Testing the Feasibility of Linear Programs with Bandit Feedback

Aditya Gangrade
Boston University, University of Michigan
gangrade@bu.edu

Aditya Gopalan
Indian Institute of Science
aditya@iisc.ac.in

Venkatesh Saligrama
Boston University
srv@bu.edu

Clayton Scott
University of Michigan
clayscot@umich.edu

Abstract

While the recent literature has seen a surge in the study of constrained bandit problems, all existing methods for these begin by assuming the feasibility of the underlying problem. We initiate the study of testing such feasibility assumptions, and in particular address the problem in the linear bandit setting, thus characterising the costs of feasibility testing for an unknown linear program using bandit feedback. Concretely, we test if $\exists x : Ax \geq 0$ for an unknown $A \in \mathbb{R}^{m\times d}$, by playing a sequence of actions $x_t \in \mathbb{R}^d$, and observing Ax_t + noise in response. By identifying the hypothesis as determining the sign of the value of a minimax game, we construct a novel test based on low-regret algorithms and a nonasymptotic law of iterated logarithms. We prove that this test is reliable, and adapts to the 'signal level,' Γ, of any instance, with mean sample costs scaling as $\widetilde{O}(d^2/\Gamma^2)$. We complement this by a minimax lower bound of $\Omega(d/\Gamma^2)$ for sample costs of reliable tests, dominating prior asymptotic lower bounds by capturing the dependence on d, and thus elucidating a basic insight missing in the extant literature on such problems.

1 Introduction

While the theory of single-objective bandit programs is well established, most practical situations of interest are multiobjective in character, e.g., clinicians trialling new treatments must balance the efficacy of the doses with the extent of their side-effects, and crowdsourcers must balance the speed of workers with the quality of their work. In cognisance of this basic fact, the recent literature has turned to the study of constrained bandit problems, wherein, along with rewards, one observes risk factors upon playing an action. For instance, along with treatment efficacy, one may measure kidney function scores using blood tests after a treatment. The goal becomes to maximise mean reward while ensuring that mean scores remain high (e.g. Nathan & DCCT/EDIC Research Group, 2014).

Many methods have been proposed for such problems, both in settings where constraints are enforced in aggregate, or in each round ('safe bandits'), see §1.1. However, every such method begins by assuming that the underlying program is feasible (or more; certain safe bandit methods require knowing a feasible ball). This is a significant assumption, since it amounts to saying that despite the fact that the risk factors are not well understood (hence the need for learning), it is known that the action space is well founded, and contains points that appropriately control the risk. This paper initiates the study of *testing this assumption*. The result of such a test bears a strong utility towards such constrained settings: if negative, it would inform practitioners of the inadequacy of their design space, and spur necessary improvements, while if positive, it would yield a cheap certificate to justify searching for optimal solutions within the space. The main challenge lies in ensuring that the tests are reliable and sample-efficient (since if testing took as many samples as finding optima, the latter question would be moot).

Concretely, we work in the linear bandit setting, i.e., in response to an action $x \in \mathcal{X} \subset \mathbb{R}^d$, we observe scores $S \in \mathbb{R}^m$ such that $\mathbb{E}[S|x] = Ax$, where A is latent, and with the constraint structured as $Ax \geq \alpha$ for a given tolerance vector α. We study the binary composite hypothesis testing problem of determining if there exists an $x : Ax \geq \alpha$ or not, with the goal of designing a sequential test that ensures that the probability of error is smaller than some given δ. Such a test is carried out for some random time τ, corresponding directly to the sample costs, which we aim to minimise. Effectively we are testing if an unknown linear program (LP) is feasible, and we may equivalently phrase the problem as testing the sign of the minimax value $\Gamma := \max_{x \in \mathcal{X}} \min_i (Ax - \alpha)^i$. Also note that by incorporating the objective as a constraint vector, and a proposed optimal value as a constraint level, this test also corresponds to solving the recognition (or decision) version of the underlying LP (e.g., Papadimitriou & Steiglitz, 1998, Ch. 15).

This problem falls within the broad purview of pure exploration bandit problems, and specifically the so-called *minimum threshold problem*, which has been studied in the *multi-armed case* for a *single constraint* (e.g. Kaufmann et al., 2018, also see §1.1). Most of this literature focuses on the asymptotic setting of $\delta \searrow 0$, and the typical result is of the form *if the instance is feasible, then there exist tests satisfying* $\lim \frac{\Gamma^2 \mathbb{E}[\tau]}{2 \log(1/\delta)} = 1$. Prima facie this is good news, in that there is a well-developed body of methods with tight instance specific costs that do not depend on the dimension of the action set, d! However, this lack of dependence should give us pause, since it does not make sense: if, e.g., $\mathcal{X}$ were a simplex, and only one corner of it were feasible, then detecting this feasibility should require us to search along each of the axes of $\mathcal{X}$ to locate *some* evidence, and so cost at least $\Omega(d)$ samples. The catch here lies in the limit, which implicitly enforces the regime $\delta = e^{-\omega(d)}$. Of course, even for modest d, such small a δ is practically irrelevant. Thus, even in the finite-armed case, the existing theory of feasibility testing does not offer a pertinent characterisation of the costs in scenarios of rich action spaces with rare informative actions.

Our contributions address this, and more. Concretely, we

- Design novel and simple tests for feasibility based on exploiting low-regret methods and laws of iterated logarithm to certify the sign of the minimax value Γ.
- Analyse these tests, and show that they are reliable and well-adapted to Γ, with stopping times scaling as $\widetilde{O}(d^2/\Gamma^2 + d \log(m/\delta)/\Gamma^2)$, thus demonstrating that the cost due to the number of constraints, m, is limited, and that testing is possible far more quickly than finding near-optimal points.
- Demonstrate a minimax lower bound of $\Omega(d/\Gamma^2)$ samples on the stopping time of reliable tests over feasible instances, thus showing that this uncaptured dependence is necessary.

We note that while the design approach of using low-regret methods for feasibility testing has appeared previously, their use arises either as subroutines in a complex method, or through modified versions of Thompson Sampling that are hard to even specify for the linear setting. Instead, our approach is directly motivated, and extremely simple, relying only on the standard technical tools of online linear regression and laws of iterated logarithms (LILs), employed in a new way to construct robust boundaries for our test statistics. Our results thus provide a new perspective on this testing problem, and more broadly on active hypothesis testing.

1.1 Related Work

Minimum Threshold testing. The single-objective finite-armed bandit setup (Lattimore & Szepesvári, 2020) posits $K < \infty$ actions, or 'arms,' and in each round, a learner may 'pull' one arm k to obtain a signal with mean $a_k \in \mathbb{R}$. The minimum threshold testing problem is typically formulated in this setup, and demands testing if $\max_{k \in [1:K]} a_k \geq \alpha$ or $< \alpha$ (notice that this is our problem, but with $\mathcal{X}$ finite and mutually orthogonal, and $m = 1$; see §D.1). The asymptotic behaviour of this problem has an asymmetric structure: if the instance is feasible, then lower bounds of the form $\liminf_{\delta \to 0} \log \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq \frac{2}{\Gamma^2}$ hold, while if the instance is infeasible, then the lower bound instead is $\sum_k \frac{2}{(\mu^k)^2}$, since each arm must be shown to have negative mean. Kaufmann et al. (2018) proposed the problem, and a 'hyper-optimistic' version of Thompson Sampling (TS) for it, called Murphy Sampling (MS), which is TS but with priors supported only on the feasible instances, and rejection boundaries based on the GLRT. We note that the resulting stopping

times were not analysed in this paper. Degenne & Koolen (2019) proposed a version of track and stop for this problem, but only showed asymptotic upper bounds on stopping behaviour; subsequently with Ménard (Degenne et al., 2019), they proposed a complex approach based on a two player game, with one of the players taking actions over the set of probability distributions on all infeasible or all feasible instances. The resulting stopping time bounds are stated in terms of the regret of the above player, and explicit forms of these for moderate δ are not derived. Further work has continued to study the single objective, finite-armed setting as $\delta \searrow 0$: Juneja & Krishnasamy (2019) extend the problem to testing if the mean vector $(a^k)_{k\in[1:K]}$ lies in a given convex set, and propose a track-and-stop method; Tabata et al. (2020) study index-based LUCB-type methods; Qiao & Tewari (2023) study testing if $0 \in (\min a_k, \max a_k)$, and propose a method that combines MS with two-arm sampling.[1]

Curiously, none of this work observes the simple fact that if only one arm were feasible, then searching for this arm must induce a $\Omega(K/\Gamma^2)$ sampling cost. This cost is significant when $1/\delta = \exp(o(K))$, which is the practically relevant scenario of moderate δ and large K. In §4, we show the the $\Omega(K/\Gamma^2)$ lower bound using the 'simulator' technique of Simchowitz et al. (2017). We note that while this method was previously applied to minimum threshold testing by Kaufmann et al. (2018), they focused on generic bounds, and only recovered a $(\log(1/\delta) + 1/K)\Gamma^{-2}$ lower bound. Instead, we show a minimax lower bound, losing this genericity, but capturing the linear dependence.

Along with demonstrating the above fact, the key distinction of our work is that we study a *multiobjective* feasibility problem in the more challenging (§D.1) *linear bandit setting*. We further note that many of the tests proposed for the finite-armed case are challenging to even define for the linear setting: MS requires sampling from the set of feasible instances $\{A \in \mathbb{R}^{m\times d} : \max_{\mathcal{X}} \min_i (Ax)^i \geq 0\}$, and the approach of Degenne et al. (2019) needs a low-regret algorithm for distributions over this highly nonconvex set. In sharp contrast, the tests we design are conceptually simple, and admit concrete bounds on sample costs. Thus, our work both extends this literature, and provides important basic insights for its nonasymptotic regime. It should be noted that one also expects statistical advantages: since the set of feasible instances is md dimensional, regret bounds on the same would vary polynomially in md, and thus one should expect stopping times to scale at best polynomially in md using the approach of Degenne et al. (2019), while our method admits bounds scaling only as $\mathrm{poly}(d, \log m)$.

In passing, we mention the parallel problem of finding either *all* feasible actions, called **thresholding bandits** (e.g. Locatelli et al., 2016), and of finding *one* feasible arm, called **good-arm identification** (e.g. Kano et al., 2017; Jourdan & Réda, 2023), assuming that they exist. Lower bounds in this line of work also focus on the asymptotic regime for finite-armed single objective cases. Of course, these problems are clearly harder than our testing problem, and so our lower bound also have implications for them.

Constrained and Safe Bandits. Multiobjective problems in linear bandit settings, amounting to bandit linear programming, are formulated as either aggregate constraint satisfaction (e.g. Badanidiyuru et al., 2013; Agrawal & Devanur, 2014, 2016) or roundwise satisfaction (called 'safe bandits', e.g. Amani et al., 2019; Katz-Samuels & Scott, 2019; Moradipari et al., 2021; Pacchiano et al., 2021; Chen et al., 2022; Wang et al., 2022; Camilleri et al., 2022). All such work assumes the feasibility of the underlying linear program to start with, and certain approaches further require knowledge of a safe point in the interior of the feasible set. Our study is directly pertinent to safe linear bandits, and to aggregate constrained bandits if $\mathcal{X}$ is convex.

Sequential Testing. Finally, some of the technical motifs in our work have previously appeared in the sequential testing literature. Most pertinently, Balsubramani & Ramdas (2015) define a test using the LIL, but without any actions (i.e., $|\mathcal{X}| = 1$). In their work, as in ours, the LIL is used to uniformly control the fluctuations of a noise process.

[1] While Qiao & Tewari (2023) *define* a very pertinent multiobjective problem, this is not analysed in their paper beyond an asymptotic lower bound that again does not capture K.

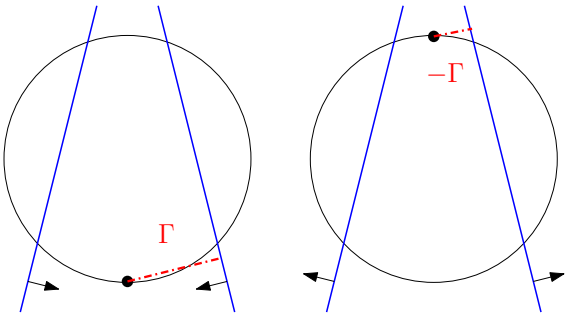

Figure 1: *Illustration of the Signal Level.* The ball is $\mathcal{X}$, and lines with arrows indicate the feasible half spaces for each constraint, and assuming that $\|A^i\| = 1$ for all i. *Left.* A feasible case; $\Gamma > 0$ is the distance of the marked point from the constraints, i.e., the length of the red dash-dotted line. *Right.* An infeasible case with $-\Gamma > 0$ shown similarly.

2 Definitions and Problem Statement

Notation. For a matrix M, M^i denotes the ith row of M, and for a vector z, z^i is the ith component of z. For a positive semidefinite matrix M, and a vector z, $\|z\|_M := \sqrt{z^\top M z}$ Standard Big-O and Big-Ω are used, and $\widetilde{O}$ further hides polylogarithmic factors of the arguments: $f(u) = \widetilde{O}(g(u))$ if $\exists c : \limsup_{u\to\infty} \frac{f(u)}{g(u)\log^c g(u)} < \infty$.

Setting. An instance of a linear bandit feasibility testing problem is determined by a domain $\mathcal{X}$, a latent constraint matrix $A \in \mathbb{R}^{m\times d}$, and a error level $\delta \in (0,1)$, to test[2]

$$\mathcal{H}_\mathsf{F} : \exists x \in \mathcal{X} : Ax \geq 0 \quad \text{vs.} \quad \mathcal{H}_\mathsf{I} : \forall x \in \mathcal{X} \exists i : (Ax)^i < 0,$$

where $\mathcal{H}_\mathsf{F}$ should be read as the 'feasibility hypothesis', and $\mathcal{H}_\mathsf{I}$ as the 'infeasibility hypothesis'. We shall also write $A \in \mathcal{H}_\mathsf{F}$ or $\in \mathcal{H}_\mathsf{I}$ if the corresponding hypothesis is true.

Information Acquisition proceeds over rounds indexed by $t \in \mathbb{N}$. For each t, the tester selects some action x_t, and observes scores $S_t \in \mathbb{R}^m$ such that $S_t = Ax_t + \zeta_t$, where ζ_t is assumed to be a subGaussian noise process. The information set of the tester after acquiring feedback in round t is $\mathsf{H}_t := \{(x_s, S_s)\}_{s\leq t}$, and the choice x_t must be adapted to the filtration generated by H_{t-1}. We let $X_{1:t} := \begin{bmatrix} x_1 & x_2 & \cdots & x_t \end{bmatrix}^\top, S_{1:t} := \begin{bmatrix} S_1 & S_2 & \cdots & S_t \end{bmatrix}$ denote the matrices whose rows are the xs and Ss up to t.

A **Test** is comprised of three components: (i) a (possibly stochastic) *action selecting algorithm* $\mathscr{A} : \mathcal{H}_{t-1} \to \mathcal{X}$, (ii) a *stopping time* τ adapted to H_t,and (iii) a *decision rule* $\mathscr{D} : \mathsf{H}_\tau \to \{\mathcal{H}_\mathsf{F}, \mathcal{H}_\mathsf{I}\}$. In each round, these are executed as follows: we begin by executing $\mathscr{A}$ to determine a new action for the round, and update the history with the feedback gained. We then check if $\tau = t$ to verify if we have accumulated enough information to reliably test, and if so, we stop, and if not, we conclude the round. Upon stopping, we evaluate the decision of $\mathscr{D}$, and return its output as the conclusion of the test. The design of $(\mathscr{A}, \tau, \mathscr{D})$ can of course depend on $(\mathcal{X}, \delta, m)$, but not on A. The basic reliability requirement for such a test is captured below.

Definition 1. *A test $(\mathscr{A}, \tau, \mathscr{D})$ is said to be reliable if for any instance $(\mathcal{X}, A, \delta)$, and $* \in \{\mathsf{F}, \mathsf{I}\}$ if $A \in \mathcal{H}_*$, then it holds that $\mathbb{P}(\mathscr{D}(\mathsf{H}_\tau) \neq \mathcal{H}_*) \leq \delta$.*

Signal level, and adaptive timescale. The hypotheses $\mathcal{H}_\mathsf{F}, \mathcal{H}_\mathsf{I}$ can equivalently be defined according to the sign of $\max_x \min_{i\in[1:m]} (Ax)^i$. We define the *signal level* of an instance as $\Gamma := \max_x \min_i (Ax)^i$. This is illustrated in Fig. 1. Notice that $|\Gamma|$ must enter the costs of testing. Indeed, even if we revealed to the tester the minimax (x^*, i_*), and the value of Γ, since the KL divergence between $\mathcal{N}(\Gamma, 1)$ and $\mathcal{N}(-\Gamma, 1)$ is Γ^2, we would still $\Omega(\Gamma^{-2}\log(1/\delta))$ samples to determine the sign $(Ax^*)^{i_*}$ (see, e.g., Lattimore & Szepesvári, 2020, Ch. 13,14). Thus, Γ^{-2} determines the minimal timescale for reliable testing, motivating

[2]notice that we have dropped the tolerance levels α in this definition. Since α is known a priori, this is without loss of generality: we can augment the dimension by appending a 1 to each action, and $-\alpha^i$ to the ith row of the constraint matrix A.

