LOW-RANK SINGLE-INDEX BANDITS WITH UNKNOWN LINKS: FROM MATRICES TO TENSORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank matrix and tensor bandits exploit structured interactions but typically assume a known reward link. Recent single-index bandit methods accommodate unknown links without directly exploiting matrix or tensor rank. We address this gap by studying stochastic matrix and tensor bandits with an unknown shared Lipschitz link and a low-rank index parameter under known regular candidate distributions and finite-variance noise. For monotone links, T-ESTOR combines robust, rank-adaptive Stein estimation with epoch-based greedy selection. Under exact selected-score access and a uniformly positive selected-design Stein signal, it achieves square-root regret with dimension dependence determined by the low-rank structure. For every admissible design, the monotone lower bound matches the rank, dimension, and horizon dependence up to logarithmic factors at large horizons, for fixed menu size and model/design constants. For nonmonotone links under a nonzero base-law Stein signal, T-BSTOR combines structured estimation with robust bin-based learning and attains the optimal $\widetilde{O}(T^{2/3})$ horizon rate for fixed dimensions, menu size, and model/design constants. Synthetic and CCLE-based experiments illustrate the benefits of structured estimation relative to vectorized and competing single-index baseline methods.

1 INTRODUCTION

Contextual bandits model decisions with feedback only for the selected action. In recommendation and experimental design, rewards often depend on interactions among users, items, and contexts, or compounds, cell lines, and doses. Matrix and tensor representations expose these interactions, while low rank reduces the number of unknown parameters. Bilinear and generalized low-rank matrix bandits exploit this structure (Jun et al., 2019; Lu et al., 2021; Kang et al., 2022); tensor bandits extend it to additional feature groups (Shi et al., 2023; Yi et al., 2024).

Low rank and knowledge of the reward response are different assumptions: interactions may admit a compact representation while their effect on the outcome is unknown. Single-index bandits model this uncertainty as $f(\langle X, \Theta_* \rangle)$, with both the index Θ_* and shared link f unknown. Kang et al. (2026) and Dey et al. (2026) develop estimation-based methods for monotone and nonmonotone rewards. Their vector formulations do not directly exploit matrix or tensor rank: a low-rank parameter can be dense, unlike the coordinate-sparse extension of Kang et al. (2026). Can regret depend on this interaction structure while retaining single-index horizon rates?

We study low-rank matrix and transformed-slice tensor bandits with an unknown Lipschitz link, known regular joint candidate laws, and finite-variance noise. Building on Stein and spectral estimation (Yang et al., 2017; Minsker, 2018; Kang et al., 2022; 2026), paired observations remove the unrestricted reward offset from moment bounds, while nuclear-norm regularization preserves dependence on matrix rank or total transformed-slice rank without knowing the ranks. We bound selected-score second moments uniformly over ranking directions, connecting these estimation guarantees to greedy decisions. The joint-law analysis permits dependence within candidates and among transformed slices, while preserving the stated rank dependence. T-ESTOR attains $\widetilde{O}(\sqrt{T})$ regret for monotone links under exact selected-score access and a uniformly positive selected-design Stein signal. For nonmonotone links, T-BSTOR controls robust scalar-bin estimates under adaptive sampling, attaining $\widetilde{O}(T^{2/3})$ regret under a nonzero base-law Stein signal and the stated distributional conditions. For each admissible design, the monotone lower bound matches rank, dimension, and

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.

horizon dependence up to logarithmic factors at sufficiently large horizons, for fixed menu size and model/design constants uniformly controlled across dimensions. The nonmonotone lower bound fixes the menu size and positive signal floor, establishing horizon optimality for fixed dimensions and model/design constants under its stated signal and noise conditions. Synthetic and CCLE-based experiments compare structured estimation with vectorized and competing single-index methods.

1.1 Related Work

Low-rank matrix bandits. Jun et al. (2019) estimate row and column spaces for bilinear rewards before regularized linear-bandit learning. Lu et al. (2021) extend this approach to matrix actions and known generalized-linear links. Kang et al. (2022) use Stein-based subspace estimation; their estimator already combines spectral robustification and a nuclear-norm-penalized quadratic objective, with a single-index extension discussed in Kang et al. (2024a). Jang et al. (2024) account for arm-set geometry through experimental design, and Kang et al. (2024b) handle heavy-tailed linear rewards by truncation and dynamic exploration.

Tensor bandits. Zhou et al. (2024) select entries or entity combinations in a low-rank reward tensor, rather than dense array actions with an unknown shared link. Shi et al. (2023) use Tucker structure and optimistic selection for linear rewards. Yi et al. (2024) use transformed t-product rank and nuclear-norm regularization with a specified generalized-linear link. Li et al. (2025) accommodate several tensor structures in a generalized-linear explore-then-commit framework.

Single-index bandits. Kang et al. (2026) study fresh menus and a shared unknown link: ESTOR uses Stein estimation and greedy epochs for monotone links under regular designs and finite-variance noise, with a sparse-vector extension; GSTOR uses kernel regression for nonmonotone links under Gaussian designs. Dey et al. (2026) combine index estimation and scalar bins for $T^{2/3}$ regret under a positive-signal condition and conditionally sub-Gaussian noise. These are the algorithmic foundations of T-ESTOR and T-BSTOR. The lower-bound construction of Dey et al. (2026) uses a growing menu and a decreasing signal floor; Appendix H.5 fixes both for each admissible design under its stated conditions. Ma & Cai (2025) study arm-specific rewards under monotonicity, smoothness, and margin conditions, with minimax regret and smoothness adaptation under self-similarity. Arya & Song (2026) study batched learning with a shared index and arm-specific links; Arya et al. (2026) develop kernel/Stein learning and inference under adaptive sampling with persistent arms and arm-specific indices and links. These models differ from our shared low-rank index and common link on fresh menus.

Structured single-index estimation. Plan & Vershynin (2016) analyze the generalized Lasso for offline nonlinear observations under Gaussian measurements. For known non-Gaussian designs, Stein methods cover sparse vectors and low-rank matrices (Yang et al., 2017), and fourth-order tensors via square matricization (Yang et al., 2018). Fan et al. (2023) study implicit regularization for sparse vectors and symmetric low-rank matrices. These offline guarantees require further analysis for bandits: selection changes the observation law, and estimation errors must be converted into sequential regret.

2 Preliminaries

2.1 Problem Setting

We study stochastic bandits with actions in $\mathbb{R}^{d_1\times\cdots\times d_h}$, where $h = 2$ for matrices and $h \geq 3$ for higher-order tensors. At round $t \in [T]$, $T \geq 1$, the learner observes a set $\mathcal{X}_t = \{X_{t,a} : a \in [K]\}$ of $K \geq 3$ candidates drawn iid from a known joint density p, independently of the past. Entries within each candidate may be dependent. Using the observed history, $\mathcal{X}_t$, and private randomness, the learner selects $X_t \in \mathcal{X}_t$ and observes only its reward, $y_t = f(\langle X_t, \Theta_*\rangle)+\eta_t$. The unknown low-rank parameter Θ_* and unknown Lipschitz link f are shared across actions and rounds; η_t is conditionally mean-zero noise with a variance bound σ^2, drawn independently from a fixed distribution depending only on X_t, and $\langle X, Z\rangle$ is the entrywise inner product. We first consider nondecreasing links, for which maximizing the index maximizes expected reward; Section 3.5 extends to nonmonotone links.

The goal is to minimize expected cumulative regret $\mathbb{E}[R_T]$ across T rounds, where R_T is defined as,

$$R_T = \sum_{t=1}^{T} \left[\max_{X \in \mathcal{X}_t} f(\langle X, \Theta_* \rangle) - f(\langle X_t, \Theta_* \rangle) \right]. \tag{1}$$

2.2 Low-Rank Structure

For a positive integer d, write $[d] = \{1, \ldots, d\}$. An order-h tensor with dimensions $d_1, \ldots, d_h$ has $N = \prod_{\ell=1}^{h} d_\ell$ entries. We use the inner product $\langle X, Z \rangle = \mathrm{vec}(X)^\top \mathrm{vec}(Z)$, the Frobenius norm $\|X\|_F = \sqrt{\langle X, X \rangle}$, the entrywise norms, and the ℓ_1-ball for some $B > 0$:

$$\|X\|_{\max} = \|\mathrm{vec}(X)\|_\infty, \qquad \|X\|_{\mathrm{ent},1} = \|\mathrm{vec}(X)\|_1, \qquad \mathcal{B}(B) = \{Z : \|Z\|_{\mathrm{ent},1} \le B\}.$$

Throughout, we assume that $\Theta_* \in \mathcal{B}(B)$, with the entrywise ℓ_1 budget imposed in the original coordinates. We next describe the low-rank structure imposed on Θ_* in the matrix and tensor settings.

Matrices. In the matrix setting, $X, \Theta_* \in \mathbb{R}^{d_1 \times d_2}$, and we write $d = \max(d_1, d_2)$. The operator norm $\|\cdot\|_{\mathrm{op}}$ is the largest singular value, while the nuclear norm $\|\cdot\|_*$ is the sum of the singular values. For an integer $1 \le r \le \min(d_1, d_2)$, we consider the class of low-rank matrices with rank at most r,

$$\mathcal{C}_{\mathrm{M}}(r, B) = \{Z \in \mathcal{B}(B) : \mathrm{rank}(Z) \le r\}, \qquad Z \in \mathbb{R}^{d_1 \times d_2}.$$

Transformed-slice tensors. Following tensor-product factorizations (Kilmer & Martin, 2011) and their transform-based extension (Kernfeld et al., 2015), we group the last $h - 2$ modes into $q = \prod_{\ell=3}^{h} d_\ell$, giving a $d_1 \times d_2 \times q$ array. A fixed, known real orthogonal matrix $U \in \mathbb{R}^{q \times q}$ defines a transform along the grouped mode through $(\mathcal{T}_U X)^{(k)}_{ij} = \sum_{a=1}^{q} U_{ka} X_{ija}$, where (k) denotes the kth frontal slice. Because U is orthogonal, $\mathcal{T}_U$ preserves inner products and Frobenius norms, and its inverse is $\mathcal{T}_{U^\top}$. We impose a rank bound on each transformed slice of Θ_*, allowing the bound to differ across slices. For integer caps $0 \le \rho_k \le \min(d_1, d_2)$, define $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)$, and consider the class of transformed-slice tensors,

$$\mathcal{C}_U(\boldsymbol{\rho}, B) = \{Z \in \mathcal{B}(B) : \mathrm{rank}((\mathcal{T}_U Z)^{(k)}) \le \rho_k \text{ for every } k \in [q]\}.$$

The ℓ_1 budget remains in the original coordinates; only rank constraints are imposed after the transform. Write $r_U = \sum_{k=1}^{q} \rho_k$ for the sum of the rank caps, so $r_U = qr$ when every slice has the same cap r. In both settings, the rank restrictions apply only to Θ_*; the actions X need not be low rank.

2.3 Score Function and Assumptions

Let $p : \mathbb{R}^{d_1 \times \cdots \times d_h} \to [0, \infty)$ be the joint Lebesgue density of the candidates. The score function $S^p : \mathbb{R}^{d_1 \times \cdots \times d_h} \to \mathbb{R}^{d_1 \times \cdots \times d_h}$ associated with the density p is defined as:

$$S^p(X) = -\nabla_X \log p(X) = -\nabla_X p(X)/p(X), \qquad p(X) > 0.$$

The gradient acts on all entries and has the same shape as X, with the p-a.e. weak interpretation in Appendix B.1. We impose the following conditions on the candidate density and its score.

Assumption 2.1 (Candidate distribution) *There exists some $A > 0$ such that $\|X\|_{\max} < A$ almost surely, and the joint Lebesgue density p is supported in $\{X : \|X\|_{\max} < A\}$. The zero extension of p has integrable weak first derivatives on $\mathbb{R}^N$. For a known finite upper bound $M > 0$,*

$$\mathbb{E}_p[\mathrm{vec}(S^p(X))\,\mathrm{vec}(S^p(X))^\top] \preceq M I_N.$$

Here $N = \prod_{\ell=1}^{h} d_\ell$ and I_N is the identity and $\preceq$ is the positive-semidefinite order.

This bounded-support model excludes Gaussian candidate laws. Under Assumption 2.1, integration by parts gives Stein's identity, justified in Appendix B.1:

$$\mathbb{E}_p[g(X) S^p(X)] = \mathbb{E}_p[\nabla_X g(X)], \tag{2}$$

for every smooth compactly supported scalar-valued function g on $\mathbb{R}^{d_1 \times \cdots \times d_h}$. The assumed weak regularity of the zero-extended density ensures that no boundary term appears. Appendix B.2 gives the product-beta example used in Section 4.

Assumption 2.2 (Link) *The unknown link f is L_f-Lipschitz on $[-AB, AB]$ for some $L_f > 0$.*

We assume that the learner knows the candidate law p and valid model bounds A, B, together with the transform U in the tensor setting. T-ESTOR additionally requires exact evaluation of the score under its selection rule, although its spectral updates do not require the rank caps or the signal lower bound; see Section 3.2. T-BSTOR uses the base score S^p and requires a structural bound and a projection certificate for its exploration schedule, as described in Section 3.5. Appendix C gives an estimated-score extension for unknown smooth product laws with slower constructive rates; its learned-selected-score epoch analysis requires additional certificates.

3 METHODS

3.1 STRUCTURED SINGLE-INDEX ESTIMATION

We use Stein's identity to estimate a scalar multiple of Θ_* without estimating the unknown link f. Fix a known sampling density p with the support and zero-extension regularity of Assumption 2.1, and retain the parameter class of Section 2.2 and the link condition of Assumption 2.2. Stein's identity gives

$$\mathbb{E}_p[yS^p(X)] = \mu_*\Theta_*, \qquad \mu_* = \mathbb{E}_p[f'(\langle X, \Theta_*\rangle)] \in [-L_f, L_f] \quad (\Theta_* \neq 0).$$

For $\Theta_* = 0$, we set $\mu_* = 0$ without evaluating $f'(0)$. Condition 3.2 ensures uniform positivity under selection. Section 3.2 applies the results conditionally with $p = p_i$.

To estimate this moment, we must account for the unrestricted offset $f(0)$ and potentially heavy-tailed observations. Although $\mathbb{E}_p S^p(X) = 0$ cancels the offset in the mean, it can still affect the second moment. We remove it by differencing independent observations. Specifically, given $2n$ iid observations $\{(X_j, y_j)\}_{j=1}^{2n}$ from the reward model, form n disjoint paired statistics

$$Z_j = \frac{y_{2j-1} - y_{2j}}{2}[S^p(X_{2j-1}) - S^p(X_{2j})], \qquad \mathbb{E}Z_j = \mu_*\Theta_*, \quad j \in [n]. \tag{3}$$

Each pair is unbiased, independent of the others, and differencing cancels the offset $f(0)$. To control their second moments, set $V = 2L_f^2A^2B^2 + \sigma^2$, and let $w > 0$ bound score second moments under the current observation law:

$$\max\{\|\mathbb{E}HH^\top\|_{\mathrm{op}}, \|\mathbb{E}H^\top H\|_{\mathrm{op}}\} \leq w,$$

Here $H = S^p(X)$ in the matrix case, or $H = (\mathcal{T}_U S^p(X))^{(k)}$ for every transformed slice k. Lemma D.1 then bounds both raw left and right second moments of Z_j, or each transformed slice of Z_j, by Vw in operator norm. This provides the moment control needed for our estimation guarantee. Section 3.2 derives a uniform choice of w under selection from the candidate-law bound M.

Matrix construction. We first reduce the influence of large singular values using $\psi(u) = \mathrm{sign}(u)\log(1 + |u| + u^2/2)$ (Minsker, 2018). For an SVD $H = P\,\mathrm{diag}(s_\ell)Q^\top$ and a tuning parameter $\nu > 0$, define

$$\widetilde{\psi}_\nu(H) = \frac{1}{\nu}P\,\mathrm{diag}(\psi(\nu s_\ell))Q^\top.$$

We apply this transformation to each pair before averaging, then use singular-value thresholding to encourage low rank. Specifically, SVT_τ replaces each singular value s by $(s - \tau)_+$. The threshold $\lambda/2$ corresponds to a nuclear-norm penalty with coefficient λ. Kang et al. (2022; 2024a) use the same influence function and penalized quadratic objective, including a single-index extension. Here pairing removes the unrestricted offset from the moment bounds.

Transformed slices, and a common estimator. For tensors, we apply the matrix construction to each slice under the supplied orthogonal transform U. For $q \geq 1$, the sum nuclear norm $\|Z\|_{*,U} = \sum_k \|(\mathcal{T}_U Z)^{(k)}\|_*$ encodes this slice geometry (Lu et al., 2019; Song et al., 2020; Lund, 2020). We first transform each paired statistic, robustify its slices, and average the corresponding slices:

$$\widehat{Z}_{\nu,k} = \frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}_\nu\big((\mathcal{T}_U Z_j)^{(k)}\big).$$

Then fit

$$\widehat{\Theta} = \underset{Z \in \mathbb{R}^{d_1 \times d_2 \times q}}{\arg\min} \left\{ \sum_{k=1}^{q} \left\| (\mathcal{T}_U Z)^{(k)} - \widehat{Z}_{\nu,k} \right\|_F^2 + \lambda \left\| Z \right\|_{*,U} \right\}. \tag{4}$$

Because both the loss and penalty separate across transformed slices, we solve this problem by thresholding each slice and applying the inverse transform:

$$\widehat{\Theta} = \mathcal{T}_U^{-1}\big([\mathrm{SVT}_{\lambda/2}(\widehat{Z}_{\nu,k})]_{k=1}^{q}\big).$$

For $q = 1$, $U = [1]$, and $r_U = r$, this is exactly the matrix estimator, with $\|\cdot\|_{*,U} = \|\cdot\|_*$ and the tuning parameters below agree with those of the matrix case.

Theorem 3.1 (Structured estimation rate) *For a fixed observation law and $2n$ iid observations satisfying the reward/noise model, Stein regularity, and score-moment bound w of Section 3.1, with $\Theta_* \in \mathcal{C}_U(\boldsymbol{\rho}, B)$ (or $\mathcal{C}_{\mathrm{M}}(r, B)$ in the matrix case), $n \geq 1$, and $\delta \in (0,1)$, in Equation* (4) *take*

$$\nu = \sqrt{\frac{2\log(2q(d_1+d_2)/\delta)}{nVw}}, \qquad \lambda = 2\sqrt{\frac{2Vw\log(2q(d_1+d_2)/\delta)}{n}}.$$

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

$$\left\| \widehat{\Theta} - \mu_* \Theta_* \right\|_F^2 \leq \frac{16 r_U V w \log(2q(d_1+d_2)/\delta)}{n} = \widetilde{O}\left(\frac{r_U V w}{n}\right).$$

3.2 T-ESTOR WITH STRUCTURED UPDATES

For T-ESTOR and its regret analysis, we assume that f is nondecreasing. This is without loss of generality within the monotone class: replacing (f, Θ_*) by $(u \mapsto f(-u), -\Theta_*)$ preserves the reward model and parameter constraints.

T-ESTOR applies the preceding estimators within the ESTOR epoch scheme (Kang et al., 2026). Epoch $i = 0, 1, \ldots$ has $2n_i = 2^{i+1}$ rounds, with $n_i = 2^i$. Starting from $\widehat{\Theta}_{-1} = 0$, the learner fixes the direction at the beginning of epoch i:

$$v_i = \begin{cases} \widehat{\Theta}_{i-1} / \left\| \widehat{\Theta}_{i-1} \right\|_F, & \widehat{\Theta}_{i-1} \neq 0, \\ 0, & \widehat{\Theta}_{i-1} = 0. \end{cases} \tag{5}$$

Throughout the epoch, it selects a candidate maximizing $\langle v_i, X \rangle$, choosing uniformly when $v_i = 0$. After completing the epoch, it computes $\widehat{\Theta}_i$ using only that epoch's observations. Conditional on the epoch-start history, the fixed ranking rule, iid candidate sets, and fresh draws from the fixed conditional noise law yield iid selected action–reward pairs.

Greedy selection changes the action distribution, so the update must use the score of the selected law. For $v_i \neq 0$, let F_i be the CDF of $\langle v_i, X \rangle$ under the candidate law p. Selecting the largest projection among K candidates gives the density p_i; differentiating its log density gives the corresponding score:

$$\begin{aligned} p_i(X) &= K p(X) F_i(\langle v_i, X \rangle)^{K-1}, \\ S^{p_i}(X) &= S^p(X) - (K-1)\frac{F_i'(\langle v_i, X \rangle)}{F_i(\langle v_i, X \rangle)} v_i \quad (p_i\text{-a.s.}). \end{aligned} \tag{6}$$

The factor F_i^{K-1} is the probability that the other $K-1$ candidates have smaller projections, while the factor K accounts for the possible selected indices. See Appendix E for the derivation. When $v_i = 0$, we draw a candidate index uniformly using fresh randomness, giving $p_i = p$ and $S^{p_i} = S^p$. We assume exact access to the selected scores in (6). To tune the epoch updates, we need a score second-moment bound that holds uniformly over the ranking directions. For $K \geq 3$, Appendix E.3 provides

$$w = M G_K(d), \qquad G_K(d) = 2Kd + \frac{K(K-1)^2}{2(K-2)}.$$

This bound applies to both matrix and transformed-slice scores, including uniform selection when $v_i = 0$. Let $Z_{i,j}$ denote the paired statistics in (3) formed from epoch i's observations using S^{p_i}. Theorem 3.1 then gives a high-probability bound on $\left\|\widehat{\Theta}_i - \mu_i \Theta_*\right\|_F$, where $\mu_i = \mathbb{E}_{p_i}[f'(\langle X, \Theta_* \rangle)]$. After each completed epoch, we compute $\widehat{\Theta}_i$ from these pairs using Theorem 3.1, with $n = n_i$, $\delta = \delta_E = (T+1)^{-2}$, and the bound w above. Algorithm 1 summarizes the procedure.

Algorithm 1 Tensor Epoched Stein's Oracle Single Index Bandit (T-ESTOR)

Input: Dimensions, $T, K, A, B, L_f, \sigma^2, M, p$; selected scores (6); U for tensors.

$t \leftarrow 1$, $i \leftarrow 0$, $\widehat{\Theta}_{-1} \leftarrow 0$, $\delta_E \leftarrow (T+1)^{-2}$
while $t \le T$ **do**
  $n_i \leftarrow 2^i$, $m_i \leftarrow \min\{2n_i, T - t + 1\}$; reset the epoch buffer
  Fix (v_i, p_i, S^{p_i}) by (5)–(6)
  for $j = 1, \ldots, m_i$ **do**
    Observe the arm set $\mathcal{X}_t$
    if $v_i = 0$ **then**
      Draw $a \sim \mathrm{Unif}([K])$;
    else
      $a \longleftarrow \arg\max_{b \in [K]} \langle v_i, X_{t,b} \rangle$;
    end if
    $X_t \leftarrow X_{t,a}$; Play X_t, observe y_t and append $(y_t, S^{p_i}(X_t))$ to the epoch buffer
    $t \leftarrow t + 1$
  end for
  if $m_i < 2n_i$ **then return** ▷ No partial-epoch update
  Form $Z_{i,1}, \ldots, Z_{i,n_i}$ by (3) from this epoch's buffer
  Fit $\widehat{\Theta}_i$ by (4) with Theorem 3.1's tuning; Epoch $i \leftarrow i + 1$
end while

For $\Theta_* \neq 0$, the estimate $\widehat{\Theta}_i$ targets $\mu_i \Theta_*$, where $\mu_i = \mathbb{E}_{p_i}[f'(\langle X, \Theta_* \rangle)]$. This multiplier varies with the sampling law. Since $\widehat{\Theta}_i$ determines the rankings in epoch $i+1$, we require a uniform positive lower bound to translate estimation accuracy into a regret guarantee.

Condition 3.2 (Selected-design signal) *For $\Theta_* \neq 0$, there exists $\mu_{\min} > 0$, possibly depending on K, such that $\mu_i \ge \mu_{\min}$ for every sampling law in* (6) *induced by a unit direction v_i, and for $p_i = p$ under uniform selection.*

For example, the condition holds with $\mu_{\min} = c$ if $f'(u) \ge c > 0$ almost everywhere on $[-AB, AB]$. The algorithm does not require knowledge of μ_i or $\mu_{\min}$.

3.3 Regret Analysis

We now derive regret bounds from estimation error under the nondecreasing-link assumption of Section 3.2 and Condition 3.2. For $\Theta_* \neq 0$, the estimate $\widehat{\Theta}_{i-1}$ used in epoch $i \ge 1$ targets $\mu_{i-1}\Theta_*$, where $\mu_{i-1} \ge \mu_{\min} > 0$. Centering candidates at their common mean preserves rankings and allows us to control error projections through the covariance $\Sigma_X = \mathrm{Cov}_p(\mathrm{vec}(X))$. Lemma F.1 then gives

$$\mathbb{E}[r_t \mid \mathcal{F}_{t-1}] \le \frac{2L_f}{\mu_{\min}} \sqrt{K \left\|\Sigma_X\right\|_{\mathrm{op}}} \left\|\widehat{\Theta}_{i-1} - \mu_{i-1}\Theta_*\right\|_F. \tag{7}$$

Here r_t is the instantaneous regret in (1), and $\mathcal{F}_{t-1}$ is the observed history. The bound also covers uniform selection when $\widehat{\Theta}_{i-1} = 0$; when $\Theta_* = 0$, regret is identically zero. Centering is used only in the proof: the algorithm operates on the original candidates and requires no covariance input.

Theorem 3.3 (Structured epoch regret) *Assume the sampling and noise model of Section 2.1, Assumptions 2.1 and 2.2, and Condition 3.2, with f nondecreasing. Let $K \ge 3$, $q \ge 1$, and $\Theta_* \in \mathcal{C}_U(\boldsymbol{\rho}, B)$, and write $r_U = \sum_{k=1}^{q} \rho_k$ and $d = \max(d_1, d_2)$. Run Algorithm 1 with exact*

selected scores and the tuning of Theorem 3.1, using $\delta_E = (T+1)^{-2}$ in each epoch. If $\Theta_* \neq 0$, then

$$\mathbb{E}R_T \le C\,\frac{L_f K}{\mu_{\min}}\sqrt{V M\,\|\Sigma_X\|_{\mathrm{op}}\; r_U(d+K)T\log\bigl(2q(d_1+d_2)T\bigr)} = \widetilde{O}(\sqrt{dr_U T}),$$

where $C > 0$ is universal. The matrix case follows by taking $q = 1$, $U = [1]$, and $\boldsymbol{\rho} = (r)$, giving $r_U = r$. When $\Theta_* = 0$, regret is zero. Every policy also satisfies $\mathbb{E}R_T \le 2L_f ABT$.

Proof idea. Conditional on the history at the start of each epoch, the frozen policy produces iid observations. The uniform selected-score moment bound $w = MG_K(d)$ from Appendix E.3 therefore lets us apply Theorem 3.1 in every epoch. Condition 3.2 and (7) convert the resulting estimation error into regret during the next epoch. Summing over doubling epochs yields the $\sqrt{T}$ dependence; the full proof appears in Appendix F.

For fixed K and model constants, the rates are $\widetilde{O}(\sqrt{drT})$ for matrices and $\widetilde{O}(\sqrt{dr_U T})$ for tensors, without rank inputs. Appendix F.1 compares these upper bounds with vectorization and unfolding: dense rank-one-slice tensors can give an order-$\sqrt{d}$ advantage over mode unfoldings for $K = O(d)$ under the same law and a linear link.

Worst-case reward and selected-score moment bounds can make the tuning conservative at finite horizons. Section 4 evaluates tuned implementations documented in Appendix I.

3.4 Lower Bounds

We next ask whether Theorem 3.3's rank and dimension dependence can be improved. A known linear link isolates the difficulty of learning the structured parameter. Fix any candidate law p satisfying Assumption 2.1, retain the fresh iid menus of Section 2.1, and take $f(z) = \beta z$ with $0 < \mu_{\min} \le \beta \le L_f$ and independent $\mathcal{N}(0, \sigma^2)$ noise, $\sigma^2 > 0$. For every such p this submodel satisfies Condition 3.2 with $\mu_{\min} = \beta$, so it lies in the model of Theorem 3.3 for the same p.

