ContrastSpanner: Learning Low-Rank Causal Contrasts to Alleviate Power-Set and Eluder Barriers

Alec Koppel[⋆], **Laixi Shi**[†]

Keywords: causal contextual bandits, combinatorial interventions, optimal design, spanners, regression oracles

Summary

Causal contextual bandits can expose a *combinatorial intervention set* (e.g., any subset of m nodes in a vision transformer or gene regulatory network, $|\mathcal{A}| \approx 2^m$). Two obstacles then dominate: a *power-set barrier* (quadratic comparisons over $|\mathcal{A}|$) and, under general function approximation, an *Eluder-dimension barrier* whose scaling is hard to certify and can be exponential (Russo & Van Roy, 2013b). We study a realizable setting where the interventional mean $\mu(x, a) = \mathbb{E}[r \mid x, (a)]$ belongs to a rich (possibly nonlinear) oracle class, and where *interventional contrasts* exhibit an unknown low-rank geometry. We propose *ContrastSpanner*, which grows a spanner/ε-net in the *causal contrast space* to *learn* an effective rank r online and to replace explicit $|\mathcal{A}|$ and opaque Eluder dependence by an explicit rank/design complexity in regret. We instantiate the framework with (i) a square-loss regression oracle and (ii) a log-loss oracle (akin to FastCB (Foster & Krishnamurthy, 2021)). We establish sublinear stochastic and minimax regret of the proposed approach holds with high probability under both oracle models, which notably mitigates the aforementioned combinatorial and Eluder scalings. We instantiate this approach in an application for test time self-assessment of perception systems, where we observe effective and scalable performance in practice, despite the parameter complexity of modern vision systems.

Contribution(s)

1. We formulate *causal contextual bandits with combinatorial interventions* under interventional semantics, and highlight the *power-set* and *Eluder-dimension* barriers that arise in causal sequential learning.
   Context: The aforementioned subordination of the SCM, via a specification of do-calculus on a chosen action and realizability, yields a reduction of causal bandits to vanilla bandits with a combinatorially large action space.
2. We introduce **ContrastSpanner**: an ε-net / spanner-growth procedure that *learns* a low-dimensional interventional *contrast span* online. We instantiate this algorithm both for square loss and log loss regression oracles in terms of how one estimates the reward, which allows for a rich class of function approximators.
   Context: Risk notions to statistical error rates/hypothesis testing are beyond scope.
3. We establish the stochastic and minimax regret of the proposed algorithms holds with high probability governed by an intrinsic rank rather than $|\mathcal{A}|$ or Eluder dimension scalings.
   Context: These results require suitable noise conditions and causal identifiability.
4. Experimental comparisons demonstrate the favorable scaling of the proposed approach in practice.
   Context: For the setting that confounding is well-captured by a linear mediator, performance of ContrastSpanner merely matches, but does not outcompete, prior art.

ContrastSpanner: Learning Low-Rank Causal Contrasts to Alleviate Power-Set and Eluder Barriers

Alec Koppel[⋆], **Laixi Shi**[†]
alec.koppel@jhuapl.edu, laixis@jhu.edu

[⋆]**Johns Hopkins University Applied Physics Laboratory, Laurel, MD**
[†]**Dept. of ECE, Johns Hopkins University, Baltimore, MD**

Abstract

Causal contextual bandits can expose a *combinatorial intervention set* (e.g., any subset of m nodes in a causal graph $|\mathcal{A}| \approx 2^m$). Two obstacles then dominate theory and practice: a *power-set barrier* (quadratic comparisons over $|\mathcal{A}|$) and, under general function approximation, an *Eluder-dimension barrier* whose scaling is hard to certify and can be exponential (Russo & Van Roy, 2013b). We study a realizable setting where the interventional mean $\mu(x, a) = \mathbb{E}[r \mid x, (a)]$ belongs to a rich oracle class, and where *interventional contrasts* exhibit an unknown low-rank geometry. We propose *ContrastSpanner*, which constructs a spanner/ε-net in the *causal contrast space* to *learn* an effective rank r online and to replace explicit $|\mathcal{A}|$ and opaque Eluder dependence by an explicit rank/design complexity in regret. We instantiate the framework with (i) a square-loss regression oracle and (ii) a log-loss oracle (akin to FastCB (Foster & Krishnamurthy, 2021)). We establish sublinear stochastic and minimax regret of the proposed approach holds with high probability under both oracle models, which notably mitigates the aforementioned combinatorial and Eluder scalings. We benchmark the proposed approach relative to alternatives both from vanilla contextual bandits and causal methodologies, demonstrating favorable performance when nonlinear confounding and large action spaces are present.

1 Contextual Bandits in the Combinatorial Causal Setting

We study a *causal contextual bandit* with a possibly combinatorial intervention set. In this setting, for times $t = 1, \ldots, T$, a learner (autonomous agent) observes a context $x_t \in \mathcal{X}$, selects an intervention $a_t \in \mathcal{A}$ (e.g., $|\mathcal{A}| \approx 2^m$ for subset interventions on m mediator variables), and a reward $r_t \in [0, R_{\max}]$ is observed with interventional mean $\mu(x, a) := \mathbb{E}[\, r_t \mid x_t{=}x, \mathrm{do}(a_t{=}a)\,]$. While the intervention space governed by a structural equation model could in general be continuous, i.e., associated with $\mathcal{U} \subset \mathbb{R}^m$, we restrict focus to the case that actions are discrete values. We study performance in terms of the benchmark defined by the context-wise optimal intervention $a^\star(x) \in \arg\max_{a \in \mathcal{A}} \mu(x, a)$, and we measure performance in terms of the difference of average rewards accumulated with respect to this context-dependent best-in-hindsight up to time T, i.e., cumulative stochastic regret

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \Big(\mu(x_t, a^\star(x_t)) - \mu(x_t, a_t) \Big). \tag{1}$$

We use the notation $\mathrm{do}(\cdot)$ to emphasize that a represents a *controlled manipulation* in Pearl's sense (Pearl, 2009, Ch. 3). In this paper the structural causal model (SCM) is subsumed into our focus on the interventional mean and a realizability condition, in particular, that the true interventional

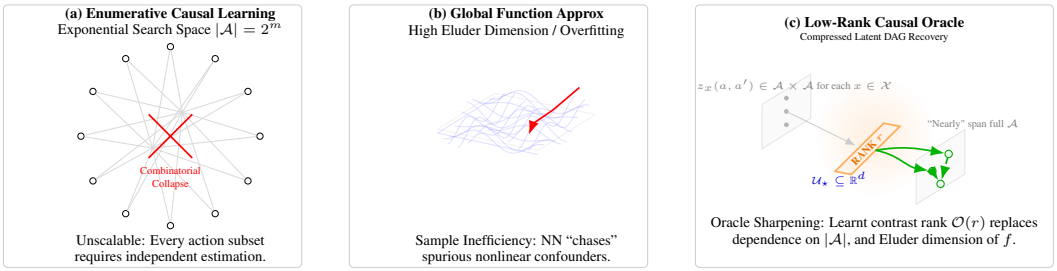


Figure 1: **Causal Learning Paradigms.** (a) Naive approaches scale exponentially in the number of interventions. (b) Generic nonlinear models fail to handle complex confounding (Eluder dimension scaling). (c) Our low-rank contrast framework projects interventions into a latent contrast space, enabling scalable rates for real-valued (square-loss) and categorical (log loss) rewards.

mean is realizable under an action-context regression oracle, similar to (Zhu et al., 2022) – see Sec. 2 for more details. This precludes causal discovery (e.g. (Malinsky & Danks, 2018), and avoids some of the impossibility results regarding identifiability that arise in SCM), although extensions to address this are possible. Further more, we focus on the case where this regression oracle satisfies certain oracle inequalities, i.e., either it exhibits sublinear regret in square-loss (Foster & Rakhlin, 2020) or log-loss (Foster & Krishnamurthy, 2021) when trained sequentially on bandit feedback data. This condition has rigorously established to hold via sequential Rademacher complexity arguments for a variety of function classes, such as linear, kernel, and neural networks, in (Foster & Rakhlin, 2020). For this setting of nonlinear function approximators and sequential causal bandits, two canonical bottlenecks manifest: **(ii) Power-set / combinatorial barrier.** When $|\mathcal{A}|$ is exponential, naive exploration or comparison analyses incur scaling with $|\mathcal{A}|$ or a union bound over $|\mathcal{A}|^2$ pairwise comparisons; **(ii) Eluder barrier.** For rich nonlinear function approximation, regret bounds often scale with Eluder dimension, an opaque quantity that can be exponentially large for natural representations (Russo & Van Roy, 2013a; Yan et al., 2024). This work addresses these barriers in a causal-intervention space via *learned* low-rank contrasts. See Appendix A for discussion of related work (also Table 1). Our **contributions** are thus as follows:

- **Causal contrast features.** We introduce a contrast feature construction tailored to *interventions*, turning causal interventional gaps into a spanner-friendly geometry (Section 2), which exhibits fundamental linkages to defining bandit policies over families of CATE estimators.
- **Rank learning via spanner growth.** We propose an ε-net/spanner-growth rule that adaptively expands the explored intervention set only when it discovers a new contrast direction, thereby *learning* an effective rank r (Section 3).
- **Two oracle-efficient algorithms.** We define ContrastSpanner (squared-loss oracle) and FastContrastSpanner (log-loss oracle), both compatible with general nonlinear reward approximators over contexts (Section 3).
- **Convergence theory.** We establish that the stochastic and minimax regret of the proposed approaches is sublinear with high probability for both oracle models. In particular, this reduces typically combinatorial dependencies to an effective rank scaling r in lieu of $|\mathcal{A}|$ or Eluder dimension of the associated function class (Section 4), *without linearity*.
- **Experiments in combinatorially large spaces.** We benchmark the proposed approach relative to alternatives both from vanilla contextual bandits and causal methodologies (Sec. 5), demonstrating favorable performance when nonlinear confounding and large action spaces are present.

2 Contrast Features, Oracle Inequalities, and Combinatorial Bottlenecks

In this work, our goal is to introduce a technical setting that enables nonlinear function approximators to be used for reward prediction in a scalable manner for causal bandits. To do so, begin then by

introducing a known (parameterized) predictor $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, e.g., a neural network or kernel regression function, which, in the causal setting, is used to estimate *differences* between candidate actions. In causal problems, decision-relevant quantities are *interventional contrasts*

$$\tau(x; a, a') := \mu(x, a) - \mu(x, a') \ , \tag{2}$$

known also the conditional average treatment effects (CATE). Observed contrasts, not merely observed conditional average rewards, are the essential element here. Thus, given reward predictor f, we define a CATE plug-in estimate $\Delta_f(x; a, a')$ the contrast vector at context x between two interventions a, a' under predictor f as

$$\Delta_f(x; a, a') := f(x, a) - f(x, a') \in \mathbb{R}^d. \tag{3}$$

Further define $z_x(a, a') \in \mathbb{R}^d$ as an (explicit or implicit) embedding induced by f, such that control of $\|z_x(a, a')\|$ suffices to control the bias of plug-in estimator $\Delta_f(x; a, a')$ in (3). In canonical references, uniform estimation of CATE across all interventional pairs a, a' may be required. Instead, here, we introduce a variation of the context-action oracle model from (Zhu et al., 2022)[Eq. (1), p. 2] for CATE and a low rank condition that permits a scalable no-regret learning under some appropriate. conditions (Assumption 1-2 in Appendix D) Then, there is structure on the true contrasts that we seek to estimate with reward predictor f. Depending on the parameterization of f or the specific oracle model, z_x may take some different tangible forms, detailed next.

(i) For the case of linear contextual bandits, i.e., if $f(x, a) = \langle \theta_\star, \phi(x, a) \rangle$, then

$$\Delta_f(x; a, a') = \langle \theta_\star, \phi(x, a) - \phi(x, a') \rangle \ , \qquad z_x(a, a') := \phi(x, a) - \phi(x, a').$$

This exactly corresponds to the action-context oracle model in (Zhu et al., 2022).

(ii) For local linearization of nonlinear models, i.e., neural networks defined in terms of parameters $\theta \in \Theta \subset \mathbb{R}^d$, write $z_x(a, a') := \nabla_\theta f_\theta(x, a) - \nabla_\theta f_\theta(x, a')$ if the reward is real-valued.

(iii) When there is a specific parameterization over categorical/ordinal outcomes, we can consider a score-based alternative, such as $z_x(a, a') := \nabla \ell_{\log}(f(x, a)) - \nabla \ell_{\log}(f(x, a'))$, which arises naturally when a softmax layer/ logits are under consideration.

(iv) An additional important case is when one explicitly can exploit a structural causal model. In particular, If interventions affect a low-dimensional mediator $m(x, a)$ and $\mu(x, a) = \tilde{h}_x(m(x, a))$, then $z_x(a, a') = m(x, a) - m(x, a')$, and rank is determined via a mediator.

Remark 3 (Appendix A) discusses the relationship between contrast features and CATE estimators.

2.1 Formalizing the power-set and Eluder bottlenecks

Power-set barrier. In settings with combinatorial interventions, action space $\mathcal{A}$ is often exponential in an underlying dimension (e.g., $|\mathcal{A}| = 2^m$ due to a collection of m potential variables one can intervene on, such as calibration of a perception model post-deployment). Whenever we enumerate actions as atomic—either algorithmically or in the analysis—there is a fundamental scalability obstacle. To be more precise, standard bandits maintain action-wise confidence bounds $\mu(x, a) \pm \text{bonus}_t(x, a)$ and select $a_t = \arg\max_a \text{UCB}_t(x_t, a)$. Even if oracle access to $\mu(x, a)$ is available, implementing or analyzing such methods requires either iterating over all $a \in \mathcal{A}$ or controlling estimation error uniformly over $\mathcal{A}$ (via union bound), which results in a $|\mathcal{A}|^2$ (or $|\mathcal{A}|$ per context) scaling in the regret. Prior work on combinatorial or causal bandits typically avoids this barrier by restricting the intervention set via either linearity, or a small set of known mediators.

Eluder barrier. When rewards are modeled by a general nonlinear function class $\mathcal{F}$, exploration is characterized by, e.g., the Eluder dimension (Russo & Van Roy, 2013a), which quantifies the length of the longest sequence of context–action pairs on which the function class can remain mutually indistinguishable while inducing different predictions elsewhere. This can be much larger than the ambient feature dimension that appears in the linear cases, i.e., regret for such settings scale at least

as $\tilde{O}(\sqrt{T \operatorname{Eluder}(\mathcal{F})})$, including recent causal extensions that replace linear structure with Lipschitz or neural function classes (Yan et al., 2024). Noticeably, Eluder dimension is notoriously difficult to compute or even upper bound sharply for modern predictors, and for common architectures it can scale exponentially with input dimension or depth. Moreover, confidence sets constructed via Eluder notions are *global*: they require controlling prediction error uniformly over the entire action space and function class, even when decision-relevant uncertainty is confined to a much lower-dimensional subspace. As a result, Eluder-dimension bounds often obscure the true statistical difficulty of the problem and lead to pessimistic exploration guarantees. These limitations motivate exploration strategies based on *design* rather than generic confidence sets: by exploiting geometric structure in how interventions affect rewards, one can replace opaque function-class complexity terms with problem-dependent quantities tied to the effective dimension of the decision-relevant subspace. We construct such an approach in the next section.

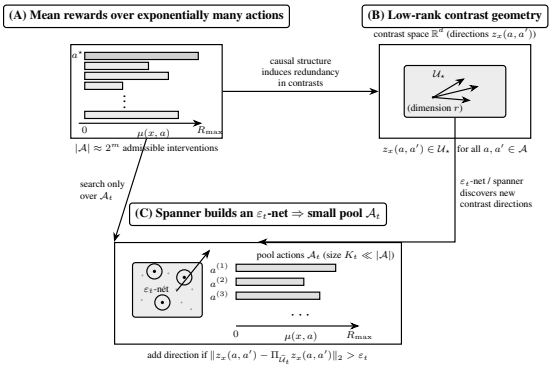


Figure 2: ContrastSpanner targets a low-rank interventional contrast span. Even when $|\mathcal{A}|$ is exponential (A), Assumption 2 posits that the number of required average reward estimates, through contrast vectors $z_x(a, a')$ from (3) lie in an unknown r-dimensional subspace $\mathcal{U}_\star \subseteq \mathbb{R}^d$ (B). The spanner/ε-net is grown online to *learn* the effective rank by adding a new direction only when it detects a residual outside the current span (C).

3 Contrast Spanners

This section describes how to operationalize the low-rank interventional contrast structure introduced in Section 2. The algorithms we present maintain a compact representation of the decision-relevant contrast geometry and use it to guide exploration and learning without enumerating the exponential action space $\mathcal{A}$. This hinges upon three core concepts that drive the algorithmic design: (i) ε-nets and spanners for contrast spaces, (ii) a residual test that allows the learner to discover new contrast directions from bandit feedback, and (iii) the use of a regression oracle to parameterize the bandit policy. To make the role of contrast growth concrete, we introduce regression oracles first.

3.1 Regression Oracles

Squared-loss oracle (Sq). SquareCB (Foster & Rakhlin, 2020, Sec. 2) assumes an online squared-loss regression oracle whose cumulative excess risk is controlled. Informally, for a sequence of examples (x_t, a_t, r_t), the oracle produces predictors $\hat{f}_t$ such that

$$\operatorname{Reg}_{\mathrm{Sq}}(T) := \sum_{t=1}^{T} \left(\hat{f}_t(x_t, a_t) - r_t\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(f(x_t, a_t) - r_t\right)^2 \;\le\; \mathsf{R}_{\mathrm{sq}}(T), \tag{4}$$

where $\mathsf{R}_{\mathrm{sq}}(T)$ is an oracle regret term (see (Foster & Rakhlin, 2020)[Eq. (2), p. 2] that grows at worst sublinearly $\mathcal{O}(\sqrt{T})$ via sequential Rademacher complexity arguments for a variety of function classes (linear statistical models, kernels, and neural networks) (Foster & Rakhlin, 2020)[Sec.3]. For this oracle at step t, an associated CATE plug-in estimate, given f_t, is:

$$\Delta_{f_t}(x_t; a) := f_t(x_t, a) - f_t(x_t, a') \tag{5}$$

Log-loss oracle (Log). For click/no-click, class labels, or vocabulary spaces in LLM response generation, rewards may be categorical $r \in \{1, \ldots, K\}$ instead of real-valued, in which case a log-loss oracle arises . This oracle returns a predictor f that parameterizes a conditional distribution via logits $\eta_f(x, a) \in \mathbb{R}^K$ which is associated with a softmax: $q_f(k \mid x, a) := \frac{\exp(\eta_{f,k}(x,a))}{\sum_{k=1}^{K} \exp(\eta_{f,k}(x,a))}$.

Typical examples include: *(i)* linear logits $\eta_f(x,a) = W\phi(x,a)$ (multinomial logistic regression with feature map $\phi(x,a) \in \mathbb{R}^d$ and parameter matrix $W \in \mathbb{R}^{K\times d}$); *(ii)* nonlinear logits $\eta_f(x,a) = g_\theta(x,a)$ produced by a neural network with parameters $\theta \in \mathbb{R}^d$; and *(iii)* mediator-based logits, $m(x,a) \in \mathbb{R}^r$ where the oracle predictor depends on a only through $m(x,a)$, in which case a matrix $A \in \mathbb{R}^{K\times r}$ exists such that $\eta_f(x,a) = A\, m(x,a)$, i.e., interventions act through a low-dimensional causal representation. We do note that logits are only defined up to an additive constant, so it makes sense to define contrasts in terms of centered logits to eliminate scaling ambiguity, i.e.,

$$\tilde{\eta}_f(x,a) = \eta_f(x,a) - \frac{1}{K}\mathbf{1}\mathbf{1}^\top \eta_f(x,a)\ , \qquad \Delta_f^{\text{logit}}(x;a,a') := \tilde{\eta}_f(x,a) - \tilde{\eta}_f(x,a'). \tag{6}$$

The plug-in for (6) is simply its evaluation at a particular oracle f_t, similar to (5). This contrast corresponds to log-likelihood ratio directions between the predicted interventional outcome distributions. In particular, FastCB (Foster & Krishnamurthy, 2021, Sec. 2) defines an online log-loss oracle to fit a conditional distribution $q(\cdot \mid x,a)$ for outcome $r \in \{1,\dots,K\}$, yielding

$$\mathrm{Reg}_{\mathrm{Log}}(T) := \sum_{t=1}^{T} \ell_{\log}\big(q_t(\cdot \mid x_t,a_t), r_t\big) - \inf_{q\in\rho(\mathcal{A})} \sum_{t=1}^{T} \ell_{\log}\big(q(\cdot \mid x_t,a_t), r_t\big) \ \le\ R_{\log}(T). \tag{7}$$

where here q and q_t are parameterized by f and f_t, respectively. This probabilistic calibration enables sharper regret than square losses under appropriate conditions (Foster & Krishnamurthy, 2021)[Eq. (2), p. 3]. Here $\rho(\mathcal{A})$ denotes the probability simplex over action space $\mathcal{A}$.

3.2 Spanners and ε-nets for contrast spaces

Fix a round t and context x. Let $\mathcal{Z}_x := \{z_x(a,a') : a,a' \in \mathcal{A}\}$ denote the (potentially enormous) set of interventional contrast features (defined following (3). Under Assumption 2, $\mathcal{Z}_x$ lies in an unknown r-dimensional linear subspace $\mathcal{U}_\star \subseteq \mathbb{R}^d$ (formalized later). A *spanner* (or ε-net) for $\mathcal{Z}_x$ is a finite set $\mathcal{S}_x \subseteq \mathcal{A}$ such that every contrast feature $z_x(a,a') \in \mathcal{Z}_x$ can be approximated, up to error ε, by a linear combination of contrast features induced by actions in $\mathcal{S}_x$, i.e., for $\widehat{\mathcal{U}} := \mathrm{span}\{z_x(a,a') : s,s' \in \mathcal{S}_x\}$, a ε-net satisfies

$$\|z_x(a,a') - \Pi_{\widehat{\mathcal{U}}} z_x(a,a')\|_2 \le \varepsilon \quad \text{for all } a,a' \in \mathcal{A}. \tag{8}$$

where parameter ε controls the tradeoff between approximation accuracy and spanner size: smaller ε may mean that more elements enter the span in order to achieve a smaller contrast approximation error. Under Assumption 2, there exists a spanner with cardinality $\mathcal{O}(r\log(1/\varepsilon))$ (Prop. 3).

Next, we provide a constructive procedure for the algorithms. ε is implemented implicitly through a time-varying residual threshold ε_t; a new action is added to the spanner only if its contrast feature cannot be approximated within tolerance ε_t by the current span. Let $\widehat{\mathcal{U}}_t$ be the current spanner span at time t. For a candidate action a, define the residual in terms of the regularized contrast variance:

$$\mathrm{Res}_t(x;a) := \|z_{x_t}(a,a') - \Pi_{\widehat{\mathcal{U}}_t} z_{x_t}(a,a')\|_{V_t^{-1}} \qquad V_t = \sum_{s\in S_t} z_{x_s}(s,s') \tag{9}$$

The residual $\mathrm{Res}_t(x;a) = \sqrt{z_{x_t}(a,a')^T V_t^{-1} z_{x_t}(a,a')}$ is defined in terms of the contrast feature $z_x(a,a')$, which is not directly observed. Instead, the learner relies on the plug-in contrast $\Delta_{f_t}(x;a,a') = f_t(x,a) - f_t(x,a')$ produced by the regression oracle. Under Assumption 1, large geometric residuals correspond to directions along which Δ_{f_t} cannot be explained by the current span, up to the oracle's estimation error.