Definition 2. *We say that a test is valid if it is reliable, and for any instance with signal level* $\Gamma > 0$*, the test eventually stops, that is,* $\mathbb{P}(\tau < \infty) = 1$*. We further say that the test is well adapted to the signal level if it holds that for fixed* d, δ*,* $\mathbb{E}[\tau] = O(\Gamma^{-2}\text{polylog}(\Gamma^{-2}))$*.*

Any well adapted and reliable test must be valid. Further, a well adapted test is *fast* compared to finding near-optimal actions for safe bandit problems in feasible instances, since Γ is determined by the 'most-feasible' point in $\mathcal{X}$. For instance, consider a crowdsourcing scenario where we want to maximise the net amount of work done in a given time period, subject to meeting a quality score constraint of Q units. Since the number of very high quality workers in the pool may be limited, optimal solutions would need to use relatively low quality workers. However, verifying that such workers meet the constraint requires time proportional to $\min_w (Q^w - Q)^{-2}$, where Q^w is the mean quality of worker w. In contrast, Γ is determined by $\max_w (Q^w - Q)$, i.e., how good the best workers are, and so Γ^{-2} is much smaller than the time scale required to find an optimal solution.

Standard Conditions. While briefly discussed above, we explicitly impose the following conditions, standard in the linear bandit literature (see, e.g., Abbasi-Yadkori et al., 2011). All results in this paper assume the following.

Assumption 3. *We assume that the instance is bounded,*[3] *that is,* $\mathcal{X} \subset \{\|x\| \leq 1\}$*, and* $\{\forall i, \|A^i\| \leq 1\}$*. We also assume the noise* ζ_t *to be conditionally* 1*-subGaussian, i.e.,*

$$\mathbb{E}[\zeta_t | \mathscr{G}_t] = 0, \forall \lambda \in \mathbb{R}^m, \mathbb{E}[\exp(\lambda^\top \zeta_t) | \mathscr{G}_t] \leq \exp(\|\lambda\|^2/2),$$

where $\mathscr{G}_t$ *is the filtration generated by* H_{t-1}, x_t*, and any algorithmic randomness used by the test.*

3 Feasibility Tests Based on Low-Regret Methods

We begin by heuristically motivating our test, and discussing the challenges arising in making this generic and formal. This is followed by an explicit description of the tests, along with main results analysing their performance.

3.1 Motivation

For simplicity, let us consider the case of $m = 1$, so that $A = a^\top$, for a vector a, and the signal level is $\Gamma = \max_{\mathcal{X}} a^\top x$. Due to the duality between testing and confidence sets (Lehmann & Romano, 2005, §3.5), a principled approach to testing the sign of Γ is to build a confidence sequence for it, i.e., processes $\ell_t \leq u_t$ such that with high probability, $\forall t, \Gamma \in (\ell_t, u_t)$. We naturally stop when $\ell_t u_t > 0$, and decide on a hypothesis using the sign of ℓ_t on stopping. Any such confidence set in turn builds an estimate of Γ itself, that is, some statistic that eventually converges to Γ, at least if we did not stop. This raises the following basic question: how can we estimate $\max a^\top x$ without knowing where the maximum lies? A simple resolution to this comes from using low-regret methods for linear bandits.

The linear bandit problem is parameterised by an objective θ, and a domain $\mathcal{X}$, and a method for it picks actions x_t sequentially with the aim to minimise the *pseudoregret* $\mathscr{R}_t := \sum \max_x \theta^\top x - \theta^\top x_t$, using feedback of the form $\theta^\top x_t + \text{noise}$. For 'good' algorithms, $\mathscr{R}_t$ scales as $\widetilde{O}(\sqrt{d^2 t})$, at least in expectation (e.g. Lattimore & Szepesvári, 2020, Ch.19). Now, notice that if we take $\mathscr{A}$ to be such an algorithm executed with the feedback $S_t = a^\top x_t + \zeta_t$, then the statistic $\mathscr{T}_t/t$, where $\mathscr{T}_t := \sum S_s$, should eventually converge to $\max_x a^\top x = \Gamma$. Indeed

$$\mathscr{T}_t = \sum_{s \leq t} S_s = \sum_{s \leq t} a^\top x_s + \sum_{s \leq t} \zeta_s,$$

[3] If we are augmenting the dimension to account for nonzero α, these conditions apply only to the unaugmented A, x.

and so the error in this estimate behaves as

$$\Gamma - \mathscr{T}_t/t = \left(t\Gamma - \sum a^\top x_s\right)/t - \sum \zeta_s/t = (\mathscr{R}_t + Z_t)/t,$$

where Z_t is a random walk, and so is typically $O(\sqrt{t})$. If $\mathscr{R}_t \in [0, \widetilde{O}(\sqrt{d^2 t})]$, we can recover the *sign* of Γ reliably if $t\Gamma \gg \mathscr{R}_t + Z_t = \widetilde{O}(\sqrt{d^2 t}) \iff t \gg d^2/\Gamma^2$.

Formalising this heuristic approach, however, requires resolving two key issues. Firstly, we need to handle the multiobjective character of our testing problem: if $A \in \mathcal{H}_\mathsf{I}$, there may be actions with only one out of m constraints violated, and detecting this may be nontrivial. Secondly, to get a reliable test requires explicit statistics that can track the fluctuations in the noise, and in the pseudoregret (which is random due to the choice of x_t) in a reliable anytime way. These factors strongly influence the design of our tests.

3.2 Background on Online Linear Regression, and on Laws of Iterated Logarithms

Before proceeding with describing our tests and results, we include a brief discussion of necessary background.

Online Linear Regression. We take the standard approach (Abbasi-Yadkori et al., 2011). The 1-regularised least squares (RLS) estimate of A using H_{t-1} is

$$\hat{A}_t := S_{1:t-1} X_{1:t-1} (X_{1:t-1}^\top X_{1:t-1} + I)^{-1}. \tag{1}$$

Let us define the signal strength as $V_t := \sum_{s<t} x_s x_s^\top + I$, and for $\delta \in (0,1)$, the m-confidence radius as

$$\omega_t(\delta) = 1 + \sqrt{\frac{1}{2} \log \frac{m\sqrt{\det V_t}}{\delta}}.$$

The main results are based on the following two concepts, which we explicitly delineate.

Definition 4. *For any time t, the* RLS confidence set *is*

$$\mathscr{C}_t(\delta) := \{\tilde{A} : \forall \text{ rows } i, \|\tilde{A}^i - \hat{A}_t^i\|_{V_t} \le \omega_t(\delta)\},$$

and the local noise-scale *is* $\rho_t(x;\delta) := 2\omega_t(\delta)\|x\|_{V_t^{-1}}$.

Evidently, the set $\mathscr{C}_t$ captures the $\tilde{A}$ that are plausible values of A given H_{t-1}, the information available at the start of round t. We shall use the following standard results on the consistency of $\mathscr{C}_t$ (Abbasi-Yadkori et al., 2011).

Lemma 5. *For any instance and sequence of actions $\{x_t\}$,*

$$\mathbb{P}(\exists t : A \notin \mathscr{C}_t(\delta)) \le \delta.$$

Further, if $A \in \mathscr{C}_t(\delta)$, then

$$\forall \tilde{A} \in \mathscr{C}_t(\delta), x \in \mathcal{X} : |\tilde{A}x - Ax| \le \rho_t(x;\delta)\mathbf{1},$$

where the inequality is interpreted row-wise. Finally, for any sequence of actions $\{x_t\}$,

$$\sum_{s \le t} \rho_t(x_t;\delta) \le \sqrt{6dt\omega_t(\delta)\log(1 + t/d)}.$$

Nonasymptotic Law of Iterated Logarithms. To the control the fluctuations introduced by the feedback noise, we use the following LIL due to Howard et al. (2021).

Algorithm 1 Ellipsoidal Optimistic-Greedy Test (EOGT)

Input: $\delta \in (0,1), N \geq 2, \mathcal{X}, m$.
Initialise: $\mathsf{H}_0 \leftarrow \varnothing, \mathscr{T}_0 \leftarrow 0, \mathscr{B}_0 \leftarrow 0$.
for $t = 1, 2, \ldots$ **do**
  $\delta_t \leftarrow \delta t^{-N}, \mathscr{D}_t \leftarrow \mathscr{C}_t(\delta_t/2)$. *(Action Selection)*
  $(x_t, i_t) \leftarrow \max_{\tilde{A} \in \mathscr{D}_t, x \in \mathcal{X}} \min_i (\tilde{A}x)^i$.
  Play x_t, and observe S_t.
  Update $\mathsf{H}_t \leftarrow \mathsf{H}_{t-1} \cup \{(x_t, S_t)\}$.
  Update $\mathscr{T}_t \leftarrow \sum_{s \leq t} S_s^{i_s}, \mathscr{B}_t(\delta)$ as per (4)
  if $|\mathscr{T}_t| > \mathscr{B}_t(\delta)$ **then**
    STOP *(Stopping Rule)*
Output $\mathscr{T}_t \underset{\mathcal{H}_1}{\overset{\mathcal{H}_\mathsf{F}}{\gtrless}} 0$ *(Decision Rule)*

Lemma 6. *For $t \in \mathbb{N}, \delta \in (0,1)$, let*

$$\mathrm{LIL}(t, \delta) := \sqrt{4t \log \frac{11 \max(\log t, 1)}{\delta}}.$$

If $\eta_t \in \mathbb{R}$ is a conditionally centred and 1-subGaussian sequence adapted to a filtration $\{\mathscr{G}_t\}$, then for $H_t := \sum \eta_t$,

$$\mathbb{P}(\exists t : |H_t| > \mathrm{LIL}(t, \delta)) \leq \delta.$$

3.3 The Ellipsoidal Optimistic-Greedy Test

We are now ready to describe our first proposed test, EOGT which is specified in Algorithm 1. The test is parametrised by δ, and a constant N, and the algorithm proceeds by constructing a confidence set $\mathscr{D}_t = \mathscr{C}_t(\delta_t/2)$ for A, which is the standard confidence set, but with a decaying confidence parameter $\delta_t = \delta t^{-N}$. It then selects both an action x_t, and a *measured constraint* i_t by solving the program[4]

$$\max_{\tilde{A} \in \mathscr{D}_t} \max_{x \in \mathcal{X}} \min_{i \in [1:m]} (\tilde{A}x)^i. \tag{2}$$

The action x_t is played, and the selected constraint i_t determines the main test statistic:

$$\mathscr{T}_t := \sum_{s \leq t} (S_s)^{i_s}. \tag{3}$$

The test stops at $\tau := \inf\{t : |\mathscr{T}_t| > \mathscr{B}_t(\delta)\}$, that is, when the magnitude of $\mathscr{T}_t$ crosses the boundary

$$\mathscr{B}_t(\delta) := \sum_{s \leq t} \rho_s(x_s; \delta_s/2) + \mathrm{LIL}(t, \delta/2). \tag{4}$$

This test can be interpreted in a game theoretic sense. Recall that Γ is the value of the zero-sum game $\max_x \min_i (Ax)^i$. We can interpret the max player as a 'feasibility-biased player', that moves first to pick an x that makes Ax large, and the min player as an 'infeasibility-biased' player that counters with a constraint that x does not meet well.

In EOGT, action selection procedure is feasibility-biased: given the lack of knowledge of A, the feasibility player chooses a plausible $\tilde{A}$ that makes the value as high as possible, and the infeasibility player must abide

[4]Note that the order of optimisation is important in (2): since $(x, \tilde{A}) \mapsto \tilde{A}x$ is not quasiconvex, this value is in general not the same as $\min_i \max_{\tilde{A}, x} (\tilde{A}x)^i$. Of course, it does hold that $\max_{\tilde{A}} \max_x \min_i (\tilde{A}x)^i = \max_{\tilde{A}} \min_i \max_x (\tilde{A}x)^i$.

by this choice of $\tilde{A}$. This is countered by the infeasibility-biased statistic $\mathscr{T}_t$, in which only the infeasibility player's choice of i_t is accounted for. This strikes a delicate balance: in the feasible case, as long as x_t converges to a feasible subset of $\mathcal{X}$, $\mathscr{T}_t$ eventually grows large and positive, while under infeasibility, if i_t captures which constraints the x_ts consistently violate, $\mathscr{T}_t$ eventually grows large and negative. Notice that while the feasibility player hedges their lack of information with optimism over the confidence ellipsoid, the infeasibility player acts greedily in the above test (and this structure inspires the name EOGT). This greediness is natural if we view the infeasibility player as learner in a contextual stochastic full-feedback game, with context $(\tilde{A}_t, x_t)$, action i_t, and noisy feedback of the losses $\{(Ax_t)^i\}$.

The reliability of the test depends strongly on the form of the boundary $\mathscr{B}_t(\delta)$ above, which in turn arises from the analysis of the approach, which we shall now sketch.

3.3.1 Analysis of Reliability

Naturally, the analysis differs if the problem is feasible or infeasible. Let us assume that $A \in \mathscr{D}_t$ for all t. Since $\mathscr{D}_t \subset \mathscr{C}_t(\delta/2)$, this occurs with probability at least $1 - \delta/2$.

Signal growth in the feasible case relies on the optimism of the feasibility player. Let $(\tilde{A}_t, x_t, i_t)$ denote a solution to (2). Since A was feasible for this program, it must hold that $(\tilde{A}_t x_t)^{i_t} \geq \max_x \min_i (Ax)^i = \Gamma$. Further, since $\tilde{A} \in \mathscr{D}_t$, using Lemma 5, it holds that $(\tilde{A}_t x_t)^{i_t} \leq (Ax_t)^{i_t} + \rho_t(x_t; \delta_t/2)$, and so $(Ax_t)^{i_t} \geq \Gamma - \rho_t(x_t; \delta_t/2)$. Defining the noise process $Z_t = \sum_{s \leq t} \zeta_s^{i_s}$ lets us conclude that

$$\mathscr{T}_t \geq t\Gamma - \sum_{s \leq t} \rho_s(x_s; \delta_s/2) + Z_t.$$

Signal growth in the infeasible case instead relies on the extremisation in i_t given x_t. Let $i_{\min}(x) := \arg\min_i (Ax)^i$. Since i is the innermost optimised variable, and since $i_{\min}(x_t)$ is feasible for the program (2), it must hold that $(\tilde{A}_t x_t)^{i_t} \leq (\tilde{A}_t x_t)^{i_{\min}(x_t)}$. But, again, using Lemma 5, $(\tilde{A}_t x_t)^{i_{\min}(x_t)} \leq (Ax_t)^{i_{\min}(x_t)} + \rho_t(x_t; \delta_t/2)$, and further, $(Ax_t)^{i_{\min}(x_t)} = \min_i (Ax_t)^i \leq \max_x \min_i (Ax)^i = \Gamma < 0$. Therefore, in the infeasible case,

$$\mathscr{T}_t \leq t\Gamma + \sum \rho_s(x_s; \delta_s/2) + Z_t.$$

Boundary design and reliability. Finally, the boundary design follows from control on the term Z_t above. Notice that since i_t is a predictable process, and ζ_t is conditionally 1-subGaussian, it follows that $\eta_t := \zeta_t^{i_t}$ constitutes a centred, conditionally 1-subGaussian process, and thus invoking the LIL (Lemma 6) immediately yields

Lemma 7. *EOGT ensures that, with probability at least $1 - \delta$, simultaneously for all $t \geq 1$,*

$$\begin{aligned} \textit{feasible case:} \quad & \mathscr{T}_t \geq t\Gamma - \mathscr{B}_t(\delta) \geq -\mathscr{B}_t(\delta), \\ \textit{infeasible case:} \quad & \mathscr{T}_t \leq -t|\Gamma| + \mathscr{B}_t(\delta) \leq \mathscr{B}_t(\delta). \end{aligned}$$

Since we stop when $|\mathscr{T}_t| > \mathscr{B}_t(\delta)$, under the above event, upon stopping, $\mathscr{T}_\tau \Gamma > 0$, making the test reliable. This leaves the question of the validity of the test, and the behaviour of $\mathbb{E}[\tau]$, which we now address.

3.3.2 Control on Stopping time

Next, we describe our main result on the validity EOGT, and the behaviour of $\mathbb{E}[\tau]$. To succinctly state this, we define

$$\begin{aligned} T(\Gamma; \delta, N) := \inf \Big\{ & t \geq 2d : t|\Gamma| > 2\mathrm{LIL}(t, \delta/2) \\ & + 4dt^{1/2} \log(2t/d) + 2(dt \log(2t/d) \log \frac{2m}{\delta t^{-N}})^{1/2} \Big\} \end{aligned}$$

Our main result, shown in §B, is

Theorem 8. *For any δ and $N > 1$, the* EOGT *is valid and well adapted. In particular,*

$$\mathbb{E}[\tau] = O(T(\Gamma/2; \delta, N) + \delta/|\Gamma|).$$

To interpret this result, in §B.1, we employ worst-case bounds on $\sum_{s \le t} \rho_s(x_s; \delta_s)$ to control $T(\Gamma; \delta, N)$.

