Tangent Stein Bandits under Predictable Common Shifts

Anonymous authors
Paper under double-blind review

Abstract

Common shifts in candidate features leave action rankings unchanged but can bias raw moment estimates of the reward direction. We study this problem in single-index contextual bandits with an unknown monotone link and Gaussian candidates, allowing the common shift to depend on past observations. Within-set centering removes the shift from action comparisons but correlates the candidates, invalidating the independent-winner score. We show that the selected centered action retains an exact Gaussian moment in the tangent space of the current direction. This identity enables learning without evaluating the link or its derivative for candidate sets containing at least two actions, including candidate counts where the full centered score has a boundary or second-moment obstruction. With finite conditional noise variance, certified scale bounds, and sufficient initialization, a coordinate-clipped tangent epoch algorithm achieves polylogarithmic-in-horizon high-probability pseudo-regret under fixed positive derivative-information and schedule constants. The sharper guarantee follows from a quadratic conditional-mean regret relation specific to Gaussian candidates. In 18 held-out simulation settings, continual updates improve mean pseudo-regret over an independently tuned, valid centered explore-then-commit learner; shared initialization comparisons isolate the benefit of subsequent updates. Six long-horizon comparisons also favor tangent updates over a tuned persistent uniform-exploration policy. Matched perturbations distinguish the distributional conditions for the tangent identity from the derivative information required for effective learning.

1 Introduction

Linear contextual bandits learn a reward direction under sequential partial feedback (Abbasi-Yadkori et al., 2011; Chu et al., 2011); generalized linear bandits retain a known nonlinear link (Filippi et al., 2010; Li et al., 2017). Unknown-link single-index bandits instead couple a high-dimensional latent direction with a one-dimensional reward link. At round t, the learner ranks a finite set of candidate actions using an estimate of the latent direction, observes only the selected reward, and must learn without knowing the link. This model separates representation—the direction that orders actions—from calibration—the unknown map from index to reward. Recent work shows that fixed-distribution Gaussian candidates admit Stein-based estimators and near-square-root regret for monotone links (Kang et al., 2026).

We extend the fixed-distribution model to predictable, round-varying common translations. Every action in round t shares a translation S_t:

$$X_{t,a} = S_t + Z_{t,a}, \qquad Z_{t,a} \overset{\text{ind}}{\sim} \mathcal{N}(0, I_d). \tag{1}$$

The translation is predictable from the past, and the learner need not observe it separately from the candidate set. A common state added to every option's features can change the response baseline while preserving their ranking. Raw moment estimators can therefore be biased toward the common shift rather than the reward direction. Demeaning restores translation-invariant comparisons, but the centered arms sum to zero and are correlated. Consequently, plugging their marginal Gaussian variance into the independent-winner score gives the wrong selected-arm law.

The useful structure is orthogonal to the current greedy direction. Let $H_{t,a} = X_{t,a} - K^{-1}\sum_b X_{t,b}$ be a centered action, let u be a unit direction fixed before the current action set, and choose $A_t \in$

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.

$\arg\max_a u^\top H_{t,a}$. The selected index depends only on axial Gaussian coordinates. Conditional on that selection, the orthogonal component remains Gaussian. Projecting away the axial direction before applying Gaussian integration by parts yields the exact tangent moment

$$\mathbb{E}\left[\frac{(I_d - uu^\top)H_{t,A_t}}{1 - 1/K} Y_t \middle| \mathcal{F}_{t-1}\right] = \mu_t(u)(I_d - uu^\top)\theta_\star, \tag{2}$$

where $\mu_t(u)$ is the conditional average derivative of the unknown link. The right-hand side is exactly the sphere-tangent direction that rotates u toward the truth.

Uniform exploration first obtains a positively aligned direction. The learner then holds its direction fixed within geometric epochs, averages coordinate-clipped tangent scores, and normalizes each tangent update. Gaussian candidate geometry turns angular accuracy into a quadratic conditional-mean regret bound. With sufficient initialization and fixed positive derivative-information and schedule constants, the resulting high-probability pseudo-regret grows polylogarithmically with the horizon. Orthogonal translations can be arbitrarily large; the parallel component affects derivative information. The algorithm evaluates neither f nor f'; it uses certified bounds on variance, Lipschitzness, and derivative information.

Our contributions are threefold.

1. We identify an exact tangent moment for the correlated winner induced by within-set centering. Its selected tangent component retains a Gaussian conditional law even though the candidates are dependent.
2. We use this moment for directional learning without constructing the axial winner score, including at $K = 2$ and $K = 3$, where the full-score argument encounters a boundary or second-moment obstruction, respectively.
3. We establish finite-variance learning guarantees and a Gaussian quadratic-regret refinement with polylogarithmic horizon dependence. Independently tuned, shared-initialization and persistent-exploration comparisons demonstrate continual-learning gains; matched controls separate score geometry from the distributional and information conditions supporting learning.

2 Related work

Single-index estimation. Regression identities (Li & Duan, 1989; Li, 1991), Gaussian linear recovery (Plan & Vershynin, 2016; Plan et al., 2017), and Stein scores (Yang et al., 2017a;b) recover index directions and subspaces from covariate structure. Isotron and monotone least squares estimate the link and index jointly (Kalai & Sastry, 2009; Kakade et al., 2011; Balabdaoui et al., 2019a). Our conditional Stein identity recovers a tangent moment after greedy selection.

Unknown-link single-index bandits. Kang et al. (2026) obtain near-square-root regret through Stein estimation for monotone links with independent, fixed-distribution candidates. Stable-arm models combine exogenous covariates with arm-wise or shared single-index structure (Ma & Cai, 2025; Arya & Song, 2025; Arya et al., 2026).

Dey et al. (2026) obtain $\widetilde{O}(T^{2/3})$ regret and a matching worst-case lower bound for general unknown, possibly nonmonotone links with fixed-distribution candidates. Their lower bound already uses Gaussian candidates. Corollary 2 instead gives high-probability $O(\log T \log\log T)$ regret for monotone links, fixed positive derivative information, and fixed K, d, δ, link, variance, and schedule constants, while allowing predictable common translations.

Epoched Stein's Oracle Single Index Bandit (ESTOR) of Kang et al. (2026) uses the independent selected density

$$p_u^{\text{ind}}(x) = Kp(x)F_u(u^\top x)^{K-1}, \tag{3}$$

where p is the candidate density and F_u the distribution function of $u^\top X$. Centering imposes $\sum_a H_a = 0$: for $h = su + v$, $v \perp u$, the winner density becomes $q_K(s)\phi_{\sigma_K^2 I_{d-1}}(v)$, where q_K is the axial winner density and $\sigma_K^2 = 1 - 1/K$ (Appendix A.6). We use its negative tangent log-density gradient v/σ_K^2 without evaluating q_K or q_K'/q_K, avoiding the full-score boundary term at $K = 2$ and axial second-moment divergence at $K = 3$.

Known-link interactive learning. Rajaraman & Han (2026) use projected gradients for learner-chosen unit-sphere queries with a known link and derivative. We recover a tangent direction from unknown-link rewards on finite exogenous candidate sets.

Orthogonalization against reward nuisances. Orthogonalized estimation (Krishnamurthy et al., 2018) and experimental design (Kim et al., 2025) handle action-independent additive reward confounders in linear bandits. Action centering estimates differential rewards against a baseline (Greenewald et al., 2017). Here, translation acts inside the unknown nonlinear link.

Greedy learning with stochastic contexts. Covariate diversity (Bastani et al., 2021) and smoothed contexts (Kannan et al., 2018; Raghavan et al., 2018) support greedy linear learning. Under local anti-concentration, Kim & Oh (2024) obtain polylogarithmic expected regret, including for Gaussian contexts. Our high-probability bound combines Gaussian index-gap geometry with tangent estimation for unknown links and predictable translations. Table 1 compares models and information; Appendix J develops further statistical connections.

Table 1: Models and estimator inputs. Our guarantee additionally requires the derivative-information bounds in Section 5.

Work	Actions and nuisance	Link	Estimator / required information
Kang et al. (2026)	Fixed independent draws	Unknown	Winner score from the independent candidate density and its projected distribution
Dey et al. (2026)	Fixed independent draws	Unknown	Candidate-density score; direction estimation and one-dimensional optimization
Rajaraman & Han (2026)	Learner-chosen sphere	Known	Projected gradient evaluating the link and its derivative
Krishnamurthy et al. (2018)	Observed contexts; additive nuisance	Linear	Orthogonalized estimation of linear reward differences
Kim et al. (2025)	Finite arms; additive nuisance	Linear	Experimental design with shift-invariant reward comparisons
This work	Gaussian draws; predictable shift inside link	Unknown	Centered Gaussian tangent moment; certified scale bounds, without the axial winner score

3 Problem setup

Let $\mathcal{F}_{t-1}$ contain all observations and learner randomness through round $t-1$. The unknown parameter $\theta_\star \in \mathbb{R}^d$ satisfies $\|\theta_\star\|_2 = 1$. Conditional on $\mathcal{F}_{t-1}$, the learner observes $K \geq 2$ actions

$$X_{t,a} = S_t + Z_{t,a}, \qquad Z_{t,1}, \ldots, Z_{t,K} \overset{\text{ind}}{\sim} \mathcal{N}(0, I_d), \tag{4}$$

where the common shift S_t is $\mathcal{F}_{t-1}$-measurable. After choosing A_t, the learner observes only

$$Y_t = f(\theta_\star^\top X_{t,A_t}) + \eta_t. \tag{5}$$

The learner observes the complete candidate set but need not observe S_t separately; every score below is computed from within-round centered actions. The unknown link f is continuously differentiable, nondecreasing, bounded by $|f(x)| \leq B_f$, and L_f-Lipschitz. The noise obeys

$$\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}, X_{t,1:K}, A_t] = 0, \tag{6}$$

$$\mathbb{E}[\eta_t^2 \mid \mathcal{F}_{t-1}, X_{t,1:K}, A_t] \leq \sigma_\eta^2. \tag{7}$$

This permits binary rewards by taking f as their conditional mean; no independence between the reward noise and the current action set is required.

Define the observed set mean, centered actions, centered variance, and tangent projector

$$M_t = \frac{1}{K}\sum_{a=1}^{K} X_{t,a}, \quad H_{t,a} = X_{t,a} - M_t, \quad \sigma_K^2 = 1 - \frac{1}{K}, \quad P_u = I_d - uu^\top. \tag{8}$$

Because the shift is common and f is nondecreasing, the oracle action is $a_t^\star \in \arg\max_a \theta_\star^\top H_{t,a}$. We measure pseudo-regret using conditional mean rewards

$$R_T = \sum_{t=1}^{T} \left[f(\theta_\star^\top X_{t,a_t^\star}) - f(\theta_\star^\top X_{t,A_t}) \right]. \tag{9}$$

This quantity remains random through the candidate sets and the learner's history; it does not average over the current candidate set.

The sample mean $M_t = S_t + K^{-1} \sum_a Z_{t,a}$ is independent of the complete centered Gaussian array $H_{t,1:K}$ conditional on $\mathcal{F}_{t-1}$. Thus centering removes S_t from action comparisons. It does not remove the parallel shift from the link derivative. For a direction u fixed before round t, define the predictable derivative information

$$\mu_t(u) = \mathbb{E}\left[f'(\theta_\star^\top X_{t,A_t}) \mid \mathcal{F}_{t-1} \right], \qquad A_t \in \arg\max_a u^\top H_{t,a}. \tag{10}$$

Monotonicity gives $\mu_t(u) \geq 0$. The regret theorem assumes that these averages remain positive during initialization and every tangent epoch; the algorithm does not observe them directly. Appendix B gives explicit lower bounds when the parallel shift is bounded and the link derivative is positive on a compact interval.

4 An exact tangent moment for the correlated winner

Fix a unit vector u using only information available before the current action set, choose $A = \arg\max_a u^\top H_a$, and write $H = H_A$. Demeaning correlates the arms, while the axial and tangent arrays remain independent. Figure 1 shows the centering, selection, and tangent-update loop.

Theorem 1 (Greedy tangent identity)**.** *Conditional on $\mathcal{F}_{t-1}$ and M_t,*

$$\mathbb{E}\left[\frac{P_u H_{t,A_t}}{\sigma_K^2} Y_t \middle| \mathcal{F}_{t-1}, M_t \right] = \mu_t(u, M_t) P_u \theta_\star, \tag{11}$$

where $\mu_t(u, M_t) = \mathbb{E}[f'(\theta_\star^\top (M_t + H_{t,A_t})) \mid \mathcal{F}_{t-1}, M_t] \geq 0$. Consequently,

$$\mathbb{E}[W_t(u) \mid \mathcal{F}_{t-1}] = \mu_t(u) P_u \theta_\star, \qquad W_t(u) = \frac{P_u H_{t,A_t}}{\sigma_K^2} Y_t. \tag{12}$$

The identity holds for every $K \geq 2$ and every predictable common shift.

Why selection does not spoil the tangent law. Rotate coordinates so that the first axis is u. The greedy index depends only on the axial centered coordinates, because $\arg\max_a u^\top H_a = \arg\max_a u^\top Z_a$. The complete orthogonal centered array is independent of the axial array. Conditional on the selected index and all axial coordinates,

$$P_u H_A \sim \mathcal{N}(0, \sigma_K^2 P_u), \tag{13}$$

and it is independent of M_t. Gaussian integration by parts (Stein, 1981) in this orthogonal subspace yields (11); the reward-noise term vanishes by iterated conditioning. Appendix A.1 gives the full argument.

