OPTIMAL REGRET BOUNDS FOR LINEAR BANDITS WITH HEAVY-TAILED REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

In heavy-tailed linear bandits, a learner chooses among N actions in the d-dimensional unit ball, the mean rewards are linear in an unknown parameter, and the rewards only have bounded centered p-th moments for some $p \in (1, 2]$. Even for polynomially many actions, the best known upper and lower regret bounds differ by a factor that grows with d. Let $q = p/(p-1)$. For fixed $B > 0$, sufficiently large d, and $2 \le N \le d^B$, we show that the minimax regret over T rounds on the hardest set of N actions is $\Theta(\min\{T, \sqrt{d}\,[\log(1 + N/d^{q/2})]^{1/q} T^{1/p}\})$. Our main contribution is the lower bound, which separates the information shared across actions from the evidence available only at the optimal action. For $N = d^a$ with $a > q/2$, it improves the previous lower bound by a factor of order $d^{1/2-1/q}(\log d)^{1/q}$ and matches the known upper bound. We also propose Hybrid FTRL, which estimates the loss of each action by a direct or a linear estimate, whichever has the smaller moment bound. It attains the rate above on every finite action set, including the range where combining existing algorithms falls short. The rate shows that linear structure reduces regret only when N exceeds $d^{q/2}$.

1 INTRODUCTION

The multi-armed bandit (MAB) problem is a fundamental model for sequential decision making under uncertainty, capturing the classical trade-off between exploration and exploitation (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2020). Because classical stochastic MABs impose no structural relation among arms, their learning complexity necessarily grows with the number of available actions. In many applications, however, actions are described by feature vectors and their expected rewards share a common structure. Linear bandits capture this structure through a linear reward model (Dani et al., 2008; Abbasi-Yadkori et al., 2011), so that observations of one action also inform the others. Such information sharing is particularly valuable when the action set is large relative to the feature dimension (Lattimore & Szepesvári, 2020).

Most analyses of stochastic linear bandits, however, rely on bounded or sub-Gaussian rewards (Dani et al., 2008; Abbasi-Yadkori et al., 2011). This assumption is often violated in practice: in financial markets and online advertising, for instance, rare but extremely large observations occur with non-negligible probability (Medina & Yang, 2016; Lu et al., 2019; Xue et al., 2020). Under such heavy-tailed rewards, empirical means lose the concentration properties on which standard confidence bounds are built, and guarantees derived for the sub-Gaussian setting no longer apply. This has made *heavy-tailed bandits*, where only weak moment conditions are imposed on the rewards, an active line of research, with robust estimators based on truncation and median-of-means as its central tools (Bubeck et al., 2013; Medina & Yang, 2016; Shao et al., 2018).

For heavy-tailed MABs, Bubeck et al. (2013) initiated a systematic study, and Robust MOSS (Wei & Srivastava, 2020) later attained the minimax-optimal regret. Heavy-tailed linear bandits have been studied through median-of-means and truncation estimators (Medina & Yang, 2016; Shao et al., 2018), with finite action sets (Xue et al., 2020) and symmetric noise (Zhong et al., 2021) also considered. These works established the optimal dependence of the regret on the time horizon. In contrast, how the regret depends *jointly* on the feature dimension and the number of actions, which is precisely where linear structure should pay off, is much less understood. Recent work (Tajdini et al., 2026; Zhao et al., 2025) has made progress on this question, yet the best known upper and lower bounds still do not match. In this paper, we study this question for linear bandits with a finite action set.

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

Specifically, we consider a known action set $\mathcal{A} = \{a_1, \dots, a_N\}$ in the Euclidean unit ball B_2^d. An instance $\nu = (\theta, P_1, \dots, P_N)$ consists of an unknown $\theta \in B_2^d$ and reward distributions $P_1, \dots, P_N$; choosing action i yields an independent draw $Y \sim P_i$ with

$$\mu_i := \mathbb{E}_{P_i}[Y] = a_i^\top \theta \qquad \text{and} \qquad \mathbb{E}_{P_i}\big[|Y - \mu_i|^p\big] \le 1, \qquad 1 < p \le 2. \tag{1.1}$$

Beyond these two conditions, the P_i may differ arbitrarily across actions. A policy π chooses $I_t \in [N]$ from past observations and incurs the pseudo-regret $R_T(\pi, \nu) = \mathbb{E}_\pi\big[\sum_{t=1}^T (\mu^\star - \mu_{I_t})\big]$, where $\mu^\star = \max_i \mu_i$. The minimax regret on a fixed action set depends on its geometry (it is small, for instance, if all actions are close to each other), so we measure the complexity of the problem by the minimax regret on the hardest set of N actions,

$$\mathfrak{R}_T(d, N, p) := \sup_{\mathcal{A} \subset B_2^d, |\mathcal{A}| = N} \inf_\pi \sup_\nu R_T(\pi, \nu), \tag{1.2}$$

where the policy may depend on $\mathcal{A}$, T, and p, and ν ranges over all instances satisfying (1.1).

Let $q = p/(p-1)$ be the Hölder conjugate of p. Two types of upper bounds on $\mathfrak{R}_T(d, N, p)$ are known. Ignoring the linear structure, Robust MOSS (Wei & Srivastava, 2020) achieves regret $O(N^{1/q} T^{1/p})$, removing a logarithmic factor from Bubeck et al. (2013). Exploiting the linear structure, Tajdini et al. (2026) obtained $\widetilde{O}(\sqrt{d}(\log N)^{1/q} T^{1/p})$, and Zhao et al. (2025) removed the extra $\log T$ factor. The best known upper bound therefore scales as $\min\{N^{1/q}, \sqrt{d}(\log N)^{1/q}\}\, T^{1/p}$. On the lower side, Tajdini et al. (2026), who also showed that several earlier lower bounds had been compared with upper bounds under mismatched moment conditions, proved the finite-action lower bound $\Omega(d^{1/q}(\log N/\log d)^{1/q} T^{1/p})$ for long horizons. These bounds fail to match even for polynomially many actions: for $N = d^a$ with fixed $a > q/2$, the upper bound scales as $\sqrt{d}(\log d)^{1/q} T^{1/p}$, whereas the lower bound scales only as $d^{1/q} T^{1/p}$, and their ratio diverges with d. This leaves a basic question open:

> *What is the minimax regret of finite-action heavy-tailed linear bandits as a joint function of the dimension d and the number of actions N?*

Beyond closing a gap between bounds, the answer determines when linear structure is statistically useful. With few actions, treating the problem as an ordinary MAB may already be optimal; with many actions, sharing information through the linear model should reduce regret. Existing results do not locate the transition between these two regimes.

In this paper, we answer this question for polynomially many actions, that is, $N \le d^B$ for an arbitrary fixed $B > 0$. Our contributions are as follows.

- We establish the minimax rate of $\mathfrak{R}_T(d, N, p)$ up to constants. For every fixed $B > 0$, all sufficiently large d, $2 \le N \le d^B$, and $T \ge 1$,

$$\mathfrak{R}_T(d, N, p) = \Theta\Big(\min\{T,\ \sqrt{d}\left[\log\Big(1 + \frac{N}{d^{q/2}}\Big)\right]^{1/q} T^{1/p}\}\Big). \tag{1.3}$$

  This rate also locates the transition discussed above at $N = d^{q/2}$: below this scale, the factor multiplying $T^{1/p}$ is of order $N^{1/q}$ and treating actions independently is already optimal; above it, the linear structure yields a strict improvement.
- We prove a new lower bound, which is the main technical contribution of this paper. For every admissible (d, N), we construct a set of exactly N unit actions on which every policy incurs the rate in (1.3), using stationary, action-dependent rewards that satisfy only (1.1). In the example above, where $N = d^a$ with $a > q/2$, it improves the previous lower bound by a factor of $d^{1/2 - 1/q}(\log d)^{1/q}$ and matches the existing linear upper bound, showing that the gap there was due to the lower bound.
- We develop a matching upper bound. Taking the better of an independent-arm and a linear algorithm is not enough: when N is of order $d^{q/2} \log d$, it yields the factor $\sqrt{d}(\log d)^{1/q}$, whereas our lower bound gives only $\sqrt{d}(\log\log d)^{1/q}$. We close this remaining gap with a hybrid estimator that chooses, action by action, between direct importance-weighted and linear estimation. Combined with clipping, bias correction, and Tsallis-regularized FTRL, the resulting algorithm, Hybrid FTRL, attains (1.3) on every finite action set.

Organization. Section 2 reviews related work. Section 3 formalizes the setting, states the main theorem, and compares it with previous bounds. Section 4 presents the lower-bound construction, and Section 5 the matching algorithm. Complete proofs of the lower and upper bounds appear in Appendices A and B, and Appendix C details the comparison with previous bounds.

2 Related Work

Dani et al. (2008) gave one of the first regret analyses for stochastic linear optimization with bandit feedback, and Abbasi-Yadkori et al. (2011) introduced OFUL, whose self-normalized confidence sets yield $\widetilde{O}(d\sqrt{T})$ regret and underlie most optimism-based methods. Thompson sampling provides a randomized alternative with frequentist guarantees (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017), while Lattimore & Szepesvári (2017) characterized the asymptotically optimal instance-dependent regret for finite-armed linear bandits and showed that neither optimism nor Thompson sampling attains it in general. Recent work pursues stronger adaptivity and sharper guarantees, including best-of-both-worlds algorithms (Lee et al., 2021), tighter confidence sequences from martingale mixtures (Flynn et al., 2023), variance-adaptive confidence sets (Jun & Kim, 2024), and ensemble sampling with an ensemble size that does not grow linearly with the horizon (Janz et al., 2024). All of these rely on light-tailed concentration, which is unavailable when only low-order moments exist.

Bubeck et al. (2013) initiated the study of heavy-tailed MABs under a bounded p-th moment, using truncated means, median-of-means, and Catoni-type estimators, and showed that the optimal horizon dependence degrades from $\sqrt{T}$ to $T^{1/p}$ when $p < 2$. Subsequent work removed the need to know the moment bound (Lee et al., 2020) and attained the minimax-optimal regret $O(N^{1/q}T^{1/p})$ with Robust MOSS (Wei & Srivastava, 2020). Recent work has extended heavy-tailed MABs beyond the stochastic setting, with best-of-both-worlds guarantees (Huang et al., 2022; Cheng et al., 2024) and parameter-free algorithms that adapt to unknown tail parameters (Chen et al., 2025). For linear rewards, Medina & Yang (2016) first extended truncation and median-of-means to heavy-tailed linear bandits, and Shao et al. (2018) obtained $\widetilde{O}(dT^{1/p})$ regret, attaining the optimal exponent of T. For finite action sets, Xue et al. (2020) obtained $\widetilde{O}(\sqrt{d}\,T^{1/p})$ regret together with an $\Omega(d^{1/q}T^{1/p})$ lower bound, showing that finite action sets admit a better dimension dependence; related ideas have been extended to generalized linear bandits (Xue et al., 2023). A more recent line replaces hard truncation with adaptive Huber regression, yielding variance-aware guarantees (Li & Sun, 2024) and one-pass algorithms with low per-round cost (Wang et al., 2025). Under symmetric noise, Zhong et al. (2021) instead reduce the problem to sub-Gaussian algorithms.