Concretely, we select the threshold ε_t so that directions with $\mathrm{Res}_t(x;a) \le \varepsilon_t$ are indistinguishable from oracle noise, while directions with $\mathrm{Res}_t(x;a) > \varepsilon_t$ represent new decision-relevant contrasts.Thus, the residual test allows the learner to *discover the rank r online* from bandit feedback, without access to the true contrasts. Then, an explicit representation of the contrast features

z_{x_t} that are retained in spanner set are defined as $\mathcal{S}_t = \{z_{x_{\kappa_t(j)}}(a_{\kappa_t(j)}, a'_{\kappa_t(j)}), j \in \mathcal{I}_t\}$, where $\kappa_t : \mathcal{I}_t \to \{1, \dots, t-1\}$ maps spanner element j to the round when it was added, with indexing set $\mathcal{I}_t = \{1, \dots, m_t\}$. Further, $(a_{\kappa_t(j)}, a'_{\kappa_t(j)})$ are the action pair recorded at step $\kappa_t(j)$. See Fig. 2 for an intuition of how this operates based upon sparse recovery principles.

Given the current spanner $\mathcal{S}_t$, the policy is parameterized by restricting attention to actions whose contrast features lie in $\widehat{\mathcal{U}}_t = \mathrm{span}\{z_{x_s}(s, s') : s, s' \in \mathcal{S}_{\tilde{t}}, \tilde{t} \le t\}$. Exploration is then carried out by sampling actions according to a distribution supported on $\mathcal{S}_t$, which can additionally be selected to minimize the variance of contrast estimation in $\widehat{\mathcal{U}}_t$ (Kiefer & Wolfowitz, 1960; Kiefer, 1974). We detail this in the following subsection.

Remark 1. For conceptual and expositional simplicity, we write this residual test for all pairs of actions (a, a') contained in the indexing set $\mathcal{S}_t$. However, in practice, computing the residual test over all pairwise comparisons is unnecessary if one uses anchoring. In particular, when a single action corresponds to no intervention or a notion of a placebo, it naturally may form an "anchor" action a_0, in which case evaluating (9) over all (a, a') pairs is not needed, but instead one only needs to search over (a, a_0). See Appendix C.0.1 for precise detail. In Appendix C.0.1 (Lemma 1), we clarify that doing so does not reduce the performance of the learnt spanner.

3.3 Policy construction: gaps, regularization, and design-solve

Inverse gap weighting. To define inverse gap weighting (Abe & Long, 1999; Foster & Rakhlin, 2020), we begin by denoting at each time-step the maximizer of our predictor as $\hat{a}_t \in \arg\max_{a \in \mathcal{A}} f_t(x_t, a)$[1]. Then, the (one-sided) plug-in gap for each action a is

$$\widehat{\Delta}_t(a)^{\mathrm{sq}} := f_t(x_t, \hat{a}_t) - f_t(x_t, a) \tag{10}$$

$$\widehat{\Delta}_t(a)^{\log} := \sum_k [r(x, a) = k] \left(q_{f_t}(k \mid x, a) - q_{f_t}(k \mid x, \hat{a})\right) \tag{11}$$

which is $\Delta_t(a) = \Delta_{f_t}(a_t^\star, a)$ in the two-action notation of (3). The later gap is simply the difference in the predicted mean associated with the softmax distribution over categorical outcomes $r_t(x, a) \in \{1, \dots, K\}$. Given a candidate set of actions (which we later refer to as pool) $\mathcal{A}_t \subseteq \mathcal{A}$ and a greedy action $\hat{a}_t \in \arg\max_{a \in \mathcal{A}_t} f_t(x_t, a)$, define IGW by

$$\pi_t(a \mid x_t) := \frac{1}{\vartheta + \gamma_t \hat{\Delta}_t(a)}, a \in \mathcal{A}_t \setminus \{\hat{a}_t\}, \qquad 1 - \sum_{b \in \mathcal{A}_t \setminus \{\hat{a}_t\}} \frac{1}{\vartheta + \gamma_t \hat{\Delta}_t(b)}, a = \hat{a}_t. \tag{12}$$

where $\gamma_t > 0$ is a step-size-like parameter that we will later select to ensure sublinear regret, and ϑ is a regularizer such that, if $\vartheta \ge |\mathcal{A}_t|$, this is a valid distribution on $\mathcal{A}_t$. The idea of IGW is to assign higher arm likelihood to near-optimal actions while ensuring exploration across $\mathcal{A}_t$. We present a slightly different form of (12) from Foster & Rakhlin (2020) so that we may unify the two different oracle models (Sec. 3.1).

Optional optimal experimental design (OED). For stochastic regret, sampling $\mathcal{A}_t$ with respect to IGW (12) is sufficient. However, optimal experimental design can be used to mitigate reward estimation variance, which is useful when the variance of each gap varies by arm, i.e., heteroskedastic noise. This reweighting is additionally useful to establish minimax regret. For these reasons, we present two possible distributions. In particular, with current spanner $\mathcal{S}_t$ (where we suppress the t subscipt in $\mathcal{S}_t$ for easier interpretation) over contrasts $z \in \mathcal{Z} \subseteq \mathbb{R}^d$, we may solve for $p_t \in \rho(\mathcal{S})$[2] to minimize a standard design criterion such as

$$\text{(G-opt)} \quad \min_{p \in \rho(\mathcal{S}_t)} \max_{z \in \mathcal{S}_t} z^\top V_t^{-1} z, \qquad \text{(A-opt)} \quad \min_{p \in \rho(\mathcal{S}_t)} \mathrm{tr}\left(V_t^{-1}\right). \tag{13}$$

[1] The reason to define $\hat{a}$ as the solution to the argmax, rather than $a^\star$ is that the later quantity is the maximizer of the true average reward, which without oracle access to the reward distribution, is not computable during runtime.

[2] Here we use $\rho(\mathcal{S})$ as the probability simplex over $\mathcal{S}$.

Algorithm 1: CONTRASTSPANNER (square-loss)

Require: Regularizer $\lambda > 0$, residual thresholds (ε_t), gap parameter/step-size (γ_t).

Initialize $\mathcal{S}_1 \leftarrow \emptyset$, $V_1 \leftarrow \lambda I$, cache $\mathcal{D}_0 \leftarrow \emptyset$.
for $t = 1, 2, \ldots, T$ **do**
  Observe context x_t.
  Oracle update: $f_t \leftarrow \mathrm{AlgSq}(\mathcal{D}_{t-1})$.
  Form contrast set at x_t and find pair (a, a') with $\mathrm{Res}_t(a, a') > \varepsilon_t$, [cf. (9)].
  if such (a, a') exists **then**
    Add to $\mathcal{S}_t$.
    Update $V_t \leftarrow V_t + z_{x_t}(a, a') z_{x_t}(a, a')^\top$.
  end if
  Compute action $a_t^\star \in \mathrm{argmax}_{a \in \mathcal{A}} f_t(x_t, a)$.
  Define IGW distribution $\pi_t(\cdot \mid x_t)$ on a finite action pool $\mathcal{A}_t$ using (10) and (12). (Optionally reweight π_t [cf.(13)])
  Sample $a_t \sim \pi_t(\cdot \mid x_t)$ and apply $\mathrm{do}(a_t)$.
  Observe reward r_t; append (x_t, a_t, r_t) to $\mathcal{D}_t$.
end for

Algorithm 2: FASTCONTRASTSPANNER (log)

Require: Regularizer $\lambda > 0$, residual thresholds (ε_t), gap parameter/step-size (γ_t).

Initialize $\mathcal{S}_1 \leftarrow \emptyset$, $V_1 \leftarrow \lambda I$, cache $\mathcal{D}_0 \leftarrow \emptyset$.
for $t = 1, 2, \ldots, T$ **do**
  Observe context x_t.
  Oracle update: $q_t \leftarrow \mathrm{AlgLog}(\mathcal{D}_{t-1})$.
  Via q_t, compute mean predictor
  $$f_t(x, a) = \sum_k [r(x, a) = k] q_{f_t}(k \mid x, a).$$
  Find contrast pair (a, a') with $\mathrm{Res}_t(a, a') > \varepsilon_t$ [cf. (9)].
  if such (a, a') exists **then**
    Add $z_{x_t}(a, a')$ to $\mathcal{S}_t$
    Update $V_t \leftarrow V_t + z_{x_t}(a, a') z_{x_t}(a, a')^\top$.
  end if
  Compute action $a_t^\star \in \mathrm{argmax}_{a \in \mathcal{A}} f_t(x_t, a)$.
  Define IGW dist. $\pi_t(\cdot \mid x_t)$ [cf. (11), (12)]. (Optionally: reweight π_t by Fisher-weighted design p_t (110).)
  Sample $a_t \sim \pi_t(\cdot \mid x_t)$ and apply $\mathrm{do}(a_t)$.
  Observe reward $r_t \in \{1, \ldots, K\}$; Append (x_t, a_t, r_t) to $\mathcal{D}_t$.
end for

One can then replace π_t in (12) by p_t (pure design), or combine them (e.g., sample from π_t but compute gaps using f_t). In the log-loss oracle setting, it is natural to weight directions by the local Fisher information, but we defer its expression for expositional clarity to Appendix E.4. In Algs. 1 - 2, when we specify that one may optionally reweight (12) by p_t, then one substitutes $\pi_t(a \mid x_t) p_t(a)$ in place of $\pi_t(a \mid x_t)$. See (26) for an expanded form of this expression.

3.4 Algorithms

We now summarize the full procedures. Both algorithms have the same skeleton: (i) use an oracle to produce f_t, (ii) form causal contrasts Δ_{f_t}, (iii) update a spanner/ε-net $\mathcal{S}_t$ that (implicitly) learns the effective contrast rank, (iv) solve an optimal-design problem on $\mathcal{S}_t$, (v) sample with IGW and feed an importance-weighted example back to the oracle. Expanded presentations ofAlgs. 1 - 2 are in supplementary (Appendix C). We also provide a concept diagram as Fig. 5 in Appendix C.1.

Next, we formalize that the proposed methods alleviate combinatorial and Eluder bottlenecks that are standard to the setting of bandit feedback with causal interventions, under appropriate conditions. In Appendix C.1, we provide an expanded discussion of the implementation of Algs. 1 - 2.

4 Stochastic and Minimax Regret Without Combinatorial or Eluder Scaling

This section summarizes the regret guarantees for CONTRASTSPANNER (square-loss oracle; Algorithm 1) and FASTCONTRASTSPANNER (log-loss oracle; Algorithm 2). To keep the main body readable, we state *succinct* bounds emphasizing the dependence on (i) oracle regret, (ii) the reward scale $R_{\max}$, and (iii) the intrinsic contrast dimension r. Expanded statements with explicit constants, as well as complete proofs, are provided in the supplementary appendices: Appendix E (square-loss), Appendix E.4 (log-loss), and Appendix F (low-rank structure, spanner growth, and minimax lower bounds).

Throughout, we work under bounded rewards and martingale noise (Assumption 4). For the square-loss results we assume an online square-loss oracle guarantee (Assumptions 3.1) and realizability

(Assumption 5). For the log-loss results we assume a log-loss oracle guarantee (Assumptions 3.2), as well as categorical realizability with clipping, together with the curvature/Fisher regularity needed for the quadratic reductions used in the analysis (Assumption 7) . The intrinsic rank condition (Assumption 2) states that all *contrast features* live (exactly or approximately) in an r-dimensional subspace, which we later rigorously justify under certain structural causal models or information-theoretic properties of their associated DAG. For minimax results, we impose a design-oracle assumption (Assumption 6) which ensures the G-optimal design is evaluated up to some numerical approximation factor over the learned spanner directions.

On the "span gap" term. Regret results below include $\sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t)$, the cumulative suboptimality incurred by restricting action selection to the learned pool $\mathcal{A}_t$. This term is *zero* if the optimal action is always contained in the pool, and more generally it is controlled by the residual/spanner certificates in Appendix F (e.g., Lemma 19 and Corollary 1), which are determined by our algorithm parameter ε, and ensured to be small under appropriate conditions (Sec. 4.3).

4.1 Square-loss oracle: CONTRASTSPANNER

Theorem 1 (Stochastic regret for CONTRASTSPANNER (Alg. 1))**.** *Under Assumptions 2, 3.1, 4, and 5, run Algorithm 1 for T rounds with parameters $(\vartheta, \gamma, \{\varepsilon_t\})$ satisfying $\vartheta \geq \max_{t\leq T} |\mathcal{A}_t|$ and with γ chosen as in Appendix E. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\mathrm{Reg}(T) \;\leq\; \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) \;+\; \widetilde{\mathcal{O}}\left(\sqrt{\vartheta T\big(\mathsf{R}_{\mathrm{sq}}(T) + R_{\max}^2 \log(1/\delta)\big)} \;+\; R_{\max}\sqrt{T\log(1/\delta)}\right), \quad (14)$$

where $\mathsf{R}_{\mathrm{sq}}(T)$ is the upper-bound on the square-loss oracle regret (Assumption 5) from (4), and $\widetilde{\mathcal{O}}(\cdot)$ suppresses absolute constants and logarithms in (T, λ) that are made explicit in Appendix E.

Full proof is in Appendix E, which adapts SQUARECB analysis Foster & Rakhlin (2020): (i) decompose regret into a span gap plus in-pool regret (Appendix F); (ii) control in-pool regret via a one-step IGW inequality and a log-barrier summation (Appendix E); (iii) convert square-loss oracle regret into prediction error using realizability (Assumption 5); and (iv) apply Freedman's inequality for bounded martingale differences (Assumption 4 (Lemma 3). Different from SQUARECB and why this removes "power-set" dependence, is that, in the place where the action set size enters the proof is through the exploration-validity constraint $\vartheta \geq |\mathcal{A}_t|$ for the IGW distribution, we substitute this dependence by the contrast geometry. In particular, $\mathcal{A}_t$ is *learned* by a contrast spanner and its size is controlled by intrinsic geometry: under Assumption 2, Proposition 3 implies $|\mathcal{A}_t| = \widetilde{\mathcal{O}}(r)$ (rather than scaling with the ambient $|\mathcal{A}|$, which can be combinatorial). Thus the SQUARECB bound carries over with "$|\mathcal{A}|$" replaced by the spanner size, making the exploration complexity depend on the intrinsic contrast dimension r instead of the full action set. The Eluder dependence is further avoided through oracle reductions, which instead depend on a Rademacher complexity argument which provably attenuates with T (Van Der Vaart & Wellner; Koltchinskii, 2006). This result, however, does not account for the "information loss" associated with restricting to the pool, which can be quantified by the gap estimation variance associated with actions outside the pool. This issue can be more sharply characterized from the perspective of performance against a worst-case sequence of noises and contexts (minimax-style), which we adopt next.

Theorem 2 (Minimax regret for CONTRASTSPANNER (Alg. 1) via optimal design)**.** *Under Assumptions 2, 3.1, 4, and 5, run Algorithm 1 with the G-optimal design distribution at each round (Assumption 6) and parameters q_t chosen as in Appendix E.3 [cf. (61)]. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over any (possibly adaptive) context sequence and any reward/noise process satisfying the stated assumptions,*

$$\mathrm{Reg}(T) \;\leq\; \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) \;+\; \widetilde{\mathcal{O}}\left(r\sqrt{T\big(\mathsf{R}_{\mathrm{sq}}(T) + R_{\max}^2 \log(1/\delta)\big)} \;+\; R_{\max}\sqrt{T\log(1/\delta)}\right), \quad (15)$$

where $r := \max_{t\leq T} \dim(\mathrm{span}(\mathcal{S}_t))$ is the intrinsic contrast dimension (Assumption 2), and the $\widetilde{\mathcal{O}}(\cdot)$ hides only absolute constants and logarithms made explicit in Appendix E.3.

The proof (Appendix E.3) replaces the stochastic IGW penalty $\sqrt{\vartheta T(\cdot)}$ in (14) by an $r\sqrt{T(\cdot)}$ term by choosing the exploration distribution using *optimal experimental design* (OED) over the spanner directions. The key technical step is a *DEC/OED certificate* showing that a (approximate) G-optimal design controls the worst-case variance of all spanner contrasts in the learned subspace; the remainder of the proof then follows the standard DEC-to-regret reduction used by SPANNERIGW Zhu et al. (2022) and related oracle-based contextual bandit analyses to alleviate Eluder dependencies. While we use the same "DEC + G-optimal design" scaffold as SPANNERIGW Zhu et al. (2022), we introduce a way for it to occur within the spanner pool. In particular, in SPANNERIGW the geometry is imposed on a fixed action embedding; instead, here the OED is performed on *causal contrast features* that are *learned* online by the spanner. Consequently, the width (and hence regret) depends on the intrinsic contrast rank r rather than on an ambient embedding dimension or on $|\mathcal{A}|$.

4.2 Log-loss oracle: FASTCONTRASTSPANNER

Theorem 3 (Stochastic regret for FASTCONTRASTSPANNER (Alg. 2))**.** *Under Assumptions 2, 3.2, 4, and 7, run Algorithm 2 for T rounds with parameters satisfying $\vartheta \geq \max_{t\leq T}|\mathcal{A}_t|$ and chosen as in Appendix E.4. Then for any $\delta \in (0,1)$, with prob. at least $1-\delta$, the following regret bound holds:*

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) \;+\; \widetilde{\mathcal{O}}\left(\sqrt{\vartheta T\big(\mathsf{R}_{\log}(T) + R_{\max}^2 \log(1/\delta)\big)} + R_{\max}\sqrt{T\log(1/\delta)}\right), \quad (16)$$

where $\mathsf{R}_{\log}(T)$ is the upper-bound on log-loss regret (Assumption 3) and the hidden constants depend on the clipping/curvature parameters in Assumptions 7 (as made explicit in Appendix E.4).

The proof (given in Appendix E.4) adapts the FASTCB analysis Foster & Krishnamurthy (2021): (i) bound regret by a "small-loss" term plus a prediction-error term; (ii) relate prediction error to *triangular discrimination* between the true and predicted categorical distributions (Topsoe, 2002; Le Cam, 2012); (iii) control the resulting martingale terms using Freedman under clipping (Assumption 2) and bounded rewards (Assumption 4); and (iv) use the contrast spanner to replace ambient-action dependence by the learned pool size. A key point of departure from prior art in this analysis is that triangular discrimination is used to control *causal contrast* errors along spanner directions rather than errors on a fixed finite action set, which is where the causal/interventional structure enters the FASTCB-style analysis: the spanner identifies a small set of interventions whose induced contrasts control all others (up to $\varepsilon_{\mathrm{span}}(t)$), and triangular discrimination provides the correct quantitative bridge from log-loss oracle regret to reward regret. Then, the oracle takes the place of typical Eluder dependencies, which can be exponentially larger than those Rademacher complexity bounds associated with oracle reduction. However, similar to the analysis for real-valued rewards and square losses, we have not accounted for the gap estimation variance associated with the pool reduction. In the worst case, however, this can yield substantial performance degradation, i.e., if noise is heavy-tailed or contexts are drawn from a worst-case distribution. We formalize that design weighting in a Fisher geometry can resolve this issue in terms of a minimax-style regret, stated next.

Theorem 4 (Minimax regret for FASTCONTRASTSPANNER (Alg. 2) via Fisher-optimal design)**.** *Under Assumptions 2, 3.2, 4, and 7, run Algorithm 2 with Fisher-rescaled G-optimal designs [cf. (110)] satisfying Assumption 6 [cf. (111)] and parameters chosen as in Appendix E.4. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over any (possibly adaptive) context sequence and any reward/noise process satisfying the stated assumptions,*

$$\mathrm{Reg}(T) \;\leq\; \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) \;+\; \widetilde{\mathcal{O}}\left(r\sqrt{T\big(\mathsf{R}_{\log}(T) + R_{\max}^2 \log(1/\delta)\big)}\right), \quad (17)$$

where $r := \max_{t\leq T}\dim(\mathrm{span}(\mathcal{S}_t))$ and the hidden constants depend on the clipping/curvature parameters in Assumptions 7 (see Appendix E.4 for the explicit dependence).

See Appendix E.4.2 for proof. As in the square-loss minimax analysis, we obtain a DEC certificate (Foster et al., 2021), but now the design is performed in the *local Fisher geometry* induced by the clipped categorical model, whose interaction effect is a unique contribution of this work, and is

inspired by (Agarwal et al., 2024). Triangular discrimination controls log-loss excess and induces a quadratic form comparable to the Fisher metric under Assumption 7, which is why Fisher-rescaled design features appear in the OED step (see Lemma 12). Morever, integrating this Fisher-weighted OED certificate with the *learned contrast span* through the weighted design problem allows us to control the gap variance, and in particular, establish a minimax rate governed by the intrinsic contrast dimension r rather than by the ambient action set.

4.3 Low-rank structure, sufficiency, and lower bounds

The regret guarantees above depend on an intrinsic contrast dimension r (Assumption 2) and on the span-gap term $\sum_t \varepsilon_{\mathrm{span}}(t)$. Appendix F provides formal conditions under which these quantities are small, in particular, the setting under which we may establish Lemma 19 (valid under Assumption 1), which ensures that $\varepsilon_{\mathrm{span}}(t) \le 2B\sqrt{\lambda}\,\varepsilon_t$. Then, under suitable choice of $\varepsilon_t = \mathcal{O}(T^{-1/2})$, the previously stated regret bounds are all sublinear. Next, we provide a concise set of sufficient structural conditions and explain how they connect to learnability.

Proposition 1 (Sufficient structural conditions for intrinsic contrast rank)**.** *Assumption 2 (intrinsic r-dimensional contrast geometry) is satisfied under either of the following settings:*

(i) ***Linear SEM with a mediator bottleneck (Pearl, 2009):*** *in a linear DAG/SEM, if all directed paths from intervened variables to reward-relevant variables pass through a mediator set M with $|M| \le r$, then the interventional mean contrasts lie in an r-dimensional linear subspace (equivalently, the associated causal transfer map has rank at most r).*

(ii) ***Nonlinear SEM with low-rank Jacobians (Spirtes et al., 2001):*** *in a differentiable nonlinear SEM, if the Jacobian of the interventional mean map has column space contained in a fixed r-dimensional subspace uniformly over the intervention domain, then all interventional mean contrasts lie in that same r-dimensional subspace.*

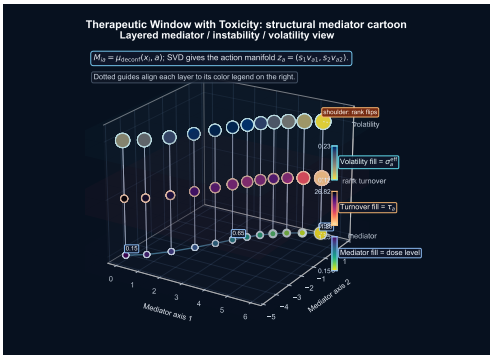
Formal statements and proofs (including precise hypotheses) are in Appendix F.

Remark 2 (Span adequacy, restricted eigenvalues, and when lower bounds are tight)**.** A low-rank *representation* does not by itself guarantee that a small spanner pool contains a near-optimal action; one additionally needs a "span adequacy" condition linking the learned contrast geometry to reward gaps. Appendix F makes this link precise via residual certificates and restricted eigenvalue/RIP-style conditions (e.g., Lemma 20). The same geometric conditions are also what make a rank-based minimax lower bound tight: under an s-RIP/restricted-eigenvalue condition on an s-dimensional contrast subspace, Theorem 5 in Appendix F yields a matching $\Omega(s\sqrt{T})$ minimax lower bound. We therefore only report the lower bound here as a supporting result and defer full details to supplementary.