Lemma 9. *For any fixed N, $T(\Gamma; \delta, N)$ is bounded as*

$$O\left(\frac{d^2 \log^2(d^2/\Gamma^2)}{\Gamma^2} + \frac{d \log(m/\delta) \log(d \log(m/\Gamma^2 \delta))}{\Gamma^2}\right).$$

Implications. The main point that the above results make is that in the moderate δ regime of $\log 1/\delta = o(d)$, the typical stopping time of EOGT is bounded as d^2/Γ^2 up to logarithmic factors. The factor of d^2 in this bound is deeply related to the analysis of online linear regression, and also commonly appears in the regret bounds (both in the worst case, $\sqrt{d^2 t}$, as well as in gapped instance-wise cases (Dani et al., 2008; Abbasi-Yadkori et al., 2011)).

Next, we note that the d^2/Γ^2 time-scale is typically much faster than that needed to approximately solve a feasible safe bandit instance: the best known method for finding a ε-optimal action for safe bandits requires $\Omega(d^2/\varepsilon^2)$ samples (Camilleri et al., 2022). However, as discussed after Definition 2, Γ is driven by the 'safest' feasible action, while, since the optima lie at a constraint boundary, obtaining reasonably safe solutions requires setting $\varepsilon \ll \Gamma$, making d^2/Γ^2 significantly smaller than d^2/ε^2. We also note that the above bound may be considerably outperformed by any run of the test: because $\mathscr{B}_t$ adapts to the trajectory, its growth can be much slower than the worst case bound that enters the definition of $T(\Gamma; \delta, N)$, allowing for fast stopping.

Finally, observe that the dependence of this time scale on the number of constraints, m, is very mild, demonstrating that from a statistical point of view, many constraints are almost as easy to handle as one constraint.

3.4 Tail Behaviour, and the Tempered EOGT

While the expected stopping time of EOGT is well behaved, its tail behaviour may be much poorer. Indeed, the best tail bound we could show, as detailed in §B.3, is

Theorem 10. *For every $(\mathcal{X}, A, \delta)$, and $\eta \in (0, \delta)$, EOGT executed with parameters (δ, N) satisfies*

$$\mathbb{P}(\tau > T(\Gamma; \delta, N)) \le \delta, \text{ and further,}$$
$$\mathbb{P}\left(\tau > (2 + 1/|\Gamma|)\left\lceil (\delta/\eta)^{1/N} \right\rceil + T(\Gamma/2; \eta, N)\right) \le \eta.$$

Notice that the tail bound above is heavy, and the η-th quantile is only bounded as $O(1/|\Gamma|\eta^{-1/N})$. It is likely that such behaviour is unavoidable due to (2), due to which, if $m = 1$, EOGT directly exploits the OFUL algorithm of Abbasi-Yadkori et al. (2011), and the pseudoregret for this method is also heavy-tailed (Simchi-Levi et al., 2023).

One way to avoid this poor behaviour is to instead select actions using variants of OFUL-type methods that achieve light-tailed pseudoregret. As summarised in Algorithm 2, we use the recently proposed approach of Simchi-Levi et al. (2023) to construct such a test. The main difference is in selecting (x_t, i_t) according to the program

$$\max_{x \in \mathcal{X}} \min_{i \in [1:m]} (\hat{A}_t x)^i + \mathrm{Rad}_t(x), \tag{5}$$
$$\text{where } \mathrm{Rad}_t(x) := (t/d)^{1/2} \|x\|^2_{V_t^{-1}} + \sqrt{d \|x\|^2_{V_t^{-1}}}.$$

As a point of comparison, the selection rule (2) can roughly be understood as (5), but with $\mathrm{Rad}'_t = \sqrt{d \log t \|x\|^2_{V_t^{-1}}}$. Thus, the effect of Rad_t is to make the method more prone to exploration than (2) if t is large and $\|x\|_{V_t^{-1}} \gg d/\sqrt{t}$. So, the rule (5) has the effect of tempering the tendency to exploitation of (2), leading to the name 'tempered EOGT' (T-EOGT). Importantly, observe that the selection rule (5) makes no explicit reference to δ.

The remaining algorithmic challenge is to define a boundary that can lead to a reliable test based on the above approach. In order to do this, we refine the techniques of Simchi-Levi et al. (2023) to construct the following anytime tail bound, shown in §C.1, for $\widetilde{\mathscr{T}_t}$. We note that this also yields an *anytime* tail bound for the regret of (5) for linear bandits.

Lemma 11. *For $\delta \in (0, 1/2)$, let*

$$\mathscr{Q}^{\mathsf{F}}_t(\delta) := 45\sqrt{dt \log^4 t}(d + \log(8m/\delta)) + \mathrm{LIL}(t, \delta/2),$$
$$\mathscr{Q}^{\mathsf{I}}_t(\delta) := 27\sqrt{dt \log^3 t}(\sqrt{d} + \log(8m/\delta)) + \mathrm{LIL}(t, \delta/2).$$

Then, for $\widetilde{\mathscr{T}_t} := \sum_{s \le t} S_s^{i_s}$ with actions picked via (5),

$$\mathbb{P}(\forall t, \widetilde{\mathscr{T}_t} \ge t\Gamma - \mathscr{Q}^{\mathsf{F}}_t(\delta)) \ge 1 - \delta \quad \textit{(feasible case)}$$
$$\mathbb{P}(\forall t, \widetilde{\mathscr{T}_t} \le t\Gamma + \mathscr{Q}^{\mathsf{I}}_t(\delta)) \ge 1 - \delta \quad \textit{(infeasible case)}$$

Naturally, we can reliably test via the stopping times

$$\widetilde{\tau} = \inf\{t : \widetilde{\mathscr{T}_t} < -\mathscr{Q}^{\mathsf{F}}_t(\delta) \text{ or } \widetilde{\mathscr{T}_t} > \mathscr{Q}^{\mathsf{I}}_t(\delta)\},$$

deciding for $\mathcal{H}_{\mathsf{F}}$ if $\widetilde{\mathscr{T}_\tau} > 0$. Using this, in §C, we show the following bounds along the lines of §3.3.1.

Theorem 12. *T-EOGT is valid and well adapted, with*

$$\mathbb{E}[\widetilde{\tau}] = \widetilde{O}(d^3/\Gamma^2 + d/\Gamma^2 \log(8m/\delta))$$

where the $\widetilde{O}$ hides logarithmic dependence on d/Γ^2, and $\log(m/\delta)$. Further, there exists a C scaling polylogarithmically in d/Γ^2 and $\log(m/\eta)$ such that for all $\eta \le \delta$,

$$\mathbb{P}(\widetilde{\tau} \ge Cd^3/\Gamma^2 + Cd/\Gamma^2 \log(1/\eta)) \le \eta.$$

To contextualise the result, as well as this tempered test, let us consider the tradeoffs expressed in the above result. Compared to EOGT, the procedure of T-EOGT suffers two main drawbacks: firstly, we see that the bound on the stopping time is significantly weaker, scaling as d^3/Γ^2 instead of d^2/Γ^2, indicating a loss of performance. While this result may just be an artefact of the analysis, a more important drawback is that the test boundaries $\mathscr{Q}^{\mathsf{F}}, \mathscr{Q}^{\mathsf{I}}$ do not adapt to the sequence of actions actually played by the method, unlike $\mathscr{B}_t$, and instead are just deterministic processes that can be seen to essentially dominate $\sum \rho_s(x_s; \delta_s)$. Even if these bounds had tight constants (which they do not), such a nonadaptive stopping criterion cannot benefit from possible discovery of good actions early in the trajectory (accumulating on which would lead to contraction of ρ_t, and thus decelaration of $\mathscr{B}_t$), and so cannot benefit from early termination that EOGT may exploit in practice.

However, this weakness is balanced by considerably stronger tail behaviour: indeed, instead of the polynomial decay in tail probabilities for EOGT, the above demonstrates exponential decay in the tails, with the decay scale further behaving as $d/\Gamma^2 \ll d^2/\Gamma^2$, meaning that typical fluctuations in the stopping time must be considerably smaller than the typical stopping time. The choice of test must depend the setting, and T-EOGT should be preferred over EOGT if rare but extreme testing delays yield strong penalties.

Algorithm 2 Tempered EOGT (T-EOGT)

Input: $\delta \in (0, 1/2), \mathcal{X}, m$.
Initialise: $\mathsf{H}_0 \leftarrow \varnothing, \widetilde{\mathscr{T}_0} \leftarrow 0$
for $t = 1, 2, \ldots$ **do**
  Compute $\hat{A}_t$. *(Arm Selection)*
  $(x_t, i_t) \leftarrow \max_{x \in \mathcal{X}} \min_i (\hat{A}_t x)^i + \mathrm{Rad}_t(x)$.
  Play x_t, and observe S_t.
  Update $\mathsf{H}_t \leftarrow \mathsf{H}_{t-1} \cup \{(x_t, S_t)\}$.
  Update $\widetilde{\mathscr{T}_t} \leftarrow \sum_{s \le t} S_s^{i_s}$, and $\mathscr{Q}^{\mathsf{F}}, \mathscr{Q}^{\mathsf{I}}$.
  if $\widetilde{\mathscr{T}_t} > \mathscr{Q}_t^{\mathsf{F}}(\delta)$ or $\widetilde{\mathscr{T}_t} < -\mathscr{Q}_t^{\mathsf{I}}(\delta)$ **then**
    STOP *(Stopping Rule)*
Output $\widetilde{\mathscr{T}_t} \underset{\mathcal{H}_{\mathsf{I}}}{\overset{\mathcal{H}_{\mathsf{F}}}{\gtrless}} 0$. *(Decision Rule)*

Finally, we would be remiss not to mention the curious difference in the boundaries $\mathscr{Q}^{\mathsf{F}}$ and $\mathscr{Q}^{\mathsf{I}}$, and in particular the weakness in $\mathscr{Q}^{\mathsf{F}}$ which is inherited in the bounds on $\mathbb{E}[\tau]$ in Theorem 12. This difference arises because when controlling $\widetilde{\mathscr{T}_t}$ from below in the feasible case, we need the means $(Ax_t)^{i_t}$ to not be too far below the minimax value Γ, which is attained at some $x^* \neq x_t$. Ensuring this requires us to have control on both the noise scale at x_t *and that at* x_*. The latter is hard to accommodate in the analysis, which instead uses a lossy application of the AM-GM inequality to avoid it, but at the cost of the extra factor of $d^{1/2}$ in $\mathscr{Q}^{\mathsf{F}}$. On the other hand, when controlling $\widetilde{\mathscr{T}_t}$ from above in the infeasible case, we only need to ensure that i_t cannot do too poor a job of locating constraints that x_t violates, which can be achieved by just considering the noise scale at x_t itself. It may be possible to improve the analysis to reduce $\mathscr{Q}^{\mathsf{F}}$ down to $\mathscr{Q}^{\mathsf{I}}$, which we leave as a direction for future work.

4 Minimax Lower Bounds

We conclude the paper by discussing minimax lower bounds that capture the necessity of the dependence on Γ^{-2}, as well as at least a linear dependence on d in generic bounds on stopping times for reliable tests. As we previously discussed in §1 and §1.1, the main point of comparison for these results are the corresponding *instance-wise* lower bounds in the literature on the minimum threshold problem, which take essentially[5] the following form (Kaufmann et al., 2018)

$$\mathbb{E}[\widetilde{\tau}] \ge 2\log(1/\delta)/\Gamma^2 + 1/K\Gamma^2 \qquad \textit{(feasible case)},$$
$$\mathbb{E}[\widetilde{\tau}] \ge 2\log(1/\delta)\sum_k (\mu^k)^{-2} + 1/K\Gamma^2 \qquad \textit{(infeasible case)}.$$

Notice that in the feasible case, the lower bound *decays* with K. While the instance specific nature of the above bounds is desirable, we focus on minimax bounds capturing a linear dependence on K (or, in our case, d) in specific instances.

Our lower bound is based on a reduction to a finite action case, through the use of a simplex. The argument underlying this bound relies on the 'simulator' technique of Simchowitz et al. (2017) for best arm identification (BAI). In fact, our main point, that the extant bounds for feasibility testing do not capture the dependence on d, is much the same as the observation of Simchowitz et al. (2017) that the analyses of 'track-and-stop' BAI methods do not capture the right dependence on K in BAI, again due to a focus on $\delta \to 0$.

[5] the terms containing $\log(1/\delta)$ are always valid. The secondary terms behaving as $1/(K\Gamma^2)$ are upper bounds on the auxiliary terms appearing in the results of Kaufmann et al. (2018).

The construction underlying the bound is natural: we take $\mathcal{X}$ to be the simplex $\{x \geq 0 : \sum x_i = 1\}$, and consider a single constraint matrix $a^\top$ for a vector $a \in [-1/2, 1/2]^d$. The noise process is as follows: upon playing an action x_t, we sample $K_t \sim x_t$, and supply the tester with $a_{K_t} + \mathcal{N}(0, 1/2)$. The vector a is selected as a uniform permutation of the entries of $(\Gamma, -\varepsilon, -\varepsilon, \cdots, -\varepsilon)$, the intuition being that in order to detect the feasibility of such an instance, the test must sample the single 'informative' extreme direction of the simplex at least $1/(\Gamma + \varepsilon)^2$ times. However, since this is selected uniformly at random, no method can generically identify this direction faster that just sampling uniformly, and so on average across the instances, $\tau = \Omega(d/\Gamma^2)$. Concretely, in §D we show the bound in a finite-armed case, and argue that the instance above must face the same costs. A technically interesting observation is that our argument relies on two uses of the simulator technique: we first compare the instance against an infeasible instance to argue that the arm with large signal must be played often, and we then use this result along with the simulator technique again to show that arms with poor signal must also be played often in an average sense across the permutations. Leaving the details to §D, this yields the following result.

Theorem 13. *For any $\Gamma, \delta \in (0, 1/2)$ and any reliable $(\mathscr{A}, \tau, \mathscr{D})$, there exists a* feasible *instance $(\mathcal{X}, A, \delta)$ with $m = 1$ and signal level Γ on which $\mathbb{E}[\tau] \geq \frac{(1-2\delta)^3}{79} \cdot \frac{d}{\Gamma^2}$.*

Note that utilising the existing results of Kaufmann et al. (2018) for the infeasible case, we can also recover a lower bound of $d/\Gamma^2 \log(1/\delta)$ if $|\Gamma| \leq 1/\sqrt{d}$, by taking the instance $(-|\Gamma|, -|\Gamma|, \cdots, -|\Gamma|)$. Thus the linear dependence on d is necessary over both feasible and infeasible cases.

We comment that the lower bound of $\Omega(d/\Gamma^2)$ remains far from the upper bounds of $O(d^2/\Gamma^2)$ in Theorem 8. This linear in d gap in the lower bound is a persistent occurrence in the theory of linear bandits, and shows up in any instance-specific control on the same, including in known regret lower bounds. As a result, resolving this is a task beyond the scope of the present paper. Nevertheless, our main point that the costs of testing depend strongly on d, unlike prior analysis suggsets, is well made by the above result.

5 Simulations

We conclude the paper by describing a heuristic implementation of EOGT, and its behaviour on the simple case of testing the feasibility of two linear constraints over the unit ball.

L_1 Confidence set. Implementing EOGT is challenging task, since the maximin program (2) is difficult to solve quickly. Indeed, even if $m = 1$, i.e., there were only a single constraint, (2) requires us to implement the OFUL iteration, which is well known to be NP-hard due to the nonconvex objective $A^1 x$ (Dani et al., 2008).

To handle this, we begin with the standard relaxation used to implement OFUL, specifically by replacing the confidence ellipsoid $\mathscr{C}_t(\delta)$ by the L_1-confidence set

$$\widetilde{\mathscr{C}}_t(\delta) := \{\tilde{A} : \text{ for all rows } i, \|(\tilde{A}^i - \hat{A}_t^i) V_t^{1/2}\|_1 \leq \sqrt{d}\omega\}.$$

Since $\|\cdot\|_2 \leq \|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$, $\widetilde{\mathscr{C}}_t \supset \mathscr{C}_t$, and thus $\widetilde{\mathscr{C}}_t$ is consistent w.h.p. Further, $\widetilde{\mathscr{C}}_t$ is in turn contained in a scaling of $\mathscr{C}_t$ by a $\sqrt{d}$-factor, and thus the noise-scales over $\mathscr{C}_t$ carries over, up to a loss of a $\sqrt{d}$ factor. This suggests that tests based on $\widetilde{\mathscr{C}}_t$ should use $\widetilde{O}(d^3/\Gamma^2)$ samples.

The main advantage, however, is that due to the L_1 structure, the set $\widetilde{\mathscr{C}}_t(\delta)$ only has $(2d)^m$ extreme points. This enables optimisation by a simple search over these extreme points, which at least for small m, leads to an implementable algorithm. In the following, we will only work with $m = 2$.

