Computation-Risk Tradeoffs for Covariance-Thresholded Regression

Dinah Shender* DSHENDER@UCHICAGO.EDU

John Lafferty*‡ LAFFERTY@UCHICAGO.EDU

Department of Statistics*, Department of Computer Science‡, University of Chicago, Chicago, IL 60637 USA

Abstract

We present a family of linear regression estimators that provides a fine-grained tradeoff between statistical accuracy and computational efficiency. The estimators are based on hard thresholding of the sample covariance matrix entries together with ℓ_2-regularizion (ridge regression). We analyze the predictive risk of this family of estimators as a function of the threshold and regularization parameter. With appropriate parameter choices, the estimate is the solution to a sparse, diagonally dominant linear system, solvable in near-linear time. Our analysis shows how the risk varies with the sparsity and regularization level, thus establishing a statistical estimation setting for which there is an explicit, smooth tradeoff between risk and computation. Simulations are provided to support the theoretical analyses.

1. Introduction

Modern data sets used for statistical analysis are often large and high dimensional. The computation required to construct standard estimators for such data may be prohibitive. In this setting it is attractive to tradeoff statistical accuracy for computational scalability—tolerating increased predictive error, or risk, in exchange for more favorable computational requirements. While several heuristics for reduced computation are often possible, including dimension reduction, sampling, and greedy algorithms, little is known about precise tradeoffs between risk and computation. In the setting of Bayesian inference using MCMC algorithms, for instance, limiting computation by early stopping of the Markov chain will introduce bias and increase risk; but a quantitative understanding of this tradeoff is generally lacking.

Linear regression is a workhorse method for many statistical problems. But without special assumptions, the method has quadratic computational cost $O(np^2)$ in the dimension p, when the sample size n is larger than p. This may be prohibitive when p is large. In this work we study a concrete, practical way to smoothly tradeoff risk for computation in linear regression, by sparsifying the sample covariance with hard thresholding.

The standard ridge regression estimator is

$$\widehat{\beta}_\lambda = \Big(\frac{1}{n}\mathbb{X}^T\mathbb{X} + \lambda_n I\Big)^{-1}\frac{1}{n}\mathbb{X}^T Y \tag{1}$$

$$= (\mathbb{S} + \lambda_n I)^{-1}\,\boldsymbol{b}_n \tag{2}$$

where $\mathbb{X}$ is the $n \times p$ design matrix, $\mathbb{S} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the sample covariance, and $\boldsymbol{b}_n = \frac{1}{n}\mathbb{X}^T Y$ is the sample marginal correlation for data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, assuming for convenience that the data are scaled to have mean zero and variance one. We consider the family of estimators

$$\widetilde{\beta}_{t,\lambda} = \big(\mathbb{S}_t + \lambda_n I\big)^{-1}\boldsymbol{b}_n \tag{3}$$

where $\mathbb{S}_t$ is a sparsified version of the sample covariance obtained by hard thresholding, to zero out the small entries. That is, $\mathbb{S}_t = T_t(\mathbb{S})$ where

$$T_t([m_{ij}]) = \big[m_{ij}\mathbb{1}(|m_{ij}| > t)\big]. \tag{4}$$

The basic intuition is that as we increase the threshold t, so that the matrix $\mathbb{S}_t$ becomes more sparse, the model degrades, but the estimator can be obtained with less computation. For sufficiently large regularization level λ_n and sparsity threshold t, the linear system

$$\big(\mathbb{S}_t + \lambda_n I\big)\beta = \boldsymbol{b}_n \tag{5}$$

is sparse and symmetric diagonally dominant (SDD). Recent research in algorithms and scientific computation has led to a breakthrough in fast solvers for

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

such systems. In particular, work of Spielman & Teng (2009) and Koutis et al. (2012) shows that sparse SDD systems can be solved in near linear time in the number of nonzero entries in the matrix. We adopt a computational model in which the sparsification $\mathbb{S}_t$ is not included in the computation cost. The calculation of $\mathbb{S}_t$ is parallelizable in a simple and direct manner, and the cost of the computation can be amortized over different regressions. We discuss this point further in the conclusion to the paper.

The main contribution of the present paper is to combine this computational analysis with a statistical analysis of the predictive risk for this family of linear models, making precise the tradeoff between computation and error. In the following section we briefly mention some previous work on risk-computation tradeoffs. In Section 3 we give a high level summary of our results, with the detailed assumptions and theorems given in Section 4. Section 5 presents numerical simulations that illustrate the methods and analysis. Details of the proofs are given in Section 6.

2. Related Work

The development of statistical methodology that provides a way of controlling tradeoffs between computation and accuracy is relatively new. However, with the growing attention on large scale data analysis in recent years, researchers have begun to focus more on this problem.

Sparse PCA is one problem that has been studied from the perspective of trading off computation for sample complexity. In the stylized setting of a sparse rank one covariance corrupted by noise, where the principal eigenvector of dimension p has only k nonzero entries, a simple thresholding algorithm has been shown to have sample complexity $O(k^2 \log(p-k))$ with computational complexity $O(np + p \log p)$. In constrast, a more expensive semidefinite relaxation algorithm is known to have lower sample complexity $O(k \log(p-k))$ at the expense of greater computational complexity $O(np^2 + p^4 \log p)$ (Johnstone & Lu, 2004; d'Aspremont et al., 2004; Amini & Wainwright, 2009). These analyses assume, however, that the solution has rank one. This problem has also been studied by Chandrasekaran & Jordan (2012).

Shalev-Shwartz et al. (2010) propose an algorithm for sparse linear prediction where the excess risk is inversely proportional to the sparsity of the estimator. Their approach is based on forward greedy selection modified to reweight the components at each step. For the regularized risk $R(w) = \mathbb{E}[L(w^T x, y)] + \frac{\lambda}{2}\|w\|_2^2$, it is shown that $R(w^{(k)}) \leq R(w^*) + \epsilon$ if the number of steps k satisfies

$$k \geq \|w^*\|_0 \frac{C}{\lambda} \log\left(\frac{R(0) - R(w^*)}{\epsilon}\right), \qquad (6)$$

where $w^{(k)}$ is the estimator after k steps and w^* is a reference model. Since the number of iterations controls both the sparsity and the computation time, this implicitly establishes a relationship between computation and excess risk.

