

$$m \geq \left(\frac{16 C_1 s^2}{\eta^2} + \frac{4 C_2 s}{\eta}\right)(\log p + 2 \log n + \log 2(s+1)).$$

Then

$$\frac{1}{\rho_m}\left\{\sqrt{\frac{\log s}{m}} + \lambda_m \left\|\left(\frac{1}{n} X_S^T X_S\right)^{-1}\right\|_\infty\right\} \to 0 \text{ implies that}$$
$$\frac{1}{\rho_m}\left\{\sqrt{\frac{\log s}{m}} + \lambda_m \left\|\left(\frac{1}{m} Z_S^T Z_S\right)^{-1}\right\|_\infty\right\} \to 0.$$

Lemma 3.10 (Gaussian Comparison): For any Gaussian random vector $(X_1, \ldots, X_n)$

$$\mathbb{E}\left(\max_{1 \leq i \leq n} |X_i|\right) \leq 3\sqrt{\log n} \max_{1 \leq i \leq n} \sqrt{\mathbb{E}(X_i^2)}. \tag{47}$$

Analysis of $\mathcal{E}(V)$. Note that for each V_j, for $j \in S^c$, $\mu_j = \mathbb{E}(V_j) = \lambda_m Z_j^T Z_S (Z_S^T Z_S)^{-1} \vec{b}$. By Proposition 3.6, we have that $\forall j \in S^c$

$$\mu_j \leq \lambda_m \|Z_{S^c}^T Z_S \left(Z_S^T Z_S\right)^{-1}\|_\infty \leq \left(1 - \frac{\eta}{2}\right) \lambda_m.$$

Let us define

$$\tilde{V}_j = Z_j^T \left\{\left[I_{m \times m} - Z_S (Z_S^T Z_S)^{-1} Z_S^T\right] \frac{\omega}{m}\right\}$$

from which we obtain

$$\begin{aligned} \max_{j \in S^c} |V_j| &\leq \lambda_m \|Z_{S^c}^T Z_S \left(Z_S^T Z_S\right)^{-1}\|_\infty + \max_{j \in S^c} |\tilde{V}_j| \\ &\leq \lambda_m \left(1 - \frac{\eta}{2}\right) + \max_{j \in S^c} |\tilde{V}_j|. \end{aligned}$$

Hence, we need to show that

$$\mathbb{P}\left(\max_{j \in S^c} |\tilde{V}_j| / \lambda_m \geq \eta/2\right) \to 0.$$

It is sufficient to show

$$\mathbb{P}\left(\max_{j \in S^c} |\tilde{V}_j| \geq \eta \frac{\lambda_m}{2}\right) \to 0.$$

By Markov's inequality and the Gaussian Comparison Lemma 3.10, we obtain that

$$\begin{aligned} \mathbb{P}\left(\max_{j \in S^c} \tilde{V}_j \geq \lambda_m \frac{\eta}{2}\right) &\leq \frac{\mathbb{E}\left(\max_{j \in S^c} \tilde{V}_j\right)}{\lambda_m \frac{\eta}{2}} \\ &\leq \frac{6\sqrt{\log(p-s)}}{\lambda_m \eta} \max_{j \in S^c} \sqrt{\mathbb{E}\left(\tilde{V}_j^2\right)}. \end{aligned}$$

Finally, let us use

$$P = Z_S (Z_S^T Z y_S)^{-1} Z_S^T = P^2$$

to represent the projection matrix; see (48a)–(48d) at the bottom of the page, where $\|Z_j\|_2^2 \leq m + m\eta/4s$ by Proposition 3.6, and for $R = \Phi\Phi^T - I$

$$\begin{aligned}
&\|R - PR - RP + PRP\| \\
&\quad \leq \|R\|_2 + \|P\|_2\|R\|_2 + \|R\|_2\|P\|_2 + \|P\|_2\|R\|_2\|P\|_2 \\
&\quad \leq 4\|R\|_2 \leq 4\|(|r_{i,j}|)\|_2 \\
&\quad \leq 4(m+2)\sqrt{\frac{2\log n}{n}}
\end{aligned}$$

given that $\|I - P\|_2 \leq 1$ and $\|P\|_2 \leq 1$ and the fact that $(|r_{i,j}|)$ is a symmetric matrix

$$\begin{aligned}
\|R\|_2 &\leq \|(|r_{i,j}|)\|_2 \leq \sqrt{\|(|r_{i,j}|)\|_\infty \|(|r_{i,j}|)\|_1} \\
&= \|(|r_{i,j}|)\|_\infty \leq (m-1)\sqrt{\frac{2\log n}{n}} \\
&\quad + \sqrt{\frac{16\log n}{n}} \leq (m+2)\sqrt{\frac{2\log n}{n}}.
\end{aligned}$$

Consequently, condition (32a) is sufficient to ensure that $\mathbb{E}\left(\max_{j\in S^c}|\tilde{V}_j|\right)/\lambda_m \to 0$. Thus, $\mathbb{P}(\mathcal{E}(V)) \to 1$ as $m \to \infty$ so long as $m \leq \sqrt{n/2\log n}$.

Analysis of $\mathcal{E}(U)$. We now show that $\mathbb{P}(\mathcal{E}(U)) \to 1$. Using the triangle inequality, we obtain the upper bound

$$\max_{i\in S}|U_i| \leq \left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\omega\right\|_\infty + \left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\|_\infty \lambda_m.$$

The second ℓ_∞-norm is a fixed value given a deterministic ΦX. Hence, we focus on the first norm. We now define, for all $i \in S$, the Gaussian random variable

$$\begin{aligned}
G_i &= e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\omega \\
&= e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\Phi\epsilon.
\end{aligned}$$

Given that $\epsilon \sim N(0, \sigma^2 I_{n\times n})$, we have for all $i \in S$ that $\mathbb{E}(G_i) = 0$ and (49) also at the bottom of the page, where $R = \Phi\Phi^T - I$. We first bound the first term of (49). By (41b), we have that for all $i \in S$

$$\begin{aligned}
\frac{\sigma^2}{m}e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}e_i &\leq \frac{\sigma^2}{m}\left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\| \\
&= \frac{\sigma^2}{m\Lambda_{\min}\left(\frac{1}{m}Z_S^TZ_S\right)} \leq \frac{4\sigma^2}{3m\eta}. \quad (50)
\end{aligned}$$

We next bound the second term of (49). Let $M = CBC/m$, where $C = \left(\frac{1}{m}Z_S^TZ_S\right)^{-1}$ and $B = Z_S^TRZ_S$. By definition, $e_i = [e_{i,1}, \ldots, e_{i,s}] = [0, \ldots, 1, 0, \ldots]$, where $e_{i,i} = 1, e_{i,j} = 0, \forall j \neq i$. Thus, $\forall i \in S$

$$e_i^T\left\{\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^TRZ_S\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\}e_i$$

$$\begin{aligned}
\mathrm{Var}(\tilde{V}_j) &= \mathbb{E}\left(\tilde{V}_j^2\right) && (48a)\\
&= \frac{\sigma^2}{m^2}Z_j^T\left\{\left[(I_{m\times m} - P)\Phi\right]\left[(I_{m\times m} - P)\Phi\right]^T\right\}Z_j && (48b)\\
&= \frac{\sigma^2}{m^2}Z_j^T\left[I_{m\times m} - P\right]Z_j + \frac{\sigma^2}{m^2}Z_j^T(R - PR - RP + PRP)Z_j \\
&\leq \frac{\sigma^2}{m^2}\|Z_j\|_2^2 + \frac{\sigma^2}{m^2}\|R - PR - RP + PRP\|_2\|Z_j\|_2^2 && (48c)\\
&\leq \left(1 + 4(m+2)\sqrt{\frac{2\log n}{n}}\right)\frac{\sigma^2(1+\frac{\eta}{4s})}{m}, && (48d)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(G_i) &= \mathbb{E}\left(G_i^2\right) \\
&= \left\{e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\Phi\right\}\left\{e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\Phi\right\}^T\mathrm{Var}(\epsilon_i) \\
&= \frac{\sigma^2}{m}e_i^T\left\{\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\Phi\Phi^TZ_S\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\}e_i \\
&= \frac{\sigma^2}{m}e_i^T\left\{\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T(I+R)Z_S\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\}e_i \\
&= \frac{\sigma^2}{m}e_i^T\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}e_i + \frac{\sigma^2}{m}e_i^T\left\{\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^TRZ_S\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\}e_i \quad (49)
\end{aligned}$$

$$= \sum_{j=1}^{s}\sum_{k=1}^{s} e_{i,j}e_{i,k}M_{j,k} = M_{i,i}. \quad (51)$$

We next require the following fact.

Claim 3.11: If m satisfies (31), then for all $i \in S$, we have $\max_i M_{i,i} \leq (1+\eta/4s)\,(4/3\eta)^2$.

The proof appears in Section VII-F. Using Claim 3.11, we have by (50), (51) that

$$\max_{1\leq i\leq s}\sqrt{\mathbb{E}\,(G_i^2)} \leq \sqrt{\left(\frac{4\sigma}{3\eta}\right)^2\frac{1}{m}\left(\frac{3\eta}{4}+1+\frac{\eta}{4s}\right)} \leq \frac{4\sigma}{3\eta}\sqrt{\frac{1}{m}\left(1+\frac{3}{4}+\frac{1}{4s}\right)}.$$

By the Gaussian Comparison Lemma 3.10, we have

$$\mathbb{E}\left(\max_{1\leq i\leq s}|G_i|\right) = \mathbb{E}\left(\left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\frac{1}{m}Z_S^T\omega\right\|_\infty\right) \leq 3\sqrt{\log s}\max_{1\leq i\leq s}\sqrt{\mathbb{E}\,(G_i^2)} \leq \frac{4\sigma}{\eta}\sqrt{\frac{2\log s}{m}}.$$

We now apply Markov's inequality to show that $\mathbb{P}\,(\mathbb{E}(U)) \longrightarrow 1$ due to condition (32*b*) in the Theorem statement and Lemma 3.9, according to the expression at the bottom of the page, which completes the proof. □

IV. Compressed Regression is Persistent

Persistence [3] is a weaker condition than sparsistency. In particular, we drop the assumption that the model is linear $\mathbb{E}(Y|X) = \beta^TX$. Moreover, we do not require any incoherence assumptions on X. Roughly speaking, persistence implies that a procedure predicts well. More precisely, consider a new pair (X,Y) and suppose we want to predict Y from X. The predictive risk using predictor β^TX is

$$R(\beta) = \mathbb{E}(Y-\beta^TX)^2. \quad (52)$$

Note that this is a well-defined quantity even though we do not assume that g$\mathbb{E}(Y|X) = \beta^TX$. The result of [3] shows that

$$R(\widehat{\beta}_n) - \inf_{\|\beta\|_1\leq L_n} R(\beta) = o_P(1) \quad (53)$$

as long as $L_n = o((n/\log n)^{1/4})$. This follows from a uniform law of large numbers on the risk over the ℓ_1 ball $B_n = \{\beta : \|\beta\|_1 \leq L_n\}$, of the form

$$\sup_{\beta\in B_n}|R(\beta)-\hat{R}_n(\beta)| = O_P\left(L_n^2\sqrt{\frac{\log n}{n}}\right). \quad (54)$$

We show a similar result in the compressed setting. We use the estimator $\widehat{\beta}_{n,m}$ based on the lasso run on the compressed data of dimension m_n; we omit the subscript n from m_n wherever we put $\{n,m\}$ together. Define $Q = (Y, X_1, \ldots, X_{p_n})$ and denote γ as

$$\gamma = (-1, \beta_1, \ldots, \beta_{p_n})^T = (\beta_0, \beta_1, \ldots, \beta_{p_n})^T. \quad (55)$$

