Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation

Dominique Guillot
Dept. of Statistics
Stanford University
Stanford, CA 94305
dguillot@stanford.edu

Bala Rajaratnam
Dept. of Statistics
Stanford University
Stanford, CA 94305
brajarat@stanford.edu

Benjamin T. Rolfs
ICME
Stanford University
Stanford, CA 94305
benrolfs@stanford.edu

Arian Maleki
Dept. of ECE
Rice University
Houston, TX 77005
arian.maleki@rice.edu

Ian Wong
Dept. of EE and Statistics
Stanford University
Stanford, CA 94305
ianw@stanford.edu

Abstract

The ℓ_1-regularized maximum likelihood estimation problem has recently become a topic of great interest within the machine learning, statistics, and optimization communities as a method for producing sparse inverse covariance estimators. In this paper, a proximal gradient method (G-ISTA) for performing ℓ_1-regularized covariance matrix estimation is presented. Although numerous algorithms have been proposed for solving this problem, this simple proximal gradient method is found to have attractive theoretical and numerical properties. G-ISTA has a linear rate of convergence, resulting in an $\mathcal{O}(\log \varepsilon)$ iteration complexity to reach a tolerance of ε. This paper gives eigenvalue bounds for the G-ISTA iterates, providing a closed-form linear convergence rate. The rate is shown to be closely related to the condition number of the optimal point. Numerical convergence results and timing comparisons for the proposed method are presented. G-ISTA is shown to perform very well, especially when the optimal point is well-conditioned.

1 Introduction

Datasets from a wide range of modern research areas are increasingly high dimensional, which presents a number of theoretical and practical challenges. A fundamental example is the problem of estimating the covariance matrix from a dataset of n samples $\{X^{(i)}\}_{i=1}^n$, drawn *i.i.d* from a p-dimensional, zero-mean, Gaussian distribution with covariance matrix $\Sigma \in \mathbb{S}_{++}^p$, $X^{(i)} \sim \mathcal{N}_p(0, \Sigma)$, where $\mathbb{S}_{++}^p$ denotes the space of $p \times p$ symmetric, positive definite matrices. When $n \geq p$ the maximum likelihood covariance estimator $\hat{\Sigma}$ is the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^n X^{(i)}X^{(i)^T}$. A problem however arises when $n < p$, due to the rank-deficiency in S. In this sample deficient case, common throughout several modern applications such as genomics, finance, and earth sciences, the matrix S is not invertible, and thus cannot be directly used to obtain a well-defined estimator for the inverse covariance matrix $\Omega := \Sigma^{-1}$.

A related problem is the inference of a Gaussian graphical model ([27, 14]), that is, a sparsity pattern in the inverse covariance matrix, Ω. Gaussian graphical models provide a powerful means of dimensionality reduction in high-dimensional data. Moreover, such models allow for discovery of conditional independence relations between random variables since, for multivariate Gaussian data, sparsity in the inverse covariance matrix encodes conditional independences. Specifically, if

$X = (X_i)_{i=1}^p \in \mathbb{R}^p$ is distributed as $X \sim \mathcal{N}_p(0, \Sigma)$, then $(\Sigma^{-1})_{ij} = \Omega_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j | \{X_k\}_{k \neq i,j}$, where the notation $A \perp\!\!\!\perp B|C$ denotes the conditional independence of A and B given the set of variables C (see [27, 14]). If a dataset, even one with $n \gg p$ is drawn from a normal distribution with sparse inverse covariance matrix Ω, the inverse sample covariance matrix S^{-1} will almost surely be a dense matrix, although the estimates for those Ω_{ij} which are equal to 0 may be very small in magnitude. As sparse estimates of Ω are more robust than S^{-1}, and since such sparsity may yield easily interpretable models, there exists significant impetus to perform sparse inverse covariance estimation in very high dimensional low sample size settings.