If $c = u^\top \theta_\star > 0$, then $P_u \theta_\star$ points along the great-circle direction that increases alignment with $\theta_\star$. The unknown scalar $\mu_t(u)$ scales this direction. A tangent step followed by normalization is a sphere retraction (Bonnabel, 2013); our moment supplies the step without estimating the link or its derivative.

5 Tangent epoch algorithm and regret

We hold u_i fixed within each epoch so that the conditional score means remain aligned with $P_{u_i} \theta_\star$. Coordinate clipping adapts truncation methods for heavy-tailed bandits (Bubeck et al., 2013; Muñoz Medina & Yang, 2016; Shao et al., 2018; Xue et al., 2023) to this moment; martingale concentration controls estimation error under finite conditional noise variance.

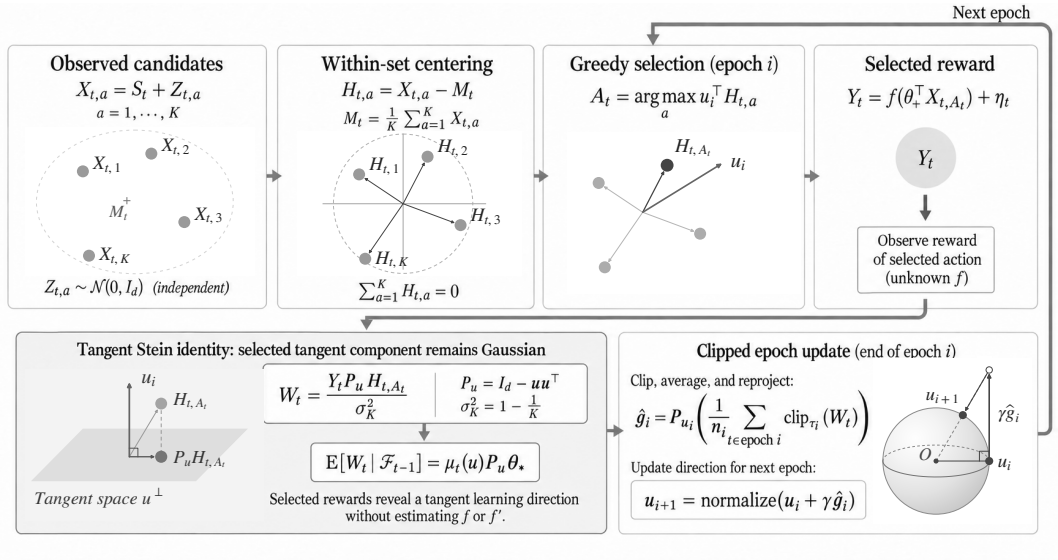

Figure 1: Overview of tangent Stein bandits. Top: observed candidates $X_{t,a} = S_t + Z_{t,a}$ share a predictable shift S_t, with independent standard Gaussian residuals $Z_{t,a}$. Subtracting the observed mean M_t gives $H_{t,a}$. The unit direction $u = u_i$ is fixed within epoch i; greedy selection maximizes $u_i^\top H_{t,a}$, after which the learner observes the selected reward Y_t under an unknown monotone link f. Bottom left: conditional on the axial coordinates, the selected tangent component remains Gaussian. With $P_u = I_d - uu^\top$ and $\sigma_K^2 = 1 - 1/K$, the reward-weighted score W_t has conditional mean $\mu_t(u)P_u\theta_\star$. Here $\mathcal{F}_{t-1}$ denotes the past, $\theta_\star$ the unknown reward direction, and $\mu_t(u)$ the conditional average link derivative. Bottom right: coordinatewise clipping at $\tau_i = \tau(n_i)$, averaging over the n_i rounds in the epoch, and reprojection produce $\widehat{g}_i$. A tangent step from u_i reaches the hollow point; radial normalization gives u_{i+1} on the unit sphere. O is its center. The outer arrow returns the updated direction to selection in the next epoch. Uniform exploration initializes u_1. Candidate locations are schematic ($K = 4$ illustrated).

For $x \in \mathbb{R}$, define $\mathrm{clip}_\tau(x) = \mathrm{sign}(x)\min\{|x|, \tau\}$ and apply it coordinatewise to vectors. Let

$$V_K = \frac{B_f^2 + \sigma_\eta^2}{\sigma_K^2}, \qquad \overline{V}_K \geq V_K, \qquad C_F = 1 + \sqrt{2} + \frac{2}{3}. \tag{14}$$

Lemma 1 (Clipped tangent mean)**.** *Within an epoch of n rounds with fixed u, set $\tau = \sqrt{\overline{V}_K n/\ell}$, average the coordinate-clipped scores $W_t(u)$, and project the average by P_u. For every $\ell > 0$, with probability at least $1 - 2d\exp(-\ell)$,*

$$\|\widehat{g} - \bar{\mu} P_u \theta_\star\|_2 \leq C_F \sqrt{\frac{\overline{V}_K d\ell}{n}}, \tag{15}$$

where $\bar{\mu} = n^{-1}\sum_t \mu_t(u)$. The same bound holds for the full centered score during uniform initialization.

The proof combines an $\overline{V}_K/\tau$ coordinatewise clipping-bias bound with a two-sided Freedman inequality (Freedman, 1975) for clipped martingale differences. Projection restores exact tangency after coordinate clipping and cannot enlarge the error.

Each round costs $O(Kd)$ for centering and selection and $O(d)$ for the score. The projection $P_u v = v - u(u^\top v)$ uses vector operations; clipped epoch sums require $O(d)$ persistent state beyond the current candidate set.

Uniform initialization avoids the antipodal stationary point. If its average derivative is at least $\mu_{\text{init}} > 0$ and its moment error is at most $\mu_{\text{init}}/3$, normalization gives

$$\sin\angle(u_1, \theta_\star) \leq \frac{1}{2}, \qquad u_1^\top \theta_\star \geq c_0 := \frac{\sqrt{3}}{2}. \tag{16}$$

Algorithm 1 Tangent epoch algorithm

Require: Initial length n_0, growth $b > 1$, step size γ, horizon T, confidence δ, and a certified bound $\overline{V}_K \geq V_K$

1: Set $N_T = 2 + \lceil \log_b \max\{T/n_0, 1\} \rceil$, $\ell_T = \log(4dN_T/\delta)$, and $\tau(n) = \sqrt{\overline{V}_K n/\ell_T}$.
2: For n_0 rounds, choose actions uniformly and independently of the current residuals; form $H_{t,A_t} Y_t/\sigma_K^2$.
3: Coordinate-clip these full centered scores at $\tau(n_0)$, average, and normalize the result to obtain u_1.
4: **for** epochs $i = 1, 2, \ldots$ with $n_i = \lceil n_0 b^{i-1} \rceil$ **do**
5: Hold u_i fixed and choose $A_t \in \arg\max_a u_i^\top H_{t,a}$ throughout the epoch.
6: Form $W_t(u_i)$ from (12), coordinate-clip at $\tau(n_i)$, average, and set $\widehat{g}_i = P_{u_i} \overline{W}_i$.
7: If the epoch is complete, set $u_{i+1} = \text{normalize}(u_i + \gamma \widehat{g}_i)$; otherwise stop at T.
8: **end for**

Assume that every tangent epoch has average predictable derivative at least $\mu_{\text{low}} > 0$. Let $\overline{L}_f \geq L_f$ be a certified Lipschitz upper bound. Since differentiability and monotonicity imply $0 \leq f' \leq L_f$, choose

$$0 < \gamma \leq \overline{L}_f^{-1}, \quad \rho = 1 - \gamma \mu_{\text{low}} c_0, \quad 1 < b < \rho^{-2}. \tag{17}$$

The last inequality couples epoch growth to angular contraction. Algorithm 1 takes $(n_0, b, \gamma, T, \delta, \overline{V}_K)$ without evaluating f, f', or $\mu_t(u)$. Certified bounds determine admissible settings; Appendix B supplies a positive-information subclass. Experiments use separately tuned initializations and fixed schedules; derivative information is recorded as a post-run diagnostic (Appendix F).

Theorem 2 (Finite-horizon pointwise-gap bound)**.** *Fix $\delta \in (0, 1)$ and an integer $1 \leq n_0 \leq T$. Define*

$$N_T = 2 + \lceil \log_b \max\{T/n_0, 1\} \rceil, \tag{18}$$

$$\ell_T = \log(4dN_T/\delta), \qquad E_T = C_F \sqrt{\overline{V}_K d \ell_T}, \tag{19}$$

$$B_T = \sqrt{\sigma_K^2} \left[\sqrt{d} + \sqrt{2 \log(2KT/\delta)} \right], \tag{20}$$

$$A_T^{\text{ceil}} = \sqrt{2} E_T \left[\frac{3}{2\mu_{\text{init}}} + \frac{2\gamma\sqrt{b}}{1 - \rho\sqrt{b}} \right], \tag{21}$$

$$C_b^{\text{ceil}} = 1 + \frac{\sqrt{2b}}{\sqrt{b} - 1}. \tag{22}$$

Under the model and information conditions above, suppose

$$n_0 \geq \frac{9E_T^2}{\mu_{\text{init}}^2}, \qquad n_0 \geq \frac{16E_T^2}{\mu_{\text{low}}^2 c_0^2}. \tag{23}$$

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

$$R_T \leq 2B_f n_0 + 2L_f B_T C_b^{\text{ceil}} A_T^{\text{ceil}} \sqrt{T}, \tag{24}$$

with initialization cost displayed explicitly.

Corollary 1 (A sufficient initialization choice)**.** *Set $\overline{N}_T = 2 + \lceil \log_b \max\{T, 1\} \rceil$, $\overline{\ell}_T = \log(4d\overline{N}_T/\delta)$, and*

$$n_0^*(T) = \left\lceil C_F^2 \overline{V}_K d \overline{\ell}_T \max\left\{ \frac{9}{\mu_{\text{init}}^2}, \frac{16}{\mu_{\text{low}}^2 c_0^2} \right\} \right\rceil. \tag{25}$$

When $n_0^*(T) \leq T$, choosing $n_0 = n_0^*(T)$ in Theorem 2 gives $R_T = \widetilde{O}(d\sqrt{T})$, treating the information, variance, link, and schedule constants as fixed. If $n_0^*(T) > T$, uniform play throughout instead gives $R_T \leq 2B_f T$.

Proof. For any $n_0 \geq 1$, $N_T \leq \overline{N}_T$, so (25) implies both inequalities in (23) without an implicit dependence on n_0. Moreover $n_0^*(T) = O(d\overline{\ell}_T)$ and $E_T, A_T^{\text{ceil}}, B_T = \widetilde{O}(\sqrt{d})$, so $B_T A_T^{\text{ceil}} \sqrt{T} = \widetilde{O}(d\sqrt{T})$. The bound $|f| \leq B_f$ gives the stated small-horizon fallback. □

Exploiting the Gaussian regret geometry. Let $m_K = \mathbb{E}\max_{a\le K} G_a$ for independent $G_a \sim \mathcal{N}(0,1)$. For a predictable unit direction u, put $c = u^\top \theta_\star$ and $A = \arg\max_a u^\top H_{t,a}$. Gaussian rotational invariance gives

$$\mathbb{E}[\theta_\star^\top (H_{t,a_t^\star} - H_{t,A}) \mid \mathcal{F}_{t-1}] = m_K(1-c). \tag{26}$$

Writing r_t for the summand in (9), monotonicity and Lipschitzness therefore imply, whenever $c \ge c_0$, $\mathbb{E}[r_t \mid \mathcal{F}_{t-1}] \le L_f m_K \sin^2 \angle(u,\theta_\star)/(1+c_0)$. This is a conditional-mean bound over the current candidates.

Corollary 2 (Gaussian quadratic-regret refinement)**.** *Under the conditions of Theorem 2, let J_T be the number of greedy epochs visited, including a final partial epoch. With probability at least $1-\delta$,*

$$R_T \le 2B_f n_0 + \frac{(e-1)L_f m_K J_T}{1+c_0}(A_T^{\text{ceil}})^2 + 2B_f \log\frac{2}{\delta}. \tag{27}$$

With $n_0 = n_0^*(T) \le T$ and fixed information, variance, link, and schedule constants, the bound has polylogarithmic horizon dependence. For fixed K, d, δ, it is $O(\log T \log\log T)$ as $T \to \infty$.

The clipped-mean event controls the predictable sum by $L_f m_K J_T (A_T^{\text{ceil}})^2/(1+c_0)$. A nonnegative exponential supermartingale (Howard et al., 2020) transfers this to the random pseudo-regret, giving (27). Appendix A.5 proves the transfer and displays the dimension and information dependence. The fixed constants include a strictly positive schedule margin $1-\rho\sqrt{b}$.

Shared angle analysis and the pointwise bound. The clipped-mean error $E_T/\sqrt{n_i}$, noiseless contraction ρ, and 2-Lipschitz normalization yield

$$s_{i+1} \le \rho s_i + \frac{2\gamma E_T}{\sqrt{n_i}}, \qquad s_i = \sin\angle(u_i,\theta_\star). \tag{28}$$

Condition (17) gives $s_i \le A_T^{\text{ceil}}/\sqrt{n_i}$ and preserves the positive hemisphere. A Gaussian norm bound (Laurent & Massart, 2000) gives pointwise regret at most $2L_f B_T s_i$. Summing the integer epochs, including a final partial epoch, proves (24); details are in Appendix A.4.