Solving the Maximin Program. Of course, even for a given A, $\max_x \min_i A^i x$ is nonobvious to solve since i is discrete. We take the natural approach via convexifying:

$$\max_{\tilde{A} \in \widetilde{\mathscr{C}}_t(\delta), x \in \mathcal{X}} \min_i \tilde{A}^i x = \max_{\tilde{A} \in \widetilde{\mathscr{C}}_t(\delta)} \max_{x \in \mathcal{X}} \min_{\pi \in \Delta} \pi^\top \tilde{A} x,$$

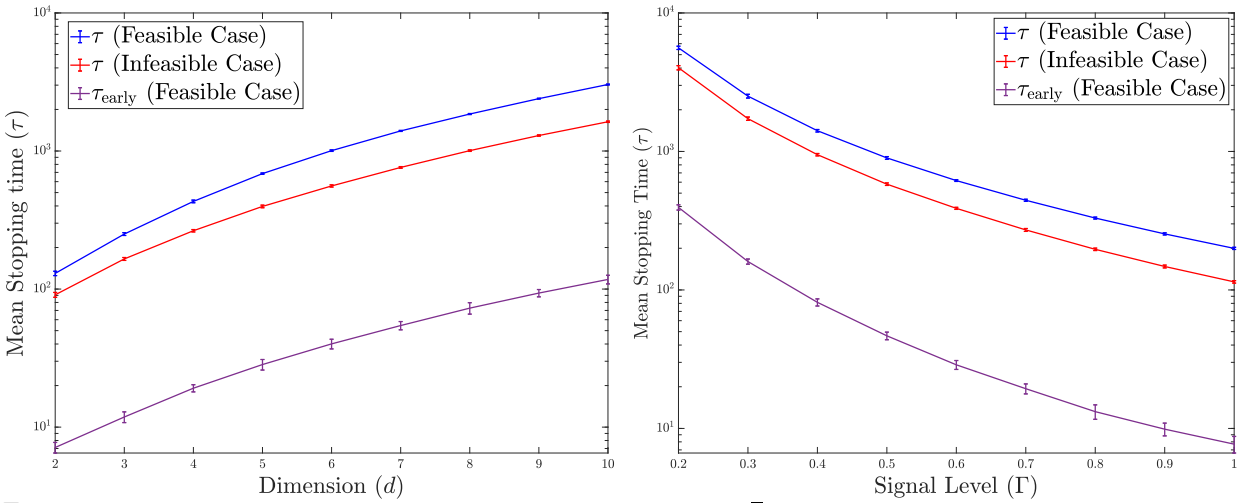

Figure 2: Behaviour of the stopping time as d is varied for fixed $\Gamma = 1/\sqrt{2}$ (*left*) and Γ is varied for fixed $d = 4$ (*right*) over the unit ball with $m = 2$. Averages and one-sigma error bars over 50 runs are reported. The test never returned an incorrect hypothesis. Notice the sharp advantage of τ_{early} in feasible cases, in that it is about a factor of 10 smaller than τ. (*best viewed zoomed-in*)

where Δ is the simplex in $\mathbb{R}^m$. Now, for a fixed $\tilde{A}$, the maximin program over (x, π) can be solved efficiently. The resulting $x, \tilde{A}$ can be used to directly minimise $(\tilde{A}x)^i$.

Procedure. Throughout the following, we will restrict attention to $\mathcal{X} = \{\|x\|_2 \leq 1\}$. This enables a further simplification by using the minimax theorem for a fixed $\tilde{A}$:

$$\max_{x \in \mathcal{X}} \min_{\pi \in \Delta} \pi^\top \tilde{A} x = \min_{\pi \in \Delta} \max_{x \in \mathcal{X}} \pi^\top \tilde{A} x = \min_{\pi \in \Delta} \|\pi^\top \tilde{A}\|_2.$$

Overall, this yields the following procedure: we enumerate the extreme points of $\widetilde{\mathscr{C}_t}$, and for each, we solve for the minimising π above, while keeping track of the maximum such value as we move over the extreme points. Upon conclusion, this yields a π_t and a $\tilde{A}_t$ that solve the above. x_t is then computed directly as $\pi_*^\top \tilde{A}_* / \|\pi_*^\top \tilde{A}_*\|$. Given $x_t, \tilde{A}_t$, we finally direclty solve for i_t by minimising $(\tilde{A}_t x_t)^i$.[6]

Early Stopping for Feasible Instances. Notice that in the feasible case, if we can ever argue that for some x, $\min_{\widetilde{\mathscr{C}_t(\delta)}} \min_i (\tilde{A}x)^i > 0$, then the test can already conclude. A natural candidate for such an x is simply the running mean over the choices of x_t played by EOGT. The potential advantage of such a procedure is that it bypasses the possibly slow growth of $\mathscr{T}_t$ when initial exploration chooses infeasible actions (which lead to a direct decrease in $\mathscr{T}_t$, but do not affect the quality of the noise estimate at x_t much). We also implement this early stopping procedure, and we will call the resulting stopping time τ_{early}.

Settings We study two scenarios: varying d for a fixed Γ, and varying Γ for a fixed d. In each case we study both feasible and infeasible instances.

In the varying d scenario, we pick the feasible instance $x_1 \geq 0, x_2 \geq 0$, and the infeasible instance $x_1 \geq 1/\sqrt{2}, x_1 \leq -1/\sqrt{2}$. Notice that in either case, $\Gamma = 1/\sqrt{2}$. With these constraints, the simulation is run for $d \in [2:10]$. In the varying Γ scenario, we fix $d = 4$, and impose the constraints $x_1 \geq 1/\sqrt{2} - \Gamma, x_2 \geq 1/\sqrt{2} - \Gamma$ for the feasible setting, and the constraints $x_1 \geq \Gamma, x_1 \leq -\Gamma$ in the infeasible case. The range $\Gamma \in [0.2, 1]$ is studied at a grid of scale 0.1.

Throughout, the feedback noise is independent Gaussian with standard deviation $\sigma = 0.1$ (the value of σ is used in the confidence radii, and in general, τ should be proportional to σ^2). The parameter δ is set to 0.1, $N = 1$, and all results are averaged over 50 runs. The code was implemented in MATLAB, and executed on a consumer grade Ryzen 5 CPU, with no multithreading, and took about 4 hours to run.

Observations As a basic observation, we find that in *all* runs, the test returns the correct hypothesis. Notice that this suggests that the testing boundary is overly conservative, and a finer analysis of the same

[6] For nonzero α, the objective is modified to $\|\pi^\top \tilde{A}\|_2 - \pi^\top \alpha$, and the final minimisation to discover i_t then studies $(\tilde{A}_t x_t - \alpha)^i$.

is thus of interest. The main observation of Figure 2 is that for feasible instances τ_{early} is typically $< \tau/10$, across all dimensions d and signal level Γ studied, indicating that this early stopping is very powerful. While the validity of stopping at time τ_{early} is easy to see from the consistency of confidence sets, nothing in our analysis indicates the sample advantage of this procedure, and the resolving this is a natural open question.

6 Discussion

The feasibility testing problem is a natural first step prior to executing constrained bandit methods, and by initiating the study of the same, our work extends the applicability of this emerging field. We presented simple tests based on existing technology of online linear regression and LILs that are effective for such problems, and further pointed out key deficiencies in the extant work on the single-constraint finite-armed theory of this problem. Naturally, this is only a first step: the real power of the finite-armed theory, and in particular the tests proposed therein, is its strong adaptation to the explicit structure of the instance at hand. A parallel theory, both in the small and moderate δ regimes, in the linear setting is critical to develop efficient tests. Naturally, the computational question of how one can implement such tests efficiently is also critical. We hope that our work will spur study on these interesting and important issues.

Acknowledgements

Aditya Gangrade was supported on AFRLGrant FA8650-22-C1039 and NSF grants CCF-2007350 and CCF-2008074. Aditya Gopalan acknowledges partial support from Sony Research India Pvt. Ltd. under the sponsored project 'Black-box Assessment of Recommendation Systems'. Clayton Scott was supported in part by NSF Grant CCF-2008074. Venkatesh Saligrama was supported by the Army Research Office Grant W911NF2110246, AFRLGrant FA8650-22-C1039, the National Science Foundation grants CCF-2007350 and CCF-1955981.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011. 5, 6, 9, 17

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29:3450–3458, 2016. 3

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014. 3

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*, 2019. 3

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013. 3

Balsubramani, A. and Ramdas, A. Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*, 2015. 3

Camilleri, R., Wagenmaker, A., Morgenstern, J. H., Jain, L., and Jamieson, K. G. Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214, 2022. 3, 9

Chen, T., Gangrade, A., and Saligrama, V. Doubly optimistic play for safe linear bandits. *arXiv preprint arXiv:2209.13694*, 2022. 3

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008. 9, 12

Degenne, R. and Koolen, W. M. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019. 3

Degenne, R., Koolen, W. M., and Ménard, P. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019. 3

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021. 6

Jourdan, M. and Réda, C. An anytime algorithm for good arm identification. *arXiv preprint arXiv:2310.10359*, 2023. 3

Juneja, S. and Krishnasamy, S. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory*, pp. 1824–1852. PMLR, 2019. 3

Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., and Sugiyama, M. Good arm identification via bandit feedback. *arXiv preprint arXiv:1710.06360*, 2017. 3

Katz-Samuels, J. and Scott, C. Top feasible arm identification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1593–1601. PMLR, 2019. 3

Kaufmann, E., Koolen, W. M., and Garivier, A. Sequential test for the lowest mean: From Thompson to Murphy sampling. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 3, 11, 12

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020. 2, 4, 5, 30, 32

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5. 5

Locatelli, A., Gutzeit, M., and Carpentier, A. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pp. 1690–1698. PMLR, 2016. 3

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. Safe linear Thompson sampling with side information. *IEEE Transactions on Signal Processing*, 2021. 3

Nathan, D. M. and DCCT/EDIC Research Group. The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: overview. *Diabetes care*, 37(1):9–16, 2014. 1

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021. 3

Papadimitriou, C. H. and Steiglitz, K. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998. 2

Qiao, G. and Tewari, A. An asymptotically optimal algorithm for the one-dimensional convex hull feasibility problem. *arXiv preprint arXiv:2302.02033*, 2023. 3

Simchi-Levi, D., Zheng, Z., and Zhu, F. Regret distribution in stochastic bandits: Optimal trade-off between expectation and tail risk. *arXiv preprint arXiv:2304.04341*, 2023. 9, 10, 22, 23, 24

Simchowitz, M., Jamieson, K., and Recht, B. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pp. 1794–1834. PMLR, 2017. 3, 11, 28, 30

Tabata, K., Nakamura, A., Honda, J., and Komatsuzaki, T. A bad arm existence checking problem: How to utilize asymmetric problem structure? *Machine learning*, 109(2):327–372, 2020. 3

Wang, Z., Wagenmaker, A. J., and Jamieson, K. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 9114–9146. PMLR, 2022. 3

A Tools from the Theory of Online Linear Regression and Linear Bandits

As is standard in the setting of linear bandits, we shall exploit tools from the theory of online linear regression to enable learning and exploration. The main tool we use is Lemma 5, stated previously in the main text, which asserts that the confidence sets $\mathscr{C}_t$ are consistent with high probability, and control the deviations of $\tilde{A}x - Ax$ for $\tilde{A} \in \mathscr{C}_t$ to the level $\rho_t(x;\delta)$ if $A \in \mathscr{C}_t(\delta)$. The latter result is almost trivial: by the triangle and Cauchy-Schwarz inequalities, for any $\tilde{A} \in \mathscr{C}_t(\delta), i \in [1:m]$,

$$|(\tilde{A}-A)x)^i| \le |(((\tilde{A}-\hat{A}_t)x)^i| \le 2 \sup_{\tilde{A}\in\mathscr{C}_t(\delta)} |(\tilde{A}^i-\hat{A}_t)^\top x| \le \sup_{\tilde{A}\in\mathscr{C}_t(\delta)} |\tilde{A}^i-\hat{A}_t^i\|_{V_t}\|x\|_{V_t^{-1}} \le \omega_t(\delta)\|x\|_{V_t^{-1}} = \rho_t(x;\delta),$$

where the final inequality is by definition of the confidence set $\mathscr{C}_t(\delta)$.

The principal way to use this bound is through the following generic control on the behaviour of $\det V_t$ and on $\sum_{s\le t}\rho_s(x_s;\delta)$. We again refer to Abbasi-Yadkori et al. (2011), although the result is older. See their paper for a historical discussion.

Lemma 14. *For any sequence of actions $\{x_t\} \subset \{\|x\| \le 1\}$, and any $t \ge 0$, it holds that*

$$\log\det V_{t+1} \le \sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \le 2\log\det V_{t+1} \le 2d\log(1+(t+1)/d).$$

As a consequence,

$$\sum_{s\le t}\rho_s(x_s;\delta)^2 \le 2\omega_t(\delta)^2 d\log(1+(t+1)/d) \le 3d^2\log^2(1+(t+1)/d) + 6d\log(1+(t+1)/d)(1+\log(m/\delta)),$$

and

$$\sum_{s\le t}\rho_s(x_s;\delta) \le \sqrt{t\sum\rho_s(x_s;\delta)^2} \le \sqrt{2dt\log(1+(t+1)/d)\omega_t(\delta)^2}.$$

We will also find it useful to state the consistency of the confidence set in the following dual way

Lemma 15. *For any sequence of actions $\{x_t\}$, and any $v > 0$, it holds that*

$$\mathbb{P}\left(\exists t, i : \|\hat{A}_t^i - A^i\|_{V_t} \ge 1 + \sqrt{\frac{d}{4}\log\left(1+\frac{t}{d}\right) + \frac{1}{2}\log m + \frac{v}{2}}\right) \le \exp(-v).$$

Proof. Since, by the first statement of Lemma 14, $\log\det V_t = \log\det V_{(t-1)+1} \le d\log(1+t/d)$, it follows that

$$\omega_t(\delta) = 1 + \sqrt{\frac{1}{2}\log\frac{m}{\delta} + \frac{1}{4}\log\det V_t} \le 1 + \sqrt{\frac{d}{4}\log(1+t/d) + \frac{1}{2}\log m + \frac{1}{2}\log(1/\delta)} =: \tilde{\omega}_t(\delta).$$

Now the claim follows by just noting that

$$\mathbb{P}(\exists t, i : \|A^i - \hat{A}_t^i\|_{V_t} \ge \tilde{\omega}_t(\delta)) \le \mathbb{P}(\exists t, i : \|A^i - \hat{A}_t^i\| \ge \omega_t(\delta)) \le \delta,$$

and inverting the form of the upper bound obtained after expressing $\tilde{\omega}_t(\delta)$ as we have above. □

B Analysis of EOGT.

We will proceed to flesh out the analysis sketched in §3.3.1, and show the relevant results.

B.1 Adpting the LIL to the Noise Process of EOGT, and Control on the Rejection Timescale Bound.

We begin arguing the following simple observation that extends the LIL to our situation.

Lemma 16. *For i_t as chosen in EOGT or T-EOGT, it holds that $\{\eta_t^{i_t}\}$ forms a conditionally centred and 1-subGaussian process with respect to the filtration generated by $\{(i_s, x_s, S_s)\}_{s\le t} \cup \{(x_t, i_t)\}$. Therefore, for $Z_t := \sum_{s\le t} \zeta_s^{i_s}$, and any $\delta \in (0,1)$, it holds that $\mathbb{P}(\exists t : |Z_t| > \mathrm{LIL}(t,\delta)) \le \delta$.*

Proof. We simply observe that (x_t, i_t) are predictable given $\mathsf{H}_{t-1} = (\{(x_s, S_s)\}_{s\le t-1})$. Thus, the sigma algebra generated by $\{(x_s, i_s, S_s)_{s\le t} \cup \{(x_t, i_t)\}$ is the same as that generated by H_{t-1}, and ζ_t is assumed to be conditionally centred and 1-subGaussian with respect to this filtration, and thus its predictable projection $\zeta_t^{i_t}$ inherits this property. The second claim is then immediate from Lemma 6. □

We further add the proof of the upper bound on $T(\Gamma;\delta,N)$, which bounds the timescale of rejection for EOGT.

Proof of Lemma 9. We note that we shall make no efforts to optimise the constants in the following argument. Recall that

$$T(\Gamma;\delta,N) = \inf\left\{ t \ge 2d : t|\Gamma| > 2\mathrm{LIL}(t,\delta/2) + 4d\log(2t/d)\sqrt{t} + 2\sqrt{dt\log(2t/d)\log\frac{2m}{\delta t^{-N}}} \right\}.$$

Now, if $t \ge \max(50, 2d)$, then

$$\begin{aligned}\frac{2\mathrm{LIL}(t,\delta/2)}{\sqrt{t}} &= 4\sqrt{\log(11\log t) + \log\frac{2}{\delta}} < 4\sqrt{\log t + \log\frac{2}{\delta}} \\ &\le 4\sqrt{N\log t + \log(2m/\delta)} \\ &\le 4\sqrt{d\log(2t/d)\log(2m/\delta t^{-N})},\end{aligned}$$

where we have used $N \ge 1$, that $\log(11\log(u)) < \log(u)$ for $u \ge 50$, and that $d\log(2t/d) > 1$ when $2t/d > 4 > e$. Thus absorbing the LIL term into the last term defining T, we conclude that

$$\begin{aligned}T(\Gamma;\delta,N) &\le \inf\left\{ t \ge \max(50,2d) : t|\Gamma| > 4d\sqrt{t}\log t + 6\sqrt{dt\log(2t/d)(\log(2m/\delta) + N\log t)} \right\} \\ &\le \inf\left\{ t \ge \max(50,2d) : t|\Gamma| > \max\left(12d\sqrt{t}\log t, 18\sqrt{dt\log(2t/d)\log(2m/\delta)}, 18\sqrt{dtN}\log t\right) \right\} \\ &\le \inf\left\{ t \ge \max(50,2d) : \frac{t}{\log^2 t} > \frac{12^2\max(d^2, 9/4 Nd)}{\Gamma^2} \text{ and } \frac{2t/d}{\log(2t/d)} > \frac{2\cdot 18^2\log(2m/\Gamma)}{\Gamma^2} \right\},\end{aligned}$$

where in the second step we used the facts that for $u, v, w \ge 0$, $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ and $(u+v+w) \le 3\max(u,v,w)$.

