OPTIMAL AND DIMENSION FREE REGRET FOR MULTI-INDEX BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce online learning in the *multi-index bandit* setting, a semiparametric generalization of d-dimensional contextual bandits in which the expected reward depends on the context through multiple (r satisfying $r \ll d$) unknown latent projections, composed with an unknown multivariate reward function. Moving from a single index to multiple latent directions captures richer dependencies, but it also introduces fundamental statistical challenges: the learner must simultaneously recover a latent subspace and learn an unknown multivariate reward function. In this work, we focus on monotone reward functions and develop algorithms with provably sublinear regret.

We first consider the class of coordinate-wise monotone functions. We propose *Coordinate-wise Monotone Multi-Index Bandits* (`CM-MIB`), a variant of the successive elimination algorithm that suitably discretizes the underlying function class, and establish a regret upper bound of $\widetilde{\mathcal{O}}\big(T^{\frac{r+1}{r+2}}\big)$. We further prove a minimax lower bound of $\widetilde{\Omega}\big(T^{\frac{r+1}{r+2}}\big)$, showing that this rate is tight up to logarithmic factors. Notably, the regret becomes nearly linear as r grows, which shows that coordinate-wise monotonicity alone is too weak a structure in high effective dimensions. To address this limitation, we impose a stronger condition on the reward function, namely *entire monotonicity*. Under this assumption, we propose and analyze *Entirely Monotone Multi-Index Bandits* (`EM-MIB`), which achieves a regret of $\widetilde{\mathcal{O}}(T^{3/4})$, where the exponent of T is independent of r and d. Moreover, by using an epoch based scheme along with inverse-gap-weighting scheme, we improve the regret to $\widetilde{\mathcal{O}}(T^{2/3})$. Our analysis employs chaining arguments based on bracketing entropy to show that the entirely monotone least squares estimator attains the near-optimal rate under random design. We further establish a minimax lower bound of $\widetilde{\Omega}(T^{2/3})$, showing that this result cannot be improved. To the best of our knowledge, this work provides the first regret guarantees for multi-index bandits and precisely characterize the trade-offs between structural assumptions and statistical efficiency.

1 INTRODUCTION

Sequential decision-making with structured reward models is a central theme in online learning and bandits. In the contextual bandit setting, a learner has K actions (a.k.a. arms). At each round $t \in [T]$, it selects an action based on observed contextual side information. The goal is to maximize cumulative reward, or equivalently, to minimize regret against the best policy in hindsight. Contextual bandits with linear payoffs are well studied; see, e.g., Chu et al. (2011); Abbasi-yadkori et al. (2011); Li et al. (2010). This work inspired nonlinear extensions through generalized linear bandits (GLBs). GLBs build on generalized linear models (GLMs), where the reward depends on the context through a known reward (or link) function. These models show that underlying structure makes sequential learning both statistically and computationally efficient; see, e.g., Filippi et al. (2010); Li et al. (2017); Zhang et al. (2025).

These models capture many real scenarios. However, even slight misspecification can cause catastrophic performance degradation Ghosh et al. (2017); Bogunovic & Krause (2021). This has driven interest, in statistics and theoretical machine learning, in richer semiparametric reward models. A prominent example is the *single-index model*. Here, the expected reward is an unknown nonlinear

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.

function of a single latent projection of the context. Such models are a fundamental paradigm in semiparametric statistics. They have recently regained attention in learning theory Brillinger (1983); Lee et al. (2024); Damian et al. (2022); Braun et al. (2025); Kovačević et al. (2026). Unlike GLMs, they do not assume a known reward function. This gives much greater modeling flexibility while keeping a simple latent structure. Recent work brings this idea to bandits as *single-index bandits*, i.e., GLBs with unknown reward functions Kang et al. (2025); Arya et al. (2026); Ma & Cai (2025). These works prove sublinear regret for both monotone and general reward functions.

Still, single-index models can be restrictive. In many decision-making problems, rewards depend on several latent factors, not one hidden direction. Examples include recommendation systems, adaptive experimentation, and representation-driven decision making. There, outcomes may arise from interactions among several hidden attributes. A single projection cannot capture this. This motivates *multi-index models*, where rewards depend on multiple latent projections. They are well studied in statistics and gaining attention in theoretical machine learning Friedman & Stuetzle (1981); Li & Duan (1989); Li (1991; 1992). But moving to multiple indices is much harder Braun et al. (2025); Bietti et al. (2023); Mousavi-Hosseini et al. (2024); Abbe et al. (2023). One must learn a latent subspace and an unknown multivariate reward function at the same time. This raises a natural question:

> *Can one design and analyze online learning algorithms with provably sublinear regret for multi-index bandits?*

In this paper, we answer the above question in affirmative under shape restrictions on the unknown reward functions, focusing on multivariate monotonicity. The learner observes a context $X_t \in \mathbb{R}^d$, while the reward may depend on X_t only through a few linear projections. This leads to the multi-index model $Y_t = g_{a_t}(LX_t) + \xi_t$, where $L \in \mathbb{R}^{r\times d}$ with $r \ll d$ and the arm-specific functions $g_a : \mathbb{R}^r \to \mathbb{R}$ are unknown. We call d the *ambient dimension* and r the *effective dimension*. Thus the learner must simultaneously recover the relevant low-dimensional coordinates and learn an unknown multivariate reward function.

Even if the correct r-dimensional coordinates were given, learning the multivariate reward functions remains difficult. We therefore ask how much can be gained from a simple shape restriction such as monotonicity. Unlike the one-dimensional case, however, there are several distinct notions of monotonicity when $r \geq 2$. We focus on two nested classes.

The first is *coordinate-wise monotonicity (CM)*: increasing one coordinate while keeping the others fixed cannot decrease the reward Han et al. (2019). For a differentiable function this means $\partial_i g \geq 0$ for every $i \in [r]$. The second is *entire monotonicity (EM)* Fang et al. (2021). Besides the first-order conditions, it requires every mixed increment across distinct coordinates to be nonnegative. For a smooth function, this means $\partial_S g \geq 0$ for every nonempty $S \subseteq [r]$, where each coordinate in S is differentiated once. Thus in two dimensions an entirely monotone function satisfies $\partial_1 g, \partial_2 g, \partial_{12} g \geq 0$. When $r = 1$, the two classes coincide.

There is a second difficulty that is specific to the unknown multi-index representation. Classical techniques like Sliced inverse regression based subspace estimation Li (1991); Fornasier et al. (2010); Djolonga et al. (2013) identifies the relevant row space, but a shape restriction depends on the basis used inside that subspace. For example, $g(z_1, z_2) = z_1 + z_2$ is nondecreasing in both coordinates. After the reflection $(z_1, z_2) \mapsto (-z_1, z_2)$ the same surface is written as $-z_1 + z_2$, which is decreasing in its first coordinate. Hence recovering the subspace up to rotation is not enough. Our algorithms therefore use a specified monotone orientation: known anchor coordinates fix the basis, while the remaining ambient featues are learned. We do not solve orientation learning from an arbitrary monotone representation.

1.1 Our contributions

Coordinate-wise monotone rewards. We introduce the learning algorithm `CM-MIB`. The reward functions are unknown over a continuous r-dimensional space, so the learner cannot run a separate bandit at every context. We therefore partition the effective space into cells and run one elimination bandit per cell. This creates an approximation error because the reward may vary within a cell. Coordinate-wise monotonicity gives the control needed here: the variation inside one cell can be

large, but the sum of the variations over all cells is small. CM−MIB achieves $\tilde{\mathcal{O}}[(KT)^{(r+1)/(r+2)}]$ regret, which is tight in the exponent of T.

Matching lower bound. We obtain a minimax lower bound showing that the exponent on the horizon T, i.e., $(r+1)/(r+2)$ is unavoidable even with known coordinates (basis). So the growth of the exponent with r comes from learning the rewards, not from estimating the subspace.

Entirely monotone rewards. The stronger shape constraint makes the reward functions statistically easier to learn. With n samples, we prove an $\tilde{\mathcal{O}}(n^{-2/3})$ squared prediction error for entirely monotone regression under random design. The polynomial exponent does not depend on r. The key observation is that an entirely monotone function can be represented as a bounded multiple of a multivariate distribution function. We combine this representation with bracketing-entropy bounds and a least-squares argument to obtain the prediction rate. To our knowledge, this is the first random-design population prediction guarantee of this form for entirely monotone regression and its transfer to bandit data. Fang et al. (2021) derive statistical rates for entirely monotone regression under a fixed design. Our problem, however, involves stochastic contexts, so we need to extend their formulation to the random-design setting. A simple explore-then-commit use of this estimator, EM−MIB, gives $\tilde{\mathcal{O}}(KT^{3/4})$ regret. We then plug the same estimator into an epoch-based inverse-gap-weighting scheme. The resulting algorithm, EM−MIB+, achieves $\tilde{\mathcal{O}}(KT^{2/3})$ regret, matching the minimax lower bound in the exponent of T.

Matching lower bound. We prove an $\Omega(K^{1/3}T^{2/3})$ minimax lower bound that holds even when the projection L is known. Hence the horizon exponent $2/3$ is minimax with known L. With an unknown L, the same holds on any class that contains the lower-bound instances and has a uniform subspace-estimation guarantee. In both cases the exponent does not grow with r.

Simulations. We complement the theory with a broad empirical study on synthetic and real-world benchmarks (Appendix E). The experiments recover the predicted dimension dependence of CM−MIB and show substantial gains over unstructured baselines. Interestingly, on several ordinary entirely monotone instances, the simpler explore–then–commit EM−MIB has lower finite-horizon regret than EM−MIB+, despite its worse worst-case horizon exponent. A separate harder regime, however, shows the opposite behavior: EM−MIB+ outperforms EM−MIB and exhibits the predicted $T^{2/3}$ scaling. The experiments also illustrate the effect of EM misspecification, chart estimation, the number of arms, and real context distributions.

2 Problem formulation and preliminaries

There are K arms, indexed by $[K] = \{1, \dots, K\}$. At round t the learner observes $X_t \in \mathbb{R}^d$, chooses $a_t \in [K]$, and observes

$$Y_t = g_{a_t}(LX_t) + \xi_t, \qquad R_T = \sum_{t=1}^{T} \left\{ \max_{a \in [K]} g_a(LX_t) - g_{a_t}(LX_t) \right\}. \tag{1}$$

Here $L \in \mathbb{R}^{r \times d}$ is common to all arms and $r \ll d$. Both L and the functions $g_a : \mathbb{R}^r \to \mathbb{R}$ are unknown. Throughout, the contexts are i.i.d. and bounded, $\|X_t\| \le C_X$ almost surely, for a fixed constant C_X; the rewards are bounded by $G_{\max}$; the projected design has a bounded density on a compact support; and the noise has a common mean-zero, conditionally σ-sub-Gaussian distribution. The formal statements are given in Assumptions 1–4 of Appendix A. Lipschitz continuity is used only when an estimated projection changes the point at which g_a is evaluated.

Coordinate-wise monotonicity. For $\mathbf{a}, \mathbf{b} \in [0,1]^r$, write $\mathbf{a} \preceq \mathbf{b}$ when $a_j \le b_j$ for every $j \in [r]$. A function $g : [0,1]^r \to \mathbb{R}$ is coordinate-wise monotone if $g(\mathbf{a}) \le g(\mathbf{b})$ whenever $\mathbf{a} \preceq \mathbf{b}$. We write $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ for the measurable coordinate-wise nondecreasing functions on $[0,1]^r$ with values in $[0, G_{\max}]$.

Entire monotonicity. For a nonempty $S \subseteq [r]$ and $\mathbf{a}, \mathbf{b} \in [0,1]^r$ with $a_j < b_j$ for $j \in S$, define the mixed finite difference

$$\Delta_S(f; \mathbf{a}, \mathbf{b}) := \sum_{\mathbf{u} \in \{0,1\}^S} (-1)^{|S| - \sum_{j \in S} u_j} f(\mathbf{z}(\mathbf{u})), \tag{2}$$

where $\mathbf{z}(\mathbf{u})_j = b_j$ if $j \in S$ and $u_j = 1$, $\mathbf{z}(\mathbf{u})_j = a_j$ if $j \in S$ and $u_j = 0$, and the coordinates outside S are fixed. A function $f : [0,1]^r \to \mathbb{R}$ is entirely monotone if $\Delta_S(f; \mathbf{a}, \mathbf{b}) \geq 0$ for every such $S, \mathbf{a}, \mathbf{b}$; the class is denoted by $\mathcal{F}^r_{\mathrm{EM}}$. For $r = 2$, the only new condition beyond coordinate-wise monotonicity is $f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2) \geq 0$.

Canonical coordinates and charts. The row space of L has many matrix representations, while both CM and EM depend on the choice of coordinates inside that row space. We therefore fix a canonical representation. After a fixed, known permutation of the ambient coordinates, take an orthonormal row basis $L_{\mathrm{raw}} = [\, V \mid U \,]$, where $V \in \mathbb{R}^{r \times r}$ is invertible. We assume $\|V^{-1}\|_{\mathrm{op}} \leq \kappa_0$, where κ_0 is a fixed bound controlling the conditioning of this coordinate system. We then define $L = V^{-1} L_{\mathrm{raw}} = [\, I_r \mid W \,]$ and call it the *canonical chart*, with the first r coordinates fixing the monotone orientation and W is unknown.

Let $Z = LX \in \mathbb{R}^r$. Its support is contained in an axis-aligned box $\mathcal{B}$, and we let $S : \mathcal{B} \to [0,1]^r$ be the increasing coordinate-wise affine rescaling. Since both CM and EM are preserved by such a rescaling, we formulate the shape constraints on $[0,1]^r$. When there is no ambiguity, we continue to write g_a for the corresponding rescaled reward function. When L is unknown, we first estimate an orthonormal row basis $\widehat{L}_{\mathrm{raw}}$ and apply the same normalization to obtain $\widehat{L} = [\, I_r \mid \widehat{W} \,]$, provided its leading block is sufficiently well conditioned. We refer to this normalization as the *canonical lock*. For a fixed radius $\rho^* > 0$, any chart $[\, I_r \mid W' \,]$ satisfying $\|W' - W\|_F \leq \rho^*$ is called a *nearby chart*.

We require two uniform conditions over such nearby charts. First, their projected supports lie in a common rectangle on which every g_a is bounded by $G_{\max}$ and satisfies the assumed shape constraint. Second, after coordinate-wise rescaling to $[0,1]^r$, the projected context has density bounded above by M_0'. The precise statements are Assumptions 8 and 9 in Appendix B.

Notation. We use $\widetilde{O}(\cdot)$ to suppress logarithmic factors in the horizon and fixed problem-dependent constants. Dependence on K is displayed explicitly; dependence on r, d and other fixed model parameters may appear in the suppressed factors unless stated otherwise.

3 Coordinate-wise Monotone Multi-Index Bandits

We begin with the function class $\mathcal{F}^r_{\mathrm{CM}}$. We propose a learning algorithm that bins the scaled projected context and runs successive elimination inside each cell.

3.1 Algorithm CM-MIB

The pseudo code is given in Algorithm 1. Fix a grid size m and cell width $h = 1/m$, and partition $[0,1]^r$ into m^r cells $\{C_{\mathbf{k}}\}_{\mathbf{k}=1}^{m^r}$. The basic idea of CM-MIB is to turn the continuous-context problem into one small bandit problem per cell. Whenever the scaled projected context falls in cell $C_{\mathbf{k}}$, the learner uses only the statistics stored for that cell: an active set of arms, a count $n_{\mathbf{k},a}$ and an empirical mean $\widehat{\mu}_a(\mathbf{k})$ for each arm. Among the active arms it plays one with the smallest count, and eliminates arm a once its upper confidence bound is below the lower confidence bound of another active arm. We use $U(n) = (\sigma + G_{\max})\sqrt{2\Lambda/n}$ with $\Lambda = \log(2KT^{r+2})$, with $U(0) = +\infty$ and initial empirical means zero.

When the coordinates are unknown, the first T_0 rounds are used only to construct the coordinate system in which this cell-wise procedure will run. One block estimates the latent row space and applies the canonical lock to obtain $\widehat{L}$; an independent second block estimates the projected support $\widehat{\mathcal{B}}$ and its coordinate-wise rescaling $\widehat{S} : \widehat{\mathcal{B}} \to [0,1]^r$. For a later context we form $\widetilde{Z}_t = \widehat{S}(\Pi_{\widehat{\mathcal{B}}}(\widehat{L}X_t))$, where $\Pi_{\widehat{\mathcal{B}}}$ clips each coordinate to the estimated box, and run the bandit associated with the cell containing $\widetilde{Z}_t$. With known coordinates and a known box, these preliminary steps disappear and we take $T_0 = 0$.

The important point is that, once a cell is fixed, the arm-selection rule uses the cell index and its stored statistics but not the exact location of the context inside that cell. Thus arm a is compared through a single cell average $\mu_a(\mathbf{k}) = \mathbb{E}[g_a(Z) \mid \widetilde{Z} \in C_{\mathbf{k}}]$, rather than by estimating the whole function g_a. The price of this compression is the variation of the rewards within a cell; the analysis below shows that coordinate-wise monotonicity controls the sum of these variations over all cells.

Algorithm 1 CM−MIB

Input: T, K, r, burn-in T_0, grid size m, confidence radius $U(\cdot)$, chart-conditioning bound κ_0.
Chart. If L is known, set $\widehat{L} = L$. Otherwise, use the first $\lceil T_0/2 \rceil$ rounds for uniform exploration and estimate $\widehat{L}_{\mathrm{raw}} = [\,\widehat{V}_{\mathrm{raw}} \mid \widehat{U}_{\mathrm{raw}}\,]$. If $\widehat{V}_{\mathrm{raw}}$ is singular or $\|\widehat{V}_{\mathrm{raw}}^{-1}\|_{\mathrm{op}} > 2\kappa_0$, play a fixed arm thereafter. Otherwise set $\widehat{L} = \widehat{V}_{\mathrm{raw}}^{-1}\widehat{L}_{\mathrm{raw}} = [\,I_r \mid \widehat{W}\,]$.
Box. If the box is known, use it. Otherwise use the remaining $\lfloor T_0/2 \rfloor$ burn-in rounds to form $\widehat{\mathcal{B}}$ from the coordinate-wise minima and maxima of $\widehat{L}X_t$, and freeze the corresponding rescaling map $\widehat{S}$.
for $t = T_0 + 1, \ldots, T$ **do**
  Compute $\widetilde{Z}_t = \widehat{S}(\Pi_{\widehat{\mathcal{B}}}(\widehat{L}X_t))$ and let $C_{\mathbf{k}}$ be its grid cell.
  if $C_{\mathbf{k}}$ is visited for the first time **then**
    Set $\mathcal{A}_{\mathbf{k}} = [K]$ and $n_{\mathbf{k},a} = 0$ for all $a \in [K]$.
  end if
  Play an arm of $\mathcal{A}_{\mathbf{k}}$ with the smallest $n_{\mathbf{k},a}$, breaking ties by index, Observe Y_t and update $n_{\mathbf{k},a_t}$ and mean $\widehat{\mu}_{a_t}(\mathbf{k})$.
  Remove every $a \in \mathcal{A}_{\mathbf{k}}$ with $\widehat{\mu}_a(\mathbf{k}) + U(n_{\mathbf{k},a}) < \max_{b \in \mathcal{A}_{\mathbf{k}}}\{\widehat{\mu}_b(\mathbf{k}) - U(n_{\mathbf{k},b})\}$.
end for

3.2 Regret upper bound

Monotonicity gives no uniform control inside a cell. A monotone function can jump by $G_{\max}$ inside a single cell. We first show that the difference between the supremum and infimum of g on cell $C_{\mathbf{k}}$, when summed over all cells is bounded.

When the projection L is unknown, we use standard inverse-regression conditions. We assume that Σ_X, the covariance of context vector, is positive definite and $\mathbb{E}[X \mid LX]$ is affine in LX. Under uniform exploration, $\Sigma_X^{-1/2}\mathrm{Cov}(\mathbb{E}[X \mid Y, A])\Sigma_X^{-1/2}$ has positive r-th eigenvalue λ_r. Moreover, from n uniform-exploration samples, the chosen subspace estimator satisfies $\mathbb{E}\|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}}^2 \le \frac{C_2 dr}{n\lambda_r}$, where, $n \ge n_{\min}$.

Here P_M denotes projection onto the row space of M, and $C_2, n_{\min}$ are oracle constants. The finite-sample bound is assumed separately from the population conditions and is used throughout the subsequent unknown-chart results.

Theorem 3.1 (CM−MIB regret)**.** *Under the bounded stationary model, anchor-coordinate and nearby-density conditions stated above, suppose the links are CM and Lipschitz, and the subspace estimator satisfies the mean-square guarantee stated above. Then, for $r \ge 1$, by choosing $T_0 \asymp T^{2/3}$, and the cell width $h \asymp (KT)^{-1/(r+2)}$ up to logarithmic factors, Algorithm 1 achieves*

$$\mathbb{E}[R_T] \le \tilde{\mathcal{O}}\Big[(KT)^{(r+1)/(r+2)}\Big].$$

The proofs are Theorems B.4 and B.10 in Appendix B.

3.3 Lower bound

We now look at the minimax lower bound over the class of all coordinate-wise monotone reward functions.

Theorem 3.2 (CM lower bound)**.** *Let Θ_{CM} denote the class of instances with known $L = I_r$, contexts uniformly distributed on $[0,1]^r$, independent Gaussian noise with fixed positive variance, and coordinate-wise monotone reward functions taking values in $[0, G_{\max}]$ and satisfying a fixed positive Lipschitz bound. Then, for every fixed $r \ge 1$ and $K \ge 2$, and for all sufficiently large T,*

$$\inf_{\pi} \sup_{\theta \in \Theta_{\mathrm{CM}}} \mathbb{E}_\theta[R_T(\pi)] \ge c\, T^{(r+1)/(r+2)},$$

where $c > 0$ depends only on the fixed problem parameters. Equivalently, for every policy π, there exists an instance $\theta \in \Theta_{\mathrm{CM}}$ for which the above lower bound holds.

Construction. We now construct a problem instance to discuss the minimax lower bound. The dependence on r is intrinsic to the CM class. We make the problem tractable by giving the learner

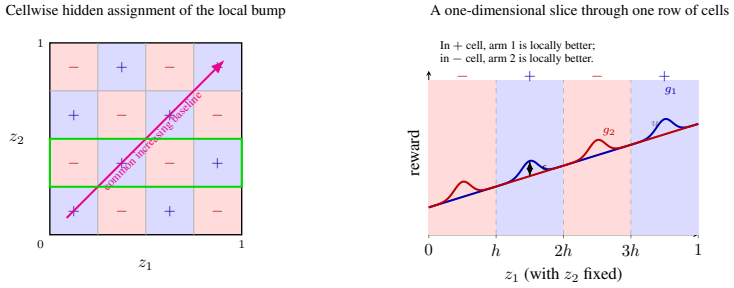


Figure 1: Illustration of the coordinate-wise monotone lower-bound construction. The domain is partitioned into cells of width h. All arms share a common increasing baseline u. Each cell carries a hidden sign: in a $+$ cell, arm 1 receives a local bump of height ϵ; in a $-$ cell, arm 2 receives that bump. Thus the identity of the better arm changes from cell to cell, while both reward functions remain coordinate-wise monotone because the common baseline dominates the local perturbations.

the projection information $d = r$, $L = I_r$, contexts uniform on $[0,1]^r$, and Gaussian noise. Partition the domain into cells of width h. All arms share the increasing baseline $u(z) = \frac{G_{\max}}{2r} \sum_j z_j$. In each cell, an independent hidden sign determines whether arm 1 or arm 2 is better by a local perturbation of height ϵ (see Fig 1). The perturbation vanishes at the cell boundary, and with $\epsilon = G_{\max} h/(8r^2)$, the slope of the baseline dominates its variation, so every arm remains monotone and Lipschitz.

A cell is visited about Th^r times, while distinguishing its hidden sign requires about σ^2/ϵ^2 samples. Choosing $\epsilon^2 Th^r \asymp \sigma^2$ gives $h \asymp T^{-1/(r+2)}$ and $\epsilon \asymp h$. The learner therefore chooses the wrong arm on a constant fraction of rounds, giving regret of order $T\epsilon \asymp T^{(r+1)/(r+2)}$. The full construction is in Appendix B.6.

Remark 3.3 (Intrinsic dimension dependence)**.** *The lower bound already holds when L is known. Hence the exponent $(r+1)/(r+2)$ is not caused by estimating the latent subspace; it is an intrinsic limitation of coordinate-wise monotonicity. In particular, the exponent approaches one as r grows.*

4 Entirely Monotone Multi-Index Bandits

Unlike coordinate-wise monotonicity, which controls only first-order increments, entire monotonicity controls all mixed increments. We show that this extra structure removes r from the polynomial exponent of the horizon.

4.1 Entirely monotone regression under random design

Fang et al. (2021) obtains statistical rates of entirely monotone regression for fixed design. However, as evident from our problem structure, we need to extend the formulation to random design. By Fang et al. (2021) (Lemma C.5), every bounded right-continuous $f \in \mathcal{F}^r_{\mathrm{EM}}$ can be written as

$$f(\mathbf{z}) = f(\mathbf{0}) + \int_{[0,1]^r} \mathbf{1}\{\mathbf{a} \preceq \mathbf{z}\}\, d\nu(\mathbf{a}), \qquad \nu \geq 0, \quad \nu([0,1]^r) = f(\mathbf{1}) - f(\mathbf{0}). \tag{3}$$

We call $\mathbf{z} \mapsto \mathbf{1}\{\mathbf{a} \preceq \mathbf{z}\}$ the *corner indicator* anchored at $\mathbf{a}$. Thus an entirely monotone function is an intercept plus a nonnegative mixture of corner indicators; normalizing ν by its total mass turns this mixture into a multivariate distribution function. Equivalently, an entirely monotone function is an intercept plus a bounded nonnegative multiple of a distribution function. We fit over

$$\mathcal{F}_V = \big\{\beta + f : |\beta| \leq G_{\max},\ f(\mathbf{0}) = 0,\ f(\mathbf{1}) \leq V_{\max},\ f \text{ right-continuous and entirely monotone}\big\},$$

where one may always take $V_{\max} = 2G_{\max}$, since $|g_a| \leq G_{\max}$, its corner difference is at most $2G_{\max}$. Given i.i.d. data $(Z_i, Y_i)_{i=1}^n$, we use any approximate least-squares fit $\widehat{f} \in \mathcal{F}_V$ with

$$\frac{1}{n}\sum_{i=1}^n \big(Y_i - \widehat{f}(Z_i)\big)^2 \leq \inf_{f \in \mathcal{F}_V} \frac{1}{n}\sum_{i=1}^n \big(Y_i - f(Z_i)\big)^2 + \eta_n. \tag{4}$$

Algorithm 2 EM−MIB

Input: T, K, exploration budget $N \le T$.
Pull arms uniformly for the first N rounds, and split these rounds into three independent blocks of about $N/3$ rounds each.
Chart. Estimate $\widehat{L}_{\mathrm{raw}}$ on the first block and apply the canonical lock; if it is rejected, play a fixed arm thereafter.
Box. Form $\widehat{\mathcal{B}}$ and $\widehat{S}$ from the second block. If the box is degenerate, play a fixed arm thereafter.
Fit. For each arm a, fit $\widehat{f}_a \in \mathcal{F}_V$ on its third-block rounds with $\widehat{L}X_t \in \widehat{\mathcal{B}}$, with tolerance $\eta_{n_a} \le n_a^{-2/3}$ (zero function if $n_a = 0$).
for $t = N+1, \dots, T$ **do**
  if $\widehat{L}X_t \in \widehat{\mathcal{B}}$ **then**
    $U_t = \widehat{S}(\widehat{L}X_t)$; play $a_t \in \arg\max_a \widehat{f}_a(U_t)$.
  else
    Play a fixed arm.
  end if
end for

The fast rate comes from *bracketing entropy*, the logarithm of the number of pairs $\ell \le u$ of L_2-width ϵ needed to sandwich every function in the class. For distribution functions it is of order ϵ^{-1} up to logarithms Gao (2013), and we show, by (3) the same holds for $\mathcal{F}_V$ (Lemma C.11). The bound is stated for Lebesgue measure, but a design density bounded by M only enlarges bracket widths by $\sqrt{M}$ (Lemma C.12). A least-squares argument then gives the following bound.