In the setting of online learning, Agarwal et al. (2012) consider model selection under a computational budget constraint, assuming that computation grows linearly with sample size. They bound excess risk for a grid search over a nested family of models. Although the relationship between computation time and risk is explicit, the linearity assumption may be unrealistic for many classes of models.

A growing body of work has investigated algorithms for scaling regression to large data sets, notably using coresets (Drineas et al., 2006; Dasgupta et al., 2009; Clarkson et al., 2013). Recently, Clarkson & Woodruff (2013) proposed a new algorithm for generating subspace embedding matrices, which yields a regression algorithm running in $O(m(X)) + \tilde{O}(p^3\epsilon^{-2})$ time, with $m(X)$ denoting the number of nonzeros of X. The procedure only increases error by a multiplicative factor of $1+\epsilon$; but the requirement of a sparse design matrix may be limiting.

Another sort of tradeoff is given by early stopping of stochastic gradient descent, which converges at rate $O(1/t)$ in the number of steps t, for strongly convex functions (Robbins & Monro, 1951; Bach & Moulines, 2011). This gives an approximate solution to the regression problem. In contrast, the approach we explore here is to approximate the original regression problem itself, which is then solved exactly.

3. Setup and Overview of Results

In this section we provide a high level overview of our assumptions, framework, and results. The following sections give more formal, precise statements of the results, together with proofs.

We consider ridge estimation in a random design setting with fixed $p < n$, so that we have $Y = \mathbb{X}\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ independent of $\mathbb{X}$ and Y, where $\mathbb{X}$ is an $n \times p$ design matrix with $\mathbb{X}_{ij}$ the jth covariate of the ith datapoint. Let $\mathbb{X}_{i-} \in \mathbb{R}^n$ refer to the ith row. Assume that the observations $\mathbb{X}_{i-}$ are iid and let $\mathbb{S} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ and $\Sigma = \mathbb{E}(\mathbb{S})$ be the sample and population covariance, respectively. Let $\mathbb{S}_t = T_t(\mathbb{S})$ be the

thresholded sample covariance, where

$$T_t([m_{ij}]) = [m_{ij}\mathbf{1}(|m_{ij}| > t)]. \tag{7}$$

When we make statements about the computational cost, we will assume that $\mathbb{S}_t + \lambda I$ is a diagonally dominant matrix. While this is true for sufficiently large λ, it will also hold (with high probability) when the population covariance $\Sigma = Cov(X)$ is diagonally dominant.

We modify the usual ridge estimator to be the solution to a SDD system

$$\widehat{\beta}_{t,\lambda} = (\mathbb{S}_t + \lambda I)^{-1}\frac{1}{n}\mathbb{X}^T Y \tag{8}$$

$$= (\mathbb{S}_t + \lambda I)^{-1}\boldsymbol{b}_n \tag{9}$$

where $\boldsymbol{b}_n = \frac{1}{n}\mathbb{X}^T Y$. We also define

$$\beta_\lambda = (\Sigma + \lambda I)^{-1}\Sigma\beta^* \tag{10}$$

$$\widetilde{\beta}_{t,\lambda} = (\mathbb{S}_t + \lambda I)^{-1}\mathbb{S}\beta^*. \tag{11}$$

These are the population ridge estimator and the conditional expectation of $\widehat{\beta}_{t,\lambda}$ given $\mathbb{X}$, respectively.

Let $\|\cdot\|_A$ denote the norm relative to a positive semidefinite matrix A, so that

$$\|x\|_A = \sqrt{x^T A x}. \tag{12}$$

We can write the excess risk over a new pair (X, Y) in terms of this norm as

$$\mathbb{E}(Y - X^T\beta)^2 - \mathbb{E}(Y - X^T\beta^*)^2 = \|\beta^* - \beta\|_\Sigma^2. \tag{13}$$

We leverage three previous results in our approach. First, we work with the family of sparse covariance matrices studied by Bickel & Levina (2008), defined by

$$U_{q,\epsilon_0} = \Big\{\Sigma \,:\, \max_i \sigma_{ii} \le M, \quad \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p) \\ \lambda_{\min}(\Sigma) > \epsilon_0 > 0\Big\} \tag{14}$$

with $0 \le q < 1$ and $\lambda_{\min}$ denoting the smallest eigenvalue. This class constrains the covariance as having rows lying in an ℓ_q ball; the matrices are sparse when $q = 0$ and $c_0(p)$ is small. It also requires the covariance to have eigenvalues bounded away from 0. Second, we adapt some of the results of Hsu et al. (2011) on ridge regression to our setting, as described in Section 4.

Finally, the computational analysis relies on recent developments in fast solvers for SDD systems. For example, for an SDD system $Ax = b$ of dimension p with m nonzero entries in A, the analysis of Koutis et al. (2012) shows that the solution x^* can be obtained to accuracy ϵ in near linear time. More precisely, the algorithm forms a chain of preconditioners that yields an ϵ-approximate solution $\widehat{x}$, so that

$$\|x^* - \widehat{x}\|_A \le \epsilon\|x^*\|_A, \tag{15}$$

in time $T(m, p, \epsilon)$ satisfying

$$T(m, p, \epsilon) = \widetilde{O}\big(m \log p \log(1/\epsilon)\big), \tag{16}$$

the notation $\widetilde{O}$ hiding a factor of order $(\log\log p)^2$.

Our main result analyzes the excess risk of the covariance-thresholded regression estimation. The assumptions required are detailed in Section 4.

Theorem 1. *Suppose that the covariance $\Sigma \in U_{q,\epsilon_0}$, and the regularization parameter satisfies $\lambda = O(n^{-1/2})$. Then the excess risk of the estimator $\widehat{\beta}_{t,\lambda}$ solving (9) is bounded in probability as*

$$\|\widehat{\beta}_{t,\lambda} - \beta^*\|_\Sigma = O_P\Big(\big(t^{1-q} + t^{-q}n^{-1/2} + \lambda\big)\|\beta^*\|\Big). \tag{17}$$

Moreover, given $\mathbb{S}$, and assuming $\mathbb{S}_t + \lambda I_p$ is diagonally dominant, the estimator can be computed in time

$$T(m_{n,t}, p) = \widetilde{O}\left(m_{n,t} \log p \log n\right) \tag{18}$$

where $m_{n,t}$ is the number of nonzero entries in the thresholded covariance matrix $\mathbb{S}_t$.

