https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–12
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Technical Note—A New Rate-Optimal Sampling Allocation for Linear Belief Models

Jiaqi Zhou,[a] **Ilya O. Ryzhov**[b,*]

[a] Applied Mathematics & Statistics, and Scientific Computation, University of Maryland, College Park, Maryland 20742; [b] Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742

*Corresponding author

Contact: emily719@umd.edu (JZ); iryzhov@rhsmith.umd.edu, https://orcid.org/0000-0002-4191-084X (IOR)

Received: September 18, 2020
Revised: December 14, 2021; March 14, 2022
Accepted: May 31, 2022
Published Online in Articles in Advance: July 6, 2023

Area of Review: Simulation

https://doi.org/10.1287/opre.2022.2337

Abstract. We derive a new optimal sampling budget allocation for belief models based on linear regression with continuous covariates, where the expected response is interpreted as the value of the covariate vector, and an "error" occurs if a lower-valued vector is falsely identified as being better than a higher-valued one. Our allocation optimizes the rate at which the probability of error converges to zero using a large deviations theoretic characterization. This is the first large deviations-based optimal allocation for continuous decision spaces, and it turns out to be considerably simpler and easier to implement than allocations that use discretization. We give a practicable sequential implementation and illustrate its empirical potential.

Funding: This work was supported by the National Science Foundation [Grant CMMI-2112828].

Keywords: optimal learning • simulation optimization • linear regression • large deviations

1. Introduction

Consider the linear regression model

$$y = \beta^\top x + \varepsilon, \tag{1}$$

where $\beta \in \mathbb{R}^d$ is a fixed, but unknown vector of regression coefficients, $x \in \mathbb{R}^d$ is a vector of data, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is residual noise. The expectation $\mathbb{E}(y \mid x) = \beta^\top x$ is interpreted as the "value" of x. For example, the elements of x could represent various attributes of a combination treatment for cancer, with the response y being the health outcome (Bertsimas et al. 2016). We assume that x is "better" if $\mathbb{E}(y \mid x)$ is larger. The set of possible x need not be discrete.

Suppose that we have the ability to choose the data vector: given a sample size of n, we may choose $x_1, \ldots, x_n$ anywhere in some compact subset of $\mathbb{R}^d$ called the "design space." This choice may be made either all at once, before any observations are collected, or sequentially, where each x_m may depend on $x_1, y_1, \ldots, x_{m-1}, y_{m-1}$, perhaps through a vector b_m of least-squares regression coefficients estimated using these previously collected data. The first, static setting has been extensively studied in the literature on *experimental design* (Dette 1997, Salagame and Barton 1997). In this literature, the expected response $\beta^\top x$ is not used to evaluate or compare designs, and the goal is to improve the overall accuracy of the least squares estimator b_n in some aggregate sense. Typically, one builds the design to optimize some summary statistic of the covariance matrix of b_n. There are many possible criteria, known by such "alphabet-optimal" names as A-optimality (Ahipaşaoglu 2015), D-optimality (Sagnol and Harman 2015, Pokhilko et al. 2019), G-optimality (Rodriguez et al. 2010), and so on. Because of the properties of linear regression models, such criteria can be computed without knowledge of β and thus do not require any information on the response.

The second, sequential setting has been considered by the community working on simulation-based optimization. This literature grew out of the *ranking and selection* problem, in which the goal is to identify the highest-valued alternative (unlike experimental design, ranking and selection always has some notion of value to maximize) from some finite set using independent samples of the value. An early effort to apply algorithmic concepts from ranking and selection to the linear regression setting was by Negoescu et al. (2011), who also assumed that each x_m could take values only in a finite set; similar settings were considered by Shen et al. (2017) and Gao et al. (2019). Han et al. (2016) provided approximation algorithms for combinatorial design spaces, whereas Brantley et al. (2013, 2014) handled low-dimensional, continuous design spaces with special structure (e.g., the value being a quadratic function of a scalar control). In the computer science literature, Dani et al. (2008), Abbasi-Yadkori et al. (2011), and others studied related "linear bandit" problems where one maximizes the total value of the sampled vectors.

However, there is a growing stream of literature that examines the problem of identifying the best x from the viewpoint of static information collection, somewhat like experimental design. In the simulation community, Glynn and Juneja (2004) used large deviations theory to derive a tractable form for the asymptotic convergence rate of the probability of incorrect selection (the event that a suboptimal alternative is erroneously estimated to have a higher value than the optimal one). Given a finite set of alternatives, one allocates the sampling budget among them to speed up this convergence; essentially, the convergence rate becomes a kind of optimality criterion admitting a new type of "design." Similar ideas motivate the literature on optimal computing budget allocation (Chen et al. 2000, 2015; Chen and Lee 2010), which uses various approximations of this error probability. Later work by Pasupathy et al. (2014), Gao et al. (2017), and Applegate et al. (2020) generalized this notion to a broader class of simulation-based optimization problems. In all these papers, both the criterion and the optimal allocation depend on the underlying unknown problem parameters (in regression, this is the vector β) that determine the value of each x. This is a significant departure from the experimental design literature, which generally refrains from including such parameters in the model, but conceptually one may think of this approach as leveraging ideas from experimental design to address other problem classes focusing on value maximization. The computer science literature has also studied similar ideas, with Soare et al. (2014) and Fiez et al. (2019) proposing variants of G-optimal design for sequential learning.

In this paper, we derive and optimize a new, large deviations theoretic optimality criterion for linear regression. We do *not* discretize the design space, unlike all the existing work on large deviations-based allocations (even Yakowitz et al. 2000, who study a continuous problem, require discretization). Rather, we allow any x on the L^2 sphere $\{x : \|x\| = 1\}$, which can be generalized to other design spaces with dimension d. The analysis requires substantial new technical developments over past work (which is limited to finite sets) and leads to a completely different interpretation of the allocation. In Glynn and Juneja (2004) and related papers, each alternative is assigned a certain nonzero proportion of the sample, which is no longer possible when x is a continuous variable. However, in the regression context, we find that samples should be allocated to the elements of an *orthonormal basis* for the design space, with β itself being one of the basis vectors. We then obtain exceptionally simple closed-form calculations for the optimal proportions to assign to each basis vector. In fact, these optimal proportions are *almost* uniform: one samples β with a certain small probability (computable in closed form) that does not depend on β itself, and otherwise chooses one of the other basis vectors uniformly at random.

In general, because optimal allocations depend on unknown parameters, they cannot be computed statically (again, unlike optimal designs), but rather must be learned over time. For problems with discrete design spaces, such as ranking and selection, this is a difficult problem, because optimal allocations require enumeration of all possible alternatives and make a special distinction between the allocation to the best alternative versus all the others. See Chen and Ryzhov (2019a, b) for algorithmic approaches in such settings. However, in the continuous setting of this paper, the optimal allocation is much easier to learn: by changing the focus to an orthonormal basis for the design space, which depends only on β, we decouple the allocation from the actual x whose values are being compared. The simplicity of this approach makes it attractive as a benchmark for continuous optimal budget allocation, and the concept of sampling a basis may be of interest for future work on other classes of continuous problems.

2. Large Deviations in Least Squares Regression

Return to (1) and assume, without loss of generality, that $\|\beta\| = 1$. Suppose that $\{x_n\}_{n=1}^{\infty}$ is a deterministic sequence satisfying

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} x_m x_m^\top = A, \tag{2}$$

where A is a symmetric, positive definite matrix. Although we will treat the data sequence $\{x_n\}$ as deterministic, intuitively one can think of (2) as a kind of "law of large numbers" for the data-generating process. For example, we could generate independent and identically distributed (i.i.d.) x_n from some distribution, independent of $\{\varepsilon_n\}_{n=1}^{\infty}$ and satisfying $\mathbb{E}(x_n x_n^\top) = A$, and satisfy (2). The results derived in this section would still hold in such a setting, as long as $\{x_n\}$ was independent of $\{\varepsilon_n\}$.