Most closely related to our work, Tajdini et al. (2026) revisited the minimax theory of heavy-tailed linear bandits under a moment bound that does not grow with the dimension. They showed that the lower bound of Shao et al. (2018) relies on a dimension-dependent moment bound, and developed experimental-design-based algorithms and new lower bounds for both finite and infinite action sets. Zhao et al. (2025) then removed the extra $\log T$ factor from the finite-action upper bound via clipped, bias-corrected FTRL. As discussed in Section 1, however, the resulting bounds still differ by a factor diverging with d even for polynomially many actions. Our lower bound connects standard ingredients, namely adaptive relative-entropy decompositions (Kaufmann et al., 2016), a list version of Fano's inequality (Sason, 2019, Section 4.1), and a polynomial code (DeVore, 2007), through a new bandit reduction; our upper bound adds to Tsallis-regularized FTRL (Zimmert & Seldin, 2021; Zhao et al., 2025) a coordinatewise interpolation between direct and linear estimates. At $p = 2$, a bounded-reward result of Ito (2024, Corollary 1) has a leading term with the same cardinality dependence as ours, whereas our lower bound concerns the larger class with a bounded centered second moment.

3 Main Result

This section makes the setting of Section 1 precise, states our main theorem, and compares it with the best previous bounds.

3.1 Setting and Notation

We work with the model (1.1). Conditionally on the entire history and on $I_t = i$, the reward at round t is a fresh draw from P_i. The laws are stationary but may vary arbitrarily across actions; in particular,

we assume neither a common additive noise law nor symmetric noise. A policy knows p, T, $\mathcal{A}$, and the moment bound, and may use its own random seed. Every round, including exploratory ones, contributes to the regret. For a fixed action set, we write

$$\mathfrak{R}_T(\mathcal{A}, p) = \inf_{\pi} \sup_{\nu} R_T(\pi, \nu), \qquad \text{so that} \qquad \mathfrak{R}_T(d, N, p) = \sup_{\mathcal{A} \subset B_2^d,\, |\mathcal{A}| = N} \mathfrak{R}_T(\mathcal{A}, p). \tag{3.1}$$

We omit p from $\mathfrak{R}_T$ when it is fixed, and all logarithms are natural. Finally, we denote the complexity factor in (1.3) by

$$F(d, N) = \sqrt{d} \left[\log\left(1 + \frac{N}{d^{q/2}}\right) \right]^{1/q}. \tag{3.2}$$

3.2 The Minimax Rate

Theorem 3.1 (Matching bounds for polynomial cardinalities)**.** *Fix $1 < p \le 2$ and $B > 0$. There exist $c_{p,B}, C_p > 0$ and $d_0(p, B)$ such that, for all integers $d \ge d_0(p, B)$, $2 \le N \le d^B$, and $T \ge 1$,*

$$c_{p,B} \min\{T, F(d, N) T^{1/p}\} \le \mathfrak{R}_T(d, N, p) \le C_p \min\{T, F(d, N) T^{1/p}\}. \tag{3.3}$$

For each such (d, N), a single set of exactly N unit actions witnesses the lower bound simultaneously for every horizon and every policy. The upper bound holds for every finite $\mathcal{A} \subset B_2^d$.

Theorem 3.1 is the formal version of (1.3). Its lower inequality, our main technical contribution, follows from Section 4: for $d \ge d_0(p, B)$, (4.6) provides a set of exactly N unit vectors, chosen independently of T, on which every policy incurs this regret. The upper inequality is Theorem 5.1 with $M = 2^p$, attained by the Hybrid FTRL algorithm of Section 5.

Since $x/2 \le \log(1 + x) \le x$ on $[0, 1]$, we have $2^{-1/q} N^{1/q} \le F(d, N) \le N^{1/q}$ for $N \le K = d^{q/2}$, the independent-arm rate; beyond K, the factor grows only logarithmically in N/K.

Scope of the theorem. The hard action set is fixed before the horizon and the policy are chosen; the hard parameters may depend on T, and the reward magnitudes may grow with T. The theorem characterizes the hardest action geometry for each (d, N); it does not assert that all action sets of size N are equally difficult. Constants need not remain uniform as B grows or as p approaches one. Finally, because the model allows action-dependent weak-moment laws, the finite-variance endpoint $p = 2$ does not yield a lower bound for the smaller classes of uniformly bounded or Gaussian rewards.

3.3 Comparison with Previous Bounds

Table 1 places our bounds next to the best previous ones. The best previous upper factor, obtained by running the better of Robust MOSS (Wei & Srivastava, 2020) and the algorithm of Zhao et al. (2025), is

$$G(d, N) = \min\{N^{1/q}, \sqrt{d} (\log N)^{1/q}\}. \tag{3.4}$$

Both algorithms assume raw-moment bounds, which our centered-moment model implies: $|\mu_i| \le 1$ gives $\mathbb{E}|Y|^p \le 2^p$. The additive $O(N)$ term of Robust MOSS is absorbed when $T \ge N$, and the linear cap covers the case $T < N$.

A common model and comparison range. Tajdini et al. (2026, Theorem 2) assume $d \le N \le 2^{\lfloor d/4 \rfloor}$ and $T \ge 4^q d^q$, and their hard instance uses at most N actions. After rescaling, their construction fits our unit-ball, centered-moment model; Appendix C verifies this transfer, including padding to exactly N actions. Our theorem, by contrast, supplies exactly N unit actions, uses one set for all horizons, and also establishes the linear cap.

For fixed powers, the gap was in the lower bound. Let $N = d^a$ with fixed $a > 1$. The previous lower complexity factor is then $\Theta_{p,a}(d^{1/q})$, and ours improves on it as follows:

Range	Our lower factor	Gain over the previous factor
$1 < a \le q/2$	$d^{a/q}$	$d^{(a-1)/q}$
$a > q/2$	$\sqrt{d} (\log d)^{1/q}$	$d^{1/2 - 1/q} (\log d)^{1/q}$.

(3.5)

Table 1: Previous and new regret bounds. Constants depending on p and on the fixed moment bounds are suppressed. The previous lower bound is shown with its explicit $\log d$ denominator and holds under the restrictions discussed in Section 3.3. Our matching lower bound holds throughout each fixed polynomial range.

Bound	Regret guarantee
Previous lower bound: Tajdini et al. (2026)	$\Omega_p\big(d^{1/q}(\log N/\log d)^{1/q}T^{1/p}\big)$
Our lower bound	$\Omega_{p,B}(\min\{T, F(d,N)T^{1/p}\})$
Best previous upper bound: Wei & Srivastava (2020); Zhao et al. (2025)	$O_p(\min\{T, G(d,N)T^{1/p}\})$
Our upper bound	$O_p(\min\{T, F(d,N)T^{1/p}\})$

These are order comparisons in the common horizon range $T \geq 4^q d^q$, with constants depending on p and a; the first range is nonempty only when $p < 2$. In the first range, the new lower bound meets the independent-arm upper rate; in the second, it meets the known linear upper order, and at $p = 2$ the gain is $\sqrt{\log d}$. In both ranges, the previous upper bound was already of the right order. For $N \leq d$, the independent-arm lower rate already follows from standard multi-armed instances, and we claim no new rate there; our claim is a single match that holds uniformly throughout the polynomial range.

Between the regimes, the upper bound must improve as well. Put $K = d^{q/2}$. For N of order $K \log d$, the best previous upper factor is $G(d,N) = \Theta(\sqrt{d}(\log d)^{1/q})$, whereas $F(d,N) = \Theta(\sqrt{d}(\log\log d)^{1/q})$; the algorithm of Section 5 attains the smaller factor. As with any capped rate, an improved factor changes the regret only when the linear cap is inactive.

4 Lower Bound: Separating Shared Evidence from Direct Hits

This section presents the lower bound, the main technical contribution of the paper. Multi-armed lower bounds hide the optimal arm among many candidates, each with its own mean. In a linear bandit, all N means are determined by a single d-dimensional parameter, so moving the optimum also shifts the means of the other actions, and observing those actions can reveal where the optimum is. A lower bound must therefore limit this shared information while counting direct discoveries of the optimum, and it must do both against a learner that uses weak observations to decide where to test next.

Our argument has three steps. First, a five-point reward family with exactly linear means routes the two kinds of evidence through separate rare channels (Section 4.2). Second, an auxiliary experiment and a pathwise coupling turn every discovery of the optimum into membership in a random list of controlled size and information (Section 4.3). These steps yield a lower bound for every low-coherence action set (Theorem 4.1). Third, finite-field polynomial graph codes supply low-coherence sets of every polynomial size (Section 4.4).

4.1 A Lower Bound from Low Coherence

Theorem 4.1 (Low-coherence lower bound)**.** *Let $\mathcal{A} = \{v_1, \dots, v_N\} \subset \mathbb{R}^d$ consist of unit vectors, where $N \geq 2$ and $|v_i^\top v_j| \leq \kappa \leq 1/2$ for $i \neq j$, with $\kappa > 0$. Set $H = \kappa^{-q}$. For every $T \geq 1$,*

$$\mathfrak{R}_T(\mathcal{A}) \geq c_p \min\left\{T, \left[H \log\left(1 + \frac{N}{H}\right)\right]^{1/q} T^{1/p}\right\}. \tag{4.1}$$

Here $c_p > 0$ depends only on p. The hard instances have stationary, action-dependent reward laws with linear means, centered p-th moments at most one, and a common five-point support.

Sections 4.2 and 4.3 prove Theorem 4.1, deferring routine calculations to Appendix A.1. For $N \leq H$, the bound is of order $\min\{T, N^{1/q}T^{1/p}\}$, the independent-arm rate. For larger N/H, its complexity factor is $\kappa^{-1}[\log(1 + N/H)]^{1/q}$.

4.2 An Admissible Hard Family with Two Reward Channels

Draw a hidden index U uniformly from $[N]$ and set $\theta_U = \rho v_U$, with $0 < \rho \le 1/32$. Then action U is the unique optimum, and every other action has gap at least $\rho/2$. We split each mean into a shared part and a diagonal part, $g_U(i) = \rho v_i^\top v_U \mathbf{1}\{i \neq U\}$ and $h_U(i) = \rho \mathbf{1}\{i = U\}$. The shared part is bounded in absolute value by $\ell = \rho\kappa$, and the diagonal part is nonzero only at U. Rewards are generated through a gate $G \in \{0, \ell, \rho\}$ drawn independently of the parameter and the action:

$$\begin{aligned} r_c &= (16c)^q, \qquad B_c = 2c/r_c, \qquad c \in \{\ell, \rho\}, \\ \Pr(G = c) &= r_c, \qquad \Pr(Y = +B_c \mid G = c, U, i) = \frac{1}{2} + \frac{m_c(U,i)}{4c}, \\ m_\ell(U,i) &= g_U(i), \qquad m_\rho(U,i) = h_U(i). \end{aligned} \tag{4.2}$$

The opposite sign receives the remaining conditional probability, and the closed gate $G = 0$ produces reward zero. The conditional sign probabilities lie between $1/4$ and $3/4$, and the gate is open with total probability at most $1/2$. The support is $\{0, \pm B_\ell, \pm B_\rho\}$, and the two sign tilts produce exactly the required linear mean. The two channels together have absolute p-th moment $2 \cdot 8^{-p}$, so the centered moment is at most $2^{1-2p} < 1$. Drawing fresh gates and signs at every pull yields stationary instances. Over the horizon, the diagonal gate opens about $\Lambda = Tr_\rho$ times, whereas the shared gate opens H times less often, since $r_\ell / r_\rho = \kappa^q = 1/H$.

Why two channels. The rarity of each channel scales with the q-th power of its signal, which makes shared observations cheap: a pull contributes at most $r_\ell/4 = r_\rho \kappa^q/4$ to the relative entropy between the auxiliary experiment and the baseline (Lemma 4.2). If the diagonal channel also carried the shared signal, a pull could contribute $r_\rho \kappa^2/4$, and the argument below would give Theorem 4.1 only with $H = \kappa^{-2}$. For the codes of Section 4.4, where κ is of order $d^{-1/2}$, the dimension factor would drop from $\sqrt{d}$ to $d^{1/q}$. The exponents agree at $p = 2$; for $p < 2$, separating the channels is what yields the factor $\sqrt{d}$ in our argument.