As the threshold t increases, the number of nonzero elements $m_{n,t}$ decreases, and the computation time scales roughly linearly in this value. The result shows how the excess risk increases as a function of this threshold for the class U_{q,ϵ_0}. The excess risk can be decomposed as

$$\|\widehat{\beta}_{t,\lambda} - \beta^*\|_\Sigma^2 \le \\ 3\left(\|\widehat{\beta}_{t,\lambda} - \widetilde{\beta}_{t,\lambda}\|_\Sigma^2 + \|\widetilde{\beta}_{t,\lambda} - \beta_\lambda\|_\Sigma^2 + \|\beta_\lambda - \beta^*\|_\Sigma^2\right). \tag{19}$$

The key step in the proof of the result is to bound the second term on the righthand side of (19), which includes both the error due to the random design and the error due to thresholding. The first term is the error due to the finite sample size, and can be bounded as $O_P\left(\sigma^2/n\right)$. This part of our analysis reuses and extends results of Hsu et al. (2011). The third term is the approximation error due to regularization.

4. Main Results

We make the following assumptions:

1. $\|\Sigma - \mathbb{S}_t\| < \lambda + \lambda_{\min}(\Sigma)$.

2. The $\mathbb{X}_{i-}$ are mean zero sub-Gaussian.

Assumption 1 insures that $\mathbb{S}_t + \lambda I$ is positive definite, since by Weyl's Theorem

$$\lambda_{\min}(\mathbb{S}_t) \geq \lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma - \mathbb{S}_t) > -\lambda. \quad (20)$$

Assumption 2 implies that for every $t \in \mathbb{R}^p$

$$\mathbb{E}\exp(t^T \mathbb{X}_{i-}) \leq e^{-\|t\|^2\alpha^2} \quad (21)$$

for some $\alpha > 0$.

Write $\lambda_j(\Sigma)$ to be the j^{th} largest eigenvalue of Σ and define

$$d_{1,\lambda} = \sum_j \frac{\lambda_j(\Sigma)}{\lambda_j(\Sigma)+\lambda} \qquad d_{2,\lambda} = \sum_j \left(\frac{\lambda_j(\Sigma)}{\lambda_j(\Sigma)+\lambda}\right)^2 \quad (22)$$

The important ingredient in bounding the second term is Theorem 1 of Bickel & Levina (2008). They define two classes of sparse covariance matrices:

$$U_q = \{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^p |\sigma_{ij}|^q \leq c_0(p), \text{ for all } i\} \quad (23)$$

$$U_{q,\epsilon_0} = \{\Sigma : \Sigma \in U_q \text{ and } \lambda_{\min}(\Sigma) \geq \epsilon_0 > 0\} \quad (24)$$

for fixed $0 \leq q < 1$.

Theorem 2. *(Bickel-Levina) In the current setting, assume that $\Sigma \in U_q$. If $\log p/n = o(1)$, then*

$$\|\mathbb{S}_t - \Sigma\| = O_P\left(c_0(p)t^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)t^{1-q}\right).$$

Moreover, if $\Sigma \in U_{q,\epsilon_0}$ then

$$\|\mathbb{S}_t^{-1} - \Sigma^{-1}\| = O_P\left(c_0(p)t^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)t^{1-q}\right).$$

Remarks:

1. Since we consider p fixed, the condition that $\log p/n = o(1)$ is simply $n \to \infty$.
2. For $\Sigma \in U_q$ and symmetric, we have the bound
$$\|\Sigma\| = \lambda_{\max}(\Sigma) \leq \max_i \sum_j |\sigma_{ij}| \leq M^{1-q}c_0(p).$$
3. Both Theorem 2 and Assumption 1 bound $\|\mathbb{S}_t - \Sigma\|$. The terms in Theorem 2 are balanced when $t = n^{-1/2}$, in which case the bound becomes $O_P\left(n^{\frac{q-1}{2}}\right)$. As we will show, the risk is minimized when $\lambda \asymp n^{\frac{q-1}{2}}$. Assumption 1 is weaker than Theorem 2 when $\lambda_{\min}(\Sigma) = \Omega\left(n^{\frac{q-1}{2}}\right)$.

We can now state the bounds for each term in (19), together with the general version of the final bound. The proofs are largely technical and are left to Section 6.

Lemma 3. *Under Assumptions 1 and 2, if $\Sigma \in U_{q,\epsilon_0}$, then the random design and thresholding error can be bounded as*

$$\frac{\|\widetilde{\beta}_{t,\lambda} - \beta_\lambda\|_\Sigma^2}{\|\beta^*\|^2} = O_P\left(2M^{3(1-q)}c_0^5(p)\left(t^{-q}\sqrt{\frac{\log p}{n}} + t^{1-q}\right)^2 + C\frac{p}{n}\right). \quad (25)$$

Recall that $\widetilde{\beta}_{t,\lambda}$ and β_λ solve

$$(\mathbb{S}_t + \lambda I)\,\widetilde{\beta}_{t,\lambda} = \mathbb{S}\beta^* \quad (26)$$

$$(\Sigma + \lambda I)\,\beta_\lambda = \Sigma\beta^*. \quad (27)$$

We can view each of these as a perturbation of the system

$$(\Sigma + \lambda I)\,\beta' = \mathbb{S}\beta^*. \quad (28)$$

The first term in the bound comes from the perturbation analysis of (26), and the second term in the bound comes from the perturbation analysis of (27). In particular, the dependence on t comes from considering $(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}$ and applying Theorem 2.

The rest of the terms are more straightforward.

Lemma 4. *Under Assumptions 1 and 2, if $\Sigma \in U_{q,\epsilon_0}$, then the stochastic error is bounded in probability by*

$$\|\widehat{\beta}_{t,\lambda} - \widetilde{\beta}_{t,\lambda}\|_\Sigma^2 = O_P\left(\frac{\sigma^2 d_{2,\lambda}}{n}\right). \quad (29)$$

Note that this is the usual order bound for Gaussian error in a model of dimension $d_{2,\lambda}$.