Banerjee et al. [1] proposed performing such sparse inverse covariance estimation by solving the ℓ_1-penalized maximum likelihood estimation problem,

$$\Theta_\rho^* = \arg \min_{\Theta \in \mathbb{S}_{++}^p} -\log\det\Theta + \langle S, \Theta \rangle + \rho \left\| \Theta \right\|_1, \tag{1}$$

where $\rho > 0$ is a penalty parameter, $\langle S, \Theta \rangle = \mathrm{Tr}\,(S\Theta)$, and $\left\|\Theta\right\|_1 = \sum_{i,j} |\Theta_{ij}|$. For $\rho > 0$, Problem (1) is strongly convex and hence has a unique solution, which lies in the positive definite cone $\mathbb{S}_{++}^p$ due to the $\log\det$ term, and is hence invertible. Moreover, the ℓ_1 penalty induces sparsity in Θ_ρ^*, as it is the closest convex relaxation of the $0-1$ penalty, $\left\|\Theta\right\|_0 = \sum_{i,j} \mathbb{I}(\Theta_{ij} \neq 0)$, where $\mathbb{I}(\cdot)$ is the indicator function [5]. The unique optimal point of problem (1), Θ_ρ^*, is both invertible (for $\rho > 0$) and sparse (for sufficiently large ρ), and can be used as an inverse covariance matrix estimator.

In this paper, a proximal gradient method for solving Problem (1) is proposed. The resulting "graphical iterative shrinkage thresholding algorithm", or `G-ISTA`, is shown to converge at a linear rate to Θ_ρ^*, that is, its iterates Θ_t are proven to satisfy

$$\left\|\Theta_{t+1} - \Theta_\rho^*\right\|_F \leq s \left\|\Theta_t - \Theta_\rho^*\right\|_F, \tag{2}$$

for a fixed worst-case contraction constant $s \in (0, 1)$, where $\left\|\cdot\right\|_F$ denotes the Frobenius norm. The convergence rate s is provided explicitly in terms of S and ρ, and importantly, is related to the condition number of Θ_ρ^*.

The paper is organized as follows. Section 2 describes prior work related to solution of Problem (1). The `G-ISTA` algorithm is formulated in Section 3. Section 4 contains the convergence proofs of this algorithm, which constitutes the primary mathematical result of this paper. Numerical results are presented in Section 5, and concluding remarks are made in Section 6.

2 Prior Work

While several excellent general convex solvers exist (for example, [11] and [4]), these are not always adept at handling high dimensional problems (i.e., $p > 1000$). As many modern datasets have several thousands of variables, numerous authors have proposed efficient algorithms designed specifically to solve the ℓ_1-penalized sparse maximum likelihood covariance estimation problem (1).

These can be broadly categorized as either primal or dual methods. Following the literature, we refer to primal methods as those which directly solve Problem (1), yielding a concentration estimate. Dual methods [1] yield a covariance matrix by solving the constrained problem,

$$\begin{array}{ll} \underset{U \in \mathbb{R}^{p \times p}}{\text{minimize}} & -\log\det(S + U) - p \\ \text{subject to} & \left\|U\right\|_\infty \leq \rho, \end{array} \tag{3}$$

where the primal and dual variables are related by $\Theta = (S + U)^{-1}$. Both the primal and dual problems can be solved using block methods (also known as "row by row" methods), which sequentially optimize one row/column of the argument at each step until convergence. The primal and dual block problems both reduce to ℓ_1-penalized regressions, which can be solved very efficiently.

2.1 Dual Methods

A number of dual methods for solving Problem (1) have been proposed in the literature. Banerjee et al. [1] consider a block coordinate descent algorithm to solve the block dual problem, which reduces each optimization step to solving a box-constrained quadratic program. Each of these quadratic programs is equivalent to performing a "lasso" (ℓ_1-regularized) regression. Friedman et al. [10] iteratively solve the lasso regression as described in [1], but do so using coordinate-wise descent. Their widely used solver, known as the graphical lasso (`glasso`) is implemented on `CRAN`. Global convergence rates of these block coordinate methods are unknown. D'Aspremont et al. [9] use Nesterov's smooth approximation scheme, which produces an ε-optimal solution in $\mathcal{O}(1/\varepsilon)$ iterations. A variant of Nesterov's smooth method is shown to have a $\mathcal{O}(1/\sqrt{\varepsilon})$ iteration complexity in [15, 16].

2.2 Primal Methods

Interest in primal methods for solving Problem (1) has been growing for many reasons. One important reason stems from the fact that convergence within a certain tolerance for the dual problem does not necessarily imply convergence within the same tolerance for the primal.

Yuan and Lin [30] use interior point methods based on the max-det problem studied in [26]. Yuan [31] use an alternating-direction method, while Scheinberg et al. [24] proposes a similar method and show a sublinear convergence rate. Mazumder and Hastie [18] consider block-coordinate descent approaches for the primal problem, similar to the dual approach taken in [10]. Mazumder and Agarwal [17] also solve the primal problem with block-coordinate descent, but at each iteration perform a partial as opposed to complete block optimization, resulting in a decreased computational complexity per iteration. Convergence rates of these primal methods have not been considered in the literature and hence theoretical guarantees are not available. Hsieh et al. [13] propose a second-order proximal point algorithm, called `QUIC`, which converges superlinearly locally around the optimum.