Let $y_m = \beta^\top x_m + \varepsilon_m$ with the residuals $\varepsilon_m \sim \mathcal{N}(0, \sigma^2)$ being independent with common variance $0 < \sigma^2 < \infty$. The ordinary least-squares estimator b_n of β, given the data (x_m, y_m) for $m = 1, \ldots, n$, is defined as $b_n = \operatorname{argmin}_b \sum_{m=1}^{n} (y_m - b^\top x_m)^2$.

2.1. Large Deviations Laws

We show that, under the given modeling assumptions, b_n obeys the large deviations law

$$\lim_{n\to\infty} \frac{1}{n} \log P(b_n \in E) = -\inf_{u\in E} I(u), \tag{3}$$

where I can be derived in closed form (Theorem 1). The set $E \subseteq \mathbb{R}^d$ satisfies $\beta \notin E$, which means that the event $\{b_n \in E\}$ represents an "error" of some sort. As $n \to \infty$, the probability of error decays exponentially, with the exponent determined by the rate function I.

The proof uses the Gärtner-Ellis theorem (Dembo and Zeitouni 2009), which requires the following steps. First,

for any n, we let $\Psi_n(\gamma) = \log \mathbb{E}(e^{\gamma^\top b_n})$ be the log-mgf of b_n. We then show that the scaled limit

$$\Psi(\gamma) = \lim_{n\to\infty} \frac{1}{n}\Psi_n(n\gamma)$$

exists, and obtain I by taking the Fenchel-Legendre transform

$$I(u) = \sup_\gamma \gamma^\top u - \Psi(\gamma) \tag{4}$$

of Ψ. Thus, the existence of Ψ is the main requirement for the analysis; additionally, as discussed in section 2.3 of Dembo and Zeitouni (2009), the interior of the domain $\{\gamma : \Psi(\gamma) < \infty\}$ should contain the origin, and Ψ should be essentially smooth and lower semicontinuous. All these conditions are satisfied in our setting.

Theorem 1. *For any $E \subseteq \mathbb{R}^d$ such that $\beta \notin E$, the least squares estimator obeys (3) with $I(u) = \frac{1}{2\sigma^2}(u-\beta)^\top A(u-\beta)$.*

Proof. For any n, b_n can be written (Lai and Wei 1982) as

$$b_n = \beta + \left(\sum_{m=1}^n x_m x_m^\top\right)^{-1} \sum_{l=1}^n x_l \varepsilon_l.$$

Using this representation, we calculate

$$\begin{aligned}\Psi_n(\gamma) &= \gamma^\top\beta + \log \mathbb{E}\left(e^{\gamma^\top(\sum_{m=1}^n x_m x_m^\top)^{-1}\sum_{l=1}^n x_l\varepsilon_l}\right)\\ &= \gamma^\top\beta + \log \mathbb{E}\left(e^{\sum_{l=1}^n[\gamma^\top(\sum_{m=1}^n x_m x_m^\top)^{-1}x_l]\varepsilon_l}\right)\\ &= \gamma^\top\beta + \sum_{l=1}^n \frac{1}{2}\sigma^2\left[\gamma^\top\left(\sum_{i=1}^n x_m x_m^\top\right)^{-1} x_l\right]^2.\end{aligned}$$

Consequently, the scaled limit Ψ exists and equals

$$\begin{aligned}\Psi(\gamma) &= \gamma^\top\beta + \lim_{n\to\infty}\sum_{l=1}^n \frac{1}{2}\sigma^2 n\left[\gamma^\top\left(\sum_{m=1}^n x_m x_m^\top\right)^{-1} x_l\right]^2\\ &= \gamma^\top\beta + \lim_{n\to\infty}\sum_{l=1}^n \frac{1}{2}\sigma^2 n\gamma^\top\left(\sum_{m=1}^n x_m x_m^\top\right)^{-1} x_l x_l^\top\left(\sum_{m=1}^n x_m x_m^\top\right)^{-1}\gamma\\ &= \gamma^\top\beta + \lim_{n\to\infty}\frac{1}{2}\sigma^2 n\gamma^\top\left(\sum_{m=1}^n x_m x_m^\top\right)^{-1}\left(\sum_{l=1}^n x_l x_l^\top\right)\left(\sum_{m=1}^n x_m x_m^\top\right)^{-1}\gamma\\ &= \gamma^\top\beta + \lim_{n\to\infty}\frac{1}{2}\sigma^2\gamma^\top\left(\frac{1}{n}\sum_{m=1}^n x_m x_m^\top\right)^{-1}\gamma\\ &= \gamma^\top\beta + \frac{1}{2}\sigma^2\gamma^\top A^{-1}\gamma.\end{aligned}$$

The domain of Ψ is all of $\mathbb{R}^d$, and Ψ is continuous, so all of the conditions needed for (3) are satisfied. Then, (4) becomes

$$I(u) = \sup_\gamma \gamma^\top(u-\beta) - \frac{1}{2}\sigma^2\gamma^\top A^{-1}\gamma.$$

The supremum is achieved at γ^* satisfying

$$\sigma^2 A^{-1}\gamma^* = u-\beta \quad\Rightarrow\quad \gamma^* = \frac{1}{\sigma^2}A(u-\beta).$$

Substituting γ^* into (4) yields $I(u) = \frac{1}{2\sigma^2}(u-\beta)^\top A(u-\beta)$, as required. Q.E.D.

It is possible for (3) to hold (with a different I) when the distribution of ε is nonnormal. In such settings, however, the rate function may have a more complicated dependence on the data-generating process, making the analysis much less tractable. For example, in the context of logistic regression, Jiang et al. (2020) derives a large deviations bound, where the exact rate in (3) is replaced by an inequality. For linear regression with normal residuals, the rate function depends on the data only through the limiting matrix A, which allows more flexibility in the data sequence. The normality assumption is predominant in classical linear regression; when the residuals are believed to be nonnormal, it is more common to transform the response (e.g., fitting the regression model to $\log y$ instead of y) than to explicitly model a nonnormal error distribution.

More explicit rate exponents may be obtained for certain choices of the error event E. In the remainder of this paper, we will primarily focus on error events of the form

$$E_v = \{u \in \mathbb{R}^d : u^\top v \le 0\} \tag{5}$$

for various fixed vectors $v \in \mathbb{R}^d$ that satisfy $\beta^\top v > 0$. As will be seen later on, such a v may be viewed as the difference between two covariate vectors, with the sign of $\beta^\top v$ indicating which vector has the better value, and $\{b_n \in E_v\}$ being the event that estimation error yields the wrong sign. The following result shows that the rate exponent for any such E_v can be computed in closed form.

Proposition 1. *Suppose that $\beta^\top v > 0$. Then,*

$$\lim_{n\to\infty}\frac{1}{n}\log P(b_n^\top v \le 0) = -\frac{1}{2\sigma^2}R(v),$$

where $R(v) = \frac{(\beta^\top v)^2}{v^\top A^{-1} v}$.

Proof. From Theorem 1, it follows that $R(v)$ is the optimal value of the convex program

$$\begin{aligned}&\min_{u\in\mathbb{R}^d} (u-\beta)^\top A(u-\beta)\\ &\text{s.t. } v^\top u \le 0.\end{aligned} \tag{6}$$

Letting λ be the Lagrange multiplier of the single linear constraint, the optimality conditions of (6) are given by

$$A(u-\beta) + \lambda v = 0, \tag{7}$$

$$v^\top u = 0, \tag{8}$$

where (8) follows because the linear constraint should be binding at optimality. Now, (7) yields

$$u = \beta - \lambda A^{-1} v, \tag{9}$$

and plugging (9) into (8) leads to

$$v^\top\beta - \lambda v^\top A^{-1} v = 0 \quad\Rightarrow\quad \lambda = \frac{v^\top\beta}{v^\top A^{-1} v}.$$

Plugging this back into (9), we obtain

$$u^* = \beta - \frac{v^\top \beta}{v^\top A^{-1} v} A^{-1} v,$$

whence

$$\begin{aligned} I(u^*) &= (u^* - \beta) A (u^* - \beta) \\ &= \left(\frac{v^\top \beta}{v^\top A^{-1} v} \right)^2 v^\top A^{-1} A A^{-1} v \\ &= \frac{(v^\top \beta)^2}{v^\top A^{-1} v}, \end{aligned}$$

as required. Q.E.D.