Lemma 5. *The approximation error can be bounded as*

$$\|\beta_\lambda - \beta\|_\Sigma^2 = \sum_{j=1}^p \beta_j^2 \frac{\lambda^2\lambda_j(\Sigma)}{(\lambda + \lambda_j(\Sigma))^2} = O(\lambda^2\|\beta^*\|^2). \quad (30)$$

Putting these bounds together leads to our main result.

Theorem 6. *Under the above assumptions, if $\Sigma \in U_{q,\epsilon_0}$, the excess risk is bounded by*

$$\|\widehat{\beta}_{t,\lambda} - \beta^*\|_\Sigma^2 = O_P\Big(\frac{\sigma^2 d_{2,\lambda}}{n} + \Big\{2M^{3(1-q)}c_0^5(p)\big(t^{-q}\sqrt{\frac{\log p}{n}} + t^{1-q}\big)^2 + C\frac{p}{n} + \lambda^2\Big\}\|\beta^*\|^2\Big). \quad (31)$$

Moreover, this is minimized for the threshold $t \asymp n^{-1/2}$ *and regularization level* $\lambda = t^{1-q} \asymp \left(\frac{1}{n}\right)^{\frac{1-q}{2}}$*, resulting in the bound*

$$\|\widehat{\beta}_{t,\lambda} - \beta^*\|_\Sigma^2 = O_P\left(\lambda^2 \|\beta^*\|^2\right). \tag{32}$$

This result shows how the excess risk increases with the threshold t and regularization level λ. As the threshold t increases, the computation time to solve the system decreases, although the precise manner in which the sparsity of $\mathbb{S}_t$ varies with t depends on the exact coefficients of the covariance Σ. As we will see in the next section, when the true covariance is near-sparse and nearly SDD, as for an AR-type covariance matrix, a very large decrease in the computation time can be obtained for a very small value of t.

5. Simulations

The simulations compare the excess risk of $\widehat{\beta}_{t,\lambda}$ for different values of n and p for both sparse and non-sparse Σ. We plot the risk against both the threshold t and the sparsity of $\mathbb{S}_t$, since the latter controls the computation time. We also make comparisons with the excess risk of the population estimator $\beta_{t,\lambda} = (\Sigma_t + \lambda I_p)^{-1}\Sigma\beta^*$. These simulations support our theoretical results and show how a moderate increase in the risk may provide a significant decrease in the computation time.

For the sparse case, we chose Σ to have $2p$ off-diagonal entries, which were generated as Uniform(-1, 1). The diagonals were set equal to $\Sigma_{ii} = 1 + \sum_{j\neq i}|\Sigma_{ij}|$ in order to guarantee that Σ was symmetric diagonally dominant. Since Σ is sparse, this corresponds to $q = 0$, i.e., $\Sigma \in U_{0,\epsilon_0}$. Theorem 6 then implies that

$$\|\widehat{\beta}_{t,\lambda} - \beta^*\|_\Sigma^2 = O_p\Big(\big(\lambda^2 + \big(n^{-1/2} + t\big)^2\big)\|\beta^*\|^2 + \frac{\sigma^2}{n}\Big). \tag{33}$$

By using a slight adaptation of Theorem 2, we can see that in the population setting

$$\|\beta_{t,\lambda} - \beta^*\|^2 = O_p((t^2 + \lambda^2)\|\beta^*\|^2). \tag{34}$$

For the non-sparse case, we took Σ to be an AR(1) covariance matrix with $\rho = .7$, i.e., $\Sigma_{ij} = \rho^{|i-j|}$. This is not a diagonally dominant matrix. However the size of the entries decreases quickly so that, even for small values of t, we do not need to take λ much larger than is optimal in order to make $\mathbb{S}_t + \lambda I$ diagonally dominant. Note that in our simulations we did not choose λ to make $\mathbb{S}_t + \lambda I$ diagonally dominant and did not use an SDD solver in either set of simulations.

For both the sparse and the AR(1) case we took sample size to be n =1000, 2000, 3000, 5000, 10,000, and 100,000 and set $\sigma^2 = 100$. For each value of n and p, we found the optimal value of $\lambda(t)$ for $\beta_{t,\lambda}$ at each value of threshold t. We averaged the excess risk of $\widehat{\beta}_{t,\lambda(t)}$ over 10 different samples of $\mathbb{X}$ and plotted this against the value of the threshold. The excess risk for $\beta_{t,\lambda(t)}$ is also shown. As n gets larger, the excess risk approaches the excess risk of the population estimator.

The computation required to solve the system depends not directly on the threshold, but on the number of nonzero entries in $\mathbb{S}_t$. In order to illustrate the tradeoff between computation and risk, for each combination of n and p we took a single trial and plotted the excess risk against the proportion of zero entries of $\mathbb{S}_t$. For smaller sample sizes, a small threshold, which only increases the risk moderately, can greatly increase the sparsity of $\mathbb{S}_t$ and therefore decrease the computation time.

6. Proofs of Technical Results

The main work is in bounding the random design and thresholding error term in (19). We prove the following version of Lemma 3:

Lemma 7. *Under Assumptions 1 and 2, if* $\Sigma \in U_{q,\epsilon_0}$

$$\frac{\|\widetilde{\beta}_{t,\lambda} - \beta_\lambda\|_\Sigma^2}{\|\beta^*\|^2} = O_P\left(\left\{\left(2M^{3(1-q)}c_0^3(p) + \frac{Cp}{n}\right)\cdot\left(c_0(p)t^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)t^{1-q}\right)^2\right\} + \frac{1}{\epsilon_0 + \lambda}\frac{p}{n}\right). \tag{35}$$

Proof. Starting with the definitions, we have

$$\begin{aligned}\|\widetilde{\beta}_{t,\lambda} - \beta_\lambda\|_\Sigma^2 &= \|(\mathbb{S}_t + \lambda I)^{-1}\mathbb{S}\beta^* - (\Sigma + \lambda I)^{-1}\Sigma\beta^*\|_\Sigma^2 \\ &\leq 3\|\left[(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}\right]\mathbb{S}\beta^*\|_\Sigma^2 \\ &\quad + 3\|(\Sigma + \lambda I)^{-1}(\mathbb{S} - \Sigma)\beta^*\|_\Sigma^2. \end{aligned}\tag{36}$$

