Optimal Experimental Design in Heavy-Tailed Linear Regression

Anonymous authors
Paper under double-blind review

Abstract

Optimal experimental design concerns the selection and allocation of measurements from a given action set for accurate estimation. In this paper, we study and essentially resolve the heavy-tailed linear regression setting, where the observation noise only has a bounded α moment for some $1 < \alpha \leq 2$. We give a novel design criterion $M_\alpha(\lambda; \mathcal{A})$ evaluating a design λ on a finite action set $\mathcal{A}$, and prove that it tightly characterizes the estimation error rate: with the design λ, a suitable estimator achieves error proportional to $M_\alpha(\lambda; \mathcal{A})^{1/\alpha}\left(\frac{\log|\mathcal{A}|}{n}\right)^{(\alpha-1)/\alpha}$, while no estimator applied to this design λ can yield a better error rate up to constants and the $\log|\mathcal{A}|$ dependence. This new quantity $M_\alpha(\lambda; \mathcal{A})$ coincides with the classical G-optimal design criterion when $\alpha = 2$, and is convex in λ with an efficiently-computable optimizer λ^*. We further demonstrate a qualitative separation between λ^* and the G-optimal design by a factor of $\tilde{\Omega}(d^{1/\alpha-1/2})$ difference in the error. Lastly, we address the problem of finding a (near-)optimal design without knowledge of α. We construct a single α-agnostic design whose error matches that of the known-α design up to an $O(\sqrt{\log d})$ factor, and we provide a lower bound showing that an $\Omega\left(\sqrt{\log d/\log\log d}\right)$ factor is unavoidable.

1 Introduction

Choosing where to collect data is a fundamental statistical problem, particularly when measurements are costly. For instance, an experimenter may need to select sensor locations, experimental conditions, or inputs at which a model will be evaluated (Huan et al., 2024). Optimal experimental design determines how to allocate a limited number of observations among these possible measurements. The allocation problem also arises in machine learning, including data selection for regression and pure exploration in linear bandits (Singh & Xie, 2020; Degenne et al., 2020).

In linear regression, measurements share a common parameter θ^*, so observations collected at some inputs are informative about the responses at others. Classical G-optimal design exploits this structure to collect data so as to minimize the largest estimation error over the inputs. With finite-variance noise, this is achieved by minimizing the largest variance of the estimates (Pukelsheim, 2006).

However, in some settings, measurement noise can be sufficiently heavy-tailed that it has infinite variance (Antypas et al., 2013; Hannsgen, 2012; Crovella & Bestavros, 1997), necessitating different strategies for achieving accurate estimation. In such settings, it is natural to assume a finite α-moment for some $\alpha \in (1, 2)$, as opposed to the finite-variance case $\alpha = 2$. The problem of robust mean estimation and regression under such assumptions has been studied in detail, e.g., see (Lugosi & Mendelson, 2019b; Cherapanamjeri et al., 2022). In contrast, the problem of *optimal data collection* in such settings has received considerably less attention.

One approach would be to simply adopt the G-optimal design despite the fact that it comes from theory for the finite-variance setting. The natural question, then, is whether this is still optimal, and if not, what the "right" design is. In this paper, we show that G-optimal design can be strictly suboptimal (by polynomial factors in the dimension) no matter what estimator it is paired with, and propose an alternative design that overcomes this limitation and provably near-optimal. We now state the problem and summarize our contributions in more detail.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Problem Setup

Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite action set spanning $\mathbb{R}^d$, and let $\theta^* \in \mathbb{R}^d$ be the unknown parameter. A design is a probability mass function $\lambda : \mathcal{A} \to [0,1]$ satisfying $\sum_{a\in\mathcal{A}} \lambda(a) = 1$. After choosing λ, we observe n independent pairs (x_i, y_i), with $x_i \sim \lambda$ and

$$y_i = x_i^\top \theta^* + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid x_i] = 0, \qquad \mathbb{E}[|\varepsilon_i|^\alpha \mid x_i] \le \nu \tag{1}$$

for some $1 < \alpha \le 2$ and $\nu > 0$. Importantly, the conditional noise distribution may depend on the action, thus allowing reward models such as scaled Bernoulli instead of only "purely additive" models. We assume that the distributions $(\epsilon_i|x_i)$ are independent across i, but not necessarily identical.

For a budget n, the objective is to choose a design λ and an estimator $\widehat{\theta}$ based on the observations to minimize the largest estimation error over the action set:

$$\max_{a\in\mathcal{A}} \left| a^\top \widehat{\theta} - a^\top \theta^* \right|. \tag{2}$$

For $\delta \in (0,1)$, we require an error bound that holds with probability at least $1-\delta$ for every parameter and noise distribution satisfying Eq. (1). Note that while this problem formulation restricts to i.i.d. designs, we will handle the case of general nonadaptive designs in Section 5.

1.2 Contributions and Paper Outline

Our main contributions are outlined as follows:

- **A separation from classical G-optimal design.** For every fixed $1 < \alpha < 2$, Example 1 constructs an action set on which the largest estimation error of every estimator using the G-optimal design exceeds the error achieved under another design by a factor $d^{1/\alpha - 1/2}$, up to logarithmic factors.
- **An optimal design criterion for heavy-tailed noise.** We introduce a design criterion quantity $M_\alpha(\lambda; \mathcal{A})$ for a design λ on an action set $\mathcal{A}$. This criterion is convex in λ, can be minimized by a finite convex program, and equals the classical G-optimal objective at $\alpha = 2$. Theorem 1 gives an estimation bound proportional to $M_\alpha(\lambda;\mathcal{A})^{1/\alpha}\left(\frac{\log|\mathcal{A}|}{n}\right)^{(\alpha-1)/\alpha}$ after optimizing the design. The lower bounds in Theorem 2 and Corollary 1 show that the joint dependence on the design value and sample size cannot be improved (i.e., tightness up to $\log|\mathcal{A}|$ dependence and constants). We further show that for every action set spanning $\mathbb{R}^d$, we have
  $$d^{(\alpha-1)/\alpha} \le M_\alpha(\mathcal{A})^{1/\alpha} \le \sqrt{d},$$
  and the dimension exponents on both sides are sharp.
- **Design for unknown α.** We construct a finite computable design whose estimation bound is within a factor $C\sqrt{\log d}$ of the bound obtained with an optimal design for each $1 < \alpha \le 2$. Conversely, we show that certain action sets require an approximation factor of at least $c\sqrt{\log d / \log\log d}$ when α is unknown.

We further extend the lower bounds to general nonadaptive designs in Section 5. In Section 6 we describe polynomial separations for classical V- and A-optimality, which are alternative classical design criteria. Experiments are provided in Section 7, and the related work is discussed in Section 8.

2 A Separation from G-Optimality

In this section, we establish the strict suboptimality of the classical G-optimal design when $\alpha \in (1,2)$, by constructing a simple example and comparing its attainable estimation accuracy against an alternative design.

For a design λ, define its *information matrix* as

$$A(\lambda) = \mathbb{E}_{x\sim\lambda}[xx^\top]. \tag{3}$$

The classical G criterion is $\max_{a\in\mathcal{A}} a^\top A(\lambda)^{-1} a$, and is based on controlling the worst-case variance among all target actions. Here and subsequently, criteria involving $A(\lambda)^{-1}$ are assigned the value

$+\infty$ when $A(\lambda)$ is singular. If $\mathcal{A}$ spans $\mathbb{R}^d$, the Kiefer–Wolfowitz equivalence theorem gives the minimum G value as d (Pukelsheim, 2006; Kiefer & Wolfowitz, 1959). Example 1 gives an example in which an allocation exists with smaller worst-case estimation error than that of the G-optimal design paired with any estimator. The proofs of its claims are given in Appendix A.6.

Example 1 (A polynomial separation from G-optimality)**.** *Fix $1 < \alpha < 2$ and let $d \geq 2$ be a power of two. Let $h_1, \ldots, h_d \in \mathbb{R}^d$ be the columns of the normalized Walsh–Hadamard matrix of order d. These vectors are orthonormal and satisfy $|\mathbf{e}_j^\top h_i| = d^{-1/2}$ for every i, j. Set*

$$\mathcal{A} = \{h_1, \ldots, h_d\} \cup \{\mathbf{e}_1/2, \ldots, \mathbf{e}_d/2\}.$$

The unique G-optimal design samples uniformly from $\{h_1, \ldots, h_d\}$. There are universal positive constants c and C such that the following statements hold. Under a conditional α moment bound on the noise, for every estimator based on the G-optimal design there exist a parameter θ^* and a noise distribution satisfying Eq. (1) such that, with probability at least $1/4$,

$$\max_{a \in \mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| \geq c\nu^{1/\alpha}\sqrt{d}\, n^{-(\alpha-1)/\alpha}.$$

The design that samples uniformly from all $2d$ actions is not G-optimal. When $n \geq Cd\log(2d/\delta)$, direct robust mean estimation on each action gives, with probability at least $1 - \delta$,

$$\max_{a \in \mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| \leq C\nu^{1/\alpha}\left(\frac{d\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha}.$$

The ratio between the lower and upper error bounds grows as $d^{1/\alpha - 1/2}$ up to logarithmic factors.

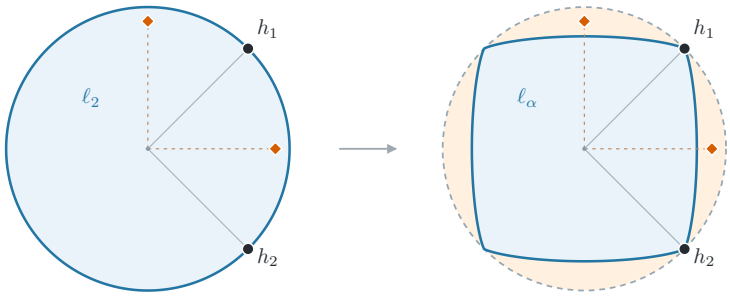


Figure 1: Error geometry under the G-optimal design. The sampled Hadamard actions (black) lie on both contours. The rescaled coordinate actions (orange) lie inside the ℓ_2 contour at $\alpha = 2$ and outside the ℓ_α contour when $\alpha < 2$.

Under uniform sampling from the Hadamard basis, the error scale for an action a is

$$d^{(\alpha-1)/\alpha}\left(\sum_{i=1}^{d} |h_i^\top a|^\alpha\right)^{1/\alpha} = \begin{cases} d^{(\alpha-1)/\alpha}, & a \in \{h_1, \ldots, h_d\}, \\ \sqrt{d}/2, & a \in \{\mathbf{e}_1/2, \ldots, \mathbf{e}_d/2\}. \end{cases}$$

Figure 1 compares this quantity for the Hadamard and coordinate actions. The coordinate-to-Hadamard ratio is $d^{1/\alpha - 1/2}/2$. It equals $1/2$ when $\alpha = 2$ and grows polynomially in d for every fixed $\alpha < 2$. Thus, the G-optimal design correctly treats the coordinate actions as easier when $\alpha = 2$, but this ordering reverses for every $\alpha < 2$.

3 Optimal Design for a Known Moment Exponent

Having established that G-optimal design is suboptimal in general, we now turn to the problem of identifying the "right" design criterion for the heavy-tailed setting. We consider the known-α setting in this section, and turn to α-agnostic designs in Section 4.

For a design λ on $\mathcal{A}$ and $1 < \alpha \le 2$, we define design criterion

$$M_\alpha(\lambda; \mathcal{A}) = \max_{a \in \mathcal{A}} \left(\inf_{u \in \mathbb{R}^d:\, u^\top a = 1} \mathbb{E}_{x \sim \lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{-(\alpha-1)}. \tag{4}$$

Theorems 1 and 2 show that, up to logarithmic factors, the dependence of the maximum estimation error on the design and the sample size is $M_\alpha(\lambda; \mathcal{A})^{1/\alpha} n^{-(\alpha-1)/\alpha}$. Thus, minimizing $M_\alpha(\lambda; \mathcal{A})$ minimizes the design-dependent term in the estimation error. The optimal design value is $M_\alpha(\mathcal{A}) = \inf_\lambda M_\alpha(\lambda; \mathcal{A})$, where the infimum is over designs on $\mathcal{A}$. At $\alpha = 2$, the criterion reduces exactly to G-optimality. If $A(\lambda)$ is nonsingular, constrained quadratic minimization gives

$$M_2(\lambda; \mathcal{A}) = \max_{a \in \mathcal{A}} \left(\inf_{u:\, u^\top a = 1} u^\top A(\lambda) u \right)^{-1} = \max_{a \in \mathcal{A}} a^\top A(\lambda)^{-1} a. \tag{5}$$

For $a \neq 0$, the infimum is attained at $u = A(\lambda)^{-1} a / (a^\top A(\lambda)^{-1} a)$. The identity in Eq. (5) and the Kiefer–Wolfowitz equivalence theorem give $M_2(\mathcal{A}) = d$ whenever $\mathcal{A}$ spans $\mathbb{R}^d$ (Pukelsheim, 2006; Kiefer & Wolfowitz, 1959).

The following lemma shows that the optimal design can be computed by a finite convex program.

Lemma 1 (Convexity and computation)**.** *For a finite action set $\mathcal{A}$ and fixed $1 < \alpha \le 2$, $M_\alpha(\lambda; \mathcal{A})$ is convex in λ, and $M_\alpha(\mathcal{A})$ can be computed by a finite convex program. If $M_\alpha(\lambda; \mathcal{A}) < \infty$, then, for every $a \in \mathcal{A}$, there exists scalars $(v_x)_{x:\, \lambda(x) > 0}$ such that*

$$\sum_{x \in \mathcal{A}:\, \lambda(x) > 0} v_x x = a, \qquad \sum_{x \in \mathcal{A}:\, \lambda(x) > 0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} \le M_\alpha(\lambda; \mathcal{A}).$$

Proof Sketch. For each $a \in \mathcal{A}$, the inner infimum in Eq. (4) is concave in λ because it is an infimum of linear functions, so inverse of it is convex. Raising the inverse to the power $\alpha - 1$ preserves convexity because this power is convex and nonincreasing. Taking the maximum over $a \in \mathcal{A}$ also preserves convexity. Hölder duality then gives for every $a \in \mathcal{A}$ that

$$\left(\inf_{u:\, u^\top a = 1} \mathbb{E}_{x \sim \lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{-(\alpha-1)} = \min_{\sum_{x:\, \lambda(x) > 0} v_x x = a} \sum_{x:\, \lambda(x) > 0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}}.$$

The right-hand side is jointly convex in (v_x) and λ; the proof in Appendix A.1 verifies this directly from convexity of $s \mapsto |s|^\alpha$. Using one copy of (v_x) for each action, together with the linear constraints $\sum_x v_x x = a$, gives the finite convex program. Minimizing over (v_x) and then taking the maximum over $a \in \mathcal{A}$ preserves convexity in λ. □

3.1 Estimation bounds

The next two theorems show that $M_\alpha(\lambda; \mathcal{A})$ characterizes the difficulty of estimation under a fixed design. The upper bound constructs an estimator for any design, while the lower bound applies to every estimator. The bounds have the same dependence on the design objective and the sample size.

Theorem 1 (Upper bound for a fixed i.i.d. design)**.** *Fix a finite action set $\mathcal{A}$, a design λ, and $\delta \in (0, 1)$. Let $B = \max_{a \in \mathcal{A}} |a^\top \theta^*|$. There is an estimator $\widehat{\theta}$ and a universal constant C such that, whenever $n \ge C \log(2|\mathcal{A}|/\delta)$, with probability at least $1 - \delta$,*

$$\max_{a \in \mathcal{A}} |a^\top (\widehat{\theta} - \theta^*)| \le C(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{\log(2|\mathcal{A}|/\delta)}{n} \right)^{(\alpha-1)/\alpha}.$$

Note that the upper bound depends on the maximum signal magnitude B as well as the noise moment ν. The estimator uses α, but requires neither B nor ν.

Proof Sketch. For every $a \in \mathcal{A}$, choose the coefficients guaranteed by Lemma 1, which satisfy

$$\sum_{x \in \mathcal{A}:\, \lambda(x) > 0} v_x(a) x = a, \qquad \sum_{x \in \mathcal{A}:\, \lambda(x) > 0} \frac{|v_x(a)|^\alpha}{\lambda(x)^{\alpha-1}} \le M_\alpha(\lambda; \mathcal{A}).$$

For each $a \in \mathcal{A}$, let $\widehat{f}_a$ be the median-of-means estimate computed from

$$\left(\frac{v_{x_i}(a)}{\lambda(x_i)} y_i\right)_{i=1}^{n}.$$

The first coefficient identity makes these transformed responses unbiased for $a^\top \theta^*$, while the second controls their centered α-moment. Finally, set

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \max_{a \in \mathcal{A}} |a^\top \theta - \widehat{f}_a|.$$

A standard scalar median-of-means guarantee, followed by this minimum-distance step, can be combined to give the stated bound. □

For the lower bound, we construct an instance in which information distinguishing two parameters appears only in rare responses. The probability of observing such a response is controlled by the projection moment in Eq. (4).

Theorem 2 (Lower bound for a fixed i.i.d. design)**.** *Fix a finite action set $\mathcal{A}$, a design λ, and parameters $\alpha \in (1, 2)$ and $\nu > 0$. There is a universal positive constant c such that, for every estimator $\widehat{\theta} = \widehat{\theta}((x_i, y_i)_{i=1}^n)$ based on n i.i.d. samples from λ, there is a reward distribution satisfying Eq. (1) with moment bound ν such that*

$$\Pr\left(\max_{a \in \mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| \geq c\nu^{1/\alpha} \min\left\{M_\alpha(\lambda; \mathcal{A})^{1/\alpha} n^{-(\alpha-1)/\alpha}, 1\right\}\right) \geq \frac{1}{4}.$$

Moreover, this instance satisfies $B^\alpha \leq \nu$ for B defined in Theorem 1, meaning that $(\nu + B^\alpha)^{1/\alpha}$ and $\nu^{1/\alpha}$ coincide to within a universal constant.

Proof Sketch. Let u be the the minimizer in Eq. (4), that is

$$\max_{a \in \mathcal{A}} |u^\top a| = 1, \qquad \mathbb{E}_{x \sim \lambda} |u^\top x|^{\alpha/(\alpha-1)} = M_\alpha(\lambda; \mathcal{A})^{-1/(\alpha-1)}.$$

We compare $\theta^* = \Delta u$ and $\theta' = -\Delta u$, with $\Delta > 0$ at the scale in the theorem. The conditional rewards are zero except for rare observations whose size ensures the required mean and moment bound. The probability of a nonzero observation is proportional to $|\Delta u^\top x|^{\alpha/(\alpha-1)}$. Its expectation therefore determines how well the two models can be distinguished from n samples. Le Cam's inequality then gives the desired lower bound. The construction, moment calculation, and constants are in Appendix A.3. □

As noted earlier, Theorems 1 and 2 match in the $M_\alpha(\lambda; \mathcal{A})^{1/\alpha}$ and n dependencies, at least in the regime of interest before the lower bound saturates. For settings where $|\mathcal{A}|$ is very large, the upper bound can be improved; specifically, Corollary 3 in Appendix A.5 gives an alternative that depends on the ambient dimension rather than the cardinality.

3.2 Dependence on the dimension

Recall that under finite variance, the optimal G-design cost is d for every action set spanning $\mathbb{R}^d$, thus establishing the dimension dependence. In the heavy-tailed setting, the dimension dependence enters through $M_\alpha(\mathcal{A})$, but it turns out to have a range of possible scalings. The following theorem characterizes this range, determining the lower endpoint exactly and the upper one up to constants.