Then we can rewrite the risk as $R(\beta) = \gamma^T\Sigma\gamma$, where $\Sigma = \mathbb{E}(QQ^T)$. The training error in the uncompressed case is then $\hat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^n(Y_i - X_i^T\beta)^2 = \gamma^T\widehat{\Sigma}^n\gamma$, where

$$\widehat{\Sigma}^n = \frac{1}{n}\mathbb{Q}^T\mathbb{Q} \quad (56)$$

and $\mathbb{Q} = (Q_1^\dagger\, Q_2^\dagger\, \cdots\, Q_n^\dagger)^T$ where

$$Q_i^\dagger = (Y_i, X_{1i}, \ldots, X_{p_ni})^T, \quad \forall i = 1, \ldots, n$$

which are i.i.d. random vectors having the same distribution as Q. Now define

$$\widehat{\Sigma}^{n,m} = \frac{1}{m_n}\mathbb{Q}^T\Phi^T\Phi\mathbb{Q}. \quad (57)$$

In the compressed case, we replace the empirical risk $\hat{R}_n$ with

$$\hat{R}_{n,m}(\beta) = \gamma^T\widehat{\Sigma}^{n,m}\gamma. \quad (58)$$

Given compressed dimension m_n, the original design matrix dimension n and p_n, let

$$B_{n,m} = \{\beta : \|\beta\|_1 \leq L_{n,m}\}. \quad (59)$$

$$1-\mathbb{P}\left(\operatorname{sgn}\left(\beta_S^* + \left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\left[\frac{1}{m}Z_S^T\omega - \lambda_m\operatorname{sgn}(\beta_S^*)\right]\right) = \operatorname{sgn}(\beta_S^*)\right) \leq \mathbb{P}\left(\max_{i\in S}|U_i| \geq \rho_m\right) \leq \mathbb{P}\left(\max_{i\in S}|G_i| + \lambda_m\left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\|_\infty \geq \rho_m\right) \leq \frac{1}{\rho_m}\left(\mathbb{E}\left(\max_{i\in S}|G_i|\right) + \lambda_m\left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\|_\infty\right) \leq \frac{1}{\rho_m}\left(\frac{4\sigma}{\eta}\sqrt{\frac{2\log s}{m}} + \lambda_m\left\|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\right\|_\infty\right) \longrightarrow 0$$

Let β_* minimize $R(\beta)$ subject to $\beta \in B_{n,m}$

$$\beta_* = \underset{\beta:\|\beta\|_1 \le L_{n,m}}{\arg\min} \; R(\beta). \tag{60}$$

Consider the compressed lasso estimator $\widehat{\beta}_{n,m}$ which minimizes $\hat{R}_{n,m}(\beta)$ subject to $\beta \in B_{n,m}$

$$\widehat{\beta}_{n,m} = \underset{\beta:\|\beta\|_1 \le L_{n,m}}{\arg\min} \; \hat{R}_{n,m}(\beta). \tag{61}$$

Assumption 1: There exist constants M and s such that for every $q \ge 2$

$$\max_{1 \le j,k \le p_n+1} \mathbb{E}\left(|Q_jQ_k - \mathbb{E}(Q_jQ_k)|^q\right) \le q! M^{q-2} s/2. \tag{62}$$

This assumption allows the use of Bernstein's inequality, and is sufficient to show persistence in the uncompressed case.

Assumption 2: Let $Q_1, \ldots, Q_{p_n+1}$ denote the columns of $\mathbb{Q}$. Let $M_1 > 0$ be a constant such that

$$\mathbb{E}\left(\|Q_j\|_2^2\right) \le M_1 n, \;\; \forall j \in \{1, \ldots, p_n + 1\}. \tag{63}$$

Theorem 4.1: Under Assumptions 1 and 2, given a sequence of sets of estimators $B_{n,m} \subset \mathbb{R}^{p_n}$ for $\log^2(np_n) \le m_n \le n$, where $B_{n,m}$ consists of all coefficient vectors β such that $\|\beta\|_1 \le L_{n,m} = o\left((m_n/\log(np_n))^{1/4}\right)$, the sequence of compressed lasso procedures as in (103) is persistent

$$R(\widehat{\beta}_{n,m}) - R(\beta_*) \xrightarrow{P} 0 \tag{64}$$

when $p_n = O\left(e^{n^c}\right)$ for some $c < 1/2$.

Proof: First note that

$$\mathbb{E}\left(\widehat{\Sigma}^{n,m}\right) = \frac{1}{m_n}\mathbb{E}\left(\mathbb{Q}^T\mathbb{E}\left(\Phi^T\Phi\right)\mathbb{Q}\right) = \frac{1}{m_n}\mathbb{E}\left(\frac{m_n}{n}\mathbb{Q}^T\mathbb{Q}\right) = \Sigma.$$

We have that

$$\begin{aligned}\sup_{\beta \in B_{n,m}} \left|R(\beta) - \hat{R}_{n,m}(\beta)\right| &= \sup_{\beta \in B_{n,m}} \left|\gamma^T(\Sigma - \widehat{\Sigma}^{n,m})\gamma\right| \\ &\le (L_{n,m}+1)^2 \max_{j,k} \left|\widehat{\Sigma}^{n,m}_{jk} - \Sigma_{jk}\right|. \end{aligned} \tag{65}$$

We claim that, given $p_n = O\left(e^{n^c}\right)$ with $c < 1/2$ chosen so that $\log^2(np_n) \le m_n \le n$ holds, then

$$\max_{j,k} \left|\widehat{\Sigma}^{n,m}_{jk} - \Sigma_{jk}\right| = O_P\left(\sqrt{\frac{\log np_n}{m_n}}\right) \tag{66}$$

where $\Sigma = \frac{1}{n}\mathbb{E}\left(\mathbb{Q}^T\mathbb{Q}\right)$ is the same as (56), but (57) defines the matrix $\widehat{\Sigma}^{n,m}$. Hence, given $p_n = O\left(e^{n^c}\right)$ for some $c < 1/2$, combining (65) and (66), we have for $L_{n,m} = o\left((m_n/\log(np_n))^{1/4}\right)$ and $n \ge m_n \ge \log^2(np_n)$

$$\sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)| = o_P(1). \tag{67}$$

By the definition of $\beta_* \in B_{n,m}$ as in (60) and $\widehat{\beta}_{n,m} \in B_{n,m}$, we immediately have

$$|R(\widehat{\beta}_{n,m}) - R(\beta_*)| \le 2 \sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)| \tag{68}$$

given that

$$\begin{aligned} R(\beta_*) &\le R(\widehat{\beta}_{n,m}) && \text{(69a)} \\ &\le \hat{R}_{n,m}(\widehat{\beta}_{n,m}) + \sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)| && \text{(69b)} \\ &\le \hat{R}_{n,m}(\beta_*) + \sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)| && \text{(69c)} \\ &\le R(\beta_*) + 2 \sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)|. && \text{(69d)} \end{aligned}$$

Thus, for every $\epsilon > 0$, event $\left\{|R(\widehat{\beta}_{n,m}) - R(\beta_*)| > \epsilon\right\}$ is contained in

$$\left\{\sup_{\beta \in B_{n,m}} \left|R(\beta) - \hat{R}_{n,m}(\beta)\right| > \epsilon/2\right\}.$$

It follows that $\forall \epsilon > 0$, given $p_n = O\left(e^{n^c}\right)$ for some $c < 1/2$, $n \ge m_n \ge \log^2(np_n)$, and $L_{n,m} = o((m_n/\log(np_n))^{1/4})$

$$\begin{aligned}&\mathbb{P}\left(|R(\widehat{\beta}_{n,m}) - R(\beta_*)| > \epsilon\right) \\ &\quad \le \mathbb{P}\left(\sup_{\beta \in B_{n,m}} |R(\beta) - \hat{R}_{n,m}(\beta)| > \epsilon/2\right) \to 0, \text{ as } n \to \infty.\end{aligned}$$

Therefore, $R(\widehat{\beta}_{n,m}) - R(\beta_*) \xrightarrow{P} 0$. The theorem follows from the definition of persistence.

It remains to show (66). We first show the following claim; note that $p_n = O\left(e^{n^c}\right)$ with $c < 1/2$ clearly satisfies the condition.

Claim 4.2: Let $C = 2M_1$. Then

$$\mathbb{P}\left(\max_j \|Q_j\|_2^2 > Cn\right) < 1/n$$

so long as $p_n \le e^{c_1 M_1^2 n}/n$ for some chosen constant c_1 and M_1 satisfying (63),

Proof: let $A = (A_1, \ldots, A_n)^T$ denote a generic column vector of $\mathbb{Q}$. Let $\mu = \mathbb{E}(A_i^2)$. Under our assumptions, there exists $c_1 > 0$ such that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n V_i > t\right) \le e^{-nc_1t^2} \tag{70}$$

where $V_i = A_i^2 - \mu$. We have $C = 2M_1 \ge \mu + \sqrt{\log(np_n)/c_1 n}$ so long as $p_n \le e^{c_1 M_1^2 n}/n$. Then

$$\begin{aligned}\mathbb{P}\left(\sum_i A_i^2 > Cn\right) &\le \mathbb{P}\left(\sum_i (A_i^2 - \mu) > n\sqrt{\frac{\log(np_n)}{c_1 n}}\right) \\ &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n V_i > \sqrt{\frac{\log(np_n)}{c_1 n}}\right) < \frac{1}{np_n}.\end{aligned}$$

We have, with probability $1 - \frac{1}{n}$, that $\|Q_j\|_2 \le 2M_1 n, \forall j = 1, \ldots, p_n + 1$. The claim follows by the union bound for $C = 2M_1$. □

Thus, we assume that $\|Q_j\|_2^2 \le Cn$ for all j, and use the triangle inequality to bound the first expression at the

bottom of the page. We first compare each entry of $\widehat{\Sigma}^{n,m}_{jk}$ with that of $\frac{1}{n}\left(\mathbb{Q}^T\mathbb{Q}\right)_{j,k}$.

Claim 4.3: Assume that $\|Q_j\|_2^2 \leq Cn = 2M_1 n, \forall j$. By taking $\epsilon = C\sqrt{8C_1 \log(np_n)/m_n}$

$$\mathbb{P}\left(\max_{j,k}\left|\frac{1}{m_n}\langle \Phi Q_j, \Phi Q_k\rangle - \frac{1}{n}\langle Q_j, Q_k\rangle\right| \geq \frac{\epsilon}{2}\right) \leq \frac{1}{n^2} \quad (71)$$

where $C_1 = 4e/\sqrt{6\pi} \approx 2.5044$ as in Lemma 3.5 and C is defined in Claim 4.2.