Thus, the convergence rate of $P(b_n \in E_v)$ is governed by the exponent $R(v)$, which depends on the specific vector v we are studying. $R(v)$ is invariant with respect to $\|v\|$, so we can assume $\|v\| = 1$ whenever it is convenient to do so. We can now study error events of the form $\bigcup_k E_{v_k}$ for countable collections $\{v_k\}_{k=1}^\infty$. A straightforward consequence of Theorem 1 and Proposition 1 is that

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(b_n \in \bigcup_k E_{v_k} \right) = -\frac{1}{2\sigma^2} \inf_k R(v_k), \tag{10}$$

provided that $\beta \notin \mathrm{cl} \bigcup_k E_{v_k}$ (or, equivalently, $\inf_k \beta^\top v_k > 0$). Intuitively, the probability that at least one error event in the collection occurs is determined by the slowest convergence rates among the individual error events.

In this work, we focus on uncountable collections of error events (with continuous-valued v_k), so $\{v_k\}$ will be dense in some such set of interest. Potentially, one could also let $\{v_k\}$ be a finite set. Then, the results of Theorem 1 and Proposition 1 will remain unchanged, and the infimum in (10) will become a minimum. Such a setting can be viewed as a generalization of Glynn and Juneja (2004). Optimal allocations for such problems are well understood and can be computed efficiently (Chen and Ryzhov 2019a). For this reason, the present work focuses on the continuous setting, which has never before been studied in the literature on optimal sampling allocation and cannot be addressed by recycling or extending well-established results from the discrete setting.

2.2. Optimal Sampling Allocations

Let $x^* \in \mathbb{R}^d$ be some fixed "reference solution," possibly obtained from some optimization problem that will not be explicitly modeled here. The value of this solution is $\beta^\top x^*$. We assume that larger values are better, so

$$\mathcal{X}(x^*) = \{x \in \mathbb{R}^d : \beta^\top (x^* - x) > 0\}$$

is interpreted as the set of all inferior solutions. If there is any $x \in \mathcal{X}(x^*)$ for which $b_n^\top (x^* - x) \le 0$, this means that the estimated coefficients b_n have led us to erroneously identify x as being superior to x^*. This is clearly an example of (5) with $v = x^* - x$ and $\{b_n \in E_v\}$ being the false identification event. The convergence rate of $P(b_n \in E_v)$ only depends on x^* and x through v.

Potentially, any $x \in \mathcal{X}(x^*)$ can generate an error. Consider a countable collection $\{x_k\}_{k=1}^\infty \subseteq \mathcal{X}(x^*)$. Each x_k corresponds to an error vector $v_k = x^* - x_k$, motivating an optimization problem of the form

$$\sup_{A \in \mathbb{S}_{++}^d} \inf_k \frac{(v_k^\top \beta)^2}{v_k^\top A^{-1} v_k}, \tag{11}$$

where $\mathbb{S}_{++}^d$ is the set of all $d \times d$ symmetric positive definite matrices. Through (10), this problem chooses the matrix A to make $P(b_n \in \bigcup_k E_{v_k})$ converge to zero at the fastest possible rate. Of course, to ensure that (11) is not unbounded, we would also need to impose a simple constraint on the magnitude of A, such as an upper bound on the trace. Such an upper bound serves as a scale factor on $R(v_k)$ for all k, but otherwise does not change the geometry of the optimal A.

However, we require $\beta \notin \bigcup_k E_{v_k}$ to use Theorem 1, which means that we cannot make $\{x_k\}$ dense in the entire set $\mathcal{X}(x^*)$. Instead, we will focus on $\{v_k\} \subseteq V_\delta$, where

$$V_\delta = \{v : \|v\| = 1, \beta^\top v \ge \delta\},$$

and $\delta > 0$ is a small constant. By introducing δ, we ensure that optimal allocation problems such as (11) are well defined. Such problems maximize the smallest rate exponent over some set, and without setting a nonzero threshold for δ, it will always be possible to find exponents arbitrarily close to zero. In terms of interpretation, we are now willing to accept $x \in \mathcal{X}(x^*)$ whose value is sufficiently close to that of x^*, and we focus on eliminating errors generated by solutions that are outside this tolerance level. Our *design space* need not be restricted to V_δ. The parameter δ only imposes restrictions on the error events that we are trying to eliminate.

With this modification, one can rewrite (11) as

$$\max_{A \in \mathbb{S}_{++}^d} \min_{v \in V_\delta} \frac{(v^\top \beta)^2}{v^\top A^{-1} v}. \tag{12}$$

Because A is symmetric and positive definite, we can write $A = \sum_{i=1}^d p_i \zeta_i \zeta_i^\top$, where $p_i > 0$ and $(\zeta_1, \ldots, \zeta_d)$ is an orthonormal basis for $\mathbb{R}^d$. We may assume that $\sum_i p_i = 1$ without loss of generality; as discussed earlier, this condition scales the optimal A without changing its geometry. Recalling the interpretation of A as an expected value, p_i can be seen as the probability of sampling ζ_i.

Thus far, (12) requires us to jointly choose both eigenvalues and eigenvectors. We will simplify this problem by setting $\zeta_1 = \beta$; that is, β itself will be an eigenvector. With this, the orthonormal basis can be straightforwardly completed, and the only remaining decision variable will be the vector p of eigenvalues. We first give some intuition for this choice. For any fixed positive definite B, the ratio $\frac{(\beta^\top v)^2}{v^\top B v}$ can in general be made arbitrarily

small. However, if we allow the positive *semidefinite* matrix $B = \beta\beta^\top$, the ratio evaluates to 1 for *any* v with $\beta^\top v \neq 0$. This suggests that, when we choose a positive definite B, its principal eigenvector should also be aligned with β.

Before providing more rigorous support for this idea, we first manipulate the problem setup as follows. Let $B = \sum_i r_i \zeta_i \zeta_i^\top$, where $(\zeta_1, \ldots, \zeta_d)$ is an orthonormal basis for $\mathbb{R}^d$, and $r_1 > r_2 \geq \cdots \geq r_d > 0$ are the eigenvalues. It can easily be seen that $\min_{v \in V_\delta} \frac{(\beta^\top v)^2}{v^\top B v}$ is attained on the boundary $\partial V_\delta = \{v : \|v\| = 1, \beta^\top v = \delta\}$. Then, the problem $\max_B \min_{v \in \partial V_\delta} \frac{(\beta^\top v)^2}{v^\top B v}$ has the same optimal solution as the problem $\min_B \max_{v \in \partial V_\delta} v^\top B v$. The inner maximization resembles an eigenvalue problem; this connection is used in the following result to bound the inner maximum below by the second-largest eigenvalue of B for *any* orthonormal basis not aligned with β.

Proposition 2. *Suppose that $\beta \neq \zeta_i$ for any i. Then, $\max_{v \in \partial V_\delta} v^\top B v > r_2$, with strict inequality continuing to hold in the regime $\delta \to 0$.*

Proof. Define $v = \delta\beta + Pw$, where $P = I - \beta\beta^\top$ is the projection onto the orthogonal complement of β. Then, the objective $\frac{v^\top B v}{v^\top v}$, which coincides with $v^\top B v$ when $v^\top v = 1$, can be rewritten in terms of w as

$$f(w) = \frac{w^\top PBPw + 2\delta w^\top PB\beta + \delta^2 \beta^\top B\beta}{w^\top Pw + \delta^2}.$$

Observe that

$$\nabla_w f = \frac{1}{w^\top Pw + \delta^2}(2PBPw + 2\delta PB\beta - 2f(w)Pw).$$

Setting the gradient equal to zero yields

$$PBPw + \delta PB\beta = f \cdot Pw. \tag{13}$$

Given any solution (f, w) of (13), we can obtain a feasible $v = \delta\beta + Pw$ whose objective value is f. Observe, however, that such a solution may be found for almost any f value: we may rewrite (13) as $(fI - PB)Pw = \delta PB\beta$, where the matrix $fI - PB$ is invertible as long as f is not equal to any of the eigenvalues $s_1 \geq \cdots \geq s_d$ of PB. Consequently, given any f satisfying $f \neq s_i$ for all i, we can obtain $Pw = \delta(fI - PB)^{-1} PB\beta$ such that $v = \delta\beta + Pw$ satisfies $\frac{v^\top B v}{v^\top v} = f$.