We can bound the first term using Theorem 2, giving

$$\begin{aligned}&\|\left[(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}\right]\mathbb{S}\beta^*\|_\Sigma^2 \\ &\leq \|\Sigma\|\|(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}\|^2\|\|\mathbb{S}\|^2\|\beta^*\|^2.\end{aligned}\tag{37}$$

Theorem 2 gives a bound for $\|\mathbb{S}_t^{-1} - \Sigma^{-1}\|$, but is easily adapted to show that $\|(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}\| = \Theta\left(\|(\mathbb{S}_t + \lambda I) - (\Sigma + \lambda I)\|\right)$, so the bound stays the same. Additionally,

$$\begin{aligned}\|\mathbb{S}\|^2 &\leq 2\left(\|\Sigma\|^2 + \|\Sigma - \mathbb{S}\|^2\right) &(38)\\ &\leq 2\left[\left(M^{1-q}c_0(p)\right)^2 + C\frac{p}{n}\right] &(39)\end{aligned}$$

for some constant C. The second inequality follows from (25) together with concentration bounds for the

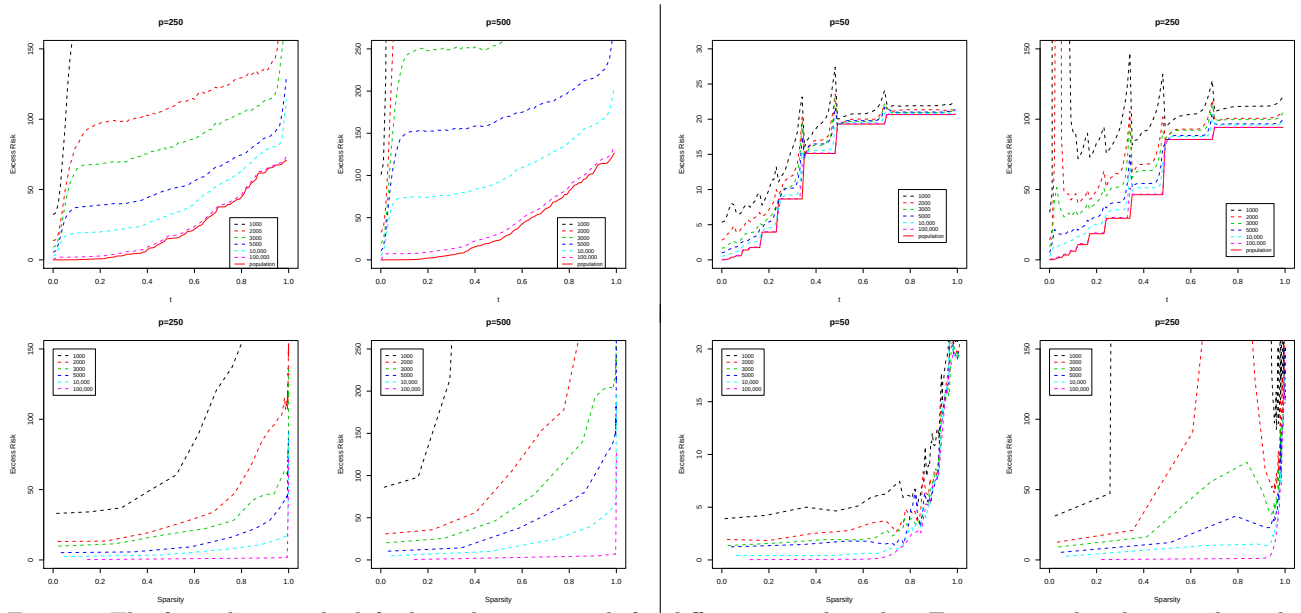

Figure 1. The four plots on the left show the excess risk for different n and p when Σ is sparse; the plots on the right show the excess risk when Σ is an AR(1) matrix with $\rho = .7$. In the first row the risk is plotted against the threshold. In the second row the risk is plotted against the proportion of zeros in $\mathbb{S}_t$.

sample covariance (Vershynin, 2012). Together these imply that

$$\begin{aligned} &\left\| \left[(\mathbb{S}_t + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1} \right] \mathbb{S}\beta^* \right\|_\Sigma^2 \qquad (40) \\ &= O_P\left(2M^{1-q} c_0(p) \left(M^{2(1-q)} c_0^2(p) + C\frac{p}{n} \right) \right. \\ &\quad \left. \cdot \left(c_0(p) t^{-q} \sqrt{\frac{\log p}{n}} + c_0(p) t^{1-q} \right)^2 \|\beta^*\|^2 \right). \end{aligned}$$

Bounding the second term in (36) is straightforward:

$$\begin{aligned} &\|(\Sigma + \lambda I)^{-1}(\mathbb{S} - \Sigma)\beta^*\|_\Sigma^2 \\ &\quad \le \|\Sigma(\Sigma + \lambda I)^{-1}\| \|(\Sigma + \lambda I)^{-1}\| \|\mathbb{S} - \Sigma\|^2 \|\beta^*\|^2 \\ &\quad = O_P\left(\frac{1}{\epsilon_0 + \lambda} \frac{p}{n} \|\beta^*\|^2 \right). \qquad (41) \end{aligned}$$

Together, (40) and (41) give the result. □

In order to handle the stochastic error term in (19), we make the following definition:

$$K_{t,\lambda} = \frac{\lambda_{\min}(\Sigma) + \lambda}{\lambda_{\min}(\Sigma) + \lambda - \|\mathbb{S}_t - \Sigma\|}. \qquad (42)$$

We can bound $K_{t,\lambda}$ in probability directly from Theorem 2.

Corollary 8. *There exists a constant ξ such that with probability $1 - \delta$,*

$$\begin{aligned} K_{t,\lambda} &= \frac{\lambda_{\min}(\Sigma) + \lambda}{\lambda_{\min}(\Sigma) + \lambda - \|\mathbb{S}_t - \Sigma\|} \qquad (43) \\ &\le \frac{\lambda_{\min}(\Sigma) + \lambda}{\lambda_{\min}(\Sigma) + \lambda - \xi \left(c_0(p) t^{-q} \sqrt{\frac{\log p}{n}} + c_0(p) t^{1-q} \right)} \\ &:= C_{t,\lambda}. \end{aligned}$$

The term $C_{t,\lambda}$ depends on δ, but that dependence is suppressed from the notation. It plays a role analogous to $K_{\lambda,\delta,n}$ in (Hsu et al., 2011), and is needed to bound the following quantity.