Theorem 3 (Dimension range of the optimal design value)**.** *Fix $1 < \alpha \leq 2$ and a finite action set $\mathcal{A}$ spanning $\mathbb{R}^d$. Then*

$$d^{(\alpha-1)/\alpha} \leq M_\alpha(\mathcal{A})^{1/\alpha} \leq \sqrt{d}. \tag{6}$$

Moreover, both dimension exponents are sharp: The lower bound in Eq. (6) is an equality for $\mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, whereas for every $d \geq 2$, there is a finite action set spanning $\mathbb{R}^d$ for which

$$M_\alpha(\mathcal{A})^{1/\alpha} \geq c\sqrt{\frac{\alpha - 1}{\alpha}}\sqrt{d},$$

where c is a universal positive constant.

Proof Sketch. Boundedness of each projection on the action set and monotonicity of moments give, for every design,

$$M_2(\lambda;\mathcal{A})^{(\alpha-1)/\alpha} \le M_\alpha(\lambda;\mathcal{A})^{1/\alpha} \le M_2(\lambda;\mathcal{A})^{1/2}.$$

The Kiefer–Wolfowitz theorem then yields both bounds on the optimal value. Further details are given in Appendix A.4, along with the two endpoint examples. □

4 Design for Unknown α

In this section, we consider the problem of finding a design with small M_α when α is unknown. We note that the M_α-dependent upper bound in Theorem 1 still requires the *estimator* to know α, and avoiding that is left for future work; here we only focus on the *design* step with unknown α.

The difficulty is that a design optimized for one exponent can then allocate too few observations to actions that are important at another exponent, in analogy with how G-optimal design is suboptimal when $\alpha < 2$. Our solution is to *mix designs across a finite subset of α values*, and we show that this gives an approximation factor $C\sqrt{\log d}$ relative to the optimal design for each $1 < \alpha \le 2$ (Theorem 4). The lower bound in Theorem 5 shows that $c\sqrt{\log d/\log\log d}$ is unavoidable, so the sufficient and necessary approximation factors differ by only $O(\sqrt{\log\log d})$.

Theorem 4 (Design for unknown α)**.** *Fix a finite action set $\mathcal{A} \subset \mathbb{R}^d$ with $d \ge 3$ and Span$(\mathcal{A}) = \mathbb{R}^d$. There is a design λ_{fin} computable by finitely many convex programs and a universal constant C such that, for every $1 < \alpha \le 2$,*

$$M_\alpha(\lambda_{\text{fin}};\mathcal{A})^{1/\alpha} \le C\sqrt{\log d}\, M_\alpha(\mathcal{A})^{1/\alpha}.$$

Proof Sketch. We solve the design problem on a geometric grid in the projection exponent $\alpha/(\alpha-1)$, with successive exponents differing by a factor $1 + 1/\log d$. We mix the resulting designs with weights that decrease exponentially in that projection exponent. The reciprocal mixture weight, raised to the power $(\alpha-1)/\alpha$, costs at most a constant multiple of $\sqrt{\log d}$. Moment monotonicity compares neighboring exponents, and Theorem 3 bounds the remaining discretization factor by a constant. We stop the grid at a projection exponent of at least $2\log(2|\mathcal{A}|)$ and assign the remaining mass to the uniform design. The uniform component controls the criterion for α smaller than the smallest grid value. The full details are given in Appendix B.1. □

The lower bound arises because different exponents favor different groups of actions, and a single design must divide its budget among these groups.

Theorem 5 (A necessary loss for unknown α)**.** *There is a universal constant $c > 0$ and infinitely many dimensions d for which there is a finite spanning action set $\mathcal{A} \subset \mathbb{R}^d$ consisting of unit vectors such that every design λ on $\mathcal{A}$ satisfies*

$$M_\alpha(\lambda;\mathcal{A})^{1/\alpha} \ge c\sqrt{\frac{\log d}{\log\log d}}\, M_\alpha(\mathcal{A})^{1/\alpha}$$

for at least one $1 < \alpha < 2$.

Proof Sketch. We construct an action set from several orthogonal groups, each containing unions of mutually unbiased orthonormal bases. The group dimensions and numbers of bases are chosen so that a different group determines the optimal allocation for each of several exponents. Any single design assigns small mass to at least one group. For the exponent associated with that group, a design chosen with knowledge of the exponent achieves a smaller cost. The details of the construction and the comparison are given in Appendix B.2. □

5 General Nonadaptive Allocations

In this section, we ask the question of how *general nonadaptive designs* compare to the class of i.i.d. designs that we have focused on so far. That is, we now allow $x_1, \dots, x_n$ to be drawn from an *arbitrary joint distribution*, including the case of an arbitrary deterministic sequence. The only

constraint is being *nonadaptive*, meaning all points must be selected (possibly randomly) prior to any data being collected. The upper bound in Theorem 1 remains valid since i.i.d. designs fall within this broader class, and we focus our attention on extending the lower bounds.

To do so, we associate a general nonadaptive design with the following distributional design:

$$\lambda(a) = \frac{1}{n}\sum_{i=1}^{n} \Pr(x_i = a), \qquad a \in \mathcal{A}. \tag{7}$$

Here $n\lambda(a)$ is the expected number of observations at action a. In the construction for Theorem 2, the two models produce identical observations unless a rare nonzero reward is observed. The probability of such an observation is bounded by the sum of the conditional probabilities of a nonzero reward. After averaging over the allocation, this sum depends only on the expected action counts, so the same indistinguishability argument applies to the allocation and to n independent draws from λ. Lemma 4 in Appendix C gives the corresponding total variation bound, and the final result is given as follows.

Corollary 1 (Lower bound over nonadaptive allocations)**.** *Fix $1 < \alpha \le 2$, a finite action set $\mathcal{A}$, and $\nu > 0$. For every nonadaptive allocation of n actions and every estimator, there is a parameter and a reward distribution satisfying Eq. (1) for which*

$$\Pr\left(\max_{a\in\mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| \ge c\nu^{1/\alpha} \min\left\{ n^{-(\alpha-1)/\alpha} M_\alpha(\mathcal{A})^{1/\alpha}, 1 \right\}\right) \ge \frac{1}{4},$$

where c is a universal positive constant.

An allocation chosen without knowing α also induces a design λ independent of α. Applying Theorem 5 to this design preserves the necessary approximation factor for unknown α.

Corollary 2 (Nonadaptive allocations with unknown α)**.** *There is a universal positive constant c and infinitely many dimensions d for which a finite spanning action set $\mathcal{A} \subset \mathbb{R}^d$ has the following property. Fix a budget $n \ge 1$, a moment bound $\nu > 0$, and a nonadaptive allocation of n actions chosen without knowing α. For every estimator, there are $1 < \alpha < 2$, a parameter, and a reward distribution satisfying Eq. (1) for which*

$$\Pr\left(\max_{a\in\mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| \ge c\nu^{1/\alpha} \min\left\{ \sqrt{\frac{\log d}{\log\log d}} n^{-(\alpha-1)/\alpha} M_\alpha(\mathcal{A})^{1/\alpha}, 1 \right\}\right) \ge \frac{1}{4}.$$

The proofs of Corollaries 1 and 2 are given in Appendix C.

Collectively, these results show that our lower bounds for i.i.d. designs also extend to general nonadaptive designs (with suitably-defined λ). Since our upper and lower bounds for i.i.d. designs are near-matching, this reveals that i.i.d. designs are already near-optimal within the broader class.

6 Other Estimation Criteria

Thus far we have focused on the worst-case estimation error described in (2), which leads to G-optimal design under finite variance and our M_α design in the heavy-tailed setting. The classical literature (outlined in Section 8) additionally handles several other objectives, such as parameter vector estimation or average-case error over actions. In this section, we explore such objectives associated with the classical A-optimal, V-optimal, and E-optimal designs, in particular showing that the former two designs again lead to suboptimality by a polynomial factor in heavy-tailed settings. We provide only an outline here, and defer the details to the appendices.

A-optimal design. For parameter estimation, the loss is $\|\widehat{\theta} - \theta^*\|_2$. This objective leads to the classical A-optimal design of minimizing $\mathrm{Tr}(A(\lambda)^{-1})$. With heavy-tailed errors, this design can allocate too few observations to a short action, whose response must be amplified to estimate its parameter coordinate. For instance, at $\alpha = 3/2$, consider

$$\mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_{d-1}, d^{-19/24}\mathbf{e}_d\}.$$

For $n \geq Cd\log(2d/\delta)$, every estimator using the classical A-optimal design has a reward distribution for which the Euclidean error is at least $c\nu^{2/3}d^{31/36}n^{-1/3}$ with probability at least $1/8$. A different allocation admits an estimator with Euclidean error at most $C\nu^{2/3}d^{5/6}(\log(2d/\delta)/n)^{1/3}$ with probability at least $1-\delta$. The ratio of these error bounds grows as $d^{1/36}$ up to logarithmic factors. Example 3 in Appendix E gives the allocations and extends the separation to every $1 < \alpha < 2$.

V-optimal design. For accuracy averaged over the actions, the loss is $(|\mathcal{A}|^{-1}\sum_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)|^2)^{1/2}$. This objective leads to the classical V-optimal design of minimizing $|\mathcal{A}|^{-1}\sum_{a\in\mathcal{A}}a^\top A(\lambda)^{-1}a$. Example 2 in Appendix D gives a separation at $\alpha = 3/2$ by placing many target actions along one coordinate, so that estimating this coordinate dominates the average error. For $n \geq Cd\log(2d/\delta)$, the classical design has error at least $c\nu^{2/3}d^{1/18}n^{-1/3}$ with probability at least $1/4$ under a suitable reward distribution, regardless of the estimator. A different allocation achieves error at most $C\nu^{2/3}(\log(2d/\delta)/n)^{1/3}$ with probability at least $1-\delta$. The error ratio therefore grows as $d^{1/18}$ up to logarithmic factors. Theorems 6 and 7 in Appendix D also determine the minimax rate over finite action sets after optimizing the design, under the signal bound $B^\alpha \leq \nu$.

E-optimal design. Classical E-optimality minimizes the largest directional variance, rather than the sum of coordinate variances. Its α-moment analogue is defined in Eq. (78) in Appendix F. Theorem 10 in Appendix F shows that a classical E-optimal design approximates the minimum of the $1/\alpha$ power of this moment criterion to within a factor of $8\sqrt{\alpha/(\alpha-1)}$. Hence, the polynomial separations for G-, A-, and V-optimality do not extend to the E-optimal design criterion.

7 Experiments

In this section, we provide a simple proof-of-concept experiment supporting the strict separation between the G-optimal design and M_α that we established in Section 2. We compare the two designs using the worst-case estimation error objective in Eq. (2). Each design receives the same budget n, with $\theta^* = \mathbf{e}_1$ and conditional noise moment bound $\nu = 1$.

We use the action set

$$\mathcal{A} = \{h_1, \ldots, h_d\} \cup \{0.99\mathbf{e}_1, \ldots, 0.99\mathbf{e}_d\},$$

where $h_1, \ldots, h_d$ form a normalized Hadamard basis. Recall from Example 1 that the unique G-optimal design is uniform on the Hadamard basis. Moreover, a symmetry argument reduces the M_α optimization to the total mass assigned to the Hadamard and coordinate bases. The two designs coincide at $\alpha = 2$. Appendix G gives the sampling allocations and estimator implementations.

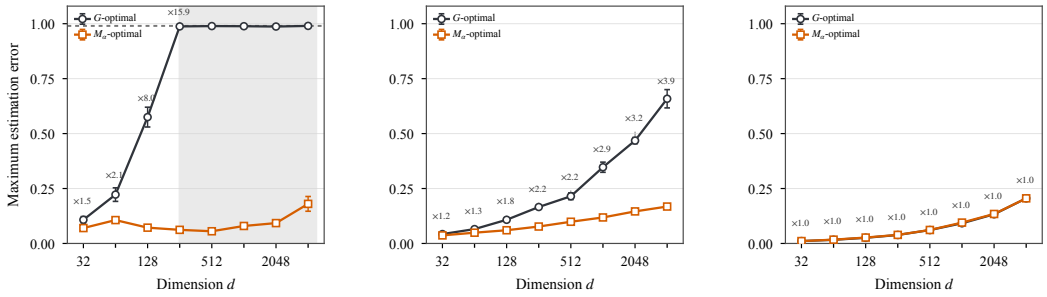
We consider three response distributions, each of which has conditional mean $x^\top\theta^*$ and conditional centered α moment at most one:

- **Rare rewards.** At action x, the response is zero with high probability and takes a nonzero value otherwise. The probability and nonzero value are calibrated to give mean $x^\top\theta^*$ and centered α moment ≤ 1.
- **Pareto.** A Pareto variable with lower endpoint one and shape $\alpha + 0.02$, centered by subtracting its expectation, gives asymmetric additive noise.
- **Student t.** The noise is an additive Student variable with $\alpha + 0.2$ degrees of freedom.

The Pareto and Student noises are divided by their centered absolute α moments to the power $1/\alpha$.

Figure 2 reports the maximum estimation error as the dimension increases. At each dimension and under each design, the plotted value is the smallest repetition-averaged maximum estimation error among the same collection of estimators. The collection includes least squares, regularized regression, robust losses, gradient and filtering methods, and robust mean reconstruction; see Appendix G for the relevant estimator definitions.

For rare rewards, the G-optimal error is at least 0.987 from $d = 256$ onward, whereas the M_α-optimal error remains below 0.19 over the entire range. Under Pareto noise, the ratio increases from 1.2 at $d = 32$ to 3.9 at $d = 4096$. The two designs perform similarly under Student t noise, showing that the moment condition alone does not always force a separation. Appendix G contains the complete protocol and additional comparisons.



(a) Rare rewards. (b) Pareto noise. (c) Student t noise.

Figure 2: Maximum estimation error on the Hadamard and coordinate action set, with $\alpha = 4/3$ and $n = 262{,}144$. Each point averages the maximum error over sixteen repetitions, and bars show empirical standard errors. The labels give the G-optimal error divided by the M_α-optimal error at each dimension. In panel (a), the dotted line is the error 0.99 of the all-zero estimator, so for $d \geq 256$, the best estimator using G-optimal collected data is no better than trivial estimator.

The designs perform comparably at the smallest dimension, and the advantage of changing the design increases with d. At $d = 4096$, median-of-means reconstruction under M_α has about one quarter of the error of the best tested estimator under G-optimal sampling. The errors under both designs increase with d, but the gap between them widens, in agreement with Example 1. Appendix G.4 includes the comparison with the design that does not use α.

8 Related Work

Classical design theory optimizes variance criteria built from the information matrix (Pukelsheim, 2006). Approximation algorithms, convex relaxations, and rounding schemes have been developed for finite and combinatorial design problems (Singh & Xie, 2020; De Castro et al., 2019). Bayesian design optimizes utilities such as expected information gain, with computational methods based on variational approximations (Rainforth et al., 2024; Dong et al., 2025). Sequential methods choose experiments using previous observations (Shen et al., 2025), while robust design also studies sensitivity to misspecified priors and measurement models (Attia et al., 2025).

Robust estimation under finite variance has been studied for scalar means, multivariate means, and regression (Catoni, 2012; Audibert & Catoni, 2011; Hsu & Sabato, 2014; 2016; Lugosi & Mendelson, 2019a; Cherapanamjeri et al., 2020). For infinite variance, Cherapanamjeri et al. (2022) give optimal multivariate mean estimation under directional moment bounds. The survey of Lugosi & Mendelson (2019b) covers robust mean and regression methods. These results concern estimation from a given distribution of observations. Here the sampling distribution is also a decision variable.

Experimental design also enters linear bandit algorithms for identification and regret minimization (Degenne et al., 2020). Tajdini et al. (2025) use a moment design criterion in a phased elimination algorithm for heavy-tailed linear bandits. Their regression estimator combines robust scalar mean estimation with minimum-distance aggregation, as does our upper-bound estimator. Their reconstruction coefficients are determined by the regularized information matrix, and their design criterion controls the moments of these reconstructions. We instead optimize over all unbiased linear reconstruction coefficients as well as the sampling design. This optimization yields a convex criterion with estimation upper and lower bounds that match in the criterion and sample size, including a lower bound for every estimator.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods for our numerical experiments and assist in the writing of proofs. We have not used generative AI tools for translation or interpretation of results, and synthetic data generation, dataset cleaning, and qualitative analysis are not applicable to this work. Additionally, we used generative AI tools for drafting parts of a research paper and editing text. We have reviewed all AI-assisted work, and we wrote and edited major parts, using the draft as an overall structural guideline. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Antonios Antypas, Phoebe Koundouri, and Nikolaos Kourogenis. Aggregational gaussianity and barely infinite variance in financial returns. *Journal of Empirical Finance*, 20:102–108, 2013.

Ahmed Attia, Sven Leyffer, and Todd S. Munson. Robust a-optimal experimental design for sensor placement in bayesian linear inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 13(2):744–774, 2025.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Annals of statistics*, 39(5):2766–2794, 2011.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. In *Annales de l'IHP Probabilités et Statistiques*, volume 48, pp. 1148–1185, 2012.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 786–806. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/cherapanamjeri19b.html.

Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *ACM SIGACT Symposium on Theory of Computing*, pp. 601–609, 2020.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, and Michael Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pp. 356–357, 2022.

M.E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997. doi: 10.1109/90.650143.

Aniket Das, Dheeraj Nagaraj, Soumyabrata Pal, Arun Suggala, and Prateek Varshney. Near-optimal streaming heavy-tailed statistical estimation with clipped sgd. *Advances in Neural Information Processing Systems*, 37:8834–8900, 2024.

Yohann De Castro, Fabrice Gamboa, Didier Henrion, Roxana Hess, and Jean-Bernard Lasserre. Approximate optimal designs for multivariate polynomial regression. *Annals of Statistics*, 47(1): 127–155, 2019.

Rémy Degenne, Pierre Menard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, volume 119, pp. 2432–2442. PMLR, 13–18 Jul 2020.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pp. 1596–1606. PMLR, 2019.

Jiayuan Dong, Christian Jacobsen, Mehdi Khalloufi, Maryam Akram, Wanjiao Liu, Karthik Duraisamy, and Xun Huan. Variational Bayesian optimal experimental design with normalizing flows. *Computer Methods in Applied Mechanics and Engineering*, 433:117457, 2025.

Greg Hannsgen. Infinite-variance, alpha-stable shocks in monetary svar. *International Review of Applied Economics*, 26(6):755–786, 2012.

Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912, 2020.

Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pp. 37–45, 2014.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024.

Jack Kiefer and Jacob Wolfowitz. Optimum designs in regression problems. *The annals of mathematical statistics*, 30(2):271–294, 1959.

Nicholas LeCompte, William J Martin, and William Owens. On the equivalence between real mutually unbiased bases and a certain class of association schemes. *European Journal of Combinatorics*, 31 (6):1499–1512, 2010.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 2020.

Jasper CH Lee and Paul Valiant. Optimal sub-Gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 672–683. IEEE, 2022.

Jasper CH Lee, Walter McKelvie, Maoyuan Song, and Paul Valiant. All-purpose mean estimation over r: Optimal sub-gaussianity with outlier robustness and low moments performance. In *Forty-second International Conference on Machine Learning*, 2025.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019a.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019b.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 4(3):253–292, 2022.

Roberto I Oliveira and Lucas Resende. Trimmed sample means for robust uniform mean estimation and regression. *The Annals of Statistics*, 53(5):2153–2178, 2025.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association*, 120(550):1002–1013, 2025.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006. doi: 10.1137/1.9780898719109. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719109.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Wanggang Shen, Jiayuan Dong, and Xun Huan. Variational sequential optimal experimental design using reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 444: 118068, 2025.

