Bernoulli **29**(1), 2023, 652–668
https://doi.org/10.3150/22-BEJ1472

Design of c-optimal experiments for high-dimensional linear models

HAMID EFTEKHARI[a], MOULINATH BANERJEE[b] and YA'ACOV RITOV[c]

Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA. [a]*hamidef@umich.edu*, [b]*moulib@umich.edu*, [c]*yritov@umich.edu*

We study randomized designs that minimize the asymptotic variance of a debiased lasso estimator when a large pool of unlabeled data is available but measuring the corresponding responses is costly. The optimal sampling distribution arises as the solution of a semidefinite program. The improvements in efficiency that result from these optimal designs are demonstrated via simulation experiments.

Keywords: Optimal design; inference; sparsity; compressed sensing

1. Introduction

Optimal design of experiments is a statistical framework for improving the efficiency of experiments with wide-ranging applications in science and industry. In the case of regression, given the covariate data (x_i's) and a fixed sample size (n), the goal is to measure the responses (y_i's) only for those x_i's that result in the most accurate estimate of the relationship between the covariates and the response. This can result in significant savings in experiment time and/or monetary expenses associated with the experiment.

The relationship between the response (y) and covariates (x) often can be modeled as a linear model:

$$y = \langle x, \beta^\star \rangle + \varepsilon,$$

where $\beta^\star \in \mathbf{R}^p$ is the true regression parameter and ε denotes a mean-zero noise term. The target of estimation is a linear combination $\langle c, \beta^\star \rangle$ of the regression parameter $\beta^\star$. In the classical framework, a c-optimal design is a design (i.e. a distribution on the covariate data $(x_i)_i$) that minimizes the variance of the ordinary least squares (OLS) estimate of $\langle c, \beta^\star \rangle$.

The optimal design problem has been thoroughly investigated in the case of low-dimensional (fixed dimension, increasing number of observations) linear regression. Various optimality criteria, usually defined as functionals of the information matrix, have been considered and their relative merits are studied in the literature. Pukelsheim [25] provides an excellent introduction to the theory of optimal design. More relevant to our work is the literature on c-optimal design, pioneered by the seminal work of Elfving [12], who provided an elegant geometric solution for the (approximate) c-optimal design problem which can also be cast as a linear program admitting an efficient solution [15]. Variations and extensions of Elfving's theorem have since been studied. Chernoff [8] generalized Elfving's result beyond the regression framework by introducing a local and asymptotic notion of optimality based on the information matrix. Sagnol [31] studied c-optimality for multi-response regression and showed that it can be formulated as a second-order cone programming (SOCP) problem. Dette et al. [9] considered a more general version of the c-optimality criterion that is robust across classes of models and proved a generalization of Elfving's result for this criterion. c-optimal designs have also been studied for models with correlated [28] and heteroscedastic [10] error terms.

In the high-dimensional regime, where the dimension of covariates is larger than the sample size, consistent estimation and inference are hopeless without further structural assumptions, the most well-studied being sparsity of $\beta^\star$. For estimation in sparse linear models, the ℓ_1-regularized least squares (also known as the lasso) estimator has been proposed [34] and shown to enjoy strong theoretical guarantees [2]. More recently, the debiasing techniques have been introduced for correcting the bias of the lasso estimator, thereby providing $\sqrt{n}$-consistent and asymptotically normal estimates of the coordinates of $\beta^\star$ and allowing the construction of confidence intervals for its finite-dimensional coordinates [18,35,37]. Variations of the debiased lasso estimator have also been devleoped for estimating linear combinations $\langle c, \beta^\star \rangle$ of $\beta^\star$ [4,5].

However, the optimal design problem for inference in high-dimensional linear models has received scant attention. Seeger [32] studied sequential design for maximizing information gain in sparse linear models in a Bayesian framework. Ravi et al. [27] introduced D-optimal designs for sparse linear models that combine the classical D-optimiality criterion applied to a perturbation of the information matrix and a term that penalizes the l_2 distance between the top eigenvectors of the full and reduced Hessians. Deng, Lin and Qian [14] used nearly orthogonal Latin hypercube designs to improve variable selection using the lasso. More recently, Huang, Kong and Ai [16] extended Keifer's notion of Φ_l-optimality to one applicable for the covariance matrix of the debiased lasso and developed an algorithm for finding local optima of the resulting non-convex optimization problem.

1.1. Contributions

Targeted inference on pre-selected parameters of high-dimensional models is of great interest in a variety of applications but has not been addressed in the existing experimental design literature. Our work bridges this critical gap between classical optimal design theory and frequentist inference in high-dimensional linear models. To our knowledge, this is the first rigorous result on this topic. Our contributions are articulated below:

1. We extend the standard convergence guarantees for the debiased lasso beyond sub-Gaussian designs to random vectors with bounded entries using Poisson sampling and a novel proof technique. In the extant literature on the debiased lasso, it is typically assumed that the rows of the design matrix are sub-Gaussian random vectors. While this assumption is convenient for concentration arguments, it is too restrictive for random vectors taking values in finite sets as are commonly encountered in the design problem. To appreciate that this is more than a mere technicality, see for example [36, Exercise 3.4.5] showing that if X is an isotropic random vector supported on a finite set $S \subset \mathbf{R}^p$ and if $\|X\|_{\psi_2} \leq M$ for a bound M not depending on p, then we must have $|S| > e^{cp}$ for a constant $c > 0$. In words, finitely-supported isotropic sub-Gaussian random vectors with bounded sub-Gaussian norms have supports that are *exponentially large* in the dimension of the vector. Clearly this rules out many interesting design problems.
2. We introduce the notion of a constrained c-optimal design that is a natural extension of the classical c-optimality criterion suitable for the high-dimensional setting. In the low-dimensional regime, the OLS estimate is unbiased and the c-optimality criterion solely focuses on minimizing the variance of the OLS estimate of $\langle c, \beta^\star \rangle$. In the high-dimensional setting, the debiased lasso estimate is not unbiased for finite samples (even though the bias is asymptotically $o(1/\sqrt{n})$ assuming a restricted eigenvalue condition), and one needs to control the bias when minimizing the variance. This naturally leads to semidefinite constraints on the covariance matrix of the design distribution. Furthermore, we illustrate through an example that the debiased lasso can have an arbitrarily large bias (with non-vanishing probability) if one minimizes the c-optimality criterion without regard for controlling the bias. The consequences of singular c-optimal designs [13,23]

and methods for regularizing them [24] have previously been studied in the low dimensional setting, but to the best of our knowledge, ours is the first work to use constrained (regularized) c-optimal designs for high-dimensional regression.

1.2. Application

Examples of the potential applications of our work include:

- **Detection of weak signals in sparse magnetic resonance images.** Due to physical limitations, measurements in MRI are time-consuming and there is substantial interest in reducing the number of measurements while preserving the reconstruction accuracy of the images [21]. In this application, optimal designs allow more efficient inference on, say, the average intensity of regions of interest (ROI) of the object under study (for example, the suspected location of a brain tumor), which enables detection of weak signals with fewer measurements than a uniform design. This example is described in detail in Section 3.
- Another application area is in controlled experiments on the web, where users or clients are directed to different versions of a product or service in order to measure the effect of an intervention on a metric of interest such as revenue or user engagement [19]. Online businesses and service providers often have large amounts of covariate data on their users (their personal information, browsing history, preferences, purchase history, etc.) that together with the intervention of interest can be encoded into high-dimensional covariates and which can be used to decide which users are enrolled into the aforementioned experiments.