However, we also require v to satisfy the normalization condition $v^\top v = 1$. Equivalently, we must have $w^\top P^2 w = 1 - \delta^2$, which becomes

$$\frac{1 - \delta^2}{\delta^2} = b^\top BP(fI - PB)^{-2} PB\beta. \tag{14}$$

Thus, the optimal value of $\max_{v \in \partial V_\delta} v^\top B v$ is the largest f for which (14) holds. Because the right-hand side of (14) has a cusp at $f = s_1$ and decreases monotonically on (s_1, ∞), the largest solution satisfies $f > s_1$. By the Courant-Fischer theorem (Horn and Johnson 2013, theorem 4.2.6), we have $r_1 > s_1 > r_2$, whence $f > r_2$. As $\delta \to 0$, the largest solution converges to s_1, which is strictly greater than r_2. Q.E.D.

In words, once we have fixed the eigenvalues of B, choosing any orthonormal basis that is not aligned with β will always result in a nonzero gap between the inner maximum $\max_{v \in \partial V_\delta} v^\top B v$ and the lower bound r_2, and this gap cannot be closed in the small-δ regime. On the other hand, if $\beta = \zeta_1$, the inner maximum is achieved by taking $v = \delta\zeta_1 + \sqrt{1 - \delta^2} \cdot \zeta_2$, yielding the optimal value $\delta^2 r_1 + (1 - \delta^2) r_2$, which converges to r_2 as $\delta \to 0$. Because we seek to minimize the value of the inner maximum, it follows that we should align the principal eigenvector with β to close the gap with the lower bound. Thus, we impose the structure

$$A = p_1 \beta\beta^\top + \sum_{i>1} p_i \zeta_i \zeta_i^\top, \tag{15}$$

where the other vectors $\zeta_2, \ldots, \zeta_d$ in the orthonormal basis are unique (up to multiplication by -1). The remainder of this paper will derive the optimal eigenvalues p_i subject to the normalization condition $\sum_i p_i = 1$. In fact, we will see that $p_1 = \min_i p_i$ in the optimal solution, confirming the intuition that β should be the principal eigenvector of A^{-1}.

3. Solving for the Optimal Allocation

Suppose that the sequence $\{v_k\}$ is dense in V_δ. Because $R(v)$ is invariant with respect to $\|v\|$, we can focus on unit vectors without loss of generality. For fixed K, we consider the problem

$$\max_p \min_{k \leq K} \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2}, \tag{16}$$

subject to the constraints $p \geq 0$, $\sum_i p_i = 1$. Equation (16) is a version of (11) with (15) plugged into the denominator. As $K \to \infty$, the inner minimum in (16) will behave like a minimum over all $v \in V_\delta$. Because we are mainly interested in this asymptotic regime, we can choose the elements of $\{v_k\}$ in any way we want, as long as the sequence remains dense in V_δ.

The objective function in (16) is concave in p and can be rewritten as $\max_{p,z} z$ subject to

$$z \leq \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2}, \qquad k = 1, \ldots, K, \tag{17}$$

in addition to the original constraints on p. The Lagrangian of this optimization problem is given by

$$L(z, p, \mu, \nu) = -z + \sum_{k=1}^K \mu_k \left(z - \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2} \right) + \nu \left(\sum_{i=1}^d p_i - 1 \right),$$

with the terms corresponding to the nonnegativity constraints on p_i omitted to ensure that A is positive definite. The optimality conditions are as follows:

1. First-order conditions:

$$\sum_{k=1}^{K} \mu_k \frac{(v_k^\top \beta)^4}{\left[\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1}\frac{1}{p_i}(v_k^\top \zeta_i)^2\right]^2} = p_1^2 \nu, \tag{18}$$

$$\sum_{k=1}^{K} \mu_k \frac{(v_k^\top \beta)^2 (v_k^\top \zeta_i)^2}{\left[\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1}\frac{1}{p_i}(v_k^\top \zeta_i)^2\right]^2} = p_i^2 \nu, \quad i = 2, \ldots, d, \tag{19}$$

$$\sum_{k=1}^{K} \mu_k = 1. \tag{20}$$

2. Primal feasibility: (17) and $\sum_i p_i = 1$, $p_i > 0$ for all i.
3. Dual feasibility: $\mu_k \geq 0$.
4. Complementary slackness:

$$\mu_k \left(z - \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1}\frac{1}{p_i}(v_k^\top \zeta_i)^2} \right) = 0, \qquad k = 1, \ldots, K. \tag{21}$$

The first-order conditions (18)–(19) can be viewed as a system of d linear equations in K variables $\mu_1, \ldots, \mu_K$. For large K, this system may have many solutions. In particular, we can construct a basic solution by taking d linearly independent vectors $v_{k_1}, \ldots, v_{k_d}$ from $\{v_k\}_{k=1}^K$ and setting $\mu_k = 0$ if $k \notin \{k_1, \ldots, k_d\}$. Because $\{v_k\}$ is dense in a set of dimension d, we can choose individual v_k to take certain values in that set without affecting the asymptotic result. For our analysis, it is convenient to take $w_1 = \beta$ and let w_j be a linear combination of β and ζ_j, for $j = 2, \ldots, d$, with $w_j^\top \zeta_i = 0$ for any $i \neq j$. We may assume that, for any j, there exists $k_j \leq K$ such that $w_j = v_{k_j}$.

With this choice of w_j, we can rewrite (18)–(19) as

$$p_1^2 \mu_{k_1} + \sum_{j>1} \mu_{k_j} \frac{(w_j^\top \beta)^4}{\left[\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2\right]^2} = p_1^2 \nu, \tag{22}$$

$$\mu_{k_j} \frac{(w_j^\top \beta)^2 (w_j^\top \zeta_j)^2}{\left[\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2\right]^2} = p_j^2 \nu, \quad j = 2, \ldots, d. \tag{23}$$

Substituting (23) into (22) yields

$$p_1^2 \mu_{k_1} + \nu \sum_{j>1} p_j^2 \frac{(w_j^\top \beta)^2}{(w_j^\top \zeta_j)^2} = p_1^2 \nu. \tag{24}$$

If we set $\mu_{k_1} = 0$, the dual variable ν cancels out of (24), yielding

$$p_1^2 = \sum_{j>1} p_j^2 \frac{(w_j^\top \beta)^2}{(w_j^\top \zeta_j)^2}. \tag{25}$$

For any p, it is easy to find $\mu_{k_j} > 0$ and ν to satisfy (23). Condition (20) can also be easily satisfied by rescaling these values. The complementary slackness condition (21) is satisfied for any $k \notin \{k_2, \ldots, k_d\}$ because the corresponding dual variables μ_k are set to zero. To satisfy the condition for the remaining values of k, it is sufficient to ensure that $R(w_i) = R(w_j)$, that is,

$$\frac{(w_i^\top \beta)^2}{\frac{1}{p_1}(w_i^\top \beta)^2 + \frac{1}{p_i}(w_i^\top \zeta_i)^2} = \frac{(w_j^\top \beta)^2}{\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2}, \qquad i, j \neq 1. \tag{26}$$

Thus, as long as p is chosen to satisfy (25)–(26), we can find feasible μ, ν to satisfy (18)–(21). Essentially, most of the optimality conditions for (16) have reduced to Conditions (25)–(26) on p, which generalize those derived in example 1 of Glynn and Juneja (2004) for large deviations of pairwise comparisons between scalar normal distributions.