3 Methodology

In this section, the *graphical iterative shrinkage thresholding algorithm* (`G-ISTA`) for solving the primal problem (1) is presented. A rich body of mathematical and numerical work exists for general iterative shrinkage thresholding and related methods; see, in particular, [3, 8, 19, 20, 21, 25]. A brief description is provided here.

3.1 General Iterative Shrinkage Thresholding (ISTA)

Iterative shrinkage thresholding algorithms (ISTA) are general first-order techniques for solving problems of the form

$$\underset{x\in\mathcal{X}}{\text{minimize}}\; F(x) := f(x) + g(x), \tag{4}$$

where $\mathcal{X}$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|$, $f : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable, convex function, and $g : \mathcal{X} \to \mathbb{R}$ is a lower semi-continuous, convex function, not necessarily smooth. The function f is also often assumed to have Lipschitz-continuous gradient ∇f, that is, there exists some constant $L > 0$ such that

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq L\,\|x_1 - x_2\| \tag{5}$$

for any $x_1, x_2 \in \mathcal{X}$.

For a given lower semi-continuous convex function g, the proximity operator of g, denoted by $\operatorname{prox}_g : \mathcal{X} \to \mathcal{X}$, is given by

$$\operatorname{prox}_g(x) = \arg\min_{y\in\mathcal{X}} \left\{ g(y) + \frac{1}{2}\|x-y\|^2 \right\}, \tag{6}$$

It is well known (for example, [8]) that $x^* \in \mathcal{X}$ is an optimal solution of problem (4) if and only if

$$x^* = \operatorname{prox}_{\zeta g}(x^* - \zeta\nabla f(x^*)) \tag{7}$$

for any $\zeta > 0$. The above characterization suggests a method for optimizing problem (4) based on the iteration

$$x_{t+1} = \mathrm{prox}_{\zeta_t g}\left(x_t - \zeta_t \nabla f(x_t)\right) \tag{8}$$

for some choice of step size, ζ_t. This simple method is referred to as an iterative shrinkage thresholding algorithm (ISTA). For a step size $\zeta_t \leq \frac{1}{L}$, the ISTA iterates x_t are known to satisfy

$$F(x_t) - F(x^*) \simeq \mathcal{O}\left(\frac{1}{t}\right), \forall t, \tag{9}$$

where x^* is some optimal point, which is to say, they converge to the space of optimal points at a sublinear rate. If no Lipschitz constant L for ∇f is known, the same convergence result still holds for ζ_t chosen such that

$$f(x_{t+1}) \leq Q_{\zeta_t}(x_{t+1}, x_t), \tag{10}$$

where $Q_\zeta(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a quadratic approximation to f, defined by

$$Q_\zeta(x, y) = f(y) + \langle x - y, \nabla f(y)\rangle + \frac{1}{2\zeta}\left\|x - y\right\|^2. \tag{11}$$

See [3] for more details.

3.2 Graphical Iterative Shrinkage Thresholding (G-ISTA)

The general method described in Section 3.1 can be adapted to the sparse inverse covariance estimation Problem (1). Using the notation introduced in Problem (4), define $f, g : \mathbb{S}^p_{++} \to \mathbb{R}$ by $f(X) = -\log\det(X) + \langle S, X\rangle$ and $g(X) = \rho\left\|X\right\|_1$. Both are continuous convex functions defined on $\mathbb{S}^p_{++}$. Although the function $\nabla f(X) = S - X^{-1}$ is not Lipschitz continuous over $\mathbb{S}^p_{++}$, it is Lipschitz continuous within any compact subset of $\mathbb{S}^p_{++}$ (See Lemma 2 of the Supplemental section).

