Experimental Design on a Budget for Sparse Linear Models and Applications

Sathya N. Ravi[†] RAVI5@WISC.EDU
Vamsi K. Ithapu[†] ITHAPU@WISC.EDU
Sterling C. Johnson[§][†] SCJ@MEDICINE.WISC.EDU
Vikas Singh[†] VSINGH@BIOSTAT.WISC.EDU
[†]University of Wisconsin Madison [§][†]William S. Middleton Memorial Veterans Hospital

Abstract

Budget constrained optimal design of experiments is a well studied problem. Although the literature is very mature, not many strategies are available when these design problems appear in the context of sparse linear models commonly encountered in high dimensional machine learning. In this work, we study this budget constrained design where the underlying regression model involves a ℓ_1-regularized linear function. We propose two novel strategies: the first is motivated geometrically whereas the second is algebraic in nature. We obtain tractable algorithms for this problem which also hold for a more general class of sparse linear models. We perform a detailed set of experiments, on benchmarks and a large neuroimaging study, showing that the proposed models are effective in practice. The latter experiment suggests that these ideas may play a small role in informing enrollment strategies for similar scientific studies in the future.

1. Introduction

Experimental Design (ED) is a problem with deep foundations dating back at least to the early 1900s (Kempthorne, 1952; Kirk, 1982). Here, given the covariates x_i's, an *experimenter* must conduct an *experiment* in order to obtain the value of the dependent (or response) variables y_i's. The focus of much of the classical work on this topic is to maximize the amount of information that the full experiment yields for a given (or least) amount of work. In many situations, each experiment or measurement may have a monetary cost associated with it. Given a fixed budget $B \in \mathbb{R}_+$, the budgeted version of the experimental design problem is to choose the set of x_i's for which we will conduct an experiment to obtain the corresponding y_i's, while satisfying the budget constraint. As a simple example, the budget constraint may restrict the number of y_i's we may request, i.e., the experimenter can perform at most B experiments. As before, the selected subset must be a good prototype for the entire dataset for model estimation purposes, e.g., $\beta : x_i \to y_i$. These problems are studied under the umbrella of optimal design (Pukelsheim, 1993).

In recent years, there is a renewed interest in ED since it provides a framework to study numerous adaptive data acquisition scenarios. While randomization offers one solution (Recht et al., 2011), recent results demonstrate that optimization based schemes yield a competitive alternative (Bertsimas et al., 2015). Solutions to a number of interesting variants of the problem have been proposed, for instance, (Horel et al., 2014) assumes that the "cost" of a response y_i depends on participant i and that they may lie about these costs and develops a *mechanism*. In vision, the availability of crowd-sourced platforms has led to scenarios where we seek to acquire low cost reliable data; trusted workers charge a premium per HIT (Li & Guo, 2013).

Apart from these results, active learning approaches also make use of ED concepts, but the selection process there is sequential. An important distinction in active learning is that the algorithm chooses the next (subset of) examples which the back-end machine learning algorithm needs labeled. The algorithm then proceeds and request labels for more examples iteratively. Most ED formulations do *not* offer this flexibility. Specifically, while these methods may try to minimize the number of queries or labelings required (for a pre-determined accuracy), we study a related but distinct question. If we fix the number of queries a priori, we study the issue of choosing a subset such that the corresponding estimator is close to the estimator inferred from the full dataset (i.e., y_i for *all* x_i were available). Further, the algorithm gets *no more chances to query the experimenter*. For this formulation to make sense, the criteria for subset selection must be closely tied to the later statistical estimation task that the subset of samples will be used for.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

To our knowledge, existing solutions to the general version of this problem do not scale up to large datasets (n large), see (Dette et al., 2011). Moreover, often we need to use sampling schemes to approximate an integral at each step which is a *nontrivial* for large p due to the *high dimensionality* (Konstantinou et al., 2011).

Application. One motivating application is conducting cost effective imaging-based neuroscience studies; this setting will be used to evaluate our proposed models. In Alzheimer's Disease (AD), the problem of *predicting* future cognitive decline is important (Landau et al., 2010; Hinrichs et al., 2011). Identifying decliners is of direct relevance in disease progression studies and also serves the goal of maximizing statistical power in trials (Ithapu et al., 2015) – both of which are based on longitudinal changes in participants (Searcey et al., 1994; Hinrichs et al., 2012). The dependent variable of interest here is *change* in cognition and/or diagnostic scores, over time. The independent variables include imaging (and imaging-derived) measures, genetic and other data acquired at the *baseline* time-point (or initial visit). Here, keeping a subject enrolled in the study (e.g., for a second visit) is expensive, but will provide the response y_i for subject i. The goal is to choose a "subset" of all subjects that can help estimate the parameters of the model, without affecting the statistical power.

Our **contributions** include : **(a)** We give two formulations for the problem. The first model is motivated geometrically while the second one involves certain algebraic manipulations. Experimentally we show that both models yield consistent results, with each other as well as with the "full" model. **(b)** We evaluate our algorithms on a large neuroimaging dataset (≈ 1000 subjects) using both qualitative and quantitative performance measures. Empirical results show that our algorithms are robust and promising for formulations involving sparse linear models.