Consider $d = \max(d_1, d_2) \ge 2$ with a positive rank cap, or total slice rank cap, so $s_{\max} \ge 2$. For the admissible ranks in Section 2.2, set $s_{\max} = dr$ and $\chi = 1$ in the matrix case. In the tensor case, assume $r_U \ge 1$ and set $s_{\max} = dr_U$. If all slice rank caps equal some $r \ge 1$, take $\chi = 1$ for any supplied orthogonal transform U; otherwise, take $\chi = \max_{k:\rho_k>0} \sum_{\ell=1}^{q} |U_{k\ell}|$. Appendix G builds s_{lb} rank-feasible directions orthogonal to $\mathbb{E}_p X$, with $s_{\mathrm{lb}} \ge \lfloor s_{\max}/2 \rfloor \ge s_{\max}/3$ and $s_{\mathrm{lb}} = s_{\max}$ for centered designs.

Theorem 3.4 (Large-horizon lower bound for every design) *Under the model above, let $\mathcal{C}$ be $\mathcal{C}_{\mathrm{M}}(r, B)$ or $\mathcal{C}_U(\boldsymbol{\rho}, B)$ with structural scale $s_{\max} \ge 2$ as defined above. There is a universal $c > 0$ such that, for every admissible candidate law p and every integer $T \ge \max\{1, T_+(p)\}$,*

$$\inf_{\pi}\sup_{\Theta_*\in\mathcal{C}} \mathbb{E}_{\Theta_*,\pi} R_T \ge \frac{c\,\sigma}{K\,M\,\|\Sigma_X\|_{\mathrm{op}}}\sqrt{s_{\max}\,T}.$$

Here $T_+(p)$ is given in Equation (29)*. The infimum ranges over all randomized adaptive policies, which may know p, the linear link, and the structural subspace used in the proof.*

For fixed K and model/design constants, uniformly controlled across dimensions, the $\sqrt{s_{\max}T}$ dependence matches Theorem 3.3 up to logarithms for $T \ge \max\{1, T_+(p)\}$. $T_+(p)$ marks a construction branch, not algorithmic burn-in or an established minimax transition.

3.5 T-BSTOR for arbitrary Links

For nonmonotone links, a larger index need not yield a higher reward, so estimating the parameter direction alone is insufficient for action selection. T-BSTOR first estimates the index up to an unknown scale using uniformly selected candidates, then learns rewards along the fitted index through the estimate-then-bin approach of Dey et al. (2026).

We retain the parameter classes, candidate-law regularity, and finite-variance noise model of Section 2, with f now any L_f-Lipschitz link on $[-AB, AB]$. For $\Theta_* \neq 0$, assume $|\mu_*| \ge \gamma > 0$, where $\mu_* = \mathbb{E}_p[f'(\langle X, \Theta_*\rangle)]$. This signal may have either sign. Uniform exploration uses only the base

score S^p, so Condition 3.2 is unnecessary. To control prediction error from the fitted index, write $m = \mathbb{E}_p X$ and let $\kappa_X \geq 1$ satisfy

$$\mathbb{E}_p e^{\lambda\langle X-m,Z\rangle} \leq \exp\left(\frac{\lambda^2 A^2 \kappa_X^2 \|Z\|_F^2}{2}\right) \quad \text{for every } Z \text{ and } \lambda \in \mathbb{R}. \tag{8}$$

Independent bounded entries admit $\kappa_X = 1$, and every bounded-entry law admits $\kappa_X = \sqrt{N}$; Appendix H.1 discusses dependent designs.

For an even exploration length $n_0 < T$, select candidates uniformly during the first n_0 rounds and fit the structured estimator using $n_0/2$ disjoint observation pairs. Project the fit in Frobenius norm onto the entrywise ball $\mathcal{B}(L_f B)$ and hold it fixed. This keeps every fitted index in $[-AL_f B, AL_f B]$, which we partition into bins of width at most w_{bin}. On each subsequent round, T-BSTOR chooses among bins containing current candidates: it prioritizes bins with too few observations and otherwise uses median-of-means upper confidence bounds to select a bin and plays a candidate in it. Algorithm 2 gives the exact policy and tuning. The concentration bound above converts estimation error into prediction error, while Lipschitz continuity controls reward variation within each bin; Appendix H.2 makes this approximation precise.

Theorem 3.5 (Horizon rate for arbitrary links) *Under these local assumptions, run Algorithm 2 with a supplied bound $D \geq 1$, $D \geq dr$ (matrices) or $D \geq dr_U$ (transformed slices), the confidence allocation* (43)*, and*

$$n_0 = \min\left\{T,\ 2\left\lceil \frac{(\kappa_X^2 D)^{1/3} T^{2/3}}{2} \right\rceil\right\}, \qquad w_{\mathrm{bin}} = 2AL_f B\, T^{-1/3}.$$

If $n_0 = T$, select uniformly throughout; otherwise, run the estimation and binning steps above. Then

$$\mathbb{E}[R_T] = \widetilde{O}\Big(\big[(\kappa_X^2 D)^{1/3} + 1\big] T^{2/3}\Big).$$

Moreover, $R_T \leq 2L_f ABT$. Here $A, B, L_f, M, \gamma, \sigma, K, h$ are held fixed, D and κ_X remain explicit, and only logarithms in dimensions and T are suppressed.

Only the exploration schedule needs the supplied D and κ_X; spectral fitting needs no rank caps. Under the same projection certificate and comparable model constants, the structured fit replaces the vectorized leading upper-bound factor $(\kappa_X^2 N)^{1/3}$ by $(\kappa_X^2 D)^{1/3}$, improving the bound when $D < N$.

For each admissible candidate law, Theorem H.6 gives an $\Omega(T^{2/3})$ lower bound at large horizons with fixed $K \geq 3$ and positive signal floor, even when a nonzero structured parameter is disclosed. Dey et al.'s lower-bound construction instead grows the menu and decreases the signal floor (Dey et al., 2026). Thus the horizon exponent is optimal for fixed dimensions, $K \geq 3$, and model/design constants, with $\sigma > 0$ and $0 < \gamma < L_f$. This comparison concerns only the horizon rate; its constants depend on the design. Links with zero base signal remain outside the guarantee, while a zero parameter has zero regret.

4 Experiments

Design. We use a shared dense $16 \times 16 \times 4$ parameter with an entrywise ℓ_1 budget of one. Its 16×64 matrix representation and each frontal slice under the identity transform have rank two, giving both representations a structural scale of 128, compared with an ambient dimension of 1024. We consider six links, three monotone and three nonmonotone, with $K = 10$ candidates per round. Candidate entries are drawn independently from $2\,\mathrm{Beta}(5,5) - 1$, and rewards are corrupted by independent scaled Student-t_3 noise. Appendix I provides parameters, links, and noise scaling.

Methods and evaluation. T-BSTOR-Vec, T-BSTOR-Mat, and T-BSTOR-Slice combine vector, matrix, and transformed-slice estimation, respectively, with the same robust scalar-bin learner. We compare these methods with GSTOR and ZoomSIB-UCB, and evaluate T-ESTOR, which exploits monotonicity, only on the monotone links. We tune influence and threshold multipliers; T-BSTOR also uses adaptive exploration (Dey et al., 2026), normalized-index bins, and tuned confidence multipliers (Appendix I). The guarantees concern the stated theoretical policies and tuning. T-ESTOR

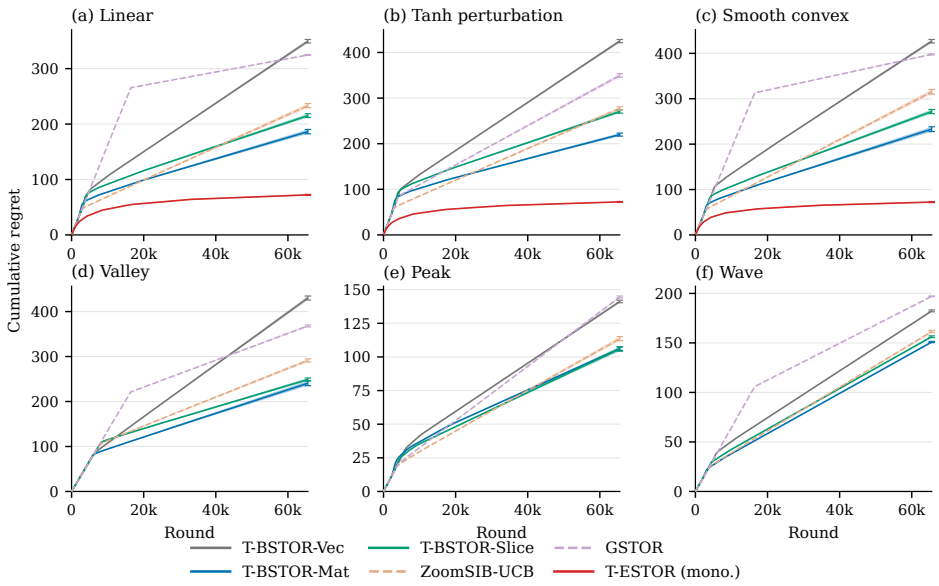


Figure 1: Mean cumulative regret over 50 paired seeds, with pointwise 95% bootstrap bands and endpoint whiskers. Top: monotone links; bottom: nonmonotone. Solid: proposed methods; dashed: external baselines.

uses numerical selected scores; Appendix I.3 reports accuracy and computational cost. Tuning, validation, and testing use separate seeds. With configurations frozen, we evaluate 50 fresh paired seeds through $T = 65{,}534$. Methods share instances but observe only their chosen rewards. Regret uses each menu's best action and includes all exploration, initialization, and fallback rounds.

Results. Both structured T-BSTOR variants achieve lower mean final regret than T-BSTOR-Vec, GSTOR, and ZoomSIB-UCB on all six links (Figure 1). All reported unadjusted paired 95% intervals for the prespecified Slice comparisons favor Slice (Table 4). Mean direction errors are 0.41–0.63 for structured fits and 0.66–0.80 for vectorization (Table 6). Exploration lengths and tuning also differ, so regret gains cannot be attributed to structure alone. On monotone links, T-ESTOR's mean final regret is approximately 72, versus 190–230 for the best binning variant.

CCLE-based evaluation. We use Cancer Cell Line Encyclopedia (CCLE) drug-response data to construct a coefficient for simulated dense actions and feedback. On Linear, T-ESTOR improves on ESTOR and achieves mean regret comparable to G-LowTESTR despite not knowing the link. On Wave, T-BSTOR-Slice-DCT has lower mean regret than vectorization, GSTOR, and ZoomSIB-UCB; G-LowTESTR and the Cell and Dose unfoldings have lower means than Slice-DCT. These comparisons are descriptive. Appendix J gives the setup and detailed results.

5 Conclusion

We studied low-rank matrix and tensor bandits with an unknown shared Lipschitz link and finite-variance noise. Under the stated assumptions, T-ESTOR achieves square-root regret for monotone links with dimension dependence determined by low-rank structure. For every admissible design, the monotone lower bound matches rank, dimension, and horizon dependence up to logarithms at large horizons, for fixed menu size and model/design constants. T-BSTOR attains the optimal $\widetilde{O}(T^{2/3})$ nonmonotone horizon rate for fixed dimensions, menu size, and model/design constants. Synthetic and CCLE-based experiments illustrate the benefits of structured estimation.

Limitations and future work. The main-policy guarantees rely on known regular candidate distributions, supplied tensor transforms, and the stated Stein-signal conditions; T-ESTOR additionally requires exact selected scores. Future work includes certifying learned selected scores, handling zero first-order signals, and establishing guarantees for the adaptive exploration used in our experiments.

AI USE STATEMENT

AI assistance was used for project setup, literature search and discovery, experiment implementation, and manuscript editing.

REPRODUCIBILITY STATEMENT

All theoretical results presented in this work are supported by complete proofs in the appendix. The experiments use frozen configurations and explicit seeds. The experiment repository retains the code, study configurations, tuning/validation/evaluation seed partitions, seed-level outcomes, numerical diagnostics, and analysis underlying the reported figures and tables. Synthetic experimental details appear in Appendix I; the CCLE data construction, preprocessing, and reused-seed bandit evaluations appear in Appendix J. The full source code used to generate our experimental results will be included as supplementary materials to facilitate reproducibility.

REFERENCES

Sakshi Arya and Hyebin Song. Batched single-index global multi-armed bandits with covariates, 2026. URL https://arxiv.org/abs/2503.00565v3. Version 3, revised May 20, 2026.

Sakshi Arya, Satarupa Bhattacharjee, and Bharath K. Sriperumbudur. Kernel single-index bandits: Estimation, inference, and learning, 2026. URL https://arxiv.org/abs/2603.18938v1. Version 1, submitted March 19, 2026.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. doi: 10.1038/nature11003. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3320027/.

Devdan Dey, Sujoy Bhore, and Avishek Ghosh. Optimal regret for single index bandits, 2026. URL https://arxiv.org/abs/2605.09454v3. Version 3, revised August 1, 2026.

Jianqing Fan, Zhuoran Yang, and Mengxin Yu. Understanding implicit regularization in overparameterized single index model. *Journal of the American Statistical Association*, 118(544): 2315–2328, 2023. doi: 10.1080/01621459.2022.2044824. URL https://doi.org/10.1080/01621459.2022.2044824.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL https://jmlr.org/papers/v6/hyvarinen05a.html.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21329–21372. PMLR, 2024. URL https://proceedings.mlr.press/v235/jang24e.html.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3163–3172. PMLR, 2019. URL https://proceedings.mlr.press/v97/jun19a.html.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19971–19983. Curran Associates, Inc., 2022. doi: 10.52202/068431-1452. URL https://doi.org/10.52202/068431-1452.

Yue Kang, Cho-Jui Hsieh, and Thomas C. M. Lee. Efficient frameworks for generalized low-rank matrix bandit problems, 2024a. URL https://arxiv.org/abs/2401.07298v1. Version 1, submitted January 14, 2024; later version of the NeurIPS 2022 paper.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Low-rank matrix bandits with heavy-tailed rewards. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1863–1889. PMLR, 2024b. URL https://proceedings.mlr.press/v244/kang24a.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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.12751v3. Author manuscript, arXiv:2506.12751v3, revised February 7, 2026.

Eric Kernfeld, Misha Kilmer, and Shuchin Aeron. Tensor–tensor products with invertible linear transforms. *Linear Algebra and its Applications*, 485:545–570, 2015. doi: 10.1016/j.laa.2015.07.021. URL https://doi.org/10.1016/j.laa.2015.07.021.

Misha E. Kilmer and Carla D. Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011. doi: 10.1016/j.laa.2010.09.020. URL https://doi.org/10.1016/j.laa.2010.09.020.

Jiannan Li, Yiyang Yang, Yao Wang, and Shaojie Tang. A unified regularization approach to high-dimensional generalized tensor bandits. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2025. doi: 10.1109/ISIT63088.2025.11195554. URL https://arxiv.org/abs/2501.10722v2. Author manuscript, arXiv:2501.10722v2, revised January 23, 2025.

Canyi Lu, Xi Peng, and Yunchao Wei. Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. doi: 10.1109/CVPR.2019.00615.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 460–468. PMLR, 2021. URL https://proceedings.mlr.press/v130/lu21a.html.

Kathryn Lund. The tensor t-function: A definition for functions of third-order tensors. *Numerical Linear Algebra with Applications*, 27(3), 2020. doi: 10.1002/nla.2288.

Wanteng Ma and T. Tony Cai. Nonparametric bandits with single-index rewards: Optimality and adaptivity, 2025. URL https://arxiv.org/abs/2512.24669v1. Version 1, submitted December 31, 2025.

Stanislav Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018. doi: 10.1214/17-AOS1642.

Yaniv Plan and Roman Vershynin. The generalized Lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016. doi: 10.1109/TIT.2016.2517008. URL https://arxiv.org/abs/1502.04071v2.

Chengshuai Shi, Cong Shen, and Nicholas D. Sidiropoulos. On high-dimensional and low-rank tensor bandits. In *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023. doi: 10.1109/ISIT54713.2023.10206552. URL https://arxiv.org/abs/2305.03884.

Guangjing Song, Michael K. Ng, and Xiongjun Zhang. Robust tensor completion using transformed tensor singular value decomposition. *Numerical Linear Algebra with Applications*, 27(3), 2020. doi: 10.1002/nla.2299.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-Gaussian single index models via thresholded score function estimation. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3851–3860. PMLR, 2017. URL https://proceedings.mlr.press/v70/yang17a.html.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. On Stein's identity and near-optimal estimation in high-dimensional index models, 2018. URL https://arxiv.org/abs/1709.08795.

Qianxin Yi, Yiyang Yang, Shaojie Tang, Jiapeng Liu, and Yao Wang. Efficient generalized low-rank tensor contextual bandits, 2024. URL https://arxiv.org/abs/2311.01771v3. Version 3, revised January 17, 2024.

Jie Zhou, Botao Hao, Zheng Wen, Jingfei Zhang, and Will Wei Sun. Stochastic low-rank tensor bandits for multi-dimensional online decision making, 2024. URL https://arxiv.org/abs/2007.15788v3. Version 3, revised February 13, 2024; initial preprint July 31, 2020.

APPENDIX CONTENTS

A Notation **14**
B Candidate Distribution and Stein Identities **15**
B.1 Joint Regularity and the Base Identity 15
B.2 Explicit Admissible Densities. 15
C Unknown Distributions and Estimated Scores. **16**
C.1 Scope and Overview. 16
C.2 Calibration Interface and Perturbed Structured Estimation 16
C.3 Calibration–Exploration–Commit 17
C.4 An Explicit Smooth Product Class 18
C.5 A Finite Calibration Algorithm and Its Proof. 18
C.6 Allocation and the Resulting Rates 20
C.7 Conditional Return to Greedy Epochs 20
D Structured Single-Index Estimation Proofs. **21**
D.1 Paired Moments 21
D.2 Matrix Estimation Proofs 22
D.3 Transformed-Tensor Estimation Proofs 23
E T-ESTOR with Structured Updates **24**
E.1 Conditional Probability Model 24
E.2 Projection Laws and Selected Scores 24
E.3 Uniform Selected-Score Moments. 25
F Regret Proofs **25**
F.1 Detailed Structural Comparisons. 27
G Lower-Bound Proofs **28**
H T-BSTOR for Arbitrary Links. **32**
H.1 Local Assumptions and Structured Index Estimation 32
H.2 Projection, Scalar Approximation, and T-BSTOR. 34
H.3 Confidence for Adaptively Sampled Bins 35
H.4 Upper Regret Bound and Exploration Schedule 37
H.5 A Fixed-Menu Lower Bound for Arbitrary Links 38
I Experimental details **41**
I.1 Dense synthetic design and six links 41
I.2 Practical adaptive T-BSTOR 42
I.3 Other algorithms and numerical implementations. 42
I.4 Registered tuning and reporting 44
I.5 Synthetic settings and diagnostic summaries 44
J CCLE-derived bandit experiments **47**
J.1 Data and preprocessing 47
J.2 Fixed bandit environment and evaluation 47
J.3 Bandit results 48

ROADMAP

Appendix A collects notation. Appendix B establishes the Stein identities and admissible designs. Appendix C develops the estimated-score interface, a constructive extension to unknown product candidate laws, and a conditional learned-score epoch analysis. Appendix D proves the structured estimation guarantees; Appendices E–F establish selected-score bounds, monotone regret, and structural comparisons. Appendix G proves the large-horizon lower bound for every admissible design, with a sharp-constant product-beta corollary. Appendix H gives the nonmonotone analysis, bin-confidence bounds, exploration-schedule sensitivity, and horizon lower bound for fixed dimensions, menu size, and model/design constants. Appendices I–J describe the synthetic and CCLE-derived experiments.

A NOTATION

Principal notation used in the main text.

Symbol	Meaning
$T,\ t,\ K$	Time horizon, round index, and candidates per round.
$\mathcal{X}_t,\ X_{t,a},\ X_t$	Candidate menu, candidate a, and chosen action at round t.
$\Theta_*,\ \widehat{\Theta}$	True parameter and structured fit targeting $\mu_*\Theta_*$.
$f,\ L_f$	Reward link and its Lipschitz bound.
$y_t,\ \eta_t,\ \sigma^2$	Reward, noise, and conditional noise-variance bound.
$R_T,\ r_t$	Cumulative regret and its round-t summand.
$A,\ B$	Action entry bound and parameter entrywise-ℓ_1 budget.
$h,\ d_\ell,\ N$	Array order, mode sizes, and ambient size $N = \prod_\ell d_\ell$.
$d_1,\ d_2,\ d,\ q$	Leading dimensions, $d = \max(d_1, d_2)$, and slice count.
$r,\ \rho_k,\ r_U$	Rank cap, slice rank caps, and their sum $r_U = \sum_k \rho_k$.
$U,\ \mathcal{T}_U,\ X^{(k)}$	Orthogonal transform, its tensor action, and frontal slice k.
$\mathcal{C}_{\mathrm{M}},\ \mathcal{C}_U$	Matrix and transformed-slice parameter classes.
$\mathcal{B}(B)$	Entrywise-ℓ_1 ball of radius B.
$p,\ p_i$	Base or locally fixed sampling density, and epoch sampling density.
$S^p,\ S^{p_i}$	Full joint scores of these densities, in array shape.
$M,\ \Sigma_X$	Known joint score-moment bound and covariance $\Sigma_X = \mathrm{Cov}_p(\mathrm{vec}(X))$.
$m,\ \kappa_X$	Candidate mean $\mathbb{E}_p X$ and centered joint-projection concentration certificate.
$\mu_*,\ \mu_i$	Actual Stein multipliers for a fixed sampling law and epoch i.
$\mu_{\min},\ \gamma$	Uniform positive selected-signal and absolute base-signal lower bounds.
$n,\ j,\ Z_j$	Disjoint pair count ($2n$ observations), pair index, and paired reward–score statistic.
$V,\ w$	Paired-reward factor $2L_f^2A^2B^2 + \sigma^2$ and score second-moment bound.
$\widehat{Z}_\nu,\ \widehat{Z}_{\nu,k}$	Robust empirical Stein moment and its transformed-slice counterpart.
$\psi,\ \widetilde{\psi}_\nu$	Scalar influence and its spectral extension including the factor $1/\nu$.
$\nu,\ \lambda$	Spectral influence tuning and nuclear-penalty coefficient.
$\delta,\ \delta_E$	Failure probability and per-epoch choice $\delta_E = (T+1)^{-2}$.
$i,\ n_i,\ v_i,\ \widehat{\Theta}_i$	Epoch, $n_i = 2^i$ pairs, frozen direction, and epoch-i fit targeting $\mu_i\Theta_*$ for epoch $i+1$.
$F_i,\ F_i',\ G_K(d)$	Projection CDF and density; selected-score inflation factor.
$\mathcal{C},\ s_{\max}$	Chosen lower-bound parameter class and hard-subspace dimension.
$c_X,\ M_0$	Exact beta-design covariance and score moments $A^2/11$ and $12/A^2$ (Appendix B.2, Corollary G.3).
$\chi,\ T_+(p)$	Entrywise cost of the hard subspace and large-horizon threshold.
$s_{\mathrm{lb}},\ \chi_{\mathrm{lb}}$	Number and entrywise cost of the mean-orthogonal hard directions.
β	Disclosed linear-link slope in the lower bound.
$n_0,\ w_{\mathrm{bin}},\ D$	Actual exploration count (even when $n_0 < T$), bin width, and a supplied bound on dr or dr_U.

B Candidate Distribution and Stein Identities

B.1 Joint Regularity and the Base Identity

Write $x = \mathrm{vec}(X)$ and extend the density p by zero to all of $\mathbb{R}^N$. The space $W^{1,1}(\mathbb{R}^N)$ consists of integrable functions whose first distributional derivatives are integrable. The zero-extension condition and weak score in Assumption 2.1 mean that

$$p \in W^{1,1}(\mathbb{R}^N), \qquad \nabla_x p(x) = -p(x)\ \mathrm{vec}(S^p(X)) \quad \text{almost everywhere.}$$

Under score integrability, these properties are equivalent to the compact-test Stein identity in Equation (2). Indeed, for every $g \in C_c^\infty(\mathbb{R}^N)$, the identity says $\int gp\ \mathrm{vec}(S^p) = \int p\nabla g$. By the definition of a distributional derivative this is precisely $\nabla p = -p\,\mathrm{vec}(S^p)$. The right-hand side is integrable, so p belongs to $W^{1,1}$. Conversely, this weak derivative equation gives the compact-test identity by integration by parts. The full joint second-moment bound implies the required score integrability by Cauchy–Schwarz. Because the derivative identity holds on all of $\mathbb{R}^N$, it includes the support boundary and permits no additional boundary measure. On $\{p > 0\}$ it identifies the weak score with $-\nabla_X \log p$; the weak derivative of a nonnegative Sobolev function is zero almost everywhere on its zero set. A smooth cutoff equal to one on the bounded support then gives $\mathbb{E}_p\,\mathrm{vec}(S^p(X)) = -\int \nabla p = 0$.

For $\Theta_* \neq 0$, orthogonal coordinates and Fubini show that $\langle X, \Theta_* \rangle$ has a density: its law is a scaled one-dimensional marginal of the joint density. Extend f as a Lipschitz function outside $[-AB, AB]$ and apply weak integration by parts to its composition with this index, multiplied by the same cutoff. The Sobolev chain rule gives gradient $f'(\langle X, \Theta_* \rangle)\,\mathrm{vec}(\Theta_*)$ almost everywhere. The index density ensures that the exceptional set where f' is undefined has probability zero. The index is bounded, and the finite score second moment makes every product integrable for each finite, unrestricted value of $f(0)$. Conditional mean-zero noise contributes zero, proving the base identity

$$\mathbb{E}_p[yS^p(X)] = \mu_* \Theta_*, \qquad \mu_* = \mathbb{E}_p[f'(\langle X, \Theta_* \rangle)] \in [-L_f, L_f].$$

When $\Theta_* = 0$, the signal part is constant and the moment vanishes by score and noise centering, without evaluating $f'(0)$.

B.2 Explicit Admissible Densities

Product-beta design. With $\mathcal{I} = [d_1] \times \cdots \times [d_h]$, a concrete admissible density is

$$p_\beta(X) = \prod_{\boldsymbol{i} \in \mathcal{I}} p_A(X_{\boldsymbol{i}}), \qquad p_A(u) = \frac{315}{256A}\left(1 - \frac{u^2}{A^2}\right)^4 \mathbf{1}\{|u| < A\}. \tag{9}$$

Write $x = \mathrm{vec}(X)$. The score and moments are

$$(S^{p_\beta}(X))_{\boldsymbol{i}} = \frac{8X_{\boldsymbol{i}}}{A^2 - X_{\boldsymbol{i}}^2}, \qquad c_X = A^2/11, \quad M_0 = 12/A^2,$$
$$\mathrm{Cov}_{p_\beta}(x) = c_X I_N, \qquad \mathbb{E}_{p_\beta}[\mathrm{vec}(S^{p_\beta}(X))\,\mathrm{vec}(S^{p_\beta}(X))^\top] = M_0 I_N.$$

The constants c_X and M_0 refer only to this beta design, including Corollary G.3, the product-beta remark in Appendix H.5, and the experiments. With $u = x_i/A$, normalization and the moments follow from $\int_{-1}^1 u^{2j}(1-u^2)^a\,du = \mathrm{B}(j + 1/2, a + 1)$; the fourth moment is $\mathbb{E}x_i^4 = 3A^4/(11 \cdot 13)$. The fourth-order taper has zero boundary trace and integrable first derivative, so the zero-extended product density is in $W^{1,1}$. Symmetry centers coordinates and scores; independence gives the exact joint covariance $c_X I_N$ and score second moment $M_0 I_N$. For this example only and any nonzero v, the density of $\langle v, X \rangle$ is the convolution over nonzero coordinates v_i of $|v_i|^{-1} p_A(u/v_i)$. It is positive in the interior of $[-A\,\|v\|_{\mathrm{ent},1}, A\,\|v\|_{\mathrm{ent},1}]$ and vanishes at its endpoints. Taking $\|\Sigma_X\|_{\mathrm{op}} = c_X$ and $M = M_0$ in Theorem 3.3 recovers the beta-design guarantee, subject to its nondecreasing-link, selected-score access, and signal assumptions.

Correlated, noncentered designs. Let z have the product-beta law (9) with A_0 in place of A, and set $x = \mathrm{vec}(X) = b + Hz$ for known b and invertible H with $\|b\|_\infty + A_0\,\|H\|_{\infty\to\infty} \leq A$.