Lemma 1 ([1, 15])**.** The solution of Problem (1), Θ^*_ρ, satisfies $\alpha I \preceq \Theta^*_\rho \preceq \beta I$, for

$$\alpha = \frac{1}{\left\|S\right\|_2 + p\rho}, \quad \beta = \min\left\{\frac{p - \alpha\,\mathrm{Tr}(S)}{\rho}, \gamma\right\}, \tag{12}$$

and

$$\gamma = \begin{cases} \min\{\mathbf{1}^T\left|S^{-1}\right|\mathbf{1}, (p - \rho\sqrt{p}\alpha)\left\|S^{-1}\right\|_2 - (p-1)\alpha\} & \text{if } S \in \mathbb{S}^p_{++} \\ 2\mathbf{1}^T\left|(S + \frac{\rho}{2}I)^{-1}\right|\mathbf{1} - \mathrm{Tr}((S + \frac{\rho}{2}I)^{-1}) & \text{otherwise,} \end{cases} \tag{13}$$

where I denotes the $p \times p$ dimensional identity matrix and $\mathbf{1}$ denotes the p-dimensional vector of ones.

Note that $f + g$ as defined is a continuous, strongly convex function on $\mathbb{S}^p_{++}$. Moreover, by Lemma 2 of the supplemental section, f has a Lipschitz continuous gradient when restricted to the compact domain $aI \preceq \Theta \preceq bI$. Hence, f and g as defined meet the conditions described in Section 3.1.

The proximity operator of $\rho\left\|X\right\|_1$ for $\rho > 0$ is the soft-thresholding operator, $\eta_\rho : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$, defined entrywise by

$$\left[\eta_\rho(X)\right]_{i,j} = \mathrm{sgn}(X_{i,j})\left(|X_{i,j}| - \rho\right)_+, \tag{14}$$

where for some $x \in \mathbb{R}$, $(x)_+ := \max(x, 0)$ (see [8]). Finally, the quadratic approximation Q_{ζ_t} of f, as in equation (11), is given by

$$Q_{\zeta_t}(\Theta_{t+1}, \Theta_t) = -\log\det(\Theta_t) + \langle S, \Theta_t\rangle + \langle\Theta_{t+1} - \Theta_t, S - \Theta_t^{-1}\rangle + \frac{1}{2\zeta_t}\left\|\Theta_{t+1} - \Theta_t\right\|_F^2. \tag{15}$$

The G-ISTA algorithm for solving Problem (1) is given in Algorithm 1. As in [3], the algorithm uses a backtracking line search for the choice of step size. The procedure terminates when a prespecified duality gap is attained. The authors found that an initial estimate of Θ_0 satisfying $[\Theta_0]_{ii} =$

$(S_{ii} + \rho)^{-1}$ works well in practice. Note also that the positive definite check of Θ_{t+1} during Step (1) of Algorithm 1 is accomplished using a Cholesky decomposition, and the inverse of Θ_{t+1} is computed using that Cholesky factor.

Algorithm 1: G-ISTA for Problem (1)

```
input : Sample covariance matrix $S$, penalty parameter $\rho$, tolerance $\varepsilon$, backtracking constant
        $c \in (0, 1)$, initial step size $\zeta_{1,0}$, initial iterate $\Theta_0$. Set $\Delta := 2\varepsilon$.
while $\Delta > \varepsilon$ do
    (1) Line search: Let $\zeta_t$ be the largest element of $\{c^j \zeta_{t,0}\}_{j=0,1,\dots}$ so that for
    $\Theta_{t+1} = \eta_{\zeta_t \rho}\left(\Theta_t - \zeta_t(S - \Theta_t^{-1})\right)$, the following are satisfied:
        $\Theta_{t+1} \succ 0 \quad \text{and} \quad f(\Theta_{t+1}) \leq Q_{\zeta_t}(\Theta_{t+1}, \Theta_t),$
    for $Q_{\zeta_t}$ as defined in (15).
    (2) Update iterate: $\Theta_{t+1} = \eta_{\zeta_t \rho}\left(\Theta_t - \zeta_t(S - \Theta_t^{-1})\right)$
    (3) Set next initial step, $\zeta_{t+1,0}$. See Section 3.2.1.
    (4) Compute duality gap:
        $\Delta = -\log\det(S + U_{t+1}) - p - \log\det\Theta_{t+1} + \langle S, \Theta \rangle + \rho \left\|\Theta_{t+1}\right\|_1,$
    where $(U_{t+1})_{i,j} = \min\{\max\{([\Theta_{t+1}^{-1}]_{i,j} - S_{i,j}), -\rho\}, \rho\}$.
end
output: $\varepsilon$-optimal solution to problem (1), $\Theta_\rho^* = \Theta_{t+1}$.
```

