Submitted to the Annals of Statistics

SPARSE HIGH DIMENSIONAL REGRESSION: EXACT SCALABLE ALGORITHMS AND PHASE TRANSITIONS

By Dimitris Bertsimas and Bart Van Parys*

Massachusetts Institute of Technology

We present a novel binary convex reformulation of the sparse regression problem that constitutes a new duality perspective. We devise a new cutting plane method and provide evidence that it can solve to provable optimality the sparse regression problem for sample sizes n and number of regressors p in the 100,000s, that is two orders of magnitude better than the current state of the art, in seconds. The ability to solve the problem for very high dimensions allows us to observe new phase transition phenomena. Contrary to traditional complexity theory which suggests that the difficulty of a problem increases with problem size, the sparse regression problem has the property that as the number of samples n increases the problem becomes easier in that the solution recovers 100% of the true signal, and our approach solves the problem extremely fast (in fact faster than `Lasso`), while for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal. Finally, our approach readily and tractably generalizes to nonlinear Kernel regression.

1. Introduction. Given input data $X = (x_1, \ldots, x_n) \in \mathrm{R}^{n\times p}$ and response data $Y = (y_1, \ldots, y_n) \in \mathrm{R}^n$, the problem of linear regression with a Tikhonov (1943) regularization term and an explicit sparsity constraint is defined as

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2\gamma}\left\|w\right\|_2^2 + \frac{1}{2}\left\|Y - Xw\right\|_2^2 \\ \text{s.t.} & \left\|w\right\|_0 \leq k, \end{array} \tag{1}$$

where $\gamma > 0$ is a given weight that controls the importance of the regularization term. The number of regression coefficients needed to explain the observations from the input data is limited to k by the ℓ_0-norm constraint on the regressor w. Tikhonov regularization helps to reduce the effect of noise in the input data. Regularization and robustness are indeed known to be

*The second author is generously supported by the Early Post.Mobility fellowship No. 165226 of the Swiss National Science Foundation.

Keywords and phrases: Best subset selection, sparse regression, kernel learning, integer optimization, convex optimization

MSC 2010 subject classifications: Primary 62J07; secondary 90C10

intimately connected as shown for instance by Bertsimas and Fertis (2009); Xu, Caramanis and Mannor (2009). Evidently in practice, both the sparsity parameter k and the Tikhonov regularization term γ must ultimately be determined from the data. Cross validation is empirically found to be an effective method to determine both hyperparameters.

Background. Despite its conceptual appeal, our sparse regression problem (1) is recognized immediately as an intrinsically discrete optimization problem, which belongs to the class of NP-hard problems. Motivated by the apparent difficulty of the sparse regression formulation (1), much of the literature until recently has largely ignored the exact discrete formulation and rather focussed on heuristic approaches. Historically, the first heuristics methods for sparse approximation seem to have arisen in the signal processing community (c.f. the work of Mallat and Zhang (1993) and references therein) and typically are of an iterative thresholding type. More recently, one popular class of sparse regression heuristics solve convex surrogates to the sparse regression formulation (1). There is an elegant theory for such schemes promising large improvements over the more myopic iterative thresholding methods. Indeed, a truly impressive amount of high-quality work Bühlmann and van de Geer (2011); Hastie, Tibshirani and Wainwright (2015); Wainwright (2009) has been written on characterizing when exact solutions can be recovered, albeit through making strong assumptions on the data.

One such heuristic based on a convex proxy related to our formulation and particularly worthy of mention is the `Elastic Net` developed by Zou and Hastie (2005). One particular canonical form of the `Elastic Net` heuristic solves the proxy convex optimization problem

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2\gamma}\left\|w\right\|_2^2 + \frac{1}{2}\left\|Y - Xw\right\|_2^2 \\ \text{s.t.} & \left\|w\right\|_1 \leq \lambda, \end{array} \tag{2}$$

where the ℓ_1-norm constraint shrinks the regressor coefficients towards zero thus encouraging sparse regressors for λ tending to zero. When disregarding the Tikhonov regularization term, the popular `Lasso` heuristic introduced by Tibshirani (1996) is recovered. An important factor in favor of heuristics such as `Lasso` and `Elastic Net` are their computational feasibility and scalability. Indeed, problem (2) can be solved efficiently and mature software implementations such as `GLMNet` by Friedman, Hastie and Tibshirani (2013) are available.

Despite all of the aforementioned positive properties, proxy based methods such as `Lasso` and `Elastic Net` do have several innate shortcomings.

These shortcomings are well known in the statistical community too. First and foremost, as argued in Bertsimas, King and Mazumder (2016) they do not recover very well the sparsity pattern. Furthermore, the `Lasso` leads to biased regression regressors, since the ℓ_1-norm penalizes both large and small coefficients uniformly. In sharp contrast, the ℓ_0-norm sparsifies the regressor without conflating the effort with unwanted shrinking.

For a few decades the exercise of trying to solve the sparse regression problem (1) at a practical scale was branded hopeless. Bixby (2012) noted however that in the last twenty-five years the computational power of Mixed Integer Optimization (MIO) solvers has increased at an astonishing rate. Riding on the explosive improvement of MIO formulations, Bertsimas, King and Mazumder (2016) achieved to solve the sparse regression problem (1) for problem instances of dimensions n, p in the 1000s. Using a big-$\mathcal{M}$ formulation of the cardinality constraint, the sparse regression problem (1) can indeed be transformed into the MIO problem

$$\begin{array}{lll} \min & \frac{1}{2\gamma}\|w\|_2^2 + \frac{1}{2}\|Y - Xw\|_2^2 & \\ \text{s.t.} & w \in \mathrm{R}^p,\ s \in \mathrm{S}_k^p & \\ & -\mathcal{M}s_j \le w_j \le \mathcal{M}s_j, & \forall j \in [p]. \end{array} \tag{3}$$

With the help of the binary set $\mathrm{S}_k^p := \left\{s \in \{0,1\}^p \ :\ \mathbb{1}^\top s \le k\right\}$, the constraint in (3) ensures that the regression coefficient w_j is nonzero only if the selection variable $s_j = 1$ for a sufficiently large constant $\mathcal{M}$. The constant $\mathcal{M}$ must be estimated from data as outlined in Bertsimas, King and Mazumder (2016) to ensure the equivalence between the sparse regression problem (1) and its MIO formulation (3). This MIO approach is significantly more scalable than the leaps and bounds algorithm outlined in Furnival and Wilson (2000), largely because of the advances in computer hardware, the improvements in MIO solvers, and the specific warm-start techniques developed by Bertsimas, King and Mazumder (2016). Even so, many problems of practical size are still far beyond the scale made tractable through this approach.

A scalable perspective. Although a direct big-$\mathcal{M}$ formulation of the sparse regression problem results in a well posed MIO problem, the constant $\mathcal{M}$ needs to be chosen with care as not to impede its numerical solution. The choice of this data dependent constant $\mathcal{M}$ indeed affects the strength of the MIO formulation (3) and is critical for obtaining solutions quickly in practice. Furthermore, as the regression dimension p grows, explicitly constructing the MIO problem (3), let alone solving it, becomes burdensome.

In order to develop an exact scalable method to the sparse regression problem (1) capable of solving problem instances of sample size n and regressor dimension in the 100,000s, a different perspective on sparse regression is needed.

The big-$\mathcal{M}$ formulation (3) of the sparse linear regression problem (1) takes on a primal perspective to regression. Like most exact as well as heuristic sparse regression formulations, the big-$\mathcal{M}$ formulation (3) indeed tries to solve for the optimal regression coefficients $w_0^\star$ in (1) directly. However, it is well known in the kernel learning community that often far deeper results can be obtained if a dual perspective is taken. We show that this dual perspective can be translated to a sparse regression context as well and offers a novel road to approach exact sparse regression. Taking this new perspective, sparse regression problem (1) can be reduced to a pure integer convex optimization problem avoiding the construction of any auxiliary constants.

Crucially, a tailored cutting plane algorithm for the resulting Convex Integer Optimization (CIO) problem renders solving the sparse regression problem (1) to optimality tractable for problem instances with number of samples and regressors in the 100,000s. That is two orders of magnitude better than the current state of art and impeaches the primary selling point of heuristic approaches such as `Elastic Net` or `Lasso`. As we will discuss subsequently, our cutting plane algorithm is often comparable or indeed even faster than the aforementioned convex proxy heuristic approaches.

Phase Transitions. Let the data come from $Y = Xw_{\text{true}} + E$ where E is zero mean noise uncorrelated with the signal Xw_{true}, then we define the accuracy and false alarm rate of a certain solution $w^\star$ in recovering the correct support as:

$$A\% := 100 \times \frac{|\text{supp}(w_{\text{true}}) \cap \text{supp}(w^\star)|}{k}$$

and

$$F\% := 100 \times \frac{|\text{supp}(w^\star) \setminus \text{supp}(w_{\text{true}})|}{|\text{supp}(w^\star)|}.$$

Perfect support recovery occurs only then when $w^\star$ tells the whole truth ($A\% = 100$) and nothing but the truth ($F\% = 0$).

