SPARSE LINEAR REGRESSION AND HYPERBOLIC POLYNOMIALS

KEVIN SHU

Abstract. We consider the sparse linear regression problem, also known as subset selection, where we seek to find a small subset of the features in some data set that best explains some dependent variable of interest. We develop a polynomial time heuristic for sparse linear regression which is motivated by the recent success of hyperbolic polynomials in many aspects of modern computer science. We give both a probabilistic formulation for this heuristic, and also show that it gives the optimum value of a convex conical program which forms a relaxation of the sparse linear regression problem. We also explore how well this heuristic performs in a number of cases. For instance, in the well studied case of the compressed sensing problem, we show that our heuristic will recover a sparse vector from an underdetermined system of equations even in the presence of noise. We will also give a result that compares the optimal sparse linear regressor with PCA based methods for regression. We will also give numerical examples comparing our methods to existing methods such as LASSO.

1. Introduction

The sparse linear regression problem is a variant of the classical linear regression problem where the number of nonzero coefficients in the final regression equation is required to be small. This problem and some of its variants is also known as the subset selection problem, [**?**, 14, 22] as it can be thougbt of as the problem of selecting k features from the data set which provide the best linear regression model for the dataset. This problem is known to be NP-hard, even to approximate within polynomial factors [21] This problem also has connections to compressed sensing, in which we are given an underdetermined system of linear equations, but promised that the resulting solution will be sparse [13]. There is an extensive literature giving heuristic and exact methods for solving certain cases of this problem [8].

We will describe an interesting heuristic to solve this problem that arises from the study of hyperbolic polynomials. Hyperbolic polynomials have been of interest in recent years because their applications in both theoretical mathematics and computer science [2, 19, 12, 20].

We now define the sparse linear regression problem formally. For $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, define the least squares error when regressing b against A by

$$\ell(A, b) = \min\{\|Ax - b\|^2 : x \in \mathbb{R}^n\}.$$

We then define the sparse least squares loss function by

$$\ell_k(A, b) = \min\{\ell(A_S, b) : |S| \leq k\} \tag{1.1}$$

Here, for $S \subseteq [n]$, A_S denotes submatrix of A obtained by removing the columns of A not in S.

We can equivalently think of $\ell_k(A, b)$ as the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \|Ax - b\|^2 \\ \text{such that} \quad & \|x\|_0 \leq k \end{aligned}$$

Here, $\|x\|_0$ denotes the number of nonzero entries of x.

For the purpose of exposition, we will present a simplified probabilistic formulation for this heuristic in the introduction, though we will give a number of generalizations and equivalent formulations in the body of the paper.

Given a subset $S \subseteq [n]$, we will let A_S denote the restriction of A to the columns in S, and we will let $A^\intercal A|_S$ denote the principal submatrix of $A^\intercal A$ whose rows and columns are both contained in S. We define the *truncated determinantal point process* (tDPP) [1] associated with A to be the probability distribution on k element subsets of $[n]$, such that for $S \in \binom{[n]}{k}$, $\Pr(S) \propto \det(A^\intercal A|_S)$. For $T \subseteq [n]$ with $|T| \leq k$, we define

$$\eta^n_{T,k}(A, b) = \mathbb{E}[\ell(A_S, b) | T \subseteq S].$$

We can interpret this as being the expected least squares error incurred when we regress using only the columns of A chosen from the tDPP.

tDPPs are common in applications related to feature selection, as they can be thought of as rewarding ‘diversity’ in the columns that are selected by the process[6]. A related idea for using determinantal point processes for sparse linear regression was considered in [18], though they did not restrict to sets of a fixed size, and they only considered problems with a small number of features.

It is clear that for any $T \subseteq [n]$ with $|T| \leq k$, $\eta^n_{T,k}(A,b) \geq \ell_k(A,b)$, since by the probabilistic method, there must exist some $S \subseteq [n]$ where $\ell(A_S,b) \leq \eta^n_{T,k}(A,b)$. A standard reduction allows us to reduce the problem of deterministically finding some S so that $\ell(A|_S,b) \leq \eta^n_{\varnothing,k}(A,b)$ to computing the value of $\eta^n_{T,k}(A,b)$ for given $T \subseteq [n]$.

Our first insight in this paper is that we can deterministically compute $\eta^n_{T,k}$, as well as a number of generalizations of this quantity, efficiently using algebraic ideas.