6 Score misspecification and small-K obstructions

The tangent projection resolves three distinct obstructions in the translated-action geometry.

Raw moments inherit the common shift. Take $\theta_\star = e_1$, $S = Be_2$, and choose an action uniformly during exploration. For $Z \sim \mathcal{N}(0, I_d)$,

$$\mathbb{E}[(S+Z)f(e_1^\top(S+Z))] = e_1\mathbb{E}[f'(Z_1)] + Be_2\mathbb{E}[f(Z_1)]. \tag{29}$$

For a noncentered positive link such as the logistic function, the second term grows linearly in B although the within-round ranking is unchanged. Thus a raw score can converge toward an irrelevant orthogonal shift.

Demean then reuse assumes the wrong winner law. Every H_a is marginally Gaussian with covariance $\sigma_K^2 I_d$, but the array satisfies $\sum_a H_a = 0$. A score derived for independent Gaussian arms therefore does not become valid by substituting the marginal variance. The constant-reward control exposes this mismatch: the correct directional moment is zero, whereas the misspecified moment is nonzero (Appendix E).

Small-K obstructions for the exact full score. For $s = u^\top H_A$, the centered winner density near zero is proportional to s^{K-2} (Appendix A.6). Thus the full score has:

- $K = 2$ a nonzero integration-by-parts boundary term;
- $K = 3$ an infinite axial second moment;
- $K \ge 4$ a finite axial second moment.

Removing the axial score makes Theorem 1 valid throughout $K \ge 2$. At $K = 4$, the full-score and tangent controls perform comparably (Section 7).

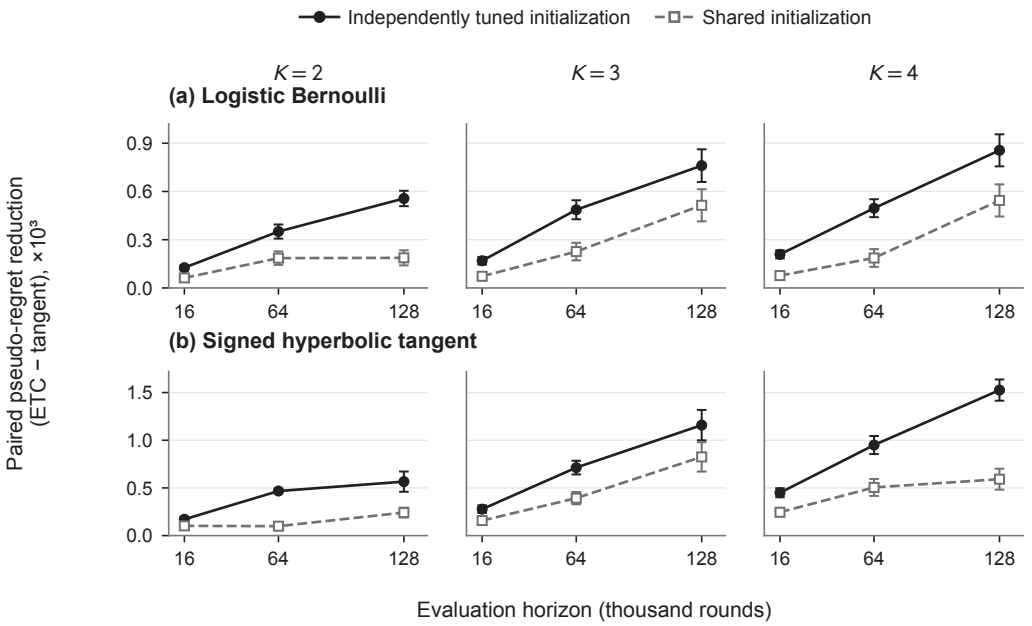

Figure 2: Continual tangent updates versus centered explore-then-commit (ETC). Rows show the two reward environments; columns give candidate count K. Points are mean paired pseudo-regret reductions, ETC minus tangent, over 48 held-out replicates at dimension six; error bars are ± 1.96 paired standard errors. Solid circles use initialization budgets independently selected for the two learners on 12 separate tuning replicates. Open squares and dashed lines use ETC's tuning-selected budget for both learners and the same clipped direction estimate, isolating subsequent updating. Positive values favor tangent updates. Horizons and regret reductions are in thousands; lines connect the three evaluated horizons. Vertical scales are shared within each reward environment.

7 Experiments

We evaluate Tangent Stein Bandits (TSB) through policy comparisons that test continual updating and controls that isolate score geometry. Within each comparison, methods share random draws for candidate residuals, potential rewards, and uniformly sampled action indices. Numerical $\pm$ values denote one standard error across independent paired replicates.

Centered explore-then-commit. Centered explore-then-commit (ETC) uses the same clipped uniform score $H_{t,A_t} Y_t / \sigma_K^2$ as our initialization and then fixes its estimated direction. It observes neither the common shift nor the link. We compare logistic Bernoulli and signed hyperbolic-tangent rewards, $d = 6$, $K \in \{2, 3, 4\}$, and three horizons from 16,000 to 128,000. Both methods independently select their initial length from the same seven-point grid using 12 tuning replicates; 48 disjoint paired replicates evaluate the selected settings. TSB uses epoch growth $b = 1.5$, with step size $\gamma = 4$ for logistic Bernoulli rewards and $\gamma = 1$ for signed hyperbolic-tangent rewards. Appendix C reports the protocol, complete comparisons, and sensitivity analysis.

All 18 independently tuned mean paired differences favor continual updates (Figure 2). At $K = 4, T = 128{,}000$, the paired ETC-minus-tangent reductions are 855.92 ± 50.83 for logistic and $1{,}525.88 \pm 56.76$ for signed rewards. Holding both initializations at ETC's tuning-selected length gives reductions of 544.32 ± 50.95 and 591.49 ± 56.07, respectively. Identical initial estimates make the latter comparison isolate subsequent updating (Table 5). Figure 3 shows the regret tradeoff across all seven initialization budgets, and Table 6 reports step-size and epoch-growth sensitivity. Appendix C gives cost decompositions and short-initialization tail variation.

Continually refreshed uniform exploration. Centered persistent exploration (CPE) starts with 128 uniform actions and continues random exploration at a fixed probability. At geometric calen-

Table 2: Dimension comparison with logistic Bernoulli rewards, common shifts, $K = 4$, and $T = 128{,}000$. Entries are mean total pseudo-regret $\pm$ one standard error over 48 paired replicates; each method selects among eight budgets on 12 separate tuning replicates. TSB denotes Tangent Stein Bandits, ETC centered explore-then-commit, and CPE centered persistent exploration.

d	TSB	ETC	CPE
6	197.7 ± 4.7	970.0 ± 56.2	$1{,}973.8 \pm 63.1$
12	367.7 ± 6.0	$1{,}425.3 \pm 43.0$	$2{,}872.9 \pm 50.9$
24	651.2 ± 6.8	$2{,}019.7 \pm 45.1$	$4{,}003.7 \pm 72.1$

dar checkpoints, it re-clips all accumulated uniform scores and reestimates its direction, discarding greedy-round rewards. It selects the expected exploration budget on 12 tuning replicates from the same seven-point grid. Table 7 uses 48 new paired replicates per cell at $T = 128{,}000$, retaining the original TSB and ETC settings. TSB has the lowest mean pseudo-regret in all six cells, followed by ETC and then CPE. The comparison evaluates complete policies, including exploration timing and update rules; Appendix D gives the protocol, selected budgets, and paired intervals.

Increasing dimension. With logistic Bernoulli rewards, common shifts, and $K = 4$, we independently retune all three policies at $d = 6, 12, 24$ using equal eight-point grids. TSB has lower regret than both baselines at every dimension (Table 2); all six paired intervals lie above zero. Appendix D.1 gives the protocol and paired results.

Direct comparison without shifts. Setting $S_t = 0$ makes the original ESTOR (Kang et al., 2026) applicable to the raw independent candidates. With equal tuning budgets, TSB has lower regret in both signed-reward cells. For logistic rewards, ESTOR has lower regret at $K = 3$, while the paired interval at $K = 4$ includes zero (Table 14). Appendix H gives the paired results and protocol.

Mechanism controls and scope. The paired score controls use identical schedules. Raw shift-unaware and demean-then-independent scores test the effect of score misspecification. With 4,096 initialization rounds, the raw shift-unaware logistic score incurs 50.22 times tangent regret at 256,000 rounds. With 2,048 initialization rounds and signed rewards, the demean-then-independent score incurs 39.53 times tangent regret at 128,000 rounds. The exact full centered score accounts for candidate dependence. At $K = 4$, its regret is comparable to that of the tangent learner (547.77 ± 2.46 versus 548.20 ± 3.87). Tangent estimation achieves this performance without constructing the axial winner score and remains valid at $K = 2$ and $K = 3$ (Section 6).

Additional checks cover 21 Gaussian identity cases ($K = 2$–16, $d = 2$–50) and ten online cases across logistic, hyperbolic-tangent, and arctangent links. Matched perturbations expose failures under non-Gaussian residuals, action-specific translations, and translations chosen from current residuals. Parallel saturation reduces derivative information, whereas orthogonal shifts leave regret unchanged (Appendices E–G).

8 Limitations and conclusion

The selected winner's Gaussian tangent law enables unknown-link directional learning under predictable common translations, including the small-K cases that obstruct a full-score analysis. Clipped estimation and quadratic Gaussian regret geometry give polylogarithmic high-probability pseudo-regret. The simulations show gains over commitment and the tested persistent uniform-exploration policy; the exact full-score control performs comparably at $K = 4$.

The guarantee requires Gaussian residuals, a common shift fixed before current residuals, positive derivative information, certified scale bounds, and sufficient initialization. Its rate is instance-dependent, not uniform over links with vanishing information. Matched perturbations separate violations of the tangent identity from saturation that reduces information. Valid tangent scores for other residual families and an observable replacement for the information lower bound are natural next steps.

AI USE STATEMENT

Generative AI tools assisted with retrieving relevant literature, revising the manuscript for clarity and consistency, refining experimental protocols, implementing and analysing simulation experiments, and developing mathematical derivations and proofs. Experimental results were obtained from executable simulations and numerical analysis. Verification included source checks, simulator and controller cross-checks, and computational checks of the reported probability laws and risk bounds. The authors take responsibility for the final content.

Reproducibility statement. The appendices contain the proofs, experimental protocols, parameter settings, and additional results. Appendix I summarizes random seed conventions, computation, and the construction of uncertainty estimates. Upon acceptance, we will release the simulation and evaluation code under an open-source license, together with configuration files, random seeds, and per-run results.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3691–3699. PMLR, 2021.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3) of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2013.

Sakshi Arya and Hyebin Song. Batched single-index global multi-armed bandits with covariates. *arXiv preprint arXiv:2503.00565*, 2025. URL https://arxiv.org/abs/2503.00565.

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

Fadoua Balabdaoui, Cécile Durot, and Hanna Jankowski. Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310, 2019a. doi: 10.3150/18-BEJ1090.

Fadoua Balabdaoui, Piet Groeneboom, and Kim Hendrickx. Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46(2):517–544, 2019b. doi: 10.1111/sjos.12361.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. doi: 10.1287/mnsc.2020.3605.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. doi: 10.1109/TIT.2013.2277869.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011.

Ran Dai, Hyebin Song, Rina Foygel Barber, and Garvesh Raskutti. Convergence guarantee for the sparse monotone single index model. *Electronic Journal of Statistics*, 16(2):4449–4496, 2022. doi: 10.1214/22-EJS2046.

Devdan Dey, Sujoy Bhore, and Avishek Ghosh. Optimal regret for single index bandits. *arXiv preprint arXiv:2605.09454*, 2026. URL https://arxiv.org/abs/2605.09454.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and Thompson sampling. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1585–1593. PMLR, 2021.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 2020.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 546–580. PMLR, 2022.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3 (1):230–261, 2013. doi: 10.1287/11-SSY032.

Kristjan Greenewald, Ambuj Tewari, Predrag Klasnja, and Susan Murphy. Action centered contextual bandits. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5973–5981, 2017.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989. doi: 10.1080/01621459.1989.10478863.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

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

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. *Operations Research*, 70(6):3261–3281, 2022. doi: 10.1287/opre.2021.2237.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sham M. Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Adam Tauman Kalai and Ravi Sastry. The Isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

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 *International Conference on Learning Representations*, 2026.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Seok-Jin Kim and Min-hwan Oh. Local anti-concentration class: Logarithmic regret for greedy linear contextual bandit. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Seok-Jin Kim, Gi-Soo Kim, and Min-hwan Oh. Experimental design for semiparametric bandits. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3215–3252. PMLR, 2025.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2776–2785. PMLR, 2018.

John Langford and Tong Zhang. The Epoch-Greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

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

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

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 2017.

Wanteng Ma and T. Tony Cai. Nonparametric bandits with single-index rewards: Optimality and adaptivity. *arXiv preprint arXiv:2512.24669*, 2025. URL https://arxiv.org/abs/2512.24669.

Andrés Muñoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1642–1650. PMLR, 2016.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013. doi: 10.1214/13-AOS1101.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016. doi: 10.1214/15-AOS1381.