1.3. Notation and definitions

For any natural number q we set $[q] := \{0, 1, \ldots, q\}$. In this work p is the dimension of covariates, N is the total number of available design points, and n denotes the sample size for which responses have been observed, so that in general we have $n \le p \le N$. Lower-case $x_i \in \mathbb{R}^p$ is used to denote potential covariate data available for sampling; the set $(x_i)_1^N$ of all these points is called the experimental domain. Upper-case $(X_i)_1^n$ denotes a sample from $(x_i)_1^N$ for which responses $(y_i)_1^n$ have been observed. The $n \times p$ matrix with X_i^T in its ith row is denoted by $\mathbb{X}$ and we let $Y = (y_1, \ldots, y_n)^T$.

The standard inner product in $\mathbb{R}^p$ is denoted by $\langle \cdot, \cdot \rangle$, whereas the ℓ_q-norm is defined by $\|a\|_q = \sqrt[q]{\sum_i |a_i|^q}$ for any $q \ge 1$. The j-th element of the standard basis of $\mathbb{R}^p$ is denoted by e_j.

For a random variable or vector Z we write $\mathbf{E}Z$ for the expected value of Z, where as $\mathbf{P}(E)$ is used to denote the probability of an event E. For a random variable Z the sub-Gaussian (respectively, sub-exponential) norm is defined by $\|Z\|_{\psi_q} := \inf\{t > 0 : \mathbf{E}\exp(|Z|^q/t^q) < 2\}$ with $q = 2$ (respectively, $q = 1$).

For a positive semidefinite matrix Σ, we use $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ to denote its smallest and largest eigenvalues. For two matrices A and B, the relation $A \preccurlyeq B$ means $B - A$ is positive semidefinite. For a matrix Σ we write Σ^+ for its pseudo-inverse and use $\mathrm{col}(\Sigma)$ to denote its column space.

For two sequence a_n, b_n we write $a_n \lesssim b_n$ to mean that there exists an absolute constant $C > 0$ (that does not depend on n, p, N) such that $a_n \le C b_n$ for all $n \ge 1$. The reverse inequality $a_n \gtrsim b_n$ means $b_n \lesssim a_n$, while $a_n \asymp b_n$ means both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold.

1.4. Problem formulation

We assume that a large pool $\mathcal{X} = (x_i)_{i=1}^N$ of covariate data (called the experimental domain) is available for which no responses have been observed yet. Upon sampling an $x \in (x_i)_{i=1}^N$, the experimenter observes the corresponding response, which we assume follows a linear model

$$y = \langle x, \beta^\star \rangle + \varepsilon \tag{Model 1}$$

where $\beta^\star \in \mathbb{R}^p$ is an unknown parameter and ε is sub-Gaussian noise independent of the choice of x. Crucially, we consider the case where the total number of observed responses is $n < p$, resulting in a high-dimensional inference problem.

We assume that the target of inference is a linear combination $\langle c, \beta^\star \rangle = \sum_{i=1}^p c_i \beta_i^\star$ of $\beta^\star$, where c is chosen apriori by the experimenter. For example, with $c = \mathrm{e}_j$ (the jth standard basis element) the aim is to conduct inference on $\beta_j^\star = \mathrm{e}_j^T \beta^\star$. Another example is the contrast $c = \mathrm{e}_i - \mathrm{e}_j$ where the goal is to conduct inference on $\beta_i^\star - \beta_j^\star$, the difference between effects of the ith and jth covariates.

We assume throughout that $N \geq p$, as it is crucial for our results that the population covariance matrix is non-singular, so that the restricted eigenvalue condition can be guaranteed for sample design matrices. In the $N < p$ case, while the restricted value condition might hold for population or sample covariance matrices, it is well-known that these conditions are NP-hard to check [11], and so equally hard to enforce on designs.

Before presenting our results for high-dimensional models, we review the c-optimality criterion in the low-dimensional setting.

1.5. c-optimality in low dimensions

Suppose that $n > p$, and we are given a fixed sample $(X_i, y_i)_{i=1}^n$ and let $\mathbb{X} \in \mathbf{R}^{n \times p}$ be the matrix with X_i^T in its ith row. Assume that c belongs to $\mathrm{span}(X_1, \ldots, X_n)$, or equivalently, $c \in \mathrm{col}(\mathbb{X}^T \mathbb{X})$. Then by the Gauss-Markov theorem, the best linear unbiased estimate of $c^T \beta^\star$ is $c^T \widehat{\beta}^{OLS}$ where the ordinary least squares (OLS) estimate of $\beta^\star$ is defined by $\widehat{\beta}^{\mathrm{OLS}} := (\mathbb{X}^T \mathbb{X})^+ \mathbb{X}^T y$. Recall that $c^T \widehat{\beta}^{\mathrm{OLS}}$ is unbiased and has variance equal to $\sigma^2 \cdot c^T (\mathbb{X}^T \mathbb{X})^+ c$. The goal in the c-optimal design of experiments is to choose $X_1, \ldots, X_n$ (among $(x_1, \ldots, x_N)$) in such a way that the above variance is minimized. Let $N_i \in [n] := \{0, 1, \ldots, n\}$ be the number of times x_i is repeated in the design. Then the exact c-optimality problem can be written as

$$\begin{aligned} \mathbf{P0}: \min_{(N_i)_1^N} \quad & c^T \Sigma^+ c \\ \text{s.t.} \quad & \Sigma = \frac{1}{n} \sum_{i=1}^N N_i x_i x_i^T, \quad c \in \mathrm{col}(\Sigma) \\ & \sum_{i=1}^N N_i = n, \quad N_i \in [n]. \end{aligned}$$

Since this *exact* design problem (that is, with the integer constraints $N_i \in [n]$) turns out to be computationally hard to solve (more specifically, it is NP-complete [6]), it is often relaxed to the so-called

approximate design problem formulated as below:

$$\mathbf{P1}: \min_{w\in\mathbb{R}^N} \quad c^T\Sigma^+ c$$
$$\text{s.t.} \quad \Sigma = \sum_{i=1}^N w_i x_i x_i^T, \quad c \in \text{col}(\Sigma)$$
$$\sum_{i=1}^N w_i = 1, \quad w \geq 0.$$

Note that given an approximate optimal design $(w_i^\star)_1^N$, it is likely that $nw_i^\star$ is not an integer. Thus one may have to resort to rounding techniques to obtain an exact design from the approximate optimal design [26]. Alternatively, one can use randomization where $\{X_1,\ldots,X_n\}$ is an iid sample from $(x_i)_1^N$ with probabilities $(w_i^\star)_1^N$ [3, section 7.5.1]. We take the latter approach in this work since randomization also allows the derivation of high-probability bounds on the bias of the debiased lasso by guaranteeing that a restricted eigenvalue condition holds with high probability.

