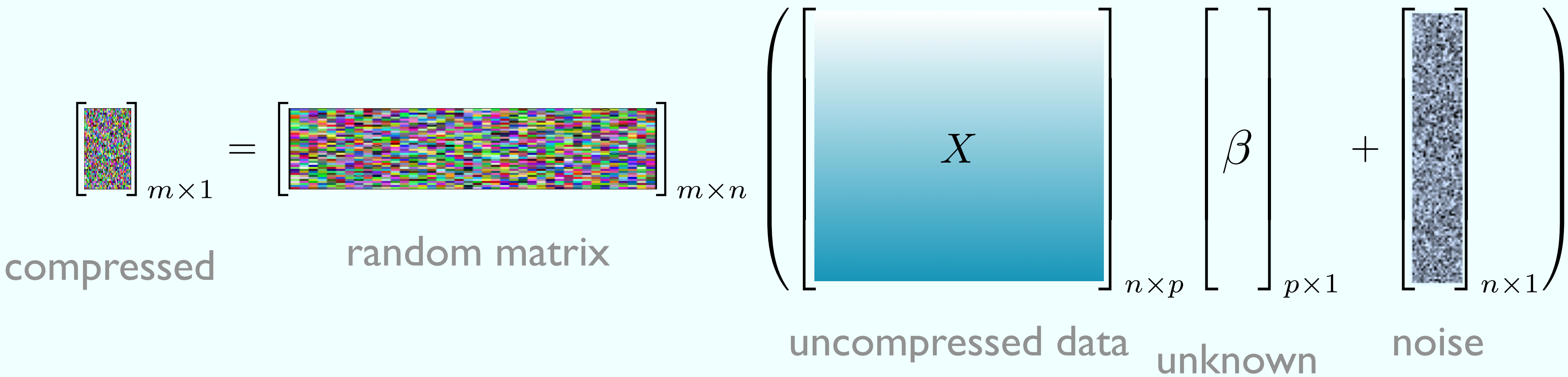
Compressed Regression

John Lafferty, Larry Wasserman, Shuheng Zhou*

School of Computer Science
Carnegie Mellon University

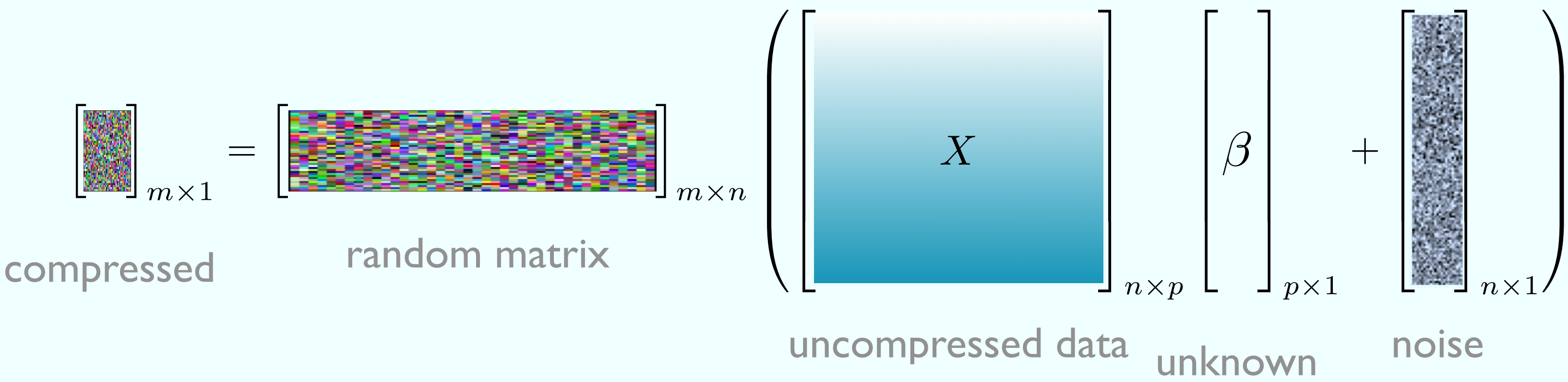
*Department of Statistics
University of Michigan

Basic Problem



Motivation: Scalability and privacy

Results



- Bounds on number of projections for accurate estimation
- Analysis of risk consistency
- Upper bounds on information rate of compressed data

