On Computational Complexity of Finding *c*-optimal Experimental Designs over a Finite Experimental Domain

How to Break RSA Using Algorithms for *c*-optimal Designs

Michal Černý, Milan Hladík, Veronika Skočdopolová

University of Economics, Prague; Charles University, Prague

Motivation. In a traditional linear regression model $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ with uncorrelated homoskedastic observations, our aim is to estimate a *linear combination of regression parameters* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{c} \neq \mathbf{0}$, with OLS as precisely as possible.

Examples. The choice $\boldsymbol{c}^{\mathrm{T}} = (1, 0, \ldots, 0)$ leads to the estimation of the first regression coefficient. In case of the Cobb-Douglas production function

$$\ln Y = \sum_{i=1}^{n-1} \beta_i \ln F_i + \beta_n,$$

where Y is output and $F_1, \ldots, F_{n-1}$ are production factors, the choice $\boldsymbol{c}^{\mathrm{T}} = (1, \ldots, 1, 0)$ leads to the estimation of returns to scale.

Experimental domain. We study the case that the experimental domain is *finite* and *rational*. Denote it $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ ($\subseteq \mathbb{R}^p$).

Definition. A regression design matrix $\boldsymbol{X}$ is $\mathcal{X}$*-correct*, if each row $\boldsymbol{x}^{\mathrm{T}}$ of $\boldsymbol{X}$ fulfills $\boldsymbol{x} \in \mathcal{X}$. It may be also described in terms of a *design vector* $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^{\mathrm{T}}$ satisfying $\boldsymbol{\xi} \geq \mathbf{0}$, $\sum_i \xi_i = 1$ with the meaning that the matrix $\boldsymbol{X}$ has $100\xi_i\%$ rows $\boldsymbol{x}_i^{\mathrm{T}}$, $i = 1, \ldots, k$.

$\boldsymbol{c}$-variance. Let $\boldsymbol{X}$ be an $\mathcal{X}$-correct matrix, let $\boldsymbol{\xi}$ its associated design and let $\widehat{\boldsymbol{\beta}}$ be the OLS-estimator of $\boldsymbol{\beta}$. Then $\mathrm{var}(\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{N} \cdot \mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi})$, where σ^2 is the variance of error terms, N stands for the number of observations and

$$\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{X}) := \mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}) := \boldsymbol{c}^{\mathrm{T}} \left(\sum_{i=1}^{k} \xi_i \cdot \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right)^{-1} \boldsymbol{c},$$

where $(\cdot)^{-1}$ stands for the matrix (pseudo)inverse. Obviously, $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi})$ measures the contribution of the design $\boldsymbol{\xi}$ to the total variance of $\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

Problem statement. Exact version. *Input:* a finite rational experimental domain $\mathcal{X}$, a rational vector $\boldsymbol{c} \neq \boldsymbol{0}$ and a natural number N.

Output: An N-row $\mathcal{X}$-correct matrix such $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{X})$ is minimal (i.e., for any N-row $\mathcal{X}$-correct matrix $\boldsymbol{X}'$ it holds $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{X}) \leq \mathrm{var}_{\boldsymbol{c}}(\boldsymbol{X}')$).

Problem statement. Approximate (or: asymptotic) version. *Input:* a finite rational experimental domain $\mathcal{X}$ and a rational vector $\boldsymbol{c} \neq \boldsymbol{0}$.

Output: A design $\boldsymbol{\xi}$ over the domain $\mathcal{X}$ such that $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi})$ is minimal (i.e., for any design $\boldsymbol{\xi}'$ over the domain $\mathcal{X}$, it holds $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}) \leq \mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}')$).

Said loosely. *Exact version:* Given N (standing for the number of observations), find "the best" design $\boldsymbol{\xi}$ such that $N\boldsymbol{\xi}$ is integral. *Approximate version:* do not care about integrality.

Theorem [Harman, Jurík, 2008]. The approximate version of the problem is solvable via linear programming.

Corollary 1. The approximate version is solvable in polynomial time.

Corollary 2. Any approximately optimal design is N-exact for some N.

(We know some estimates on such N, but they do not seem to be very useful in practice; for example, N can be exponential in the size of the experimental domain; but, possibly, in special cases this can be improved.)