Theorem 4.1 (Entirely monotone regression)**.** *Under a bounded design density and conditionally sub-Gaussian noise, for a bounded truth $g \in \mathcal{F}_V$,*

$$\mathbb{E}\|\widehat{f} - g\|^2_{L_2(P_Z)} \le Cn^{-2/3}\{\log(en)\}^{q_r} + C\eta_n, \qquad q_r = \max\left\{2, \frac{4r-2}{3}\right\}. \tag{5}$$

Remark 4.2 (Dependence on the effective dimension)**.** *The polynomial rate $n^{-2/3}$ does not depend on r. The dimension still enters through logarithmic factors and constants, so the theorem is not uniform in r. This is the main statistical distinction from the coordinate-wise monotone class.*

Remark 4.3. *The regression guarantee is also stable under an estimated chart. If $\|\widehat{W} - W\|_F \le \epsilon_W$, then Lipschitz continuity and $\|X\| \le C_X$ give $|g_a(LX) - g_a(\widehat{L}X)| \le L_g C_X \epsilon_W$. Hence chart estimation contributes a squared misspecification term of order $L_g^2 C_X^2 \epsilon_W^2$. The precise statement is Lemma C.14.*

Computability. The fit in (4) is computationally explicit for fixed r. On a sample of size n, the problem reduces to a finite convex program over corner-indicator functions on the sample grid. A Frank–Wolfe procedure reaches the required optimisation tolerance $\eta_n \le n^{-2/3}$ after $O(n^{2/3})$ iterations, and therefore gives polynomial-time computation for fixed r. The finite reduction, algorithm, and arithmetic complexity are given in Appendix D; see in particular Proposition D.2.

4.2 A SIMPLE EXPLORE-THEN-COMMIT ALGORITHM: EM−MIB

We first use the regression result in a simple explore-then-commit scheme. EM−MIB spends the first N rounds on uniform exploration and splits them into three independent blocks. The first block estimates the latent subspace and converts it to the canonical chart $\widehat{L}$. The second estimates the projected support and the increasing rescaling $\widehat{S} : \widehat{\mathcal{B}} \to [0,1]^r$. On the third block, each arm is pulled uniformly and an entirely monotone regression function $\widehat{f}_a \in \mathcal{F}_V$ is fitted to its observations. After these three steps the estimates are frozen. For every later context inside $\widehat{\mathcal{B}}$, the learner computes $U_t = \widehat{S}(\widehat{L}X_t)$ and plays $a_t \in \arg\max_a \widehat{f}_a(U_t)$. Contexts outside the estimated box are rare and are handled by a fixed fallback arm. Algorithm 2 summarizes the procedure.

The three blocks separate the estimation of the coordinates, the scaling box, and the reward functions. Up to constants, each arm receives N/K regression samples, so Theorem 4.1 gives prediction error of order $(N/K)^{-1/3}$. Thus the exploration and exploitation costs satisfy $\mathbb{E}[R_T] \lesssim N + K^{4/3}TN^{-1/3}$.

Balancing them gives $N \asymp KT^{3/4}$.The chart step uses the same assumed mean-square subspace guarantee as CM-MIB.

Throughout this section, we use the bounded stationary model and canonical coordinates described above. For unknown L, we additionally assume the mean-square subspace guarantee $\mathbb{E}\|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}}^2 \le C_2 dr/(n\lambda_r)$ for $n \ge n_{\min}$.

Theorem 4.4 (EM−MIB regret)**.** *Under the standing conditions above, suppose the reward functions are Lipschitz and entirely monotone and the projected design has bounded density. Then EM−MIB, with $N \asymp KT^{3/4}$ up to logarithmic factors, satisfies*

$$\mathbb{E}[R_T] \le \widetilde{O}\Big(KT^{3/4} + \sqrt{K}\,T^{5/8}\Big).$$

Remark 4.5. *The $T^{3/4}$ rate is a consequence of the one-shot explore–then–commit conversion, not of the regression rate itself. The estimator has squared prediction error $\widetilde{O}(n^{-2/3})$; the loss occurs because EM-MIB freezes the fitted functions after the initial exploration phase.*

4.3 An epoch-based algorithm: EM−MIB+

EM−MIB+ avoids the one-shot limitation of EM−MIB by repeatedly refitting the reward functions and continuing to explore arms that remain competitive.

The chart and scaling box are estimated once at the beginning. A later round is called *retained* when $\widehat{L}X_t \in \widehat{\mathcal{B}}$; rounds outside the box use the fallback arm and are discarded from regression. Retained rounds are divided into epochs of doubling length. At the beginning of epoch m, each arm is refitted using its observations from epoch $m-1$. For a new context, let $\widehat{a}_t$ be the arm with the largest fitted reward. Every other arm a is sampled with probability $p_t(a) = 1/(K + \gamma_m[\widehat{f}_{m,\widehat{a}_t}(U_t) - \widehat{f}_{m,a}(U_t)])$, and the remaining probability is assigned to $\widehat{a}_t$. Hence arms with small estimated gaps continue to be explored, while clearly inferior arms are sampled less often. This is the inverse-gap-weighting rule used in Algorithm 3.

Let $\Xi(n)$ be a known upper envelope for the prediction error; up to logarithmic and fixed model-dependent factors, $\Xi(n) = \widetilde{O}(Kn^{-2/3} + \beta^2)$, where β accounts for chart-estimation error. Inverse-gap weighting converts this squared prediction error into epoch regret of order $2^m\sqrt{K\Xi(2^{m-1})}$, which is $O(K2^{2m/3})$. Summing the doubling epochs therefore gives $\widetilde{O}(KT^{2/3})$ regret.

Two technical issues arise because the data are now collected adaptively. First, the chart estimate must hold with high probability over all epochs. We obtain this by running the mean-square subspace estimator on $O(\log T)$ independent blocks and selecting an estimate supported by a majority of the others. Second, an arm played with probability q_a receives about nq_a samples, while conditioning on that arm can inflate the design density by a factor of order $1/q_a$. The EM regression bound therefore gives conditional squared error of order $q_a^{-1/3}(nq_a)^{-2/3} = q_a^{-1}n^{-2/3}$. In the joint prediction error this is weighted by q_a, so the dependence on q_a cancels. Summing over the K arms gives $\widetilde{\mathcal{O}}(Kn^{-2/3})$ without an arm-conditional density assumption; see Proposition C.30.

Theorem 4.6 (EM−MIB+ regret)**.** *Under the same model conditions as Theorem 4.4, with known upper bounds on the constants used to calibrate Ξ, and with $T_1 \asymp T^{2/3}$, we have $\mathbb{E}[R_T] \le \widetilde{\mathcal{O}}(KT^{2/3})$.*

Thus the epoch-based procedure fully exploits the $n^{-2/3}$ entirely monotone regression rate, yielding $\widetilde{O}(KT^{2/3})$ regret under the standing EM conditions. The polynomial exponent in T remains $2/3$; estimating the chart affects only the dependence on the fixed dimension and model parameters. Under the additional arm-conditional design condition, the dependence on the number of arms improves from K to $K^{5/6}$. The full statements and proofs of both bounds are given in Appendix C.5; see C.6.

4.4 Minimax lower bound

We next show that the $T^{2/3}$ horizon exponent is unavoidable. As in the CM lower bound, we make the problem easier by giving the projection to the learner. Thus the hardness below comes entirely from learning the reward functions.

Algorithm 3 EM−MIB+

Input: T, K, initial length T_1, error envelope $\Xi(\cdot)$.
Chart. Pull arms uniformly on a subspace block of $n_{\mathrm{sub}} \asymp T_1$ rounds. Split it into $k = \lceil 16 \log T \rceil$ sub-blocks, estimate a projector on each, and keep the one closest to a majority of the others. Apply the canonical lock as in Step 3 of Algorithm 2.
Box. Form $\widehat{\mathcal{B}}$ and $\widehat{S}$ from an independent block, as in Step 4 of Algorithm 2. On every later context outside the box, play a fixed arm and discard the observation.
Sample uniformly in the first epoch of two retained rounds.
for epochs $m = 2, 3, \ldots$ of length 2^m **do**
  Fit $\widehat{f}_{m,a}$ for each arm as in Step 5 of Algorithm 2, but on its retained data from epoch $m-1$; set $\gamma_m = \frac{1}{2}\sqrt{K/\Xi(2^{m-1})}$.
  for each retained round t of epoch m **do**
    $U_t = \widehat{S}(\widehat{L}X_t), \quad \widehat{a}_t \in \arg\max_a \widehat{f}_{m,a}(U_t)$.
    For $a \neq \widehat{a}_t$, set $p_t(a) = 1/\big(K + \gamma_m[\widehat{f}_{m,\widehat{a}_t}(U_t) - \widehat{f}_{m,a}(U_t)]\big)$; give the rest to $\widehat{a}_t$. Sample $a_t \sim p_t$.
  end for
end for

Theorem 4.7 (EM minimax lower bound)**.** *Let Θ_{EM} denote the bounded Lipschitz EM instances satisfying the standing model, chart, variation, linearity, and population SIR-identifiability conditions. No subspace-estimation oracle is assumed. Then, for $K \geq 2$ and all sufficiently large T,*

$$\inf_{\pi} \sup_{\theta \in \Theta_{\mathrm{EM}}} \mathbb{E}_\theta[R_T(\pi)] \geq cK^{1/3}T^{2/3},$$

where $c > 0$ depends only on the fixed problem parameters. The bound already holds in the easier setting where L is revealed to the learner.

Construction. Take $L = [\, I_r \mid 0 \,]$ and rescale the projected context to $Z \in [0,1]^r$. All arms share the increasing additive baseline $\sum_{i=1}^r w_i Z_i$ with distinct positive weights w_i. We partition the first coordinate into N slabs. Each slab j has an independent hidden label $\omega_j \in [K]$ that names the better arm there. That arm receives a trapezoidal bump of height $\epsilon \asymp 1/N$ supported inside the slab, and the other arms keep only the baseline.

The bump is chosen so that baseline plus bump remains nondecreasing in z_1. Every reward is additive across coordinates, so all mixed differences of order at least two vanish, and hence every reward is entirely monotone. The distinct weights play no role in the testing argument. They keep every latent direction visible to population inverse regression, so the SIR-identifiability condition also holds. Section C.4 verifies all model conditions.

Each slab is visited about T/N times. On a constant fraction of these visits, the learner cannot tell which of the K arms carries the Gaussian mean shift ϵ when $\epsilon^2 T/N \lesssim K\sigma^2$. With $\epsilon \asymp 1/N$, this gives the hard scale $N \asymp (T/K)^{1/3}$. The learner loses order ϵ on each ambiguous visit, so the regret is of order $T\epsilon \asymp K^{1/3}T^{2/3}$.

The formal lower-bound construction and testing argument are given in Theorem C.23. Together with Theorem 4.6, this shows that the upper and lower bounds match in the horizon exponent $2/3$, independently of r. For known L, this yields the minimax horizon rate $\widetilde{\Theta}(T^{2/3})$. For unknown L, the same conclusion holds conditional on the stated uniform subspace-estimation guarantee. The dependence on K remains open: $K^{1/3}$ in the lower bound versus K in the general upper bound.

5 Conclusion and Future Work

We propose the setup of multi-index bandits and for special case of monotone functions, obtain regret bounds. The bounds are corroborated by matching minimax lower bounds. One immediate step is to remove the assumption of monotonicity and obtain results for general reward functions. Moreover, we want to study special structures of multi-index functions (like sum of single index functions) in the bandit setting.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of experimental code and plotting results in Section E. The research questions, theoretical models and conceptual framework, mathematical claims and proofs, hypotheses, experimental design, choice of experiments, and interpretation of the results were developed by the authors without generative AI assistance. Generative AI was not used to generate synthetic datasets, design or provide feedback on the research methodology or experiments, support qualitative or thematic data analysis, or clean and reformat datasets. Translation and analysis of qualitative research material are not applicable to this work.

Additionally, we used generative AI tools for editing and paraphrasing manuscript text to improve readability. The choice of quantities to plot, experimental comparisons, and scientific conclusions drawn from the experiments were determined by the authors.

All AI-assisted code was manually reviewed, verified, and executed by one of the authors, who also checked the resulting numerical outputs and figures. All AI-assisted text was reviewed by the authors. We take responsibility for the final content of this work, including all text, code, results, and artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *ArXiv*, abs/2302.11055, 2023. URL https://api.semanticscholar.org/CorpusID:257078637.

Sakshi Arya, Satarupa Bhattacharjee, and Bharath K Sriperumbudur. Kernel single-index bandits: Estimation, inference, and learning. *arXiv preprint arXiv:2603.18938*, 2026.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *ArXiv*, abs/2310.19793, 2023. URL https://api.semanticscholar.org/CorpusID:264805764.

Ron Blei, Fuchang Gao, and Wenbo V Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society*, 135(12):4009–4018, 2007.

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. *Advances in neural information processing systems*, 34:3004–3015, 2021.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002. ISSN 1631-073X. doi: https://doi.org/10.1016/S1631-073X(02)02292-6. URL https://www.sciencedirect.com/science/article/pii/S1631073X02022926.

Guillaume Braun, Minh Ha Quang, and Masaaki Imaizumi. Learning a single index model from anisotropic data with vanilla stochastic gradient descent. *arXiv preprint arXiv:2503.23642*, 2025.

David R. Brillinger. A generalized linear model with "Gaussian" regressor variables. In P. J. Bickel, K. A. Doksum, and J. L. Hodges (eds.), *A Festschrift for Erich L. Lehmann*, pp. 97–114. Wadsworth, Pacific Grove, CA, 1983.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Josip Djolonga, Andreas Krause, and Volkan Cevher. High-dimensional gaussian process bandits. *Advances in neural information processing systems*, 26, 2013.

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *The Annals of Statistics*, 49(2):769 – 792, 2021. doi: 10.1214/20-AOS1977. URL https://doi.org/10.1214/20-AOS1977.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Massimo Fornasier, Karin Schnass, and Jan Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12:229 – 262, 2010. URL https://api.semanticscholar.org/CorpusID:1051.

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. doi: 10.1080/01621459.1981.10477729.

Fuchang Gao. Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI: The Banff Volume*, pp. 3–17. Springer, 2013.

Fuchang Gao and Jon A Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

Dongming Huang, Songtao Tian, and Qian Lin. On the structural dimension of sliced inverse regression. *The Annals of Statistics*, 54(1):25–47, 2026. doi: 10.1214/25-AOS2505.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. 28(1):427–435, 17–19 Jun 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Yue Kang, Mingshuo Liu, Bongsoo Yi, Jing Lyu, Zhi Zhang, Doudou Zhou, and Yao Li. Single index bandits: Generalized linear contextual bandits with unknown reward functions. *arXiv preprint arXiv:2506.12751*, 2025.

Filip Kovačević, Hong Chang Ji, Denny Wu, Mahdi Soltanolkotabi, and Marco Mondelli. Full-batch gradient descent outperforms one-pass sgd: Sample complexity separation in single-index learning. *arXiv preprint arXiv:2602.02431*, 2026.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. doi: 10.1080/01621459.1992.10476258.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Annals of Statistics*, 17(3): 1009–1052, 1989. doi: 10.1214/aos/1176347254.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Wanteng Ma and T Tony Cai. Nonparametric bandits with single-index rewards: Optimality and adaptivity. *arXiv preprint arXiv:2512.24669*, 2025.

Alireza Mousavi-Hosseini, Denny Wu, and Murat A. Erdogdu. Learning multi-index models with neural networks via mean-field Langevin dynamics. *arXiv preprint arXiv:2408.07254*, 2024.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Sara van de Geer. The entropy bound for monotone functions. *Techn. Report TW*, pp. 91–10, 1991.

Sara A. Van De Geer. *Empirical Processes in M-estimation*. Cambridge university press Cambridge, 2009.

A. van der vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 1996. ISBN 9781475725452. URL https://books.google.co.in/books?id=zdDkBwAAQBAJ.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. *arXiv preprint arXiv:2507.11847*, 2025.

A ASSUMPTIONS

We use the bounded-feature and stationary-noise conditions throughout. Lipschitz continuity and the subspace oracle are required only for unknown-chart upper bounds; linearity and population identifiability describe the inverse-regression model. Known-chart results and lower bounds do not assume the subspace oracle. The notation $\|\cdot\|$ is used to denote the Euclidean norm.

Assumption 1 (Bounded Features)**.** *The feature vectors $\{X_t\}_{t\geq 1}$ are i.i.d. and $\|X_t\| \leq C_X$ almost surely for some constant $C_X < \infty$.*

The boundedness assumption is ubiquitous in contextual, linear and generalized bandits (Abbasi-yadkori et al. (2011); Chu et al. (2011); Filippi et al. (2010)).

Assumption 2 (Lipschitz reward)**.** *Each $g_a : \mathbb{R}^r \to \mathbb{R}$ is L_g-Lipschitz and bounded: $|g_a(\mathbf{z}) - g_a(\mathbf{z}')| \leq L_g \|\mathbf{z} - \mathbf{z}'\|$ and $|g_a(\mathbf{z})| \leq G_{\max}$ for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^r$ and all $a \in [K]$.*

Even for the study of single index bandits, bounded Lipschitz assumption is often used for tractability (Kang et al. (2025); Ma & Cai (2025)).

Assumption 3 (Bounded design density)**.** *The projected variable $Z := LX$ has a density f_Z on a compact set $\mathcal{Z} \subset \mathbb{R}^r$ satisfying $0 \leq f_Z(\mathbf{z}) \leq M_0 < \infty$ for all $\mathbf{z} \in \mathcal{Z}$.*

Assumption 4 (Sub-Gaussian noise)**.** *The noise variables have a common zero-mean distribution. At each round, ξ_t is independent of $(X_t, A_t, \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ denotes the history before round t. They satisfy*

$$\mathbb{E}[\xi_t \mid X_t, A_t, \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[e^{\lambda \xi_t} \mid X_t, A_t, \mathcal{F}_{t-1}] \leq e^{\lambda^2 \sigma^2/2}, \qquad \lambda \in \mathbb{R}.$$

The following two conditions concern population inverse regression. They do not by themselves provide a finite-sample subspace-estimation rate.

Assumption 5 (Linearity condition)**.** *The covariance matrix $\Sigma_X := \mathrm{Cov}(X)$ is positive definite, and $\mathbb{E}[X \mid LX]$ is an affine function of LX.*

Assumption 6 (SIR identifiability)**.** *Define the population exploration distribution by drawing X from the context distribution, drawing A uniformly from $[K]$ independently of X, and setting $Y = g_A(LX) + \xi$. Here ξ has the distribution in Assumption 4 and is independent of (X, A). The matrix*

$$\Sigma_X^{-1/2}\mathrm{Cov}\big(\mathbb{E}[X \mid Y, A]\big)\Sigma_X^{-1/2}$$

has a strictly positive r-th eigenvalue, denoted by λ_r.

The symmetric matrix in Assumption 6 has the same eigenvalues as $\Sigma_X^{-1}\mathrm{Cov}(\mathbb{E}[X \mid Y, A])$. The symmetric form makes their ordering explicit.

Assumption 5 is required only along the specified index space. It holds for elliptically symmetric features. It also holds for a coordinate-aligned product design when the remaining coordinates are independent of the index coordinates. Assumption 6 requires inverse regression to detect every direction of the index space. Neither condition follows from boundedness or Lipschitzness of the reward functions.

Statistical guarantee for subspace estimation. The unknown-chart upper bounds require a quantitative subspace-estimation guarantee. We state this separately from the model assumptions. It is not imposed by the known-chart results or by the lower-bound constructions.

Assumption 7 (Mean-square subspace oracle)**.** *From n independent observations drawn from the uniform exploration distribution of Assumption 6, the chosen subspace estimator returns an r-dimensional row space represented by $\widehat{L}_{\mathrm{raw}}$. There are constants $C_2 < \infty$ and $n_{\min} < \infty$ such that, for every deterministic $n \geq n_{\min}$,*

$$\mathbb{E}\Big[\|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}}^2\Big] \leq \frac{C_2 dr}{n\lambda_r}. \tag{6}$$

Here P_M denotes the orthogonal projector onto the row space of M. The constants C_2 and $n_{\min}$ may depend on the exploration distribution and the chosen estimator, but not on n. Any such dependence is retained in the upper-bound constants.

Theorem 6 of Huang et al. (2026) establishes a related rate under its stated distributional and sliced-stability conditions. We do not claim that it directly proves Assumption 7 for our context–action data.

The exploration blocks used below have predetermined sizes. Applying (6) to a data-dependent sample pool would require a separate justification.

For an unmodified estimator, the mean-square bound alone need not imply a logarithmic high-probability bound. The final EM-MIB+ algorithm instead runs that estimator on independent blocks and selects a majority-supported projector; Markov's inequality on each block and concentration of the number of good blocks yield the needed high-probability guarantee. Thus that algorithm requires the mean-square oracle, not an additional high-probability assumption.

B Coordinate-wise Monotone Multi-Index Bandits

We study a contextual bandit in which the dependence on the ambient context is low-dimensional and the arm-specific reward functions are monotone in a particular low-dimensional coordinate system. There are K arms, indexed by $[K] = \{1, \ldots, K\}$. At round t the learner observes a context $X_t \in \mathbb{R}^d$, chooses an arm $a_t \in [K]$, and observes

$$Y_t = g_{a_t}(LX_t) + \xi_t, \qquad R_T = \sum_{t=1}^{T}\left\{\max_{a\in[K]} g_a(LX_t) - g_{a_t}(LX_t)\right\}. \tag{7}$$

Here $L \in \mathbb{R}^{r\times d}$, with $r \ll d$, is a common projection matrix and $g_a : \mathbb{R}^r \to \mathbb{R}$ is the reward function of arm a. In the full problem both L and the functions g_a are unknown. We use the standing bounded-feature and noise assumptions, Assumptions 1 and 4, throughout. Thus $\|X_t\| \leq C_X$ almost surely, and the noise is conditionally centred and sub-Gaussian. The standing Lipschitz assumption (Assumption 2) will be invoked only when the projection has to be estimated; the known-projection analysis below does not use smoothness.

Two geometric objects will appear repeatedly. The projected context is $Z = LX \in \mathbb{R}^r$. Its *scaling box* is the smallest axis-aligned rectangle containing the support of Z; the associated map S rescales that rectangle coordinate by coordinate onto $[0,1]^r$. Thus binning always takes place in a fixed unit cube, even when the original projected coordinates have different ranges.

Roadmap. We first fix the correct monotone coordinate system and prove the structural oscillation bound that drives the analysis. We then study the easier benchmark in which the canonical projection L is known: a simple successive-elimination procedure is run independently in each grid cell and achieves $\tilde{\mathcal{O}}(T^{(r+1)/(r+2)})$ regret. We next return to the full problem, estimate the row space of L, convert the arbitrary estimated basis into the same canonical coordinates, estimate the scaling box, and run the same cell-wise algorithm. The additional price of learning the chart is of smaller order in T (and of the same $T^{2/3}$ order when $r = 1$), so the horizon exponent is unchanged. Finally, we give a matching lower bound for the known-chart benchmark. That lower bound isolates the intrinsic decision hardness; a formal minimax statement for a more restricted unknown-chart class also requires the lower-bound instances to satisfy the particular subspace estimation assumptions imposed on that class.

Throughout this section, $\tilde{\mathcal{O}}_{\mathrm{par}}(\cdot)$ suppresses polylogarithmic factors and multiplicative constants depending only on fixed problem or oracle parameters. When a dependence such as the number of arms K is displayed explicitly, it is not absorbed into this notation.

B.1 Model and coordinate system

For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^r$, write $\mathbf{a} \preceq \mathbf{b}$ when $a_j \le b_j$ for every j. We use $\mathbf{0} = (0, \dots, 0)$ and $\mathbf{1} = (1, \dots, 1)$.

Definition B.1 (Coordinate-wise monotonicity)**.** *A function g is coordinate-wise monotone if $g(z) \le g(z')$ whenever $z \preceq z'$. We write $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ for the measurable functions on $[0,1]^r$ that are coordinate-wise nondecreasing and take values in $[0, G_{\max}]$. When the range bound is understood, we write $\mathcal{F}^r_{\mathrm{CM}}$.*

The phrase "the projection L" needs some care because a subspace has many matrix representations. Let L_{raw} be any orthonormal row basis of the unknown r-dimensional index space. After a fixed, known permutation of the ambient coordinates, write $L_{\mathrm{raw}} = [\, V \mid U \,]$ with $V \in \mathbb{R}^{r \times r}$ invertible. Multiplying by V^{-1} removes the arbitrary basis choice and gives the canonical representative $L = V^{-1} L_{\mathrm{raw}} = [\, I_r \mid W \,]$, where $W = V^{-1}U$. The reward functions g_a in (7) are always understood in these canonical coordinates.

This normalization is necessary because coordinate-wise monotonicity is not a property of a row space alone. For example, $g(z_1, z_2) = z_1 + z_2$ is nondecreasing in both coordinates, whereas the same surface written after the reflection $(z_1, z_2) \mapsto (-z_1, z_2)$ becomes $-z_1 + z_2$, which is decreasing in the first coordinate. Hence a subspace estimator that is accurate only up to rotation is not yet enough for a monotone model.

Assumption 8 (Identifiable monotone chart)**.** *After a known permutation of the ambient coordinates, the common index space has an orthonormal row representative $[\, V \mid U \,]$ with V invertible and $\|V^{-1}\|_{\mathrm{op}} \le \kappa_0$. With $W = V^{-1}U$, the canonical chart is $L = [\, I_r \mid W \,]$.*

There exist $\rho^* > 0$ and an axis-aligned rectangle $\mathcal{R} \subset \mathbb{R}^r$ that contains the supports of $[\, I_r \mid W' \,]X$ for every $\|W' - W\|_F \le \rho^*$. On $\mathcal{R}$, every g_a takes values in $[0, G_{\max}]$ and is coordinate-wise nondecreasing.

The neighbourhood in this assumption is needed only when L is estimated: it ensures that the same reward function can be evaluated at both the true and a nearby estimated projection. The condition itself is a shape constraint, not a smoothness condition; discontinuities are allowed.

The standing density Assumption 3 controls the true projected design. For the unknown-chart argument we also need this control to be stable under the nearby canonical charts that the learner may produce.

Assumption 9 (Design stability near the canonical chart)**.** *For every $L' = [\, I_r \mid W' \,]$ with $\|W' - W\|_F \le \rho^*$, the smallest axis-aligned rectangle containing the support of $L'X$ has positive side lengths. After this rectangle is mapped coordinate-wise and increasingly to $[0,1]^r$, the resulting density is bounded above by a common constant $M_0' < \infty$.*

Only the upper density bound is used in the regret proofs; no positive lower density bound is imposed.

B.2 Oscillation and decision error

Let $m \geq 1$, set the cell width $h = 1/m$, and partition $[0,1]^r$ into $M = m^r$ axis-aligned cells $C_{\mathbf{k}}$, indexed by $\mathbf{k} \in \{0, \ldots, m-1\}^r$. The last interval in each coordinate is closed at its right endpoint, so every point of the cube belongs to exactly one cell. For a bounded function g, define its oscillation on a cell by $\mathrm{osc}_{\mathbf{k}}(g) = \sup_{z \in C_{\mathbf{k}}} g(z) - \inf_{z \in C_{\mathbf{k}}} g(z)$.

The next lemma is the structural reason monotonicity is useful. It does *not* say that every cell has small variation; a discontinuous monotone function can jump by $G_{\max}$ inside one cell. It says that such large jumps cannot occur independently in many cells: the *sum* of all cell oscillations is controlled.