Mohit Singh and Weijun Xie. Approximation algorithms for d-optimal design. *Mathematics of Operations Research*, 45(4):1512–1534, 2020. doi: 10.1287/moor.2019.1041.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Artin Tajdini, Jonathan Scarlett, and Kevin Jamieson. Improved regret bounds for linear bandits with heavy-tailed rewards. *arXiv preprint arXiv:2506.04775*, 2025.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pp. 299–303, 1965.

A Proofs for a Known Moment Exponent

A.1 Convexity and computation

Consider the inner minimization problem in Eq. (4). We first give an equivalent finite minimization for every $a \in \mathcal{A}$. This equivalence is used both to compute the design and to construct the estimator in Theorem 1.

Lemma 2 (Equivalent finite minimization)**.** *Fix $1 < \alpha \leq 2$, a design λ, and a nonzero $a \in$ Span$(\{x \in \mathcal{A} : \lambda(x) > 0\})$. Then*

$$\min_{\sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x x = a} \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} = \left(\inf_{u\in\mathbb{R}^d:\,u^\top a=1} \mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{-(\alpha-1)}. \tag{8}$$

Proof. Let the real numbers $(v_x)_{x\in\mathcal{A}:\,\lambda(x)>0}$ satisfy $\sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x x = a$. For every u with $u^\top a = 1$, Hölder's inequality gives

$$\begin{aligned} 1 &= \left| \sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x u^\top x \right| \\ &\leq \left(\sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} \right)^{1/\alpha} \left(\mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}. \end{aligned}$$

Taking the infimum over u and then the minimum over (v_x) proves the "$\geq$" direction of Eq. (8). Orthogonally projecting u onto Span$(\{x \in \mathcal{A} : \lambda(x) > 0\})$ leaves $u^\top a$ and every $u^\top x$ with $\lambda(x) > 0$ unchanged. On the unit sphere of this space, $\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}$ is continuous and strictly positive. Its minimum on the unit sphere is therefore positive, and scaling u shows that the moment tends to infinity with $\|u\|_2$. Since $u^\top a = 1$ defines a closed nonempty set, the infimum is attained. Moreover, since $\alpha/(\alpha-1) > 1$, the map $u \mapsto \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}$ is differentiable. At a minimizer, for every $w \in \mathbb{R}^d$ satisfying $w^\top a = 0$, we have

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \mathbb{E}_{x\sim\lambda} |(u+tw)^\top x|^{\alpha/(\alpha-1)} \\ &= \frac{\alpha}{\alpha-1} w^\top \sum_{x\in\mathcal{A}} \lambda(x) \operatorname{sign}(u^\top x) |u^\top x|^{1/(\alpha-1)} x. \end{aligned}$$

Hence, for every $w \in a^\perp$,

$$w^\top \sum_{x\in\mathcal{A}} \lambda(x) \operatorname{sign}(u^\top x) |u^\top x|^{1/(\alpha-1)} x = 0,$$

and therefore

$$\sum_{x\in\mathcal{A}} \lambda(x) \operatorname{sign}(u^\top x) |u^\top x|^{1/(\alpha-1)} x \in \operatorname{Span}(\{a\}).$$

Taking the inner product with u and using $u^\top a = 1$ gives

$$\sum_{x\in\mathcal{A}} \lambda(x) \operatorname{sign}(u^\top x) |u^\top x|^{1/(\alpha-1)} x = \left(\mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)} \right) a. \tag{9}$$

In accordance with this identity, we consider the choice

$$v_x = \frac{\lambda(x) \operatorname{sign}(u^\top x) |u^\top x|^{1/(\alpha-1)}}{\mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)}}, \qquad x \in \mathcal{A} \text{ with } \lambda(x) > 0. \tag{10}$$

Substituting Eq. (9) into Eq. (10) gives $\sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x x = a$. Moreover,

$$\begin{aligned} \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} &= \frac{\sum_{x\in\mathcal{A}:\,\lambda(x)>0} \lambda(x) |u^\top x|^{\alpha/(\alpha-1)}}{\left(\mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^\alpha} && \text{(by Eq. (10))} \\ &= \left(\mathbb{E}_{x\sim\lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{-(\alpha-1)}. \end{aligned}$$

This proves the "$\leq$" direction of Eq. (8). □

Proof of Lemma 1. For $a = 0$, take $v_x = 0$ for every $x \in \mathcal{A}$. If a nonzero a does not lie in $\mathrm{Span}(\{x \in \mathcal{A} : \lambda(x) > 0\})$, there is a vector u such that $u^\top a = 1$ and $u^\top x = 0$ whenever $\lambda(x) > 0$. The infimum in Eq. (4) is then zero, while the finite minimization has no feasible point, so both expressions equal $+\infty$. Otherwise, Lemma 2 gives

$$\left(\inf_{u:\, u^\top a = 1} \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{-(\alpha-1)} = \min_{\sum_{x:\,\lambda(x)>0} v_x x = a} \sum_{x:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}}.$$

When $M_\alpha(\lambda;\mathcal{A}) < \infty$, Eq. (8) shows that the minimum is finite and no larger than $M_\alpha(\lambda;\mathcal{A})$ for every $a \in \mathcal{A}$. When $\lambda(x) = 0$, define

$$\frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} := \begin{cases} 0, & v_x = 0, \\ +\infty, & v_x \neq 0. \end{cases}$$

For $t \in [0,1]$, $\lambda(x), \lambda'(x) > 0$, and real numbers v_x, v'_x, convexity of $s \mapsto |s|^\alpha$ gives

$$\begin{aligned}
&\frac{|tv_x + (1-t)v'_x|^\alpha}{(t\lambda(x) + (1-t)\lambda'(x))^{\alpha-1}} \\
&\quad = \big(t\lambda(x) + (1-t)\lambda'(x)\big)\left|\frac{t\lambda(x)}{t\lambda(x) + (1-t)\lambda'(x)}\frac{v_x}{\lambda(x)} + \frac{(1-t)\lambda'(x)}{t\lambda(x) + (1-t)\lambda'(x)}\frac{v'_x}{\lambda'(x)}\right|^\alpha \\
&\quad \le t\frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} + (1-t)\frac{|v'_x|^\alpha}{\lambda'(x)^{\alpha-1}}.
\end{aligned}$$

The same inequality holds under the stated convention when either design weight is zero. Convexity is automatic if either minimum is infinite. Otherwise, let (v_x) and (v'_x) be minimizers at λ and λ'. Then $(tv_x + (1-t)v'_x)$ is feasible at $t\lambda + (1-t)\lambda'$. Summing the inequality over $x \in \mathcal{A}$ proves that the minimum in Eq. (8) is convex in λ. Taking the maximum over $a \in \mathcal{A}$ proves that $M_\alpha(\lambda;\mathcal{A})$ is convex in λ. Moreover, the optimal design value is given by the finite convex program

$$M_\alpha(\mathcal{A}) = \min_{\substack{\lambda(x)\ge 0,\ \sum_{x\in\mathcal{A}}\lambda(x)=1 \\ \sum_{x\in\mathcal{A}} v_x(a)x = a \text{ for every } a\in\mathcal{A}}} \max_{a\in\mathcal{A}} \sum_{x\in\mathcal{A}} \frac{|v_x(a)|^\alpha}{\lambda(x)^{\alpha-1}}. \tag{11}$$

The uniform design, together with $v_a(a) = 1$ and $v_x(a) = 0$ for $x \neq a$, is feasible and has finite objective value. On any finite sublevel set, the design weights belong to the probability simplex and $|v_x(a)|^\alpha \le |v_x(a)|^\alpha/\lambda(x)^{\alpha-1}$ whenever $\lambda(x) > 0$. The convention at $\lambda(x) = 0$ makes these sublevel sets closed. They are therefore compact, so the minimum in Eq. (11) is attained.

□

A.2 The estimation upper bound

Proof of Theorem 1. If $M_\alpha(\lambda;\mathcal{A}) = \infty$, the right-hand side of the asserted inequality is infinite. We therefore suppose that $M_\alpha(\lambda;\mathcal{A}) < \infty$. For $a = 0$, set $\widehat{f}_a = 0$. For every nonzero $a \in \mathcal{A}$, Lemma 2 gives real numbers $(v_x)_{x:\,\lambda(x)>0}$ satisfying

$$\sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x x = a, \qquad \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} \le M_\alpha(\lambda;\mathcal{A}). \tag{12}$$

Partition the observations into $\lceil 8\log(2|\mathcal{A}|/\delta)\rceil$ blocks. Within each block, average $v_{x_i}y_i/\lambda(x_i)$, and let $\widehat{f}_a$ be the median of these block averages. Define

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \max_{a\in\mathcal{A}} |a^\top\theta - \widehat{f}_a|.$$

Since x_i lies in the support of λ, the denominator is positive. Each transformed response has expectation $a^\top\theta^*$ because

$$\mathbb{E}\left[\frac{v_{x_i}}{\lambda(x_i)}y_i\right] = \sum_{x\in\mathcal{A}:\,\lambda(x)>0} v_x x^\top\theta^* \qquad (\mathbb{E}[y_i \mid x_i] = x_i^\top\theta^*)$$

$$= a^\top \theta^*.$$

For $s, t \in \mathbb{R}$, convexity gives

$$|s+t|^\alpha \le 2^{\alpha-1}\big(|s|^\alpha + |t|^\alpha\big). \tag{13}$$

Together with $y_i = x_i^\top \theta^* + \varepsilon_i$ and $\max_{x\in\mathcal{A}} |x^\top \theta^*| \le B$, this gives

$$\begin{aligned}
\mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i\right|^\alpha &\le 2^{\alpha-1}\mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)}\varepsilon_i\right|^\alpha + 2^{\alpha-1}\mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} x_i^\top \theta^*\right|^\alpha && \text{(by Eq. (13))} \\
&\le 2^{\alpha-1}(\nu + B^\alpha) \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} \\
&\le 2(\nu + B^\alpha) \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} && (2^{\alpha-1} \le 2). \qquad (14)
\end{aligned}$$

Jensen's inequality and $\mathbb{E}[v_{x_i} y_i / \lambda(x_i)] = a^\top \theta^*$ give

$$|a^\top \theta^*|^\alpha = \left|\mathbb{E}\left[\frac{v_{x_i}}{\lambda(x_i)} y_i\right]\right|^\alpha \le \mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i\right|^\alpha. \tag{15}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i - a^\top \theta^*\right|^\alpha &\le 2^{\alpha-1}\left(\mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i\right|^\alpha + |a^\top \theta^*|^\alpha\right) && \text{(by Eq. (13))} \\
&\le 2^\alpha \mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i\right|^\alpha && \text{(by Eq. (15))} \\
&\le 8(\nu + B^\alpha) \sum_{x\in\mathcal{A}:\,\lambda(x)>0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} && \text{(by Eq. (14) and } 1 < \alpha \le 2) \qquad (16) \\
&\le 8(\nu + B^\alpha) M_\alpha(\lambda; \mathcal{A}) && \text{(by Eq. (12))}. \qquad (17)
\end{aligned}$$

Consider a block containing m observations. The random variables $v_{x_i} y_i / \lambda(x_i) - a^\top \theta^*$ are independent, have mean zero, and satisfy Eq. (17). The von Bahr–Esseen inequality (von Bahr & Esseen, 1965) gives

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{m}\sum_{i=1}^m \left(\frac{v_{x_i}}{\lambda(x_i)} y_i - a^\top \theta^*\right)\right|^\alpha \\
&\quad\le \frac{2}{m^\alpha}\sum_{i=1}^m \mathbb{E}\left|\frac{v_{x_i}}{\lambda(x_i)} y_i - a^\top \theta^*\right|^\alpha && \text{(von Bahr–Esseen inequality)} \\
&\quad\le \frac{16(\nu + B^\alpha) M_\alpha(\lambda; \mathcal{A})}{m^{\alpha-1}}. && \text{(by Eq. (17))}
\end{aligned}$$

Markov's inequality therefore yields

$$\begin{aligned}
&\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m \frac{v_{x_i}}{\lambda(x_i)} y_i - a^\top \theta^*\right| > 64^{1/\alpha}(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} m^{-(\alpha-1)/\alpha}\right) \\
&\quad\le \frac{1}{4}. && \text{(Markov's inequality)}
\end{aligned}$$

The block sizes differ by at most one. When $n \ge 20\log(2|\mathcal{A}|/\delta)$, each contains at least $n/(20\log(2|\mathcal{A}|/\delta))$ observations. Hence, each block estimate has error at most

$$\begin{aligned}
&64^{1/\alpha}(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{20\log(2|\mathcal{A}|/\delta)}{n}\right)^{(\alpha-1)/\alpha} \\
&\quad\le 64(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{\log(2|\mathcal{A}|/\delta)}{n}\right)^{(\alpha-1)/\alpha} && (20^{\alpha-1} \le 64^{\alpha-1}) \qquad (18)
\end{aligned}$$

with probability at least $3/4$. Hoeffding's inequality for the block indicators then gives

$$\Pr\left(\text{at least half of the block estimates for } a \text{ violate Eq. (18)}\right) \le \exp\left(-\frac{1}{8}\left\lceil 8\log\frac{2|\mathcal{A}|}{\delta}\right\rceil\right) \le \frac{\delta}{2|\mathcal{A}|}.$$

A median can violate Eq. (18) only if at least half of its block estimates do. A union bound therefore gives this error bound for every $\widehat{f}_a$ simultaneously with probability at least $1-\delta$. On this event, the definition of $\widehat{\theta}$ gives

$$\max_{a\in\mathcal{A}} |a^\top\widehat{\theta} - \widehat{f}_a| \le \max_{a\in\mathcal{A}} |a^\top\theta^* - \widehat{f}_a|.$$

Therefore, for every $a \in \mathcal{A}$,

$$\begin{aligned} |a^\top(\widehat{\theta}-\theta^*)| &\le |a^\top\widehat{\theta} - \widehat{f}_a| + |\widehat{f}_a - a^\top\theta^*| \\ &\le 2\max_{a'\in\mathcal{A}} |\widehat{f}_{a'} - a'^\top\theta^*|. \end{aligned} \tag{19}$$

Combining Eq. (19) with Eq. (18) proves Theorem 1 with $C = 128$. □

A.3 The estimation lower bound

Lemma 3 (Two-point response construction). *Fix $1 < \alpha \le 2$, $\nu > 0$, $u \in \mathbb{R}^d$, and $\Delta > 0$ such that $|\Delta u^\top x|^\alpha \le \nu/4$ for every $x \in \mathcal{A}$. For the parameters $\theta^* = \Delta u$ and $\theta' = -\Delta u$, there are conditional response distributions satisfying Eq. (1) with moment bound ν and*

$$\mathrm{TV}\left(\mathcal{L}_{\theta^*}(y(x)\mid x), \mathcal{L}_{\theta'}(y(x)\mid x)\right) = \left(\frac{4\Delta^\alpha|u^\top x|^\alpha}{\nu}\right)^{1/(\alpha-1)}. \tag{20}$$

Proof. When $u^\top x = 0$, set $y(x) = 0$. Otherwise, under θ^*, let

$$y(x) = \begin{cases} \left(\frac{\nu}{4\Delta}\right)^{1/(\alpha-1)} \frac{u^\top x}{|u^\top x|^{\alpha/(\alpha-1)}}, & \text{with probability } \left(\frac{4\Delta^\alpha|u^\top x|^\alpha}{\nu}\right)^{1/(\alpha-1)}, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

The assumption $|\Delta u^\top x|^\alpha \le \nu/4$ gives

$$\left(\frac{4\Delta^\alpha|u^\top x|^\alpha}{\nu}\right)^{1/(\alpha-1)} \le 1.$$

Moreover, the conditional mean is

$$\mathbb{E}[y(x)\mid x] = \left(\frac{4\Delta^\alpha|u^\top x|^\alpha}{\nu}\right)^{1/(\alpha-1)} \left(\frac{\nu}{4\Delta}\right)^{1/(\alpha-1)} \frac{u^\top x}{|u^\top x|^{\alpha/(\alpha-1)}} \tag{22}$$

$$= \Delta u^\top x = x^\top\theta^*, \tag{23}$$

and the uncentered conditional moment satisfies

$$\begin{aligned} \mathbb{E}[|y(x)|^\alpha \mid x] &= \left(\frac{4\Delta^\alpha|u^\top x|^\alpha}{\nu}\right)^{1/(\alpha-1)} \left(\frac{\nu}{4\Delta}\right)^{\alpha/(\alpha-1)} |u^\top x|^{-\alpha/(\alpha-1)} \\ &= \frac{\nu}{4}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[|y(x) - x^\top\theta^*|^\alpha \mid x] &\le 2^{\alpha-1}\left(\mathbb{E}[|y(x)|^\alpha \mid x] + |x^\top\theta^*|^\alpha\right) && \text{(by Eq. (13))} \\ &\le 2^\alpha \mathbb{E}[|y(x)|^\alpha \mid x] && (|\mathbb{E}[y(x)\mid x]|^\alpha \le \mathbb{E}[|y(x)|^\alpha \mid x]) \\ &\le \nu && (1 < \alpha \le 2). \end{aligned} \tag{24}$$

Under θ', use the negative of the response in Eq. (21). Then $\mathbb{E}[y(x)\mid x] = -\Delta u^\top x = x^\top\theta'$ and the distribution of $|y(x) - x^\top\theta'|$ is the same as that of $|y(x) - x^\top\theta^*|$ under θ^*. Hence, Eq. (24) holds under both parameters. The two conditional distributions place the same probability on nonzero outcomes of opposite signs, which proves Eq. (20). □

Proof of Theorem 2. If $M_\alpha(\lambda;\mathcal{A}) = \infty$, there is a direction u that vanishes on the support of λ and satisfies $\max_{a\in\mathcal{A}} |u^\top a| = 1$. Then, the two parameters $\theta^* = \nu^{1/\alpha}u/8$ and $\theta' = -\nu^{1/\alpha}u/8$, together with $y(x) = 0$ on the support of λ, produce identical observations. Outside the support, set the response equal to its conditional mean under each parameter. An action attaining $|u^\top a| = 1$ has values separated by $\nu^{1/\alpha}/4$, so Le Cam's inequality proves Theorem 2 in this case. It remains to consider the case of finite $M_\alpha(\lambda;\mathcal{A})$.