2. Preliminaries

Consider the linear model $y_i = x_i^T\beta + \epsilon$ where $x_i, \beta \in \mathbb{R}^p, y_i \in \mathbb{R}$ and $\epsilon \sim \mathcal{N}(0,1)$. The regression task for β is,

$$\beta^* := \arg\min_{\beta} \frac{1}{2}||X\beta - y||_u^v + \epsilon g(\beta) \tag{1}$$

where the rows of $X \in \mathbb{R}^{n\times p}$ correspond to samples (or data instances, subjects). Here, g is a penalty function that specifies desired properties or characteristics of the optimal regressor β^* and ϵ is the Tikhonov regularization parameter. We assume that $u = v = 2$ unless otherwise specified which corresponds to the standard linear regression loss function. Recall that when $g(\beta) = \beta^T M\beta$ for some $M \succ 0$, then β^* has a closed form solution $\beta^* = \left(X^T X + \epsilon M\right)^{-1} X^T y$ also known as *ridge* regression. Ridge regression is particularly useful when $p > n$ because $X^T X$ is singular or when the covariates are highly correlated, where a typical regressor may overfit rather than explaining the underlying process. So, the ability to adjust the regularizer enables the estimation process. There are some obvious choices for M, e.g., $M = I$ corresponds to the least norm least squares solution. On the other hand, when $p < n$ and if the rank of X is p, (Li, 2008) shows that the ridge regression is robust to noise or outliers.

Ridge Regression ED: The ED problem for ridge regression can be written as the following (combinatorial) problem,

$$S^* := \arg\max_{|S|\le B} f\left(\sum_{i\in\mathcal{S}} x_i x_i^T + \epsilon I\right)$$

where we identify $\mathcal{S}$ with the set of selected subjects for a budget B. This problem can be equivalently formulated as,

$$S^* = \arg\max_{\mu\in\{0,1\}^n} f\left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right) \text{ s.t. } \mathbf{1}^T\mu \le B \tag{2}$$

where $\mathbf{1}\in \mathbb{R}^n$ is the vector of all 1s. The choice of f determines the nature of the regressor from the selected subset, for example, $f(\cdot) = \log\det(\cdot)$ is referred to as the $D-$optimality criterion. Intuitively, a $D-$optimal design corresponds to the set of subjects that maximizes the information gain. There are other choices for the objective, see (Das; Pukelsheim, 1993; Chaloner & Verdinelli, 1995). A common feature of many optimality criteria is that they lead to convex problems when the integrality constraints are relaxed. There are efficient ways to solve this relaxed problem — in particular, when Frank Wolfe type methods are employed, the number of nonzero entries of μ has a relationship with the number of iterations (Jaggi, 2013). Once the relaxed problem is solved, pipage rounding schemes yield a solution without sacrificing the objective function value much (Ageev & Sviridenko, 2004).

The case for ℓ_1: While ridge regression has many attractive properties, the solutions from ridge regression may have many nonzero entries that are close to zero (Tibshirani, 1996). Recent results suggest that in various cases it may be more appropriate to use the ℓ_1-norm instead – it induces sparsity in the optimal regressor and hence the model or the regressor may be more interpretable (Candes & Tao, 2005; Candès & Plan, 2009). When the ℓ_1-regularization is used, coordinates with nonzero entries in the optimal regressor correspond to features that are responsible in the "linearity" of the model, and explains the "selection" aspect of LASSO. After this procedure, a ridge regression problem is solved only on this reduced set of features that were selected by LASSO as described in (Tibshirani, 1996). The problem of interest is $\beta_1^* \in \arg\min_\beta \frac{1}{2}||X\beta - y||_2^2 + \epsilon||\beta||_1$. Under some mild conditions, we can assume that β_1^* is unique and so replace the containment operator with equality. Unlike ridge regression, iterative procedures are needed here since β_1^* does not have a closed form expression.

3. Our proposed formulations

Some basic assumptions. We first clarify a few basic assumptions to simplify the presentation. If necessary, we will add ϵI to the covariance matrix so that the corresponding inverse and the $\log\det(\cdot)$ operations are meaningful. We also assume without loss of generality that $||x_i||_2 \leq 1$, in other words, X is divided by the maximum sample norm. The constraint $\mathbf{1}^T\mu \leq B$ can be replaced by a more general $d^T\mu \leq B$ for $d > 0$, i.e., where the cost of selecting different subjects is different. Finally, we fix $f(\cdot)$ to be $\log\det(\cdot)$ (corresponding to the $D-$optimality criterion) since it is conceptually simpler. Our algorithms remain unchanged if f is replaced by another smooth convex function.

We now describe the two models for the ED problem (for LASSO): the first formulation in Sec. 3.1 is motivated geometrically whereas the next one in Sec. 3.2 involves certain algebraic manipulations but offers some efficiency benefits.

3.1. ED-S: Spectral Experimental Design

The ED-S approach is driven by a simple geometric interpretation of the LASSO. Consider the following two equivalent formulations of ridge regression and LASSO,

$$\begin{aligned}\beta^* &= \arg\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 + \lambda||\beta||_2^2 \quad \text{(RIDGE)}\\ &\equiv \arg\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 \text{ s.t. } ||\beta||_2^2 \leq \tau\end{aligned} \tag{3}$$

$$\begin{aligned}\beta_1^* &= \arg\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 + \lambda||\beta||_1 \quad \text{(LASSO)}\\ &\equiv \arg\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 \text{ s.t. } ||\beta||_1 \leq \tau_1\end{aligned} \tag{4}$$

for positive scalars τ and τ_1. The optimal solution in both the cases is where the objective function (identical for both problems) touches the feasible set: the ℓ_1 (and ℓ_2 norm) balls respectively. The difference between the problems is that the $\{\beta : ||\beta||_1 \leq \tau_1\}$ is polyhedral (compact) and so has a finite number of extreme points given by $\{\pm e_i\}$, where e_i is the standard basis vectors in appropriate dimensions, see Fig. 3.1. In (4), the objective function is likely to touch a vertex of the ℓ_1 ball, and so yields sparse solutions.