In fact, there is only one optimality condition for (16) that has not yet been treated, namely (17). Our choice of p must also imply $R(w_i) \leq R(v_k)$ for all $i = 2, \ldots, d$ and $k = 1, \ldots, K$. Recalling that we have the freedom to pick w_j, we further suppose that $(w_j^\top \beta)^2 = \delta^2$ for $j = 2, \ldots, d$. Because each w_j is a unit vector, it follows that $(w_j^\top \zeta_j)^2 = 1 - \delta^2$. Consequently, (26) now implies that $p_i = p_j = c$ for $i, j \neq 1$ and some constant c. Then, for any $v \in V_\delta$, the rate exponent $R(v)$ simplifies to

$$R(v) = \frac{(v^\top \beta)^2}{\frac{1}{p_1}(v^\top \beta)^2 + \frac{1}{c}\sum_{i>1}(v^\top \zeta_i)^2}.$$

Because $(v^\top \beta)^2 \geq \delta^2$ for any $v \in V_\delta$, we must also have $\sum_{i>1}(v^\top \zeta_i)^2 \leq 1 - \delta^2$ because v is a unit vector. Consequently,

$$R(v) \geq \frac{\delta^2}{\frac{1}{p_1}\delta^2 + \frac{1}{c}(1 - \delta^2)} = R(w_j)$$

for any $j = 2, \ldots, d$. Thus, our choice of w has caused (17) to be satisfied for *any* $v \in V_\delta$. Therefore, the solution p^* of (25)–(26), for this choice of w, is optimal for *any* arbitrarily large K, and therefore

$$p^* = \arg\max_{p:\sum_i p_i = 1} \min_{v \in V_\delta} R(v)$$

also optimizes the convergence rate of the probability that an error arises from any $v \in V_\delta$.

It remains to calculate p^*. Letting $\Delta = \frac{\delta^2}{1-\delta^2}$, we find that (25) reduces to

$$p_1^2 = (d-1)\Delta c^2.$$

At the same time, $p_1 = 1 - (d-1)c$, whence

$$1 - (d-1)c = c\sqrt{(d-1)\Delta},$$

leading to the closed-form solution

$$p_1^* = \frac{\sqrt{(d-1)\Delta}}{(d-1)+\sqrt{(d-1)\Delta}}, \tag{27}$$

$$p_i^* = \frac{1}{(d-1)+\sqrt{(d-1)\Delta}}, \qquad i = 2,\dots,d. \tag{28}$$

Recalling our earlier interpretation of A as an expected value, (15) allows us to view the allocation as a discrete probability distribution where each p_i represents the probability of collecting a data point using ζ_i as the covariate vector. The solution (27)–(28) indicates that the optimal distribution is *almost* uniform: any basis vector that is orthogonal to β can be sampled with the same probability. However, the probability assigned to the first eigenvector β is different from the others; as δ becomes smaller, this probability is reduced, which means that A^{-1} will correspondingly place *more* weight on $\beta\beta^\top$, as expected.

One especially striking aspect of this solution is that the probabilities p_i^* are completely deterministic. Thus, the only unknown quantity in (15) is β itself, as suitable ζ_i can be straightforwardly computed if β is known. In other words, the budget is being allocated to an orthonormal basis that depends only on β, not x^*. Another way to interpret our results is that, for *any* x^*, the probability that $b_n^\top(x^*-x)>0$ for *all* x satisfying $\beta^\top(x^*-x)\geq\delta$ converges to 1 at the fastest possible rate.

The interpretation of the eigenvalues p_i^* as probabilities also allows us to apply the previous results to other d-dimensional design spaces. For example, suppose that the design space is a hyper-rectangle, allowing us to sample scalar multiples $a_1\beta, a_2\zeta_2, \dots, a_d\zeta_d$ of the original orthonormal basis vectors. We simply renormalize the scaled probabilities $\frac{p_i^*}{a_i^2}$ and obtain a new data-generating distribution with support $(a_1\beta,\dots,a_d\zeta_d)$. The geometry of the optimal A will be preserved, although the rate exponent itself will be scaled by a constant factor depending on the multipliers a_i.

4. Practical Implementation and Numerical Examples

Section 4.1 gives a simple sequential implementation of our new optimal allocation and describes three benchmarks. Section 4.2 describes the generation of test instances, and Section 4.3 presents numerical results.

4.1. Description of Algorithms

Algorithm 1 states a very simple algorithm (which we call "LD-optimal") for sequentially implementing our new optimal allocation. Essentially, we use the least-squares estimator b_n in place of β. The estimator itself can be updated recursively, but in every iteration, we have to extend it to an orthonormal basis. A simple way to do this is to take d arbitrary, prespecified linearly independent vectors $(\xi_1,\dots,\xi_d)$ and apply the Gram-Schmidt process to $(b_n,\xi_1,\dots,\xi_d)$. Again, the algorithm does not need to know or estimate x^*. In our numerical experiments, we implemented this procedure together with three benchmarks, which we now briefly describe.

Algorithm 1 (LD-Optimal Algorithm for Sequential Implementation of the Optimal Allocation)

Step 0: Let $n=1$, initialize $b_1\in\mathbb{R}^d$ and $A_1\in\mathbb{S}^d_{++}$.

Step 1: Calculate vectors $\zeta_{n,i}$ such that $\left(\frac{b_n}{\|b_n\|},\zeta_{n,2},\dots,\zeta_{n,d}\right)$ is an orthonormal basis for $\mathbb{R}^d$.

Step 2: Set

$$x_{n+1} = \begin{cases} \dfrac{b_n}{\|b_n\|} & \text{w.p. } p_1^* \\ \zeta_{n,i} & \text{w.p. } p_i^*. \end{cases}$$

Step 3: Observe $y_{n+1}=\beta^\top x_{n+1}+\varepsilon_{n+1}$ and update

$$b_{n+1} = b_n + \frac{y_{n+1}-b_n^\top x_{n+1}}{1+x_{n+1}^\top A_n x_{n+1}} A_n x_{n+1},$$

$$A_{n+1} = A_n - \frac{A_n x_{n+1} x_{n+1}^\top A_n}{1+x_{n+1}^\top A_n x_{n+1}^\top}.$$

Increment n by 1 and return to step 1.

4.1.1. Randomized Adaptive Gap Elimination (Fiez et al. 2019).

The Randomized Adaptive Gap Elimination (RAGE) method assumes that the sampling decision is restricted to a prespecified finite set of vectors $\{z_\ell\}$ (unlike our algorithm, which can sample any vector on the unit sphere). RAGE proceeds in "phases." In each phase, some vectors z_ℓ are removed (screened out), and each of the remaining vectors is sampled a number of times that is determined dynamically. The procedure terminates when only one element is left, and the screening and sampling steps are constructed to ensure that a certain type of accuracy guarantee is achieved at termination. The number of phases and samples needed for termination is not known ahead of time. Instead, RAGE runs until its termination criterion is satisfied. As a result, RAGE cannot be run with a prespecified fixed sample size. We run RAGE using the creators' publicly available code.

4.1.2. Uniformly Random (D-Optimal) Design.

This method samples a vector that is uniformly generated on the L^2 sphere. It turns out that this simple procedure is equivalent, in a certain sense, to a classical design of experiments method known as D-optimal (Mitchell 2000). Traditionally, D-optimal design chooses $x_1,\dots,x_n$ to maximize $\log\det\left(\sum_{m=1}^n x_m x_m^\top\right)$. In many classical settings, the design space is discretized, and the optimal design is interpreted in terms of proportions of the total simulation budget assigned to each vector in the discrete set. These proportions can be obtained using convex optimization methods (Lu et al. 2018).

Because (2) assumes $\frac{1}{n}\sum_{m=1}^{n} x_m x_m^\top \to A$, we can simply use the limiting matrix A in the D-optimality criterion. Because A is positive definite, one can write $A = \sum_i p_i \zeta_i \zeta_i^\top$ for some orthonormal basis $\zeta_1, \ldots, \zeta_d$. It is easy to see that the optimality criterion is unaffected by the choice of orthonormal basis, because $\log\det(A) = \sum_i \log p_i$. Under the normalization condition $\sum_i p_i = 1$, it is readily seen that the objective is maximized by setting $p_i \equiv \frac{1}{d}$. It is then straightforward to show that sampling x from a uniform distribution on the L^2 sphere makes the matrix $\mathbb{E}(xx^\top)$ D-optimal.