4.3 Adaptive Reduction and Proof of Theorem 4.1

Consider an auxiliary experiment Q_U that removes the diagonal signal h_U but retains the shared signal g_U, and a baseline Q_0 that removes both. Only the true experiment P_U needs linear means; Q_U serves purely as a reference measure. Let $\mathcal{D}$ consist of the auxiliary trajectory, the policy's private seed, and an independent round J drawn uniformly from $[T]$ and not revealed to the policy. Because the gates can be read off from the reward magnitudes, we can form the list of actions played while the diagonal gate was open, augmented by I_J^Q:

$$\begin{aligned} \Lambda &= Tr_\rho, \qquad I_Q(U; \mathcal{D}) \le Tr_\ell/4 = \Lambda/(4H), \\ L(\mathcal{D}) &= \{I_t^Q : G_t = \rho,\ 1 \le t \le T\} \cup \{I_J^Q\}, \\ \mathbb{E}_Q|L| &\le \Lambda + 1. \end{aligned} \tag{4.3}$$

The two lemmas below show that this list captures everything an adaptive policy can learn about U.

Lemma 4.2 (Information budget and pathwise coupling)**.** *The information and list-size bounds in* (4.3) *hold. Moreover, P_U and Q_U can be coupled so that $\{I_J^P = U\} \subseteq \{U \in L(\mathcal{D})\}$ pathwise.*

Proof sketch. A χ^2 calculation gives $\mathrm{KL}(Q_{u,i} \| Q_0) \le r_\ell/4$ for every action i and index u. The chain rule evaluates this bound at the actions the policy actually selects, so it survives adaptive sampling; averaging over U bounds the mutual information by $Tr_\ell/4$. Each round enters the list with probability r_ρ. For the coupling, share the hidden index, the policy seed, the gates, the sign uniforms, and J. The two experiments have the same reward law unless the diagonal gate is open at action U, so the trajectories coincide until the first such round, which already places U in the list. If no such round occurs, they coincide throughout, and $I_J^P = U$ implies $I_J^Q = U \in L(\mathcal{D})$. Appendix A.1 gives the details. □

Lemma 4.3 (Testing with a list of random size). *Let $v = \mathbb{E}_Q|L|/N$. Then*

$$\begin{aligned} \Pr_P(I_J^P = U) \le s &:= \Pr_Q\{U \in L(\mathcal{D})\}, \\ I_Q(U;\mathcal{D}) &\ge \mathrm{kl}(s\|v) \ge s\log(1/v) - \log 2, \\ s &\le \frac{\Lambda/(4H) + \log 2}{\log(N/(\Lambda+1))} \qquad (\Lambda + 1 < N). \end{aligned} \tag{4.4}$$

Proof. The first inequality is the coupling of Lemma 4.2. Under the product of the Q-marginals of U and $\mathcal{D}$, the index is uniform and independent of the list, so the event $\{U \in L(\mathcal{D})\}$ has probability v. Data processing applied to the indicator of this event gives the middle line, and combining it with Lemma 4.2 and $v \le (\Lambda+1)/N$ gives the last line. □

The list may depend on U through the adaptive observations; the lemma, a Fano inequality for a list of random size (cf. Sason, 2019), uses only its expected size.

Proof of Theorem 4.1. To calibrate, set $c_0 = 1/64$ and $\rho = \min\{1/32, (1/16)[c_0 H \log(1 + N/H)/T]^{1/q}\}$. Then $\Lambda \le c_0 H \log(1 + N/H)$ for every horizon. When $N/H \ge 16$, (4.4) bounds the success probability by $c_0 + 1/2 < 3/4$. When $N/H \le 16$, the list bound alone does not suffice, and we instead compare the baseline directly with the true experiment: $N^{-1}\sum_u \mathrm{KL}(Q_0^{\mathcal{D}} \| P_u^{\mathcal{D}}) \le (\Lambda/2)(1/H + 1/N)$. This uses the fact that, under the baseline, the actions are independent of the uniformly averaged index. Pinsker's and Jensen's inequalities then bound the success probability by $1/N + \sqrt{(\Lambda/4)(1/H + 1/N)} \le 5/8$. In both cases the probability of a mistake at the random round is at least $1/4$. Since this round is uniform and independent of the policy, the prior-averaged regret is at least $\rho T/8$, so some fixed index incurs at least this regret. Substituting ρ yields (4.1), including its linear cap, for every horizon. □

4.4 Polynomially Many Actions in a Fixed Dimension

We now instantiate Theorem 4.1 with the normalized polynomial graph construction of DeVore (2007). Fix $B > 0$, let $b = \lceil 2B \rceil$, and choose a power of two k with $\sqrt{d}/2 < k \le \sqrt{d}$. Polynomials of degree at most b over $\mathbb{F}_k$ give

$$\begin{aligned} v_f(x,y) &= k^{-1/2}\mathbf{1}\{y = f(x)\}, \\ \#\{v_f\} &= k^{b+1}, \qquad 0 \le v_f^\top v_g \le b/k \quad (f \ne g). \end{aligned} \tag{4.5}$$

The inner-product bound follows from the root bound for polynomials (Appendix A.2). For d sufficiently large, depending on B, there are at least d^B such vectors and $k \ge 2b$. Retain any $2 \le N \le d^B$ of them and pad them with zeros to dimension d. Then $H = (k/b)^q$ is comparable to $d^{q/2}$, with constants depending on p and B, and Theorem A.2 gives

$$\mathfrak{R}_T(d,N) \ge c_{p,B} \min\left\{T, \sqrt{d}\left[\log\left(1 + \frac{N}{d^{q/2}}\right)\right]^{1/q} T^{1/p}\right\}. \tag{4.6}$$

The set has exactly N elements and does not depend on the horizon or the policy. Adding arbitrary actions, however, need not preserve the coherence bound or the direct-hit mechanism, so the guarantee concerns this particular set.

5 A Matching Upper Bound via Hybrid Estimation

Section 4 shows that regret of order $\min\{T, F(d,N)T^{1/p}\}$ is unavoidable; we now show that it is attained on every finite action set. Since running the better of an independent-arm and a linear algorithm is not enough when N is of order $d^{q/2}\log d$ (Section 3.3), our policy, Hybrid FTRL, combines both estimation strategies and chooses between them action by action. It is Tsallis-regularized FTRL (Zimmert & Seldin, 2021) with the clipping and bias correction of Zhao et al. (2025), mixed with a fixed G-optimal design in the spirit of the geometric exploration of Bubeck et al. (2012). We work with losses; the guarantee holds for any mean sequence fixed in advance by an oblivious adversary.

Theorem 5.1 (Upper bound for every finite action set). *Let $\mathcal{A} = \{a_1, \dots, a_N\} \subset B_2^d$ with $N \geq 2$. Before play, an oblivious adversary fixes vectors $\vartheta_t \in \mathbb{R}^d$ and loss laws $P_{t,i}$ with means $\mu_{t,i} = a_i^\top \vartheta_t \in [-1, 1]$ and raw p-th moments at most a known $M > 0$. When action i is selected at round t, the learner observes a fresh draw $W_t \sim P_{t,i}$. Then Hybrid FTRL (Algorithm 1) satisfies, for every fixed action j,*

$$\mathbb{E} \sum_{t=1}^{T} (\mu_{t,I_t} - \mu_{t,j}) \leq C_{p,M} \min\{T, F(d, N) T^{1/p}\}. \tag{5.1}$$

Appendix B proves Theorem 5.1. The stationary model (1.1) is the special case $W_t = -Y_t$ and $\vartheta_t = -\theta$: the centered-moment bound and $|\mu_i| \leq 1$ give $\mathbb{E}|W_t|^p \leq M = 2^p$, so the constant depends only on p (Corollary B.2).

5.1 Choosing the Cheaper Estimate

Given a sampling distribution π_t with full support, let $Q_t = \sum_i \pi_{t,i} a_i a_i^\top$ and $\ell_{t,i} = a_i^\top Q_t^\dagger a_i$, where the inverse is taken on the span of the actions. A single observation (I_t, W_t) yields two estimates of each mean loss:

$$D_{t,i} = \frac{\mathbf{1}\{I_t = i\} W_t}{\pi_{t,i}}, \qquad Z_{t,i} = (a_i^\top Q_t^\dagger a_{I_t}) W_t. \tag{5.2}$$

Both are conditionally unbiased, and their conditional p-th moments are at most $M \pi_{t,i}^{1-p}$ and $M \ell_{t,i}^{p/2}$, respectively. The direct estimate pays nothing for the dimension, whereas the linear estimate can be much cheaper for an action that is rarely sampled. For each action, we select the estimate with the smaller moment bound; the choice depends only on the past and on π_t. Because the selection is predictable, it preserves unbiasedness, and the selected estimate has moment multiplier $m_{t,i} = \min\{\pi_{t,i}^{1-p}, \ell_{t,i}^{p/2}\}$.

5.2 The Algorithm

For $\beta \in (0, 1)$ and $\alpha = 1 - \beta$, let $\Psi(w) = -(\sum_i w_i^\alpha - 1)/\alpha$ be the Tsallis regularizer on the simplex Δ_N. Follow-the-regularized-leader (FTRL) computes $w_t = \arg\min_{w \in \Delta_N} \{\Psi(w) + \eta \sum_{s<t} \langle w, z_s \rangle\}$, whose coordinates are strictly positive. With the clipping threshold $s_{t,i} = \lambda w_{t,i}^{-\beta}$ and the bonus $b_{t,i} = M m_{t,i} s_{t,i}^{1-p}$, the loss estimates passed to FTRL are

$$z_{t,i} = \widetilde{\mu}_{t,i} \mathbf{1}\{|\widetilde{\mu}_{t,i}| \leq s_{t,i}\} - b_{t,i}, \qquad \eta = \frac{\beta}{8\lambda}. \tag{5.3}$$

The bonus offsets the clipping bias in the direction that matters for a fixed comparator. Algorithm 1 gives the complete, tuned policy; when $N \leq K$, direct estimates alone suffice.

5.3 Proof Sketch of Theorem 5.1

Both the stability term and the bias term of FTRL are controlled by the same quantity $S_t = \sum_i w_{t,i}^{1+\beta(p-1)} m_{t,i}$, which we call the effective moment cost. In the small branch, $\beta = 1/q$ gives $S_t \leq N^{1/q}$. The key step is the corresponding bound in the hybrid branch.