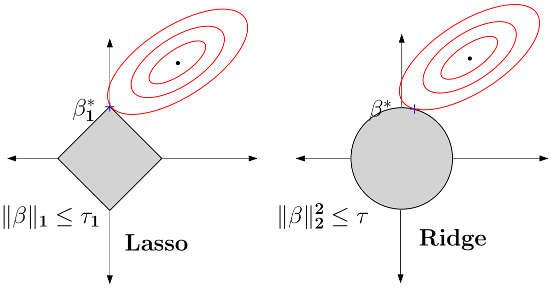


Figure 1. Variable selection of Lasso estimate β_1^*

Our proposal is motivated by the following simple observation: consider the full setting where we have access to the entire set of y_i's and the reduced setup where y_i's are available only for a subset. Intuitively, if the objective function of the full and the reduced setups behave similarly, then the corresponding regression coefficients will also be similar. To obtain this desired property, we may ask that the *reduced objective function to have approximately the same curvature as the full one*. Recall that the Hessian carries most of the curvature information and the optimal value. That is, eigenvalues of the Hessian are called "principal curvatures" in differential geometry and play a critical role in analyzing first order methods (Kingma & Ba, 2014). So, ED-S may offer strong guarantees directly if, the spectrum of full and reduced sets are the same; then, the iterates generated (and optimal solutions) will be similar. To that end, we require that the eigen vectors of the Hessian to be close to each other which may be accomplished by making sure that the geometry of the reduced setup is preserved relative to the full one. The unknowns, μ, will correspond to the selection of samples for the reduced setup. Succinctly, the ED problem can be formulated as ($\lambda \geq 0$),

$$\begin{aligned}&\max_{\mu\in\{0,1\}^n} \log\det\left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right) + \lambda R_\Gamma(\Gamma_\mu)\\ &\qquad \text{s.t. } \mathbf{1}^T\mu \leq B\end{aligned} \tag{5}$$

where Γ contains the eigen vectors of the entire X^TX and Γ_μ represents the eigen vectors of the chosen subjects given by μ. $R_\Gamma(\cdot)$ is a (smooth) concave function that encourages similarity between the eigenvectors of the full Hessian and the reduced Hessian. While the $\log\det$ term captures the linearity ($D-$optimal) in the regression problem, R_Γ captures the geometry. That is, the $\log\det$ term corresponds to D-optimality for linear regression as mentioned earlier.

We can now proceed to write the explicit formulation of the problem. For simplicity, we only promote similarity between the top eigen vector between the Hessians noting that the top $k-$eigen vectors case ($k \leq p$) is easy as well. Let γ be the eigenvector corresponding to the largest eigenvalue and u be the decision variable for the largest eigenvector of the reduced Hessian. With this notation, taking $R(\cdot)$ to be the squared loss and ($\lambda \geq 0$), we seek to solve,

$$\begin{aligned}&\min_{\mu,u} \log\det\left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right)^{-1} + \lambda||\gamma - u||_2^2\\ &\text{s.t. } 0 \leq \mu \leq 1, \mathbf{1}^T\mu \leq B,\\ &u \in \arg\max_v \left\{ v^T\left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right) v \text{ s.t. } v^Tv = 1\right\}\end{aligned} \tag{6}$$

3.1.1. Algorithm

We will now use some simple manipulations to obtain an equivalent formulation of the above problem for which efficient algorithms can be designed.

The largest eigenvector of a symmetric positive definite matrix can be written as an optimization problem,

$$\arg \max_{u:||u||_2^2=1} u^T M u = \arg \min_{u:||u||_2^2=1} u^T M^{-1} u$$

for a given symmetric positive definite matrix M. Therefore, our formulation can be written as,

$$\begin{aligned} \min_{\mu,u} \log\det & \left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right)^{-1} \qquad (7) \\ & + \lambda||\gamma - u||_2^2 + u^T \left(\sum_{i=1}^n \mu_i x_i x_i^T + \epsilon I\right)^{-1} u \\ & \text{s.t. } 0 \le \mu \le 1, \quad \mathbf{1}^T\mu \le B, \quad ||u||_2^2 = 1 \end{aligned}$$

The above problem is nonconvex because of the squared norm constraint. But we note two important aspects of (7). First, if we fix μ, we obtain a subproblem with a convex objective with one orthogonality constraint, we will call this problem S_O. Second, when we fix u, we will get a convex optimization problem, which we will call S_{μ_i}. We will see that these sub-problems can be solved efficiently suggesting that an Alternating Minimization or a Batch coordinate descent algorithm (Gorski et al., 2007) can be used to solve (7). We now provide details about the sub-problems.

Algorithm 1 Alternating Minimization Algorithm

```
Pick arbitrary starting point $\mu$, initialize $u$ such that $||u||_2 = 1$.
for $t = 1, 2, \cdots, T$ do
    Update $\mu \leftarrow \arg\min S_\mu$
    Update $u \leftarrow \arg\min S_O$
end for
```

3.1.2. Subproblem S_O

For a fixed μ, we define $M := \left(\sum_{i=1}^k \mu_i x_i x_i^T + \epsilon I\right)^{-1}$. Our model can be written as,

$$\min_{u:||u||_2^2=1} \lambda||\gamma - u||_2^2 + u^T M u \qquad (8)$$

Expanding the ℓ_2 penalty term, we get, $||\gamma - u||_2^2 = \gamma^T\gamma - 2u^T\gamma + u^Tu = 1 - 2\gamma^Tu + u^Tu$. The last equality is true since γ is a unit norm eigenvector. So our subproblem is,

$$\min_{u:||u||_2^2=1} u^T(\lambda I + M)u - 2\gamma^T u \qquad (9)$$

Note that this is *almost* an eigenvalue problem, that is, we are interested in the largest eigenvalue of $(\lambda I + M)^{-1}$ *except* that we also have the $-2\gamma^T u$ term in the objective. In any case, since the objective is differentiable, we can run a projected gradient method with the projection step being the simple normalization of the vector u at each step. When the eigenvalue spectrum of the Hessian matrix is large, we should make sure that the top k eigenvectors of the reduced and the full Hessian are close. In this case, we solve,

$$\min_{U \in \mathbb{R}^{p\times k}} \sum_{j=1}^k ||\gamma_j - u_j||_2^2 + tr\left(U^T M U\right) \quad \text{s.t. } U^TU = I \qquad (10)$$

where $u_j, 1 \le j \le k$ is the $j-$th column of U. While in medium scale datasets, projected gradient algorithms tend to be efficient, an algorithm that stays in the feasible set (Wen & Yin, 2013; Collins et al., 2014) is better suited for large convex problems with orthogonality constraints.