We also considered a more traditional implementation of *D*-optimal in which the design space was obtained by discretizing the L^2 sphere. However, the resulting allocations were very close to uniform, and performance was nearly identical to the previously described setup. Thus, we proceed with uniform design, because it offers more flexibility in the choice of design points and is also easy to implement sequentially.

4.1.3. Oracle Allocation. This method is given access to an orthonormal basis $(\beta, \zeta_2, \ldots \zeta_d)$, which includes the true value of β, and implements the true optimal allocation: in each iteration, it samples β with probability (27) and ζ_i with probability (28). The true β is only used for sampling; the recursive least squares parameters (b_n, A_n) from Algorithm 1 are still calculated and used to evaluate performance. Such a method is not implementable in practice, but we include it here to study whether any loss in performance is incurred by using an estimator in place of β.

We did not compare against such sequential simulation optimization procedures as the knowledge gradient method of Han et al. (2016) or the Thompson sampling method of Russo and Van Roy (2014), because these methods all focus on identifying a particular x^* value. By contrast, our method eliminates false identification events of the form $\{b_n^\top(x^* - x) \le 0\}$ for x, x^* pairs satisfying $\beta^\top(x^* - x) > 0$. This goal is related to simulation optimization, because x^* could be the desired optimal solution; however, because the large deviations rates depend on x^* only through pairwise differences $x^* - x$, essentially our allocation is optimizing convergence rates for all such x, x^* pairs simultaneously (the performance metric used in Section 4.2 is also based on all possible pairs). A similar insight is built into the RAGE algorithm, which is why we see it as the most natural benchmark.

4.2. Test Problems and Performance Evaluation

First, we describe the metric used to evaluate performance. For $k = 1, \ldots, 100$, we generate vectors x_k^* and a normalized vector β. After n observations have been collected, we report

$$F_n = \frac{\sum_{k,k'} 1_{\{b_n^\top(x_k^* - x_{k'}^*) > 0\}} 1_{\{\beta^\top(x_k^* - x_{k'}^*) > 0\}}}{\sum_{k,k'} 1_{\{\beta^\top(x_k^* - x_{k'}^*) > 0\}}}, \tag{29}$$

the proportion of pairs $x_k^*, x_{k'}^*$ that satisfy $\beta^\top(x_k^* - x_{k'}^*) > 0$ and are correctly ordered under the estimated coefficients b_n. The problem is more challenging (i.e., F_n tends to be smaller) if there are more pairs with $\beta^\top(x_k^* - x_{k'}^*)$ close to zero. In such cases, the sign of $\beta^\top(x_k^* - x_{k'}^*)$ will be more difficult to identify. If the problem is not sufficiently difficult in this sense, it is likely that virtually any algorithm will perform well.

The literature has not worked out a standard for generating difficult test problems. One can find toy examples for discretized problems: Soare et al. (2014) gives one example where sampling is restricted to $d + 1$ predetermined vectors, which are also used as the test vectors x_k^*. The simulation literature uses some standard test settings, such as the slippage configuration (Shen et al. 2021), but these settings grew out of the simpler ranking and selection problem and do not reflect the full richness of covariates that one would likely see in an application of regression. For these reasons, we developed a new schema for generating difficult, but diverse test instances.

We first choose β uniformly on the unit sphere in $\mathbb{R}^d$. Then, for each k, we generate a vector $\tilde{x}_k$ by sampling each component independently from a uniform distribution on $[-0.5, 0.5]$. We then calculate $x_k^* = (I - \beta\beta^\top)\tilde{x}_k + u_k$, where u_k is a vector whose components are i.i.d. uniform on a small interval (e.g., on $[-0.01, 0.01]$). Thus, $\beta^\top x_k^*$ is very close (but not identical) to zero, and there is a much larger number of $(x_k^*, x_{k'}^*)$ pairs for which the sign of $\beta^\top(x_k^* - x_{k'}^*)$ is difficult to identify. At the same time, the individual components of each x_k^* may be very different from zero, allowing for considerable diversity between test vectors. Our results will demonstrate that this schema can produce exceptionally difficult test instances.

We also comment on the implementation of the methods from Section 4.1. Both LD-optimal and D-optimal can be run for any arbitrary n, neither of them requires any knowledge of the x_k^* vectors, and both can sample anywhere on the L^2 sphere. Thus, we can run them sequentially and evaluate (29) for each n. On the other hand, the RAGE algorithm requires the vectors x_k^* as inputs, and furthermore is restricted to sampling from a finite set $\{z_\ell\}$ of unit vectors, which we generate independently and uniformly on the unit sphere. We run the method with these inputs and two different values of the accuracy (error tolerance) parameter: the default value of 0.01 and a higher value of 0.05. As explained previously, RAGE runs until its termination criterion is met, and we do not know ahead of time how long this will take. Thus, we cannot directly compare it against the other methods for fixed n. Instead, we report the results of RAGE separately from the other methods and give a qualitative discussion of the differences.

4.3. Numerical Examples

The space of possible regression problems is very rich: performance will depend on β, the x_k^* vectors that are being compared, the dimensionality d, and the level σ of

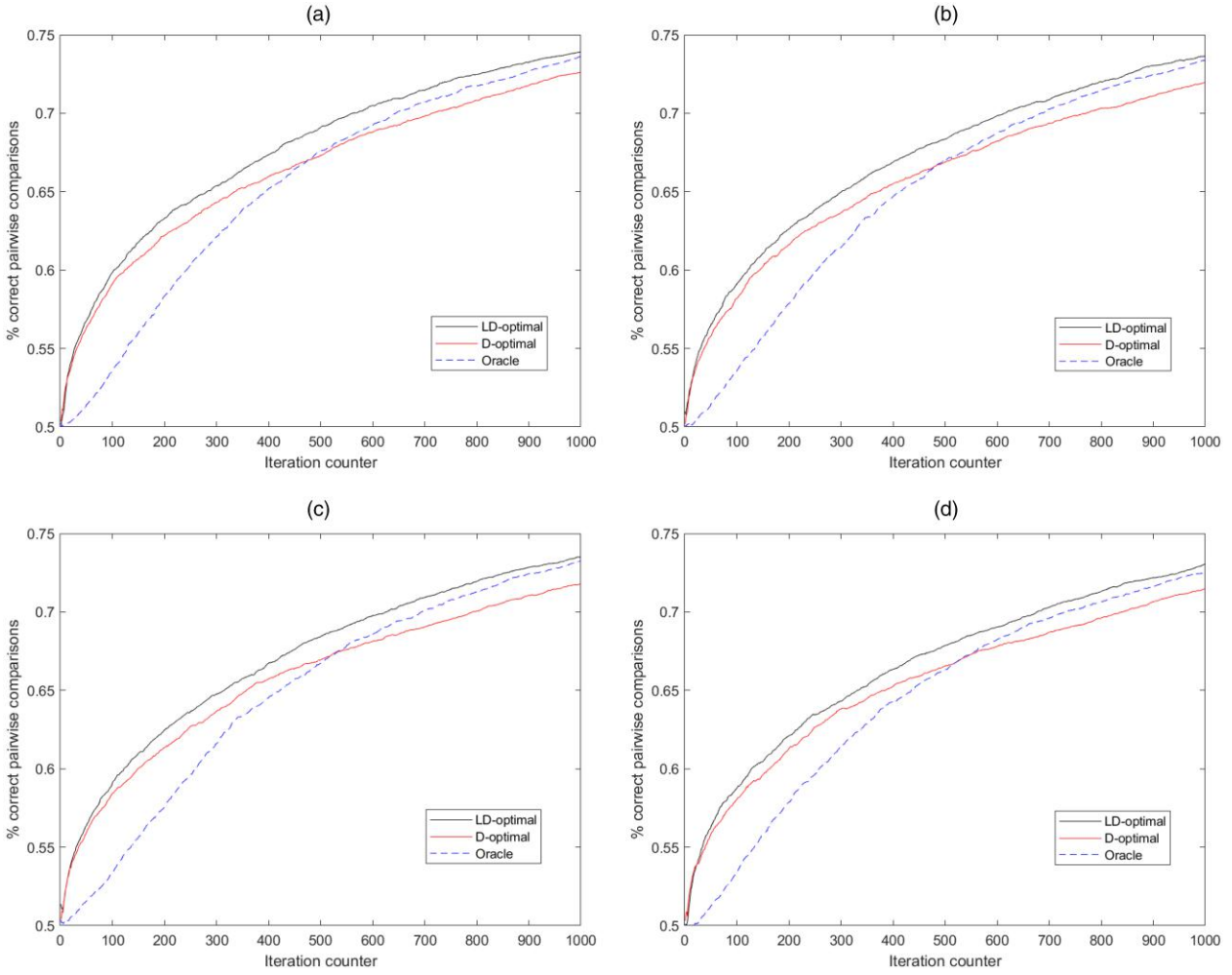
noise. In our experience, even small differences in these inputs may significantly change the difficulty of the problem. In the following, we examine some individual instances, all following the general schema described in Section 4.2, to illustrate situations in which the LD-optimal allocation adds value.