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.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017. doi: 10.1093/imaiai/iaw015.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1724–1738. PMLR, 2018.

Nived Rajaraman and Yanjun Han. Interactive learning of single-index models via stochastic gradient descent. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lZY0uluCzl.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 54–66, 2010.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022. doi: 10.1287/moor.2021.1193.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70880–70891. Curran Associates, Inc., 2023. doi: 10.52202/075280-3105.

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, 2017a.

Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, and Han Liu. Estimating high-dimensional non-Gaussian multiple index models via Stein's lemma. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Chenlu Ye, Yujia Jin, Alekh Agarwal, and Tong Zhang. Catoni contextual bandits are robust to heavy-tailed rewards. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71998–72040. PMLR, 2025.

A Proofs

A.1 Conditional Gaussian decomposition and tangent identity

We first make the conditioning in Theorem 1 explicit. Fix a round and suppress its index. Conditional on $\mathcal{F}_{t-1}$, S is fixed and $Z_1, \dots, Z_K$ are independent standard Gaussian vectors. The Gaussian sample mean $\bar{Z} = K^{-1}\sum_a Z_a$ is independent of the centered array $H_a = Z_a - \bar{Z}$. Hence $M = S + \bar{Z}$ is independent of $H_{1:K}$.

Fix a unit vector u before observing the current residuals and decompose

$$H_a = u\alpha_a + V_a, \qquad \alpha_a = u^\top H_a, \qquad V_a = P_u H_a. \tag{30}$$

The Gaussian arrays $\alpha_{1:K}$ and $V_{1:K}$ are independent. The selected index $A = \arg\max_a \alpha_a$ is measurable with respect to the axial array. Conditional on A, the full axial array, and M, the vector V_A is therefore Gaussian with mean zero and covariance $\sigma_K^2 P_u$. Although the vectors $V_1, \dots, V_K$ are mutually correlated, this marginal conditional law is all that the proof needs.

Write $r = \theta_\star^\top M + (u^\top\theta_\star)\alpha_A$. Gaussian integration by parts (Stein, 1981) in the range of P_u gives

$$\mathbb{E}[V_A f(r + \theta_\star^\top V_A) \mid \mathcal{F}_{t-1}, M, \alpha_{1:K}, A] = \sigma_K^2 P_u\theta_\star \mathbb{E}[f'(r + \theta_\star^\top V_A) \mid \mathcal{F}_{t-1}, M, \alpha_{1:K}, A]. \tag{31}$$

For the noise component, the tower property and (6) imply

$$\mathbb{E}[V_A\eta \mid \mathcal{F}_{t-1}, M] = 0. \tag{32}$$

Averaging (31) over the axial array proves (11); averaging over M proves (12). Monotonicity gives nonnegative $\mu_t(u, M)$ and $\mu_t(u)$. Ties have probability zero.

During uniform initialization, choose A uniformly and independently of the current residual array. Conditional on $\mathcal{F}_{t-1}$, $H_A \sim \mathcal{N}(0, \sigma_K^2 I_d)$ and H_A is independent of M. Applying the same argument in all coordinates gives

$$\mathbb{E}\left[\frac{H_A Y}{\sigma_K^2}\middle|\mathcal{F}_{t-1}\right] = \mu_{0,t}\theta_\star \tag{33}$$

for a nonnegative predictable derivative coefficient $\mu_{0,t}$.

A.2 Concentration of clipped score means

For any coordinate vector e_j, conditional mean zero of η_t gives

$$\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}, X_{t,1:K}, A_t] \le B_f^2 + \sigma_\eta^2. \tag{34}$$

The selected tangent coordinate has conditional second moment at most σ_K^2. Therefore

$$\mathbb{E}[(e_j^\top W_t(u))^2 \mid \mathcal{F}_{t-1}] \le \frac{B_f^2 + \sigma_\eta^2}{\sigma_K^2} = V_K. \tag{35}$$

The same inequality holds for every coordinate of the uniform-initialization score.

For a scalar random variable Q with second moment at most V_K,

$$|\mathbb{E}[Q - \mathrm{clip}_\tau(Q)]| \le \mathbb{E}[Q^2]/\tau \le V_K/\tau. \tag{36}$$

Set $\tau = \sqrt{\overline{V}_K n/\ell}$. Since $V_K \le \overline{V}_K$, the centered clipped martingale differences have conditional variance at most $\overline{V}_K$ and absolute value at most 2τ. A two-sided Freedman inequality (Freedman, 1975), followed by a union bound over d coordinates, bounds each coordinate's clipping bias plus stochastic error by $(1 + \sqrt{2} + 2/3)\sqrt{\overline{V}_K \ell/n}$ with failure probability at most $2d\exp(-\ell)$. Taking the Euclidean norm yields (15). Applying P_u after averaging cannot increase this error.

A.3 Initialization and angular contraction

Let the average uniform-initialization population moment be $\bar{\mu}_0\theta_\star$ with $\bar{\mu}_0 \ge \mu_{\text{init}}$. If the estimated moment is $\bar{\mu}_0\theta_\star + e_0$ and $\|e_0\|_2 \le \epsilon_0 \le \mu_{\text{init}}/3$, then

$$\sin\angle(u_1, \theta_\star) \le \frac{\epsilon_0}{\mu_{\text{init}} - \epsilon_0} \le \frac{1}{2}. \tag{37}$$

The inner product of the unnormalized moment with $\theta_\star$ is positive, so $u_1^\top\theta_\star \ge \sqrt{3}/2 = c_0$.

For a tangent epoch, put $c = u^\top\theta_\star$, $s = \sqrt{1-c^2}$, and $g = \mu P_u\theta_\star$. Direct calculation gives

$$\sin\angle(\text{normalize}(u + \gamma g), \theta_\star) = \frac{s(1-\gamma\mu c)}{\sqrt{1 + (\gamma\mu s)^2}}. \tag{38}$$

When $c \ge c_0$, $\mu \ge \mu_{\text{low}}$, and $\gamma\mu \le 1$, this is at most ρs.

Now let $\widehat{g} = g + e$ with $e \perp u$ and $\|e\|_2 \le \epsilon$. Both $u + \gamma g$ and $u + \gamma\widehat{g}$ have norm at least one. Normalization is therefore 2-Lipschitz between them, giving

$$s_+ \le \rho s + 2\gamma\epsilon. \tag{39}$$

If $s \le 1/2$ and $\epsilon \le \mu_{\text{low}} c_0/4$, the right-hand side is at most $1/2$. The unnormalized update retains positive inner product with $\theta_\star$, so the normalized update remains in the same hemisphere and has cosine at least c_0.

A.4 Geometric epochs and regret

Let $\mathcal{G}$ be the event that all clipped means used to choose a subsequent direction obey Lemma 1. At most N_T estimators are used, so $\Pr(\mathcal{G}) \ge 1 - 2dN_T\exp(-\ell_T) = 1 - \delta/2$. The following angular bounds hold on $\mathcal{G}$, without any bound on the realized action norms.

The initialization conditions (23) imply both $\epsilon_0 \le \mu_{\text{init}}/3$ and $E_T/\sqrt{n_i} \le \mu_{\text{low}} c_0/4$ for every tangent epoch. Combining Lemma 1 with (39) gives

$$s_{i+1} \le \rho s_i + 2\gamma E_T n_i^{-1/2}. \tag{40}$$

Because $n_i = \lceil n_0 b^{i-1} \rceil$ and $n_0 \ge 1$,

$$n_0 b^{i-1} \le n_i \le 2n_0 b^{i-1}. \tag{41}$$

Iterating (40), using $\rho\sqrt{b} < 1$, and converting between n_i and its real geometric envelope gives

$$s_i \le \frac{A_T^{\text{ceil}}}{\sqrt{n_i}}, \tag{42}$$

with A_T^{ceil} in (21). The factor $\sqrt{2}$ follows from the integer-ceiling bound.

To prove Theorem 2, additionally bound all centered-action norms by B_T. A Gaussian norm tail (Laurent & Massart, 2000) and union bound over at most KT actions cost at most $\delta/2$, so the intersection has probability at least $1-\delta$. For any round in epoch i, let $a^\star$ maximize the latent index and A maximize $u_i^\top H_a$. The greedy inequality and positive hemisphere give

$$\theta_\star^\top (H_{a^\star} - H_A) \le (P_{u_i}\theta_\star)^\top (H_{a^\star} - H_A) \tag{43}$$

$$\le 2B_T s_i. \tag{44}$$

Lipschitzness of f turns this into a reward gap at most $2L_f B_T s_i$. Summing (42) over complete epochs and a final partial epoch yields

$$\sum_i n_i s_i \le C_b^{\text{ceil}} A_T^{\text{ceil}} \sqrt{T}, \tag{45}$$

where $C_b^{\text{ceil}} = 1 + \sqrt{2b}/(\sqrt{b}-1)$. Uniform initialization costs at most $2B_f n_0$. Combining these terms proves Theorem 2.

A.5 Gaussian quadratic regret and high-probability transfer

Lemma 2 (Gaussian conditional index gap)**.** *Under* (4)*, let u be an $\mathcal{F}_{t-1}$-measurable unit vector, $c = u^\top \theta_\star$, and $A = \arg\max_a u^\top H_{t,a}$. Then* (26) *holds. If $c \ge c_0$, the predictable reward regret $a_t := \mathbb{E}[r_t \mid \mathcal{F}_{t-1}]$ satisfies*

$$0 \le a_t \le L_f m_K (1-c) = \frac{L_f m_K (1-c^2)}{1+c} \le \frac{L_f m_K}{1+c_0} \sin^2 \angle(u, \theta_\star). \tag{46}$$

Proof. Centering and the common translation cancel in action differences and leave the maximizers unchanged. Conditional on $\mathcal{F}_{t-1}$, rotational invariance gives $\mathbb{E}\max_a \theta_\star^\top Z_{t,a} = m_K$. Decompose $\theta_\star = cu + P_u\theta_\star$. Selection by u depends only on the axial coordinates $u^\top Z_{t,1:K}$, which are independent of their orthogonal Gaussian components. The selected orthogonal component has conditional mean zero, while the selected axial coordinate has mean m_K. Thus $\mathbb{E}[\theta_\star^\top Z_{t,A} \mid \mathcal{F}_{t-1}] = cm_K$, proving the index-gap identity. The latent index gap is nonnegative. Monotonicity and Lipschitzness bound r_t by L_f times this gap, and $1-c = (1-c^2)/(1+c)$ proves (46). □

Proof of Corollary 2. On the clipped-mean event $\mathcal{G}$ of Appendix A.4, $c_i = u_i^\top \theta_\star \ge c_0$ and $s_i \le A_T^{\text{ceil}}/\sqrt{n_i}$ for every visited epoch. If epoch i uses $m_i \le n_i$ rounds, including a possible final partial epoch, Lemma 2 gives

$$\sum_{t=n_0+1}^{T} a_t \le \frac{L_f m_K}{1+c_0} \sum_{i=1}^{J_T} m_i s_i^2 \le \frac{L_f m_K J_T}{1+c_0} (A_T^{\text{ceil}})^2. \tag{47}$$

To transfer the predictable-sum bound to realized pseudo-regret, we use the exponential-supermartingale method (Howard et al., 2020). Since $0 \le r_t \le 2B_f$, set $z_t = r_t/(2B_f)$, $p_t = \mathbb{E}[z_t \mid \mathcal{F}_{t-1}]$ for $t > n_0$. The positive-information assumption implies $B_f > 0$. Convexity gives $e^z \le 1 + (e-1)z$ for $z \in [0,1]$, and hence

$$\mathbb{E}[e^{z_t} \mid \mathcal{F}_{t-1}] \le 1 + (e-1)p_t \le \exp((e-1)p_t). \tag{48}$$

Consequently

$$L_s = \exp\left(\sum_{t=n_0+1}^{s} z_t - (e-1) \sum_{t=n_0+1}^{s} p_t \right), \qquad s \geq n_0, \tag{49}$$

is a nonnegative supermartingale with $L_{n_0} = 1$. Markov's inequality at the fixed horizon T implies, with probability at least $1 - \delta/2$,

$$\sum_{t=n_0+1}^{T} r_t \leq (e-1) \sum_{t=n_0+1}^{T} a_t + 2B_f \log(2/\delta). \tag{50}$$

This probability statement is established without conditioning on $\mathcal{G}$. A union bound with $\Pr(\mathcal{G}^c) \leq \delta/2$ and the deterministic initialization cost $2B_f n_0$ proves (27).