3.1.3. Subproblem S_μ

Proposition 1. *Denoting the objective function of (7) as f, f is convex w.r.t. μ for all $\mu \in [0, 1]^n$.*

Remark: Note that if we use $A-$ or $E-$optimal designs instead of the $D-$optimal design, we can reformulate this subproblem S_μ as a Semidefinite programming problem with second order cone constraints which can be solved efficiently using standard optimization solvers.

Corollary 2. *Alg. (1) constructs a monotonically decreasing sequence of iterates in the objective.*

Synopsis: Even though the geometric formulation mentioned in the previous section provides a clear intuition to the problem formulation, the number of decision variables in (6) is $pk+n$ whereas the ED problem for ridge regression (2) only had n decision variables. This might become problematic especially when $p \gg n$ which is typically where variable selection is essential. To remedy this issue, we propose an alternative formulation next.

3.2. ED-I: Incoherent Experimental Design

Our second formulation utilizes a result related to the LASSO and the well known Restricted Isometry Property (RIP) which we formally define here and then review the statement of a theorem that will be useful in our model.

Definition 3 (Restricted Isometry Property (Candès & Plan, 2009)). Let $X \in \mathbb{R}^{n\times p}$. For $s \ge 0$, the s-restricted isometry constant δ_s of X is the smallest nonnegative number δ such that $(1-\delta)||\beta||_2 \le ||X\beta||_2 \le (1+\delta)||\beta||_2$ for all $s-$sparse β, i.e., $||\beta||_0 \le s$. If $\delta_s < 1$, then X is a $s-$restricted isometry ($s-$RIP).

With this definition in hand, the next theorem provides a guarantee on the quality of variable selection.

Theorem 4. *(Candès & Plan, 2009) Suppose X has $4s-$RIP constant $\delta_{4s} \le \frac{1}{4}$. Let $\beta_0 \in \arg\min_\beta\{||\beta||_0$:*

$X\beta = y\}$ and $\beta_1 \in \arg\min_\beta\{||\beta||_1 : X\beta = y\}$. If $||\beta_0||_0 \leq s$, then $\beta_1 = \beta_0$.

The theorem suggests that if the matrix X *satisfies* the RIP, then using ℓ_1 instead of ℓ_0 is not a relaxation — the variable selection done by LASSO is exactly equal to that from the ℓ_0 problem. Using this property, we can write a combinatorial form of the ED problem for the LASSO model. More formally, we seek to solve,

$$\arg\max_{\mu\in\{0,1\}^n} \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T + \epsilon I\right)$$
$$\text{s.t. } \mathbf{1}^T\mu \leq B, X_{[\mu],:} \text{ is } 4s-\text{RIP}$$

where $X_{[\mu],:}$ denotes the selected subset of rows of X, that is, row i is chosen if $\mu_i = 1$. As before, the objective drives the inclusion of subjects based on the D-optimality criterion. But the constraints require that the data matrix for the selected set satisfy RIP — this will ensure that the variable selection aspect of LASSO works exactly as intended.

Unfortunately, recent results show that checking $4s-$RIP is NP-Hard (Bandeira et al., 2012). Whence, even if a black box returns an optimal X, we cannot verify the optimality. As a result, other measures that are easy to check have been developed as surrogates to RIP. We will utilize a common alternative which will lead to a tractable formulation.

Leverage scores: (Juditsky & Nemirovski, 2011; Drineas et al., 2012) and others have noted that RIP is a strong assumption and in practice, a less conservative requirement may be as effective. As a surrogate, one typically uses *Incoherence* which is easier to compute, this is defined next.

In statistics, the hat matrix and leverage scores determine how much information a data sample carries with respect to the linear model. The hat matrix is defined as $\hat{H} := X(X^TX + \epsilon I)^{-1}X^T$. The leverage score l_i of a particular sample $i \in \{1, .., n\}$ is defined as the $i-$th diagonal element of $\hat{H}$. With each set of leverage scores, we may associate a quantity known as *coherence* defined as $\mathsf{c} := \max_i l_i$ where a higher value of c implies that the samples are highly correlated. There are various approaches in machine learning that use coherence, rather *incoherence*, see (Chen et al., 2014) . We can now provide a formulation (analogous to the previous section) that selects the (feasible) set of samples that have the least value of c.

$$\min_\mu \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T + \epsilon I\right)^{-1}$$
$$+ \lambda \max_{i=1,\ldots,n}\left\{\mu_i e_i^T X\left(\sum_{i=1}^{n}\mu_i x_i x_i^T + \epsilon I\right)^{-1} X^T e_i\right\}$$
$$\text{subject to } \mu_i \in \{0,1\}, \quad \mathbf{1}^T\mu \leq B \tag{11}$$

Observe that since $\mu_i \in \{0,1\} \iff \mu_i^2 \in \{0,1\}, \mu \geq 0$, we have an equivalent form of the selection problem as,

$$\min_\mu \log\det\left(\sum_{i=1}^{n}\mu_i x_i x_i^T + \epsilon I\right)^{-1}$$
$$+ \lambda \max_{i=1,\ldots,n}\left\{\mu_i^2 e_i^T X\left(\sum_{i=1}^{n}\mu_i x_i x_i^T + \epsilon I\right)^{-1} X^T e_i\right\}$$
$$\text{subject to } \mu_i \in \{0,1\}, \quad \mathbf{1}^T\mu \leq B \tag{12}$$

3.2.1. Algorithm

We may solve the optimization problem using a randomized coordinate descent method shown in Alg. 2, see (Bertsekas & Tsitsiklis, 1989) for details. Here, $\Pi_{\mathcal{C}}$ denotes the projection onto $\mathcal{C} := \{\mu : 0 \leq \mu \leq 1, \mathbf{1}^T\mu \leq B\}$.