For complexity-theoretic classification we need decision versions of the problems.

Exact version (EOD): Given N, $\boldsymbol{c}$, $\mathcal{X}$ and S^2, is there an N-row $\mathcal{X}$-exact matrix $\boldsymbol{X}$ satisfying $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{X}) \leq S^2$? Or: is it possible to design an N-exact experiment with $\boldsymbol{c}$-variance at most S^2?

Approximate version (AOD): Given $\boldsymbol{c}$, $\mathcal{X}$ and S^2, is there a design $\boldsymbol{\xi}$ satisfying $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}) \leq S^2$? Equivalently: is it possible to find an N and an N-exact experiment with $\boldsymbol{c}$-variance at most S^2?

Theorem [Černý, Hladík, 2010] *EOD* is $\boldsymbol{NP}$-complete.

Theorem [Černý, Hladík, 2010] *AOD* is $\boldsymbol{P}$-complete.

To recall: a set A is $\boldsymbol{P}$*-complete*, if any set in $\boldsymbol{P}$ (the class of sets decidable in Turing polynomial time) is reducible to A via a function computable in Turing logarithmic space.

A set A is $\boldsymbol{NP}$*-complete*, if any set in $\boldsymbol{NP}$ (the class of sets decidable in Turing nondeterministic polynomial time) is reducible to A via a function computable in Turing polynomial time.

Consequences of $\boldsymbol{P}$-completeness of AOD (under some broadly-accepted complexity-theoretic conjectures).

- The problem is not in the $\boldsymbol{NC}$-hierarchy. (Recall that $\boldsymbol{NC}$, the Nick's Class, is the class of problems that are said to be "well-computable in parallel", i.e. problems decidable with circuits of polynomial size and polylogarithmic depth.) Hence, *AOD is not well-computable in parallel*. So we cannot expect that the problem could be solvable by parallel systems much faster than by sequential computers.

- General linear programming is reducible to AOD, i.e. any algorithm for AOD is able to solve any general linear program. So, any designer of an algorithm for AOD is, in fact, designing a general-purpose algorithm for linear programming. (This gives some limits to such a designer. On the other hand: could this approach bring some new ideas to the theory of linear programming algorithms?)

Consequences of *NP*-completeness of EOD (under some broadly-accepted complexity-theoretic conjectures).

- The problem is not decidable in polynomial time.
- A nice example: any algorithm for EOD is able to break the RSA cryptographic protocol.

How to do that? The RSA protocol relies on the following belief. Given two primes p_1 and p_2, let $n := p_1 p_2$. The problem *given* n*, find* p_1 *and* p_2 is believed to be extremely difficult.

We can do this. It is easy to write down a boolean formula $f(p_1, p_2, n)$ (where p_1, p_2, n are regarded as bit strings) such as f is true if and only if $n = p_1 p_2$. We substitute the bits of n into f as constants and leave the bits p_1 and p_2 as free variables. Then, breaking RSA is equivalent to finding any satisfying assignment (p_1, p_2) to f.

We can convert f into an instance of EOD. We can show that *from the optimal design found by any algorithm for EOD we can recover the satisfying assignment to* f, and hence to find the two prime factors.

By the way: this is a nice testing instance for any such algorithm.

Unnatural instances of the design problem. The statement of the problem EOD is so general such that it admits instances that "a statistician would never think of", here called "unnatural". For example: the instance for factoring an n-bit integer requires dimension $\approx 16n^2$ (dimension = number of regression parameters).

It is a usual situation in complexity theory: from the large space of all instances, the theory selects a (usually small) subset, sometimes called *complexity core*, making the problem difficult. Often it happens that the core instances are unnatural for the theory which motivated the formulation of the problem.

At present, we cannot find an instance of the design problem that would be both hard (i.e. sufficient to prove ***NP***-completeness) and natural for statistics.

Question. *Is it possible to define, in an exact sense, what the "natural instance" of the design problem is?* Is it possible to define a restriction of the design problem that would rule out unnaturalness? (Of course, we cannot e.g. restrict dimension, as complexity theory always studies asymptotic behaviour.) Then, is the problem restricted to the natural instances ***NP***-complete again? Is the property "being natural" decidable in polynomial time?

Thank you for attention.