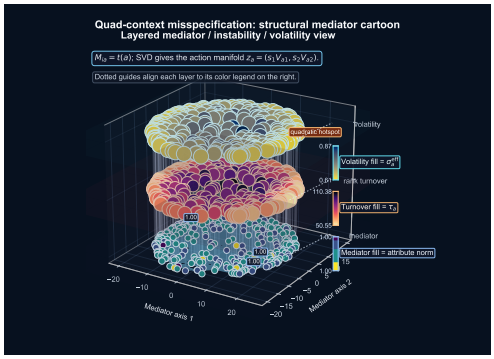
5 Experiments

We evaluate ContrastSpanner and ContrastSpanner+Design against a broad set of baselines. Our *contextual* comparators include regression-oracle contextual bandits (e.g., SquareCB, FastCB, SpannerIGW, CODA) (Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021; Zhu et al., 2022; Li et al., 2022), with linear and Thompson-sampling baselines (Li et al., 2010; Russo & Van Roy, 2013a; Lattimore & Szepesvári, 2020), as well as variants with oracle access to the true mediator. Our *causal* comparators include SEM-based or do-operator bandits (Lattimore et al., 2016; Sen et al., 2017; Varici et al., 2023; Yan et al., 2024).

For each trial we report: (i) *cumulative regret* $\sum_{t=1}^{T}\big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\big)$; (ii) *BAI rate* $\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{a_t = a_t^\star\}$ (fraction of rounds playing the true best arm); and (iii) *coverage* $\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbf{1}\{N_a(T) > 0\}$ with $N_a(T) = \sum_{t=1}^{T}\mathbf{1}\{a_t = a\}$, which diagnoses exploration breadth. All curves show mean $\pm$ one standard error. We consider two families of contextual intervention problems. *Therapeutic Window with Toxicity (TWT)* models a context-dependent dose–response where the best dose is determined by a latent therapeutic window and a toxicity penalty, motivated by ACTG175 (Hammer et al., 1996). *Quad-context misspecification* models a low-rank bilinear

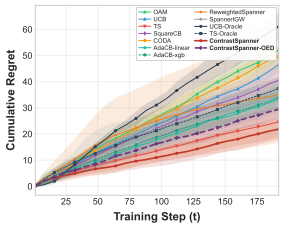


(a) Therapeutic Window with Toxicity (TWT).

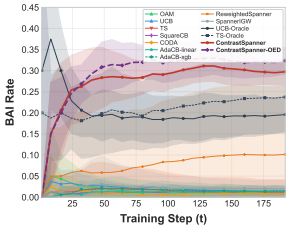


(b) Quad-context misspecification (Quad-miss).

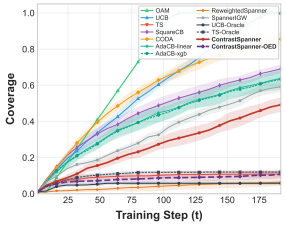
Figure 3: **Problem setup visualizations.** The top layers show a ground-truth deconfounded reward surface over contexts and actions projected onto low-dimensional mediator proxies as quantified by the top two singular values of a stacked SVD. Lower layers visualize decision-relevant structure: as rank instability and heteroskedasticity. Appendix C.3 visualize learnt proxy geometry.



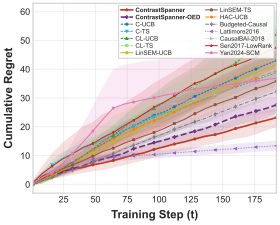
(a) Cumulative regret (contextual).



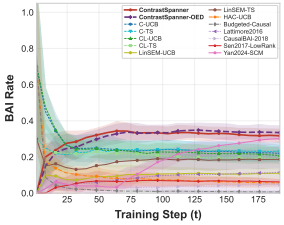
(b) BAI rate (contextual).



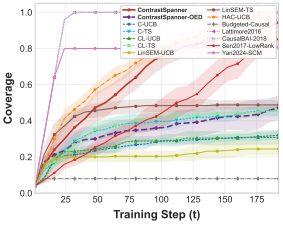
(c) Coverage (contextual).



(d) Cumulative regret (causal).



(e) BAI rate (causal).



(f) Coverage (causal).

Figure 4: **Therapeutic Window with Toxicity: time traces.** Top row: contextual comparators on $K = 100$ actions. Bottom row: causal comparators on $K = 25$ actions. Across both regimes, CONTRASTSPANNER tracks strong contextual baselines with oracle access to the mediator, and remains competitive with causal comparators.

trend with a localized quadratic misspecification that can overturn the best-arm ranking. Figure 3 visualizes the induced mediator geometry of each scenario – see Appendix C.3. Figure 4 reports results for TWT at horizon $T = 192$ averaged over 10 trials. The top row uses a larger action set ($K = 100$) to stress contextual baselines; the bottom row uses a smaller action set ($K = 25$) to compare with causal baselines. Additional time traces for quad-context misspecification, as well as full tables across multiple action-set sizes, are in Appendix C.3.

6 Conclusion and Future Work

We studied causal contextual bandits in the realizable setting, in which the learner has access to realizations of the true interventional mean. In this work, we adopted an approach based on the one hand on adapting low rank representations from compressive sensing to the causal DAG setting, and on the other, modern oracle reductions of such functional complexity measures. The result is the first causal bandit algorithm to mitigate dependence on combinatorially large action spaces at the same time as it may permit nonlinear function approximators.

References

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

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning (ICML)*, pp. 3–11, 1999.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. See Definition 1 for the low-rank transition model formulation.

Alekh Agarwal, Jian Qian, Alexander Rakhlin, and Tong Zhang. The non-linear f-design and applications to interactive learning. In *Forty-first International Conference on Machine Learning*, 2024.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000.

Anthony C. Atkinson, Alexander N. Donev, and Randall D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, 2007.

Alex Ayoub, David Szepesvari, Alireza Bakhtiari, Csaba Szepesvari, and Dale Schuurmans. Rectifying regression in reinforcement learning. In *Reinforcement Learning Conference*.

Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique*, 346(9–10):589–592, 2008.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. DOI: 10.1109/TIT.2006.871582.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. DOI: 10.1214/14-STS500.

Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Partial structure discovery is sufficient for no-regret learning in causal bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c50e3c72bf45a361afc7c16d26c21a1a-Abstract-Conference.html.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 3199–3210, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2107.02237.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021. URL https://arxiv.org/abs/2112.13487.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Heidi Gundacker, Robert T. Schooley, Richard H. Haubrich, William K. Henry, Michael M. Lederman, John P. Phair, Ming Niu, Martin S. Hirsch, and Thomas C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996. DOI: 10.1056/NEJM199610103351501.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. DOI: 10.1080/01621459.1952.10483446.

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

Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, pp. 849–879, 1974.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. 2006.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *PMLR*, pp. 535–543, 2015.

Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure, 2019. URL https://arxiv.org/abs/1901.09490.

Finnian Lattimore, Tor Lattimore, Mark D. Reid, and Csaba Szepesvári. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/b4288d9c0ec0a1841b3b3728321e7088-Abstract.html.

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

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *arXiv preprint arXiv:1003.0146*, 2010. URL https://arxiv.org/abs/1003.0146.

Zhaoqi Li, Lillian Ratliff, Houssam Nassif, Kevin Jamieson, and Lalit Jain. Instance-optimal pac algorithms for contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/f4821075019a058700f6e6738eea1365-Paper-Conference.pdf.

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

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

Zakaria Mhammedi, Adam Block, Dylan J. Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/d2dc4d6c7b102d05f111c02a32e7c6bc-Abstract-Conference.html.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. DOI: 10.1093/biomet/asaa076.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling, 2013a. URL https://arxiv.org/abs/1301.2609.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013b. Introduces the eluder dimension and connects it to sample complexity for exploration.

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sanjay Shakkottai. Contextual bandits with latent confounders: An NMF approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *PMLR*, pp. 518–527, 2017. URL https://proceedings.mlr.press/v54/sen17a.html.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2 edition, 2001.

Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2002.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pp. 16–28. Springer.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *Journal of Machine Learning Research*, 24(297):1–59, 2023. URL https://jmlr.org/papers/v24/22-0969.html.

Andrew Wagenmaker and Dylan J. Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *Conference on Learning Theory (COLT)*, volume 195 of *PMLR*, pp. 5077–5133, 2023. URL https://proceedings.mlr.press/v195/wagenmaker23a/wagenmaker23a.pdf.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1113–1122, 2015.

Zirui Yan, Dennis Wei, Dmitriy A. Katz, Prasanna Sattigeri, and Ali Tajer. Causal bandits with general causal models and interventions. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pp. 4609–4617, 2024. URL https://proceedings.mlr.press/v238/yan24a.html.

Yinglun Zhu, Dylan J. Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning (ICML)*, pp. 27428–27453. PMLR, 2022.

Supplementary Materials for: ContrastSpanner: Learning Low-Rank Causal Contrasts to Alleviate Power-Set and Eluder Barriers

A Further Discussion of Related Work

Ref.	Algorithm / setting	Causal?	Policy / oracle model	Rank r	What is (and is not) resolved
This work	**ContrastSpanner / Fast-ContrastSpanner** (contextual bandit; combinatorial $\mathcal{A}$)	**Yes**	**General nonlinear realizable oracle for $\mu(x,a) = \mathbb{E}[r \mid x, \mathrm{do}(a)]$ (Sq or Log)**	**Learned online**	**Introduces *interventional contrast features* and grows a spanner/ε-net to *learn* an effective low-rank contrast span. Breaks the power-set barrier and replaces opaque Eluder($\mathcal{F}$) dependence by an explicit rank/design complexity $\mathfrak{C}(r)$ appearing in regret.**
Zhu et al. (2022)	SpannerIGW (large-action contextual bandit)	No	Sq oracle; *associational* contrasts	N/A	Removes explicit $\lvert\mathcal{A}\rvert$ dependence via spanners/design, but not causal and does not target interventional gaps.
Foster & Rakhlin (2020)	SquareCB (finite-action contextual bandit)	No	Sq oracle reduction	N/A	Minimax-style regret, but does not address combinatorial $\lvert\mathcal{A}\rvert$ and retains function-class complexity dependence.
Foster & Krishnamurthy (2021)	FastCB (finite-action contextual bandit)	No	Log-loss oracle reduction	N/A	Small-loss regret, but explicit $\lvert\mathcal{A}\rvert$ dependence and no spanner-based mitigation.
(Lattimore et al., 2016)	Causal bandits (known graph; restricted interventions)	Yes	Typically parametric SCM / restricted arms	Given	Avoids power-set scaling by restricting interventions to small set of mediators; does not treat large combinatorial $\mathcal{A}$ with rich function approximation.
Sen et al. (2017)	Causal contextual bandit with latent confounders	Yes	Low-rank matrix/NMF structure (parametric)	Given	Low-rank helps, but rank is model-side, not learned as a contrast span for general nonlinear μ.
Yan et al. (2024;?)	General SCM causal bandits	Yes	Nonlinear SCM classes; complexity via Eluder/covering	N/A	Handles nonlinear SCMs but regret scales with opaque complexity notions; no spanner-based rank learning.
Elahi et al. (2024)	Causal bandits with partial discovery	Yes	Discover-then-learn; graph-learning overhead	N/A	Addresses causal discovery; orthogonal to our focus on contrast geometry and combinatorial/Eluder barriers.
Kveton et al. (2015); Wen et al. (2015)	Combinatorial semi-bandits (noncausal)	No	Linear-in-items reward; subset actions	N/A	Breaks power-set scaling via *linearity* & semi-bandit feedback; incompatible with general nonlinear μ.
Jun et al. (2019); Lu et al. (2021); Lale et al. (2019)	Low-rank/latent subspace bandits (noncausal)	No	Parametric low-rank or linear-subspace structure	Often learned	Learn low-dimensional representations, but not interventional contrast features and not designed to remove Eluder dependence for nonlinear causal models.
Agarwal et al. (2020); Mhammedi et al. (2023)	Low-rank MDP exploration (noncausal)	No	Spanners/designs for representation learning	Learned	Inspires our *rank discovery* viewpoint: grow a basis only when new contrasts are "significant."

Table 1: Where ContrastSpanner sits relative to prior art. The throughline is *learning a low-rank interventional contrast geometry* to address **both** the *power-set* (combinatorial $|\mathcal{A}|$) and *Eluder/function-class* barriers.

Related Work [cf. Table 1]. SpannerIGW (Zhu et al., 2022) shows that *spanner/optimal-design* ideas can remove explicit $|\mathcal{A}|$ dependence in non-causal contextual bandits under a squared-loss reward predictor. On the causal side, known-graph causal bandits often focus on restricted intervention families (e.g., (Lattimore et al., 2016)). If one does not restrict focus to a small number of mediators, then linearity/specific structures may be exploited, as in combinatorial semi-bandits where ground-set/additive structure is deployed (Kveton et al., 2015; Wen et al., 2015). Low-rank causal bandits typically posit parametric factorizations or linear structural equation models (SEM) with rank built into the model class (Sen et al., 2017; Varici et al., 2023) also exhibit promise. However, these approaches assume a small known number of mediators, or that the relationships are monotone/additive. Beyond these settings, general-SCM analyses exhibit complexity scaling with respect to covering numbers or Eluder dimension (Yan et al., 2024), as originally identified in (Russo & Van Roy, 2013b). To mitigate this issue, we take inspiration from spanner and sparsification ideas (Batson et al., 2012), and by the philosophy of learning low-dimensional structure via exploration and optimal design in low-rank MDPs (Agarwal et al., 2020; Mhammedi et al., 2023), as well as compressive sensing (Donoho, 2006). In doing so, we construct *causal contrast features* that make spanner-based exploration applicable to *interventions*, rather than observational actions, through a novel parameterization of a bandit policy. In particular, these contrast features may be interpreted as conditional average treatment effect (CATE) estimators in the parameter space of a regression oracle. When operating in tandem with an ε-net spanner-growth mechanism that learns an effective rank,

we can *learn online* a small number of mediators that is sufficient for no-regret in an appropriate sense, thereby bypassing both the power-set and Eluder-dimension barriers. See Table 1.

Extension to observational confounding. If feedback is observational rather than interventional, $\mathbb{E}[r \mid x, a]$ may differ from $\mathbb{E}[r \mid x, \mathrm{do}(a)]$ due to confounding, and plug-in contrasts Δ_f can be biased even if f predicts associational outcomes well. In that case, one can retain the ContrastSpanner framework by replacing the reward-regression oracle with a *causal contrast oracle* that outputs orthogonalized/DR (Dudík et al., 2014) contrasts: for example, use DR-style scores (Dudík et al., 2014) that combine an outcome model $\hat{\mu}(x, a)$ with a propensity model $\hat{e}(a \mid x)$, or an R-learner-style residualization objective, to obtain a contrast estimator whose error is robust to first-order nuisance misspecification. Algorithmically, the spanner and design steps are unchanged; only the oracle (and its guarantee) is strengthened to ensure that the contrasts being spanned correspond to $\mathrm{do}(\cdot)$ semantics rather than mere conditioning.

Remark 3. ***(Differences in the goal of CATE estimation vs. causal bandits)*** The CATE literature studies statistical estimators $\hat{\tau}$ of (2) from i.i.d. data, typically under either randomized treatment assignment or observational data with overlap and unconfoundedness, using inverse propensity weighting (IPW) (Horvitz & Thompson, 1952), doubly robust (DR) estimators (Dudík et al., 2014), or meta-learning reductions (Wager & Athey, 2018) (e.g., R-learner (Nie & Wager, 2021), X-learner (Künzel et al., 2019)). This connects to the primitive under consideration, a regression oracle, that outputs a predictor $f \in \mathcal{F}$ intended to approximate $\mu(x, a)$ (in squared loss or log loss), and with induced contrast (3), in the sense that Δ_f may be defined via a plug-in estimator of the CATE (2). CATE and contrast features coincide in the following setting: (i) the causal semantics are fully captured by $\mu(x, a)$ (e.g., μ is identifiable from the data-generating process, or data are generated under $\mathrm{do}(a_t)$), and (ii) the oracle returns f satisfying an oracle inequality for predicting μ. Then $\Delta_f(x; a, a')$ is a valid CATE estimator with error controlled by the regression error of f. In particular, under uniform prediction control,

$$\sup_{x,a} |f(x, a) - \mu(x, a)| \le \epsilon \quad \Longrightarrow \quad \sup_{x,a,a'} |\Delta_f(x; a, a') - \tau(x; a, a')| \le 2\epsilon. \tag{18}$$

However, our use of contrasts differs from typical CATE estimation: contrasts are not the end goal, but rather define a *decision-relevant geometry* over (a, a') that drives spanner/ε-net growth over a combinatorial set of possible interventions, to enable efficient no-regret learning. To realize this goal, we focus on an *interventional* (bandit) protocol: data are generated under $\mathrm{do}(a_t)$, so the target is directly $\mu(x, a) = \mathbb{E}[r \mid x, \mathrm{do}(a)]$. In this regime, the plug-in contrast $\Delta_f(x; a, a') := f(x, a) - f(x, a')$ (3) is the natural estimator of τ induced by a reward predictor f. Moreover, it alleviates the need for uniform (global) accuracy of Δ_f over all $(a, a') \in \mathcal{A}^2$, but instead only on *decision-relevant* contrasts—those needed to distinguish near-optimal actions in the realized contexts. Under a low rankness (Assumption 2), this permits us to discern online via bandit feedback, which are required via a spanner/ε-net mechanism (Zhu et al., 2022). Extensions that permit observational confounding are discussed in Appendix A.

Relation to Zhu et al. (2022). Zhu et al. assume a realizable oracle class of the form $f_g(x, a) = \langle \phi(x, a), g(x) \rangle$, where $\phi(x, a) \in \mathbb{R}^d$ is a known action embedding and $g(x) \in \mathbb{R}^d$ is an unknown context-dependent weight. If the true mean reward satisfies $\mu(x, a) = \langle \phi(x, a), g^\star(x) \rangle$, then pairwise contrasts obey

$$\mu(x, a) - \mu(x, a') = \langle g^\star(x),\, \phi(x, a) - \phi(x, a') \rangle,$$

which is a special case of our contrast realizability assumption with $z_x(a, a') = \phi(x, a) - \phi(x, a')$ and $w_x = g^\star(x)$. Thus Zhu et al.'s model fits within our framework without the low-rank restriction. Our formulation differs in two respects: (i) we interpret $\mu(x, a)$ as an *interventional* mean $\mathbb{E}[r \mid x, \mathrm{do}(a)]$, and (ii) we assume that the contrast family $\{z_x(a, a')\}$ lies in an unknown r-dimensional subspace that is learned online via spanners. This low-rank contrast structure is not imposed in Zhu et al. (2022), and is central to our reduction of exponential action dependence.