Algorithm 2 Randomized coordinate descent algorithm for solving (12)

```
Pick an arbitrary starting point $\mu$
for $t = 1, 2, \cdots, T$ do
  for $k = 1, 2, \cdots n$ do
    $i \in \{0, \cdots, n\}, \quad \mu_i \leftarrow \mu_i - \eta\nabla_{\mu_i} f$
  end for
  Update $\mu \leftarrow \Pi_{\mathcal{C}}(\mu)$
end for
```

Proposition 5. *(12) is a convex optimization problem.*

Corollary 6. *Denote the objective function of (12) as f. Given an accuracy $\rho > 0$, Alg. (2) outputs a $\bar{\mu} \in \mathcal{C}$ such that $|f(\mu^*) - f(\bar{\mu})| \leq \rho$ where μ^* is the optimal solution.*

The above statements assert that we can find the global optimum of the integrality relaxed problem.

3.2.2. Pipage Rounding

Pipage rounding scheme was introduced in (Ageev & Sviridenko, 2004) to round fractional solutions producing a feasible solution without incurring a substantial loss in the value of the objective function (Harvey & Olver, 2014), (Chekuri et al., 2009). For this technique to work in the present context, we need three conditions to be satisfied: **(i)** For any $\mu \in \mathcal{C}$, we need a vector v and $\delta, \tau > 0$ such that $\mu+\delta v$ or $\mu-\tau v$ have strictly more integral coordinates; **(ii)** For all μ, the objective function f must be convex in the direction of v; and **(iii)** Most importantly, we need a starting fractional μ with a guarantee that $f(\mu) \leq \varpi\cdot$**opt** where **opt** is the optimal value of the (discrete) optimization problem for a known constant ϖ. With some work (using similar techniques as in (Horel et al., 2014)), by choosing a suitable algorithm to solve the relaxed form of (12), we can show an approximation ratio for the problem. However, this needs interior point methods and we found that the overall procedure becomes impractical for our datasets. The rounding scheme, if applicable, is still powerful, independent of how the relaxed form of (12) is solved.

Applicability of Rounding. It is clear from Prop. 5 and Cor. 6 that conditions (ii) and (iii) are satisfied by the objective f in (12). So, we only need to verify condition (i). Suppose that μ is a non-integral vector in $\mathcal{C}$. Also, assume that there are at least two fractional coordinates μ_k, μ_l. Then, let us set $v = e_k - e_l$ where e_k, e_l are standard basis vectors in k, l coordinates respectively. Letting $\delta := \min(1-\mu_k, \mu_l)$ and $\tau := \min(1-\mu_l, \mu_k)$ we immediately have that $\mu + \delta v$ and $\mu - \tau v$ are vectors in $\mathcal{C}$ with *strictly more integral coordinates*. Observe that *at least* one of the two vectors stated above is feasible. When both are feasible, if $f(\mu + \delta v) \geq f(\mu)$, we set $\mu \leftarrow \mu + \delta v$, otherwise we set $\mu \leftarrow \mu - \tau v$ and repeat until $\mu \in \{0,1\}^n$. The procedure terminates in at most n steps and its complexity is determined entirely by evaluating the objective function which involves a determinant and an inverse computation.

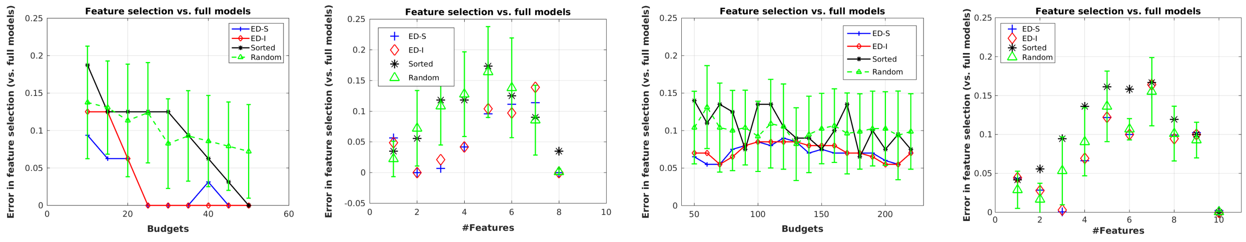
Implementation. We now briefly explain the implementation using simple numerical techniques. The key observation is that each step of the procedure involves at most a (symmetric) rank-2 update of $M^* = \sum_{i=1}^n \mu_i^* x_i x_i^T$ where μ_i^* is the output of Alg. (2). If the determinant and inverse of M^* are computed, at each step of the rounding procedure, we need to compute the determinant and inverse of $M^* + \Delta$ where Δ is symmetric and has rank at most 2. Using an inductive technique described in (Saigal, 1993), the inverse can be updated in $4n^2 + 4n + 1$ operations as opposed to $\mathcal{O}(n^3)$. Then, the update of the determinant only involves computation of four dot products between vectors of length p (Chap. 6 in (Nocedal & Wright, 2006)).

4. Experiments

Data summary. We first evaluated the overall efficacy of our proposed formulations and algorithms on two standard LASSO datasets (*prostate*, (Tibshirani, 1996) and *lars*, (Efron et al., 2004)) and compared their performance to baseline/alternative schemes. After these proof of principle experiments, we performed a set of statistical analyses related to the motivating neuroscience application involving imaging and cognitive data from Alzheimer's Disease Neuroimaging Initiative (ADNI) (*neuro*). These experiments were designed to assess the extent to which statistical power will be compromised when using a smaller fraction of the subjects for estimating linear models associating imaging/clinical covariates with cognitive outcomes. The benchmark data, *prostate* and *lars*, include 8 and 10 features respectively with one dependent variable each, and are well-studied for feature selection. The *neuro* data contains 118 features for image-derived Region-of-Interest (ROI) summaries from Positron Emission Tomography (PET) images. Two cognitive scores were used as dependent variables: Alzheimer's Disease Assessment Score (ADAS) and the diagnostic rating Clinical Dementia Rating (CDR). The appendix includes additional details.