3.2.1 Choice of initial step size, ζ_0

Each iteration of Algorithm 1 requires an initial step size, ζ_0. The results of Section 4 guarantee that any $\zeta_0 \leq \lambda_{\min}(\Theta_t)^2$ will be accepted by the line search criteria of Step 1 in the next iteration. However, in practice this choice of step is overly cautious; a much larger step can often be taken. Our implementation of Algorithm 1 chooses the Barzilai-Borwein step [2]. This step, given by

$$\zeta_{t+1,0} = \frac{\operatorname{Tr}\left((\Theta_{t+1} - \Theta_t)(\Theta_{t+1} - \Theta_t)\right)}{\operatorname{Tr}\left((\Theta_{t+1} - \Theta_t)(\Theta_t^{-1} - \Theta_{t+1}^{-1})\right)}, \tag{16}$$

is also used in the SpaRSA algorithm [29], and approximates the Hessian around Θ_{t+1}. If a certain number of maximum backtracks do not result in an accepted step, G-ISTA takes the safe step, $\lambda_{\min}(\Theta_t)^2$. Such a safe step can be obtained from $\lambda_{\max}(\Theta_t^{-1})$, which in turn can be quickly approximated using power iteration.

4 Convergence Analysis

In this section, linear convergence of Algorithm 1 is discussed. Throughout the section, Θ_t $(t = 1, 2, \dots)$ denote the iterates of Algorithm 1, and Θ_ρ^* the optimal solution to Problem (1) for $\rho > 0$. The minimum and maximum eigenvalues of a symmetric matrix A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively.

Theorem 1. *Assume that the iterates Θ_t of Algorithm 1 satisfy $aI \preceq \Theta_t \preceq bI, \forall t$ for some fixed constants $0 < a < b$. If $\zeta_t \leq a^2, \forall t$, then*

$$\left\|\Theta_{t+1} - \Theta_\rho^*\right\|_F \leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \left\|\Theta_t - \Theta_\rho^*\right\|_F. \tag{17}$$

Furthermore,

1. *The step size ζ_t which yields an optimal worst-case contraction bound $s(\zeta_t)$ is $\zeta = \frac{2}{a^{-2}+b^{-2}}$.*
2. *The optimal worst-case contraction bound corresponding to $\zeta = \frac{2}{a^{-2}+b^{-2}}$ is given by*

$$s(\zeta) := 1 - \frac{2}{1 + \frac{b^2}{a^2}}$$

Proof. A direct proof is given in the appendix. Note that linear convergence of proximal gradient methods for strongly convex objective functions in general has already been proven (see Supplemental section). □

It remains to show that there exist constants a and b which bound the eigenvalues of $\Theta_t, \forall t$. The existence of such constants follows directly from Theorem 1, as Θ_t lie in the bounded domain $\{\Theta \in \mathbb{S}^p_{++} \ : \ f(\Theta) + g(\Theta) < f(\Theta_0) + g(\Theta_0)\}$, for all t. However, it is possible to specify the constants a and b to yield an explicit rate; this is done in Theorem 2.

Theorem 2. *Let $\rho > 0$, define α and β as in Lemma 1, and assume $\zeta_t \leq \alpha^2, \forall t$. Then the iterates Θ_t of Algorithm 1 satisfy $\alpha I \preceq \Theta_t \preceq b' I, \forall t$, with $b' = \left\|\Theta^*_\rho\right\|_2 + \left\|\Theta_0 - \Theta^*_\rho\right\|_F \leq \beta + \sqrt{p}(\beta + \alpha)$.*

Proof. See the Supplementary section. □

Importantly, note that the bounds of Theorem 2 depend explicitly on the bound of Θ^*_ρ, as given by Lemma 1. These eigenvalue bounds on Θ_{t+1}, along with Theorem 1, provide a closed form linear convergence rate for Algorithm 1. This rate depends only on properties of the solution.

Theorem 3. *Let α and β be as in Lemma 1. Then for a constant step size $\zeta_t := \zeta < \alpha^2$, the iterates of Algorithm 1 converge linearly with a rate of*

$$s(\zeta) = 1 - \frac{2\alpha^2}{\alpha^2 + (\beta + \sqrt{p}(\beta - \alpha))^2} < 1 \tag{18}$$