The ability to recover the support of the ground truth w_{true} of the `Lasso` heuristic (2) was shown by Donoho and Tanner (2009) to experience a phase transition. The phase transition described by Donoho and Tanner (2009) concerns the ability of the `Lasso` solution $w_1^\star$ to coincide in support with the ground truth w_{true}. This accuracy phase transition for the `Lasso` has

been extensively studied in Bühlmann and van de Geer (2011); Hastie, Tibshirani and Wainwright (2015); Wainwright (2009) and is considered well understood by now. That being said, the assumptions made on the data needed for a theoretical justification of such phase transition are quite stringent and often of limited practical nature. For instance, Wainwright (2009) showed that for uncorrupted observations Y and independent Gaussian input data X a phase transition occurs at the phase transition curve

$$n = 2k \log(p - k). \tag{4}$$

In the regime $n > 2k \log(p - k)$ exact recovery of the support occurs with high-probability, while on the other side of the transition curve the probability for successful recovery drops to zero. Nonetheless, this phase transition from accurate discovery to statistical meaninglessness has been widely observed empirically Donoho and Tanner (2009); Donoho and Stodden (2006) even under conditions in which these assumptions are severely violated.

For exact sparse regression (1) a similar phase transition has been observed by Zheng et al. (2015) and Wang, Xu and Tang (2011), although this transition is far less understood from a theoretical perspective than the similar transition for its heuristic counterpart. It is however known that the accuracy phase transition for exact sparse regression must occur even sooner than that of any heuristic approach. That is, exact sparse regression (1) yields statistically more meaningful optima than for instance the convex `Lasso` heuristic (2) does. Empirical verification of this phase transition was historically hindered due to the lack of exact scalable algorithms. Our novel cutting plane algorithm lifts this hurdle and opens the way to show the benefits of exact sparse regression empirically. We will show that exact regression is significantly better than `Lasso` in discovering all true relevant features ($A\% = 100$), while truly outperforming its ability to reject the obfuscating ones ($F\% = 0$).

More importantly, we present strong empirical evidence that a computational phase transition occurs as well. Specifically, there is a phase transition concerning our ability to solve the sparse regression problem (1) efficiently. Importantly, we did not find any evidence that the transition from accurate to unreliable discovery and from fast to slow running times are distinct. In other words, there is a phase transition in our ability to recover the true coefficients of the sparse regression problem and most surprisingly in our ability to solve it. This complexity phase transition does not seem to be reported before and sheds a new light on the complexity of sparse linear regression. Contrary to traditional complexity theory which suggests that the difficulty of a problem increases with size, the sparse regression problem

(1) has the property that for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal. However, for a large number of samples n, our approach solves the problem extremely fast and perfectly recovers the support of the true regressor w_{true}.

Contributions and structure.

1. In Section 2, we propose a novel binary convex reformulation of the sparse regression problem (1) that represents a new dual perspective to the problem. The reformulation does not use the big-$\mathcal{M}$ constant present in the primal formulation (3). In Section 3, we devise a novel cutting plane method and provide evidence that it can solve the sparse regression problem for sizes of n and p in the 100,000s. That is two orders of magnitude than what was achieved in Bertsimas, King and Mazumder (2016). The empirical computational results in this paper do away with the long held belief that exact sparse regression for practical problem sizes is a lost cause.
2. The ability to solve the sparse regression problem (1) for very high dimensional problems allows us to observe properties of the problem that demonstrate new phase transition phenomena. Specifically, we demonstrate experimentally in Section 4 that there is a threshold n_0 such that if $n/k \geq n_0$, then $w_0^\star$ recovers the true support ($A\% = 100$ for $F\% = 0$) and the time to solve problem (1) is seconds (for n and p in 100,000s) and it only grows only linear in n. Remarkably, these times are less than the time to solve `Lasso` for similar sizes. Moreover, if $n/k < n_0$, then $w_0^\star$ is statistically meaningless ($A\% = 0$ and $F\% = 100$) and the time to solve problem (1) grows proportional to $\binom{p}{k}$. In other words, there is a phase transition in our ability to recover the true coefficients of the sparse regression problem and most surprisingly in our ability to solve it. Contrary to traditional complexity theory that suggests that the difficulty of a problem increases with dimension, the sparse regression problem (1) has the property that for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal. However, for a large number of samples n, our approach solves the problem extremely fast and recovers $A\% = 100$ of the support of the true regressor w_{true}. Significantly, the threshold n_0 for the phase transition for full recovery of exact sparse regression is significantly smaller than the corresponding threshold for `Lasso`. Whereas `Lasso` tends to include many irrelevant features as

well, exact sparse regression furthermore achieves this full recovery at almost $F\% = 0$ false alarm rate.

3. We are able to generalize in Section 5 our approach to sparse kernel regression. We believe that this nonlinear approach can become a fierce and more disciplined competitor compared to "black box" approaches such as neural networks.

Notation. Denote with $[n]$ the set of integers ranging from one to n. The set S_k^p denotes the set

$$\mathrm{S}_k^p := \left\{ s \in \{0,1\}^p \ : \ \mathbb{1}^\top s \leq k \right\},$$

which contains all binary vectors s selecting k components from p possibilities. Assume that $(y_1, \ldots, y_p)$ is a collection of elements and suppose that s is an element of S_k^p, then y_s denotes the sub-collection of y_j where $s_j = 1$. We use $\|x\|_0$ to denote the number of elements of a vector x in R^p which are nonzero. Similarly, we use $\mathrm{supp}(x) = \{s \in \{0,1\}^p \ : \ s_i = 1 \iff x_i \neq 0\}$ to denote those indices of a vector x which are nonzero. Finally, we denote by S_+^n (S_{++}^n) the cone of $n \times n$ positive semidefinite (definite) matrices.

2. A convex binary reformulation of sparse linear regression. Sparse regression taken at face value is recognized as a mixed continuous and discrete optimization problem. Indeed, the sparse regressor w as an optimization variable in (1) takes values in a continuous subset of R^p. The ℓ_0-norm sparsity constraint, however, adds a discrete element to the problem. The support s of the sparse regressor w is discrete as it takes values in the binary set $\mathrm{S}_k^p = \left\{ s \in \{0,1\}^p \ : \ \mathbb{1}^\top s \leq k \right\}$. It should not come as a surprise then that the reformulation (3) developed by Bertsimas, King and Mazumder (2016) formulates the sparse regression problem as a MIO problem.

For the reasons outlined in the introduction of this paper, we take a different approach to the sparse regression problem (1) entirely. To that end we first briefly return to the ordinary regression problem for which any sparsity considerations are ignored and in which a linear relationship between input data X and observations Y is determined through solving the least squares regression problem

$$\begin{array}{rll} c := \min & \frac{1}{2\gamma} \|w\|_2^2 + \frac{1}{2} \|Y - Xw\|_2^2 \\ \text{s.t.} & w \in \mathrm{R}^p. \end{array} \tag{5}$$

We will refer to the previously defined quantity c as the regression loss. The quantity c does indeed agree with the regularized empirical regression

loss for the optimal linear regressor corresponding to the input data X and response Y. We point out now that the regression loss function c is convex as a function of the outer product $XX^\top$ and furthermore show that it admits an explicit characterization as a semidefinite representable function.

LEMMA 1 (The regression loss function c). *The regression loss c admits the following explicit characterizations*

$$c = \frac{1}{2} Y^\top \left(\mathbb{I}_n - X \left(\mathbb{I}_p/\gamma + X^\top X \right)^{-1} X^\top \right) Y, \tag{6}$$

$$= \frac{1}{2} Y^\top \left(\mathbb{I}_n + \gamma X X^\top \right)^{-1} Y. \tag{7}$$

Futhermore, the regression loss c as a function of the kernel matrix $XX^\top$ is conic representable using the formulation

$$c(XX^\top) = \min \left\{ \eta \in \mathrm{R}_+ \ : \ \begin{pmatrix} 2\eta & Y^\top \\ Y & \mathbb{I}_n + \gamma X X^\top \end{pmatrix} \in \mathrm{S}_+^{n+1} \right\}. \tag{8}$$

PROOF. As the minimization problem (5) over w in R^p is an unconstrained Quadratic Optimization Problem (QOP), the optimal value $w^\star$ satisfies the linear relationship $(\mathbb{I}_p/\gamma + X^\top X) w^\star = X^\top Y$. Substituting the expression for the optimal linear regressor $w^\star$ back into optimization problem, we arrive at

$$c = \frac{1}{2} Y^\top Y - \frac{1}{2} Y^\top X \left(\mathbb{I}_p/\gamma + X^\top X \right)^{-1} X^\top Y$$

establishing the first explicit characterization (6) of the regression function c. The second characterization (7) can be derived from the first with the help of the matrix inversion lemma found in Hager (1989) stating the identity

$$\left(\mathbb{I}_n + \gamma X X^\top \right)^{-1} = \mathbb{I}_n - X \left(\mathbb{I}_p/\gamma + X^\top X \right)^{-1} X^\top.$$

The Schur complement condition discussed at length in Zhang (2006) guarantees that as $\mathbb{I}_n + \gamma X X^\top$ is strictly positive definite, we have the equivalence

$$2\eta \geq Y^\top \left(\mathbb{I}_n + \gamma X X^\top \right)^{-1} Y \iff \begin{pmatrix} 2\eta & Y^\top \\ Y & \mathbb{I}_n + \gamma X X^\top \end{pmatrix} \in \mathrm{S}_+^{n+1}.$$

Representation (8) is thus an immediate consequence of expression (7) as well. □

We next establish that the sparse regression problem (1) can in fact be represented as a pure binary optimization problem. The following result provides a novel perspective on the sparse regression problem (1) and is of central importance in the paper.