B Summary of Notation

Symbol	Def. / Ref	Meaning	Origin
Problem setup and causal quantities			
T	Sec. 1	Time horizon (number of rounds).	Bandits (Lattimore & Szepesvári, 2020)
t	Sec. 1	Round index $t \in \{1, \dots, T\}$.	—
$\mathcal{X}$	Sec. 1	Context space.	This work
$\mathcal{A}$	Sec. 1	Intervention/action space (possibly $\lvert\mathcal{A}\rvert \approx 2^m$ for subset interventions on m mediators).	Causal bandits (Lattimore et al., 2016)
$\mathrm{do}(\cdot)$	Sec. 1	Causal intervention operator (Pearl's do-calculus).	(Pearl, 2009)
x_t, a_t	Sec. 1	Context and action at round t.	—
$Y_t \in [K]$	Sec. E.4	Categorical outcome for the log-loss model; alphabet $[K] = \{1, \dots, K\}$.	FastCB (Foster & Krishnamurthy, 2021)
r_t	Sec. 1, Ass. 4	Observed scalar reward (bandit feedback).	Bandits (Lattimore & Szepesvári, 2020)
$R_{\max}$	Ass. 4	Known bound with $\lvert r_t\rvert \le R_{\max}$ a.s.	Standard
$\mu(x, a)$	Sec. 1; Sec. E.4	Interventional mean reward $\mathbb{E}[r_t \mid X_t{=}x, \mathrm{do}(A_t{=}a)]$.	Causality (Pearl, 2009)
$\mathrm{Reg}(T)$	Sec. 1	Cumulative (bandit) regret $\sum_{t=1}^{T}(\max_a \mu(x_t, a) - \mu(x_t, a_t))$.	Bandits (Lattimore & Szepesvári, 2020)
$\tau(x; a, a')$	Eq. (2)	True interventional contrast / CATE between a and a'.	CATE lit. (Wager & Athey, 2018)
$\Delta_f(x; a, a')$	Eq. (3)	Plug-in contrast induced by predictor f (CATE estimator in the model class).	This work
$z_x(a, a')$	Ass. 1	Contrast feature map used to represent/approximate contrasts.	SpannerIGW (Zhu et al., 2022)
w_x	Ass. 1	Context-dependent linear weight for the contrast realizability model.	This work
$\mathcal{Z}_x$	Sec. 3.2	Set of all contrast features at context x: $\mathcal{Z}_x = \{z_x(a, a') : a, a' \in \mathcal{A}\}$.	This work
$\mathcal{F}$	Sec. 2, Sec. E.3	Function/predictor class used by the oracle (appears in DEC).	(Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021)
$\mathcal{U}_\star$	Ass. 2	Unknown r-dimensional subspace containing all contrasts $\mathcal{Z}_x$ (for all x).	Low-rank models (Candès & Recht, 2009)
r	Ass. 2	Intrinsic contrast rank ($r \ll \lvert\mathcal{A}\rvert^2$).	This work
$m(x, a)$	Sec. 3.1	(Optional) low-dimensional mediator representation of an intervention.	Causality (Pearl, 2009)
Algorithmic objects: spanner, pool, residuals			
$\mathcal{S}_t$	Def. 5; Alg. 1	Spanner state: stored contrast directions $u_s = z_{x_s}(a_s, a'_s)$.	This work; sparsification (Batson et al., 2012)
$\mathcal{A}_t$	Sec. F.3.1	Finite action pool induced by the spanner at round t.	This work
K_t	Sec. F.3.1	Pool size $K_t := \lvert\mathcal{A}_t\rvert$.	—
$\widehat{\mathcal{U}}_t$	Def. 5	Learned contrast span $\widehat{\mathcal{U}}_t = \mathrm{span}\{u_s : s \in \mathcal{S}_t\}$.	This work
Π_U	Def. 5	Orthogonal projector onto subspace $U \subseteq \mathbb{R}^d$.	Linear algebra
λ	Def. 5	Regularization for design/Gram matrices.	Bandits (Abbasi-Yadkori et al., 2011)
V_t	Eq. (122)	Regularized design matrix $V_t = \lambda I + \sum_{s \in \mathcal{S}_t} u_s u_s^\top$.	(Abbasi-Yadkori et al., 2011)
$\lVert v\rVert_{V_t^{-1}}$	Def. 5	Mahalanobis (self-normalized) norm $\sqrt{v^\top V_t^{-1} v}$.	(Abbasi-Yadkori et al., 2011)
$\mathrm{Res}_t(z)$	Def. 6	Residual score: $\mathrm{Res}_t(z) = \lVert z\rVert_{V_{t-1}^{-1}}$.	This work
ε_t	Def. 6	Residual threshold controlling net accuracy vs. spanner size.	This work
$\varepsilon_{\mathrm{span}}(t)$	Lem. 14	Span gap (loss from restricting to pool): $\max_{a \in \mathcal{A}} \mu(x_t, a) - \max_{a \in \mathcal{A}_t} \mu(x_t, a)$.	This work
Oracles, gaps, and IGW sampling			
AlgSq, AlgLog	Sec. 3.1	Square-loss and log-loss regression oracles.	(Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021)
$\mathrm{Reg}_{\mathrm{Sq}}(T)$	Eq. (4)	Square-loss oracle regret (cumulative excess square loss).	(Foster & Rakhlin, 2020)
$\mathsf{R}_{\mathrm{sq}}(T)$	Eq. (4)	Deterministic upper bound on $\mathrm{Reg}_{\mathrm{Sq}}(T)$ assumed in analysis.	(Foster & Rakhlin, 2020)
$\mathrm{Reg}_{\mathrm{Log}}(T)$	Eq. (7)	Log-loss oracle regret (cumulative excess log loss).	(Foster & Krishnamurthy, 2021)
$\mathsf{R}_{\log}(T)$	Ass. 3	Deterministic upper bound on $\mathrm{Reg}_{\mathrm{Log}}(T)$ assumed in analysis.	(Foster & Krishnamurthy, 2021)
$\hat{a}_t$	Sec. 3.3	Predicted best action, e.g. $\hat{a}_t \in \arg\max_{a \in \mathcal{A}_t} f_{t-1}(x_t, a)$.	(Foster & Rakhlin, 2020)
$\widehat{\Delta}_t(\cdot)$	Eqs. (10),(11)	Plug-in one-sided gaps used by IGW (square/log).	(Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021)
ϑ	Eq. (12)	IGW regularizer ensuring π_t is a distribution (typically $\vartheta \ge K_t$).	(Foster & Rakhlin, 2020)
γ_t	Eq. (12)	IGW gap parameter (tuned for regret).	(Foster & Rakhlin, 2020)
$\pi_t(\cdot \mid x_t)$	Eq. (12)	Inverse-gap-weighting (IGW) sampling distribution over $\mathcal{A}_t$.	IGW (Abe & Long, 1999; Foster & Rakhlin, 2020)
Optimal design, DEC, and minimax analysis			
$\rho(\mathcal{S})$	Sec. 3.3	Probability simplex over a finite set $\mathcal{S}$.	—
δ_a	Sec. 3.3	Dirac mass at action a (point mass distribution).	—
$u_t(a)$	Sec. E.3	Contrast direction used in minimax analysis, e.g. $u_t(a) = z_{x_t}(\hat{a}_t, a)$.	This work
$M_t(q)$	Sec. E.3	Design (information) matrix induced by q and $\{u_t(a)\}_{a \in \mathcal{A}_t}$.	OED (Kiefer, 1974)
p_t, q_t	Eqs. (13)	Design distributions for OED / minimax sampling (G-opt and A-opt).	OED (Kiefer, 1974; Atkinson et al., 2007)
$\kappa_t(q)$	Sec. E.3; Eq. (62)	G-opt width $\max_a u_t(a)^\top M_t(q)^{-1} u_t(a)$ controlling worst-case variance.	(Kiefer, 1974; Lattimore & Szepesvári, 2020)
C_{opt}	Ass. 6	Approximation factor of the design solver.	This work
$\mathrm{dec}_{\bar{\gamma}}$	Eq. (66)	Decision-estimation coefficient used to certify minimax rates.	(Foster & Rakhlin, 2020; Zhu et al., 2022)
$\bar{\gamma}$	Sec. E.3; Sec. E.4	DEC / design-weighted IGW tuning parameter (minimax).	This work
Concentration and log-loss geometry			
$\mathbb{P}, \mathbb{E}$	Sec. E.1	Probability and expectation operators.	Standard
$\mathcal{F}_t$	Sec. E.1	Filtration generated by past observations (history).	—
$\mathbb{E}_t[\cdot]$	Sec. E.1	Conditional expectation given $\mathcal{F}_{t-1}$.	—
δ	Throughout Sec. E, E.4	Confidence level for high-probability events ($1 - \delta$).	Standard
ξ_t	Ass. 4	Martingale noise $\xi_t = r_t - \mu(x_t, a_t)$.	—
σ^2	Ass. 4	Conditional second-moment bound $\mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}] \le \sigma^2$.	—
$\mathrm{conv}(\cdot)$	Sec. 3.3	Convex hull operator (e.g. $\mathrm{conv}(\mathcal{S}_t)$).	Convex analysis
$\mathrm{tr}(\cdot)$	Sec. E.3	Matrix trace operator.	Linear algebra
$q^\star(\cdot \mid x, a)$	Sec. E.4	True categorical outcome distribution under $\mathrm{do}(a)$.	(Foster & Krishnamurthy, 2021)
$q_t(\cdot \mid x, a)$	Sec. 3.1, Sec. E.4	Oracle-predicted categorical distribution.	(Foster & Krishnamurthy, 2021)
$\ell_{\log}(q; y)$	Sec. E.4	Log loss $\ell_{\log}(q; y) = -\log q(y)$.	(Foster & Krishnamurthy, 2021)
$\mathrm{KL}(p \Vert q)$	Sec. E.4	Kullback–Leibler divergence.	Information theory
$\mathrm{TD}(p, q)$	Sec. E.4	Triangular discrimination (curvature proxy for log loss).	(Topsoe, 2002; Le Cam, 2012)
ν	Ass. 7	Clipping level ensuring bounded log-loss increments and stable curvature.	(Foster & Krishnamurthy, 2021)
κ	Ass. 7	Curvature constant relating KL/TD to squared error.	(Amari & Nagaoka, 2000; Foster & Krishnamurthy, 2021)
$w_t(a)$	Sec. E.4	Fisher weight for log-loss in OED.	(Amari & Nagaoka, 2000)

C Implementation Details for Algorithms 1 - 2

Notation and conventions. We write $\rho(\mathcal{A})$ for the probability simplex over a finite set $\mathcal{A}$. When we use a point-mass distribution, we write δ_a for the Dirac distribution concentrated at action a, i.e. $\delta_a(a) = 1$ and $\delta_a(a') = 0$ for $a' \neq a$. The learner maintains a bandit feedback cache $\mathcal{D}_t = \{(x_s, a_s, r_s)\}_{s \leq t}$. All contrasts are indexed by unordered pairs (a, a') (no anchoring is assumed).

Plug-in contrasts and estimated contrast features. Given any reward predictor $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, define the plug-in contrast

$$\Delta_f(x; a, a') := f(x, a) - f(x, a') \qquad \text{(cf. (3))}. \tag{19}$$

To drive spanner growth we require an *estimated contrast-feature vector* $\widehat{z}_t(a, a') \in \mathbb{R}^d$ whose geometry controls the uncertainty of Δ_f. In the linear case, $\widehat{z}_t(a, a')$ reduces to a feature difference. For general nonlinear predictors, $\widehat{z}_t(a, a')$ is an abstract embedding (e.g. local linearization, score/logit differences, or mediator differences) as formalized in the problem setup (cf. (3)).

ε-net/spanner growth via a residual test. Let $\mathcal{S}_t$ denote the *set of stored contrast-feature vectors* accumulated so far: $\mathcal{S}_t = \{\widehat{z}_{\kappa(j)}(a_{\kappa(j)}, a'_{\kappa(j)})\}_{j=1}^{m_t}$, where $m_t = |\mathcal{S}_t| \ll t$ is the current spanner size and $\kappa(j)$ indexes the round at which the j-th stored vector was added. Let $U_t := \mathrm{span}(\mathcal{S}_t)$ and define the (Euclidean) residual

$$\mathrm{Res}_t(x_t; a, a') := \left\| \widehat{z}_t(a, a') - \Pi_{U_t} \widehat{z}_t(a, a') \right\|_2. \tag{20}$$

A new contrast direction is added if $\mathrm{Res}_t(x_t; a, a') > \varepsilon_t$ for some candidate pair (a, a'). (One may replace (20) by an elliptic-potential test $\widehat{z}_t(a, a')^\top V_t^{-1} \widehat{z}_t(a, a') > \varepsilon_t^2$ if desired; see Prop. 3 in the theory section.)

C.0.1 Anchoring Actions and Other Practical Matters

Lemma 1 (Anchor representation of pairwise contrasts)**.** *Let $\phi_x : \mathcal{A} \to \mathbb{R}^d$ be such that for all $a, a' \in \mathcal{A}$,*

$$z_x(a, a') = \phi_x(a) - \phi_x(a').$$

Fix any reference action $a_0 \in \mathcal{A}$. Then

$$\begin{aligned} z_x(a, a') &= z_x(a, a_0) - z_x(a', a_0), \qquad \textit{and} \\ \mathrm{span}\{z_x(a, a') : a, a' \in \mathcal{A}\} &= \mathrm{span}\{z_x(a, a_0) : a \in \mathcal{A}\}. \end{aligned} \tag{21}$$

r

Anchored contrasts and spanner construction. Throughout we assume contrasts are differences of a single action embedding, i.e. $z_x(a, a') = \phi_x(a) - \phi_x(a')$ (this covers both mean contrasts $\Delta_f(x; a, a') = f(x, a) - f(x, a')$ and centered-logit contrasts). Fix any reference action $a_0 \in \mathcal{A}$. By simple algebra every pairwise contrast decomposes as $z_x(a, a') = z_x(a, a_0) - z_x(a', a_0)$, hence $\mathrm{span}\{z_x(a, a') : a, a' \in \mathcal{A}\} = \mathrm{span}\{z_x(a, a_0) : a \in \mathcal{A}\}$. Algorithmically we therefore maintain a spanner/basis of the anchored vectors $\{z_x(a, a_0)\}_{a \in \mathcal{A}}$, which reduces candidate comparisons from $|\mathcal{A}|^2$ to $|\mathcal{A}|$ and simplifies the residual test used to grow the spanner. (See Lemma 1 for a formal statement.)

IGW (inverse-gap weighting) policy. Fix a finite *candidate action pool* $\mathcal{A}_t \subseteq \mathcal{A}$ (how this pool is formed is specified below). Given predictor f_t, define the predicted best action in the pool

$$\hat{a}_t \in \arg\max_{a \in \mathcal{A}_t} f_t(x_t, a), \tag{22}$$

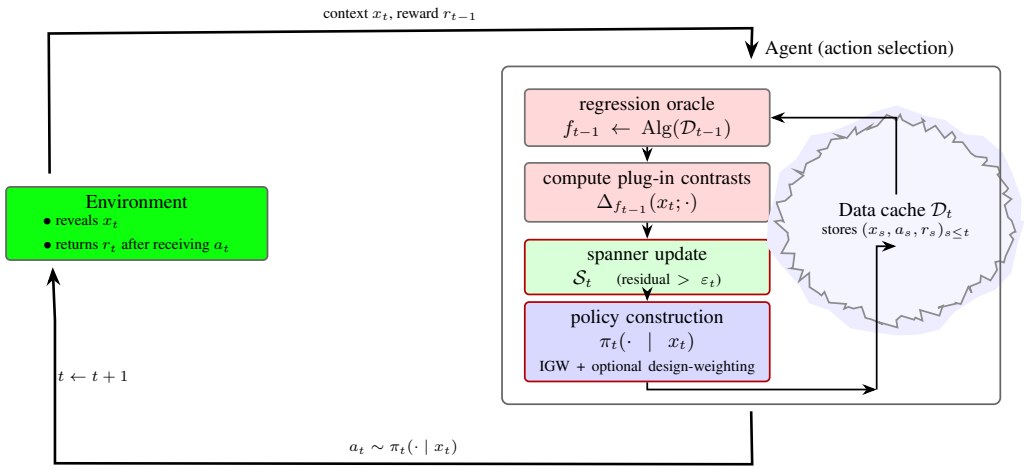


Figure 5: Visualization of how contrast spanners interact with bandit feedback loop consisting of context, action, and reward. The agent's policy is constructed through a combination of (i) oracle $\rightarrow$, (ii) contrasts $\rightarrow$, (iii) spanner $\rightarrow$, and (iv) IGW/design-weighting. The cache $\mathcal{D}_t$ is a compact cloud storing bandit feedback triples.

and the (one-sided) estimated gaps

$$\widehat{\Delta}_t(a) := f_t(x_t, \hat{a}_t) - f_t(x_t, a) \geq 0, \qquad a \in \mathcal{A}_t. \tag{23}$$

Given a stepsize $\gamma_t > 0$ and regularizer $\vartheta > 0$, define the unnormalized IGW weights

$$w_t(a) := \frac{1}{\vartheta + \gamma_t \, \widehat{\Delta}_t(a)}, \qquad a \in \mathcal{A}_t, \tag{24}$$

and the corresponding policy $\pi_t \in \rho(\mathcal{A}_t)$ by normalization:

$$\pi_t(a \mid x_t) := \frac{w_t(a)}{\sum_{b \in \mathcal{A}_t} w_t(b)}. \tag{25}$$

Optional design-solve reweighting (OED). One may optionally compute a design distribution $q_t \in \rho(\mathcal{A}_t)$ by solving an approximate optimal-design problem over the *current stored directions* (or a proxy set of directions), e.g. G-optimal design:

$$p_t \in \arg \min_{q \in \rho(\mathcal{A}_t)} \max_{(a,a') \in \mathcal{C}_t} \widehat{z}_t(a,a')^\top \Big(\lambda I + \sum_{u \in \mathcal{A}_t} q(u) \, \widehat{z}_t(u, \hat{a}_t) \widehat{z}_t(u, \hat{a}_t)^\top \Big)^{-1} \widehat{z}_t(a,a'), \tag{26}$$

where $\mathcal{C}_t$ is a finite contrast candidate set. For *stochastic regret*, this step can be viewed as optional variance control under heteroskedasticity; for *minimax-style* analyses it is often essential to obtain uniform control (cf. standard design theory).

A simple way to combine IGW with design is multiplicative reweighting:

$$\tilde{\pi}_t(a \mid x_t) \propto \pi_t(a \mid x_t) \, q_t(a), \qquad a \in \mathcal{A}_t, \tag{27}$$

followed by normalization. (If design is omitted, take q_t uniform on $\mathcal{A}_t$.)

Why the "Dirac mix" sometimes appears (and when it can be dropped). Some FastCB-style reductions add a small point-mass on $\hat{a}_t$ to prevent degenerate allocations when (i) the action pool $\mathcal{A}_t$ is a restricted proxy for the full action space and/or (ii) gap estimates are noisy early on. Formally, for $\alpha_t \in [0,1]$ one defines

$$\tilde{\pi}_t := (1 - \alpha_t)\pi_t + \alpha_t \delta_{\hat{a}_t}. \tag{28}$$

Algorithm 1 CONTRASTSPANNER (restated)

Require: Regularizer $\lambda > 0$, residual thresholds (ε_t), smoothing (γ_t).
Initialize $\mathcal{S}_1 \leftarrow \emptyset$, $V_1 \leftarrow \lambda I$, data cache $\mathcal{D}_0 \leftarrow \emptyset$.
for $t = 1, 2, \ldots, T$ **do**
  Observe context x_t.
  Oracle update: $f_t \leftarrow \mathrm{AlgSq}(\mathcal{D}_{t-1})$.
  Form candidate set of contrasts at x_t (e.g. via a candidate pool) and find any pair (a, a') with $\mathrm{Res}_t(a, a') > \varepsilon_t$, where Res_t is defined in (9).
  if such (a, a') exists **then**
    Add $z_{x_t}(a, a')$ to $\mathcal{S}_t$ and update $V_t \leftarrow V_t + z_{x_t}(a, a') z_{x_t}(a, a')^\top$.
  end if
  Predicted best action $a_t^\star \in \arg\max_{a \in \mathcal{A}} f_t(x_t, a)$.
  Define IGW distribution $\pi_t(\cdot \mid x_t)$ on a finite action pool $\mathcal{A}_t$ using (10) and (12).
  (Optionally: re-weight π_t by p_t [cf. (13)])
  Sample $a_t \sim \pi_t(\cdot \mid x_t)$ and apply $\mathrm{do}(a_t)$.
  Observe reward r_t and append (x_t, a_t, r_t) to $\mathcal{D}_t$.
end for

In the *square-loss* SquareCB-style analysis this extra mixing is typically unnecessary. In the *log-loss* / first-order setting it can be a practical stabilization device, but if one chooses ϑ in (24) sufficiently large (or enforces minimum probability floors), then (28) is often dispensable in practice.

C.1 Data Structure Considerations for Algorithms 1–2

This appendix provides a detailed, implementation-focused description of Algorithms 1 and 2. We clarify data structures, numerical updates, statistical tests, and parameter roles.

1. At time t, the agent maintains:

   (i) Data cache

   $$\mathcal{D}_t = \{(x_s, a_s, r_s)\}_{s \le t}.$$

   This cache is passed to the oracle update routine AlgSq or AlgLog.

   (ii) Spanner set

   $$\mathcal{S}_t = \{z^{(1)}, \ldots, z^{(m_t)}\} \subseteq \mathbb{R}^d, \qquad m_t := |\mathcal{S}_t|.$$

   Each $z^{(j)}$ is a previously selected contrast vector $z_{x_{i_j}}(a_{i_j}, a'_{i_j})$. Under Assumption 2 and the residual test (9), $m_t \ll t$.

   (iii) Gram matrix

   $$V_t = \lambda I + \sum_{j=1}^{m_t} z^{(j)} (z^{(j)})^\top, \qquad \lambda > 0.$$

   When a new contrast vector z is added to $\mathcal{S}_t$, the update is

   $$V_{t+1} \leftarrow V_t + z z^\top.$$

   For efficiency, V_t^{-1} may be maintained via a Sherman–Morrison update.

2. **Residual Test and Spanner Growth.** Given a candidate contrast pair (a, a') at context x_t, compute

   $$z := z_{x_t}(a, a'), \qquad \mathrm{Res}_t(a, a') := \|z - \Pi_{\mathrm{span}(\mathcal{S}_t)} z\|_2.$$

   If $\mathrm{Res}_t(a, a') > \varepsilon_t$, add z to $\mathcal{S}_t$ and update V_t. This procedure adaptively learns the effective contrast rank.

Algorithm 2 FASTCONTRASTSPANNER (restated)

Require: Regularizer $\lambda > 0$, residual thresholds (ε_t), smoothing (γ_t).
Initialize $\mathcal{S}_1 \leftarrow \emptyset$, $V_1 \leftarrow \lambda I$, data cache $\mathcal{D}_0 \leftarrow \emptyset$.
for $t = 1, 2, \ldots, T$ **do**
  Observe context x_t.
  Oracle update: $q_t \leftarrow \mathrm{AlgLog}(\mathcal{D}_{t-1})$.
  Induce mean predictor $f_t(x, a) = \sum_k [r(x, a) = k] q_{f_t}(k \mid x, a)$ from q_t
  Find any contrast pair (a, a') with $\mathrm{Res}_t(a, a') > \varepsilon_t$
  if such (a, a') exists **then**
    add $z_{x_t}(a, a')$ to $\mathcal{S}_t$ and update V_t as in Algorithm 1.
  end if
  Predicted best action $a_t^\star \in \arg\max_{a \in \mathcal{A}} f_t(x_t, a)$.
  Define IGW distribution $\pi_t(\cdot \mid x_t)$ using (11) and (12).
  (Optionally: reweight π_t by Fisher-weighted design p_t analogue of (110).)
  Sample $a_t \sim \pi_t(\cdot \mid x_t)$ and apply $\mathrm{do}(a_t)$.
  Observe outcome/reward $r_t \in \{1, \ldots, K\}$; append (x_t, a_t, r_t) to $\mathcal{D}_t$.
end for

3. **Gap Estimation and IGW Policy.** Let f_t denote the oracle predictor at round t, and the predicted best action as

$$a_t^\star \in \arg\max_{a \in \mathcal{A}_t} f_t(x_t, a),$$

where $\mathcal{A}_t$ is a finite candidate pool. The

Estimated gaps

$$\widehat{\Delta}_t^{\mathrm{sq}}(a) = f_t(x_t, a_t^\star) - f_t(x_t, a).$$
$$\widehat{\Delta}_t(a)^{\log} := \sum_k [r(x, a) = k] \left(q_{f_t}(k \mid x, a) - q_{f_t}(k \mid x, a^\star)\right)$$

IGW distribution

$$\pi_t(a \mid x_t) = \frac{1}{\vartheta + \gamma_t \widehat{\Delta}_t(a)},$$

Optional Design Reweighting When enabled, solve a convex optimal-design problem over $\mathcal{S}_t$ (e.g. A-opt or G-opt as in (13)) to obtain $p_t \in \rho(\mathcal{A}_t)$. Then combine multiplicatively:

$$\tilde{\pi}_t(a \mid x_t) \propto \pi_t(a \mid x_t)\, p_t(a),$$

followed by normalization.

C.2 Parameter Summary

Symbol	Meaning
λ	Gram matrix regularizer
ε_t	Residual threshold for spanner growth
ϑ	IGW smoothing constant
γ_t	IGW gap scaling parameter
$\mathcal{S}_t$	Spanner set of stored contrast vectors
V_t	Regularized Gram matrix over $\mathcal{S}_t$
$\mathcal{A}_t$	Candidate action pool at time t
π_t	Sampling policy at time t
AlgSq	Square-loss regression oracle
AlgLog	Log-loss oracle

Table 2: Algorithm parameters and maintained quantities.

C.3 Experiments

C.3.1 Implementation Details

Protocol and plotting. Unless stated otherwise, we run each algorithm for a fixed horizon T and report mean $\pm$ one standard error over independent trials. We report (i) cumulative regret $\sum_{t=1}^{T}\big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\big)$, (ii) BAI rate $\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{a_t = a_t^\star\}$, and (iii) coverage $\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbf{1}\{N_a(T) > 0\}$.

Therapeutic Window with Toxicity (TWT). Each action $a \in \mathcal{A}$ corresponds to a dose level $d(a) \in [0.15, 1.25]$. We partition the context into efficacy, baseline, and toxicity blocks $x = (x_g, x_b, x_t)$. The mean reward is a saturating therapeutic-window function

$$\beta(x) = 0.85 + 0.30\tanh(w_b^\top x_b), \qquad d_0(x) = 0.70 + 0.25\tanh(w_g^\top x_g), \tag{29}$$

$$\kappa(x) = 0.60 + 0.30\big(1 + \tanh(w_t^\top x_t)\big), \tag{30}$$

$$\mu_{\mathrm{twt}}(x, a) = r_0 + s \cdot \tanh\Big(\beta(x)d(a) - \kappa(x)\big(d(a) - d_0(x)\big)^2\Big), \tag{31}$$

with fixed scalars r_0, s and fixed weight vectors w_g, w_b, w_t. In the main text we plot $T = 192$ over 10 trials.

Quad-context misspecification. This scenario is designed to stress instance dependence and rank turnover. Each action a has an attribute vector $t(a) \in \mathbb{R}^m$ (randomly sampled and then held fixed). The mean reward is a low-rank bilinear trend in $(x, t(a))$ perturbed by a localized quadratic "hotspot" that can overturn rankings; Figure 3(b) visualizes the resulting action manifold and rank turnover. We plot $T = 2600$ over 5 trials (Figure 7).

C.3.2 Empirical mediator geometry cartoons

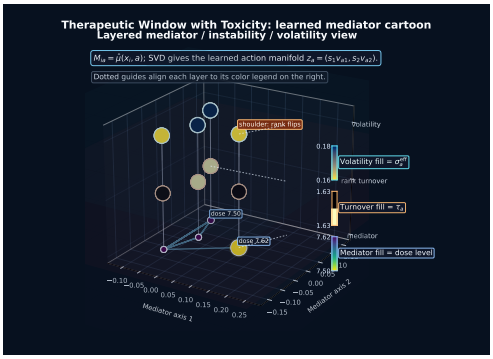
Figure 3 in the main text uses the known generative model to visualize decision-relevant geometry. Figure 6 instead computes the same layered visualizations from *learned* proxy quantities produced by ContrastSpanner, illustrating that the learned pool can recover an approximately low-dimensional mediator geometry.