Evaluations Setup. The evaluations were two-fold. First, we compare the performance of ED-S and ED-I to existing baseline experimental design algorithms including: (i) a random design where a given budget number of instances are selected uniformly at random, and (ii) a sampling procedure that approximates the distribution of the observations (referred to as "1-mean"), where we first compute the mean of the (current) samples and the standard deviation σ and filter the samples lying inside the ball centered at the mean with radius $\varrho||\sigma||_2$ for a given $\varrho > 0$. This process is repeated after removing the points lying inside the ball. After k steps, we will be left with k bags of samples with varying sizes. From each of these bags, samples are selected proportionally such that they sum to the budget B. The latter scheme is popular in optimality design where the general idea is to *cover* the dataset by repeatedly selecting 'representer' observations. Our algorithms are compared to these baselines in terms of their ability to consistently pick the correct features (which are defined via the full model i.e., LASSO on *all* instances).

The second set of evaluations deal with the model-fit of reduced models (linear models learned using ED-I and ED-S for a given budget) versus the full model. These goodness-of-fit criteria include consistency of zeros and signs of model coefficients, Bayesian Information Criterion (BIC), Akaike information criterion (AIC) and adjusted R^2. Recall that the two linear models we seek to compare do *not* use the same set of observations, and they do not necessarily sparsify the same set of features. Therefore, unlike the classical non-nested hypothesis testing there is no direct way (e.g., using F or χ^2 statistics) to compare them (Pesaran & Weeks, 2001). To address this problem, we *generate* samples from the full and reduced setups (in a bootstrap sense) and perform a two-sample t-test. The null hypothesis is that the full and reduced setups give 'similar' responses (e.g., ADAS) for a given set of covariates (ROI values). In the ideal case, where the reduced setup captures all the variation of the full one, a *non-significant* (or high) p-value is desired. This is similar to providing "insignificant evidence" against the null (i.e., insignificant evidence that the linear models are different) (Berger & Sellke, 1987). A single workstation with 8 cores and 32GB RAM is used for experiments. We ran 1000 epochs of ED-I, and 50 main iterations of ED-S (with 20 iterations for each of its subproblems). For a fixed budget B, ED-I and ED-S take approximately 7 and 10 min respectively. This is followed by rounding μ_is, and if the rounding generates an infeasible solution (i.e., $\mathbf{1}^T\mu > B$) we randomly drop some of the fractional μ_i subjects. In all the experiments, $\mathbf{1}^T\mu$ overshot B by at most 3 (i.e., most μ_is were binary).



(a) (b) (c) (d)

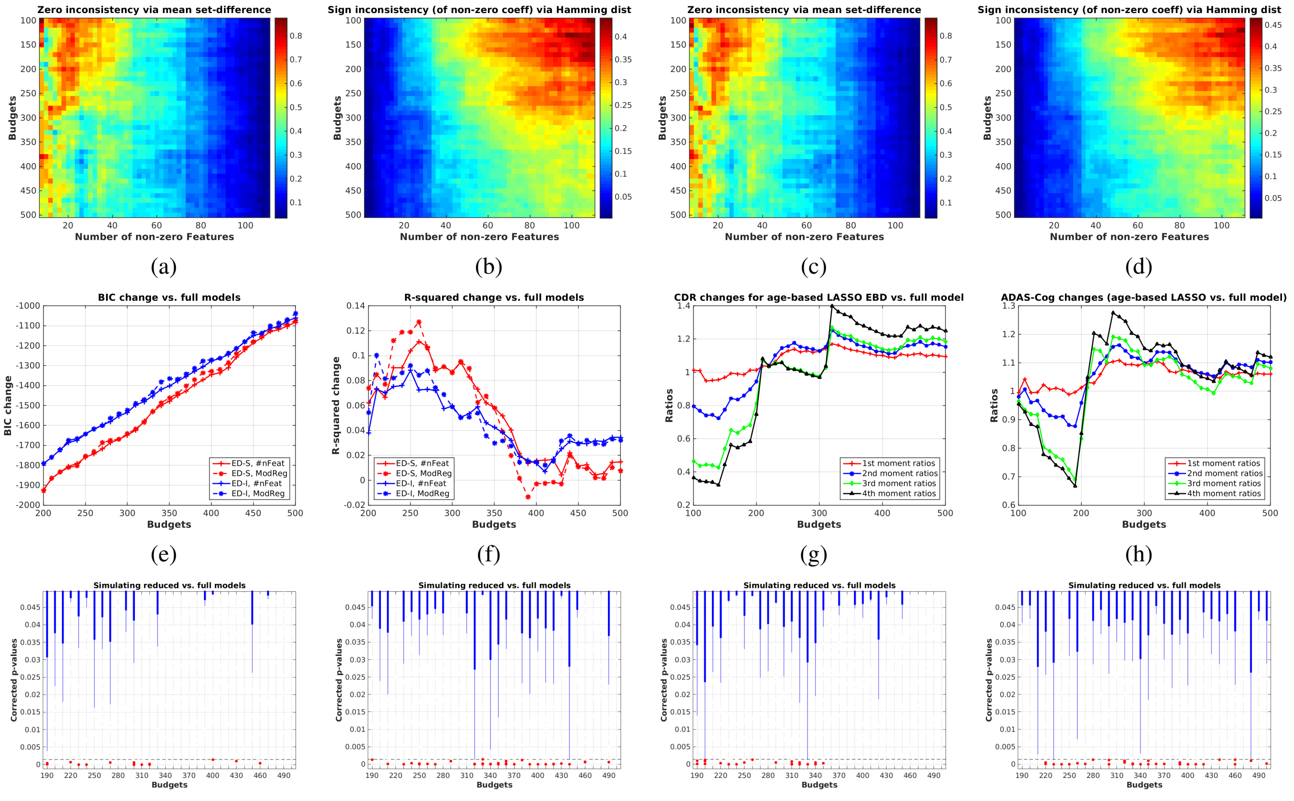
Figure 2. Errors in consistent selection of correct features (derived from full model) for *prostate* (a,b) and *lars* (c,d) datasets.