We will also show quantity $\eta^n_{T,k}(A,b)$ also arises naturally as the optimal value for a relaxation of the sparse linear regression problem. Specifically, we will show that

$$\eta^n_{T,k}(A,b) = \begin{array}{cc} \text{minimize} & \|b\|^2 - y \\ \text{such that} & A^\intercal A y - A^\intercal b b^\intercal A \in H^n_{T,k}, \end{array}$$

where $H^n_{T,k}$ is a convex cone, which is specifically the hyperbolicity cone of the polynomial

$$c^n_{T,k}(X) = \sum_{S \in \binom{[n]}{k} : T \subseteq S} \det(X|_S).$$

We will give the definition of hyperbolicity cones and hyperbolic polynomials in Section 3, but for now, it suffices to note that interior point methods can optimize over these cones efficiently [16].

Having established that this quantity arises naturally in a number of settings, we will also examine the quality of this heuristic in a number of settings.

We show that $\eta^n_{\varnothing,k}(A,b)$ is basis invariant in the sense that if U is an orthgonal transformation, then $\eta^n_{\varnothing,k}(A,b) = \eta^n_{\varnothing,k}(UA,Ub)$. Using this fact, we are able to compare the result of sparse linear regression to regression according to the top k singular vectors of A when A is sufficiently close to being low rank. This seems to be deeply related to a result shown in [6].

We are also able to show a result which was previously shown for ℓ_1 regression methods that if A satisfies the restricted isometry property (RIP), then our method can recover a k-sparse vector f from the value of $Af + \epsilon$, where ϵ is a sufficiently small random noise.

In practice, we are able to compare our method to regression with ℓ_1 regression methods in cases when n is in the thousands and k is small.

The structure of this paper is as follows: we begin with some preliminaries on the types of polynomials that we will use to solve this problem, and how they relate to the sparse linear regression problem and hyperbolic optimization. We then give some precise statements of the results we show in this paper. The last half of this paper is devoted to proofs of these results.

2. Effeciently Computing $\eta_{T,k}$.

We will be making use of a number of theorems from linear algebra, such as the matrix-determinant lemma for rank 1 updates to the determinant, and the Schur complement lemma. All of these results can be found in any standard reference on linear algebra, and we recommend [17].

2.1. Linear Principal Minor Polynomials.

In this section, we will define a family of polynomials whose properties we will be considering in detail throughout this paper.

Let $\vec{a} = (a_S : S \in \binom{[n]}{k}) \in \mathbb{R}^{\binom{n}{k}}$ be a vector of coefficients, and let X be a symmetric matrix of indeterminants.

We will associate to $\vec{a}$ the *linear principal minor* (lpm) polynomial $p_{\vec{a}}(X)$,

$$p_{\vec{a}}(X) = \sum_{S \in \binom{[n]}{k}} a_S \det(X|_S),$$

where $X|_S$ denotes the principal submatrix of X indexed by S.

One key family of examples of such polynomials are the *characteristic coefficients*, which are defined as

$$c^n_k(X) = p_{\vec{1}}(X) = \sum_{S \in \binom{[n]}{k}} \det(X|_S).$$

Here, $\vec{1}$ denotes the all 1's vector.

These arise naturally when considering the eigenvalues of a matrix, as we have the following formula for the characteristic polynomial of an $n \times n$ matrix:

$$\det(X + tI) = \sum_{k=0}^{n} c_{n-k}^{n}(X)t^k.$$

In particular, we see that $c_k^n(X)$ invariant under change of basis, i.e for any orthogonal matrix U, $c_k^n(X) = c_k^n(U^\intercal X U)$.

One operation that we will need to perform on these polynomials is what we call *restriction*, which we will use in the next subsection to compute conditional expectations for weighted determinantal point processes.

Given a lpm polynomial $p_{\vec{a}}$ of degree k, we define the restriction of $p_{\vec{a}}$ to a set $T \subseteq [n]$ with $|T| \leq k$ by

$$p_{\vec{a},T}(X) = \sum_{S \in \binom{[n]}{k} : T \subseteq S} a_S \det(X|_S)$$

We will want a formula for $p_{\vec{a},T}$ in terms of $p_{\vec{a}}$. To do this, we will need to define the Schur complement of X when X is a symmetric matrix, and $T \subseteq [n]$. Suppose we have the following block structure for X,

$$X = \begin{pmatrix} X|_T & X_{T,[n]\setminus T} \\ X_{[n]\setminus T,T} & X|_{[n]\setminus T} \end{pmatrix}$$

then the Schur complement of X with respect to T is

$$X \setminus T = X|_{[n]-T} - X_{[n]\setminus T,T} X|_T^{-1} X_{T,[n]\setminus T}.$$

Lemma 2.1. *Suppose that $T \subseteq [n]$ with $|T| = t \leq k$, then*

$$p_{\vec{a},T}(X) = \det(X|_T)\left(\sum_{i\in T} \frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X \setminus T)$$

Proof. Firstly, it is possible to see from the polynomial expansion of the determinant that

$$\frac{\partial}{\partial X_{ii}} \det(X|_S) = \begin{cases} 0 \text{ if } i \notin S \\ \det(X|_{S\setminus i}) \text{ otherwise.} \end{cases}$$