Lemma B.2 (Summability of oscillations)**.** *For every $g \in \mathcal{F}^r_{\mathrm{CM}}(G_{\max})$,*

$$\sum_{\mathbf{k}} \mathrm{osc}_{\mathbf{k}}(g) \leq r G_{\max} m^{r-1} = r G_{\max} h^{-(r-1)}. \tag{8}$$

Proof. Let $z^-_{\mathbf{k}} = h\mathbf{k}$ and $z^+_{\mathbf{k}} = z^-_{\mathbf{k}} + h\mathbf{1}$. Monotonicity gives $\mathrm{osc}_{\mathbf{k}}(g) \leq g(z^+_{\mathbf{k}}) - g(z^-_{\mathbf{k}})$. Move from $z^-_{\mathbf{k}}$ to $z^+_{\mathbf{k}}$ one coordinate at a time. The resulting r increments telescope. For a fixed coordinate j and fixed values of the other $r-1$ cell indices, summing over the m cells along that coordinate telescopes to at most $G_{\max}$. There are m^{r-1} choices of the remaining indices. Summing over $j = 1, \ldots, r$ yields (8). □

Under a density bounded above by M_0', the lemma implies an expected within-cell variation of order $M_0' r G_{\max} h$. This is the approximation term that will later be balanced against the statistical cost of learning a separate bandit problem in each cell.

B.3 Relation to isotonic estimation and subspace recovery

Our use of monotonicity is different from classical multivariate isotonic regression. Han et al. Han et al. (2019) study least-squares estimation of the entire coordinate-wise monotone regression surface. In the random-design setting, their upper and lower bounds give the familiar dimension-dependent isotonic risk, of order $n^{-1/r}$ for $r \geq 2$ up to logarithmic factors under bounded signals and a design density bounded above and away from zero. We do not estimate the functions g_a globally. Instead, we estimate only one mean per arm and grid cell and use monotonicity through Lemma B.2 to control the aggregate discretisation error. This is why the bandit analysis needs no isotonic least-squares optimisation and, for the known-chart result, no lower density bound.

A closer comparison for the unknown projection is the high-dimensional Gaussian-process bandit method of Djolonga et al. (2013). Their SI-BO procedure first approximates a matrix of gradients by finite differences, recovers that low-rank matrix with a matrix Dantzig selector, and then takes its leading singular vectors to estimate the active subspace before running GP-UCB. The recovered matrix is a basis for the correct subspace only up to rotation, which is sufficient there because the model and isotropic kernel are rotation-invariant. Their subspace-identification argument uses substantially stronger smoothness, including second-order differentiability, Lipschitz-continuous second derivatives and a nondegeneracy condition on the Hessian. Our monotone model cannot ignore the rotation because monotonicity is coordinate dependent; this is the reason for the canonical lock $[\, I_r \mid W \,]$. Conversely, our subspace step is supplied by an SIR-type oracle and therefore relies on inverse-regression identifiability rather than finite-difference smoothness. Thus neither set of assumptions uniformly dominates the other. The advantage of the present route is that the decision problem with known L requires no smoothness at all, while the unknown-chart extension uses only the standing Lipschitz condition to translate projection error into reward error.

B.4 Binned elimination with a known chart

We first isolate the decision problem from the difficulty of estimating the projection. In this subsection the canonical chart $L = [\, I_r \mid W \,]$ and the scaling box of $Z = LX$ are known. Let S be the increasing

coordinate-wise affine map from that box to $[0,1]^r$, and write $\widetilde{Z} = S(Z)$. The context therefore determines a grid cell, and after the cell is identified the algorithm ignores the exact position inside that cell.

For each arm a and cell $\mathbf{k}$, define the true cell mean and its gap by

$$\mu_a(\mathbf{k}) = \mathbb{E}[g_a(Z) \mid \widetilde{Z} \in C_{\mathbf{k}}], \qquad \Delta_a(\mathbf{k}) = \max_{b\in[K]} \mu_b(\mathbf{k}) - \mu_a(\mathbf{k}). \tag{9}$$

Thus $\mu_a(\mathbf{k})$ is the average reward arm a would obtain when the context falls in cell $\mathbf{k}$, and $\Delta_a(\mathbf{k})$ is the loss from using arm a instead of the best *cell-average* arm. These are not the same as the pointwise rewards $g_a(Z_t)$; the difference between cell averages and pointwise rewards is the approximation error controlled later by monotonicity.

When arm a has been played $n_{\mathbf{k},a} > 0$ times in cell $\mathbf{k}$, write $\widehat{\mu}_a(\mathbf{k})$ for the empirical average of those observed rewards. We use this notation, rather than $\widehat{\mu}_{\mathbf{k},a}$, so that the empirical quantity mirrors its target $\mu_a(\mathbf{k})$. Define the confidence radius $U(n) = (\sigma + G_{\max})\sqrt{2\Lambda/n}$ with $\Lambda = \log(2KT^{r+2})$.

Algorithm 4 `CM-MIB, known chart`

1: **Input:** horizon T, arms K, dimension r, chart L, scaling map S, and grid size m.
2: Set $h = 1/m$. For each visited cell $\mathbf{k}$, initialise $\mathcal{A}_{\mathbf{k}} = [K]$, $n_{\mathbf{k},a} = 0$, and $\widehat{\mu}_a(\mathbf{k}) = 0$.
3: **for** $t = 1, \dots, T$ **do**
4: Observe X_t, compute $\widetilde{Z}_t = S(LX_t)$, and locate its cell $\mathbf{k}(t)$.
5: Among the active arms $\mathcal{A}_{\mathbf{k}(t)}$, play one with the smallest count $n_{\mathbf{k}(t),a}$; resolve count ties by a fixed arm order.
6: Observe Y_t and update $n_{\mathbf{k}(t),a_t}$ and $\widehat{\mu}_{a_t}(\mathbf{k}(t))$.
7: Once every active arm in this cell has been sampled, eliminate every arm a for which

$$\widehat{\mu}_a(\mathbf{k}(t)) + U(n_{\mathbf{k}(t),a}) < \max_{b\in\mathcal{A}_{\mathbf{k}(t)}} \{\widehat{\mu}_b(\mathbf{k}(t)) - U(n_{\mathbf{k}(t),b})\}.$$

8: **end for**

The fixed tie-breaking rule is not cosmetic. Conditional on the past and on the current cell, the selected arm depends only on the cell statistics, not on the exact position of the context inside that cell. Consequently, rewards collected for arm a in cell $\mathbf{k}$ are centred at the same target $\mu_a(\mathbf{k})$.

Define the simultaneous confidence event

$$\mathcal{G} = \{|\widehat{\mu}_a(\mathbf{k}) - \mu_a(\mathbf{k})| \le U(n) \text{ whenever arm } a \text{ in cell } \mathbf{k} \text{ has count } n \le T\}.$$

The next lemma says that, on this event, each cell behaves like an ordinary successive-elimination bandit: a suboptimal arm can survive to count s only if its cell gap is of order $1/\sqrt{s}$.

Lemma B.3 (Per-cell elimination)**.** *If $K \ge 2$, $T \ge 2$ and $m \le T$, then $\mathbb{P}(\mathcal{G}^c) \le 1/T$. On $\mathcal{G}$, simultaneously for every cell $\mathbf{k}$,*

$$\sum_{t\le T:\, \mathbf{k}(t)=\mathbf{k}} \Delta_{a_t}(\mathbf{k}) \le 16(\sigma + G_{\max})\sqrt{K n_{\mathbf{k}} \Lambda}, \tag{10}$$

where $n_{\mathbf{k}}$ is the number of visits to cell $\mathbf{k}$.

Proof. Conditional on the past, the cell, and the selected arm, the reward has mean $\mu_a(\mathbf{k})$. This holds because selection uses only the cell and its past statistics. The centred reward has a conditional sub-Gaussian proxy at most $(\sigma + G_{\max})^2$. Apply the exponential supermartingale bound at the time of the n-th observation of a fixed cell-arm pair. It gives

$$\mathbb{P}\big(n \text{ observations are collected and } |\widehat{\mu}_a(\mathbf{k}) - \mu_a(\mathbf{k})| > U(n)\big) \le 2e^{-\Lambda}.$$

A union bound over at most Km^rT triples gives the claimed failure probability. No independence of the adaptively selected sample counts is required.

Work now on $\mathcal{G}$. A cell-optimal arm a^* is never eliminated: for every active b, its lower confidence bound is at most $\mu_b(\mathbf{k}) \le \mu_{a^*}(\mathbf{k})$, while the upper confidence bound of a^* is at least $\mu_{a^*}(\mathbf{k})$.

After every surviving arm has been sampled once, suppose the algorithm next selects arm a with current count s. Because the rule selects a least-pulled active arm, the surviving optimal arm has count $n_{\mathbf{k},a^*} \geq s$. Since a has not been eliminated, its upper confidence bound overlaps the lower confidence bound of a^*. Combining this overlap with the confidence event gives $\Delta_a(\mathbf{k}) \leq 2U(s)+2U(n_{\mathbf{k},a^*}) \leq 4U(s)$. Summing these $s^{-1/2}$ bounds over pulls and then over arms yields (10); the first sample of each arm contributes only the same order. This is the standard successive-elimination calculation. □

The lemma itself uses no monotonicity. Monotonicity enters only when we compare the cell-average problem to the true contextual regret.

Theorem B.4 (Known-chart regret)**.** *Suppose the chart L and its scaling box are known, and Assumptions 8 and 9 hold. Choose the grid width by balancing the learning and approximation terms below (for fixed parameters, $h \asymp T^{-1/(r+2)}$, up to logarithmic and K factors). Then Algorithm 4 satisfies*

$$\mathbb{E}[R_T] = \tilde{\mathcal{O}}_{\mathrm{par}}\left(K^{(r+1)/(r+2)}T^{(r+1)/(r+2)}\right). \tag{11}$$

The proof does not use the Lipschitz part of Assumption 2; the same conclusion therefore holds for bounded coordinate-wise monotone rewards without any smoothness assumption.

Proof. Let $\overline{\mathrm{osc}}_{\mathbf{k}}$ be the largest oscillation, over arms, of $g_a \circ S^{-1}$ on cell $\mathbf{k}$, and let $R^{\mathrm{bin}} = \sum_t \Delta_{a_t}(\mathbf{k}(t))$ be the regret of the artificial cell-average bandit problems.

For a context in cell $\mathbf{k}(t)$, both the pointwise reward $g_a(Z_t)$ and the cell mean $\mu_a(\mathbf{k}(t))$ lie between the minimum and maximum of the same function on that cell. Hence $|g_a(Z_t) - \mu_a(\mathbf{k}(t))| \leq \overline{\mathrm{osc}}_{\mathbf{k}(t)}$. Applying this once to the pointwise-optimal arm and once to the selected arm gives the central decomposition

$$\mathbb{E}[R_T] \leq \mathbb{E}[R^{\mathrm{bin}}] + 2\mathbb{E}\left[\sum_{\mathbf{k}} n_{\mathbf{k}}\overline{\mathrm{osc}}_{\mathbf{k}}\right]. \tag{12}$$

The first term is the price of *learning the best cell-average arm*; the second is the price of *representing an entire cell by one average*.

Lemma B.3 and Cauchy–Schwarz give $\mathbb{E}[R^{\mathrm{bin}}] \lesssim (\sigma+G_{\max})\sqrt{KT\Lambda}\,h^{-r/2}+G_{\max}$. For the second term, Assumption 9 gives $\mathbb{E}[n_{\mathbf{k}}] \leq M_0' T h^r$, while Lemma B.2 gives $\sum_{\mathbf{k}} \overline{\mathrm{osc}}_{\mathbf{k}} \leq KrG_{\max}h^{-(r-1)}$. Therefore

$$\mathbb{E}[R_T] \lesssim (\sigma + G_{\max})\sqrt{KT\Lambda}\,h^{-r/2} + KM_0' r G_{\max} T h + G_{\max}. \tag{13}$$

This equation contains the whole tradeoff. Smaller cells reduce the approximation error Th but create more independent cell-bandit problems, which increases the learning term $\sqrt{T}\,h^{-r/2}$. Balancing the two terms gives (11); the trivial bound $R_T \leq G_{\max}T$ handles short-horizon and boundary choices of the grid. □

If $K = 1$ or $G_{\max} = 0$, regret is zero. If every arm mean is constant on the relevant rectangle, one may take $m = 1$ and recover the usual $\mathcal{O}((\sigma + G_{\max})\sqrt{KT\log(eKT)})$ finite-armed rate.

Remark B.5 (An estimated scaling box)**.** *Knowing the rectangle is a convenience rather than a necessity, and in particular it is not what makes Theorem B.4 free of a smoothness hypothesis. Suppose the chart is known but the rectangle is not. Form $\widehat{\mathcal{B}}$ from the coordinate-wise minima and maxima of n independent contexts, and let $\widehat{S}$ be the increasing coordinate-wise affine map carrying $\widehat{\mathcal{B}}$ onto $[0,1]^r$. On a subsequent round with $LX_t \notin \widehat{\mathcal{B}}$, play an arbitrary arm and discard the observation; otherwise proceed as in Algorithm 4 with $\widehat{S}$ in place of S.*

No clipping occurs, so the binned law acquires no boundary atoms. Since $\widehat{\mathcal{B}}$ is contained in the rectangle of the support and $\widehat{S}$ is increasing and coordinate-wise affine, the retained observations still satisfy $g_a \circ \widehat{S}^{-1} \in \mathcal{F}^r_{\mathrm{CM}}(G_{\max})$, and by Assumption 9 the cell probabilities are still at most $M_0' h^r$. Each projected coordinate has a continuous marginal law, so exchangeability of the n construction contexts with a fresh one, together with a union bound over coordinates, gives

$$\mathbb{E}\left[\mathbb{P}\{LX \notin \widehat{\mathcal{B}} \mid \widehat{\mathcal{B}}\}\right] \leq \frac{2r}{n+1}. \tag{14}$$

The construction rounds and the discarded rounds each cost at most $G_{\max}$, so the total additional regret is at most

$$G_{\max} n + \frac{2rG_{\max}T}{n+1}.$$

Taking $n = \min\{\lceil\sqrt{2rT}\,\rceil, T\}$ makes this $\mathcal{O}(G_{\max}\sqrt{rT})$, which for every $r \geq 1$ is of smaller order in T than (11)*. The conclusion of Theorem B.4 therefore persists with an estimated box, still with no smoothness hypothesis. This concerns the scaling box alone and says nothing about an unknown projection.*

B.5 An unknown chart

We now return to the full problem in which L is unknown. The learner bins $\widehat{L}X$ rather than LX, so two new geometric errors appear. First, the estimated projection moves the point from LX to $\widehat{L}X$. Second, the scaling box is built from finitely many projected samples, so a future projection may fall outside that box. From this subsection onward we invoke the standing Lipschitz Assumption 2: it converts these geometric displacements into reward errors. This is the only reason smoothness is needed in the unknown-chart analysis.

For an estimated box $\widehat{\mathcal{B}} = \prod_{j=1}^{r}[\widehat{m}_j, \widehat{M}_j]$, let $\Pi_{\widehat{\mathcal{B}}}(z)$ denote coordinate-wise clipping, i.e. replacing each z_j by its nearest point in $[\widehat{m}_j, \widehat{M}_j]$. Let $\widehat{S}$ be the increasing affine map from $\widehat{\mathcal{B}}$ to $[0,1]^r$. Thus the binned context used after burn-in is $\widetilde{Z}_t = \widehat{S}(\Pi_{\widehat{\mathcal{B}}}(\widehat{L}X_t))$.

Two-phase construction. Use a burn-in of T_0 rounds. One half estimates the row space of L; the other half, after the chart estimate is frozen, estimates its scaling box. Future rounds then run exactly the same cell-wise successive elimination as in Algorithm 4. Splitting the burn-in keeps the box sample independent of the chart estimator, and future contexts are independent of both frozen objects.

Algorithm 5 `CM-MIB, estimated chart`

1: **Input:** horizon T, arms K, dimensions d, r, burn-in T_0, and grid size m. Use the same Λ and confidence radius $U(n)$ as in Algorithm 4.
2: Pull arms independently and uniformly for the first $\lceil T_0/2 \rceil$ rounds. Use the remaining burn-in rounds to collect an independent box sample. If $T_0 = T$, stop.
3: From the first $n = \lceil T_0/2 \rceil$ rounds, compute an orthonormal-row subspace estimate $\widehat{L}_{\mathrm{raw}}$ satisfying the oracle bound (16). Let $\widehat{V}_{\mathrm{raw}}$ be its leading $r \times r$ block.
4: If $\sigma_{\min}(\widehat{V}_{\mathrm{raw}}) < 1/(2\kappa_0)$, play a fixed arm thereafter. Otherwise freeze the canonical estimate $\widehat{L} = \widehat{V}_{\mathrm{raw}}^{-1}\widehat{L}_{\mathrm{raw}} = [\, I_r \mid \widehat{W}\,]$.
5: From the remaining $n_{\mathrm{box}} = \lfloor T_0/2 \rfloor$ exploration contexts, form $\widehat{\mathcal{B}}$ from the coordinate-wise minima and maxima of $\widehat{L}X_t$. If any side length is zero, play a fixed arm thereafter; otherwise freeze $\widehat{S}$.
6: Initialise the same cell statistics and active sets as in Algorithm 4.
7: **for** $t = T_0 + 1, \ldots, T$ **do**
8: Observe X_t, compute $\widetilde{Z}_t = \widehat{S}(\Pi_{\widehat{\mathcal{B}}}(\widehat{L}X_t))$, locate its cell, and perform the same least-pulled selection, update and elimination steps as in Algorithm 4.
9: **end for**

B.5.1 From subspace error to canonical-chart error

The subspace estimator returns a row space, not the particular basis in which monotonicity is valid. The canonical lock removes this arbitrary basis choice. If the raw estimate were replaced by $Q\widehat{L}_{\mathrm{raw}}$ for an orthogonal Q, then its leading block would become $Q\widehat{V}_{\mathrm{raw}}$ and $(Q\widehat{V}_{\mathrm{raw}})^{-1}Q\widehat{L}_{\mathrm{raw}} = \widehat{V}_{\mathrm{raw}}^{-1}\widehat{L}_{\mathrm{raw}}$. Hence the canonical estimate is basis invariant.

Let P_L and $P_{\widehat{L}}$ denote the orthogonal projectors onto the true and estimated row spaces, and define the subspace error $\delta = \|P_{\widehat{L}} - P_L\|_{\mathrm{op}}$.

Lemma B.6 (Perturbation of the canonical chart)**.** *Suppose* $[\,V \mid U\,]$ *and* $\widehat{L}_{\mathrm{raw}}$ *have orthonormal rows,* $\|V^{-1}\|_{\mathrm{op}} \le \kappa_0$*, and* $\delta \le 1/(2\kappa_0)$*. Then* $\widehat{V}_{\mathrm{raw}}$ *is invertible,* $\|\widehat{V}_{\mathrm{raw}}^{-1}\|_{\mathrm{op}} \le 2\kappa_0$*, and*

$$\epsilon_W := \|\widehat{W} - W\|_F \le 2\kappa_0^2\sqrt{r}\,\delta. \tag{15}$$

Proof. For $v \in \mathbb{R}^r$, embed v as $(v,0) \in \mathbb{R}^d$. Orthonormality gives $\|\widehat{V}_{\mathrm{raw}}v\| = \|P_{\widehat{L}}(v,0)\| \ge (\kappa_0^{-1} - \delta)\|v\|$, proving the inverse bound.

The columns of $\binom{-W}{I_{d-r}}$ are orthogonal to the true row space, and one obtains the identity

$$\widehat{W} - W = \widehat{V}_{\mathrm{raw}}^{-1}\widehat{L}_{\mathrm{raw}}(I_d - P_L)\begin{pmatrix}-W\\ I_{d-r}\end{pmatrix}.$$

Moreover $\|\binom{-W}{I_{d-r}}\|_{\mathrm{op}} \le \kappa_0$, while $\widehat{L}_{\mathrm{raw}}(I_d - P_L) = \widehat{L}_{\mathrm{raw}}(P_{\widehat{L}} - P_L)$ has Frobenius norm at most $\sqrt{r}\,\delta$. Combining these three bounds proves (15). If $d = r$, both canonical representatives equal I_r. □

We invoke Assumption 7. For a subspace block of deterministic size $n \ge n_{\min}$, write $\delta := \|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}}$. Then

$$\mathbb{E}[\delta^2] \le \frac{C_2 dr}{n\lambda_r}. \tag{16}$$

This is an assumed statistical guarantee. Lemma B.6 converts it into a bound on the canonical-chart error on the good-chart event. Thus the row-space error is root-n in mean square. The constants and the validity of this oracle inequality depend on the chosen inverse-regression construction and exploration distribution; monotonicity alone does not imply it.

Compatibility with the lower-bound family. The lower-bound construction in Section B.6 can be modified, when $K \ge r+2$, so that the inverse-regression matrix has a fixed positive eigengap. Replace r duplicate arms by $g_{2+i}(z) = c\langle \mathbf{1} - \frac{1}{2}e_i, z\rangle$, $i = 1,\dots,r$, and keep any remaining arms as duplicates. These added arms are bounded above by the baseline used in the lower-bound construction, so they do not change its binary testing argument. Under uniform exploration over arms, if a vector v had zero variance against $\mathbb{E}[Z \mid Y, A]$, then for every added arm $0 = \mathrm{Cov}(v^\top Z, Y \mid A = 2+i) = (c/12)v^\top(\mathbf{1} - \frac{1}{2}e_i)$. The r displayed directions are linearly independent, forcing $v = 0$. Thus the eigengap component of an SIR assumption can be made compatible with the lower-bound family. Other hypotheses of a particular SIR theorem still require separate verification, so this observation does not by itself establish a matching minimax lower bound for every SIR-restricted class.

Define the good-chart event

$$\mathcal{L} = \{\delta \le \tau\}, \qquad \tau = \min\left\{\frac{1}{2\kappa_0}, \frac{\rho^*}{2\kappa_0^2\sqrt{r}}\right\}. \tag{17}$$

The first threshold guarantees that the canonical lock is stable; the second keeps $\widehat{W}$ inside the neighbourhood where the monotonicity and density assumptions are valid. Lemma B.6 and (16) imply $\mathbb{E}[\epsilon_W \mathbf{1}_{\mathcal{L}}] \lesssim_{\mathrm{par}} n^{-1/2}$ and $\mathbb{P}(\mathcal{L}^c) \lesssim_{\mathrm{par}} n^{-1}$.

B.5.2 What the estimated chart does to rewards

Write $Z = LX$ and $\widehat{Z} = \widehat{L}X$. On $\mathcal{L}$, $\|Z - \widehat{Z}\| \le C_X\epsilon_W$. By the standing Lipschitz assumption, for every arm a,

$$|g_a(LX) - g_a(\widehat{L}X)| \le L_g C_X \epsilon_W. \tag{18}$$

This is the first genuinely new error relative to the known-chart proof: even a perfect cell-wise bandit algorithm would be evaluating the reward function at a slightly displaced point.

Lemma B.7 (Fibre averaging)**.** *Fix a realised chart on* $\mathcal{L}$*. For any arm* a *whose observation noise is conditionally centred,*

$$\left|\mathbb{E}[Y \mid \widehat{L}X = \widehat{z}, A = a] - g_a(\widehat{z})\right| \le L_g C_X \epsilon_W \tag{19}$$

for almost every $\widehat{z}$*.*

Proof. Every context on the fibre $\widehat{L}x = \widehat{z}$ satisfies $\|Lx - \widehat{z}\| = \|(\widehat{W} - W)x_{\mathrm{rest}}\| \le C_X \epsilon_W$. Apply Assumption 2 pointwise and then average conditional on $\widehat{L}X = \widehat{z}$ and $A = a$. □

The conditional mean above need not itself be monotone as a function of $\widehat{z}$. This causes no problem: Lemma B.3 needs only a well-defined cell-conditional mean, not monotonicity of that conditional mean. Monotonicity is used later only to control the oscillations of the original functions g_a in canonical coordinates.

Conditional on the frozen chart and box, define $\mu_a(\mathbf{k})$ and $\Delta_a(\mathbf{k})$ exactly as in (9), but with the clipped-and-scaled $\widetilde{Z}$ defined above. The empirical mean $\widehat{\mu}_a(\mathbf{k})$ is again centred at this cell target.

B.5.3 Cost of estimating the scaling box

A fresh projected context can lie outside the sample box. Let $\varrho = \|\widehat{Z} - \Pi_{\widehat{\mathcal{B}}}(\widehat{Z})\|$ be its distance to the clipped point. The next lemma shows that both the probability and the average size of such an overshoot are of order $1/n_{\mathrm{box}}$.

Lemma B.8 (Cost of the frozen box)**.** *On $\mathcal{L}$, for an independent fresh context,*

$$\mathbb{P}(\widehat{Z} \notin \widehat{\mathcal{B}} \mid \widehat{L}) \le \frac{2r}{n_{\mathrm{box}} + 1}, \qquad \mathbb{E}[\varrho \mid \widehat{L}] \le \frac{8r\kappa_0 C_X}{n_{\mathrm{box}} + 1}. \tag{20}$$

For a fixed realised box, each interior grid cell has probability at most $M_0' h^r$, and $|g_a(\widehat{Z}) - g_a(\Pi_{\widehat{\mathcal{B}}}\widehat{Z})| \le L_g \varrho$.

The probability and overshoot bounds above average over the random box-building sample as well as the fresh context; they are not asserted uniformly after conditioning on a particular realised box. The interior-cell mass bound is the statement that is used conditionally on a fixed box.

Proof. For each coordinate, among the n_{box} box-building samples and one fresh sample, exchangeability gives probability $1/(n_{\mathrm{box}} + 1)$ that the fresh sample is the largest and the same probability that it is the smallest. A union bound over r coordinates gives the first statement. On $\mathcal{L}$, $\|\widehat{L}\|_{\mathrm{op}} \le 2\kappa_0$, so each support width is at most $4\kappa_0 C_X$; multiplying this width by the probability of exceeding a sample extreme gives the overshoot bound. The cell-mass bound follows from Assumption 9, and the reward-displacement bound follows from Assumption 2. □

Lemma B.9 (Retaining observations in the frozen box)**.** *Fix a chart on the good event $\mathcal{L}$, and construct its sample box from $s \ge 2$ independent contexts. Write $p_{\mathcal{B}}$ for the conditional probability that a fresh projected context belongs to this box. For every $u \in (0, 1)$,*

$$\mathbb{P}(1 - p_{\mathcal{B}} > u \mid \widehat{L}) \le 2r \exp\left(-\frac{su}{2r}\right).$$

On $p_{\mathcal{B}} \ge 1/2$, the law of the rescaled projection conditional on retention has a density bounded above by $2M_0'$. The retained observations form an i.i.d. sequence whenever the original observations under the fixed sampling kernel are i.i.d.

Proof. For each coordinate, the population probability below the sample minimum exceeds v with probability at most $(1 - v)^s \le e^{-sv}$. The same holds above the sample maximum. Continuity of the marginal laws follows from the design-density assumption. The probability outside the box is at most the sum of these $2r$ tail probabilities. Set $v = u/(2r)$ and take a union bound.

Let $\mathcal{B}_*$ be the true support rectangle for the fixed chart. The density bound after rescaling $\mathcal{B}_*$ implies that the unscaled density is at most $M_0'/\operatorname{vol}(\mathcal{B}_*)$. The sample box is contained in $\mathcal{B}_*$. Its affine rescaling therefore produces an unnormalised density at most M_0'. Conditioning on retention divides this density by $p_{\mathcal{B}}$. Finally, retention depends only on each fresh context and the frozen box. Independent thinning gives the last assertion. □

B.5.4 REGRET UPPER BOUND

The proof is best read as the known-chart decomposition plus three new costs: burn-in, projection error, and box/clipping error. Conditional on the frozen chart and box, the per-cell learning term is still governed by Lemma B.3.