C.3.3 Additional Experiments

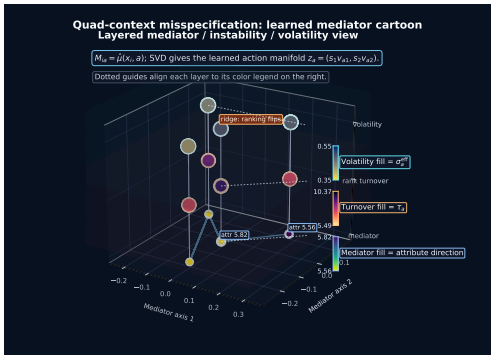
Figure 7 reports time traces for quad-context misspecification.

C.3.4 Full tables of results

Tables 3–6 summarize TWT performance across multiple action-set sizes.

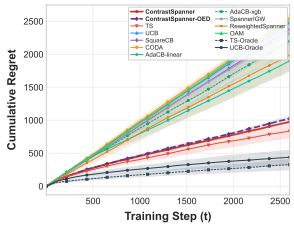


(a) TWT: learned proxy geometry.

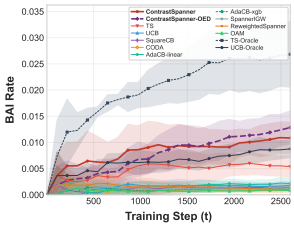


(b) Quad-context: learned proxy geometry.

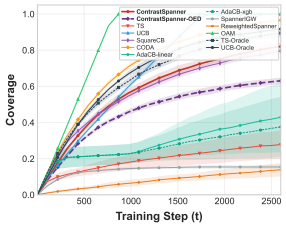
Figure 6: **Empirical mediator-geometry diagnostics (learned).** Each panel computes a low-dimensional action embedding (via an SVD of the learned proxy matrix) and overlays layers such as a deconfounded reward estimate, a rank-turnover proxy, and a volatility proxy.



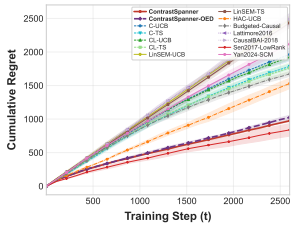
(a) Cumulative regret (contextual comparators).



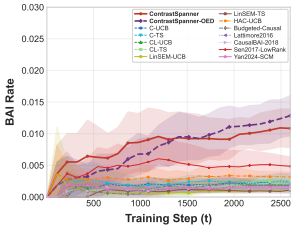
(b) BAI rate (contextual comparators).



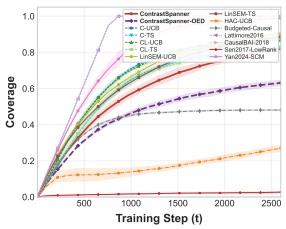
(c) Coverage (contextual comparators).



(d) Cumulative regret (causal comparators).



(e) BAI rate (causal comparators).



(f) Coverage (causal comparators).

Figure 7: **Quad-context misspecification: time traces.** Top row: contextual comparators on $K = 800$ actions. Bottom row: causal comparators on $K = 25$ actions.

D Formalization of the Technical Setting

Assumption 1 (Contrast Realizability)**.** *For each context x, there exists a representation $z_x : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^d$ and a vector $w_x \in \mathbb{R}^d$ such that*

$$\mu(x, a) - \mu(x, a') = \langle w_x, z_x(a, a') \rangle \quad \textit{for all } a, a' \in \mathcal{A},$$

Assumption 2 (Low-rankness)**.** *For each context x, the collection of representations $\{z_x(a, a')\}_{a,a'}$ lies in a linear subspace $\mathcal{U}_\star \subseteq \mathbb{R}^d$ of dimension at most $r \ll |\mathcal{A}|^2$. In addition, assume the contrast features are uniformly bounded: there exists $L_z > 0$ such that $\|z_x(a, a')\|_2 \leq L_z$ for all x and $a, a' \in \mathcal{A}$.*

Table 3: Therapeutic Window with Toxicity, grouped across comparator families at $K = 25$, $T = 192$, context dimension $p = 8$, action-attribute dimension $m = 4$, and 10 Monte Carlo trials. The additional column identifies whether each algorithm is a contextual or causal comparator. We report Cumulative Regret (lower is better), BAI Rate (higher is better), and Coverage (higher is better). Bold marks the best algorithm for each metric at this K.

Comparator Family	Algorithm	Cumulative Regret	BAI Rate	Coverage	Wall-clock (ms)
Causal	Lattimore2016	$\mathbf{13.419} \pm 0.517$	0.116	**1.000**	5.7 ± 0.1
Causal	ContrastSpanner	23.175 ± 1.012	0.317	0.944	105.8 ± 0.3
Causal	ContrastSpanner+Design	27.588 ± 0.895	**0.335**	0.472	339.1 ± 1.8
Causal	ConfoundedBudgetedCB	32.208 ± 0.440	0.007	0.080	63.5 ± 0.3
Causal	LinSEM-TS	35.869 ± 0.923	0.187	0.488	80.3 ± 0.3
Causal	Yan2024-SCM	36.866 ± 3.837	0.302	0.808	1055.8 ± 61.8
Causal	CausalBAI-2018	37.243 ± 0.667	0.047	**1.000**	5.9 ± 0.1
Causal	C-TS	38.824 ± 1.573	0.232	0.456	80.8 ± 0.3
Causal	HAC-UCB	39.438 ± 1.787	0.056	0.996	73.0 ± 0.2
Causal	LinSEM-UCB	41.642 ± 1.330	0.109	0.244	60.9 ± 0.2
Causal	C-UCB	42.968 ± 1.100	0.220	0.320	61.4 ± 0.1
Causal	Sen2017-LowRank	47.455 ± 2.194	0.062	0.852	57.5 ± 1.2
Causal	CL-TS	49.277 ± 2.030	0.191	0.456	82.7 ± 0.3
Causal	CL-UCB	52.108 ± 2.599	0.208	0.308	62.2 ± 0.3
Contextual	TS	21.871 ± 2.407	0.063	0.336	175.2 ± 1.0
Contextual	ContrastSpanner	23.175 ± 1.012	0.317	0.944	105.8 ± 0.3
Contextual	ContrastSpanner+Design	27.588 ± 0.895	**0.335**	0.472	339.1 ± 1.8
Contextual	AdaCB-linear	28.144 ± 1.593	0.076	0.924	276.3 ± 11.0
Contextual	AdaCB-xgb	29.016 ± 1.346	0.055	0.928	12364.9 ± 312.7
Contextual	ReweightedSpanner	35.090 ± 7.947	0.033	0.256	2078.0 ± 3.2
Contextual	SquareCB	35.575 ± 1.004	0.056	0.992	16988.1 ± 694.8
Contextual	TS-Oracle	38.807 ± 1.237	0.224	0.472	60.7 ± 0.1
Contextual	SpannerIGW	42.926 ± 2.889	0.049	0.988	343.3 ± 0.8
Contextual	UCB	50.941 ± 0.955	0.040	**1.000**	33.6 ± 0.2
Contextual	CODA	52.751 ± 1.390	0.039	**1.000**	244.1 ± 0.5
Contextual	UCB-Oracle	62.995 ± 0.883	0.205	0.244	64.4 ± 0.2
Contextual	OAM	71.800 ± 9.520	0.031	**1.000**	25.0 ± 0.1

This assumption does not require $\mu(x, a)$ to be linear in a; it only posits that pairwise interventional contrasts admit a low-dimensional linear representation.

E Proof of Theorems 1 - 2

This section gives a didactic, step-by-step convergence analysis for the two oracle models. We first treat the square-loss oracle (stochastic and minimax regret), then the log-loss oracle. To proceed with the high-probability regret analysis for CONTRASTSPANNER under a square-loss regression oracle, following the reduction logic of SQUARECB Foster & Rakhlin (2020), (especially its inverse-gap weighting step), there are some adaptations required in the proof. Specifically, we sample from a *data-dependent* finite action pool $\mathcal{A}_t$ induced by the learned contrast spanner; this is where the dependence on the ambient number of actions is replaced by a dependence on the learned effective rank. We provide an independent analysis on when the effective rank permits sublinear regret, which allows us to establish how sharp the rates are achieved by the proposed approaches.

E.1 Preliminaries, standing notation, and assumptions

Recall the definition of the interventional mean reward is $\mu(x, a)$ and the (plug-in) contrast is $\Delta_f(x; a, a')$ as defined in (3), respectively, from Sections 1–3. The learner maintains the bandit-feedback cache $\mathcal{D}_t = \{(x_s, a_s, r_s)\}_{s \le t}$. We consider the following assumptions.

Table 4: Therapeutic Window with Toxicity, grouped across comparator families at $K = 50$, $T = 192$, context dimension $p = 8$, action-attribute dimension $m = 4$, and 10 Monte Carlo trials. The additional column identifies whether each algorithm is a contextual or causal comparator. We report Cumulative Regret (lower is better), BAI Rate (higher is better), and Coverage (higher is better). Bold marks the best algorithm for each metric at this K.

Comparator Family	Algorithm	Cumulative Regret	BAI Rate	Coverage	Wall-clock (ms)
Causal	Lattimore2016	$\mathbf{19.706} \pm 0.553$	0.044	**1.000**	6.4 ± 0.0
Causal	ContrastSpanner	23.236 ± 1.479	0.291	0.768	156.8 ± 1.0
Causal	ContrastSpanner+Design	27.723 ± 1.422	**0.305**	0.210	438.8 ± 2.2
Causal	Yan2024-SCM	28.559 ± 4.384	0.285	0.414	1029.9 ± 41.0
Causal	ConfoundedBudgetedCB	31.813 ± 0.313	0.005	0.040	113.5 ± 0.4
Causal	LinSEM-TS	33.709 ± 0.969	0.172	0.240	132.6 ± 0.4
Causal	C-TS	39.647 ± 1.613	0.212	0.228	131.2 ± 0.5
Causal	CausalBAI-2018	42.199 ± 0.445	0.027	**1.000**	5.9 ± 0.0
Causal	LinSEM-UCB	42.554 ± 1.330	0.107	0.110	108.6 ± 0.3
Causal	HAC-UCB	43.353 ± 1.019	0.047	0.726	131.0 ± 0.4
Causal	C-UCB	44.999 ± 0.450	0.231	0.156	108.8 ± 0.4
Causal	CL-TS	48.310 ± 1.894	0.215	0.228	134.4 ± 0.7
Causal	CL-UCB	50.098 ± 2.164	0.228	0.162	111.4 ± 0.4
Causal	Sen2017-LowRank	56.000 ± 3.138	0.059	0.366	6.2 ± 0.0
Contextual	ContrastSpanner	23.236 ± 1.479	0.291	0.768	156.8 ± 1.0
Contextual	TS	25.495 ± 3.285	0.030	0.182	318.0 ± 1.0
Contextual	ContrastSpanner+Design	27.723 ± 1.422	**0.305**	0.210	438.8 ± 2.2
Contextual	AdaCB-linear	29.417 ± 0.918	0.031	0.812	394.5 ± 8.3
Contextual	ReweightedSpanner	30.491 ± 8.974	0.067	0.116	6162.4 ± 12.8
Contextual	AdaCB-xgb	32.383 ± 1.461	0.018	0.812	15142.8 ± 282.0
Contextual	SquareCB	35.904 ± 1.399	0.024	0.886	21889.6 ± 357.2
Contextual	TS-Oracle	38.300 ± 1.577	0.214	0.240	89.2 ± 0.2
Contextual	SpannerIGW	41.389 ± 2.864	0.026	0.862	531.3 ± 1.7
Contextual	CODA	48.767 ± 1.370	0.018	0.988	455.6 ± 1.1
Contextual	UCB	50.951 ± 0.858	0.018	**1.000**	57.6 ± 0.2
Contextual	OAM	51.150 ± 0.880	0.024	**1.000**	37.9 ± 0.3
Contextual	UCB-Oracle	59.498 ± 1.468	0.202	0.116	116.7 ± 0.3

Assumption 3 (Oracle guarantees (square and log))**.** *Fix any sequence of adaptively chosen examples* (x_t, a_t, r_t).

1. ***Square oracle.*** *The subroutine* AlgSq *returns predictors* $f_t \in \mathcal{F}$ *whose square-loss oracle regret* $\mathrm{Reg}_{Sq}(T)$ *(defined in Section 3) satisfies* $\mathrm{Reg}_{Sq}(T) \le \mathsf{R}_{\mathrm{sq}}(T)$ *almost surely for a known nondecreasing bound* $\mathsf{R}_{\mathrm{sq}}(\cdot)$.
2. ***Log oracle.*** *For categorical outcomes (as in Algorithm 2), the subroutine* AlgLog *returns predictive distributions* $q_t(\cdot \mid x, a)$ *whose log-loss oracle regret* $\mathrm{Reg}_{Log}(T)$ *(defined in Section 3) satisfies* $\mathrm{Reg}_{Log}(T) \le \mathsf{R}_{\log}(T)$ *almost surely for a known nondecreasing bound* $\mathsf{R}_{\log}(\cdot)$.

While this assumption may appear arcane, it automatically holds across choice of linear, kernel, and neural network estimators for the average reward, as established in Foster & Rakhlin (2020); Foster & Krishnamurthy (2021), where it is shown to range between sublinear, e.g. $\mathcal{O}(\sqrt{T})$ and constant, related to the cardinality of the parameter space of the estimator.

We adopt the filtration $\mathcal{F}_t$ generated by $(x_1, a_1, r_1, \ldots, x_t, a_t, r_t)$, and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}, x_t]$.

Assumption 4 (Bounded rewards and martingale noise)**.** *For each round* t*, after choosing* a_t*, the learner observes a reward*

$$r_t = \mu(x_t, a_t) + \xi_t,$$

Table 5: Therapeutic Window with Toxicity, grouped across comparator families at $K = 100$, $T = 192$, context dimension $p = 8$, action-attribute dimension $m = 4$, and 10 Monte Carlo trials. The additional column identifies whether each algorithm is a contextual or causal comparator. We report Cumulative Regret (lower is better), BAI Rate (higher is better), and Coverage (higher is better). Bold marks the best algorithm for each metric at this K.

Comparator Family	Algorithm	Cumulative Regret	BAI Rate	Coverage	Wall-clock (ms)
Causal	ConfoundedBudgetedCB	$\mathbf{20.673} \pm 0.468$	0.044	0.020	210.8 ± 0.9
Causal	ContrastSpanner	21.828 ± 1.318	0.297	0.492	182.8 ± 1.0
Causal	ContrastSpanner+Design	29.416 ± 1.112	**0.323**	0.108	523.8 ± 4.8
Causal	Yan2024-SCM	32.975 ± 3.752	0.286	0.214	1115.1 ± 55.6
Causal	Lattimore2016	33.419 ± 0.821	0.014	**1.000**	7.7 ± 0.1
Causal	LinSEM-TS	36.790 ± 1.884	0.152	0.120	238.7 ± 2.2
Causal	C-TS	37.674 ± 1.262	0.249	0.110	238.9 ± 2.7
Causal	HAC-UCB	42.222 ± 1.541	0.031	0.409	245.1 ± 0.7
Causal	C-UCB	44.027 ± 1.992	0.230	0.073	205.0 ± 1.0
Causal	CL-TS	46.309 ± 2.173	0.205	0.117	236.1 ± 1.3
Causal	Sen2017-LowRank	46.509 ± 3.784	0.044	0.187	6.4 ± 0.1
Causal	LinSEM-UCB	47.821 ± 1.399	0.095	0.056	206.8 ± 1.2
Causal	CausalBAI-2018	49.446 ± 0.773	0.006	**1.000**	5.7 ± 0.0
Causal	CL-UCB	53.181 ± 2.333	0.193	0.077	207.0 ± 0.7
Contextual	ContrastSpanner	21.828 ± 1.318	0.297	0.492	182.8 ± 1.0
Contextual	TS	24.589 ± 1.877	0.012	0.109	619.0 ± 2.6
Contextual	ContrastSpanner+Design	29.416 ± 1.112	**0.323**	0.108	523.8 ± 4.8
Contextual	AdaCB-linear	33.471 ± 1.084	0.014	0.638	510.6 ± 13.3
Contextual	AdaCB-xgb	34.009 ± 1.417	0.015	0.633	18569.2 ± 396.1
Contextual	SpannerIGW	34.502 ± 1.958	0.020	0.592	1254.4 ± 7.5
Contextual	ReweightedSpanner	34.916 ± 6.012	0.102	0.064	21008.0 ± 42.2
Contextual	TS-Oracle	37.414 ± 1.447	0.237	0.120	151.6 ± 1.4
Contextual	SquareCB	40.500 ± 1.686	0.013	0.691	27776.6 ± 352.1
Contextual	UCB	46.518 ± 0.889	0.013	**1.000**	107.4 ± 0.7
Contextual	CODA	49.632 ± 1.335	0.008	0.855	913.8 ± 1.4
Contextual	OAM	52.064 ± 1.118	0.009	**1.000**	60.0 ± 0.2
Contextual	UCB-Oracle	60.979 ± 1.629	0.196	0.058	217.8 ± 0.6

where $\mu(x, a)$ is the (unknown) conditional mean reward. Assume there exists $R_{\max} > 0$ such that

$$|\mu(x, a)| \le R_{\max} \quad \textit{and} \quad |r_t| \le R_{\max} \textit{ a.s.}$$

The noise sequence $(\xi_t)_{t\ge1}$ is a martingale difference sequence: $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}, x_t, a_t] = 0$ a.s., and is bounded and square-integrable in the sense that

$$|\xi_t| \le 2R_{\max} \textit{ a.s.}, \qquad \mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t] \le \sigma^2 \textit{ a.s.}$$

for some $\sigma^2 \in (0, \infty)$.

Assumption 5 (Square-loss oracle and realizability)**.** *Let $\{\hat{f}_t\}_{t\ge1}$ be the sequence of predictors returned by the square-loss regression oracle defined in Section 3.1. We assume the oracle predictions are clipped to the range $[-R_{\max}, R_{\max}]$. Moreover, we assume realizability: there exists $f^\star \in \mathcal{F}$ such that $f^\star(x, a) = \mu(x, a)$ for all (x, a). Equivalently, we may take the comparator in the oracle regret bound (defined in Section 3.1) to be $f^\star = \mu$.*

E.1.1 Inverse-gap weighting on a pool: lemmas and technical results

Fix a round t. To simplify the proof exposition, we temporarily suppress dependence on the context, i.e., we write $\hat{y}(a) = f(x, a)$. Given scores $\hat{y}(a)$ for $a \in \mathcal{A}_t$, define

$$\hat{a} := \arg\max_{a\in\mathcal{A}_t} \hat{y}(a), \qquad \hat{\Delta}(a) := \hat{y}(\hat{a}) - \hat{y}(a) \ \ge 0.$$

Table 6: Therapeutic Window with Toxicity, grouped across comparator families at $K = 250$, $T = 192$, context dimension $p = 8$, action-attribute dimension $m = 4$, and 10 Monte Carlo trials. The additional column identifies whether each algorithm is a contextual or causal comparator. We report Cumulative Regret (lower is better), BAI Rate (higher is better), and Coverage (higher is better). Bold marks the best algorithm for each metric at this K. For transparency and statistical integrity, this caption reports the exact experimental setting, metric directionality, and Monte Carlo replication count, and the table should be interpreted descriptively rather than as a thresholded significance claim. At this K, the regret winner is ReweightedSpanner, the BAI winner is Yan2024-SCM, and the coverage winner is Lattimore2016. Within the ContrastSpanner family, the stronger variant at this K is ContrastSpanner for regret and ContrastSpanner for BAI. At larger action-space sizes, a plausible explanation is that the learner must separate many near-by doses, so representation quality and targeted contrast selection matter more, but the extra design layer can still trade lower identification error against noisier regret.

Comparator Family	Algorithm	Cumulative Regret	BAI Rate	Coverage	Wall-clock (ms)
Causal	Yan2024-SCM	18.757 ± 1.309	$\mathbf{0.324}$	0.084	1226.2 ± 51.0
Causal	ConfoundedBudgetedCB	20.793 ± 0.240	0.046	0.008	503.0 ± 1.9
Causal	ContrastSpanner	23.596 ± 1.567	0.314	0.218	260.3 ± 0.5
Causal	ContrastSpanner+Design	30.740 ± 1.149	0.305	0.047	633.4 ± 3.2
Causal	LinSEM-TS	32.352 ± 0.991	0.215	0.048	537.1 ± 2.0
Causal	C-TS	39.043 ± 1.334	0.236	0.044	537.5 ± 1.3
Causal	HAC-UCB	42.956 ± 1.303	0.023	0.201	582.9 ± 1.4
Causal	Sen2017-LowRank	43.199 ± 3.862	0.055	0.074	6.9 ± 0.0
Causal	C-UCB	44.040 ± 1.075	0.223	0.028	500.9 ± 1.5
Causal	Lattimore2016	47.153 ± 0.829	0.006	$\mathbf{0.768}$	9.3 ± 0.0
Causal	CausalBAI-2018	47.153 ± 0.829	0.006	$\mathbf{0.768}$	6.3 ± 0.0
Causal	LinSEM-UCB	49.110 ± 1.598	0.072	0.021	497.7 ± 1.1
Causal	CL-TS	51.544 ± 1.797	0.181	0.044	545.5 ± 3.2
Causal	CL-UCB	53.580 ± 1.608	0.201	0.030	502.2 ± 1.7
Contextual	ReweightedSpanner	$\mathbf{13.270} \pm 3.388$	0.206	0.024	125223.3 ± 659.8
Contextual	TS	20.649 ± 1.866	0.000	0.044	1509.7 ± 4.6
Contextual	ContrastSpanner	23.596 ± 1.567	0.314	0.218	260.3 ± 0.5
Contextual	SpannerIGW	27.458 ± 1.322	0.082	0.239	6836.3 ± 19.8
Contextual	ContrastSpanner+Design	30.740 ± 1.149	0.305	0.047	633.4 ± 3.2
Contextual	TS-Oracle	36.982 ± 0.872	0.198	0.047	319.5 ± 0.5
Contextual	AdaCB-xgb	38.550 ± 1.457	0.004	0.414	9663.1 ± 231.1
Contextual	AdaCB-linear	39.115 ± 1.366	0.006	0.416	585.5 ± 13.6
Contextual	SquareCB	39.825 ± 1.438	0.006	0.412	35958.5 ± 600.2
Contextual	UCB	43.951 ± 1.493	0.012	0.417	255.4 ± 0.8
Contextual	OAM	46.470 ± 0.997	0.015	0.740	135.2 ± 0.2
Contextual	CODA	48.189 ± 1.242	0.004	0.546	2732.6 ± 12.6
Contextual	UCB-Oracle	61.977 ± 0.997	0.200	0.024	531.0 ± 0.8

This matches the notation in Section 3 (see (12) and (10)).

Definition 1 (Pool IGW distribution)**.** *Fix parameters $\vartheta > 0$ and $\gamma > 0$. Define a distribution p on $\mathcal{A}_t$ by*

$$\pi(a) := \frac{1}{\vartheta + \gamma\, \hat{\Delta}(a)} \text{ for } a \neq \hat{a}, \qquad \pi(\hat{a}) := 1 - \sum_{b \in \mathcal{A}_t \setminus \{\hat{a}\}} \pi(b). \tag{32}$$

We will impose that the regularizer satisfies $\vartheta \geq K_t$ so that $\pi(\hat{a}) \geq 0$ (indeed $\sum_{b \neq \hat{a}} \pi(b) \leq (K_t - 1)/\vartheta \leq 1 - 1/\vartheta$).