Therefore, we can see by induction that

$$\left(\sum_{i\in T} \frac{\partial}{\partial X_{ii}}\right)^t \det(X|_S) = \begin{cases} 0 \text{ if } T \not\subseteq S \\ \det(X|_{S\setminus T}) \text{ otherwise} \end{cases}$$

We then have that

$$\left(\sum_{i\in T} \frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X) = \sum_{S \in \binom{[n]}{k} : T \subseteq S} a_S \det(X|_{S\setminus T}).$$

Next, we will need to use the following two facts about Schur complements: firstly, there is the Schur complement lemma, which states that

$$\det(X) = \det(X|_T)\det(X \setminus T).$$

Secondly, we have that Schur complements commute with taking submatrices, in that if $T \subseteq S \subseteq [n]$, then

$$(X|_S) \setminus T = (X \setminus T)|_{S\setminus T}.$$

Therefore, we see that

$$\begin{aligned} \det(X|_T)\left(\sum_{i\in T} \frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X \setminus T) &= \sum_{S \in \binom{[n]}{k} : T \subseteq S} a_S \det(X|_T)\det((X \setminus T)|_{S\setminus T}) \\ &= \sum_{S \in \binom{[n]}{k} : T \subseteq S} a_S \det(X|_S) \\ &= p_{\vec{a},T}(X) \end{aligned}$$

□

2.2. **Weighted Determinantal Point Processes.** We now let $\vec{a} = (a_S : S \in \binom{[n]}{k}) \in \mathbb{R}_+^{\binom{n}{k}}$, so that these coefficients are assumed to be nonnegative. Let M be a fixed positive semidefinite definite matrix.

We define the weighted determinantal point process for this set of coefficients to be a proability distribution on $S \in \binom{[n]}{k}$ with probability mass function given by

$$\mu_{M,\vec{a}}(S) \propto a_S \det(M|_S).$$

To relate this to sparse linear regression, we can imagine using the following randomized variant of sparse linear regression. Given the data A and b, first draw a random subset S according to $\mu_{AA^\intercal,\vec{a}}$, and then perform the regression using only the columns of A in S.

We can then notate the expected error produced by this randomized linear regression.

$$\eta_{\vec{a},T,k}(A,b) = \mathbb{E}_{\mu_{AA^\intercal,\vec{a}}}[\ell(A_S,b)|T \subseteq S].$$

Once again, this expectation serves as an upper bound on the sparse linear regression problem.

We can then relate this quantity $\eta_{\vec{a},T,k}(A,b)$ to the lpm polynomial $p_{\vec{a}}$:

Theorem 2.1. *If $AA^\intercal$ has rank at least k, then*

$$\eta_{\vec{a},T,k}(A,b) = \|b\|^2 - \frac{p_{\vec{a},T}(A^\intercal A + A^\intercal bb^\intercal A)}{p_{\vec{a},T}(AA^\intercal)} + 1$$

Proof. We first recall the so-called matrix determinant lemma, which states that for any invertible $X \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$

$$\det(X + vv^\intercal) = (1 + v^\intercal X^{-1} v)\det(X).$$

This implies that

$$\begin{aligned} p_{\vec{a},T}(A^\intercal A + A^\intercal bb^\intercal A) &= \sum_{S\in\binom{[n]}{k}, T\subseteq S} a_S \det(A^\intercal A|_S + A^\intercal bb^\intercal A|_S) \\ &= \sum_{S\in\binom{[n]}{k}, T\subseteq S} a_S \det(A^\intercal A|_S)\big(1 + b^\intercal A_S (A^\intercal A|_S)^{-1} A_S^\intercal b\big) \end{aligned}$$

We also recall the closed form formula for $\ell(A,b)$, given by

$$\ell(A,b) = \|b\|^2 - b^\intercal A(A^\intercal A)^{-1}A^\intercal b.$$

We can thus simplify the above expression and see that

$$\begin{aligned} p_{\vec{a},T}(A^\intercal A + A^\intercal bb^\intercal A) &= \sum_{S\in\binom{[n]}{k}, T\subseteq S} a_S \det(A^\intercal A|_S)\big(1 + \|b\|^2 - \ell(A_S,b)\big) \\ &= (1+\|b\|^2)\left(\sum_{S\in\binom{[n]}{k}, T\subseteq S} a_S \det(A^\intercal A)\right) + \sum_{S\in\binom{[n]}{k}} \det(A^\intercal A|_S)\ell(A|_S,b) \\ &= (1+\|b\|^2)p_{\vec{a}}(A^\intercal A) + \sum_{S\in\binom{[n]}{k}, T\subseteq S} \det(A^\intercal A|_S)\ell(A_S,b) \end{aligned}$$

We then have that

$$\begin{aligned} \frac{p_{\vec{a},T}(A^\intercal A + A^\intercal bb^\intercal A)}{p_{\vec{a},T}(A^\intercal A)} &= (1+\|b\|^2) + \big(\sum_{S\in\binom{[n]}{k}, T\subseteq S} \Pr(S)\ell(A_S,b)\big) \\ &= \|b\|^2 - \eta_{\vec{a},T,k}(A,b) + 1 \end{aligned}$$

Rearranging, we obtain the result.