Lemma 5.2 (Effective moment cost). *In every round of the hybrid branch, $S_t \leq 4d^{p\alpha/2}$.*

Proof. Drop the index t and write $v_i = \pi_i \ell_i$. Standard covariance identities give $0 \leq v_i \leq 1$ and $\sum_i v_i \leq d$, and $\gamma \leq 1/4$ gives $w_i \leq 2\pi_i$. Interpolating between the two moment bounds and applying Hölder's inequality yields

$$\begin{aligned} S_t &\leq 4 \sum_i \min\{\pi_i^{2-p+\beta(p-1)}, \pi_i^{1-p/2+\beta(p-1)} v_i^{p/2}\} \\ &\leq 4 \sum_i \pi_i^{1-p\alpha/2} v_i^{p\alpha/2} \leq 4 d^{p\alpha/2}, \end{aligned} \tag{5.4}$$

where the middle inequality uses $\min\{x, y\} \leq x^\beta y^\alpha$. □

Algorithm 1 Hybrid FTRL

Require: $\mathcal{A}, p, T, M$; set $q = p/(p-1)$, $K = d^{q/2}$.
if $N \le K$ **then**
  $\beta = 1/q$, $\lambda = 8M^{1/p}T^{1/p}$, $\gamma = 0$.
else
  $L = \log(eN/K)$, $\beta = \min\{1/(2q), 1/L\}$, $h = 1 - q\beta$.
  If $T < L$ or the action span has rank zero, play a fixed action in every round and stop.
  Choose a fixed G-optimal design ν on $\mathcal{A}$: $\max_i a_i^\top Q_\nu^\dagger a_i \le \operatorname{rank}(\mathcal{A})$.
  $\lambda = 8M^{1/p}d^{h/2}(T/L)^{1/p}$, $\gamma = \frac{1}{4}(L/T)^{2/(ph)}$.
end if
$\alpha = 1 - \beta$, $\eta = \beta/(8\lambda)$; initialize w_1 uniformly.
for $t = 1, \dots, T$ **do**
  If $N \le K$: set $\pi_t = w_t$, $m_{t,i} = w_{t,i}^{1-p}$, and select the direct estimate $D_{t,i}$ for every i.
  Otherwise: set $\pi_t = (1-\gamma)w_t + \gamma\nu$; compute $Q_t, \ell_{t,i}$ and select $D_{t,i}$ if $\pi_{t,i}^{1-p} \le \ell_{t,i}^{p/2}$, else $Z_{t,i}$. Set $m_{t,i}$ to the smaller bound.
  Draw $I_t \sim \pi_t$, observe W_t, and evaluate the selected estimates.
  Compute $s_{t,i}, b_{t,i}, z_{t,i}$ using (5.3) and update w_{t+1} by FTRL.
end for

With direct estimates alone, S_t can grow polynomially in N, as in the independent-arm analysis; the minimum in the first line of (5.4) trades the two costs off action by action. This coordinatewise interpolation lets the upper bound attain $F(d, N)$, including when N is of order $K \log d$ (Section 3.3).

From the effective moment cost to regret. The exploration design and the two moment bounds guarantee $b_{t,i} \le s_{t,i}$ even for arbitrarily small positive $w_{t,i}$, so $\eta|z_{t,i}| \le (\beta/4)w_{t,i}^{-\beta}$, the condition required by the Tsallis stability analysis. For every fixed action j, the virtual regret of the FTRL iterates satisfies

$$\mathbb{E}\sum_{t=1}^{T}\langle w_t - e_j, \mu_t\rangle \le 8\lambda\frac{N^\beta - 1}{\alpha\beta} + 4M\lambda^{1-p}\sum_{t=1}^{T}\mathbb{E}S_t. \tag{5.5}$$

Mixing in the design distribution adds at most $2\gamma T$ to this nonnegative upper bound. In the large branch, $(N^\beta - 1)/(\alpha\beta) \le C_p K^\beta L$, and the identities $h + q\beta = 1$ and $h(1-p) + p\alpha = 1$ make both terms of (5.5) of order $M^{1/p}\sqrt{d}\,L^{1/q}T^{1/p}$ under the stated tuning. Since $\log(1 + N/K) \le L \le 2\log(1 + N/K)$ for $N > K$, this is the factor in (3.2). The small branch gives $O_p(M^{1/p}N^{1/q}T^{1/p})$, and bounded means supply the cap $2T$, which also covers the short-horizon branch. Appendix B gives the details without assuming bounded observations.

6 Conclusion

We determined the minimax regret of finite-action heavy-tailed linear bandits jointly in the dimension and the number of actions, for every fixed polynomial range of cardinalities. The key is a lower bound that separates information shared across actions from direct evidence at the optimum, complemented by a hybrid estimator on the upper side. The rate shows that linear structure helps only once N exceeds $d^{q/2}$.

Our characterization concerns the worst action geometry under stationary, action-dependent weak-moment laws, with constants depending on p and B. Several questions remain open: extending the characterization beyond polynomial cardinalities, determining the minimax regret of each action geometry separately, and establishing the same lower bound under common additive, bounded, or Gaussian noise.

AI USE STATEMENT

In this work, we used generative AI tools mainly for polishing the writing, searching the literature, and assisting with proofs. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

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

Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniINF: Best-of-both-worlds algorithm for parameter-free heavy-tailed MABs. In *International Conference on Learning Representations*, 2025.

Duo Cheng, Xingyu Zhou, and Bo Ji. Taming heavy-tailed losses in adversarial bandits and the best-of-both-worlds setting. In *Advances in Neural Information Processing Systems*, 2024.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Ronald A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4–6):918–925, 2007.

Hamish Flynn, David Reeb, Melih Kandemir, and Jan R. Peters. Improved algorithms for stochastic linear bandits using tail bounds for martingale mixtures. In *Advances in Neural Information Processing Systems*, 2023.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, 2022.

Shinji Ito. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice. In *Advances in Neural Information Processing Systems*, 2024.

David Janz, Alexander E. Litvak, and Csaba Szepesvári. Ensemble sampling for linear bandits: Small ensembles suffice. In *Advances in Neural Information Processing Systems*, 2024.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to Bayesian optimization. In *International Conference on Machine Learning*, 2024.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Tor Lattimore and Csaba Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, 2021.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems*, 2020.

Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, 2019.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, 2016.

Igal Sason. On data-processing and majorization inequalities for f-divergences with applications. *Entropy*, 21(10):1022, 2019.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, 2018.

Artin Tajdini, Jonathan Scarlett, and Kevin Jamieson. Improved regret bounds for linear bandits with heavy-tailed rewards. arXiv preprint arXiv:2506.04775v3, 2026.

Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-tailed linear bandits: Huber regression with one-pass update. In *International Conference on Machine Learning*, 2025.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. arXiv preprint arXiv:2007.10493v2, 2020.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *International Joint Conference on Artificial Intelligence*, 2020.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems*, 2023.

Canzhe Zhao, Shinji Ito, and Shuai Li. Heavy-tailed linear bandits: Adversarial robustness, best-of-both-worlds, and beyond. arXiv preprint arXiv:2508.13679v1, 2025.

Han Zhong, Jiayi Huang, Lin F. Yang, and Liwei Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, 2021.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A A CUMULATIVE-REGRET LOWER BOUND BY ADAPTIVE LIST TESTING

Throughout this appendix, $1 < p \le 2$ and $q = p/(p-1) \ge 2$. We work in the stationary, action-dependent kernel model: each reward is drawn with fresh randomness, has mean $a^\top \theta$, and has centered p-th moment at most one. All actions and parameters constructed below have Euclidean norm at most one. The action set is fixed independently of the horizon and the policy, whereas the magnitude of the hard parameter and the reward support may depend on the horizon.

The argument separates two sources of information. An off-diagonal channel has uniformly small forward relative entropy, even after adaptive histories. A diagonal channel can distinguish the hidden index only when a rare gate, independent of the parameter, opens while the selected action is the one with that index. The actions played while this gate is open form a random list, whose expected cardinality enters the information inequality.

A.1 A LOWER BOUND FOR ANY LOW-COHERENCE CODE

This subsection proves Theorem 4.1, restated below, and supplies the calculations omitted from Lemmas 4.2 and 4.3.

Theorem A.1 (Adaptive list lower bound). *Let $d \ge 1$ and $N \ge 2$ be integers, and let $v_1, \dots, v_N \in \mathbb{R}^d$ be unit vectors satisfying*

$$|v_i^\top v_j| \le \kappa \le \frac{1}{2} \quad (i \ne j), \qquad \kappa > 0. \tag{A.1}$$

Put $H = \kappa^{-q}$ and $\mathcal{A} = \{v_1, \dots, v_N\}$. For every integer $T \ge 1$,

$$\mathfrak{R}_T(\mathcal{A}) \ge c_p \min\left\{T, \left[H \log\left(1 + \frac{N}{H}\right)\right]^{1/q} T^{1/p}\right\}. \tag{A.2}$$

The constant $c_p > 0$ depends only on p. The hard reward laws can be chosen with a common five-point support across all actions and hard parameters, and the conclusion holds against every adaptive randomized policy.

For $N \le H$, the right-hand side of (A.2) is of order $\min\{T, N^{1/q} T^{1/p}\}$. As N/H grows, its nontrivial factor behaves like $\kappa^{-1}[\log(N\kappa^q)]^{1/q}$. The proof below retains the linear cap throughout.

Proof. Fix a policy and let U be uniform on $[N] = \{1, \dots, N\}$. Set $\theta_U = \rho v_U$, where $0 < \rho \le 1/32$ is specified below. Write $I_t \in [N]$ for the index of the action selected at round t, so that $X_t = v_{I_t}$.

The five-point reward kernel. For an action index i, split the mean into

$$g_U(i) = \rho v_i^\top v_U \, \mathbf{1}\{i \ne U\}, \qquad h_U(i) = \rho \mathbf{1}\{i = U\}. \tag{A.3}$$

Thus $g_U(i) + h_U(i) = v_i^\top \theta_U$, $|g_U(i)| \le \rho\kappa$, and $0 \le h_U(i) \le \rho$. Define

$$\ell = \rho\kappa, \qquad r_\ell = (16\ell)^q, \qquad r_\rho = (16\rho)^q, \qquad B_\ell = \frac{2\ell}{r_\ell}, \qquad B_\rho = \frac{2\rho}{r_\rho}. \tag{A.4}$$

At action v_i, the true reward kernel $P_{U,i}$ is

reward value	probability
$+B_\ell$	$r_\ell/2 + g_U(i)/(2B_\ell)$
$-B_\ell$	$r_\ell/2 - g_U(i)/(2B_\ell)$
$+B_\rho$	$r_\rho/2 + h_U(i)/(2B_\rho)$
$-B_\rho$	$r_\rho/2 - h_U(i)/(2B_\rho)$
0	$1 - r_\ell - r_\rho$.

(A.5)

Each nonzero probability lies between one quarter and three quarters of the mass of the corresponding channel. Moreover, $r_\rho \le 1/4$, $r_\ell \le r_\rho$, and the zero probability is at least $1/2$. Since $0 < \kappa < 1$ and $q > 1$, the two nonzero magnitudes are distinct, so the support has exactly five points and is common to all U, i.

The mean is $g_U(i) + h_U(i) = v_i^\top \theta_U$. Sign tilting leaves the absolute moment unchanged, and the identity $q(p-1) = p$ gives

$$\mathbb{E}_{P_{U,i}}|Y|^p = r_\ell B_\ell^p + r_\rho B_\rho^p = 2 \cdot 8^{-p}. \tag{A.6}$$

Consequently, by Jensen's inequality and the elementary p-power inequality,

$$\mathbb{E}|Y - \mathbb{E}Y|^p \le 2^{p-1}\big(\mathbb{E}|Y|^p + |\mathbb{E}Y|^p\big) \le 2^p \mathbb{E}|Y|^p = 2^{1-2p} < 1. \tag{A.7}$$

Drawing from (A.5) afresh at each pull therefore yields an admissible stationary instance with $\|\theta_U\|_2 = \rho \le 1$. The action v_U is the unique optimum, and every other action has gap at least

$$\rho - v_i^\top \theta_U \ge \rho(1-\kappa) \ge \rho/2 \qquad (i \ne U). \tag{A.8}$$

Removing the diagonal signal (information budget in Lemma 4.2). Let $Q_{U,i}$ be the kernel in (A.5) with $h_U(i)$ replaced by zero and $g_U(i)$ retained, and let Q_0 be the kernel with both signals set to zero; Q_0 depends neither on the action nor on the hidden index. The Q_U experiment need not have linear means: it serves only as an auxiliary probability measure. Every true instance P_U, over which the supremum in the model is taken, satisfies the linear-mean and moment constraints.