Proof. By Theorem 2, for $\zeta < \alpha^2$, the iterates Θ_t satisfy

$$\alpha I \preceq \Theta_t \preceq \left(\left\|\Theta^*_\rho\right\|_2 + \left\|\Theta_0 - \Theta^*_\rho\right\|_F\right) I$$

for all t. Moreover, since $\alpha I \preceq \Theta^* \preceq \beta I$, if $\alpha I \preceq \Theta_0 \preceq \beta I$ (for instance, by taking $\Theta_0 = (S + \rho I)^{-1}$ or some multiple of the identity) then this can be bounded as:

$$\begin{aligned} \left\|\Theta^*_\rho\right\|_2 + \left\|\Theta_0 - \Theta^*_\rho\right\|_F &\leq \beta + \sqrt{p}\left\|\Theta_0 - \Theta^*_\rho\right\|_2 & (19) \\ &\leq \beta + \sqrt{p}(\beta - \alpha). & (20) \end{aligned}$$

Therefore,

$$\alpha I \preceq \Theta_t \preceq (\beta + \sqrt{p}(\beta - \alpha))\, I, \tag{21}$$

and the result follows from Theorem 1. □

Remark 1. Note that the contraction constant (equation 18) of Theorem 3 is closely related to the condition number of Θ^*_ρ,

$$\kappa(\Theta^*_\rho) = \frac{\lambda_{\max}(\Theta^*_\rho)}{\lambda_{\min}(\Theta^*_\rho)} \leq \frac{\beta}{\alpha}$$

as

$$1 - \frac{2\alpha^2}{\alpha^2 + (\beta + \sqrt{p}(\beta - \alpha))^2} \geq 1 - \frac{2\alpha^2}{\alpha^2 + \beta^2} \geq 1 - 2\kappa(\Theta^*_\rho)^{-2}. \tag{22}$$

Therefore, the worst case bound becomes close to 1 as the conditioning number of Θ^*_ρ increases.

5 Numerical Results

In this section, we provide numerical results for the G-ISTA algorithm. In Section 5.2, the theoretical results of Section 4 are demonstrated. Section 5.3 compares running times of the G-ISTA, glasso [10], and QUIC [13] algorithms. All algorithms were implemented in C++, and run on an Intel $i7 - 2600k$ 3.40GHz × 8 core with 16 GB of RAM.

5.1 Synthetic Datasets

Synthetic data for this section was generated following the method used by [16, 17]. For a fixed p, a p dimensional inverse covariance matrix Ω was generated with off-diagonal entries drawn *i.i.d* from a uniform$(-1, 1)$ distribution. These entries were set to zero with some fixed probability (in this case, either 0.97 or 0.85 to simulate a very sparse and a somewhat sparse model). Finally, a multiple of the identity was added to the resulting matrix so that the smallest eigenvalue was equal to 1. In this way, Ω was insured to be sparse, positive definite, and well-conditioned. Datsets of n samples were then generated by drawing *i.i.d.* samples from a $\mathcal{N}_p(0, \Omega^{-1})$ distribution. For each value of p and sparsity level of Ω, $n = 1.2p$ and $n = 0.2p$ were tested, to represent both the $n < p$ and $n > p$ cases.

	ρ	0.03	0.06	0.09	0.12
problem	**algorithm**	**time/iter**	**time/iter**	**time/iter**	**time/iter**
$p = 2000$ $n = 400$ nnz$(\Omega) = 3\%$	nnz(Ω^*_ρ)/$\kappa(\Omega^*_\rho)$	27.65%/48.14	15.08%/20.14	7.24%/7.25	2.39%/2.32
	glasso	1977.92/11	831.69/8	604.42/7	401.59/5
	QUIC	1481.80/21	257.97/11	68.49/8	15.25/6
	G-ISTA	**145.60**/437	**27.05**/9	**8.05**/27	**3.19**/12
$p = 2000$ $n = 2400$ nnz$(\Omega) = 3\%$	nnz(Ω^*_ρ)/$\kappa(\Omega^*_\rho)$	14.56%/10.25	3.11%/2.82	0.91%/1.51	0.11%/1.18
	glasso	667.29/7	490.90/6	318.24/4	233.94/3
	QUIC	211.29/10	24.98/7	5.16/5	**1.56**/4
	G-ISTA	**14.09**/47	**3.51**/13	**2.72**/10	2.20/8
$p = 2000$ $n = 400$ nnz$(\Omega) = 15\%$	nnz(Ω^*_ρ)/$\kappa(\Omega^*_\rho)$	27.35%/64.22	15.20%/28.50	7.87%/11.88	2.94%/2.87
	glasso	2163.33/11	862.39/8	616.81/7	48.47/7
	QUIC	1496.98/21	318.57/12	96.25/9	23.62/7
	G-ISTA	**251.51**/714	**47.35**/148	**7.96**/28	**3.18**/12
$p = 2000$ $n = 2400$ nnz$(\Omega) = 15\%$	nnz(Ω^*_ρ)/$\kappa(\Omega^*_\rho)$	19.98%/17.72	5.49%/4.03	65.47%/1.36	0.03%/1.09
	glasso	708.15/6	507.04/6	313.88/4	233.16/3
	QUIC	301.35/10	491.54/17	4.12/5	1.34/4
	G-ISTA	**28.23**/88	**4.08**/16	**1.95**/7	**1.13**/4