For $n_0 = n_0^*(T) \leq T$, fix $\mu_{\text{init}}, \mu_{\text{low}}, \overline{V}_K, \overline{L}_f, \gamma, b$ with $1 - \rho\sqrt{b} > 0$. Then $n_0^* = O(d\bar{\ell}_T)$, $(A_T^{\text{ceil}})^2 = O(d\bar{\ell}_T)$, and $J_T = O(1 + \log T)$. For $\lambda > 0$, the Gaussian moment-generating function gives $m_K \leq \log K/\lambda + \lambda/2$; optimizing yields $m_K \leq \sqrt{2 \log K}$. Thus the bound is

$$O\left(C_{\text{inst}} d \log \frac{d(1+\log T)}{\delta} (1 + \log T) \sqrt{\log K} + B_f \log \frac{2}{\delta} \right), \tag{51}$$

where C_{inst} retains the fixed information, variance, link, and schedule factors. This gives $O(\log T \log\log T)$ horizon dependence for fixed K, d, δ. When $n_0^*(T) > T$, the uniform-play bound $2B_f T$ still applies. □

A.6 Centered-winner boundary

Let $G_m(r)$ be the probability that the maximum of the centered array formed from m independent standard Gaussian scalars is at most r. For the selected axial coordinate $s = u^\top H_A \geq 0$, conditioning one centered coordinate and requiring all other coordinates to be smaller gives the density

$$q_K(s) = K \phi_{\sigma_K^2}(s) G_{K-1}\left(\frac{Ks}{K-1} \right), \qquad s \geq 0, \tag{52}$$

where $\phi_{\sigma_K^2}$ is the centered Gaussian density with variance σ_K^2. Near zero,

$$q_K(s) = C_K s^{K-2}(1 + O(s^2)) \tag{53}$$

for a positive constant C_K. An integration-by-parts score based on the full winner therefore has a nonvanishing boundary contribution at $K = 2$. At $K = 3$ its axial score behaves as $1/s$ and has second-moment integrand proportional to s^{-1} near zero. For $K \geq 4$ that integral is finite. The tangent score removes this axial term entirely.

B A concrete derivative-information subclass

Lemma 3 (Bounded parallel shifts)**.** *Under the model assumptions, suppose $|\theta_\star^\top S_t| \leq B$ almost surely and, for some $r > 0$, $f'(x) \geq m_r > 0$ whenever $|x| \leq B + r$. Let $p_r = 2\Phi(r) - 1$, where Φ is the standard normal cumulative distribution function. For every predictable unit direction u, $\mu_t(u) \geq m_r p_r^K$. The mean derivative under uniform initialization obeys $\mu_t^{\text{init}} \geq m_r p_r$.*

Proof. Conditional on $\mathcal{F}_{t-1}$, the variables $\theta_\star^\top Z_{t,a}$ are independent standard normals. The event $\max_{a \leq K} |\theta_\star^\top Z_{t,a}| \leq r$ has probability p_r^K. On this event every action, including the selected one, has latent index in $[-B - r, B + r]$. Since $f' \geq 0$ elsewhere, conditional expectation gives the first bound. Under an independent uniform action, the selected residual has a standard normal projection, giving the second bound. □

The bounds hold conditional on $\mathcal{F}_{t-1}$ and allow arbitrary orthogonal shifts. The derivative envelope specifies the link class.

For the logistic link, take $m_r = e^{-(B+r)}/(1 + e^{-(B+r)})^2$. At $B = 1.15$, $r = 1.5$, and $K = 4$, the bounds give $\mu_{\text{low}} = 0.0347271$ and $\mu_{\text{init}} = 0.0533992$. With $\gamma = 4$, the operational choice

Table 3: Matched unknown-link comparison. Each mean pseudo-regret and parenthesized standard error uses 48 held-out paired replicates. Explore-then-commit (ETC) freezes the centered uniform-exploration moment. The last column is ETC minus tangent, with its paired standard error. Initial lengths are tangent/ETC and were selected on 12 independent tuning replicates.

Environment	K	T	n_0	Tangent	Centered ETC	Difference
Logistic	2	16k	128/1024	86.48 (2.82)	212.53 (9.01)	126.05 (8.24)
Logistic	2	64k	128/2048	110.09 (2.48)	460.96 (22.41)	350.88 (22.36)
Logistic	2	128k	128/4096	131.09 (3.13)	687.53 (24.65)	556.44 (24.41)
Logistic	3	16k	256/1024	122.84 (2.96)	292.04 (12.37)	169.19 (12.07)
Logistic	3	64k	128/2048	146.53 (4.00)	632.78 (28.27)	486.25 (30.09)
Logistic	3	128k	128/2048	178.18 (4.35)	938.93 (51.33)	760.74 (51.77)
Logistic	4	16k	128/1024	121.67 (3.06)	330.63 (12.35)	208.96 (12.86)
Logistic	4	64k	128/2048	171.70 (3.46)	667.62 (28.45)	495.93 (28.48)
Logistic	4	128k	128/2048	191.70 (4.42)	1,047.61 (52.33)	855.92 (50.83)
Signed	2	16k	128/1024	65.37 (9.06)	236.21 (11.47)	170.84 (13.29)
Signed	2	64k	128/4096	61.83 (5.24)	529.12 (11.45)	467.29 (11.40)
Signed	2	128k	128/4096	112.77 (44.82)	678.57 (29.01)	565.81 (54.15)
Signed	3	16k	128/1024	95.20 (15.71)	372.51 (15.23)	277.31 (20.17)
Signed	3	64k	128/2048	81.27 (6.89)	794.29 (35.11)	713.02 (36.57)
Signed	3	128k	128/2048	71.33 (4.05)	1,229.83 (81.13)	1,158.50 (81.27)
Signed	4	16k	128/1024	84.83 (6.53)	534.40 (26.01)	449.56 (24.03)
Signed	4	64k	128/2048	91.11 (6.00)	1,040.72 (48.17)	949.61 (48.24)
Signed	4	128k	128/4096	108.17 (17.56)	1,634.05 (58.48)	1,525.88 (56.76)

$b = 1.5$ gives $\rho\sqrt{b} = 1.07741$; this particular lower bound does not certify that schedule. Choosing $b = 1.1$ gives $\rho\sqrt{b} = 0.92264$ and satisfies (17). For $d = 6$, $\delta = 0.05$, $\overline{V}_K = 5/3$, and $T = 128{,}000$, however, (25) gives $n_0^*(T) = 18{,}486{,}784 > T$. The simulations in Appendix F use shorter initializations than this sufficient bound.

C Matched uniform-exploration comparison

Protocol and information. The two environments use the logistic and signed hyperbolic-tangent links of the mechanism studies, with $d = 6$, $K \in \{2, 3, 4\}$, and $T \in \{16{,}000, 64{,}000, 128{,}000\}$. The logistic reward is Bernoulli with mean $f(x) = (1 + e^{-x})^{-1}$ and parallel shift $1.15 \sin(0.011t + 0.4)$; the signed reward is $f(x) = \tanh(x)$ with parallel shift -2 and no additional noise. Both use common orthogonal components $7 \operatorname{sign} \sin(0.0063t + 0.2)$ and $4.5 \cos(0.0027t + 0.7)$ on the last two coordinates. The direction is proportional to $(0.61, -0.44, 0.52, 0.39, 0, 0)$. Learners receive the shifted candidate set and their selected reward, not the link, direction, shift, or derivative. Both use the same reward-variance bound and clipping rule; the tangent schedule uses the declared Lipschitz bound.

Comparator and selection. The centered explore-then-commit (ETC) baseline uses exactly the uniform initialization score $H_{t,A_t} Y_t / \sigma_K^2$, coordinate-clips each score, averages and normalizes, and then commits to that direction. This moment remains valid under common shifts. For each of the 18 cells, each method selects n_0 from $\{128, 256, 512, 1{,}024, 2{,}048, 4{,}096, 8{,}192\}$ by mean pseudo-regret on 12 tuning replicates, before its 48 held-out evaluation replicates. The initial grid ended at 4,096; four tuning selections reached that endpoint, so it was extended before held-out regret summaries were inspected. The extension did not change any selected configuration. Both methods receive the same tuning budget. Tangent growth is $b = 1.5$, with step 4 for logistic and 1 for signed rewards.

Potential outcomes, Gaussian residuals, and uniform action choices are paired within each evaluation replicate. Tuning and evaluation use disjoint seed rules given in Appendix I. The reported paired difference is $D_j = R_j^{\text{ETC}} - R_j^{\text{tan}}$; its standard error is the sample standard deviation of D_j divided by $\sqrt{48}$. The intervals in the main figure are $\bar{D} \pm 1.96$ SE, where SE denotes standard error. These are pointwise simulation intervals. Table 3 reports all paired comparisons; Table 4 separates initialization from later regret.

Table 4: Mean pseudo-regret decomposition at 128,000 rounds, using the configurations selected before the 48 paired evaluation replicates. Explore-then-commit (ETC) uses the same clipped centered score during uniform initialization; "initial" denotes this phase and "later" denotes all remaining rounds.

Environment	K	Tangent initial	Tangent later	ETC initial	ETC later
Logistic	2	13.34	117.75	440.84	246.68
Logistic	3	19.11	159.08	330.80	608.13
Logistic	4	22.16	169.53	401.44	646.18
Signed	2	12.15	100.61	402.70	275.87
Signed	3	22.04	49.29	362.61	867.22
Signed	4	31.16	77.01	999.00	635.06

Table 5: Shared initialization at $K = 4$ candidates, dimension six, and $T = 128{,}000$ rounds. The common n_0 is selected for centered explore-then-commit (ETC) on its original 12 tuning replicates. Entries are mean pseudo-regret with one standard error in parentheses over 48 paired held-out replicates. The last column is ETC minus tangent with its paired standard error. Both methods share the same initialization.

Environment	Shared n_0	Tangent	Centered ETC	Difference
Logistic	2,048	$503.29\ (3.31)$	$1,047.61\ (52.33)$	$544.32\ (50.95)$
Signed	4,096	$1,042.56\ (3.78)$	$1,634.05\ (58.48)$	$591.49\ (56.07)$

Shared initialization isolates subsequent updates. For each cell we also evaluate both methods at the initialization length already selected for ETC on its original tuning seeds. The shared residual arrays, potential rewards, uniform action choices, variance bound, and clipping threshold produce exactly the same initial score and direction. Only the tangent learner subsequently updates this direction. Table 5 reports the two four-candidate, longest-horizon cells discussed in the main text; Figure 2 shows all 18 shared-initialization comparisons.

Operational sensitivity. We evaluate all seven initialization budgets on the same held-out replicates. Figure 3 plots total pseudo-regret against initialization cost for all seven budgets. This displays the objective used to select the starred configurations on the separate tuning sample. At four candidates and 128,000 rounds, we additionally vary the logistic step between 2 and 4 and growth among 1.25, 1.5, and 2, holding $n_0 = 128$ fixed (Table 6). The signed two-candidate cell at 128,000 rounds has tangent mean 112.77 with standard error 44.82, median 54.96, and 95th percentile 117.99 (linear interpolation between ordered replicates). Its maximum is $2{,}137.52$, showing the tail variation associated with a short initialization. Its first and third quartiles are 43.70 and 65.91. The paired ETC-minus-tangent difference is 565.81 ± 54.15: 46 of the 48 differences are positive and two are negative. Quartiles also use linear interpolation.

Table 6: Operational schedule sensitivity at four candidates and 128,000 rounds in the logistic environment. Initial length is 128. Entries are mean pseudo-regret with one standard error over the same 48 evaluation seeds.

Step γ	Growth b	Pseudo-regret
2	1.25	149.66 (3.59)
2	1.5	138.38 (4.64)
2	2	183.02 (11.96)
4	1.25	252.84 (4.48)
4	1.5	191.70 (4.42)
4	2	174.10 (4.85)

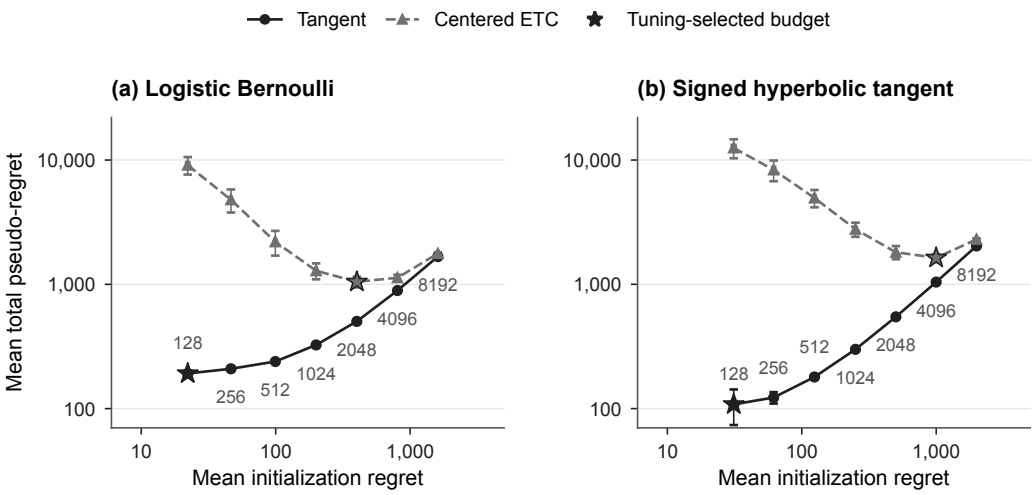

Figure 3: Total pseudo-regret versus initialization cost at $K = 4$ candidates and horizon $T = 128{,}000$. Each point represents one of seven predeclared initialization budgets for the tangent learner or centered explore-then-commit (ETC). Horizontal coordinates are mean initialization regret, and vertical coordinates are mean total regret, including both initialization and subsequent rounds. Both coordinates are computed from the same 48 held-out replicates; error bars give marginal intervals of mean plus or minus 1.96 standard errors for each coordinate. Gray labels along the tangent curve give the initialization budget n_0 in rounds; the same ordered seven budgets are used for ETC. Stars identify budgets independently selected for each learner using 12 separate tuning replicates to minimize total regret. Both axes are logarithmic, and lines connect the tested budgets.