Rescaling u in Eq. (4) gives

$$M_\alpha(\lambda;\mathcal{A}) = \sup_{u\in\mathbb{R}^d:\, u\neq 0} \frac{\sup_{a\in\mathcal{A}} |u^\top a|^\alpha}{\left(\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{\alpha-1}}.$$

Finiteness of $M_\alpha(\lambda;\mathcal{A})$ implies that the support of λ spans $\mathrm{Span}(\mathcal{A})$; otherwise a direction orthogonal to the support but not to $\mathcal{A}$ would make the ratio infinite. We may therefore restrict u to $\mathrm{Span}(\mathcal{A})$. On this space, the set $\{u : \max_{a\in\mathcal{A}} |u^\top a| = 1\}$ is compact because $\mathcal{A}$ spans this space. The continuous function $u \mapsto \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}$ therefore attains its minimum on this set. A minimizer satisfies

$$\max_{a\in\mathcal{A}} |u^\top a| = 1, \qquad \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} = M_\alpha(\lambda;\mathcal{A})^{-1/(\alpha-1)}.$$

After replacing u by $-u$ if necessary, there is an action $a \in \mathcal{A}$ with $u^\top a = 1$. Let

$$\Delta = \frac{1}{8}\nu^{1/\alpha} \min\left\{ M_\alpha(\lambda;\mathcal{A})^{1/\alpha} n^{-(\alpha-1)/\alpha}, 1 \right\}. \tag{25}$$

For every $x \in \mathcal{A}$, the normalization gives $|u^\top x| \le 1$. Hence,

$$|\Delta u^\top x|^\alpha \le 8^{-\alpha}\nu \le \frac{\nu}{4}.$$

Lemma 3 gives response distributions satisfying the required conditional means and moment bound. By Eq. (25),

$$\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} = 2^{-3-1/(\alpha-1)} \min\left\{ \frac{M_\alpha(\lambda;\mathcal{A})^{1/(\alpha-1)}}{n}, 1 \right\}. \tag{26}$$

In particular,

$$\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} \le 2^{-3-1/(\alpha-1)} \frac{M_\alpha(\lambda;\mathcal{A})^{1/(\alpha-1)}}{n}. \tag{27}$$

Averaging Eq. (20) over $x \sim \lambda$ and using Eq. (27) bounds the one-sample total variation distance by

$$\begin{aligned}
&\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} \\
&\le 2^{-3-1/(\alpha-1)} M_\alpha(\lambda;\mathcal{A})^{-1/(\alpha-1)} \frac{M_\alpha(\lambda;\mathcal{A})^{1/(\alpha-1)}}{n} \\
&= 2^{-3-1/(\alpha-1)} \frac{1}{n} \\
&\le \frac{1}{16n} \qquad (1 < \alpha \le 2).
\end{aligned} \tag{28}$$

The total variation distance between the n-sample laws is at most n times the one-sample distance and hence at most $1/16$. Since $a^\top\theta^* = \Delta$ and $a^\top\theta' = -\Delta$, Le Cam's two-point inequality (Tsybakov & Zaiats, 2009, Chapter 2) implies error at least Δ with probability at least $1/4$ under at least one of θ^* and θ'. This establishes the desired lower bound with a constant of $c = 1/8$. □

A.4 Dimension range of the optimal design value

Proof of Theorem 3. Fix a design λ. If $A(\lambda)$ is singular, choose $u \neq 0$ in its kernel. Then

$$\mathbb{E}_{x\sim\lambda}|u^\top x|^2 = u^\top A(\lambda) u = 0.$$

Hence, every action in the support of λ is orthogonal to u. Since $\mathcal{A}$ spans $\mathbb{R}^d$, $\max_{a\in\mathcal{A}} |u^\top a| > 0$. Rescaling u in the definitions of M_α and M_2 shows that both values are infinite, so the lower bound holds.

Suppose now that $A(\lambda)$ is nonsingular. Rescaling the directions in the definitions gives

$$M_\alpha(\lambda;\mathcal{A}) = \sup_{u\neq 0} \frac{\max_{a\in\mathcal{A}} |u^\top a|^\alpha}{\left(\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{\alpha-1}}, \tag{29}$$

$$M_2(\lambda;\mathcal{A}) = \sup_{u\neq 0} \frac{\max_{a\in\mathcal{A}} |u^\top a|^2}{\mathbb{E}_{x\sim\lambda}|u^\top x|^2}. \tag{30}$$

For each u, the numerator in both expressions is determined by $\max_{a\in\mathcal{A}} |u^\top a|$. Since every sampled action belongs to $\mathcal{A}$ and $\alpha/(\alpha-1) \geq 2$,

$$\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} \leq \max_{a\in\mathcal{A}} |u^\top a|^{(2-\alpha)/(\alpha-1)} \mathbb{E}_{x\sim\lambda}|u^\top x|^2. \tag{31}$$

Substituting Eq. (31) into Eq. (29) and taking the supremum over u gives

$$M_\alpha(\lambda;\mathcal{A}) \geq M_2(\lambda;\mathcal{A})^{\alpha-1} \geq d^{\alpha-1}. \qquad \text{(Kiefer–Wolfowitz lower bound)}$$

Taking the infimum over λ and then the $1/\alpha$ power gives $M_\alpha(\mathcal{A})^{1/\alpha} \geq d^{(\alpha-1)/\alpha}$.

Let λ be a G-optimal design, and observe that for every $a \in \mathcal{A}$, we have

$$\begin{aligned}
&\left(\inf_{u:\ u^\top a=1} \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{-(\alpha-1)} \\
&\quad\leq \left(\inf_{u:\ u^\top a=1} \mathbb{E}_{x\sim\lambda}|u^\top x|^2\right)^{-\alpha/2} && \text{(Lyapunov's inequality)} \\
&\quad= \left(a^\top A(\lambda)^{-1} a\right)^{\alpha/2}. && \left(\inf_{u:\ u^\top a=1} u^\top A(\lambda) u = (a^\top A(\lambda)^{-1} a)^{-1}\right)
\end{aligned}$$

Taking the maximum over a and using $M_2(\lambda;\mathcal{A}) = d$ gives $M_\alpha(\lambda;\mathcal{A}) \leq d^{\alpha/2}$. Since this is a feasible design, $M_\alpha(\mathcal{A})^{1/\alpha} \leq \sqrt{d}$.

The equality $M_\alpha(\mathcal{A})^{1/\alpha} = d^{(\alpha-1)/\alpha}$ holds for $\mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$. For $a = \mathbf{e}_j$, the inner infimum in Eq. (4) equals $\lambda(\mathbf{e}_j)$. Hence,

$$M_\alpha(\lambda;\mathcal{A}) = \max_{1\leq j\leq d} \lambda(\mathbf{e}_j)^{-(\alpha-1)},$$

whose minimum is $d^{\alpha-1}$ at the uniform design.

For the upper endpoint in Eq. (6), take any finite Euclidean $1/2$-net of the unit sphere as the action set. Fix an arbitrary design on this set, let u be uniform on the unit sphere, and let g be a standard Gaussian vector. The direction $g/\|g\|_2$ is uniform on the sphere and independent of $\|g\|_2$ (Vershynin, 2019, Exercise 3.3.7). For every unit action x, this gives

$$\begin{aligned}
&\left(\mathbb{E}_u|u^\top x|^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} \\
&\quad= \frac{\left(\mathbb{E}|g_1|^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}}{\left(\mathbb{E}\|g\|_2^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}} && \text{(independence of radius and direction)} \\
&\quad\leq \frac{\sqrt{\alpha/(\alpha-1)}}{\left(\mathbb{E}\|g\|_2^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}} && \text{(Gaussian moment bound)} \\
&\quad\leq \frac{\sqrt{\alpha/(\alpha-1)}}{\sqrt{d}}.
\end{aligned} \tag{32}$$

Here $\left(\mathbb{E}\|g\|_2^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} \geq (\mathbb{E}\|g\|_2^2)^{1/2} = \sqrt{d}$. Averaging first over $x \sim \lambda$ and then over the uniform direction gives

$$\begin{aligned}
&\inf_{u\in\mathbb{R}^d:\ \|u\|_2=1} \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} \\
&\quad\leq \mathbb{E}_u \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} && \text{(the minimum is at most the average)}
\end{aligned}$$

$$= \mathbb{E}_{x\sim\lambda}\mathbb{E}_u|u^\top x|^{\alpha/(\alpha-1)} \qquad \text{(Tonelli's theorem)}$$
$$\le \mathbb{E}_{x\sim\lambda}\left(\frac{\alpha}{(\alpha-1)d}\right)^{\alpha/(2(\alpha-1))} \qquad \text{(by Eq. (32))}$$
$$= \left(\frac{\alpha}{(\alpha-1)d}\right)^{\alpha/(2(\alpha-1))}.$$

Thus, some unit vector satisfies

$$\left(\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} \le \sqrt{\frac{\alpha}{(\alpha-1)d}}. \tag{33}$$

Since $\mathcal{A}$ is a Euclidean $1/2$-net, $\max_{a\in\mathcal{A}}|u^\top a| \ge 1/2$. Combining this inequality with Eq. (29) and Eq. (33) yields

$$M_\alpha(\lambda;\mathcal{A})^{1/\alpha} \ge \frac{1}{2}\sqrt{\frac{\alpha-1}{\alpha}}\sqrt{d}.$$

Since the design was arbitrary, the spherical-net lower bound in Theorem 3 follows with $c = 1/2$. □

A.5 A BOUND INDEPENDENT OF THE ACTION SET CARDINALITY

The proof of the following corollary estimates $a_j^\top\theta^*$ at no more than 5^d points a_j in the symmetric convex hull of $\mathcal{A}$. Their estimation errors control Eq. (2) over all of $\mathcal{A}$.

Corollary 3 (Upper bound independent of $|\mathcal{A}|$)**.** *Under the assumptions of Theorem 1, there is an estimator $\widehat{\theta}$ and a universal constant C such that, whenever*

$$n \ge C\left(d + \log(2/\delta)\right),$$

with probability at least $1-\delta$,

$$\max_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)| \le C(\nu + B^\alpha)^{1/\alpha}M_\alpha(\lambda;\mathcal{A})^{1/\alpha}\left(\frac{d+\log(2/\delta)}{n}\right)^{(\alpha-1)/\alpha}.$$

Proof of Corollary 3. When $M_\alpha(\lambda;\mathcal{A}) = \infty$, the asserted upper bound has an infinite right-hand side, so we may assume that $M_\alpha(\lambda;\mathcal{A}) < \infty$. Then the support of λ spans $\mathrm{Span}(\mathcal{A})$. The volumetric covering bound for a d-dimensional normed space (Vershynin, 2019, Section 4.2) gives $a_1,\dots,a_{5^d} \in \mathrm{conv}(\mathcal{A}\cup-\mathcal{A})$ such that

$$\mathrm{conv}(\mathcal{A}\cup-\mathcal{A}) \subseteq \bigcup_{j=1}^{5^d}\left(a_j + \frac{1}{2}\mathrm{conv}(\mathcal{A}\cup-\mathcal{A})\right). \tag{34}$$

For every $a\in\mathcal{A}$ and $w\in\mathbb{R}^d$, Eq. (34) gives some j for which

$$|a^\top w| \le |a_j^\top w| + \frac{1}{2}\max_{x\in\mathcal{A}}|x^\top w|.$$

Taking the maximum over a and rearranging gives

$$\max_{a\in\mathcal{A}}|a^\top w| \le 2\max_{1\le j\le 5^d}|a_j^\top w|. \tag{35}$$

Since each a_j lies in the symmetric convex hull of $\mathcal{A}$, $|u^\top a_j| \le \max_{a\in\mathcal{A}}|u^\top a|$. Consequently, Lemma 2 gives, for each j, real numbers $(v_x(j))_{x:\,\lambda(x)>0}$ satisfying

$$\sum_{x:\,\lambda(x)>0} v_x(j)x = a_j, \qquad \sum_{x:\,\lambda(x)>0}\frac{|v_x(j)|^\alpha}{\lambda(x)^{\alpha-1}} \le M_\alpha(\lambda;\mathcal{A}). \tag{36}$$

Let $\widehat{f}_{a_j}$ be the median-of-means estimate formed from $v_{x_i}(j)y_i/\lambda(x_i)$, and choose

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\max_{1\le j\le 5^d}|a_j^\top\theta - \widehat{f}_{a_j}|.$$

By Eq. (16) and Eq. (36), the centered α moment of each transformed response is at most $8(\nu + B^\alpha)M_\alpha(\lambda; \mathcal{A})$. Using $\lceil 8 \log(2 \cdot 5^d/\delta) \rceil$ blocks, the median-of-means bound Eq. (18) and a union bound over the 5^d estimates give, with probability at least $1 - \delta$,

$$\begin{aligned}
\max_{a \in \mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)| &\le 2 \max_{1 \le j \le 5^d} |a_j^\top(\widehat{\theta} - \theta^*)| && \text{(Eq. (35))} \\
&\le 4 \max_{1 \le j \le 5^d} |\widehat{f}_{a_j} - a_j^\top \theta^*| && \text{(definition of } \widehat{\theta} \text{ and the triangle inequality)} \\
&\le 256(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{\log(2 \cdot 5^d/\delta)}{n} \right)^{(\alpha-1)/\alpha} \\
&\le 512(\nu + B^\alpha)^{1/\alpha} M_\alpha(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{d + \log(2/\delta)}{n} \right)^{(\alpha-1)/\alpha},
\end{aligned}$$

where we also applied $\log(2 \cdot 5^d/\delta) \le 2(d + \log(2/\delta))$. The constant $C = 1024$ therefore satisfies both the sample-size and error bounds. □

A.6 The separation from classical G-optimality

Proof of Example 1. We first identify the G-optimal design. Let λ be any design whose information matrix $A(\lambda)$ is nonsingular and satisfies $\max_{a \in \mathcal{A}} a^\top A(\lambda)^{-1} a \le d$. The uniform design on $\{h_1, \dots, h_d\}$ has information matrix $d^{-1} I$ and G value d, so such designs exist. Since each h_i belongs to $\mathcal{A}$, we have

$$\begin{aligned}
\operatorname{Tr}(A(\lambda)^{-1}) &= \sum_{i=1}^{d} h_i^\top A(\lambda)^{-1} h_i && (h_1, \dots, h_d \text{ is an orthonormal basis}) \\
&\le d^2. && (h_i^\top A(\lambda)^{-1} h_i \le d \text{ for every } i)
\end{aligned}$$

The trace satisfies $\operatorname{Tr}(A(\lambda)) \le 1$, because the h_i have norm one and the actions $\mathbf{e}_i/2$ have norm $1/2$. Moreover,

$$\begin{aligned}
\operatorname{Tr}(A(\lambda)^{-1}) &\ge \frac{d^2}{\operatorname{Tr}(A(\lambda))} && \text{(Cauchy–Schwarz)} \\
&\ge d^2. && (\operatorname{Tr}(A(\lambda)) \le 1)
\end{aligned}$$

Hence, all inequalities are equalities, so $\operatorname{Tr}(A(\lambda)) = 1$, $A(\lambda) = d^{-1} I$, and λ assigns no mass to the shorter actions $\mathbf{e}_i/2$. Since $h_1, \dots, h_d$ are an orthonormal basis, the only design supported on them with information matrix $d^{-1} I$ is the uniform design. The claimed G-optimal design follows.

For the statistical lower bound, fix a coordinate j. Apply Lemma 3 on the support of the design with $u = \mathbf{e}_j$ and $\Delta = \nu^{1/\alpha} \sqrt{d}\, n^{-(\alpha-1)/\alpha}/8$. On every sampled action, $|\Delta u^\top x| = \nu^{1/\alpha} n^{-(\alpha-1)/\alpha}/8$, so the lemma applies. At actions outside the support of the design, set the response equal to its conditional mean.

For the two parameters $\theta^* = \Delta \mathbf{e}_j$ and $\theta' = -\Delta \mathbf{e}_j$, we have

$$\begin{aligned}
&\operatorname{TV}\left(\mathcal{L}_{\theta^*}((x_i, y_i)_{i=1}^n), \mathcal{L}_{\theta'}((x_i, y_i)_{i=1}^n)\right) \\
&\le n \left(\frac{4\Delta^\alpha}{\nu} \right)^{1/(\alpha-1)} \mathbb{E}_{x \sim \lambda} |\mathbf{e}_j^\top x|^{\alpha/(\alpha-1)} && \text{(tensorization and Eq. (20))} \\
&= \left(4 \cdot 8^{-\alpha}\right)^{1/(\alpha-1)} && \text{(uniform Hadamard projections and choice of } \Delta) \\
&= 2^{-3-1/(\alpha-1)} \le \frac{1}{16}.
\end{aligned}$$

The coordinate action $\mathbf{e}_j/2$ has values $\Delta/2$ and $-\Delta/2$. Relabel the parameter selected by Le Cam's inequality as θ^*. Then

$$\Pr\left(|(\mathbf{e}_j/2)^\top(\widehat{\theta} - \theta^*)| \ge \frac{1}{16} \nu^{1/\alpha} \sqrt{d}\, n^{-(\alpha-1)/\alpha} \right) \ge \frac{1}{4}.$$

For the competing uniform design, a Chernoff bound gives

$$\Pr\left(\text{some action receives fewer than } n/(4d) \text{ observations}\right) \le 2d\exp(-n/(16d)) \le \frac{\delta}{2}$$

when $n \ge 512d\log(2d/\delta)$. Conditional on the sampled actions, responses at each action have centered α moment at most ν. Use $\lceil 8\log(4d/\delta)\rceil$ median-of-means blocks for each action. The scalar median-of-means bound (Lugosi & Mendelson, 2019b) and a union bound give

$$\begin{aligned}
\max_{a\in\mathcal{A}} |\widehat{f}_a - a^\top\theta^*| &\le 8^{1/\alpha}\nu^{1/\alpha}\left(\frac{80d\log(4d/\delta)}{n}\right)^{(\alpha-1)/\alpha} \\
&\le 8^{1/\alpha}160^{(\alpha-1)/\alpha}\nu^{1/\alpha}\left(\frac{d\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha} && (\log(4d/\delta)\le 2\log(2d/\delta)) \\
&\le 64\nu^{1/\alpha}\left(\frac{d\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha} && (8^{1/\alpha}160^{(\alpha-1)/\alpha} = 8\cdot 20^{(\alpha-1)/\alpha} < 64)
\end{aligned}$$

with conditional probability at least $1-\delta/2$. Choose

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \max_{a\in\mathcal{A}} |a^\top\theta - \widehat{f}_a|.$$

Then

$$\max_{a\in\mathcal{A}} |a^\top(\widehat{\theta}-\theta^*)| \le 2\max_{a\in\mathcal{A}} |\widehat{f}_a - a^\top\theta^*|.$$

Finally, the uniform design on all actions has information matrix $5I/(8d)$ and G value $8d/5 > d$. The lower and upper bounds have the dimension ratio asserted in Example 1. The choices $c = 1/16$ and $C = 512$ satisfy all statements. □

B Proofs for an Unknown Moment Exponent

B.1 A finite mixture of designs

Proof of Theorem 4. Let J be the smallest nonnegative integer satisfying

$$2\left(1+\frac{1}{\log d}\right)^J \ge 2\log(2|\mathcal{A}|). \tag{37}$$

For every $j \in \{0,\dots,J\}$, let $\alpha_j \in (1,2]$ be the value satisfying

$$\frac{\alpha_j}{\alpha_j - 1} = 2\left(1+\frac{1}{\log d}\right)^j. \tag{38}$$

In particular, $\alpha_0 = 2$. For each α_j, compute a design λ_j minimizing $M_{\alpha_j}(\lambda;\mathcal{A})$ over designs λ on $\mathcal{A}$. Let λ_{unif} be the uniform distribution on $\mathcal{A}$, and set

$$\lambda_{\text{fin}} = \sum_{j=0}^{J}\left(8\sqrt{\log d}\right)^{-\alpha_j/(\alpha_j-1)}\lambda_j + \left(1-\sum_{j=0}^{J}\left(8\sqrt{\log d}\right)^{-\alpha_j/(\alpha_j-1)}\right)\lambda_{\text{unif}}.$$

We first verify that the mixture weights sum to less than one. Bernoulli's inequality gives

$$\frac{\alpha_j}{\alpha_j-1} = 2\left(1+\frac{1}{\log d}\right)^j \ge 2\left(1+\frac{j}{\log d}\right).$$

Consequently,

$$\sum_{j=0}^{J}\left(8\sqrt{\log d}\right)^{-\alpha_j/(\alpha_j-1)} \le \frac{1}{64\log d}\sum_{j=0}^{\infty}\exp\left(-\frac{2j}{\log d}\log\left(8\sqrt{\log d}\right)\right)$$

(Bernoulli's inequality)

$$= \frac{1}{64 \log d} \left(1 - \exp\left(-\frac{2}{\log d} \log\left(8\sqrt{\log d}\right)\right)\right)^{-1}$$

(geometric series)

$$\leq \frac{1}{64 \log d} + \frac{1}{128 \log\left(8\sqrt{\log d}\right)} \quad ((1-e^{-t})^{-1} \leq 1 + t^{-1} \text{ for } t > 0)$$

$$< \frac{5}{256} < \frac{1}{2}. \quad (d \geq 3, \text{ so } \log d > 1 \text{ and } \log 8 > 2)$$

Thus the uniform component has mass at least $1/2$.