THEOREM 1 (Sparse linear regression). *The sparse regression problem (1) can be reformulated as the nonlinear optimization problem*

$$\begin{array}{ll} \min & \frac{1}{2} Y^\top \left(\mathbb{I}_n + \gamma \sum_{j \in [p]} s_j K_j \right)^{-1} Y \\ \text{s.t.} & s \in \mathrm{S}_k^p, \end{array} \tag{9}$$

where the micro kernel matrices K_j in S_+^n are defined as the dyadic products

$$K_j := X_j X_j^\top. \tag{10}$$

PROOF. We start the proof by separating the optimization variable w in the sparse regression problem (1) into its support $s := \operatorname{supp} w$ and the corresponding non-negative entries w_s. Evidently, we can now write the sparse regression problem (1) as the bilevel minimization problem

$$\min_{s \in \mathrm{S}_k^p} \left[\min_{w_s \in \mathrm{R}^k} \frac{1}{2\gamma} \|w_s\|_2^2 + \frac{1}{2} \|Y - X_s w_s\|_2^2 \right]. \tag{11}$$

It now remains to be shown that the inner minimum can be found explicitly as the objective function of the optimization problem (9). Using Lemma 1, the minimization problem can be reduced to the binary minimization problem $\min_s \left\{ c(X_s X_s^\top) \ : \ s \in \mathrm{S}_k^p \right\}$. We finally remark that the outer product can be decomposed as the sum

$$X_s X_s^\top = \textstyle\sum_{j \in [p]} s_j X_j X_j^\top,$$

thereby completing the proof. □

The optimization problem (9) is a pure binary formulation of the sparse regression problem directly over the support s instead of the regressor w itself. It should be remarked that as the objective function in (9) is convex in the vector s, problem (9) casts the sparse regression problem as a CIO problem. Nevertheless, we will never explicitly construct the CIO formulation as such and rather develop in Section 3 an efficient cutting plane algorithm. We finally discuss here how the sparse regression formulation in Theorem 1 is related to kernel regression and admits an interesting dual relaxation.

2.1. *The kernel connection.* In ordinary linear regression a linear relationship between input data X and observations Y is determined through solving the least squares regression problem (5). The previous optimization problem is known as Ridge regression as well and balances the least-squares prediction error with a Tikhonov regularization term. One can solve the Ridge regression problem in the primal space – the space of parameters w – directly. Ridge regression is indeed easily recognized to be a convex QOP. Ordinary linear regression problems can thus be formulated as QOPs of size linear in the number of regression coefficients p.

Correspondingly, the big-$\mathcal{M}$ formulation (3) can be regarded as a primal perspective on the sparse regression problem (1). Formulation (3) indeed attempts to solve the sparse regression problem in the primal space of parameters w directly.

However, it is well known in the kernel learning community that far deeper results can be obtained if one approaches regression problems from its convex dual perspective due to Vapnik (1998a). Indeed, in most of the linear regression literature the dual perspective is often preferred over its primal counterpart. We state here the central result in this context to make the exposition self contained.

THEOREM 2 (Vapnik (1998a)). *The primal regression problem* (5) *can equivalently be formulated as the unconstrained maximization problem*

$$\begin{aligned} c = \max \quad & -\tfrac{\gamma}{2}\alpha^\top K\alpha - \tfrac{1}{2}\alpha^\top\alpha + Y^\top\alpha \\ \text{s.t.} \quad & \alpha \in \mathrm{R}^n, \end{aligned} \tag{12}$$

where the kernel matrix $K = XX^\top$ *in* S^n_+ *is a positive semidefinite matrix.*

The dual optimization problem (12) is a convex QOP as well and, surprisingly, scales only with the number of samples n and is insensitive to the input dimension p. This last surprising observation is what gives the dual perspective its historical dominance over its primal counterpart in the context of kernelized regression discussed in Schölkopf and Smola (2002). When working with high dimensional data for which the number of inputs p is vastly bigger than the number of samples n, the dual optimization problem (12) is smaller and often easier to solve.

For any i and j, the kernel matrix entry $K(i,j)$ corresponds to the inner product between input samples x_i and x_j in R^p. The matrix K is usually referred to as the kernel matrix or Gram matrix and is always positive definite and symmetric. Since the kernel specifies the inner products between all

pairs of sample points in X, it completely determines the relative positions of those points in the embedding space.

Our CIO formulation (9) of the sparse optimization problem (1) can be seen to take a dual perspective on the sparse regression problem (1). That is, our novel optimization formulation (9) is recognized as a subset selection problem in the space of kernels instead of regressors. It can indeed be remarked that when the sparsity constraint is omitted the kernel matrix reduces to the standard kernel matrix

$$K = \sum_{j\in[p]} X_j X_j^\top = XX^\top.$$

2.2. *A second-order cone relaxation.* Many heuristics approach the sparse regression problem (1) through a continuous relaxation. Indeed, a continuous relaxation of the big-$\mathcal{M}$ formulation (3) of the sparse regression problem is immediately recognized as the convex QOP

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2\gamma}\|w\|_2^2 + \frac{1}{2}\|Y - Xw\|_2^2 \\ \text{s.t.} & \|w\|_\infty \leq \mathcal{M}, \;\; \|w\|_1 \leq \mathcal{M}k \end{array} \tag{13}$$

which Bertsimas, King and Mazumder (2016) recognized as a slightly stronger relaxation than the `Elastic Net` (2). It thus makes sense to look at the continuous relaxation of the sparse kernel optimization problem (9) as well. Note that both the big-$\mathcal{M}$ (13) and `Elastic Net` (2) relaxation provide lower bounds to the exact sparse regression problem (1) in terms of a QOP. However, neither of the these relaxations is very tight. In Theorem 3 we will indicate that a more intuitive and comprehensive lower bound based on our CIO formulation (9) can be stated as a Second-Order Cone Problem (SOCP).

A naive attempt to state a continuous relaxation of the CIO formulation (9) in which we would replace the binary set S_k^p with its convex hull would result in a large but convex Semidefinite Optimization (SDO) problem. Indeed, the convex hull of the set S_k^p is the convex polytope $\left\{s \in [0,1]^p \;:\; \mathbb{1}^\top s \leq k\right\}$. It is, however, folklore that large SDOs are notoriously difficult to solve in practice. For this reason, we reformulate here the continuous relaxation of (9) as a small SOCP for which very efficient solvers do exist. This continuous relaxation provides furthermore some additional insight towards the binary formulation of the sparse regression problem (1).

Using Theorem 2, we can equate the continuous relaxation of problem (9) to the following saddle point problem

$$\min_{s\in\mathrm{conv}(\mathrm{S}_k^p)} \max_{\alpha\in\mathrm{R}^n} -\frac{\gamma}{2}\sum_{j\in[p]} s_j \cdot \left[\alpha^\top K_j \alpha\right] - \tfrac{1}{2}\alpha^\top\alpha + \alpha^\top y. \tag{14}$$

Note that the saddle point function is linear in α for any fixed s in the compact set $\text{conv}(\text{S}_k^p)$ and concave continuous in s for any fixed α. It then follows (see Sion (1958)) that we can exchange the minimum and maximum operators. By doing so, the continuous relaxation of our CIO problem satisfies

$$\begin{aligned}\min_{s\in\text{conv}(\text{S}_k^p)} & c(\textstyle\sum_{j\in[p]} s_j K_j) = \\ & \max_{\alpha\in\text{R}^n} -\frac{1}{2}\alpha^\top\alpha + \alpha^\top y - \frac{\gamma}{2}\max_{s\in\text{conv}(\text{S}_k^p)} \textstyle\sum_{j\in[p]} s_j\cdot\alpha^\top K_j\alpha.\end{aligned} \tag{15}$$

The inner maximization problem admits an explicit representation as the sum of the k-largest components in the vector with components $\alpha^\top K_j\,\alpha$ ranging over j in $[p]$. It is thus worth noting that this continuous relaxation has a discrete element to it. The continuous relaxation of the MIO problem (9) can furthermore be written down as a tractable SOCP.

THEOREM 3. *The continuous relaxation of the sparse kernel regression problem* (9) *can be reduced to the following SOCP*

$$\begin{aligned}\min_{s\in\text{conv}(\text{S}_k^p)} c(\textstyle\sum_{j\in[p]} s_j K_j) = \max\quad & -\frac{1}{2}\alpha^\top\alpha + \alpha^\top y - \mathbb{1}^\top u - kt \\ \text{s.t.}\quad & \alpha\in\text{R}^n,\ t\in\text{R},\ u\in\text{R}^p_+, \\ & \frac{2}{\gamma}u_j \geq \alpha^\top K_j\alpha - \frac{2}{\gamma}t, \quad \forall j\in[p].\end{aligned} \tag{16}$$

PROOF. The continuous relaxation of the optimization problem (9) was already identified as the optimization problem (15). We momentarily focus on the inner maximization problem in (15) and show it admits a closed form expression. As the only constraint on the (continuous) selection vector s is a knapsack constraint, the inner maximum is nothing but the sum of the k-largest terms in the objective. Hence, we have

$$\max_{s\in\text{conv}(\text{S}_k^p)} \textstyle\sum_{j\in[p]} s_j\cdot\alpha^\top K_j\alpha = \max_{[k]}([\alpha^\top K_1\alpha,\ldots,\alpha^\top K_p\alpha]),$$

where $\max_{[k]}$ is defined as the convex function mapping its argument to the sum of its k-largest components. Using standard linear optimization duality we have

$$\begin{aligned}\max_{[k]}(x) = \max\quad & x^\top s && = \quad \min\quad && kt + \mathbb{1}^\top u \\ \text{s.t.}\quad & s\in\text{R}_+ && \quad \text{s.t.}\quad && t\in\text{R},\ u\in\text{R}^p_+ \\ & s\leq\mathbb{1},\ \mathbb{1}^\top s = k && && u_j \geq x_j - t, \quad \forall j\in[p].\end{aligned}$$

where t and u are the dual variables corresponding to the constraints in the maximization characterization of the function $\max_{[k]}$. Making use of the dual characterization of $\max_{[k]}$ in expression (15) gives us the desired result. □

The continuous relaxation (16) of the sparse regression problem (1) discussed in this section is thus recognized as selecting the k-largest terms $\alpha^\top K_j \alpha$ to construct the optimal dual lower bound. We shall find that the dual offers an excellent warm start when attempting to solve the sparse linear regression problem exactly.