Translation and an invertible linear change of variables preserve the zero-extension $W^{1,1}$ regularity, and

$$p_X(x) = \frac{p_\beta(H^{-1}(x-b))}{|\det H|}, \qquad \mathrm{vec}(S^{p_X}(X)) = H^{-\top}\,\mathrm{vec}(S^{p_\beta}(Z)),$$

where $\mathrm{vec}(Z) = H^{-1}(x-b)$. Hence $\Sigma_X = (A_0^2/11)HH^\top$ and $M = (12/A_0^2)\left\|H^{-1}\right\|_{\mathrm{op}}^2$, so $\|\Sigma_X\|_{\mathrm{op}} M = (12/11)\,\mathrm{cond}_2(H)^2$. Hoeffding's lemma applied to z gives (8) with $\kappa_X = \max\{1, (A_0/A)\,\|H\|_{\mathrm{op}}\}$. For example, take block-diagonal H with blocks

$$\frac{1}{1+\rho}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad 0 < \rho < 1,$$

with at least one two-coordinate block and an identity block for an unmatched coordinate. Its row-sum and operator norms are one, and its condition number is $(1+\rho)/(1-\rho)$, independent of N. With $A_0 = A/2$, any $\|b\|_\infty \le A/2$ is allowed; paired coordinates have covariance $2\rho A_0^2/\{11(1+\rho)^2\} > 0$, and $\kappa_X = 1$ with a dimension-free M for fixed ρ. The structured parameter is not transformed; selected-score access and the signal conditions remain assumptions.

C Unknown Distributions and Estimated Scores

C.1 Scope and Overview

This appendix treats an unknown product candidate law on $(-A, A)^N$ with zero-boundary, absolutely continuous marginal densities and square-integrable scores. Retain the support, entrywise budgets, and absolutely continuous L_f-Lipschitz link regularity of Section 2, together with the fixed conditional noise kernel used in Section 3.2. For the monotone regret results below, impose the non-decreasing orientation of Section 3.2 and require $\mu_* := \mathbb{E}_p f'(\langle X, \Theta_* \rangle) \ge \mu_{\min} > 0$ for $\Theta_* \ne 0$. This is a positive *base-law* signal condition. The learner sees all menu candidates but knows neither p nor its score.

Independent unlabeled calibration yields score-error and fitted-score covariance certificates. Theorem C.1 transfers them to matrix and transformed-slice estimation, and Sections C.3–C.6 give a calibration–exploration–commit policy with a finite score-matching construction (Hyvärinen, 2005) on an explicit smooth product class. For fixed class, model and structural quantities, its regret is

$$\mathbb{E}R_T = \widetilde{O}\Big(T^{2/3} + T^{(2s+14)/(3s+14)}\Big). \tag{10}$$

An exactly specified fixed-dimensional score family gives $\widetilde{O}(T^{2/3})$. These are constructive, not asserted optimal, unknown-density rates. They concern the separate calibration–exploration–commit policy; extending greedy T-ESTOR also requires direction-valid selected-score certificates (Section C.7). A zero parameter has zero regret without evaluating $f'(0)$.

C.2 Calibration Interface and Perturbed Structured Estimation

Let p be a fixed observation density whose zero extension lies in $W^{1,1}(\mathbb{R}^N)$ with square-integrable score, as in Appendix B.1. Retain the support, Lipschitz regularity, budget and noise assumptions above, but impose neither monotonicity nor a nonzero Stein signal for the estimation result; here p need not be a product law and may be a frozen selected law. Conditional on a calibration history $\mathcal{G}$, p and $\widehat{S}$ are fixed, and subsequent action–reward observations are iid with that law and reuse no calibration observations. On a calibration event of probability at least $1-\delta_{\mathrm{cal}}$, suppose deterministic certificates $\varepsilon \ge 0$, $\widehat{J} > 0$ satisfy

$$\mathrm{Cov}_p(\mathrm{vec}(\widehat{S} - S^p)) \preceq \varepsilon^2 I_N, \qquad \mathrm{Cov}_p(\mathrm{vec}\,\widehat{S}) \preceq \widehat{J} I_N. \tag{11}$$

All expectations below condition on $\mathcal{G}$. Write $V = 2L_f^2 A^2 B^2 + \sigma^2$ and, for $\Theta_* \ne 0$, put $\mu_* = \mathbb{E}_p f'(\langle X, \Theta_* \rangle) \in [-L_f, L_f]$; otherwise set $\mu_* = 0$. The estimation bound allows signed or zero multipliers.

Theorem C.1 (Structured error with estimated scores) *Conditional on successful calibration, form* $\widehat{Z}_j = \frac{1}{2}(y_{2j-1} - y_{2j})[\widehat{S}(X_{2j-1}) - \widehat{S}(X_{2j})]$ *from* $n \geq 1$ *fresh pairs. For* $\Theta_* \in \mathcal{C}_{\mathrm{M}}(r, B)$ *or* $\mathcal{C}_U(\boldsymbol{\rho}, B)$*, apply the corresponding spectral estimator with the tuning in Equation* (17) *or* (22)*, using* $w = d\widehat{J}$ *and* $d = \max(d_1, d_2)$*. For* $\delta \in (0, 1)$*, with conditional probability at least* $1 - \delta$*, respectively,*

$$\left\|\widehat{\Theta} - \mu_*\Theta_*\right\|_F \leq \begin{cases} 4\sqrt{Vrd\widehat{J}\log(2(d_1+d_2)/\delta)/n} + L_f AB\varepsilon, & \textit{matrix}, \\ 4\sqrt{Vr_U d\widehat{J}\log(2q(d_1+d_2)/\delta)/n} + L_f AB\varepsilon, & \textit{transformed slices}. \end{cases}$$

Proof. Write $F(X) = f(\langle X, \Theta_* \rangle)$ and $e = \widehat{S} - S^p$. Conditional mean-zero noise and Stein integration by parts give $\mathbb{E}\widehat{Z}_j = \mathrm{Cov}_p(F, \widehat{S}) = \mu_*\Theta_* + B_p$, where $B_p = \mathrm{Cov}_p(F, e)$. For a unit array D, Cauchy–Schwarz and the range bound for F imply

$$|\langle B_p, D\rangle| \leq \sqrt{\mathrm{Var}_p F}\sqrt{\mathrm{Var}_p\langle e, D\rangle} \leq L_f AB\varepsilon.$$

Thus $\|B_p\|_F \leq L_f AB\varepsilon$ without a dimension factor. An independent score difference has second moment twice its covariance; Lemma D.1's paired-moment proof therefore gives

$$\mathbb{E}[\mathrm{vec}(\widehat{Z}_j)\,\mathrm{vec}(\widehat{Z}_j)^\top] \preceq V\,\mathrm{Cov}_p(\mathrm{vec}\,\widehat{S}) \preceq V\widehat{J}I_N.$$

The left and right raw moments of each matrix or transformed slice are bounded by $dV\widehat{J}$. Spectral concentration controls the influenced average about $\mu_*\Theta_* + B_p$. Subtract B_p from that average only for analysis and apply the thresholding argument of Appendices D.2–D.3 to the structured target. Frobenius nonexpansiveness of the nuclear-norm proximal map bounds the change on restoring B_p by $\|B_p\|_F$, without a rank assumption on the bias. This comparison is analytical: the learner never subtracts B_p. Pairing cancels constant reward or score offsets.

C.3 Calibration–Exploration–Commit

Return to the nondecreasing link, the unknown product *base* law p, and its positive signal $\mu_* \geq \mu_{\min}$ for $\Theta_* \neq 0$; write $\Sigma_X = \mathrm{Cov}_p(\mathrm{vec}\,X)$, used only in the regret analysis. Choose deterministic integers $b, n \geq 1$ with $b + 2n \leq T$. During b uniform-selection calibration rounds, discard rewards and fit scores from all menus. For the next $2n$ rounds, choose uniform indices independently of menus and past, giving n disjoint iid pairs under p independent of calibration. Fit once by Theorem C.1, then commit greedily to this fixed fit for $H_T = T - b - 2n$ rounds, selecting uniformly if the fit is zero. Commit rewards are not reused.

For this subsection and Section C.7, let $e_{\mathrm{fit}}(n, \delta, J, \varepsilon)$ be the deterministic right-hand side of the norm bound in Theorem C.1 for the chosen fit, with J in place of $\widehat{J}$. It includes both the statistical term and its stated bias constant. Put $\Delta = 2L_f AB$. If calibration certifies $(\varepsilon_b, \widehat{J})$ with failure probability δ_{cal}, and the reward fit uses confidence δ_{rew}, then

$$\begin{aligned}\mathbb{E}R_T \leq \min\Big\{&\Delta T,\ \Delta(b+2n) + \Delta H_T(\delta_{\mathrm{cal}} + \delta_{\mathrm{rew}}) \\ &+ \frac{2L_f\sqrt{K\,\|\Sigma_X\|_{\mathrm{op}}}}{\mu_{\min}} H_T e_{\mathrm{fit}}(n, \delta_{\mathrm{rew}}, \widehat{J}, \varepsilon_b)\Big\}.\end{aligned} \tag{12}$$

The certificates are deterministic bounds valid on the calibration event.

To prove the bound, center a fresh menu by $\overline{X} = \mathbb{E}_p X$ in the transfer argument. This preserves fitted rankings and true index differences, giving

$$\mathbb{E}[r_t \mid \mathcal{F}_{t-1}] \leq \frac{2L_f}{\mu_{\min}}\sqrt{K\,\|\Sigma_X\|_{\mathrm{op}}}\left\|\widehat{\Theta} - \mu_*\Theta_*\right\|_F.$$

The transfer uses the nondecreasing link and positive base multiplier; no selected-law signal condition is needed because the commit-stage fit is not updated. The learner need not know or subtract $\overline{X}$. The first $b + 2n$ rounds cost at most Δ each. Conditioning on calibration and the reward fit bounds failure by $\delta_{\mathrm{cal}} + \delta_{\mathrm{rew}}$. Apply the norm bound on success and Δ on failure in each commit round; the deterministic cap ΔT supplies the minimum.

C.4 An Explicit Smooth Product Class

Fix known constants $0 < c \le 1 \le C$, $s > 0$, and $R \ge 0$. Let ϕ_k, $k = 0, 1, \ldots$, be the Jacobi $(4, 4)$ polynomials normalized to an orthonormal basis in $L^2(\omega)$ and positive at 1, where $\omega(u) = (315/256)(1 - u^2)^4 \mathbf{1}\{|u| < 1\}$. The unknown marginal j has density

$$p_j(x) = A^{-1}\omega(x/A)r_j(x/A), \quad c \le r_j \le C, \quad \int_{-1}^{1} \omega r_j = 1,$$

with $\log r_j$ locally absolutely continuous and

$$h_j = -(\log r_j)' = \sum_{k\ge 0} a_{jk}\phi_k \quad \text{in } L^2(\omega), \qquad \sum_{k\ge 0}(1+k)^{2s}a_{jk}^2 \le R^2. \tag{13}$$

Equation (13) specifies the approximation class; the positive lower bound is on the density ratio, not the density at the boundary. For $c < 1 < C$ and $R > 0$ the class is infinite-dimensional: sufficiently small, rapidly decaying coefficient sequences give $r(u) \propto \exp(-\int_0^u h(v)\,dv)$ with the required ratio bounds.

Put $s_{\mathrm{ref}}(u) = 8u/(1 - u^2)$. The jth coordinate of the full score S^p is the scalar function $s_j(x) = A^{-1}(s_{\mathrm{ref}}(x/A) + h_j(x/A))$. Known certificates are

$$\|\Sigma_X\|_{\mathrm{op}} \le \min(A^2, CA^2/11), \qquad \mathbb{E}s_j(x_j)^2 \le \frac{C}{A^2}(\sqrt{12} + R)^2.$$

Reference moments, Minkowski's inequality and independence give these bounds and the joint score certificate $M = C(\sqrt{12} + R)^2/A^2$. The ratio bound makes the boundary trace zero, and the score bound makes $p_j' = -s_j p_j$ integrable. Thus the zero extension is absolutely continuous, the score is centered and Stein integration by parts is valid, including for distinct marginals with nonzero means.

Known basis bounds. Writing $P_k = P_k^{(4,4)}$, the Rodrigues norm and endpoint value give

$$\int P_k^2\omega = \frac{630}{2k+9}\frac{(k+4)!^2}{k!(k+8)!}, \qquad \|\phi_k\|_\infty^2 = \frac{(2k+9)\prod_{i=1}^{8}(k+i)}{9!}.$$

The endpoint maximum follows from the beta-moment identity

$$\frac{P_k^{(a,a)}(u)}{P_k^{(a,a)}(1)} = \mathbb{E}(u + i\sqrt{1-u^2}Z_a)^k,$$

where Z_a has density proportional to $(1 - z^2)^{a-1/2}$ for $a > 0$. The complex quantity has modulus at most one. Using the derivative identity for Jacobi polynomials then gives $\|\phi_k'\|_\infty \le k(k + 9)\|\phi_k\|_\infty/10$. For a dictionary size $d_{\mathrm{cal}} \ge 1$, set $H_{d_{\mathrm{cal}}} = \max_{k<d_{\mathrm{cal}}}\|\phi_k\|_\infty$; thus

$$H_{d_{\mathrm{cal}}}^2 \le d_{\mathrm{cal}}^9, \qquad \max_{k<d_{\mathrm{cal}}}\|\phi_k'\|_\infty \le d_{\mathrm{cal}}^2 H_{d_{\mathrm{cal}}}.$$

The dictionary size is distinct from the matrix side length d and the lower-bound dimension $s_{\max}$ in Section 3.4.

C.5 A Finite Calibration Algorithm and Its Proof

For one marginal, write $U_0 = X/A$, $\Phi = (\phi_0, \ldots, \phi_{d_{\mathrm{cal}}-1})^\top$, and define known moment integrands

$$G = \mathbb{E}[\Phi(U_0)\Phi(U_0)^\top], \qquad (\boldsymbol{b})_a = \mathbb{E}[\phi_a'(U_0) - s_{\mathrm{ref}}(U_0)\phi_a(U_0)].$$

Integration of $(\phi_a\omega r)'$ shows $\boldsymbol{b} = \mathbb{E}[h(U_0)\Phi(U_0)]$, while $cI \preceq G \preceq CI$. Therefore $\beta^* = G^{-1}\boldsymbol{b}$ is the $L^2(\omega r)$ projection coefficient of h. With $Q = \sqrt{C/c}\,R$,

$$\|\beta^*\|_2 \le Q, \qquad \left\|h - \Phi^\top\beta^*\right\|_{L^2(\omega r)} \le \sqrt{C}R\,d_{\mathrm{cal}}^{-s}.$$

Projection and the lower Gram eigenvalue give the first bound; comparison with the first d_{cal} reference coefficients gives the second.

Given m_{cal} iid scalar samples, choose an odd k_{blk} with $m_{\mathrm{cal}} \geq 2k_{\mathrm{blk}}$ and split the first $k_{\mathrm{blk}}\lfloor m_{\mathrm{cal}}/k_{\mathrm{blk}} \rfloor$ samples into equal blocks. Estimate every upper-triangular entry of G and every coordinate of $\boldsymbol{b}$ by the median of its block means, then reflect the Gram entries to obtain a symmetric $\widehat{G}$. If $\lambda_{\min}(\widehat{G}) \geq c/2$, set $\widetilde{\beta} = \widehat{G}^{-1}\widehat{\boldsymbol{b}}$; otherwise set it to zero. Project onto the Euclidean ball of radius Q to obtain $\widehat{\beta}$, and return

$$\widehat{s}(x) = A^{-1}\big(s_{\mathrm{ref}}(x/A) + \Phi(x/A)^{\top}\widehat{\beta}\big).$$

The eigenvalue check is necessary because entrywise medians need not preserve positive semidefiniteness. This score corresponds to a genuine density proportional to

$$\omega(u)\exp\left(-\sum_{a<d_{\mathrm{cal}}}\widehat{\beta}_a\int_0^u \phi_a(v)\,dv\right).$$

Its normalizing constant is not needed for scoring.

Theorem C.2 (Finite-sample calibration) *Fit J_{marg} marginals, with m_{cal} iid samples for each. For odd $k_{\mathrm{blk}} \geq 8\log(2J_{\mathrm{marg}}(d_{\mathrm{cal}}+1)^2/\delta_{\mathrm{cal}})$, assume*

$$m_{\mathrm{cal}} \geq (32C/c^2)k_{\mathrm{blk}}d_{\mathrm{cal}}^{11}.$$

Then, with probability at least $1-\delta_{\mathrm{cal}}$, simultaneously,

$$\begin{aligned}\|\widehat{s}_j - s_j\|_{L^2(p_j)} &\leq e_{m_{\mathrm{cal}},d_{\mathrm{cal}}}\\ &:= \frac{1}{A}\left[\sqrt{C}Rd_{\mathrm{cal}}^{-s} + \frac{2C}{c}\sqrt{\frac{8k_{\mathrm{blk}}}{m_{\mathrm{cal}}}}H_{d_{\mathrm{cal}}}\left\{\sqrt{2d_{\mathrm{cal}}(d_{\mathrm{cal}}^4+12)} + d_{\mathrm{cal}}Q\right\}\right] \qquad (14)\\ &\leq \frac{1}{A}\left[\sqrt{C}Rd_{\mathrm{cal}}^{-s} + Dd_{\mathrm{cal}}^{7}\sqrt{k_{\mathrm{blk}}/m_{\mathrm{cal}}}\right], \qquad D = \frac{2C}{c}\sqrt{8}(\sqrt{26}+Q).\end{aligned}$$

On every calibration history, including failures,

$$\mathbb{E}_{p_j}\widehat{s}_j(x_j)^2 \leq M_{\mathrm{fit}} := \frac{C}{A^2}(\sqrt{12}+Q)^2. \qquad (15)$$

On the same calibration event of probability at least $1-\delta_{\mathrm{cal}}$, stacking the fitted scalar functions into $\widehat{S}$ gives, for a fresh product array, Equation (11) *with $\varepsilon = e_{m_{\mathrm{cal}},d_{\mathrm{cal}}}$, $\widehat{J} = M_{\mathrm{fit}}$.*

Proof. A block mean of a variable with variance at most v has variance at most $2vk_{\mathrm{blk}}/m_{\mathrm{cal}}$. Chebyshev bounds the probability of error greater than $\sqrt{8vk_{\mathrm{blk}}/m_{\mathrm{cal}}}$ by $1/4$. The median exceeds that error only if at least half the independent blocks are bad, with probability at most $e^{-k_{\mathrm{blk}}/8}$. Union over at most $J_{\mathrm{marg}}(d_{\mathrm{cal}}+1)^2$ moment estimates gives simultaneous accuracy.

For $a, b < d_{\mathrm{cal}}$, the needed integrand bounds are

$$\mathbb{E}_{\omega r}(\phi_a\phi_b)^2 \leq CH_{d_{\mathrm{cal}}}^2, \qquad \mathbb{E}_{\omega r}(\phi_a' - s_{\mathrm{ref}}\phi_a)^2 \leq 2CH_{d_{\mathrm{cal}}}^2(d_{\mathrm{cal}}^4+12).$$

The singular reference score is handled by $\mathbb{E}_{\omega}s_{\mathrm{ref}}^2 = 12$, not a nonexistent supremum bound. Put $t_m = 8k_{\mathrm{blk}}/m_{\mathrm{cal}}$. On the simultaneous event,

$$\begin{aligned}\left\|\widehat{G} - G\right\|_{\mathrm{op}} &\leq \sqrt{t_m}\,d_{\mathrm{cal}}\sqrt{C}H_{d_{\mathrm{cal}}},\\ \left\|\widehat{\boldsymbol{b}} - \boldsymbol{b}\right\|_2 &\leq \sqrt{2t_mCd_{\mathrm{cal}}H_{d_{\mathrm{cal}}}^2(d_{\mathrm{cal}}^4+12)}.\end{aligned}$$

The sample floor makes the first bound at most $c/2$, so the inverse branch is used, with $\left\|\widetilde{\beta} - \beta^*\right\|_2 \leq (2/c)(\left\|\widehat{\boldsymbol{b}} - \boldsymbol{b}\right\|_2 + Q\left\|\widehat{G} - G\right\|_{\mathrm{op}})$. Projection onto the radius-Q ball cannot increase the error. Multiplying by $\sqrt{C}/A$ and adding the approximation error proves the first bound in Equation (14). The inequalities $H_{d_{\mathrm{cal}}} \leq d_{\mathrm{cal}}^{9/2}$ and $d_{\mathrm{cal}}^4 + 12 \leq 13d_{\mathrm{cal}}^4$ give the second. On every history, $\left\|\widehat{\beta}\right\|_2 \leq Q$ proves Equation (15) by reference orthonormality and Minkowski.

Conditional on calibration, each fitted score and error coordinate is a fixed function of its own fresh independent coordinate. Their covariance matrices are therefore diagonal even when fits share training data; the scalar squared-error and fitted second-moment bounds control the diagonals, proving the joint certificates without a factor N.

C.6 Allocation and the Resulting Rates

For distinct marginal laws use $J_{\mathrm{marg}} = N$, $m_{\mathrm{cal}} = Kb$. Only when all coordinates have an explicitly common marginal may they be pooled, giving $J_{\mathrm{marg}} = 1$, $m_{\mathrm{cal}} = NKb$. Write $m_{\mathrm{cal}} = c_{\mathrm{samp}} b$, with $c_{\mathrm{samp}} = K$ or NK respectively. Product structure alone does not justify the latter sample count.

Set $\delta_{\mathrm{cal}} = \delta_{\mathrm{rew}} = (T+1)^{-2}$, $d_{\max} = \max(1, \lceil c_{\mathrm{samp}} T \rceil)$, and let k_T be the smallest positive odd integer at least $8 \log(2 J_{\mathrm{marg}} (d_{\max}+1)^2 / \delta_{\mathrm{cal}})$. Define

$$\alpha = \frac{s}{2s+14}, \qquad \rho_0 = \max\{2, (32C/c^2)^{(2s+14)/(2s+3)}\},$$
$$b_0 = \lceil k_T \rho_0 / c_{\mathrm{samp}} \rceil, \quad E_s = 2^s \sqrt{C} R + D.$$

For $b_0 \le b \le T$, choose $d_{\mathrm{cal}} = \lfloor (c_{\mathrm{samp}} b / k_T)^{1/(2s+14)} \rfloor$ and $k_{\mathrm{blk}} = k_T$. Then

$$\varepsilon_b \le \frac{E_s}{A} \left(\frac{k_T}{c_{\mathrm{samp}} b} \right)^{\alpha}, \qquad \widehat{J} = M_{\mathrm{fit}}. \tag{16}$$

Indeed, $z = c_{\mathrm{samp}} b / k_T \ge \rho_0$ implies $z \ge (32C/c^2) z^{11/(2s+14)} \ge (32C/c^2) d_{\mathrm{cal}}^{11}$ and $z \ge 2$. Thus the sample floor and block requirements hold. Using $\lfloor u \rfloor \ge u/2$ for $u \ge 1$ in Equation (14) proves the bound, hence $\varepsilon_b = \widetilde{O}(b^{-s/(2s+14)})$ for fixed class quantities.

For $T \ge 8$, take

$$n = \lfloor T^{2/3}/4 \rfloor, \qquad b = \max\{b_0, \lfloor T^{1/(1+\alpha)}/4 \rfloor\}.$$

Use this allocation if $b + 2n \le T$; otherwise select uniformly throughout. For $T < 8$ also select uniformly. For fixed class and model constants, $b_0 = O(\log(T+1))$, so the allocation eventually fits. Substituting Equations (16) and (15) into Equation (12) gives exploration costs, a statistical term of order $Tn^{-1/2}$, and a bias term of order $Tb^{-\alpha}$, with the explicit prefactors in Equation (12). The failure cost is at most $2\Delta T/(T+1)^2$. This proves Equation (10). The fallback has the linear cap without invoking a calibration certificate whose sample floor fails.

The exponent 14 comes from squaring the conservative d_{cal}^7 estimation bound; it is not asserted optimal. For example, $s = 1$ gives $T^{16/17}$. The bias has no additional rank multiplier, while the statistical terms retain their matrix or transformed-slice dimensions.

Exactly specified finite-dimensional families. If every h_j belongs exactly to the known span of the first d_0 basis functions, with fixed known d_0, take $d_{\mathrm{cal}} = d_0$. The approximation term is then zero, and the same algorithm gives $\varepsilon_b = O(\sqrt{\log(NT)/(c_{\mathrm{samp}} b)})$, subject to the fixed-$d_0$ sample floor and displayed constants. Taking b, n of order $T^{2/3}$ proves $\widetilde{O}(T^{2/3})$ regret. Choosing to fit d_0 coefficients does not itself establish exact inclusion.

For example, $p_{\beta_j,\lambda_j}(x) \propto p_A(x) \exp(\beta_j (x/A)^2/2 + \lambda_j x/A)$ has $h_j(u) = -\beta_j u - \lambda_j$ in the span of $\phi_0 = 1$, $\phi_1 = \sqrt{11}u$. Known bounds $|\beta_j| \le H_0$, $|\lambda_j| \le H_1$ give $C = c^{-1} = \exp(H_0/2 + 2H_1)$ and $R^2 = H_1^2 + 2^{2s} H_0^2/11$. Both coefficients are fitted by the specified two-dimensional system. If only the linear tilt is unknown and β_j is known, its constant score contribution cancels in uniform pairs, so calibration can be omitted. This cancellation does not identify the greedy projection distribution.

C.7 Conditional Return to Greedy Epochs

Retain the nondecreasing-link assumption for the greedy regret analysis. Consider deterministic planned epoch lengths and deterministic scored-pair counts $n_i \ge 1$. Freeze the direction throughout epoch i, including calibration, and construct $\widehat{S}_i$ before collecting those fresh action–reward pairs from the selected law p_i. Conditional on the post-calibration history, these observations are iid and reuse no calibration observations. The completed fit determines the next epoch's direction. Assume Stein regularity and, for $\Theta_* \ne 0$, $\mu_i = \mathbb{E}_{p_i} f'(\langle X, \Theta_* \rangle) \ge \mu_{\min} > 0$. This selected-law condition is additional to the base-law condition in Sections C.3–C.6.

Suppose deterministic certificates $(\varepsilon_i, \widehat{J}_i)$ satisfy Equation (11) under p_i, except on a calibration event of marginal probability at most β_i, and use conditional reward-fit failure probability δ_i. Recall e_{fit} from Section C.3, including its score-bias term, and put $\Delta = 2 L_f A B$. If h_0 is the initial planned

length and h_{i+1} the actual following epoch length, including calibration and a partial final epoch, then

$$\mathbb{E}R_T \le \Delta \min\{T, h_0\} + \frac{2L_f\sqrt{K\,\|\Sigma_X\|_{\mathrm{op}}}}{\mu_{\min}} \sum_i h_{i+1} e_{\mathrm{fit}}(n_i, \delta_i, \widehat{J}_i, \varepsilon_i) + \Delta \sum_i h_{i+1}(\beta_i + \delta_i).$$

Here Σ_X is the covariance of the base menu law. The sums concern completed fits used in a following epoch. The bound may also be capped by ΔT; $\Theta_* = 0$ has zero regret. To prove it, condition on successful calibration and apply Theorem C.1. The fit event is measurable before the following epoch; Lemma F.1 controls regret on success using the preceding target $\mu_i\Theta_*$, and Δ controls failure.

For doubling epochs, a fixed fraction of each completed epoch allocated to scored pairs, uniformly bounded $\widehat{J}_i$, and $\beta_i, \delta_i \le (T+1)^{-2}$, the statistical term is $\widetilde{O}(\sqrt{T})$ for fixed structural and model quantities. A separately established certificate $\varepsilon_i = O(n_i^{-\alpha})$, $0 < \alpha < 1$, adds $O(T^{1-\alpha})$; $\alpha = 1/2$ preserves the square-root horizon rate up to logarithms. These certificates must cover the adaptively chosen direction through a uniform argument or independent calibration after freezing it. The base-score construction of Sections C.3–C.6 does not itself certify the projection correction F_i'/F_i or the fitted selected-score covariance.

D Structured Single-Index Estimation Proofs

We prove the paired-moment identity and Theorem 3.1, first for matrices and then for transformed slices. The result applies conditionally to each fixed-policy T-ESTOR batch with $n = n_i$, $w = MG_K(d)$, and $\mu_* = \mu_i$; the fit from completed epoch i is used in epoch $i+1$.