Comparison to baseline designs. Figure 2 shows the discrepancy in feature selection versus the baseline algorithms. The y-axis in these plots measures this mismatch where the 'correct' features (i.e., ground truth) are assumed to come from the full LASSO. The x-axis lists different budgets. Clearly, both ED-I and ED-S have consistently smaller errors compared to the two baselines, and achieve zero error in some cases as budget increases (Figure 2(a)). We see that the proposed models outperform the 1-mean baseline, although the latter approximates the modes of the data distribution efficiently. Similar behaviour is observed for error plots vs. number of LASSO features (see Figure 2(b,d)). We note that the increase in error as the number of LASSO features increases is due to the correlation in the input data. Unlike the baselines, ED-I and ED-S models select the covariates consistently, even when the number of LASSO features is small (left ends of the x-axis).

Do reduced models approximate the full model? Figure 3 summarizes some of the model-fit measures comparing ED-I and ED-S to full models on *neuro* data (complete set of plots are in the appendix). First observe that the zero inconsistency in Figures 3(a,c) decrease gradually as the budget (y-axis) and/or number of allowed nonzero coefficients of LASSO (x-axis) increases. The zero inconsistency refers to the proportion of nonzero coefficients in the reduced setup that are *absent* in the full one. The input ROI features in *neuro* are strongly correlated, and therefore when the number of allowed nonzero coefficients is small (top-left in Figures 3(a,c)) LASSO picks few of the many 'similar looking' features making zero inconsistency larger. Such a monotonic trend is also evident for sign inconsistency in Figure 3(b,d). However, unlike the previous case, the sign inconsistencies are high for smaller budgets with a large number of nonzero features (top-right in Figure 3(b,d)). This follows directly from the fact that, at the top-right corners LASSO gradually approaches Ridge regression where it is allowed to pick $> 75\%$ of features. Most of these nonzero coefficients will be very small in magnitude, and due to the correlations in the data, the signs of these coefficients are prone to mismatch. The strong linear trends of the inconsistency plots suggest that both ED-I and ED-S are robust to noise and behave well with changing budgets and regularizers.

We see in Figure 3(e,f) that the reduced setup has much smaller BIC compared to the full one. The red and blue curves correspond to ED-S and ED-I respectively and the plots are averaged across multiple choices of model and LASSO regularizers. Clearly, the magnitude of change decreases monotonically as the budget increases. Further, the adjusted R^2 of reduced setup is larger (Figures 3(f)) compared to full ones, although the trends are not as monotonic as was seen for BIC change. This implies a better log-likelihood model-fit for reduced setups, which follows from the fact that D-optimality objective of ED-I (and hence ED-S) maximizes the variation among the selected subjects. Any input feature and/or dependent variable noise (like corrupted observations, sampling noise) from the unselected subjects (which are now linear combinations of the selected ones), does not propagate into the linear model estimation. This interpretation is also supported by noticing that the gain in R^2 is higher for smaller budgets and reduces as the budget increases (Figure 3(f)). Note that, the trend in Figure 3(f) also implies that the optimal choices of budgets for the *neuro* dataset are ~ 400 ($\sim 40\%$ of the total population). The R^2 change of reduced vs. full model (3(f)) was used to pick a "good" budget. Indeed, full dataset is always better than the reduced, but here we refer to the smallest budget that approximates the full model in R^2 change as "optimal". Figures 3(a-f) also show that the two proposed algorithms (ED-I and ED-S) yield very similar results.

Building upon these model-fit measures Figure 3(g,h) shows the ratios of 1$^{\text{st}}$ to 4$^{\text{th}}$ moments of the samples *generated* from the full and reduced setups. Most of these ratios (even for higher order moments) are centered around 1, suggesting that reduced setups are *excellent* approximations of the full one. The appendix includes a discussion of the mismatch of selected subjects between ED-I and ED-S and sparsistency of the algorithms, another desired property of interest for sparse models. We also quantify the observations in Figure 3(g,h) by performing hypothesis testing of the reduced vs. full models. This is shown in the box plots of p-values in Figure 3(i,j) and 3(k,l) for ED-I



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 3. Zero (a,c) and Sign inconsistency (b,d) of ED-I. Change in BIC (e) and R^2 (f) vs. full model. (g,h) Dependent variable moments from reduced vs. full models. p-values of ED-I vs. full (i,j) and ED-S vs. full (k,l) hypothesis testing.

and ED-S respectively for different budgets (x-axis). Recall that the null hypothesis is that the samples from the two setups are *similar*. Since we perform multiple testing (tests across different model and LASSO regularizers) for each budget, the p-values are Bonferroni corrected (the dotted line in each plot). The 'red' dots *below* this Bonferroni threshold are the significant p-values implying that the samples from reduced and full models are different. Clearly, these significant ones are *very few* and scattered across all budgets, and much smaller compared to the non-significant ones (denoted by blue boxes). This means that the number of budget and regularizer combinations that reject the null is extremely small. Also, such cases are much smaller for ED-I (Figure 3(i,j)) compared to ED-S (Figure 3(k,l)). Note that depending on the number of samples (for computing t-statistics), these scattered red points will further reduce. Overall, we see that the reduced setups capture all modeling/distributional characteristics of the full one for almost all choices of budget, LASSO and/or λ's.