Lemma 9. *Under assumption 1,*

$$\left\| (\Sigma + \lambda I)^{1/2} (\mathbb{S}_t + \lambda I)^{-1} (\Sigma + \lambda I)^{1/2} \right\| \le K_{t,\lambda} \qquad (44)$$

and so with probability $1 - \delta$

$$\left\| (\Sigma + \lambda I)^{1/2} (\mathbb{S}_t + \lambda I)^{-1} (\Sigma + \lambda I)^{1/2} \right\| \le C_{t,\lambda}. \qquad (45)$$

Proof.

$$\begin{aligned} &(\Sigma + \lambda I)^{-1/2} (\mathbb{S}_t + \lambda I) (\Sigma + \lambda I)^{-1/2} \qquad (46) \\ &\quad = I + (\Sigma + \lambda I)^{-1/2} (\mathbb{S}_t + \lambda I - \Sigma - \lambda I) (\Sigma + \lambda I)^{-1/2} \\ &\quad = I + (\Sigma + \lambda I)^{-1/2} (\mathbb{S}_t - \Sigma) (\Sigma + \lambda I)^{-1/2} \end{aligned}$$

Then we have

$$\lambda_{\min}(I + (\Sigma+\lambda I)^{-1/2}(\mathbb{S}_t - \Sigma)(\Sigma+\lambda I)^{-1/2}) \quad (47)$$
$$= 1 + \lambda_{\min}((\Sigma+\lambda I)^{-1/2}(\mathbb{S}_t - \Sigma)(\Sigma+\lambda I)^{-1/2}) \quad (48)$$
$$\geq 1 - \|(\Sigma+\lambda I)^{-1/2}(\mathbb{S}_t - \Sigma)(\Sigma+\lambda I)^{-1/2}\| \quad (49)$$
$$\geq 1 - \frac{\|\mathbb{S}_t - \Sigma\|}{\lambda_{\min}(\Sigma)+\lambda}. \quad (50)$$

By Assumption 1, this lower bound is always positive. Note that the first inequality is strict unless the matrix product is not positive definite, in which case $\lambda_{\min}$ might be equal in absolute value to the norm.

Since assumption 1 implies that $(\mathbb{S}_t+\lambda I)^{-1}$ is positive definite, we have that

$$\|(\Sigma+\lambda I)^{1/2}(\mathbb{S}_t+\lambda I)^{-1}(\Sigma+\lambda I)^{1/2}\| \quad (51)$$
$$= \lambda_{\max}\left((\Sigma+\lambda I)^{1/2}(\mathbb{S}_t+\lambda I)^{-1}(\Sigma+\lambda I)^{1/2}\right) \quad (52)$$
$$= \frac{1}{\lambda_{\min}\left((\Sigma+\lambda I)^{-1/2}(\mathbb{S}_t+\lambda I)(\Sigma+\lambda I)^{-1/2}\right)} \quad (53)$$
$$\leq \frac{\lambda_{\min}(\Sigma)+\lambda}{\lambda_{\min}(\Sigma)+\lambda-\|\mathbb{S}_t-\Sigma\|}. \quad \square$$

The final ingredient in the bound of the stochastic error term is a bound on the Frobenius norm of the λ-whitened version of $\mathbb{S}-\Sigma$. This bound is stated and proved as Lemma 9 of Hsu et al. (2011).

Lemma 10. *With probability at least* $1-\delta$,

$$\|(\Sigma+\lambda I)^{-1/2}(\mathbb{S}-\Sigma)(\Sigma+\lambda I)^{-1/2}\|_F \quad (54)$$
$$\leq (1+\sqrt{8\log(1/\delta)})\sqrt{\frac{\mathbb{E}[\|(\Sigma+\lambda I)^{-1/2}\mathbb{X}_{1-}\|^4]-d_{2,\lambda}}{n}} + \frac{4/3(\rho_\lambda^2 d_{1,\lambda}+\sqrt{d_{2,\lambda}})\log(1/\delta)}{n}. \quad (55)$$

We now prove a version of Lemma 4.

Lemma 11.

$$P\Big[\|\widehat{\beta}_{t,\lambda}-\widetilde{\beta}_{t,\lambda}\|_\Sigma^2 \quad (56)$$
$$\leq \sigma^2 tr(M_t) + 2\sigma^2\sqrt{tr(M_t)\|M_t\|\log(1/\delta)} + 2\sigma^2\|M_t\|\log(1/\delta) \mid X\Big] \geq 1-\delta$$

where,

$$M_t = \frac{1}{n^2}\mathbb{X}(\mathbb{S}_t+\lambda I)^{-1}\Sigma(\mathbb{S}_t+\lambda I)^{-1}\mathbb{X}^T. \quad (57)$$

Under Assumptions 1 and 2, with probability 1,

$$tr(M_t) \leq \frac{K_{\lambda,t}^2}{n}\Big(d_{2,\lambda} + \sqrt{d_{2,\lambda}\|(\Sigma+\lambda I)^{-1/2}(\mathbb{S}-\Sigma)(\Sigma+\lambda I)^{-1/2}\|_F^2}\Big) \quad (58)$$

$$\|M_t\| \leq \frac{Ad_{1,\lambda}K_{\lambda,t}^2}{n} \quad (59)$$

where A is a constant such that with probability at least $1-\delta$, $\|\Sigma^{-1/2}\mathbb{X}_{i-}\|^2 \leq A$.