□

Remark 2.2. *There is a rich literature concerning the proiblem of efficiently sampling from truncated determinantal point processes* [1], *but it is worth noting that sampling based approaches will not be effective at computing this quantity to the accuracy that we wish to compute it.*

For instance, consider the unweighted case where $\vec{a} = \vec{1}$, and suppose that $A^\intercal A$ is such that for any $S \subseteq [n]$ with $|S| = k$, $\det(A^\intercal A|_S) = 1$. In this case, we see that the determinantal point process selects elements from $\binom{[n]}{k}$ uniformly at random. So, we can consider two cases: one in which b is chosen to lie in the image of A_S, and one in which b is chosen to be orthogonal to the image of A. If we attempted to distinguish these two cases using sampling, we would need to take $\Omega(\binom{[n]}{k})$ samples before we could find that there is some S so that $\ell(A_S, b) < \|b\|^2$. On the other hand, by computing the expectation in these two cases, we could immediately detect which situation we are in.

We have thus reduced the problem of computing $\eta_{\vec{a},T,k}(A, b)$ to that of computing $p_{\vec{a},T}$, which the following lemma makes formal.

Lemma 2.3. *If we can compute $p_{\vec{a}}(X)$ in $O(\tau)$ time for any complex symmetric matrix X, then we can compute $\eta_{\vec{a},T,k}(A, b)$ in $O(k\tau + k\log(k) + n^\omega)$ time, where ω denotes the matrix multiplication constant.*

Proof. We have seen that in order to compute $\eta_{\vec{a},T,k}(A, b)$, it suffices to compute $p_{\vec{a},T}$ at $A^\intercal A$ and $A^\intercal A + A^\intercal bb^\intercal A$.

Let $|T| = t$. We have seen in lemma 2.1 that

$$p_{\vec{a},T}(X) = \det(X|_T)\left(\sum_{i\in T}\frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X \setminus T)$$

For symmetric matrices X, We can easily compute $X \setminus T$ and $\det(X|_T)$ in n^ω time, and so it remains to argue that we can compute

$$\left(\sum_{i\in T}\frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X)$$

in the desired time.

If we let D be the diagonal matrix so that $D_{ii} = \begin{cases} 1 \text{ if } i \in T \\ 0 \text{ otherwise} \end{cases}$, we can see that we can rewrite this expression as a directional derivative:

$$\left(\sum_{i\in T}\frac{\partial}{\partial X_{ii}}\right)^t p_{\vec{a}}(X) = \frac{d^t}{dt^t}p_{\vec{a}}(X + tD)|_{t=0}.$$

Notice that $p_{\vec{a}}(X + tD)$ is a univariate polynomial in t of degree k. Therefore, if we can compute this function at k distinct values of t, then we can uniquely recover all of the coefficients of this univariate polynomial via interpolation. Computing this at any k distinct values only requires $k\tau$ time.

In fact, using the Fast-Fourier transform, it is in fact possible to reconstruct all of the coefficients of this polnymial this $k\log(k)$ time from its evaluations at all the r^{th} roots of unity where $r = 2^{\lceil \log(k) \rceil}$.

Once we have all of the coefficients of $p_{\vec{a}}(X + tD)$, we can compute $\frac{d^t}{dt^t}$ in constant time by reading of the t^{th} coefficient of this polynomial. This gives the final time complexity given in the theorem. □

2.3. **Characteristic Coefficients.** We will see that we can in fact obtain much faster methods when the polynomial in question is the characteristic coefficient defined above.

In fact, there are a number of algorithms for computing the characteristic coefficients of a symmetric matrix. Here, we will use the the Fadeev-LeVerrier method for computing the characteristic coefficients, which has the advantage of not needing to compute all of the characteristic coefficients at once.