5. Conclusions

We addressed the problem of experimental design in sparse linear models common in many applications. We proposed two novel formulations and derived efficient algorithms for experimental design problems on a budget. We presented detailed analysis along with the optimization schemes for ℓ_1 regularized linear models. Our technical results hold for a more general class of sparse linear models as well as optimal design criteria other than $D-$optimality (as long as the relaxation yields a convex model). We showed an extensive set of experiments providing strong evidence for the robustness and efficiency of these formulations. The ideas described here have applications in experiment design problems in neuroscience leading to potential cost savings in longitudinal studies aimed at clinical trials.

Acknowledgements

The authors are supported by NIH R01 AG040396 (VS), NSF CAREER RI 1252725, NIH R01 AG027161 (SCJ), NSF CCF 1320755, University of Wisconsin ADRC (P50 AG033514) and UW CPCP (U54 AI117924). We thank Karl Rohe, Paul J. Rathouz and Thomas D. Cook for various discussions related to this work. The code is publicly available at https://github.com/sravi-uwmadison/Exp_design_sparse.

References

Ageev, Alexander A and Sviridenko, Maxim I. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004.

Bandeira, Afonso S, Dobriban, Edgar, Mixon, Dustin G, and Sawin, William F. Certifying the restricted isometry property is hard. *arXiv preprint arXiv:1204.1580*, 2012.

Berger, James O and Sellke, Thomas. Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

Bertsekas, Dimitri P and Tsitsiklis, John N. *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., 1989.

Bertsimas, Dimitris, Johnson, Mac, and Kallus, Nathan. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 2015.

Candès, Emmanuel J and Plan, Yaniv. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

Candes, Emmanuel J and Tao, Terence. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

Chaloner, Kathryn and Verdinelli, Isabella. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Chekuri, Chandra, Vondrák, Jan, and Zenklusen, Rico. Dependent randomized rounding for matroid polytopes and applications. *arXiv preprint arXiv:0909.4348*, 2009.

Chen, Yudong, Bhojanapalli, Srinadh, Sanghavi, Sujay, and Ward, Rachel. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 674–682, 2014.

Collins, Maxwell D, Liu, Ji, Xu, Jia, Mukherjee, Lopamudra, and Singh, Vikas. Spectral clustering with a convex regularizer on millions of images. In *Computer Vision–ECCV 2014*, pp. 282–298. Springer, 2014.

Das, Ashish. An introduction to optimality criteria and some results on optimal block design.

Dette, Holger, Pepelyšev, Andrej, and Žigljavskij, Anatolij A. Optimal design for linear models with correlated observations. 2011.

Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W, and Woodruff, David P. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Efron, Bradley, Hastie, Trevor, Johnstone, Iain, Tibshirani, Robert, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Gorski, Jochen, Pfeuffer, Frank, and Klamroth, Kathrin. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

Harvey, Nicholas JA and Olver, Neil. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 926–945. SIAM, 2014.

Hinrichs, Chris, Singh, Vikas, Xu, Guofan, Johnson, Sterling C, Initiative, Alzheimers Disease Neuroimaging, et al. Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population. *Neuroimage*, 55(2):574–589, 2011.

Hinrichs, Chris, Dowling, N Maritza, Johnson, Sterling C, and Singh, Vikas. Mkl-based sample enrichment and customized outcomes enable smaller ad clinical trials. In *Machine Learning and Interpretation in Neuroimaging*, pp. 124–131. Springer, 2012.

Horel, Thibaut, Ioannidis, Stratis, and Muthukrishnan, S. Budget feasible mechanisms for experimental design. In *LATIN 2014: Theoretical Informatics*, pp. 719–730. Springer, 2014.

Ithapu, Vamsi K, Singh, Vikas, Okonkwo, Ozioma C, Chappell, Richard J, Dowling, N Maritza, Johnson, Sterling C, Initiative, Alzheimer's Disease Neuroimaging, et al. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer's & Dementia*, 11(12): 1489–1499, 2015.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 427–435, 2013.

Juditsky, Anatoli and Nemirovski, Arkadi. On verifiable sufficient conditions for sparse signal recovery via 1 minimization. *Mathematical programming*, 127(1):57–88, 2011.

Kempthorne, Oscar. The design and analysis of experiments. 1952.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirk, Roger E. *Experimental design*. Wiley Online Library, 1982.

Konstantinou, Maria, Biedermann, Stefanie, and Kimber, Alan. Optimal designs for two-parameter nonlinear models with application to survival models. 2011.

Landau, SM, Harvey, D, Madison, CM, Reiman, EM, Foster, NL, Aisen, PS, Petersen, RC, Shaw, LM, Trojanowski, JQ, Jack, CR, et al. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–238, 2010.

Li, Jia. Linear, ridge regression, and principal component analysis, 2008.

Li, Xin and Guo, Yuhong. Adaptive active learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 859–866. IEEE, 2013.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Pesaran, M Hashem and Weeks, Melvyn. Non-nested hypothesis testing: an overview. *A Companion to Theoretical Econometrics*, pp. 279–309, 2001.

Pukelsheim, Friedrich. *Optimal design of experiments*, volume 50. siam, 1993.

Recht, Benjamin, Re, Christopher, Wright, Stephen, and Niu, Feng. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 693–701, 2011.

Saigal, Romesh. On the inverse of a matrix with several rank one. *Ann Arbor*, 1001:48109–2117, 1993.

Searcey, Terena, Bierer, Linda, and Davis, Kenneth L. A longitudinal study of alzheimers disease: measurement, rate, and predictors of cognitive deterioration. *Am J Psychiatry*, 1:51, 1994.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Wen, Zaiwen and Yin, Wotao. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.