Time

52.5 minutes = one μ-century

Goal for this talk: $\frac{1}{2}\,\mu$-century

Linear Regression

$$\begin{bmatrix} \\ Y \\ \\ \end{bmatrix}_n = \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{n \times p} \begin{bmatrix} \\ \beta \\ \\ \end{bmatrix}_p + \begin{bmatrix} \\ \epsilon \\ \\ \end{bmatrix}_n$$

Without compression

- The design matrix X is $n \times p$, where p grows with n
- The response vector $Y = X\beta + \epsilon$ is in $\mathbb{R}^n$. Lasso solves:

$$(P0) \qquad \min \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1$$

Compressed Linear Regression

$$\begin{bmatrix} \mathcal{Y} \end{bmatrix}_m = \begin{bmatrix} & \mathcal{X} & \end{bmatrix}_{m \times p} \begin{bmatrix} \beta \end{bmatrix}_p + \begin{bmatrix} \mathcal{E} \end{bmatrix}_m$$

Let $\Phi_{m \times n}$ be a (hidden) random Gaussian matrix. Observe

- compressed design matrix $\mathcal{X} = \Phi X$ in $\mathbb{R}^{m \times p}$ and
- compressed response $\mathcal{Y} = \Phi Y = \Phi X \beta + \Phi \epsilon$ in $\mathbb{R}^m$.

$$(P1) \qquad \min \frac{1}{2m} \|\mathcal{Y} - \mathcal{X}\beta\|_2^2 + \lambda_m \|\beta\|_1$$

- Complication: elements in noise vector $\varepsilon = \Phi \epsilon$ not i.i.d.

Sparsistency: Model selection consistency

Given the set of optimal solutions Ω_m to (P1)

$$\Omega_m = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2m} \|\mathcal{Y} - \mathcal{X}\beta\|_2^2 + \lambda_m \|\beta\|_1$$

Definition: A set of estimators Ω_m is **sparsistent** if

$$\mathbb{P}(\exists \beta_m \in \Omega_m,\ s.t.\ \mathrm{supp}(\beta_m) = \mathrm{supp}(\beta)) \to 1 \text{ as } m \to \infty.$$

Stronger condition: sign consistency

$$\mathbb{P}\left(\exists \beta_m \in \Omega_m\ s.t.\ \mathrm{sign}(\beta_m) = \mathrm{sign}(\beta)\right) \to 1 \text{ as } m \to \infty$$

Sparsistency: S-Incoherence

Sign consistency for compressed sparse linear regression is possible when the design matrix $\mathcal{X}$ is "sufficiently nice"

Let β be the true model, $S = \mathrm{supp}(\beta)$, and $S^c = \{1, .., p\} \backslash S$

S-Incoherence:

$$\left\| \frac{1}{n} \mathcal{X}_{S^c}^T \mathcal{X}_S \right\|_\infty + \left\| \frac{1}{n} \mathcal{X}_S^T \mathcal{X}_S - \mathcal{I}_{|S|} \right\|_\infty \leq 1 - \eta, \quad \text{some } \eta \in (0, 1]$$

Sparsistency Result

Theorem. Suppose that before compression, we have

$$Y = X\beta^* + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2 I_n),$$

- $X_{n\times p}$ is S-incoherent, where $S = \text{supp}(\beta^*)$, $\rho_m = \min_{i\in S} |\beta_i^*|$, and
- columns $\|X_j\|_2^2 = n, \forall j \in \{1, ..., p\}$.

Let $s = |S|$ and $\Phi_{m\times n}$ consist of i.i.d. $\Phi_{ij} \sim N(0, \frac{1}{n})$. Suppose that

$$\left(\frac{16C_1 s^2}{\eta^2} + \frac{4sC_2}{\eta}\right)\log 2pn^2(s+1) \le m \le \sqrt{\frac{n}{16\log n}}$$

with $C_1 \approx 2.5044$ and $C_1 \approx 7.6885$, and $\lambda_m \to 0$ satisfies

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

Then the compressed Lasso is sparsistent.

Sparsistency: Ingredients

By excluding the bad events, we can consider $\mathcal{X}_{m \times p}$ as a fixed matrix

- Similar conditions imposed on deterministic design matrix X for (P0) in Wainwright (2006), and Zhao and Yu (2007).
- The S-Incoherence condition is stronger.
- But we are in (P1), where $\varepsilon = \Phi\epsilon$, unlike ϵ in (P0), is not i.i.d.