Lemma 2.4. *We can compute $\eta_{\vec{1},T,k}(A, b)$ in $O(kn^\omega)$ time where ω is the matrix multiplication constant.*

Proof. We will once again make use of our result that

$$p_{\vec{1},T}(X) = \det(X|_T)\left(\sum_{i\in T}\frac{\partial}{\partial X_{ii}}\right)^{|T|} p_{\vec{1}}(X \setminus T)$$

We will note that for the characteristic coefficients,

$$\left(\sum_{i\in T}\frac{\partial}{\partial X_{ii}}\right)^{|T|} c_k^n(X) = c_{k-|T|}^{n-|T|}(X|_{[n]\setminus T})$$

Therefore,

$$p_{\vec{1},T}(X) = \det(X|_T)c_k^n(X \setminus T)$$

The Fadeev-LeVerrier algorithm computes c_k^n in $O(kn^\omega)$ time [5], and the result follows. □

2.4. A Heuristic for Sparse Regression.

We will focus this section on the unweighted determinantal point process, as that illustrates all of the relevant ideas, and is what we will use in our simulations.

We can now present our heuristic for finding a sparse regressor. We can think of method as follows: we will iteratively construct the set which we will use for regression. At each step, we will choose the element from $[n]$ which minimizes our expected error, if we condition on taking the elements we have already selected.

Algorithm 1 The η-greedy method

```
$T \leftarrow \varnothing$
for $t = 1 \dots k$ do
    $j \leftarrow \operatorname{argmin} \eta_{T+j,k}(A, b)$
    $T \leftarrow T + j$
end for
    return T
```

3. Connections to Hyperbolic Optimization

3.1. Conical Optimization and Sparse Linear Regression.

In [3], the authors introduced a conical optimization formulation for sparse regression and also introduced the idea of finding tractible convex relaxations of this formulation. This formulation was simplified and more relaxations were found in later work [4], and [7] gave an elegant formulation for this problem in terms of the factor-width-k cone:

$$\begin{aligned} \alpha = \text{minimize} \quad & \|b\|^2 - \operatorname{tr}(XA^\intercal bb^\intercal A) \\ \text{such that} \quad & \operatorname{tr}(AA^\intercal X) = 1 \\ & X \in \mathcal{FW}^{n,k} \end{aligned} \tag{3.1}$$

Here, $\mathcal{FW}^{n,k}$ denotes the factor width k cone, which is defined as

$$\mathcal{FW}_k^n = \operatorname{conv}\{xx^\intercal : x \in \mathbb{R}^n, \|x\|_0 \leq k\}.$$

This cone has been the subject of much study, for example in [11]. In particular, its dual is known to be

$$S_k^n = \{X \in \operatorname{Sym}(\mathbb{R}^n) : \forall S \in \binom{[n]}{k}, X|_S \succeq 0\}.$$

Here, $X|_S \succeq 0$ denotes the fact that $X|_S$ is PSD.

In [9], a relaxation was given for S_k^n, which happens to be a hyperbolicity cone. We define

$$H^{n,k} = \{X \in \operatorname{Sym}(\mathbb{R}^n) : \forall t \geq 0, c_n^k(X + tI) \geq 0\}.$$

This relaxation was used to fully give bounds for the possible eigenvalues of matrices of S_k^n.

In the next section, we will give some general theory about hyperbolic polynomials and their hyperbolicity cones, and then show how this theory relates to the sampling theory given in the previous part of this paper.

3.2. Hyperbolicity Cones.

A multivariate polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ is said to be *hyperbolic* with respect to fixed $e \in \mathbb{R}^n$ if $p(e) > 0$, and for all $x \in \mathbb{R}^n$, the univariate polynomial $g(t) = p(x + te)$ is real rooted, in the sense that the only complex numbers t where $g(t) = 0$ are real. A multivariate polynomial p is said to be *stable* if p is hyperbolic with respect to e for all $e \in \mathbb{R}_+^n$.

Hyperbolic polynomials were originally defined in the work of Garding in differential equations [15], and they have been the subject of much study because of their connections to combinatorics and optimization [20].

One object of particular interest attached to a polynomial p is the *hyperbolicity cone* of p with respect to e.

If p is hyperbolic with respsect to e, then the hyperbolicity cone of p with respect to e is defined as

$$\Lambda_e(p) = \{x \in \mathbb{R}^n : \forall t \geq 0, p(x + te) \geq 0\}.$$

The hyperbolicity cone of a polynomial is known to be convex, and moreover, it is known that the function $\log(p(x))$ is a self-concordant barrier function for $\Lambda_e(p)$ [16]. For our purposes, it is sufficient to note that $p(x) > 0$ for $x \in \operatorname{relint} \Lambda_e(p)$, and $p(x) = 0$ for $x \in \partial \Lambda_e(p)$.

An important example of a hyperbolic polynomial is the determinant of a symmetric matrix, which is hyperbolic with respect to the identity matrix. Hyperbolicity in this case follows immediately from the spectral theorem. The associated hyperbolicity cone is precisely the positive semidefinite cone.

It is known that if p is hyperbolic with respect to e, then $D_e p$ is also hyperbolic with respect to e, where D_e denotes the directional derivative of p with respect to e. In the case of the determinant, $D_I^k \det(X) = (n-k)! c_{n-k}^n(X)$, which shows that the characteristic coefficients are all hyperbolic.