Theorem B.10 (Regret with an estimated chart)**.** *Suppose Assumptions 1, 2, 4,7, 8 and 9 hold. Choose the same grid order as in the known-chart case and take $T_0 \asymp T^{2/3}$ up to fixed problem/oracle factors. Assume T is sufficiently large that $T_0 \le T$ and the subspace block contains at least $n_{\min}$ observations. Then Algorithm 5 satisfies*

$$\mathbb{E}[R_T] \le \widetilde{O}_{\mathrm{par}}\Big(K^{(r+1)/(r+2)}T^{(r+1)/(r+2)} + T^{2/3}\Big). \tag{21}$$

For fixed K, the horizon exponent is $(r+1)/(r+2)$. When $r = 1$, both contributions have exponent $2/3$. The constants include those of the stated subspace oracle.

Proof. Assume first $4 \le T_0 < T$ and work on the good-chart event $\mathcal{L}$ from (17). Let R^{bin} be the sum of the cell gaps after burn-in and let $\overline{\mathrm{osc}}_{\mathbf{k}}$ be the largest arm oscillation in cell $\mathbf{k}$.

For a context in cell $\mathbf{k}$, the difference between the true pointwise reward and its cell mean now has three sources: variation inside the cell, displacement from L to $\widehat{L}$, and clipping at the estimated box. The previous lemmas give, up to the corresponding conditional average for the cell,

$$|g_a(Z_t) - \mu_a(\mathbf{k})| \le \overline{\mathrm{osc}}_{\mathbf{k}} + 2L_g C_X \epsilon_W + L_g\{\varrho_t + \mathbb{E}[\varrho \mid \widetilde{Z} \in C_{\mathbf{k}}, \widehat{L}, \widehat{\mathcal{B}}]\}.$$

Applying this once to the pointwise-optimal arm and once to the selected arm, then summing and averaging, yields the conceptual decomposition

$$\begin{aligned}\mathbb{E}[R_T] \lesssim\ & \underbrace{G_{\max}T_0}_{\text{burn-in}} + \underbrace{\mathbb{E}[R^{\mathrm{bin}}\mathbf{1}_{\mathcal{L}}]}_{\text{cell learning}} + \underbrace{\mathbb{E}\left[\mathbf{1}_{\mathcal{L}}\sum_{\mathbf{k}} n_{\mathbf{k}}\overline{\mathrm{osc}}_{\mathbf{k}}\right]}_{\text{cell approximation}} \\ & + \underbrace{L_g C_X T\mathbb{E}[\epsilon_W \mathbf{1}_{\mathcal{L}}]}_{\text{chart displacement}} + \underbrace{L_g T\mathbb{E}[\varrho \mathbf{1}_{\mathcal{L}}]}_{\text{box/clipping}} + \underbrace{G_{\max}T\mathbb{P}(\mathcal{L}^c)}_{\text{bad-chart event}}.\end{aligned} \tag{22}$$

The displayed constants are immaterial for the rate; the exact preceding lemmas provide fixed numerical factors if desired.

The first two post-burn-in terms are the known-chart terms. Lemma B.3 gives $\mathbb{E}[R^{\mathrm{bin}}\mathbf{1}_{\mathcal{L}}] \lesssim_{\mathrm{par}} \sqrt{KT\log T}\, h^{-r/2}$. For contexts inside the frozen box, Lemma B.2 and the density bound give an approximation cost $\lesssim_{\mathrm{par}} KTh$. Exterior visits occur with probability $\mathcal{O}_{\mathrm{par}}(1/T_0)$ and contribute another $\mathcal{O}_{\mathrm{par}}(T/T_0)$ term.

The first half of burn-in has $n \asymp T_0$ samples, so Lemma B.6 and the SIR oracle give $\mathbb{E}[\epsilon_W \mathbf{1}_{\mathcal{L}}] = \mathcal{O}_{\mathrm{par}}(T_0^{-1/2})$ and $\mathbb{P}(\mathcal{L}^c) = \mathcal{O}_{\mathrm{par}}(T_0^{-1})$. Lemma B.8 likewise gives $\mathbb{E}[\varrho \mathbf{1}_{\mathcal{L}}] = \mathcal{O}_{\mathrm{par}}(T_0^{-1})$. Consequently

$$\mathbb{E}[R_T] \lesssim_{\mathrm{par}} T_0 + \sqrt{KT\log T}\, h^{-r/2} + KTh + TT_0^{-1/2} + T/T_0 + 1. \tag{23}$$

This is the unknown-chart analogue of (13). The terms involving h are unchanged, so the same balancing gives the $T^{(r+1)/(r+2)}$ contribution. Balancing the burn-in cost T_0 with the chart-displacement cost $T/\sqrt{T_0}$ gives $T_0 \asymp T^{2/3}$ and a $T^{2/3}$ contribution; the remaining T/T_0 terms are then only $T^{1/3}$. This proves (21). The trivial bound $R_T \le G_{\max}T$ covers short horizons and the fallback cases. □

The role of Lipschitz continuity should be kept separate from the role of monotonicity. Monotonicity controls the aggregate within-cell oscillation and is what makes the known-chart approximation term small without any smoothness. Lipschitz continuity is used here only when the same reward function is evaluated at two different points: LX versus $\widehat{L}X$, and $\widehat{Z}$ versus its clipped version. The second use can be avoided by discarding out-of-box rounds; the first is intrinsic to this particular unknown-chart argument. We do not claim that Lipschitz continuity is necessary for every possible unknown-chart method.

B.6 A MATCHING LOWER BOUND WITH A KNOWN CHART

We now consider $d = r$, $L = I_r$ known to the learner and X_t uniform on $[0,1]^r$. The noise is independent Gaussian with variance $\sigma^2 > 0$. The construction below already uses Lipschitz monotone functions. It therefore gives a lower bound for the larger bounded monotone class as well.

Use the grid defined in Section B.2. Let $\delta = 1/(4r)$ and define, for $u \in \mathbb{R}$,

$$\chi(u) = \max\{0, \min\{1, u/\delta, (1-u)/\delta\}\}, \qquad \varphi(u) = \prod_{j=1}^{r} \chi(u_j).$$

The function χ vanishes outside $[0,1]$ and equals one on $[\delta, 1-\delta]$. Set $\psi_{\mathbf{k}}(z) = \varphi((z - h\mathbf{k})/h)$ and let $B_{\mathbf{k}} = \{z \in C_{\mathbf{k}} : \psi_{\mathbf{k}}(z) = 1\}$. The bumps vanish on cell boundaries. Almost everywhere,

$$|\partial_j \psi_{\mathbf{k}}| \le 4r/h, \qquad |B_{\mathbf{k}}| = \left(1 - \frac{1}{2r}\right)^r h^r \ge \frac{1}{2} h^r. \tag{24}$$

The last inequality follows from Bernoulli's inequality.

Let $c = G_{\max}/(2r)$ and $u(z) = c\sum_{j=1}^{r} z_j$. For $\omega \in \{-1,+1\}^M$, define

$$\begin{aligned} g_1^\omega &= u + \epsilon \sum_{\mathbf{k}} \frac{1 - \omega_{\mathbf{k}}}{2} \psi_{\mathbf{k}}, \\ g_2^\omega &= u + \epsilon \sum_{\mathbf{k}} \frac{1 + \omega_{\mathbf{k}}}{2} \psi_{\mathbf{k}}. \end{aligned} \tag{25}$$

Arms $3, \dots, K$ duplicate arm 1. On $B_{\mathbf{k}}$, arm 2 is optimal if $\omega_{\mathbf{k}} = +1$, and the other arms are optimal if $\omega_{\mathbf{k}} = -1$. Choosing the wrong group loses exactly ϵ. The duplicates make this binary comparison immediate.

Lemma B.11 (Feasibility)**.** *If*

$$0 < \epsilon \le \frac{G_{\max} h}{8r^2}, \tag{26}$$

then every g_a^ω belongs to $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ and has Lipschitz constant at most $G_{\max}/\sqrt{r}$.

Proof. At any point, at most one bump is nonzero. Almost everywhere,

$$0 \le c - \frac{4r\epsilon}{h} \le \partial_j g_a^\omega \le c + \frac{4r\epsilon}{h} \le 2c.$$

Along each coordinate line the function is continuous and piecewise linear with nonnegative slope, so it is nondecreasing. Also, $0 \le u \le G_{\max}/2$ and $\epsilon \le G_{\max}/(8r^2) \le G_{\max}/2$, which proves the range bound. Finally, $\|\nabla g_a^\omega\| \le 2c\sqrt{r} = G_{\max}/\sqrt{r}$ almost everywhere. Integrating along line segments proves the Lipschitz bound. □

Theorem B.12 (Known-chart lower bound)**.** *Fix $r \ge 1$, $K \ge 2$, $G_{\max} > 0$ and $\sigma > 0$. Define*

$$h^* = \left(\frac{8r^2\sigma}{G_{\max}\sqrt{T}}\right)^{2/(r+2)}, \qquad n_0^* = T(h^*)^r. \tag{27}$$

Suppose

$$T \ge \max\left\{\left(\frac{8r^2\sigma}{G_{\max}}\right)^2, 2^{r(r+2)/2}\left(\frac{G_{\max}}{8r^2\sigma}\right)^r\right\}. \tag{28}$$

For every policy π,

$$\sup_\omega \mathbb{E}_\pi^\omega[R_T] \ge \frac{1}{48}\left(\frac{G_{\max}}{8r^2}\right)^{r/(r+2)} \sigma^{2/(r+2)} T^{(r+1)/(r+2)}, \tag{29}$$

where the instances have $d = r$, known $L = I_r$, uniform contexts and Gaussian noise of variance σ^2. Their reward functions lie in $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ and are Lipschitz with constant $G_{\max}/\sqrt{r}$. The bound therefore applies verbatim to every class of reward functions containing them, in particular to $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ itself, to the L_g-Lipschitz functions with values in $[0, G_{\max}]$ for every $L_g \ge G_{\max}/\sqrt{r}$, and to the intersection of the two.

Proof. Set $m = \lceil 1/h^* \rceil$, $h = 1/m$ and $\epsilon = G_{\max} h/(8r^2)$. The threshold implies $h^* \le 1$ and $n_0^* \ge 2^r$. Hence $h^*/2 \le h \le h^*$ and, writing $n_0 = Th^r$,

$$n_0 \ge 1, \qquad \epsilon^2 n_0 = \frac{G_{\max}^2}{64r^4} T h^{r+2} \le \sigma^2. \tag{30}$$

Feasibility follows from Lemma B.11.

Let $S_{\mathbf{k}}(\omega)$ count the rounds in $B_{\mathbf{k}}$ on which the policy selects the wrong group of arms. Since regret is nonnegative outside these plateaus,

$$\mathbb{E}_\omega[R_T] \ge \epsilon \sum_{\mathbf{k}} \mathbb{E}_\omega[S_{\mathbf{k}}(\omega)]. \tag{31}$$

Fix $\mathbf{k}$, and let ω' differ from ω only at that cell. Conditional on the complete context sequence $\mathcal{X} = (X_1, \dots, X_T)$, the action kernels are the same under the two instances. Reward means agree outside $C_{\mathbf{k}}$ and differ in absolute value by at most ϵ within it, regardless of the chosen arm. The chain rule for Gaussian relative entropy gives

$$\mathrm{KL}(\mathbb{P}_\omega^{\mathcal{X}} \,\|\, \mathbb{P}_{\omega'}^{\mathcal{X}}) \le \frac{\epsilon^2}{2\sigma^2} N_{\mathbf{k}}, \qquad N_{\mathbf{k}} = \sum_{t=1}^{T} \mathbf{1}\{X_t \in C_{\mathbf{k}}\}. \tag{32}$$

Conditioning here is only on the exogenous contexts, whose law is common to the two instances.

At each round in $B_{\mathbf{k}}$, the two error events are complementary. The Bretagnolle–Huber inequality therefore gives

$$\mathbb{E}_\omega[S_{\mathbf{k}}(\omega) \mid \mathcal{X}] + \mathbb{E}_{\omega'}[S_{\mathbf{k}}(\omega') \mid \mathcal{X}] \ge \frac{N_{B_{\mathbf{k}}}}{2} \exp\left(-\frac{\epsilon^2 N_{\mathbf{k}}}{2\sigma^2}\right), \tag{33}$$

where $N_{B_{\mathbf{k}}}$ counts visits to the plateau.

Put $\alpha = \epsilon^2/(2\sigma^2)$. The variables $N_{B_{\mathbf{k}}}$ and $N_{\mathbf{k}}$ are dependent, so we average their product directly. Independence of the contexts gives the exact identity

$$\mathbb{E}[N_{B_{\mathbf{k}}} e^{-\alpha N_{\mathbf{k}}}] = T|B_{\mathbf{k}}| e^{-\alpha} \mathbb{E}[e^{-\alpha N'}], \qquad N' \sim \mathrm{Bin}(T-1, h^r).$$

By Jensen's inequality, $\mathbb{E}[e^{-\alpha N'}] \ge e^{-\alpha n_0}$. Also $\alpha n_0 \le 1/2$, $\alpha \le 1/2$, and $T|B_{\mathbf{k}}| \ge n_0/2$. Thus

$$\mathbb{E}_\omega[S_{\mathbf{k}}(\omega)] + \mathbb{E}_{\omega'}[S_{\mathbf{k}}(\omega')] \ge \frac{n_0}{4e}. \tag{34}$$

Average (31) over all sign vectors and pair each vector with its $\mathbf{k}$-flip. Since $Mn_0 = T$,

$$2^{-M} \sum_{\omega} \mathbb{E}_\omega[R_T] \ge \frac{\epsilon M n_0}{8e} = \frac{\epsilon T}{8e} \ge \frac{\epsilon T}{24}.$$

Finally, $\epsilon \ge G_{\max} h^*/(16r^2)$, and

$$\frac{G_{\max}}{8r^2} h^* T = \left(\frac{G_{\max}}{8r^2}\right)^{r/(r+2)} \sigma^{2/(r+2)} T^{(r+1)/(r+2)}.$$

This proves (29). □

Remark B.13 (Smaller Lipschitz constants)**.** *Fix $L_g > 0$ and put*

$$c_{L_g} = \min\left\{\frac{G_{\max}}{2r}, \frac{L_g}{2\sqrt{r}}\right\}, \qquad A = \min\left\{\frac{G_{\max}}{8r^2}, \frac{L_g}{8r^{3/2}}\right\} = \frac{c_{L_g}}{4r}. \tag{35}$$

Running the construction of (25) *with c_{L_g} in place of c and with $\epsilon \le Ah$ in place of* (26) *gives $\|\nabla g_a^\omega\| \le 2c_{L_g}\sqrt{r} \le L_g$ and $u \le c_{L_g} r \le G_{\max}/2$, so the functions are L_g-Lipschitz and still lie in $\mathcal{F}_{\mathrm{CM}}^r(G_{\max})$. The proof of Lemma B.11 applies unchanged with these constants. Replacing $G_{\max}/(8r^2)$ by A in* (27)*,* (28) *and* (29) *then gives the same conclusion with lower-bound coefficient proportional to $A^{r/(r+2)}\sigma^{2/(r+2)}$, for every fixed $L_g > 0$. The case $L_g = 0$ is excluded: every arm function is then constant on the rectangle and the construction degenerates.*

Corollary B.14 (Minimax rate for a known chart)**.** *For fixed $r \geq 1$, $K \geq 2$, $G_{\max} > 0$ and $\sigma > 0$, consider uniform contexts on $[0,1]^r$, a known identity chart, independent Gaussian noise of variance σ^2, and bounded coordinate-wise monotone arm functions. The minimax regret satisfies*

$$\mathfrak{R}^*(T) = \Omega(T^{(r+1)/(r+2)}), \qquad \mathfrak{R}^*(T) = \tilde{\mathcal{O}}(T^{(r+1)/(r+2)}).$$

The same conclusion holds after restricting the class to those functions that are also L_g-Lipschitz, for every fixed $L_g > 0$.

Proof. The upper bound is Theorem B.4 with $M_0' = 1$ and S the identity; restricting the class can only decrease the minimax value. For the lower bound over $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$, and over its intersection with the L_g-Lipschitz functions when $L_g \geq G_{\max}/\sqrt{r}$, use Theorem B.12, whose functions lie in both classes. For smaller L_g, use Remark B.13, which supplies the same statement with $G_{\max}/(8r^2)$ replaced by the constant A of (35). □

Remark B.15 (The role of monotonicity)**.** *Theorem B.4 uses only the range bound and the ordering. Its approximation step is Lemma B.2, which bounds a sum of cell oscillations rather than each oscillation separately. A single cell may have oscillation $G_{\max}$, and functions in $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ need not be continuous.*

Were one instead to assume that each g_a is L_g-Lipschitz, the same elimination argument would apply with an oscillation bound in place of Lemma B.2. That bound must be stated in the binned coordinates. Writing $\ell_{\max}$ for the largest side length of the rectangle carried onto $[0,1]^r$ by S, the inverse map S^{-1} is coordinate-wise affine with Lipschitz constant $\ell_{\max}$, so

$$\operatorname{osc}_{\mathbf{k}}(g_a \circ S^{-1}) \leq L_g \ell_{\max} \sqrt{r}\, h, \tag{36}$$

the factor $\ell_{\max}$ being one exactly when S is the identity on the unit cube. Balancing this against the elimination cost again gives the exponent $(r+1)/(r+2)$. The two hypotheses are not comparable: $\mathcal{F}^r_{\mathrm{CM}}(G_{\max})$ contains discontinuous functions, which are Lipschitz for no constant, and the Lipschitz class contains functions that are not coordinate-wise monotone. Since the construction of Theorem B.12 lies in the intersection, and a lower bound over a class is inherited by every class containing it, the exponent is minimax for each of the three classes.

On the intersection the two approximation bounds may be combined. The bound (36) *holds cell by cell and uniformly over arms, so summing it against $\sum_{\mathbf{k}} \mathbb{E}[n_{\mathbf{k}}] = T$ carries no factor K, whereas Lemma B.2 is summed over arms and does. The cell term of* (13) *therefore becomes*

$$2Th \min\left\{ K M_0' r G_{\max},\ L_g \ell_{\max} \sqrt{r} \right\},$$

and monotonicity is the sharper of the two when $L_g \ell_{\max} > K\sqrt{r}\, M_0' G_{\max}$, that is, for steep functions of small range, on a grid with many arms. The comparison thus depends on K and on the scaling as well as on r, M_0' and $G_{\max}$. The exponent $(r+1)/(r+2)$ is not specific to a shape constraint; what is specific to monotonicity is that it delivers the exponent with no smoothness at all.

B.7 Optimality and interpretation

The lower bound uses two informative arm profiles and therefore does not match the upper bound's dependence on K. The present result matches the horizon exponent for fixed K and r. In the upper bound, the estimate $\sum_{\mathbf{k}} \max_a \operatorname{osc}_{\mathbf{k}} \leq \sum_a \sum_{\mathbf{k}} \operatorname{osc}_{\mathbf{k}}$ can be loose. No sharp minimax dependence on K is claimed.

It is also useful to compare elimination with a fixed exploration schedule. Consider policies that allocate a deterministic budget $T_{\exp}$ equally among arms, independently of the contexts, and then use a decision rule fitted only to that exploration data. Subsequent rewards are not used to revise the decisions. In the known-chart class, one arm of constant mean $G_{\max}/4$ and $K-1$ arms of mean zero force at least $G_{\max} T_{\exp}/8$ exploration regret, when the equal allocation is exact. Integer allocation changes this only by a bounded initial cost.

For the exploitation cost, put $A = G_{\max}/(8r^2)$ and use the bump family with

$$h^\flat = \min\left\{ 1, \left(\frac{\sigma}{A\sqrt{\max\{T_{\exp}, 1\}}} \right)^{2/(r+2)} \right\}, \qquad h = \frac{1}{\lceil 1/h^\flat \rceil}, \qquad \epsilon = Ah.$$

The relative entropy between adjacent exploration experiments is at most $\epsilon^2 T_{\exp} h^r/(2\sigma^2) \le 1/2$. A future context is independent of those experiments. Applying the same two-point inequality conditional on that future context and averaging over the plateaus gives exploitation regret at least an absolute constant times

$$(T - T_{\exp}) \min\left\{A,\ A^{r/(r+2)}\sigma^{2/(r+2)} \max\{T_{\exp}, 1\}^{-1/(r+2)}\right\}.$$

For fixed positive parameters and sufficiently large T, balancing this expression against the exploration cost yields

$$\sup \mathbb{E}_\pi[R_T] \ge c\, G_{\max}^{(r+1)/(r+3)} \sigma^{2/(r+3)} r^{-2r/(r+3)} T^{(r+2)/(r+3)}. \tag{37}$$

If $T_{\exp} > T/2$, the exploration cost alone is linear. For smaller budgets the preceding balance applies; its truncated branch also covers budgets with too little data for a nontrivial grid. This explains why learning within each cell continues during the second phase. Elimination stops sampling an arm once it is inferior in that cell.

The improvement of the exponent $(r+1)/(r+2)$ over the two-phase guarantee $(r+2)/(r+3)$ is a consequence of adaptive elimination rather than of the shape constraint. As noted in Remark B.15, the exponent itself is attainable under a Lipschitz hypothesis with no shape constraint; what monotonicity supplies is the same exponent with no smoothness at all.

For $r = 1$, the known-chart minimax exponent is $2/3$. The model still has arm-specific functions, so the best arm need not be constant along the index. This differs from Kang et al. (2025), where the arms have different feature vectors but share a common link and a common parameter. For a common nondecreasing link, maximising the scalar index selects an optimal arm. Equality of the two sets of maximisers requires further strictness when the link has flat intervals.

After the initial subspace estimation, the computation per round is $\mathcal{O}(rd + K)$, using direct projection and an elimination scan. Storing statistics only for visited cells uses $\mathcal{O}(rd + (K + r)\min\{M, T\})$ memory, including the chart and the coordinate keys of visited cells. No shape-constrained optimisation is required. This computational simplicity should not be confused with hardness of isotonic regression. Finite-sample isotonic least squares on a general partial order is a convex quadratic optimisation problem and admits polynomial-time methods Han et al. (2019); ordinary pool-adjacent-violators applies directly to a total order.

The horizon exponent approaches one as r increases. Thus the known-chart lower bound already exhibits a curse of effective dimension. The upper bound separates this decision cost from the statistical cost of estimating the chart. Any sharper claim for the full unknown-chart model must also control the oracle constants and verify that its lower-bound instances satisfy the same assumptions.

C Details of Entirely Monotone Reward Function

We drop the arm index throughout this section for readability. We use reward function and link function interchangeably.

Definition C.1 (Corner indicator function)**.** *For any $\mathbf{a} \in [0,1]^r$, the corner indicator above $\mathbf{a}$ is*

$$\mathbf{1}_{[\mathbf{a},\mathbf{1}]}(\mathbf{z}) := \prod_{j=1}^{r} \mathbf{1}[z_j \ge a_j] = \begin{cases} 1 & \text{if } \mathbf{a} \preceq \mathbf{z}, \\ 0 & \text{otherwise.} \end{cases}$$

Definition C.2 (Mixed finite difference)**.** *For a non-empty $S \subseteq \{1, \dots, r\}$, points $\mathbf{a}, \mathbf{b} \in [0,1]^r$ with $a_j < b_j$ for all $j \in S$, and a function $f : [0,1]^r \to \mathbb{R}$, define*

$$\Delta_S(f; \mathbf{a}, \mathbf{b}) := \sum_{\mathbf{u} \in \{0,1\}^S} (-1)^{|S| - \sum_{j \in S} u_j} f\big(\mathbf{z}(\mathbf{u})\big),$$

where,

$$z(\mathbf{u})_j = \begin{cases} b_j & \text{if } j \in S \text{ and } u_j = 1, \\ a_j & \text{if } j \in S \text{ and } u_j = 0, \\ a_j = b_j & \text{if } j \notin S. \end{cases}$$

Remark C.3. *For $r = 2$ and $S = \{1, 2\}$, we have*

$$\Delta_{\{1,2\}}(f; a, b) = f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2),$$

the second-order mixed difference over the rectangle $[a, b]$.

Definition C.4 (Entire monotonicity)**.** *A function $f : [0, 1]^r \to \mathbb{R}$ is entirely monotone if $\Delta_S(f; \mathbf{a}, \mathbf{b}) \geq 0$ for every non-empty $S \subseteq \{1, \ldots, r\}$ and every $\mathbf{a}, \mathbf{b}$ with $a_j < b_j$ for all $j \in S$.*

The class of all entirely monotone functions on $[0, 1]^r$ is denoted $\mathcal{F}^r_{\mathrm{EM}}$.

The definition C.4 of monotonicity above matches the mixed-difference formulation used in Fang et al. (2021).

Coordinates and sampling. We use Assumption 8 with entire monotonicity on its rectangle $\mathcal{R}$. The variation bound below is imposed on that rectangle, so it also holds on every subrectangle used for fitting. We use Assumption 9 for nearby canonical charts. Entire monotonicity is preserved by increasing coordinatewise affine maps. It is not preserved by arbitrary changes of basis.

For the regression-based procedures in this section we assume stationary stochastic rewards. Under each fixed action kernel, fresh observations are i.i.d. and the noise satisfies the Assumption 4. This requirement is additional to a conditional moment bound that allows the noise law to vary arbitrarily over time. All conditional regression arguments below use fresh data after the chart and box have been frozen.

The following lemma is the structural foundation of the entire estimation route. It shows that $\mathcal{F}^r_{\mathrm{EM}}$ is, in a precise sense, no more complex than the one-dimensional class of non-decreasing functions.

Lemma C.5 (Indicator representation, see Fang et al. (2021) Lemma 2.2)**.** *Every bounded, right-continuous $f \in \mathcal{F}^r_{\mathrm{EM}}$ with $f(\mathbf{0}) = 0$ satisfies*

$$f(\mathbf{z}) = \int_{[0,1]^r} \mathbf{1}_{[\mathbf{a},\mathbf{1}]}(\mathbf{z})\, d\nu(\mathbf{a}), \tag{38}$$

for a unique non-negative Borel measure ν on $[0, 1]^r$ with total mass $\nu([0, 1]^r) = V^* := f(\mathbf{1}) - f(\mathbf{0})$. Conversely, every function of this form with $\nu \geq 0$ belongs to $\mathcal{F}^r_{\mathrm{EM}}$.

Assumption 10 (Variation budget)**.** *Each link has corner difference at most the known constant $V_{\max}$ on $\mathcal{R}$. One may always take $V_{\max} = 2G_{\max}$ under Assumption 2.*

Definition C.6 (EM-LSE)**.** *The* Entirely Monotone Least Squares Estimator *is any solution to*

$$\hat{f}_{\mathrm{EM}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}^r_{\mathrm{EM}}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(Z_i)\big)^2. \tag{39}$$

Existence follows from Fang et al. (2021).

Definition C.7 (Variation-constrained least squares)**.** *Set*

$$\mathcal{F}_V = \left\{ \beta + f \;\middle|\; \begin{array}{l} |\beta| \leq G_{\max}, \quad f(0) = 0, \quad f(1) \leq V_{\max}, \\ f \text{ is right-continuous and entirely monotone} \end{array} \right\}.$$

Its elements are bounded in absolute value by $B_0 = G_{\max} + V_{\max}$. An estimator with optimisation tolerance $\eta_n \geq 0$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{f}(Z_i))^2 \leq \inf_{f \in \mathcal{F}_V} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 + \eta_n.$$

We use measurable estimators and a separable version of the class. Exact minimisation is the case $\eta_n = 0$, when a minimiser is attained.

C.1 Entropy and regression error

Proposition C.8 (Bracketing entropy of r-dimensional distribution functions, Gao (2013))**.** *Let $\mathcal{D}_r$ denote the class of distribution functions of probability measures on $[0, 1]^r$, $r \geq 2$, and let λ*

denote Lebesgue measure on $[0,1]^r$. For every $p \in [1,\infty)$ there is a constant $c_{r,p}$ such that for every $\eta \in (0, e^{-1})$,

$$\log N_{[]}\big(\eta,\ \mathcal{D}_r,\ L_p(\lambda)\big)\ \le\ c_{r,p}\,\frac{1}{\eta}\big|\log \eta\big|^{2(r-1)}.$$

We use the case $p = 2$ throughout.