Such an A will exist because $\mathbb{X}_{i-}$ is sub-Gaussian. Then the stochastic bounds on $\mathrm{tr}(M_t)$ and $\|M_t\|$ involve only terms for which we already have stochastic bounds via Lemmas 9 and 10. In particular, with probability $1-3\delta$,

$$\mathrm{tr}(M_t) \leq \frac{C_{t,\lambda}^2 d_{2,\lambda}}{n} \qquad \text{and} \qquad \|M_t\| \leq \frac{Ad_{1,\lambda}C_{t,\lambda}^2}{n}. \quad (60)$$

Then with probability at least $1-4\delta$

$$\|\widehat{\beta}_{t,\lambda}-\widetilde{\beta}_{t,\lambda}\|_\Sigma^2 \quad (61)$$
$$\leq \frac{\sigma^2 C_{t,\lambda}^2 d_{2,\lambda}}{n} + \frac{2\sigma^2 C_{t,\lambda}^2\rho_\lambda\sqrt{d_{2,\lambda}d_{1,\lambda}\log(1/\delta)}}{n} + \frac{2\sigma^2\log(1/\delta)\rho_\lambda^2 d_{1,\lambda}C_{t,\lambda}^2}{n}$$
$$= O_P\left(\frac{\sigma^2 d_{2,\lambda}}{n}\right).$$

Proof. The proof is structured like the proof of Lemma 12 in Hsu et al. (2011). We have that

$$\|\widehat{\beta}_{t,\lambda}-\widetilde{\beta}_{t,\lambda}\|_\Sigma^2 \quad (62)$$
$$= \|(\mathbb{S}_t+\lambda I)^{-1}(\mathbb{S}\beta^* + \frac{1}{n}\mathbb{X}^T\varepsilon) - (\mathbb{S}_t+\lambda I)^{-1}\mathbb{S}\beta^*\|_\Sigma^2$$
$$= \|(\mathbb{S}_t+\lambda I)^{-1}\frac{1}{n}\mathbb{X}^T\varepsilon\|_\Sigma^2 = \|M_t^{1/2}\varepsilon\|^2. \quad (63)$$

Then by Lemma 14 of Hsu et al. (2011), a general lemma on sub-Gaussian quadratic forms, conditional on X, the following bound holds with probability at least $1-\delta$

$$\|\widehat{\beta}_{t,\lambda}-\widetilde{\beta}_{t,\lambda}\|_\Sigma^2 \leq \sigma^2\mathrm{tr}(M_t) + 2\sigma^2\sqrt{\mathrm{tr}(M_t)\|M_t\|\log(1/\delta)} + 2\sigma^2\|M_t\|\log(1/\delta). \quad (64)$$

We can bound $\|M_t\|$ by

$$\|M_t\| = \frac{1}{n^2}\|\Sigma^{1/2}(\mathbb{S}_t+\lambda I)^{-1}\mathbb{X}^T\|^2 \quad (65)$$
$$\leq \frac{1}{n^2}\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}\|^2 \quad (66)$$
$$\cdot \|(\Sigma+\lambda I)^{1/2}(\mathbb{S}_t+\lambda I)^{-1}(\Sigma+\lambda I)^{1/2}\|^2$$
$$\cdot \|(\Sigma+\lambda I)^{-1/2}\Sigma^{1/2}\|^2\|\Sigma^{-1/2}\mathbb{X}^T\|^2$$
$$\leq \frac{K_{\lambda,t}^2}{n^2}\|\Sigma^{-1/2}\mathbb{X}^T\|_F^2 \quad (67)$$
$$= \frac{K_{\lambda,t}^2}{n^2}\sum_{j=1}^n\|\Sigma^{-1/2}\mathbb{X}_{i-}\|^2 \quad (68)$$
$$\leq \frac{Ad_{1,\lambda}K_{\lambda,t}^2}{n}. \quad (69)$$

The last inequality follows from Assumption 2. Bounding $\mathrm{tr}(M_t)$ can be done in the same way as in Hsu et al. (2011), except where they use their Lemma 13, we would use Lemma 9 above. To make the notation simpler, define:

$$\Sigma_w = (\Sigma + \lambda I)^{-1/2}\Sigma(\Sigma + \lambda I)^{-1/2} \quad (70)$$

$$\widehat{\Sigma}_w = (\Sigma + \lambda I)^{-1/2}\mathbb{S}(\Sigma + \lambda I)^{-1/2} \quad (71)$$

$$\widehat{\Sigma}_{t,\lambda,w} = (\Sigma + \lambda I)^{-1/2}(\mathbb{S}_t + \lambda I)(\Sigma + \lambda I)^{-1/2} \quad (72)$$

Then

$$\mathrm{tr}(M_t) = \frac{1}{n^2}\mathrm{tr}(\mathbb{X}(\mathbb{S}_t + \lambda I)^{-1}\Sigma(\mathbb{S}_t + \lambda I)^{-1}\mathbb{X}^T) \quad (73)$$

$$= \frac{1}{n}\mathrm{tr}((\mathbb{S}_t + \lambda I)^{-1}\Sigma(\mathbb{S}_t + \lambda I)^{-1}\mathbb{S}) \quad (74)$$

$$= \frac{1}{n}\mathrm{tr}(\widehat{\Sigma}_{t,\lambda,w}^{-1}\widehat{\Sigma}_w\widehat{\Sigma}_{t,\lambda,w}^{-1}\Sigma_w). \quad (75)$$

Von Neumann's theorem gives that

$$\mathrm{tr}(\widehat{\Sigma}_{t,\lambda,w}^{-1}\widehat{\Sigma}_w\widehat{\Sigma}_{t,\lambda,w}^{-1}\Sigma_w) \quad (76)$$
$$\leq \sum_j \lambda_j(\widehat{\Sigma}_{t,\lambda,w}^{-1}\widehat{\Sigma}_w\widehat{\Sigma}_{t,\lambda,w}^{-1})\lambda_j(\Sigma_w)$$

Under Assumption 1, $\widehat{\Sigma}_{t,\lambda,w}^{-1}$ will be positive definite and so we can use Ostrowski's theorem to say that

$$\lambda_j(\widehat{\Sigma}_{t,\lambda,w}^{-1}\widehat{\Sigma}_w\widehat{\Sigma}_{t,\lambda,w}^{-1}) \leq \lambda_{\max}(\widehat{\Sigma}_{t,\lambda,w}^{-2})\lambda_j(\widehat{\Sigma}_w). \quad (77)$$