3. A cutting plane algorithm. We have formulated the sparse regression problem (1) as a pure binary convex optimization problem in Theorem 1. Unfortunately, no commercial solvers are available which are targeted to solve CIO problems of the type (9). In this section, we discuss a tailored solver largely based on the algorithm described by Duran and Grossmann (1986). The algorithm is a cutting plane approach which iteratively solves increasingly better MIO approximations to the CIO formulation (9). Furthermore, the cutting plane algorithm avoids constructing the CIO formulation (9) explicitly which can prove burdensome when working with high-dimensional data. We provide numerical evidence in Section 4 that the algorithm described here is indeed extremely efficient.

3.1. *Outer approximation algorithm.* In order to solve the CIO problem (9), we follow the outer approximation approach introduced by Duran and Grossmann (1986). The algorithm described by Duran and Grossmann (1986) proceeds to find a solution to the CIO problem (9) by constructing a sequence of MIO approximations based on cutting planes. In pseudocode, it can be seen to construct a piece-wise affine lower bound to the convex regression loss function c defined in equation (8).

Algorithm 1: The outer approximation process

input : $Y \in \mathrm{R}^n$, $X \in \mathrm{R}^{n \times p}$ and $k \in [1, p]$
output: $s^\star \in \mathrm{S}^p_k$ and $w^\star \in \mathrm{R}^p$
$s_1 \leftarrow$ warm start
$\eta_1 \leftarrow 0$
$t \leftarrow 1$
while $\eta_t < c(s_t)$ **do**
  $s_{t+1},\ \eta_{t+1} \leftarrow \arg\min_{s,\,\eta} \{\, \eta \in \mathrm{R}_+ \text{ s.t. } s \in \mathrm{S}^p_k,\ \ \eta \geq c(s_t) + \nabla c(s_t)(s - s_t),\ \forall i \in [t] \}$
  $t \leftarrow t + 1$
$s^\star \leftarrow s_t$
$w^\star \leftarrow 0, \quad w^\star_{s^\star} \leftarrow \left(\mathbb{I}_p / \gamma + X^\top_{s^\star} X_{s^\star} \right)^{-1} X^\top_{s^\star} Y$

At each iteration, the cutting plane added $\eta \geq c(s_t) + \nabla c(s_t)(s - s_t)$ cuts off the current binary solution s_t unless it happened to be optimal in (9). As the algorithm progresses, the outer approximation function c_t thus constructed

$$c_t(s) := \max_{i \in [t]} c(s_t) + \nabla c(s_t)(s - s_t)$$

becomes an increasingly better approximation to the regression loss function c of interest. Unless the current binary solution s_t is optimal, a new cutting plane will refine the feasible region of the problem by cutting off the current feasible binary solution.

In general, outer approximation methods are known as "multi-tree" methods because every time a linearization is added, a new MIO problem must be solved anew. Over the course of the iterative cutting plane Algorithm 1, multiple branch and bound trees are built in order to solve the successive MIO problems. We implement a "single tree" way of solving the iteration algorithm (1) by using dynamic constraint generation, known in the optimization literature as lazy constraint callbacks, which dynamically add cutting planes to the model whenever a binary feasible solution is found.

Lazy constraint callbacks are a relatively new type of callback. To date, the only MIO solvers which provide lazy constraint callback functionality are `CPLEX`, `Gurobi` and `GLPK`. The outer approximation method for solving CIO problems does not require lazy constraint callbacks, but if we do exploit their functionality, only one branch and bound tree needs to be built. This saves the rework of rebuilding a new branch and bound tree every time a new binary solution is found in Algorithm 1.

In what follows, we discuss two tailored adjustments to the general outer approximation method which render the overall method more efficient. The first concerns an efficient way to evaluate both the regression loss function c and its subgradient ∇c efficiently. The second discusses a heuristic to compute a warm start s_0 to ensure that the first cutting plane added is of high quality, causing the outer approximation algorithm to converge more quickly.

3.2. *Efficient dynamic constraint generation.* In the outer approximation method considered in this document to solve the CIO problem (9) linear constraints of the type

$$\eta \geq c(\bar{s}) + \nabla c(\bar{s})(s - \bar{s}) \tag{17}$$

at $\bar{s}$ a given iterate, are considered as cutting planes at every iteration. As such constraints need to be added dynamically, it is essential that we can

evaluate both the regression loss function c and its subgradient components efficiently.

LEMMA 2 (Derivatives of the optimal regression loss c). *Suppose the kernel matrix K is differentiable function of the parameter s. Then, we have that the gradient of the regression loss function $c(K) = \frac{1}{2}\alpha^\star(K)Y$ can be stated as*

$$\nabla c(s) = \alpha^\star(K)^\top \cdot \frac{\gamma}{2}\frac{\mathrm{d}K}{\mathrm{d}s} \cdot \alpha^\star(K),$$

where $\alpha^\star(K)$ maximizes (12) and hence is the solution to the linear system

$$\alpha^\star(K) = (\mathbb{I}_n + \gamma K)^{-1} Y.$$

We note that the naive numerical evaluation of the convex loss function c or any of its subgradients would require the inversion of the regularized kernel matrix $\mathbb{I}_n + \gamma \sum_{j\in[p]} \bar{s}_j K_j$. The regularized kernel matrix is dense in general and always of full rank. Unfortunately, matrix inversion of general matrices presents work in the order of $\mathcal{O}(n^3)$ floating point operations and quickly becomes excessive for sample sizes n in the order of a few 1,000s. Bear in mind that such an inversion needs to take place for each cutting plane added in the outer approximation Algorithm 1.

It would thus appear that computation of the regression loss c based on its explicit characterization (7) is very demanding. Fortunately, the first explicit characterization (6) can be used to bring down the work necessary to $\mathcal{O}(k^3 + nk)$ floating point operations as we will show now. Comparing equalities (6) and (7) results immediately in the identity

$$\alpha^\star\Big(\textstyle\sum_{j\in[p]} s_j K_j\Big) = \big(\mathbb{I}_n - X_s(\mathbb{I}_k/\gamma + X_s^\top X_s)^{-1} X_s\big)\, Y. \tag{18}$$

The same result can also be obtained by applying the matrix inversion lemma stated in Hager (1989) to the regularized kernel matrix by noting that the micro kernels K_j are rank one dyadic products. The main advantage of the previous formula is the fact that it merely requires the inverse of the much smaller capacitance matrix $C := \mathbb{I}_k/\gamma + X_s^\top X_s$ in S^k_{++} instead of the dense full rank regularized kernel matrix in S^n_{++}.

Using expression (18), both the regression loss function c and any of its subgradients can be evaluated using $\mathcal{O}(k^3 + nk)$ instead of $\mathcal{O}(n^3)$ floating point operations. When the number of samples n is significantly larger than k, the matrix inversion lemma provides a significant edge over a vanilla matrix inversion. We note that from a statistical perspective this always must be the case if there is any hope that sparse regression might yield statistically meaningful results.

Pseudocode implementing the ideas discussed in this section is provided in Algorithm 2.

Algorithm 2: Regression function and subgradients

input : $Y \in \mathrm{R}^n$, $X \in \mathrm{R}^{n\times p}$, $s \in \mathrm{S}_k^p$ and $\gamma \in \mathrm{R}_{++}$
output: $c \in \mathrm{R}_+$ and $\nabla c \in \mathrm{R}^p$
$\alpha^\star \leftarrow Y - X_s(\mathbb{I}_k/\gamma + X_s^\top X_s)^{-1} X_s Y$
$c \leftarrow \frac{1}{2} Y^\top \alpha^\star$
for j **in** $[p]$ **do**
  $\nabla c_j \leftarrow \frac{\gamma}{2}(X_j^\top \alpha^\star)^2$

3.3. *Dual warm starts.* Regardless of the initial selection s_1, the outer approximation Algorithm 1 will eventually return the optimal subset solution $s^\star$ to the sparse regression formulation in Theorem 1. Nevertheless, to improve computational speed in practice it is often desirable to start with a high-quality warm start rather than any arbitrary feasible point in S_k^p.