Fix the true α, and suppose that

$$\frac{\alpha}{\alpha - 1} \leq 2\left(1 + \frac{1}{\log d}\right)^{J}.$$

Choose $j \in \{0, \ldots, J\}$ with the smallest index such that

$$\frac{\alpha}{\alpha - 1} \leq \frac{\alpha_j}{\alpha_j - 1} \leq \left(1 + \frac{1}{\log d}\right) \frac{\alpha}{\alpha - 1}. \tag{39}$$

For every u satisfying $\sup_{a \in \mathcal{A}} |u^\top a| = 1$, the mixture definition and $|u^\top x| \leq 1$ give

$$\mathbb{E}_{x \sim \lambda_{\text{fin}}} |u^\top x|^{\alpha/(\alpha-1)} \geq \left(8\sqrt{\log d}\right)^{-\alpha_j/(\alpha_j - 1)} \mathbb{E}_{x \sim \lambda_j} |u^\top x|^{\alpha/(\alpha-1)}$$

$$(\lambda_{\text{fin}}(x) \geq (8\sqrt{\log d})^{-\alpha_j/(\alpha_j-1)} \lambda_j(x))$$

$$\geq \left(8\sqrt{\log d}\right)^{-\alpha_j/(\alpha_j - 1)} \mathbb{E}_{x \sim \lambda_j} |u^\top x|^{\alpha_j/(\alpha_j - 1)}. \quad \text{(Eq. (39) and } |u^\top x| \leq 1)$$

Raising both sides to the power $-(\alpha - 1)/\alpha$ gives

$$\left(\mathbb{E}_{x \sim \lambda_{\text{fin}}} |u^\top x|^{\alpha/(\alpha-1)}\right)^{-(\alpha-1)/\alpha}$$

$$\leq \left(8\sqrt{\log d}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}} \left[\left(\mathbb{E}_{x \sim \lambda_j} |u^\top x|^{\alpha_j/(\alpha_j-1)}\right)^{-(\alpha_j-1)/\alpha_j}\right]^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}}.$$

Taking the supremum over u such that $\sup_{a \in \mathcal{A}} |u^\top a| = 1$ yields

$$M_\alpha(\lambda_{\text{fin}}; \mathcal{A})^{1/\alpha} \leq \left(8\sqrt{\log d}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}} \left(M_{\alpha_j}(\lambda_j; \mathcal{A})^{1/\alpha_j}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}}. \tag{40}$$

Let λ_α minimize $M_\alpha(\lambda; \mathcal{A})$. Using the fact that λ_j minimizes $M_{\alpha_j}(\lambda; \mathcal{A})$, along with Lyapunov's inequality, we have

$$M_{\alpha_j}(\lambda_j; \mathcal{A})^{1/\alpha_j} \leq M_{\alpha_j}(\lambda_\alpha; \mathcal{A})^{1/\alpha_j}$$

$$\leq M_\alpha(\lambda_\alpha; \mathcal{A})^{1/\alpha}. \tag{41}$$

Combining Eq. (40) and Eq. (41) gives

$$M_\alpha(\lambda_{\text{fin}}; \mathcal{A})^{1/\alpha} \leq \left(8\sqrt{\log d}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}} M_\alpha(\lambda_\alpha; \mathcal{A})^{1/\alpha} \left(M_\alpha(\lambda_\alpha; \mathcal{A})^{1/\alpha}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)} - 1}. \tag{42}$$

We know from Theorem 3 that $M_\alpha(\lambda_\alpha; \mathcal{A})^{1/\alpha} \leq \sqrt{d}$. Since $(\alpha_j/(\alpha_j - 1))/(\alpha/(\alpha - 1)) - 1 \leq 1/\log d$, it follows that

$$\left(M_\alpha(\lambda_\alpha; \mathcal{A})^{1/\alpha}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)} - 1} \leq d^{\frac{1}{2}\left(\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)} - 1\right)}$$

$$\leq e^{1/2} < 2. \tag{43}$$

Since $\log d > 1$ and $\log \log d \leq \log d$, the factor $(8\sqrt{\log d})^{(\alpha_j/(\alpha_j-1))/(\alpha/(\alpha-1))}$ in Eq. (42) satisfies

$$\left(8\sqrt{\log d}\right)^{\frac{\alpha_j/(\alpha_j-1)}{\alpha/(\alpha-1)}} \leq \left(8\sqrt{\log d}\right)^{1 + 1/\log d}$$

$$\leq 8\sqrt{\log d}\,\exp\left(\log 8 + \frac{1}{2}\right)$$
$$< 128\sqrt{\log d}. \tag{44}$$

Combining Eq. (42), Eq. (43), and Eq. (44) gives

$$M_\alpha(\lambda_{\text{fin}};\mathcal{A})^{1/\alpha} \leq 256\sqrt{\log d}\, M_\alpha(\lambda_\alpha;\mathcal{A})^{1/\alpha}. \tag{45}$$

It remains to consider the case where $\alpha/(\alpha-1) > 2(1+1/\log d)^J$. For every u satisfying $\sup_{a\in\mathcal{A}}|u^\top a| = 1$, there is an $a \in \mathcal{A}$ with $|u^\top a| = 1$. Since $\lambda_{\text{fin}}(a) \geq 1/(2|\mathcal{A}|)$, we have

$$\mathbb{E}_{x\sim\lambda_{\text{fin}}}|u^\top x|^{\alpha/(\alpha-1)} \geq \frac{1}{2|\mathcal{A}|}.$$

The definition of M_α and Eq. (37) therefore give

$$M_\alpha(\lambda_{\text{fin}};\mathcal{A})^{1/\alpha} \leq (2|\mathcal{A}|)^{(\alpha-1)/\alpha}$$
$$< e^{1/2} < 2. \qquad (\alpha/(\alpha-1) > 2\log(2|\mathcal{A}|))$$

Moreover, for every design λ', one has $M_\alpha(\lambda';\mathcal{A})^{1/\alpha} \geq 1$. Indeed, every normalized projection has absolute value at most one on $\mathcal{A}$, so its moment under λ' is at most one. Since $\sqrt{\log d} > 1$, the bound $M_\alpha(\lambda_{\text{fin}};\mathcal{A})^{1/\alpha} < 2$ is at most the right-hand side of Theorem 4 with $C = 256$. Hence, $C = 256$ is valid in both cases. □

The construction uses $J+1$ convex programs, where J is specified in Eq. (37).

B.2 The necessary approximation factor

Proof of Theorem 5. Fix an integer $J \geq 21$ and set

$$d = \sum_{j=1}^{J} 4^{J\lceil\log J\rceil - j}.$$

In the following, we use a standard definition: In D dimensions, two orthonormal bases are said to be *mutually unbiased* if every cross-basis inner product has absolute value at most $D^{-1/2}$. We will use this definition shortly with $D = 4^{J\lceil\log J\rceil - j}$.

Decompose $\mathbb{R}^d$ into an orthogonal direct sum of J subspaces, with the jth subspace having dimension $4^{J\lceil\log J\rceil - j}$. The j-th subspace admits $\frac{1}{2}4^{J\lceil\log J\rceil - j} + 1$ pairwise mutually unbiased real orthonormal bases (LeCompte et al., 2010, Theorem 3.2(ii)), which is lower bounded by 4^{2j} when $J \geq 21$ (and hence $\lceil\log J\rceil \geq 4$):

$$3j \leq 3J \leq J\lceil\log J\rceil - \frac{1}{2},$$
$$4^{2j} \leq \frac{1}{2}4^{J\lceil\log J\rceil - j} + 1.$$

Accordingly, in the j-th subspace, we take 4^{2j} pairwise mutually unbiased bases and place all their vectors in $\mathcal{A}$. Hence, vectors in the same basis are orthogonal, while vectors from distinct bases satisfy the stated inner product bounds.

The action set spans every subspace and hence spans $\mathbb{R}^d$. The jth subspace contains

$$4^{2j}4^{J\lceil\log J\rceil - j} = 4^{J\lceil\log J\rceil + j}$$

actions. For each $j \in \{1,\ldots,J\}$, let $\alpha_j \in (1,2)$ satisfy

$$\frac{\alpha_j}{\alpha_j - 1} = 2\frac{J\lceil\log J\rceil + j}{J\lceil\log J\rceil - j}. \tag{46}$$

Eq. (46) gives

$$\left(4^{J\lceil\log J\rceil - j}\right)^{\frac{\alpha_j}{2(\alpha_j-1)}} = 4^{J\lceil\log J\rceil + j}. \tag{47}$$

Fix any design λ on $\mathcal{A}$. One of the J subspaces has total design mass at most $1/J$; fix its index j. Since this subspace contains $4^{J\lceil \log J\rceil + j}$ actions, it contains an action a satisfying

$$\lambda(a) \le \frac{1}{J 4^{J\lceil \log J\rceil + j}}. \tag{48}$$

The action a is orthogonal to the other vectors in its basis and to every other subspace. Using the mass bound for its subspace and Eq. (47) gives

$$\begin{aligned}
\mathbb{E}_{x\sim\lambda}|a^\top x|^{\alpha_j/(\alpha_j-1)} &\le \lambda(a) + \frac{1}{J}\left(4^{J\lceil \log J\rceil - j}\right)^{-\frac{\alpha_j}{2(\alpha_j-1)}} \\
&\qquad \text{(orthogonality and mutual unbiasedness)} \\
&\le \frac{2}{J 4^{J\lceil \log J\rceil + j}}. \qquad \text{(Eq. (47) and Eq. (48))}
\end{aligned}$$

Since all actions have norm one and $a \in \mathcal{A}$, $\max_{x\in\mathcal{A}} |a^\top x| = 1$. Rescaling the direction in Eq. (4) and substituting $u = a$ therefore yields

$$M_{\alpha_j}(\lambda;\mathcal{A})^{1/\alpha_j} \ge \left(\frac{J 4^{J\lceil \log J\rceil + j}}{2}\right)^{(\alpha_j-1)/\alpha_j}. \tag{49}$$

To bound the known-α_j optimum, assign to the kth subspace total mass proportional to

$$\min\left\{\left(4^{J\lceil \log J\rceil - k}\right)^{\frac{\alpha_j}{2(\alpha_j-1)}}, 4^{J\lceil \log J\rceil + k}\right\},$$

and distribute this mass uniformly among the actions in that subspace. Fix $u \in \mathbb{R}^d$ with $\max_{a\in\mathcal{A}} |u^\top a| = 1$. Choose $a \in \mathcal{A}$ such that $|u^\top a| = 1$, and let k be the index of the subspace containing a. Under the uniform distribution on the $4^{J\lceil \log J\rceil + k}$ actions in this subspace, the projection moments satisfy

$$\begin{aligned}
\frac{1}{4^{J\lceil \log J\rceil + k}} \sum_{\substack{x\in\mathcal{A}: \\ x \text{ in subspace } k}} |u^\top x|^{\alpha_j/(\alpha_j-1)} &\ge \frac{|u^\top a|^{\alpha_j/(\alpha_j-1)}}{4^{J\lceil \log J\rceil + k}} \\
&= 4^{-J\lceil \log J\rceil - k}, \qquad (50) \\
\frac{1}{4^{J\lceil \log J\rceil + k}} \sum_{\substack{x\in\mathcal{A}: \\ x \text{ in subspace } k}} |u^\top x|^{2} &\ge 4^{-J\lceil \log J\rceil + k}. \qquad (51)
\end{aligned}$$

For Eq. (51), Parseval's identity shows that the average over any orthonormal basis in the kth subspace equals the squared norm of the component of u in that subspace divided by $4^{J\lceil \log J\rceil - k}$. The component has norm at least one because an action in the subspace has an inner product of absolute value one with u. Since $\alpha_j/(\alpha_j - 1) \ge 2$, Lyapunov's inequality and Eq. (51) imply

$$\begin{aligned}
&\frac{1}{4^{J\lceil \log J\rceil + k}} \sum_{\substack{x\in\mathcal{A}: \\ x \text{ in subspace } k}} |u^\top x|^{\alpha_j/(\alpha_j-1)} \\
&\ge \left(\frac{1}{4^{J\lceil \log J\rceil + k}} \sum_{\substack{x\in\mathcal{A}: \\ x \text{ in subspace } k}} |u^\top x|^{2}\right)^{\alpha_j/(2(\alpha_j-1))} \qquad \text{(Lyapunov's inequality)} \\
&\ge \left(4^{J\lceil \log J\rceil - k}\right)^{-\frac{\alpha_j}{2(\alpha_j-1)}} \qquad \text{(by Eq. (51)).} \qquad (52)
\end{aligned}$$

Combining Eq. (50) and Eq. (52) gives

$$\begin{aligned}
&\frac{1}{4^{J\lceil \log J\rceil + k}} \sum_{\substack{x\in\mathcal{A}: \\ x \text{ in subspace } k}} |u^\top x|^{\alpha_j/(\alpha_j-1)} \\
&\ge \max\left\{4^{-J\lceil \log J\rceil - k}, \left(4^{J\lceil \log J\rceil - k}\right)^{-\frac{\alpha_j}{2(\alpha_j-1)}}\right\}
\end{aligned}$$

$$= \left[\min \left\{ 4^{J\lceil \log J \rceil + k}, \left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}} \right\} \right]^{-1}.$$

Multiplying this average by the mass assigned to the kth subspace bounds the projection moment under the constructed design from below by

$$\left[\sum_{k=1}^{J} \min \left\{ \left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}}, 4^{J\lceil \log J \rceil + k} \right\} \right]^{-1}.$$

The constructed design therefore satisfies

$$M_{\alpha_j}(\mathcal{A})^{1/\alpha_j} \leq \left[\sum_{k=1}^{J} \min \left\{ \left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}}, 4^{J\lceil \log J \rceil + k} \right\} \right]^{(\alpha_j - 1)/\alpha_j}. \tag{53}$$

We now consider three cases:

- At $k = j$, Eq. (47) gives

$$\left(4^{J\lceil \log J \rceil - j} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}} = 4^{J\lceil \log J \rceil + j}.$$

- For $k < j$, we have

$$4^{J\lceil \log J \rceil + k} \leq \left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}},$$

  and hence

$$\sum_{k=1}^{j-1} 4^{J\lceil \log J \rceil + k} \leq \frac{1}{3} 4^{J\lceil \log J \rceil + j}.$$

- For $k > j$, we have

$$\left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}} \leq 4^{J\lceil \log J \rceil + k}.$$

  Moreover,

$$\begin{aligned} J\lceil \log J \rceil + j - \frac{(J\lceil \log J \rceil - k)(J\lceil \log J \rceil + j)}{J\lceil \log J \rceil - j} \\ = \frac{(J\lceil \log J \rceil + j)(k - j)}{J\lceil \log J \rceil - j} \geq k - j, \end{aligned}$$

  and consequently,

$$\sum_{k=j+1}^{J} \left(4^{J\lceil \log J \rceil - k} \right)^{\frac{\alpha_j}{2(\alpha_j - 1)}} \leq \frac{1}{3} 4^{J\lceil \log J \rceil + j}.$$

The sum in Eq. (53) is therefore at most $5\, 4^{J\lceil \log J \rceil + j}/3$, and hence

$$M_{\alpha_j}(\mathcal{A})^{1/\alpha_j} \leq \left(\frac{5}{3} 4^{J\lceil \log J \rceil + j} \right)^{(\alpha_j - 1)/\alpha_j}. \tag{54}$$

Dividing Eq. (49) by Eq. (54) gives

$$\frac{M_{\alpha_j}(\lambda; \mathcal{A})^{1/\alpha_j}}{M_{\alpha_j}(\mathcal{A})^{1/\alpha_j}} \geq \left(\frac{3J}{10} \right)^{(\alpha_j - 1)/\alpha_j}. \tag{55}$$

To simplify this expression, we proceed to suitably bound both J and the exponent $(\alpha_j - 1)/\alpha_j$. Starting with the latter, by Eq. (46), we have

$$\begin{aligned}\frac{\alpha_j - 1}{\alpha_j} &= \frac{J\lceil \log J\rceil - j}{2(J\lceil \log J\rceil + j)} \\ &\geq \frac{\lceil \log J\rceil - 1}{2(\lceil \log J\rceil + 1)} = \frac{1}{2} - \frac{1}{\lceil \log J\rceil + 1}.\end{aligned}$$

Then, since $J^{1/(\lceil \log J\rceil + 1)} \leq e$, it follows that $(3J/10)^{(\alpha_j - 1)/\alpha_j} \geq e^{-1}\sqrt{3J/10}$.

Summing the subspace dimensions gives

$$\frac{1}{4}4^{J\lceil \log J\rceil} \leq d \leq \frac{1}{3}4^{J\lceil \log J\rceil}. \tag{56}$$

For $J \geq 21$, Eq. (56) gives

$$\begin{aligned}\log d &\leq J\lceil \log J\rceil \log 4 && \text{(Eq. (56))} \\ &\leq 2J \log J \log 4 && (\lceil \log J\rceil \leq 2\log J) \\ &\leq 2J \log\log d \log 4. && (\log d \geq (J\lceil \log J\rceil - 1)\log 4 \geq J)\end{aligned}$$

Rearranging then gives

$$J \geq \frac{\log d}{2(\log 4)\log\log d}. \tag{57}$$

Combining Eq. (57) with $(3J/10)^{(\alpha_j - 1)/\alpha_j} \geq e^{-1}\sqrt{3J/10}$ and Eq. (55) proves Theorem 5 with $c = e^{-1}\sqrt{3/(20\log 4)}$. □

C Lower Bounds for General Nonadaptive Allocations

Lemma 4 (Total variation under a nonadaptive allocation)**.** *Let $x_1, \ldots, x_n$ be selected jointly before the responses are observed, independently of the unknown parameter, and let λ be defined by Eq. (7). Suppose that, conditional on $x_1, \ldots, x_n$, the responses are independent and, for $\theta \in \{\theta^*, \theta'\}$,*

$$\mathcal{L}_\theta(y_i \mid x_1, \ldots, x_n) = \mathcal{L}_\theta(y(x_i) \mid x_i).$$

Then λ is a valid design, and

$$\begin{aligned}&\mathrm{TV}\left(\mathcal{L}_{\theta^*}((x_i, y_i)_{i=1}^n), \mathcal{L}_{\theta'}((x_i, y_i)_{i=1}^n)\right) \\ &\quad \leq n\mathbb{E}_{x\sim\lambda}\left[\mathrm{TV}\left(\mathcal{L}_{\theta^*}(y(x) \mid x), \mathcal{L}_{\theta'}(y(x) \mid x)\right)\right].\end{aligned} \tag{58}$$

Proof of Lemma 4. Every term in Eq. (7) is nonnegative, and

$$\sum_{a\in\mathcal{A}} \lambda(a) = \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} \Pr(x_i = a) = 1.$$

Hence, λ is a valid design. The selected actions have the same joint law under θ^* and θ'. Conditioning on their joint value and applying tensorization of total variation (Lindvall, 2002) gives

$$\begin{aligned}&\mathrm{TV}\left(\mathcal{L}_{\theta^*}((x_i, y_i)_{i=1}^n), \mathcal{L}_{\theta'}((x_i, y_i)_{i=1}^n)\right) \\ &\quad \leq \mathbb{E}\left[\mathrm{TV}\left(\mathcal{L}_{\theta^*}((y_i)_{i=1}^n \mid x_1, \ldots, x_n), \mathcal{L}_{\theta'}((y_i)_{i=1}^n \mid x_1, \ldots, x_n)\right)\right] \\ &\qquad\qquad \text{(conditioning on the selected actions)} \\ &\quad \leq \sum_{i=1}^n \mathbb{E}\left[\mathrm{TV}\left(\mathcal{L}_{\theta^*}(y_i \mid x_i), \mathcal{L}_{\theta'}(y_i \mid x_i)\right)\right] \qquad \text{(tensorization of total variation)} \\ &\quad = n\mathbb{E}_{x\sim\lambda}\left[\mathrm{TV}\left(\mathcal{L}_{\theta^*}(y(x) \mid x), \mathcal{L}_{\theta'}(y(x) \mid x)\right)\right]. \qquad \text{(Eq. (7))}\end{aligned}$$

This establishes Eq. (58). □

C.1 Known moment exponent

Proof of Corollary 1. Let λ be the design in Eq. (7). If $M_\alpha(\lambda;\mathcal{A})=\infty$, choose u that vanishes on the support of λ and satisfies $\max_{a\in\mathcal{A}}|u^\top a|=1$. The parameters $\nu^{1/\alpha}u/8$ and $-\nu^{1/\alpha}u/8$ then both give all-zero responses with probability one for all sampled actions. At all non-sampled actions, we set the response to deterministically equal its conditional mean. A standard application of Le Cam's inequality then gives the desired result.