Table 1: Timing comparisons for $p = 2000$ dimensional datasets, generated as in Section 5.1. Above, nnz(A) is the percentage of nonzero elements of matrix A.

5.2 Demonstration of Convergence Rates

The linear convergence rate derived for G-ISTA in Section 4 was shown to be heavily dependent on the conditioning of the final estimator. To demonstrate these results, G-ISTA was run on a synthetic dataset, as described in Section 5.1, with $p = 500$ and $n = 300$. Regularization parameters of $\rho = 0.75$, 0.1, 0.125, 0.15, and 0.175 were used. Note as ρ increases, Θ^*_ρ generally becomes better conditioned. For each value of ρ, the numerical optimum was computed to a duality gap of 10^{-10} using G-ISTA. These values of ρ resulted in sparsity levels of 81.80%, 89.67%, 94.97%, 97.82%, and 99.11%, respectively. G-ISTA was then run again, and the Frobenius norm argument errors at each iteration were stored. These errors were plotted on a log scale for each value of ρ to demonstrate the dependence of the convergence rate on condition number. See Figure 1, which clearly demonstrates the effects of conditioning.

5.3 Timing Comparisons

The G-ISTA, glasso, and QUIC algorithms were run on synthetic datasets (real datasets are presented in the Supplemental section) of varying p, n and with different levels of regularization, ρ. All algorithms were run to ensure a fixed duality gap, here taken to be 10^{-5}. This comparison used efficient C++ implementations of each of the three algorithms investigated. The implementation of G-ISTA was adapted from the publicly available C++ implementation of QUIC Hsieh et al. [13]. Running times were recorded and are presented in Table 1. Further comparisons are presented in the Supplementary section.

Remark 2. The three algorithms variable ability to take advantage of multiple processors is an important detail. The times presented in Table 1 are wall times, not CPU times. The comparisons were run on a multicore processor, and it is important to note that the Cholesky decompositions and