It follows from Elfving's theorem [12,33] that the above problem is equivalent to the following ℓ_1 minimization problem:

$$\mathbf{P1'}: \min_{b\in\mathbb{R}^N} \quad \|b\|_1$$
$$\text{s.t.} \quad c = \sum_{i=1}^N b_i x_i \tag{1}$$

If $b^\star$ is an optimal solution of **P1′**, then $w^\star = (|b_i^\star|/\|b^\star\|_1)_{i=1}^N$ is an optimal solution for **P1**. It can be shown that $w^\star$ has at most p non-zero entries if **P1′** is feasible.

Remark 1. In a low-dimensional setting, we have $n > p$ and therefore it is generally possible to write c as a linear combination of $p < n$ points. One may then use rounding techniques on $(nw_i^\star)_1^N$ to obtain deterministic exact designs (i.e. with integer values close to $nw_i^\star$ for the number of times x_i is selected) from the approximate design $(w_i^\star)_1^N$ [26]. However, in a high-dimensional setting where $p \gg n$, it is in general not possible to write c as a linear combination of n of the x_i's. This means that we may have $|\{i \leq N : w_i^\star \neq 0\}| > n$ (and possibly $\gg n$). Furthermore, in this case we typically obtain $nw_i^\star$'s that are all close to zero, rendering the rounding techniques inapplicable. This motivates the use of randomized designs in high dimension, along with fact that randomization allows probabilistic guarantees of the restricted eigenvalue condition needed for the analysis of the lasso estimator [29].

2. Debiased inference of parameters

Various methods have been proposed for estimation and inference of linear functionals $\gamma := \langle c,\beta\rangle$ in the high-dimensional inference literature. Cai and Guo [5] provided minimax optimal rates for the lengths of confidence intervals depending on the sparsity structure of c. Javanmard and Lee [17] studied hypothesis testing for general null hypothesis of the form $\beta \in \Omega_0$ for arbitrary Ω_0. Cai, Cai and Guo [4] studied inference of individualized treatment effects $\langle c,\beta_1-\beta_2\rangle$ for a two-sample problem and linear functionals $\langle c,\beta\rangle$ for the one-sample problem. The common theme among all these works is the use of debiasing procedures, where one starts with a biased estimate (typically variants of the lasso) and then uses a projection method to correct its bias. We describe a variant of the method proposed by Cai, Cai and Guo [4] in the following.

Suppose that $(Y_i, X_i)_{i=1}$ is an i.i.d. sample drawn according to P. To obtain an initial estimate for $\beta^\star$, use the Lasso estimator defined by

$$\widehat{\beta} := \arg\min_{\beta' \in \mathbb{R}^p} \left\{ L_\lambda(\beta') := \frac{\|Y - \mathbb{X}\beta'\|_2^2}{2n} + \lambda \sum_{j=1}^{p} \widehat{W}_j |\beta'_j| \right\}. \tag{2}$$

where $\widehat{W}_j = \sqrt{n^{-1} \sum_{i=1}^n X_{ij}^2}$ and with the tuning parameter of order[1] $\lambda \asymp \widetilde{\sigma}_\varepsilon \sqrt{\frac{\log p}{n}}$ where $\widetilde{\sigma}_\varepsilon^2 = \mathbf{E}\varepsilon^2$ is the noise variance.

In order to correct the bias of $\langle c, \widehat{\beta} \rangle$, Cai, Cai and Guo [4] propose to use the following estimator

$$\widehat{\gamma} = \langle c, \widehat{\beta} \rangle + \frac{1}{n} \widehat{u}^T \mathbb{X}^T (Y - \mathbb{X}\widehat{\beta}),$$

where

$$\begin{aligned} \widehat{u} := \arg\min_u \quad & u^T \widehat{\Sigma} u \\ \text{s.t.} \quad & \|\widehat{\Sigma} u - c\|_\infty \le \|c\|_2 \lambda, \\ & |c^T \widehat{\Sigma} u - \|c\|_2^2| \le \|c\|_2^2 \lambda. \end{aligned}$$

The above minimization problem effectively estimates $u = \Sigma^{-1} c$ in the typical setting where Σ is unknown. In our setting, however, the population of covariates is fully known (after determining a design) and thus one can directly use $u = \Sigma^{-1} c$ in the debiasing procedure. Thus we will be using the debiased estimator

$$\widehat{\gamma} := \langle c, \widehat{\beta} \rangle + \frac{1}{n} u^T \mathbb{X}^T (Y - \mathbb{X}\widehat{\beta}), \tag{3}$$

with $u = \Sigma^{-1} c$.

Before we state our main theorem regarding the consistency and asymptotic normality of $\widehat{\gamma}$ we describe and motivate the Poisson sampling scheme used in our work.

Poisson Sampling. Suppose that $w = (w_i)_1^N$ is a probability distribution on $(x_i)_1^N$. Given an *iid* sample $X_1, \ldots, X_n$ according to probabilities w, define $\widetilde{N}_i$ to be the number of times x_i is repeated among $X_1, \ldots, X_n$:

$$\widetilde{N}_i := |\{k \in \{1, \ldots, n\} : X_k = x_i\}|$$

Then random objects such as $\widehat{\Sigma} = n^{-1} \sum_1^n X_i X_i^T$ that appear in our theoretical analysis can be viewed as functions of the multinomial random variables $(\widetilde{N}_i)_1^N$ as

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{N} \widetilde{N}_i \, x_i x_i^T .$$

[1] This is a theoretical (oracle) value for the tuning parameter since $\widetilde{\sigma}_\varepsilon$ is typically unknown. In practice one can use an estimate of $\widetilde{\sigma}_\varepsilon$ or use cross-validation.

This shifts the focus from high-dimensional random vectors $(X_i)_1^n$ to random variables $(\widetilde{N}_i)_1^N$ which are more amenable to theoretical analysis. Studying the concentration properties of $\widehat{\Sigma}$ therefore necessitates dealing with the $\widetilde{N}_i$'s. However, the dependence among the $\widetilde{N}_i$'s leads to difficulties with concentration arguments which are typically based on independence (note that under iid sampling, the distribution of $(\widetilde{N}_i)_1^N$ is multinomial with probabilities $(w_i)_1^N$ and sum equal to n). In order to use standard concentration results based on independence, it is useful to break this dependence among the $\widetilde{N}_i$'s, which can be achieved by using Poisson sampling as follows: Let $(N_i)_1^N$ be independent random variables with $N_i \sim \text{Poisson}(nw_i)$ and set $K = \sum_{i=1}^N N_i$. A sample $X_1, \dots, X_K$ is then created where x_j appears N_j times in this sample[2]. It is easy to see that:

1. The total number of samples drawn is close to n (relative to n):
$$\frac{K}{n} \to_p 1 \text{ as } n \to \infty.$$
2. Conditioned on $K = k$, the distribution of $(N_i)_{i=1}^N$ is multinomial with parameters $\left(k, (w_i)_1^N\right)$.

Thus the sample obtained from Poisson sampling is similar to sampling with replacement according to w. This Poisson sampling scheme precludes the need for sub-Gaussianity of the vectors X_i.