D.1 Paired Moments

Lemma D.1 (Paired mean and second moment) *Let p be a fixed known observation density, supported in $\|X\|_{\max} < A$, whose zero extension belongs to $W^{1,1}(\mathbb{R}^N)$. Write $S^p = -\nabla_X \log p$ and assume its score has a finite joint second moment. Take $2n$ independent action–reward observations with this density, the budget of Section 2.2, and the link condition of Assumption 2.2 and the conditional noise bounds of Section 2.1, where $n \ge 1$. The n paired statistics in Equation (3) are iid. For $\Theta_* \ne 0$, put $\mu_* = \mathbb{E}_p f'(\langle X, \Theta_*\rangle) \in [-L_f, L_f]$. For $\Theta_* = 0$, set $\mu_* = 0$ without evaluating $f'(0)$. With $V = 2L_f^2A^2B^2 + \sigma^2$,*

$$\mathbb{E}Z_j = \mu_*\Theta_*,$$
$$\mathbb{E}[\mathrm{vec}(Z_j)\,\mathrm{vec}(Z_j)^\top] \preceq V\,\mathbb{E}_p[\mathrm{vec}(S^p)\,\mathrm{vec}(S^p)^\top].$$

For every fixed linear map $\mathcal{L}$ from arrays to matrices,

$$\mathbb{E}[\mathcal{L}(Z_j)\mathcal{L}(Z_j)^\top] \preceq V\mathbb{E}[\mathcal{L}(S^p)\mathcal{L}(S^p)^\top],$$
$$\mathbb{E}[\mathcal{L}(Z_j)^\top\mathcal{L}(Z_j)] \preceq V\mathbb{E}[\mathcal{L}(S^p)^\top\mathcal{L}(S^p)].$$

These conclusions also hold conditionally on a history fixing the law and score, including each frozen T-ESTOR epoch with $p = p_i$.

For the appendix proofs, let $J_{\mathrm{obs}} > 0$ bound the joint score moment by $J_{\mathrm{obs}}I_N$. The second-moment bound $w > 0$ from Section 3.1 bounds both left and right score moments by w times their identity matrices. The latter applies to the matrix score or, for the supplied transform, each transformed slice of the full score. These bounds need not be equal or minimal. Appendix E.3 verifies them for $p = p_i$ with $w = MG_K(d)$ and $J_{\mathrm{obs}} = \overline{J} = MG_K(1)$, including the uniform fallback.

Proof of Lemma D.1. The weak integration-by-parts argument of Appendix B.1 applies verbatim to p: $\mathbb{E}S^p = 0$ and $\mathbb{E}[yS^p(X)] = \mu_*\Theta_*$. Its bounded support gives the cutoff, and its finite score second moment ensures integrability. For independent observations $(X_1, y_1), (X_2, y_2)$, put

$S_j = S^p(X_j)$. Expanding $(y_1 - y_2)(S_1 - S_2)/2$ gives $\mathbb{E}Z = \mathbb{E}[yS] - \mathbb{E}y\,\mathbb{E}S = \mu_*\Theta_*$. Conditional independence of the noise draws gives

$$\mathbb{E}[(y_1 - y_2)^2 \mid X_1, X_2] = (f(\langle X_1, \Theta_* \rangle) - f(\langle X_2, \Theta_* \rangle))^2 + \mathbb{E}[\eta_1^2 \mid X_1] + \mathbb{E}[\eta_2^2 \mid X_2] \le 2V.$$

For any unit array D this implies $\mathbb{E}\langle Z, D\rangle^2 \le (V/2)\mathbb{E}\langle S_1 - S_2, D\rangle^2 = V\mathbb{E}\langle S, D\rangle^2$, proving the joint PSD bound. For a matrix-valued linear map, apply the same conditional inequality to $(\mathcal{L}(S_1) - \mathcal{L}(S_2))(\mathcal{L}(S_1) - \mathcal{L}(S_2))^\top$; score centering removes its cross terms. The right second moment follows by transposition. These bounds control raw second moments of Z, without subtracting its population target. Pairing cancels the unrestricted reward level $f(0)$. The nonzero-parameter index has a density under any such p; the zero-parameter case follows directly by centering. The moment statement requires neither monotonicity nor a nonzero Stein signal: μ_* may be negative or zero.

D.2 Matrix Estimation Proofs

For the matrix case of Theorem 3.1, set $q = 1$, $U = [1]$, and $r_U = r$ in Equation (4). Its tuning specializes to

$$\nu = \sqrt{\frac{2\log(2(d_1 + d_2)/\delta)}{nVw}}, \qquad \lambda = 2\sqrt{\frac{2Vw\log(2(d_1 + d_2)/\delta)}{n}}. \tag{17}$$

We prove, with probability at least $1 - \delta$,

$$\left\|\widehat{Z}_\nu - \mu_*\Theta_*\right\|_{\mathrm{op}} \le \lambda/2, \qquad \left\|\widehat{\Theta} - \mu_*\Theta_*\right\|_F^2 \le \frac{16rVw\log(2(d_1 + d_2)/\delta)}{n}. \tag{18}$$

The bound may use the true rank, but does not justify substituting a fitted rank into a confidence bound or assert exact rank recovery.

Dilation and concentration. The rectangular influence from Section 3.1 is equivalently expressed by the symmetric dilation

$$\mathcal{H}(H) = \begin{pmatrix} 0 & H \\ H^\top & 0 \end{pmatrix}, \qquad \widetilde{\psi}_\nu(H) = \nu^{-1}[\psi(\nu\mathcal{H}(H))]_{12}. \tag{19}$$

The subscript 12 selects the upper-right block, and the symmetric matrix function applies ψ to eigenvalues. Thus the robust average is $\widehat{Z}_\nu = n^{-1}\sum_{j=1}^n \widetilde{\psi}_\nu(Z_j)$; the factor ν^{-1} is already part of the scaled influence map. The scalar influence function defined in Section 3.1 is odd and satisfies

$$-\log(1 - x + x^2/2) \le \psi(x) \le \log(1 + x + x^2/2). \tag{20}$$

For $x \ge 0$ the upper inequality is equality, and the lower follows from $(1+x+x^2/2)(1-x+x^2/2) = 1 + x^4/4 \ge 1$; oddness gives $x < 0$. In an SVD basis the dilation has eigenvalues $\pm s_\ell$, which proves its off-diagonal influence formula.

For n independent self-adjoint $d_0 \times d_0$ matrices A_j with common mean and $\left\|\mathbb{E}A_j^2\right\|_{\mathrm{op}} \le v_0$, the matrix influence inequality of Minsker (2018) gives

$$\mathbb{P}\left\{\left\|\frac{1}{n\nu}\sum_{j=1}^n \psi(\nu A_j) - \mathbb{E}A_1\right\|_{\mathrm{op}} > \varepsilon\right\} \le 2d_0\exp(-n\nu\varepsilon + n\nu^2 v_0/2). \tag{21}$$

Indeed, functional calculus applies Equation (20) to each matrix. The trace exponential bound for independent self-adjoint sums controls the upper tail by $d_0\exp(-n\nu\varepsilon + n\nu^2 v_0/2)$; applying it to the negative sum gives the other tail.

Take $A_j = \mathcal{H}(Z_j)$, where $\mathcal{H}$ denotes the symmetric dilation in Equation (19). Lemma D.1 gives $v_0 = Vw$ and $d_0 = d_1 + d_2$. With $\varepsilon = \sqrt{2v_0\log(2(d_1 + d_2)/\delta)/n}$ and the stated ν, Equation (21) gives the operator bound in Equation (18).

For completeness, let $\widehat{Z}_\nu$ have operator error at most ε from a target $\mu_*\Theta_*$ of rank at most r, and soft threshold it at ε. The $(r+1)$st singular value of $\widehat{Z}_\nu$ is at most ε, so $\mathrm{rank}(\widehat{\Theta}) \le r$ on this

event. Soft thresholding changes a matrix by operator norm at most ε, hence $\left\|\widehat{\Theta} - \mu_* \Theta_*\right\|_{\mathrm{op}} \le 2\varepsilon$. The difference has rank at most $2r$, giving $\left\|\widehat{\Theta} - \mu_* \Theta_*\right\|_F^2 \le 8r\varepsilon^2$. Substitution proves the constant 16 in Equation (18). This proof requires neither a lower bound on the nonzero singular values nor knowledge of r in the tuning.

Arithmetic cost. Conditional on score access, the robust moment uses one thin SVD per paired $d_1 \times d_2$ matrix, followed by one final SVD for soft thresholding. The dense arithmetic cost is $O((n+1)d_1 d_2 \min(d_1, d_2))$, with $O(d_1 d_2)$ accumulation storage apart from the SVD workspace. This count does not include computing the score or claim exact finite-bit evaluation of the logarithmic influence.

D.3 Transformed-Tensor Estimation Proofs

For clarity, define the per-pair transformed influence map

$$\Psi_\nu^U(Z) = \mathcal{T}_U^{-1}\left(\left[\widetilde{\psi}_\nu\big((\mathcal{T}_U Z)^{(k)}\big)\right]_{k=1}^q\right).$$

The moment in Section 3.1 is equivalently $\widehat{Z}_\nu = n^{-1}\sum_{j=1}^n \Psi_\nu^U(Z_j)$; linearity of $\mathcal{T}_U$ gives $(\mathcal{T}_U \widehat{Z}_\nu)^{(k)} = \widehat{Z}_{\nu,k}$ for every k.

For the transformed-slice case of Theorem 3.1, Equation (4) uses

$$\nu = \sqrt{\frac{2\log(2q(d_1+d_2)/\delta)}{nVw}}, \qquad \lambda = 2\sqrt{\frac{2Vw\log(2q(d_1+d_2)/\delta)}{n}}. \tag{22}$$

We prove the simultaneous bounds, with probability at least $1-\delta$,

$$\begin{aligned} \max_{k\in[q]}\left\|(\mathcal{T}_U(\widehat{Z}_\nu - \mu_*\Theta_*))^{(k)}\right\|_{\mathrm{op}} &\le \lambda/2, \\ \left\|\widehat{\Theta} - \mu_*\Theta_*\right\|_F^2 &\le \frac{16 V r_U w \log(2q(d_1+d_2)/\delta)}{n}. \end{aligned} \tag{23}$$

The dual of the sum nuclear norm is the maximum slice operator norm, because $\mathcal{T}_U$ preserves inner products. Its Frobenius isometry also gives the square-completion identity

$$\|Z\|_F^2 - 2\left\langle Z, \widehat{Z}_\nu\right\rangle = \sum_{k=1}^q \left\|(\mathcal{T}_U Z)^{(k)} - \widehat{Z}_{\nu,k}\right\|_F^2 - \left\|\widehat{Z}_\nu\right\|_F^2.$$

The last term is independent of Z, so adding the sum nuclear penalty separates the objective into slice problems with quadratic coefficient one. Each problem has solution $\mathrm{SVT}_{\lambda/2}(\widehat{Z}_{\nu,k})$; the inverse transform then gives the estimator in Section 3.1. Define the block-diagonal dilation

$$\mathcal{H}_U(Z) = \mathrm{diag}_{k=1}^q \mathcal{H}((\mathcal{T}_U Z)^{(k)}).$$

Its dimension is $q(d_1+d_2)$ and the operator norm of its raw second moment is at most Vw. Applying Equation (21) to this one block matrix gives the simultaneous dual bound in Equation (23); independence between its blocks is not required. Apply the soft-thresholding argument of Appendix D.2 in slice k with rank at most ρ_k, then sum squared errors. Frobenius isometry proves Equation (23).

One can accumulate influenced slices in transformed coordinates and invert only after fitting. Dense multiplication by U and q thin SVDs cost $O(d_1 d_2 q^2 + q d_1 d_2 \min(d_1, d_2))$ per pair; the final fit uses q more SVDs and one inverse transform. This is an arithmetic-operation statement conditional on scores, not an exact finite-bit claim for transcendental function evaluation. For $q = 1$, $U = [1]$ and $r_U = r$, both the estimator and tuning reduce to the matrix case.

Epoch buffering and streaming. Algorithm 1 literally retains an epoch buffer, using $O(n_i N)$ observation storage. Its fixed epoch length and tuning also permit an equivalent streaming implementation: retain one pending observation, form each disjoint pair when the next arrives, and add its influenced statistic to the running matrix or slice sums. At the epoch end, divide by the prescribed pair count n_i and apply the same threshold. Observation and accumulator storage is then $O(N)$, apart from score-oracle and linear-algebra workspace. This equivalence concerns the theoretical policy with tuning fixed before sampling; it does not automatically apply to the adaptive experimental procedure, which recomputes tuning from stored data at checkpoints.

E T-ESTOR WITH STRUCTURED UPDATES

We verify the conditional sampling model, selected-score identities, and moment bounds used in Section 3.2.

E.1 CONDITIONAL PROBABILITY MODEL

Include the learner's private random seed in the history. Conditional on each epoch-start history, the direction and score rule are fixed. Independent menus and fresh noise from the fixed conditional kernel used in Section 3.2 then give iid selected action–reward observations. Concentration holds after every such history. Within-epoch reward-dependent policy changes would invalidate this reduction and are not part of Algorithm 1.

E.2 PROJECTION LAWS AND SELECTED SCORES

Fix any epoch-start history and any unit-Frobenius direction v_i; the following argument is uniform over all such directions. Choose an orthonormal complement Q to $\mathrm{vec}(v_i)$ and write $x = u\,\mathrm{vec}(v_i) + Qz$. If F_i is the CDF of $\langle v_i, X\rangle$ under p, orthogonal change of variables gives its density and weak derivative:

$$F_i'(u) = \int p(u\,\mathrm{vec}(v_i) + Qz)\,dz,$$
$$F_i''(u) = \int \mathrm{vec}(v_i)^\top \nabla p(u\,\mathrm{vec}(v_i) + Qz)\,dz.$$

Fubini and the zero-extension regularity $p \in W^{1,1}$ in Assumption 2.1 imply $F_i' \in W^{1,1}(\mathbb{R})$; use its continuous, absolutely continuous representative. The weak score identity implies, for almost every u with $F_i'(u) > 0$,

$$-\frac{F_i''(u)}{F_i'(u)} = \mathbb{E}_p[\langle v_i, S^p(X)\rangle \mid \langle v_i, X\rangle = u].$$

Jensen's inequality and the full joint score certificate give $\int_{\{F_i'>0\}} (F_i'')^2/F_i' \le M$, and Cauchy–Schwarz gives $\int |F_i''| \le \sqrt{M}$. The density is supported on $[-A\,\|v_i\|_{\mathrm{ent},1}, A\,\|v_i\|_{\mathrm{ent},1}]$, so it vanishes outside this interval. An absolutely continuous nonnegative density must rise from and return to zero; hence

$$\|F_i'\|_\infty \le \frac{1}{2}\int |F_i''| \le \frac{\sqrt{M}}{2}. \tag{24}$$

No strict positivity throughout that interval is required.

For a nonzero vectorized direction $v \in \mathbb{R}^N$, let F_v be the CDF of $\langle v, \mathrm{vec}(X)\rangle$. Rescaling Equation (24) gives

$$\|F_v'\|_\infty \le \frac{\sqrt{M}}{2\,\|v\|_2}, \qquad \mathbb{E}_p\big[F_v'(\langle v, \mathrm{vec}(X)\rangle)^2\big] \le \frac{M}{4\,\|v\|_2^2} \le \frac{M}{4} \quad \text{if } \|v\|_\infty = 1.$$

Thus our joint-score assumption implies the projection-density moment condition in Assumption 3.3 of Kang et al. (2026). For dependent laws, the joint certificate is stronger than coordinatewise score-moment control.

The selected density in Equation (6) is $p_i(X) = Kp(X)F_i(\langle v_i, X\rangle)^{K-1}$. Since F_i' is bounded, the CDF factor is a bounded Lipschitz multiplier, so the zero extension of p_i remains in $W^{1,1}$. Its score is

$$S^{p_i}(X) = S^p(X) - (K-1)\alpha_i(\langle v_i, X\rangle)v_i, \qquad \alpha_i(u) = \frac{F_i'(u)}{F_i(u)} \quad \text{on } \{F_i > 0\}.$$

Set $\alpha_i = 0$ on $\{F_i = 0\}$, a set of zero probability under the selected projection law. Thus the formula is defined p_i-almost surely, including when the support has gaps. Applying the weak integration-by-parts argument of Appendix B.1 gives

$$\mathbb{E}_{p_i} S^{p_i} = 0, \qquad \mathbb{E}_{p_i}[yS^{p_i}(X)] = \mu_i \Theta_*, \qquad \mu_i = \mathbb{E}_{p_i}[f'(\langle X, \Theta_*\rangle)]$$

for $\Theta_* \neq 0$. The selected law is absolutely continuous, and its score is square integrable by the bounds below, so the same index and integrability arguments apply. For the uniform fallback $v_i = 0$,

set $p_i = p$ and $S^{p_i} = S^p$; no projection density is needed. For a zero parameter the moment is zero directly without evaluating $f'(0)$.

E.3 Uniform Selected-Score Moments

For a unit direction, the selected projection has density $KF_i'(u)F_i(u)^{K-1}$. For $K \geq 3$, Equation (24) and the probability integral transform give

$$\begin{aligned}\mathbb{E}_{p_i}\alpha_i(\langle v_i, X\rangle)^2 &= K\int_{\{F_i>0\}} F_i'(u)^3 F_i(u)^{K-3}\,du \\ &\leq \frac{KM}{4}\int F_i'(u)F_i(u)^{K-3}\,du = \frac{KM}{4(K-2)}.\end{aligned}$$

This substitution is valid for any continuous CDF, including one with flat segments. At $K = 3$, the power F_i^{K-3} is simply one; this is why we require $K \geq 3$. Since $p_i \leq Kp$, the base-score contribution in any unit array direction is at most KM. Applying $(a-b)^2 \leq 2a^2 + 2b^2$ to the score formula gives

$$\mathbb{E}_{p_i}[\mathrm{vec}(S^{p_i})\,\mathrm{vec}(S^{p_i})^\top] \preceq \overline{J} I_N, \qquad \overline{J} = MG_K(1), \qquad G_K(u) = 2Ku + \frac{K(K-1)^2}{2(K-2)}.$$

For a $d_1 \times d_2$ base score and a unit $a \in \mathbb{R}^{d_1}$,

$$a^\top \mathbb{E}_p[S^p(S^p)^\top]a = \sum_{b=1}^{d_2} \mathbb{E}_p\left\langle S^p, ae_b^\top\right\rangle^2 \leq d_2 M.$$

Thus the left and right base moments are bounded by $d_2 M I_{d_1}$ and $d_1 M I_{d_2}$, respectively. Since $v_i v_i^\top \preceq I_{d_1}$ and $v_i^\top v_i \preceq I_{d_2}$, the PSD inequality $(C-D)(C-D)^\top \preceq 2CC^\top + 2DD^\top$ gives selected left and right moment bounds wI_{d_1} and wI_{d_2}, where $d = \max(d_1, d_2)$ and $w = MG_K(d)$. These are certificates, not isotropic moment equalities.

For a transformed slice, put $B_k = (\mathcal{T}_U S^p)^{(k)}$. The joint base moment bound MI_N is preserved by the full orthogonal transform. Extracting a slice and applying the row and column calculation gives $\mathbb{E}B_k B_k^\top \preceq d_2 M I_{d_1}$ and $\mathbb{E}B_k^\top B_k \preceq d_1 M I_{d_2}$. Neither transformed entries nor slices need be independent. Moreover, $\left\|(\mathcal{T}_U v_i)^{(k)}\right\|_F \leq 1$. The same PSD argument slice by slice proves the same w bound, without a factor of q. Here B_k is a slice of the transformed full joint score, not a marginal slice score. Orthogonality also preserves the selected joint vector moment bound. For the uniform fallback, the smaller base moments are covered by all these certificates.

Why the selected-signal condition is separate. Strict monotonicity alone gives no uniform signal floor over the parameter class: for $f(u) = u^3$ the base multiplier is $3\mathbb{E}_p\langle X, \Theta_*\rangle^2$, equal to $(3A^2/11)\|\Theta_*\|_F^2$ under the beta design, and tends to zero along nonzero parameters.

Menu size can also affect the signal. For $\Theta_* \neq 0$ and $f(u) = L_f \max\{u, 0\}$, selection in direction $v_i = -\Theta_*/\|\Theta_*\|_F$ chooses the minimum true index among the K candidates. Its derivative is nonzero only when all K indices are positive, so

$$\mu_i = L_f\big[\Pr_p\{\langle X, \Theta_*\rangle > 0\}\big]^K.$$

The index has a density, so the derivative's value at zero is immaterial under our a.e. convention. If the admissible candidate law is centrally symmetric about zero, this signal is $L_f 2^{-K}$. A uniform floor in Condition 3.2 can therefore be at most $L_f 2^{-K}$ in this example, making its possible dependence on K explicit.

F Regret Proofs

Under the assumptions and tuning of Theorem 3.3, the matrix bound with explicit constants is

$$\begin{aligned}\mathbb{E}R_T \leq 100\min\Bigg\{ & L_f B\sqrt{K\|\Sigma_X\|_{\mathrm{op}}}T, \\ & \frac{L_f}{\mu_{\min}}\sqrt{VK\|\Sigma_X\|_{\mathrm{op}}\,MrTG_K(d)\log(2(d_1+d_2)T)}\Bigg\}.\end{aligned} \tag{25}$$

For transformed slices, replace $r\log(2(d_1+d_2)T)$ by $r_U\log(2q(d_1+d_2)T)$. The proof below also gives the smaller constant 72. The relation $G_K(d) \asymp K(d+K)$ for $K \geq 3$ gives the compact main-text bound with a universal constant.

Lemma F.1 (Fresh-menu transfer) *Under the distribution and link assumptions of Section 2.3 and the noise model of Section 2.1, with f nondecreasing as in Section 3.2, $\Theta_* \in \mathcal{B}(B)$, and Condition 3.2, let $\widehat{\Theta}$ and $\mu \geq \mu_{\min}$ be measurable before the current menu. Select greedily by $\widehat{\Theta}$, with uniform selection if it is zero. Then*

$$\mathbb{E}[r_t \mid \mathcal{F}_{t-1}] \leq \frac{2L_f}{\mu_{\min}}\sqrt{K\left\|\Sigma_X\right\|_{\mathrm{op}}}\left\|\widehat{\Theta}-\mu\Theta_*\right\|_F.$$

For every policy, including on histories where estimation failed, $\mathbb{E}[r_t \mid \mathcal{F}_{t-1}] \leq \overline{\Delta} := 2L_fB\sqrt{K\left\|\Sigma_X\right\|_{\mathrm{op}}}$. This is a conditional mean bound; the deterministic reward-gap bound is $\Delta = 2L_fAB$.

If $\Theta_* = 0$, all candidates have the same mean reward and regret is zero. Below assume $\Theta_* \neq 0$.

Fresh-menu transfer. Let $\overline{X} = \mathbb{E}_pX$ and $X^c_{t,a} = X_{t,a} - \overline{X}$. Subtracting the same array from every candidate preserves both fitted rankings and true index differences. Let a_* maximize the true index and put $X^c_t = X_t - \overline{X}$ for the chosen action. For $E = \widehat{\Theta} - \mu\Theta_*$, greedy optimality gives

$$\mu\left\langle X_{t,a_*} - X_t, \Theta_*\right\rangle \leq \left\langle X^c_t - X^c_{t,a_*}, E\right\rangle \leq 2\max_{a\leq K}\left|\left\langle X^c_{t,a}, E\right\rangle\right|.$$

If $\widehat{\Theta} = 0$, every index is a fitted maximizer, so the same statement holds for uniform selection. Nondecreasingness and Lipschitz continuity convert this into a reward gap. Conditional on the past, E is fixed and

$$\mathbb{E}\max_a\left|\left\langle X^c_{t,a}, E\right\rangle\right| \leq \sqrt{\sum_{a=1}^K \mathbb{E}\left\langle X^c_{t,a}, E\right\rangle^2} \leq \sqrt{K\left\|\Sigma_X\right\|_{\mathrm{op}}}\left\|E\right\|_F.$$

Since $\mu \geq \mu_{\min}$, this proves Lemma F.1. For an arbitrary policy, the reward gap is at most $2L_f\max_a\left|\left\langle X^c_{t,a}, \Theta_*\right\rangle\right|$; the same calculation with $\|\Theta_*\|_F \leq B$ gives the conditional mean cap $\overline{\Delta} = 2L_fB\sqrt{K\left\|\Sigma_X\right\|_{\mathrm{op}}}$. Entrywise budgets also give the deterministic cap $\Delta = 2L_fAB$. Centering serves only this proof. The algorithm uses the original actions and need not know or subtract $\overline{X}$; the model and budgets remain in those coordinates.

Consequences of the Stein assumptions. Write $\bar{x} = \mathbb{E}_px$. Applying weak integration by parts to $x_i - \bar{x}_i$, with a cutoff equal to one on the support, gives

$$\mathbb{E}_p[(x_i - \bar{x}_i)(\mathrm{vec}(S^p(X)))_i] = 1.$$

Cauchy–Schwarz, $\mathrm{Var}(x_i) \leq \mathbb{E}x_i^2 \leq A^2$, and $\mathbb{E}(\mathrm{vec}(S^p(X)))_i^2 \leq M$ imply $MA^2 \geq 1$. For every unit $u \in \mathbb{R}^N$, the same calculation gives $\mathbb{E}[(u^\top(x-\bar{x}))(u^\top\mathrm{vec}(S^p(X)))] = 1$ and hence $M\left\|\Sigma_X\right\|_{\mathrm{op}} \geq 1$. These inequalities follow from the general regularity and moment certificates; they impose no new assumptions.

Epoch summation and numerical constants. For epoch confidence $\delta_E = (T+1)^{-2}$, put $\ell_{\mathrm{M},E} = \log(2(d_1+d_2)(T+1)^2)$ and $\ell_{U,E} = \log(2q(d_1+d_2)(T+1)^2)$. Valid error coefficients are

$$a_{\mathrm{M},E} = 4\sqrt{VrMG_K(d)\ell_{\mathrm{M},E}}, \quad a_{U,E} = 4\sqrt{Vr_UMG_K(d)\ell_{U,E}}.$$

Each conditional estimation failure probability is at most δ_E after every start history, and hence also unconditionally. In epoch $i \geq 1$, the preceding estimate $\widehat{\Theta}_{i-1}$ and its target $\mu_{i-1}\Theta_*$ are measurable before the current menu. On success, the per-round conditional mean regret is at most $(2L_f/\mu_{\min})\sqrt{K\left\|\Sigma_X\right\|_{\mathrm{op}}}a/\sqrt{n_{i-1}}$; on failure use $\overline{\Delta}$. This avoids conditioning on the accuracy of a fit that uses the current or a future menu. If $\mathcal{B}_i$ is the failure event for the fit from epoch i, it is measurable before epoch $i+1$. Its contribution over the actual following length h_{i+1} is at most

$\overline{\Delta} h_{i+1} \mathbb{P}(\mathcal{B}_i)$. Summing these terms gives at most $\overline{\Delta} T \delta_E$, with no additional number-of-epochs factor.

An epoch $i \geq 1$ has planned length 2^{i+1} and uses 2^{i-1} pairs from its predecessor. Its final time is at most $2^{i+2} - 2 < 8 \cdot 2^{i-1}$. Thus $1/\sqrt{n_{\mathrm{prev}}(t)} \leq \sqrt{8/t}$ and summation gives

$$\mathbb{E}R_T \leq \overline{\Delta} \min(T, 2) + \frac{8\sqrt{2} L_f}{\mu_{\min}} \sqrt{K \left\|\Sigma_X\right\|_{\mathrm{op}}} \, a\sqrt{T} + \overline{\Delta} T \delta_E. \tag{26}$$

This also holds when the final epoch is incomplete. The same policy separately satisfies $\mathbb{E}R_T \leq \overline{\Delta} T$.

For the matrix or slice method, respectively, set $R = r$ or $R = r_U$; in either case $R \geq 1$. Let ℓ_E denote the corresponding $\ell_{\mathrm{M},E}$ or $\ell_{U,E}$, and define

$$W_T = (L_f/\mu_{\min}) \sqrt{VRK \left\|\Sigma_X\right\|_{\mathrm{op}} M G_K(d) \ell_E T}.$$

The leading term in Equation (26) is $32\sqrt{2} W_T$. Using $V \geq 2L_f^2 A^2 B^2$, $\mu_{\min} \leq L_f$, $MA^2 \geq 1$, $R \geq 1$, $G_K(d) \geq 8$ and $\ell_E \geq 1$ gives

$$\frac{\overline{\Delta}}{W_T} = \frac{2\mu_{\min} B}{\sqrt{VRMG_K(d)\ell_E T}} \leq \frac{\sqrt{2}}{\sqrt{MA^2 R G_K(d) \ell_E T}} \leq 1.$$

Initialization and failure add at most $(9/4)W_T$, since $T/(T+1)^2 \leq 1/4$. The total is at most $48 W_T$. For dilation dimension $d_0 \geq 2$, $\log(2d_0(T+1)^2) \leq 2\log(2d_0 T)$. Combining this with the separate linear cap gives the constant 72 in the matrix and slice cases of Equation (25).