As already briefly hinted upon, a high-quality warm start can be obtained by solving the continuous relaxation (16). More specifically, we take as warm start s_1 to the outer approximation algorithm the solution to

$$s_1 \in \arg\max_{s\in \mathrm{S}_k^p} \ \textstyle\sum_{j\in[p]} s_j \cdot \alpha^{\star\top} K_j \alpha^\star, \tag{19}$$

where $\alpha^\star$ is optimal in (16). Note that the solution to problem (19) can be found explicitly as the vector indicating the k largest components of $(\alpha^{\star\top} K_1 \alpha^\star, \ldots, \alpha^{\star\top} K_p \alpha^\star)$. We finally remark that the `Lasso` or the solution found by the first order heuristic developed in Bertsimas, King and Mazumder (2016) could have been used equally well.

4. Scalability and phase transitions. To evaluate the effectiveness of the cutting plane algorithm developed in Section 3, we report its ability to recover the correct regressors as well as its running time. In this section, we present empirical evidence on two critically important observations. The first observation is that our cutting plane algorithm scales to provable optimality in seconds for large regression problems with n and p in the 100,000s. That is two orders of magnitude larger than the known exact sparse regressor methods in Bertsimas, King and Mazumder (2016) and takes away the main propelling justification for heuristic approaches for many regression instances in practice.

Theoretical considerations and empirical evidence suggests that the ability to recover the support of the correct regressor w_{true} from noisy data

using the `Lasso` heuristic experiences a phase transition. While it is theoretically understood that a similar phase transition must occur in case of exact sparse regression, due to a lack of scalable algorithms such a transition was never empirically reported. The scalable cutting plane algorithm developed in Section 3 offers us the means to do so however. We will show that exact regression is significantly better than convex heuristics such as `Lasso` in discovering all true relevant features ($A\% = 100$), while truly outperforming their ability to reject the obfuscating ones ($F\% = 0$).

We witnessed furthermore a phase transition not only in the ability to recover the correct regressors, but in the time it takes to solve the sparse regression problem as well. More importantly, we did not find any evidence that both transitions are distinct. Hence, the second critical observation we want to advance here is that the sparse regression problem has the property that as the number of samples increases the problem becomes easier in that the solution $w_0^\star$ perfectly recovers the true support, and our approach solves the problem extremely fast (in fact faster than `Lasso`), while for small number of samples, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution $w_0^\star$ does not coincide with the true support at all.

All algorithms in this document are implemented in `Julia` and executed on a standard `Intel(R) Xeon(R) CPU E5-2690 @ 2.90GHz` running `CentOS release 6.7`. All optimization was done with the help of the commercial mathematical optimization distribution `Gurobi version 6.5` interfaced through the `JuMP` package developed by Lubin and Dunning (2015).

4.1. *Data description.* Before we present the empirical results, we first describe the properties of the synthetic data which shall be used throughout this section. The input and response data are generated synthetically with the observations Y and input data X satisfying the linear relationship

$$Y = Xw_{\text{true}} + E.$$

The unobserved true regressor w_{true} has exactly k-nonzero components at indices selected uniformly without replacement from $[f]$. Likewise, the nonzero coefficients in w_{true} are drawn uniformly at random from the set $\{-1, +1\}$. The observation Y consists of the signal $S := Xw_{\text{true}}$ corrupted by the noise vector E. The noise components E_i for i in $[n]$ are drawn independent identically distributed (i.i.d.) from a normal distribution and scaled such that the signal-to-noise ratio equals

$$\sqrt{\text{SNR}} = \|S\|_2 / \|E\|_2 .$$

Evidently as the signal-to-noise ratio SNR increases, recovery of the unobserved true regressor w_{true} from the noisy observations can be done with higher precision.

We have yet to specify how the input matrix X is chosen. We assume here that the input data samples $X = (x_1, \ldots, x_n)$ are drawn from an i.i.d. source with Gaussian distribution; that is

$$x_i \sim N(0, \Sigma), \quad \forall i \in [n].$$

The variance matrix Σ will be parametrized by the correlation coefficient $\rho \in [0, 1)$ as $\Sigma(i, j) := \rho^{|i-j|}$ for all i and j in $[p]$. As the ρ tends to 1, the columns of the data matrix X become more alike which should impede the discovery of nonzero components of the true regressor w_{true} by obfuscating them with highly correlated look-a-likes. In the extreme case in which $\rho = 1$, all columns of X are the same at which point there is no hope of discovering the true regressor w_{true} even in the noiseless case.

4.2. *Scalability.* We provide evidence that the cutting plane Algorithm 1 represents a truly scalable algorithm to the exact sparse regression problem (1) for n and p in the 100,000s. As many practical regression problems are within reach of our exact cutting plane Algorithm 1, the need for convex surrogate regressors such as `Elastic Net` and `Lasso` is greatly diminished.

We note that an effective regression must find all relevant features ($A\% = 100$) while at the same time reject those that are irrelevant ($F\% = 0$). To separate both efforts, we assume in this and the following section that true number k of nonzero components of the ground truth w_{true} is known. In this case $A\% + F\% = 100$ which allows us to focus entirely on the the accuracy $A\%$ of the obtained regressors. Evidently, in most practical regression instances k needs to be inferred from the data as well. Incorrect determination of this number can indeed lead to high false alarm rates $F\%$. We will return to the issue of variable selection and false alarm rates in Section 4.4.

For the sake of comparison, we will also come to discuss the time it takes to solve the `Lasso` heuristic (2) as implemented by the `GLMNet` implementation of Friedman, Hastie and Tibshirani (2013). Contrary to exact sparse regression, no direct way exists to obtain a sparse regressor from solving the convex surrogate heuristic (2). In order to facilitate a fair comparison however, we shall take that `Lasso` regressor along a path of optimal solutions in (2) for varying λ which is the least regularized but has exactly k nonzero coefficients as a heuristic sparse solution.

In Table 1 we discuss the timing results for exact sparse linear regression as well as for the `Lasso` heuristic applied to noisy ($\sqrt{\text{SNR}} = 20$) and lightly

		Exact T [s]			`Lasso` T [s]		
		$n = 10$k	$n = 20$k	$n = 100$k	$n = 10$k	$n = 20$k	$n = 100$k
$k = 10$	$p = 50$k	21.2	34.4	310.4	69.5	140.1	431.3
	$p = 100$k	33.4	66.0	528.7	146.0	322.7	884.5
	$p = 200$k	61.5	114.9	NA	279.7	566.9	NA
$k = 20$	$p = 50$k	15.6	38.3	311.7	107.1	142.2	467.5
	$p = 100$k	29.2	62.7	525.0	216.7	332.5	988.0
	$p = 200$k	55.3	130.6	NA	353.3	649.8	NA
$k = 30$	$p = 50$k	31.4	52.0	306.4	99.4	220.2	475.5
	$p = 100$k	49.7	101.0	491.2	318.4	420.9	911.1
	$p = 200$k	81.4	185.2	NA	480.3	884.0	NA

TABLE 1

A comparison between exact sparse regression using our cutting plane algorithm and the `Lasso` *heuristic with respect to their solution time in seconds applied to noisy* ($\sqrt{\text{SNR}} = 20$) *and lightly correlated data* ($\rho = 0.1$) *explained by either* $k = 10$, $k = 20$ *or* $k = 30$ *relevant features. These problem instances are truly large scale as for the largest instance counting* $n = 100,000$ *samples for* $p = 200,000$ *regressors a memory exception was thrown when building the data matrices* Y *and* X*. Remarkably, even on this scale the cutting plane algorithm can be significantly faster than the* `Lasso` *heuristic.*

correlated ($\rho = 0.1$) synthetic data. We do not report the accuracy of the obtained solution as this specific data is in the regime where exact discovery of the support occurs for both the `Lasso` heuristic and exact sparse regression. We discuss phase transition phenomena in the subsequent section.

Remarkably, the timing results in Table 1 suggests that using an exact method does not impede our ability to obtain the solution fast. The problem instances displayed are truly large scale as indeed for the largest problem instance a memory exception was thrown when building the data matrices X and Y. In fact, even in this large scale setting our cutting plane algorithm can be significantly faster than the `Lasso` heuristic. Admittedly though, the `GLMNet` implementation returns an entire solution path for varying λ instead of a single regression model. Nevertheless, the results in Table 1 do refute the widely held belief that exact sparse regression is not feasible at large scales.

Although a hard theoretical picture is not yet available as for why the cutting plane Algorithm 1 proves so efficient, we hope that these encouraging results spur an interest in exact approaches towards sparse regression. In the subsequent section, we will come to see that the scalability of exact sparse regression entails more than meets the eye.

4.3. *Phase transition phenomena.* We have established that the cutting plane Algorithm 1 scales to provable optimality for problems with number of samples and regressor dimension in the 100,000s. Let us remark that for

the results presented in Table 1, both the exact and heuristic algorithms returned a sparse regressor with correct support and otherwise were of similar precision. In cases where the data does not allow a statistically meaningful recovery of the ground truth w_{true} an interesting phenomenon occurs. We present and discuss in this part of the paper two remarkable phase transition phenomena. The first will concern the statistical power of sparse regression, whereas the second will concern our ability to find the optimal sparse regressor efficiently. We will refer to the former transition as the accuracy transition, while referring to the latter as the complexity transition. We will argue here using strong empirical evidence that both transitions are in fact intimately related.