The following theorem describes the asymptotic distribution of the debiased estimator $\widehat{\gamma}$ on data obtained using Poisson sampling. The theorem is similar in spirit to the results of Cai, Cai and Guo [4] and Javanmard and Lee [17], with the main difference being that it does not assume the covariates are sub-Gaussian vectors. Recall that as discussed in subsection 1.1, this is important in the design setting because it is not clear how to enforce the sub-Gaussianity property when finding an optimal design. We bypass this problem via Poisson sampling along with a novel proof technique that does not rely on the sub-Gaussianity of the design. The asymptotic regime in the following theorem is that of a triangular array where all the parameters (e.g. $N_n, p_n, s_n, M_n, \Sigma_n$, etc.) can be taken to depend on n, with n diverging to infinity. Moreover, as is conventional in high-dimensional statistics, we suppress the dependence on n and write for example N, p, s, M, Σ instead of $N_n, p_n, s_n, M_n, \Sigma_n$ to lighten the notation.

Theorem 1. *Suppose that $\mathbb{X} = (X_i^T)_1^K$ is a Poisson sample according to a distribution w on $(x_i)_1^N$ with $\mathbf{E}K = n$ and that $(y_i, X_i)_1^K$ follow (Model 1) with a s-sparse regression parameter $\beta^\star$. Also assume that the following conditions are satisfied:*

1. $\max_{1 \le i \le N} \|x_i\|_\infty \le M = o\left(\sqrt{n/\log(p)}\right)$.
2. $\lambda_\star I \preccurlyeq \Sigma_w := \sum_1^N w_i x_i x_i^T \preccurlyeq \lambda^\star I$ *where* $0 < \lambda_\star, \lambda^\star < \infty$ *do not depend on* n.
3. *The noise terms ε_j are i.i.d. mean-zero sub-Gaussian random variables with $\|\varepsilon_j\|_{\psi_2} \le \sigma_\varepsilon$ and a variance $\mathbf{E}\varepsilon_n^2$ that is bounded away from* 0 *and* ∞.
4. $s \log^{\frac{3}{2}}(p) = o(\sqrt{n})$.

Then the debiased estimate (3) satisfies

$$\sqrt{n}(\widehat{\gamma} - \gamma) = \frac{1}{\sqrt{n}} c^T \Sigma_w^{-1} \mathbb{X}^T \varepsilon + b_n$$

where

[2]Note the distinction between N and K: The former is the total number of available (potential) design points $x_1, \dots, x_N$ before any sampling is done, whereas the latter is the number of points in the selected random sample $X_1, \dots, X_K$. Also note that K, being a sum of indpendent Poisson random variables, is itself Poisson distributed with mean equal to n.

- *(Bias Bound) As $n \to \infty$, with probability $1 - o(1)$, we have*

$$|b_n| \lesssim \frac{M\sigma_\varepsilon \sqrt{c^T \Sigma_w^{-1} c}}{\lambda_\star} \cdot \frac{s \log^{\frac{3}{2}}(p)}{\sqrt{n}}.$$

- *(Asymptotic Normality) Let $v^2 := (c^T \Sigma_w^{-1} \widehat{\Sigma} \Sigma_w^{-1} c) \mathbf{E}\varepsilon_n^2$ where $\widehat{\Sigma} = n^{-1} \sum_1^K X_i X_i^T$. Then:*

$$\frac{1}{v\sqrt{n}} c^T \Sigma_w^{-1} \mathbb{X}^T \varepsilon \to_d N(0,1) \quad \text{as } n \to \infty.$$

- *(Variance Approximation) The variance of the noise term can be approximated (asymptotically) by $c^T \Sigma^{-1} c$:*

$$\frac{c^T \Sigma_w^{-1} \widehat{\Sigma} \Sigma_w^{-1} c}{c^T \Sigma_w^{-1} c} \to_p 1 \quad \text{as } n \to \infty.$$

Discussion. 1. The first assumption imposes a uniform bound on the coordinates of all design points. Note that we can permit M to depend on n, p, N, and thus slowly grow with n, provided the sparsity constraint (in the fourth assumption) is modified accordingly. For example, if the experimental domain $(x_i)_1^N$ is itself a sample from a population such that all coordinates x_{ij} are sub-Gaussian with $\|x_{ij}\|_{\psi_2} = O(1)$, then it is well-known that $\max_{i,j} |x_{ij}| \lesssim \sqrt{\log(Np)}$ with probability $1 - o(1)$. Thus in this case we could set $M \asymp \sqrt{\log(Np)}$ and replace assumption 4 with $s \log^{3/2}(p) \sqrt{\log(Np)} = o(\sqrt{n})$.

2. The second assumption in Theorem 1 is used in our analysis to ensure that the restricted eigenvalue condition is satisfied (with a non-negligible restricted eigenvalue) for the sample design matrix with high probability, thereby guaranteeing the fast rate of convergence for the lasso estimator. Furthermore, the semidefinite nature of this constraint leads to computationally tractable optimization problems as opposed to constraints directly involving restricted eigenvalues which are themselves NP-hard to compute in general [11]. Also, note that this condition implies that $c \in \text{span}(\{x_i : w_i^\star > 0\})$, which is a necessary condition for unbiasedness in the low-dimensional setting. As illustrated in the following example, in the absence of this assumption the population covariance matrix Σ resulting from $P1$ can have a vanishing restricted eigenvalue, leading to the inconsistency of the debiased estimator.

3. The third assumption regarding the sub-Gaussianity of the noise term is standard in the literature on inference for high-dimensional linear models and is assumed to allow concentration arguments.

4. The last assumption imposes a constraint on the sparsity of the regression parameter $\beta^\star$ in terms of n, p. Note that this is a stronger condition than the "ultra-sparsity" condition assumed in high-dimensional inference[18,35,37]. The extra factor of $\sqrt{\log(p)}$ in this assumption is the price we pay for dispensing with the sub-Gaussianity of covariate vectors (as is typically assumed in the literatture) and only assuming uniformly bounded entries.

Example 1. We provide a simple example where the low-dimensional approach of minimizing the (asymptotic) variance $c^T \Sigma^+ c$ without controlling the bias of the debiased lasso estimator leads to arbitrarily large biases in estimation of $\langle c, \hat{\beta} \rangle$

Let m be a large positive integer. In this example we take $N = p = 2m$ (that is, the number N of available design points is equal to the dimension of covariates p). Let $(x_i)_1^p$ be an orthogonal basis of R^p that satisfies the following conditions:

- $\|x_i\|_\infty \le M$ for some $M > 0$ not depending on n, p.

- The first entry of x_i is given by

$$x_{i1} = \begin{cases} 0 & 1 \le i \le m \\ \sqrt{2} & m+1 \le i \le p. \end{cases}$$

- The x_i's are orthogonal: $p^{-1}\sum_{i=1}^{p} x_i x_i^T = I_p$.