Remark (sub-Gaussian menus). If the base law also satisfies the projection certificate (8), a conditional log-sum-exp bound gives

$$\mathbb{E}\left[\max_{a \leq K} |\, \langle X_{t,a} - m, E \rangle \,| \;\middle|\; \mathcal{F}_{t-1}\right] \leq A\kappa_X \sqrt{2\log(2K)} \, \|E\|_F$$

for every history-measurable E. This gives an alternative fresh-menu transfer coefficient to $\sqrt{K \left\|\Sigma_X\right\|_{\mathrm{op}}}$ in Lemma F.1. Substituting it into the epoch argument gives the corresponding regret refinement, with the selected-score factor $G_K(d)$ unchanged and initialization and failures controlled as above. The assumption concerns fresh candidate projections, not reward noise or selected-law projections; the separate deterministic regret cap remains valid.

F.1 Detailed Structural Comparisons

The bound has two menu-size factors: $G_K(d) \asymp K(d+K)$ from the selected-score moments and $\sqrt{K}$ from Lemma F.1, so estimation and transfer contribute $\sqrt{G_K(d)}\sqrt{K} \asymp K\sqrt{d+K}$. Against the lower bound's $1/K$ prefactor this leaves a gap of order $K^2\sqrt{1+K/d}$, which we do not claim is necessary. The design constants leave a further gap of $(M \left\|\Sigma_X\right\|_{\mathrm{op}})^{3/2}$ relative to Theorem G.2, which Remark G.4 reduces to $\sqrt{M \left\|\Sigma_X\right\|_{\mathrm{op}}}$ for the designs of Appendix B.2.

Product-beta specialization. For the density in Appendix B.2, Equation (9), take $\left\|\Sigma_X\right\|_{\mathrm{op}} = c_X = A^2/11$ and $M = M_0 = 12/A^2$ in Theorem 3.3. The general guarantee then recovers the product-design bounds with $\left\|\Sigma_X\right\|_{\mathrm{op}} M = 12/11$, under the same exact-score and selected-signal assumptions.

Vectorization and ESTOR. Representing the same matrix or tensor as an $N \times 1$ rank-one matrix gives the vectorized specialization

$$\widetilde{O}\Big(K\sqrt{(N+K)T}\Big)$$

of Theorem 3.3, with signal and model/design constants held fixed. ESTOR's stated Theorem 3.5 gives $\widetilde{O}(NK^{3/2}\sqrt{T})$ (Kang et al., 2026). This compares published upper bounds on a shared admissible bounded product-design subclass, with comparable norm budgets and signal floors and

uniformly controlled model/design constants. It is neither an empirical separation nor an algorithm-independent lower bound, and it makes no dominance claim over all ESTOR designs.

Comparison with unfolding. Any matricization preserves the original entrywise budgets and the joint density assumptions (it permutes coordinates), so the matrix specialization supplies a valid baseline using the rank and larger side length of the chosen unfolding.

Here let $d \geq 3$ be the side length of a cubic tensor and let $a_1, \ldots, a_d$ be the columns of $H = \mathbf{1}\mathbf{1}^\top - 2I_d$. Its eigenvalues are $d-2$ on the span of $\mathbf{1}$ and -2 on its orthogonal complement, so H is an invertible sign matrix. Set

$$\Theta_* = \frac{B}{d^3} \sum_{k=1}^{d} a_k \otimes a_k \otimes e_k, \qquad U = I_d.$$

Every entry has magnitude B/d^3, so the tensor is dense and $\|\Theta_*\|_{\mathrm{ent},1} = B$. Each frontal slice is the rank-one matrix $(B/d^3)a_k a_k^\top$, giving $r_U = d$ and slice structural scale d^2. In the first unfolding, fixing the second index j gives the invertible $d \times d$ submatrix $(B/d^3)H\,\mathrm{diag}(H_{j1}, \ldots, H_{jd})$. The second unfolding has the same property by symmetry. A relation with coefficients c_k among the rows of the third unfolding would give $H\,\mathrm{diag}(c)H^\top = 0$, hence $c = 0$. All three unfoldings therefore have rank d and larger side d^2, giving matrix structural scale d^3.

Use the same candidate law and parameter, and a fixed increasing linear link $f(z) = \beta z$, $0 < \beta \leq L_f$, for both representations. This gives the shared selected-design signal floor $\mu_{\min} = \beta$. For the explicit G_K bounds, including confidence factors, the ratio of the unfolding upper root term to the slice upper root term is

$$\left(\frac{G_K(d^2)\log(2(d+d^2)T)}{G_K(d)\log(4d^2T)} \right)^{1/2}.$$

Before these logarithms, the ratio is $\sqrt{G_K(d^2)/G_K(d)} \asymp \sqrt{(d^2+K)/(d+K)}$. The compact bound in Theorem 3.3 has structural-factor ratio $\sqrt{(d^2+K)/(d+K)}$. Thus the slice bound is smaller by order $\sqrt{d}$ when $K = O(d)$, but this gain does not hold uniformly as K grows. This compares upper bounds, not all unfolding-based algorithms or empirical performance.

G Lower-Bound Proofs

Throughout this appendix p is any candidate law satisfying Assumption 2.1, including the zero-extension regularity explained in Appendix B.1, with score certificate M, covariance $\Sigma_X = \mathrm{Cov}_p(\mathrm{vec}(X))$ and mean $m = \mathbb{E}_p \,\mathrm{vec}(X)$. Candidate entries may be dependent and noncentered. The product-beta law (9) appears only in Corollary G.3, where its independence and exact moments sharpen the constants. Arrays are identified with their vectorizations, so inner products with m and Frobenius norms are Euclidean.

We use the original-coordinate classes from Section 2.2, and the same realized-menu comparator as the upper bounds. Restrict the reward model to a disclosed linear link $f(z) = \beta z$, where $0 < \mu_{\min} \leq \beta \leq L_f$, and independent Gaussian noise of variance $\sigma^2 > 0$. Define $\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) = \inf_\pi \sup_{\Theta_* \in \mathcal{C}} \mathbb{E}_{\Theta_*,\pi} R_T$, with the infimum over all randomized adaptive policies, which may use p and the disclosed link.

Lemma G.1 (Menu comparison for every design) *Let p satisfy Assumption 2.1 with score certificate M. For every nonzero $v \in \mathbb{R}^N$, the law of $\langle v, \mathrm{vec}(X)\rangle$ has a continuous density bounded by $\sqrt{M}/(2\|v\|_2)$. For $K \geq 2$ iid candidates and $g_K^c(v) = \mathbb{E}\max_{a \leq K} \langle v, \mathrm{vec}(X_a) - m\rangle$,*

$$\frac{\|v\|_2}{8\sqrt{M}} \leq g_K^c(v) \leq \sqrt{K\, v^\top \Sigma_X v} \leq \sqrt{K\,\|\Sigma_X\|_{\mathrm{op}}}\,\|v\|_2. \tag{27}$$

Proof. *(a) Density bound.* Equation (24) of Appendix E is derived for an arbitrary unit direction. Its derivation uses Fubini's theorem, the zero-extension $W^{1,1}$ regularity, the conditional-score identity (E.2), the bounded support and $\mathbb{E}_p[\mathrm{vec}(S^p)\,\mathrm{vec}(S^p)^\top] \preceq MI_N$; it uses no selection rule and no product structure. Applied to the unit direction $v/\|v\|_2$, it bounds the continuous density of

$\langle v/\|v\|_2, \mathrm{vec}(X)\rangle$ by $\sqrt{M}/2$. Rescaling by $\|v\|_2$ gives a continuous density q_v of $\langle v, \mathrm{vec}(X)\rangle$ with $\|q_v\|_\infty \le \sqrt{M}/(2\|v\|_2)$.

(b) Lower bound. Let $Y_a = \langle v, \mathrm{vec}(X_a) - m\rangle$ and $W = Y_1 - Y_2$. Given $Y_2 = y$, W is the translate $Y_1 - y$, so W has density $\int q_v(w+y+\langle v, m\rangle)\,\mathbb{P}_{Y_2}(dy) \le \|q_v\|_\infty$: a bounded density convolved with a probability law. Hence $\mathbb{P}(|W| \le \delta) \le \delta\sqrt{M}/\|v\|_2$. Taking $\delta = \|v\|_2/(2\sqrt{M})$ gives $\mathbb{P}(|W| > \delta) \ge 1/2$ and $\mathbb{E}|W| \ge \delta\,\mathbb{P}(|W| > \delta) \ge \|v\|_2/(4\sqrt{M})$. Since $\max(Y_1, Y_2) = (Y_1+Y_2)/2 + |W|/2$ and $\mathbb{E}Y_a = 0$, $g_2^c(v) = \mathbb{E}|W|/2 \ge \|v\|_2/(8\sqrt{M})$. For $K \ge 2$, the maximum over all K candidates dominates the maximum over the first two, so $g_K^c \ge g_2^c$ pointwise.

(c) Upper bound. Since $\max_a Y_a \le (\sum_a Y_a^2)^{1/2}$, Jensen's inequality gives $g_K^c(v) \le (\sum_a \mathbb{E}Y_a^2)^{1/2} = (Kv^\top \Sigma_X v)^{1/2}$, and $v^\top \Sigma_X v \le \|\Sigma_X\|_{\mathrm{op}} \|v\|_2^2$.

Unlike the product-beta bound (32), this argument uses no independence, symmetry, or centering of p.

Rank-feasible orthonormal families. For matrices with $d_1 \ge d_2$, use coordinate matrices in the first r columns; for $d_2 > d_1$, use the first r rows. Their span has dimension dr, rank at most r, and coefficient entrywise cost one. For transformed slices, use the first ρ_k columns (or rows) in slice k, then pull back the coordinate basis through $\mathcal{T}_U^{-1}$. The family remains orthonormal and has dimension dr_U. A basis vector from slice k has original-coordinate entrywise norm $\sum_a |U_{ka}|$, so the triangle inequality gives cost χ. For uniform caps r, use original-coordinate tensors supported in the same first r matrix columns (or rows) across all trailing coordinates. Transforming only mixes those trailing coordinates, preserving the zero columns or rows. This alternative span has dimension drq and cost one. No invariance of the candidate law under U is invoked.

Each construction supplies $E_1, \dots, E_{s_{\max}}$ whose entire span obeys the rank constraints and satisfies $\left\|\sum_j t_j E_j\right\|_{\mathrm{ent},1} \le \chi \sum_j |t_j|$. Every active subset of $s \le s_{\max}$ directions therefore permits $\Theta_\omega = \epsilon \sum_{j=1}^s \omega_j E_j$ with independent fair signs, $\|\Theta_\omega\|_F = \epsilon\sqrt{s}$ and $\|\Theta_\omega\|_{\mathrm{ent},1} \le \chi s\epsilon$.

Mean-orthogonal hard directions. For a noncentered design, the common reward offset $\beta\langle m, \Theta_\omega\rangle$ would reveal the signs through every selected reward, and the per-coordinate relative-entropy bound below cannot absorb it. We therefore make every hard direction orthogonal to m. Let $\mu_j = \langle m, E_j\rangle$. If $\mu_j = 0$ for all j, as for every centered design including (9), set $u_j = E_j$, $s_{\mathrm{lb}} = s_{\max}$ and $\chi_{\mathrm{lb}} = \chi$. Otherwise, for $k \le \lfloor s_{\max}/2\rfloor$ set

$$u_k = \frac{\mu_{2k}E_{2k-1} - \mu_{2k-1}E_{2k}}{(\mu_{2k-1}^2 + \mu_{2k}^2)^{1/2}} \text{ if } (\mu_{2k-1}, \mu_{2k}) \ne 0, \qquad u_k = E_{2k-1} \text{ otherwise,}$$

and put $s_{\mathrm{lb}} = \lfloor s_{\max}/2\rfloor$ and $\chi_{\mathrm{lb}} = \sqrt{2}\,\chi$. The u_k are orthonormal, because they have disjoint supports in the E-basis. Each satisfies $\langle m, u_k\rangle = 0$. Their span lies in the span of the E_j, so it is rank-feasible. Finally, $\|\sum_k t_k u_k\|_{\mathrm{ent},1} \le \chi_{\mathrm{lb}} \sum_k |t_k|$, because $|a| + |b| \le \sqrt{2}$ when $a^2 + b^2 = 1$. We require $s_{\mathrm{lb}} \ge 1$, which holds automatically when $s_{\max} \ge 2$ or the design is centered; then $s_{\max}/3 \le s_{\mathrm{lb}} \le s_{\max}$.

Theorem G.2 (Same-menu lower bound for every design) *Let p satisfy Assumption 2.1, let $\mathcal{C}$, $s_{\max}$ and χ be as in Section 3.4, and let $s_{\mathrm{lb}} \ge 1$ and χ_{lb} be as above. Put*

$$\gamma_{\mathrm{lb}} = \frac{\beta B}{16\,\chi_{\mathrm{lb}}\sqrt{M}}, \qquad \tau_{\mathrm{lb}} = \frac{\sigma}{256\sqrt{2}\,K\,M\,\|\Sigma_X\|_{\mathrm{op}}}.$$

Then for every $T \ge 1$,

$$\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) \ge 2^{-1/4} \min\{\gamma_{\mathrm{lb}} T,\ \sqrt{\gamma_{\mathrm{lb}}\tau_{\mathrm{lb}}}\,T^{3/4},\ \tau_{\mathrm{lb}}\sqrt{s_{\mathrm{lb}}\,T}\}. \tag{28}$$

The bound still holds if the learner knows p and the subspace used in the proof.

The envelope crosses at

$$T_-(p) = \left(\frac{\tau_{\mathrm{lb}}}{\gamma_{\mathrm{lb}}}\right)^2, \qquad T_+(p) = \left(\frac{\tau_{\mathrm{lb}}\,s_{\mathrm{lb}}}{\gamma_{\mathrm{lb}}}\right)^2 = \frac{\sigma^2\chi_{\mathrm{lb}}^2 s_{\mathrm{lb}}^2}{512\,K^2\beta^2B^2M\,\|\Sigma_X\|_{\mathrm{op}}^2}. \tag{29}$$

For $T \geq \max\{1, T_+(p)\}$ the last branch is active, so

$$\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) \geq \frac{c\,\sigma\sqrt{s_{\mathrm{lb}}\,T}}{K\,M\,\|\Sigma_X\|_{\mathrm{op}}}, \qquad c = \frac{2^{-1/4}}{256\sqrt{2}} \geq \frac{1}{431}.$$

For $s_{\max} \geq 2$, the construction gives $s_{\mathrm{lb}} \geq \lfloor s_{\max}/2 \rfloor \geq s_{\max}/3$. Thus Theorem G.2 implies Theorem 3.4 with the universal constant reduced by a factor $1/\sqrt{3}$; the certificate above gives $1/(431\sqrt{3})$ for the main theorem.

A prior on $s \leq s_{\mathrm{lb}}$ signed mean-orthogonal directions balances the budget constraint against the information in selected rewards, and optimizing s gives (28). For a general slice profile, $1 \leq \chi \leq \sqrt{q}$ accounts for the inverse transform in the original-coordinate budget. The crossings are construction thresholds, not established minimax transitions; growing χ_{lb} changes the matching horizon. For $T \geq T_+(p)$ the square-root branch matches the dimension and horizon powers of Theorem 3.3 up to logarithms. This holds for every fixed design, fixed K, and fixed link, budget, noise and design constants, including $M\,\|\Sigma_X\|_{\mathrm{op}}$; since $s_{\max}/3 \leq s_{\mathrm{lb}} \leq s_{\max}$, the matrix and slice dimensions dr and dr_U enter up to a factor of at most three. Not matched are the menu size K; the ratios $L_f/\mu_{\min}$ and V/σ^2; the design factor, $\sqrt{M\,\|\Sigma_X\|_{\mathrm{op}}}$ in the upper bound versus $1/(M\,\|\Sigma_X\|_{\mathrm{op}})$ here (Remark G.4); and horizons below $T_+(p)$, including the $T^{3/4}$ branch. If $M\,\|\Sigma_X\|_{\mathrm{op}}$ grows with N along a family of designs, the two bounds differ by the growing factor $(M\,\|\Sigma_X\|_{\mathrm{op}})^{3/2}$, and neither bound is shown to have sharp dimension dependence for that family.

Proof of Theorem G.2. *Prior.* Fix $s \in [s_{\mathrm{lb}}]$ and take

$$\epsilon = \min\left\{\frac{B}{\chi_{\mathrm{lb}} s},\ \frac{\sigma}{16\sqrt{2}\,K\beta\,\|\Sigma_X\|_{\mathrm{op}}\sqrt{MT}}\right\}, \qquad \rho = \epsilon\sqrt{s}.$$

For independent fair signs $\omega \in \{-1,1\}^s$ put $\Theta_\omega = \epsilon\sum_{k=1}^{s}\omega_k u_k$. Then $\|\Theta_\omega\|_F = \rho$, $\langle m, \Theta_\omega\rangle = 0$ and $\|\Theta_\omega\|_{\mathrm{ent},1} \leq \chi_{\mathrm{lb}} s\epsilon \leq B$; the span of the u_k is rank-feasible, so $\Theta_\omega \in \mathcal{C}$.

Menu comparison. Lemma G.1 supplies both menu bounds below. The product-beta computation (32) is used only for Corollary G.3.

Adaptive information. Let P_ω^t be the history law through round t, including all observed menus and the parameter-independent policy seed. At a common history and menu, flipping sign j changes only the Gaussian reward mean, by $2\beta\epsilon\omega_j\langle X_u, u_j\rangle = 2\beta\epsilon\omega_j\langle X_u - m, u_j\rangle$, because $\langle m, u_j\rangle = 0$. Menus and the conditional policy kernel add no parameter-dependent likelihood factor. The chosen action is one of the K fresh candidates, so $\mathbb{E}[\langle X_u - m, u_j\rangle^2 \mid \mathcal{F}_{u-1}] \leq \sum_{a\leq K}\mathbb{E}\langle X_{u,a} - m, u_j\rangle^2 \leq K\,\|\Sigma_X\|_{\mathrm{op}}$. The chain rule gives

$$\mathrm{KL}(P_\omega^t \| P_{\omega^{(j)}}^t) = \frac{2\beta^2\epsilon^2}{\sigma^2}\sum_{u=1}^{t}\mathbb{E}_\omega\langle X_u - m, u_j\rangle^2 \leq \frac{2K\beta^2\epsilon^2 t\,\|\Sigma_X\|_{\mathrm{op}}}{\sigma^2}.$$

Selected actions need not have the same marginal law in the two experiments.

Posterior mean. Let $\bar\omega_j = \mathbb{E}[\omega_j \mid \mathcal{F}_t]$ and $\overline{\Theta}_t = \epsilon\sum_j \bar\omega_j u_j$. Pinsker's inequality for a fair binary sign gives $\mathbb{E}\bar\omega_j^2 \leq 2I(\omega_j; \mathcal{F}_t)$. Prior independence implies $I(\omega_j; \mathcal{F}_t) \leq I(\omega_j; \mathcal{F}_t \mid \omega_{-j})$. For fixed other signs, convexity of relative entropy bounds this last quantity by one quarter of the two directed KL divergences. Therefore

$$\mathbb{E}\bar\omega_j^2 \leq \frac{2K\beta^2\epsilon^2 t\,\|\Sigma_X\|_{\mathrm{op}}}{\sigma^2}, \qquad \mathbb{E}\left\|\overline{\Theta}_t\right\|_F \leq \rho\,\frac{\beta\epsilon\sqrt{2Kt\,\|\Sigma_X\|_{\mathrm{op}}}}{\sigma},$$

and $\langle m, \overline{\Theta}_t\rangle = 0$. No posterior independence is used.

Bayes regret. A fresh menu leaves the posterior mean unchanged. Since $\langle m, \overline{\Theta}_{t-1}\rangle = 0$, candidate a has posterior expected reward $\beta\langle X_{t,a} - m, \overline{\Theta}_{t-1}\rangle$, so the best posterior expected reward has mean $\beta g_K^c(\overline{\Theta}_{t-1}) \leq \beta\sqrt{K\,\|\Sigma_X\|_{\mathrm{op}}}\left\|\overline{\Theta}_{t-1}\right\|_F$ by Lemma G.1. Likewise $\langle m, \Theta_\omega\rangle = 0$, so the oracle's

expected reward is $\beta g_K^c(\Theta_\omega) \geq \beta\rho/(8\sqrt{M})$. The comparator is the realized-menu maximum, so a common mean offset would cancel from the regret in any case. Hence

$$\mathbb{E}r_t \geq \beta\rho\left[\frac{1}{8\sqrt{M}} - \frac{\sqrt{2}\,K\,\|\Sigma_X\|_{\mathrm{op}}\,\beta\epsilon\sqrt{T}}{\sigma}\right] \geq \frac{\beta\rho}{16\sqrt{M}}$$

by the choice of ϵ. Summing over rounds, the Bayes, and hence minimax, regret is at least

$$\frac{\beta\rho T}{16\sqrt{M}} = \min\{\gamma_{\mathrm{lb}}T/\sqrt{s},\ \tau_{\mathrm{lb}}\sqrt{sT}\}.$$

Optimizing s. For $A_0, B_0 > 0$, the continuous maximum over $1 \leq s \leq s_{\mathrm{lb}}$ of $\min(A_0/\sqrt{s}, B_0\sqrt{s})$ is $\min(A_0, \sqrt{A_0B_0}, B_0\sqrt{s_{\mathrm{lb}}})$. Rounding the crossing to either adjacent integer loses at most $2^{-1/4}$: if $j \leq u = A_0/B_0 \leq j+1$, the ratio is at least $\max(\sqrt{j/u}, \sqrt{u/(j+1)}) \geq (j/(j+1))^{1/4}$. With $A_0 = \gamma_{\mathrm{lb}}T$ and $B_0 = \tau_{\mathrm{lb}}\sqrt{T}$, this proves Equation (28).

The two crossings are $T_\pm(p)$ in Equation (29); for $T \geq T_+(p)$ the last branch is selected. The argument needs $\sigma > 0$ and concerns the disclosed linear submodel, for any p.

Corollary G.3 (Product-beta same-menu lower bound) *Let p be the product-beta law (9), and let $\mathcal{C}$, $s_{\max}$ and χ be as in Theorem G.2. For $T \geq 1$ set $\gamma_{\mathrm{pb}} = \beta B\sqrt{c_X}/\chi$ and $\tau_{\mathrm{pb}} = \sigma\sqrt{c_X M_0}$. Then*

$$\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) \geq \frac{1}{20K}\min\{\gamma_{\mathrm{pb}}T,\ \sqrt{\gamma_{\mathrm{pb}}\tau_{\mathrm{pb}}}\,T^{3/4},\ \tau_{\mathrm{pb}}\sqrt{s_{\max}T}\}. \tag{30}$$

The bound still holds if the learner is told the structural subspace used in its proof.

The envelope crosses at

$$T_- = \frac{\tau_{\mathrm{pb}}^2}{\gamma_{\mathrm{pb}}^2}, \qquad T_+ = \frac{\tau_{\mathrm{pb}}^2 {s_{\max}}^2}{\gamma_{\mathrm{pb}}^2} = \frac{\sigma^2\chi^2 M_0 {s_{\max}}^2}{\beta^2 B^2}. \tag{31}$$

The moment in this corollary is the actual $M_0 = 12/A^2$, not a freely enlarged certificate M.

Proof of Corollary G.3. Law (9) is centered, so every $\mu_j = 0$, $u_j = E_j$, $s_{\mathrm{lb}} = s_{\max}$ and $\chi_{\mathrm{lb}} = \chi$. Its independence and exact moments sharpen Lemma G.1. For a fresh menu, put $g_K(v) = \mathbb{E}\max_{a\leq K}\langle v, X_{t,a}\rangle$, which equals $g_K^c(v)$ here. Independence and the beta coordinate moments in Appendix B.2 give

$$\mathbb{E}\langle v, X\rangle^2 = c_X\|v\|_F^2, \qquad \mathbb{E}\langle v, X\rangle^4 = 3c_X^2\|v\|_F^4 - \frac{6}{13}c_X^2\sum_i v_i^4 \leq 3c_X^2\|v\|_F^4.$$

For $W = \langle v, X_{t,1} - X_{t,2}\rangle$, these imply $\mathbb{E}W^2 = 2c_X\|v\|_F^2$ and $\mathbb{E}W^4 \leq 12c_X^2\|v\|_F^4$. Hölder gives $\mathbb{E}|W| \geq \sqrt{2/3}\sqrt{c_X}\|v\|_F$. The sum of the three pairwise distances of three scalars equals twice their range. Central symmetry then gives $g_3(v) = 3\mathbb{E}|W|/4$. Monotonicity in K and the second-moment upper bound yield

$$\frac{\sqrt{6}}{4}\sqrt{c_X}\|v\|_F \leq g_K(v) \leq \sqrt{Kc_X}\|v\|_F. \tag{32}$$

This does not identify $g_K(v)$ with a direction-independent multiple of $\|v\|_F$.

Rerun the proof of Theorem G.2 with Equation (32) in place of Lemma G.1, with $\|\Sigma_X\|_{\mathrm{op}} = c_X$, and with

$$\epsilon = \min\left\{\frac{B}{\chi s}, \frac{\sqrt{3}\,\sigma}{8K\beta\sqrt{Tc_X}}\right\}.$$

The information and posterior steps give $\mathbb{E}\left\|\overline{\Theta}_t\right\|_F \leq \rho\sqrt{2K\beta^2\epsilon^2 tc_X/\sigma^2} \leq \sqrt{6}\rho/(8\sqrt{K})$. The best posterior expected reward therefore has mean at most $(\sqrt{6}/8)\beta\sqrt{c_X}\rho$, whereas the oracle's expected reward is at least $(\sqrt{6}/4)\beta\sqrt{c_X}\rho$. The Bayes, and hence minimax, regret is at least

$$\frac{\sqrt{6}}{8}\beta\sqrt{c_X}\rho T \geq \frac{1}{16K}\min\{\gamma_{\mathrm{pb}}T/\sqrt{s},\ \sigma\sqrt{sT}\}.$$

The rounding argument with $A_0 = \gamma_{\mathrm{pb}} T$ and $B_0 = \sigma\sqrt{T}$ gives a lower envelope with coefficient $2^{-1/4}/(16K)$ and σ in place of τ_{pb}. Since $\sqrt{c_X M_0} = \sqrt{12/11}$ and $2^{-1/4}\sqrt{11/12} \geq 4/5$, replacing σ by τ_{pb} gives Equation (30) with coefficient $1/(20K)$. The score moment enters only through this fixed-law numerical conversion.