We chose eight instances to illustrate both small-sample and large-sample behaviors. All instances use $d = 5$. The first four instances use $\sigma = 0.2$ and independently generate each component of the random perturbations u_k from a uniform distribution on $[-0.01, 0.01]$. These instances can be said to be moderately difficult; one can make significant progress with a few hundred samples. The second set of four instances uses $\sigma = 0.25$ and intervals $[-0.005, 0.005]$. This seemingly minor difference significantly increases the difficulty of these instances relative to the first four.

Figure 1 examines the performance of the three sequential methods (LD-optimal, D-optimal, and Oracle) on Instances 1–4 over a horizon of 10^3 samples. Results are averaged over 1,000 macro-replications. The four graphs are very similar, although we will show later (Table 1) that Instances 1–4 are quite different from the viewpoint of the RAGE algorithm. Here, we observe that LD-optimal consistently outperforms D-optimal in all four instances. The gap between them widens after an initial learning period (in the first 100 samples, no method can do much better than random guessing). Typically, one can obtain an improvement of about 2% in the metric (29) by using LD-optimal rather than D-optimal. Because each instance has 4,950 distinct $(x_k^*, x_{k'}^*)$ pairs, this translates to about 100 more correct pairwise comparisons.

Figure 2 reports performance on Instances 5–8 over a larger horizon of 10^5 samples. Again, LD-optimal consistently outperforms D-optimal throughout the time horizon. The typical improvement obtained is 2%–2.5%. The increased difficulty of these problems is demonstrated

Figure 1. (Color online) Comparison of Sequential Methods in Small-Sample Settings



Notes. (a) Instance 1. (b) Instance 2. (c) Instance 3. (d) Instance 4.

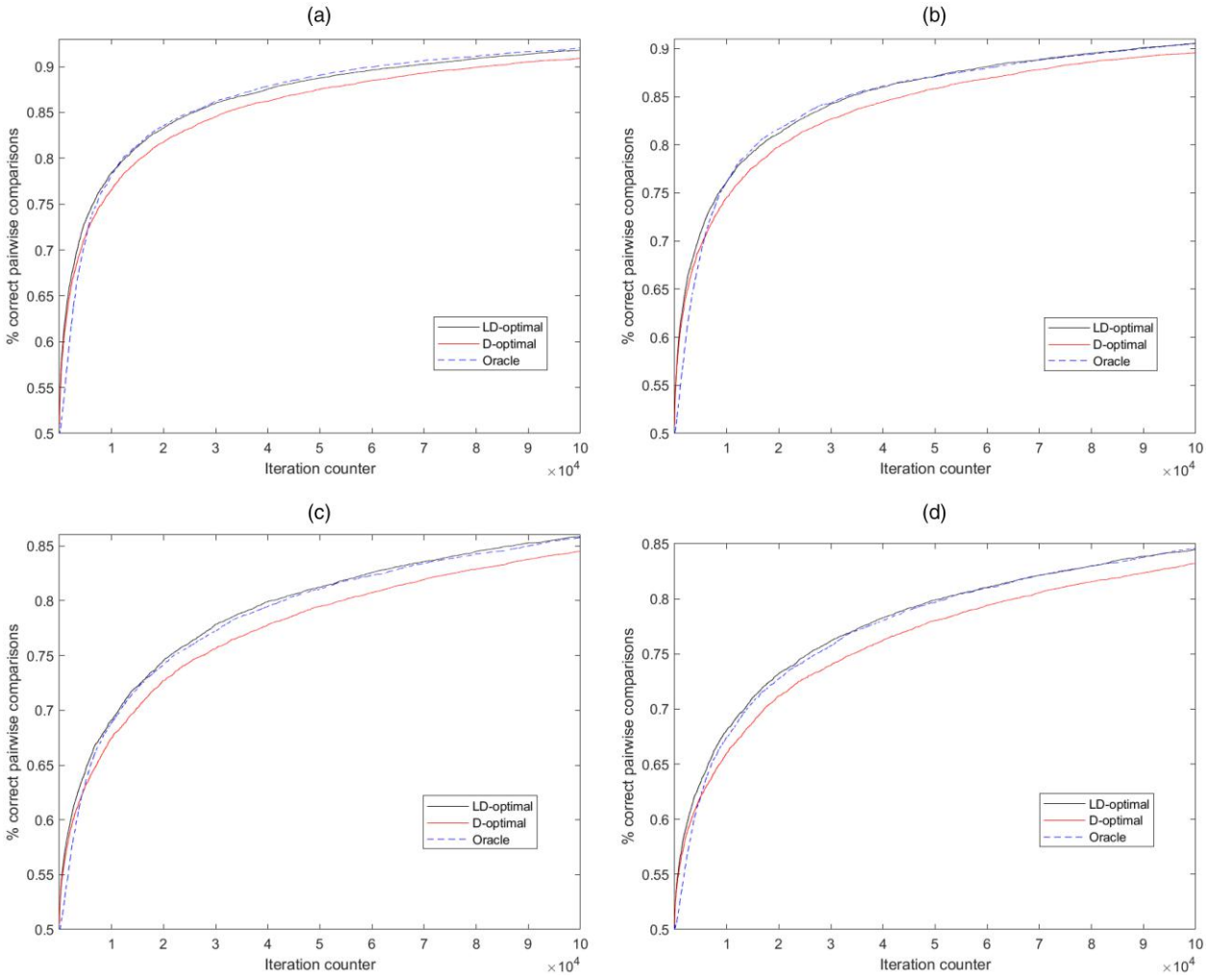
Table 1. Performance of RAGE

	Accuracy		Sample size	
Instance	Tolerance 0.01	Tolerance 0.05	Tolerance 0.01	Tolerance 0.05
1	0.9909	0.9798	6,207,952	5,337,158
2	0.9952	0.9966	11,225,936	10,329,391
3	0.9986	0.9899	107,693,007	96,912,107
4	0.9962	0.9949	44,182,938	40,090,358
5	0.9962	0.9960	39,008,844	36,699,609
6	0.9960	0.9984	300,467,296	273,618,006
7	0.9962	0.9929	236,558,223	212,688,891
8	0.9962	0.9903	208,700,241	198,886,657

by the fact that the initial learning period is now much longer for all methods; their accuracy after 10^3 samples is much lower than in Instances 1–4.

It is interesting to observe that the Oracle method, which knows the true β value, actually lags behind LD-optimal in all eight instances. Similar behavior has been observed before by Chen et al. (2006) for optimal computing budget allocation methods. In our setting, this happens because the optimal allocation assigns a much lower proportion of the budget to the basis vector β. As a result, this vector receives very few samples early on, making the matrix A_n in the computation of the

Figure 2. (Color online) Comparison of Sequential Methods in Large-Sample Settings



Notes. (a) Instance 5. (b) Instance 6. (c) Instance 7. (d) Instance 8.

recursive least squares estimator unstable. By contrast, when LD-optimal uses the estimator b_n to make sampling decisions, the random error in this estimator actually improves the stability of A_n. However, once sufficiently many samples have been collected, LD-optimal and Oracle behave identically.