An explicit construction of such a basis is given in the Supplementary Material [20]. Define c to be a scaled sum of the first $m+1$ of x_i's:

$$c = \frac{1}{\sqrt{m+1}} \sum_{i=1}^{m+1} x_i,$$

so that we have $\|c\|_2 = \sqrt{p}$. It follows from Elfving's Theorem that the c-optimal design obtained from problem $P1$ in this case is characterized by

$$w_i^\star = \begin{cases} \dfrac{1}{m+1} & : 1 \le i \le m+1, \\ 0 & : m+1 < i \le p. \end{cases}$$

It is clear that $w^\star$ leads to a singular population covariance matrix[3]

$$\Sigma_{w^\star} = \sum_{i=1}^{p} w_i^\star x_i x_i^T = \frac{1}{m+1} \sum_{i=1}^{m+1} x_i x_i^T.$$

In a high-dimensional setting, we have a sample size budget $n \ll p$. For concreteness, let $n = \lfloor \sqrt{p} \rfloor$. First note that using rounding techniques to obtain a deterministic exact design from $w^\star$ is not an option here because $nw_i^\star \ll 1$. More precisely, $\|nw^\star\|_\infty \to 0$ as $m \to \infty$. Therefore we need to consider randomized designs, where we take i.i.d. samples from $(x_i)_1^p$ with probabilities $(w_i^\star)_1^p$.

Given a $w^\star$-weighted sample $X_1, \ldots, X_n$, let E be the event that x_{m+1} is not among $X_1, \ldots, X_n$. Then we have

$$\mathbf{P}(E) = \mathbf{P}(x_{m+1} \notin \{X_1, \ldots, X_n\}) = \left(1 - \frac{1}{m+1}\right)^n \to 1, \text{ as } m \to \infty.$$

On this event E we have $X_{i1} = 0$ for $1 \le i \le n$. It is clear that in this case we can not hope to estimate $\beta_1^\star$, while $c^T\beta^\star$ can depend $\beta_1^\star$. For a simple example, take $\beta^\star = (\theta\sqrt{(m+1)/2}, 0, \ldots, 0)^T$, so that $\gamma = \langle c, \beta^\star \rangle = \theta$. Assuming a $N(0,1)$ noise distribution for ε_i, it is clear that on E we have $y_i = \varepsilon_i$. Given E, we can show that $\widehat{\beta} = 0_p$ is a solution of the lasso problem (2) with probability $1 - o(1)$ (with $\lambda = \sqrt{(2+\eta)\log(p)/n}$ for a small $\eta > 0$; see the last section in the Supplementary Material [20] for a proof). A natural extension of the debiased estimator (3) with $\Sigma_{w^\star}^+$ substituted for $\Sigma_{w^\star}^{-1}$ is given by

$$\widehat{\gamma} = \langle \widehat{\beta}, c \rangle + \frac{1}{n} c^T \Sigma_{w^\star}^+ \mathbb{X}^T (Y - \mathbb{X}\widehat{\beta}).$$

[3] That is, the population covariance matrix under i.i.d. sampling according to weights $w^\star$.

After conditioning on E we have

$$\mathbf{P}\left(\widehat{\gamma} = c^T \Sigma_{w^\star}^{+} \mathbb{X}^T \varepsilon / n \mid E\right) \to 1.$$

This implies that with probability $\mathbf{P}(E) \to 1$, the debiased estimator $\widehat{\gamma}$ is not consistent. Finally, note that this result does not depend on the specific choice of the relaxed inverse used in place of $\Sigma_{w^\star}^{-1}$.

The reason for the bias in this example is that the restricted eigenvalue of $\widehat{\Sigma}$ is zero with non-negligible probability. A simple solution to this problem is to constrain the population covariance matrix to have eigenvalues that are bounded away from zero. This motivates considering designs where for pre-specified $0 < \lambda_\star \le \lambda^\star < \infty$ the covariance matrix is constrained to have eigenvalues that are bounded away from 0 and ∞:

$$\begin{aligned} \mathbf{P2}: \min_{w \in \mathbb{R}^N} \quad & c^T \Sigma^{-1} c \\ \text{s.t.} \quad & \Sigma = \sum_{1 \le i \le N} w_i x_i x_i^T \\ & \lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I \\ & \sum_{1 \le i \le N} w_i = 1, \quad w \ge 0. \end{aligned} \tag{4}$$

While the original problem was equivalent to an LP, we show (in Proposition 2 in the supplement [20]) that the optimization problem $P2$ can be recast as the following semidefinite program (SDP):

$$\begin{aligned} \mathbf{P2'}: \min_{t \in \mathbb{R}, w \in \mathbb{R}^N} \quad & t \\ \text{s.t.} \quad & \Sigma = \sum_{i=1}^N w_i x_i x_i^T, \quad \sum_{i=1}^N w_i = 1 \\ & \lambda_\star I \preccurlyeq \Sigma \preccurlyeq \lambda^\star I, \quad \begin{bmatrix} t & c^T \\ c & \Sigma \end{bmatrix} \succcurlyeq 0 \\ & w \ge 0. \end{aligned}$$

Remark 2 (The advantanges of randomization). Randomized designs have important computational and statistical advantages in the high-dimensional setting. To see this, let us consider a natural notion of c-optimality for deterministic experiments. If $n_i \in \{0, 1, \ldots, n\}$ is the number of times x_i is repeated in the sample, then the variance of the debiased lasso estimator is approximately equal to

$$\frac{1}{n} c^T \widehat{\Omega} \widehat{\Sigma} \widehat{\Omega}^T c,$$

where $\widehat{\Sigma} = n^{-1} \sum_1^N n_i x_i x_i^T$ and $\widehat{\Omega}$ is a relaxed inverse of $\widehat{\Sigma}$ that is typically required to satisfy $\|c^T \widehat{\Omega} \widehat{\Sigma} - c^T\|_\infty \lesssim \sqrt{\log(p)/n}$ to ensure that the bias of the debiased lasso estimator remains $o(1/\sqrt{n})$. It is then natural to attempt to minimize the the above variance subject to the aforementioned constraint (see for

example a similar formulation for D-optimality in the work of Huang, Kong and Ai [16]):

$$\begin{aligned} \min_{\substack{(n_i)_1^N \in [n]^N \\ \widehat{\Omega} \in \mathbf{R}^{p\times p}}} \quad & c^T \widehat{\Omega}\widehat{\Sigma}\widehat{\Omega} c \\ \text{s.t.} \quad & \|c^T \widehat{\Omega}\widehat{\Sigma} - c^T\|_\infty \lesssim \sqrt{\frac{\log(p)}{n}} \\ & \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^N n_i x_i x_i^T \\ & \sum_1^N n_i \le n. \end{aligned}$$

Designs obtained using the above formulation suffer from two problems:

- **Computational feasibility.** The above is a non-convex problem in a high-dimensional space ($\mathbf{R}^{N+p^2}$) for which it can be very difficult to find global optima (even when we ignore the integer constraints on n_i). Randomization here allows to approximate $c^T\widehat{\Omega}\widehat{\Sigma}\widehat{\Omega}c$ with $c^T\widehat{\Omega}c$ for $\widehat{\Omega} := \Sigma^{-1}$, which is a convex function of the design $(w_i)_1^N$.
- **Statistical accuracy.** Even if a sufficiently good (locally) optimal solution of the above problem is found, guaranteeing statistical accuracy for the resulting design may not be possible, as the restricted eigenvalue condition may not be satisfied for it. Note that verifying the RE condition is NP-hard [11], so it is not known how to constrain the design in the above problem to satisfy the RE condition. Randomization solves this problem by using linear matrix inequalities (LMI) on the population covariance matrix, casting the problem as a semidefinite problem. The RE condition then is guaranteed to hold with high probability for the sample design matrix.