The accuracy phase transition describes the ability of the sparse regression formulation (1) to uncover the ground truth w_{true} from corrupted measurements alone. The corresponding phase transition for the `Lasso` has been extensively studied in the literature by amongst many others Bühlmann and van de Geer (2011); Hastie, Tibshirani and Wainwright (2015) and Wainwright (2009) and is considered well understood by now. As mentioned, in the noiseless case (SNR $\to \infty$) with uncorrelated input data ($\rho = 0$) a phase transition occurs at the curve (4). In the regime $n > 2k \log(p-k)$ exact recovery occurs with high-probability, whereas otherwise the probability for successful recovery drops to zero. A similar phase transition has been observed by Zheng et al. (2015) and Wang, Xu and Tang (2011) for exact sparse regression as well, although this transition is far less understood from a theoretical perspective than the similar transition for its heuristic counterpart. It is known though that the accuracy phase transition for exact sparse regression must occur even sooner than that of any heuristic approach. Empirical verification of this phase transition has been historically hindered due to the lack of exact scalable algorithms. With the help of our cutting plane algorithm we shall find direct empirical evidence of this transition.

In Figure 1, we show computational results for noiseless uncorrelated synthetically generated data counting $p = 20,000$ regressors of which only $k = 10$ are relevant. The accuracy $A\%$ of the obtained regressors using exact sparse regression as well as the `Lasso` heuristic and time T in seconds necessary to obtain either one are taken as the median values of fifteen independent synthetic datasets. Please note that when the optimal solution is not found in less than two minutes we take the best solution found up to that point. The error bars give an indication of the inter-sample variance among these fifteen independent experiments. The colored horizontal lines indicate that number of samples n after which either method returned more often than not a full recovery ($A\% = 100$) of the support of the ground

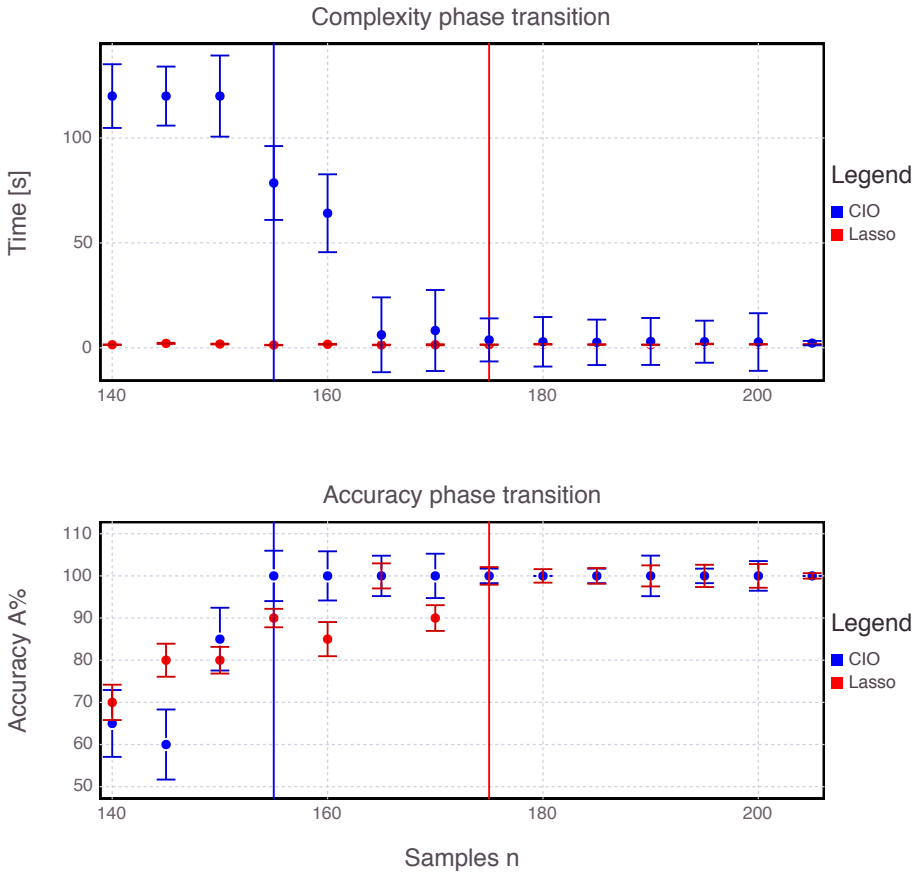

FIG 1. *A comparison between exact sparse regression using our cutting plane algorithm and the approximate* `Lasso` *heuristic noiseless* ($\mathrm{SNR} \to \infty$) *and uncorrelated data* ($\rho = 0$) *counting* $p = 20,000$ *regressors of which only* $k = 10$ *are relevant. In the figure above we depict the time in seconds necessary to solve the sparse regression problem using either method as a function of the number of samples. The figure below gives the corresponding accuracy* $A\%$ *of the regressors as a function of the number of samples. The red vertical line at* $n = 175$ *samples depicts the accuracy phase transition concerning the ability of the* `Lasso` *heuristic to recover the support of the ground truth* w_{true}. *The blue vertical line at* $n = 155$ *does the same for exact sparse regression. It can thus be seen that exact sparse regression does yields more statistically meaningful regressors than the* `Lasso` *heuristic. Furthermore, a complexity phase transition can be recognized as well. Most remarkably, we did not find any evidence that the accuracy and complexity phase transitions are distinct. Either exact sparse regression is easy and its optimal sparse regressor statistically meaningful, or very hard and unreliable.*

truth. The `Lasso` heuristic is empirically found to require $n = 175$ samples to recover the true support completely which corresponds rather well with the theoretically predicted $n = 198$ necessary samples by Wainwright (2009). Unsurprisingly, the accuracy phase transition of exact sparse regression using Algorithm 1 is found empirically to occur even earlier at $n = 155$ samples.

We now discuss the second transition which indicates that the time it takes to solve the sparse regression (1) using the cutting plane Algorithm 1 experiences a phase transition as well. We seem to be the first to have seen this complexity phase transition likely due to the fact that scalable algorithms for exact sparse regression have historically been lacking. Most importantly, in Figure 1 we do not find any evidence that the transition from accurate to unreliable discovery and from fast to slow running times are distinct. The accuracy and complexity phase transitions seem to occur simultaneously. In other words, there is a phase transition in our ability to recover the true coefficients of the sparse regression problem and most surprisingly in our ability to solve it. Contrary to traditional complexity theory which suggests that the difficulty of a problem increases with problem size, the sparse regression problem has the property that as the number of samples n increases the problem becomes easier in that the solution recovers 100% of the true signal, and our approach solves the problem extremely fast (in fact faster than `Lasso`), while for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal.

To investigate the effect of each of the parameters discussed in this section even further, we use synthetic data with the properties presented in Table 2. In order to be able to separate the effect of each parameter individually, we present the accuracy $A\%$ and solution time T of our cutting plane algorithm as a function of the number of samples n for each parameter value separately while keeping all other parameters fixed to their nominal value. Again all results are obtained as the median values of fifteen independent experiments. The figures in the remainder of this section indicate that the accuracy and complexity phase transitions persist for a wide variety of properties of the synthetic data and occur simultaneously.

Feature dimension p. As the phase transition curve (4) described by Wainwright (2009) for the statistical power of the `Lasso` heuristic depends only logarithmically on the regressor dimension, we do not expect the transition curve of exact sparse regression to be very sensitive to the regressor dimension either. Indeed, in Figure 2 only a minor influence on the point of

Synthetic data	Parameter	Value
Sparsity	k	$\{10, 15, 20^\star\}$
Dimension	p	$\{5000, 10000, 20000^\star\}$
Signal-to-noise ratio	$\sqrt{\text{SNR}}$	$\{3, 7, 20^\star\}$
Correlation	ρ	$\{0.1^\star, 0.9\}$

TABLE 2
Parameters describing the synthetic data used in Section 4.3. The starred values denote the nominal values of each parameter.

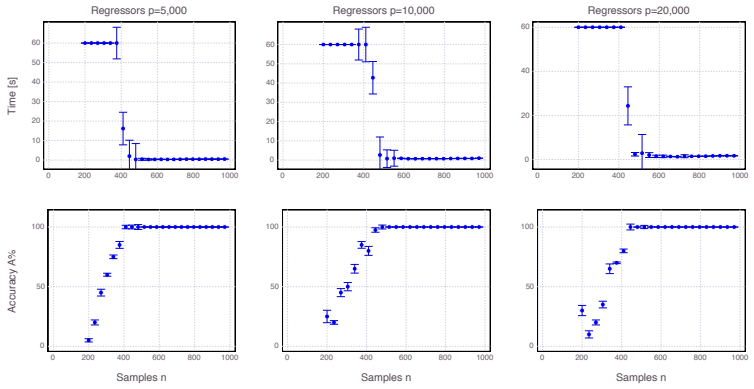

FIG 2. *Each figure in the top row shows the time it takes to solve the sparse regression problem in seconds using the cutting plane method for data with $p = 5,000$, $p = 10,000$ or $p = 20,000$ regressors and the nominal properties given in Table 2 as a function of the number of samples. Note that when the optimal solution is not found in less than one minute we take the best solution found up to that point. Each figure in the bottom row shows the accuracy A% of the found regressors. Only a minor influence on the point of transition between statistically meaningful and efficient sparse regression to unreliable and intractable regression is observed as a function of the regression dimension p.*

transition between statistically meaningful and efficient sparse regression to unreliable and intractable regressors is observed as a function of the regression dimension p.