One particular class of polynomials which are of interest to us are the hyperbolic lpm polynomials. In [10], it was shown that

Theorem 3.1. *The lpm polynomial $p_{\vec{a}}(X)$ with nonnegative coefficients is hyperbolic with respect to I if and only if*

$$\sum_{S \in \binom{[n]}{k}} a_S \prod_{i \in S} x_i$$

is stable.

Any lpm polynomial gives rise to a family of relaxations for S_k^n, due to the following fact:

Theorem 3.2. *Suppose that $p_{\vec{a}}(X)$ is a degree k lpm polynomial with nonnegative coefficients which is hyperbolic with respect to I, then*

$$S_k^n \subseteq \Lambda_I(p_{\vec{a}}).$$

Proof. If $X \in S_k^n$, then $X|_S \succeq 0$, so that for any $t \geq 0$,

$$\det((X + tI)|_S) \geq 0.$$

In particular, for $t \geq 0$,

$$p_{\vec{a}}(X + tI) = \sum_{S \in \binom{[n]}{k}} a_S \det((X + tI)|_S) \geq 0.$$

This shows that $X \in \Lambda_I(p_{\vec{a}})$. □

3.3. **Optimization and Sampling.** We now connect the probabilistic formula of η_k^n to this optimization perspective.

We can dualize Equation 3.1, and apply Slater's condition to see that strong duality holds.

$$\ell_k(A, b) = \begin{array}{cc} \text{maximize} & \|b\|^2 - y \\ \text{such that} & A^\intercal A y - A^\intercal b b^\intercal A \in S^{n,k} \end{array}. \tag{3.2}$$

We may then relax this problem as follows: let $p_{\vec{a}}$ be any lpm polynomial with nonnegative coefficients, and which is hyperbolic with respect to I. Then,

$$\ell_k(A, b) \leq \begin{array}{cc} \text{maximize} & \|b\|^2 - y \\ \text{such that} & A^\intercal A y - A^\intercal b b^\intercal A \in \Lambda(p_{\vec{a}}) \end{array}. \tag{3.3}$$

Theorem 3.3.

$$\eta_{\vec{a}, \varnothing, k}(A, b) = \begin{array}{cc} \textit{maximize} & \|b\|^2 - y \\ \textit{such that} & A^\intercal A y - A^\intercal b b^\intercal A \in \Lambda(p_{\vec{a}}) \end{array}.$$

Proof. It is clear that the optimal value of y,

$$A^\intercal A y - A^\intercal b b^\intercal A \in \partial \Lambda(p_{\vec{a}})$$

At this point, we have that $y > 0$, and

$$p_{\vec{a}}(A^\intercal A y - A^\intercal b b^\intercal A) = 0.$$

Regard

$$g(y) = p_{\vec{a}}(A^\intercal A y - A^\intercal b b^\intercal A)$$

as a univariate polynomial in y.

When $y = 0$, we see that

$$g(0) = p_{\vec{a}}(A^\intercal b b^\intercal A) = 0,$$

and indeed, because $A^\intercal b b^\intercal A$ is of rank 1, this root is of multiplicity $k - 1$.

This implies that $g(y)$ must have the form

$$g(y) = c_1 y^k + c_2 y^{k-1}$$

Here,

$$c_1 = \lim_{y \to \infty} \frac{g(y)}{y} = p_{\vec{a}}(A^\intercal A),$$

and

$$c_2 = g(1) - c_1 = p_{\vec{a}}(A^\intercal A + A^\intercal b b^\intercal A) - p_{\vec{a}}(A^\intercal A).$$

Therefore, the only nonzero root of $g(y)$ is when

$$y = 1 - \frac{p_{\vec{a}}(A^\intercal A + A^\intercal b b^\intercal A)}{p_{\vec{a}}(A^\intercal A)}.$$

Combining, we see that

$$\eta_{\vec{a},\varnothing,k}(A,b) = \begin{array}{ll} \text{maximize} & \|b\|^2 - y \\ \text{such that} & A^\intercal A y - A^\intercal b b^\intercal A \in \Lambda(p_{\vec{a}}) \end{array}.$$

As desired. □

4. Sparse Regression and PCA

In this section, we will consider

$$\eta_{k,\varnothing}(A,b) = \|b\|^2 - \frac{c_k^n(A^\intercal A + A^\intercal b b^\intercal A)}{c_k^n(A^\intercal A)} + 1.$$

We have seen that this provides an upper bound on the sparse least squares loss. On the other hand, $c_k^n(X)$ is a basis invariant polynomial. This suggests that we may be able ot relate some basis invariant features of A, such as the singular values of A, to the sparse linear regression loss.