Table 1 reports the performance of RAGE. We find that this method is able to achieve extremely high accuracy according to the metric (29), but this comes at the cost of extremely large sample sizes. Even the easiest among Instances 5–8 requires more than 35 million samples before the termination criterion is met; the others require upward of 300 million. Increasing the tolerance parameter of RAGE yields only a modest reduction in the sample sizes. The numbers in Table 1 were obtained by running RAGE ten times (on each instance and tolerance parameter) with different randomly generated z_ℓ (as described in Section 4.1) and taking the result with the *smallest* sample size among these ten. As an aside, we found that sample sizes were highly sensitive to the choice of z_ℓ. For example, on Instance 6 with tolerance 0.01, the smallest sample size observed in ten runs was approximately 300 million, but the largest was more than 484 million. There is also a great deal of variation between individual instances, despite that Instances 1–4 (and, respectively, 5–8) were generated from the same specifications.

Thus, although RAGE nominally achieves the highest accuracy, the cost of this is impractical. If one truly has the ability to collect hundreds of millions of samples, there seems to be little need for a sequential algorithm. We acknowledge that RAGE has strong guarantees on the error probability at the moment of termination; however, it has often been observed in the past (Wang and Kim 2012) that such "fixed-precision" guarantees often result in very conservative empirical behavior. If a very large sampling budget is infeasible, LD-optimal may offer a powerful alternative.

5. Conclusion

We derived a new optimal sampling allocation for linear regression based on a large deviations theoretic analysis of error probability. Our result has several novel characteristics relative to previous work. First, in the linear regression setting, it is not necessary to specify or estimate a particular "optimal" solution that we are trying to select. The asymptotic behavior of the error probability depends only on the size on the suboptimality gap, so our allocation simultaneously learns about any gaps, between any two solutions, in excess of a given threshold δ. As a result, it becomes optimal to allocate the budget to an orthonormal basis for the solution space that depends only on β rather than on specific x^* as in discrete problems. This makes the allocation very easy to implement, offering a natural computational benchmark for this problem class that can perform well under limited sampling budgets.

Acknowledgments

The authors thank Tianyu Zhang for coding assistance.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY), 2312–2320.

Ahipaşaoglu SD (2015) A first-order algorithm for the A-optimal experimental design problem: A mathematical programming approach. *Statist. Comput.* 25(6):1113–1127.

Applegate EA, Feldman G, Hunter SR, Pasupathy R (2020) Multi-objective ranking and selection: Optimal sampling laws and tractable approximations via SCORE. *J. Simulation* 14(1):21–40.

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016) An analytics approach to designing combination chemotherapy regimens for cancer. *Management Sci.* 62(5):1511–1531.

Brantley MW, Lee LH, Chen CH, Chen A (2013) Efficient simulation budget allocation with regression. *IIE Trans.* 45(3):291–308.

Brantley MW, Lee LH, Chen CH, Xu J (2014) An efficient simulation budget allocation method incorporating regression for partitioned domains. *Automatica* 50(5):1391–1400.

Chen CH, Lee LH (2010) *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation* (World Scientific, Belmont, MA).

Chen CH, He D, Fu MC (2006) Efficient dynamic simulation allocation in ordinal optimization. *IEEE Trans. Automatic Control* 51(12):2005–2009.

Chen CH, Chick SE, Lee LH, Pujowidianto NA (2015) Ranking and selection: Efficient simulation budget allocation. Fu MC, ed. *Handbook of Simulation Optimization* (Springer, Berlin), 45–80.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10(3):251–270.

Chen Y, Ryzhov IO (2019a) Balancing optimal large deviations in ranking and selection. Mustafee N, Bae KH, Lazarova-Molnar S, Rabe M, Szabo C, Haas P, Son YJ, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3368–3379.

Chen Y, Ryzhov IO (2019b) Complete expected improvement converges to an optimal budget allocation. *Adv. Appl. Probability* 51(1):209–235.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. *Proc. 21st Annual Conf. on Learn. Theory* (Omnipress, Madison, WI), 355–366.

Dembo A, Zeitouni O (2009) *Large Deviations Techniques and Applications*, 2nd ed. (Springer, Berlin).

Dette H (1997) Designing experiments with respect to 'standardized' optimality criteria. *J. Royal Statist. Soc. B* 59(1):97–110.

Fiez T, Jain L, Jamieson KG, Ratliff L (2019) Sequential experimental design for transductive linear bandits. Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY), vol. 32, 10667–10677.

Gao S, Chen W, Shi L (2017) A new budget allocation framework for the expected opportunity cost. *Oper. Res.* 65(3):787–803.

Gao S, Du J, Chen CH (2019) Selecting the optimal system design under covariates. *Proc. 15th Internat. Conf. on Automation Sci. and Engrg.* (IEEE, Piscataway, NJ), 547–552.

Glynn PW, Juneja S (2004) A large deviations perspective on ordinal optimization. Ingalls R, Rossetti MD, Smith JS, Peters BA, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 577–585.

Han B, Ryzhov IO, Defourny B (2016) Optimal learning in linear regression with combinatorial feature selection. *INFORMS J. Comput.* 28(4):721–735.

Horn RA, Johnson CJ (2013) *Matrix Analysis*, 2nd ed. (Cambridge University Press, Cambridge, UK).

Jiang G, Hong LJ, Nelson BL (2020) Online risk monitoring using offline simulation. *INFORMS J. Comput.* 32(2):356–375.

Lai TL, Wei CZ (1982) Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* 10(1):154–166.

Lu H, Freund RM, Nesterov Y (2018) Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.* 28(1):333–354.

Mitchell TJ (2000) An algorithm for the construction of "D-optimal" experimental designs. *Technometrics* 42(1):48–54.

Negoescu DM, Frazier PI, Powell WB (2011) The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS J. Comput.* 23(3):346–363.

Pasupathy R, Hunter SR, Pujowidianto NA, Lee LH, Chen CH (2014) Stochastically constrained ranking and selection via SCORE. *ACM Trans. Modeling Comput. Simulation* 25(1):1:1–1:26.

Pokhilko V, Zhang Q, Kang L, Mays DP (2019) D-optimal design for network A/B testing. *J. Statist. Theory Practice* 13(4):61:1–61:23.

Rodriguez M, Jones B, Borror CM, Montgomery DC (2010) Generating and assessing exact G-optimal designs. *J. Quality Tech.* 42(1):3–20.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.

Sagnol G, Harman R (2015) Computing exact D-optimal designs by mixed integer second-order cone programming. *Ann. Statist.* 43(5):2198–2224.

Salagame RR, Barton RR (1997) Factorial hypercube designs for spatial correlation regression. *J. Appl. Statist.* 24(4):453–474.

Shen H, Hong LJ, Zhang X (2017) Ranking and selection with covariates. Chan WKV, D'Ambrogio A, Zacharewicz G, Mustafee N, Wainer G, Page E, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 2137–2148.

Shen H, Hong LJ, Zhang X (2021) Ranking and selection with covariates for personalized decision making. *INFORMS J. Comput.* 33(4):1500–1519.

Soare M, Lazaric A, Munos R (2014) Best-arm identification in linear bandits. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY), vol. 27, 828–836.

Wang H, Kim SH (2012) Reducing the conservativeness of fully sequential indifference-zone procedures. *IEEE Trans. Automatic Control* 58(6):1613–1619.

Yakowitz S, L'Ecuyer P, Vazquez-Abad F (2000) Global stochastic optimization with low-dispersion point sets. *Oper. Res.* 48(6):939–950.

Jiaqi Zhou is an assistant vice president at Citi. Her doctoral research dealt with simulation optimization and queueing analysis.

Ilya O. Ryzhov is a professor of operations management and management science in the Robert H. Smith School of Business at the University of Maryland. He works on models, theory, and applications in business analytics. He coauthored the book *Optimal Learning*, received the 2017 Outstanding Simulation Publication Award from the INFORMS Simulation Society, and was recognized on three separate occasions in the Best Theoretical Paper award competition at the Winter Simulation Conference.