In the rest of the analysis, suppose that $M_\alpha(\lambda;\mathcal{A})<\infty$. The support of λ spans $\mathrm{Span}(\mathcal{A})$, and the set $\{u:\max_{a\in\mathcal{A}}|u^\top a|=1\}$ is compact on this span. The continuous expectation in Eq. (4) therefore attains its minimum, so there is a vector u satisfying

$$\max_{a\in\mathcal{A}}|u^\top a|=1,\qquad \mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}=M_\alpha(\lambda;\mathcal{A})^{-1/(\alpha-1)}.$$

Set

$$\Delta=\frac{1}{8}\nu^{1/\alpha}\min\left\{M_\alpha(\lambda;\mathcal{A})^{1/\alpha}n^{-(\alpha-1)/\alpha},1\right\}.$$

Lemma 3 supplies two response distributions for the parameters Δu and $-\Delta u$. Lemma 4 and Eq. (20) then give

$$\begin{aligned}&\mathrm{TV}\left(\mathcal{L}_{\Delta u}((x_i,y_i)_{i=1}^n),\mathcal{L}_{-\Delta u}((x_i,y_i)_{i=1}^n)\right)\\&\le n\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)}\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\\&\le 2^{-3-1/(\alpha-1)}\le\frac{1}{16}.\end{aligned}$$

An action attaining $|u^\top a|=1$ separates the two parameters by 2Δ. Le Cam's inequality therefore gives

$$\Pr\left(\max_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)|\ge\frac{1}{8}\nu^{1/\alpha}\min\left\{M_\alpha(\lambda;\mathcal{A})^{1/\alpha}n^{-(\alpha-1)/\alpha},1\right\}\right)\ge\frac{1}{4}.\tag{59}$$

Since $M_\alpha(\lambda;\mathcal{A})\ge M_\alpha(\mathcal{A})$, this proves Corollary 1 with $c=1/8$. □

C.2 Allocations independent of the moment exponent

Proof of Corollary 2. Let λ be defined by Eq. (7). It is independent of α, so Theorem 5 gives an $\alpha\in(1,2)$ for which

$$M_\alpha(\lambda;\mathcal{A})^{1/\alpha}\ge e^{-1}\sqrt{\frac{3}{20\log 4}}\sqrt{\frac{\log d}{\log\log d}}M_\alpha(\mathcal{A})^{1/\alpha}.\tag{60}$$

Applying Eq. (59) with this α gives

$$\Pr\left(\max_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)|\ge\frac{1}{8}\nu^{1/\alpha}\min\left\{M_\alpha(\lambda;\mathcal{A})^{1/\alpha}n^{-(\alpha-1)/\alpha},1\right\}\right)\ge\frac{1}{4}.\tag{61}$$

Combining Eq. (60) and Eq. (61) proves Corollary 2 with $c=e^{-1}\sqrt{3/(20\log 4)}/8$. □

D V-Optimality

V-optimality concerns accuracy averaged over the target actions. This means that an inaccurate estimate at one action contributes only its share of the total error, unlike the maximum loss studied in the main text. We specifically measure this average accuracy using the root mean squared error:

$$\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)|^2\right)^{1/2}.\tag{62}$$

For least squares under finite variance, the design enters the expected square of Eq. (62) through

$$\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} a^\top A(\lambda)^{-1}a. \tag{63}$$

This is the classical V criterion (Pukelsheim, 2006; Huan et al., 2024). Note that here and in Appendix E and Appendix F, normalization factors containing the noise level and sample size are omitted from the design criteria, as they only amount to rescaling.

With heavy-tailed errors, minimizing the classical V criterion can be polynomially suboptimal for the same estimation loss in Eq. (62), as the following example shows.

Example 2 (Classical V-optimality is polynomially suboptimal)**.** *Fix $\alpha = 3/2$, $\nu > 0$, and let $d \geq 8$ be a perfect cube. Consider the action set*

$$\mathcal{A} = \{\mathbf{e}_1, \dots, \mathbf{e}_{d-1}\} \cup \left\{ \left(1 + \frac{\ell}{d^{5/3}}\right)\mathbf{e}_d : \ell = 1, \dots, d^{5/3} \right\}.$$

The unique classical V-optimal design is supported on $\{\mathbf{e}_1, \dots, \mathbf{e}_{d-1}, 2\mathbf{e}_d\}$. There are universal positive constants c and C such that, whenever $n \geq Cd$, every estimator based on n i.i.d. samples from this classical design admits a parameter θ^* and a response distribution satisfying Eq. (1) for which

$$\Pr\left(\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)|^2 \right)^{1/2} \geq c\nu^{2/3} d^{1/18} n^{-1/3} \right) \geq \frac{1}{4}.$$

The design supported on the same d actions with

$$\begin{aligned} \lambda(\mathbf{e}_j) &= \frac{1}{2d-1}, \qquad j < d, \\ \lambda(2\mathbf{e}_d) &= \frac{d}{2d-1} \end{aligned} \tag{64}$$

is not classically V-optimal. Under independent sampling from this design, for every $\delta \in (0,1)$, direct robust coordinate estimation gives the following with probability at least $1-\delta$ when $n \geq Cd\log(2d/\delta)$:

$$\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} |a^\top(\widehat{\theta} - \theta^*)|^2 \right)^{1/2} \leq C\nu^{2/3} \left(\frac{\log(2d/\delta)}{n} \right)^{1/3}.$$

The ratio between the two bounds has dimension dependence $d^{1/18}$ up to logarithmic factors.

Proof. Since $1 \leq 1 + \ell/d^{5/3} \leq 2$, the squared lengths of the actions parallel to $\mathbf{e}_d$ satisfy

$$d^{5/3} \leq \sum_{\ell=1}^{d^{5/3}} \left(1 + \frac{\ell}{d^{5/3}}\right)^2 \leq 4d^{5/3}. \tag{65}$$

Moving positive mass from a shorter action parallel to $\mathbf{e}_d$ to $2\mathbf{e}_d$ increases the last diagonal entry of $A(\lambda)$ and leaves the other entries unchanged. It therefore strictly decreases Eq. (63). Hence, the classical design is supported on the stated d actions. On this support, Eq. (63) equals

$$\frac{1}{d - 1 + d^{5/3}} \left(\sum_{j=1}^{d-1} \frac{1}{\lambda(\mathbf{e}_j)} + \frac{\sum_{\ell=1}^{d^{5/3}} (1 + \ell/d^{5/3})^2}{4\lambda(2\mathbf{e}_d)} \right).$$

Strict convexity and the first order conditions give equal masses on the first $d-1$ coordinates and

$$\frac{\lambda(2\mathbf{e}_d)}{\lambda(\mathbf{e}_1)} = \frac{1}{2}\left(\sum_{\ell=1}^{d^{5/3}} (1 + \ell/d^{5/3})^2 \right)^{1/2}. \tag{66}$$

Together with $\sum_{x\in\mathcal{A}} \lambda(x) = 1$, Eq. (65) and Eq. (66) imply

$$\begin{aligned} \frac{1}{2d} \leq \lambda(\mathbf{e}_j) \leq \frac{2}{d}, &\qquad j < d, \\ \frac{1}{4}d^{-1/6} \leq \lambda(2\mathbf{e}_d) \leq 2d^{-1/6}. & \end{aligned} \tag{67}$$

For the lower bound, take $\theta^* = \Delta \mathbf{e}_d$ and $\theta' = -\Delta \mathbf{e}_d$, where

$$\Delta = \frac{1}{16}\nu^{2/3}(n\lambda(2\mathbf{e}_d))^{-1/3}.$$

Since $n\lambda(2\mathbf{e}_d) \geq 1$ when $n \geq 256d$, Lemma 3 applies with $u = \mathbf{e}_d$. Only observations at $2\mathbf{e}_d$ distinguish the parameters. The total variation distance between the laws of the n observations is at most

$$n\lambda(2\mathbf{e}_d)\left(\frac{4(2\Delta)^{3/2}}{\nu}\right)^2 = \frac{1}{32}.$$

Le Cam's inequality therefore gives, under one of these parameters,

$$\Pr\left(|\mathbf{e}_d^\top(\widehat{\theta} - \theta^*)| \geq \Delta\right) \geq \frac{1}{4}.$$

On this event, the contribution of the actions parallel to $\mathbf{e}_d$ gives

$$\begin{aligned}\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}|a^\top(\widehat{\theta} - \theta^*)|^2\right)^{1/2} &\geq \frac{1}{\sqrt{2}}|\mathbf{e}_d^\top(\widehat{\theta} - \theta^*)| && \text{(Eq. (65))}\\ &\geq \frac{1}{64}\nu^{2/3}d^{1/18}n^{-1/3}. && (\lambda(2\mathbf{e}_d) \leq 2d^{-1/6})\end{aligned}$$

Therefore, the lower bound in Example 2 holds.

For the competing design, Eq. (64) gives $\lambda(\mathbf{e}_j) \geq 1/(2d)$ for $j < d$ and $\lambda(2\mathbf{e}_d) \geq 1/2$. Combining these bounds with Eq. (65) yields

$$\begin{aligned}&\frac{1}{d - 1 + d^{5/3}}\left(\sum_{j=1}^{d-1}\lambda(\mathbf{e}_j)^{-2/3} + \frac{1}{4}\sum_{\ell=1}^{d^{5/3}}(1 + \ell/d^{5/3})^2\lambda(2\mathbf{e}_d)^{-2/3}\right)\\ &\quad\leq \frac{2d^{5/3} + 4d^{5/3}}{d^{5/3}} \leq 6.\end{aligned} \tag{68}$$

The median-of-means calculation in Example 1, applied to the d coordinate rewards, gives simultaneously

$$\begin{aligned}|\widehat{\theta}_j - \theta^*_j| &\leq 64\nu^{2/3}\left(\frac{\log(2d/\delta)}{n\lambda(\mathbf{e}_j)}\right)^{1/3}, \qquad j < d,\\ |\widehat{\theta}_d - \theta^*_d| &\leq 32\nu^{2/3}\left(\frac{\log(2d/\delta)}{n\lambda(2\mathbf{e}_d)}\right)^{1/3}.\end{aligned} \tag{69}$$

These inequalities hold simultaneously with probability at least $1 - \delta$ whenever $n \geq 160d\log(2d/\delta)$. Substituting Eq. (69) and Eq. (68) into Eq. (62) proves the upper bound. The classical mass ratio has order $d^{5/6}$, whereas the competing mass ratio equals d. The choices $c = 1/64$ and $C = 256$ therefore satisfy the claims of Example 2. □

D.1 Worst case minimax rate

The separation from classical V-optimality leaves open how accurately an experimenter can estimate on the hardest action set after choosing the design. Theorem 6 and Theorem 7 answer this question when $\max_{a\in\mathcal{A}}|a^\top\theta^*| \leq \nu^{1/\alpha}$.

Theorem 6 (Minimax upper bound for the V loss)**.** *Fix $1 < \alpha \leq 2$, a finite action set $\mathcal{A}$ spanning $\mathbb{R}^d$, $\nu > 0$, and $\delta \in (0, 1)$. There is a design and an estimator based on n i.i.d. samples from that design such that, for every response distribution satisfying Eq. (1) and every θ^* satisfying $\max_{a\in\mathcal{A}}|a^\top\theta^*| \leq \nu^{1/\alpha}$, whenever $n \geq C(d + \log(2/\delta))$, with probability at least $1 - \delta$,*

$$\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}|a^\top(\widehat{\theta} - \theta^*)|^2\right)^{1/2} \leq C\nu^{1/\alpha}\left(\frac{d + \log(2/\delta)}{n}\right)^{(\alpha-1)/\alpha},$$

where C is a universal constant.

Proof. Take the uniform design on $\mathcal{A}$. Its information matrix is positive definite because $\mathcal{A}$ spans $\mathbb{R}^d$. Fix a Euclidean $1/2$-net of the unit sphere containing at most 5^d points. For every point u in the net, use $A(\lambda)^{1/2}u$ as a target vector. If $v^\top A(\lambda)^{1/2}u = 1$, then

$$\begin{aligned}\mathbb{E}_{x\sim\lambda}|v^\top x|^{\alpha/(\alpha-1)} &\geq (v^\top A(\lambda)v)^{\alpha/(2(\alpha-1))} && \text{(Lyapunov's inequality)}\\ &\geq |v^\top A(\lambda)^{1/2}u|^{\alpha/(\alpha-1)} = 1. && \text{(Cauchy–Schwarz)}\end{aligned}$$

Lemma 2 therefore gives real numbers $(v_x)_{x\in\mathcal{A}}$ satisfying

$$\sum_{x\in\mathcal{A}} v_x x = A(\lambda)^{1/2}u, \qquad \sum_{x\in\mathcal{A}} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} \leq 1.$$

Apply the median-of-means estimator and final minimization from Theorem 1 to these net targets. The bound on $\max_{a\in\mathcal{A}}|a^\top\theta^*|$ and a union bound over the net give, with probability at least $1-\delta$,

$$\max_{u \text{ in the net}} |u^\top A(\lambda)^{1/2}(\widehat{\theta}-\theta^*)| \leq 256\nu^{1/\alpha}\left(\frac{\log(2\cdot 5^d/\delta)}{n}\right)^{(\alpha-1)/\alpha}.$$

The net property and uniformity of the design imply

$$\begin{aligned}\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}|a^\top(\widehat{\theta}-\theta^*)|^2\right)^{1/2} &= \|A(\lambda)^{1/2}(\widehat{\theta}-\theta^*)\|_2\\ &\leq 2\max_{u \text{ in the net}}|u^\top A(\lambda)^{1/2}(\widehat{\theta}-\theta^*)|.\end{aligned}$$

Since $\log(2\cdot 5^d/\delta) \leq 2(d+\log(2/\delta))$, Theorem 6 holds with $C = 1024$ in both the sample size condition and the error bound. □

Theorem 7 (Minimax lower bound for the V loss)**.** *Fix $1 < \alpha \leq 2$, $\nu > 0$, $\delta \in (0, 1/16]$, and $n \geq d + \log(1/\delta)$. For the action set $\mathcal{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, every design and every estimator based on n i.i.d. samples from that design admit a parameter satisfying $\max_{a\in\mathcal{A}}|a^\top\theta^*| \leq \nu^{1/\alpha}$ and a reward distribution satisfying Eq. (1) such that*

$$\Pr\left(\left(\frac{1}{d}\sum_{j=1}^{d}|\mathbf{e}_j^\top(\widehat{\theta}-\theta^*)|^2\right)^{1/2} \geq c\nu^{1/\alpha}\left(\frac{d+\log(1/\delta)}{n}\right)^{(\alpha-1)/\alpha}\right) \geq \delta,$$

where c is a universal positive constant.

Proof. Fix the design λ. At least $d/2$ coordinates satisfy $\lambda(\mathbf{e}_j) \leq 2/d$. Set

$$\Delta = \frac{1}{16}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{(\alpha-1)/\alpha}.$$

On these coordinates, consider parameters with $\theta_j^* \in \{-\Delta, \Delta\}$, and set the remaining coordinates to zero. Conditional on selecting $\mathbf{e}_j$ from this collection, let

$$y(\mathbf{e}_j) = \begin{cases}\operatorname{sign}(\theta_j^*)\left(\dfrac{\nu}{4\Delta}\right)^{1/(\alpha-1)}, & \text{with probability } \left(\dfrac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)},\\ 0, & \text{otherwise.}\end{cases}$$

For the remaining actions, set $y(\mathbf{e}_j) = 0$. By Lemma 3, the conditional mean is $\mathbf{e}_j^\top\theta^*$ and the centered conditional α moment is at most ν. Moreover,

$$\begin{aligned}\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} &= \left(\frac{4}{16^\alpha}\right)^{1/(\alpha-1)}\frac{d}{n}\\ &\leq \frac{d}{64n}. && (1 < \alpha \leq 2)\end{aligned}$$

For two parameters differing in one coordinate, tensorization bounds the total variation distance by

$$n\lambda(\mathbf{e}_j)\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} \le \frac{2n}{d}\frac{d}{64n} = \frac{1}{32}.$$

Assouad's lemma (Tsybakov & Zaiats, 2009, Chapter 2) now gives, under the uniform distribution over these parameters,

$$\mathbb{E}\left[\sum_{j:\ \lambda(\mathbf{e}_j)\le 2/d} \mathbb{I}\{\widehat{\theta}_j\theta_j^* \le 0\}\right] \ge \frac{d}{8}.$$

The sum is bounded by d, so it is at least $d/16$ with probability at least $1/16$. On this event, the V loss is at least $\Delta/4$. Hence one fixed parameter in the family satisfies

$$\Pr\left(\left(\frac{1}{d}\sum_{j=1}^{d}|\widehat{\theta}_j - \theta_j^*|^2\right)^{1/2} \ge \frac{1}{64}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{(\alpha-1)/\alpha}\right) \ge \frac{1}{16}.$$

For the confidence term, compare the two parameters

$$\theta^* = \Delta\sum_{j=1}^{d}\mathbf{e}_j \qquad \text{and} \qquad \theta' = -\Delta\sum_{j=1}^{d}\mathbf{e}_j, \qquad \Delta = \left(\frac{\nu}{4}\right)^{1/\alpha}\left(\frac{\log(1/(2\delta))}{2n}\right)^{(\alpha-1)/\alpha}.$$

For every selected action, let the reward distribution be either the following or its negation:

$$y(\mathbf{e}_j) = \begin{cases} \left(\dfrac{\nu n}{2\log(1/(2\delta))}\right)^{1/\alpha}, & \text{with probability } \dfrac{\log(1/(2\delta))}{2n}, \\ 0, & \text{otherwise,} \end{cases}$$

with the sign matching that of the parameter. The conditional means are Δ and $-\Delta$, respectively. In either case, the uncentered conditional α moment is $\nu/4$, and Eq. (24) bounds the centered conditional moment by ν. Since $n \ge \log(1/\delta)$, the probability of observing only zero rewards satisfies

$$\left(1 - \frac{\log(1/(2\delta))}{2n}\right)^n \ge \exp\left(-\log(1/(2\delta))\right) = 2\delta. \qquad (1 - t \ge e^{-2t} \text{ for } 0 \le t \le 1/2)$$

On this event, the observations have the same distribution under the two parameters, whose separation in Eq. (62) is 2Δ. Averaging over the two models shows that one of them has error at least Δ with probability at least δ. For $\delta \le 1/16$,

$$\Delta \ge \frac{1}{8}\nu^{1/\alpha}\left(\frac{\log(1/\delta)}{n}\right)^{(\alpha-1)/\alpha}. \qquad (\log(1/(2\delta)) \ge 3\log(1/\delta)/4)$$

The two thresholds satisfy

$$\max\left\{\frac{1}{64}\left(\frac{d}{n}\right)^{(\alpha-1)/\alpha}, \frac{1}{8}\left(\frac{\log(1/\delta)}{n}\right)^{(\alpha-1)/\alpha}\right\} \ge \frac{1}{128}\left(\frac{d+\log(1/\delta)}{n}\right)^{(\alpha-1)/\alpha}.$$

When the dimension term is larger, its probability $1/16$ is at least δ; otherwise, use the construction for the confidence term. It follows that Theorem 7 holds with $c = 1/128$. Both constructions satisfy its signal bound. □

Together, Theorem 6 and Theorem 7 show that the worst case minimax rate over finite action sets, after optimizing the design, is $\nu^{1/\alpha}((d+\log(1/\delta))/n)^{(\alpha-1)/\alpha}$ up to universal constants, under the sample-size and confidence conditions in the two theorems.