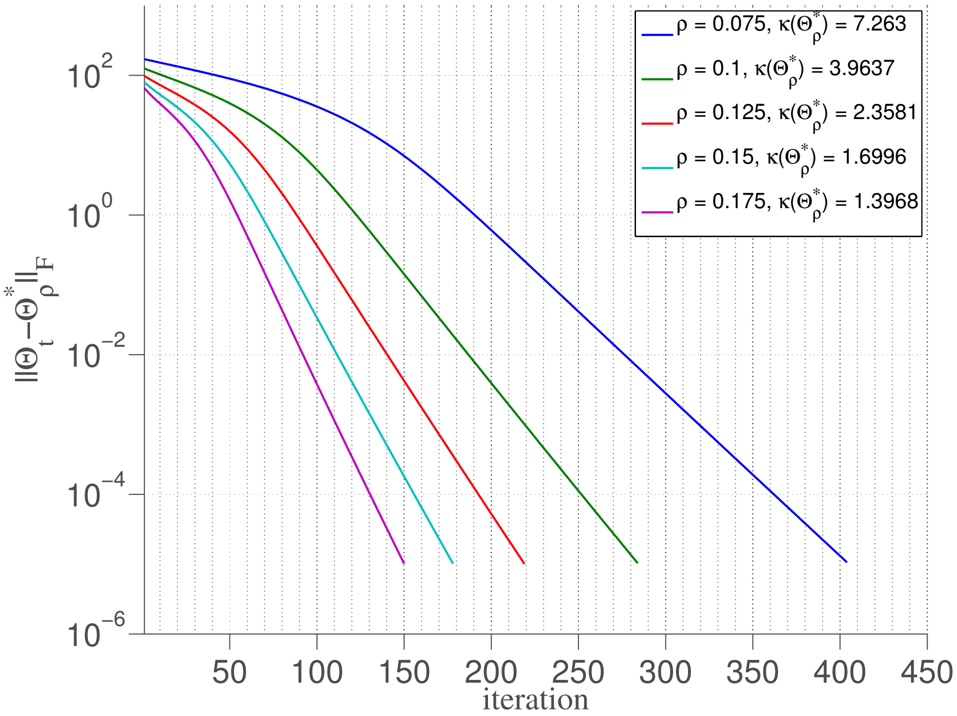

Figure 1: Semilog plot of $\left\|\Theta_t - \Theta^*_\rho\right\|_F$ vs. iteration number t, demonstrating linear convergence rates of G-ISTA, and dependence of those rates on $\kappa(\Theta^*_\rho)$.

inversions required by both G-ISTA and QUIC take advantage of multiple cores. On the other hand, the p^2 dimensional lasso solve of QUIC and p-dimensional lasso solve of glasso do not. For this reason, and because Cholesky factorizations and inversions make up the bulk of the computation required by G-ISTA, the CPU time of G-ISTA was typically greater than its wall time by a factor of roughly 4. The CPU and wall times of QUIC were more similar; the same applies to glasso.

6 Conclusion

In this paper, a proximal gradient method was applied to the sparse inverse covariance problem. Linear convergence was discussed, with a fixed closed-form rate. Numerical results have also been presented, comparing G-ISTA to the widely-used glasso algorithm and the newer, but very fast, QUIC algorithm. These results indicate that G-ISTA is competitive, in particular for values of ρ which yield sparse, well-conditioned estimators. The G-ISTA algorithm was very fast on the synthetic examples of Section 5.3, which were generated from well-conditioned models. For poorly conditioned models, QUIC is very competitive. The Supplemental section gives two real datasets which demonstrate this. For many practical applications however, obtaining an estimator that is well-conditioned is important ([23, 28]). To conclude, although second-order methods for the sparse inverse covariance method have recently been shown to perform well, simple first-order methods cannot be ruled out, as they can also be very competitive in many cases.

References

[1] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivarate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[2] Jonathan Barzilai and Jonathan M. Borwein. Two-Point Step Size Gradient Methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009. ISSN 1936-4954.

[4] S. Becker, E.J. Candes, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3:165–218, 2010.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] P. Brohan, J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, 111, 2006.

[7] George H.G. Chen and R.T. Rockafellar. Convergence rates in forward-backward splitting. *Siam Journal on Optimization*, 7:421–444, 1997.

[8] Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[9] Alexandre D'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, April 2011.

[12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[13] Cho-Jui Hsieh, Matyas A. Sustik, Inderjit S. Dhillon, and Pradeep K. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems 24*, pages 2330–2338. 2011.

[14] S.L. Lauritzen. *Graphical models*. Oxford Science Publications. Clarendon Press, 1996.

[15] Zhaosong Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009. ISSN 1052-6234. doi: http://dx.doi.org/10.1137/070695915.

[16] Zhaosong Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31:2000–2016, 2010.

[17] Rahul Mazumder and Deepak K. Agarwal. A flexible, scalable and efficient algorithmic framework for the *Primal* graphical lasso. *Pre-print*, 2011.

[18] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Pre-print*, 2011.

[19] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[20] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[21] Yurii Nesterov. Gradient methods for minimizing composite objective function. CORE discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

[22] Jennifer Pittman, Erich Huang, Holly Dressman, Cheng-Fang F. Horng, Skye H. Cheng, Mei-Hua H. Tsou, Chii-Ming M. Chen, Andrea Bild, Edwin S. Iversen, Andrew T. Huang, Joseph R. Nevins, and Mike West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431–8436, 2004.

[23] Benjamin T. Rolfs and Bala Rajaratnam. A note on the lack of symmetry in the graphical lasso. *Computational Statistics and Data Analysis*, 2012.

[24] Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems 23*, pages 2101–2109. 2010.

[25] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[26] Lieven Vandenberghe, Stephen Boyd, and Shao-Po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1996.

[27] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

[28] J. Won, J. Lim, S. Kim, and B. Rajaratnam. Condition number regularized covariance estimation. *Journal of the Royal Statistical Society Series B*, 2012.

[29] Stephen J. Wright, Robert D. Nowak, and Mário A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[30] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94 (1):19–35, 2007.

[31] X.M. Yuan. Alternating direction method of multipliers for covariance selection models. *Journal of Scientific Computing*, pages 1–13, 2010.