Proof: Following arguments that appear before (117a), and by Lemma 3.5, it is straightforward to verify:

$$\mathbb{P}\left(|\frac{1}{m_n}\langle \Phi Q_j, \Phi Q_k\rangle - \frac{1}{n}\langle Q_j, Q_k\rangle| \geq \varepsilon\right) \leq 2\exp\left(\frac{-m_n\varepsilon^2}{C_1C^2 + C_2C\varepsilon}\right)$$

where $C_2 = \sqrt{8e} \approx 7.6885$ as in Lemma 3.5. There are at most $(p_n+1)p_n/2$ unique events given that both matrices are symmetric; the claim follows by the union bound. □

Using Bernstein's inequality, under Assumption 1, we have that

$$\mathbb{P}\left(\max_{j,k}\left|\widehat{\Sigma}^n_{jk} - \Sigma_{jk}\right| > \epsilon\right) \leq p_n^2 e^{-cn\epsilon^2} \leq e^{-cn\epsilon^2/2}. \quad (72)$$

We have by the union bound and (72), (71), Claim 4.2, and Claim 4.3, the second expression at the bottom of the page. Hence, given $p_n = O\left(e^{n^c}\right)$ with $c < 1/2$, by taking

$$\epsilon = \epsilon_{m,n} = O\left(\sqrt{\log(np_n)/m_n}\right)$$

we have

$$\mathbb{P}\left(\max_{jk}\left|\widehat{\Sigma}^{n,m}_{jk} - \Sigma_{jk}\right| > \epsilon\right) \leq \frac{2}{n} \to 0 \quad (73)$$

which completes the proof of the theorem. □

Remark 4.4: We can interpret the above result as quantifying the "cost of compression" in terms of the rate at which the excess risk converges to zero. For simplicity, suppose here that $L_n = O(1)$, $L_{n,m} = O(1)$, and $p_n = O(n^c)$. The result of [3] implies that the excess risk converges to zero at the rate

$$R(\widehat{\beta}_n) - \inf_{\|\beta\|_1 \leq L_n} R(\beta) = O_P\left(\sqrt{\frac{\log n}{n}}\right). \quad (74)$$

Our result above shows that, with compression in the regime where $m = \Omega(\log^2 n)$, the excess risk converges to zero at the much slower rate

$$R(\widehat{\beta}_{n,m}) - \inf_{\|\beta\|_1 \leq L_{n,m}} R(\beta) = O_P\left(\sqrt{\frac{1}{\log n}}\right). \quad (75)$$

The ratio of the uncompressed to compressed excess risk convergence rates is $\sqrt{m/n}$.

Remark 4.5: 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 $\Omega(\log^2(np_n))$ to n; hence for each fixed n

$$\left\{\widehat{\beta}_{n,m},\ \forall m_n \text{ such that } \log^2(np_n) < m_n \leq n\right\} \quad (76)$$

defines a subsequence of estimators. In Section VI, we run simulations that compare the empirical risk to the oracle risk on such a subsequence for a fixed n, to illustrate the compressed lasso persistency property.

V. Information-Theoretic Analysis of Privacy

In this section, we derive bounds on the rate at which the compressed data $\tilde{X}$ 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.

$$\max_{jk}|\widehat{\Sigma}^{n,m}_{jk} - \Sigma_{jk}| \leq \max_{jk}\left|\widehat{\Sigma}^{n,m}_{jk} - \left(\frac{1}{n}\mathbb{Q}^T\mathbb{Q}\right)_{jk}\right| + \max_{jk}\left|\left(\frac{1}{n}\mathbb{Q}^T\mathbb{Q}\right)_{jk} - \Sigma_{jk}\right| \quad \text{where}$$

$$\widehat{\Sigma}^{n,m} = \frac{1}{m_n}\begin{bmatrix} \|\Phi Y\|_2^2 & \langle \Phi Y, \Phi X_1\rangle & \dots & \langle \Phi Y, \Phi X_{p_n}\rangle \\ \langle \Phi X_1, \Phi Y\rangle & \|\Phi X_1\|_2^2 & \dots & \langle \Phi X_1, \Phi X_{p_n}\rangle \\ \dots & & & \\ \langle \Phi X_p, \Phi Y\rangle & \langle \Phi X_p, \Phi X_1\rangle & \dots & \|\Phi X_{p_n}\|_2^2 \end{bmatrix}_{(p_n+1)\times(p_n+1)}.$$

$$\begin{aligned}\mathbb{P}\left(\max_{jk}|\widehat{\Sigma}^{n,m}_{jk} - \Sigma_{jk}| > \epsilon\right) &\leq \mathbb{P}\left(\max_{jk}\left|\frac{\left(\mathbb{Q}^T\mathbb{Q}\right)_{jk}}{n} - \Sigma_{jk}\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\max_j \|Q_j\|_2^2 > Cn\right) \\ &\quad + \mathbb{P}\left(\max_{j,k}\left|\frac{\langle \Phi Q_j, \Phi Q_k\rangle}{m_n} - \frac{\langle Q_j, Q_k\rangle}{n}\right| \geq \frac{\epsilon}{2} \mid \max_j \|Q_j\|_2^2 \leq Cn\right) \\ &\leq e^{-cn\epsilon^2/8} + \frac{1}{n} + \frac{1}{n^2}.\end{aligned}$$

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

$$r_{n,m} = \sup_{p(X)} \frac{I(X;\tilde{X})}{np}. \tag{77}$$

Thus, we seek bounds on the capacity of this channel, where several independent blocks are coded. A privacy guarantee is given in terms of bounds on the rate $r_{n,m} \to 0$ decaying to zero. Intuitively, if $I(X;\tilde{X}) = H(X) - H(X \mid \tilde{X}) \approx 0$, then the compressed data $\tilde{X}$ reveal, on average, no more information about the original data X than could be obtained from an independent sample.

Our analysis yields the rate bound $r_{n,m} = O(m/n)$. Under the lower bounds on m in our sparsistency and persistence analyses, this leads to the information rates

$$r_{n,m} = O\left(\frac{\log(np)}{n}\right) \text{ (sparsistency)}$$
$$r_{n,m} = O\left(\frac{\log 2(np)}{n}\right) \text{ (persistence.)}$$

It is important to note, however, 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. Thus, it may be possible to show a faster rate of convergence to zero. We make this simplification since the capacity of the underlying communication channel does not have a closed form, and appears difficult to analyze in general. Conditioning on Φ yields the familiar Gaussian channel in the case of nonzero additive noise Δ.

In the following subsection, we first consider the case where additive noise Δ is allowed; this is equivalent to a multiple-antenna model in a Rayleigh flat-fading environment. While our sparsistency and persistence analysis has only considered $\Delta = 0$, additive noise is expected to give greater privacy guarantees. Thus, extending our regression analysis to this case is an important direction for future work. In Section V-B, we consider the case where $\Delta = 0$ with a direct analysis. This special case does not follow from analysis of the multiple antenna model.

A. Privacy Under the Multiple Antenna Channel Model

In the multiple-antenna model for wireless communication [4], [5], 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 pm transmitted signals. The authors of [4] prove that the capacity for $n > p$ is equal to the capacity for $n = p$, and is achieved when X factors as a product of a $p \times p$ isotropically distributed unitary matrix and a $p \times n$ random matrix that is diagonal, with nonnegative entries. They also show that as p gets large, the capacity approaches the capacity obtained as if the matrix of propagation coefficients Φ were known. Intuitively, this is because the transmitter could send several "training" messages used to estimate Φ, and then send the remaining information based on this estimate.

More formally, the channel is modeled as

$$Z = \Phi X + \gamma\Delta \tag{78}$$

where $\gamma > 0$, $\Delta_{ij} \sim N(0,1)$, $\Phi_{ij} \sim N(0, \frac{1}{n})$, and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_{ij}^2] \leq P$, where the latter is a power constraint. The compressed data are then conditionally Gaussian, with

$$\mathbb{E}(Z \mid X) = 0 \tag{79a}$$
$$\mathbb{E}(Z_{ij}Z_{kl} \mid X) = \delta_{ik}\left(\gamma^2\delta_{jl} + \sum_{t=1}^{n} X_{tj}X_{tl}\right). \tag{79b}$$

Thus, the conditional density $p(Z \mid X)$ is given by

$$p(Z \mid X) = \frac{\exp\left\{-\mathrm{tr}\left[\left(\gamma^2 \mathrm{I_p} + \mathrm{X^T X}\right)^{-1} \mathrm{Z^T Z}\right]\right\}}{(2\pi)^{pm/2}\det^{m/2}(\gamma^2 I_p + X^T X)} \tag{80}$$

which completely determines the channel. Note that this distribution does not depend on Φ, and the transmitted signal affects only the variance of the received signal.

The channel capacity is difficult to compute or accurately bound in full generality. However, an upper bound is obtained by assuming that the multiplicative coefficients Φ are known to the receiver. In this case, we have that $p(Z, \Phi \mid X) = p(\Phi)p(Z \mid \Phi, X)$, and the mutual information $I(Z, \Phi; X)$ is given by

$$I(Z,\Phi;X) = \mathbb{E}\left[\log \frac{p(Z,\Phi \mid X)}{p(Z,\Phi)}\right] \tag{81a}$$
$$= \mathbb{E}\left[\log \frac{p(Z \mid X,\Phi)}{p(Z \mid \Phi)}\right] \tag{81b}$$
$$= \mathbb{E}\left[\mathbb{E}\left[\log \frac{p(Z \mid X,\Phi)}{p(Z \mid \Phi)}\middle| \Phi\right]\right]. \tag{81c}$$

Now, conditioned on Φ, the compressed data $Z = \Phi X + \gamma\Delta$ can be viewed as the output of a standard additive noise Gaussian channel. We thus obtain the upper bound

$$\sup_{p(X)} I(Z;X) \leq \sup_{p(X)} I(Z,\Phi;X) \tag{82a}$$
$$= \mathbb{E}\left[\sup_{p(X)} \mathbb{E}\left[\log \frac{p(Z \mid X,\Phi)}{p(Z \mid \Phi)}\middle| \Phi\right]\right] \tag{82b}$$
$$\leq p\mathbb{E}\left[\log\det\left(I_m + \frac{P}{\gamma^2}\Phi\Phi^T\right)\right] \tag{82c}$$
$$\leq pm\log\left(1 + \frac{P}{\gamma^2}\right) \tag{82d}$$

where inequality (82c) comes from assuming the p columns of X are independent, and inequality (82d) uses Jensen's inequality and concavity of $\log\det S$. Summarizing, we've shown the following result.

Proposition 5.1: Suppose that $E[X_j^2] \leq P$ and the compressed data are formed by

$$Z = \Phi X + \gamma\Delta \tag{83}$$

where Φ is $m \times n$ with independent entries $\Phi_{ij} \sim N(0, \frac{1}{n})$ and Δ is $m \times p$ with independent entries $\Delta_{ij} \sim N(0,1)$. Then the information rate $r_{n,m}$ satisfies

$$r_{n,m} = \sup_{p(X)} \frac{I(X;Z)}{np} \leq \frac{m}{n} \log\left(1 + \frac{P}{\gamma^2}\right). \tag{84}$$

B. Privacy Under Multiplicative Noise

When $\Delta = 0$, or equivalently $\gamma = 0$, the above analysis yields the trivial bound $r_{n,m} \leq \infty$. Here we derive a separate bound for this case; the resulting asymptotic order of the information rate is the same, however.

Consider first the case where $p = 1$, so that there is a single column X in the data matrix. The entries are independently sampled as $X_i \sim F$ where F has mean zero and bounded variance $\mathrm{Var}(F) \leq P$. Let $Z = \Phi X \in \mathbb{R}^m$. An upper bound on the mutual information $I(X;Z)$ again comes from assuming the compression matrix Φ is known. In this case

$$I(Z, \Phi; X) = H(Z \,|\, \Phi) - H(Z \,|\, X, \Phi) \tag{85a}$$
$$= H(Z \,|\, \Phi) \tag{85b}$$

where the second conditional entropy in (85a) is zero since $Z = \Phi X$. Now, the conditional variance of $Z = (Z_1, \ldots, Z_m)^T$ satisfies

$$\mathrm{Var}(Z_i \,|\, \Phi) = \sum_{j=1}^{n} \Phi_{ij}^2 \,\mathrm{Var} X_j \leq P \sum_{j=1}^{n} \Phi_{ij}^2. \tag{86}$$

Therefore

$$I(Z, \Phi; X) = H(Z \,|\, \Phi) \tag{87a}$$
$$\leq \sum_{i=1}^{m} H(Z_i \,|\, \Phi) \tag{87b}$$
$$\leq \sum_{i=1}^{m} \left[\frac{1}{2} \log\left(2\pi e P \sum_{j=1}^{n} \Phi_{ij}^2\right)\right] \tag{87c}$$
$$\leq \sum_{i=1}^{m} \frac{1}{2} \log\left(2\pi e P \sum_{j=1}^{n} \mathbb{E}(\Phi_{ij}^2)\right) \tag{87d}$$
$$= \frac{m}{2} \log(2\pi e P) \tag{87e}$$

where inequality (87b) follows from the chain rule and the fact that conditioning reduces entropy, inequality (87c) is achieved by taking $F = N(0, P)$, a Gaussian, and inequality (87d) uses concavity of $\log\det S$. In the case where there are p columns of X, taking each column to be independently sampled from a Gaussian with variance P gives the upper bound

$$I(Z, \Phi; X) \leq \frac{mp}{2} \log(2\pi e P). \tag{88}$$

Summarizing, we have the following result.