D Continually refreshed uniform-exploration comparison

Policy and information. Centered persistent exploration (CPE) reestimates a direction throughout the horizon using the valid uniform-action score, then chooses greedily between exploration rounds. It uses the two environments of Appendix C, $d = 6$, $K \in \{2, 3, 4\}$, and $T = 128{,}000$. The first 128 actions are uniform. Thereafter, an independent Bernoulli coin with probability $p = (q - 128)/(T - 128)$ selects uniform exploration; otherwise the action maximizes $u^\top H_{t,a}$. Thus q is the expected total number of uniform rounds, including initialization.

The initial direction is computed after round 128. Subsequent calendar segments have lengths $\lceil 128(1.5)^j \rceil$, $j = 0, 1, \ldots$. After each complete segment with future actions remaining, CPE reestimates using all m uniform samples collected so far. It stores the raw vectors $H_{t,A_t} Y_t / \sigma_K^2$, clips every stored vector at $\tau(m) = \sqrt{\overline{V}_K m / \ell}$, averages and normalizes. The fixed logarithmic factor is $\ell = \log(4d\lceil 2 + \lceil \log_{1.5}(T/128) \rceil \rceil / 0.05)$. The bound $\overline{V}_K$ is $1.25/\sigma_K^2$ for logistic rewards and $1/\sigma_K^2$

Table 7: Paired policy comparison at $d = 6$ and $T = 128{,}000$: mean total pseudo-regret $\pm$ one standard error over 48 replicates. TSB denotes Tangent Stein Bandits, ETC centered explore-then-commit, and CPE centered persistent exploration. The last column is a paired difference; positive values favor TSB.

Link	K	TSB	ETC	CPE	CPE − TSB
Logistic	2	133.6 ± 3.4	656.1 ± 17.5	$1{,}405.8 \pm 29.0$	$1{,}272.2 \pm 28.9$
Logistic	3	172.1 ± 4.3	868.7 ± 40.5	$1{,}702.7 \pm 84.0$	$1{,}530.5 \pm 82.9$
Logistic	4	201.9 ± 4.4	$1{,}086.2 \pm 74.6$	$2{,}009.8 \pm 80.0$	$1{,}808.0 \pm 79.3$
Signed	2	52.1 ± 4.1	676.5 ± 19.6	$1{,}288.1 \pm 54.2$	$1{,}236.1 \pm 52.9$
Signed	3	81.0 ± 5.9	$1{,}143.8 \pm 64.5$	$2{,}413.9 \pm 66.0$	$2{,}332.9 \pm 64.1$
Signed	4	95.3 ± 9.2	$1{,}570.3 \pm 46.6$	$3{,}335.9 \pm 135.6$	$3{,}240.6 \pm 132.8$

Table 8: Pointwise paired intervals (mean ± 1.96 standard errors) for centered persistent exploration (CPE) and centered explore-then-commit (ETC) minus Tangent Stein Bandits (TSB), over 48 paired replicates at $d = 6, T = 128{,}000$. The last column counts positive CPE-minus-TSB differences.

Link	K	CPE − TSB	ETC − TSB	Positive
Logistic	2	[1215.6, 1328.8]	[489.8, 555.2]	48/48
Logistic	3	[1368.1, 1692.9]	[617.7, 775.4]	48/48
Logistic	4	[1652.5, 1963.5]	[738.9, 1029.7]	48/48
Signed	2	[1132.4, 1339.7]	[587.3, 661.6]	48/48
Signed	3	[2207.2, 2458.6]	[938.9, 1186.8]	48/48
Signed	4	[2980.4, 3500.8]	[1386.0, 1564.0]	48/48

for signed rewards, as for the other learners. Greedy-round rewards do not enter this estimator. A moment norm at most 10^{-14} retains the previous direction, initialized to $(1, \ldots, 1)/\sqrt{d}$. There are 16 estimates including initialization; no final unused update is performed. At $q = 128$, CPE reduces to centered explore-then-commit (ETC) with 128 initial observations.

Selection and evaluation. Before evaluation, CPE independently selects q from $\{128, 256, 512, 1{,}024, 2{,}048, 4{,}096, 8{,}192\}$ using mean total pseudo-regret on 12 tuning replicates per cell, with ties toward smaller q. All six decisions are frozen together before any new test run. Tangent Stein Bandits (TSB) and ETC retain their original tuning choices: TSB uses $n_0 = 128$ in all six cells; ETC uses $(4096, 2048, 2048)$ for logistic and $(4096, 2048, 4096)$ for signed rewards, ordered by $K = 2, 3, 4$. The methods share candidate sets, potential rewards and uniform arm choices within each of 48 evaluation replicates. CPE coins use a separate random stream. Development, tuning and evaluation seed namespaces are disjoint from each other and from the other studies; Appendix I gives the seed rules.

The grid size and tuning-replicate budget match the original search. Logistic $K = 2$ and signed $K = 3$ select the upper endpoint $q = 8192$; the other four cells select $q = 4096$. The grid was fixed before tuning.

Results and interpretation. Table 7 reports all six comparisons. TSB has lower mean pseudo-regret than both CPE and ETC in every cell. The CPE-minus-TSB paired reductions range from 1236.1 to 3240.6; all six descriptive intervals lie above zero (Table 8). CPE also incurs more mean regret than ETC in all six cells. Continually collecting uniform scores therefore does not by itself improve this finite-horizon tradeoff. At $K = 4$, logistic CPE incurs mean exploration and greedy regret 802.1 and 1207.7, respectively; signed CPE incurs 993.8 and 2342.1 (Table 9). Final CPE mean cosines with the truth range from 0.9666 to 0.9896. Accurate final directions do not recover the regret incurred during exploration and earlier greedy rounds. No CPE normalization fallback occurs.

These comparisons evaluate complete policies that differ in initialization, exploration timing, estimation, and updating. The shared-initialization experiment in Appendix C isolates the gain from continued updates by holding the initial direction fixed.

Table 9: Centered persistent exploration (CPE) settings and regret decomposition at $T = 128{,}000$, over 48 evaluation replicates. q is the tuning-selected expected total exploration count; observed counts are mean $\pm$ standard deviation (SD), with the full range. Regret columns are means $\pm$ one standard error (SE). Both include all replicates.

Link	K	q	Count $\pm$ SD	Range	Exploration	Greedy
Logistic	2	8192	8187.9 ± 100.1	7927–8415	884.6 ± 2.7	521.2 ± 28.4
Logistic	3	4096	4091.9 ± 54.3	3936–4223	660.8 ± 2.4	$1{,}041.9 \pm 84.2$
Logistic	4	4096	4102.0 ± 67.7	3925–4235	802.1 ± 2.4	$1{,}207.7 \pm 79.5$
Signed	2	4096	4091.7 ± 64.0	3960–4265	403.1 ± 2.2	885.0 ± 54.6
Signed	3	8192	8186.6 ± 97.3	8010–8427	$1{,}456.6 \pm 4.9$	957.3 ± 65.4
Signed	4	4096	4087.5 ± 72.6	3925–4374	993.8 ± 3.8	$2{,}342.1 \pm 136.1$

Implementation and computational cost. CPE retains $O(N_{\text{exp}}d)$ raw uniform-score history for re-clipping, where N_{exp} is the realized exploration count. An update costs $O(md)$ for its m accumulated scores. Algorithm 1, by contrast, can stream clipped epoch sums in $O(d)$ persistent state. The shared simulation harness preallocates candidates and potential outcomes for the entire horizon, using $O(TKd)$ storage for all methods. An independent scalar implementation agrees with the batched CPE actions in 18 development cases; tests also verify the $q = 128$ ETC reduction, re-clipping, unused-reward exclusion, seed separation, and fallback behavior. The comparison comprises 504 tuning and 864 evaluation method runs, or $175{,}104{,}000$ method-rounds. All tables are reconstructed from their frozen per-replicate records.

D.1 Paired dimension comparison

We compare all three policies at $d \in \{6, 12, 24\}$, $K = 4$, and $T = 128{,}000$ with logistic Bernoulli rewards. For each dimension, the Gram–Schmidt construction in Appendix G gives unit vectors $\theta_\star, v, w$. The predictable common shift is

$$S_t = 1.15 \sin(0.011t + 0.4)\theta_\star + 7 \operatorname{sign} \sin(0.0063t + 0.2)v + 4.5 \cos(0.0027t + 0.7)w.$$

Residual candidates are independent $\mathcal{N}(0, I_d)$, and potential rewards are independent Bernoulli draws conditional on their indices. The same basis construction, shift amplitudes, and policy constants apply at every dimension.

Each policy independently selects an exploration budget using 12 tuning replicates and the common grid

$$\{128, 256, 512, 1{,}024, 2{,}048, 4{,}096, 8{,}192, 16{,}384\}.$$

The budget is n_0 for TSB and ETC, and the expected total uniform exploration count q for CPE. All three policies receive the same grid, including its predeclared additional point at 16,384. TSB uses $\gamma = 4$, growth 1.5, and the dimension-specific clipping threshold; CPE retains the policy above. All nine selected configurations are frozen before the 48 new paired evaluation replicates per dimension. Candidates, potential rewards, and uniform arm indices are shared within each replicate; CPE exploration coins use an independent stream.

Table 2 reports total regret, and Table 10 reports selected budgets and paired reductions. TSB has lower regret than both baselines in all three dimensions. At $d = 24$, its mean regret is 651.2 ± 6.8, compared with $2{,}019.7 \pm 45.1$ for ETC and $4{,}003.7 \pm 72.1$ for CPE. Mean regret increases with dimension for all three policies, while continual tangent learning retains its advantage. All six paired intervals lie above zero. No selected budget reaches the grid's upper endpoint.

Seeds are $s = B + 10^8 d + 100003j$, where j is the zero-based replicate index and $B = 7 \cdot 10^{12}, 8 \cdot 10^{12}, 9 \cdot 10^{12}$ for development, tuning, and evaluation, respectively. CPE coins use $s + 2 \cdot 10^{13}$. The study comprises 864 tuning and 432 evaluation method runs, or $165{,}888{,}000$ method-rounds. An independent scalar controller matches the batched policies in 18 development cases spanning all three dimensions and policies. The implementation also checks exclusion of unobserved rewards and diagnostic means from policy decisions, and the reduction of CPE at $q = 128$ to ETC with 128 initial rounds.

Table 10: Selected budgets and paired regret reductions for the dimension comparison. TSB denotes Tangent Stein Bandits, ETC centered explore-then-commit, and CPE centered persistent exploration. Budgets are initial rounds for TSB and ETC, and expected total exploration rounds for CPE. Differences are baseline minus TSB; brackets give pointwise mean differences plus or minus 1.96 paired standard errors over 48 replicates.

d	TSB n_0	ETC n_0	CPE q	ETC − TSB	CPE − TSB
6	128	2048	4096	772.3 ± 56.2	$1{,}776.1 \pm 61.7$
				$[662.2, 882.4]$	$[1{,}655.1, 1{,}897.1]$
12	256	4096	8192	$1{,}057.6 \pm 42.5$	$2{,}505.2 \pm 50.3$
				$[974.3, 1{,}140.9]$	$[2{,}406.5, 2{,}603.8]$
24	128	4096	8192	$1{,}368.5 \pm 44.6$	$3{,}352.6 \pm 73.0$
				$[1{,}281.0, 1{,}456.0]$	$[3{,}209.6, 3{,}495.5]$

E Mechanism controls

We compare raw, misspecified, exact centered, and tangent scores under matched schedules, and test how distributional changes affect the tangent identity. Figure 4 shows mean paired excess pseudo-regret per round relative to the tangent learner, with intervals of plus or minus 1.96 paired standard errors. Its separate linear vertical ranges resolve both the large misspecification effects and the near-zero valid-score comparisons. The numerical $\pm$ values below report one standard error of cumulative pseudo-regret. Appendix F gives parameters and complete aggregates; Appendix I gives the seed rules. The raw shift-unaware control embeds the fixed-distribution Stein moment in the same initialization and epoch schedule as our learner to isolate the effect of the score.

Logistic score comparison. We first use $K = 4$, $d = 6$, binary logistic rewards, $4{,}096$ uniform initialization rounds, step size $\gamma = 4$, and epoch factor $b = 1.5$. The predictable shift combines a bounded parallel component with two large orthogonal components. Across five horizons from $16{,}000$ to $256{,}000$ and 48 paired replicates, the tangent method's mean pseudo-regret increases from 826.53 ± 2.20 to 899.35 ± 4.27. The raw shift-unaware score increases from $2{,}974.66 \pm 8.93$ to $45{,}161.55 \pm 154.29$. At the largest horizon the ratio is 50.22. All realized tangent epochs retain derivative information and remain in the positive hemisphere.

Independent-winner misspecification. The second test uses $K = 4$, $d = 6$, a hyperbolic-tangent link, a predictable parallel shift of -2, $2{,}048$ initialization rounds, and 48 paired replicates at horizons $8{,}000$ through $128{,}000$. At a pre-specified 20° population cell, the current direction has cosine 0.93969 with the truth. The exact tangent moment has cosine 1.00000 with the projected truth, while the demean-then-independent reset has cosine 0.23515 despite average derivative 0.44605. At horizon $128{,}000$, mean pseudo-regret is 548.20 ± 3.87 for the tangent method and $21{,}670.11 \pm 28.64$ for the misspecified reset, a ratio of 39.53. The exact full-score and oracle-shift controls remain within 2.5% of tangent regret. Under a constant reward, the misspecified population moment has norm 0.44967, versus 0.00059 for the tangent moment at the same Monte Carlo size, isolating selection bias rather than weak information.