Remark C.9 (The univariate case)**.** *Define the logarithmic exponent in the bracketing bound by $\alpha_r := 2(r-1)$. For $r = 1$, a stronger statement holds. The class $\mathcal{D}_1$ is contained in the class of nondecreasing functions from $[0,1]$ to $[0,1]$. For this larger class, the bracketing bound is $\log N_{[]}(\eta, \mathcal{D}_1, L_2(Q)) \le C\eta^{-1}$, where $0 < \eta < 1$, and C is independent of the probability measure Q. See van de Geer (1991); Van De Geer (2009). The same order holds for functions with total variation at most one and a uniformly bounded value at a fixed point. The latter restriction is necessary to control additive constants.*

Thus $\alpha_1 = 0$. The entropy bound holds directly under the design distribution, without a density assumption. Lemma C.12 is therefore unnecessary when $r = 1$.

Remark C.10 (Why we use bracketing entropy)**.** *For $r \ge 2$, the metric entropy of distribution functions under Lebesgue measure satisfies $\log N(\eta, \mathcal{D}_r, L_2(\lambda)) \lesssim \eta^{-1}\{\log(1/\eta)\}^{r-1/2}$, where $0 < \eta < e^{-1}$. This order is sharp when $r = 2$ Blei et al. (2007). However, this covering bound does not directly supply the bracketing bound required by the empirical-process inequality in Lemma C.14. In general, a small covering number does not imply a small bracketing number.*

A bounded design density allows both covers and brackets to transfer from Lebesgue measure to the population design measure. Indeed, if the density is bounded above by M, then

$$\|f - g\|_{L_2(P)} \le \sqrt{M}\,\|f - g\|_{L_2(\lambda)}.$$

The additional difficulty is control under the random empirical measure. A population L_2 cover alone does not provide such control. Brackets provide pointwise bounds $\ell \le f \le u$. These bounds allow the bracketing maximal inequality to control the empirical process using population bracket widths. Proposition C.8, obtained from Gao (2013), supplies precisely this input.

Supremum-norm covering is not a useful substitute for the fitted class. It contains the distribution functions

$$h_t(z) = \mathbf{1}\{z_1 \ge t\}, \qquad 0 < t < 1.$$

Distinct thresholds satisfy $\|h_t - h_s\|_\infty = 1$. Consequently, $N(\eta, \mathcal{D}_r, \|\cdot\|_\infty) = \infty$, where $0 < \eta < \frac{1}{2}$. Uniform L_2 covering entropy is a different notion. It controls $\sup_Q \log N(\eta, \mathcal{D}_r, L_2(Q))$, where Q ranges over finitely supported probability measures. Such a bound could support an alternative empirical-process argument. However, it does not follow from the displayed Lebesgue-measure covering bound. Our proof instead uses the available bracketing bound. We do not claim that every proof of the statistical rate must use brackets.

Lemma C.11 (Entropy of the constrained class)**.** *Put $\alpha_r = 2(r-1)$. There are constants $A \ge 1$ and $D \ge \max\{1, 2B_0\}$, depending only on the range, variation, dimension, and density bounds, such that*

$$\log N_{[]}(v, \mathcal{F}_V, L_2(P)) \le \frac{A}{v}\left(\log\frac{eD}{v}\right)^{\alpha_r}, \qquad 0 < v \le 2B_0,$$

whenever P has density bounded above by the specified constant. The assertion includes the zero-variation class of constant functions.

Proof. For positive mass, Lemma C.5 writes the nonconstant part as v_*F, with $0 < v_* \le V_{\max}$ and $F \in \mathcal{D}_r$. The zero-mass function is treated separately. For $V_{\max} > 0$, bracket F at resolution $\epsilon/(4V_{\max})$, partition the mass interval into pieces of length at most $\epsilon/4$, and partition the intercept interval into pieces of length at most $\epsilon/2$. Truncate CDF bracket endpoints to $[0,1]$. For $v_* \in [v_-, v_+]$ and $\ell \le F \le u$,

$$v_-\ell \le v_*F \le v_+u, \qquad \|v_+u - v_-\ell\|_2 \le V_{\max}\|u - \ell\|_2 + (v_+ - v_-) \le \epsilon/2.$$

The numbers of mass and intercept intervals are bounded by $1 + 4V_{\max}/\epsilon$ and $1 + 4G_{\max}/\epsilon$. Their logarithms add to the CDF entropy. For $V_{\max} = 0$, only the intercept partition is needed. Enlarging A, D covers all resolutions in the stated interval. Finally apply Lemma C.12. □

Lemma C.12 (Transfer to the design distribution)**.** *Suppose P_Z admits a density bounded above by M_0, as in Assumption 3. Then for every class $\mathcal{F}$ and every $v > 0$,*

$$N_{[]}(\sqrt{M_0}\, v, \mathcal{F}, L_2(P_Z)) \le N_{[]}(v, \mathcal{F}, L_2(\lambda)).$$

The same statement holds for any probability measure on $[0,1]^r$ with density bounded above by M, with M in place of M_0. Consequently, in Lemma C.11 the constants may be taken as $A = A_\lambda\sqrt{M}$ and $D = D_\lambda\sqrt{M}$, where A_λ, D_λ are the constants for Lebesgue measure.

Proof. For any measurable h we have $\|h\|^2_{L_2(P_Z)} = \int h^2\, dP_Z \le M_0 \int h^2\, d\lambda = M_0\|h\|^2_{L_2(\lambda)}$. A bracket of width v in $L_2(\lambda)$ is therefore a bracket of width at most $\sqrt{M_0}\, v$ in $L_2(P_Z)$, and the same collection of brackets serves both purposes. □

Remark C.13 (The gap between coordinate-wise and entire monotonicity)**.** *Entire monotonicity imposes nonnegativity of every mixed difference, whereas coordinate-wise monotonicity controls only first-order differences. This additional structure reduces the complexity of the function class.*

For uniformly bounded coordinate-wise monotone functions on $[0,1]^2$, the $L_2(\lambda)$ bracketing entropy is bounded above by $C\epsilon^{-2}\{\log(1/\epsilon)\}^2$ Gao & Wellner (2007). The polynomial entropy exponent is 2, at the boundary of integrability of the square-root entropy. The corresponding global least-squares rate in squared L_2 loss is $n^{-1/2}$ up to logarithmic factors under the usual design and noise conditions.

In contrast, Lemma C.5 represents a right-continuous entirely monotone function as an additive constant plus a nonnegative multiple of a distribution function. Under the intercept and variation bounds, the bracketing entropy has polynomial exponent 1. This yields the $n^{-2/3}$ statistical term in Theorem C.15. Such a rate cannot hold uniformly over the full bounded bivariate coordinate-wise monotone class. For $r = 1$, the two monotonicity notions coincide.

C.1.1 The improved estimation rate

From empirical fit to prediction error. A small training error does not by itself guarantee accurate predictions on fresh data. We therefore compare the empirical and population squared losses. The conditional mean of the response need not belong to our fitted class. This occurs, for example, when the fitted coordinates come from an estimated chart. We assume instead that the class contains a function uniformly close to that conditional mean.

The lemma below separates three sources of prediction error. These are the error from a finite sample, the approximation error of the fitted class, and the error from approximate optimisation. Its proof controls the whole loss difference. This treats the noise fluctuation and the fluctuation of the squared prediction error together, without a separate empirical-to-population norm conversion.

Lemma C.14 (Least squares with bounded approximation error)**.** *Let $(U_i, Y_i)_{i=1}^n$ be independent observations with a common distribution. Here U_i is the regression input and Y_i is the response. Write P_U for the distribution of a fresh input U, and define the conditional mean by $m(u) := \mathbb{E}[Y \mid U = u]$. Thus $Y_i = m(U_i) + \varepsilon_i$. Assume that, conditionally on U_i, the noise ε_i is centred and sub-Gaussian with variance proxy s^2, as in Assumption 4. (In the bandit applications below, $s^2 \le \sigma^2 + 4G^2_{\max}$.)*

Let $f_0 \in \mathcal{F}_V$ be a fixed comparator satisfying $\|m - f_0\|_\infty \le \beta$, for some $\beta \ge 0$. Let $\widehat{f} \in \mathcal{F}_V$ be a measurable η_n-approximate least-squares estimator in the sense of Defn. C.7. Write $B := B_0 = G_{\max} + V_{\max}$ for the uniform bound of $\mathcal{F}_V$. Assume $|m(u)| \le B$ for P_U-almost every u. In our bandit applications this follows from $|g_a| \le G_{\max} \le B$. Take the optimisation tolerance η_n to be deterministic. Recall the bracketing entropy constants $A \ge 1$, $D \ge \max\{1, 2B\}$, and logarithmic exponent $\alpha \ge 0$ from Lemma C.11, applied with $P = P_U$; this requires the density of P_U to be bounded above, and A, D depend on that bound as in Lemma C.12. For $0 < \delta < 1$, define

$$\ell_\delta := \log \frac{8enD(1+A)}{\delta}, \qquad q_\alpha := \max\left\{2, \frac{2\alpha+2}{3}\right\}.$$

Then, with probability at least $1 - \delta$,

$$\|\widehat{f} - f_0\|^2_{L_2(P_U)} \le C\left\{A^{2/3}n^{-2/3}\ell_\delta^{q_\alpha} + \beta^2 + \eta_n\right\}.$$

Moreover,

$$\mathbb{E}\|\widehat{f} - f_0\|^2_{L_2(P_U)} \le C\left\{A^{2/3}n^{-2/3}\{\log(8enD(1+A))\}^{q_\alpha} + \beta^2 + \eta_n\right\}.$$

The constant C depends only on B, σs, and α. The same bounds, with a different constant, hold with m in place of f_0 on the left.

Proof. Throughout the proof, C denotes a finite constant depending only on B, s, and α. Its value may change between displays. Write $\ell = \ell_\delta$.

Bounding the responses for the analysis. We first clip the noise at the level $H = s\sqrt{2\ell}$. Define

$$\varepsilon^H := \max\{-H, \min\{\varepsilon, H\}\}, \qquad Y^H := m(U) + \varepsilon^H.$$

When $s = 0$, the noise is zero almost surely and we set $\varepsilon^H = 0$. The estimator still uses the original responses. Clipping is only a device used in this proof.

The original and clipped samples agree unless at least one noise value exceeds H in absolute value. The sub-Gaussian tail bound and a union bound give

$$\mathbb{P}\big(\exists i \le n : \varepsilon_i^H \ne \varepsilon_i\big) \le 2ne^{-\ell}.$$

Clipping may introduce a small conditional bias. Write $b_H(u) := \mathbb{E}[\varepsilon^H \mid U = u]$. Since the original noise is conditionally centred, tail integration gives, for $s > 0$,

$$\|b_H\|_\infty \le \int_H^\infty 2e^{-v^2/(2s^2)}\,dv \le \frac{2s^2}{H}e^{-\ell}.$$

In particular, $\|b_H\|_\infty \le Ce^{-\ell}$.

The population loss controls prediction error. For this proof, let P denote expectation under the joint law of (U, Y^H). Let P_n denote the empirical average over the clipped sample. Explicitly, for a measurable function h,

$$Ph := \mathbb{E}[h(U, Y^H)], \qquad P_n h := \frac{1}{n}\sum_{i=1}^n h(U_i, Y_i^H).$$

For $f \in \mathcal{F}_V$, define its population distance from the comparator by $\rho(f) := \|f - f_0\|_{L_2(P_U)}$. Define its loss difference by

$$Q_f(U, Y^H) := (Y^H - f(U))^2 - (Y^H - f_0(U))^2.$$

Thus PQ_f is the difference in population squared loss, and P_nQ_f is the difference in empirical squared loss.

Expanding the squares and taking the conditional expectation of the noise gives

$$PQ_f = \rho(f)^2 - 2\mathbb{E}[\{m(U) - f_0(U) + b_H(U)\}\{f(U) - f_0(U)\}].$$

Cauchy–Schwarz and the inequality $2ab \le a^2/2 + 2b^2$ imply

$$PQ_f \ge \frac{1}{2}\rho(f)^2 - 2\{\beta + \|b_H\|_\infty\}^2.$$

Hence a function far from f_0 has a larger population loss, apart from the approximation and clipping errors.

Controlling the empirical loss fluctuation. We next bound $(P_n - P)Q_f$ uniformly over functions with $\rho(f) \le t$. Put $L_H := 2B + H$. This quantity controls the size of the clipped loss difference. Indeed,

$$|Q_f| \le CBL_H, \qquad PQ_f^2 \le CL_H^2 t^2 \quad \text{when } \rho(f) \le t.$$

To transfer brackets to the loss class, first clip their endpoints to $[-B, B]$. This preserves coverage of $\mathcal{F}_V$ and cannot increase their widths. For a fixed observation, the squared loss is CL_H-Lipschitz

in its fitted value on this interval. Taking the minimum and maximum of the loss over each bracket therefore gives a bracket for Q_f.

Define the bracketing entropy integral

$$J(t) := \int_0^t \sqrt{1 + \log N_{[]}(v, \mathcal{F}_V, L_2(P_U))}\, dv.$$

The assumed entropy bound gives

$$J(t) \le C \left\{ \sqrt{At} \left(\log \frac{eD}{t} \right)^{\alpha/2} + t \right\}, \qquad 0 < t \le 2B.$$

This is the entropy-integral calculation obtained by substituting $v = tu$. The additional t accounts for the constant one under the square root.

Apply the bounded bracketing maximal inequality (van der vaart & Wellner, 1996, Lemma 3.4.2) to the loss-difference class. Using the preceding envelope and second-moment bounds gives

$$\mathbb{E} \sup_{\rho(f) \le t} |(P_n - P)Q_f| \le C \left\{ \frac{L_H J(t)}{\sqrt{n}} + \frac{L_H^2 J(t)^2}{nt^2} \right\}.$$

The bounded empirical-process concentration inequality of Bousquet (2002), applied also to the negative class, then gives the following statement. For each fixed t and $x > 0$, outside an event of probability at most $2e^{-x}$,

$$\sup_{\rho(f) \le t} |(P_n - P)Q_f| \le C \left\{ \frac{L_H J(t)}{\sqrt{n}} + \frac{L_H^2 J(t)^2}{nt^2} + L_H t \sqrt{\frac{x}{n}} + \frac{L_H^2 x}{n} \right\}.$$

Choosing the statistical error scale. Choose a sufficiently large constant C_0, depending only on B, s, and α, and define

$$a_n^2 := C_0 A^{2/3} n^{-2/3} \ell^{q_\alpha}.$$

This is the squared statistical error scale. If $a_n \ge 2B$, the asserted probability bound follows immediately from boundedness.

Otherwise partition the possible distances in $(a_n, 2B]$ into dyadic shells. Let

$$M := \left\lceil \log_2 \frac{2B}{a_n} \right\rceil, \qquad t_j := \min\{2^j a_n, 2B\}, \quad j = 1, \dots, M.$$

Every distance $\rho \in (a_n, 2B]$ belongs to a shell whose upper endpoint satisfies $\rho \le t_j \le 2\rho$. Also $M \le C \log(en)$.

Apply the concentration bound at all these endpoints with $x = \log(4M/\delta)$. A union bound makes the resulting inequalities simultaneous, with failure probability at most $\delta/2$.

At these radii, $\log(eD/t_j) \le C\ell$, $L_H^2 \le C\ell$, and $x \le C\ell$. The terms in the concentration bound are at most $t_j^2/32$ once

$$t_j^3 \ge CA\ell^{\alpha+1}/n, \qquad t_j^2 \ge C\ell^2/n.$$

Both inequalities follow from $t_j \ge a_n$ and the definition of q_α, after increasing C_0. Consequently, on the simultaneous event,

$$|(P_n - P)Q_f| \le \frac{1}{8}\rho(f)^2 \qquad \text{for every } f \in \mathcal{F}_V \text{ with } \rho(f) > a_n.$$

Applying approximate minimisation. On the event that the original and clipped samples agree, the estimator satisfies $P_n Q_{\widehat{f}} \le \eta_n$, because $f_0 \in \mathcal{F}_V$. If $\rho(\widehat{f}) > a_n$, the population loss lower bound and the fluctuation bound therefore imply

$$\frac{3}{8}\rho(\widehat{f})^2 \le \eta_n + 2\{\beta + \|b_H\|_\infty\}^2.$$

If $\rho(\widehat{f}) \le a_n$, no further argument is needed. Combining these cases, and absorbing the exponentially small clipping bias, gives

$$\rho(\widehat{f})^2 \le C\{a_n^2 + \beta^2 + \eta_n\}.$$

The total failure probability is at most δ, since $2ne^{-\ell} \le \delta/2$. This proves the probability bound.

For the expectation bound, take $\delta = (en)^{-2}$. On the failure event use $\rho(\widehat{f})^2 \le 4B^2$. The resulting logarithm is at most a constant multiple of $\log(8enD(1+A))$, and the failure contribution is absorbed by the statistical term.

Finally,

$$\|\widehat{f} - m\|_{L_2(P_U)}^2 \le 2\|\widehat{f} - f_0\|_{L_2(P_U)}^2 + 2\beta^2.$$

This proves the corresponding bounds for prediction of the conditional mean. □

Theorem C.15 (Estimation rate)**.** *Under a fixed upper design-density bound, the estimator of Definition C.7 satisfies*

$$\mathbb{E}\|\widehat{f} - g\|_{L_2(P_Z)}^2 \le Cn^{-2/3}\{\log(en)\}^{q_r} + C\eta_n, \qquad q_r = \max\left\{2, \frac{4r-2}{3}\right\},$$

for a bounded entirely monotone truth in $\mathcal{F}_V$ and i.i.d. conditionally sub-Gaussian observations. With confidence $1-\delta$, replace $\log(en)$ by $\log(Cen/\delta)$. The constant depends on the fixed model parameters.

Proof. Apply Lemmas C.11 and C.14 with $f_0 = m = g$ and $\beta = 0$. □

Remark C.16 (Interpretation and dependence on dimension)**.** *The polynomial sample-size exponent $2/3$ does not depend on the effective dimension r. The dimension enters through the logarithmic power $q_r = \max\{2, (4r-2)/3\}$ and through the constant C. Thus the rate is dimension free only in its polynomial exponent. It is not a bound uniform over dimensions. The ambient dimension d does not enter this regression theorem directly, although it can affect the separate chart-estimation error.*

Up to logarithmic factors, the rate has the same $n^{-2/3}$ order as the fixed uniform lattice results of Fang et al. (2021). The present theorem gives a population prediction bound under random design with a bounded density. Lemma C.12 transfers the Lebesgue bracketing bound to the design distribution. Lemma C.14 then controls the population prediction error directly.

The exponent comes from the bracketing entropy of the variation-constrained class, whose polynomial entropy exponent is one. It does not require an infinite-atom implementation. An approximate empirical minimiser retains the statistical rate whenever its optimisation error satisfies $\eta_n = O(n^{-2/3})$. The additional logarithmic factors therefore leave the polynomial regret exponent in Theorem C.20 unchanged, under that theorem's assumptions.

Lemma C.17 (Basic inequality with slack)**.** *Let $\mathcal{F}$ be a class of functions on $[0,1]^r$ with $g \in \mathcal{F}$, let $Y_i = g(Z_i) + \varepsilon_i$, and let $\tilde{f}^{\mathrm{approx}} \in \mathcal{F}$ satisfy $\frac{1}{n}\|Y - \tilde{f}^{\mathrm{approx}}\|^2 \le \min_{f\in\mathcal{F}} \frac{1}{n}\|Y - f\|^2 + \delta_n$. Then*

$$\|\tilde{f}^{\mathrm{approx}} - g\|_{P_n}^2 \;\le\; \frac{2}{n}\sum_{i=1}^n \varepsilon_i\big(\tilde{f}^{\mathrm{approx}} - g\big)(Z_i) \;+\; \delta_n.$$

Proof. Since $g \in \mathcal{F}$, the minimum is at most $\frac{1}{n}\|Y - g\|^2 = \frac{1}{n}\sum \varepsilon_i^2$. Writing $Y_i - \tilde{f}^{\mathrm{approx}}(Z_i) = \varepsilon_i + (g - \tilde{f}^{\mathrm{approx}})(Z_i)$ and expanding,

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n (g - \tilde{f}^{\mathrm{approx}})^2(Z_i) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(g - \tilde{f}^{\mathrm{approx}})(Z_i) \;\le\; \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \delta_n,$$

and cancelling and rearranging gives the claim. □

C.2 Estimated coordinates and an exploration baseline

Assume Assumption 7 holds. Use the canonical lock, its rejection rule, and the good-chart event $\mathcal{L}$ from Section B.5.1. Lemma B.6 gives the required chart perturbation bound. For the mean-square oracle in (16), put

$$D_0 = \frac{C_2 dr}{\lambda_r}, \qquad C_W = 2\kappa_0^2\sqrt{rD_0}.$$

With n_{sub} subspace observations,

$$\mathbb{E}[\epsilon_W^2 \mathbf{1}_{\mathcal{L}}] \le C_W^2/n_{\mathrm{sub}}, \qquad \mathbb{P}(\mathcal{L}^c) \le D_0/(n_{\mathrm{sub}}\tau^2).$$

Here τ is defined in (17).

Construct an independent sample box after freezing the chart. A subsequent context outside this box is assigned an arbitrary arm and its observation is discarded. We never clip a regression design point. On the event $p_{\mathcal{B}} \ge 1/2$, Lemma B.9 gives a retained design density bounded by $2M_0'$. For a retained context write

$$U = \widehat{S}(\widehat{L}X), \qquad f_{0,a}(u) = g_a(\widehat{S}^{-1}u), \qquad \beta = L_g C_X \epsilon_W.$$

On $\mathcal{L}$, the comparator belongs to $\mathcal{F}_V$, and $|g_a(LX) - f_{0,a}(U)| \le \beta$. The conditional mean given U is within β of this comparator. Its centred noise is sub-Gaussian with proxy at most $\sigma^2 + 4G_{\max}^2$. This follows by conditioning first on the full context and then applying the bounded-variable moment bound to its reward mean. Thus Lemma C.14 applies without requiring monotonicity of the conditional mean.

Lemma C.18 (Prediction after estimating the chart)**.** *Adopt the assumptions and sampling scheme of Section C.2. Condition on a frozen chart $\widehat{L}$ and box $\widehat{\mathcal{B}}$ for which the good-chart event $\mathcal{L}$ holds and the retention probability satisfies $p_{\mathcal{B}} \ge 1/2$.*

Fix an arm a. Conditional on its retained sample size $n_a \ge 1$, suppose its fitting observations are i.i.d. from uniform exploration. Let $\widehat{f}_a \in \mathcal{F}_V$ be their least-squares fit with deterministic optimisation tolerance η_{n_a}. Recall

$$U = \widehat{S}(\widehat{L}X), \qquad f_{0,a}(u) = g_a(\widehat{S}^{-1}u), \qquad \beta = L_g C_X \epsilon_W.$$

For an independent test context drawn conditional on retention,

$$\mathbb{E}\left[\,|\widehat{f}_a(U) - g_a(LX)| \,\middle|\, \widehat{L}, \widehat{\mathcal{B}}, n_a\right] \le C n_a^{-1/3}\{\log(en_a)\}^{q_r/2} + C\beta + C\sqrt{\eta_{n_a}},$$

where $q_r = \max\{2, (4r-2)/3\}$. The expectation averages over the fitting sample and the independent retained test context. The constant C depends only on the fixed model parameters.

Proof. All expectations below are conditional on the frozen chart, the frozen box, and the sample size n_a. Let Q denote the distribution of the rescaled input U for a fresh retained context. Uniform exploration makes Q the input distribution for the retained observations of arm a as well. Lemma B.9 bounds its density by $2M_0'$.

The conditional mean response for this arm is $m_a(u) := \mathbb{E}[g_a(LX) \mid U = u,\ X \text{ retained}]$. On the good-chart event, $f_{0,a} \in \mathcal{F}_V$. The Lipschitz condition gives the pointwise bound $|g_a(LX) - f_{0,a}(U)| \le \beta$. Taking a conditional expectation therefore yields $\|m_a - f_{0,a}\|_{L_\infty(Q)} \le \beta$. The centred response noise has the conditional sub-Gaussian bound established in Section C.2. Hence Lemma C.14, together with Lemma C.11, gives

$$\mathbb{E}\|\widehat{f}_a - f_{0,a}\|_{L_2(Q)}^2 \le C\left\{n_a^{-2/3}\{\log(en_a)\}^{q_r} + \beta^2 + \eta_{n_a}\right\}.$$

For the test context, the triangle inequality separates estimation error from chart error. Independence of this context from the fitting sample and Cauchy–Schwarz give

$$\mathbb{E}|\widehat{f}_a(U) - g_a(LX)| \le \left(\mathbb{E}\|\widehat{f}_a - f_{0,a}\|_{L_2(Q)}^2\right)^{1/2} + \beta.$$

Substituting the squared-risk bound and using $\sqrt{x+y+z} \le \sqrt{x} + \sqrt{y} + \sqrt{z}$ proves the claim. □

Remark C.19 (When chart error is smaller than regression error)**.** *Under the assumed mean-square subspace guarantee, the expected chart contribution on the good-chart event is bounded by* $L_g C_X C_W n_{\mathrm{sub}}^{-1/2}$. *Write* $C_{\mathrm{chart}} := L_g C_X C_W$. *With three exploration blocks of comparable size,* $n_{\mathrm{sub}} \asymp T_{\mathrm{exp}}$. *On the event that the box retains probability at least one half, each arm receives order* T_{exp}/K *fitting observations with high probability, provided this quantity is sufficiently large.*

Ignoring logarithmic factors and fixed regression constants, the chart-error bound is no larger than the regression-error bound once $C_{\mathrm{chart}} T_{\mathrm{exp}}^{-1/2} \lesssim (T_{\mathrm{exp}}/K)^{-1/3}$, *equivalently* $T_{\mathrm{exp}} \gtrsim C_{\mathrm{chart}}^{6} K^{-2}$. *For the exploration choice* $T_{\mathrm{exp}} \asymp K T^{3/4}$, *up to logarithmic factors, the corresponding comparison threshold is* $T \gtrsim C_{\mathrm{chart}}^{8} K^{-4}$.

This threshold compares only the two prediction-error terms. The subspace oracle's sample requirements, the retained-count conditions, and the feasibility condition $T_{\mathrm{exp}} \le T$ *must also hold. The regret contributions from bad-chart events and discarded contexts are controlled separately.*

If the oracle constants remain fixed as K *varies, increasing* K *under this exploration schedule enlarges the pooled subspace sample while keeping the per-arm fitting sample of the same order. This explains the decreasing comparison threshold. Above it, the chart term does not change the leading polynomial regret exponent. The constants and the required horizon can still depend on the ambient dimension.*

C.3 The Algorithm and Regret Bound

We use an explore-then-commit algorithm called `EM-MIB`. We split the exploration sample into three blocks. The first estimates the common subspace and fixes its canonical coordinates. The second constructs a box and freezes its coordinatewise rescaling. The third fits one variation-constrained entirely monotone estimator per arm, as in Definition C.7, using only contexts inside the box. During exploitation, we choose the arm with the largest fitted reward for each in-box context and a fixed arm otherwise. If the canonical lock is rejected or the box is degenerate, we use a fixed arm for the remaining rounds. Algorithm 6 gives the complete procedure.