For this, let the singular value decomposition of A be

$$A = U\Sigma V$$

where U and V are orthogonal, and Σ is diagonal, where $\Sigma_{ii} = \sigma_i(A)$, and $\sigma_1(A) \geq \sigma_2(A) \dots$.

The *principal components analysis* (PCA) problem is to find the k largest singular values of A, and the corresponding k singular directions, $v_1, \dots, v_k$. A common property of real world data matrices is that $\sigma_i(A)$ is very small for i much smaller than n, so that A is actually well approximated by the first $k < n$ singular vectors.

We define the spectral sparsification of A to be

$$A^{(k)} = U\Sigma^{(k)}V$$

Here, $\Sigma^{(k)}$ is obtained by setting all but the first k columns of Σ to 0. Out of all rank k matrices, $A^{(k)}$ is known to be the one that is closest to A in Frobenius norm [**?**].

Consider $\ell(A^{(k)}, b)$, which is the result of using the top k singular vectors of A to perform linear regression on b. This quantity is natural because we may wish to first perform dimension reduction by spectrally sparsifying A, and then performing linear regression on b.

We will show that if the top k singular values of A are significantly larger than the bottom k singular values, then in fact, sparse linear regression performs comparably to regression according to the top k singular vectors.

Lemma 4.1. *Let*

$$p = \frac{\prod_{i=1}^k \sigma_i^2}{c_k^n(A^\intercal A)}$$

Then,

$$\eta_{k,\varnothing}(A,b) \leq p\ell(A^{(k)}, b) + (1-p)\|b\|^2.$$

Proof. it can be seen from our formula that

$$\eta_{k,\varnothing}(A,b) = \eta_{k,\varnothing}(U^\intercal A V^\intercal, U^\intercal b) = \eta_{k,\varnothing}(\Sigma, U^\intercal b)$$

Now, we note that the tDDP μ_{Σ^2} assigns to the set S probability $\frac{\prod_{i \in S} \sigma_i(A)^2}{c_k^n(A^\intercal A)}$.

Therefore,

$$\eta_{k,\varnothing}(\Sigma, U^\intercal b) = \mathbb{E}_{\mu_{\Sigma^2}}[\ell(\Sigma|_S, U^\intercal b)].$$

We will write

$$\mathbb{E}_{\mu_{\Sigma^2}}[\ell(\Sigma|_S, U^\intercal b)] = \Pr(S = \{1,\dots,k\})\ell(\Sigma^{(k)}, U^\intercal b) + (1 - \Pr(S = \{1,\dots,k\}))\mathbb{E}_{\mu_{\Sigma^2}}[\ell(\Sigma|_S, U^\intercal b)|S \neq \{1,\dots,k\}].$$

Now, we have that $\Pr(S = \{1,\dots,k\}) = \frac{\prod_{i \in S} \sigma_i(A)^2}{c_k^n(A^\intercal A)} = p$. We also have that $\ell(\Sigma|_S, U^\intercal b) \leq \|b\|^2$, so that

$$\eta_{k,\varnothing}(A,b) \leq p\ell(\Sigma^{(k)}, U^\intercal b) + (1-p)\|b\|^2 = p\ell(A^{(k)}, b) + (1-p)\|b\|^2,$$

as desired. □

We now want to give some natural conditions under which $p = \frac{\prod_{i=1}^k \sigma_i^2}{c_k^n(A^\intercal A)}$ is close to 1. A strong such condition is that A is close to low rank, in the following sense: A exhibits (ϵ, k)-spectral decay if $\epsilon\sigma_k(A) \geq \sigma_{k+1}(A)$.

Lemma 4.2. *If A exhibits (ϵ, k) spectral decay, then for all b, we have that*

$$\eta_{k,\varnothing}(A,b) \leq (1 - \frac{1}{(1+\epsilon^2)^{nk}})\|b\|^2 + \frac{1}{(1+\epsilon^2)^{nk}}\ell_k(A,b).$$

Proof. It suffices here to show that

$$p = \frac{\prod_{i=1}^k \sigma_i^2}{c_k^n(A^\intercal A)} \geq (1+\epsilon^2)^{nk}.$$

Equivalently, we can show that

$$\frac{c_k^n(\Sigma^2)}{\det(\Sigma|_{[k]})^2} \leq (1+\epsilon)^{nk},$$

We expand c_k^n out as a sum over $S \in \binom{[n]}{k}$, and reindex the summation to be over $|S\Delta[k]|$ first.

$$c_k^n(\Sigma^2) = \sum_{\ell=0}^{k} \sum_{S\in\binom{[n]}{k}:|S\Delta[k]|=\ell} \prod_{i\in S} \sigma_i^2$$

Now, notice that if $|S\Delta[k]| = \ell$, then $\prod_{i\in S}\sigma_i^2 \leq \epsilon^{2\ell}\prod_{i=1}^k \sigma_i^2$. There are $\binom{k}{\ell}\binom{n}{\ell}$ sets with the property that $|S\Delta[k]|$, so that

$$c_k^n(\Sigma^2) \leq \prod_{i=1}^k \sigma_i^2 \left(\sum_{\ell=0}^k \epsilon^{2\ell}\binom{k}{\ell}\binom{n}{\ell}\right)$$

Now, we note that

$$\binom{k}{\ell}\binom{n}{\ell} \leq \binom{nk}{\ell},$$

which follows from simply expanding out both sides.