For each fixed u, i, a direct calculation gives

$$\begin{aligned} \mathrm{KL}(Q_{u,i} \| Q_0) \le \chi^2(Q_{u,i}, Q_0) &= \frac{g_u(i)^2}{B_\ell^2 r_\ell} \\ &= \frac{r_\ell g_u(i)^2}{4\ell^2} \le \frac{r_\ell}{4}. \end{aligned} \tag{A.9}$$

This bound holds pointwise in both the action and the hidden index.

Run the given policy under Q_U. Let $\mathcal{D}$ consist of its complete private random seed, all actions and rewards up to time T, and an independent uniform round $J \in [T]$ that is not revealed to the policy. Write $Q_u^{\mathcal{D}}$ for the law of this transcript given $U = u$, and $Q_0^{\mathcal{D}}$ for the law generated by the same policy under baseline rewards. Put

$$\Lambda = T r_\rho. \tag{A.10}$$

The adaptive chain rule evaluates (A.9) at the actions the policy actually selects, which depend on the history. Decomposing the average relative entropy into the mutual information plus a nonnegative divergence between marginals then yields

$$\begin{aligned} I_Q(U; \mathcal{D}) &\le \frac{1}{N} \sum_{u=1}^{N} \mathrm{KL}(Q_u^{\mathcal{D}} \| Q_0^{\mathcal{D}}) \\ &\le \frac{T r_\ell}{4} = \frac{\Lambda \kappa^q}{4} = \frac{\Lambda}{4H}. \end{aligned} \tag{A.11}$$

In particular, this step does not require the posterior distribution of U to remain uniform after an adaptive history.

A random list and a coupling (Lemma 4.2). Both the true and the auxiliary rewards can be generated by first drawing a channel $G_t \in \{0, \ell, \rho\}$ with

$$\mathbb{P}(G_t = \ell) = r_\ell, \qquad \mathbb{P}(G_t = \rho) = r_\rho, \tag{A.12}$$

independently across rounds and of the parameter, the past, and the action, and then drawing the sign according to (A.5), with or without its diagonal signal as appropriate. The channel is determined by the absolute value of the observed reward, so adjoining the channels to $\mathcal{D}$ does not change its information content.

On the auxiliary trajectory, define the index set

$$L(\mathcal{D}) = \{I_t^Q : G_t = \rho,\ 1 \le t \le T\} \cup \{I_J^Q\}. \tag{A.13}$$

Its expected cardinality satisfies

$$\mathbb{E}_Q|L(\mathcal{D})| \le T r_\rho + 1 = \Lambda + 1. \tag{A.14}$$

The list itself may depend on U through the adaptive observations.

Couple the true and auxiliary trajectories by using the same hidden index, policy seed, channels, sign uniforms, and J. Their conditional reward distributions agree whenever the channel is not ρ, and also when it is ρ but the common selected index differs from U. Let τ be the first time on the auxiliary trajectory with $G_t = \rho$ and $I_t^Q = U$, with $\tau = \infty$ if no such time exists. By induction, the two trajectories, including their selected actions, agree before τ. If $\tau \le T$, then $U \in L(\mathcal{D})$ already. If $\tau > T$, the complete trajectories agree, so $I_J^P = U$ implies $I_J^Q = U \in L(\mathcal{D})$. This proves the pathwise inclusion and its probabilistic consequence

$$\{I_J^P = U\} \subseteq \{U \in L(\mathcal{D})\}, \qquad \mathbb{P}_P(I_J^P = U) \le \mathbb{P}_Q\{U \in L(\mathcal{D})\}. \tag{A.15}$$

Nothing needs to be assumed about the true trajectory after the two trajectories separate.

An information inequality for a list of random size (Lemma 4.3). Let $s = \mathbb{P}_Q\{U \in L(\mathcal{D})\}$. Under the product of the reference marginals of U and $\mathcal{D}$, the same event has probability

$$v = \frac{\mathbb{E}_Q|L(\mathcal{D})|}{N} \le \frac{\Lambda + 1}{N}. \tag{A.16}$$

Applying data processing to the indicator of this event gives

$$I_Q(U; \mathcal{D}) \ge \mathrm{kl}(s\|v) \ge s \log(1/v) - \log 2. \tag{A.17}$$

The second inequality follows by expanding the binary relative entropy, bounding the binary entropy by $\log 2$, and using $-(1-s)\log(1-v) \ge 0$. Combining (A.11)–(A.17) yields, whenever $\Lambda + 1 < N$,

$$s \le \frac{\Lambda/(4H) + \log 2}{\log(N/(\Lambda + 1))}. \tag{A.18}$$

This calculation uses the expected size of the random list directly; it requires neither a deterministic bound on its size nor independence between the list and the hidden index under the reference joint law.

Calibration. Take $c_0 = 1/64$ and set

$$\rho = \min\left\{\frac{1}{32}, \frac{1}{16}\left(\frac{c_0 H \log(1 + N/H)}{T}\right)^{1/q}\right\}. \tag{A.19}$$

In both branches of this minimum, $\Lambda = T(16\rho)^q \le c_0 H \log(1 + N/H)$. Put $R = N/H$.

First suppose that $R \ge 16$. Since $H \ge 4$ and $\log(1 + R) \le \sqrt{R}$,

$$\Lambda + 1 \le H\{c_0 \log(1 + R) + 1/H\} \le H\sqrt{R}, \qquad \log \frac{N}{\Lambda + 1} \ge \frac{1}{2} \log R. \tag{A.20}$$

The middle inequality follows, for example, from $1/H \le 1/4 \le \sqrt{R}/16$ and $c_0 + 1/16 < 1$. Hence (A.18) applies, and using $\log(1 + R) \le 2 \log R$ and $\log R \ge 4 \log 2$ in that bound gives

$$\mathbb{P}_P(I_J^P = U) \le s \le c_0 + \frac{1}{2} < \frac{3}{4}. \tag{A.21}$$

Now suppose that $0 < R \le 16$. In this case we compare the true transcript directly with the parameter-independent baseline. At a common action i, the lower bounds on the atom probabilities in (A.5) imply

$$\begin{aligned} \mathrm{KL}(Q_0\|P_{u,i}) &\le \chi^2(Q_0, P_{u,i}) \\ &\le \frac{r_\ell g_u(i)^2}{2\ell^2} + \frac{r_\rho h_u(i)^2}{2\rho^2} \\ &\le \frac{r_\rho}{2}\left(\kappa^q + \mathbf{1}\{i = u\}\right). \end{aligned} \tag{A.22}$$

Under the baseline experiment, every adaptively selected action is independent of a fresh, uniformly averaged hidden index. The chain rule in the baseline-to-true direction therefore gives

$$\frac{1}{N}\sum_{u=1}^{N} \mathrm{KL}(Q_0^{\mathcal{D}}\|P_u^{\mathcal{D}}) \le \frac{\Lambda}{2}\left(\frac{1}{H} + \frac{1}{N}\right). \tag{A.23}$$

For each u, compare the probability of the event $\{I_J = u\}$ under the two transcript laws. Pinsker's inequality and Jensen's inequality give

$$\begin{aligned}\mathbb{P}_P(I_J^P = U) &\le \frac{1}{N} + \frac{1}{N}\sum_{u=1}^{N} \mathrm{TV}(Q_0^{\mathcal{D}}, P_u^{\mathcal{D}}) \\ &\le \frac{1}{N} + \sqrt{\frac{\Lambda}{4}\left(\frac{1}{H} + \frac{1}{N}\right)}.\end{aligned} \tag{A.24}$$

Reverse relative entropy is used here only to bound total variation. The function $(1+1/R)\log(1+R)$ is increasing, since its derivative is $(R - \log(1+R))/R^2 > 0$, and its value at $R = 16$ is less than four. Thus the square-root term in (A.24) is at most $\sqrt{c_0} = 1/8$. Since $N \ge 2$,

$$\mathbb{P}_P(I_J^P = U) \le \frac{5}{8} < \frac{3}{4}. \tag{A.25}$$

In both cases, the prior-averaged probability of a mistake at the independent uniform round is at least $1/4$. By (A.8),

$$\mathbb{E}_U \mathbb{E}_{P_U} R_T \ge \frac{\rho T}{2} \mathbb{P}_P(I_J^P \ne U) \ge \frac{\rho T}{8}. \tag{A.26}$$

At least one fixed index u incurs regret no smaller than this prior average. Substituting (A.19) proves (A.2), including the linear cap, for every horizon. The action set and the prior family are chosen before the policy; only the final choice of a hard member may depend on the policy. □

A.2 Polynomial-Size Codes from Finite-Field Graphs

We use the normalized polynomial graph construction of DeVore (2007, Section 3, Theorem 3.1) and include the elementary counting and coherence argument so that the bandit application is self-contained. Fix an integer $b \ge 1$, and let k be a prime power with $b < k$. Index the coordinates of $\mathbb{R}^{k^2}$ by pairs $(x, y) \in \mathbb{F}_k^2$. For each polynomial f over $\mathbb{F}_k$ of degree at most b, define

$$v_f(x, y) = k^{-1/2} \mathbf{1}\{y = f(x)\}. \tag{A.27}$$

There are k^{b+1} such polynomials. Each vector has exactly k nonzero coordinates and hence unit norm. Distinct polynomials give distinct vectors, because their difference is a nonzero polynomial of degree less than k and therefore cannot vanish on all of $\mathbb{F}_k$. The root bound for polynomials yields, for $f \ne g$,

$$0 \le v_f^\top v_g = \frac{1}{k} \#\{x \in \mathbb{F}_k : f(x) = g(x)\} \le \frac{b}{k}. \tag{A.28}$$

This gives k^{b+1} unit vectors in dimension k^2 with coherence at most b/k. If $k \ge 2b$, Theorem A.1 applies to any subset of $N \le k^{b+1}$ of these vectors, with

$$H = (k/b)^q. \tag{A.29}$$

In the application below, the polynomial degree is fixed, and its contribution to the coherence constant enters only through the dependence of the constants on the cardinality exponent.

A.3 The Lower Bound over a Fixed Polynomial Cardinality Range

Theorem A.2 (Polynomial-cardinality lower bound)**.** *Fix $B > 0$. There are constants $c_{p,B} > 0$ and an integer $d_0(B)$ such that, for every integer $d \ge d_0(B)$, every integer $2 \le N \le d^B$, and every integer $T \ge 1$,*

$$\mathfrak{R}_T(d, N) \ge c_{p,B} \min\left\{T, \sqrt{d}\left[\log\left(1 + \frac{N}{d^{q/2}}\right)\right]^{1/q} T^{1/p}\right\}. \tag{A.30}$$

For each such d, N, the lower bound is witnessed by a fixed action set of exactly N unit vectors, independent of the horizon and policy.

Proof. Take $b = \lceil 2B \rceil$, and let k be the largest power of two not exceeding $\sqrt{d}$. Then

$$\sqrt{d}/2 < k \le \sqrt{d}.$$

For all sufficiently large d, depending only on B, we have $k \ge 2b$ and $k^{b+1} \ge d^B$. Indeed, $(b+1)/2 - B \ge 1/2$, so $k^{b+1} > 2^{-(b+1)} d^{(b+1)/2} \ge d^B$ once d is large enough. Choose $d_0(B)$ to enforce both inequalities. A field with k elements exists because k is a power of two.