Lemma 2 (High-probability IGW one-step inequality on a pool)**.** *Fix a round t and a pool $\mathcal{A}_t$ of size K_t. Let $a^\circ \in \arg\max_{a \in \mathcal{A}_t} \mu(x_t, a)$ denote the best-in-pool action. Let π be the IGW distribution from Definition 1 with $\vartheta \geq K_t$. Then, for any (deterministic) score vector $\hat{y} : \mathcal{A}_t \to \mathbb{R}$ and any*

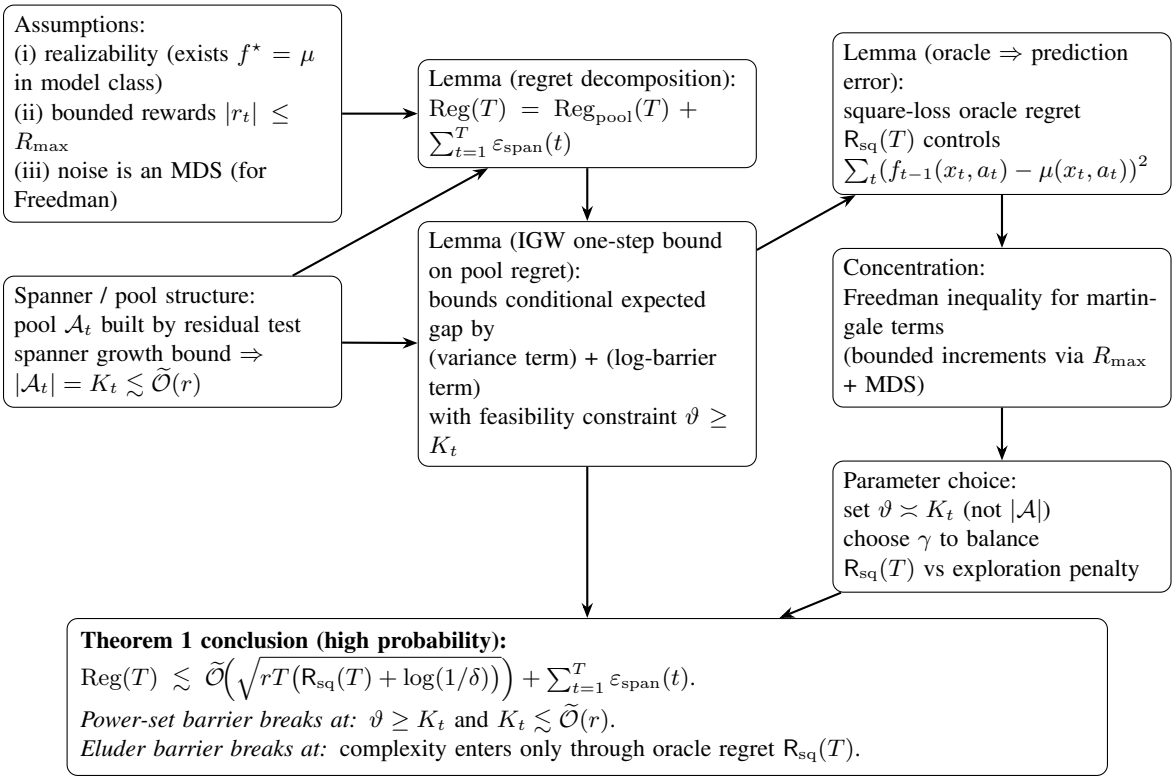


Figure 8: Proof logic flowchart for Theorem 1 (square-loss oracle, stochastic high-probability regret).

(realizable) mean reward vector $y^\star(a) := \mu(x_t, a)$ *(Assumption 5 holds),*

$$\mathbb{E}_{a\sim\pi}\big[y^\star(a^\circ) - y^\star(a)\big] \le \frac{2\vartheta}{\gamma} + \frac{\gamma}{4}\,\mathbb{E}_{a\sim\pi}\Big[\big(\hat{y}(a) - y^\star(a)\big)^2\Big]. \tag{33}$$

Proof. This inequality is the pool-restricted analogue of the key inverse-gap-weighting step in SQUARECB; see Eq. (10) in Foster & Rakhlin (2020). We provide a self-contained proof.

Fix any $a^\circ \in \arg\max_{a\in\mathcal{A}_t} y^\star(a)$ and define the prediction error $\mathrm{err}(a) := \hat{y}(a) - y^\star(a)$. Let $\hat{a} \in \arg\max_{a\in\mathcal{A}_t} \hat{y}(a)$ and $\hat{\Delta}(a) := \hat{y}(\hat{a}) - \hat{y}(a) \ge 0$. Let π be the IGW distribution from (32) (with $\vartheta \ge K_t$). This is the key set that allows us to remove the power set barrier – dependence on $|\mathcal{A}_t|$ replaces $|\mathcal{A}|$.

If $\pi(a^\circ) = 1$, then $a = a^\circ$ almost surely and the left-hand side of (33) is 0, so the claim is trivial. Hence assume $\pi(a^\circ) \in (0, 1)$ below.

Step 1: A four-term decomposition. By conditioning on whether $a = a^\circ$,

$$\begin{aligned}
\mathbb{E}_{a\sim\pi}\big[y^\star(a^\circ) - y^\star(a)\big] &= \sum_{a\in\mathcal{A}_t\setminus\{a^\circ\}} \pi(a)\,\big(y^\star(a^\circ) - y^\star(a)\big) \\
&= (1 - \pi(a^\circ))\big(y^\star(a^\circ) - \hat{y}(a^\circ)\big) + (1 - \pi(a^\circ))\big(\hat{y}(a^\circ) - \hat{y}(\hat{a})\big) \\
&\quad + \sum_{a\in\mathcal{A}_t\setminus\{a^\circ\}} \pi(a)\,\big(\hat{y}(\hat{a}) - \hat{y}(a)\big) + \sum_{a\in\mathcal{A}_t\setminus\{a^\circ\}} \pi(a)\,\big(\hat{y}(a) - y^\star(a)\big).
\end{aligned} \tag{34}$$

Denote these four terms by T_1, T_2, T_3, T_4 in the order they appear.

Step 2: Convert T_1 and T_4 into squared errors (AM–GM). For any real x and any $\eta > 0$, the inequality

$$x \le \frac{1}{\eta} + \frac{\eta}{4}x^2 \tag{35}$$

holds (since $(\frac{\sqrt{\eta}}{2}x - \frac{1}{\sqrt{\eta}})^2 \ge 0$).

Apply (35) to T_4 with $\eta = \gamma$ and $x = \hat{y}(a) - y^\star(a) = \mathrm{err}(a)$ for each $a \ne a^\circ$:

$$T_4 = \sum_{a \ne a^\circ} \pi(a)\,\mathrm{err}(a) \le \frac{1-\pi(a^\circ)}{\gamma} + \frac{\gamma}{4}\sum_{a \ne a^\circ} \pi(a)\,\mathrm{err}(a)^2. \tag{36}$$

Apply (35) to T_1 with $\eta = \gamma\,\frac{\pi(a^\circ)}{1-\pi(a^\circ)}$ and $x = y^\star(a^\circ) - \hat{y}(a^\circ) = -\mathrm{err}(a^\circ)$:

$$T_1 = (1-\pi(a^\circ))\big(y^\star(a^\circ) - \hat{y}(a^\circ)\big) \le \frac{(1-\pi(a^\circ))^2}{\pi(a^\circ)\gamma} + \frac{\gamma}{4}\,\pi(a^\circ)\,\mathrm{err}(a^\circ)^2. \tag{37}$$

Step 3: Cancellation and a first reduction. Substituting (36) and (37) into (34) and subtracting $\frac{\gamma}{4}\mathbb{E}_{a\sim p}[\mathrm{err}(a)^2]$ from both sides yields

$$\mathbb{E}_{a\sim\pi}\big[y^\star(a^\circ) - y^\star(a)\big] - \frac{\gamma}{4}\mathbb{E}_{a\sim\pi}[\mathrm{err}(a)^2] \le \frac{1-\pi(a^\circ)}{\gamma} + \Big(T_2 + \frac{(1-\pi(a^\circ))^2}{\pi(a^\circ)\gamma}\Big) + T_3. \tag{38}$$

Step 4: Bound the (T_2-residual) term by ϑ/γ. We claim that

$$T_2 + \frac{(1-\pi(a^\circ))^2}{\pi(a^\circ)\gamma} \le \frac{\vartheta}{\gamma}. \tag{39}$$

There are two cases.

Case 1: $a^\circ = \hat{a}$. Then $T_2 = (1-\pi(\hat{a}))(\hat{y}(\hat{a}) - \hat{y}(\hat{a})) = 0$. Also $\pi(\hat{a}) \ge 1/\vartheta$ because $\sum_{b\ne\hat{a}}\pi(b) \le (K_t - 1)/\vartheta \le 1 - 1/\vartheta$. Hence

$$\frac{(1-\pi(\hat{a}))^2}{\pi(\hat{a})\gamma} \le \frac{1}{\pi(\hat{a})\gamma} \le \frac{\vartheta}{\gamma}.$$

Case 2: $a^\circ \ne \hat{a}$. Let $\Delta := \hat{\Delta}(a^\circ) = \hat{y}(\hat{a}) - \hat{y}(a^\circ) > 0$. Then $\hat{y}(a^\circ) - \hat{y}(\hat{a}) = -\Delta$ and, by (32), $\pi(a^\circ) = 1/(\vartheta + \gamma\Delta)$ as given in Def. 1. Writing $\pi := \pi(a^\circ)$ for brevity, we have

$$\begin{aligned} T_2 + \frac{(1-p)^2}{p\gamma} &= (1-\pi)(-\Delta) + \frac{(1-\pi)^2}{p\gamma} = \frac{\vartheta+\gamma\Delta-1}{\vartheta+\gamma\Delta}\Big(-\Delta + \frac{\vartheta+\gamma\Delta-1}{\gamma}\Big) \\ &= \frac{\vartheta+\gamma\Delta-1}{\vartheta+\gamma\Delta}\cdot\frac{\vartheta-1}{\gamma} \le \frac{\vartheta}{\gamma}, \end{aligned}$$

where we used $\vartheta \ge 1$. This proves (39).

Step 5: Bound T_3 by $(K_t - 1)/\gamma$. Since $\hat{\Delta}(\hat{a}) = 0$, we may restrict the sum to $a \ne \hat{a}$:

$$T_3 = \sum_{a\ne a^\circ} \pi(a)\,\hat{\Delta}(a) \le \sum_{a\in\mathcal{A}_t\setminus\{\hat{a}\}} \pi(a)\,\hat{\Delta}(a).$$

For $a \ne \hat{a}$, $\pi(a) = 1/(\vartheta + \gamma\hat{\Delta}(a))$, so

$$\pi(a)\,\hat{\Delta}(a) = \frac{\hat{\Delta}(a)}{\vartheta + \gamma\hat{\Delta}(a)} \le \frac{1}{\gamma}.$$

There are at most $K_t - 1$ actions in $\mathcal{A}_t \setminus \{\hat{a}\}$, hence

$$T_3 \le \frac{K_t - 1}{\gamma}. \tag{40}$$

Step 6: Combine bounds. Using (39) and (40) in (38) gives

$$\mathbb{E}_{a\sim\pi}[y^\star(a^\circ) - y^\star(a)] - \frac{\gamma}{4}\mathbb{E}_{a\sim\pi}[\mathrm{err}(a)^2] \le \frac{1-\pi(a^\circ)}{\gamma} + \frac{\vartheta}{\gamma} + \frac{K_t - 1}{\gamma} \le \frac{\vartheta + K_t}{\gamma} \le \frac{2\vartheta}{\gamma},$$

where the second inequality uses $1 - \pi(a^\circ) \le 1$ and the last uses $\vartheta \ge K_t$. Rearranging yields (33). □

Lemma 3 (Freedman inequality)**.** *(Lattimore & Szepesvári, 2020, Theorem 20.1) Let $(X_t)_{t\ge1}$ be a real-valued martingale difference sequence with respect to a filtration $(\mathcal{H}_t)_{t\ge0}$. Assume $|X_t| \le b$ a.s. for all t. Define the predictable quadratic variation $V_T := \sum_{t=1}^T \mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^T X_t \le \sqrt{2V_T \log(\delta^{-1})} + \frac{b}{3}\log(\delta^{-1}). \tag{41}$$

E.1.2 From square-loss oracle regret to prediction error

Define the prediction error on the played actions

$$e_t := \hat{f}_{t-1}(x_t, a_t) - \mu(x_t, a_t).$$

By Assumptions 4–5 and clipping, $|e_t| \le 2R_{\max}$ almost surely.

Lemma 4 (Prediction error from square-loss oracle regret)**.** *Assume Assumptions 4 and 5. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^T \big(\hat{f}_{t-1}(x_t,a_t) - \mu(x_t,a_t)\big)^2 \le 2\mathsf{R}_{\mathrm{sq}}(T) + \Big(8\sigma^2 + \frac{16}{3}R_{\max}^2\Big)\log(\delta^{-1}). \tag{42}$$

Proof. Let $f_t := \hat{f}_{t-1}$ for brevity. By the oracle regret definition in Section 3.1 and realizability (Assumptions 3.1 and 5, comparator $f^\star = \mu$), we have

$$\mathrm{Reg}_{\mathrm{Sq}}(T) = \sum_{t=1}^T \big(f_t(x_t,a_t) - r_t\big)^2 - \sum_{t=1}^T \big(\mu(x_t,a_t) - r_t\big)^2 \le \mathsf{R}_{\mathrm{sq}}(T). \tag{43}$$

Using $r_t = \mu(x_t,a_t) + \xi_t$ and expanding,

$$\big(f_t - r_t\big)^2 - \big(\mu - r_t\big)^2 = \big(f_t - \mu - \xi_t\big)^2 - \xi_t^2 = e_t^2 - 2e_t\xi_t.$$

Summing over t and substituting into (43) gives

$$\sum_{t=1}^T e_t^2 \le \mathsf{R}_{\mathrm{sq}}(T) + 2\sum_{t=1}^T e_t\xi_t. \tag{44}$$

Let $X_t := e_t\xi_t$. Then (X_t) is a martingale difference sequence because e_t is measurable with respect to $\mathcal{F}_{t-1} \cup \{x_t, a_t\}$ and $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}, x_t, a_t] = 0$. Moreover, $|X_t| \le |e_t|\,|\xi_t| \le (2R_{\max})(2R_{\max}) = 4R_{\max}^2$. Also,

$$\mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}] = e_t^2\,\mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t] \le \sigma^2 e_t^2.$$

Thus the predictable variation satisfies $V_T \le \sigma^2 \sum_{t=1}^T e_t^2$. Applying Lemma 3 with $b = 4R_{\max}^2$ shows that with probability at least $1-\delta$,

$$\sum_{t=1}^T e_t\xi_t \le \sqrt{2\sigma^2\Big(\sum_{t=1}^T e_t^2\Big)\log(\delta^{-1})} + \frac{4R_{\max}^2}{3}\log(\delta^{-1}). \tag{45}$$

Substituting (45) into (44) yields

$$\sum_{t=1}^{T} e_t^2 \le \mathsf{R}_{\mathrm{sq}}(T) + 2\sqrt{2\sigma^2\Big(\sum_{t=1}^{T} e_t^2\Big)\log(\delta^{-1}) + \frac{8R_{\max}^2}{3}\log(\delta^{-1})}.$$

Finally, use the inequality $\sqrt{uv} \le \frac{1}{2}u + \frac{1}{2}v$ with $u := \sum_{t=1}^{T} e_t^2$ and $v := 8\sigma^2\log(\delta^{-1})$ to obtain

$$2\sqrt{2\sigma^2\Big(\sum_{t=1}^{T} e_t^2\Big)\log(\delta^{-1})} = \sqrt{8\sigma^2\Big(\sum_{t=1}^{T} e_t^2\Big)\log(\delta^{-1})} \le \frac{1}{2}\sum_{t=1}^{T} e_t^2 + 4\sigma^2\log(\delta^{-1}).$$

Rearranging yields (42). □

Lemma 5 (Controlling predictable squared error by realized squared error)**.** *Assume $|e_t| \le 2R_{\max}$ a.s. for all t. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_t[e_t^2] \le 2\sum_{t=1}^{T} e_t^2 + \frac{32}{3}R_{\max}^2\log(\delta^{-1}). \tag{46}$$

Proof. Define the martingale difference sequence $X_t := \mathbb{E}_t[e_t^2] - e_t^2$. Then $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ and, since X_t and $\bar{X}_t = \mathbb{E}_t[e_t^2]$ are less than $4R_{\max}^2$

$$|X_t| = |\bar{X}_t - e_t^2| \le 4R_{\max}^2.$$

Moreover,

$$\mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] \le \mathbb{E}_t[e_t^4] \le (4R_{\max}^2)\,\mathbb{E}_t[e_t^2],$$

because $e_t^4 \le (4R_{\max}^2)e_t^2$ whenever $e_t^2 \le 4R_{\max}^2$. Thus the predictable variation satisfies $V_T \le 4R_{\max}^2\sum_{t=1}^{T}\mathbb{E}_t[e_t^2]$. Applying Lemma 3 (with $b = 4R_{\max}^2$) gives that with prob. at least $1-\delta$,

$$\sum_{t=1}^{T}\big(\mathbb{E}_t[e_t^2] - e_t^2\big) \le \sqrt{8R_{\max}^2\Big(\sum_{t=1}^{T}\mathbb{E}_t[e_t^2]\Big)\log(\delta^{-1}) + \frac{4R_{\max}^2}{3}\log(\delta^{-1})}.$$

Use $\sqrt{uv} \le \frac{1}{2}u + \frac{1}{2}v$ with $u := \sum_{t=1}^{T}\mathbb{E}_t[e_t^2]$ and $v := 8R_{\max}^2\log(\delta^{-1})$ to upper bound the square-root term by $\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_t[e_t^2] + 4R_{\max}^2\log(\delta^{-1})$. Rearranging then yields (46). □

E.2 High-probability regret for Algorithm 1 (CONTRASTSPANNER)

Recall the (random) cumulative bandit regret $\mathrm{Reg}(T)$ from Section 1 and the exact identity (Section F)

$$\mathrm{Reg}(T) = \underbrace{\sum_{t=1}^{T}\Big(\max_{a\in\mathcal{A}}\mu(x_t,a) - \max_{a\in\mathcal{A}_t}\mu(x_t,a)\Big)}_{\text{span gap term } \sum_t \varepsilon_{\mathrm{span}}(t)} + \underbrace{\sum_{t=1}^{T}\Big(\max_{a\in\mathcal{A}_t}\mu(x_t,a) - \mu(x_t,a_t)\Big)}_{\text{in-pool regret}}.$$

The first term is controlled by the low-rank / span-adequacy analysis in Section F; in this square-loss subsection we treat it as an explicit additive term and focus on bounding the in-pool regret.

Theorem (**Restatement of Theorem 1**): High-probability regret of CONTRASTSPANNER (Alg. 1) with a square-loss oracle)**.** *Fix $T \ge 1$ and $\delta \in (0,1)$. Assume 2, 3.1, 4, and 5 so that we may invoke Proposition 3 [cf. (129)] Let K be the uniform pool-size upper bound in (129). Choose the exploration parameter ϑ in (32) so that*

$$\vartheta \ge \max_{t\le T} K_t \;\; \textit{(in particular, it suffices to take } \vartheta := K\textit{)}.$$

Set the learning-rate parameter

$$\gamma := \sqrt{\frac{2\vartheta T}{\mathsf{R}_{\mathrm{sq}}(T) + \left(4\sigma^2 + \frac{16}{3}R_{\max}^2\right)\log(4\delta^{-1})}}. \tag{47}$$

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

$$\begin{aligned}\mathrm{Reg}(T) &\le \sum_{t=1}^{T}\varepsilon_{\mathrm{span}}(t) + \frac{2\vartheta T}{\gamma} + \gamma\mathsf{R}_{\mathrm{sq}}(T) + \gamma\left(4\sigma^2 + \frac{16}{3}R_{\max}^2\right)\log(4\delta^{-1}) + 2R_{\max}\sqrt{2T\log(4\delta^{-1})} \\ &\le \sum_{t=1}^{T}\varepsilon_{\mathrm{span}}(t) + 2\sqrt{2\vartheta T\left(\mathsf{R}_{\mathrm{sq}}(T) + \left(4\sigma^2 + \tfrac{16}{3}R_{\max}^2\right)\log(4\delta^{-1})\right)} + 2R_{\max}\sqrt{2T\log(4\delta^{-1})}.\end{aligned} \tag{48}$$

In particular, if $\sum_{t=1}^{T}\varepsilon_{\mathrm{span}}(t) = o(\sqrt{T})$ *(see Lemma 19) and* ϑ *grows sublinearly in* T *(e.g.* $\vartheta = \tilde{\mathcal{O}}(r)$ *from (129)), then the bound is* $\tilde{\mathcal{O}}(\sqrt{T})$*.*

Proof. Let $a_t^\circ \in \arg\max_{a\in\mathcal{A}_t}\mu(x_t,a)$ denote the best-in-pool action. Define the in-pool regret random variable

$$R_t^{\mathrm{pool}} := \mu(x_t,a_t^\circ) - \mu(x_t,a_t) \ \in [0, 2R_{\max}].$$

Then

$$\mathrm{Reg}(T) = \sum_{t=1}^{T}\varepsilon_{\mathrm{span}}(t) + \sum_{t=1}^{T}R_t^{\mathrm{pool}}. \tag{49}$$

which is a similar regret decomposition as that which appears for the square-loss oracle.