Theorem C.20 (Exploration-baseline regret)**.** *Under the chart, design, stationary reward, and mean-square subspace-oracle conditions above, for sufficiently large exploration blocks,*

$$\mathbb{E}[R_T] \le C \left\{ N + K^{4/3} T N^{-1/3} \{\log(eN)\}^{q_r/2} + K L_g C_X C_W T N^{-1/2} + T/N \right\}.$$

Constants in C *may include* D_0/τ^2. *If the budget fits within the horizon, choosing*

$$N \asymp K T^{3/4} \{\log(eT)\}^{3q_r/8}$$

gives

$$\mathbb{E}[R_T] \le \widetilde{O}\left(K T^{3/4} + \sqrt{K}\, L_g C_X C_W T^{5/8} \right).$$

Otherwise the trivial linear regret bound applies.

Proof. Exploration costs at most $2G_{\max} N$. Conditional on a good chart and a box of retained mass at least one half, each fitting count is binomial with success probability at least $1/(2K)$. Splitting at half its mean and using a Chernoff bound gives

$$\mathbb{E}[(n_a \vee 1)^{-1/3}] \le C(K/N)^{1/3}$$

when $N \ge K$; bounded prediction error covers smaller budgets and zero counts. Use Lemma C.18 and bound the maximum arm error by its sum. Averaging the chart term gives the displayed $K L_g C_X C_W T/\sqrt{N}$ contribution. Bad charts cost at most $2G_{\max} T D_0/(n_{\mathrm{sub}} \tau^2)$. Lemma B.9 bounds the probability of retained mass below one half by $2re^{-n_{\mathrm{box}}/(4r)}$. Exchangeability bounds the expected number of discarded post-exploration contexts by $2rT/(n_{\mathrm{box}} + 1)$. These terms are absorbed in CT/N. Balance the first two terms and substitute the resulting budget. □

The polynomial horizon exponent $3/4$ does not depend on the ambient dimension d. However, d enters the chart-estimation term through C_W and can affect the required exploration length and horizon threshold. Under the conditions of Remark C.19, the chart contribution is dominated by the regression contribution. This comparison does not establish a regret bound with constants uniform in d.

Algorithm 6 *Entirely Monotone Multi-Index Bandits* (EM−MIB)

Input Horizon T, exploration budget $6 \le N \le T$, arms $[K]$, and class $\mathcal{F}_V$.
Set $b = \lfloor N/3 \rfloor$ and

$$I_{\mathrm{sub}} = \{1, \dots, b\}, \quad I_{\mathrm{box}} = \{b+1, \dots, 2b\}, \quad I_{\mathrm{fit}} = \{2b+1, \dots, N\}.$$

for $t = 1, \dots, N$ **do**
  Observe X_t, draw $A_t \sim \mathrm{Unif}([K])$ independently of X_t and the past, and observe Y_t.
end for
Compute $\widehat{L}_{\mathrm{raw}}$ from $\{(X_t, A_t, Y_t) : t \in I_{\mathrm{sub}}\}$ using the estimator assumed in (16).
Let $\widehat{V}_{\mathrm{raw}}$ be its first r columns.
if $\sigma_{\min}(\widehat{V}_{\mathrm{raw}}) < (2\kappa_0)^{-1}$ **then**
  Set $A_t = 1$ for $t > N$ and **stop**.
end if
Apply the canonical lock of Section B.5.1 and Lemma B.6

$$\widehat{L} = \widehat{V}_{\mathrm{raw}}^{-1} \widehat{L}_{\mathrm{raw}}.$$

For $j \in [r]$, set $\widehat{\ell}_j = \min_{t \in I_{\mathrm{box}}} (\widehat{L}X_t)_j$, $\widehat{u}_j = \max_{t \in I_{\mathrm{box}}} (\widehat{L}X_t)_j$.
if $\widehat{u}_j = \widehat{\ell}_j$ for some $j \in [r]$ **then**
  Set $A_t = 1$ for $t > N$ and **stop**.
end if
Freeze the box and rescaling map of Lemma B.9

$$\widehat{\mathcal{B}} = \prod_{j=1}^{r} [\widehat{\ell}_j, \widehat{u}_j], \qquad \widehat{S}(z)_j = \frac{z_j - \widehat{\ell}_j}{\widehat{u}_j - \widehat{\ell}_j}.$$

for $a \in [K]$ **do**
  Form

$$\mathcal{D}_a = \left\{ \left(\widehat{S}(\widehat{L}X_t), Y_t \right) : t \in I_{\mathrm{fit}},\ A_t = a,\ \widehat{L}X_t \in \widehat{\mathcal{B}} \right\}, \qquad n_a = |\mathcal{D}_a|.$$

  if $n_a = 0$ **then**
    Set $\widehat{f}_a = 0$.
  else
    Compute $\widehat{f}_a$ from $\mathcal{D}_a$ by Definition C.7, with tolerance $\eta_{n_a} \le n_a^{-2/3}$.
  end if
end for
for $t = N+1, \dots, T$ **do**
  Observe X_t and set $\widehat{Z}_t = \widehat{L}X_t$.
  if $\widehat{Z}_t \in \widehat{\mathcal{B}}$ **then**
    Choose $A_t \in \arg\max_{a \in [K]} \widehat{f}_a(\widehat{S}(\widehat{Z}_t))$, breaking ties by the smallest arm index.
  else
    Set $A_t = 1$.
  end if
  Play A_t.
end for

C.4 Lower Bound

We construct entirely monotone rewards for which every policy incurs regret of order $K^{1/3}T^{2/3}$. The projection is known to the learner. Thus the lower bound concerns learning the reward functions. The construction satisfies the linearity and population SIR identifiability assumptions. It does not rely on a subspace-estimation oracle.

The construction. Assume $V_{\max}, G_{\max}, L_g > 0$, and define

$$\Lambda_{\mathrm{lb}} := \min\{V_{\max}, G_{\max}, L_g/2\}, \qquad w_i := \frac{Lambda_{\mathrm{lb}}}{2r}\left(1 + \frac{i}{r}\right), \quad i \in [r]. \tag{40}$$

The weights are positive and distinct. They satisfy $w_i \leq$ $Lambda_{\mathrm{lb}}/r$ and $\sum_i w_i \leq$ $Lambda_{\mathrm{lb}}$.

Fix $d \geq r$. Let X be uniform on $[-1/2, 1/2]^d$, take $L = [I_r \mid 0]$, and use independent Gaussian noise with variance $\sigma^2 > 0$. Write $Z = LX$. To define the rewards on all of $\mathbb{R}^r$, set $u_i(z) := \min\{1, \max\{0, z_i + 1/2\}\}$. On the support of Z, the vector $u(Z)$ is uniform on $[0,1]^r$.

Partition the first coordinate of this unit cube into N intervals of length $1/N$. For interval j, let Ψ_j be the continuous trapezoidal bump given by

$$\Psi_j(t) := \max\left\{0,\, \min\left\{t - \frac{j-1}{N},\, \frac{1}{4N},\, \frac{j}{N} - t\right\}\right\}.$$

It rises with slope one, is constant on the middle half of the interval, and falls with slope minus one. Its height is $1/(4N)$. It vanishes outside its interval.

An instance is indexed by $\omega \in [K]^N$. The entry ω_j specifies the preferred arm in interval j. Define

$$F_a^\omega(t) := t + \sum_{j:\, \omega_j = a} \Psi_j(t), \qquad g_a^\omega(z) := w_1 F_a^\omega(u_1(z)) + \sum_{i=2}^{r} w_i u_i(z). \tag{41}$$

Every arm shares the same additive baseline. Only its bumps depend on ω. Because each bump returns to zero, changing ω_j changes the rewards only in the corresponding slab.

Shape, smoothness, and design conditions. The slopes of F_a^ω lie in $[0, 2]$. Also $F_a^\omega(0) = 0$ and $F_a^\omega(1) = 1$. Hence each reward is a sum of nondecreasing univariate functions. Its first-order differences are nonnegative, and its mixed differences of order at least two vanish. Thus it is entirely monotone.

In the coordinates u, the reward vanishes at the origin and has corner difference $\sum_i w_i \leq$ $Lambda_{\mathrm{lb}}$. Consequently, it satisfies the variation constraint. On all of $\mathbb{R}^r$, its values lie in [0, $Lambda_{\mathrm{lb}}$]. Its almost-everywhere gradient satisfies $\|\nabla g_a^\omega\| \leq ($ $Lambda_{\mathrm{lb}}/r)\sqrt{r+3} \leq 2$ $Lambda_{\mathrm{lb}} \leq L_g$. Since the function is continuous and piecewise affine, this proves the global Lipschitz bound.

The feature and design assumptions hold whenever $C_X \geq \sqrt{d}/2$ and $M_0 \geq 1$. Moreover, $\Sigma_X = I_d/12$ and $\mathbb{E}[X \mid LX] = (LX, 0)$. This verifies Assumption 5. The true canonical chart has leading block I_r, so its conditioning requirement holds for $\kappa_0 \geq 1$.

The rewards are entirely monotone on every rectangle, including a common rectangle containing all nearby projected supports. The nearby design condition also holds with compatible constants. Indeed, conditional on the last $d - r$ coordinates of X, the variable $[I_r \mid W']X$ is uniform on a translated unit cube. Its unconditional density is therefore at most one. Its support rectangle has side lengths $1 + \sum_{\ell=1}^{d-r} |W'_{i\ell}|$. After rescaling this rectangle, the density is bounded above by

$$\prod_{i=1}^{r}\left(1 + \sum_{\ell=1}^{d-r} |W'_{i\ell}|\right) \leq \left(1 + \sqrt{d-r}\,\rho^*\right)^r \qquad \text{when } \|W'\|_F \leq \rho^*.$$

Thus Assumption 9 holds whenever its constant M_0' is at least this bound. For $d = r$, the bound is one.

Population identifiability. Distinct weights are needed to make every index direction visible to inverse regression. We first verify this for the common baseline.

Lemma C.21 (Identifiability of the baseline)**.** *Let Z be uniform on $[-1/2, 1/2]^r$ and let $Y_0 = \sum_i w_i Z_i + \xi$, where $\xi \sim N(0, \sigma^2)$ is independent of Z. Then $M_{\mathrm{base}} := \mathrm{Cov}(\mathbb{E}[Z \mid Y_0])$ is positive definite. In particular, $\lambda_* := 12\lambda_{\min}(M_{\mathrm{base}}) > 0$ depends only on r,* $Lambda_{\mathrm{lb}}$*, and σ.*

Proof. Suppose $v^\top M_{\mathrm{base}} v = 0$. Since $\mathbb{E}[Z] = 0$, this implies $\mathbb{E}[v^\top Z \mid Y_0] = 0$. Therefore $\mathbb{E}[(v^\top Z)e^{tY_0}] = 0$ for every $t > 0$. Independence of the Gaussian noise allows its positive moment-generating function to be cancelled.

For a uniform variable on $[-1/2, 1/2]$, the moment-generating function is $m(s) = \sinh(s/2)/(s/2)$. Independence of the coordinates then gives

$$\sum_{i=1}^{r} v_i q(tw_i) = 0, \qquad q(s) := \frac{m'(s)}{m(s)} = \frac{1}{2} - \frac{1}{s} + \frac{1}{e^s - 1}.$$

Letting t tend to infinity first gives $\sum_i v_i = 0$. Multiplying the remaining identity by t and taking the limit gives $\sum_i v_i/w_i = 0$. Hence $\sum_i v_i/(e^{tw_i} - 1) = 0$ for every $t > 0$.

Multiply this identity by e^{tw_1} and let t tend to infinity. Since w_1 is strictly smaller than every other weight, the limit gives $v_1 = 0$. Repeat with the remaining weights to obtain $v_i = 0$ for every i. Thus M_{base} is positive definite. □

The baseline in (41) differs from Y_0 only by the constant $\sum_i w_i/2$. This shift does not change its inverse-regression covariance.

Lemma C.22 (Identifiability after adding the bumps)**.** *Put $\varepsilon := w_1/(4N)$. Under uniform independent arm selection, every constructed instance satisfies*

$$\lambda_r \geq \lambda_* - \frac{9r\varepsilon}{\sigma}.$$

Proof. The last $d - r$ feature coordinates are independent of (Z, A, Y) and have mean zero. Since $\Sigma_X = I_d/12$, the nonzero eigenvalues in Assumption 6 are those of $12\mathrm{Cov}(\mathbb{E}[Z \mid Y, A])$.

Each arm's reward differs from the baseline by at most ε. The Gaussian divergence formula therefore bounds the Kullback–Leibler divergence between the baseline and perturbed laws of (Z, A, Y) by $\varepsilon^2/(2\sigma^2)$. Pinsker's inequality bounds their total variation distance by $\varepsilon/(2\sigma)$.

Fix a unit vector v and put $W = v^\top Z$. Then $|W| \leq \sqrt{r}/2$ and $\mathbb{E}[W] = 0$. Under either law,

$$\mathbb{E}[\mathbb{E}[W \mid Y, A]^2] = \sup_{|h| \leq \sqrt{r}/2} \mathbb{E}[2Wh(Y, A) - h(Y, A)^2].$$

Every integrand on the right has absolute value at most $3r/4$. Changing the joint law therefore changes this supremum by at most $3r\varepsilon/(4\sigma)$. This bounds the change in every unit-vector quadratic form. Multiplying by twelve proves the claimed eigenvalue bound. □

The regret bound. The next theorem concerns the model assumptions verified above. Neither Assumption 7 nor Assumption 11 is imposed.

Theorem C.23 (Lower bound for entirely monotone multi-index bandits)**.** *Let $d \geq r \geq 1$, $K \geq 2$, and $\sigma > 0$. Suppose $V_{\max}, G_{\max}, L_g > 0$, $C_X \geq \sqrt{d}/2$, and $M_0 \geq 1$. Take compatible chart constants as described above. Let*
$Lambda_{\mathrm{lb}}, w_1$ be defined in (40)*, and let λ_* be given by Lemma C.21. Define*

$$T_{\mathrm{lb}} := \max\left\{ \frac{2K\sigma^2}{w_1^2},\ \frac{24^{3/2} w_1}{\sqrt{2K}\,\sigma},\ \frac{729 K r^3 w_1}{4\lambda_*^3 \sigma} \right\}. \tag{42}$$

For every $T \geq T_{\mathrm{lb}}$ and every policy π,

$$\sup_{\omega} \mathbb{E}_\omega[R_T(\pi)] \geq \frac{1}{128} w_1^{1/3} \sigma^{2/3} K^{1/3} T^{2/3} \geq \frac{1}{128} \left(\frac{\Lambda_{\mathrm{lb}}}{2r} \right)^{1/3} \sigma^{2/3} K^{1/3} T^{2/3}.$$

The instances satisfy Assumptions 1, 2, 3, 4, 5, and 6, with $\lambda_r \geq \lambda_*/2$. They also satisfy the entire-monotonicity, variation, and chart conditions verified above. The bound holds when $L = [I_r \mid 0]$ is known to the learner.

Proof. Choose $N = \lceil N_* \rceil$, where $N_* = (w_1^2 T/(2K\sigma^2))^{1/3}$, and draw ω uniformly from $[K]^N$. A lower bound on the average regret is also a lower bound on the worst-instance regret.

Mistakes within a slab. Fix a slab j. Let m_j count the contexts whose first rescaled coordinate lies in its interval. Let s_j count those in the middle half of that interval. For each arm a, let $m_{j,a}$ and $s_{j,a}$ count the corresponding rounds on which the policy selects a.

On the middle half, every incorrect arm incurs regret $\varepsilon = w_1/(4N)$. The additive baseline cancels from all reward gaps. Unlike the product construction, no restriction on the other coordinates is needed.

Fix the entire context path and the other labels ω_{-j}. Let $\mathbb{P}_a$ denote the conditional history law when $\omega_j = a$. Let $\mathbb{P}_0$ denote the comparison law obtained by removing the bump in slab j from every arm. The policy uses the same decision rule under these laws. The chain rule for Gaussian divergence gives

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_a) \le \frac{\varepsilon^2}{2\sigma^2} \mathbb{E}_0[m_{j,a}].$$

The contexts have the same exogenous law under every instance, so conditioning on their path is valid.

Since $s_{j,a}$ lies between zero and the fixed count s_j, Pinsker's inequality gives $\mathbb{E}_a[s_{j,a}] \le \mathbb{E}_0[s_{j,a}] + s_j \varepsilon \sqrt{\mathbb{E}_0[m_{j,a}]}/(2\sigma)$. Average over a. Using $\sum_a \mathbb{E}_0[s_{j,a}] = s_j$ and Cauchy–Schwarz yields

$$\frac{1}{K} \sum_{a=1}^{K} \mathbb{E}_a[s_j - s_{j,a}] \ge s_j \left(1 - \frac{1}{K} - \frac{\varepsilon}{2\sigma} \sqrt{\frac{m_j}{K}}\right).$$

For $m_j = 0$, both sides vanish. For $m_j > 0$, the right side is at least $s_j/4$ whenever $\varepsilon \le (\sigma/2)\sqrt{K/m_j}$.

Averaging over the contexts. Write $\mu = T/N$ and retain only context paths with $m_j \le 2\mu$. The count m_j is binomial with mean μ, while $\mathbb{E}[s_j] = \mu/2$. For $\mu \ge 12$, Cauchy–Schwarz and the binomial Chernoff bound imply

$$\mathbb{E}[s_j \mathbf{1}_{\{m_j \le 2\mu\}}] \ge \frac{\mu}{2} - \mathbb{E}[m_j \mathbf{1}_{\{m_j > 2\mu\}}] \ge \frac{\mu}{2} - \sqrt{2}\, \mu e^{-\mu/6} \ge \frac{\mu}{4}.$$

On these paths, the preceding testing bound applies if $\varepsilon \le (\sigma/2)\sqrt{KN/(2T)}$. Our choice $N \ge N_*$ ensures this inequality.

Average also over ω_{-j} and sum over the N slabs. Regret is nonnegative outside their middle halves. Thus

$$\frac{1}{K^N} \sum_{\omega} \mathbb{E}_\omega[R_T] \ge \frac{\varepsilon}{4} \sum_{j=1}^{N} \mathbb{E}[s_j \mathbf{1}_{\{m_j \le 2\mu\}}] \ge \frac{\varepsilon T}{16}.$$

Checking the parameter choices. The first condition in (42) gives $N_* \ge 1$, so $N \le 2N_*$. The second gives $T \ge 24 N_*$ and hence $T/N \ge 12$. The third gives $N_* \ge 9 r w_1/(2\lambda_* \sigma)$. Consequently, $\varepsilon \le \lambda_* \sigma/(18r)$, and Lemma C.22 ensures $\lambda_r \ge \lambda_*/2$ uniformly over ω.

Finally, $\varepsilon \ge w_1/(8N_*) \ge w_1^{1/3} \sigma^{2/3} (K/T)^{1/3}/8$. Substituting this into the preceding regret bound proves the theorem. □

Remark C.24 (Scope of the lower bound)**.** *The explicit dependence on the number of arms is $K^{1/3}$. The horizon exponent $2/3$ is independent of r and d. The prefactor and the required horizon are not uniform in the effective dimension. In particular, the threshold depends on λ_*^{-3}.*

The lower bound verifies population SIR identifiability. Theorem C.32 removes the two additional assumptions used in Theorem C.27. It retains the mean-square subspace-oracle assumption already used by `EM-MIB`*. The horizon exponent is therefore matched for a known chart. For an unknown chart, the matching minimax conclusion requires a uniform oracle guarantee on a class containing the lower-bound family, as stated in Corollary C.34.*

C.5 Improved Upper Bound via an Offline Regression Oracle

The EM−MIB upper bound in Theorem C.20 has horizon exponent $3/4$, whereas the lower bound in Theorem C.23 has exponent $2/3$. The regression estimator itself has squared-error rate $\widetilde{O}(n^{-2/3})$. We now reuse this estimator within an epoch-based procedure. Under the additional subspace and sampling-design assumptions stated below, this gives a regret bound with horizon exponent $2/3$.

Remark C.25 (Where the EM−MIB exponent comes from)**.** *Suppose each armwise estimator has squared prediction error of order $n^{-2\gamma}$. Ignoring logarithmic factors, chart error, and fixed constants, our explore-then-commit analysis gives*

$$\mathbb{E}[R_T] \lesssim T_{\exp} + KT(K/T_{\exp})^{\gamma}.$$

The second term uses about $T_{\exp}/K$ samples per arm and sums the absolute prediction errors over the arms. Balancing the two terms gives $T_{\exp} \asymp KT^{1/(1+\gamma)}$, provided this budget fits within the horizon. For $\gamma = 1/3$, the resulting bound is of order $KT^{3/4}$.

This calculation explains the exponent in the explore-then-commit bound. The procedure stops collecting useful fitting data after exploration. The method below continues to update its estimates and allocates exploration according to the estimated reward gaps.

We use FALCON+, Algorithm 2 and Theorem 2 of Simchi-Levi & Xu (2022). It divides the retained observations into epochs of doubling length. At the start of each epoch, it fits the reward functions using the preceding epoch's data. These fits remain fixed throughout the new epoch.

For each context, the algorithm identifies an arm with the largest fitted reward. Each other arm receives probability $1/(K + \gamma_m \widehat{\Delta}_a)$, where $\widehat{\Delta}_a$ is its fitted reward gap and γ_m controls exploration. The fitted best arm receives the remaining probability. Thus arms with smaller estimated gaps receive more exploration.

The fitting procedure is called an *offline regression oracle*. Here its error must be controlled under the joint context–action distribution generated in the preceding epoch. A regression guarantee under the marginal context distribution alone is insufficient for this application.

Proposition C.26 below establishes the required joint prediction bound. Writing $\Xi(n)$ for its calibrated error envelope, the reduction gives an epoch contribution of order $|\text{epoch } m|\sqrt{K\Xi(n_m)}$, where n_m is the preceding epoch's sample size. The statistical part of our envelope is $\widetilde{O}(K^{2/3}n^{-2/3})$. Summing over epochs therefore gives $\widetilde{O}(K^{5/6}T^{2/3})$. The chart approximation error and the initial exploration cost are accounted for separately below. The two assumptions stated next are used in Theorem C.27. Section C.6 shows that both can be removed, at the cost of replacing $K^{5/6}$ by K.

Assumption 11 (High-probability subspace oracle)**.** *For the uniform first-stage sample, a known constant D_{hp} satisfies, for each prescribed confidence level δ,*

$$\mathbb{P}\left(\|P_{\widehat{L}} - P_L\|_{\mathrm{op}} > D_{\mathrm{hp}}\sqrt{\frac{\log(2/\delta)}{n_{\mathrm{sub}}}}\right) \le \delta.$$

The oracle is used only above its stated minimum sample size. This assumption is separate from the expectation bound (16)*.*

Assumption 12 (Conditional design bound within retained epochs)**.** *Condition on the frozen chart and box, and on the history before an epoch. On the good-chart and good-box events, the rescaled retained context, conditional on selection of any arm with positive probability, has a Lebesgue density bounded by a known constant M_{arm}. The bound is uniform over the epochs.*

This is an additional assumption on the sampling designs. It does not follow from Assumption 9. The latter bounds the context density before conditioning on the selected arm. We impose Assumption 12 only on retained contexts, so it is not contradicted by clipping atoms. No clipping is used here.

Proposition C.26 (Joint regression guarantee)**.** *Fix a good chart and a box of retained mass at least one half. Let n retained observations be generated using a fixed action kernel. Fit each arm by Definition C.7, with tolerance at most $(n_a \vee 1)^{-2/3}$. Use zero for an arm with no observations. Under*

Algorithm 7 Entirely monotone bandits with FALCON+

Split T_1 initial rounds into independent subspace and box blocks. Sample arms independently and uniformly in the subspace block.
Estimate the chart and apply the canonical lock and its rejection rule. Form a sample box from the second block. Freeze the chart and box. Use a fixed arm thereafter if the lock is rejected or the box is degenerate.
On every later out-of-box context, play an arbitrary arm and discard its observation. Count only in-box rounds in the following epochs.
Use epochs of lengths $2, 4, 8, \ldots$ on this retained sequence. Sample uniformly in the first epoch.
for each later epoch m **do**
  Fit each arm using only its retained observations from epoch $m-1$. Fit each arm as in Proposition C.30, with tolerance $\eta_{n_a} \le n_a^{-2/3}$ when $n_a \ge 1$. Use the zero function when $n_a = 0$.
  Let n_m be the preceding epoch length and set $\gamma_m = \frac{1}{2}\sqrt{K/\Xi(n_m)}$.
  for each retained round t of epoch m **do**
    Compute $U_t = \widehat{S}(\widehat{L}X_t)$ and $\widehat{a}_t \in \arg\max_a \widehat{f}_{m,a}(U_t)$.
    For $a \neq \widehat{a}_t$, set

$$p_t(a) = \frac{1}{K + \gamma_m[\widehat{f}_{m,\widehat{a}_t}(U_t) - \widehat{f}_{m,a}(U_t)]}.$$

    Assign the remaining probability to $\widehat{a}_t$. Sample an arm and record its reward.
  end for
end for

Assumption 12, with probability at least $1-\delta$,

$$\mathbb{E}_{X|\text{retained},\, A\sim p(\cdot|X)}[\{\widehat{f}_A(U) - g_A(LX)\}^2] \le C\left\{K^{2/3}n^{-2/3}\left(\log\frac{CenK}{\delta}\right)^{q_r} + \beta^2\right\}.$$

Here $\beta = L_g C_X \epsilon_W$*, and the constant depends on* M_{arm} *and the fixed regression parameters.*

Proof. Let q_a be the marginal selection probability for arm a, and put $u = \log(CnK/\delta)$. Its count is binomial with parameters n, q_a. Conditional on the count, its observations are i.i.d. from the arm-conditional design. The comparator and residual-noise argument preceding Lemma C.18 applies to this design. For arms with $nq_a \ge cu$, a Chernoff bound and a union bound ensure $n_a \ge nq_a/2$. Lemma C.14, with confidence allocated over the arms, therefore gives a weighted error at most

$$Cn^{-2/3}u^{q_r}\sum_a q_a^{1/3} + C\beta^2 \le CK^{2/3}n^{-2/3}u^{q_r} + C\beta^2.$$

The same bound holds for the squared error against the original full-context reward, by the pointwise comparator error and the squared triangle inequality. For the remaining arms, boundedness gives total error at most $C\min\{1, Ku/n\}$. Since $q_r \ge 2$, this is bounded by a constant multiple of $K^{2/3}n^{-2/3}u^{q_r}$. This also covers zero counts. Allocate the remaining confidence to the count events. □

For a known horizon, take one confidence budget sufficiently small for all arm fits and epochs. Replace the sample-dependent logarithm in the last proposition by the larger deterministic value $\log(CeKT^3)$. It follows that a known constant C_* gives a nonincreasing valid error envelope

$$\Xi(n) = \min\left\{4B_0^2, C_*\left[K^{2/3}n^{-2/3}\{\log(CeKT^3)\}^{q_r} + \overline{\beta}^{\,2}\right]\right\}.$$

The cap follows from the predictor and reward bounds. Degenerate zero-reward cases are treated separately. All constants used to calibrate this envelope are assumed available to the learner, as required by the oracle-based reduction.

Theorem C.27 (Improved expected regret)**.** *Suppose Assumptions 1, 2, 4, 10, 8, 9, 11, and 12 hold. Assume also the stationary sampling condition stated in Section C.For any fixed arm-selection*

rule $p(a \mid x)$, where $p(a \mid x)$ is the probability of choosing arm a at context x, assume that fresh context–action–reward observations are i.i.d. Let

$$C_{W,\mathrm{hp}} = 2\kappa_0^2\sqrt{r}\, D_{\mathrm{hp}}, \qquad D_* = L_g C_X C_{W,\mathrm{hp}}.$$

Choose T_1 large enough for the oracle sample threshold, for its prescribed deviation bound to be at most τ, and for $n_{\mathrm{box}} \geq Cr\log(eT)$. Use

$$\overline{\beta} = CD_*\sqrt{\frac{\log(eT)}{T_1}}.$$

For a sufficiently large known constant in Ξ, Algorithm 7 satisfies

$$\mathbb{E}[R_T] \leq \widetilde{O}\left(G_{\max}T_1 + K^{5/6}T^{2/3} + \sqrt{K}\, D_* T T_1^{-1/2} + G_{\max}T/T_1 + 1\right).$$

If the resulting initial stage fits within the horizon, balancing its first and third terms gives

$$T_1 \asymp \left(\frac{\sqrt{K}\, D_* T\sqrt{\log(eT)}}{G_{\max}}\right)^{2/3}$$

up to the stated minimum sample requirements, and hence

$$\mathbb{E}[R_T] \leq \widetilde{O}\left(K^{5/6}T^{2/3} + G_{\max}^{1/3}K^{1/3}D_*^{2/3}T^{2/3}\right).$$

This is a conditional upper bound. It uses the two additional assumptions stated in this subsection. No minimax claim is made here.