Concentration Lemma. $\mathbb{E}(\Phi\Phi^T) = \mathcal{I}$; with high probability, each entry of $\Phi\Phi^T - \mathcal{I}_{m \times m}$ is at most $O\left(\sqrt{\frac{\log n}{n}}\right)$.

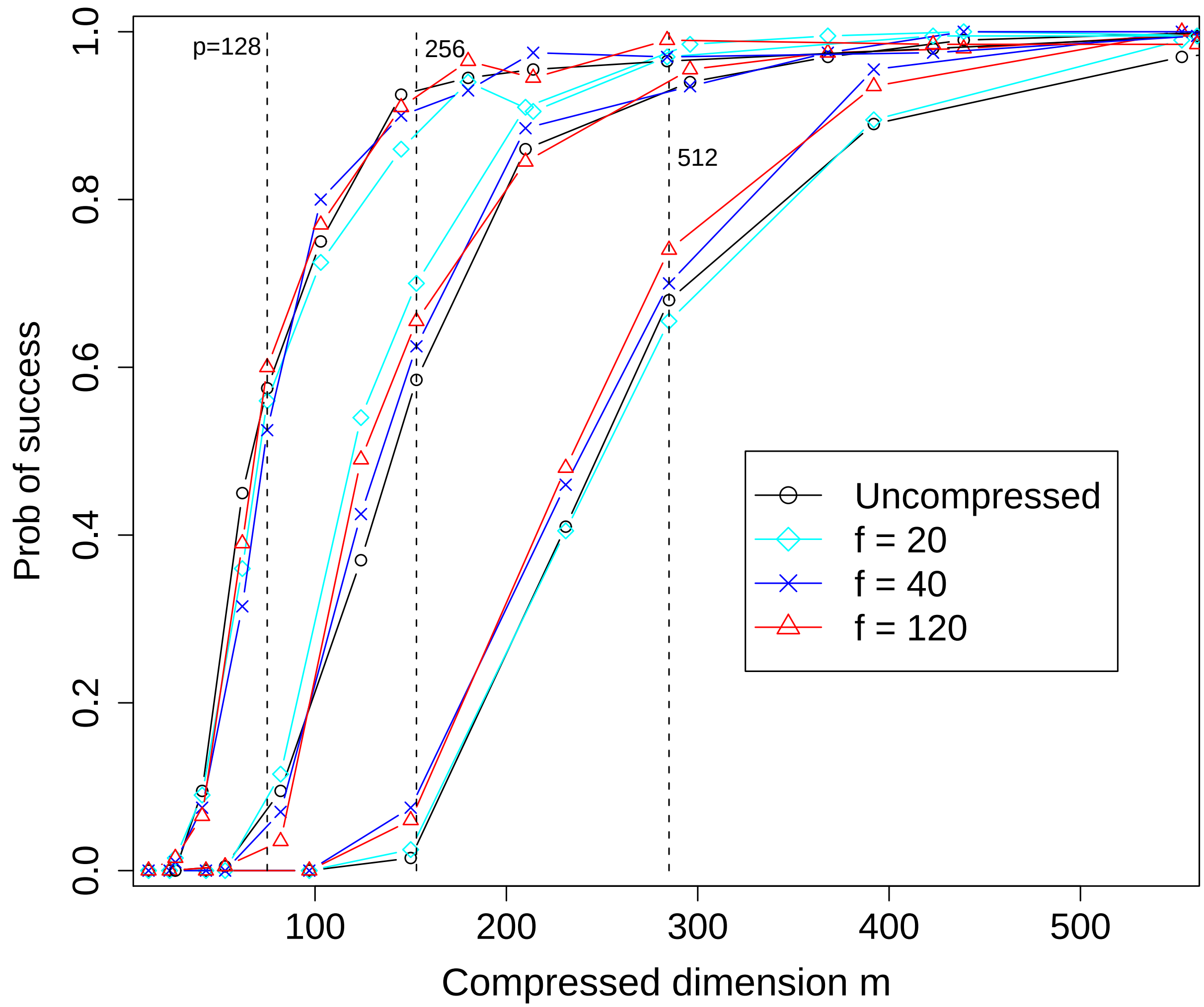
- Important in adapting Wainwright's proof in the (P0) setting for a fixed design to the compressed setting of (P1).