Now, we observe the following elementary properties.

1. The map $u \mapsto u/\log(u)$ is increasing for $u \ge 3$. Thus, if $t > 2z\log 2z$ for some $z \ge 1.5$ (which implies $2z\log 2z \ge 3$), then
$$\frac{t}{\log t} > \frac{2z\log(2z)}{\log 2z + \log\log(2z)} \ge z,$$
where we have used that $2z > 1$ for $z \ge 1.5$. Since $\frac{2\cdot 12^2\log(2m/\delta)}{\Gamma^2} > 2\cdot 12^2\cdot\log(2) > 1.5$,
$$2t/d > \frac{4\cdot 18^2\log(2m/\delta)}{\Gamma^2}\log\frac{4\cdot 18^2\log(2m/\delta)}{\Gamma^2} \implies \frac{2t/d}{\log(2t/d)} > \frac{2\cdot 18^2\log(2m/\delta)}{\Gamma^2}.$$

2. For $u > 1, v > 0$,

$$\frac{u}{\log^2 u} \geq v \iff \left(\frac{\sqrt{u}}{2\log\sqrt{u}}\right)^2 \geq v \iff \frac{\sqrt{u}}{\log\sqrt{u}} \geq \sqrt{4v}.$$

But, as detailed above, if $\sqrt{4v} > 3/2 \iff v > 9/16$, then it holds for any u such that

$$\sqrt{u} > 2 \cdot \sqrt{4v}\log(2 \cdot \sqrt{4v}) \iff u > 4v\log^2(16v).$$

Setting $u = t, v = \frac{12^2\max(d^2, 9/4Nd)}{\Gamma^2} > 9/16$, we conclude that

$$t > \frac{4 \cdot 12^2\max(d^2, 9/4Nd)}{\Gamma^2}\log^2\frac{16 \cdot 12^2\max(d^2, 9/4Nd)}{\Gamma^2} \implies \frac{t}{\log^2 t} > \frac{12^2\max(d^2, 9/4Nd)}{\Gamma^2}.$$

Incorporating the above analysis into the bound on $T(\Gamma; \delta, N)$, we conclude that

$$T(\Gamma; \delta, N) \leq \max\left(50, 2d, \frac{576\max(d^2, 9/4Nd)}{\Gamma^2}\log^2\frac{2304\max(d^2, 9/4Nd)}{\Gamma^2}, \frac{648d\log(2m/\delta)}{\Gamma^2}\log\frac{1296\log(2m/\delta)}{\Gamma^2}\right).$$

□

B.2 Signal growth under consistency of confidence sets, and reliability

The growth of $\mathscr{T}_t$ was detailed in the main text in §3.3.1, the only informal aspect of this section being the treatment of Z_t, which can be accounted for immediately using Lemma 16. Thus, we have already shown Lemma 7. As briefly mentioned in the main text, this immediately yields reliability.

Proposition 17. *EOGT is reliable.*

Proof. Suppose that $\mathcal{H}_\mathsf{F}$ is true, and the event of Lemma 7 holds. Then since $\tau = \inf\{t : |\mathscr{T}_t| > \mathscr{B}_t(\delta)\}$, and since $\mathscr{T}_t \geq -\mathscr{B}_t(\delta)$, it follows that upon stopping, $\mathscr{T}_\tau > \mathscr{B}_t(\delta)$. Since $\mathscr{D}(\mathsf{H}_\tau) = \mathcal{H}_\mathsf{F}$ if $\mathscr{T}_\tau > 0$, it follows that this decision is correct. Hence, the only way for the decision to be incorrect is if $\exists t : \mathscr{T}_t < t\Gamma - \mathscr{B}_t(\delta)$, which can occur with probability at most δ. The same argument can be repeated mutatis mutandis for $\mathcal{H}_\mathsf{I}$. □

B.3 Control on the Stopping Time of EOGT in Mean and Tails

We shall prove the stronger result, Theorem 10. Note that expectation result follows from this directly.

Proof of Theorem 8 assuming Theorem 10. The reliability has already been shown in Proposition 17. To control the expectation, let us define, for naturals $k \geq 2$, $T_k = T(\Gamma/2; \delta/2^{k^3-1}; N) + \lceil 2^{(k^3-1)/N}\rceil(2 + 1/|\Gamma|)$, and define $T_1 = T(\Gamma; \delta, N)$. Then by Theorem 10, $\mathbb{P}(\tau > T_k) \leq 2^{-(k^3-1)}\delta$. As a consequence,

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{t\geq 0}\mathbb{P}(\tau > t) \\
&\leq \sum_{t\leq T_1} P(\tau > t) + \sum_{k=2}^{\infty}\sum_{t\in[T_{k-1}+1:T_k]}\mathbb{P}(\tau > t) \\
&\leq T_1 + 1 + \sum_{k=2}^{\infty}\delta 2^{1-(k-1)^3}(T_k - T_{k-1}) \leq T_1 + 1 + \delta\sum_{k=2}^{\infty}T_k 2^{1-(k-1)^3}.
\end{aligned}$$

To control the above, we shall show that $T(\Gamma; \delta^{k^3}, N)$ is bounded from above by $k^6 T(\Gamma; \delta, N)$ for $k \geq 2$. To this end, recall that

$$T(\Gamma; \eta, N) = \inf\left\{t \geq 2d : t|\Gamma| > 4\sqrt{t\log\log t + t\log\frac{22m}{\delta}} + 4d\sqrt{t}\log(2t/d) + \sqrt{2dt\log(2t/d)\log(2m/\delta t^{-N})}\right\}.$$

Now, first observe that if $t \geq 16$, and $k \geq 2$, then $\log\log(k^6 t) \leq k^6 \log\log t$. Indeed, if $k \geq t$, then[7] $\log\log(k^6 t) \leq \log\log(k^7) \leq k < k^6$. If instead $k \leq t$, then $\log(7) < 2 < (k^6 - 1) \implies \log\log t^7 = \log 7 + \log\log t < k^6 - 1 + \log\log t < k^6 \log\log t$, which exploits that $\log\log t > 1$ for $t \geq 16 > e^e$.

Next, if $t \geq 2d$, and $k \geq 2$, then $\log(2k^6 t/d) \leq k^3 \log(2t/d)$. Again, if $2t/d < k$, then[8] $\log(k^7) < 7(k-1) < k^3 \log(4) < k^3 \log(2t/d)$, and if $2t/d \geq k$, then $7\log(2t/d) < k^3 \log(2t/d)$ since $k^3 \geq 8$. Similarly, if $k \geq 2, t \geq 16$ then $\log(k^6 t) \leq k^3 \log t$.

It follows from the above that if $t \geq \max(2d, 16)$, and $t \geq T(\Gamma; \delta, N)$, then $k^6 t \geq T(\Gamma; \delta/2^{k^2-1}, N)$. Indeed, since $t > T(\Gamma; \delta, N)$, we have

$$t|\Gamma| > 4\sqrt{t\log\log t + \log(22m/\delta)} + 4d\sqrt{t}\log(2t/d) + \sqrt{(2d\log(2t/d)(\log(2m/\delta) + N\log t)}.$$

Multiplying through by k^6, and using $t \geq \max(2d, 16)$, we observe that

$$\begin{aligned}
k^6 t|\Gamma| &> 4\sqrt{k^6 t(k^6\log\log t + k^6 \log\frac{22m}{\delta})} + 4d\sqrt{k^6 t}\cdot k^3\log(2t/d) \\
&\qquad + \sqrt{2d(k^6 t)\cdot k^3\log(2t/d)(k^3\log(2m/\delta) + 2Ndk^3\log t)} \\
&\geq 4\sqrt{(k^6 t)\log\log(k^6 t) + \log\frac{22m}{\delta^{k^6}}} + 4d\sqrt{k^6 t}\log(2k^6 t/d) \\
&\qquad + \sqrt{2d(k^6 t)\log(2k^6 t/d)(\log(2m/\delta^{k^3}) + 2Nd\log(k^6 t))},
\end{aligned}$$

where we have used that $m \geq 1$. Since $\delta \leq 1/2$,

$$\delta^{k^6} \leq \delta^{k^3} = \delta\cdot\delta^{k^3-1} \leq \delta\cdot 2^{-(k^3-1)}.$$

Thus, we conclude that for $k \geq 2$,

$$T_k - \lceil 2^{(k^3-1)/N}\rceil(2 + 1/|\Gamma|) = T(\Gamma/2; \delta/2^{k^3-1}, N) \leq \max(2d, 16, k^6 T(\Gamma/2; \delta; N)).$$

Plugging this into the bound on $\mathbb{E}[\tau]$, we conclude using numerical estimates of the quickly converging series $\sum_{k\geq 2} 2^{1-(k-1)^3} \leq 1.01$ and $\sum_{k\geq 2} k^6 2^{1-(k-1)^3} \leq 70$ that

$$\begin{aligned}
\mathbb{E}[\tau] &\leq T_1 + 1 + \delta\sum_{k\geq 2} 2^{1-(k-1)^3} T_k \\
&\leq T_1 + 1 + \delta\sum_{k\geq 2} 2^{1-(k-1)^3}(2d+16) + \delta T(\Gamma/2;\delta,N)\sum_{k\geq 2} k^6 2^{1-(k-1)^3} \\
&\qquad + \delta(2+1/|\Gamma|)\left(\sum_{k\geq 2} 2^{-((k-1)^3 - 1 - (1-1/N)k^3)} + 2^{1-(k-1)^3}\right) \\
&\leq T(\Gamma;\delta,N) + 1 + 70\delta T(\Gamma/2;\delta,N) + (20+3d)\delta + O(1)\delta(3+1/|\Gamma|),
\end{aligned}$$

where the $O(1)$ term is $\leq 1.01 + \sum_{k\geq 2} 2^{1-(k-1)^3 + (k^3(1-1/N))}$, which is summable since $N > 1$. □

Let us now proceed with the

Proof of Theorem 10. First notice by Lemma 14, if $t \geq 2d$, then

$$\sum_{s\leq t}\rho_s(x_s;\delta_s/2) \leq \sum_{s\leq t}\rho_s(x_s;\delta_t/2) \leq \omega_t(\delta_t/2)\sqrt{2dt\log(1+(t+1)/d)} \leq \omega_t(\delta_t/2)\sqrt{2dt\log(2t/d)}.$$

[7] $\log\log k^7 = \log 7 + \log\log k \leq \log 7 + \log k - 1 \leq \log 7 - 2 + k$, and $e^2 > 7.3$.

[8] $k^3 - 7k + 7$ is growing for $k \geq \sqrt{7/3} \approx 1.52$, and $2^3 - 14 + 7 = 1 > 0$. Of course, $\log(4) > 1$.

Consequently, we have that for $t \geq 2d$,

$$\mathscr{B}_t(\delta) \leq \omega_t(\delta_t/2)\sqrt{2dt\log(1+(t+1)/d)} \leq \omega_t(\delta_t/2)\sqrt{2dt\log(2t/d)}.$$

If $\mathcal{H}_{\mathsf{F}}$ is true, then we know by Lemma 7 that with probability at least $1-\delta$,

$$\forall t, \mathscr{T}_t \geq t\Gamma - \mathscr{B}_t(\delta),$$

and so we conclude that under this event,

$$\tau = \inf\{t : t\Gamma > 2\mathscr{B}_t(\delta)\}.$$

But due to the deterministic upper bound on $\mathscr{B}_t(\delta)$ under the same event,

$$\tau \leq \inf\{t : t\Gamma > 2\omega_t(\delta_t/2)\sqrt{2dt\log(2t/d)} + 2\mathrm{LIL}(t;\delta/2)\}.$$

But, for $t \geq 2d, N > 1$, $\omega_t(\delta_t/2) \leq 1 + \sqrt{\frac{1}{2}\log(2m/\delta t^{-N}) + \frac{d}{4}\log(2t/d)}$, and so,

$$\begin{aligned}
2\omega_t(\delta_t/2)\sqrt{2dt\log 2t/d} &\leq 2\sqrt{2dt\log(2t/d)} + 2\sqrt{\frac{d^2}{2}t\log^2(2t/d)} + 2\sqrt{(dt\log(2t/d))(\log(2m/\delta t^{-N}))} \\
&\leq (\sqrt{8/\log(4)d} + \sqrt{2}d)\sqrt{t}\log(2t/d) + 2\sqrt{(d\log(2t/d))(\log(2m/\delta t^{-N}))} \\
&< 4d\sqrt{t}\log(2t/d) + 2\sqrt{(d\log(2t/d))(\log(2m/\delta t^{-N}))},
\end{aligned}$$

where the first line uses $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$, and the final line uses the fact that $t \geq 2d \implies \log(2t/d) \geq \log(4)$, and that for $u \geq 1$, $\sqrt{8u/\log 4} + \sqrt{2}u < 4u$. But this implies that

$$\tau \leq \inf\left\{t : t|\Gamma| > 2\mathrm{LIL}(t,\delta/2) + 4d\sqrt{t}\log(2t/d) + 2\sqrt{d\log(2t/d)\log\frac{2m}{\delta t^{-N}}}\right\} = T(\Gamma;\delta,N).$$

In fact, this is precisely why $T(\Gamma;\delta,N)$ was so defined. Thus, in the feasible case, with probability at least $1-\delta, \mathbb{P}(\tau > T(\Gamma;\delta,N)) \leq \delta$. The argument is identical in the infeasible case, barring sign flips.

Control on the tail can be obtained by essentially bootstrapping the above result along with our choice of $\mathscr{D}_t = \mathscr{C}_t(\delta_t)$, the key idea being that since $\delta_t \to 0$, for large enough $t, A \in \mathscr{D}_t$ must actually occur with near-certainty. Formally, let us define $T_\eta = \inf\{t : \delta_t < \eta\} = \lceil(\delta/\eta)^{1/N}\rceil$. Then notice that for every $t \geq T_\eta$, it holds that $\mathscr{D}_t \subset \mathscr{C}_t(\eta)$, and so $\mathbb{P}(\forall t \geq T_\eta, A \in \mathscr{D}_t) \geq 1-\eta$. Therefore, repeating the proof of Lemma 7, we conclude that in the feasible case, for all $t \geq T_\eta$,

$$\mathscr{T}_t \geq -T_\eta + (t - T_\eta)\Gamma - \sum_{T_\eta \leq s \leq t} \rho_s(x_s;\delta_s) - \mathrm{LIL}(t,\eta/2),$$

where we have used the fact that $\|x\| \leq 1, \|A^i\| \leq 1$ to conclude that $|(Ax)^{i_t}| \leq 1$ in order to handle the times $t \in [1 : T_\eta - 1]$. In particular, if $t > 2T_\eta + T_\eta/\Gamma$, then $\mathscr{T}_t \geq t\frac{\Gamma}{2} - \mathscr{B}_t(\eta)$.

But we know that we must stop before time t if $\mathscr{T}_t \geq \mathscr{B}_t(\delta)$, and since $\mathscr{B}_t(\delta) \leq \mathscr{B}_t(\eta)$ uniformly, we conclude that under the event that $A \in \mathscr{D}_t$ for all $t \geq T_\eta$, then it must hold that

$$\tau \leq \max\left((2 + 1/\Gamma)T_\eta, T(\Gamma/2,\eta,N)\right).$$

Since this occurs with probability at least $1-\eta$, the conclusion follows for the feasible case. Again, the argument is identical for the infeasible case, barring sign flips. □

C Analysis of T-EOGT

The main result follows simply from the key control offered in Lemma 11, and showing the latter will form the bulk of this section. We proceed by first showing the stopping time bounds.

Proof of Theorem 12. Let us consider the feasible case; the infeasible case follows similarly. For reliability, observe that via Lemma 11, it holds with probability at least $1-\delta$ that for all t,

$$\widetilde{\mathscr{T}_t} \geq t\Gamma - \mathscr{Q}_t^{\mathsf{F}}(\delta) > -\mathscr{Q}_t^{\mathsf{F}}(\delta).$$

Since the stopping time is

$$\widetilde{\tau} = \inf\{t : \widetilde{\mathscr{T}_t} < -\mathscr{Q}_t^{\mathsf{F}}(\delta) \text{ or } \widetilde{\mathscr{T}_t} > \mathscr{Q}_t^{\mathsf{I}}(\delta)\},$$

it follows that if the preceding event occurs, then if the test stops, it must be correct. But, since $\mathscr{Q}^{\mathsf{I}} + \mathscr{Q}^{\mathsf{F}}$ grows sublinearly in t, under the same event the test must eventually stop. Therefore, the probability that we stop and make an error is bounded by δ, making the test reliable.