Proof. Use confidence of order T^{-2} for the subspace and box events. Lemma B.6 gives the stated $\overline{\beta}$. Lemma B.9 ensures retained mass at least one half except with the allocated probability. On these events, the retained contexts form an i.i.d. stochastic bandit with a fixed context law. The fits are fixed during each epoch. Proposition C.26 therefore supplies the preceding-kernel error guarantee required by `FALCON+`.

Apply Theorem 2 of Simchi-Levi & Xu (2022) through its expected-regret argument in the appendix subsection on bounding true regret. Its proof also covers known misspecification because the oracle error is measured against the true full-context mean. Affine reward rescaling supplies the source's mean-gap normalisation. Observed rewards need not be bounded for this expected-regret conclusion. The error cap allows uniform initialisation with a learning rate no larger than the later learning rates.

For the retained sequence, the bound is

$$C\sum_m |\text{epoch } m|\sqrt{K\Xi(n_m)} \leq \widetilde{O}\left(K^{5/6}T^{2/3} + \sqrt{K}\,\overline{\beta}T\right).$$

For completeness, its length at the physical horizon is random. Extend the retained sequence to T virtual observations and stop the algorithm at the physical horizon. Pseudo-regret is nonnegative, so the stopped sum is bounded by the extended sum. This justifies using the source's deterministic-horizon guarantee.

The initial stage costs at most $2G_{\max}T_1$. Exchangeability bounds discarded-round regret by $4rG_{\max}T/(n_{\mathrm{box}}+1)$. All simultaneous failure events have total probability $O(T^{-1})$ when each arm-epoch regression confidence is chosen of order $(KT^2)^{-1}$. Their expected cost is bounded. Adding these terms and balancing proves the result. □

Remark C.28 (Dependence on K and additional assumptions)**.** *The joint regression bound contains a factor $K^{2/3}$. The reduction takes its square root and multiplies by $\sqrt{K}$, giving $K^{5/6}$. The corresponding explicit factors are $K^{1/3}$ in the lower bound and K in the leading explore-then-commit upper bound. These comparisons hold with the other parameters fixed. They do not establish the optimal dependence on K.*

The improved upper bound uses the high-probability subspace guarantee of Assumption 11 and Assumption 12. The mean-square guarantee does not imply Assumption 11 for the same *estimator. However, Section C.6 shows that a median-of-means version of that estimator satisfies it, and that Assumption 12 can be dropped at the price of replacing $K^{5/6}$ by K.*

C.6 Removing the two additional assumptions

We now show that the horizon exponent $2/3$ is attained under exactly the assumptions already used by `EM-MIB` in Theorem C.20. Two changes are needed. First, the mean-square subspace oracle is converted into a high-probability one by a median-of-means selection. Second, the arm-conditional design bound is replaced by an explicit calculation: an arm that is rarely chosen has a more concentrated design, but it also carries a smaller weight in the joint error.

A high-probability subspace estimate from the mean-square oracle.

Lemma C.29 (Median-of-means subspace selection)**.** *Suppose Assumption 7 holds. Fix $\delta \in (0,1)$ and put $k = \lceil 8\log(1/\delta) \rceil$. Let $n_{\mathrm{sub}} \geq k \max\{n_{\min}, 2\}$ observations from the uniform exploration distribution be split into k disjoint blocks of deterministic size $b = \lfloor n_{\mathrm{sub}}/k \rfloor$; any remaining observations are discarded. Let $\widehat{P}_j$ be the projector onto the row space returned by the subspace estimator on block j. For each $j \in [k]$ define*

$$\rho_j := \min\left\{\rho \geq 0 : \ \left|\{i \in [k] : \|\widehat{P}_i - \widehat{P}_j\|_{\mathrm{op}} \leq \rho\}\right| > k/2\right\},$$

choose $j^* \in \arg\min_j \rho_j$, and let $\widehat{L}_{\mathrm{raw}}$ be an orthonormal row basis of the range of $\widehat{P}_{j^*}$; the canonical lock is applied to this orthonormalized basis. If the subspace estimator is randomized, it uses independent internal randomness on different blocks. Neither convention changes the selected projector. Then, with probability at least $1-\delta$,

$$\|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}} \leq D_{\mathrm{hp}} \sqrt{\frac{\log(2/\delta)}{n_{\mathrm{sub}}}}, \qquad D_{\mathrm{hp}} := 34\sqrt{\frac{C_2 dr}{\lambda_r}}.$$

In particular, the selected estimator satisfies Assumption 11.

Proof. Put $M := C_2 dr/(b\lambda_r)$ and $\delta_j := \|\widehat{P}_j - P_L\|_{\mathrm{op}}$. Since $b \geq n_{\min}$ is deterministic, Assumption 7 and Markov's inequality give $\mathbb{P}(\delta_j > 2\sqrt{M}) = \mathbb{P}(\delta_j^2 > 4M) \leq 1/4$. Call block j good if $\delta_j \leq 2\sqrt{M}$. The blocks are independent, so the number of bad blocks is a sum of independent indicators with means at most $1/4$. Hoeffding's inequality gives $\mathbb{P}(\text{at least } k/2 \text{ bad blocks}) \leq e^{-2k(1/4)^2} = e^{-k/8} \leq \delta$.

Work on the complementary event, so the set $\mathcal{G}$ of good blocks has more than $k/2$ elements. For $j \in \mathcal{G}$, every $i \in \mathcal{G}$ satisfies $\|\widehat{P}_i - \widehat{P}_j\|_{\mathrm{op}} \leq \delta_i + \delta_j \leq 4\sqrt{M}$. Hence $\rho_j \leq 4\sqrt{M}$, and therefore $\rho_{j^*} \leq 4\sqrt{M}$. More than $k/2$ blocks lie within ρ_{j^*} of $\widehat{P}_{j^*}$, and more than $k/2$ blocks are good. Some good block i belongs to both sets, so

$$\delta_{j^*} \leq \|\widehat{P}_{j^*} - \widehat{P}_i\|_{\mathrm{op}} + \delta_i \leq 4\sqrt{M} + 2\sqrt{M} = 6\sqrt{M}.$$

Finally, $n_{\mathrm{sub}}/k \geq 2$ gives $b \geq n_{\mathrm{sub}}/(2k)$, and $k \leq 8\log(1/\delta) + 1 \leq 8\log(e/\delta) \leq 16\log(2/\delta)$. Thus $M \leq 32\, C_2 dr \log(2/\delta)/(\lambda_r n_{\mathrm{sub}})$ and $6\sqrt{32} \leq 34$. □

The selection rule needs no knowledge of C_2, λ_r or M. It only compares the k candidate projectors with one another. The price is a logarithmic factor in the block size.

A joint regression guarantee without an arm-conditional design bound. Assumption 12 was used only to bound the design density of each arm's data. Without it, the density of arm a's design can be as large as a constant multiple of $1/q_a$, where q_a is the probability of selecting arm a. By Lemma C.12, this enlarges the entropy constant by the factor $q_a^{-1/2}$. The next proposition shows that this loss is exactly offset in the joint error, where arm a carries weight q_a.

Proposition C.30 (Joint regression guarantee under the context density alone)**.** *Fix a chart in the good-chart event $\mathcal{L}$ and a box of retained mass at least one half. Let n retained observations (X_i, A_i, Y_i) be i.i.d., with A_i drawn from a fixed action kernel $p(\cdot \mid X_i)$. Fit each arm by Definition C.7 on its own observations, with tolerance at most $(n_a \vee 1)^{-2/3}$, and use zero for an arm with no observations. Then, with probability at least $1-\delta$,*

$$\mathbb{E}_{X|\text{retained},\ A\sim p(\cdot|X)}\left[\{\widehat{f}_A(U) - g_A(LX)\}^2\right] \leq C\left\{K\, n^{-2/3}\left(\log\frac{CenK}{\delta}\right)^{q_r} + \beta^2\right\},$$

where $\beta = L_g C_X \epsilon_W$ and $U = \widehat{S}(\widehat{L}X)$. The constant C depends only on M_0', $G_{\max}$, $V_{\max}$, σ and r. It does not depend on K or on the action kernel. No assumption on the arm-conditional designs is made.

Proof. Put $u = \log(CenK/\delta)$ and $q_a = \mathbb{P}(A = a)$. Arms with $q_a = 0$ contribute zero to the joint risk and are omitted below. For $q_a > 0$, let Q_a be the law of U given $A = a$. By Lemma B.9, the retained input U has a density $f_U \le 2M_0'$ on $[0,1]^r$. Hence Q_a has density $f_U(u)\,\mathbb{P}(A = a \mid U = u)/q_a \le 2M_0'/q_a$. By Lemma C.11 and Lemma C.12, the entropy bound under $L_2(Q_a)$ holds with constants

$$A_a = A_\lambda \sqrt{2M_0'/q_a}, \qquad D_a = D_\lambda \sqrt{2M_0'/q_a},$$

where A_λ, D_λ are the Lebesgue-measure constants of Lemma C.11.

Conditional on its count n_a, the observations of arm a are i.i.d. with input law Q_a. Every retained X satisfies $|g_a(LX) - f_{0,a}(U)| \le \beta$ on $\mathcal{L}$, whatever the arm selected. Hence the conditional mean of Y given $(U, A = a)$ is within β of $f_{0,a} \in \mathcal{F}_V$. The residual noise is conditionally sub-Gaussian with proxy at most $\sigma^2 + 4G_{\max}^2$, as in Section C.2.

Call arm a frequent if $nq_a \ge 8u$. For a frequent arm, the Chernoff bound gives $n_a \ge nq_a/2$ except with probability $e^{-nq_a/8} \le \delta/(2K)$. Also $q_a \ge 8u/n \ge n^{-1}$, so $\log(1 + A_a)$ and $\log D_a$ are at most a constant multiple of $\log(en)$. Lemma C.14, with confidence $\delta/(2K)$, therefore gives

$$q_a \|\widehat{f}_a - f_{0,a}\|^2_{L_2(Q_a)} \le C q_a \left\{ (2M_0')^{1/3} A_\lambda^{2/3} q_a^{-1/3} \Big(\frac{nq_a}{2}\Big)^{-2/3} u^{q_r} + \beta^2 + \Big(\frac{nq_a}{2}\Big)^{-2/3} \right\} \le C \left\{ n^{-2/3} u^{q_r} + q_a \beta^2 \right\}.$$

The factor q_a cancels exactly: $q_a \cdot q_a^{-1/3} \cdot q_a^{-2/3} = 1$, and $q_a^{1/3} \le 1$ in the tolerance term. Summing over the at most K frequent arms gives $C\{Kn^{-2/3}u^{q_r} + \beta^2\}$.

For an infrequent arm, both $\widehat{f}_a$ and g_a are bounded by B_0, so its weighted squared error is at most $4B_0^2 q_a \le 32 B_0^2 u/n$. The total over at most K such arms is $CKu/n \le CKn^{-2/3}u^{q_r}$, because $q_r \ge 2$. This also covers arms with no observations.

Finally, for every arm, $\{\widehat{f}_a(U) - g_a(LX)\}^2 \le 2\{\widehat{f}_a(U) - f_{0,a}(U)\}^2 + 2\beta^2$. Averaging over $A \sim p(\cdot \mid X)$ and X gives the claim. The failure probabilities of the count events and of the regression events add to at most δ. □

Remark C.31 (Why the factor is K rather than $K^{2/3}$)**.** *Under Assumption 12 the design constant does not depend on q_a. The weighted sum is then $n^{-2/3}\sum_a q_a^{1/3} \le K^{2/3} n^{-2/3}$, as in Proposition C.26. Without it, each frequent arm contributes the same amount $n^{-2/3}$, whatever its selection probability. This costs a factor $K^{1/3}$ in the joint error and $K^{1/6}$ in regret. It does not change the power of n.*

The algorithm and its regret. We run Algorithm 7 with two changes. In step 2, the chart is estimated by the median-of-means selection of Lemma C.29 with $\delta = T^{-2}$, using the same subspace estimator as `EM-MIB`. The error envelope is replaced by

$$\Xi_{\text{free}}(n) = \min\left\{ 4B_0^2,\ C_* \left[K\, n^{-2/3} \{\log(CeKT^3)\}^{q_r} + \overline{\beta}^{\,2} \right] \right\}, \tag{43}$$

and $\gamma_m = \frac{1}{2}\sqrt{K/\Xi_{\text{free}}(n_m)}$. We call the resulting procedure `EM-MIB+`.

Theorem C.32 (Improved regret under the `EM-MIB` assumptions)**.** *Suppose the assumptions of Theorem C.20 hold, namely Assumptions 1, 2, 4, 7, 10, 8 with entire monotonicity, 9, and the stationary sampling condition of Section C. Assumptions 11 and 12 are* not *assumed. Let*

$$D_* = 2\kappa_0^2 \sqrt{r}\, L_g C_X D_{\text{hp}}, \qquad \overline{\beta} = D_* \sqrt{\frac{\log(2T^2)}{n_{\text{sub}}}},$$

with D_{hp} from Lemma C.29, $n_{\text{sub}} = \lfloor T_1/2 \rfloor$ and $n_{\text{box}} = T_1 - n_{\text{sub}}$. Choose T_1 large enough that $n_{\text{sub}} \ge \lceil 16 \log T \rceil \max\{n_{\min}, 2\}$, $D_{\text{hp}}\sqrt{\log(2T^2)/n_{\text{sub}}} \le \tau$, and $n_{\text{box}} \ge 4r \log(2rT^2)$. For a sufficiently large constant C_* in (43)*, `EM-MIB+` satisfies*

$$\mathbb{E}[R_T] \le \widetilde{O}\left(G_{\max} T_1 + K\, T^{2/3} + \sqrt{K}\, D_* T\, T_1^{-1/2} + G_{\max} T/T_1 + 1 \right).$$

For fixed positive model parameters and sufficiently large T, suppose the balanced choice $T_1 \asymp (\sqrt{K}\, D_* T / G_{\max})^{2/3}$, up to logarithmic factors, satisfies all the preceding minimum sample requirements and $T_1 \le T$. Then

$$\mathbb{E}[R_T] \le \widetilde{O}\left(K\, T^{2/3} + G_{\max}^{1/3} K^{1/3} D_*^{2/3} T^{2/3}\right).$$

Proof. The proof of Theorem C.27 uses the two additional assumptions in exactly two places. We replace each of them.

Chart. Lemma C.29 with $\delta = T^{-2}$ gives $\|P_{\widehat{L}_{\mathrm{raw}}} - P_L\|_{\mathrm{op}} \le D_{\mathrm{hp}} \sqrt{\log(2T^2)/n_{\mathrm{sub}}}$, except with probability T^{-2}. The condition on T_1 makes this at most τ, so the good-chart event $\mathcal{L}$ holds. Lemma B.6 then gives $\beta = L_g C_X \epsilon_W \le \overline{\beta}$. This is the conclusion that Assumption 11 supplied.

Regression oracle. On $\mathcal{L}$ and on the event that the box retains mass at least one half, the retained contexts are i.i.d., and the fits are fixed during each epoch. Proposition C.30, with confidence of order $(KT^2)^{-1}$ per epoch, shows that Ξ_{free} is a valid error envelope under the action kernel of the preceding epoch. This is the conclusion that Proposition C.26 supplied under Assumption 12.

For the uniform first epoch, take $\widehat{f}_1 \equiv 0$ and $\gamma_1 = \sqrt{K}/(4B_0)$; inverse-gap weighting with zero gaps is uniform sampling. Since $\Xi_{\mathrm{free}}(n) \le 4B_0^2$ is nonincreasing, the subsequent learning rates are nondecreasing and are at least γ_1. The initial mean regret is at most $2G_{\max} \le K/\gamma_1$. This supplies the initialisation needed by the induction in Simchi-Levi & Xu (2022).

The remainder of the proof of Theorem C.27 applies verbatim. The only change is in the epoch sum. With epochs of length 2^m on the retained sequence,

$$\begin{aligned} C \sum_m 2^m \sqrt{K\, \Xi_{\mathrm{free}}(2^{m-1})} &\le C\sqrt{K} \sum_{m \le \log_2 T + 1} 2^m \left\{ \sqrt{K}\, 2^{-(m-1)/3} \{\log(CeKT^3)\}^{q_r/2} + \overline{\beta} \right\} \\ &\le \widetilde{O}\left(K\, T^{2/3} + \sqrt{K}\, \overline{\beta} T\right). \end{aligned}$$

Substituting $\overline{\beta}$ gives the term $\sqrt{K}\, D_* T T_1^{-1/2}$ up to a logarithmic factor. The initial stage, the discarded out-of-box rounds and the failure events are bounded exactly as before. Balancing $G_{\max} T_1$ with $\sqrt{K}\, D_* T T_1^{-1/2}$ proves the second display. □

Remark C.33 (Calibration constants)**.** *The algorithm uses known upper bounds for the constants C_* and D_* in its error envelope. Its initial sample size must also satisfy the stated oracle and chart requirements. For a uniform regret guarantee over a class, these bounds must hold uniformly over that class. We do not claim adaptation to unknown calibration constants. If the constants are replaced by slowly growing multiples of $\log T$, the envelope becomes valid only eventually, for fixed parameters, and the regret bound gains additional logarithmic factors.*

Corollary C.34 (Minimax horizon rate for the EM class)**.** *Fix $K \ge 2$, r, d, and compatible positive problem parameters as in Theorem C.23.*

(i) Known chart. *If L is known, no subspace estimation is needed and $\beta = 0$. Use an independent box block of $n_{\mathrm{box}} \asymp T^{1/2}$ observations. For sufficiently large T, this satisfies the box requirement. The box exploration and discarded-context costs are both $O_{\mathrm{par}}(T^{1/2})$ and are absorbed by the stated bound. Every policy then has worst-case regret at least $cK^{1/3}T^{2/3}$ by Theorem C.23, and* `EM-MIB+` *with the chart step omitted has regret $\widetilde{O}(K\, T^{2/3})$. Thus the minimax regret is $\widetilde{\Theta}(T^{2/3})$ in the horizon, with explicit dependence on K between $K^{1/3}$ and K.*

(ii) Unknown chart. *Let Θ be a class satisfying the assumptions of Theorem C.32. Use one fixed subspace-estimation procedure. Assume that the model bounds, the oracle constant $C_2 dr/\lambda_r$, and the minimum sample size $n_{\min}$ have common bounds over Θ. Calibrate the algorithm using these common bounds. Then, for sufficiently large T uniformly over Θ,*

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[R_T(\texttt{EM-MIB+})] \le \widetilde{O}_{\mathrm{par}}(KT^{2/3}).$$

If Θ contains the lower-bound family of Theorem C.23, then every policy has worst-case regret at least $cK^{1/3}T^{2/3}$. For fixed $K \geq 2$, the minimax horizon rate over this class is therefore $\widetilde{\Theta}(T^{2/3})$.

The lower-bound instances satisfy linearity and population SIR identifiability with $\lambda_r \geq \lambda_*/2$ (Theorem C.23). Whether they belong to Θ in part (ii) depends on the chosen estimator meeting Assumption 7 on them, with constants common to the whole class. Population identifiability alone does not establish this guarantee, and the median-of-means step only amplifies an existing mean-square guarantee. An unconditional minimax statement for the whole unknown-chart EM class is therefore not claimed. The dependence on K is not matched: the lower bound has $K^{1/3}$ and the upper bound has K.

D COMPUTING THE ENTIRELY MONOTONE ESTIMATOR

The statistical results allow an empirical loss tolerance of order $n^{-2/3}$. We show how to attain this tolerance over the full class $\mathcal{F}_V$ of Definition C.7. First, we reduce the optimisation to a finite convex problem. We then give a Frank–Wolfe algorithm with an explicit iteration bound. The construction applies to the retained, rescaled observations after the chart and box have been frozen. It also applies when the chart is known. All statements in this section are deterministic conditional on the data.

A finite representation. Consider one arm with observations $(U_i, Y_i)_{i=1}^n$, where $n \geq 1$ and $U_i \in [0,1]^r$. Recall that $B_0 = G_{\max} + V_{\max}$ bounds every function in $\mathcal{F}_V$. Write $\mathcal{Q}_n(f) = n^{-1}\sum_{i=1}^n (Y_i - f(U_i))^2$ for the empirical loss. For each coordinate j, let $\mathcal{G}_j = \{0, U_{1j}, \ldots, U_{nj}\}$, with duplicates removed. Define $\mathcal{A} = (\mathcal{G}_1 \times \cdots \times \mathcal{G}_r) \setminus \{\mathbf{0}\}$ and $h_a(u) = \mathbf{1}\{a_j \leq u_j \text{ for all } j\}$ for $a \in \mathcal{A}$. The intercept is treated separately. Excluding $\mathbf{0}$ ensures that each nonconstant component vanishes at the origin.

Proposition D.1 (Finite constrained least squares)**.** *The minimum empirical loss over $\mathcal{F}_V$ equals the value of the finite convex quadratic programme*

$$\min_{\substack{|b|\leq G_{\max},\ \theta_a \geq 0 \\ \sum_{a\in\mathcal{A}} \theta_a \leq V_{\max}}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - b - \sum_{a\in\mathcal{A}} \theta_a h_a(U_i) \right)^2 . \tag{44}$$

A minimiser exists. Its coefficients define a function in $\mathcal{F}_V$. The dictionary has at most $(n+1)^r - 1$ atoms.

Proof. By Lemma C.5, every member of $\mathcal{F}_V$ has the form $b + \int h_a \, d\nu(a)$, where $|b| \leq G_{\max}$, ν is nonnegative, $\nu(\{\mathbf{0}\}) = 0$, and $\nu([0,1]^r) \leq V_{\max}$. If an atom vanishes at every observation, discard its mass. For each remaining atom, leave every zero coordinate unchanged. Round each positive coordinate a_j up to the smallest observed coordinate U_{ij} that is at least a_j. Such a coordinate exists because the atom is nonzero at some observation. This operation preserves every comparison $a_j \leq U_{ij}$. It therefore preserves the atom's values on the whole sample. The rounded atom belongs to $\mathcal{A}$. Collecting its mass gives feasible coefficients in (44), with the same fitted values and no larger total mass. Conversely, every feasible coefficient vector defines a function in $\mathcal{F}_V$. The feasible coefficient set is compact, so the minimum is attained. This is a constrained version of the finite representation used in Fang et al. (2021). □

The Cartesian grid is essential for this reduction. For arbitrary multivariate designs, restricting atom locations to $U_1, \ldots, U_n$ need not preserve the minimum. No bound on the number of nonzero coefficients is imposed in (44).

Approximate optimisation. Adjoin the zero function to the dictionary and write $\mathcal{H} = \{0\} \cup \{h_a : a \in \mathcal{A}\}$. Algorithm 8 implements Frank–Wolfe on the feasible set of (44). Its linear optimisation step selects an intercept and an atom separately. The zero function allows the variation constraint to be slack. Use a fixed ordering to break ties and set $\operatorname{sign}(0) = 0$.

Algorithm 8 Approximate variation-constrained EM least squares

Require: Data $(U_i, Y_i)_{i=1}^n$, bounds $G_{\max}, V_{\max}$, and iteration count $k \geq 1$.
1: Form $\mathcal{G}_1, \ldots, \mathcal{G}_r$ and $\mathcal{H}$ as above.
2: Set $f^{(0)} = 0$.
3: **for** $s = 1, \ldots, k$ **do**
4: $R_i \leftarrow Y_i - f^{(s-1)}(U_i)$ for $i = 1, \ldots, n$.
5: $b_s \leftarrow G_{\max} \operatorname{sign}(\sum_{i=1}^n R_i)$.
6: $h_s \in \arg\max_{h \in \mathcal{H}} \sum_{i=1}^n R_i h(U_i)$.
7: $v_s \leftarrow b_s + V_{\max} h_s$ and $\gamma_s \leftarrow 2/(s+1)$.
8: $f^{(s)} \leftarrow (1-\gamma_s) f^{(s-1)} + \gamma_s v_s$.
9: **end for**
10: **return** $f^{(k)}$.

Proposition D.2 (Optimisation tolerance)**.** *For every sample and every $k \geq 1$, Algorithm 8 returns a function in $\mathcal{F}_V$ satisfying*

$$\mathcal{Q}_n(f^{(k)}) - \min_{f \in \mathcal{F}_V} \mathcal{Q}_n(f) \leq \frac{16 B_0^2}{k+2}. \tag{45}$$

Its representation uses an intercept and at most k atoms. In particular, $k_n = \max\{1, \lceil 16 B_0^2 n^{2/3} \rceil\}$ iterations give tolerance at most $n^{-2/3}$.

Proof. Each v_s is feasible, so all iterates are feasible by convexity. The selected v_s minimises the linearised loss at $f^{(s-1)}$. Indeed, the intercept maximises $b \sum_i R_i$ over its interval. The nonnegative coefficients maximise a linear function over a simplex, so all available mass can be placed on a best atom, or no mass is used.

Let $e_s = \mathcal{Q}_n(f^{(s)}) - \min_{f \in \mathcal{F}_V} \mathcal{Q}_n(f)$. Expanding the squared loss, using the linear minimisation property, and then convexity gives

$$e_s \leq (1-\gamma_s) e_{s-1} + 4 B_0^2 \gamma_s^2.$$

The last term follows from $|v_s(U_i) - f^{(s-1)}(U_i)| \leq 2B_0$. Since $\gamma_1 = 1$, we have $e_1 \leq 4B_0^2$. Induction with $\gamma_s = 2/(s+1)$ yields $e_s \leq 16 B_0^2/(s+2)$. Each iteration adds at most one atom. This is the standard Frank–Wolfe argument specialised to our loss and intercept constraint Jaggi (2013). □

Arithmetic cost. Let $q = \prod_{j=1}^r |\mathcal{G}_j| \leq (n+1)^r$. The atom search can be performed without storing an $n \times q$ matrix. Sort the coordinate grids and store each observation's coordinate ranks. At each iteration, place the residuals in an array indexed by the grid, adding residuals when observations coincide. Successive suffix sums along the r axes give $\sum_{i: a \leq U_i} R_i$ at every grid point a. Exclude the origin and compare the largest score with the score zero of the zero function. This computes the required maximiser exactly, even when residuals have both signs.

Sorting costs $O(rn \log(n+1))$ operations. Each iteration costs $O(r(q+n))$ operations, including the update of the fitted values. Thus k iterations cost

$$O\big(rn \log(n+1) + kr(q+n)\big).$$

The required storage is $O(q + rn + rk)$. For example, the final sparse representation can be assembled from the selected atoms after the iterations, without repeatedly updating all stored coefficients. For fixed r and fixed model bounds, choosing $k = k_n$ gives $O_r(n \log(n+1) + n^{r+2/3})$ operations per fit. Evaluation at a new point costs $O(1 + rk_n)$ operations. These are arithmetic operation bounds, not bit complexity bounds. They establish polynomial computation for fixed r. They do not establish a runtime that is polynomial jointly in n and r.

Use in the bandit algorithms. Run Algorithm 8 with $k = k_{n_a}$ for each arm having $n_a \geq 1$ retained observations. Use the zero function when $n_a = 0$. In the epoch algorithm, apply this rule separately to each fit using only the preceding epoch's retained observations. Fixing all tie rules makes the

construction measurable. Proposition D.2 supplies the required tolerance for every realised sample, so no additional failure event is needed.