Thus

$$\mathrm{tr}(\widehat{\Sigma}_{t,\lambda,w}^{-1}\widehat{\Sigma}_w\widehat{\Sigma}_{t,\lambda,w}^{-1}\Sigma_w)$$
$$\leq \lambda_{\max}(\widehat{\Sigma}_{t,\lambda,w}^{-2})\sum_j \lambda_j(\widehat{\Sigma}_w)\lambda_j(\Sigma_w) \quad (78)$$

$$\leq K_{\lambda,t}^2\sum_j \left(\lambda_j(\Sigma_w)^2 + \lambda_j(\Sigma_w)(\lambda_j(\widehat{\Sigma}_w) - \lambda_j(\Sigma_w))\right) \quad (79)$$

$$\leq K_{\lambda,t}^2\Bigg(\sum_j \lambda_j(\Sigma_w)^2 \quad (80)$$
$$+ \sqrt{\sum_j \lambda_j(\Sigma_w)^2}\sqrt{\sum_j(\lambda_j(\widehat{\Sigma}_w) - \lambda_j(\Sigma_w))^2}\Bigg).$$

For the final step, notice that

$$\sum_j \lambda_j(\Sigma_w)^2 = \sum_j \left(\frac{\lambda_j(\Sigma)}{\lambda_j(\Sigma) + \lambda}\right)^2 = d_{2,\lambda} \quad (81)$$

and by Mirsky's theorem

$$\sum_j(\lambda_j(\widehat{\Sigma}_w) - \lambda_j(\Sigma_w))^2 \leq \|\widehat{\Sigma}_w - \Sigma_w\|_F^2 \quad (82)$$

$$= \|(\Sigma + \lambda I)^{-1/2}(\mathbb{S} - \Sigma)(\Sigma + \lambda I)^{-1/2}\|_F^2. \quad (83)$$

Then we get that

$$\mathrm{tr}(M) \leq \frac{K_{\lambda,t}^2}{n}\Bigg(d_{2,\lambda} \quad (84)$$
$$+ \sqrt{d_{2,\lambda}\|(\Sigma + \lambda I)^{-1/2}(\mathbb{S} - \Sigma)(\Sigma + \lambda I)^{-1/2}\|_F^2}\Bigg).$$

□

7. Conclusion

We have presented a framework for trading off risk for computational efficiency in linear regression. Our analysis shows how the predictive risk degrades as a function of a hard threshold parameter and a regularization parameter. As the sparsity level of the thresholded sample covariance increases, the computation decreases, as analyzed in the recent literature on fast solvers for SDD systems. This establishes a setting where a tuning parameter provides a fine-grained way to tradeoff accuracy for computation.

We have adopted a computational model where the thresholded covariance $\mathbb{S}_t$ is given as input, as a sparse matrix. Of course, straightforward algorithms require $O(np^2)$ computation to compute $\mathbb{S} = \frac{1}{n}\sum_i X_i X_i^T$. But this is easily parallelizable, and the cost of the computation can be amortized across many regressions. Alternatively, provided a sparsity pattern of m entries, the actual matrix $\mathbb{S}_t$ can be computed in time $O(nm)$.

The approach we have introduced here leaves several interesting possibilities for further work. In particular, the computation-risk tradeoff we have studied for ridge regression can be leveraged for other learning problems. For example, the alternating direction method of multipliers (ADMM) procedure has been shown to be an effective algorithm for the optimizations required in many learning problems, including the lasso, elastic net, and Gaussian graph estimation (Boyd et al., 2010). In many ADMM algorithms, the proximal procedure leads to a form of ridge regression in the iterative step. An interesting future direction is to study sparsification of these linear systems to obtain faster solvers, and to analyze the resulting tradeoff in statistical risk.

Acknowledgements

Research supported in part under NSF Grant IIS-1116730, AFOSR grant FA9550-09-1-0373 and ONR grant N000141210762.

References

Agarwal, Alekh, Bartlett, Peter L., and Duchi, John C. Oracle inequalities for computationally adaptive model selection. arxiv: 1208.0129, 2012.

Amini, Arash A. and Wainwright, Martin J. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5):2877–2921, 2009.

Bach, Francis and Moulines, Eric. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Bickel, Peter J. and Levina, Elizaveta. Covariance regularization by thresholding. *The Annals of Statistics*, 2008.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

Chandrasekaran, Venkat and Jordan, Michael. Computational and statistical tradeoffs via convex relaxation. Technical report, University of California, Berkeley, 2012. arXiv:1211.1073.

Clarkson, Kenneth L. and Woodruff, David P. Low rank approximation and regression in input sparsity time. Technical report, IBM Almaden Research Center, 2013. arXiv:1207.6365.

Clarkson, Kenneth L., Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W., Meng, Xiangrui, and Woodruff, David P. The fast cauchy transform and faster robust linear regression. Technical report, IBM Almaden Research Center, 2013. arXiv:1207.6365.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. Sampling algorithms and coresets for *lp* regression. *SIAM J. Computing*, 38:2060–2078, 2009.

d'Aspremont, A., Ghaoui, L. El, Jordan, M. I., and Lanckriet, G. A direct formulation for sparse PCA using semidefinite programming. In *In S. Thrun, L. Saul, and B. Schoelkopf (Eds.), Advances in Neural Information Processing Systems (NIPS) 16, 2004.*, 2004.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for l_2 regression and applications. In *Proc. of the 17-th Annual SODA*, pp. 1127–1136, 2006.

Hsu, Daniel, Kakade, Sham M., and Zhang, Tong. An analysis of random design linear regression. arxiv:1106.2363, 2011.

Johnstone, Iain M. and Lu, Arthur Yu. Sparse principal components analysis. Technical report, Stanford University, 2004. arXiv:0901.4392.

Koutis, Ioannis, Miller, Gary, and Peng, Richard. A nearly-$m \log n$ time solver for sdd linear systems. Technical report, Carnegie Mellon University, 2012. arXiv:cs/1102.4842; FOCS 2011.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. 22(3):400–407, 1951.

Shalev-Shwartz, Shai, Srebro, Nathan, and Zhang, Tong. Trading accuracy for sparsity in optimization problems with sparsity constraints. *Siam Journal on Optimization*, 2010.

Spielman, Daniel A. and Teng, Shang-Hua. Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. Technical report, Yale University, 2009. arXiv:cs/0607105.

Vershynin, Roman. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 2012.