Cost of Compression

$$n = \Omega(s \log p) \quad \text{(uncompressed)}$$

$$m = \Omega(s^2 \log pn) \quad \text{(compressed)}$$

Compressed Lasso Sparsistency



Probability of correctly recovering true sparsity pattern, $p = 126, 256, 512$.

Risk Consistency

Roughly speaking, persistence means that the procedure predicts well. Given a sequence of sets of estimators B_n, the sequence of estimators $\widehat{\beta}_n \in B_n$ is called *persistent* (Greenshtein and Ritov, 2004) if

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

where $R(\beta) = \mathbb{E}(Y - X^T\beta)^2$ is the prediction risk of a new pair (X, Y).

- Linear model not assumed correct
- Answers the asymptotic question: How large may the set B_n be, so that it is still possible to empirically select a predictor whose risk is close to that of the best predictor in the set?
- Lasso is persistent when the order of magnitude for ℓ_1 radius L_n of B_n is restricted to $o\left((n/\log n)^{1/4}\right)$.

Compressed Lasso is Persistent

Theorem. Suppose $p = O(e^{n^c})$, $c < \frac{1}{2}$ and $\log^2(np) \leq m \leq n$. Let

$$L_{n,m} = o\left(\frac{m}{\log(np_n)}\right)^{1/4}.$$

Then the sequence of compressed lasso estimators

$$\widehat{\beta}_{n,m} = \underset{\|\beta\|_1 \leq L_{n,m}}{\arg\min} \|\mathcal{Y} - \mathcal{X}\beta\|_2^2$$

is persistent with respect to $B_{n,m} = \{\beta \ : \ \|\beta\|_1 \leq L_{n,m}\}$:

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

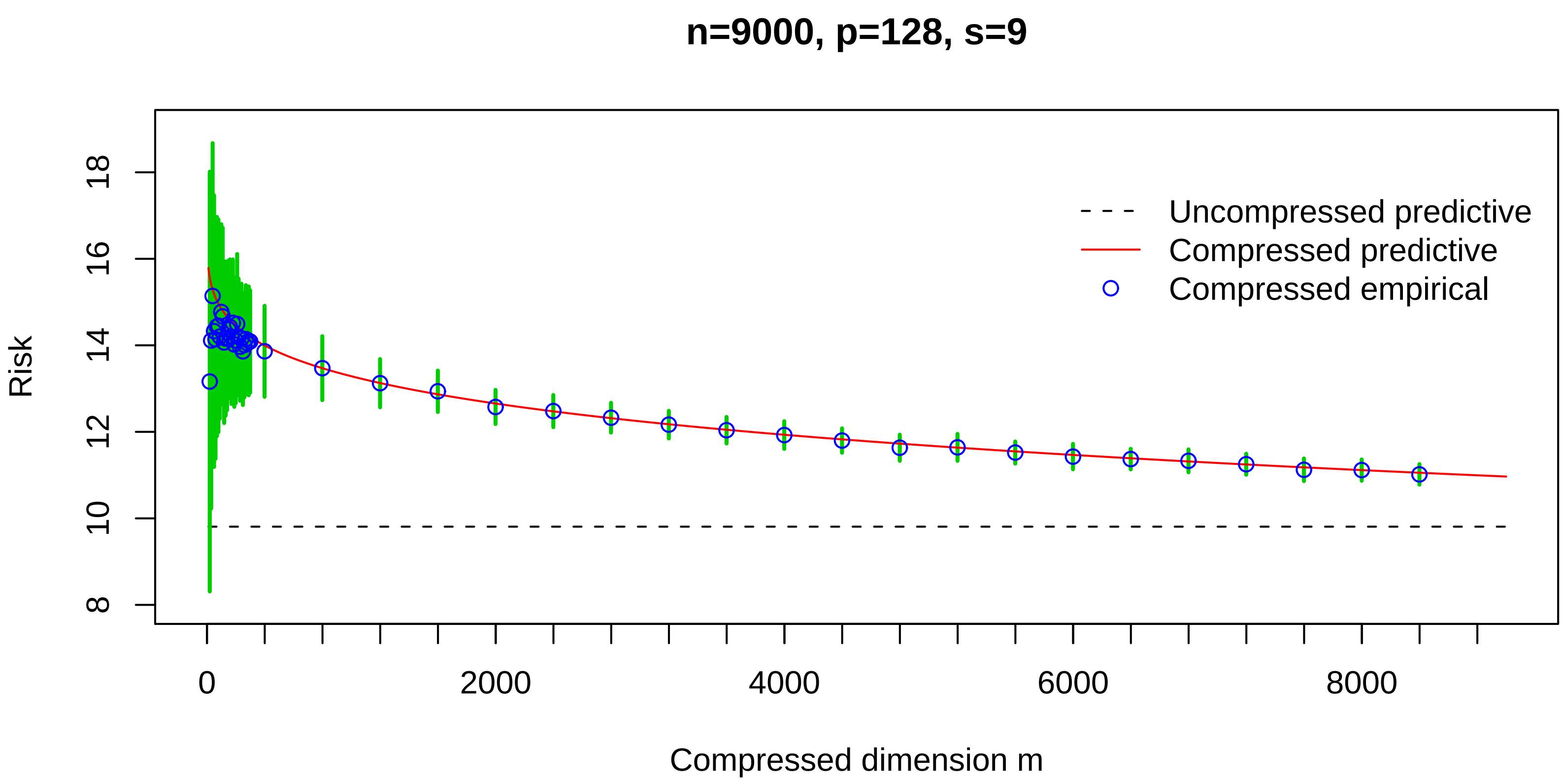
Cost of Compression

For simplicity take $L_n = O(1)$, $L_{n,m} = O(1)$, $p = n^c$ and $m = \Omega(\log^2 n)$. Then

$$\begin{aligned} R(\widehat{\beta}_n) - \inf_{\|\beta\|_1 \leq L_n} R(\beta) &= O_P\left(\sqrt{\frac{\log n}{n}}\right) \\ R(\widehat{\beta}_{n,m}) - \inf_{\|\beta\|_1 \leq L_{n,m}} R(\beta) &= O_P\left(\sqrt{\frac{1}{\log n}}\right) \end{aligned}$$

Ratio of compressed to uncompressed excess risks is $O(\sqrt{m/n})$.

Compressed Lasso Persistence



Each point corresponds to the mean empirical risk, over 100 trials. For each trial, randomly draw $X_{n \times p}$ with $x_i \sim N(0, T(0.1))$, with $T(\rho)_{i,j} = \rho^{|i-j|}$.

Privacy Analysis

General “matrix masking” takes the form $\mathcal{X} = AXB + C$

- Represents many possible schemes: subsampling, adding noise...
- Limited analysis of such schemes in privacy literature.

Multiple Wireless Antenna Model

Our setup corresponds to standard model for multiple antenna wireless communication (Marzetta and Hochwald, 1999).

- Have n transmitter and m receiver antennas over p time periods
- Allows model $\widetilde{X} = \Phi X + \Delta$
- When capacity of channel decays to zero, little information is conveyed about original data X from the compressed data $\mathcal{X}$

Privacy Analysis

Theorem. If $\mathbb{E}(X_j^2) \leq P$, the maximum information rate satisfies

$$r_{n,m} = \sup_{p(X)} \frac{I(X;\mathcal{X})}{np} \leq \frac{m}{2n} \log\left(2\pi e P\right)$$

- With $m = O(\log np)$ this gives the upper bound

$$r_{n,m} = O\left(\frac{\log np}{2n}\right) \to 0$$

- If compression matrix Φ is “leaked,” compressed sensing may allow reconstruction of sparse variables.
- Average case analysis.

Summary of Tradeoffs

- Variable selection: extra factor of s in sample complexity
- Excess risk rates: $O(\sqrt{m/n})$ uncompressed to compressed
- Information per symbol: $O(m/n)$

Summary

- Compressing the design matrix across rows has little impact on effectiveness of sparse regression
- Expect similar results hold for nonparametric regression
- Privacy guarantees are information-theoretic, average case.

For all the details, please see S. Zhou, J. Lafferty and L. Wasserman, "Compressed and privacy-sensitive sparse regression," IEEE Trans. Info. Theory, Vol 55, No. 2, 2009