E A-OPTIMALITY

A-optimality concerns recovery of the parameter vector itself. Errors across parameter coordinates contribute to the Euclidean loss, including coordinates that are weakly represented in the available measurements. Under finite variance, its expected square is

$$\mathbb{E}\left[\|\widehat{\theta}-\theta^*\|_2^2\right]=\sum_{j=1}^{d}\mathbb{E}\left[|\mathbf{e}_j^\top(\widehat{\theta}-\theta^*)|^2\right].$$

For least squares, the classical A criterion is (Pukelsheim, 2006; Huan et al., 2024)

$$\mathrm{Tr}(A(\lambda)^{-1})=\sum_{j=1}^{d}\mathbf{e}_j^\top A(\lambda)^{-1}\mathbf{e}_j, \tag{70}$$

where the trace sums the variances of the parameter coordinates.

The trace criterion can allocate too few observations to a short action when the errors are heavy-tailed. The following example makes this failure explicit.

Example 3 (Classical A-optimality is polynomially suboptimal)**.** *Fix $1<\alpha<2$, $\nu>0$, and $d\geq 2$, and let*

$$\mathcal{A}=\left\{\mathbf{e}_1,\dots,\mathbf{e}_{d-1},d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right\}.$$

The classical A-optimal design assigns mass

$$\lambda(\mathbf{e}_j)=\frac{1}{d-1+d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}},\qquad j<d,$$
$$\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)=\frac{d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}}{d-1+d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}}.$$

There are universal positive constants c and C such that, whenever $n\geq 2d$, every estimator based on n i.i.d. samples from this classical design admits a parameter θ^* and a response distribution satisfying Eq. (1) for which

$$\Pr\left(\|\widehat{\theta}-\theta^*\|_2\geq c\nu^{1/\alpha}d^{\frac{5\alpha^2-\alpha-2}{4\alpha^2}}n^{-(\alpha-1)/\alpha}\right)\geq\frac{1}{8}.$$

The design assigning mass

$$\lambda(\mathbf{e}_j)=\frac{1}{d-1+d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}}},\qquad j<d,$$
$$\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)=\frac{d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}}}{d-1+d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}}}$$

minimizes the factor multiplying the coordinatewise robust Euclidean error bound and is not classically A-optimal. Under independent sampling from this design, for every $\delta\in(0,1)$, coordinatewise robust mean estimation gives, when $n\geq Cd\log(2d/\delta)$ and with probability at least $1-\delta$,

$$\|\widehat{\theta}-\theta^*\|_2\leq C\nu^{1/\alpha}d^{\frac{3\alpha-2}{2\alpha}}\left(\frac{\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha}.$$

The ratio between the two bounds has dimension dependence $d^{(2-\alpha)(\alpha-1)/(4\alpha^2)}$ up to logarithmic factors.

Proof. We first show that the exponent of the short action lies strictly between the two values needed for the separation:

$$\frac{\alpha^2+3\alpha-2}{4\alpha}-\frac{\alpha}{2}=\frac{(2-\alpha)(\alpha-1)}{4\alpha}>0,$$

$$\frac{3\alpha-2}{2\alpha}-\frac{\alpha^2+3\alpha-2}{4\alpha}=\frac{(2-\alpha)(\alpha-1)}{4\alpha}>0.$$

For a diagonal design, the objective in Eq. (70) is

$$\sum_{j=1}^{d-1}\frac{1}{\lambda(\mathbf{e}_j)}+\frac{d^{\frac{\alpha^2+3\alpha-2}{2\alpha}}}{\lambda(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d)}.$$

The objective is strictly convex and tends to infinity if any design mass tends to zero. At its minimizer, varying one mass while keeping the sum of the masses fixed therefore gives

$$\frac{1}{\lambda(\mathbf{e}_j)^2}=\frac{d^{\frac{\alpha^2+3\alpha-2}{2\alpha}}}{\lambda(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d)^2},\qquad j<d.$$

Thus the first $d-1$ weights are equal, and

$$\frac{\lambda(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d)}{\lambda(\mathbf{e}_j)}=d^{\frac{\alpha^2+3\alpha-2}{4\alpha}},\qquad j<d.$$

The constraint that the weights sum to one gives the classical design in Example 3.

Squaring the coordinatewise robust mean bounds and summing shows that the design-dependent factor in the squared Euclidean error bound is

$$\sum_{j=1}^{d-1}\lambda(\mathbf{e}_j)^{-\frac{2(\alpha-1)}{\alpha}}+d^{\frac{\alpha^2+3\alpha-2}{2\alpha}}\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)^{-\frac{2(\alpha-1)}{\alpha}}. \tag{71}$$

Under the classical design, the contribution of the short coordinate alone shows that the square root of Eq. (71) is at least

$$\left(d-1+d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}\right)^{\frac{\alpha-1}{\alpha}}d^{\frac{\alpha^2+3\alpha-2}{4\alpha^2}}. \tag{72}$$

The objective in Eq. (71) is strictly convex in the design weights. Equating its derivatives under the same constraint gives

$$\lambda(\mathbf{e}_j)^{-\frac{3\alpha-2}{\alpha}}=d^{\frac{\alpha^2+3\alpha-2}{2\alpha}}\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)^{-\frac{3\alpha-2}{\alpha}},\qquad j<d.$$

Consequently,

$$\frac{\lambda(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d)}{\lambda(\mathbf{e}_j)}=d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}},\qquad j<d.$$

Normalization gives the competing design in Example 3. Substitution followed by taking the square root then gives

$$\left(d-1+d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}}\right)^{\frac{3\alpha-2}{2\alpha}}. \tag{73}$$

The exponents in the two allocations are unequal when $\alpha<2$, so the designs are distinct.

It remains to verify the statistical claims. Both exponents appearing in the design weights are smaller than one because

$$1-\frac{\alpha^2+3\alpha-2}{4\alpha}=\frac{(2-\alpha)(\alpha+1)}{4\alpha}>0,$$
$$1-\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}=\frac{(2-\alpha)(\alpha-1)}{2(3\alpha-2)}>0.$$

Consequently, every mass in either design is at least $1/(2d)$. When $n\geq 2d$, apply Lemma 3 with $u=d^{(\alpha^2+3\alpha-2)/(4\alpha)}\mathbf{e}_d$ and

$$\Delta=\frac{1}{8}\nu^{1/\alpha}\left[n\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)\right]^{-(\alpha-1)/\alpha}.$$

Here $u^\top x$ is one at the short action and zero elsewhere, and $n\lambda(x) \geq 1$ ensures $|\Delta u^\top x|^\alpha \leq \nu/4$. By Eq. (20), the total variation distance between the laws of the n observations is at most

$$n\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)} = \left(\frac{4}{8^\alpha}\right)^{1/(\alpha-1)} \leq \frac{1}{16}.$$

The Euclidean distance between Δu and $-\Delta u$ is $2\Delta d^{(\alpha^2+3\alpha-2)/(4\alpha)}$. Le Cam's inequality therefore implies that one of the two models satisfies

$$\Pr\left(\|\widehat{\theta}-\theta^*\|_2 \geq \frac{1}{8}\nu^{1/\alpha}d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}\left[n\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)\right]^{-\frac{\alpha-1}{\alpha}}\right) \geq \frac{1}{4}. \tag{74}$$

Under the classical design, $\lambda(d^{-(\alpha^2+3\alpha-2)/(4\alpha)}\mathbf{e}_d) \leq 2d^{(\alpha^2+3\alpha-2)/(4\alpha)-1}$. Thus the threshold in Eq. (74) satisfies

$$\begin{aligned}&\frac{1}{8}\nu^{1/\alpha}d^{\frac{\alpha^2+3\alpha-2}{4\alpha}}\left[n\lambda\left(d^{-\frac{\alpha^2+3\alpha-2}{4\alpha}}\mathbf{e}_d\right)\right]^{-\frac{\alpha-1}{\alpha}}\\ &\qquad\geq \frac{1}{16}\nu^{1/\alpha}d^{\frac{5\alpha^2-\alpha-2}{4\alpha^2}}n^{-\frac{\alpha-1}{\alpha}}.\end{aligned} \tag{75}$$

Eq. (74) and Eq. (75) prove the lower bound with $c = 1/16$. For the competing design, the Chernoff and median-of-means calculation in Example 1, applied to the reward divided by the length of the corresponding diagonal action, gives the following simultaneously for all coordinates:

$$|\widehat{\theta}_j - \theta^*_j| \leq 64\nu^{1/\alpha}\frac{1}{|a^\top \mathbf{e}_j|}\lambda(a)^{-(\alpha-1)/\alpha}\left(\frac{\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha}, \tag{76}$$

where a is the unique action parallel to $\mathbf{e}_j$. The inequalities in Eq. (76) hold with probability at least $1-\delta$ when $n \geq 160d\log(2d/\delta)$. Squaring and summing the coordinate inequalities and using Eq. (73) yields

$$\begin{aligned}\|\widehat{\theta}-\theta^*\|_2 &\leq 64\nu^{1/\alpha}\left(d-1+d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}}\right)^{\frac{3\alpha-2}{2\alpha}}\left(\frac{\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha}\\ &\leq 128\nu^{1/\alpha}d^{\frac{3\alpha-2}{2\alpha}}\left(\frac{\log(2d/\delta)}{n}\right)^{(\alpha-1)/\alpha}. \qquad (d^{\frac{\alpha^2+3\alpha-2}{2(3\alpha-2)}} \leq d)\end{aligned}$$

The choice $C = 160$ then satisfies the upper bound. Subtracting the two dimension exponents gives

$$\frac{5\alpha^2-\alpha-2}{4\alpha^2} - \frac{3\alpha-2}{2\alpha} = \frac{(2-\alpha)(\alpha-1)}{4\alpha^2},$$

which proves the claimed separation. □

F E-Optimality

Classical E-optimality protects the least accurately estimated direction of the parameter. Under finite variance, it minimizes the largest expected squared error among unit directions:

$$\sup_{u\in\mathbb{R}^d:\ \|u\|_2=1}\mathbb{E}\left[|u^\top(\widehat{\theta}-\theta^*)|^2\right].$$

For least squares, the classical E criterion (Pukelsheim, 2006; Huan et al., 2024) is

$$\lambda_{\min}(A(\lambda))^{-1} = \sup_{u\in\mathbb{R}^d:\ \|u\|_2=1} u^\top A(\lambda)^{-1}u. \tag{77}$$

With heavy-tailed errors, we seek a high probability guarantee over all unit directions. Controlling their maximum error is equivalent to controlling $\|\widehat{\theta}-\theta^*\|_2$. Define the E design criterion by

$$M_{E,\alpha}(\lambda;\mathcal{A}) = \left(\inf_{u\in\mathbb{R}^d:\ \|u\|_2=1}\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)}\right)^{-(\alpha-1)}. \tag{78}$$

At $\alpha = 2$, this simplifies to

$$M_{E,2}(\lambda; \mathcal{A}) = \left(\inf_{u \in \mathbb{R}^d:\ \|u\|_2 = 1} u^\top A(\lambda) u \right)^{-1} = \lambda_{\min}(A(\lambda))^{-1},$$

so Eq. (78) recovers classical E-optimality exactly. The inner infimum in Eq. (78) is concave in λ, since the expectation is linear in λ for each fixed u. The map $s \mapsto s^{-(\alpha-1)}$ is convex and decreasing on $(0, \infty)$, so $M_{E,\alpha}$ is convex in λ. For $\alpha < 2$, evaluating the infimum over the Euclidean sphere remains a separate optimization problem; convexity in the design does not by itself give a polynomial time oracle for this objective.

Theorem 8 (Upper bound for E-optimality)**.** *Fix $1 < \alpha \le 2$, a finite action set $\mathcal{A}$ spanning $\mathbb{R}^d$, $\nu > 0$, $\delta \in (0, 1)$, and a design λ with finite $M_{E,\alpha}(\lambda; \mathcal{A})$. Let $B = \max_{a \in \mathcal{A}} |a^\top \theta^*|$. There is an estimator $\widehat{\theta}$ based on n i.i.d. samples from λ and a universal constant C such that, for every response distribution satisfying Eq. (1), whenever $n \ge C(d + \log(2/\delta))$, with probability at least $1 - \delta$,*

$$\|\widehat{\theta} - \theta^*\|_2 \le C(\nu + B^\alpha)^{1/\alpha} M_{E,\alpha}(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{d + \log(2/\delta)}{n} \right)^{(\alpha-1)/\alpha}.$$

Proof. Fix a Euclidean $1/2$-net of the unit sphere containing at most 5^d points. For a point a in this net, $u^\top a = 1$ implies $\|u\|_2 \ge 1$. Hence

$$\begin{aligned} \mathbb{E}_{x \sim \lambda} |u^\top x|^{\alpha/(\alpha-1)} &= \|u\|_2^{\alpha/(\alpha-1)} \mathbb{E}_{x \sim \lambda} \left| \left(\frac{u}{\|u\|_2} \right)^\top x \right|^{\alpha/(\alpha-1)} \\ &\ge \|u\|_2^{\alpha/(\alpha-1)} M_{E,\alpha}(\lambda; \mathcal{A})^{-1/(\alpha-1)} && \text{(Eq. (78))} \\ &\ge M_{E,\alpha}(\lambda; \mathcal{A})^{-1/(\alpha-1)}. && (\|u\|_2 \ge 1) \end{aligned}$$

Lemma 2 therefore gives

$$\begin{aligned} \min_{\sum_{x:\, \lambda(x) > 0} v_x x = a} \sum_{x:\, \lambda(x) > 0} \frac{|v_x|^\alpha}{\lambda(x)^{\alpha-1}} &= \left(\inf_{u:\, u^\top a = 1} \mathbb{E}_{x \sim \lambda} |u^\top x|^{\alpha/(\alpha-1)} \right)^{-(\alpha-1)} \\ &\le M_{E,\alpha}(\lambda; \mathcal{A}). \end{aligned} \tag{79}$$

Apply the median-of-means estimator in Theorem 1 to every point of the net. A union bound gives value estimates satisfying

$$|\widehat{f}_a - a^\top \theta^*| \le 64(\nu + B^\alpha)^{1/\alpha} M_{E,\alpha}(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{\log(2 \cdot 5^d/\delta)}{n} \right)^{(\alpha-1)/\alpha}$$

simultaneously over the net with probability at least $1 - \delta$. As in Theorem 1, choose $\widehat{\theta}$ to minimize $\max_{a \text{ in the net}} |a^\top \theta - \widehat{f}_a|$ over $\theta \in \mathbb{R}^d$. The net property and Eq. (19) give

$$\begin{aligned} \|\widehat{\theta} - \theta^*\|_2 &\le 2 \max_{a \text{ in the net}} |a^\top (\widehat{\theta} - \theta^*)| && (1/2\text{-net}) \\ &\le 4 \max_{a \text{ in the net}} |\widehat{f}_a - a^\top \theta^*| && \text{(Eq. (19))} \\ &\le 256(\nu + B^\alpha)^{1/\alpha} M_{E,\alpha}(\lambda; \mathcal{A})^{1/\alpha} \left(\frac{\log(2 \cdot 5^d/\delta)}{n} \right)^{(\alpha-1)/\alpha}. \end{aligned}$$

Finally, $\log(2 \cdot 5^d/\delta) \le 2(d + \log(2/\delta))$ and $(\alpha - 1)/\alpha \le 1/2$. Thus $C = 512$ is valid both in the sample size condition and in the error bound. □

Theorem 9 (Lower bound for E-optimality)**.** *Fix $1 < \alpha \le 2$, a finite action set $\mathcal{A}$ spanning $\mathbb{R}^d$, a design λ, and $\nu > 0$. There is a universal positive constant c such that every estimator based on n i.i.d. samples from λ admits a parameter θ^* and a response distribution satisfying Eq. (1) for which*

$$\Pr\left(\|\widehat{\theta} - \theta^*\|_2 \ge c\nu^{1/\alpha} \min\left\{ M_{E,\alpha}(\lambda; \mathcal{A})^{1/\alpha} n^{-(\alpha-1)/\alpha}, \frac{1}{\max_{a \in \mathcal{A}} \|a\|_2} \right\} \right) \ge \frac{1}{4}.$$

Proof. If $M_{E,\alpha}(\lambda;\mathcal{A}) = \infty$, there is a unit vector u with $u^\top x = 0$ whenever $\lambda(x) > 0$. In this case, take the two parameters

$$\theta^* = \frac{\nu^{1/\alpha}}{8\max_{a\in\mathcal{A}}\|a\|_2}u \qquad \text{and} \qquad \theta' = -\frac{\nu^{1/\alpha}}{8\max_{a\in\mathcal{A}}\|a\|_2}u.$$

Set $y(x) = 0$ on the support of λ and set it equal to its conditional mean elsewhere. The two observation laws are identical, so Le Cam's inequality proves Theorem 9 in this case. It remains to consider finite $M_{E,\alpha}(\lambda;\mathcal{A})$.

Since $\mathcal{A}$ is finite, the projection moment is continuous in u. Compactness of the unit sphere therefore gives a unit vector u satisfying

$$\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} = M_{E,\alpha}(\lambda;\mathcal{A})^{-1/(\alpha-1)}.$$

Set

$$\Delta = \frac{1}{8}\nu^{1/\alpha}\min\left\{M_{E,\alpha}(\lambda;\mathcal{A})^{1/\alpha}n^{-(\alpha-1)/\alpha}, \frac{1}{\max_{a\in\mathcal{A}}\|a\|_2}\right\}.$$

The definition of Δ gives $|\Delta u^\top x|^\alpha \le \nu/4$. Use the two parameters $\theta^* = \Delta u$ and $\theta' = -\Delta u$ with the conditional reward laws from Lemma 3, which satisfy Eq. (1).