3. Application

Our example application[4] is inspired by sparse magnetic resonance imaging (MRI) [21]. At a high level, the measurements in an MRI application form a noisy Fourier transform of an underlying quantity (typically proton density of water molecules in a cross-section of body):

$$y = \mathcal{F} \cdot \mathrm{vec}(\theta^\star) + \varepsilon,$$

where $\beta^\star := \mathrm{vec}(\theta^\star)$ is a vectorized (i.e. one-dimensional) representation of the two-dimensional underlying truth $\theta^\star$, and $\mathcal{F}$ is the matrix of the discrete Fourier transform (in the same basis that $\beta^\star$ is represented), and ε is measurement noise taken to have a $N(0,1)$ distribution. For an $n \times p$ matrix A, the vectorized representation $\mathrm{vec}(A)$ is obtained by stacking the columns of A in an np vector as follows

$$\mathrm{vec}(A) = (A_{11}, \ldots, A_{n1}, A_{12}, \ldots, A_{n2}, \ldots, A_{1p}, \ldots, A_{np})^T.$$

In practice, obtaining the MRI measurements is slow and time-consuming, a problem that has inspired a vast literature (under the umbrella of "compressed sensing") on methods allowing reduction of the number of such measurements while preserving the quality of the reconstructed image. If $\beta^\star$ is sparse, one of the most well-known methods of reconstruction is the Lasso

$$\widehat{\beta} := \arg\min_{\beta'} \left\{ \frac{1}{2n}\|y - \mathcal{F}\beta'\|_2^2 + \lambda\|\beta'\|_1 \right\}.$$

[4]Python code for both numerical experiments is available at https://github.com/ehamid/HD_DoE

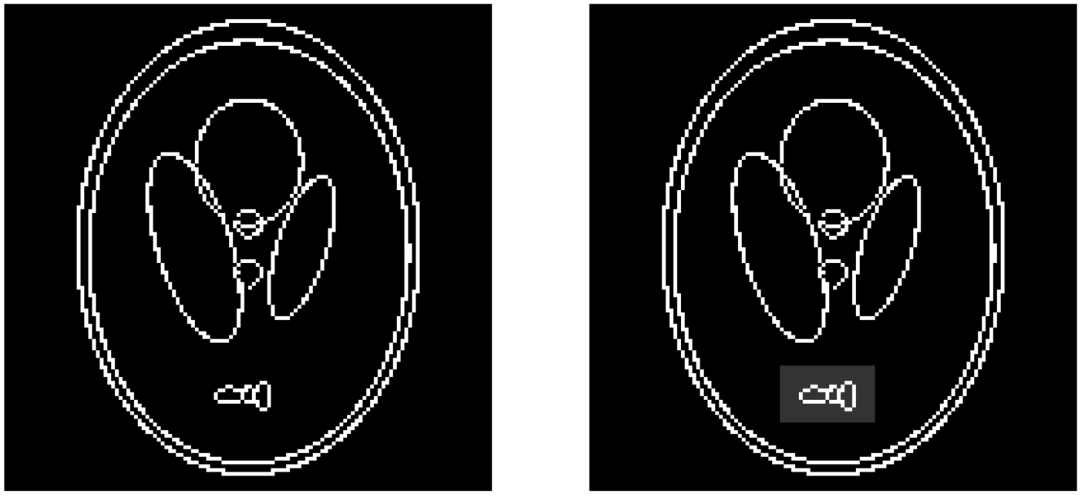

Figure 1. The sparsified Shepp-Logan phantom as the underlying truth for our experiments (left). The gray rectangle shows the region whose average intensity (γ) is the target of inference.(right)

Often, $\beta^\star$ is sparse, not in the original basis, but rather in a different basis, e.g. the wavelet basis W. In this case one solves the transformed problem [7]:

$$\widehat{\beta} := \arg\min_{\beta'} \left\{ \frac{1}{2n} \|y - \mathcal{F}\beta'\|_2^2 + \lambda \|W\beta'\|_1 \right\}.$$

Alternatively, if β has a small total variation, for example if the image comprises large piece-wise constant parts, one can use a total variation penalty term [30]:

$$\widehat{\beta} := \arg\min_{\beta'} \left\{ \frac{1}{2n} \|y - \mathcal{F}\beta'\|_2^2 + \lambda \|\nabla\beta'\|_1 \right\}.$$

where ∇ is a discrete gradient operator that yields sparse representations of the class of images under study.

In order to make the discussion of our experiments more straightforward, we make the following simplifications:

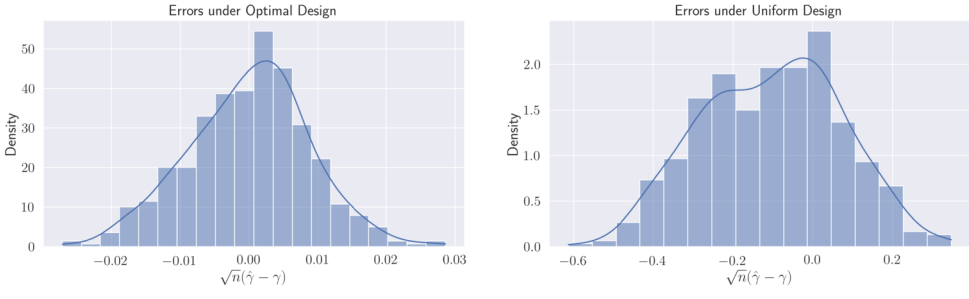

Figure 2. Comparison of $\sqrt{n}$-scaled estimation errors $\sqrt{n}(\widehat{\gamma} - \gamma)$ for the optimal and uniform designs

- We start with a sparse image (Figure 1[left]), so that we do not need to take the sparsifying transforms (wavelet, gradient, etc.) into account. If we denote the original image (in our case, the Shepp-Logan phantom) by ι, then the sparse image is taken to be the sign of the Laplacian of the original image, that is, $\theta^\star = \mathbf{1}(\Delta\iota > 0)$ where $\Delta\iota := \partial_x^2\iota + \partial_y^2\iota$ is the Laplacian operator typically used for edge detection in image processing and where the ∂_x, ∂_y are Sobel derivatives (see [22, Chapter 10] for details). In the resulting image only 6.3% of the pixels are non-zero while in the original image 43% are non-zero.
- We use the discrete cosine transform (DCT) [1] to avoid complex-valued Fourier measurements that are obtained in MRI applications. Alternatively, one could also consider the Fourier transform as pairs of cosine and sine transforms, resulting in a design matrix with twice as many rows as the (complex) Fourier transform matrix.

The number of pixels in the image on the left of Figure 1 is 128×128. Thus for this application $N = p = 128^2 = 16384$ and we set $n = N/2$, so that we are under-sampling the DCT coefficients by 50%. The design matrix is the matrix of 2-dimensional discrete cosine transform of type 2, given by $X = \sqrt{N} \cdot (D \otimes D)^T$ where $\otimes$ denotes the Kronecker product[5] and

$$D_{ij} = \begin{cases} \dfrac{1}{\sqrt{m}} & : i = 1 \\ \sqrt{\dfrac{2}{m}} \cos\left(\dfrac{\pi(i-1)(2j-1)}{2m}\right) & : 2 \le i \le m = \sqrt{p}. \end{cases} \tag{5}$$

The scaling by $\sqrt{N}$ is done to ensure $X^T X/N = I_N$, meaning that under uniform sampling, the population design covariance is the identity matrix (and thus has well-behaved extreme eigenvalues).