Proposition 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 $\Phi_{ij} \sim N(0, \frac{1}{n})$. Then the information rate $r_{n,m}$ satisfies

$$r_{n,m} = \sup_{p(X)} \frac{I(X;Z)}{np} \leq \frac{m}{2n} \log(2\pi e P). \tag{89}$$

Note that our results from Section IV imply that the ratio of the uncompressed excess risk to the compressed excess risk decays at the rate $O\left(\sqrt{m/n}\right)$. Here we see that the information rate, assuming Φ is known, decays at the faster rate $O(m/n)$

.

VI. Experiments

In this section, we report the results of simulations designed to validate the theoretical analysis presented in the previous sections. We first present results that indicate the compressed lasso is comparable to the uncompressed lasso in recovering the sparsity pattern of the true linear model, in accordance with the analysis in Section III. We then present experimental results on persistence that are in close agreement with the theoretical results of Section IV.

A. 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 shown in [6]. A design parameter is the compression factor

$$f = \frac{n}{m} \tag{90}$$

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 (16) and (17) 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 3.4 in case X is deterministic.

In more detail, we consider the Gaussian ensemble for the projection matrix Φ, where $\Phi_{i,j} \sim N(0, \frac{1}{n})$ are independent. The noise vector is always composed of i.i.d. Gaussian random variables $\epsilon \sim N(0, \sigma^2)$, 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

$$T(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{p-1} & \rho^{p-1} \\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \rho & \ldots & \rho^{p-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho^{p-1} & \ldots & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}_{p \times p}. \tag{91}$$

We use $\rho = 0.1$. Both I and $T(0.1)$ satisfy conditions (107a) and (107b) and (109) [9]. For $\Sigma = I$, $\theta_u = \theta_\ell = 1$, while for $\Sigma = T(0.1)$, $\theta_u \approx 1.84$ and $\theta_\ell \approx 0.46$ [6], for the uncompressed lasso in (16) and in (17).

In the following simulations, we carry out the lasso using procedure $\mathsf{lars}(Y, X)$ that implements the LARS algorithm of [33] to calculate the full regularization path; the parameter λ is then selected along this path to match the appropriate condition specified by the analysis. For the uncompressed case, we run $\mathsf{lars}(Y, X)$ such that

$$Y = X\beta^* + \epsilon \tag{92}$$

and for the compressed case we run $\mathsf{lars}(\Phi Y, \Phi X)$ such that

$$\Phi Y = \Phi X \beta^* + \Phi \epsilon. \tag{93}$$

In each individual plot shown below, the covariance $\Sigma = \frac{1}{n}\mathbb{E}\left(X^T X\right)$ and model β^* are fixed across all curves in the plot. For each curve, a compression factor $f \in \{5, 10, 20, 40, 80, 120\}$ is chosen for the compressed lasso, and we show the probability of success for recovering the signs of β^* as the number of compressed observations m increases, where $m = 2\theta\sigma^2 s \log(p-s) + s + 1$ for $\theta \in [0.1, u]$, for $u \geq 3$. Thus, the number of compressed observations is m, and the number of uncompressed observations is $n = fm$. Each point on a curve, for a particular θ or m, is an average over 200 trials; for each trial, we randomly draw $X_{n\times p}$, $\Phi_{m\times n}$, and $\epsilon \in \mathbb{R}^n$. However, β^* remains the same for all 200 trials, and is fixed across different sets of experiments for the same sparsity level.

We consider two sparsity regimes

$$\text{Sublinear sparsity}: \quad s(p) = \frac{\alpha p}{\log(\alpha p)}, \quad \text{for } \alpha \in \{0.1, 0.2, 0.4\} \tag{94a}$$

$$\text{Fractional power sparsity}: \quad s(p) = \alpha p^{\gamma}, \quad \text{for } \alpha = 0.2 \text{ and } \gamma = 0.5. \tag{94b}$$

The coefficient vector β^* is selected to be a prefix of a fixed vector

$$\beta^\star = (-0.9, -1.7, 1.1, 1.3, 0.9, 2, -1.7, -1.3, -0.9, -1.5, 1.3, -0.9, 1.3, 1.1, 0.9)^T.$$

That is, if s is the number of nonzero coefficients, then

$$\beta_i^* = \begin{cases} \beta_i^\star & \text{if } i \leq s \\ 0 & \text{otherwise.} \end{cases} \tag{95}$$

As an exception, for the case $s = 2$, we set $\beta^* = (0.9, -1.7, 0, \ldots, 0)^T$.

After each trial, $\mathsf{lars}(Y, X)$ outputs a "regularization path," which is a set of estimated models $\mathcal{P}_m = \{\beta\}$ such that each $\beta \in \mathcal{P}_m$ is associated with a corresponding regularization parameter $\lambda(\beta)$, which is computed as

$$\lambda(\beta) = \frac{\|Y - X\tilde{\beta}\|_2^2}{m\|\tilde{\beta}\|_1}. \tag{96}$$

The coefficient vector $\tilde{\beta} \in \mathcal{P}_m$ for which $\lambda(\tilde{\beta})$ is closest to the value λ_m is then evaluated for sign consistency, where

$$\lambda_m = c\sqrt{\frac{\log(p-s)\log s}{m}}. \tag{97}$$

If $\operatorname{sgn}(\tilde{\beta}) = \operatorname{sgn}(\beta^*)$, the trial is considered a success, otherwise, it is a failure. We allow the constant c that scales λ_m to change with the experimental configuration (covariance Σ, compression factor f, dimension p, and sparsity s), but c is a fixed constant across all m along the same curve. The plots in Fig. 1 show the empirical probability of the event $\mathcal{E}(\operatorname{sgn}(\tilde{\beta}) = \operatorname{sgn}(\beta^*))$, which is a lower bound for that of the event $\{\operatorname{supp}(\tilde{\beta}) = \operatorname{supp}(\beta^*)\}$, for the sublinear sparsity regime with $\alpha = 0.2$. The results for other sparsity regimes are qualitatively the same. The plots clearly demonstrate that the compressed lasso recovers the true sparsity pattern as well as the uncompressed lasso.

B. Persistence

We now study the behavior of predictive and empirical risks under compression. In this section, we refer to $\mathsf{lasso2}(Y \sim X, L)$ as the code that solves the following ℓ_1-constrained optimization problem directly, based on algorithms described by [34]:

$$(P_3) \quad \tilde{\beta} = \arg\min \|Y - X\beta\|_2 \tag{98a}$$

$$\text{such that } \|\beta\|_1 \leq L. \tag{98b}$$

Let us first define the following ℓ_1-balls B_n and $B_{n,m}$ for a fixed uncompressed sample size n and dimension p, and a varying compressed sample size m. By [3], given a sequence of sets of estimators

$$B_n = \{\beta : \|\beta\|_1 \leq L_n\}, \quad \text{where } L_n = \frac{n^{1/4}}{\sqrt{\log n}} \tag{99}$$

the uncompressed Lasso estimator $\widehat{\beta}_n$ is persistent over B_n. Given n, p, Theorem 4.1 shows that, given a sequence of sets of estimators

$$B_{n,m} = \{\beta : \|\beta\|_1 \leq L_{n,m}\}, \quad \text{where } L_{n,m} = \frac{m^{1/4}}{\sqrt{\log(np)}} \tag{100}$$

for $\log^2(np) \leq m \leq n$, the compressed Lasso estimator $\widehat{\beta}_{n,m}$ as in (61) is persistent over $B_{n,m}$.

We use simulations to illustrate how close the compressed empirical risk computed through (105) is to that of the best compressed predictor $\beta_{*,n}$ as in (60) for a given set $B_{n,m}$, the size of which depends on the data dimension n, p of an uncompressed design matrix X, and the compressed dimension m; we also illustrate how close these two type of risks are to that of the best uncompressed predictor defined over a given set B_n for all $\log np \leq m \leq n$.

We let the row vectors of the design matrix be independent identical copies of a random vector $X \sim N(0, \Sigma)$. For simplicity, we generate $Y = X^T\beta^* + \epsilon$, where X and $\beta^* \in \mathbb{R}^p$, $\mathbb{E}(\epsilon) = 0$, and $\mathbb{E}(\epsilon^2) = \sigma^2$; note that $\mathbb{E}(Y|X) = X^T\beta^*$, although the persistence model need not assume this. Note that for all $m \leq n$