Lemma C.14 already includes the additive optimisation tolerance. Its contribution is $O(n_a^{-2/3})$ to squared prediction error. Consequently, this implementation preserves the rates in Theorems C.15, C.20, and C.32, under their respective assumptions. The same observation applies to the joint regression guarantee in Proposition C.30. No separate sparse approximation theorem is required. The chart estimation and sample-box rules are unchanged. In particular, contexts outside the frozen box are discarded rather than clipped. The computational bounds above concern the regression step. The cost of the chosen subspace estimator must be added separately.

E Detailed Experiments

E.1 Experimental protocol and interpretation

The experiments are designed to answer separate structural questions rather than to produce a single aggregate ranking. The coordinatewise-monotone experiments test the dependence on the effective dimension r, the benefit of exploiting monotonicity relative to unstructured baselines, and the effect of estimating an unknown chart. The entirely-monotone experiments compare the two algorithms obtained from the stronger shape restriction, including the explore–then–commit algorithm `EM-MIB` and the epoch-based algorithm `EM-MIB+`. We also deliberately evaluate the EM procedures on functions that are coordinatewise monotone but not entirely monotone in order to make the distinction between the two shape classes visible empirically.

Unless stated otherwise, rewards are corrupted by independent Gaussian noise with standard deviation 0.05. Every curve reports mean cumulative regret across the indicated independent seeds, with shaded bands showing a 95% normal confidence interval for the mean. The rate annotations are always computed from cumulative regret itself. In particular, if $\bar{R}_T$ denotes the mean regret at horizon T, the displayed log–log exponent is the least-squares slope in $\log \bar{R}_T = a + \widehat{\alpha} \log T$. When a figure says *tail log–log slope*, the same regression is applied only to the three largest tested horizons. We do not fit slopes to regret-per-round or to a normalized regret quantity.

For the synthetic comparisons we use paired context/noise streams across algorithms whenever the implementation permits it. The practical constant used by `EM-MIB+` in the $K = 10$ and real-context experiments is fixed at $c = 5 \times 10^{-4}$ from a disjoint synthetic development sweep; the final seeds reported below are untouched test seeds. No further tuning is performed on the final test runs or on the real-context data.

E.2 CM-MIB: horizon scaling and baseline comparison

Hard-family scaling. The first experiment isolates the dependence on the intrinsic dimension r. We use the lower-bound-inspired hard family with $K = 3$ and 20 seeds. For $r = 1, 2$, horizons range from 5×10^3 to 2×10^5; for $r = 3$ we extend the experiment to 1.6×10^6. Figure 2 overlays the observed `CM-MIB` curve with the corresponding theoretical power law. The empirical slopes are 0.666, 0.758, and 0.792 for $r = 1, 2, 3$, respectively, compared with $2/3$, $3/4$, and $4/5$.

Comparison with unstructured baselines. We next use fixed synthetic CM instances and retain two unstructured baselines, Global-UCB and Uniform.

In all three dimensions, `CM-MIB` has substantially lower regret at the largest horizon: 497 versus 5563 and 11104 for $r = 1$, 1179 versus 16195 and 17827 for $r = 2$, and 2391 versus 10761 and 17775 for $r = 3$. The baseline log–log slopes are essentially one, whereas the observed `CM-MIB` slopes are 0.496, 0.600, and 0.701 on these fixed instances. These fixed-instance exponents are not expected to equal the worst-case exponents of the hard family; their role is to show the practical benefit of the structural elimination rule on benign instances.

E.3 Synthetic function classes and cross-class behavior

A central purpose of the experiments is to make the difference between coordinatewise monotonicity and entire monotonicity visible. Figure 4 shows the optimal-arm regions for five fixed $r = 2$, $K = 3$

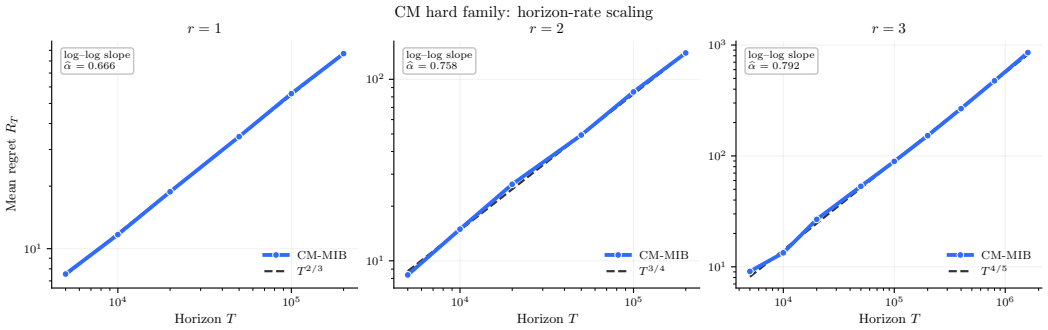


Figure 2: **CM hard-family scaling.** Each panel plots mean regret against the horizon on log–log axes. The dashed line is the theoretical reference power $T^{(r+1)/(r+2)}$. The fitted exponents closely track the predicted horizon dependence.

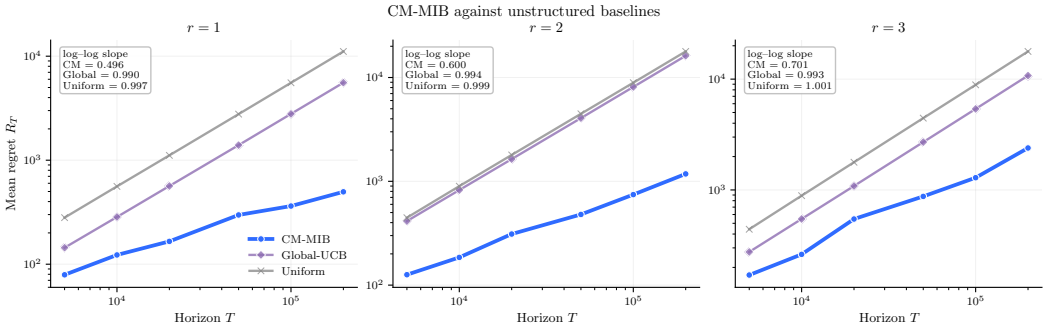


Figure 3: **CM-MIB versus retained unstructured baselines.** Mean regret on fixed synthetic instances for $r = 1, 2, 3$. `CM-MIB` grows markedly more slowly than Global-UCB and Uniform on these instances.

synthetic families. The first three belong to the EM class. The CDF benchmark is a fixed nonnegative mixture of 12 product-ramp CDF components per arm. The complements family uses

$$g_1(x, y) = 0.08 + 0.55x + 0.08y + 0.18xy,$$
$$g_2(x, y) = 0.08 + 0.08x + 0.55y + 0.18xy,$$
$$g_3(x, y) = 0.26 + 0.18x + 0.18y + 0.15xy,$$

so all mixed interactions are nonnegative. The context-gap family is monotone and additive apart from a monotone step in the first coordinate, so its mixed differences are zero.

The final two families are coordinatewise monotone but not EM. In the max-substitutes family the first arm is $0.05 + 0.75 \max\{x, y\}$, whose mixed finite difference over the unit square is -0.75. In the negative-interaction family the first two arms contain the term $-0.05xy$, providing a direct negative mixed interaction while preserving coordinatewise monotonicity.

Entirely monotone classes. Figure 5 compares `CM-MIB`, `EM-MIB`, and `EM-MIB+` on the three EM families. At the largest tested horizon, `EM-MIB` has the smallest regret on all three: 196 on the CDF mixture, 153 on complements, and 1178 on context-gap. Thus the epoch-based refinement of `EM-MIB+` should not be interpreted as a pointwise finite-horizon dominance statement. On these relatively easy fixed instances, the lower exploration burden of explore–then–commit is advantageous. The context-gap CM curve is non-monotone because the CM grid changes discretely with the horizon; we report the raw data rather than smoothing this transition.

Table 1: Predicted and empirical horizon exponents. The CM fits use all displayed hard-family horizons. For the $K = 10$ EM experiment we report the last-three-horizon slope because the experiment is designed to probe the asymptotic separation between the two EM algorithms after their initial transients.

Method / setting	Theory	Empirical	Fit window
CM-MIB, $r = 1$	0.667	**0.666**	all horizons
CM-MIB, $r = 2$	0.750	**0.758**	all horizons
CM-MIB, $r = 3$	0.800	**0.792**	all horizons
EM-MIB, $K = 10$	0.750	**0.743**	last three
EM-MIB+, $K = 10$	0.667	**0.666**	last three

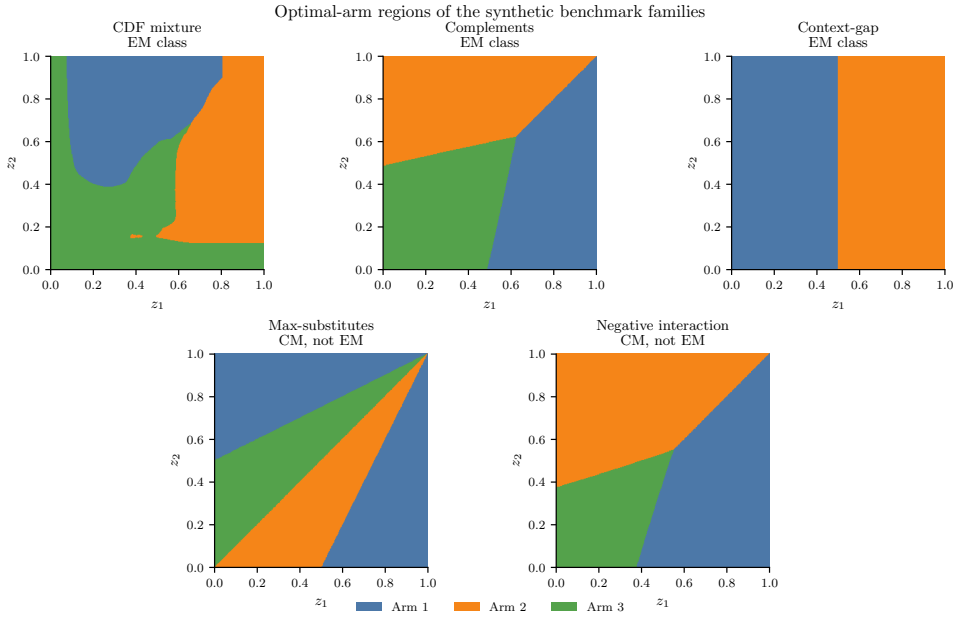


Figure 4: **Synthetic benchmark geometry.** Colors indicate the optimal arm on $[0, 1]^2$. The upper row contains entirely monotone examples; the lower row contains coordinatewise-monotone examples that violate entire monotonicity.

Coordinatewise monotone but non-EM classes. The misspecification experiment in Figure 6 gives two qualitatively different outcomes. On max-substitutes, `CM-MIB` is clearly best: at $T = 128{,}000$ its mean regret is 1205, compared with 3273 for `EM-MIB` and 2657 for `EM-MIB+`; the tail slopes are 0.656, 0.990, and 0.898. This is a clean example where the stronger EM assumption is materially wrong and the CM-specific method is preferable. On the milder negative-interaction family, however, `EM-MIB` has the smallest finite-horizon regret (326 versus 846 for `CM-MIB` and 1120 for `EM-MIB+` at $T = 128{,}000$), despite being used outside its assumed class. Class membership determines the validity of the guarantee, but it does not force a pointwise ordering on every finite instance.

As an additional robustness check, the coverage-substitutes family has mixed interaction -0.64 in its first arm. At $T = 128{,}000$ all three methods are close (1228, 1192, and 1193 for `CM-MIB`, `EM-MIB`, and `EM-MIB+`), while the respective tail slopes are 0.682, 0.910, and 0.781. We include the full curve in Figure 7 rather than selecting only the CM-favorable substitute example.

E.4 When does EM-MIB+ improve on EM-MIB?

The preceding fixed instances show that `EM-MIB+` need not have smaller finite-horizon regret. We therefore use a separate $K = 10$, $r = 3$ entirely monotone fixed-margin family to expose the regime

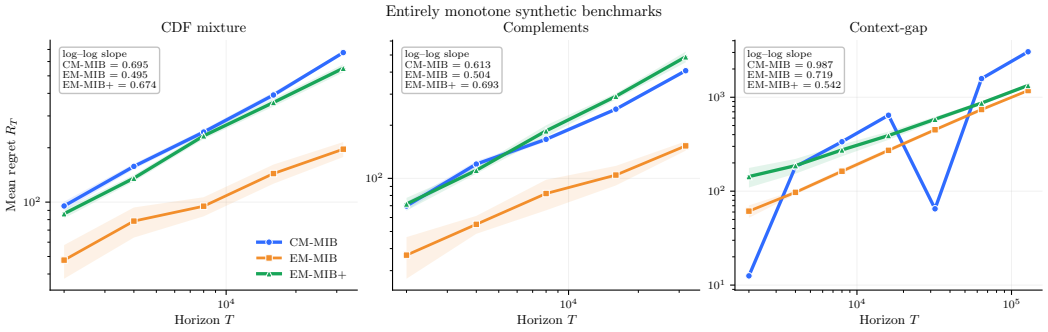


Figure 5: **Three entirely monotone synthetic classes.** Mean regret on log–log axes. EM−MIB is the best finite-horizon method on these three fixed instances.

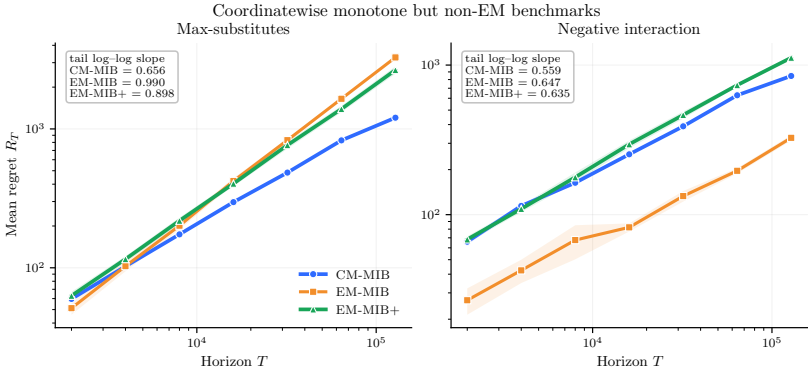


Figure 6: **CM-but-not-EM benchmarks.** Max-substitutes exposes the cost of EM misspecification, while the negative-interaction family shows that mild misspecification need not make EM−MIB empirically poor.

in which continued refitting and adaptive exploration become useful. The common component is

$$0.08 + 0.06(z_1 + z_2 + z_3) + 0.04 z_1 z_2 z_3,$$

and the arms add distinct nonnegative multiples of $(z_1 + 2z_2 + 3z_3)/6$ together with fixed offsets. Arm 1 is uniformly optimal with minimum gap 0.20. The practical EM−MIB+ exploration constant is chosen on disjoint development seeds and then frozen at 5×10^{-4}; the results below use 20 untouched test seeds.

Figure 8 gives the clean separation. EM−MIB+ has lower regret at every tested horizon. At $T = 1{,}048{,}576$, its mean regret is 8862 compared with 13643 for EM−MIB, a reduction of approximately 35%. More importantly for the theoretical comparison, the last-three-horizon slopes are 0.743 for EM−MIB and 0.666 for EM−MIB+, closely matching $3/4$ and $2/3$.

E.5 Sensitivity to the number of arms

We next vary $K \in \{2, 3, 5, 9, 17\}$ at fixed $T = 64{,}000$ on a nested $r = 2$ family, using 10 seeds. Figure 9 and Table 3 show that regret increases with K for every method, while CM−MIB is lowest on this particular family for every tested arm count. This experiment is a robustness check with respect to the number of arms; because T is fixed and the nested instance changes with K, we do *not* interpret the fitted dependence on K as an empirical verification of the exact theorem-level K exponent.

Table 2: Mean regret at the largest tested horizon for five representative synthetic families. Green marks the lowest regret in each row and yellow the second-lowest.

Family	$T_{\max}$	CM-MIB	EM-MIB	EM-MIB+
CDF mixture	32,000	672.6	**196.3**	551.0
Complements	32,000	407.4	**152.8**	488.3
Context-gap	128,000	3050.1	**1178.4**	1336.6
Max-substitutes	128,000	**1204.6**	3273.3	2656.8
Negative interaction	128,000	845.7	**326.3**	1119.7

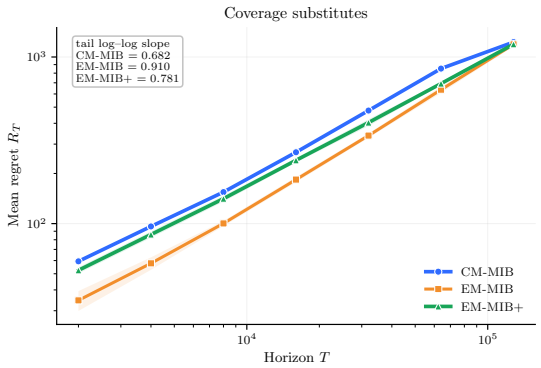


Figure 7: Coverage-substitutes robustness experiment.

E.6 Unknown-chart robustness for CM-MIB

To isolate the cost of learning the low-dimensional chart, we use $r = 2$ with ambient dimensions $d \in \{5, 10, 20\}$, 20 seeds, and horizons 20,000 through 200,000. The implementation uses an independent subspace-estimation block, canonical locking, and a separate box-estimation block. Figure 10 reports the relative regret overhead of the estimated chart compared with the oracle chart and the subspace error δ. The canonical lock succeeds in every recorded run. At $T = 200{,}000$, the estimated-chart overhead is about 0.8%, 2.2%, and 4.8% for $d = 5, 10, 20$, while the mean subspace errors are 0.060, 0.091, and 0.184. The experiment therefore shows a moderate dimension-dependent constant cost rather than a qualitative failure of the CM procedure.

E.7 Real-context semi-synthetic evaluation

Finally, we replace the uniform synthetic context distribution by real feature distributions derived from KDD99 and CoverType. The cached features are projected and scaled to two dimensions exactly as in the CM real-data pipeline. Rewards remain synthetic so that the structural class and the optimal arm are known; these experiments should therefore be read as *real-context semi-synthetic* evaluations, not as claims about naturally observed rewards. We use 20 seeds and horizons from 5,000 to 200,000.

Entirely monotone rewards on real contexts. For the in-class EM experiment we use the complements reward family. Figure 11 shows that `EM-MIB` has lower finite-horizon regret than `EM-MIB+` on both real context pools. At $T = 200{,}000$, the mean regrets are 925 versus 2338 on KDD99 and 284 versus 1582 on CoverType. The corresponding tail slopes are 0.765 versus 0.853 on KDD99 and 0.694 versus 0.821 on CoverType. Together with the $K = 10$ experiment, this reinforces that the improved worst-case exponent of `EM-MIB+` is not a universal finite-horizon dominance statement.

CM-but-not-EM rewards on the same real contexts. We then use the same negative-interaction reward family as the existing CM real-context experiment, which allows a direct three-method comparison with matched context/noise streams. On KDD99, `CM-MIB` is best by a wide margin at $T = 200{,}000$: 246 versus 769 for `EM-MIB` and 1566 for `EM-MIB+`. On CoverType, however, `EM-MIB` is best in raw regret (296 versus 1057 and 1255), despite the reward being outside the EM

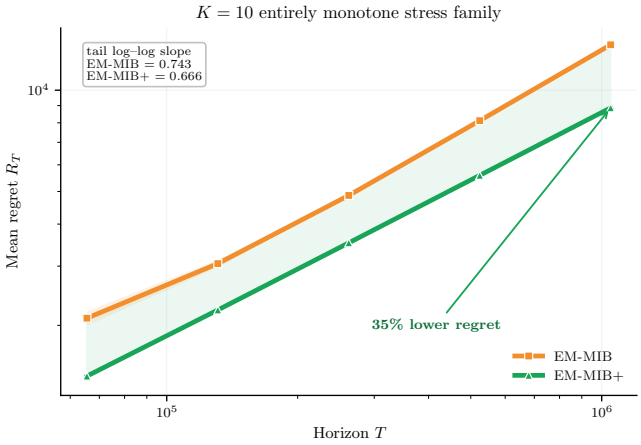


Figure 8: **A clean finite-horizon EM-MIB+ win.** The highlighted region is the regret gap between EM−MIB and EM−MIB+. The annotated exponents are slopes of $\log R_T$ against $\log T$ over the last three horizons.

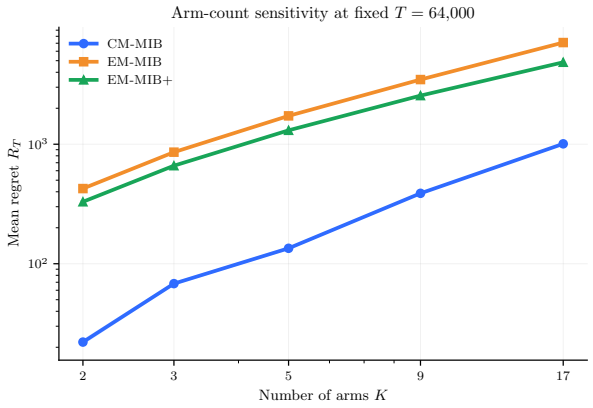


Figure 9: Arm-count sensitivity at fixed horizon.

class. The CM tail slope on CoverType is smaller (0.550 versus 0.710 and 0.778), but the observed horizon range does not justify claiming a crossover.

E.8 Parameter sensitivity

For completeness, Figure 13 varies the two numerical CM constants on the $r = 2$ fixed synthetic instance. The confidence scale has the larger effect: smaller values trigger elimination in a larger fraction of cells and reduce regret on this benchmark. The grid scale changes the constant more mildly over the tested range. This diagnostic is kept separate from the final comparisons; the final experiment constants are frozen before evaluating the reported test curves.

E.9 What the experiments support

The experiments provide four complementary pieces of evidence. First, the CM hard-family exponents track the predicted $T^{(r+1)/(r+2)}$ dependence very closely. Second, on ordinary fixed CM instances, structural elimination yields much lower regret than the retained unstructured baselines. Third, the cross-class experiments show that the distinction between CM and EM is operational: a strongly substitutable CM-only instance can make the EM procedures perform poorly, while mild misspecification can still leave EM−MIB competitive. Fourth, the $K = 10$ experiment exhibits the

Table 3: Mean regret at $T = 64{,}000$ as the number of arms varies. Green marks the lowest regret and yellow the second-lowest in each row.

K	CM-MIB	EM-MIB	EM-MIB+
2	**22.1**	425.5	329.9
3	**68.2**	858.1	661.1
5	**134.6**	1726.8	1307.0
9	**388.3**	3475.9	2551.2
17	**1007.7**	7091.2	4849.5

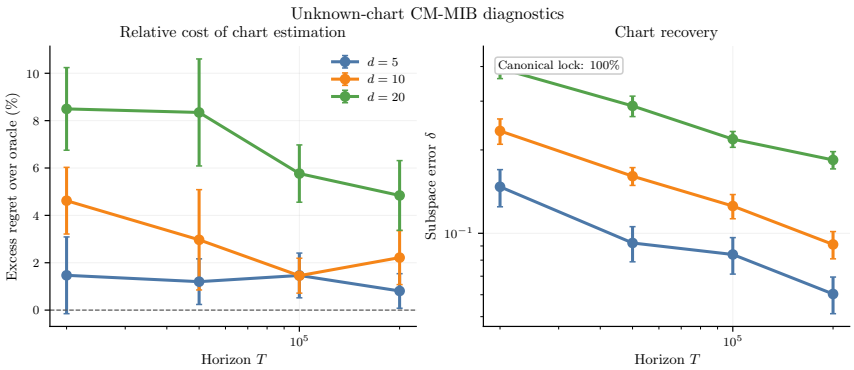


Figure 10: **Unknown-chart diagnostics.** Left: relative regret overhead of using the estimated chart instead of the oracle chart. Right: subspace recovery error.

regime targeted by `EM-MIB+`, with both lower finite-horizon regret and a tail exponent essentially equal to $2/3$.

These plots are empirical checks, not proofs of the minimax statements. In particular, the fixed-instance slopes can be better or worse than the worst-case upper-bound exponents, and the fixed-T arm-count experiment does not identify the theorem-level dependence on K.

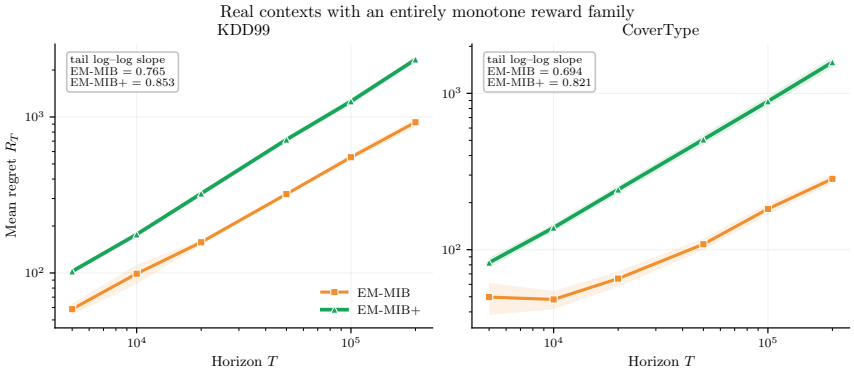


Figure 11: Real-context evaluation with an entirely monotone reward family.

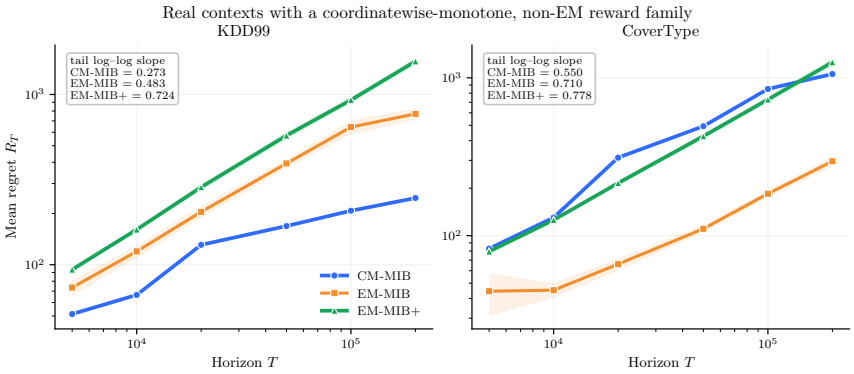


Figure 12: Real-context evaluation with a coordinatewise-monotone reward family that is not entirely monotone.

Table 4: Real-context mean regret at $T = 200{,}000$. Green denotes the lowest regret among the methods evaluated on that row; yellow denotes the second-lowest.

Dataset	Reward class	CM-MIB	EM-MIB	EM-MIB+
KDD99	EM class	–	**924.8**	2337.9
CoverType	EM class	–	**284.0**	1581.5
KDD99	CM, not EM	**246.4**	769.1	1565.6
CoverType	CM, not EM	1057.2	**296.0**	1255.2

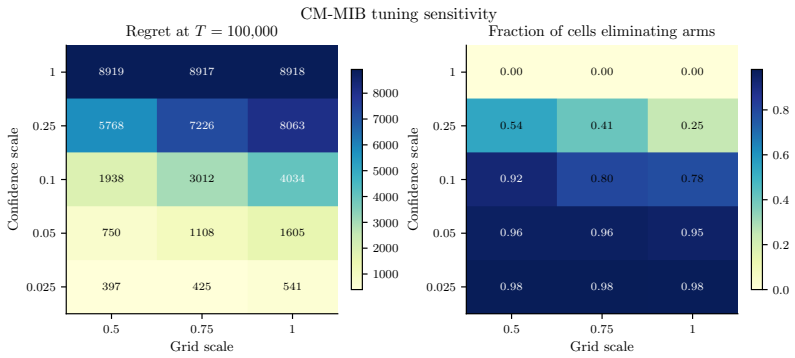


Figure 13: CM-MIB parameter sensitivity at $T = 100{,}000$. Only CM-MIB is shown; no additional baseline is used in this diagnostic.