Step 1: Concentration from realized in-pool regret to its conditional expectation. The sequence $\left(R_t^{\mathrm{pool}} - \mathbb{E}_t[R_t^{\mathrm{pool}}]\right)_{t\le T}$ is a martingale difference sequence, and $|R_t^{\mathrm{pool}} - \mathbb{E}_t[R_t^{\mathrm{pool}}]| \le 2R_{\max}$ almost surely. By Azuma–Hoeffding, with probability at least $1-\delta/4$,

$$\sum_{t=1}^{T}R_t^{\mathrm{pool}} \le \sum_{t=1}^{T}\mathbb{E}_t[R_t^{\mathrm{pool}}] + 2R_{\max}\sqrt{2T\log(4\delta^{-1})}. \tag{50}$$

Step 2: Bound the conditional expectation by IGW + squared prediction error. Condition on $\mathcal{F}_{t-1}, x_t$. On round t, ContrastSpanner samples $a_t \sim p_t$ from the pool-IGW rule (32) (this is exactly the pool restriction of (12)). Applying Lemma 2 with $\hat{y}(a) = \hat{f}_{t-1}(x_t,a)$ and $y^\star(a) = \mu(x_t,a)$ gives

$$\mathbb{E}_t[R_t^{\mathrm{pool}}] \le \frac{2\vartheta}{\gamma} + \frac{\gamma}{4}\,\mathbb{E}_t\left[\left(\hat{f}_{t-1}(x_t,a_t) - \mu(x_t,a_t)\right)^2\right] = \frac{2\vartheta}{\gamma} + \frac{\gamma}{4}\,\mathbb{E}_t[e_t^2]. \tag{51}$$

Summing (51) over $t \le T$ and substituting into (74) yields: with probability at least $1-\delta/4$,

$$\sum_{t=1}^{T}R_t^{\mathrm{pool}} \le \frac{2\vartheta T}{\gamma} + \frac{\gamma}{4}\sum_{t=1}^{T}\mathbb{E}_t[e_t^2] + 2R_{\max}\sqrt{2T\log(4\delta^{-1})}. \tag{52}$$

Step 3: Control $\sum_t \mathbb{E}_t[e_t^2]$ using Freedman and the oracle regret. This is a standard Freedman step controlling the gap between conditional quadratic variation and realized error. In particular: apply Lemma 5 with failure probability $\delta/4$ to obtain

$$\sum_{t=1}^{T}\mathbb{E}_t[e_t^2] \le 2\sum_{t=1}^{T}e_t^2 + \frac{32}{3}R_{\max}^2\log(4\delta^{-1}). \tag{53}$$

Next apply Lemma 4 with failure probability $\delta/4$ to obtain

$$\sum_{t=1}^{T} e_t^2 \le 2\mathsf{R}_{\mathrm{sq}}(T) + \Big(8\sigma^2 + \frac{16}{3}R_{\max}^2\Big)\log(4\delta^{-1}). \tag{54}$$

Combining (53) and (54) yields, on the intersection of the two events,

$$\sum_{t=1}^{T} \mathbb{E}_t[e_t^2] \le 4\mathsf{R}_{\mathrm{sq}}(T) + \Big(16\sigma^2 + \frac{64}{3}R_{\max}^2\Big)\log(4\delta^{-1}). \tag{55}$$

Step 4: Substitute back and simplify. Substituting (55) into (52) gives

$$\begin{aligned}\sum_{t=1}^{T} R_t^{\mathrm{pool}} &\le \frac{2\vartheta T}{\gamma} + \frac{\gamma}{4}\Big(4\mathsf{R}_{\mathrm{sq}}(T) + \big(16\sigma^2 + \tfrac{64}{3}R_{\max}^2\big)\log(4\delta^{-1})\Big) + 2R_{\max}\sqrt{2T\log(4\delta^{-1})} \\ &= \frac{2\vartheta T}{\gamma} + \gamma\mathsf{R}_{\mathrm{sq}}(T) + \gamma\Big(4\sigma^2 + \frac{16}{3}R_{\max}^2\Big)\log(4\delta^{-1}) + 2R_{\max}\sqrt{2T\log(4\delta^{-1})}.\end{aligned}$$

Finally, combine this with the decomposition (49) to obtain the first line of (48).

Step 5: Choice of γ. The first line of (48) has the form

$$\sum_{t=1}^{T}\varepsilon_{\mathrm{span}}(t) + \frac{2\vartheta T}{\gamma} + \gamma B + 2R_{\max}\sqrt{2T\log(4\delta^{-1})},$$

where $B := \mathsf{R}_{\mathrm{sq}}(T) + \big(4\sigma^2 + \frac{16}{3}R_{\max}^2\big)\log(4\delta^{-1})$. The choice (47) minimizes $\frac{2\vartheta T}{\gamma} + \gamma B$ and yields

$$\frac{2\vartheta T}{\gamma} + \gamma B = 2\sqrt{2\vartheta T B},$$

which is the second line of (48).

Union bound. We used three concentration events (Azuma plus two Freedman applications), each at failure probability $\delta/4$. A union bound gives total failure probability at most δ. □

Remark 4 (Parameter selection and barrier mitigation (square-loss))**.** *The high-probability bound in Theorem 1 has the form*

$$\mathrm{Reg}(T) \;\le\; B\sqrt{\lambda}\sum_{t=1}^{T}\varepsilon_t + \frac{2R_{\max}}{\gamma}\sum_{t=1}^{T}|\mathcal{A}_t| + \frac{\gamma}{4R_{\max}}\Big(2\mathsf{R}_{\mathrm{sq}}(T) + c_{\mathrm{hp}}\log\tfrac{1}{\delta}\Big) + c'_{\mathrm{hp}}R_{\max}\sqrt{T\log\tfrac{1}{\delta}}, \tag{56}$$

where $c_{\mathrm{hp}}, c'_{\mathrm{hp}}$ are the explicit numerical constants from the theorem, which subsume dependence on the reward boundedness constant and other constant scaling factors. The four terms in (56) isolate the fundamental tradeoffs:

- $\sum_t \varepsilon_t$ *controls the* span approximation *bias (via Lemma 19);*
- $\big(\sum_t |\mathcal{A}_t|\big)/\gamma$ *is the* exploration *penalty from IGW (Lemma 2);*
- $\gamma\,\mathsf{R}_{\mathrm{sq}}(T)$ *is the* oracle/estimation *penalty (Lemma 4);*
- *the final term is the martingale concentration penalty from converting conditional to realized regret.*

Choosing ϑ. *The IGW rule (12) defines a valid distribution provided $\vartheta \ge \max_{t\le T}|\mathcal{A}_t|$ (this is the same normalization condition as in* SQUARECB *Foster & Rakhlin (2020) with the global action count K replaced by the* learned *pool size $|\mathcal{A}_t|$). By the spanner growth bound (cf. (129)),*

$|\mathcal{A}_t|$ is controlled by the spanner size and hence by the intrinsic contrast rank r up to logarithmic factors; consequently one may set ϑ proportional to this spanner-based upper bound rather than proportional to $|\mathcal{A}|$.

Choosing γ (constructive balancing). *Ignoring only the span term and the final martingale term in (56), the remaining dependence on γ is*

$$\frac{2R_{\max}}{\gamma}\sum_{t=1}^{T}|\mathcal{A}_t| + \frac{\gamma}{4R_{\max}}\left(2\mathsf{R}_{\mathrm{sq}}(T) + c_{\mathrm{hp}}\log\tfrac{1}{\delta}\right),$$

which is minimized by the explicit choice

$$\gamma^\star = 2R_{\max}\sqrt{\frac{\sum_{t=1}^{T}|\mathcal{A}_t|}{\left(2\mathsf{R}_{\mathrm{sq}}(T) + c_{\mathrm{hp}}\log\tfrac{1}{\delta}\right)}}. \tag{57}$$

Substituting (57) yields the balanced contribution

$$\frac{2R_{\max}}{\gamma^\star}\sum_{t=1}^{T}|\mathcal{A}_t| + \frac{\gamma^\star}{4R_{\max}}\left(2\mathsf{R}_{\mathrm{sq}}(T) + c_{\mathrm{hp}}\log\tfrac{1}{\delta}\right) = 2\sqrt{\sum_{t=1}^{T}|\mathcal{A}_t|\,\left(2\mathsf{R}_{\mathrm{sq}}(T) + c_{\mathrm{hp}}\log\tfrac{1}{\delta}\right)}. \tag{58}$$

Thus, whenever S_T grows at most polylogarithmically in T times T through the spanner bound (e.g. $|\mathcal{A}_t| \lesssim r\log(1+t/\lambda)$), and the oracle has sublinear square-loss regret $\mathsf{R}_{\mathrm{sq}}(T) = o(T)$, the term (58) is $o(T)$ and in particular is $\tilde{\mathcal{O}}(\sqrt{Tr})$ under the standard regime $\mathsf{R}_{\mathrm{sq}}(T) = \tilde{\mathcal{O}}(\sqrt{T})$.

Choosing ε_t. *The span term is bounded by $B\sqrt{\lambda}\sum_t \varepsilon_t$ (Lemma 19). A sufficient condition for this term to be $\tilde{\mathcal{O}}(\sqrt{T})$ is $\sum_{t=1}^{T}\varepsilon_t = \tilde{\mathcal{O}}(\sqrt{T})$ (e.g. constant $\varepsilon_t \equiv 0$ in the exactly realizable case, or $\varepsilon_t \equiv T^{-1/2}$ for a simple schedule). This choice interacts with the pool complexity through (129): smaller ε reduces the span bias but increases the spanner/pool size, hence increases S_T and the balanced term (58). The theorem makes this tradeoff explicit.*

Barrier mitigation. *The ambient action set $\mathcal{A}$ may be enormous, but the regret bound depends on it only through $|\mathcal{A}_t|$, which is controlled by the spanner growth bound in terms of r (not $|\mathcal{A}|$). This removes "power-set" dependence on $|\mathcal{A}|$. Furthermore, oracle complexity enters only through $\mathsf{R}_{\mathrm{sq}}(T)$ and exploration geometry enters only through r (via $|\mathcal{A}_t|$), thereby separating approximation/estimation from exploration geometry and avoiding Eluder-style scaling with a global complexity parameter (Russo & Van Roy, 2013b).*

Remark 5 (Parameter choice and linear-bandit rate)**.** *If the low-rank contrast model is* exactly *realizable in the spanner (so $\varepsilon = 0$), one may set a constant threshold ε (e.g. $\varepsilon = 1$) to keep $|\mathcal{A}_t| = \tilde{\mathcal{O}}(r\log(1+t/\lambda))$ by (129), and choose $\gamma_t \equiv \sqrt{T}$. Then the IGW exploration penalty in (48) is at most $\tilde{\mathcal{O}}\big(r\sqrt{T}\,\log(1+T/\lambda)\log(1+R_{\max}\sqrt{T}/\vartheta)\big)$, and the remaining term matches the standard $\sqrt{T}$ scaling of linear contextual bandits up to logarithmic factors. The dependence on $|\mathcal{A}_t|$ enters only through the* spanner size *bound (129), eliminating power-set dependence on $|\mathcal{A}|$ and separating oracle complexity (in $\mathsf{R}_{\mathrm{sq}}(T)$) from exploration geometry (in r), thereby mitigating Eluder-style scaling.*

Remark 6 (SquareCB versus SpannerIGW: why the SquareCB-style strategy above?)**.** *The regret analysis underlying Theorem 1 deliberately adapts the* SQUARECB *template (Foster & Rakhlin, 2020): once the spanner/residual test controls the effective pool size $|\mathcal{A}_t|$, the standard inverse-gap weighting (IGW) argument yields a clean reduction from contextual bandit regret to the square-loss oracle regret $\mathrm{Reg}_{Sq}(T)$ plus a log-barrier term. Conceptually, this is the same reduction that* SPANNERIGW *uses, but with a* trivial design *(uniform / finite-action covariance control), which leads to the crude complexity scaling $\mathrm{dec}_{\tilde{\gamma}} \lesssim |\mathcal{A}_t|/\tilde{\gamma}$, where $\mathrm{dec}_{\tilde{\gamma}}$ is the variant of decision estimation coefficient as defined in (66), dating back to the closely related disagreement coefficient Hanneke (2014); Li et al. (2022), and $\tilde{\gamma}$ is a scaled version of the IGW gap coefficient to be defined in the next subsection.*

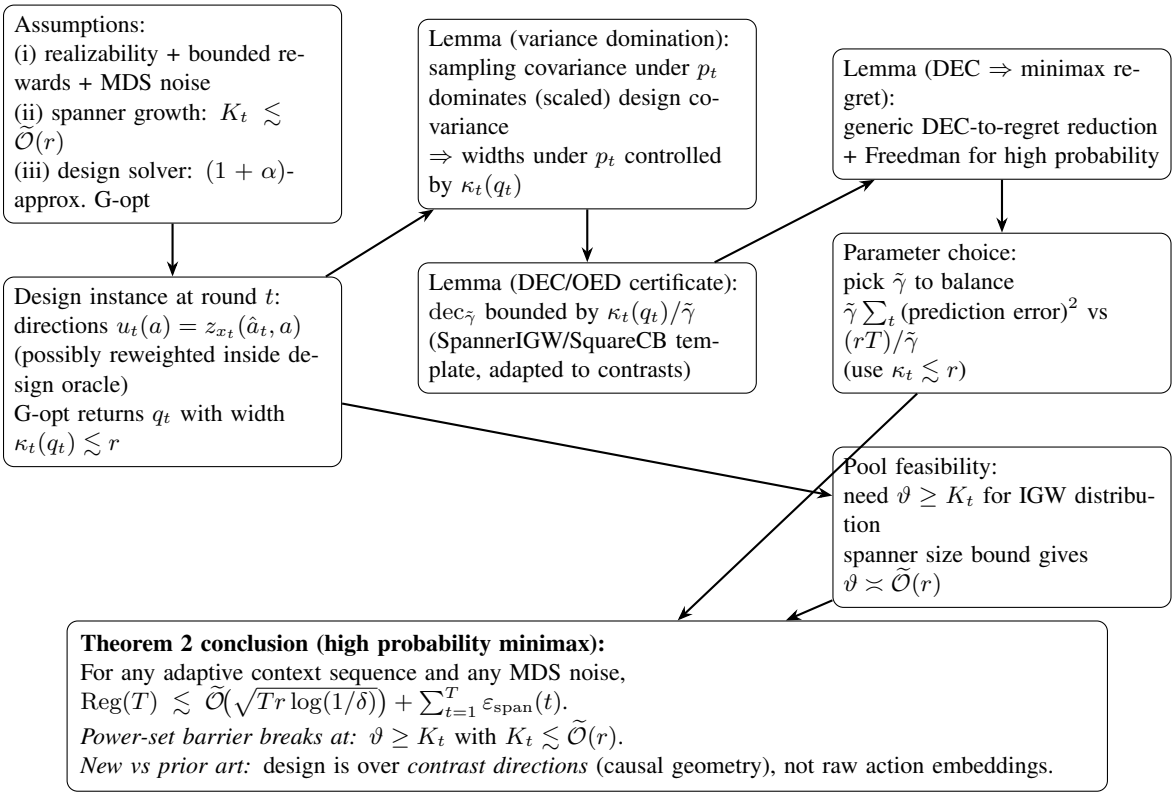


Figure 9: Proof logic flowchart for Theorem 2 (square-loss oracle, minimax high-probability regret via OED/DEC).

E.3 Square-loss oracle: minimax regret and the role of optimal design

We now turn to *minimax* (adversarial) regret guarantees. In contrast to the stochastic analysis above, minimax bounds must hold uniformly over all context and reward sequences satisfying Assumption 4. A key difficulty is that the SquareCB/IGW analysis in Section E controls the variance of plug-in gaps through the pool size $|\mathcal{A}_t|$ (via the constant ϑ in (12)). When $\mathcal{A}_t$ can be large, this produces loose worst-case bounds even if the *contrast geometry* is low-rank. Optimal experimental design (OED) sharpens this step by choosing an exploration distribution that makes *all* pool contrasts well-conditioned, so variance scales with a rank parameter rather than with $|\mathcal{A}_t|$.

Notation. Recall the notation of Algorithm 1 and Section E. At round t, let

$$\hat{a}_t \in \arg\max_{a\in\mathcal{A}_t} f_{t-1}(x_t, a), \qquad \hat{\Delta}_t(a) := f_{t-1}(x_t, \hat{a}_t) - f_{t-1}(x_t, a)$$

be the greedy action and plug-in gaps from (10). Let $a_t \sim \pi_t$ be the played action, and let $a_t^\star \in \arg\max_{a\in\mathcal{A}_t} \mu(x_t, a)$ be an optimal action *within the current pool*.

Optimal design on contrast directions. The object that OED conditions is the set of *contrast feature vectors* induced by the greedy action, for which we subsequently introduce the shorthand notation:

$$u_t(a) := z_{x_t}(\hat{a}_t, a) \in \mathbb{R}^d, \qquad a \in \mathcal{A}_t. \tag{59}$$

For any distribution $q \in \Delta(\mathcal{A}_t)$, define the (regularized) information matrix and the associated G-optimal width

$$M_t(q) := \lambda I + \sum_{a\in\mathcal{A}_t} q(a)\, u_t(a)u_t(a)^\top, \qquad \kappa_t(q) := \max_{a\in\mathcal{A}_t} u_t(a)^\top M_t(q)^{-1} u_t(a). \tag{60}$$

Definition 2 (Optimal experimental design)**.** *Let $\mathcal{U} = \{u(a) : a \in \mathcal{A}\} \subset \mathbb{R}^d$ be a finite set of vectors and let $p \in \rho(\mathcal{A})$. Define the (regularized) design matrix Kiefer & Wolfowitz (1960); Kiefer (1974)*

$$V(p) \coloneqq \lambda I + \sum_{a \in \mathcal{A}} p(a)\, u(a) u(a)^\top.$$

The G-optimal *and* A design *values, respectively, are*

$$\kappa^\star \coloneqq \inf_{p \in \rho(\mathcal{A})} \max_{a \in \mathcal{A}} \|u(a)\|^2_{V(p)^{-1}}, \qquad \tau^\star \coloneqq \inf_{p \in \rho(\mathcal{A})} \operatorname{tr}\big(V(p)^{-1}\big).$$

We say that p is an α-approximate G-optimal design if $\max_{a \in \mathcal{A}} \|u(a)\|^2_{V(p)^{-1}} \le (1+\alpha)\kappa^\star$, and an α-approximate A-optimal design if $\operatorname{tr}(V(p)^{-1}) \le (1+\alpha)\tau^\star$.

Assumption 6 (Design oracle)**.** *There exists a procedure that, given a finite design set $\mathcal{U}$ and regularization $\lambda \ge 0$, returns an α-approximate G-optimal design (or A-optimal design) in the sense of Definition 2. We write $C_{\mathrm{opt}} \coloneqq 1 + \alpha$.*

Let $\kappa_t^\star \coloneqq \inf_{q \in \rho(\mathcal{A}_t)} \kappa_t(q)$. Assumption 6 returns a C_{opt}-approximate design q_t with $C_{\mathrm{opt}} \coloneqq 1+\alpha$:

$$\kappa_t(q_t) \le C_{\mathrm{opt}}\, \kappa_t^\star. \tag{61}$$

When the contrast directions $\{u_t(a) : a \in \mathcal{A}_t\}$ lie in an r_t-dimensional subspace, classical G-optimal design theory implies $\kappa_t^\star \le r_t$ (and equality when $\lambda = 0$); see, e.g., (Lattimore & Szepesvári, 2020, Chapter 21). Under Assumption 2, $r_t \le r$, and thus

$$\kappa_t(q_t) \le C_{\mathrm{opt}}\, r. \tag{62}$$

This is a consequence of the classical results on optimal experimental design (Kiefer & Wolfowitz, 1960; Kiefer, 1974), and is the point at which the low-rank contrast structure enters the minimax analysis: it lets OED replace a $|\mathcal{A}_t|$-dependence by a rank dependence.

Design-weighted IGW without Dirac notation. For a given gap coefficient $\gamma > 0$ associated with IGW, as defined in Sec. 3.3, it turns out to be useful for the analysis to define the rescaled gap coefficient:

$$\tilde{\gamma} \coloneqq \frac{\gamma}{C_{\mathrm{opt}} r}. \tag{63}$$

Given the design q_t, define $\lambda_t \in [\frac{1}{2}, 1]$ as the (unique) solution to

$$\frac{1}{2\lambda_t} + \sum_{a \ne \hat{a}_t} \frac{q_t(a)}{2(\lambda_t + \tilde{\gamma}\, \hat{\Delta}_t(a))} = 1. \tag{64}$$

Then define the sampling distribution $p_t \in \rho(\mathcal{A}_t)$ by

$$p_t(\hat{a}_t) \coloneqq \frac{1}{2\lambda_t}, \qquad p_t(a) \coloneqq \frac{q_t(a)}{2(\lambda_t + \tilde{\gamma}\, \hat{\Delta}_t(a))} \quad (a \ne \hat{a}_t). \tag{65}$$

Equation (64) ensures $\sum_{a \in \mathcal{A}_t} p_t(a) = 1$, and in particular $p_t(\hat{a}_t) = 1/(2\lambda_t) \ge 1/2$. This is algebraically equivalent to the common description "mix q_t with a point mass on $\hat{a}_t$ and normalize," but (65) avoids introducing Dirac measure notation. Relative to writing the same construction as an explicit mixture with a point mass, (65) only changes constants by at most a factor of 2 and keeps the normalizer λ_t in the convenient range $[1/2, 1]$.

Decision–estimation coefficient. Following Zhu et al. (2022), for any $\hat{f} \in \mathcal{F}$ and any context x, define the decision–estimation coefficient (see also (Wagenmaker & Foster, 2023))

$$\mathrm{dec}_{\tilde{\gamma}}(\mathcal{F}; \hat{f}, x) \coloneqq \inf_{p \in \rho(\mathcal{A})} \sup_{f \in \mathcal{F}} \sup_{a^\star \in \mathcal{A}} \mathbb{E}_{a \sim p}\Big[f(x, a^\star) - f(x, a) - \tilde{\gamma}\big(\hat{f}(x, a) - f(x, a)\big)^2 \Big]. \tag{66}$$

In our application $x = x_t$ and $\mathcal{A} = \mathcal{A}_t$ is the current pool. The next result is very similar to prior art, except that we are able to replace the combinatorial dependence $|\mathcal{A}|$ by intrinsic contrast rank r.

Lemma 6 (Variance domination for OED–IGW)**.** *Fix a round t and abbreviate $\hat{a} = \hat{a}_t$, $\hat{\Delta}(a) = \hat{\Delta}_t(a)$, and $u(a) = u_t(a)$. Let q_t be the design distribution and let p_t be defined by (65). Define the sub-probability measure*

$$\check{q}_t(a) \;:=\; \frac{q_t(a)}{2\big(\lambda_t + \tilde{\gamma}\,\hat{\Delta}(a)\big)} \qquad (a \neq \hat{a}),$$

and $\check{q}_t(\hat{a}) = 0$, where $\hat{a}$ is the best action within the pool. For any (sub-)distribution ν over $\mathcal{A}_t$, define the (unregularized) sampling covariance

$$V(\nu) \;:=\; \sum_{a \in \mathcal{A}_t} \nu(a)\, u(a) u(a)^\top.$$

Define also the design-weighted *covariance*

$$\widetilde{V}(q_t) \;:=\; \sum_{a \in \mathcal{A}_t \setminus \{\hat{a}\}} \frac{q_t(a)}{\lambda_t + \tilde{\gamma}\,\hat{\Delta}(a)}\, u(a) u(a)^\top. \tag{67}$$

(Equivalently, if $\tilde{u}(a) := u(a)/\sqrt{\lambda_t + \tilde{\gamma}\,\hat{\Delta}(a)}$, then $\widetilde{V}(q_t) = \sum_{a \neq \hat{a}} q_t(a)\, \tilde{u}(a)\tilde{u}(a)^\top$; i.e. the design oracle may be viewed as operating on the reweighted features $\{\tilde{u}(a)\}$.)

Then the following hold:

1. *$V(p_t) = V(\check{q}_t)$ (in particular $V(p_t) \succeq V(\check{q}_t)$).*
2. *$V(\check{q}_t) = \frac{1}{2}\,\widetilde{V}(q_t)$, and hence $V(p_t) = \frac{1}{2}\,\widetilde{V}(q_t)$.*
3. *Consequently, on $\mathrm{span}\{u(a) : a \in \mathcal{A}_t\}$ we have,*

$$V(p_t)^{-1} \preceq 2\,\widetilde{V}(q_t)^{-1}\,,$$

where we interpret this expression as restricted to the range where both matrices are invertible. Moreover, for any ridge $\lambda > 0$, defining

$$M(\nu) := \lambda I + V(\nu), \qquad \widetilde{M}(q_t) := \lambda I + \widetilde{V}(q_t),$$

we have $M(p_t) \succeq \frac{1}{2}\,\widetilde{M}(q_t)$ and hence

$$M(p_t)^{-1} \preceq 2\,\widetilde{M}(q_t)^{-1} \qquad \textit{(as PSD matrices on } \mathbb{R}^d\textit{)}.$$

Proof. By the definition of p_t in (65), for every $a \neq \hat{a}$ we have $p_t(a) = \check{q}_t(a)$, and $p_t(\hat{a}) = 1 - \sum_{b \neq \hat{a}} \check{q}_t(b)$. Since $u(\hat{a}) = u_t(\hat{a}) = z_{x_t}(\hat{a}_t, \hat{a}_t) = 0$, it follows that

$$V(p_t) = \sum_{a \neq \hat{a}} p_t(a)\, u(a) u(a)^\top = \sum_{a \neq \hat{a}} \check{q}_t(a)\, u(a) u(a)^\top = V(\check{q}_t),$$

proving item 1.