Construct the vectors in (A.27), keep any N of them, and pad their coordinates with zeros to place them in $\mathbb{R}^d$. The coherence proxy $\kappa = b/k$ satisfies $0 < \kappa \le 1/2$ and gives

$$H = (k/b)^q \ge (2b)^{-q} d^{q/2}. \tag{A.31}$$

For fixed N, the function $x \log(1 + N/x)$ is increasing on $(0, \infty)$, since its derivative is $\log(1 + N/x) - N/(x + N) > 0$. Put $c = (2b)^{-q} \le 1$. Then (A.31) implies

$$\begin{aligned} H \log(1 + N/H) &\ge c d^{q/2} \log\left(1 + \frac{N}{c d^{q/2}}\right) \\ &\ge c d^{q/2} \log\left(1 + \frac{N}{d^{q/2}}\right). \end{aligned} \tag{A.32}$$

Applying Theorem A.1 and absorbing $c^{1/q}$ into the constant proves (A.30), with its linear cap unchanged. □

Small cardinalities and scope. When $N \le d^{q/2}$, the bound in (A.30) is of order $\min\{T, N^{1/q} T^{1/p}\}$, since $x/2 \le \log(1 + x) \le x$ for $0 \le x \le 1$. The constants in Theorem A.2 may depend on the fixed exponent B; the theorem does not provide constants that are uniform in a cardinality exponent growing with the dimension. Moreover, the construction concerns the specified finite action set. Adding arbitrary actions, for instance the whole Euclidean ball, can destroy both the off-diagonal projection bound and the single-index diagonal mechanism, so the same logarithmic lower bound does not follow for the enlarged action set.

B Hybrid FTRL: Complete Proof of the Upper Bound

This appendix gives a self-contained proof of the upper bound, including the small-cardinality branch. Clipped, bonus-corrected FTRL and Tsallis regularization with an adjustable exponent have antecedents in Ito (2024); Zhao et al. (2025). The small-cardinality rate is known from the heavy-tailed multi-armed bandit literature (Bubeck et al., 2013; Wei & Srivastava, 2020); our proof of it does not invoke an external MOSS guarantee. The step responsible for the cardinality transition is the coordinatewise interpolation between direct and linear estimates in (B.31).

B.1 Precise Adversarial Statement and Stationary Consequence

We work with losses throughout this appendix. Let $p \in (1, 2]$, $q = p/(p-1)$, $T \ge 1$, and $\mathcal{A} = \{a_1, \dots, a_N\} \subset B_2^d$, where $N \ge 2$ and $d \ge 1$. An oblivious adversary fixes, before play, probability kernels $P_{t,i}$ and vectors $\vartheta_t \in \mathbb{R}^d$ such that

$$\mu_{t,i} := \mathbb{E}_{P_{t,i}} W = a_i^\top \vartheta_t \in [-1, 1], \qquad \mathbb{E}_{P_{t,i}} |W|^p \le M \quad (t \le T,\ i \le N). \tag{B.1}$$

Here $M > 0$ is known. If the index selected at round t is I_t, the observation has conditional law

$$\mathcal{L}(W_t \mid \mathcal{H}_{t-1}, I_t) = P_{t,I_t},$$

with fresh exogenous sampling randomness. The history contains the policy's past actions, observations, and internal randomness. Thus the assumption is that the displayed moment bounds hold conditionally on the full past and the selected action, not merely marginally. No bound on $\|\vartheta_t\|_2$ is needed beyond the bounded means in (B.1).

The following theorem restates Theorem 5.1; the policy is Hybrid FTRL (Algorithm 1), whose two branches are specified in full in Appendices B.3 and B.4.

Theorem B.1 (Oblivious adversarial upper bound)**.** *Under* (B.1)*, there is a policy knowing $p, M, T, \mathcal{A}$ such that, for every fixed index $j \in \{1, \dots, N\}$,*

$$\mathbb{E}\sum_{t=1}^{T}(\mu_{t,I_t} - \mu_{t,j}) \le C_{p,M} \min\left\{T, \sqrt{d}\,[\log(1 + N/d^{q/2})]^{1/q} T^{1/p}\right\}. \tag{B.2}$$

The same policy and constant work for every fixed index. In particular, the bound applies to $j_* \in \arg\min_j \sum_{t=1}^{T} \mu_{t,j}$, with ties broken deterministically, and hence bounds the expected regret against the best fixed action. The constant depends only on p and M.

The comparator in this theorem is fixed before the randomness of the policy and of the rewards is realized; we make no claim for a comparator chosen from the realized mean sequence of an adaptive adversary. When $M = 0$, all losses and means vanish and the conclusion is immediate. All inverses used below are restricted to the linear span of $\mathcal{A}$.

Corollary B.2 (Stationary centered-moment model)**.** *Suppose $\mathcal{A} \subset B_2^d$ is finite with $|\mathcal{A}| = N \ge 2$, and $\theta \in \mathbb{R}^d$ satisfies $\|\theta\|_2 \le 1$. At each pull of a_i, let the reward be a fresh draw from an arbitrary stationary, action-dependent kernel with mean $a_i^\top \theta$ and centered p-th moment at most one. Then there is a policy knowing $p, T, \mathcal{A}$ whose expected cumulative pseudo-regret satisfies*

$$\mathbb{E}\sum_{t=1}^{T}\left(\max_{a\in\mathcal{A}} a^\top\theta - a_{I_t}^\top\theta\right) \le C_p \min\left\{T, \sqrt{d}\,[\log(1 + N/d^{q/2})]^{1/q} T^{1/p}\right\}. \tag{B.3}$$

Proof. Set $W_t = -Y_t$ and $\vartheta_t = -\theta$. Conditionally on any history and on $I_t = i$, the centered-moment and bounded-mean assumptions give

$$\begin{aligned}\mathbb{E}[|W_t|^p \mid \mathcal{H}_{t-1}, I_t = i] &\le 2^{p-1}\left(\mathbb{E}[|Y_t - a_i^\top\theta|^p \mid \mathcal{H}_{t-1}, I_t = i] + |a_i^\top\theta|^p\right)\\ &\le 2^p.\end{aligned}$$

Apply Theorem B.1 with $M = 2^p$ and a fixed mean-optimal action. The reward kernel may vary arbitrarily across actions; no common additive-noise law is required. □

B.2 A Common Tsallis Stability and Clipping Calculation

Write $\Delta_N = \{w \in [0,1]^N : \sum_i w_i = 1\}$. For $\beta \in (0,1)$ and $\alpha = 1 - \beta$, we use the regularizer

$$\Psi(w) = -\frac{\sum_{i=1}^{N} w_i^\alpha - 1}{\alpha}, \qquad D_\Psi(x, w) = \Psi(x) - \Psi(w) - \langle \nabla\Psi(w), x - w\rangle. \tag{B.4}$$

The same formula extends Ψ to the nonnegative orthant, and the Bregman divergence is defined whenever w has positive coordinates. For estimated loss vectors z_t and a step size $\eta > 0$, let

$$w_t = \arg\min_{w\in\Delta_N}\left\{\Psi(w) + \eta\sum_{s<t}\langle w, z_s\rangle\right\}. \tag{B.5}$$

Every minimizer has positive coordinates: moving a small mass ε to a zero coordinate decreases the regularizer by an amount of order ε^α, which dominates every linear change of order ε. Strict convexity then gives uniqueness, and in particular w_1 is uniform.

Lemma B.3 (Tsallis one-step stability)**.** *Let w be a positive probability vector. If*

$$\eta|z_i| \le \frac{\beta}{4} w_i^{-\beta} \qquad (i \le N), \tag{B.6}$$

then for every nonnegative vector w',

$$\langle w - w', z\rangle - \frac{1}{\eta} D_\Psi(w', w) \le \frac{4\eta}{\beta}\sum_{i=1}^{N} w_i^{1+\beta} z_i^2. \tag{B.7}$$

Consequently, if (B.6) *holds at every round of* (B.5)*, then every $u \in \Delta_N$ satisfies*

$$\sum_{t=1}^{T}\langle w_t - u, z_t\rangle \le \frac{\Psi(u) - \Psi(w_1)}{\eta} + \frac{4\eta}{\beta}\sum_{t=1}^{T}\sum_{i=1}^{N} w_{t,i}^{1+\beta} z_{t,i}^2. \tag{B.8}$$

For a unit-vector comparator $u = e_j$, the regularizer term is

$$\Psi(e_j) - \Psi(w_1) = \frac{N^\beta - 1}{\alpha}. \tag{B.9}$$

Proof. The left side of (B.7) is at most the sum of separate scalar suprema over $x \geq 0$. Let $f(x) = -x^\alpha/\alpha$, so that $f'(x) = -x^{-\beta}$ and $f''(x) = \beta x^{-1-\beta}$ for $x > 0$. For a coordinate with current value $w > 0$, the scalar objective is

$$(w - x)z - \frac{1}{\eta}\{f(x) - f(w) - f'(w)(x - w)\}.$$

Its unique global maximizer x_* satisfies

$$\frac{x_*}{w} = (1 + \eta z w^\beta)^{-1/\beta}.$$

Indeed, (B.6) ensures $w^{-\beta} + \eta z > 0$, and the objective has infinite positive derivative at zero and tends to minus infinity at infinity. Writing $r = \eta z w^\beta$, we have $|r| \leq \beta/4$ and

$$-\frac{1}{4} \leq -\frac{\log(1 + r)}{\beta} \leq \frac{1}{4(1 - \beta/4)} \leq \frac{1}{3}.$$

The lower and upper inequalities follow from $\log(1 + x) \leq x$ and $-\log(1 - x) \leq x/(1 - x)$, respectively. Hence $w/2 \leq x_* \leq 2w$, so throughout the segment joining w and x_*,

$$f''(x) \geq \frac{\beta}{4w^{1+\beta}}.$$

Taylor's theorem with integral remainder therefore bounds the scalar supremum by

$$\begin{aligned}(w - x_*)z - \frac{\beta(x_* - w)^2}{8\eta w^{1+\beta}} &\leq \sup_{v \in \mathbb{R}} \left\{ -vz - \frac{\beta v^2}{8\eta w^{1+\beta}} \right\} \\ &= \frac{2\eta}{\beta} w^{1+\beta} z^2.\end{aligned}$$

The boundary $x = 0$ is covered by continuity. Summing over coordinates and weakening the constant from two to four gives (B.7).

For completeness, we recall the FTRL telescoping argument. Let $G_t = \sum_{s=1}^t z_s$, $G_0 = 0$, and $\Phi_t = \min_{w \in \Delta_N}\{\Psi(w) + \eta\langle w, G_t\rangle\}$. The optimality condition for w_t in the interior of the simplex gives

$$\Phi_t - \Phi_{t-1} = \eta\langle w_{t+1}, z_t\rangle + D_\Psi(w_{t+1}, w_t).$$

The same condition for w_{T+1} gives, for $u \in \Delta_N$, $\Phi_T = \Psi(u) + \eta\langle u, G_T\rangle - D_\Psi(u, w_{T+1})$. Consequently, we have the exact telescoping identity

$$\begin{aligned}\sum_{t=1}^T \langle w_t - u, z_t\rangle = &\frac{\Psi(u) - \Psi(w_1) - D_\Psi(u, w_{T+1})}{\eta} \\ &+ \sum_{t=1}^T \left\{ \langle w_t - w_{t+1}, z_t\rangle - \frac{1}{\eta} D_\Psi(w_{t+1}, w_t) \right\}.\end{aligned}$$

Discarding the nonnegative terminal divergence and applying (B.7) to each summand yields (B.8). Finally, $\Psi(e_j) = 0$ and $\Psi(w_1) = -(N^\beta - 1)/\alpha$. □

Suppose now that, conditionally on $\mathcal{H}_{t-1}$, the estimators $\widetilde{\mu}_{t,i}$ satisfy

$$\mathbb{E}[\widetilde{\mu}_{t,i} \mid \mathcal{H}_{t-1}] = \mu_{t,i}, \qquad \mathbb{E}[|\widetilde{\mu}_{t,i}|^p \mid \mathcal{H}_{t-1}] \leq M m_{t,i}, \tag{B.10}$$

where each $m_{t,i}$ is determined before the current observation. For a fixed $\lambda > 0$, define

$$\begin{aligned}s_{t,i} &= \lambda w_{t,i}^{-\beta}, & \widehat{\mu}_{t,i} &= \widetilde{\mu}_{t,i}\mathbf{1}\{|\widetilde{\mu}_{t,i}| \leq s_{t,i}\}, \\ b_{t,i} &= M m_{t,i} s_{t,i}^{1-p}, & z_{t,i} &= \widehat{\mu}_{t,i} - b_{t,i}, \qquad \eta = \frac{\beta}{8\lambda}.\end{aligned} \tag{B.11}$$

Assume $b_{t,i} \le s_{t,i}$ for all coordinates and rounds. Then $|z_{t,i}| \le 2s_{t,i}$, so (B.6) holds.