This mild dependence is what ultimately grants sparse regression extraordinary statistical power. Not the feature dimension p determines the statistical power of the method as in ordinary regression, but the sparsity parameter k which we discuss next. We will use this observation in Section 5 to justify a nonlinear extension to standard linear regression.

Sparsity level k. The sparsity level k is an important parameter that in practice must be inferred from the data. Indeed, choosing k too large

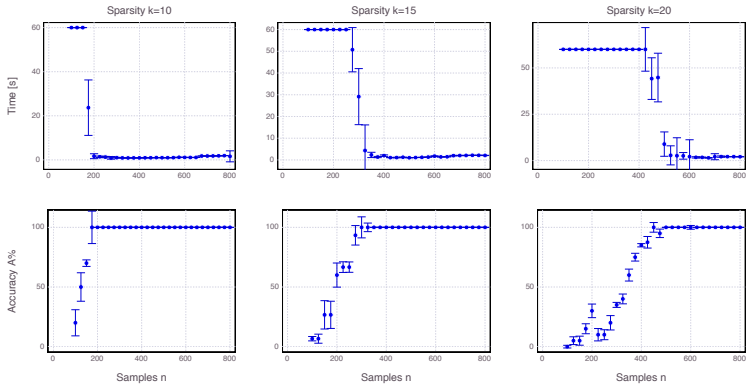

FIG 3. *Each figure in the top row shows the time it takes as a function of the number of samples to solve the sparse regression problem in seconds using the cutting plane method for data with* $p = 20,000$ *regressors of which only* $k = 10$, $k = 15$ *or* $k = 20$ *are relevant and further nominal properties as given in Table 2. Note that when the optimal solution is not found in less than one minute we take the best solution found up to that point. Each figure below shows the accuracy* $A\%$ *of the found regressors. These results suggest that the quantity* n/k *is a major factor in the phase transition curve of exact sparse regression.*

can lead to overfitting whereas choosing it too small can result in biased regressors. Figure 3 suggests that k has an important influence of the phase transition curve. The experiments suggest that there is a threshold n_0 such that if $n/k \geq n_0$, then full support recovery $A\% = 100$ occurs and the time to solve problem (1) is in the order of seconds and only grows linear in n. Furthermore, if $n/k < n_0$, then support recovery $A\%$ drops to zero while the time to solve problem (1) grows combinatorially as $\binom{p}{k}$.

As mentioned, we expect the threshold n_0 denoting the ratio between samples n and effective regressors k necessary for meaningful regression to depend only logarithmically on the regressor dimension. We do point out in the subsequent discussion that the signal-to-noise ratio (SNR) has an important influence too.

*Signal-to-noise ratio (*SNR*).* From an information theoretic point of view, the SNR must play an important role as well. Indeed, the statistical power of any method is questionable when the noise exceeds the signal in the data. In Figure 4 this effect of noise is clearly observed as for noisy data the phase transition occurs later than for more accurate data.

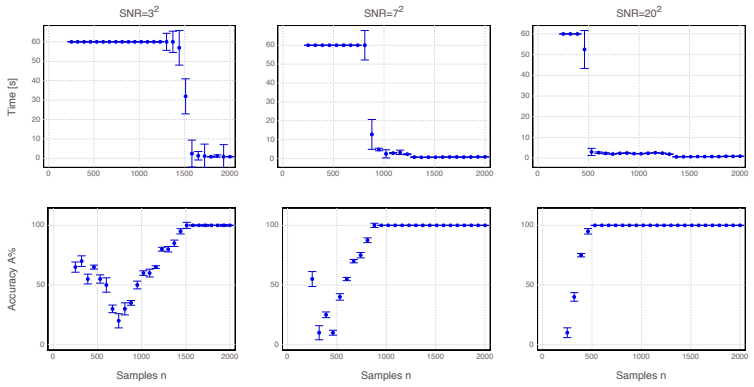

FIG 4. *Each figure in the top row shows the time it takes as a function of the number of samples to solve the sparse regression problem in seconds using the cutting plane method for data with signal-to-noise level* $\sqrt{\mathrm{SNR}} = 3$, $\sqrt{\mathrm{SNR}} = 7$ *or* $\sqrt{\mathrm{SNR}} = 20$ *and further nominal properties as given in Table 2. Note that when the optimal solution is not found in less than one minute we take the best solution found up to that point. Each figure in the bottom row shows the accuracy* $A\%$ *of the found regressors. As can be expected from an information theoretic point of view, meaningful regression is harder when the data is noisy.*

4.4. *Variable selection.* In all the experiments conducted up to this point, we assumed that the number of non-zero regressor coefficients k of the ground truth w_{true} underlying the data was given. Evidently, in most practical applications the sparsity parameter k needs to be inferred from the data as well. In essence thus, any practical sparse regression procedure must pick those regressors contributing to the response out of the obfuscating bulk. To that end, we introduced the false alarm rate $F\%$ of a certain solution $w^\star$ as the percentage of regressors selected which are in fact unfitting. The ideal method would of course find all contributing regressors ($A\% = 100$) and not select any further ones ($F\% = 0$). In practice clearly, a trade-off must sometimes be made. We shall study here exact sparse regression from this variable selection perspective and will come to find that it significantly outperforms the `Lasso` heuristic here as well.

Historically, cross validation has been empirically found to be an effective way to infer the sparsity parameter k from data. Hence, for both exact sparse regression and the `Lasso` heuristic, we select that number of non-zero coefficients which generalizes best to the validation sets constructed using ten fold cross validation with regards to prediction performance. In case of exact sparse regression, we let k range between ten and thirty whereas the

	Exact $F\%$		`Lasso` $F\%$	
	$\rho = 0.1$	$\rho = 0.80$	$\rho = 0.1$	$\rho = 0.80$
$\sqrt{\text{SNR}} = 10$	4.8	9.9	86.2	84.0
$\sqrt{\text{SNR}} = 10$	0	0	85.2	89.9
$\sqrt{\text{SNR}} = 10$	0	0	71.8	80.6
$\sqrt{\text{SNR}} = 10$	0	0	0	0

TABLE 3

False alarm rates of exact versus heuristic sparse regression for $n = 1,000$ data samples and $p = 20,000$ regressors of which only $k = 20$ are relevant. The `Lasso` heuristic has difficulty keeping a low false alarm rate $F\%$ with noisy data. Exact sparse regression does yield sparser models as it avoids including regressors that do not contribute to the observations. Although increased correlation does increase the false alarm rate, its effect is observed as marginal.

true unknown number of non-zero regressors was in fact twenty.

We consider now $n = 1,000$ data samples for $p = 20,000$ regressors of which only $k = 20$ are relevant. Both exact and heuristic sparse regression are in this configuration perfectly accurate ($A\% = 100$) as all contributing regressors are found by either method. The false alarm rates tell a different story though. In Table 3 we present the false alarm rates for synthetic data as a function of the signal-to-noise ratio SNR and the correlation coefficient ρ. The values stated are the median of seven independent experiments. As can be seen, the `Lasso` heuristic has difficulty keeping a low false alarm rate with noisy data. Exact sparse regression does indeed yield sparser models as it avoids including regressors that do not contribute to the observations. Although increased correlation does elevate the false alarm rate, its effect is deemed marginal over a fairly large range from lightly to heavily correlated input data.

4.5. *A remark on complexity.* The empirical results in this paper suggest that the traditional complexity point of view might be misleading towards a better understanding of the complexity of the sparse regression problem (1). Indeed, contrary to traditional complexity theory which suggests that the difficulty of a problem increases with dimension, the sparse regression problem (1) has the property that for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal. However, for a large number of samples n, our approach solves the problem extremely fast and recovers 100% of the support of the true regressor w_{true}.

5. The road towards nonlinear feature discovery. The most striking property of sparse regression is surely its defiance to the curse of dimen-

Transformations $x \mapsto$	x	$\sqrt{\lvert x \rvert}$	$\log \lvert x \rvert$	x^2	x^3	$\cos(10\pi x)$	$\sin(x)$	$\tanh(2x)$

TABLE 4

Nonlinear transformations considered in the nonlinear regression example discussed in Section 5. The first transformation is included to ensure the method is more powerful than sparse linear regression. Evidently, the transformations determine to which nonlinearities the regression method (20) is sensitive.

sionality. The curse of dimensionality first coined by Bellman describes here the phenomenon that classical regression methods in the face of many regressors tend to have extremely limited statistical power. Put in terms data collection, an extremely large quantity of samples needs to be accumulated in order to aspire any hope of obtaining a significant regressor. Even in the age of Big Data a pretty grim outlook. Modern sparse regression methods miraculously exploit the folklore that in real life only a few things really matter. As indeed empirically observed in Figure 2 as well as theoretically suggested by Wainwright (2009), both the computational complexity and the statistical power of exact sparse regression depends only very mildly on the number of regressors p. This defiance of the curse of dimensionality is arguably what renders sparse regression so powerful in an era of Big Data. We briefly illustrate the power of this observation and the efficacy of our devised cutting plane method by considering a nonlinear extension to the sparse regression problem discussed.