Remark G.4 (Scale-free constant). The factor $1/(M \|\Sigma_X\|_{\mathrm{op}})$ in Theorem G.2 comes from measuring the hard directions in Euclidean rather than covariance geometry. Let $\mathcal{U} = \mathrm{span}\{u_1, \ldots, u_{s_{\mathrm{lb}}}\}$, of dimension s_{lb}, and let $\Sigma_{\mathcal{U}}$ be the compression of Σ_X to $\mathcal{U}$. The identity $\mathbb{E}[(u^\top(x - \bar{x}))(u^\top \mathrm{vec}(S^p(X)))] = 1$ of Appendix F and Cauchy–Schwarz give $u^\top \Sigma_X u \geq 1/M$ for every unit u, so $\lambda_{\min}(\Sigma_{\mathcal{U}}) \geq 1/M > 0$. Choose the hard directions $w_1, \ldots, w_{s_{\mathrm{lb}}} \in \mathcal{U}$ orthonormal for $\langle a, b\rangle_\Sigma = a^\top \Sigma_X b$, write $\|v\|_\Sigma = \langle v, v\rangle_\Sigma^{1/2}$, and define

$$a_p = \inf_{0 \neq v \in \mathcal{U}} \frac{\mathbb{E}|\langle v, X_1 - X_2\rangle|}{2\sqrt{v^\top \Sigma_X v}} \in (0, 2^{-1/2}].$$

The upper end follows from $\mathbb{E}|W| \leq (\mathbb{E}W^2)^{1/2}$. Run the proof of Theorem G.2 with $s = s_{\mathrm{lb}}$, prior $\Theta_\omega = \epsilon \sum_k \omega_k w_k$ and three changes. First, $\|\Sigma_X\|_{\mathrm{op}}$ becomes 1 in the information and upper-menu steps, because $w_j^\top \Sigma_X w_j = 1$ and $g_K^c(v) \leq \sqrt{K}\|v\|_\Sigma$; the posterior step then bounds $\mathbb{E}\left\|\overline{\Theta}_t\right\|_\Sigma$, with $\rho = \|\Theta_\omega\|_\Sigma = \epsilon\sqrt{s_{\mathrm{lb}}}$. Second, the oracle floor $1/(8\sqrt{M})$ becomes a_p, because $g_K^c \geq g_2^c = \mathbb{E}|W|/2 \geq a_p \|v\|_\Sigma$ on $\mathcal{U}$. Third, $\epsilon = a_p\sigma/(2\sqrt{2}K\beta\sqrt{T})$, which makes the per-round bracket $a_p/2$. For the budget, every $\Theta \in \mathcal{U}$ has $\|\Theta\|_{\mathrm{ent},1} \leq \chi_{\mathrm{lb}}\sqrt{s_{\mathrm{lb}}}\|\Theta\|_F \leq \chi_{\mathrm{lb}}\sqrt{s_{\mathrm{lb}}}\|\Theta\|_\Sigma / \sqrt{\lambda_{\min}(\Sigma_{\mathcal{U}})}$, so $\|\Theta_\omega\|_{\mathrm{ent},1} \leq \chi_{\mathrm{lb}} s_{\mathrm{lb}}\epsilon/\sqrt{\lambda_{\min}(\Sigma_{\mathcal{U}})}$, which is at most B once $T \geq T_+^\Sigma(p)$ below. Hence

$$\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) \geq \frac{a_p^2 \sigma\sqrt{s_{\mathrm{lb}} T}}{4\sqrt{2}\, K} \quad \text{for} \quad T \geq T_+^\Sigma(p) = \frac{a_p^2 \sigma^2 \chi_{\mathrm{lb}}^2 s_{\mathrm{lb}}^2}{8K^2\beta^2 B^2 \lambda_{\min}(\Sigma_{\mathcal{U}})}.$$

Two sufficient conditions bound a_p below. Lemma G.1(b) gives $\mathbb{E}|W| \geq \|v\|_2/(4\sqrt{M})$, hence $a_p \geq 1/(8\sqrt{M\lambda_{\max}(\Sigma_{\mathcal{U}})})$, which recovers the large-horizon coefficient of Theorem G.2 up to constants, under the threshold above. If $\mathbb{E}\langle v, X - m\rangle^4 \leq L(v^\top \Sigma_X v)^2$ on $\mathcal{U}$, then Hölder's inequality $\mathbb{E}|W| \geq (\mathbb{E}W^2)^{3/2}(\mathbb{E}W^4)^{-1/2}$ and $\mathbb{E}W^4 = 2\mathbb{E}Y^4 + 6(\mathbb{E}Y^2)^2$, with $Y = \langle v, X - m\rangle$, give $a_p^2 \geq 1/(L + 3)$.

Independent coordinates with coordinate kurtosis at most L_0 give $L \leq \max\{L_0, 3\}$. Law (9) has $L_0 = 363/143$, so $L \leq 3$. For the correlated designs of Appendix B.2, $m = b$ and $\langle \mathrm{vec}(X) - b, v\rangle = \langle z, H^\top v\rangle$, where z has independent product-beta coordinates. Hence $L \leq 3$ for every invertible H, and $\mathfrak{R}_T(\mathcal{C}; \beta, \sigma, p) \geq \sigma\sqrt{s_{\mathrm{lb}} T}/(24\sqrt{2}K)$ for $T \geq T_+^\Sigma(p)$. The constant does not depend on H; the threshold does, through $\lambda_{\min}(\Sigma_{\mathcal{U}})$ and the construction of $\mathcal{U}$. For this family Theorem 3.3 carries $\sqrt{M\|\Sigma_X\|_{\mathrm{op}}} = \sqrt{12/11}\,\mathrm{cond}_2(H)$. The remaining design gap therefore lies on the upper-bound side: it is a factor of the analysis of Theorem 3.3 that may be loose, not a demonstrated suboptimality of T-ESTOR.

H T-BSTOR FOR ARBITRARY LINKS

This appendix treats Lipschitz links that need not be monotone. T-BSTOR first estimates a signed multiple of the structured parameter from uniform observations, then learns rewards along the frozen estimated index; its scalar confidence bounds allow adaptive sampling and finite-variance noise. Matrix and transformed-slice fits are treated together.

H.1 LOCAL ASSUMPTIONS AND STRUCTURED INDEX ESTIMATION

Keep the parameter classes, entrywise budgets, fresh iid menus, known base density and score, and conditional noise kernel from Section 2. Retain absolute continuity and L_f-Lipschitz continuity on $[-AB, AB]$, without imposing monotonicity. There is no restriction on the additive reward level $f(0)$. For a nonzero parameter assume only

$$\mu_* = \mathbb{E}_p f'(\langle X, \Theta_*\rangle), \qquad |\mu_*| \geq \gamma > 0.$$

In particular, $|\mu_*| \le L_f$, and no selected-design signal condition is imposed. A zero parameter has identically zero regret and is handled separately without evaluating $f'(0)$.

Candidate entries may be dependent and have arbitrary means; write $m = \mathbb{E}_p X$, and retain $\|X\|_{\max} < A$, the zero-extension regularity $p \in W^{1,1}(\mathbb{R}^N)$, and the joint score certificate $\mathbb{E}_p[\mathrm{vec}(S^p(X))\,\mathrm{vec}(S^p(X))^\top] \preceq MI_N$ from Assumption 2.1. The prediction step uses the projection certificate $\kappa_X \ge 1$ of Equation (8), which concerns candidate projections rather than reward noise. A covariance bound alone does not establish that exponential inequality with the same constant. Every bounded law admits $\kappa_X = \sqrt{N}$: $\langle X, Z\rangle$ lies in an interval of length at most $2A\,\|Z\|_{\mathrm{ent},1} \le 2A\sqrt{N}\,\|Z\|_F$, so Hoeffding's lemma (Hoeffding, 1963) gives Equation (8) whatever the dependence. Independent bounded entries have $\kappa_X = 1$ for any means, by applying the same lemma coordinatewise; Appendix B.2 gives correlated, noncentered designs with $\kappa_X = 1$. Stronger dependence can force κ_X to grow: Equation (8) implies $A^2\kappa_X^2 \ge \|\Sigma_X\|_{\mathrm{op}}$, which is of order N when all coordinates share a random factor of constant variance. Write $V = 2L_f^2A^2B^2 + \sigma^2$ and $\Delta = 2L_fAB$; every realized-menu mean-reward gap is at most Δ.

Let $n_0 \ge 2$ be an even number of uniform-selection rounds. Its $n = n_0/2$ disjoint reward–score pairs are

$$Z_j = \frac{y_{2j-1} - y_{2j}}{2}[S^p(X_{2j-1}) - S^p(X_{2j})], \qquad j \in [n_0/2].$$

Uniform selection independently of each menu and the past gives iid selected observations with density p. This independence is between observations; entries of one observation need not be independent. Denote the resulting structured fit, before the additional projection below, by $\widehat{\Theta}^0$. This is the fitted matrix or transformed-slice estimator, not the unregularized robust moment.

Lemma H.1 (Signed Stein estimation from uniform observations) *Under the local assumptions above, with no signal lower bound needed for this lemma, the paired mean is $\mathbb{E}Z_j = \mu_*\Theta_*$, and the paired second-moment inequalities of Lemma D.1 hold with the base score and the same V. For $\delta_1 \in (0,1)$, let $d = \max(d_1, d_2)$. Use Equation* (4)*, with its matrix specialization when $q = 1$, and the following tuning and radii:*

$$\begin{aligned}
\textit{Matrix:}\quad \nu &= \sqrt{\frac{4\log(2(d_1+d_2)/\delta_1)}{n_0 VdM}}, \qquad \lambda = 4\sqrt{\frac{VdM\log(2(d_1+d_2)/\delta_1)}{n_0}},\\
&\qquad e_{n_0}^2 = \frac{32VMrd\log(2(d_1+d_2)/\delta_1)}{n_0};\\
\textit{Transformed slices:}\quad \nu &= \sqrt{\frac{4\log(2q(d_1+d_2)/\delta_1)}{n_0 VdM}}, \qquad \lambda = 4\sqrt{\frac{VdM\log(2q(d_1+d_2)/\delta_1)}{n_0}},\\
&\qquad e_{n_0}^2 = \frac{32VMdr_U\log(2q(d_1+d_2)/\delta_1)}{n_0}.
\end{aligned} \tag{33}$$

Then

$$\mathbb{P}\{\left\|\widehat{\Theta}^0 - \mu_*\Theta_*\right\|_F \le e_{n_0}\} \ge 1 - \delta_1.$$

The matrix and slice tuning does not use their rank caps.

Proof. Weak integration by parts gives $\mathbb{E}S^p(X) = 0$ and $\mathbb{E}[yS^p(X)] = \mu_*\Theta_*$ without using the sign of f', centering of X, or independence of its entries. Thus the proof of Lemma D.1 applies with the base score: $\mathbb{E}Z_j = \mu_*\Theta_*$, and the raw second moments of Z_j are at most V times those of S^p. The row and column calculation of Appendix E.3 bounds the left and right base-score moments of every matrix or transformed slice by dM, so $w = dM$, without slice independence or a slice-count factor. The concentration and thresholding arguments of Appendices D.2–D.3 use only these raw moments and upper bounds on the target ranks, which remain valid for negative or zero μ_*. Applying Theorem 3.1 with $n = n_0/2$ and $w = dM$ gives Equation (33). The nonzero base-signal condition is needed for the subsequent prediction and regret analysis, not for this estimation lemma.

H.2 Projection, Scalar Approximation, and T-BSTOR

For the estimation branch with even $2 \le n_0 < T$, project the fitted index onto a closed convex entrywise ball:

$$\widehat{\Theta} = \operatorname*{arg\,min}_{Z:\,\|Z\|_{\mathrm{ent},1} \le L_f B} \left\| Z - \widehat{\Theta}^0 \right\|_F^2, \qquad \mathcal{I} = [-AL_f B, AL_f B], \quad W = 2AL_f B. \tag{34}$$

Since $\|\mu_* \Theta_*\|_{\mathrm{ent},1} \le |\mu_*| B \le L_f B$, the Euclidean projection property gives $\left\| \widehat{\Theta} - \mu_* \Theta_* \right\|_F \le \left\| \widehat{\Theta}^0 - \mu_* \Theta_* \right\|_F$. Projection may change the fitted ranks; no subsequent step uses those ranks. For every candidate, $|\left\langle X, \widehat{\Theta} \right\rangle| \le AL_f B$, even when index estimation fails.

Condition on the exploration history and put $E = \widehat{\Theta} - \mu_* \Theta_*$ and $c_E = \langle m, E \rangle$. For analysis only, define the shifted link

$$g_E(u) = f\left(\mathrm{clip}_{[-AB, AB]} \left(\frac{u - c_E}{\mu_*} \right) \right), \qquad G = \frac{L_f}{|\mu_*|} \le \frac{L_f}{\gamma}.$$

The function g_E is fixed conditional on exploration and is G-Lipschitz. Since

$$\frac{\left\langle X, \widehat{\Theta} \right\rangle - c_E}{\mu_*} = \langle X, \Theta_* \rangle + \frac{\langle X - m, E \rangle}{\mu_*},$$

and the true index belongs to $[-AB, AB]$, nonexpansiveness of clipping around that index gives

$$\left| f(\langle X, \Theta_* \rangle) - g_E(\left\langle X, \widehat{\Theta} \right\rangle) \right| \le G |\langle X - m, E \rangle|. \tag{35}$$

This holds for either sign of μ_*. The candidate mean contributes a common horizontal shift of the unknown scalar link. The learner does not compute m, E, c_E, μ_*, or g_E: it continues to bin the uncentered fitted indices in the same interval $\mathcal{I}$. No feature centering, whitening, extra intercept, or score change is required.

The algorithm takes $T, K, A, B, L_f, \sigma^2, M$, exact base-score access, the chosen structural fit and its required structural inputs, a deterministic actual exploration count $n_0 \le T$, with either $n_0 = T$ or even $2 \le n_0 < T$, a bin width $w_{\mathrm{bin}} > 0$, and confidence allocations $\delta_1, \delta_2, \delta_3 \in (0, 1)$ with $\delta_1 + \delta_2 + \delta_3 < 1$. The allocation δ_2 is used only to report the prediction bound. The radii e_{n_0}, γ, μ_*, g_E, and the approximation allowance below are not action-rule inputs. If $n_0 = T$, the algorithm selects uniformly throughout and performs no fit or bin construction. This branch permits odd T. The following definitions apply only when $n_0 < T$.

Put $Q = T - n_0 > 0$ and $J_{\mathrm{bin}} = \lceil W / w_{\mathrm{bin}} \rceil$. Partition $\mathcal{I}$ into J_{bin} equal bins I_b of length $W / J_{\mathrm{bin}} \le w_{\mathrm{bin}}$, with centers c_b. The bins are left-closed and right-open, except that $I_{J_{\mathrm{bin}}}$ includes the right endpoint. Let k_{blk} be the smallest odd positive integer satisfying

$$k_{\mathrm{blk}} \ge \max\left\{ 1, \frac{2}{\log 2} \log \frac{J_{\mathrm{bin}} T}{\delta_3} \right\}. \tag{36}$$

For second-phase round t, let b_t denote the chosen bin and $n_b(t) = \sum_{u=n_0+1}^{t} \mathbf{1}\{b_u = b\}$ its selection count through round t, with $n_b(n_0) = 0$. For $s \ge k_{\mathrm{blk}}$, its confidence score is

$$U_{b,s} = \widehat{m}_{b,s} + \mathrm{rad}(s), \qquad \mathrm{rad}(s) = 4\sigma \sqrt{k_{\mathrm{blk}} / s}. \tag{37}$$

Here $\ell_s = \lfloor s / k_{\mathrm{blk}} \rfloor$ and the median uses the first $k_{\mathrm{blk}} \ell_s$ stored rewards in k_{blk} consecutive equal blocks:

$$\widehat{m}_{b,s} = \mathrm{median}_{1 \le j \le k_{\mathrm{blk}}} \left\{ \frac{1}{\ell_s} \sum_{a=(j-1)\ell_s + 1}^{j\ell_s} y_{b,a} \right\}. \tag{38}$$

The remainder is ignored for this evaluation and retained in the reward list. Uniform draws below are independent of the current menu and past.

Only currently available bins are compared. Initialization selects a bin with fewer than k_{blk} rewards, so each bin receives at most k_{blk} such selections. Their total count is at most $\min\{T - n_0, J_{\mathrm{bin}} k_{\mathrm{blk}}\}$ selections; these selections need not form a contiguous phase because bins become available at different times.

Algorithm 2 T-BSTOR: uniform structured exploration and robust available-bin learning.

Input: Model, structural fit, and tuning inputs specified above.

if $n_0 = T$ **then**
  Select uniformly and independently from each fresh menu for all T rounds
  return
end if
Clear the exploration buffer
for $t = 1, \dots, n_0$ **do**
  Observe $\mathcal{X}_t$; draw $a \sim \mathrm{Unif}([K])$ and play $X_{t,a}$
  Observe y_t and store $(y_t, S^p(X_{t,a}))$
end for
Form $n_0/2$ disjoint pairs by (3) with the base score
Compute the regularized fit $\widehat{\Theta}^0$ with tuning (33)
Project $\widehat{\Theta}^0$ by (34) and freeze $\widehat{\Theta}$
Set Q, J_{bin}, bins I_b, and k_{blk} as above; initialize empty reward lists and $n_b(n_0) = 0$
for $t = n_0 + 1, \dots, T$ **do**
  Observe $\mathcal{X}_t$ and compute $z_{t,a} = \left\langle X_{t,a}, \widehat{\Theta} \right\rangle$ for every $a \in [K]$
  $\mathcal{A}_t \leftarrow \{b : \text{some } z_{t,a} \in I_b\}$ ▷ Currently available bins
  if some $b \in \mathcal{A}_t$ has $n_b(t-1) < k_{\mathrm{blk}}$ **then**
    $b_t \leftarrow \min\{b \in \mathcal{A}_t : n_b(t-1) < k_{\mathrm{blk}}\}$
  else
    Evaluate $U_{b,n_b(t-1)}$ by (37)–(38) for $b \in \mathcal{A}_t$
    Choose b_t maximizing this score, breaking ties by smallest bin number
  end if
  $a \leftarrow \min\{a : z_{t,a} \in I_{b_t}\}$; play $X_{t,a}$ and observe y_t
  Append y_t to bin b_t's list and increment only its count
end for

Lemma H.2 (Uniform prediction and bin approximation) *Let $Q \geq 1$ and use a frozen projected estimate satisfying $\mathbb{P}\{\left\|\widehat{\Theta} - \mu_* \Theta_*\right\|_F \leq e_{n_0}\} \geq 1 - \delta_1$, as provided by Lemma H.1 and projection. Set*

$$\varepsilon_{\mathrm{pred}} = \frac{L_f A \kappa_X}{\gamma} e_{n_0} \sqrt{2 \log \frac{2KQ}{\delta_2}}, \qquad \epsilon_{\mathrm{bin}} = \varepsilon_{\mathrm{pred}} + \frac{G w_{\mathrm{bin}}}{2}, \qquad m_b = g_E(c_b).$$

With probability at least $1 - \delta_1 - \delta_2$, every future candidate whose fitted index falls in bin b satisfies

$$|f(\langle X, \Theta_* \rangle) - m_b| \leq \epsilon_{\mathrm{bin}}.$$

Proof. Conditional on the exploration history $\mathcal{F}_{n_0}$, the error $E = \widehat{\Theta} - \mu_* \Theta_*$ is fixed and future candidates are independent of that history with their original law p. Applying Equation (8) in the fixed direction E and optimizing the exponential Markov bound gives, for $E \neq 0$,

$$\mathbb{P}\{|\langle X - m, E \rangle| > u \mid \mathcal{F}_{n_0}\} \leq 2 \exp\left(- \frac{u^2}{2A^2 \kappa_X^2 \|E\|_F^2} \right).$$

If $E = 0$, the projection error is zero directly. On the index-confidence event $\|E\|_F \leq e_{n_0}$, a union bound over the KQ future candidates gives $\max_{t,a} |\langle X_{t,a} - m, E \rangle| \leq A \kappa_X e_{n_0} \sqrt{2 \log(2KQ/\delta_2)}$, except with conditional probability at most δ_2. Equation (35) bounds the resulting reward-prediction error by $\varepsilon_{\mathrm{pred}}$, since $G \leq L_f/\gamma$. Every bin center is at most $w_{\mathrm{bin}}/2$ from its indices. Adding this within-bin error and the estimation failure probability δ_1 proves the assertion. The union is over candidate arrays, not directions: E is fixed by conditioning. No covering argument or dimension factor beyond the supplied κ_X is needed.

H.3 Confidence for Adaptively Sampled Bins

Condition only on $\mathcal{F}_{n_0}$, which fixes the fitted index, partition, shifted link g_E, and reference means $m_b = g_E(c_b)$. Let $\mathcal{H}_t$ contain the past, current menu, action randomization, and selected action,

before observing its reward. The retained causal noise kernel implies

$$\mathbb{E}[\eta_t \mid \mathcal{H}_t] = 0, \qquad \mathbb{E}[\eta_t^2 \mid \mathcal{H}_t] \le \sigma^2.$$

The rewards assigned to a bin need not be iid or have a constant conditional mean. On the approximation event $\mathcal{E}_{\mathrm{app}}$ of Lemma H.2, their representation is

$$y_{b,a} = m_b + d_{b,a} + \eta_{b,a}, \qquad |d_{b,a}| \le \epsilon_{\mathrm{bin}}.$$

The drift is determined before its own reward noise and may vary with the adaptively selected candidate.

Lemma H.3 (Adaptive-bin confidence) *For the block count in Equation* (36)*, there is an event* $\mathcal{E}_{\mathrm{noise}}$ *satisfying* $\mathbb{P}(\mathcal{E}_{\mathrm{noise}}^c \mid \mathcal{F}_{n_0}) \le \delta_3$ *such that, on* $\mathcal{E}_{\mathrm{noise}} \cap \mathcal{E}_{\mathrm{app}}$*,*

$$|\widehat{m}_{b,s} - m_b| \le \epsilon_{\mathrm{bin}} + 4\sigma\sqrt{k_{\mathrm{blk}}/s} \tag{39}$$

simultaneously for every bin and every actual observed prefix $k_{\mathrm{blk}} \le s \le n_b(T)$*.*

Proof: adaptive times and padding. For each bin b, consider its chronological noise martingale and count

$$W_t^b = \sum_{u=n_0+1}^{\min(t,T)} \mathbf{1}\{b_u = b\}\eta_u, \qquad C_t^b = \sum_{u=n_0+1}^{\min(t,T)} \mathbf{1}\{b_u = b\}.$$

Conditioning first on $\mathcal{H}_t$ proves that W_t^b is a square-integrable martingale and $(W_t^b)^2 - \sigma^2 C_t^b$ is a supermartingale with respect to the chronological post-reward history. The selection indicator need not be measurable before the current menu; its measurability in $\mathcal{H}_t$ is sufficient.

A bin may never receive a prescribed number of observations. For the proof only, after time T pad its sequence with rewards m_b, zero drift, and zero noise, one per artificial round. The padded ath selection time $\tau_{b,a}$ is a stopping time bounded by $T + a$; the post-horizon history contains no new randomness. Take $\tau_{b,0} = n_0$ and $\mathcal{G}_{b,a} = \mathcal{F}_{\tau_{b,a}}$. Optional sampling of the two processes above gives

$$\mathbb{E}[\eta_{b,a} \mid \mathcal{G}_{b,a-1}] = 0, \qquad \mathbb{E}[\eta_{b,a}^2 \mid \mathcal{G}_{b,a-1}] \le \sigma^2. \tag{40}$$

Indeed, $W^b_{\tau_{b,a}} - W^b_{\tau_{b,a-1}}$ is this noise increment, and C^b increases by at most one between these times; martingale centering removes the cross term in the squared increment. All intervening menus and other bins' rewards are included in the stopped history. No conditioning on the eventual number of pulls is used, and padding never changes an actual observed prefix.

Proof: deterministic prefixes and dependent blocks. Fix a bin and a deterministic integer $s \in [k_{\mathrm{blk}}, T]$. The block size $\ell_s = \lfloor s/k_{\mathrm{blk}} \rfloor$ is deterministic. For the noise average $\bar{\eta}_j$ in block j, iterating Equation (40) cancels all cross terms conditional on the block-start history, so

$$\mathbb{E}[\bar{\eta}_j^2 \mid \mathcal{G}_{b,(j-1)\ell_s}] \le \frac{\sigma^2}{\ell_s}.$$

Conditional Chebyshev bounds the probability of $|\bar{\eta}_j| > \sqrt{8\sigma^2/\ell_s}$ by $1/8$. When $\sigma = 0$, all actual and padded noises are zero almost surely, and this conclusion holds directly. For any specified set of a bad blocks, iterated conditioning in reverse temporal order bounds their joint probability by 8^{-a}. Thus, without block independence,

$$\mathbb{P}\{\text{at least } (k_{\mathrm{blk}}+1)/2 \text{ bad blocks} \mid \mathcal{F}_{n_0}\} \le 2^{k_{\mathrm{blk}}} 8^{-k_{\mathrm{blk}}/2} = 2^{-k_{\mathrm{blk}}/2}.$$

There are at most $J_{\mathrm{bin}}T$ pairs (b, s), all defined by the padded sequences. A union bound and Equation (36) give an event of conditional failure probability at most δ_3 on which every such prefix has a strict majority of good blocks. This event covers the random counts used by the algorithm because it was constructed for deterministic prefix sizes.

Now intersect with $\mathcal{E}_{\mathrm{app}}$. Every actual block's average drift has magnitude at most ϵ_{bin}, so a strict majority of its reward block averages lie within $\epsilon_{\mathrm{bin}} + \sqrt{8\sigma^2/\ell_s}$ of m_b. Their median does too. Since $s \ge k_{\mathrm{blk}}$ implies $\ell_s \ge s/(2k_{\mathrm{blk}})$, this radius is at most that in Equation (39). The noise event was proved conditional only on exploration history, not on $\mathcal{E}_{\mathrm{app}}$ or all future menus. The final intersection uses a union bound, not independence of the two events.

H.4 Upper Regret Bound and Exploration Schedule

Theorem H.4 (Uniform exploration and robust-bin regret) *Under the local assumptions in Appendix H.1, with $\Theta_* \neq 0$, run Algorithm 2 with any deterministic even n_0 satisfying $2 \le n_0 < T$, width $w_{\mathrm{bin}} > 0$, and the stated confidence allocations. Use any deterministic valid radius e_{n_0} with $\mathbb{P}\{\left\|\widehat{\Theta}^0 - \mu_*\Theta_*\right\|_F \le e_{n_0}\} \ge 1-\delta_1$, as in Equation (33). With probability at least $1-\delta_1-\delta_2-\delta_3$,*

$$R_T \le \mathcal{U}_T := \Delta n_0 + \Delta \min\{Q, J_{\mathrm{bin}} k_{\mathrm{blk}}\} + 16\sigma\sqrt{k_{\mathrm{blk}} J_{\mathrm{bin}} Q} + 4Q\varepsilon_{\mathrm{pred}} + 2GQw_{\mathrm{bin}}.$$

Consequently,

$$\mathbb{E}R_T \le \min\{\Delta T,\ \mathcal{U}_T + \Delta T(\delta_1+\delta_2+\delta_3)\}. \tag{41}$$

The unknown G can be replaced by its upper bound L_f/γ when reporting these bounds. The all-exploration branch $n_0 = T$ has $R_T \le \Delta T$ directly; it uses no Q-dependent logarithm or undefined estimator when $Q = 0$.

Proof. If $n_0 = T$, the per-round gap bound gives $R_T \le \Delta T$ directly. Now take even $2 \le n_0 < T$, so $Q > 0$. Lemma H.2 bounds the approximation-event failure probability by $\delta_1 + \delta_2$. The conditional noise bound in Lemma H.3, integrated over the exploration history, contributes at most δ_3. Work on their intersection. The exploration phase costs at most Δn_0. At most k_{blk} initialization selections are charged to each bin, hence their total count is at most $\min\{Q, J_{\mathrm{bin}} k_{\mathrm{blk}}\}$.

On any other round, every available bin has at least k_{blk} observations. Let $b_\star$ contain a true reward-maximizing current candidate, let b_t be selected, and write $s = n_{b_t}(t-1)$. The confidence event and maximization of Equation (37) give

$$m_{b_\star} \le U_{b_\star, n_{b_\star}(t-1)} + \epsilon_{\mathrm{bin}} \le U_{b_t,s} + \epsilon_{\mathrm{bin}} \le m_{b_t} + 2\,\mathrm{rad}(s) + 2\epsilon_{\mathrm{bin}}.$$

Both actual candidate rewards are within ϵ_{bin} of their respective references, so $r_t \le 2\,\mathrm{rad}(s) + 4\epsilon_{\mathrm{bin}}$. The allowance ϵ_{bin} is common to every bin, which is why the algorithm need not add it to its scores or know its value. For each bin the pre-selection count takes each integer value at most once. Therefore

$$\sum_{\text{noninitialization } t} n_{b_t}(t-1)^{-1/2} \le \sum_{b=1}^{J_{\mathrm{bin}}} \sum_{s=1}^{n_b(T)} s^{-1/2} \le 2\sum_{b=1}^{J_{\mathrm{bin}}} \sqrt{n_b(T)} \le 2\sqrt{J_{\mathrm{bin}} Q}.$$

The confidence-width sum is at most $16\sigma\sqrt{k_{\mathrm{blk}} J_{\mathrm{bin}} Q}$. Finally, $4Q\epsilon_{\mathrm{bin}} = 4Q\varepsilon_{\mathrm{pred}} + 2GQw_{\mathrm{bin}}$. This proves the realized-menu mean-regret bound. The deterministic cap $R_T \le \Delta T$ adds at most ΔT times the combined failure probability when taking expectation, proving Equation (41).

Explicit schedule and structural inputs. For a nonzero parameter class, supply $D \ge 1$ satisfying

$$D \ge rd \quad \text{for the matrix fit,} \qquad D \ge dr_U \quad \text{for the transformed-slice fit.} \tag{42}$$

Set

$$\begin{aligned} n_0 &= \min\left\{T,\ 2\left\lceil \frac{(\kappa_X^2 D)^{1/3} T^{2/3}}{2} \right\rceil\right\}, \\ w_{\mathrm{bin}} &= WT^{-1/3}, \qquad \delta_1 = \delta_2 = \delta_3 = \frac{1}{3(T+1)^2}. \end{aligned} \tag{43}$$

If $n_0 = T$, select uniformly throughout without fitting or initializing bins, even when T is odd. Otherwise $2 \le n_0 < T$ is even, so the algorithm fits using $n_0/2$ pairs and runs the bin policy for the remaining rounds. In particular, $T = 1, 2$ use the all-exploration branch. A known zero parameter class may use any policy and has zero regret.