$$L_{n,m} = \frac{m^{1/4}}{\sqrt{\log(np)}} \leq L_n. \tag{101}$$

Hence, the risk of the model constructed on the compressed data

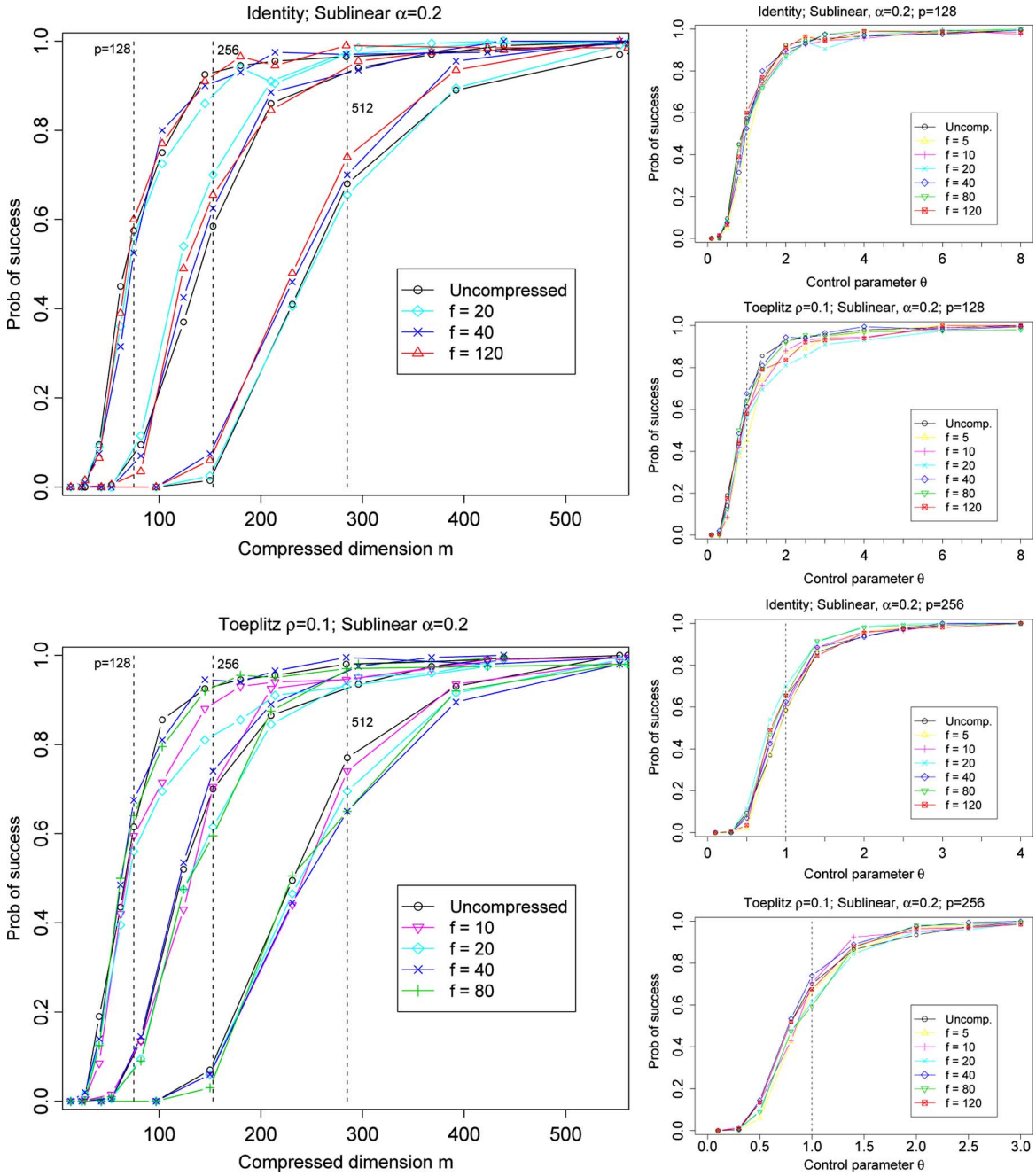

Fig. 1. Plots of the number of samples versus the probability of success. The three sets of curves on the left panel map to $p = 128, 256$ and 512, with vertical dashed lines marking $m = 2\theta s \log(p - s) + s + 1$ for $\theta = 1$, and $s = 5, 9$ and 15, respectively.

over $B_{n,m}$ is necessarily no smaller than the risk of the model constructed on the uncompressed data over B_n, for all $m \leq n$.

For $n = 9000$ and $p = 128$, we set $s(p) = 9$, following the sublinear sparsity (94a) with $\alpha = 0.4$; the following set of coefficients is chosen for β^* so that $\|\beta^*_a\|_1 > L_n$ and $\beta^*_a \notin B_n$: $\beta^*_a = (-0.9, -1.7, 1.1, 1.3, -0.5, 2,$

$$-1.7, -1.3, -0.9, 0, \ldots, 0)^T.$$

To find β_* that minimizes the predictive risk $R(\beta) = \mathbb{E}\left((Y - X^T\beta)^2\right)$, we first derive the following expression for the risk. With $\Sigma = A^TA$, a simple calculation shows that

$$\begin{aligned}
&\mathbb{E}(Y - X^T\beta)^2 - \mathbb{E}(Y^2) \\
&\quad = -\beta^{*T}\Sigma\beta^* + \|A\beta^* - A\beta\|_2^2 \quad \text{hence} \\
R(\beta) &= \mathbb{E}(Y^2) - \beta^{*T}\Sigma\beta^* + \|A\beta^* - A\beta\|_2^2 \\
&= \mathbb{E}(Y^2) - \beta^{*T}\mathbb{E}\left(XX^T\right)\beta^* + \|A\beta^* - A\beta\|_2^2 \\
&= \sigma^2 + \|A\beta^* - A\beta\|_2^2.
\end{aligned}$$

For the first set of simulations, we fix $n = 9000$ and $p = 128$. To generate the uncompressed predictive (oracle) risk curve, we let

$$\beta_{*,n} = \underset{\|\beta\|_1 \leq L_n}{\arg\min}\, R(\beta) = \underset{\|\beta\|_1 \leq L_n}{\arg\min}\, \|A\beta^* - A\beta\|_2^2. \tag{102}$$

Hence we obtain $\beta_{*,n}$ by running $\mathsf{lasso2}(\Sigma^{1/2}\beta^* \sim \Sigma^{1/2}, L_n)$. To generate the compressed predictive (oracle) curve, for each m, we let

$$\begin{aligned}
\beta_{*,n,m} &= \underset{\|\beta\|_1 \leq L_{n,m}}{\arg\min}\, R(\beta) \\
&= \underset{\|\beta\|_1 \leq L_{n,m}}{\arg\min}\, \|A\beta^* - A\beta\|_2^2.
\end{aligned} \tag{103}$$

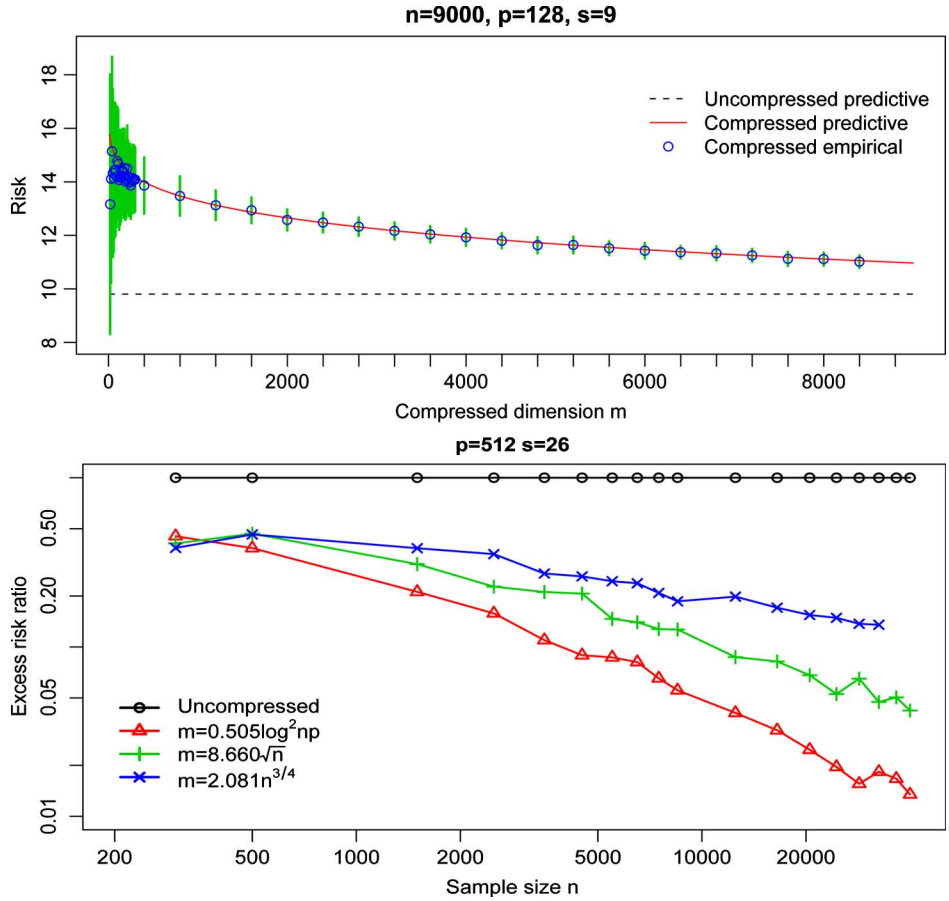

Fig. 2. Top plot: Risk versus compressed dimension for $\beta^* = \beta_a^*$; the uncompressed oracle risk is $R = 9.81$. $L_n = 2.6874$. Each vertical bar shows one standard deviation over 100 trials. Bottom plot: The ratio between the uncompressed mean excess risk and the compressed mean excess risks for each function of m in $\log - \log$ scale as sample size n grows. The uncompressed and compressed oracle risks are $R = 40.40234$ for $L_n = L_{m,n} = 1$ and $\beta^* = \beta_c^*$. Each mean excess risk is computed over 100 trials. The starting point for all three curves are $m = 150, n = 300$.

Hence we obtain $\beta_{*,n,m}$ for each m by running $\mathsf{lasso2}(\Sigma^{1/2}\beta^* \sim \Sigma^{1/2}, L_{n,m})$. We then compute oracle risks for both cases with $\beta_* = \beta_{*,n}$ or $\beta_* = \beta_{*,n,m}$, and

$$R(\beta_*) = (\beta_* - \beta^*)^T \Sigma (\beta_* - \beta^*) + \sigma^2. \tag{104}$$

For each chosen value of m, we compute the corresponding empirical risk, its sample mean, and sample standard deviation by averaging over 100 trials (see Fig. 2). For each trial, we randomly draw $X_{n\times p}$ with independent row vectors $x_i \sim N(0, T(0.1))$, and $Y = X\beta_a^* + \epsilon$. If β is the coefficient vector returned by $\mathsf{lasso2}(\Phi Y \sim \Phi X, L_{n,m})$, then the empirical risk is computed as

$$\hat{R}(\beta) = \gamma^T \widehat{\Sigma} \gamma \tag{105}$$

where $\widehat{\Sigma} = \frac{1}{m}\mathbb{Q}^T\Phi^T\Phi\mathbb{Q}$, for $\mathbb{Q}_{n\times(p+1)} = [Y, X]$ and $\gamma = (-1, \beta_1, \ldots, \beta_p)$.

The second set of simulations aims to verify the excess risk ratios as defined in Remark 4.4. We fix $p = 512$ and $s = 26$ that correspond to the sublinear sparsity model with $\alpha = 0.4$. Let

$$\begin{aligned}\beta_c^* = \{&-0.9, -1.7, 1.1, 1.3, -0.5, 2, -1.7, -1.3, -0.9, -1.5,\\ &1.3, -0.9, 1.3, 1.1, 0.9, -1.1, 1.1, 1.3, -0.5, 2,\\ &-1.7, -1.3, -0.9, -1.5, 1.3, 0.9, 0, \ldots, 0\}\end{aligned}$$

We show the excess risk ratios for each function of m. The mean excess risks are computed by averaging over 100 trials. For each trial, we randomly draw $X_{n\times p}$ with independent row vectors $x_i \sim N(0, T(0.1))$, $\Phi_{m\times n}$ for the compressed cases, and $\epsilon \in \mathbb{R}^n$. We let $Y = X\beta_c^* + \epsilon$ and hence $\Phi Y = \Phi X\beta_c^* + \Phi\epsilon$. The ratios for each function of m are obtained through the following:

$$\begin{aligned}&\text{excess risk ratio } (n, m)\\ &= \frac{\text{mean of excess risks with uncompressed sample size } n}{\text{mean of excess risks with compressed sample size } m}.\end{aligned}$$

VII. Proofs of Technical Results

A. Connection to the Gaussian Ensemble Result

First, let us state the following slightly relaxed conditions that are imposed on the design matrix by [6], and also by [9], when X is deterministic:

$$\left\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\right\|_\infty \le 1 - \eta, \text{ for some } \eta \in (0, 1] \tag{106a}$$

$$\Lambda_{\min}\left(\frac{1}{n} X_S^T X_S\right) \ge C_{\min} > 0 \tag{106b}$$

where $\Lambda_{\min}(A)$ is the smallest eigenvalue of A. In Section VII-B, Proposition 7.4 shows that S-incoherence implies the conditions in (106a) and (106b).