Arbitrary nonlinear regressors can be obtained rather painlessly by augmenting the input data X with potentially a great amount of auxiliary nonlinear transformations. We have certainly not been the first to make this powerful observation. The idea of nonlinear regression as linear regression to lifted data is what underpins kernel methods as popularized by Vapnik (1998b) as well. Kernel methods can in a primal perspective be viewed as Tikhonov regularization between the observations Y and transformed versions $\psi(x_i)$ of the original data samples. The feature map $\psi(\cdot)$ encodes which nonlinearities should be detected and is ultimately a design parameter. Unfortunately, kernel regression inherits the curse of dimensionality through its application of standard regression to the lifted data. Here will illustrate that this deficiency can be overcome by considering sparse regression instead.

We explore here this idea of nonlinear regression with sparsity as an illustration of the methods developed in this work. We augment each of the p original regressors with the nonlinear transformations given in Table 4. The method could be made even more general by allowing for nonlinear products between variables but we abstain from doing so for the sake of simplicity. To enforce a sparse regression model, we demand that the final regressor can

only depend on k different (potentially nonlinear) features.

Instead of solving the sparse regression problem (1), we then solve its nonlinear version

$$\begin{array}{ll} \min & \frac{1}{2\gamma}\|\tilde{w}\|_2^2 + \frac{1}{2}\|Y - \psi(X)\tilde{w}\|_2^2 \\ \text{s.t.} & \|\tilde{w}\|_0 \leq k, \end{array} \tag{20}$$

where the matrix $\psi(X)$ in $\mathrm{R}^{n\times f}$ consists of the application of the transformations in Table 4 to the input matrix X. The nonlinear sparse regression problem (20) can be dealt with in an identical manner as its linear counterpart (1). Notice that the dimension of the nonlinear regressor $\tilde{w}$ is potentially much larger than its linear counterpart w. The resilience of the power of exact sparse regression to the regressor dimension is thus of great importance for the nonlinear regression method to be reliable.

COROLLARY 1 (Sparse nonlinear regression). *The sparse regression problem* (20) *can be reformulated as the nonlinear optimization problem*

$$\begin{array}{ll} \min & \frac{1}{2}Y^\top \left(\mathbb{I}_n + \gamma \sum_{j\in[f]} s_j K_j\right)^{-1} Y \\ \text{s.t.} & s \in \mathrm{S}_k^f, \end{array}$$

where the micro kernel matrices K_j *in* S_+^n *are defined as the dyadic products*

$$K_j := \psi_j(X)\psi_j(X)^\top.$$

The only material difference between the pure integer reformulation for the nonlinear sparse regression problem found in Corollary 1 to its linear counterpart Theorem 1 is the definition of the micro kernel matrices K_j.

As an illustration of the nonlinear approach described above, consider observations and data coming from the following nonlinear model

$$Y = 3\sqrt{|X_4|} - 2X_2^2 + 4\tanh(2X_3) + 3\cos(2\pi X_2) - 2X_1 + aX_1X_2 + E. \tag{21}$$

We assume again that the input data X and noise E is generated using the method outlined in Section 4.1. That is, the signal-to-noise ratio was chosen to be $\sqrt{\mathrm{SNR}} = 20$ to simulate the effect of noisy data. For the sake of simplicity we assume the original data X to be uncorrelated ($\rho = 0$). It should be remarked, however, that after lifting the input data $\psi(X)$ might be correlated making exact recovery more difficult. An additional 16 regressors are added to obfuscate the four relevant regressors in the nonlinear model (21). The input data after the nonlinear transformations in Table 4

	Quality $w^\star$	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
$a = 0$	$A\%$	100	100	100	100	100
	$F\%$	38	0	0	0	0
$a = 1$	$A\%$	80	100	100	100	100
	$F\%$	50	17	17	28	17

TABLE 5

Nonlinear regression results on nonlinear data from the nonlinear model (21). In the unbiased case ($a = 0$), a mere 200 samples suffice to identify the correct nonlinearities and features. With a sufficient amount of samples, we do eventually discover all detectable nonlinearities also in the biased case ($a = 1$) but nevertheless suffer a nonzero false alarm rate $F\%$.

comprised a total of $f = 160$ nonlinear features. We consider two distinct nonlinear models for corresponding parameter values $a = 0$ and $a = 1$. Notice that for the biased case $a = 1$, the term aX_1X_2 will prevent our nonlinear regression approach to find the true underlying nonlinear model (21) exactly.

We state the results of our nonlinear regression approach applied to the nonlinear model (21) for both $a = 0$ and $a = 1$ in Table 5. All reported results are the median values of five independent experiments. Cross validation on k ranging between one and ten was used to determine the number of regressors considered. Determining the best regressor for each k took around ten seconds, thus making a complete regression possible in a little under two minutes. As currently outlined though, our nonlinear regression approach is not sensitive to nonlinearities appearing as feature products and consequently it will treat the term aX_1X_2 as noise. Hence, the number of underlying regressors we can ever hope to discover is five. In the unbiased case ($a = 0$), a mere 200 samples suffice to identify the correct nonlinearities and features. Evidently, we do expect the quality of the regressors in the biased case ($a = 1$) to suffer. Although the results in Table 5 do indeed report an increased false alarm rate compared to the unbiased case, the quality of the biased regressors seem to deteriorate gracefully. Given a sufficient amount of samples, we do eventually discover all detectable nonlinearities $A\% = 100$ but nevertheless suffer a nonzero false alarm rate $F\%$.

Evidently, the method proposed here serves only as an illustration. Much more nonlinearities should be considered as well as products between variables eliminating the bias phenomenon discussed in our illustration. It must also be said however that no method can aspire to discover arbitrary nonlinearities without sacrificing its statistical power. Hence in practice, a tradeoff between the types of nonlinearities for which the method is sensitive and its statistical power much be bore in mind. Despite the previous issues, we

believe that this constitutes a promising new road towards nonlinear feature discovery in data. With additional research, we believe that it can become a fierce and more disciplined competitor towards the more "black box" approaches such as neural networks.

6. Conclusions. In this paper, we tried to do away with the notion that exact sparse regression is impractical for only but the smallest of problem instances. We did so by presenting a novel binary convex reformulation of the sparse regression problem that constitutes a new duality perspective. Our novel cutting plane algorithm can solve to provable optimality exact sparse regression problems for instances with sample sizes and regressor dimensions well in the 100,000s. This by itself presents an improvement of two orders of magnitude compared to known exact sparse regression approaches. In fact, it can be argued that the improvement takes away the computational edge attributed to sparse regression heuristics such as the `Lasso` or `Elastic Net`.

The ability to solve sparse regression problems for very high dimensions allows us to observe new phase transition phenomena. Contrary to traditional complexity theory which suggests that the difficulty of a problem increases with problem size, the sparse regression problem is shown to possess the property that as the number of samples n increases, the problem becomes easier in that the solution perfectly the support of the true signal, and our approach solves the problem extremely fast (in fact faster than `Lasso`), whereas for small number of samples n, our approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal. We further provide preliminary evidence that the results described in this paper open a new road towards nonlinear feature discovery based on sparse selection from a potentially huge amount of desired nonlinearities.

References.

Bertsimas, D. and Fertis, A. (2009). On the equivalence of robust optimization and regularization in statistics Technical Report, Massachusetts Institute of Technology. Working paper.

Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics* **44** 813-852.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2014). Julia: A fresh approach to numerical computing. *arXiv preprint arXiv:1411.1607*.

Bixby, R. E. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica* 107–121.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Donoho, D. and Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. In *International Joint Conference on Neural Networks* 1916–1921. IEEE.

Donoho, D. and Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **367** 4273–4293.

Duran, M. A. and Grossmann, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming* **36** 307–339.

Friedman, J., Hastie, T. and Tibshirani, R. (2013). GLMNet: Lasso and elastic-net regularized generalized linear models. R package version 1.9–5.

Furnival, G. M. and Wilson, R. W. (2000). Regressions by leaps and bounds. *Technometrics* **42** 69–79.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review* **31** 221–239.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press.

Lubin, M. and Dunning, I. (2015). Computing in Operations Research Using Julia. *INFORMS Journal on Computing* **27** 238-248.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41** 3397–3415.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics* **8** 171–176.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B* **58** 267-288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR* **39** 195–198.

Vapnik, V. (1998a). The support vector method of function estimation. In *Nonlinear Modeling* 55–85. Springer.

Vapnik, V. N. (1998b). *Statistical Learning Theory* **1**. Wiley.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55** 2183–2202.

Wang, M., Xu, W. and Tang, A. (2011). On the performance of sparse recovery via ℓ_p-minimization ($0 \leq p \leq 1$). *IEEE Transactions on Information Theory* **57** 7255–7278.

Xu, H., Caramanis, C. and Mannor, S. (2009). Robustness and regularization of support vector machines. *The Journal of Machine Learning Research* **10** 1485–1510.

Zhang, F. (2006). *The Schur Complement and Its Applications* **4**. Springer Science & Business Media.

Zheng, L., Maleki, A., Wang, X. and Long, T. (2015). Does ℓ_p-minimization outperform ℓ_1-minimization? *arXiv*. arXiv:1501.03704.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society* **67** 301–320.

Sloan School of Management
Cambridge, MA 02139, USA
E-mail: dbertsim@mit.edu
URL: http://www.mit.edu/~dbertsim

Operations Research Center
Cambridge, MA 02139, USA
E-mail: vanparys@mit.edu
URL: http://www.mit.edu/~vanparys