Therefore,

$$c_k^n(\Sigma^2) \leq \prod_{i=1}^k \sigma_i^2 \left(\sum_{\ell=0}^{nk} \epsilon^{2\ell}\binom{nk}{\ell}\right) = \det(\Sigma|_{[k]}^2)(1+\epsilon^2)^{nk}.$$

This gives the desired result. □

5. Compressed Sensing

In the compressed sensing problem, we consider $A \in \mathbb{R}^{m\times n}$, where $m < n$, and we are given the value

$$Af + \epsilon,$$

where $\|f\|_0 \leq k \leq m$, and ϵ is a noise term, chosen from $\mathcal{N}(0, \sigma^2 I)$.

Our goal is to recover f from this information.

In the existing literature, it is common to assume that A has the restricted isometry property (RIP), which states that for all $S \in \binom{[n]}{2k}$,

$$1 - \delta \leq \lambda_{min}(A^\intercal A|_S) \leq 1 - \delta.$$

For instance, if the entries of A are chosen from a normal distribution, it has been shown that with high probability A will satisfy the RIP.

TODO: Prove that in this setting, this algorithm succeeds with high probability.

References

[1] Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *Conference on Learning Theory*, pages 103–115. PMLR, 2016.

[2] Nima Anari, Shayan Oveis Gharan, Amin Saberi, and Mohit Singh. Nash social welfare, matrix permanent, and stable polynomials. *arXiv preprint arXiv:1609.07056*, 2016.

[3] Alper Atamturk and Andres Gomez. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

[4] Francis Bach, Selin Damla Ahipasaoglu, and Alexandre d'Aspremont. Convex relaxations for subset selection. *arXiv preprint arXiv:1006.3601*, 2010.

[5] Christian Bär. The faddeev-leverrier algorithm and the pfaffian. *Linear Algebra and its Applications*, 630:39–55, 2021.

[6] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. A determinantal point process for column subset selection. *arXiv preprint arXiv:1812.09771*, 2018.

[7] Walid Ben-Ameur and José Neto. New bounds for subset selection from conic relaxations. *European Journal of Operational Research*, 298(2):425–438, 2022.

[8] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

[9] Grigoriy Blekherman, Santanu S Dey, Kevin Shu, and Shengding Sun. Hyperbolic relaxation of k-locally positive semidefinite matrices. *arXiv preprint arXiv:2012.04031*, 2020.

[10] Grigoriy Blekherman, Mario Kummer, Raman Sanyal, Kevin Shu, and Shengding Sun. Linear principal minor polynomials: Hyperbolic determinantal inequalities and spectral containment, 2021.

[11] Erik G Boman, Doron Chen, Ojas Parekh, and Sivan Toledo. On factor width and symmetric h-matrices. *Linear algebra and its applications*, 405:239–248, 2005.

[12] Julius Borcea and Petter Brändén. Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized fischer products. *Duke Mathematical Journal*, 143(2):205–223, 2008.

[13] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[14] Alberto Del Pia, Santanu S. Dey, and Robert Weismantel. Subset selection in sparse matrices. *SIAM Journal on Optimization*, 30(2):1173–1190, 2020.

[15] Lars Gårding. Linear hyperbolic partial differential equations with constant coefficients. *Acta Mathematica*, 85:1–62, 1951.

[16] Osman Güler. Hyperbolic polynomials and interior point methods for convex programming. *Mathematics of Operations Research*, 22(2):350–377, 1997.

[17] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[18] Mutsuki Kojima and Fumiyasu Komaki. Determinantal point process priors for bayesian variable selection in linear regression. *Statistica Sinica*, 26(1):97–117, 2016.

[19] James Saunderson. Certifying polynomial nonnegativity via hyperbolic optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(4):661–690, 2019.

[20] David Wagner. Multivariate stable polynomials: theory and applications. *Bulletin of the American Mathematical Society*, 48(1):53–84, 2011.

[21] William J Welch. Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15(1):17–25, 1982.

[22] Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117–33123, 2020.

Department of Mathematics, Georgia Institute of Technology, Atlanta, GA
Email address: `kshu8@gatech.edu`
URL: `www.kevinshu.me`