The c vector is taken to be $c = \mathrm{vec}(\tilde{c})$ where:

$$\tilde{c}_{ij} = \begin{cases} \dfrac{1}{375} & : 95 \le i < 110 \text{ and } 50 \le j < 75 \\ 0 & : \text{otherwise.} \end{cases}$$

This choice of c results in $\gamma = \langle c, \beta^\star \rangle = \langle \tilde{c}, \theta^\star \rangle$ being the average intensity of the image on the support of $\tilde{c}$ (the gray rectangle in Figure 1[right]), that is, $\gamma = \sum_{95 \le i < 110} \sum_{50 \le j < 75} \theta^\star_{ij}/375$. With the above choice of c, the parameter of interest $\gamma = \langle c, \beta^\star \rangle = 0.10667$. The optimal design is then obtained by solving the problem[6] $P1'$.

Figure 2 shows the histograms of estimation errors over 500 i.i.d. realizations of the estimates under uniform and optimal designs. The scaled bias and standard errors are presented in Table 1. As can be seen in the table, the optimal design leads to significant improvements in both bias and variance of the estimator. In practice, this translates to higher detection rates of weak signals in images and higher power for hypothesis tests.

Simulated Data Example. In our second experiment, we compare the performance of the debiased estimator under uniform, optimal and n-Nearest-Neighbors[7] designs for an experimental domain generated from a multivariate normal distribution. For each combination of problem parameters, $N = 1000$

[5] Here D is the matrix of DCT (type 2) for 1-dimensional signals, and the Kronecker product computes the 2-dimensional DCT matrix for vectorized images.

[6] Linear programs can be solved much faster than semidefinite programs. See remark 3 at the end of this section for an explanation of why we we use the optimization problem $P1'$ rather than $P2$.

[7] By the n-Nearest-Neighbors design we mean a deterministic design consisting of the n closest points to c in Euclidean distance.

Design / Measure	$\sqrt{n}\times$ bias	$\sqrt{n}\times$ standard error
Optimal Design	4.1564×10^{-5}	0.0088
Uniform Design	−0.1042	0.1751

Table 1. Comparison of bias and standard error for the optimal and uniform designs

observations $(x_i)_1^N$ were drawn from $N(0_p, \Sigma_\kappa)$ where $(\Sigma_\kappa)_{ij} = \kappa^{|i-j|}$ and $\kappa \in \{0.1, 0.9\}$ and $p = 500$. The regression parameter $\beta^\star$ with two sparsity levels $s \in \{5, 10\}$ was taken to be

$$\beta_j^\star = \begin{cases} \dfrac{s+1-j}{s} & : j \le s, \\ 0 & : s < j \le p. \end{cases}$$

Two different choice of $c \in \{e_s, e_{s+1}\}$ were considered that correspond to estimating $\gamma = \beta_s = 1/s$ and $\gamma = \beta_{s+1} = 0$. The optimal design was found in each case using the linear program (1) (resulting in an optimal design with non-singular population covariance matrix). To approximate the distribution of the debiased lasso estimator for each specification of the problem (i.e. choices of κ, s and c), 500 Monte Carlo samples $\widehat{\gamma}$ were generated as follows. Each time a sample $(X_i)_1^n$ of size $n = 200$ was drawn from $(x_i)_1^N$ according to the uniform or optimal design, the responses were generated according to $y_i = \langle \beta^\star, X_i \rangle + \varepsilon_i$ with $\varepsilon_i \overset{iid}{\sim} N(0,1)$, and finally the debiased 95% confidence intervals were generated using

$$\widehat{\gamma} \pm z_{0.025} \cdot \sqrt{\frac{c^T \Sigma^{-1} \widehat{\Sigma} \Sigma^{-1} c}{n}}$$

with $z_{0.025}$ denoting the upper 2.5 percentile of a $N(0,1)$ distribution and where $\Sigma = N^{-1} \sum_1^N x_i x_i^T$ in the case of uniform design and $\Sigma = \sum_1^N w_i^\star x_i x_i^T$ under the optimal design $w^\star$. In the case $\gamma = \beta_s = s^{-1}$ (which we call the "signal" scenario), power was computed as the proportion of times (out of the 500 realizations) the 95% confidence interval did not include zero. In the case $\gamma = \beta_{s+1} = 0$ (which we call the "noise" scenario), the false positive rate is computed as the proportion of times (out of the 500 realizations) the 95% confidence interval does not include zero. In both cases the coverage is the proportion of times the confidence interval covers the parameter of interest. The standard errors were computed as the sample standard deviation of the 500 realizations of $\widehat{\gamma}$ in each case.

The results of the experiment are presented in Table 2 and Table 3 for the signal and noise scenarios, respectively. As can be seen in Table 2, the optimal design results in improved power (compared to the uniform design) while controlling the coverage close to the nominal level. It is also observed in Table 2 that in the non-zero signal scenario the n-Nearest-Neighbors design has very poor coverage, far from the nominal 95% rate. The n-Nearest-Neighbors design also fails to control the false positive rate as is evident from Table 3. These shortcomings of the n-Nearest-Neighbors design are caused by the non-negligible bias of the estimates compared to their standard errors when using such singular designs.

Remark 3. For both of our experiments, we have used the linear program $P1'$ to find the optimal design, which in both cases resulted in designs with non-singular population covariance matrices. In practice, for large-scale problems it is often much faster to solve the linear program $P1'$ versus the semidefinite program $P2$. Therefore one can first quickly solve $P1'$, and resort to $P2$ only when the covariance matrix resulting from $P1'$ is (close to) singular.

(κ, s)	Design	Power	Coverage	$\sqrt{n}\times$Standard Error
(0.1, 5)	nNN	0.678	0.4	0.890
	U	0.496	0.952	1.42
	O	0.806	0.928	1.00
(0.1, 10)	nNN	0.432	0.78	0.746
	U	0.152	0.936	1.49
	O	0.29	0.926	1.05
(0.9, 5)	nNN	0.428	0.76	0.759
	U	0.096	0.97	4.02
	O	0.156	0.938	2.96
(0.9, 10)	nNN	0.238	0.816	0.604
	U	0.064	0.938	4.46
	O	0.078	0.944	2.85

Table 2. The non-zero signal scenario : Performance of de-bisaed lasso under uniform and optimal designs for estimating $\gamma = \beta_s = s^{-1}$. The abbreviations nNN, U and O signify n-Nearest-Neighbors, Uniform and Optimal design, respectively.