It remains to control the behaviour of $\widetilde{\tau}$. To this end, again observe that for any $\eta \in (0,1)$, with probability at least $1-\eta$, it holds for all time that

$$\widetilde{\mathscr{T}_t} \geq t\Gamma - \mathscr{Q}_t^{\mathsf{F}}(\eta).$$

Thus, we conclude that with probability at least $1-\eta$,

$$\tau \leq \inf\{t : t\Gamma \geq \mathscr{Q}_t^{\mathsf{F}}(\eta) + \mathscr{Q}_t^{\mathsf{I}}(\delta)\} \leq T_\eta := \inf\{t : t\Gamma \geq \mathscr{Q}_t^{\mathsf{F}}(\eta) + \mathscr{Q}_t^{\mathsf{I}}(\eta)\}.$$

But notice that

$$\mathscr{Q}_t^{\mathsf{F}}(\eta) + \mathscr{Q}_t^{\mathsf{I}}(\eta) \leq 50t^{1/2}\log^2(t)\left(d^{3/2} + d^{1/2}\log(8m/\eta)\right) + 2\mathrm{LIL}(t,\eta/2).$$

Following the approach in the proof of Lemma 9 as presented in §B.1,[9] we immediately get that there exists a constant C such that with probability at least $1-\eta$,

$$\tau \leq \frac{C\log(C\log(\Gamma^{-2})/\delta)}{\Gamma^2} + \frac{Cd^3}{\Gamma^2}\log^4\frac{Cd^3}{\Gamma^2} + \frac{Cd\log(8m/\eta)}{\Gamma^2}\log^4\frac{d\log(8m/\eta)}{\Gamma^2}.$$

The expectation bound is immediate upon integrating the tail. ∎

It remains then to show Lemma 11, which is the subject of the next section.

C.1 Proof of Anytime Behaviour of $\widetilde{\mathscr{T}_t}$

We begin with setting up some notation, and then proceed by explicitly describing key observations underlying the argument, encapsulated as lemmata. The key aspects of this argument follow the analysis of Simchi-Levi et al. (2023).

C.1.1 Notation

Let (x^*, i^*) denote any solution to the program $\max_x \min_i (Ax)^i$, which we shall fix for the remainder of this section. Of course, $(Ax^*)^{i^*} = \Gamma$. Recall that $i_{\min}(x) = \arg\min_i (Ax)^i$. We further define

$$i_t(x) = \arg\min_i (\hat{A}_t x)^i, \text{ and } i_t^* = i_t(x^*).$$

[9] the only new information needed being that $4\log\log z \leq \log z$ for all $z \geq 2$

We denote the estimation error in $\hat{A}_t$ as

$$B_t = \hat{A}_t - A.$$

Next, we define the random quantity

$$\Delta_t = (\Gamma - (Ax_t)^{i_t})\text{sign}(\Gamma) = \begin{cases} \Gamma - (Ax_t)^{i_t} & \text{if } \Gamma > 0, \text{ i.e., under feasibility} \\ (Ax_t)^{i_t} - \Gamma & \text{if } \Gamma < 0, \text{ i.e., under infeasibility.} \end{cases},$$

and the cumulative pseduoregret-like object

$$\mathscr{R}_t = \sum_{s \le t} \Delta_s.$$

The point here is that we may decompose

$$\widetilde{\mathscr{T}_t} = \sum_{s \le t} (Ax_s)^{i_s} + Z_t = t\Gamma - \mathscr{R}_t + Z_t, \quad \textit{(feasible case)}$$
$$\widetilde{\mathscr{T}_t} = \sum_{s \le t} (Ax_s)^{i_s} + Z_t = t\Gamma + \mathscr{R}_t + Z_t, \quad \textit{(infeasible case)}$$

and thus in either case, if we show that $\mathscr{R}_t$ is not too large, then $\widetilde{\mathscr{T}_t}$ has favourable behaviour. Observe that if we were working in a single objective setting, $m = 1$, then in the feasible case $\mathscr{R}_t$ would be the pseudoregret of a linear bandit instance.

Since these quantities will appear often in the argument, we further define

$$N_t = \|x_t\|^2_{V_t^{-1}} \textit{ and } N_t^* = \|x^*\|^2_{V_t^{-1}},$$

and for $v \ge 0$,

$$\mathscr{W}_t(v) := 1 + \sqrt{\frac{d}{4}\log(1 + t/d) + \frac{1}{2}\log m + \frac{v}{2}}$$

Finally, notice that with the above notation, Lemma 15 can be expressed as

$$\forall v > 0, \mathbb{P}\left(\exists t, i : \|B_t^i\|_{V_t} \ge \mathscr{W}_t(v)\right) \le e^{-v}.$$

Further,

$$\text{Rad}_t(x_t) = (t/d)^{1/2} N_t + \sqrt{dN_t}, \textit{ and } \text{Rad}_t(x^*) = (t/d)^{1/2} N_t^* + \sqrt{dN_t^*}.$$

C.1.2 Structural Observations

The following two results constitute basic structural observations due to Simchi-Levi et al. (2023) that enable the subsequent analysis. The first argues that in each round, some quantity of the form $(B_t x)^i$ for some (x, i) is large in absolute value.

Lemma 18. *For the sequence of actions $\{x_t\}$ selected by* T-EOGT*, the following hold.*

- *In the feasible case, at each time, either the first or the second of the following hold:*

$$(B_t x_t)^{i_t} \ge \Delta_t/2 - (t/d)^{1/2} N_t - \sqrt{dN_t}$$
$$\textit{or} \quad -(B_t x^*)^{i_t^*} \ge \Delta_t/2 + (t/d)^{1/2} N_t^* + \sqrt{dN_t^*}$$

- *In the infesible case, at each time t, either the first or the second of the following hold:*

$$-(B_t x_t)^{i_t} \ge \Delta_t/2$$
$$\textit{or} \quad (B_t x_t)^{i_{\min}(x_t)} \ge \Delta_t/2.$$

Proof. In the feasible case, due to the optimistic selection, it must hold that

$$(\hat{A}_t x_t)^{i_t} + \mathrm{Rad}_t(x_t) \geq (\hat{A}_t x^*)^{i_t^*} + \mathrm{Rad}_t^*.$$

Now, we may write $\hat{A}_t = A + B_t$, and so get

$$(B_t x_t)^{i_t} + \mathrm{Rad}(x_t) \geq \left((Ax^*)^{i_t^*} - (Ax_t)^{i_t}\right) + \mathrm{Rad}_t(x^*).$$

But note that $(Ax^*)^{i_t^*} \geq \min_i (Ax^*)^i = \Gamma$, and so $(Ax^*)^{i_t^*} - (Ax_t)^{i_t} \geq \Delta_t$ in the feasible case. Thus, we have

$$(B_t x_t)^{i_t} + \mathrm{Rad}(x_t) \geq \Delta_t + (B_t x^*)^{i_t^*} + \mathrm{Rad}_t(x^*).$$

But, since if $A \geq B + C$, then either $A \geq B/2$ or $-C \geq B/2$, it follows that at least one of the following must hold:

$$(B_t x_t)^{i_t} \geq \Delta_t/2 - \mathrm{Rad}_t(x_t) \text{ or } -(B_t x^*)^{i_t^*} \geq \Delta_t/2 + \mathrm{Rad}_t(x^*).$$

The conclusion follows upon incorporating the form of $\mathrm{Rad}_t(x_t)$ and $\mathrm{Rad}_t(x^*)$ indicated before the statement of the lemma.

In the infesible case, we note that it must hold that

$$(\hat{A}_t x_t)^{i_t} \leq (\hat{A}_t x_t)^{i_{\min}(x_t)} \iff (B_t x_t)^{i_t} - (B_t x_t)^{i_{\min}(x_t)} \geq (Ax_t)^{i_t}) - (Ax_t)^{i_{\min}(x_t)}.$$

But, $(Ax_t)^{i_{\min}(x_t)} = \min_i (Ax_t)^i \leq \max_x \min_i (Ax)^i = \Gamma$, and so noting that $\Delta_t = (Ax_t)^{i_t} - \Gamma$ in the infeasible case, we have

$$(B_t x_t)^{i_t} - (B_t x_t)^{i_{\min}(x_t)} \geq \Delta_t,$$

which again yields the conclusion. ☐

The next observation essentially yields a condition for low $\mathscr{R}_t$ in terms of (Δ_t, N_t), and forms a refinement of the key observation of Simchi-Levi et al. (2023) that allows us to extend their results to yield anytime bounds.

Lemma 19. *For any nondecreasing sequence of positive reals u_t, it holds that*

$$\{\exists t : \mathscr{R}_t > u_t(1 + \log(t+1))\} \subset \{\exists t : \Delta_t \geq u_t/3t, N_t < d/t\} \cup \{\exists t : \Delta_t/N_t \geq u_t/3d, N_t \geq d/t\}.$$

Proof. Suppose that for all t, $N_t < d/t \implies \Delta_t < u_t/3t$ and $N_t \geq d/t \implies \Delta_t/N_t < u_t/3d$. Then

$$\begin{aligned}
\mathscr{R}_t = \sum_{s \leq t} \Delta_s &= \sum_{s \leq t} \Delta_s \mathbb{1}\{N_s < d/t\} + \sum_{s \leq t} \frac{\Delta_s}{N_s} \cdot N_s \mathbb{1}\{N_s \geq d/t\} \\
&< \sum_{s \leq t} u_s/3s + \sum_{s \leq t} \frac{u_s}{3d} N_s \\
&\leq \frac{u_t}{3} \sum_{s \leq t} 1/s + \frac{u_t}{3d} \sum_{s \leq t} N_s \\
&\leq \frac{u_t(\log(t)+1)}{3} + \frac{u_t}{3d} \cdot 2d \log(1 + t/d) \\
&\leq u_t(1 + \log(t+1)),
\end{aligned}$$

where the second inequality is because $u_s \leq u_t$ for all $s \leq t$, and the third uses the bound on $\sum_{s \leq t} N_s = \sum_{s \leq t} \|x_s\|^2_{V_s^{-1}}$ from Lemma 14, and the standard bound on harmonic numbers $\sum_{s \leq t} 1/s \leq \log(t) + 1$. ☐

This sets up the basic approach: the two events in Lemma 18 along with the two events in Lemma 19 set up four potential ways that high $\mathscr{R}_t$ can arise in either the feasible or the infeasible case. We will separately bound the probabilities of these events by repeated reduction to the key result of Lemma 15.

C.1.3 Controlling the Chance of Poor Events

We now proceed to execute the strategy we described at the end of the previous section. We will separate the arguments for the feasilbe and the infeasible cases.

Feasible Case We shall further separate the analysis into two cases, depending on if we control the event with $|(B_t x_t)^{i_t}|$ being large, or $|(B_t x^*)^{i_t^*}|$ being large.

Lemma 20. *For any $v \geq 0$, define*

$$U_t^{\mathsf{F},\mathsf{A}}(v) := 6\sqrt{dt} + 6d\sqrt{t} + 6\sqrt{dt}\mathscr{W}_t(v).$$

Then both of the following inequalities hold true:

$$\mathbb{P}(\exists t : \Delta_t \geq U_t^{\mathsf{F},\mathsf{A}}(v)/3t, N_t < d/t, (B_t x_t)^{i_t} \geq \Delta_t/2 - (t/d)^{1/2} N_t - \sqrt{dN_t}) \leq e^{-v},$$
$$\mathbb{P}(\exists t : \Delta_t/N_t \geq U_t^{\mathsf{F},\mathsf{A}}(v)/3d, N_t \geq d/t, (B_t x_t)^{i_t} \geq \Delta_t/2 - (t/d)^{1/2} N_t - \sqrt{dN_t}) \leq e^{-v}.$$

Proof. We argue the two inequalities using slightly different, but ultimatly similar approaches. The key observation we will need is that by the Cauchy-Schwarz inequality, and since $N_t = \|x_t\|^2_{V_t^{-1}}$, $|(B_t x_t)^{i_t}| = |(B_t^{i_t} V_t^{1/2} V_t^{-1/2} x_t)| \leq \|B_t^{i_t}\|_{V_t} \sqrt{N_t}$. Throughout, we will let u_t denote an arbitrary nondecreasing sequence, and derive the form of $U_t^{\mathsf{F},\mathsf{A}}$ at the end.

Case (i). Suppose $\Delta_t \geq u_t/3t$ and $N_t < d/t$. Then

$$\begin{aligned}
(B_t x_t)^{i_t} &\geq \frac{\Delta_t}{2} - \sqrt{\frac{t}{d}} N_t - \sqrt{dN_t} \\
&\geq \frac{u_t}{6t} - \sqrt{\frac{d}{t}} - \frac{d}{\sqrt{t}} \\
\implies \sqrt{N_t}\|B_t^{i_t}\|_{V_t} &\geq \frac{u_t - 6\sqrt{dt} - d\sqrt{t}}{6t} \\
\implies \|B_t^{i_t}\|_{V_t} &\geq \frac{u_t - 6\sqrt{dt} - 6d\sqrt{t}}{6\sqrt{dt}}.
\end{aligned}$$

Case (ii). If instead, $\Delta_t/N_t \geq u_t/3d$ and $N_t \geq d/t$, then

$$\begin{aligned}
(B_t x_t)^{i_t} &\geq \frac{\Delta_t}{2} - \sqrt{\frac{t}{d}} N_t - \sqrt{dN_t} \\
\iff (B_t x_t)^{i_t}/N_t &\geq \frac{\Delta_t}{2N_t} - \sqrt{\frac{t}{d}} - \sqrt{d/N_t} \\
\implies \|B_t^{i_t}\|_{V_t}/\sqrt{N_t} &\geq \frac{\Delta_t}{2N_t} - \sqrt{\frac{t}{d}} - \sqrt{d/N_t} \\
\implies \|B_t^{i_t}\|_{V_t} &\geq \frac{u_t}{6d} \cdot \sqrt{d/t} - 1 - \sqrt{t} \\
&= \frac{u_t - 6\sqrt{dt} - 6d\sqrt{t}}{6\sqrt{dt}}.
\end{aligned}$$

Now observe that due to the form of $U_t^{\mathsf{F},\mathsf{A}}$, it holds that

$$\frac{U_t^{\mathsf{F},\mathsf{A}}(v) - 6\sqrt{dt} - 6d\sqrt{t}}{6\sqrt{dt}} = \mathscr{W}_t(v),$$

and so we have

$$\mathbb{P}\left(\exists t : \|B_t^{i_t}\|_{V_t} \geq \frac{U_t^{\mathsf{F},\mathsf{A}}(v) - 6\sqrt{dt} - 6d\sqrt{t}}{6\sqrt{dt}}\right) \leq \mathbb{P}\left(\exists t, i : \|B_t^i\|_{V_t} \geq \mathscr{W}_t(v)\right),$$

and the claim follows by Lemma 15. ☐

Lemma 21. *For any $v \geq 0$, define*

$$U_t^{\mathsf{F},\mathsf{B}}(v) := \frac{3\sqrt{dt}}{2}(\mathscr{W}_t(v) - \sqrt{d})_+^2,$$

where $(z)_+^2 = (\max(z,0))^2$. Then it holds that

$$\mathbb{P}(\exists t : \Delta_t \geq U_t^{\mathsf{F},\mathsf{B}}(v)/3t, N_t < d/t, -(B_t x^*)^{i_t^*} \geq \Delta_t/2 + \sqrt{t/d}N_t^* + \sqrt{dN_t^*}) \leq e^{-v}$$
$$\mathbb{P}(\exists t : \Delta_t/N_t \geq U_t^{\mathsf{F},\mathsf{B}}(v)/3d, N_t \geq d/t, -(B_t x^*)^{i_t^*} \geq \Delta_t/2 + \sqrt{t/d}N_t^* + \sqrt{dN_t^*}) \leq e^{-v}$$

Proof. As in the proof of Lemma 20, let $u_t \geq 0$ be any sequence. Then observe that $\Delta_t/N_t \geq u_t/3d, N_t \geq d/t \implies \Delta_t \geq u_t/3t$. Further, by the AM-GM inequality,

$$\frac{u_t}{6t} + \sqrt{\frac{t}{d}}N_t^* \geq 2\sqrt{\frac{u_t}{6\sqrt{dt}}N_t^*}.$$

But, if $\Delta_t \geq u_t/3t$, then

$$-(B_t x^*)^{i_t^*} \geq \frac{u_t}{6t} + \frac{t}{d}N_t^* + \sqrt{dN_t^*} \geq 2\sqrt{\frac{u_t}{6\sqrt{dt}}N_t^*} + \sqrt{dN_t^*}$$
$$\implies \|B_t^{i_t^*}\|_{V_t} \geq \sqrt{\frac{2u_t}{3\sqrt{dt}}} + \sqrt{d}.$$

Now, $U_t^{\mathsf{F},\mathsf{B}}$ is chosen so that

$$\sqrt{\frac{2U_t^{\mathsf{F},\mathsf{B}}(v)}{3\sqrt{dt}}} + \sqrt{d} = \mathscr{W}_t(v),$$

therefore, both of the probabilities in the claim are bounded from above by $\mathbb{P}(\exists t, i : \|B_t^i\|_{V_t} \geq \mathscr{W}_t(v))$, and we may conclude using Lemma 15. ☐

Infeasible Case Turning now to the infeasible case, we have the somewhat simpler bound below.