The exact bounds above display every score, noise, signal, concentration, and confidence factor. To read their rate, hold $A, B, L_f, M, \gamma, \sigma, K, h$ fixed, retain D and κ_X explicitly, and suppress logarithms in dimensions and T. For the estimation branch with even $2 \le n_0 < T$, Equation (33) gives $e_{n_0} = \widetilde{O}(\sqrt{D/n_0})$ and the upper bound becomes

$$\mathbb{E}R_T = \widetilde{O}\Big(n_0 + T\kappa_X\sqrt{D/n_0} + Tw_{\mathrm{bin}} + \sqrt{T/w_{\mathrm{bin}}} + 1/w_{\mathrm{bin}}\Big).$$

Under Equation (43) in this branch, $J_{\mathrm{bin}} = \lceil T^{1/3} \rceil$, $k_{\mathrm{blk}} = O(\log T)$, and hence

$$\mathbb{E}R_T = \widetilde{O}\Big(\min\{T, [(\kappa_X^2 D)^{1/3} + 1]T^{2/3}\}\Big). \tag{44}$$

Before capping, rounding adds fewer than two rounds. If $n_0 = T$, then $2\lceil (\kappa_X^2 D)^{1/3} T^{2/3}/2 \rceil \geq T$; use $R_T \leq \Delta T$ directly. Either $T \leq 3$ or $(\kappa_X^2 D)^{1/3} T^{2/3} > T - 2 \geq T/2$, so its linear cost obeys the same rate. The initialization cost is of order $T^{1/3}$ times a logarithm under this schedule and is explicitly retained in the finite-sample theorem.

Corollary H.5 (Sensitivity to the exploration schedule) *Retain the local model assumptions of Theorem H.4. Let $D \geq 1$ satisfy Equation* (42) *and let $\kappa_X \geq 1$ be a valid projection certificate. These are analytical bounds; the schedule may use a misspecified multiple of their product, without supplying rank caps to the fit. For an integer horizon $T \geq 1$ and $\varrho > 0$, set*

$$a = \varrho^{1/3} (\kappa_X^2 D)^{1/3} T^{2/3}, \qquad n_0 = \min\{T,\, 2\lceil a/2 \rceil\}.$$

If $n_0 < T$, keep the bin width and confidence allocations of Equation (43)*. If $n_0 = T$, select uniformly throughout without fitting or constructing bins, including for odd T. Then*

$$\mathbb{E}R_T = \widetilde{O}\Big(\min\Big\{T,\, \Big[(\varrho^{1/3} + \varrho^{-1/6})(\kappa_X^2 D)^{1/3} + 1\Big] T^{2/3}\Big\}\Big).$$

As in Equation (44)*, the model quantities $A, B, L_f, M, \gamma, \sigma, K, h$ are held fixed and logarithms in dimensions and T are suppressed; dependence on ϱ, D, κ_X is explicit. A zero parameter still has zero regret.*

Proof. In the estimation branch, $\max\{2, a\} \leq n_0 < a + 2$. Thus the exploration cost contributes at most $\Delta(a + 2)$, and

$$T\kappa_X \sqrt{D/n_0} \leq \varrho^{-1/6} (\kappa_X^2 D)^{1/3} T^{2/3}.$$

The general bound above Equation (44), with the unchanged bin settings, therefore gives the displayed rate: exploration contributes $\varrho^{1/3}$, estimation contributes $\varrho^{-1/6}$, and the scalar-learning and rounding terms are absorbed by the $+1$. The deterministic cap $R_T \leq \Delta T$ supplies the minimum with T. If $n_0 = T$, use that direct cap without constructing an estimator or bins. Either $T \leq 3$, or $2\lceil a/2 \rceil \geq T$ implies $a > T - 2 \geq T/2$. Hence the linear cost is absorbed by the same rate, with the $+1$ covering the small horizons. This proves both branches.

For fixed $\varrho > 0$ and fixed model quantities, schedule misspecification changes the prefactor but preserves the $T^{2/3}$ horizon exponent. This is not uniform over factors that deteriorate with T: a bounded exploration count can leave an order-T estimation term. Scheduling without κ_X corresponds to $\varrho = \kappa_X^{-2}$ and gives leading factor $(1 + \kappa_X) D^{1/3}$. Scheduling without either input corresponds to $\varrho = (\kappa_X^2 D)^{-1}$ and gives $1 + \kappa_X \sqrt{D}$, in each case with the existing binning contribution and linear cap. This corollary does not relax the input assumptions of Theorem 3.5.

The schedule uses D and κ_X only to choose n_0; the fits need no rank caps. The analogous vectorized analysis has leading factor $(\kappa_X^2 N)^{1/3}$ in place of $(\kappa_X^2 D)^{1/3}$, a ratio of $(D/N)^{1/3}$ under the same certificate and comparable model constants. With the universal $\kappa_X = \sqrt{N}$, the structured rate is $\widetilde{O}(\min\{T, [(ND)^{1/3} + 1]T^{2/3}\})$. These comparisons concern upper bounds, and the exponent $2/3$ treats the dimensions, κ_X and the model and signal constants as fixed. The bounds do not cover links with zero base first-order signal.

H.5 A Fixed-Menu Lower Bound for Arbitrary Links

The lower bound below fixes the structured parameter and varies only the link. It holds for every candidate law satisfying the local assumptions of Appendix H.1; the proof uses only the support bound, the zero-extension $W^{1,1}$ regularity and the score certificate M, through the projection-density bound (24). It therefore matches the horizon exponent of Theorem 3.5 for each design at fixed dimensions, menu size, and model/design constants, though not its dependence on κ_X, D, or the law of the disclosed index.

Theorem H.6 (Fixed-K horizon lower bound) *Fix finite dimensions, $K \geq 3$, $A, B, L_f, \sigma > 0$, and $0 < \gamma < L_f$, all independent of T. Let p satisfy the local assumptions of Appendix H.1. Let $\mathcal{C}$ be any of the parameter classes in Section 2.2 containing a nonzero parameter, and fix $0 \neq \Theta_0 \in \mathcal{C}$. There is a fixed class $\mathcal{F}_0 = \mathcal{F}_0(p, \Theta_0)$ of C^∞ links whose derivatives change sign on the actual index support, such that every $f \in \mathcal{F}_0$ is L_f-Lipschitz on $[-AB, AB]$ and satisfies*

$$\mathbb{E}_p f'(\langle X, \Theta_0 \rangle) \geq \gamma.$$

For rewards $y_t = f(\langle X_t, \Theta_0 \rangle) + \eta_t$ with independent $\mathcal{N}(0, \sigma^2)$ noise, there exist $c > 0$ and an integer $T_0 \geq 1$, depending only on the fixed quantities above, p and Θ_0, such that

$$\inf_{\pi} \sup_{f \in \mathcal{F}_0} \mathbb{E}_{\Theta_0, f, \pi} R_T \geq c T^{2/3} \qquad (T \geq T_0). \tag{45}$$

The policies may know Θ_0, the candidate law, and the link class. The class can additionally satisfy $\inf_{v : \|v\|_F = 1 \text{ or } v = 0} \mathbb{E}_{p^v} f'(\langle X, \Theta_0 \rangle) \geq \gamma$, where p^v denotes the law in Equation (6) for a frozen direction v, and $p^0 = p$. No uniform bound on higher derivatives is imposed.

The projected density. Every class of Section 2.2 with a nonzero cap contains a nonzero parameter: a coordinate spike for matrices, or the pull-back through $\mathcal{T}_U^{-1}$ of a single nonzero transformed entry for transformed slices, scaled to meet the budget. The proof applies to any fixed $0 \neq \Theta_0 \in \mathcal{C}$. Write $\theta_0 = \mathrm{vec}(\Theta_0)$, $r_0 = \|\Theta_0\|_F > 0$, $b_0 = A \|\Theta_0\|_{\mathrm{ent},1} \leq AB$, and $Z = \langle X, \Theta_0 \rangle$, so $|Z| \leq b_0$; let $b^* = \operatorname{ess\,sup} Z \leq b_0$. By Equation (24), applied to the unit direction θ_0 / r_0 and rescaled as in Lemma G.1, Z has a continuous density p_Z with

$$0 < p_+ := \|p_Z\|_\infty \leq \frac{\sqrt{M}}{2 r_0}.$$

Since Z has no atoms, $\Pr(Z \leq u) < 1$ for every $u < b^*$ and $\Pr(Z \leq u) \uparrow 1$ as $u \uparrow b^*$. For every $u < b^*$, $\Pr(Z > u) > 0$, so continuity gives a point $z^* > u$ with $p_Z(z^*) > 0$ and some $\eta > 0$ such that $z^* - 2\eta > u$ and $p_Z \geq p_Z(z^*)/2$ on $[z^* - 2\eta, z^* + 2\eta]$. This interval lies in $(u, b^*]$, since $\Pr(Z > b^*) = 0$. Under the product-beta law (9), $p_Z > 0$ on $(-b_0, b_0)$, so $b^* = b_0$, and $p_+ \leq 315/(256 A \max_i |\theta_{0,i}|)$.

A flat baseline, signed bumps, and an anchor. Choose a slope $d_0 \in (\gamma, L_f)$, then a point $u_0 < b^*$ such that $d_0 \alpha^K > \gamma$, where $\alpha = \Pr(Z \leq u_0)$. This is possible because $\Pr(Z \leq u) \uparrow 1$ as $u \uparrow b^*$. Then choose $u_1 \in (u_0, b^*)$. Choose a smooth cutoff $\zeta : \mathbb{R} \to [0, 1]$ equal to one on $[-AB, u_0]$ and zero on $[u_1, AB]$, and define the baseline $\mathfrak{b}(z) = d_0 \int_{u_1}^{z} \zeta(s)\, ds$. The cutoff can have compact support and remain zero for $z \geq u_1$. Thus $\mathfrak{b}' = d_0$ on $[-AB, u_0]$, $\mathfrak{b}' \geq 0$, and $\mathfrak{b} = 0$ on $[u_1, AB]$.

Apply the preceding paragraph with $u = u_1$, and set $\mathcal{I}_{\mathrm{b}} = [z^* - 2\eta, z^* - \eta]$ for bumps and $\mathcal{I}_{\mathrm{a}} = [z^* + \eta, z^* + 2\eta]$ for anchors. These disjoint nondegenerate intervals lie in $(u_1, b^*]$, where the baseline is zero, and

$$\ell_{\mathrm{b}} = |\mathcal{I}_{\mathrm{b}}| = \eta, \qquad p_- = \min_{z \in \mathcal{I}_{\mathrm{b}}} p_Z(z) \geq \frac{p_Z(z^*)}{2} > 0, \qquad q_{\mathrm{a}} = \Pr(Z \in \mathcal{I}_{\mathrm{a}}) \geq \frac{\eta\, p_Z(z^*)}{2} > 0.$$

Take a fixed smooth bump φ with $0 \leq \varphi \leq 1$, supported on $[-1/2, 1/2]$, and equal to one on $[-1/4, 1/4]$. Choose $a_0 > 0$ with $a_0 \|\varphi'\|_\infty \leq L_f$. For $0 < w \leq \ell_{\mathrm{b}}/2$, place $m_w = \lfloor \ell_{\mathrm{b}}/w \rfloor \geq \ell_{\mathrm{b}}/(2w)$ intervals I_j of length w with disjoint interiors in $\mathcal{I}_{\mathrm{b}}$, with centers z_j and cores $J_j = [z_j - w/4, z_j + w/4]$. For $\omega \in \{-1, 1\}^{m_w}$, set

$$B_{\omega,w}(z) = a_0 w \sum_{j=1}^{m_w} \omega_j \varphi\big((z - z_j)/w\big), \qquad f_{\omega,w}(z) = \mathfrak{b}(z) + B_{\omega,w}(z).$$

At most one bump is nonzero at each point, and the baseline is flat throughout their intervals, so $\|B_{\omega,w}\|_\infty \leq a_0 w$ and $\|f'_{\omega,w}\|_\infty \leq \max\{d_0, a_0 \|\varphi'\|_\infty\} \leq L_f$. Each link has both derivative signs on the actual support, since the baseline is flat on every nonzero signed bump. Its value on J_j is $\omega_j a_0 w$, whereas its value on the anchor interval is zero. For any fixed upper cutoff on w, the links are uniformly bounded on $[-AB, AB]$. Their higher derivatives need not be uniformly bounded as $w \downarrow 0$.

Preserving the signal without normalizing the parameter. For any projection-greedy law p^v, including $p^0 = p$, let S^{p^v} be its full joint score. Weak integration by parts gives

$$\mathbb{E}_{p^v} B'_{\omega,w}(Z) = \frac{1}{r_0^2}\mathbb{E}_{p^v}[B_{\omega,w}(Z)\left\langle S^{p^v}(X), \Theta_0\right\rangle].$$

Indeed, the directional derivative along θ_0 of $B_{\omega,w}(\theta_0^\top x)$ is $r_0^2 B'_{\omega,w}(Z)$. The zero-extended density is in $W^{1,1}$, and a smooth ambient cutoff equal to one on the support justifies integration by parts. Thus no boundary condition on individual projected fibers is being assumed. The joint score certificates in Appendix E.3 give

$$|\mathbb{E}_p B'_{\omega,w}(Z)| \le \frac{a_0 w\sqrt{M}}{r_0}, \qquad \sup_v |\mathbb{E}_{p^v} B'_{\omega,w}(Z)| \le \frac{a_0 w\sqrt{\overline{J}}}{r_0}, \qquad \overline{J} = MG_K(1).$$

The factor r_0^{-1} is required when the revealed parameter is not unit-Frobenius. On the event that all K candidate indices satisfy $Z \le u_0$, every selection rule chooses such an index; this event has probability α^K. Since the baseline derivative is nonnegative, $\mathbb{E}_{p^v}\mathfrak{b}'(Z) \ge d_0\alpha^K$ uniformly in v, including $v = 0$. Set

$$w_0 = \min\left\{\ell_{\mathrm{b}}/2,\ 1,\ \frac{(d_0\alpha^K - \gamma)r_0}{a_0\sqrt{\overline{J}}}\right\} > 0.$$

For every $w \le w_0$ and sign vector, $\mathbb{E}_{p^v} f'_{\omega,w}(Z) \ge \gamma$ uniformly in v. In particular the base signal is positive. This stronger conclusion concerns projection-greedy laws; it does not assert a signal floor under every nonlinear selection policy. Define the fixed class $\mathcal{F}_0 = \{f_{\omega,w} : 0 < w \le w_0,\ \omega \in \{-1,1\}^{m_w}\}$.

Adaptive information is limited by menu exposure. Let H_t contain the observed menus, actions, rewards, and private policy randomness through round t. For neighboring signs ω and $\omega^{(j)}$, the menu laws and conditional decision kernels agree at each common history. Only the selected Gaussian reward contributes to their relative entropy. Writing $Z_{t,a} = \langle X_{t,a}, \Theta_0\rangle$ and $Z_t = \langle X_t, \Theta_0\rangle$, the chain rule therefore gives

$$\begin{aligned}\mathrm{KL}(P_\omega^{H_T} \| P_{\omega^{(j)}}^{H_T}) &= \frac{1}{2\sigma^2}\mathbb{E}_\omega \sum_{t=1}^{T}\{f_{\omega,w}(Z_t) - f_{\omega^{(j)},w}(Z_t)\}^2 \\ &\le \frac{2a_0^2 w^2}{\sigma^2}\sum_{t=1}^{T}\Pr_\omega(Z_t \in I_j) \le \frac{2a_0^2 K p_+}{\sigma^2} T w^3.\end{aligned} \tag{46}$$

The last inequality holds for every adaptive strategy: it can select a candidate in I_j only when its current menu contains one, an event of probability at most Kp_+w. Observing all the other coordinates does not change this argument. Take

$$c_0 = \left(\frac{\sigma^2}{16a_0^2 K p_+}\right)^{1/3}, \qquad w = c_0 T^{-1/3}, \qquad T_0 = \max\{1, \lceil (c_0/w_0)^3\rceil\}.$$

For $T \ge T_0$, every constructed link belongs to $\mathcal{F}_0$ and the KL in Equation (46) is at most $1/8$; replacing p_+ by its bound $\sqrt{M}/(2r_0)$ in c_0 preserves this. Marginalization and Pinsker's inequality imply $\mathrm{TV}(P_\omega^{H_{t-1}}, P_{\omega^{(j)}}^{H_{t-1}}) \le 1/4$ at every round.

Testing conditional on the current informative menu. Let $E_{t,j}$ be the event that precisely one candidate index is in J_j and all other indices are in $\mathcal{I}_{\mathrm{a}}$. These events are disjoint over j, and

$$\Pr(E_{t,j}) = K\Pr(Z \in J_j) q_{\mathrm{a}}^{K-1} \ge c_E w, \qquad c_E = Kp_- q_{\mathrm{a}}^{K-1}/2 > 0.$$

There is at least one anchor because $K \ge 3$. Fix signs ω^+, ω^- differing only at j, with that sign positive and negative respectively. For each fixed full menu $\mathcal{X} \in E_{t,j}$, let $\pi_t(H, \mathcal{X}) \in [0,1]$ be the policy's probability of selecting the unique bump candidate. Current menus are independent of past histories and have the same law under both environments. Hence

$$\mathbb{E}_{\omega^+}[1 - \pi_t(H_{t-1}, \mathcal{X})] + \mathbb{E}_{\omega^-}[\pi_t(H_{t-1}, \mathcal{X})] \ge 1 - \mathrm{TV}(P_{\omega^+}^{H_{t-1}}, P_{\omega^-}^{H_{t-1}}) \ge 3/4.$$

On this menu, a wrong choice has regret $a_0 w$: the bump is better than the anchors under ω^+ and worse under ω^-. Integrating the last inequality over these current menus gives

$$\mathbb{E}_{\omega^+}[r_t \mathbf{1}_{E_{t,j}}] + \mathbb{E}_{\omega^-}[r_t \mathbf{1}_{E_{t,j}}] \geq \tfrac{3}{4} a_0 w \Pr(E_{t,j}) \geq \tfrac{3}{4} a_0 c_E w^2.$$

The event probability multiplies a constant testing error conditional on the current menu. Subtracting a constant unconditional total variation bound from the rare-event probability would not establish this result.

Average over the 2^{m_w} sign patterns, pairing neighbors for each j. Nonnegative regret and disjointness of the informative-menu events yield

$$\begin{aligned} 2^{-m_w} \sum_{\omega} \mathbb{E}_\omega R_T &\geq \sum_{t=1}^{T} \sum_{j=1}^{m_w} 2^{-m_w} \sum_{\omega} \mathbb{E}_\omega[r_t \mathbf{1}_{E_{t,j}}] \\ &\geq \tfrac{3}{8} a_0 c_E m_w T w^2 \geq \tfrac{3}{16} a_0 c_E \ell_{\mathrm{b}} c_0 T^{2/3}. \end{aligned}$$

This proves Equation (45) with $c = 3a_0 c_E \ell_{\mathrm{b}} c_0/16$. All choices defining c and T_0 depend only on the fixed problem quantities, the law p and the fixed parameter; they depend on p through p_+, p_-, q_{a}, ℓ_{b} and α. Although the hardest link may depend on T, the class $\mathcal{F}_0$ does not. Revealing Θ_0 only makes learning easier, so for the same p the lower bound applies to the full model with an unknown structured index.

The bound matches the horizon exponent only, for fixed dimensions, K, and positive noise and signal constants; c_E contains q_{a}^{K-1}, so it gives no uniform statement in K. A zero-parameter class or $K = 1$ has zero regret. The constructed links are individually smooth, but the class has no common bound on higher derivatives.

I Experimental details

I.1 Dense synthetic design and six links

Let $a_1 \in \{\pm 1\}^{16}$ and $b_1 \in \{\pm 1\}^{64}$ have independent uniform signs. Set $a_2 = a_1 \odot q_a$, where q_a is a uniformly permuted balanced sign vector, and $b_2 = b_1 \odot q_b$, where each of the four length-16 blocks of q_b is independently balanced and permuted. Define

$$M_* = \frac{a_1 b_1^\top + \frac{1}{2} a_2 b_2^\top}{1024}, \qquad \Theta_{*,:,:,k} = M_{*,:,16(k-1)+1:16k}, \quad k = 1, \ldots, 4.$$

The two components are orthogonal globally and within each frontal slice. Every entry is nonzero, $\|\Theta_*\|_1 = 1$, and $\|\Theta_*\|_F^2 = 1.25/1024$. The matrix singular values are $1/32$ and $1/64$; each slice also has rank two and a $2:1$ singular-value ratio. The slice transform is $U = I_4$. The matrix scale $2\max\{16, 64\}$ and slice scale $8\max\{16, 16\}$ are both 128, so this family does not imply a matrix-versus-slice separation. Matrices horizontally concatenate frontal slices without interleaving their entries. All three representations preserve scalar inner products exactly.

Each round supplies $K = 10$ independent candidate arrays with independent entries $X_j = 2\,\mathrm{Beta}(5,5) - 1$. The covariance and score moments are $\mathbb{E}[\mathrm{vec}(X)\,\mathrm{vec}(X)^\top] = I/11$ and $\mathbb{E}[\mathrm{vec}(S(X))\,\mathrm{vec}(S(X))^\top] = 12I$, where $S_j(x) = 8x_j/(1 - x_j^2)$. The candidate bound is $A = 1$ and coefficient budget is $B = 1$. Write $s_0 = \sqrt{1.25/(11 \cdot 1024)} \simeq 0.01053437$ and $v = u/s_0$. The six fixed links are

$$\begin{aligned} f_{\mathrm{linear}}(u) &= u, & f_{\mathrm{valley}}(u) &= \tfrac{u}{2} + s_0(\sqrt{1+v^2} - 1), \\ f_{\tanh}(u) &= u + \tfrac{s_0}{2}\tanh(v), & f_{\mathrm{peak}}(u) &= \tfrac{u}{2} - s_0(\sqrt{1+v^2} - 1), \\ f_{\mathrm{convex}}(u) &= u + \tfrac{s_0}{2}\log\cosh(v), & f_{\mathrm{wave}}(u) &= \tfrac{u}{2} + \tfrac{s_0}{2}(\cos(2v) - 1). \end{aligned}$$

All vanish at zero, are 1.5-Lipschitz, and are bounded in magnitude by 1.5 on $[-1, 1]$. The first three are strictly increasing. Symmetry gives base Stein multipliers 1, a value in $[1, 1.5]$, 1, $1/2$, $1/2$, and $1/2$, respectively. Valley and peak turn at $-s_0/\sqrt{3}$ and $s_0/\sqrt{3}$; wave derivatives vanish at standardized indices $\pi/12 + k\pi$ and $5\pi/12 + k\pi$. These turns occur on the index's standard-deviation scale. Noise is Student-t_3 multiplied by $s_0/(2\sqrt{3})$, giving standard deviation $s_0/2$ and finite variance.

Coefficient, context, potential-noise, and uniform-action streams are independent. Within each seed, all six links and methods share the coefficient, candidate menus, potential-noise realizations, and uniform choices. Policies receive only their candidate features and chosen rewards. Hidden coefficients, true means, derivatives, and signal checks stay in the simulator. Regret is

$$R_T = \sum_{t=1}^{T} \left[\max_{a \le K} f(\langle X_{t,a}, \Theta_* \rangle) - f(\langle X_{t,A_t}, \Theta_* \rangle) \right],$$

including exploration, bin initialization, and fallback rounds.

I.2 Practical adaptive T-BSTOR

The three T-BSTOR variants use the same uniform exploration and robust scalar learner, with radial vector influence or spectral matrix/slice influence in their paired base-score fit. At checkpoints $n = 100, 200, \ldots$, each fit uses all $n/2$ consecutive observation pairs and recomputes its sample-size-dependent influence and threshold parameters. The reconstructed estimate $\widehat{\Theta}_n^0$ is projected onto the original-coordinate entrywise ℓ_1 ball of radius $L_f B$, with $L_f = 1.5$, to give $\widehat{\Theta}_n$.

For numerically valid nonzero fits, form $v_n = \mathrm{vec}(\widehat{\Theta}_n)/\|\widehat{\Theta}_n\|_F$ and compare $\min\{\|v_n - v_{n-100}\|_2, \|v_n + v_{n-100}\|_2\}$ with the tuned tolerance. Exploration stops after two consecutive successful comparisons and at least 400 observations. An invalid or zero fit resets the success count and breaks successive-fit comparisons. The cap is $n_{\max} = 2\lfloor 0.4T/2 \rfloor$; fitting at this even cap occurs even when it is not a multiple of 100. A zero/invalid capped fit triggers uniform selection for the remaining horizon. This stability rule is a practical stopping heuristic, not a certificate of estimator accuracy.

At the final fit freeze $s = \|\widehat{\Theta}\|_F$ and $v = \mathrm{vec}(\widehat{\Theta})/s$. Scalar decisions use $\langle x, v \rangle$; the unit direction is not projected again onto an ℓ_1 ball. Divide $[-R, R]$ into $J_{\mathrm{bin}} = \lceil c_J T^{1/3} \rceil$ equal bins, where

$$R = \sqrt{\frac{2}{11} \log \frac{2KT}{\delta_R}}, \qquad \delta_R = (T+1)^{-2}.$$

The registered $c_J \in \{0.5, 1, 2\}$ gives $J_{\mathrm{bin}} = 21, 41, 81$ at the final synthetic horizon. The right endpoint belongs to the last bin. Indices outside the interval map to its edge bins; candidate and selected-action overflows are counted. Only chosen rewards collected after exploration initialize or update the bins. The minimum odd block count, consecutive equal-size median-of-means blocks, confidence allocation, initialization priority among available bins, and tie breaking are shared across all three representations. The radius is $4c_{\mathrm{conf}}\sigma\sqrt{k_{\mathrm{blk}}/m}$ for m observations in a bin. Tuned confidence multipliers do not remove mandatory bin initialization.

Index range. After normalization the bin-approximation term scales with the fitted norm s: if g is G-Lipschitz, $z \mapsto g(sz)$ is Gs-Lipschitz, giving $Gs(2R/J_{\mathrm{bin}})$. Each beta coordinate has moment-generating function at most $\exp(\lambda^2/22)$, so every unit projection of a fresh candidate has this bound. Conditional on the exploration stopping history and frozen direction, future menus remain independent. A union bound over at most KT candidates keeps indices in $[-R, R]$ except with probability δ_R, costing at most $\Delta T \delta_R$ in expectation when gaps are bounded by Δ. This controls the index range only; extending the deterministic-exploration theorem to adaptive stopping would require a simultaneous estimation argument.

I.3 Other algorithms and numerical implementations

T-ESTOR retains doubling epochs $2, 4, 8, \ldots$, latest-epoch paired spectral estimation, and projection-greedy selection. A frozen direction v determines the projection CDF F_v and density F_v' under the product-beta base law, yielding the exact selected score

$$S^{p_v}(X) = S^p(X) - (K-1)\frac{F_v'(\langle X, v \rangle)}{F_v(\langle X, v \rangle)} v.$$

The numerical implementation uses double-precision Fourier convolution with all nonzero direction coefficients, 4,096 grid points, a 5% support margin, and CDF floor 10^{-12}. For multicoordinate

Table 2: Oracle discrepancies against a 16,384-point numerical reference over all 27 frozen directions. Actual: the first up to 256 chosen projections per selected epoch; tails: 17 points from $-8/\sqrt{11}$ to $8/\sqrt{11}$. Entries are absolute; $g_v = F_v'$ and $c = (K-1)g_v/\max\{F_v, 10^{-12}\}$. Floor counts refer to the compared grid.

Grid	Queries	Max $\lvert\Delta g\rvert$	Max $\lvert\Delta F\rvert$	95th $\lvert\Delta c\rvert$	Max $\lvert\Delta c\rvert$	Floors
4,096	Actual	3.80×10^{-4}	1.26×10^{-4}	1.81×10^{-3}	5.85×10^{-3}	0/5284
4,096	Tails	1.66×10^{-4}	1.17×10^{-4}	1.16×10^{0}	1.65×10^{0}	37/459
8,192	Actual	1.03×10^{-4}	2.88×10^{-5}	4.11×10^{-4}	1.72×10^{-3}	0/5284
8,192	Tails	4.04×10^{-5}	2.77×10^{-5}	2.38×10^{-1}	3.27×10^{-1}	37/459

directions, construction multiplies the coordinate characteristic functions, inverts the Fourier transform, normalizes the density, and integrates it to obtain the CDF. Subsequent scores project each action, interpolate the density and CDF, apply the floored ratio, and subtract the directional correction from the explicit beta base score. Oracle construction and evaluation are separate from the paired robust fits and their SVDs. A zero fit uses uniform play and base scores. T-ESTOR is evaluated only on monotone links. The numerical scores approximate the theorem's exact oracle. Grid comparisons establish neither uniform score-error certificates nor the selected-score and epoch conditions of Appendix C.7.