4. Discussion and future work

In this work we studied the problem of optimal design under the c-optimality criterion in the high-dimensional setting where $n < p$. We have proposed a semidefinite program that minimizes the variance of debiased estimators while allowing the derivation of theoretical guarantees for the resulting design. We have shown that in contrast to the low-dimensional setting where the OLS estimator is unbiased, in the high-dimensional setting one needs to control the bias when minimizing the variance of debi-

(κ, s)	Design	False Positive Rate	Coverage	$\sqrt{n}\times$Standard Error
(0.1, 5)	nNN	0.062	0.938	0.560
	U	0.054	0.946	1.44
	O	0.048	0.952	0.97
(0.1, 10)	nNN	0.056	0.944	0.518
	U	0.058	0.942	1.44
	O	0.064	0.936	1.00
(0.9, 5)	nNN	0.1	0.9	0.466
	U	0.048	0.952	4.19
	O	0.042	0.958	2.84
(0.9, 10)	nNN	0.064	0.936	0.461
	U	0.046	0.954	4.25
	O	0.056	0.944	2.91

Table 3. The noise scenario: Performance of debisaed lasso under uniform and optimal designs for estimating $\gamma = \beta_{s+1} = 0$. The abreviations nNN, U and O signify n-Nearest-Neighbors, Uniform and Optimal design, respectively.

ased estimators. The practical efficiency gains from these optimal designs are presented in simulation experiments and are quite impressive.

Our work focuses on the setting where $p \leq N$. However the $N < p$ case can also arise in applications where the experimental domain is not as large and poses interesting challenges (such as the difficulty of enforcing RE conditions on the design as discussed in the introduction) for future work. Other directions for further research include the study of optimal designs for generalized linear models in a high-dimensional setting and under other optimality criteria than c-optimality. Yet another direction for extending our results is a study optimal designs for $A\beta$ for a matrix $A \in \mathbb{R}^{q \times p}$ with q a fixed integer.

Supplementary Material

Supplement to "Design of c-optimal experiments for high-dimensional linear models" (DOI: 10.3150/22-BEJ1472SUPP; .pdf). The proofs of the results in the paper and backgorund material are provided in the Supplementary Material [20].

References

[1] Ahmed, N., Natarajan, T. and Rao, K.R. (1974). Discrete cosine transform. *IEEE Trans. Comput.* **C-23** 90–93. MR0356555 https://doi.org/10.1109/t-c.1974.223784

[2] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[3] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge Univ. Press. MR2061575 https://doi.org/10.1017/CBO9780511804441

[4] Cai, T., Cai, T. and Guo, Z. (2019). Individualized treatment selection: An optimal hypothesis testing approach in high-dimensional models. arXiv preprint. Available at arXiv:1904.12891.

[5] Cai, T.T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

[6] Černý, M. and Hladík, M. (2012). Two complexity results on c-optimality in experimental design. *Comput. Optim. Appl.* **51** 1397–1408. MR2891943 https://doi.org/10.1007/s10589-010-9377-8

[7] Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159. MR1854649 https://doi.org/10.1137/S003614450037906X

[8] Chernoff, H. (1953). Locally optimal designs for estimating parameters. *Ann. Math. Stat.* **24** 586–602. MR0058932 https://doi.org/10.1214/aoms/1177728915

[9] Dette, H. et al. (1993). Elfving's theorem for D-optimality. *Ann. Statist.* **21** 753–766. MR1232517 https://doi.org/10.1214/aos/1176349149

[10] Dette, H. and Holland-Letz, T. et al. (2009). A geometric characterization of c-optimal designs for heteroscedastic regression. *Ann. Statist.* **37** 4088–4103. MR2572453 https://doi.org/10.1214/09-AOS708

[11] Dobriban, E. and Fan, J. (2016). Regularity properties for sparse regression. *Commun. Math. Stat.* **4** 1–19. MR3475839 https://doi.org/10.1007/s40304-015-0078-6

[12] Elfving, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Stat.* **23** 255–262. MR0047998 https://doi.org/10.1214/aoms/1177729442

[13] Fan, S.K. and Chaloner, K. (2003). A geometric method for singular c-optimal designs. *J. Statist. Plann. Inference* **113** 249–257. MR1963044 https://doi.org/10.1016/S0378-3758(01)00289-0

[14] Gu, L. and Yang, J.-F. (2013). Construction of nearly orthogonal Latin hypercube designs. *Metrika* **76** 819–830. MR3085834 https://doi.org/10.1007/s00184-012-0417-5

[15] Harman, R. and Jurík, T. (2008). Computing c-optimal experimental designs using the simplex method of linear programming. *Comput. Statist. Data Anal.* **53** 247–254. MR2649082 https://doi.org/10.1016/j.csda.2008.06.023

[16] Huang, Y., Kong, X. and Ai, M. (2020). Optimal designs in sparse linear models. *Metrika* **83** 255–273. MR4057422 https://doi.org/10.1007/s00184-019-00722-9

[17] Javanmard, A. and Lee, J.D. (2020). A flexible framework for hypothesis testing in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 685–718. MR4112781
[18] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[19] Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R.M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Min. Knowl. Discov.* **18** 140–181. MR2469594 https://doi.org/10.1007/s10618-008-0114-1
[20] Eftekhari, H., Banerjee, M. and Ritov, Y. (2023). Supplement to "Design of c-optimal experiments for high-dimensional linear models." https://doi.org/10.3150/22-BEJ1472SUPP
[21] Lustig, M., Donoho, D. and Pauly, J.M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **58** 1182–1195.
[22] Misra, S. and Wu, Y. (2020). *Machine Learning Assisted Segmentation of Scanning Electron Microscopy Images of Organic-Rich Shales with Feature Extraction and Feature Ranking*. Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-817736-5.00010-7
[23] Pázman, A. and Pronzato, L. (2006). On the irregular behavior of LS estimators for asymptotically singular designs. *Statist. Probab. Lett.* **76** 1089–1096. MR2269278 https://doi.org/10.1016/j.spl.2005.12.010
[24] Pronzato, L. (2009). On the regularization of singular c-optimal designs. *Math. Slovaca* **59** 611–626. MR2557313 https://doi.org/10.2478/s12175-009-0151-2
[25] Pukelsheim, F. (2006). *Optimal Design of Experiments*. *Classics in Applied Mathematics* **50**. Philadelphia, PA: SIAM. MR2224698 https://doi.org/10.1137/1.9780898719109
[26] Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika* **79** 763–770. MR1209476 https://doi.org/10.1093/biomet/79.4.763
[27] Ravi, S.N., Ithapu, V., Johnson, S. and Singh, V. (2016). Experimental design on a budget for sparse linear models and applications. In *International Conference on Machine Learning* 583–592.
[28] Rodríguez-Díaz, J.M. (2017). Computation of c-optimal designs for models with correlated observations. *Comput. Statist. Data Anal.* **113** 287–296. MR3662408 https://doi.org/10.1016/j.csda.2016.10.019
[29] Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory* 10–1.
[30] Rudin, L.I., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom.* **60** 259–268.
[31] Sagnol, G. (2011). Computing optimal designs of multiresponse experiments reduces to second-order cone programming. *J. Statist. Plann. Inference* **141** 1684–1708. MR2763200 https://doi.org/10.1016/j.jspi.2010.11.031
[32] Seeger, M.W. (2008). Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.* **9** 759–813. MR2417254 https://doi.org/10.1017/s0143385700005320
[33] Stoica, P. and Babu, P. (2010). Algebraic derivation of elfving theorem on optimal experiment design and some connections with sparse estimation. *IEEE Signal Process. Lett.* **17** 743–745. https://doi.org/10.1109/LSP.2010.2053533
[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[35] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
[36] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge: Cambridge Univ. Press. MR3837109 https://doi.org/10.1017/9781108231596
[37] Zhang, C.-H. and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

Received September 2021