Lemma 22. *For $v \geq 0$, let*

$$U_t^{\mathsf{I}}(v) := 6\sqrt{dt}\mathscr{W}_t(v).$$

It holds that

$$\mathbb{P}(\exists t, i : \Delta_t \geq U_t^{\mathsf{I}}(v)/3t, N_t < d/t, |(B_t x_t)^i| \geq \Delta_t/2) \leq e^{-v}$$
$$\mathbb{P}(\exists t, i : \Delta_t/N_t \geq U_t^{\mathsf{I}}(v)/3d, N_t \geq d/t, |(B_t x_t)^i| \geq \Delta_t/2) \leq e^{-v}$$

Proof. The argument is similar to that underlying Lemma 20. Let u_t be any positive real. Then *Case (i)* If $\Delta_t \geq u_t/3t, N_t < d/t$, then for any i,

$$|(B_t x_t)^i| \geq \Delta_t/2$$
$$\implies \|B_t^i\|_{V_t}\sqrt{N_t} \geq \frac{u_t}{6t}$$
$$\implies \sqrt{d/t}\|B_t^i\|_{V_t} \geq \frac{u_t}{6t} \iff \|B_t^i\|_{V_t} \geq \frac{u_t}{6\sqrt{dt}}.$$

Case (ii) If instead $\Delta_t \geq u_t/3d, N_t \geq d/t$, then note that

$$\Delta_t/\sqrt{N_t} = \Delta_t/N_t \cdot \sqrt{N_t} \geq \frac{u_t}{3d} \cdot \sqrt{d/t} = \frac{u_t}{3\sqrt{dt}},$$

and thus

$$|(B_t x_t)^i| \geq \frac{\Delta_t}{2} \implies \|B_t^i\|_{V_t} \geq \frac{u_t}{6\sqrt{dt}}.$$

Since $U_t^{\mathsf{I}}(v) = 6\sqrt{dt}\mathscr{W}_t(v)$, it again follows that either of the probaiblities in the claim are bounded by $\mathbb{P}(\exists t, i : \|B_t^i\|_{V_t} \geq \mathscr{W}_t(v))$, and we are done upon applying Lemma 15. □

C.2 Proof of Tail Bounds

We are now ready to prove the claim. We begin by summarising the previous section through the lemma below. Note that setting $m = 1$, the bound for the feasible instance yields an anytime regret bound for the tempered action selection rule (5) over linear bandit instances.

Lemma 23. *For any $\delta \in (0,1)$, the following hold for the actions of* T-EOGT

- *For any feasible instance,*

$$\mathbb{P}(\forall t, \mathscr{R}_t \leq \log(t+t) \cdot \max(U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta)), U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta))) \geq 1 - \delta/2.$$

- *For any infeasible instance,*

$$\mathbb{P}(\forall t, \mathscr{R}_t \leq U_t^{\mathsf{I}}(\log(8/\delta))(1 + \log(t+1))) \geq 1 - \delta/2.$$

Proof. In the feasible case, let $u_t := \max(U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta)), U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta))$. Since $\mathscr{W}_t$ is nondecreasing, and the $U_t^{\mathsf{F},\cdot}$ are defined as nondecreasing functions of $\mathscr{W}_t$, it follows that u_t is nondecreasing. By Lemma 19, it follows that

$$\mathbb{P}(\exists t : \mathscr{R}_t > U_t^{\mathsf{F}} \cdot (1 + \log(t+1))) \leq \mathbb{P}(\exists t : \Delta_t \geq u_t/3t, N_t < d/t) + \mathbb{P}(\exists t : \Delta_t/N_t \geq u_t/3d, N_t \geq d/t).$$

But since the events in Lemma 18 must occur with certainty, we have

$$\begin{aligned}\mathbb{P}(\exists t : \Delta_t \geq u_t/3t, N_t < d/t) \leq{}& \mathbb{P}(\exists t : \Delta_t \geq u_t/3t, N_t < d/t, (B_t x_t)^{i_t} \geq \Delta_t/2 - (t/d)^{1/2} N_t - \sqrt{dN_t}) \\ &+ \mathbb{P}(\exists t : \Delta_t \geq u_t/3t, N_t < d/t, (B_t x^*)^{i_t^*} \geq \Delta_t/2 + (t/d)^{1/2} N_t^* + \sqrt{dN_t^*}).\end{aligned}$$

But, since $u_t \geq U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta))$, by Lemma 20, the first term is at most $\delta/8$, and similarly since $u_t \geq U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta 0))$, by Lemma 21, the second term is at most $\delta/8$, controlling the above to $\delta/4$. Of course, the same argument may be repeated to bound $\mathbb{P}(\exists t : \Delta_t/N_t \geq u_t/3d, N_t \geq d/t)$, giving the first bound. The infeasible case follows the same template, but uses the alternate result in Lemma 18, and Lemma 22 to control probabilities instead. We omit the details. □

To concretise the bounds above, we next show an auxiliary lemma controlling the sizes of $U_t^{\mathsf{F},\mathsf{A}}, U_t^{\mathsf{F},\mathsf{B}}$ and U_t^{I}.

Lemma 24. *Suppose $\delta \leq 1/2$. Then*

$$\begin{aligned}U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta)) &\leq 12d\sqrt{t\log(t+1)} + 15\sqrt{dt\log(8m/\delta)} \\ U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta)) &\leq 2d^{3/2}\sqrt{t\log^2(t+1)} + 3\sqrt{dt}\log(8m/\delta) \\ U_t^{\mathsf{I}}(\log(8/\delta)) &\leq 6\sqrt{d^2t(1+\log(t+1)) + 2dt\log(8m/\delta)}\end{aligned}$$

Proof. First, we note that if $\delta \leq 1/2$, then $\frac{1}{2}\log(8/\delta) \geq 3\log(2) > 1$. Thus, we have

$$\begin{aligned}\mathscr{W}_t(\log(4/\delta)) &= 1 + \sqrt{\frac{d}{4}\log(1+t/d) + \frac{1}{2}(\log m + \log(8/\delta))} \\ &\leq 2\sqrt{\frac{d}{4}\log(1+t) + \frac{1}{2}\log\frac{8m}{\delta}} \\ &\leq \sqrt{d\log(t+1) + 2\log(8m/\delta)} \\ &\leq \sqrt{d\log(t+1)} + \frac{3}{2}\sqrt{\log(8m/\delta)}.\end{aligned}$$

Thus,

$$U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta)) \leq 6\sqrt{dt} + 6d\sqrt{t} + 6d\sqrt{t(1+\log(t+1))} + 9\sqrt{dt\log(8m/\delta)},$$

and further,

$$U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta)) \leq \frac{3\sqrt{dt}}{2} \cdot \sqrt{d\log(t+1) + 2\log(8m/\delta)},$$

and finally,

$$U_t^{\mathsf{I}}(\log(4/\delta)) \leq 6\sqrt{dt} \cdot \sqrt{d\log(t+1) + 2\log(8m/\delta)},$$

yielding the claimed bounds. □

With these in hand, we can conclude.

Proof of Lemma 11. We shall only show the feasible case; the infeasible is identical, and thus the details are omitted. Recall from §C.1.1 that in the feasible case,

$$\widetilde{\mathscr{T}_t} \geq t\Gamma - \mathscr{R}_t + Z_t.$$

By Lemma 16, with probability at least $1 - \delta/2, Z_t \geq \mathrm{LIL}(t, \delta/2)$ for all t. Further, by Lemma 23, with probability at least $1 - \delta/2$, at all times

$$\mathscr{R}_t \leq (1 + \log(t+1)) \cdot \max(U_t^{\mathsf{F},\mathsf{A}}(\log(8/\delta)), U_t^{\mathsf{F},\mathsf{B}}(\log(8/\delta)).$$

Finally, opening up the form of the same via , we have

$$(1+\log(t+1)) \cdot \max\left(12d\sqrt{t\log(t+1)} + 15\sqrt{dt\log(8m/\delta)}, 2d^{3/2}\sqrt{t\log^2(t+1)} + 3\sqrt{dt}\log(8m/\delta)\right),$$

and $1 + \log(t+1) \leq 3\log(t+1)$ for $t \geq 1$ But note that $d^{3/2}\sqrt{t\log^2(t+1)} \geq d\sqrt{t(1+\log(t+1))}$, and $\sqrt{dt}\log(8m/\delta) \geq \sqrt{dt\log(8m/\delta)}$ since $\delta \leq 1/2$. So, we may simply adjust the constants, and conclude that with probability at least $1 - \delta/2$,

$$\mathscr{R}_t \leq 36d^{3/2}\sqrt{t}\log^2(t+1) + 45\sqrt{dt\log^2(t+1)}\log(8m/\delta) \leq \mathscr{Q}_t^{\mathsf{F}}(\delta) - \mathrm{LIL}(t, \delta/2).$$

But now the result is obvious. □

D Proof of the Lower Bound

We conclude the appendix by presenting the proof of the lower bound of Theorem 13. We will first show that it suffices to show a $\Omega(K/\Gamma^2)$ lower bound for the minimum threshold problem (which we shall also formally specify) in order to show the claimed result. We then give a brief summary of the 'simulator' technique of Simchowitz et al. (2017), and proceed to show the aforementioned bound.

D.1 The Finite-Armed Single Objective Feasibility Testing Problem, and a Reduction to Feasibility Testing of LPs over a Simplex

We start by explicitly defining the finite-armed single objective feasibility testing problem, also known as the minimum threshold testing problem as discussed in §1

Problem Definition An instance of this problem is defined by a natural $K < \infty$, and a set of K probability distributions, $\{\mathbb{P}_k\}_{k\in[1:K]}$, each supported over $\mathbb{R}$, and a real $\delta \in (0,1)$. Let $a_k := \mathbb{E}_{S\sim\mathbb{P}_k}[S]$, and let a denote the K-dimensional vector collecting these means. We will assume that $a \in [-1/2, 1/2]^K$. The aim of the test is to distinguish the hypotheses

$$\mathcal{H}_{\mathsf{F}}^K : \max_k a_k > 0 \quad \text{versus} \quad \mathcal{H}_{\mathsf{I}}^K : \max_k a_k < 0.$$

The tester chooses an arm K_t in round t, and if $K_t = k$, then it observes in response a score $S \sim \mathbb{P}_k$, independently of the history. We shall assume that each $\mathbb{P}_k$ is σ^2-subGaussian about its mean, with $\sigma^2 \leq 1$. As in the linear setting considered in the main text, a test for this finite-armed single objective setting consists of an arm selection policy, a stopping time, and a decision rule, which we summarise as $(\mathscr{A}, \tau, \mathscr{D})$ in line with §2. The goal is reliability in the sense of Definition 1, and a good test should be valid and well adapted in the sense of Definition 2.

We now specify reductions of the above problem to the linear feasibility testing problem that is the subject of our paper. The key observation is that the finite-armed problem can either be directly interpreted as a LP feasibility testing problem over a discrete action set, or can, with a small loss in the noise strength, be expressed as a LP feasibility testing problem over a continuous $\mathcal{X}$, the critical implication being that lower bounds for the finite-armed setting extend to our problem of testing feasibility of linear programs. This enables us to only concentrate on showing a lower bound for the finite-armed single objective problem in the subsequent.

Reduction to General LP Feasibility Testing Note that in effect, the problem above reduces to feasibility testing for the linear case if we set $d = K$, $A = a^\top \in \mathbb{R}^{1\times d}$ and set $\mathcal{X} = \{e_i\}_{i=1}^d$, where the e_i are the standard basis elements for $\mathbb{R}^d$: $e_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots 0 \end{pmatrix}^\top$, where the 1 occurs in the ith position. Indeed, in this case, upon playing $x = e_i$, we observe feedback $S \sim \mathbb{P}_k$. But we can write $S = \mathbb{E}[S] + (S - \mathbb{E}[S]) = a_k + \zeta = Ax + \zeta$, where $\zeta = S - \mathbb{E}[S]$ is conditionally σ^2-subGaussian due to our assumption that each $\mathbb{P}_k$ is σ^2-subGaussian, so the reduction is valid if $\sigma^2 \leq 1$.

Reduction to LP Feasiblity Testing Over the Simplex We further observe that if $\sigma^2 \leq 1/2$, then the finite case also reduces to single constraint feasibility testing over the simplex. Indeed, suppose that we set d, A as above, and take $\mathcal{X} = \{x \in [0,1]^d : \sum x_i = 1\}$, and let $(\mathscr{A}, \tau, \mathscr{D})$ be a reliable test for this instance over 1-subGaussian noise. Then we can get a corresponding reliable test for the d-armed setting as follows:

- At each t, we first execute $\mathscr{A}$ to obtain a putative action x_t.
- Next, we *draw* a random index $K_t \sim x_t$, which is meaningful since x_t lies in the simplex, and so is a distribution over $[1:d]$.
- Then, we pull arm K_t in the finite-armed instance and we supply the feedback S_t to the linear algorithm to enable testing.

To argue that the ensuing test is reliable, we need to verify that the feedback obeys the structure we demand, in particular, that $S_t = Ax_t + \zeta_t$ for 1-subGaussian ζ_t. But notice that

$$S_t = a_{K_t} + \eta_t$$

for η_t σ^2-subGaussian, and further,

$$\mathbb{E}[S_t] = \mathbb{E}[a_{K_t}] = \sum x_t^k a_k = a^\top x_t = Ax_t,$$

as required. Further, since each $\mathbb{P}_k$ is supported on $[-1/2, 1/2]$, the random variable a_{K_t} is also supported on $[-1/2, 1/2]$, and so is 1/2-subGaussian by Hoeffding's inequality. Due to the independence of K_t and η_t, it follows that the feedback noise is $(1/2 + \sigma^2)$-subGaussian, and so the reduction holds if $\sigma^2 \leq 1/2$.

Improved Costs for Finite Arms. Prima facie the above reduction implies an $\widetilde{O}(K^2/\Gamma^2)$ stopping cost for our test employed on finite-armed settings. However, if $K < d^2$, then this may be improved to $\widetilde{O}(K/\Gamma^2)$, either by coupling the EOGT approach with direct UCB-based constructions as commonly employed for finite arm bandits, or by directly analysing EOGT whilst exploiting standard analyses that enable proofs of improved costs for the OFUL scheme over finite-armed settings (Lattimore & Szepesvári, 2020).

D.2 The Simulator Argument

For an execution of a feasibility test over a finite-armed setting, let N_t^k denote the number of times arm k has been pulled up to time t, and correspondingly let N_τ^k be the number of times the arm k has been pulled at stopping. Notice that in a distributional sense, we can view the behaviour of the tester over a fixed *transcript*, defined as a set of K sequences $\{S_i^k\}_{i=1}^\infty$, one for each k, each comprising of values drawn independently and identically from $\mathbb{P}_k$, the idea being that for each t such that $K_t = k$, we can just supply the learner with $S_{N_t^k}^k$ in response. This maintains the feedback distributions, and thus the probability of any event in the filtration induced by $\{\mathsf{H}_t\}_{t\geq 1}$. The main utility of the transcript view is that it allows manipulation of the distributions underlying an instance *after some number of arm pulls*, and exploiting such distribution shifts is the key insight of the simulator argument of Simchowitz et al. (2017).

Let us succinctly denote a transcript as $\{S_i^k\}_{k\in[1:K], i\in[1:\infty)}$. Further, let us write $\mathbf{P} = (\mathbb{P}_1, \cdots, \mathbb{P}_k)$ to compactly denote an instance, and write $\mathbf{P}(\cdot)$ to denote the probability of an event when the instance is $\mathbf{P}$. Throughout, we work with the natural filtration of the tester $\mathscr{F}_t$, which is the sigma algebra over H_t *and* any algorithmic randomness used by the tester. A *simulator* $\mathfrak{S}$ is a randomised map from transcripts to transcripts. Notice that this induces a new distribution over the behaviour of the algorithm, which we denote by $\mathbf{P}_\mathfrak{S}$. Let us say that an event $W \in \mathscr{F}_\tau$ is truthful for an instance $\mathbf{P}$ under a simulator $\mathfrak{S}$ if it holds that for every $E \in \mathscr{F}_\tau$,

$$\mathbf{P}(W \cap E) = \mathbf{P}_\mathfrak{S}(W \cap E).$$

In words, given any truthful event, the simulator does not modify the behaviour of the test up to the time it stops. We shall succinctly specify the simulator and distribution with respect to which an event is truthful by saying that 'W is $(\mathbf{P}, \mathfrak{S})$-truthful.'