Oracle accuracy and cost. For three evaluation seeds and the three monotone links, we use the first, middle, and last epochs with nonzero incoming fits: 27 directions. Against a 16,384-point numerical reference, the 4,096-point oracle's maximum discrepancies at selected actions are 3.8×10^{-4} in density, 1.3×10^{-4} in CDF, and 5.9×10^{-3} in score correction (Table 2). Tail discrepancies are larger: all three grids activate the floor on 37 of 459 tail queries, but none of 5,284 selected queries; nine full evaluation runs have no activations in 587,668 nonzero-fit queries. These checks do not certify uniform error.

On an Intel i7-1065G7 shared desktop, sequential double-precision diagnostics with one BLAS/OpenMP thread give median construction cost 1.52 s per direction and correction evaluation cost $0.344\,\mu$s per query from precomputed 256-projection batches. Construction is fresh and uncached after one warm-up; evaluation uses five warm-ups and seven repeats of 100 batches. Both exclude reconstruction, feature projection, base scores, and robust fitting/SVD. Construction accounts for 43.1% of pooled online time (31.6–51.6% per run); the combined score timer does not isolate correction evaluation.

ZoomSIB-UCB follows Algorithm 1 and Appendix G of Dey et al. (2026): clipped reward-score exploration, entrywise ℓ_1 normalization, adaptive normalized-estimate comparisons every 100 rounds, and sample-mean UCB over available bins. Its minimum is 200 observations with two checkpoints; its cap is $0.4T$. Its tuned stability criterion is unchanged and differs from the two-success Frobenius rule used by the T-BSTOR variants. At each checkpoint clipping is recomputed using all observations and the current sample size, an explicit implementation choice where the adaptive presentation is underspecified. The frozen direction uses the published scalar procedure; exploration rewards do not initialize bins. Its synthetic index window is $[-1, 1]$.

GSTOR follows Kang et al. (2026), including Appendix K: two independent uniform stages of $T_1 = \lceil c_n T^{3/4} \rceil$ observations, each capped at $\lfloor (T-1)/2 \rfloor$. The first fits coordinatewise clipped reward-score products, optionally soft-thresholded by the declared penalty; the second fits the uniform-kernel predictor with bandwidth $c_h T_1^{-1/3}$. The coefficient normalizer is $\|\Sigma^{1/2}\widehat{\theta}\|_1$, with $\Sigma^{1/2} = I/\sqrt{11}$ and synthetic window $[-\sqrt{11}, \sqrt{11}]$. Both estimators remain frozen during greedy commitment. Sorted indices and prefix sums evaluate the same closed kernel windows exactly. Empty and out-of-window predictions equal zero with first-candidate ties, including for negative rewards. The feasible set is the full vector space; selected nonzero penalties are identified as regularized GSTOR implementations. None of these baselines receives a true coefficient or link. The beta design differs from GSTOR's Gaussian theorem setting, and Student-t_3 noise is not sub-Gaussian as assumed by ZoomSIB-UCB; this comparison does not assert all external guarantees under the common environment.

I.4 Registered tuning and reporting

All arbitrary-link methods have 48 distinct configurations. For matrix and slice T-BSTOR, cross stability tolerances $\{0.05, 0.1\}$, bin-count multipliers $\{0.5, 1, 2\}$, confidence multipliers $\{1/30, 0.1\}$, and thresholds $c_\tau \in \{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$, fixing the influence multiplier $c_\nu = 16$. T-BSTOR-Vec has no low-rank threshold and instead crosses $c_\nu \in \{1, 16, 64, 256\}$ with the same first three sets. ZoomSIB-UCB crosses tolerances $\{0.025, 0.05, 0.1\}$, width multipliers $\{0.0625, 0.125, 0.5, 1\}$, confidence multipliers $\{0.25, 1\}$, and truncation multipliers $\{0.1, 1\}$. GSTOR retains the previous 24 settings with $c_n \in \{0.125, 0.5, 2\}$, bandwidth multipliers $\{0.25, 1\}$, penalties $\{0, 0.005\}$, and truncation multipliers $\{0.1, 1\}$, adding 24 that use bandwidth multipliers $\{0.0625, 0.125\}$ and penalties $\{0, 0.0001\}$ with the same other sets. T-ESTOR retains its 24 settings crossing $c_\nu \in \{1, 16, 256\}$ with $c_\tau \in \{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$.

Synthetic screening uses three seeds at $T = 8{,}190$ and advances eight configurations per method and link. Second-stage validation uses ten separate seeds at $T = 32{,}766$; final evaluation uses 50 fresh paired seeds at $T = 65{,}534$. Each selection minimizes mean final regret, with prespecified configuration order breaking ties. The monotone method screens fewer configurations and has the same second-stage budget. All choices freeze before evaluation; held-out results do not change grids or selected settings.

Figures use 10,000 paired whole-seed percentile bootstrap draws. Bands are pointwise 95% intervals, not simultaneous bands. Final-regret contrasts resample the same paired seedwise differences; these intervals are unadjusted for multiple comparisons. Direction errors average the sign-invariant distance $\min\{\|v - v_*\|_2, \|v + v_*\|_2\}$ over valid nonzero fits, where $v_* = \mathrm{vec}(\Theta_*)/\|\Theta_*\|_F$. They are undefined for zero fits and do not certify T-ESTOR's orientation. No structural rank is assigned to the trivial vector matrix encoding. Tables 5 and 6 summarize exploration, initialization, bins, and fits. Diagnostics include stopping times, phase regret, within-selected-bin loss, norms, pre/post-projection ranks, overflows, and fallbacks; baseline phase summaries do not separately identify bin initialization.

Online times exclude shared environment construction and offline preparation/calibration; cached fits/projections and parallel runs make these operational timings, not hardware-independent speed comparisons. In an earlier signed-diagonal design, no arbitrary-link T-BSTOR variant had the lowest mean regret in any panel: monotone means were 1,645–1,658, versus 813 for ZoomSIB-UCB and 127 for T-ESTOR; GSTOR had the lowest means on the old nonmonotone panel. These descriptive means are not direct improvement estimates: the design, signal/noise scale, and practical algorithm changed. The revision was designed to test whether dense low-rank structure and the revised practical implementation yield a finite-horizon advantage.

I.5 Synthetic settings and diagnostic summaries

Table 3 lists the selected hyperparameters for every method and link. Table 4 gives the prespecified paired contrasts; Tables 5 and 6 report exploration, initialization, and fit diagnostics.

Both structured T-BSTOR variants have smaller mean sign-invariant direction errors than T-BSTOR-Vec on each link: their means range from 0.413 to 0.626, compared with 0.661 to 0.804 for vectorization. Their mean later-scalar regret and within-selected-bin loss are also lower on every link. Initialization still requires about 1,100–2,072 rounds. These descriptive totals span different numbers of scalar rounds; their magnitudes alone do not compare per-round prediction quality or identify a sole cause of the regret gap. The fitted ranks are estimated rather than supplied from the true tensor, and the direction diagnostic does not certify the stopping rule.

Table 3: Frozen synthetic configurations. Every tuned field appears in its method's tuple order. c_ν is the influence multiplier; c_w and c_h are width and bandwidth multipliers, λ_{reg} is the GSTOR penalty, and c_{trunc} is the truncation multiplier. Nonzero λ_{reg} denotes regularized GSTOR.

Panel	Policy	Parameter tuple
T-BSTOR: $(\varepsilon, c_J, c_{\mathrm{conf}}, c_\nu, c_\tau)$		
Linear	T-BSTOR-Vec	$(0.1,\ 0.5,\ \frac{1}{30},\ 64,\ 0)$
Linear	T-BSTOR-Mat	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Linear	T-BSTOR-Slice	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Tanh	T-BSTOR-Vec	$(0.1,\ 0.5,\ \frac{1}{30},\ 256,\ 0)$
Tanh	T-BSTOR-Mat	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Tanh	T-BSTOR-Slice	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Convex	T-BSTOR-Vec	$(0.1,\ 1,\ \frac{1}{30},\ 256,\ 0)$
Convex	T-BSTOR-Mat	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Convex	T-BSTOR-Slice	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Valley	T-BSTOR-Vec	$(0.1,\ 1,\ \frac{1}{30},\ 256,\ 0)$
Valley	T-BSTOR-Mat	$(0.05,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Valley	T-BSTOR-Slice	$(0.05,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Peak	T-BSTOR-Vec	$(0.1,\ 0.5,\ \frac{1}{30},\ 1,\ 0)$
Peak	T-BSTOR-Mat	$(0.1,\ 1,\ \frac{1}{30},\ 16,\ 0.001)$
Peak	T-BSTOR-Slice	$(0.1,\ 0.5,\ 0.1,\ 16,\ 0.001)$
Wave	T-BSTOR-Vec	$(0.1,\ 0.5,\ 0.1,\ 256,\ 0)$
Wave	T-BSTOR-Mat	$(0.1,\ 0.5,\ 0.1,\ 16,\ 0.001)$
Wave	T-BSTOR-Slice	$(0.1,\ 0.5,\ \frac{1}{30},\ 16,\ 0.001)$
ZoomSIB-UCB: $(\varepsilon, c_w, c_{\mathrm{conf}}, c_{\mathrm{trunc}})$		
Linear	ZoomSIB-UCB	$(0.1,\ 0.0625,\ 0.25,\ 0.1)$
Tanh	ZoomSIB-UCB	$(0.1,\ 0.0625,\ 0.25,\ 0.1)$
Convex	ZoomSIB-UCB	$(0.1,\ 0.0625,\ 0.25,\ 0.1)$
Valley	ZoomSIB-UCB	$(0.05,\ 0.0625,\ 0.25,\ 0.1)$
Peak	ZoomSIB-UCB	$(0.1,\ 0.125,\ 0.25,\ 0.1)$
Wave	ZoomSIB-UCB	$(0.1,\ 0.125,\ 0.25,\ 0.1)$
GSTOR: $(c_n, c_h, \lambda_{\mathrm{reg}}, c_{\mathrm{trunc}})$		
Linear	GSTOR	$(2,\ 1,\ 0,\ 0.1)$
Tanh	GSTOR	$(0.5,\ 0.25,\ 0,\ 0.1)$
Convex	GSTOR	$(2,\ 1,\ 0,\ 0.1)$
Valley	GSTOR	$(2,\ 1,\ 0,\ 0.1)$
Peak	GSTOR	$(0.5,\ 1,\ 0,\ 0.1)$
Wave	GSTOR	$(2,\ 1,\ 0,\ 0.1)$
T-ESTOR (monotonicity-aware): (c_ν, c_τ)		
Linear	T-ESTOR	$(16,\ 0.0003)$
Tanh	T-ESTOR	$(1,\ 0.0003)$
Convex	T-ESTOR	$(256,\ 0.0003)$

Table 4: Paired final-regret differences: the predeclared structured reference minus each comparator. Negative values favor the reference. Brackets give unadjusted paired 95% percentile bootstrap intervals. The reference is T-BSTOR-Slice.

Panel	Versus T-BSTOR-Vec	Versus ZoomSIB-UCB	Versus GSTOR
Linear	$-134.2\,[-137.6, -130.5]$	$-17.9\,[-22.9, -13.4]$	$-109.5\,[-112.9, -106.0]$
Tanh	$-154.5\,[-158.5, -150.5]$	$-6.8\,[-11.9, -2.0]$	$-78.9\,[-84.6, -73.7]$
Convex	$-154.9\,[-159.9, -149.8]$	$-43.4\,[-50.2, -36.8]$	$-126.1\,[-130.9, -121.1]$
Valley	$-181.2\,[-185.0, -177.5]$	$-42.3\,[-47.5, -37.0]$	$-119.1\,[-123.4, -114.8]$
Peak	$-35.2\,[-37.1, -33.4]$	$-7.5\,[-10.0, -5.2]$	$-38.6\,[-40.4, -36.9]$
Wave	$-26.1\,[-27.0, -25.1]$	$-5.2\,[-6.6, -3.7]$	$-40.9\,[-41.8, -39.9]$

Table 5: T-BSTOR diagnostics averaged over evaluation seeds. n_0 is exploration length; m_{init} counts scalar-bin initialization rounds. The three regret columns charge exploration, initialization, and later scalar rounds. D_{bin} sums loss relative to the best candidate in the selected bin. Overflow is the percentage of binned candidate indices outside the frozen interval. Uniform-fallback rounds are excluded from the later-scalar column and retained separately in the accompanying summary.

Panel	Policy	n_0	m_{init}	R_{exp}	R_{init}	R_{later}	D_{bin}	Overflow %
Linear	T-BSTOR-Vec	4138.0	1102.0	67.2	20.6	261.6	145.5	0.000
Linear	T-BSTOR-Mat	2062.0	2072.3	33.5	40.4	112.4	53.6	0.000
Linear	T-BSTOR-Slice	2972.0	2069.6	48.2	40.0	126.9	56.5	0.000
Tanh	T-BSTOR-Vec	4018.0	1102.9	83.4	25.9	315.6	180.3	0.000
Tanh	T-BSTOR-Mat	2392.0	2070.5	49.7	50.2	120.0	61.3	0.000
Tanh	T-BSTOR-Slice	3134.0	2068.3	65.0	49.7	155.8	69.9	0.000
Convex	T-BSTOR-Vec	4198.0	2061.4	80.3	39.2	307.2	95.8	0.000
Convex	T-BSTOR-Mat	2126.0	2070.8	40.7	40.0	152.3	71.3	0.000
Convex	T-BSTOR-Slice	3054.0	2067.8	58.3	39.8	173.6	75.3	0.000
Valley	T-BSTOR-Vec	5002.0	2060.8	67.7	22.3	340.2	82.4	0.000
Valley	T-BSTOR-Mat	4920.0	2061.1	66.5	20.0	154.1	62.3	0.000
Valley	T-BSTOR-Slice	6914.0	2053.5	93.4	19.8	135.7	57.7	0.000
Peak	T-BSTOR-Vec	5102.0	1099.2	25.7	11.7	103.8	48.6	0.000
Peak	T-BSTOR-Mat	2590.0	2067.6	13.1	25.0	68.1	25.0	0.000
Peak	T-BSTOR-Slice	3632.0	1105.0	18.3	13.8	73.8	46.0	0.000
Wave	T-BSTOR-Vec	5140.0	1098.8	33.3	9.4	139.6	54.0	0.000
Wave	T-BSTOR-Mat	2518.0	1106.8	16.3	9.9	124.7	46.6	0.000
Wave	T-BSTOR-Slice	3512.0	1104.8	22.7	9.8	123.7	46.3	0.000

Table 6: Frozen-fit and bin diagnostics for T-BSTOR policies. s is the fitted Frobenius norm before unit normalization. r_{th} and r_{pr} are the rank after thresholding and after entrywise projection: matrix rank or sum of slice ranks, omitted for vectors. Bins counts occupied scalar bins. Numeric summaries average observed runs; zero and fallback columns count final zero fits and uniform-fallback runs out of all evaluation seeds. Dashes denote unavailable or inapplicable quantities. d is sign-invariant direction error averaged over valid nonzero fits; parentheses give their count.

Panel	Policy	s	r_{th}	r_{pr}	Bins	Zero	Fallback	d (count)
Linear	T-BSTOR-Vec	0.04509	–	–	12.8	0/50	0/50	0.673 (50)
Linear	T-BSTOR-Mat	0.02027	2.8	2.8	23.7	0/50	0/50	0.444 (50)
Linear	T-BSTOR-Slice	0.02395	17.4	17.4	23.8	0/50	0/50	0.481 (50)
Tanh	T-BSTOR-Vec	0.05618	–	–	12.8	0/50	0/50	0.661 (50)
Tanh	T-BSTOR-Mat	0.03267	6.6	6.6	23.7	0/50	0/50	0.413 (50)
Tanh	T-BSTOR-Slice	0.03615	24.8	24.8	23.8	0/50	0/50	0.483 (50)
Convex	T-BSTOR-Vec	0.04427	–	–	23.6	0/50	0/50	0.677 (50)
Convex	T-BSTOR-Mat	0.02051	3.1	3.1	23.8	0/50	0/50	0.445 (50)
Convex	T-BSTOR-Slice	0.02421	18.0	18.0	23.7	0/50	0/50	0.486 (50)
Valley	T-BSTOR-Vec	0.02533	–	–	23.9	0/50	0/50	0.804 (50)
Valley	T-BSTOR-Mat	0.007656	1.9	1.9	23.7	0/50	0/50	0.505 (50)
Valley	T-BSTOR-Slice	0.009414	8.4	8.4	23.7	0/50	0/50	0.478 (50)
Peak	T-BSTOR-Vec	0.02574	–	–	12.9	0/50	0/50	0.797 (50)
Peak	T-BSTOR-Mat	0.005133	1.1	1.1	24.0	0/50	0/50	0.620 (50)
Peak	T-BSTOR-Slice	0.007434	7.5	7.5	12.9	0/50	0/50	0.626 (50)
Wave	T-BSTOR-Vec	0.02476	–	–	12.8	0/50	0/50	0.777 (50)
Wave	T-BSTOR-Mat	0.004731	1.1	1.1	12.9	0/50	0/50	0.617 (50)
Wave	T-BSTOR-Slice	0.007078	7.0	7.0	12.8	0/50	0/50	0.626 (50)

J CCLE-DERIVED BANDIT EXPERIMENTS

We conduct semisynthetic bandit experiments using a fixed cell–compound–dose response block from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). T-ESTOR has the lowest mean Linear regret among the compared methods, including G-LowTESTR with the correct linear link.

J.1 DATA AND PREPROCESSING

Cohort and preprocessing. We use Broad's legacy pharmacological release dated February 24, 2015, retrieved on September 20, 2026.[1] The file contains 93,251 observed activity values for 504 cell lines, 24 compounds, and eight concentrations. Restricting to the 282 cell lines observed at every compound–dose combination gives a $282 \times 24 \times 8$ complete-case tensor without imputation. Concentrations are ordered as $0.0025, 0.008, 0.025, 0.08, 0.25, 0.8, 2.53, 8$ μM. We preserve raw reported activity, including negative values, without clipping, sign reversal, or replacement by an IC50 summary. We make no therapeutic interpretation of the response sign.

J.2 FIXED BANDIT ENVIRONMENT AND EVALUATION

Empirical coefficient and simulated feedback. We use a fixed, reproducible subset of 16 complete-case cell lines, selected before model fitting and policy comparison. Cell and compound axes are ordered lexicographically, and doses follow the concentration order above. For the resulting $16 \times 24 \times 8$ raw response block Y, set

$$\Theta_* = \frac{Y - \bar{Y}}{\|Y - \bar{Y}\|_1}, \qquad s_0 = \frac{\|\Theta_*\|_F}{\sqrt{11}} = 0.006945307344.$$

Here $\bar{Y}$ is the grand mean of the selected block, and the denominator is the entrywise ℓ_1 norm. This coefficient is not projected onto an exact low-rank class. The DCT methods use a fixed orthonormal DCT-II transform along the dose axis, based on dose order without assuming equally spaced physical concentrations. Each of the $K = 10$ fresh candidate tensors has independent entries distributed as $2\,\mathrm{Beta}(5,5) - 1$. Rewards have independent Gaussian noise with standard deviation $s_0/2$. We use

$$f_{\mathrm{Linear}}(u) = u, \qquad f_{\mathrm{Wave}}(u) = \tfrac{1}{2}u + \tfrac{1}{2}s_0\{\cos(2u/s_0) - 1\}.$$

These actions are dense measurements of a CCLE-derived coefficient, not drug prescriptions; feedback is simulated, not replayed laboratory outcomes. The base score is $S_j(x) = 8x_j/(1 - x_j^2)$. Policies receive their current menus and selected noisy feedback; coefficients, true means, and direction diagnostics remain evaluator-only.

Methods and pairing. We compare 12 methods on Linear and ten on Wave at $T = 20{,}000$ on ten shared trajectory seeds, totaling 220 trajectories. Methods use the same coefficient, scale, candidate menus, and potential-noise realizations within each seed. Cumulative pseudo-regret uses the best true mean in each realized menu and includes every exploration, initialization, and fallback round. The matrix representations unfold the cell, compound, or dose axis; Mat-Drug is the declared primary unfolding, and all three are reported.

The Gaussian sample-mean T-BSTOR variants use 4,000 uniform observations (2,000 independent pairs), 28 scalar bins, and confidence multiplier $c_{\mathrm{conf}} = 0.25$. The common scalar configuration was selected on DCT Slice and transferred to the other representations; each representation fits its own direction. The dagger (†) in the table and figure marks the DCT Slice variant using median-of-means scalar bins; the other T-BSTOR variants use the Gaussian sample-mean learner. GSTOR and ZoomSIB-UCB retain their own estimators and policies (Kang et al., 2026; Dey et al., 2026). G-LowTESTR retains tensor GLM estimation and full-coordinate LowGLM-UCB, including complementary-subspace regularization (Yi et al., 2024). The method assumes a supplied GLM link. We supply identity on both tasks: it is correct on Linear and misspecified on Wave; the true Wave link is not given to this policy. T-ESTOR and dense ESTOR use monotonicity and are evaluated only on Linear. Tuning opportunities and exploration schedules differ across families.

[1] Broad Institute distribution: CCLE legacy pharmacological data. Release counts need not equal those in the original article.

Table 7: CCLE-calibrated cumulative pseudo-regret at $T = 20{,}000$: means over ten shared trajectory seeds, with every round included (lower is better). T-ESTOR and ESTOR are monotone-only and inapplicable on Wave (—). G-LowTESTR assumes the identity link, correctly specified on Linear and misspecified on Wave. The seed panel was previously inspected; results are conditional on the fixed coefficient and settings.

Method	Linear	Wave
T-ESTOR	57.139	—
ESTOR	90.608	—
G-LowTESTR (identity link)	59.173	42.648
T-BSTOR-Slice-DCT	102.385	46.390
T-BSTOR-Slice-DCT†	110.678	49.549
T-BSTOR-Slice-Identity	107.808	50.074
T-BSTOR-Mat-Cell	101.514	45.905
T-BSTOR-Mat-Drug	106.483	48.126
T-BSTOR-Mat-Dose	104.045	46.188
T-BSTOR-Vec	122.918	54.333
GSTOR	116.139	55.453
ZoomSIB-UCB	159.043	65.639

Epoch methods. T-ESTOR uses the DCT slice estimator with $c_\eta = 16$ and $c_\tau = 0.00015$. ESTOR uses the dense unpaired estimator of Kang et al. (2026), with initial epoch size $T_0 = 50$ and coordinatewise clipping multiplier $c_\tau = 1$. Each method retains its own doubling-epoch schedule. Their last completed fits use 8,192 observations (4,096 pairs) for T-ESTOR and 6,400 unpaired observations for ESTOR; terminal partial epochs continue with the previous fit. Both methods evaluate the selected-law score using their own estimated direction, an 8,192-point numerical projection oracle, and CDF floor 10^{-12}. Numerical refinement checks do not provide a uniform tail-error certificate.

Selection and uncertainty. Development uses three seeds. Settings were fixed before each reported run; the ten-seed evaluation panel had been inspected during development and does not constitute independent confirmation. All ten seeds are retained. Bandit intervals use 2,000 paired trajectory-seed bootstrap resamples, conditional on the fixed coefficient and settings, without adjustment for multiple comparisons or adaptive development. Curve bands are pointwise, not simultaneous.

J.3 Bandit results

On Linear, T-ESTOR's mean regret is 57.139, versus 59.173 for G-LowTESTR with the correct linear link and 90.608 for ESTOR (Table 7). Its paired difference from G-LowTESTR is -2.034 with interval $[-3.373, -0.746]$: a 3.44% lower matched mean and lower regret on seven of ten seeds. Relative to ESTOR, the difference is -33.470 with interval $[-35.335, -31.775]$, a 36.94% reduction in matched mean regret. These contrasts remain descriptive under the reused-panel qualifications.

Figure 2 shows regret on both links. T-BSTOR-Slice-DCT has lower mean final regret than Vec, GSTOR, and ZoomSIB-UCB on both tasks. On Wave, its mean is 46.390, versus 55.453 for GSTOR and 65.639 for ZoomSIB-UCB, reductions of 16.3% and 29.3%. These gains support its usefulness against the two external unknown-link baselines in this setting. Cell unfolding (45.905), Dose unfolding, and identity-link G-LowTESTR (42.648) have lower means. T-ESTOR and ESTOR require monotonicity and are not applied to Wave.

G-LowTESTR's lower Wave regret does not imply knowledge of the true link. The Wave mean decomposes as $f_{\mathrm{Wave}}(u) = \frac{1}{2}u - s_0 \sin^2(u/s_0)$. Under the centrally symmetric exploration distribution, the even nonlinear term has zero linear cross-moment with the action. Thus the unpenalized population least-squares coefficient is $\Theta_*/2$, retaining the true index direction despite misspecification. This identity concerns the base sampling distribution, not the finite-sample regularized or adaptive fit, and does not guarantee the correct nonlinear action ranking. G-LowTESTR also uses 2,000 initial exploration rounds and continued online coefficient updates, whereas T-BSTOR-Slice-DCT uses 4,000 and then fixes its direction while learning scalar bins. These algorithm differences may contribute to the observed ranking; the experiment does not isolate their causal effects.

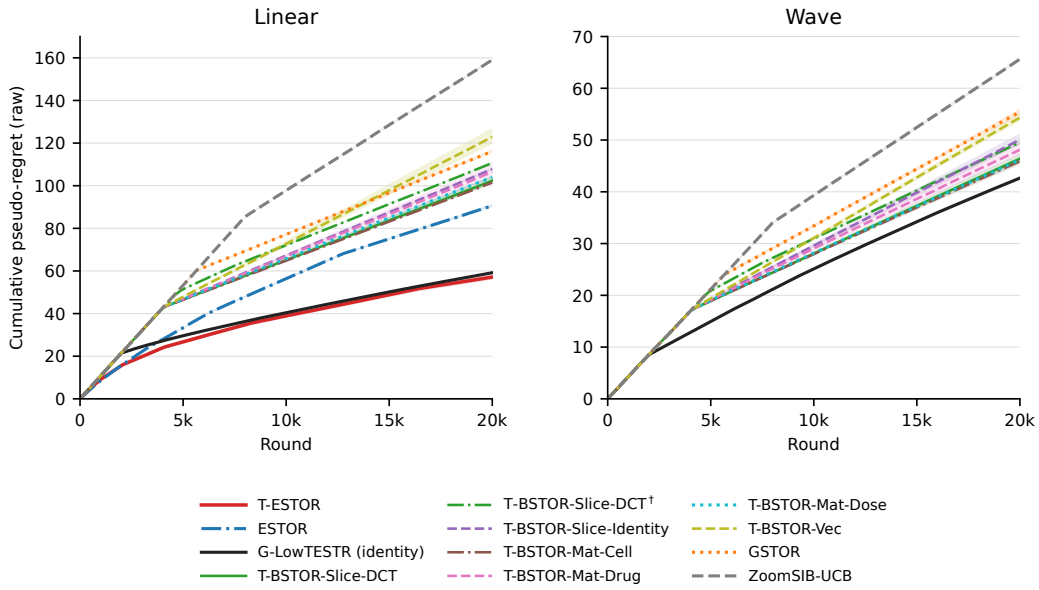


Figure 2: CCLE-calibrated mean cumulative pseudo-regret over ten shared trajectory seeds on Linear and Wave. Shading gives pointwise 95% paired-seed percentile intervals, not simultaneous bands. T-ESTOR and ESTOR are monotone-only; G-LowTESTR assumes identity on both links. Curves connect recorded cumulative-regret values. Results are conditional on the fixed coefficient and settings; intervals do not adjust for the previously inspected seed panel or adaptive configuration selection.

These results concern one complete-case CCLE coefficient that was not projected onto an exact low-rank class; they establish neither population generalization nor a uniform tensor-over-matrix advantage, and do not directly instantiate a small-rank regret theorem. Numerical scores, Gaussian sample-mean bins, and reduced confidence multipliers do not automatically inherit the exact-score and finite-variance MOM guarantees. Differences in estimators, tuning, and update schedules prevent attributing every regret gap solely to structure. The simulated feedback supports no claim about clinical decisions.