# Near-Optimal Discrete Optimization for Experimental Design: A Regret Minimization Approach[*]

**Zeyuan Allen-Zhu**
Microsoft Research Redmond
zeyuan@csail.mit.edu

**Yuanzhi Li**
Princeton University
yuanzhil@cs.princeton.edu

**Aarti Singh**
Carnegie Mellon University
aarti@cs.cmu.edu

**Yining Wang**
Carnegie Mellon University
yiningwa@cs.cmu.edu

## Abstract

The experimental design problem concerns the selection of $k$ points from a potentially large design pool of $p$-dimensional vectors, so as to maximize the statistical efficiency regressed on the selected $k$ design points. Statistical efficiency is measured by *optimality criteria*, including A(verage), D(eterminant), T(race), E(igen), V(ariance) and G-optimality.

We propose a poly-time regret minimization framework to achieve a $(1+\varepsilon)$ approximation with $O(p/\varepsilon^2)$ design points, for all the optimality criteria above.

In contrast, to the best of our knowledge, before our work, no polynomial-time algorithm achieves $(1+\varepsilon)$ approximations for D/E/G-optimality, and the best poly-time algorithm achieving $(1+\varepsilon)$-approximation for A/V-optimality requires $k=\Omega(p^2/\varepsilon)$ design points.

## 1 Introduction

Let $x_1,\ldots,x_n \in \mathbb{R}^p$ be $p$-dimensional vectors and $f : \mathbb{S}_p^+ \to \mathbb{R}^+$ be a non-negative function defined over $\mathbb{S}_p^+$, the class of all $p$-dimensional positive definite matrices. We focus on the design of polynomial-time algorithms for approximately solving the following *discrete* optimization problem:

$$\min_{s\in\mathcal{S}_k} F(s) = \min_{s\in\mathcal{S}_k} f\left(\sum_{i=1}^n s_i \cdot x_i x_i^\top\right) \quad \text{where} \quad \mathcal{S}_k := \Big\{s\in\{0,1\}^n,\ \sum_{i=1}^n s_i \le k\Big\}\ . \tag{1.1}$$

In other words, we wish to select a subset $S\subset[n]$ of cardinality at most $k$, so that its covariance matrix $\Sigma_S=\sum_{i\in S} x_i x_i^\top$ has the smallest function value $f(\Sigma_S)$. The main challenge of solving Problem (1.1) is the discrete constraint $s\in\{0,1\}^n$.

**Classical experimental design** The *(classical) experimental design* problem concerns the selection of $k$ points from a potentially very large design pool $\{x_1,\ldots,x_n\}$ so as to maximize the *statistical efficiency* regressed on the selected $k$ design points.

For example, consider a clinical study application where $n$ is the number of patients; $p$ is the number of parameters (e.g., blood pressure, low-density lipoprotein, etc.) that are hypothesized to affect some disease; and $x_1,\ldots,x_n\in\mathbb{R}^p$ are the parameters for all the patients. Since determining whether or not a patient has a certain disease may be expensive or time-consuming, one wishes to select $k \ll n$ patients that are the most *statistically efficient* for establishing a regression model that connects experimental parameters to the disease.

[*] Full version available at http://arxiv.org/abs/1711.05174.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

This experimental design problem reduces to Problem (1.1), where the evaluation of statistical efficiency is reflected in the choice of the objective function $f$, known as the *optimality criterion* [1]. Popular choices of $f$ include

- *A(verage)-optimality* $f_A(\Sigma) = \mathrm{tr}(\Sigma^{-1})/p$,
- *D(eterminant)-optimality* $f_D(\Sigma) = (\det|\Sigma|)^{-1/p}$,
- *T(race)-optimality* $f_T(\Sigma) = p/\mathrm{tr}(\Sigma)$,
- *E(igen)-optimality* $f_E(\Sigma) = \|\Sigma^{-1}\|_2$,
- *V(araince)-optimality* $f_V(\Sigma) = \frac{1}{n}\mathrm{tr}(X\Sigma^{-1}X^\top)$, and
- *G-optimality* $f_G(\Sigma) = \max\mathrm{diag}(X\Sigma^{-1}X^\top)$.

We refer the readers to [1] for a complete list and discussion of various optimality criteria used in the experimental design literature.

**Other applications** In the full version we shall also discuss applications to Bayesian experimental design, active learning and graph signa processing.

## References

[1] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.