In the next calculations, we suppress t and condition all expectations on $\mathcal{H}_{t-1}$. The clipping error $e_i = \mu_i - \mathbb{E}\widehat{\mu}_i$ satisfies

$$|e_i| \le \mathbb{E}\big[|\widetilde{\mu}_i|\mathbf{1}\{|\widetilde{\mu}_i| > s_i\}\big] \le s_i^{1-p}\mathbb{E}|\widetilde{\mu}_i|^p \le b_i.$$

Since $\mu_i - \mathbb{E}z_i = e_i + b_i \in [0, 2b_i]$, every fixed $u \in \Delta_N$ satisfies

$$\langle w - u, \mu\rangle \le \mathbb{E}\langle w - u, z\rangle + 2\sum_i w_i b_i. \tag{B.12}$$

This one-sided correction for the comparator does not require any regret to be nonnegative. In addition,

$$\begin{aligned}\mathbb{E}\widehat{\mu}_i^2 &\le s_i^{2-p}\mathbb{E}|\widetilde{\mu}_i|^p \le M m_i s_i^{2-p},\\ b_i^2 &\le b_i s_i = M m_i s_i^{2-p},\\ \mathbb{E}z_i^2 &\le 2\mathbb{E}\widehat{\mu}_i^2 + 2b_i^2 \le 4M m_i s_i^{2-p}.\end{aligned} \tag{B.13}$$

We define the quantity that controls both bias and stability by

$$S_t = \sum_i w_{t,i}^{1+\beta(p-1)} m_{t,i}. \tag{B.14}$$

Indeed, (B.11) and (B.13) imply

$$2\sum_i w_{t,i} b_{t,i} = 2M\lambda^{1-p} S_t,$$

$$\frac{4\eta}{\beta}\sum_i w_{t,i}^{1+\beta}\mathbb{E}[z_{t,i}^2 \mid \mathcal{H}_{t-1}] \le \frac{16\eta M\lambda^{2-p}}{\beta} S_t = 2M\lambda^{1-p} S_t.$$

Taking expectations in (B.8) and using (B.12) gives, for every fixed $u = e_j$,

$$\mathbb{E}\sum_{t=1}^T \langle w_t - u, \mu_t\rangle \le 8\lambda\frac{N^\beta - 1}{\alpha\beta} + 4M\lambda^{1-p}\sum_{t=1}^T \mathbb{E}S_t. \tag{B.15}$$

These expectations are well defined. To verify integrability directly, write $G_{t-1} = \sum_{s<t} z_s$. The simplex optimality condition, applied with a coordinate satisfying $w_{t,k} \ge 1/N$, implies

$$w_{t,i}^{-\beta} \le N^\beta + 2\eta\|G_{t-1}\|_\infty, \qquad \|z_t\|_\infty \le 2\lambda\big(N^\beta + 2\eta\|G_{t-1}\|_\infty\big).$$

By induction, these quantities, as well as the clipping thresholds and bonuses, admit a finite deterministic bound at every finite round. Together with (B.10) and $b_{t,i} \le s_{t,i}$, this justifies taking the expectations above.

B.3 Small Cardinality: The Direct-Estimate Branch

Let $K = d^{q/2}$, and first suppose that $N \le K$. Set

$$\beta = \frac{1}{q}, \qquad \alpha = \frac{1}{p}, \qquad \lambda = 8M^{1/p}T^{1/p}, \qquad \eta = \frac{\beta}{8\lambda}. \tag{B.16}$$

Run (B.5), sample I_t directly from w_t, and apply (B.11) with

$$\widetilde{\mu}_{t,i} = \frac{\mathbf{1}\{I_t = i\}W_t}{w_{t,i}}, \qquad m_{t,i} = w_{t,i}^{1-p}. \tag{B.17}$$

This completely specifies the small-cardinality policy, which requires neither a design distribution nor a matrix inverse. Conditional unbiasedness and the moment bound in (B.10) follow by summing over I_t. Since $1 - p + \beta p = 0$,

$$\frac{b_{t,i}}{s_{t,i}} = M\lambda^{-p} w_{t,i}^{1-p+\beta p} = \frac{M}{\lambda^p} = \frac{1}{8^p T} \le 1. \tag{B.18}$$

Furthermore, by concavity of $x \mapsto x^{1/p}$,

$$S_t = \sum_i w_{t,i}^{2-p+\beta(p-1)} = \sum_i w_{t,i}^{1/p} \le N^{1/q}. \tag{B.19}$$

Thus (B.15) gives

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}\langle w_t - u, \mu_t\rangle &\le 8\lambda\frac{N^{1/q}-1}{(1/p)(1/q)} + 4MT\lambda^{1-p}N^{1/q} \\ &\le C_p M^{1/p}N^{1/q}T^{1/p}.\end{aligned} \tag{B.20}$$

Since the actual sampling distribution is w_t, this already bounds the expected regret. Means in $[-1,1]$ also give the deterministic upper bound $2T$ on the cumulative mean-loss difference. Finally, $x/2 \le \log(1+x) \le x$ for $0 \le x \le 1$ yields

$$\frac{N}{2} \le K\log(1+N/K) \le N, \qquad [K\log(1+N/K)]^{1/q} = \sqrt{d}\,[\log(1+N/K)]^{1/q}. \tag{B.21}$$

Combining (B.20) and (B.21) proves (B.2) in this branch. The argument uses only a finite set of arms and conditional moment bounds; in particular, the small-cardinality result does not require linear means.

B.4 Large Cardinality: The Hybrid Branch

Now suppose that $N > K$, and define

$$L = \log(en/K), \qquad \beta = \min\left\{\frac{1}{2q}, \frac{1}{L}\right\}, \qquad \alpha = 1-\beta, \qquad h = 1 - q\beta. \tag{B.22}$$

Then $L \ge 1$, $\alpha \ge 3/4$, and $1/2 \le h < 1$. If $T < L$, play any fixed action throughout; the resulting regret is at most $2T$, and

$$T \le \sqrt{d}\,L^{1/q}T^{1/p}.$$

Since

$$\log(1+N/K) \le L \le 2\log(1+N/K) \qquad (N \ge K), \tag{B.23}$$

this case already achieves the required rate. Hence we assume $T \ge L$ from now on.

Let $V = \mathrm{span}(\mathcal{A})$ and $r = \dim V$. If $r = 0$, all means are zero and so is the regret. Otherwise, choose a distribution $\nu \in \Delta_N$ maximizing $\log\det Q_\nu$ on V, where $Q_\nu = \sum_i \nu_i a_i a_i^\top$. Since some distribution has a positive-definite covariance on V, every determinant maximizer is positive definite there. The directional derivative toward the point mass at a_i gives

$$a_i^\top Q_\nu^{-1} a_i - r \le 0, \qquad \max_i a_i^\top Q_\nu^{-1} a_i \le r \le d. \tag{B.24}$$

This also provides a self-contained justification of the design used by the algorithm. Fix such a maximizer before play.

Set

$$\lambda = 8M^{1/p}d^{h/2}(T/L)^{1/p}, \qquad \gamma = \frac{1}{4}(L/T)^{2/(ph)}, \qquad \eta = \frac{\beta}{8\lambda}. \tag{B.25}$$

Run (B.5) with the parameters in (B.22), and at each round t perform the following steps:

1. Form $\pi_t = (1-\gamma)w_t + \gamma\nu$, and compute

$$Q_t = \sum_i \pi_{t,i} a_i a_i^\top, \qquad \ell_{t,i} = a_i^\top Q_t^\dagger a_i, \qquad m_{t,i} = \min\{\pi_{t,i}^{1-p}, \ell_{t,i}^{p/2}\}. \tag{B.26}$$

   Here the dagger denotes the inverse on V, extended by zero on $V^\perp$. Since $\gamma > 0$, Q_t is positive definite on V.

2. Draw $I_t \sim \pi_t$, observe W_t, and form both estimates

$$D_{t,i} = \frac{\mathbf{1}\{I_t = i\}W_t}{\pi_{t,i}}, \qquad Z_{t,i} = (a_i^\top Q_t^\dagger a_{I_t})W_t. \tag{B.27}$$

   Use $\widetilde{\mu}_{t,i} = D_{t,i}$ if $\pi_{t,i}^{1-p} \le \ell_{t,i}^{p/2}$, and use $\widetilde{\mu}_{t,i} = Z_{t,i}$ otherwise.

3. Compute $s_{t,i}, \widehat{\mu}_{t,i}, b_{t,i}, z_{t,i}$ from (B.11), and pass z_t to the next FTRL update.

The choice between the two estimators depends only on the history and the current sampling distribution, and is made before the reward is observed. Every coordinate can be computed from the single observation (I_t, W_t). The resulting vector need not be a linear function of the features; coordinatewise unbiasedness before clipping suffices.

B.5 Conditional Moments and Effective Dimension

In this subsection, we suppress t and condition on the complete past. The direct estimator has mean μ_i. For the linear estimator,

$$\mathbb{E}[a_I W] = \sum_j \pi_j a_j (a_j^\top \vartheta_t) = Q\vartheta_t, \qquad \mathbb{E}Z_i = a_i^\top Q^\dagger Q \vartheta_t = a_i^\top \vartheta_t = \mu_i.$$

The last equality holds even if ϑ_t has a component outside V. Using the conditional raw-moment bound and $p \le 2$,

$$\begin{aligned}
\mathbb{E}|D_i|^p &\le M\pi_i^{1-p}, \\
\mathbb{E}|Z_i|^p &\le M \sum_j \pi_j |a_i^\top Q^\dagger a_j|^p \\
&\le M \left(\sum_j \pi_j (a_i^\top Q^\dagger a_j)^2 \right)^{p/2} = M\ell_i^{p/2}.
\end{aligned} \tag{B.28}$$

The second inequality is the monotonicity of moments on a probability space. Hence the selected estimator satisfies (B.10) with m_i as in (B.26).

The design yields the following leverage identities and bounds:

$$0 \le v_i := \pi_i \ell_i \le 1, \qquad \sum_i v_i = r \le d, \qquad \ell_i \le d/\gamma, \qquad w_i \le 2\pi_i. \tag{B.29}$$

To verify them, note that $Q \succeq \pi_i a_i a_i^\top$ implies $\pi_i a_i^\top Q^\dagger a_i \le 1$, and summing gives $\sum_i \pi_i \ell_i = \mathrm{tr}(Q^\dagger Q) = r$. The inequality $Q \succeq \gamma Q_\nu$, together with (B.24), gives $\ell_i \le r/\gamma$. Finally $\gamma \le 1/4$ and $\pi_i \ge (1-\gamma) w_i$ imply $w_i \le 2\pi_i$.