We first observe that with X fixed, each row of $\tilde{X}$ is chosen as an i.i.d. Gaussian random vector $\sim N(0, \Sigma)$ with covariance matrix $\Sigma = \frac{1}{n}X^T X$. Hence, the design matrix $\tilde{X} = \Phi X$ is exactly a Gaussian ensemble that [6] analyzes, except that in our case, Σ is a singular matrix while his current analysis assumes

that Σ is nonsingular. In the following, let $\Lambda_{\min}(\Sigma_{SS})$ be the minimum eigenvalue of Σ_{SS} and $\Lambda_{\max}(\Sigma)$ be the maximum eigenvalue of Σ. By imposing the S-incoherence condition on $X_{n\times p}$, we obtain the following two conditions on the covariance matrix Σ, which are also required by [6] for deriving the threshold conditions (16) and (17), when the design matrix is a Gaussian ensemble:

$$\|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty \leq 1-\eta, \quad \text{for } \eta\in(0,1], \quad \text{and} \tag{107a}$$

$$\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0. \tag{107b}$$

When we apply this to $\tilde{X}=\Phi X$, where Φ is an Gaussian ensemble, X is deterministic, and $\mathbb{E}\left(\frac{1}{m}X^T\Phi^T\Phi X\right)=\frac{1}{n}X^TX$. This condition requires that

$$\|X_{S^c}^TX_S(X_S^TX_S)^{-1}\|_\infty \leq 1-\eta, \quad \text{for } \eta\in(0,1], \quad \text{and} \tag{108a}$$

$$\Lambda_{\min}\left(\frac{1}{n}X_S^TX_S\right) \geq C_{\min} > 0. \tag{108b}$$

In addition, it is assumed in [6] that there exists a constant $C_{\max}$ such that

$$\Lambda_{\max}(\Sigma) \leq C_{\max} \quad \text{and} \quad \theta_u = \frac{C_{\max}}{\eta^2 C_{\min}} \tag{109}$$

is a constant, where θ is defined in (16). This condition need not hold for $\frac{1}{n}X^TX$; in more detail, given $\Lambda_{\max}(\frac{1}{n}X^TX)=\frac{1}{n}\Lambda_{\max}(X^TX)=\frac{1}{n}\|X\|_2^2$, we first obtain a loose upper and lower bound for $\|X\|_2^2$ through the Frobenius norm $\|X\|_F$ of X. Given that $\|X_j\|_2^2=n, \forall j\in\{1,\ldots,p\}$, we have

$$\|X\|_F^2=\sum_{j=1}^p\sum_{i=1}^n|X_{ij}|^2=pn.$$

Thus, by $\|X\|_2\leq\|X\|_F\leq\sqrt{\min\{n,p\}}\|X\|_2$, we obtain for $n<p$

$$p=\frac{1}{n}\|X\|_F^2\leq\|X\|_2^2\leq\|X\|_F^2=pn$$

which implies that

$$\frac{p}{n}\leq\Lambda_{\max}\left(\frac{1}{n}X^TX\right)\leq p.$$

Since we allow p to grow with n, (109) need not hold. In particular, $\theta_u=\Omega\left(p/nC_{\min}\right)$. In summary, by imposing the S-incoherence condition on a deterministic $X_{n\times p}$ with all columns of X having ℓ_2-norm n, when m satisfies the lower bound in (31), we have shown that the probability of sparsity recovery through lasso approaches one, given λ_m satisfies (32*a*)), when the design matrix $\tilde{X}$ is a Gaussian ensemble with a singular covariance matrix generated through ΦX. Directly applying Wainwright's result as in (16) to our scenario will be impossible. We do not have a comparable result for the failure of recovery given (17).

B. S-incoherence

We first state some generally useful results about matrix norms.

Theorem 7.1: [35, p. 301] If $|||\cdot|||$ is a matrix norm and $|||A||| < 1$, then $I+A$ is invertible and $(I+A)^{-1}=\sum_{k=0}^\infty(-A)^k$.

Proposition 7.2: If the matrix norm $\|\cdot\|$ has the property that $\|I\|=1$, and if $A\in M_n$ is such that $\|A\|<1$, we have

$$\frac{1}{1+\|A\|}\leq\|(I+A)^{-1}\|\leq\frac{1}{1-\|A\|}. \tag{110}$$

Proof: The upper bound follows from Theorem 7.1 and triangle inequality

$$\|(I+A)^{-1}\|=\left\|\sum_{k=0}^\infty(-A)^k\right\|\leq\sum_{k=0}^\infty\|-A\|^k$$
$$=\sum_{k=0}^\infty\|A\|^k=\frac{1}{1-\|A\|}.$$

The lower bound follows that general inequality $\|B^{-1}\|\geq 1/\|B\|$, given that $\|I\|\leq\|B\|\|B^{-1}\|$ and the triangle inequality: $\|A+I\|\leq\|A\|+\|I\|=\|A\|+1$, that is, $\|(A+I)^{-1}\|\geq 1/\|A+I\|\geq 1/1+\|A\|$. □

Let us define the following symmetric matrices, that we use throughout the rest of this section:

$$A=\frac{1}{n}X_S^TX_S-I_{|S|} \tag{111a}$$

$$\tilde{A}=\frac{1}{m}(\Phi X)_S^T(\Phi X)_S-I_s=\frac{1}{m}Z_S^TZ_S-I_s. \tag{111b}$$

We next show the following consequence of the S-Incoherence condition.

Proposition 7.3: Let X be an $n\times p$ that satisfies the S-incoherence condition. Then for the symmetric matrix A in (111a) , we have $\|A\|_\infty=\|A\|_1\leq 1-\eta$, for some $\eta\in(0,1]$, and

$$\|A\|\leq\sqrt{\|A\|_\infty\|A\|_1}\leq 1-\eta. \tag{112}$$

and, hence, $\Lambda_{\min}(\frac{1}{n}X_S^TX_S)\geq\eta$, i.e., the S-incoherence condition implies condition (106b).

Proof: Given that $\|A\|_2<1$, $\|I\|_2=1$, and by Proposition 7.2

$$\Lambda_{\min}\left(\frac{1}{n}X_S^TX_S\right)=\frac{1}{\|(\frac{1}{n}X_S^TX_S)^{-1}\|_2}$$
$$=\frac{1}{\|(I+A)^{-1}\|_2}\geq 1-\|A\|_2\geq\eta>0.$$

□

Proposition 7.4: The S-incoherence condition on an $n\times p$ matrix X implies conditions (106a) and (106b).

Proof: It remains to show (106a) given Proposition 7.3. Now suppose that the incoherence condition holds for some $\eta\in(0,1]$, i.e., $\|\frac{1}{n}X_{S^c}^TX_S\|_\infty+\|A\|_\infty\leq 1-\eta$, we must have

$$\frac{\|\frac{1}{n}X_{S^c}^TX_S\|_\infty}{1-\|A\|_\infty}\leq 1-\eta \tag{113}$$

given that

$$\|\frac{1}{n}X_{S^c}^TX_S\|_\infty+\|A\|_\infty(1-\eta)\leq 1-\eta$$

and $1 - \|A\|_\infty \geq \eta > 0$. Next observe that, given $\|A\|_\infty < 1$, by Proposition 7.2

$$\left\| \left(\frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty = \|(I+A)^{-1}\|_\infty \leq 1/1 - \|A\|_\infty.$$

Finally, we have

$$\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\|_\infty \leq \|\frac{1}{n} X_{S^c}^T X_S\|_\infty \left\| \left(\frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \leq \|\frac{1}{n} X_{S^c}^T X_S\|_\infty / 1 - \|A\|_\infty \leq 1 - \eta. \quad \square$$

C. Proof of Proposition 3.6

We use Lemma 3.5, except that we now have to consider the change in absolute row sums of $\|\frac{1}{n} X_{S^c}^T X_S\|_\infty$ and $\|A\|_\infty$ after multiplication by Φ. We first prove the following claim.

Claim 7.5: Let X be a deterministic matrix that satisfies $\tau = \frac{\eta}{4s}$, the incoherence condition. If for any two columns X_i, X_j of X that are involved in (40b)

$$\left| \frac{1}{m} \langle \Phi X_i, \Phi X_j \rangle - \frac{1}{n} \langle X_i, X_j \rangle \right| \leq \tau, \quad \text{then}$$

$$\left\| \frac{1}{m} (\Phi X)_{S^c}^T (\Phi X)_S \right\|_\infty + \|\tilde{A}\|_\infty \leq 1 - \eta + 2s\tau \quad (114)$$

$$\Lambda_{\min} \left(\frac{1}{m} Z_S^T Z_S \right) \geq \eta - s\tau. \quad (115)$$

Proof: It is straightforward to show the first inequality in (114). Since each row in $\frac{1}{m}(\Phi X)_{S^c}^T (\Phi X)_S$ and A has s entries, where each entry changes by at most τ compared to those in $\frac{1}{n} X^T X$, the absolute sum of any row can change by at most $s\tau$

$$\left| \left\| \frac{1}{m} (\Phi X)_{S^c}^T (\Phi X)_S \right\|_\infty - \left\| \frac{1}{n} X_{S^c}^T X_S \right\|_\infty \right| \leq s\tau$$

$$\left| \|\tilde{A}\|_\infty - \|A\|_\infty \right| \leq s\tau.$$

Hence

$$\left\| \frac{1}{m} (\Phi X)_{S^c}^T (\Phi X)_S \right\|_\infty + \|\tilde{A}\|_\infty \leq \left\| \frac{1}{n} X_{S^c}^T X_S \right\|_\infty + \|A\|_\infty + 2s\tau \leq 1 - \eta + 2s\tau.$$

We now prove the second inequality. Defining $E = \tilde{A} - A$, we have $\|E\|_2 \leq s \max_{i,j} |\tilde{A}_{i,j} - A_{i,j}| \leq s\tau$, given that each entry of $\tilde{A}$ deviates from that of A by at most τ. Thus, we have that

$$\|\tilde{A}\|_2 = \|A + E\|_2 \quad (116a)$$

$$\leq \|A\|_2 + \|E\|_2 \quad (116b)$$

$$\leq \|A\|_2 + s \max_{i,j} |E_{i,j}| \quad (116c)$$

$$\leq 1 - \eta + s\tau \quad (116d)$$

where $\|A\|_2 \leq 1 - \eta$ is due to Proposition 7.3. Given that $\|I\|_2 = 1$ and $\|A\|_2 < 1$, by Proposition 7.2

$$\Lambda_{\min} \left(\frac{1}{m} Z_S^T Z_S \right) = \frac{1}{\left\| \left(\frac{1}{m} Z_S^T Z_S \right)^{-1} \right\|_2} = \frac{1}{\|(I+\tilde{A})^{-1}\|_2} \geq 1 - \|\tilde{A}\|_2 \geq \eta - s\tau.$$

$\square$

We let $\mathcal{E}$ represents union of the following events, where $\tau = \eta/4s$:

1) $\exists i \in S, j \in S^c$, such that
$$\left| \frac{1}{m} \langle \Phi X_i, \Phi X_j \rangle - \frac{1}{n} \langle X_i, X_j \rangle \right| \geq \tau.$$
2) $\exists i, i' \in S$, such that
$$\left| \frac{1}{m} \langle \Phi X_i, \Phi X_{i'} \rangle - \frac{1}{n} \langle X_i, X_{i'} \rangle \right| \geq \tau$$
3) $\exists j \in S^c$, such that
$$\left| \frac{1}{m} \langle \Phi X_j, \Phi X_j \rangle - \frac{1}{n} \langle X_j, X_j \rangle \right| = \left| \frac{1}{m} \|\Phi X_j\|_2^2 - \frac{1}{n} \|X_j\|_2^2 \right| > \tau.$$