By Eq. (20), the total variation distance between one sample under θ^* and one sample under θ' is

$$\begin{aligned}\left(\frac{4\Delta^\alpha}{\nu}\right)^{1/(\alpha-1)}\mathbb{E}_{x\sim\lambda}|u^\top x|^{\alpha/(\alpha-1)} &\le \frac{1}{n}\left(\frac{4}{8^\alpha}\right)^{1/(\alpha-1)} && \text{(definitions of } \Delta \text{ and } u)\\ &= \frac{1}{n}2^{-3-1/(\alpha-1)} \le \frac{1}{16n}. && (1<\alpha\le 2)\end{aligned}$$

Tensorization bounds the distance between the n-sample laws by $1/16$. The two parameters are separated by 2Δ in Euclidean norm. Le Cam's two-point inequality therefore gives error at least Δ with probability at least $1/4$ under at least one of θ^* and θ'. The choice $c = 1/8$ is thus valid. □

When the first term is the minimum in Theorem 9, the upper and lower bounds have the same dependence on $M_{E,\alpha}(\lambda;\mathcal{A})^{1/\alpha}$. The additional factor $(d+\log(2/\delta))^{(\alpha-1)/\alpha}$ in Theorem 8 comes from the union bound over the sphere net; the lower bound does not establish that this factor is necessary. Theorem 10 compares the minimizer of $M_{E,\alpha}$ with the classical E design.

Theorem 10 (Bounds for the optimal E design value)**.** *Fix $1 < \alpha \le 2$ and a finite action set spanning $\mathbb{R}^d$. If λ is a classical E-optimal design, then*

$$\begin{aligned}&\max\left\{\frac{d^{(\alpha-1)/\alpha}}{\max_{a\in\mathcal{A}}\|a\|_2}, \frac{1}{8}\sqrt{\frac{\alpha-1}{\alpha}}M_{E,2}(\lambda;\mathcal{A})^{1/2}\right\}\\ &\le \inf_{\lambda'\colon \lambda' \text{ is a design on } \mathcal{A}} M_{E,\alpha}(\lambda';\mathcal{A})^{1/\alpha}\\ &\le M_{E,2}(\lambda;\mathcal{A})^{1/2}.\end{aligned} \tag{80}$$

In particular, if $\max_{a\in\mathcal{A}}\|a\|_2 \le 1$ and the action set admits a design whose information matrix is I/d, then

$$\begin{aligned}&\max\left\{d^{(\alpha-1)/\alpha}, \frac{1}{8}\sqrt{\frac{\alpha-1}{\alpha}}\sqrt{d}\right\}\\ &\le \inf_{\lambda'\colon \lambda' \text{ is a design on } \mathcal{A}} M_{E,\alpha}(\lambda';\mathcal{A})^{1/\alpha}\\ &\le \sqrt{d}.\end{aligned} \tag{81}$$

Proof. Lyapunov's inequality gives

$$M_{E,\alpha}(\lambda;\mathcal{A})^{1/\alpha} \le M_{E,2}(\lambda;\mathcal{A})^{1/2}.$$

For every design λ', boundedness of each projection and the trace bound give

$$\inf_{u\in\mathbb{R}^d\colon \|u\|_2=1}\mathbb{E}_{x\sim\lambda'}|u^\top x|^{\alpha/(\alpha-1)}$$

$$\le \left(\max_{a\in\mathcal{A}} \|a\|_2\right)^{(2-\alpha)/(\alpha-1)} \lambda_{\min}(A(\lambda'))$$
$$\le \frac{1}{d}\left(\max_{a\in\mathcal{A}} \|a\|_2\right)^{\alpha/(\alpha-1)}. \qquad (\lambda_{\min}(A(\lambda')) \le \mathrm{Tr}(A(\lambda'))/d)$$

Taking the power $-(\alpha-1)/\alpha$ proves the first term in the lower bound in Eq. (80).

To prove the second term in the lower bound, write $\lambda_{\min}(A(\lambda')) = \min_{H\succeq 0:\,\mathrm{Tr}(H)=1} \mathrm{Tr}(HA(\lambda'))$ for each design λ'. Classical optimality of λ gives

$$M_{E,2}(\lambda;\mathcal{A})^{-1} = \max_{\lambda'} \min_{H\succeq 0:\,\mathrm{Tr}(H)=1} \sum_{x\in\mathcal{A}} \lambda'(x) x^\top H x$$
$$= \min_{H\succeq 0:\,\mathrm{Tr}(H)=1} \max_{\lambda'} \sum_{x\in\mathcal{A}} \lambda'(x) x^\top H x \qquad \text{(minimax theorem)}$$
$$= \min_{H\succeq 0:\,\mathrm{Tr}(H)=1} \max_{a\in\mathcal{A}} a^\top H a.$$

Here λ' ranges over designs on $\mathcal{A}$, and $H \in \mathbb{R}^{d\times d}$ ranges over positive semidefinite matrices with trace one. These sets are compact and convex, and the expression is linear in each variable, which justifies exchanging the minimum and maximum. Choose a minimizing H and let $g \sim \mathcal{N}(0, H)$. For any design λ',

$$\mathbb{E}_g \mathbb{E}_{x\sim\lambda'} |g^\top x|^{\alpha/(\alpha-1)} = \mathbb{E}|\mathcal{N}(0,1)|^{\alpha/(\alpha-1)} \mathbb{E}_{x\sim\lambda'} (x^\top H x)^{\alpha/(2(\alpha-1))} \qquad (g^\top x \sim \mathcal{N}(0, x^\top H x))$$
$$\le \mathbb{E}|\mathcal{N}(0,1)|^{\alpha/(\alpha-1)} M_{E,2}(\lambda;\mathcal{A})^{-\alpha/(2(\alpha-1))}. \qquad \text{(choice of } H)$$

Moreover,

$$\mathbb{E}_g \|g\|_2^2 = 1, \qquad \mathbb{E}_g \|g\|_2^4 = (\mathrm{Tr}H)^2 + 2\mathrm{Tr}(H^2) \le 3.$$

Paley–Zygmund's inequality and Markov's inequality then imply

$$\Pr_g(\|g\|_2^2 \ge 1/2) \ge \frac{1}{12},$$
$$\Pr_g\left(\mathbb{E}_{x\sim\lambda'} |g^\top x|^{\alpha/(\alpha-1)} \le 24 \mathbb{E}_g \mathbb{E}_{x\sim\lambda'} |g^\top x|^{\alpha/(\alpha-1)}\right) \ge \frac{23}{24}.$$

Their intersection has probability at least $1/12 - 1/24 = 1/24$. For a realization g in this intersection, $g/\|g\|_2$ is a unit vector, so

$$\inf_{u\in\mathbb{R}^d:\ \|u\|_2=1} \mathbb{E}_{x\sim\lambda'} |u^\top x|^{\alpha/(\alpha-1)}$$
$$\le \frac{\mathbb{E}_{x\sim\lambda'} |g^\top x|^{\alpha/(\alpha-1)}}{\|g\|_2^{\alpha/(\alpha-1)}}$$
$$\le 24 \cdot 2^{\alpha/(2(\alpha-1))} \mathbb{E}|\mathcal{N}(0,1)|^{\alpha/(\alpha-1)} M_{E,2}(\lambda;\mathcal{A})^{-\alpha/(2(\alpha-1))}.$$

Taking the power $-(\alpha-1)/\alpha$ gives

$$M_{E,\alpha}(\lambda';\mathcal{A})^{1/\alpha} \ge \frac{M_{E,2}(\lambda;\mathcal{A})^{1/2}}{\sqrt{2}\, 24^{(\alpha-1)/\alpha} \left(\mathbb{E}|\mathcal{N}(0,1)|^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}}.$$

We now use the following Gaussian moment bound (Vershynin, 2019, Section 2.5):

$$\left(\mathbb{E}|\mathcal{N}(0,1)|^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} \le \sqrt{\frac{\alpha}{\alpha-1}}.$$

Since $\sqrt{2}\, 24^{(\alpha-1)/\alpha} \le \sqrt{48} < 8$, substituting this moment bound and taking the infimum over λ' proves Eq. (80). Under the additional normalization, $\mathrm{Tr}(A(\lambda')) \le 1$ for every design, so $\lambda_{\min}(A(\lambda')) \le 1/d$. The isotropic design attains equality, and therefore its classical E value is d. Substitution proves Eq. (81). □

For each fixed α, Theorem 10 bounds the ratio of $M_{E,\alpha}(\lambda;\mathcal{A})^{1/\alpha}$ for a classical E-optimal design to its minimum independently of d. On any fixed interval separated from one, this ratio is uniformly bounded for every exponent in that interval. The dimension-independent comparison concerns design selection, whereas constructing an estimator that adapts to an unknown moment exponent is a separate statistical problem.

G Experimental Details and Additional Results

G.1 Action sets and sampling designs

Both action sets contain individual-coordinate measurements and measurements that combine several coordinates:

- **Hadamard and coordinate measurements.** For d a power of two, let $h_1, \dots, h_d$ be the normalized Sylvester Hadamard basis from Example 1. We use
$$\mathcal{A} = \{h_1, \dots, h_d\} \cup \{0.99\mathbf{e}_1, \dots, 0.99\mathbf{e}_d\}.$$
The unique G-optimal design is uniform on $h_1, \dots, h_d$. For the M_α design, probabilities are equal within each basis, and we minimize Eq. (4) over the total mass assigned to the Hadamard basis. Both bases receive positive mass at $\alpha = 4/3$.
- **Time and coefficient measurements.** For even d, define $x(t) \in \mathbb{R}^d$, for $t \in [0, 2\pi)$, by
$$x(t) = \frac{1}{\sqrt{d}}\Big(1, \sqrt{2}\cos t, \sqrt{2}\sin t, \dots, \sqrt{2}\cos((d/2-1)t), \sqrt{2}\sin((d/2-1)t), \cos(dt/2)\Big)^\top.$$
We use the finite action set
$$\mathcal{A} = \{x(\pi i/d) : i = 0, \dots, 2d-1\} \cup \{0.99\mathbf{e}_1, \dots, 0.99\mathbf{e}_d\}.$$
The d actions $x(2\pi i/d)$, for $i = 0, \dots, d-1$, form an orthonormal basis. Their norms are one, whereas the remaining time actions have squared norm $1 - 1/d$. The trace argument in Example 1 gives the unique G-optimal design as uniform on this basis. We compute the M_α design with the convex program in Lemma 1, reserving total probability 10^{-6} for the uniform design.

At $\alpha = 2$, both comparisons use the exact G-optimal allocation. Design probabilities are rounded to integer action counts summing to n, and reconstruction coefficients are recomputed for the rounded allocation. Responses are independent conditional on these actions.

G.2 Responses and evaluation

For both action sets and all three response distributions, $\theta^* = \mathbf{e}_1$.

- **Rare rewards.** If $x^\top\theta^* = 0$, the response is identically zero. Otherwise, let $p \in (0, 1/2]$ be the unique solution of
$$(1-p)|x^\top\theta^*|^\alpha + p\left|\frac{x^\top\theta^*}{p} - x^\top\theta^*\right|^\alpha = 1. \tag{82}$$
The response equals $x^\top\theta^*/p$ with probability p and zero otherwise. Thus its conditional mean is $x^\top\theta^*$ and its conditional centered α moment is exactly one.
- **Pareto.** The response is $x^\top\theta^*$ plus a Pareto variable with lower endpoint one and shape $\alpha + 0.02$, centered by subtracting its expectation $(\alpha + 0.02)/(\alpha - 0.98)$ and divided by its centered absolute α moment to the power $1/\alpha$. The normalizing moment is evaluated by numerical integration, independently of the observations.
- **Student t.** The response is $x^\top\theta^*$ plus a Student variable with $\alpha + 0.2$ degrees of freedom, divided by its absolute α moment to the power $1/\alpha$. Its expectation is zero, and the normalizing moment is evaluated using the gamma-function formula.

Consequently, every sampled action satisfies
$$\mathbb{E}[y(x) \mid x] = x^\top\theta^*, \qquad \mathbb{E}[|y(x) - x^\top\theta^*|^\alpha \mid x] \le 1.$$

The dimension and exponent sweeps use $n = 262{,}144$. The dimension sweep fixes $\alpha = 4/3$. The exponent sweep fixes $d = 32$ and uses $\alpha \in \{1.1, 1.2, 4/3, 1.5, 1.7, 1.9, 2\}$. The budget sweep fixes $d = 32$ and $\alpha = 4/3$, with $n \in \{65{,}536, 131{,}072, 262{,}144, 1{,}048{,}576\}$.

Figure 2 uses $d \in \{32, 64, 128, 256, 512, 1024, 2048, 4096\}$ on the Hadamard action set. For each repetition, error is evaluated on all actions, including those not sampled, using Eq. (2). The curves average these errors over sixteen independent repetitions; the bars give empirical standard errors. Under each design, we compute the average error of each of the twenty-two estimators in Appendix G.3 and report the smallest average. All estimators under a given design receive the same observations.

G.3 Estimators

The comparison in Figure 2 includes the following 21 estimators; this broad collection is used in order to mitigate the possibility of separations arising merely due to a poor choice of estimator. The tuning rules and iteration budgets are fixed across sampling designs. All regression estimators use the given model without incorporating an intercept term.

Least squares and robust losses.

- **Ordinary least squares.** We minimize the sum of squared residuals without regularization.
- **Ridge regression.** The quadratic penalty is selected by four-fold cross-validation using squared prediction loss.
- **Lasso.** The absolute penalty is selected using the same folds and validation loss as ridge regression.
- **Least absolute deviations.** We minimize the sum of absolute residuals. Note that under asymmetric noise, the conditional median need not equal the conditional mean.
- **Huber regression.** We use the scikit-learn `HuberRegressor` objective, which jointly estimates the regression coefficients and scale, with default loss parameter 1.35 and penalty 10^{-4}. The solver permits 1000 iterations with tolerance 10^{-8}. Large structured instances use the equivalent objective without expanding repeated observations.
- **Adaptive Huber regression.** Following the threshold calibration of Sun et al. (2020), we minimize Huber loss with fixed threshold $(n/\log((2d+1)/0.05))^{1/\alpha}$, using the true α and moment bound $\nu = 1$.

Block regression and loss comparisons.

- **Hsu–Sabato ridge regression.** We fit ridge regression with penalty $1/n$ on each of nine random blocks and apply the robust selection rule of Hsu & Sabato (2016). Each candidate is scored by its median squared distance to the other candidates, measured using their block information matrices with the ridge penalty added.
- **Minmax median-of-means regression.** We use the alternating optimization procedure of Lecué & Lerasle (2020) on squared-loss differences over nine random blocks, with at most 1000 iterations.
- **Trimmed regression.** We use Algorithm 1 of Oliveira & Resende (2025), trimming five observations from each tail of the loss differences. The two initial candidates are obtained by independently multiplying each least-squares coefficient by a uniform draw from $[0, 1]$. After 20 alternating iterations, we return the candidate with the smallest trimmed squared loss.
- **Cross-validated trimmed regression.** We use the same procedure and select the amount of trimming by cross-validation.

Gradients and filtering.

- **Robust gradient descent.** We use the method of Prasad et al. (2018), aggregating block gradients by their Euclidean geometric median. 25 blocks are reused over at most 100 iterations, with step size equal to the reciprocal of the largest eigenvalue of the design information matrix and stopping tolerance 10^{-7}.
- **SEVER.** We use the strategy of Diakonikolas et al. (2019) with ordinary least squares as the regression subroutine. Four rounds alternate regression with filtering by projections of the centered gradients onto their leading singular direction. The removal fraction is 0.025.

- **Filtered least trimmed squares.** We whiten the covariates using the design information matrix, apply the covariate filter of Pensia et al. (2025), and then perform iterative least trimmed squares with a trimming fraction of 0.05.
- **Clipped stochastic gradient descent.** We use one pass through a random permutation of the observations, starting at zero, and return the last iterate (Das et al., 2024). Gradients are clipped at Euclidean norm 10. At iteration t, the step size is $4/((t+8d)\lambda_{\min}(A(\lambda)))$ for the Hadamard action set.

Robust mean reconstruction. The first three methods estimate the response mean separately at each sampled action. A linear program then returns the parameter minimizing the maximum discrepancy from those estimates. The last two methods first form the reconstruction estimates in Appendix A.2 for every target action, then estimate their means and apply the same minimum-distance fit.

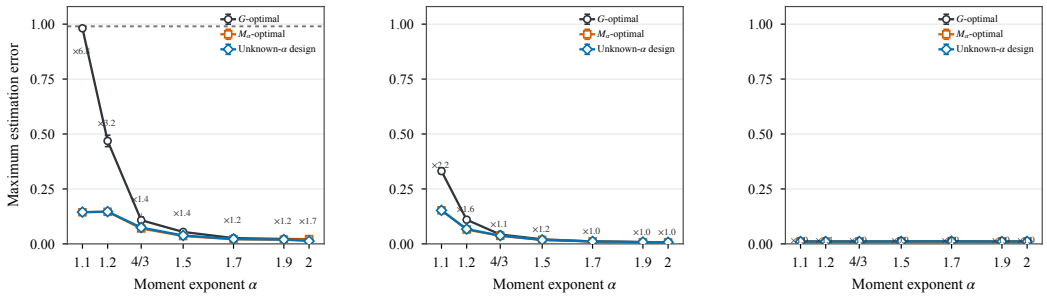
- **Actionwise trimmed mean.** At each sampled action, we remove $\lceil \log(2|\mathcal{A}|/0.05) \rceil$ observations from each tail and average the remaining responses (Oliveira & Resende, 2025).
- **Actionwise truncated mean.** We use the scalar truncation estimator of Bubeck et al. (2013). To obtain the threshold at an action, we multiply its observation count by $2^{\alpha}/\log(2|\mathcal{A}|/0.05)$ and raise this product to the power $1/\alpha$. Responses whose absolute values exceed this threshold are replaced by zero before averaging over the original observation count. The raw moment bound 2^{α} follows from $\nu = 1$ and $\max_{a \in \mathcal{A}} |a^{\top}\theta^*| \le 1$.
- **All-purpose mean estimation.** We apply the scalar estimator of Lee & Valiant (2022), analyzed under weak moments by Lee et al. (2025), to the observations at each sampled action. Its confidence input is 0.05 divided by the number of sampled actions. It does not require α or ν.
- **Median-of-means reconstruction (Moment MOM).** We divide the observations into blocks balanced within each sampled action. The number of blocks is the smallest odd integer at least $\log_2(2|\mathcal{A}|/0.05)$. We reconstruct each target value within every block and take the median across blocks (Lugosi & Mendelson, 2019b).
- **Trimmed-mean reconstruction.** We instead apply a scalar trimmed mean to the individual reconstruction estimates, removing $\lceil \log(3|\mathcal{A}|/0.05) \rceil$ observations from each tail (Oliveira & Resende, 2025). The final minimum-distance fit is the Chebyshev regression indicated in the estimator label.

Both reconstruction methods compute their coefficients using the true exponent α, including when the sampling design is independent of α.

Vector mean reconstruction (two methods). In this approach, we estimate the mean of the transformed response-covariate products and map the estimate back to the regression parameter, following Mourtada et al. (2022). The two mean estimators are the spectral weighting method of Hopkins et al. (2020) and the semidefinite descent method of Cherapanamjeri et al. (2019).

G.4 Experiment for design without knowledge of the moment exponent

We compare the designs in Figure 2b with λ_{fin} from Theorem 4. For both action sets, this design combines the grid in Eq. (38) with a uniform remainder and uses the stopping rule in Eq. (37). Its sampling probabilities remain fixed as the noise exponent varies. For the time measurements, weights smaller than 10^{-8} are redistributed uniformly; the redistributed mass is less than 10^{-7}.



(a) rare rewards. (b) Pareto. (c) Student t.

Figure 3: Comparison of the G-optimal, known-α, and unknown-α designs as the response exponent varies, with $d = 32$ and $n = 262{,}144$. Each curve reports the maximum estimation error averaged over sixteen repetitions, and each annotation gives the G-optimal error divided by the error under the unknown-α design. At every point and under every design, the plotted error is the smallest repetition average among the twenty-three estimators. Only the sampling design labeled unknown α is independent of the exponent.