The simulator approach to lower bounds, presented in Proposition 2 of Simchowitz et al. (2017), is summarised through the following bound. Fix an algorithm, and consider a pair of instances $\mathbf{P}^1$ and $\mathbf{P}^2$. Then, if W_1 is $(\mathbf{P}^1, \mathfrak{S})$-truthful, and W_2 is $(\mathbf{P}^2, \mathfrak{S})$-truthful, it holds that

$$\mathbf{P}^1(W_1^c) + \mathbf{P}^2(W_2^c) \geq \sup_{E\in\mathscr{F}_\tau} |\mathbf{P}^1(E) - \mathbf{P}^2(E)| - \mathrm{TV}(\mathbf{P}^1_\mathfrak{S}, \mathbf{P}^2_\mathfrak{S}), \tag{6}$$

where TV is the total variation distance $\mathrm{TV}(\mu\|\nu) := \sup_E \mu(E) - \nu(E)$. The idea thus is that if we construct a simulator that makes the algorithm behave similarly in either instance, i.e., such that $\mathrm{TV}(\mathbf{P}^1_\mathfrak{S}\|\mathbf{P}^2_\mathfrak{S}) \approx 0$, but the instances themselves are fundamentally quite different, so that $\sup_{E\in\mathscr{F}_\tau} |\mathbf{P}^1(E) - \mathbf{P}^2(E)|$ is large, then we can show lower bounds on how likely truthful events are to not occur.

The bound itself is easy to show: for any $E \in \mathscr{F}_\tau$, we have

$$|\mathbf{P}^1(E) - \mathbf{P}^2(E)| \leq |\mathbf{P}^1_\mathfrak{S}(E) - \mathbf{P}^2_\mathfrak{S}(E)| + |\mathbf{P}^1_\mathfrak{S}(E) - \mathbf{P}^1(E)| + |\mathbf{P}^2_\mathfrak{S}(E) - \mathbf{P}^2(E)|.$$

Since W_1 is $(\mathbf{P}^1, \mathfrak{S})$-truthful, the second term may be refined as

$$|\mathbf{P}^1_{\mathfrak{S}}(E) - \mathbf{P}^1(E)| = |\mathbf{P}^1_{\mathfrak{S}}(E \cap W_1) - \mathbf{P}^1(E \cap W_1) + \mathbf{P}^1_{\mathfrak{S}}(E \cap W_1^c) - \mathbf{P}^1(E \cap W_1^c)| = |\mathbf{P}^1_{\mathfrak{S}}(E \cap W_1^c) - \mathbf{P}^1(E \cap W_1^c)|,$$

and we may similarly bound $|\mathbf{P}^2_{\mathfrak{S}}(E) - \mathbf{P}^2(E)|$. The difference $|\mathbf{P}^1_{\mathfrak{S}}(E) - \mathbf{P}^2_{\mathfrak{S}}(E)|$ can in turn be bounded by the total variation distance. We conclude that

$$\sum_{i=1}^{2} \sup_{E \in \mathscr{F}_\tau} |\mathbf{P}^i_{\mathfrak{S}}(E \cap W_i^c) - \mathbf{P}^i(E \cap W_i^c)| + \mathrm{TV}(\mathbf{P}^1_{\mathfrak{S}}, \mathbf{P}^2_{\mathfrak{S}}) \geq \sup_{E \in \mathscr{F}_\tau} |\mathbf{P}^1(E) - \mathbf{P}^2(E)|,$$

and the left hand side can be resolved by just taking $E = W_i^c \in \mathscr{F}_\tau$.

We will utilise the above twice in our argument below, with the main trick being that if we only modify the transcript to affect arm k after T pulls, that is, we only change S_i^k for $i > T$, then the event $\{N_\tau^k \leq T\}$ is truthful under this simulator, letting us lower bound the probability that N_τ^k is small in some instance. We shall succinctly call such simulators *post-T simulators*.

D.3 A Lower Bound for Finite-Armed Single Constraint Feasibility Testing

We shall show the following

Theorem 25. *For any $\Gamma \in (0, 1/2], \delta \leq 1/4$, and $K < \infty$, and for any reliable test, there exists a finite-armed single objective feasibility testing instance that is feasible, with signal level at least Γ, $\sigma^2 = 1/2$-subGaussian noise, and $\sum a_k^2 \leq 1$, on which the algorithm must admit*

$$\mathbb{E}[\tau] \geq \frac{(1-2\delta)^3 K}{79\Gamma^2}.$$

Theorem 13 is immediate from the above.

Proof of Theorem 13. Setting $K = d$, and constructing either of the reductions from the finite-armed case to the linear program feasibility testing problem detailed in §D.1, which is possible because $\sigma^2 = 1/2$ and since $\|a\|_2 \leq 1$. But then the lower bound of Theorem 25 must apply. □

Without further ado, let us launch into proving the finite-armed lower bound.

Proof of Theorem 25. Fix $\Gamma \in (0, 1/2]$, and for $k \in [0 : K]$, and an $\varepsilon \in (0, \sqrt{1 - \Gamma^2/4K})$, define the following instance

$$\mathbf{P}^k = (\mathbb{P}^k_1, \cdots, \mathbb{P}^k_K),$$

where

$$\mathbb{P}^k_\ell = \begin{cases} \mathcal{N}(-\varepsilon, 1/2) & \ell \neq k \\ \mathcal{N}(\Gamma, 1/2) & \ell = k \end{cases}.$$

Observe that for $k > 0$, in instance $\mathbf{P}^k$, the kth arm is the only feasible action, while the rest are infeasible, while in instance $\mathbf{P}^0$, all arms are infeasible, with the tiny signal level $-\varepsilon$. Of course, each $\mathbf{P}^k$ defines an instance for us. We implicitly reveal to the test that the instance must lie in one of the $\mathbf{P}^k$ as the argument does not change even if the test is allowed to use this fact. Notice that the mean vector for $\mathbf{P}^k$ is some permutation of $(\Gamma, -\varepsilon, \cdots, -\varepsilon)$, and so has 2-norm $\Gamma^2 + (K-1)\varepsilon^2 \leq \Gamma^2 + (1 - \Gamma^2)/4 \leq 1$, since $\Gamma \in (0, 1/2]$. We shall, at the end of the proof, send $\varepsilon \to 0$, so the precise size of it is not important to the argument.

Now, the first key observation is that since $\mathbf{P}^k$ is feasible for each $k > 0$, but $\mathbf{P}^0$ is infeasible, it must be the case that under $\mathbf{P}^k$ for arm $k > 0$, the test verifies the feasibility of the instance by pulling arm k at least $\Omega(\Gamma^{-2})$ times. We will need a slightly refined form of this statement, as seen below.

Lemma 26. *Under the above instance structure, for every $k \in [1:K]$ and any $T \in \mathbb{N}$, it holds that*

$$\mathbf{P}^k(N_\tau^k > T) \geq 1 - 2\delta - \sqrt{T(\Gamma+\varepsilon)^2/2}.$$

Proof. Consider a post-T simulator $\mathfrak{S}^k$ such that $\{\hat{S}_i^k\} = \mathfrak{S}^k(\{S_i^k\})$, has the form

$$\hat{S}_i^{k'} = \begin{cases} S_i^{k'} & k' \neq k \text{ or } i \leq T \\ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(-\varepsilon, 1/2) & k' = k \text{ and } i > T \end{cases}.$$

First notice that for the KL divergence[10] $\mathrm{KL}(\mathbf{P}^k_{\mathfrak{S}^k} \| \mathbf{P}^0_{\mathfrak{S}^k})$, using the data processing inequality, this is bounded by the KL-divergence between the laws of the transcript under the two distrubtions, which in turn is only driven by the the first T entries of the transcript for T. Since, by a standard calculation,[11]

$$\mathrm{KL}(\mathcal{N}(\mu, 1/2) \| \mathcal{N}(\nu, 1/2)) = (\mu - \nu)^2,$$

we conclude that

$$\mathrm{KL}(\mathbf{P}^k_{\mathfrak{S}^k} \| \mathbf{P}^0_{\mathfrak{S}^k}) \leq T(\Gamma+\varepsilon)^2,$$

and in turn by an application of Pinsker's inequality[12] (see, e.g., Lattimore & Szepesvári, 2020, Chs.13, 14),

$$\mathrm{TV}(\mathbf{P}^k_{\mathfrak{S}^k}, \mathbf{P}^0_{\mathfrak{S}^k}) \leq \sqrt{T(\Gamma+\varepsilon)^2/2}.$$

Next, observe that the event $W_k := \{N_\tau^k \leq T\}$ is $(\mathbf{P}^k, \mathfrak{S}^k)$-truthful since the transcript for arm i is only modified after T pulls, and further, every event is $(\mathbf{P}^0, \mathfrak{S}^k)$-truthful since the simulator does not modify the arm distributions for $\mathbf{P}^0$, and so in particular $W_0 = \{N_\tau^k \leq \infty\}$ is truthful (and of course $\mathbf{P}^0(N_\tau^k > \infty) = 0$ trivially).

Finally, observe that since the instance $\mathbf{P}^k$ is feasible, and since the test is reliable, it holds that $\mathbf{P}^k(\mathscr{D}(\mathsf{H}_\tau) = \mathcal{H}_\mathsf{F}^K) \geq 1 - \delta$. But by the same coin, since $\mathbf{P}^0$ is infeasible, $\mathbf{P}^0(\mathscr{D}(\mathsf{H}_\tau) = \mathcal{H}_\mathsf{I}^K) \leq \delta$. Of course, $\{\mathscr{D}(\mathsf{H}_\tau) = \mathcal{H}_\mathsf{F}^K\}$ is an $\mathscr{F}_\tau$-event.

So, we may proceed to populate the inequality (6) with the above selections to conclude that

$$\mathbf{P}^k(N_\tau^K > T) + 0 \geq |1 - \delta - \delta| - \sqrt{T(\Gamma+\varepsilon)^2/2}.$$ □

With the above in hand, observe that since $\tau = \sum_{k=1}^K N_\tau^k$, the above already shows that $\mathbb{E}_{\mathbf{P}^k}[\tau] = \Omega(\Gamma^{-2})$. To extend this, we employ the following result.

Lemma 27. *Under the same setting as Lemma 26, for any $k, k' \in [1:K]$,*

$$\mathbf{P}^k(N_\tau^{k'} > T) + \mathbf{P}^{k'}(N_\tau^k > T) \geq \frac{1-2\delta}{2} - \frac{1+1/\sqrt{2}}{2}\sqrt{T(\Gamma+\varepsilon)^2}.$$

Proof. If $k = k'$, the claim is true due to Lemma 26. Without loss of generality, let us set $k = 1, k' = 2$. Define the simulator $\mathfrak{S}^{1\to 2}$ so that $\{\hat{S}_i^k\} = \mathfrak{S}^{1\to 2}(\{S_i^k\})$ has the form

$$\hat{S}_i^k = \begin{cases} S_i^k & k \notin \{1,2\} \text{ or } i \leq T \\ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\Gamma, 1/2) & k \in \{1,2\} \text{ and } i > T \end{cases}.$$

As in the proof of Lemma 26, the only difference between $\mathbf{P}^1_{\mathfrak{S}^{1\to 2}}$ and $\mathbf{P}^2_{\mathfrak{S}^{1\to 2}}$ is induced by the first T entries of the $k = 1$ and $k = 2$ rows, and thus

$$\mathrm{KL}(\mathbf{P}^1_{\mathfrak{S}^{1\to 2}} \| \mathbf{P}^2_{\mathfrak{S}^{1\to 2}}) \leq 2 \cdot T(\Gamma+\varepsilon)^2.$$

[10] which we measure in nats, i.e., $\mathrm{KL}(P\|Q) = \int \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}Q$, where the logarithm is natural

[11] $\int \frac{e^{-(x-\mu)^2}}{\sqrt{\pi}}((x-\nu)^2 - (x-\mu)^2)\mathrm{d}x = \int \frac{e^{-(x-\mu)^2}}{\sqrt{\pi}}(\mu-\nu)(2x-\mu-\nu)\mathrm{d}x = (\mu-\nu)\cdot(2\mu-\mu-\nu)$

[12] which says $\mathrm{TV}(P,Q) \leq \sqrt{\mathrm{KL}(P\|Q)/2}$.

Further, again, $W_1 := \{N_\tau^2 \le T\}$ is $(\mathbf{P}^1, \mathfrak{S}^{1\to 2})$-truthful, since for $\mathbf{P}^1$, the simulator $\mathfrak{S}^{1\to 2}$ only modifies the the law of arm 2, and does this only after T pulls of the same. Similarly, $W_2 := \{N_\tau^1 \le T\}$ is $(\mathbf{P}^2, \mathfrak{S}^{1\to 2})$-truthful. Now set $E = \{N_\tau^2 > T\}$. Then by (6), we have

$$\mathbf{P}^1(N_\tau^2 > T) + \mathbf{P}^2(N_\tau^1 > T) \ge |\mathbf{P}^1(N_\tau^2 > T) - \mathbf{P}^2(N_\tau^2 > T)| - \sqrt{T(\Gamma+\varepsilon)^2}.$$

Now observe that if $\mathbf{P}^1(N_\tau^2 > T) \ge \mathbf{P}^2(N_\tau^2 > T)$, then we are already done since by Lemma 26,

$$\mathbf{P}^2(N_\tau^2 > T) \ge 1 - 2\delta - \sqrt{T(\Gamma+\varepsilon)^2/2} > \frac{1-2\delta}{2} - \frac{1+1/\sqrt{2}}{2}\sqrt{T(\Gamma+\varepsilon)^2}.$$

So, we may assume that $\mathbf{P}^1(N_\tau^2 > T) \le \mathbf{P}^2(N_\tau^2 > T)$. But then we conclude that

$$2\mathbf{P}^1(N_\tau^2 > T) + \mathbf{P}^2(N_\tau^1 > T) \ge \mathbf{P}^2(N_\tau^2 > T) - \sqrt{T(\Gamma+\varepsilon)^2} \ge 1 - 2\delta - (1+1/\sqrt{2})\sqrt{T(\Gamma+\varepsilon)^2},$$

and the conclusion follows since $2\mathbf{P}^1(N_\tau^2 > T) + \mathbf{P}^2(N_\tau^1 > T) \le 2\mathbf{P}^1(N_\tau^2 > T) + 2\mathbf{P}^2(N_\tau^1 > T)$. □

With the above in hand, observe that since an arm is pulled at each t, $\tau = \sum_k N_\tau^k$. Thus, for any $T > 0$,

$$\begin{aligned}
\frac{1}{K}\sum_k \mathbb{E}_{\mathbf{P}^k}[\tau] &= \frac{1}{K}\sum_k \sum_{k'} \mathbb{E}_{\mathbf{P}^k}[N_\tau^{k'}] \\
&\ge \frac{1}{K}\sum_k \sum_{k'} T\mathbf{P}^{k'}(N_\tau^k > T) \\
&= \frac{T}{K}\left(\sum_k \mathbf{P}^k(N_\tau^k > T) + \frac{1}{2}\sum_{k,k'\neq k} \mathbf{P}^k(N_\tau^{k'} > T) + \mathbf{P}^{k'}(N_\tau^k > T)\right).
\end{aligned}$$

Now employing Lemma 26 and Lemma 27, we have

$$\begin{aligned}
\frac{1}{K}\sum_k \mathbb{E}_{\mathbf{P}^k}[\tau] &\ge \frac{T}{K}\left(1 - 2\delta - \sqrt{T(\Gamma+\varepsilon)^2/4}\right) + \frac{TK(K-1)}{2K}\left(\frac{1-2\delta}{2} - \frac{1+1/\sqrt{2}}{2}\sqrt{T(\Gamma+\varepsilon)^2}\right) \\
&\ge \frac{TK}{4}\left((1-2\delta) - (1+1/\sqrt{2})\sqrt{(T(\Gamma+\varepsilon)^2)}.\right)
\end{aligned}$$

Since the bound holds for every T, we can optimise the same[13] to conclude that

$$\max_k \mathbb{E}_{\mathbf{P}^k}[\tau] \ge \frac{1}{K}\sum_k \mathbb{E}_{\mathbf{P}^k}[\tau] \ge \frac{(1-2\delta)^3}{27(1+1/2+\sqrt{2})} \cdot \frac{(1-2\delta)^3 K}{(\Gamma+\varepsilon)^2} \ge \frac{(1-2\delta)^3 K}{79(\Gamma+\varepsilon)^2}.$$

If $\delta \le \frac{1}{4}$, this can be further lower bounded by $\frac{K}{632(\Gamma+\varepsilon)^2}$. Since the above inequality holds true for every $\varepsilon > 0$ small enough, the claimed result follows upon sending $\varepsilon \to 0$. □

[13] For $f(x) = ux - vx^{3/2}$, the derivative is $u - \frac{3v}{2}\sqrt{x}$, while the second derivative is negative over $[0,\infty)$, yielding the global maxima at $(2u/3v)^2$, with the maximum evaluating to $4u^3/9v^2 - 8u^3/27v^2 = \frac{4u^3}{27v^2}$. Setting $u = (1-2\delta)$, $v = (1+1/\sqrt{2})(\Gamma+\varepsilon)$, this evaluates to $\frac{4}{27} \cdot \frac{(1-2\delta)^3}{(1+1/\sqrt{2})^2(\Gamma+\varepsilon)^2}$.