Gaussian identity checks and distributional perturbations. We evaluate the identity in 21 Gaussian conditions using $300{,}000$ action sets per condition: $K \in \{2, 3, 4, 8, 16\}$, dimensions from 2 to 50, logistic, hyperbolic-tangent, and arctangent links, several angles, and predictable shifts. The minimum cosine with the projected truth is 0.99546; the maximum relative identity error is 0.10069. Ten online cases at horizon $64{,}000$ and 20 replicates have minimum final cosine 0.98084 and maximum mean regret per round 0.03026. A paired orthogonal-shift control produces identical mean regret to machine precision (both 507.20919 at amplitudes zero and 20), as predicted.

The perturbations affect both moment direction and magnitude. The Gaussian control has relative identity error 0.00981. Rademacher residuals raise it to 1.48809; an action-specific shift reverses the moment (cosine -0.999999); and a common shift selected from current residuals reduces cosine to 0.60983. Finally, moving a logistic index into saturation reduces average derivative from 0.20070 at the center to 0.00660 and 0.00206 at the two tested extremes. These perturbations distinguish changes to the tangent identity from loss of learning information under saturation.

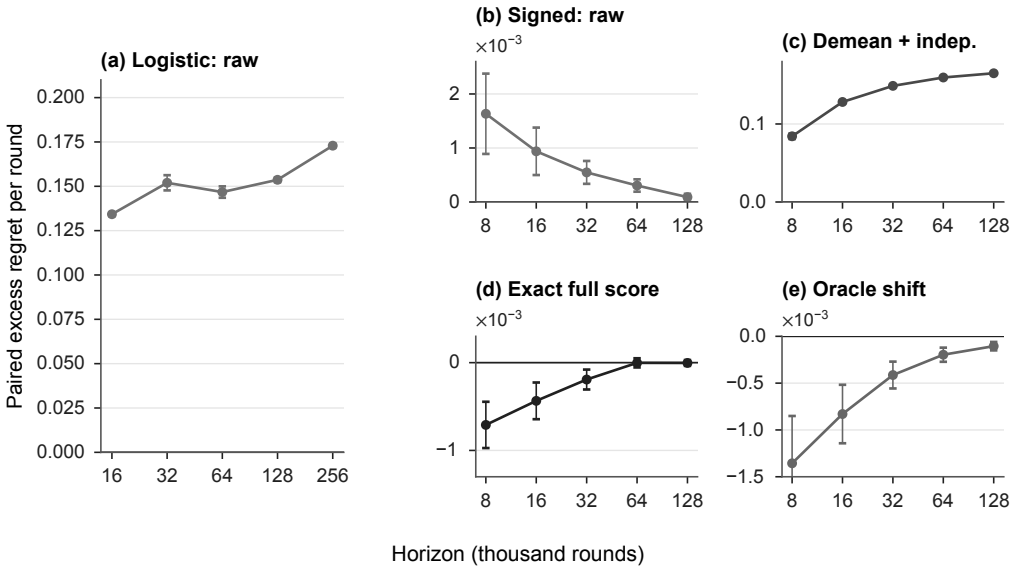

Figure 4: Paired score-control effects relative to the tangent learner. Each point is the mean of $(R_T^{\text{control}} - R_T^{\text{tangent}})/T$ over 48 matched replicates, where R_T is cumulative pseudo-regret and T is horizon; error bars are plus or minus 1.96 paired standard errors. Positive values favor tangent, and zero denotes equality. Panel (a) uses logistic rewards with 4,096 initialization rounds. Panels (b)–(e) use signed hyperbolic-tangent rewards with 2,048 initialization rounds. Both environments have four candidates and dimension six. Raw shift-unaware scores in (a,b) and the demean-then-independent score in (c) are misspecified controls; "indep." abbreviates independent. The exact full centered score in (d) is valid at this candidate count, and the oracle-shift control in (e) observes the common shift. Each panel has its own explicitly labeled linear vertical range; the horizontal axes are logarithmic, and lines connect tested horizons. The fine-scale panels show that exact full-score and oracle differences approach zero as the horizon increases.

F Experimental protocol and complete settings

All simulations use paired random streams within each replicate: methods see the same Gaussian residual actions, potential-outcome uniforms or reward noise, and initialization choices. Aggregate uncertainty is the standard error across independent replicates. The learning algorithms never receive $\theta_\star$, f', or a population derivative; these quantities appear only in post-run diagnostics and do not influence action selection.

F.1 Logistic score comparison

The logistic simulation uses $K = 4$, $d = 6$, binary rewards with a logistic conditional mean, $n_0 = 4{,}096$, $\gamma = 4$, $b = 1.5$, and declared derivative range $[\mu_{\text{low}}, \mu_{\text{high}}] = [0.08, 0.25]$. The resulting schedule values are

$$c_0 = 0.8660, \qquad \rho = 0.7229, \qquad \rho^{-2} = 1.9137, \qquad \rho\sqrt{b} = 0.8853. \tag{54}$$

Thus $\gamma \le 1/\mu_{\text{high}}$, $1 < b < \rho^{-2}$, and $\rho\sqrt{b} < 1$. Each of the five horizons has 48 paired replicates.

The smallest realized initialization derivative average is 0.1853 and the smallest realized greedy-epoch average is 0.1710. These are sample averages of $f'(\theta_\star^\top X_{t,A_t})$, not certified lower bounds on the conditional population information in (10). The minimum cosine after any initialization or update is 0.9323. The maximum ratio between observed tangent-moment error and the theoretical envelope is 0.1514.

The worst-case sufficient sizes in (23) range from roughly 0.47 to 0.52 million for initial alignment and 2.52 to 2.76 million for the hemisphere induction over the tested horizon grid. The executed $n_0 = 4{,}096$ lies below these sufficient bounds. Table 11 reports pseudo-regret at each tested horizon.

Table 11: Logistic score comparison. Values are mean pseudo-regret with one standard error in parentheses over 48 paired replicates.

Horizon	Tangent epoch	Raw shift-unaware
16,000	826.53 (2.20)	2,974.66 (8.93)
32,000	839.51 (2.50)	5,702.67 (70.22)
64,000	861.90 (3.31)	10,253.91 (104.62)
128,000	881.12 (2.66)	20,549.98 (89.30)
256,000	899.35 (4.27)	45,161.55 (154.29)

F.2 Independent-winner misspecification

The comparison uses $K = 4$, $d = 6$, a hyperbolic-tangent link, predictable parallel shift -2, $n_0 = 2{,}048$, and 48 replicates. The five methods are the tangent epoch learner; the misspecified demean-then-independent-winner reset; an exact full-score control, which is finite-variance here because $K = 4$; an oracle that knows and removes the shift; and a raw learner used as a secondary control. All methods share the same epoch schedule and paired randomness where their estimators permit it.

The population study covers 24 combinations of current angle and parallel shift. The pre-specified 20° cell with parallel shift -2 has derivative average 0.44605, misspecified-reset cosine 0.23515, and tangent cosine 1.00000. Deterministic quadrature gives population initialization information 0.188697, above the pre-specified 0.18 threshold. Table 12 compares the five methods at the longest horizon.

Table 12: Independent-winner score comparison at horizon $128{,}000$. Values are mean pseudo-regret with one standard error in parentheses over 48 paired replicates.

Method	Pseudo-regret
Tangent epoch	548.20 (3.87)
Demean then independent-winner reset	21,670.11 (28.64)
Exact centered full score	547.77 (2.46)
Oracle shift	534.76 (2.20)
Raw secondary control	559.81 (2.38)

F.3 Gaussian identity checks and distributional perturbations

Each Gaussian identity cell uses $300{,}000$ independently generated action sets. Online cases use horizon $64{,}000$ and 20 replicates. Each matched distributional perturbation uses $600{,}000$ sets. Table 13 relates the assumptions to their role in the proof and the perturbation results.

Table 13: Theorem assumptions, their role, and the consequence or matched stress result when each is removed.

Assumption	Role in proof	Consequence or matched stress result
Gaussian residuals	Tangent conditional law and integration by parts	Rademacher relative error 1.488; Laplace and Student-t scale errors
Action-common shift	Centering preserves action comparisons	Action-specific shift reverses the moment (cosine -1.000)
Predictable shift	Shift fixed before current Gaussian residuals	Residual-dependent common shift gives cosine 0.610
Monotone link	Greedy index is reward comparator and $\mu_t \geq 0$	Comparator equivalence relies on monotonicity
Positive derivative information	Angular contraction	Logistic saturation reduces derivative to 0.00660 and 0.00206
Uniform conditional noise second-moment bound	Clipped Freedman bound	Without this bound, the stated concentration guarantee does not apply.

G Additional scope and stress-test results

For the variable-dimension experiments, define vectors with coordinates $a_j = \sin(0.73j) + 0.4\cos(1.17j)$, $b_j = \cos(0.41j) - 0.3\sin(1.31j)$, and $c_j = \sin(1.91j) + 0.2\cos(0.29j)$ for $j = 1, \ldots, d$. Gram–Schmidt normalization of (a, b, c) gives the reward direction $\theta_\star$ and orthogonal directions v, w; at $d = 2$, set $w = v$. The current direction at angle β is $u = \cos(\beta)\theta_\star + \sin(\beta)v$. Identity checks apply the constant shift $s\theta_\star + qw$, where s and q are the stated parallel and orthogonal shifts.

The Gaussian identity grid spans action counts 2, 3, 4, 8, and 16; dimensions 2, 6, 20, and 50; three monotone link families; current angles from 30° to 75°; parallel shifts from -2 to 2; and a large orthogonal shift of magnitude 10. Across all 21 cells, the minimum cosine with $P_u\theta_\star$ is 0.99545565 and the maximum relative error is 0.10069297.

Figure 5(b) groups the conditions by the factor varied. Cases 1–5 vary $K = (2, 3, 4, 8, 16)$; 6–9 vary $d = (2, 6, 20, 50)$; 10–18 use logistic, hyperbolic-tangent, and arctangent links, each at parallel shifts $(-2, 0, 2)$; 19–21 vary the angle $(30, 50, 75)$ degrees. Unless varied, $K = 4$, $d = 6$, the angle is 55 degrees, the link is logistic, and the orthogonal shift is 10. The parallel shift is 1 for cases 1–5 and 19–21, and -1 for 6–9.

The ten online cases span action counts 2, 4, and 8; dimensions 2, 6, 12, and 24; logistic, hyperbolic-tangent, and arctangent links; and orthogonal-shift amplitudes 0, 7, and 20. They use $n_0 = \max\{2048, 384d\}$, growth $b = 1.5$, and step sizes 4, 1, and 1.5 for the three links, respectively. The parallel shift is $1.15\sin(0.011t + 0.4)$, except for the hyperbolic-tangent case, which uses -2. The orthogonal shift is $q\,\mathrm{sign}\sin(0.0063t + 0.2)v$. Their minimum final cosine is 0.98083833 and maximum mean pseudo-regret per round is 0.03025827. The paired equality between orthogonal amplitudes 0 and 20, both 507.20919 to the reported precision, is expected: an orthogonal common shift changes neither the centered action array nor any latent reward index.

The stress matrix uses a Gaussian control and changes one assumption at a time, with $K = 4$, $d = 3$, a logistic link, angle 50°, and zero baseline shift. Laplace and Student-t_3 coordinates are scaled to unit variance; Rademacher coordinates take values -1 and 1 with equal probability. The action-specific translations are $(-3, -1, 1, 3)v$ for the four actions. The residual-dependent common translation is $\mathrm{sign}(w^\top H_A)\theta_\star$. Figure 5(a) displays alignment and magnitude separately for these six conditions. The Rademacher cell has cosine 0.97632 but target-norm ratio 2.44861, so direction alone would hide a severe scale error. The action-specific shift cell has cosine -0.999999 and relative error 7.75027. The residual-dependent common shift has cosine 0.60983 and relative error 1.31020.