Consider first the implication of $\mathcal{E}^c$, i.e., when none of the events in $\mathcal{E}$ happens. We immediately have that (40b), (115), and (41b) all simultaneously hold by Claim 7.5; and (40b) implies that the incoherence condition is satisfied for $Z = \Phi X$ by Proposition 7.4. We first bound the probability of a single event counted in $\mathcal{E}$. Consider two column vectors $x = \frac{X_i}{\sqrt{n}}, y = \frac{X_j}{\sqrt{n}} \in \mathbb{R}^n$ in matrix $\frac{X}{\sqrt{n}}$, we have $\|x\|_2 = 1, \|y\|_2 = 1$, and for $\tau = \frac{\eta}{4s}$

$$\mathbb{P}\left(\left| \frac{1}{m} \langle \Phi X_i, \Phi X_j \rangle - \frac{1}{n} \langle X_i, X_j \rangle \right| \geq \tau \right) \quad (117a)$$

$$= \mathbb{P}\left(\left| \frac{n}{m} \langle \Phi x, \Phi y \rangle - \langle x, y \rangle \right| \geq \tau \right) \quad (117b)$$

$$\leq 2 \exp\left(\frac{-m\tau^2}{C_1 + C_2 \tau} \right) \quad (117c)$$

$$\leq 2 \exp\left(-\frac{m\eta^2/16s^2}{C_1 + C_2 \eta/4s} \right). \quad (117d)$$

We can now bound the probability that any such large-deviation event happens. Recall that p is the number of columns of X and $s = |S|$; the total number of events in $\mathcal{E}$ is less than $p(s+1)$. Thus

$$\mathbb{P}(\mathcal{E}) \leq p(s+1) \mathbb{P}\left(\left| \frac{1}{m} \langle \Phi X_i, \Phi X_j \rangle - \frac{1}{n} \langle X_i, X_j \rangle \right| \geq \frac{\eta}{4s} \right)$$

$$\leq 2p(s+1) \exp\left(-\frac{m\eta^2/16s^2}{C_1 + C_2 \eta/4s} \right)$$

$$= 2p(s+1) \exp\left(-(\ln p + c \ln n + \ln 2(s+1)) \right) \leq \frac{1}{n^c}$$

given that $m \geq \left(\frac{16 C_1 s^2}{\eta^2} + \frac{4 C_2 s}{\eta} \right) (\ln p + c \ln n + \ln 2(s+1))$. Note that this is where the dependence on s^2 arises in the lower bound on the compressed sample size m. $\square$

D. Proof of Lemma 3.8

Recall that $Z = \tilde{X} = \Phi X, W = \tilde{Y} = \Phi Y$, and $\omega = \tilde{\epsilon} = \Phi \epsilon$, and we observe $W = Z\beta^* + \omega$. First observe that the KKT conditions imply that $\tilde{\beta} \in \mathbb{R}^p$ is optimal, i.e., $\tilde{\beta} \in \tilde{\Omega}_m$ for $\tilde{\Omega}_m$ as defined in (25), if and only if there exists a subgradient

$$\tilde{z} \in \partial \|\tilde{\beta}\|_1 = \{ z \in \mathbb{R}^p \,|\, z_i = \operatorname{sgn}(\tilde{\beta}_i) \text{ for } \tilde{\beta}_i \neq 0, |z_j| \leq 1 \text{ otherwise} \} \quad (118)$$

such that $\frac{1}{m}Z^TZ\tilde{\beta}-\frac{1}{m}Z^TW+\lambda_m\tilde{z}=0$, which is equivalent to the following linear system by substituting $W=Z\beta^*+\omega$ and rearranging:

$$\frac{1}{m}Z^TZ(\tilde{\beta}-\beta^*)-\frac{1}{m}Z^T\omega+\lambda_m\tilde{z}=0. \tag{119}$$

Hence, given Z,β^*,ω and $\lambda_m>0$ the event $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m)=\mathrm{sgn}(\beta^*)\right)$ holds if and only if

1) there exist a point $\tilde{\beta}\in\mathbb{R}^p$ and a subgradient $\tilde{z}\in\partial\|\tilde{\beta}\|_1$ such that (119) holds, and
2) $\mathrm{sgn}(\tilde{\beta}_S)=\mathrm{sgn}(\beta^*_S)$ and $\tilde{\beta}_{S^c}=\beta^*_{S^c}=0$, which implies that $\tilde{z}_S=\mathrm{sgn}(\beta^*_S)$ and $|\tilde{z}_{S^c}|\le 1$ by definition of $\tilde{z}$.

Plugging $\tilde{\beta}_{S^c}=\beta^*_{S^c}=0$ and $\tilde{z}_S=\mathrm{sgn}(\beta^*_S)$ in (119) shows that the event $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m)=\mathrm{sgn}(\beta^*)\right)$ holds if and only

1) there exists a point $\tilde{\beta}\in\mathbb{R}^p$ and a subgradient $\tilde{z}\in\partial\|\tilde{\beta}\|_1$ such that

$$\frac{Z^T_{S^c}Z_S(\tilde{\beta}_S-\beta^*_S)}{m}-\frac{Z^T_{S^c}\omega}{m}=-\lambda_m\tilde{z}_{S^c} \tag{120a}$$

$$\frac{Z^T_{S}Z_S(\tilde{\beta}_S-\beta^*_S)}{m}-\frac{Z^T_{S}\omega}{m}=-\lambda_m\mathrm{sgn}(\beta^*_S); \tag{120b}$$

2) and $\mathrm{sgn}(\tilde{\beta}_S)=\mathrm{sgn}(\beta^*_S)$ and $\tilde{\beta}_{S^c}=\beta^*_{S^c}=0$.

Using invertability of $Z^T_SZ_S$, we can solve for $\tilde{\beta}_S$ and $\tilde{z}_{S^c}$ using (120a) and (120b) to obtain

$$-\lambda_m\tilde{z}_{S^c}=Z^T_{S^c}Z_S(Z^T_SZ_S)^{-1}\left[\frac{Z^T_S\omega}{m}-\lambda_m\mathrm{sgn}(\beta^*_S)\right]-\frac{Z^T_{S^c}\omega}{m} \tag{121a}$$

$$\tilde{\beta}_S=\beta^*_S+\frac{(Z^T_SZ_S)^{-1}}{m}\left[\frac{1}{m}Z^T_S\omega-\lambda_m\mathrm{sgn}(\beta^*_S)\right]. \tag{121b}$$

Thus, given invertability of $Z^T_SZ_S$, the event $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m)=\mathrm{sgn}(\beta^*)\right)$ holds if and only if

1) there exists simultaneously a point $\tilde{\beta}\in\mathbb{R}^p$ and a subgradient $\tilde{z}\in\partial\|\tilde{\beta}\|_1$ such that (121a) and (121b) hold;
2) $\mathrm{sgn}(\tilde{\beta}_S)=\mathrm{sgn}(\beta^*_S)$ and $\tilde{\beta}_{S^c}=\beta^*_{S^c}=0$.

The last set of necessary and sufficient conditions for the event $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m)=\mathrm{sgn}(\beta^*)\right)$ to hold implies that there exists simultaneously a point $\tilde{\beta}\in\mathbb{R}^p$ and a subgradient $\tilde{z}\in\partial\|\tilde{\beta}\|_1$ such that the first two equations at the bottom of the page hold, given that $|\tilde{z}_{S^c}|\le 1$ by definition of $\tilde{z}$. Thus, (43) and (44) hold for the given Z,β^*,ω and $\lambda_m>0$. Thus, we have shown the lemma in one direction.

For the reverse direction, given Z,β^*,ω, and supposing that (43) and (44) hold for some $\lambda_m>0$, we first construct a point $\tilde{\beta}\in\mathbb{R}^p$ by letting $\tilde{\beta}_{S^c}=\beta^*_{S^c}=0$ and

$$\tilde{\beta}_S=\beta^*_S+\left(\frac{1}{m}Z^T_SZ_S\right)^{-1}\left[\frac{1}{m}Z^T_S\omega-\lambda_m\mathrm{sgn}(\beta^*_S)\right]$$

which guarantees that

$$\begin{aligned}&\mathrm{sgn}(\tilde{\beta}_S)\\&\quad=\mathrm{sgn}\left(\beta^*_S+\left(\frac{1}{m}Z^T_SZ_S\right)^{-1}\left[\frac{1}{m}Z^T_S\omega-\lambda_m\mathrm{sgn}(\beta^*_S)\right]\right)\\&\quad=\mathrm{sgn}(\beta^*_S)\end{aligned}$$

by (44). We simultaneously construct $\tilde{z}$ by letting $\tilde{z}_S=\mathrm{sgn}(\tilde{\beta}_S)=\mathrm{sgn}(\beta^*_S)$ and define $\tilde{z}_{S^c}$ in (122), also shown at the bottom of the page, which guarantees that $|\tilde{z}_{S^c}|\le 1$ due to (44); hence, $\tilde{z}\in\partial\|\tilde{\beta}\|_1$. Thus, we have found a point $\tilde{\beta}\in{}^p$ and a subgradient $\tilde{z}\in\partial\|\tilde{\beta}\|_1$ such that $\mathrm{sgn}(\tilde{\beta})=\mathrm{sgn}(\beta^*)$ and the set of (121a) and (121b) is satisfied. Hence, assuming the invertability of $Z^T_SZ_S$, the event $\mathcal{E}\left(\mathrm{sgn}(\tilde{\beta}_m)=\mathrm{sgn}(\beta^*)\right)$ holds for the given $Z,\beta^*,\omega,\lambda_m$. □

E. Proof of Lemma 3.9

Given that $\frac{1}{m}Z^T_SZ_S=\tilde{A}+I_s$, we bound $\|(\frac{1}{m}Z^T_SZ_S)^{-1}\|_\infty$ through $\|(\tilde{A}+I_s)^{-1}\|$.

First, we have for $m\ge\left(\frac{16C_1s^2}{\eta^2}+\frac{4C_2s}{\eta}\right)(\ln p+c\ln n+\ln 2(s+1))$

$$\|\tilde{A}\|_\infty\le\|A\|_\infty+\frac{\eta}{4}\le 1-\eta+\eta/4=1-3\eta/4 \tag{123a}$$

where $\eta\in(0,1]$, due to (29) and (40a). Hence, given that $\|I\|_\infty=1$ and $\|\tilde{A}\|_\infty<1$, by Proposition 7.2

$$\left\|\left(\frac{1}{m}Z^T_SZ_S\right)^{-1}\right\|_\infty=\left\|(\tilde{A}+I_s)^{-1}\right\|_\infty\le\frac{1}{1-\|\tilde{A}\|_\infty}\le\frac{4}{3\eta}. \tag{124a}$$

Similarly, given $\|A\|_\infty<1$, we have

$$\begin{aligned}\frac{1}{1+\|A\|_\infty}&\le\left\|\left(\frac{1}{n}X^T_SX_S\right)^{-1}\right\|_\infty\\&=\|(A+I_s)^{-1}\|_\infty\le\frac{1}{1-\|A\|_\infty}.\end{aligned}$$

$$\begin{aligned}&\left|Z^T_{S^c}Z_S(Z^T_SZ_S)^{-1}\left[\frac{Z^T_S\omega}{m}-\lambda_m\mathrm{sgn}(\beta^*_S)\right]-\frac{Z^T_{S^c}\omega}{m}\right|\\&\quad=|-\lambda_m\tilde{z}_{S^c}|\le\lambda_m\\&\mathrm{sgn}(\tilde{\beta}_S)\quad=\mathrm{sgn}\left(\beta^*_S+\left(\frac{1}{m}Z^T_SZ_S\right)^{-1}\left[\frac{1}{m}Z^T_S\omega-\lambda_m\mathrm{sgn}(\beta^*_S)\right]\right)\\&\quad=\mathrm{sgn}(\beta^*_S)\end{aligned}$$

$$\tilde{z}_{S^c}=-\frac{1}{\lambda_m}\left(Z^T_{S^c}Z_S(Z^T_SZ_S)^{-1}\left[\frac{1}{m}Z^T_S\omega-\lambda_m\mathrm{sgn}(\beta^*_S)\right]-\frac{1}{m}Z^T_{S^c}\omega\right) \tag{122}$$

Given that $\frac{\lambda_m}{\rho_m}\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\|_\infty \to 0$, we have $\frac{\lambda_m}{\rho_m}\frac{1}{1+\|A\|_\infty} \to 0$, and thus

$$\frac{\lambda_m}{\rho_m}\frac{1}{1-\|\tilde{A}\|_\infty} = \frac{\lambda_m}{\rho_m}\frac{1}{1+\|A\|_\infty}\frac{1+\|A\|_\infty}{1-\|\tilde{A}\|_\infty} \tag{125a}$$

$$\leq \frac{\lambda_m}{\rho_m}\frac{1}{1+\|A\|_\infty}\left(\frac{4(2-\eta)}{3\eta}\right) \to 0 \tag{125b}$$

by (124a) and the fact that by (29) $1+\|A\|_\infty \leq 2-\eta$. □

F. Proof of Claim 3.11

We first prove the following.