For the effective sum (B.14), write

$$A = 2 - p + \beta(p-1), \qquad B = 1 - p/2 + \beta(p-1).$$

Since $1 + \beta(p-1) \le 2$, (B.29) yields

$$S \le 4 \sum_i \min\{\pi_i^A, \pi_i^B v_i^{p/2}\}. \tag{B.30}$$

For nonnegative x, y, we have $\min\{x, y\} \le x^\beta y^\alpha$. Moreover, $\beta A + \alpha B = 1 - p\alpha/2$. Therefore Hölder's inequality, with $0 < p\alpha/2 < 1$, gives

$$\begin{aligned}
S &\le 4 \sum_i \pi_i^{1-p\alpha/2} v_i^{p\alpha/2} \\
&\le 4 \left(\sum_i \pi_i \right)^{1-p\alpha/2} \left(\sum_i v_i \right)^{p\alpha/2} \le 4d^{p\alpha/2}.
\end{aligned} \tag{B.31}$$

This estimate uses both estimators through the minimum in (B.30), and it proves Lemma 5.2.

B.6 The Bonus Is Dominated by the Clipping Threshold

It remains to verify the condition $b_i \le s_i$ required by the common analysis. By (B.11), it is equivalent to $M m_i w_i^{\beta p} \le \lambda^p$. Since $\pi_i \ge w_i/2$ and $\ell_i \le d/\gamma$,

$$m_i w_i^{\beta p} \le \min\left\{ 2^{p-1} w_i^{-ph/q}, (d/\gamma)^{p/2} w_i^{\beta p} \right\}. \tag{B.32}$$

Here $1 - p + \beta p = -ph/q$. Raise the first entry of the minimum to the power $q\beta$ and the second to the power $h = 1 - q\beta$, and multiply; these nonnegative weights sum to one and cancel the power of w_i. Substituting (B.25) then gives

$$\begin{aligned} M m_i w_i^{\beta p} &\le M 2^{p\beta} (d/\gamma)^{ph/2} \\ &= M 2^{p(\beta+h)} d^{ph/2}\, T/L \\ &\le M 2^p d^{ph/2}\, T/L \le \lambda^p. \end{aligned} \tag{B.33}$$

The last line uses $\beta + h = 1 - (q-1)\beta \le 1$. This establishes threshold domination even for arbitrarily small positive action probabilities, so the stability lemma applies at every round of the hybrid branch.

B.7 From Estimated Losses to Regret and Final Tuning

Substituting (B.31) into (B.15) gives, for every fixed $u = e_j$,

$$\mathbb{E} \sum_{t=1}^{T} \langle w_t - u, \mu_t \rangle \le 8\lambda \frac{N^\beta - 1}{\alpha\beta} + 16 M T \lambda^{1-p} d^{p\alpha/2} =: B_T. \tag{B.34}$$

The deterministic quantity B_T is nonnegative. Because the adversary's sequence μ_t and the comparator u are fixed, the actual sampling mixture satisfies the exact relation

$$\begin{aligned} \mathbb{E} \sum_{t=1}^{T} \langle \pi_t - u, \mu_t \rangle &= (1-\gamma)\mathbb{E} \sum_{t=1}^{T} \langle w_t - u, \mu_t \rangle + \gamma \sum_{t=1}^{T} \langle \nu - u, \mu_t \rangle \\ &\le (1-\gamma) B_T + 2\gamma T \le B_T + 2\gamma T. \end{aligned} \tag{B.35}$$

This argument does not assume that $\langle w_t - u, \mu_t \rangle$ is nonnegative, either round by round or in expectation. Moreover, by conditional sampling, $\mathbb{E}(\mu_{t,I_t} - \mu_{t,j}) = \mathbb{E}\langle \pi_t - u, \mu_t \rangle$.

Since $\beta \log(N/K) \le 1$, $\alpha \ge 3/4$, and $\beta^{-1} = \max\{2q, L\} \le 2qL$, we have

$$\frac{N^\beta - 1}{\alpha\beta} \le \frac{eK^\beta}{\alpha\beta} \le C_p K^\beta L. \tag{B.36}$$

The two exponent identities needed to tune (B.34) are

$$h + q\beta = 1, \qquad h(1-p) + p\alpha = 1. \tag{B.37}$$

With $K = d^{q/2}$, they give the explicit bounds

$$\begin{aligned} 8\lambda \frac{N^\beta - 1}{\alpha\beta} &\le C_p M^{1/p} d^{(h+q\beta)/2} L^{1-1/p} T^{1/p} = C_p M^{1/p} \sqrt{d}\, L^{1/q} T^{1/p}, \\ 16 M T \lambda^{1-p} d^{p\alpha/2} &= 16 \cdot 8^{1-p} M^{1/p} d^{[h(1-p)+p\alpha]/2} L^{(p-1)/p} T^{1/p} \\ &\le C_p M^{1/p} \sqrt{d}\, L^{1/q} T^{1/p}. \end{aligned} \tag{B.38}$$

Finally, $ph \le 2$ and $T \ge L$ imply

$$2\gamma T = \frac{1}{2} L (L/T)^{2/(ph)-1} \le \frac{L}{2} \le \frac{1}{2} \sqrt{d}\, L^{1/q} T^{1/p}. \tag{B.39}$$

Equations (B.35)–(B.39), followed by (B.23), establish the second term of (B.2). The same policy always has cumulative mean-loss difference at most $2T$, since every mean lies in $[-1, 1]$. Combining these two upper bounds gives the minimum in the theorem, without any sign assumption on the virtual regret. Together with the small-cardinality and $T < L$ branches, this completes the proof of Theorem B.1.

Attribution and range of the conclusion. At $p = 2$, the leading cardinality dependence agrees with the bounded-feedback upper bound of Ito (2024, Corollary 1), $O(\sqrt{dT \log_+(N/d)})$ with $\log_+ x = \max\{1, \log x\}$, combined with a multi-armed rate. That corollary assumes a sufficiently large horizon and contains a design-dependent exploration remainder, so the comparison concerns

only its leading $\sqrt{T}$ term in the noncontextual case. Zhao et al. (2025) provide the heavy-tailed clipped, bonus-corrected FTRL antecedent and the finite-set upper dependence $\sqrt{d}(\log N)^{1/q}T^{1/p}$. The direct-estimate branch above reproves a known multi-armed rate for completeness. The hybrid estimate (B.31) and its tuning locate the transition at $N = d^{q/2}$. The theorem is an upper bound for every finite action set and supplies the algorithmic half of the polynomial-cardinality match in Theorem 3.1. Its proof uses the action geometry only through the linear means and the covariance matrices on the action span, while the bounded means control exploration and the linear regret cap.

C Details of the Comparison with Previous Bounds

C.1 Transferring the Previous Lower Bound to Our Model

We use Tajdini et al. (2026, Theorem 2 and Appendix E), specifically arXiv:2506.04775v3. Their lower bound has order $d^{1/q}(\log N/\log d)^{1/q}T^{1/p}$ for $d \le N \le 2^{\lfloor d/4 \rfloor}$ and $T \ge 4^q d^q$. The following normalization shows that this rate is a valid baseline under our assumptions of unit-norm actions, unit-norm parameter, and centered moments.

Let k denote the number of blocks, $m \ge 4$ their size, and $D = km \le d$ the dimension of the construction before zero padding. Every action has exactly one entry equal to one in each block, so $\|a\|_2 = \sqrt{k}$. The parameter has one entry 2Δ per block and all remaining entries equal to Δ, where

$$\Delta = \frac{1}{8}k^{(p-2)/p}(T/m)^{-1/q} = \frac{1}{8}k^{2/q-1}m^{1/q}T^{-1/q}.$$

Thus $\|\theta\|_2^2 = k(m+3)\Delta^2$. Set $\bar{a} = a/\sqrt{k}$ and $\bar{\theta} = \sqrt{k}\,\theta$. The actions then have unit norm, and all means and gaps are preserved. The rescaled parameter satisfies

$$\|\bar{\theta}\|_2^q = \frac{8^{-q}k^2 m(m+3)^{q/2}}{T} \le 1 \qquad \text{when } T \ge 4^q d^q. \tag{C.1}$$

Indeed, since $q \ge 2$ and $m \ge 4$,

$$\frac{k^2 m(m+3)^{q/2}}{D^q} = k^{2-q}m^{1-q/2}(1+3/m)^{q/2} \le (7/4)^{q/2}.$$

The same horizon assumption also guarantees the probability condition $\Delta \le 1/(4k)$ of the source construction, since it implies $T \ge 2^{-q}k^2 m$.

The construction provides the raw-moment bound $\mathbb{E}|Y|^p \le 1$. Finally, scale the rewards and the normalized parameter by $1/2$. Jensen's inequality and the elementary power bound give

$$\mathbb{E}|Y/2 - \mathbb{E}(Y/2)|^p \le 2^{-p}2^{p-1}\{\mathbb{E}|Y|^p + |\mathbb{E}Y|^p\} \le 1.$$

This enforces our centered-moment normalization at the cost of a factor of two in regret, and the parameter norm only decreases.

Exact-cardinality transfer. The source theorem provides a set of at most N actions. If exact cardinality is required, add distinct convex combinations of pairs of original actions until the set has N elements. Assign each new action the same convex mixture of the two corresponding original reward laws, before the final reward scaling above. The mean of each new action remains linear, its raw moment remains at most one, its mean never exceeds the best original mean, and its norm is at most one. A policy on the enlarged set can be simulated on the original set: whenever it chooses a new action, sample a mixture component, pull that original action once, and reveal only the resulting reward to the simulated policy. This preserves the observation law and the expected regret, and uses exactly one physical pull per round. Consequently, adding these actions cannot defeat the original lower bound. The resulting set has exactly N actions in the unit ball, although the added actions need not have unit norm. By contrast, our construction in Appendix A directly provides exactly N unit vectors and uses the same set for all horizons.

C.2 Upper Benchmarks and the Explicit Gain

Under our model, $\mathbb{E}|Y|^p \le 2^p$ because the means are bounded by one. Wei & Srivastava (2020, Theorem 7) give an upper bound of order $N^{1/q}T^{1/p} + N$ under a raw-moment assumption, with

constants depending on p and the moment bound. For $T \geq N$, the additive term is absorbed; for $T < N$, the linear cap suffices. Zhao et al. (2025, Theorem 2) give $O_p(\sqrt{d}(\log N)^{1/q}T^{1/p})$, with no logarithm of the horizon. Taking the better of these two guarantees yields the factor G in (3.4). For comparison, Theorem 3 and Lemma 2 of Tajdini et al. (2026) give the explicit upper expression

$$O_p\Big(1 + \sqrt{d}[\log(NT\log^2 T)]^{1/q}T^{1/p}\Big) \qquad (T \geq 2).$$

Table 1 therefore uses the stronger combined benchmark.

For fixed $a > 1$ and $N = d^a$ (or an integer rounding of it), the previous lower factor is $\Theta_{p,a}(d^{1/q})$. If $a \leq q/2$, then $F(d, N) = \Theta_{p,a}(d^{a/q})$, which gives the first gain in (3.5). If $a > q/2$, then $\log(1 + N/d^{q/2}) = \Theta_{p,a}(\log d)$, which gives the second. For sufficiently large d, these comparisons lie within the previous cardinality range, and $T \geq 4^q d^q$ makes the new linear cap inactive. They therefore compare the regret itself, not merely coefficients under an active cap. The new theorem itself applies to every integer $T \geq 1$.