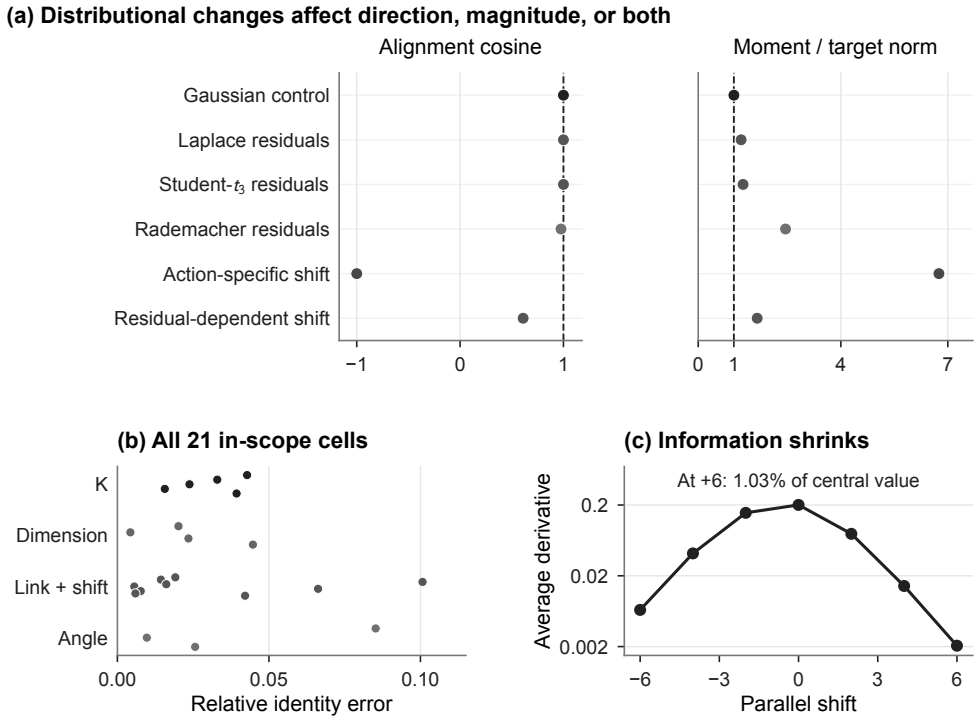

Figure 5: Distributional validity and learning information are separate conditions. (a) Six matched conditions, each with 600,000 action sets, show the cosine between the empirical tangent moment and the Gaussian target and the ratio of their norms. The reference target is $\mu P_u \theta_\star$, where u is the current unit direction, $P_u = I - uu^\top$, $\theta_\star$ is the reward direction, and μ is average link derivative. The dashed value one marks exact agreement for each metric. Good alignment alone can conceal a magnitude error. (b) All 21 in-scope Gaussian conditions, each with 300,000 action sets, are shown by relative error $\|\widehat{m} - \mu P_u \theta_\star\| / \|\mu P_u \theta_\star\|$. Rows identify the factor varied: candidate count K, dimension, link family with parallel shift, or current angle. Vertical offsets only separate points. (c) Seven predictable parallel shifts, each with 300,000 action sets, reduce logistic derivative information near saturation. At shift $+6$, the mean derivative is 1.03% of its zero-shift value. These remain Gaussian, predictable-shift conditions of the identity. Small target moments amplify finite Monte Carlo error.

H Direct comparison with ESTOR without shifts

Environments and policies. We compare Tangent Stein Bandits (TSB) with Epoched Stein's Oracle Single Index Bandit (ESTOR) (Kang et al., 2026) in four cells with $S_t = 0$, $d = 6$, $K \in \{3, 4\}$, and $T = 128{,}000$. The normalized direction, logistic Bernoulli rewards, and noiseless signed hyperbolic-tangent rewards are those of Appendix C, with all translation components set to zero. TSB centers the candidates; ESTOR uses the raw independent standard Gaussian candidates. Both receive the candidate set and their selected reward.

We implement ESTOR's Algorithm 2 with its original doubling schedule: n_0 uniform rounds followed by greedy epochs of lengths $2n_0, 4n_0, 8n_0, \ldots$. Each new estimate uses the preceding epoch alone. For a unit greedy direction u, its selected density is $K\phi_d(x)\Phi(u^\top x)^{K-1}$, where ϕ_d is the standard d-dimensional Gaussian density and Φ is the standard normal distribution function. The corresponding score is

$$S_u^{\mathrm{ESTOR}}(x) = x - (K-1)\frac{\phi(u^\top x)}{\Phi(u^\top x)}u, \tag{55}$$

where ϕ is the standard normal density. Uniform initialization uses the score x. With regularization set to zero, ESTOR averages the coordinate-clipped reward-times-score vectors and uses that estimate

Table 14: No-shift comparison at $d = 6$ and $T = 128{,}000$: mean cumulative pseudo-regret $\pm$ one standard error over 48 paired evaluation replicates. TSB denotes Tangent Stein Bandits; ESTOR denotes Epoched Stein's Oracle Single Index Bandit. Both methods select among 16 configurations on 12 separate tuning replicates. Positive paired differences favor TSB.

Link	K	TSB	ESTOR	ESTOR − TSB
Logistic	3	186.43 ± 3.98	162.77 ± 3.83	-23.66 ± 4.96
Logistic	4	212.50 ± 4.24	220.11 ± 6.22	7.61 ± 5.50
Signed	3	66.65 ± 1.33	73.73 ± 1.31	7.08 ± 1.39
Signed	4	85.27 ± 1.71	111.66 ± 2.71	26.38 ± 2.19

for the next epoch. We evaluate the density ratio through the log normal distribution function and normalize estimates for numerical stability.

Tuning and evaluation. Both methods evaluate 16 configurations, combining initialization lengths and clipping multipliers:

$$n_0 \in \{50, 128, 256, 512, 1{,}024, 2{,}048, 4{,}096, 8{,}192\}, \qquad \alpha \in \{1, 2\}.$$

The grid includes ESTOR's original implemented $n_0 = 50$ and the existing TSB initialization grid. TSB uses growth 1.5 and step size 4 for logistic rewards or 1 for signed rewards, multiplying its threshold in Section 5 by α. Following Appendix L.1 of Kang et al. (2026), ESTOR uses $\tau(n) = \alpha\sqrt{3n/\log(2d\log_2(T)/\delta)}$, with $\delta = 0.05$. Each policy selects minimum mean total pseudo-regret over 12 tuning replicates; all eight choices are frozen before 48 paired evaluation replicates per cell. Candidate arrays, potential rewards, and uniform initial actions are shared within each replicate.

For logistic rewards at $K = 3, 4$, respectively, the selected (n_0, α) pairs are $(128, 1), (50, 1)$ for TSB and $(128, 1), (128, 1)$ for ESTOR. For signed rewards, both candidate counts select $(50, 2)$ for TSB and $(50, 1)$ for ESTOR. The comparison comprises 1,536 tuning and 384 evaluation method-runs. An independent scalar action loop matches both implementations, and numerical differentiation verifies (55).

Results. Table 14 reports the four comparisons. For signed rewards, TSB reduces mean regret by 7.08 ± 1.39 at $K = 3$ and 26.38 ± 2.19 at $K = 4$, where $\pm$ denotes one paired standard error. For logistic rewards, ESTOR has lower regret at $K = 3$ (162.77 ± 3.83 versus 186.43 ± 3.98); the paired ESTOR-minus-TSB difference at $K = 4$ is 7.61 ± 5.50, with a pointwise interval of $[-3.17, 18.39]$ using the mean difference plus or minus 1.96 paired standard errors. Thus the signed-reward comparisons favor TSB, whereas the logistic comparisons favor ESTOR at $K = 3$ and do not resolve a difference at $K = 4$.

I Reproducibility

Simulation and estimation. Synthetic environments, policy settings, and tuning grids are given in Appendices C, D, D.1, and H. Simulation time starts at zero when evaluating the shift functions. NumPy's `default_rng` generates full-horizon candidate arrays, Bernoulli reward uniforms when applicable, and uniform action indices in that order. Policies within a replicate share these draws. The signed hyperbolic-tangent rewards are noiseless. Identity checks and the online cases in Appendix G use conditional mean rewards directly, with arctangent link $f(x) = 2\arctan(x)/\pi$. TSB clipping uses $\delta = 0.05$ and $\overline{V}_K = 1.25/\sigma_K^2$ for Bernoulli logistic rewards or $1/\sigma_K^2$ for noiseless bounded rewards, together with the epoch-length threshold in Algorithm 1. The ESTOR comparison also tunes the multiplier α specified in Appendix H.

Seeds for the policy comparisons. Let j denote the zero-based replicate index and let $c = 0$ for logistic rewards and $c = 1$ for signed rewards. The 18-cell comparison uses

$$s = B + 10^8 c + 10^6 K + T + 100003j,$$

with $B = 2026090511$ for tuning and $B = 9096090501$ for evaluation. The persistent-exploration and no-shift comparisons use

$$s = B + 10^8 c + 10^6 K + 100003j.$$

For persistent exploration, $B = 10^{12}, 2 \cdot 10^{12}, 3 \cdot 10^{12}$ for development, tuning, and evaluation, respectively; exploration coins use a separate generator with seed $s + 10^{13}$. The no-shift comparison uses $B = 4 \cdot 10^{12}, 5 \cdot 10^{12}, 6 \cdot 10^{12}$ for the same three stages. The 12 tuning replicates select each configuration before its 48 paired evaluation replicates. Ties favor smaller initialization or exploration budgets, then the smaller clipping multiplier when applicable.

Seeds for the mechanism studies. The logistic and signed online score controls use seeds $B + 100003j + 17T$, with $B = 2026080529$ and $B = 2026080541$, respectively. The signed population comparison uses seed 2026080541. For the Gaussian identity grid, case $i = 0, \ldots, 20$ in the order given in Appendix G uses $2026080553 + 1009i$. The four residual-distribution checks use $2026080553 + 70001 + i$, $i = 0, 1, 2, 3$, ordered Gaussian, Laplace, Student-t_3, and Rademacher. The action-specific and residual-dependent shifts use offsets 80001 and 90001 from the same base. The seven saturation shifts, ordered from -6 to 6, use offsets $100001 + i$, $i = 0, \ldots, 6$. The identity-grid and perturbation moments are accumulated in batches of $30{,}000$ action sets.

Statistical reporting. Pseudo-regret is computed from the known conditional mean rewards in the simulator; policies observe their selected rewards. Reported means retain all replicates. For n independent replicates, standard errors equal the sample standard deviation, with denominator $n - 1$, divided by $\sqrt{n}$. Paired comparisons apply this calculation to within-replicate regret differences. Plotted error bars and paired intervals use pointwise means or mean differences plus or minus 1.96 standard errors. Quantiles use linear interpolation between ordered values. The online dimension-and-link grid uses 20 replicates; other online comparisons use 48.

Computation. The simulations use CPU NumPy arrays. The no-shift comparison ran on an Apple M4 with 16 GB memory, using four workers and one BLAS thread per worker. The paired dimension comparison used the same hardware and software, with two workers and one BLAS thread per worker. The recorded versions are Python 3.12.14, NumPy 2.3.5, and SciPy 1.18.1.

J Connections to statistical estimation and contextual bandits

Known-link contextual bandits. Linear contextual bandits estimate a common reward parameter using upper confidence bounds (Abbasi-Yadkori et al., 2011; Chu et al., 2011) or posterior sampling (Agrawal & Goyal, 2013). Generalized linear bandits extend this model through a known response function (Filippi et al., 2010; Li et al., 2017). Within that model, online-to-confidence-set conversion (Jun et al., 2017) and stochastic-gradient updates with Thompson sampling (Ding et al., 2021) reduce the cost of repeatedly fitting the reward parameter. Logistic-bandit analyses sharpen the dependence on the link's nonlinearity and local curvature (Faury et al., 2020; Abeille et al., 2021); online sufficient statistics can retain these statistical improvements while reducing computation (Faury et al., 2022). Our tangent moment updates the ordering direction from selected rewards with an unknown response function.

Further single-index estimators. Average-derivative estimation connects the index to derivatives of the regression surface (Härdle & Stoker, 1989). For monotone links, score equations yield parametric-rate index estimation under regularity conditions (Balabdaoui et al., 2019b). Sparse orthogonal descent alternates projected direction steps and isotonic regression in high-dimensional models (Dai et al., 2022). For action selection in our model, estimating the ordering direction suffices.

Nonparametric rewards and regression reductions. Nonparametric contextual bandits partition the covariate space to learn action-specific rewards (Rigollet & Zeevi, 2010; Perchet & Rigollet, 2013). Smoothness and margin conditions determine how local estimation translates into regret; local polynomial methods extend this analysis across smoothness regimes (Hu et al., 2022). Reductions to online regression (Foster & Rakhlin, 2020) or offline regression (Simchi-Levi & Xu, 2022) turn prediction guarantees into contextual-bandit guarantees. Here the Gaussian tangent identity supplies a direct estimating equation for the ordering direction. The algorithm need not solve a regression problem over the unknown link class.

Natural exploration and epoch policies. Forced sampling combined with least-squares decisions gives logarithmic regret in linear response bandits (Goldenshluger & Zeevi, 2013). Epoch-Greedy separates uniformly sampled training observations from greedy decisions (Langford & Zhang, 2007), while batched-bandit theory quantifies the effect of restricting policy updates (Perchet et al., 2016). Our geometric epochs keep the score means aligned while collecting information during greedy play. The centered explore-then-commit comparison follows the usual fixed exploration-then-commit structure (Lattimore & Szepesvári, 2020), with its estimator adapted to the centered moment in (33).

Other robust estimators. Catoni-based contextual bandits extend variance-sensitive estimation to general reward function classes (Ye et al., 2025); robust parameter estimation under low-order moment assumptions is also developed by Hsu & Sabato (2016). Lemma 1 controls the average conditional tangent moment along an adaptive trajectory, where predictable common shifts change the conditional reward mean across rounds.

Spherical updates and Gaussian concentration. The distinction between obtaining initial alignment and subsequently learning the direction also appears in high-dimensional stochastic-gradient theory (Ben Arous et al., 2021). Our initialization uses an estimable centered moment, after which the exact tangent drift yields the angular recursion. The Gaussian norm and maximum bounds used to translate this recursion into regret are standard concentration tools (Vershynin, 2018); the selected-arm decomposition and the quadratic index-gap identity determine how they enter the present analysis.