Claim 7.6: If m satisfies (31), then $\frac{1}{m}\max_{i,j}(B_{i,j}) \leq 1+\eta/4s$.

Proof: Let us denote the ith column in Z_S with $Z_{S,i}$. Let $x = Z_{S,i}$ and $y = Z_{S,j}$ be $m \times 1$ vectors. By Proposition 3.6, $\|x\|_2^2, \|y\|_2^2 \leq m\left(1+\frac{\eta}{4s}\right)$. We have by function of x, y

$$\begin{aligned}
B_{ij} &= Z_{S,i}^T R Z_{S,j} = \sum_{i=1}^m\sum_{j=1}^m x_i y_j R_{ij} \\
&\leq \sum_{i=1}^m\sum_{j=1}^m \left|x_i\right|\left|y_j\right|\left|R_{ij}\right| \\
&\leq \max_{i,j}|R_{ij}| \sum_{i=1}^m\sum_{j=1}^m \left|x_i\right|\left|y_j\right| \\
&= \max_{i,j}|R_{ij}| \left(\sum_{i=1}^m \left|x_i\right|\right)\left(\sum_{j=1}^m \left|y_j\right|\right) \\
&\leq \max_{i,j}|R_{ij}| m\|x\|_2\|y\|_2 \leq \max_{i,j}|R_{ij}| m^2\left(1+\frac{\eta}{4s}\right).
\end{aligned}$$

Thus, the claim follows given that $\max_{i,j}|R_{ij}| \leq 4\sqrt{\frac{\log n}{n}}$ and $4m \leq \sqrt{\frac{n}{\log n}}$. □

Finally, to finish the proof of Claim 3.11 we have

$$\begin{aligned}
\max_i M_{ii} &= \max_i \frac{C_i^T B C_i}{m} = \frac{\max_i C_i^T B C_i}{m} \\
&= \frac{1}{m}\max_i\left(\sum_{j=1}^m\sum_{k=1}^m C_{ij}C_{ik}B_{jk}\right) \\
&\leq \frac{1}{m}\max_{i,j}|B_{ij}|\max_i\left(\sum_{j=1}^m\left|C_{ij}\right|\sum_{k=1}^m\left|C_{ik}\right|\right) \\
&\leq \left(1+\frac{\eta}{4s}\right)\max_i\left(\sum_{j=1}^m\left|C_{ij}\right|\right)^2 \\
&\leq \left(1+\frac{\eta}{4s}\right)\left(\max_i\sum_{j=1}^m\left|C_{ij}\right|\right)^2 \\
&\leq \left(1+\frac{\eta}{4s}\right)\|C\|_\infty^2 \leq \left(1+\frac{\eta}{4s}\right)\left(\frac{4}{3\eta}\right)^2
\end{aligned}$$

where $\|C\|_\infty = \|\left(\frac{1}{m}Z_S^TZ_S\right)^{-1}\|_\infty \leq \frac{4}{3\eta}$ as in (124a) for $m \geq \left(\frac{16C_1s^2}{\eta^2}+\frac{4C_2s}{\eta}\right)(\ln p + c\ln n + \ln 2(s+1))$.

Remark 7.7: In fact, $\max_{i,j} M_{ij} = \max_{i,i} M_{ii}$. □

VIII. Discussion

The results presented here suggest several directions for future work. Most immediately, our current sparsity analysis holds for compression using random linear transformations. However, compression with a random affine mapping $X \mapsto \Phi X + \Delta$ may have stronger privacy properties; we expect that our sparsity results can be extended to this case. While we have studied data compression by random projection of columns of X to low dimensions, one also would like to consider projection of the rows, reducing p to a smaller number of effective variables. However, simulations suggest that the strong sparsity recovery properties of ℓ_1 regularization are not preserved under projection of the rows.

It would be natural to investigate the effectiveness of other statistical learning techniques under compression of the data. For instance, logistic regression with ℓ_1-regularization has recently been shown to be effective in isolating relevant variables in high-dimensional classification problems [36]; we expect that compressed logistic regression can be shown to have similar theoretical guarantees to those shown in the current paper. It would also be interesting to extend this methodology to nonparametric methods. As one possibility, the rodeo is an approach to sparse nonparametric regression that is based on thresholding derivatives of an estimator [37]. Since the rodeo is based on kernel evaluations, and Euclidean distances are approximately preserved under random projection, this nonparametric procedure may still be effective under compression.

The formulation of privacy in Section V is, arguably, weaker than the cryptographic-style guarantees sought through, for example, differential privacy [26]. In particular, our analysis in terms of average mutual information may not preclude the recovery of detailed data about a small number of individuals. For instance, suppose that a column X_j of X is very sparse, with all but a few entries zero. Then the results of compressed sensing [11] imply that, given knowledge of the compression matrix Φ, this column can be approximately recovered by solving the compressed sensing linear program

$$\min \|X_j\|_1 \tag{126a}$$

$$\text{such that } Z_j = \Phi X_j. \tag{126b}$$

However, crucially, this requires knowledge of the compression matrix Φ; our privacy protocol requires that this matrix is not known to the receiver. Moreover, this requires that the column is sparse; such a column cannot have a large impact on the predictive accuracy of the regression estimate. If a sparse column is removed, the resulting predictions should be nearly as accurate as those from an estimator constructed with the full data. We leave the analysis of this case as an interesting direction for future work.

Acknowledgment

The authors wish to thank Avrim Blum, Steve Fienberg, and Pradeep Ravikumar for helpful comments on this work.

References

[1] G. Duncan and R. Pearson, "Enhancing access to microdata while protecting confidentiality: Prospects for the future," *Statisti. Sci.*, vol. 6, no. 3, pp. 219–232, Aug. 1991.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[3] E. Greenshtein and Y. Ritov, "Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization," *Bernoulli*, vol. 10, pp. 971–988, 2004.
[4] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.
[5] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Europ. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.
[6] M. Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Recovery of Sparsity," Dept. Statist., Univ. Calif. Berkeley, Berkeley, CA, 2006, Tech. Rep. 709.
[7] N. Meinshausen and B. Yu, "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," Dept. Statist., Univ. Calif. Berkeley, Berkeley, CA, 2006, Tech. Rep. 720.
[8] N. Meinshausen and P. Buhlmann, "High dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
[9] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learn. Res.*, vol. 7, pp. 2541–2567, 2007.
[10] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[11] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
[12] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
[13] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.
[14] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk, "Sparse signal detection from incoherent projections," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. III-305–308.
[15] M. Davenport, M. Wakin, and R. Baraniuk, "Detection and Estimation With Compressive Measurements," Elec. Comp. Eng. Dept., Rice Univ., Houston, TX, 2006, TREE 0610, Tech. Rep..
[16] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh, "Compressive sampling for signal classification," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct.\Nov. 2006, pp. 1430–1434.
[17] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proc. Computational Imaging V*, San Jose, CA, Feb. 2007, vol. 6498, p. 153.
[18] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Proc. Conf. Modern Analysis and Probability*, R. Beals, A. Beck, A. Bellow, and A. Hajian, Eds., New Haven, CT, Jun. 1984, pp. 189–206.
[19] T. Dalenius, "Privacy transformations for statistical information systems," *J. Statis. Plann. Inference*, vol. 1, pp. 73–86, 1977.
[20] A. P. Sanil, A. Karr, X. Lin, and J. P. Reiter, "Privacy preserving regression modelling via distributed computation," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Seattle, WA, Aug. 2004, pp. 677–682.
[21] D. Ting, S. E. Fienberg, and M. Trottini, "Random orthogonal matrix masking methodology for microdata release," *Int. J. Inf. Comp.Security*, vol. 2, no. 1, pp. 86–105, Jan. 2008.
[22] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, Oct. 1949.
[23] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.
[24] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 92–106, Jan. 2006.
[25] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, Santa Barbara, CA, May 2001, pp. 247–255.
[26] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloquium on Automata, Languages and Programming*, Venice, Italy, 2006, pp. 1–12.
[27] T. Dalenius, "Towards a methodology for statistical disclosure control," *Statistik Tidskrift*, vol. 15, pp. 429–444, 1977.
[28] C. Dwork, F. McSherry, and K. Talwar, "The price of privacy and the limits of LP decoding," in *Proc. Symp. Theory of Computing (STOC)*, San Diego, CA, Jun. 2007, pp. 85–94.
[29] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[30] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[31] I. Johnstone, "Chi-square oracle inequalities," in *State of the Art in Probability and Statistics, Festchrift for Willem R. van Zwet*, M. de Gunst, C. Klaassen, and A. van der Waart, Eds. Beachwood, OH: Inst. Math. Statist., 2001, vol. 36, IMS Lecture Notes–Monographs, pp. 399–418.
[32] I. Johnstone and A. Y. Lu, "Sparse Principal Components Analysis," Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep., 2004.
[33] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[34] M. Osborne, B. Presnell, and B. Turlach, "On the lasso and its dual," *J. Computat. Graph. Statist.*, vol. 9, no. 2, pp. 319–337, 2000.
[35] R. Horn and C. Johnson, *Matrix Analysis*, reprint ed. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[36] M. Wainwright, P. Ravikumar, and J. Lafferty, "High-dimensional graphical model selection using ℓ_1 -regularized logistic regression," in *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, 2007.
[37] J. Lafferty and L. Wasserman, "Rodeo: Sparse, greedy nonparametric regression," *Ann. Statist.*, vol. 36, no. 1, pp. 28–63, 2008.

Shuheng Zhou received the Ph.D. degree in electrical and computer engineering, from Carnegie Mellon University (CMU), Pittsburgh, PA, in August 2006

She was a Postdoctoral Fellow in the Computer Science Department at CMU from September 2006 till July 2008, working with Prof. John Lafferty and Prof. Larry Wasserman on statistical learning theory and algorithms. She is currently a Postdoctoral Researcher in the Seminar für Statistik at ETH Zurich, Zuriich, Switzerland.

John Lafferty (M'96–SM'00–F'07) received the Ph.D. in Mathematics from Princeton University, Pronceton, NJ, where he was a member of the Program in Applied and Computational Mathematics.

He has been a member of the faculty at Carnegie Mellon University, Pittsburgh, PA, since 1994. He is now a Professor in the Computer Science Department at Carnegie Mellon with a joint appointment in the Machine Learning Department and the Department of Statistics. His recent research interests include statistical and computational aspects of machine learning, with a focus on nonparametric methods for high-dimensional data and applications in information retrieval.

Larry Wasserman received the Ph.D. degree in biostatistics from the University of Toronto, Toronto, ON, Canada, in 1988.

He has been a member of the faculty at Carnegie Mellon University, Pittsburgh, PA, since 1988. He is now a Professor in the Statistics Department in Carnegie Mellon with a joint appointment in the Machine Learning Department. His recent research interests include statistical and computational aspects of machine learning, with a focus on nonparametric methods for high-dimensional data.