Next, expanding the definition of $\check{q}_t$ gives

$$V(\check{q}_t) = \sum_{a \neq \hat{a}} \frac{q_t(a)}{2(\lambda_t + \tilde{\gamma}\,\hat{\Delta}(a))}\, u(a) u(a)^\top = \frac{1}{2} \sum_{a \neq \hat{a}} \frac{q_t(a)}{\lambda_t + \tilde{\gamma}\,\hat{\Delta}(a)}\, u(a) u(a)^\top = \frac{1}{2}\,\widetilde{V}(q_t),$$

which is item 2 (and together with item 1 yields $V(p_t) = \frac{1}{2}\,\widetilde{V}(q_t)$).

For item 3, note that $V(p_t) = \frac{1}{2}\,\widetilde{V}(q_t)$ implies the two PSD matrices have the same range, namely $\mathrm{span}\{u(a) : a \in \mathcal{A}_t\}$. Restricting to this subspace where both operators are invertible, scaling gives $V(p_t)^{-1} = 2\,\widetilde{V}(q_t)^{-1}$, hence $V(p_t)^{-1} \preceq 2\,\widetilde{V}(q_t)^{-1}$ on the span.

Finally, for the ridge-regularized matrices,

$$M(p_t) = \lambda I + V(p_t) = \lambda I + \tfrac{1}{2}\,\widetilde{V}(q_t) = \tfrac{1}{2}\big(\widetilde{M}(q_t) + \lambda I\big) \succeq \tfrac{1}{2}\,\widetilde{M}(q_t),$$

and PSD inverse monotonicity yields $M(p_t)^{-1} \preceq 2\,\widetilde{M}(q_t)^{-1}$. □

Lemma 7 (OED–IGW certifies a minimax DEC bound)**.** *Assume* (62) *(which is valid under Assumption 6) and define* p_t *by* (65) *with* $\tilde{\gamma}$ *as in* (63)*. Then for every* t*,*

$$\sup_{f\in\mathcal{F}} \sup_{a^\star\in\mathcal{A}_t} \mathbb{E}_{a\sim p_t}\Big[f(x_t,a^\star) - f(x_t,a) - \tilde{\gamma}\big(f_{t-1}(x_t,a) - f(x_t,a)\big)^2\Big] \le \frac{2C_{\mathrm{opt}} r}{\tilde{\gamma}}. \tag{68}$$

Consequently, $\mathrm{dec}_{\tilde{\gamma}}(\mathcal{F}; f_{t-1}, x_t) \le 2C_{\mathrm{opt}} r/\tilde{\gamma}$.

Proof of Lemma 7. Fix t and abbreviate $x = x_t$, $\hat{a} = \hat{a}_t$, and $\hat{f} = f_{t-1}$. Write $a^\star$ for the maximizer of $f^\star(x,\cdot)$. By definition,

$$\mathrm{dec}_{\tilde{\gamma}}(\mathcal{F}; \hat{f}, x) = \inf_{p\in\Delta(\mathcal{A}_t)} \sup_{a^\star\in\mathcal{A}_t} \sup_{f^\star\in\mathcal{F}} \mathbb{E}_{a\sim p}\Big[f^\star(x,a^\star) - f^\star(x,a) - \tilde{\gamma}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2\Big].$$

We bound the quantity for the specific $p = p_t$ from (65).

Step 1: Decomposition. Add and subtract $\hat{f}$ and group terms:

$$f^\star(x,a^\star) - f^\star(x,a) = \underbrace{\hat{f}(x,\hat{a}) - \hat{f}(x,a)}_{\hat{\Delta}(a)} + \underbrace{\big(f^\star(x,a^\star) - \hat{f}(x,a^\star)\big)}_{T_3} + \underbrace{\big(\hat{f}(x,a) - f^\star(x,a)\big)}_{T_2} + \underbrace{\big(\hat{f}(x,a^\star) - \hat{f}(x,\hat{a})\big)}_{T_4}.$$

Taking expectation under p_t yields

$$\mathbb{E}_{a\sim p_t}\big[f^\star(x,a^\star) - f^\star(x,a)\big] = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \mathbb{E}_{a\sim p_t}\hat{\Delta}(a), \quad T_2 = \mathbb{E}_{a\sim p_t}\big[\hat{f}(x,a) - f^\star(x,a)\big],$$
$$T_3 = f^\star(x,a^\star) - \hat{f}(x,a^\star), \quad T_4 = \hat{f}(x,a^\star) - \hat{f}(x,\hat{a}) \le 0.$$

(The inequality uses $\hat{a} = \arg\max_a \hat{f}(x,a)$.)

Step 2: Bound T_1. By the definition of p_t,

$$T_1 = \sum_{a\neq\hat{a}} \frac{q_t(a)}{2(\lambda_t + \tilde{\gamma}\hat{\Delta}(a))}\,\hat{\Delta}(a) \le \sum_{a\neq\hat{a}} \frac{q_t(a)}{2\tilde{\gamma}} \le \frac{1}{2\tilde{\gamma}}.$$

Step 3: AM–GM on T_2. By AM–GM,

$$T_2 \le \frac{\tilde{\gamma}}{2}\,\mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2 + \frac{1}{2\tilde{\gamma}}.$$

Step 4: Control T_3 via variance domination and design width (where Assumption 6 helps). Using Cauchy–Schwarz and AM–GM in the standard self-normalized form, which is valid under contrast realizability (Assumption 1),

$$T_3 = f^\star(x,a^\star) - \hat{f}(x,a^\star) \le \frac{1}{2\tilde{\gamma}} \|u_t(a^\star)\|^2_{\tilde{V}(p_t)^{-1}} + \frac{\tilde{\gamma}}{2} \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2.$$

By Lemma 6 and the assumed design-width bound $\kappa_t(q_t) \le C_{\mathrm{opt}} r$ [cf. (62)] which holds true via Assumption 6,

$$\|u_t(a^\star)\|^2_{\tilde{V}(p_t)^{-1}} \le 2\,\|u_t(a^\star)\|^2_{\tilde{V}(q_t)^{-1}} \le 2\,\kappa_t(q_t) \le 2C_{\mathrm{opt}} r.$$

Hence,

$$T_3 \le \frac{C_{\mathrm{opt}} r}{\tilde{\gamma}} + \frac{\tilde{\gamma}}{2} \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2.$$

Step 5: Combine and cancel. Summing T_1–T_4 (and using $T_4 \le 0$) yields

$$\mathbb{E}_{a\sim p_t}\big[f^\star(x,a^\star) - f^\star(x,a)\big] \le \frac{2C_{\mathrm{opt}} r}{\tilde{\gamma}} + \tilde{\gamma}\, \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2,$$

which is exactly (68). □

It turns out to be useful for the minimax analysis to consider a slightly different error decomposition of the regret than in the vanilla high probability stochastic regret setting [cf. (34)]. We present this decomposition next as a lemma.

Lemma 8 (One-step decomposition)**.** *Fix t and condition on $\mathcal{F}_{t-1}$ and x_t. Let $\hat{a}_t \in \arg\max_{a\in\mathcal{A}_t} f_{t-1}(x_t,a)$ be the predicted best action in the current pool, and define the plug-in gaps $\hat{\Delta}_t(a) := f_{t-1}(x_t,\hat{a}_t) - f_{t-1}(x_t,a)$ for $a \in \mathcal{A}_t$. Then for any chosen action distribution $\pi_t(\cdot \mid x_t)$ supported on $\mathcal{A}_t$,*

$$\mathbb{E}[\mu(x_t,a_t^\star) - \mu(x_t,a_t) \mid \mathcal{F}_{t-1}, x_t] \le \underbrace{\big(\mu(x_t,a_t^\star) - \mu(x_t,\hat{a}_t)\big)}_{\textit{span/approximation}} \tag{69}$$

$$+ \underbrace{\mathbb{E}\Big[\hat{\Delta}_t(a_t) \mid \mathcal{F}_{t-1}, x_t\Big]}_{\textit{exploitation controlled by IGW}} + \underbrace{2 \sup_{a\in\mathcal{A}_t} \big|f_{t-1}(x_t,a) - \mu(x_t,a)\big|}_{\textit{estimation}}. \tag{70}$$

Proof. Begin by adding and subtracting $\mu(x_t,\hat{a}_t)$:

$$\mu(x_t,a_t^\star) - \mu(x_t,a_t) = \big(\mu(x_t,a_t^\star) - \mu(x_t,\hat{a}_t)\big) + \big(\mu(x_t,\hat{a}_t) - \mu(x_t,a_t)\big).$$

Next, consider the grouped terms inside the second parentheses of the preceding line, to which we add and subtract f_{t-1}:

$$\mu(x_t,\hat{a}_t) - \mu(x_t,a_t) = \underbrace{f_{t-1}(x_t,\hat{a}_t) - f_{t-1}(x_t,a_t)}_{=\hat{\Delta}_t(a_t)} + \big(\mu(x_t,\hat{a}_t) - f_{t-1}(x_t,\hat{a}_t)\big) + \big(f_{t-1}(x_t,a_t) - \mu(x_t,a_t)\big).$$

Taking conditional expectations and upper bounding the last two terms by the uniform deviation $\sup_{a\in\mathcal{A}_t} |f_{t-1}(x_t,a) - \mu(x_t,a)|$ gives (70). □

The lemma isolates three requirements for sublinear regret: (i) a *span adequacy* bound for the pool (term "span/approximation"), (ii) a *variance/efficiency* bound for IGW (term "exploitation"), which exploits the OED reweighting (Kiefer & Wolfowitz, 1960; Kiefer, 1974), and (iii) a way to control the prediction error of f_{t-1} on actions actually relevant to the policy (term "estimation") using only bandit feedback.

Next, we establish that under these conditions, with design-weighting, Algorithm 1 can achieve sublinear regret against an adaptively or adversarially chosen sequence of noise/reward/context triples. That is, for any class of problems, possibly adversarially chosen, satisfying Assumption 4.[3] To ease the notation, we omit computing an explicit sup over such noise and context distributions.

Theorem (Restatement of Theorem 2): High-probability minimax regret for ContrastSpanner)**.** *Under Assumptions 2, 3.1, 4, and 5, run Algorithm 1 with the G-optimal design distribution p_t defined in (65) at each round satisfying (Assumption 6) Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for any class of noise processes satisfying Assumptions 4, and any possibly adaptive/adversarially chosen series of contexts x_t, we have the following regret bound*

$$\begin{aligned}\mathrm{Reg}(T) \le B\sum_{t=1}^{T}\epsilon_t + \frac{2C_{\mathrm{opt}} r\, T}{\gamma} + 4\gamma\, \mathsf{R}_{\mathrm{sq}}(T) \\ + 32\gamma\,\sigma^2 \log\frac{6}{\delta} + \frac{88}{3}\gamma\, R_{\max}^2 \log\frac{6}{\delta} + 2R_{\max}\sqrt{2T\log\frac{6}{\delta}}.\end{aligned} \tag{71}$$

Moreover, setting

$$\gamma = \sqrt{\frac{C_{\mathrm{opt}} r\, T}{2\mathsf{R}_{\mathrm{sq}}(T) + 16\sigma^2\log(6/\delta) + \frac{44}{3}R_{\max}^2\log(6/\delta)}}$$

balances the leading γ-dependent terms in (71) and yields the minimax rate

$$\mathrm{Reg}(T) = \mathcal{O}\left(B\sum_{t=1}^{T}\epsilon_t \;+\; r\sqrt{T\left(\mathsf{R}_{\mathrm{sq}}(T) + \sigma^2\log\tfrac{1}{\delta} + R_{\max}^2\log\tfrac{1}{\delta}\right)} \;+\; R_{\max}\sqrt{T\log\tfrac{1}{\delta}}\right).$$

Proof. By Lemma 8, it suffices to bound the within-pool regret $\sum_{t=1}^{T}\big(\mu(x_t,a_t^\star) - \mu(x_t,a_t)\big)$ and then add the approximation term $B\sum_{t=1}^{T}\epsilon_t$.

Fix $\gamma > 0$. Apply Lemma 7 with $f = \mu$ and $\hat{f} = f_{t-1}$. Conditioning on $\mathcal{F}_{t-1}\cup\{x_t\}$ and using realizability $\mu \in \mathcal{F}$ (Assumption 5) gives (with using the shorthand $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}, x_t]$)

$$\mathbb{E}_t[\mu(x_t,a_t^\star) - \mu(x_t,a_t)] \le \frac{2C_{\mathrm{opt}} r}{\gamma} + \gamma\,\mathbb{E}_t\left[(f_{t-1}(x_t,a_t) - \mu(x_t,a_t))^2\right]. \tag{72}$$

Summing (72) over $t \le T$ yields

$$\sum_{t=1}^{T}\mathbb{E}_t\Big[\mu(x_t,a_t^\star) - \mu(x_t,a_t)\Big] \le \frac{2C_{\mathrm{opt}} r\, T}{\gamma} + \gamma\sum_{t=1}^{T}\mathbb{E}_t\Big[(f_{t-1}(x_t,a_t) - \mu(x_t,a_t))^2\Big]. \tag{73}$$

Step 3: A high-probability bound from the conditional bound. Define the martingale difference sequence

$$D_t := \Big(\mu(x_t,a_t^\star) - \mu(x_t,a_t)\Big) - \mathbb{E}_t[\mu(x_t,a_t^\star) - \mu(x_t,a_t)]\,.$$

Then $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}, x_t] = 0$ and D_t is adapted to $(\mathcal{F}_{t-1}, x_t)_{t\ge1}$. Moreover, Assumption 4 implies $|\mu(x,a)| \le R_{\max}$ for all (x,a), so and since $a_t^\star \in \arg\max_{a\in\mathcal{A}_t}\mu(x_t,a)$ we have $0 \le \mu(x_t,a_t^\star) - \mu(x_t,a_t) \le 2R_{\max}$, so $|D_t| \le 2R_{\max}$ almost surely. Therefore, by the Hoeffding–Azuma inequality for martingale differences (see, e.g., Lattimore & Szepesvári (2020)), with probability at least $1-\delta/3$,

$$\sum_{t=1}^{T}\Big(\mu(x_t,a_t^\star) - \mu(x_t,a_t)\Big) \le \sum_{t=1}^{T}\mathbb{E}[\mu(x_t,a_t^\star) - \mu(x_t,a_t) \mid \mathcal{F}_{t-1}, x_t] + 2R_{\max}\sqrt{2T\log\frac{3}{\delta}}. \tag{74}$$

[3] This is not fully minimax style in the sense of distribution-free, adversarial settings that are often studied, but rather "minimax-style" within a class of noise processes and context distributions.

Step 4: Controlling the predictable squared-error term. Let $e_t := f_{t-1}(x_t, a_t) - \mu(x_t, a_t)$ and $X_t := e_t^2$. Since $|f_{t-1}(x_t, a)| \le R_{\max}$ by the clipping step and $|\mu(x_t, a)| \le R_{\max}$ by Assumption 4, we have $0 \le X_t \le 4R_{\max}^2$ almost surely. Write $\bar{X}_t := \mathbb{E}[X_t \mid \mathcal{H}_{t-1}]$ and define the martingale differences $Z_t := \bar{X}_t - X_t$, so that $\sum_{t=1}^T \bar{X}_t = \sum_{t=1}^T X_t + \sum_{t=1}^T Z_t$ and $\mathbb{E}[Z_t \mid \mathcal{H}_{t-1}] = 0$, where $\mathcal{H}_{t-1} = \mathcal{F}_t \cup \{x_t\}$.

Apply Freedman's inequality (Theorem 20.1 in Lattimore & Szepesvári (2020)) to (Z_t) with the bound $|Z_t| \le 4R_{\max}^2$. The predictable quadratic variation satisfies

$$\mathbb{E}[Z_t^2 \mid \mathcal{H}_{t-1}] \le \mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}] \le (4R_{\max}^2)\, \mathbb{E}[X_t \mid \mathcal{H}_{t-1}] = (4R_{\max}^2)\, \bar{X}_t,$$

where we used $X_t^2 \le (4R_{\max}^2) X_t$ since $0 \le X_t \le 4R_{\max}^2$. Let $Y := \sum_{t=1}^T \bar{X}_t$ and $S := \sum_{t=1}^T X_t$. Freedman's inequality therefore yields that with probability at least $1 - \delta/3$,

$$Y - S = \sum_{t=1}^T Z_t \le \sqrt{2(4R_{\max}^2)\, Y \log \frac{3}{\delta}} + \frac{4R_{\max}^2}{3} \log \frac{3}{\delta}.$$

Rearranging gives

$$Y \le S + \sqrt{8R_{\max}^2\, Y \log \frac{3}{\delta}} + \frac{4R_{\max}^2}{3} \log \frac{3}{\delta}.$$

Let $b := 4R_{\max}^2$ and $L := \log(3/\delta)$, and set $t := \sqrt{Y}$. Then the previous display is equivalent to $t^2 \le S + \sqrt{2bL}\, t + \frac{b}{3} L$, which implies $t \le \sqrt{2bL} + \sqrt{S + \frac{b}{3} L}$. Using $(u+v)^2 \le 2u^2 + 2v^2$ with $u = \sqrt{2bL}$ and $v = \sqrt{S + \frac{b}{3} L}$ yields the convenient bound

$$\sum_{t=1}^T \mathbb{E}[e_t^2 \mid \mathcal{H}_{t-1}] = Y \le 2S + \frac{14}{3} bL = 2 \sum_{t=1}^T e_t^2 + \frac{56}{3} R_{\max}^2 \log \frac{3}{\delta}. \tag{75}$$

Finally, apply Lemma 4 with failure probability $\delta/3$ to the (adaptive) sequence (x_t, a_t). With probability at least $1 - \delta/3$, it gives

$$\sum_{t=1}^T e_t^2 \le 2\, \mathsf{R}_{\mathrm{sq}}(T) + 16\sigma^2 \log \frac{6}{\delta} + \frac{16}{3} R_{\max}^2 \log \frac{6}{\delta}.$$

Combining this with (75) and the inequality $\log(3/\delta) \le \log(6/\delta)$ yields

$$\sum_{t=1}^T \mathbb{E}[e_t^2 \mid \mathcal{H}_{t-1}] \le 4\, \mathsf{R}_{\mathrm{sq}}(T) + 32\sigma^2 \log \frac{6}{\delta} + \frac{88}{3} R_{\max}^2 \log \frac{6}{\delta}. \tag{76}$$

Step 5: Combine the pieces. On the intersection of the three high-probability events used above (each failing with probability at most $\delta/3$), we may plug (76) into (73) and then use (74) to obtain

$$\sum_{t=1}^T \Big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\Big) \le \frac{2C_{\mathrm{opt}} r\, T}{\gamma} + 4\gamma\, \mathsf{R}_{\mathrm{sq}}(T) + 32\gamma\, \sigma^2 \log \frac{6}{\delta} + \frac{88}{3} \gamma\, R_{\max}^2 \log \frac{6}{\delta} + 2R_{\max} \sqrt{2T \log \frac{3}{\delta}}.$$

Using $\log(3/\delta) \le \log(6/\delta)$ and adding the approximation term $B \sum_{t=1}^T \epsilon_t$ (from Lemma 8) completes the proof of (71). Combining this inequality with (73) and (74), and taking a union bound over the two events, gives the within-pool part of (71). Finally, add $B \sum_{t=1}^T \epsilon_t$ from Lemma 8. □

Remark 7 (Relation to SpannerIGW and what's new)**.** *The structure of Lemma 7 and Theorem 2 follows the minimax analysis of SpannerIGW Zhu et al. (2022). The key substitutions are: (i) their action-embedding dimension is replaced by the* contrast rank r *(the dimension of the span of learned contrast directions), and (ii) their spanner is built from an action embedding, whereas ours is built from causal contrast features $z_x(a, a')$. The proof itself is unchanged once one views the pool through the lens of the contrast directions $u_t(a) = z_{x_t}(\hat{a}_t, a)$: optimal design bounds the largest Mahalanobis norm of these directions and therefore controls the variance of the plug-in gap estimator under importance weighting.*

This also clarifies the stochastic-versus-minimax comparison. In Section E, the IGW/SquareCB analysis controls variance in a cardinality-based *way through the pool size $|\mathcal{A}_t|$ (via ϑ) and yields gap-dependent improvements whenever $\hat{\Delta}_t$ is accurate. In the minimax setting we cannot rely on such favorable structure: gaps can be tiny, and early gap estimates can be noisy. OED mitigates this by choosing q_t so that* every *contrast direction has bounded leverage, which is exactly what is needed to make the negative square-loss term in (68) dominate the estimation error.*

Remark on gap-weighted designs (instance dependence). *SpannerIGW refines the above minimax argument into instance-dependent rates by applying optimal design to a gap-weighted geometry (sometimes described as an "embedding reweighting"), downweighting actions with large predicted gaps as in ReweightedSpanner in (Zhu et al., 2022). An analogous extension is possible here by designing over suitably scaled contrasts $z_{x_t}(\hat{a}_t, a)/\sqrt{1 + \tilde{\gamma}\,\hat{\Delta}_t(a)}$. We omit this refinement in the main analysis and focus on the minimax guarantee (71). This would operationally be akin to trying to merge together the problem with causal discovery with regret minimization through an instance-dependent mechanism. We defer this to future work.*

E.4 Log-loss oracle: stochastic and minimax regret

This section analyzes FASTCONTRASTSPANNER (Algorithm 2), which replaces the square-loss oracle in CONTRASTSPANNER with a log-loss oracle that predicts a categorical distribution over outcomes. The analysis mirrors the square-loss case in Section E–E.3, but the key technical step is now to convert the log-loss oracle guarantee into a *high-probability* bound on squared *mean-reward* prediction error, using (i) triangular discrimination as a curvature proxy (Topsoe, 2002; Le Cam, 2012) and (ii) a Freedman inequality for martingale difference sequences.

Categorical outcome model and mean reward. Fix an outcome alphabet $[K] := \{1, 2, \ldots, K\}$. For each context–action pair (x, a), let $q^\star(\cdot \mid x, a) \in \rho([K])$ denote the true conditional distribution of the observed categorical outcome $Y \in [K]$.

$$r_t := r(Y_t), \qquad \text{with} \qquad |r(k)| \le R_{\max} \ \text{ for all } k \in [K].$$

At round t, after choosing a_t, the learner observes an outcome $Y_t \in [K]$ and receives reward $R_t := r(Y_t)$. Here we denote an outcome alphabet as $[K] := \{1, 2, \ldots, K\}$, i.e., a collection of K labels in multi-class classification. Let $q^\star(\cdot \mid x, a) \in \rho([K])$ denote the true conditional law of Y under $do(a)$, and define the interventional mean reward

$$\mu(x, a) := \mathbb{E}_{Y \sim q^\star(\cdot \mid x, a)}[r(Y)]. \tag{77}$$

This convention reconciles the fact that the observation is categorical while the bandit objective is scalar; it also ensures Assumption 4 (bounded martingale noise for the scalar reward) remains applicable.

Assumption 7 (Clipping, log-loss realizability, and curvature)**.** *There exists a clipping level $\nu \in (0, 1/K]$ such that:*

1. ***(Clipped model)*** *The true conditional distributions satisfy $q^\star(k \mid x, a) \ge \nu$ for all x, a, k. Moreover, the log-loss oracle (Assumption 3.2) returns predictors q_t satisfying $q_t(k \mid x, a) \ge \nu$ for all*

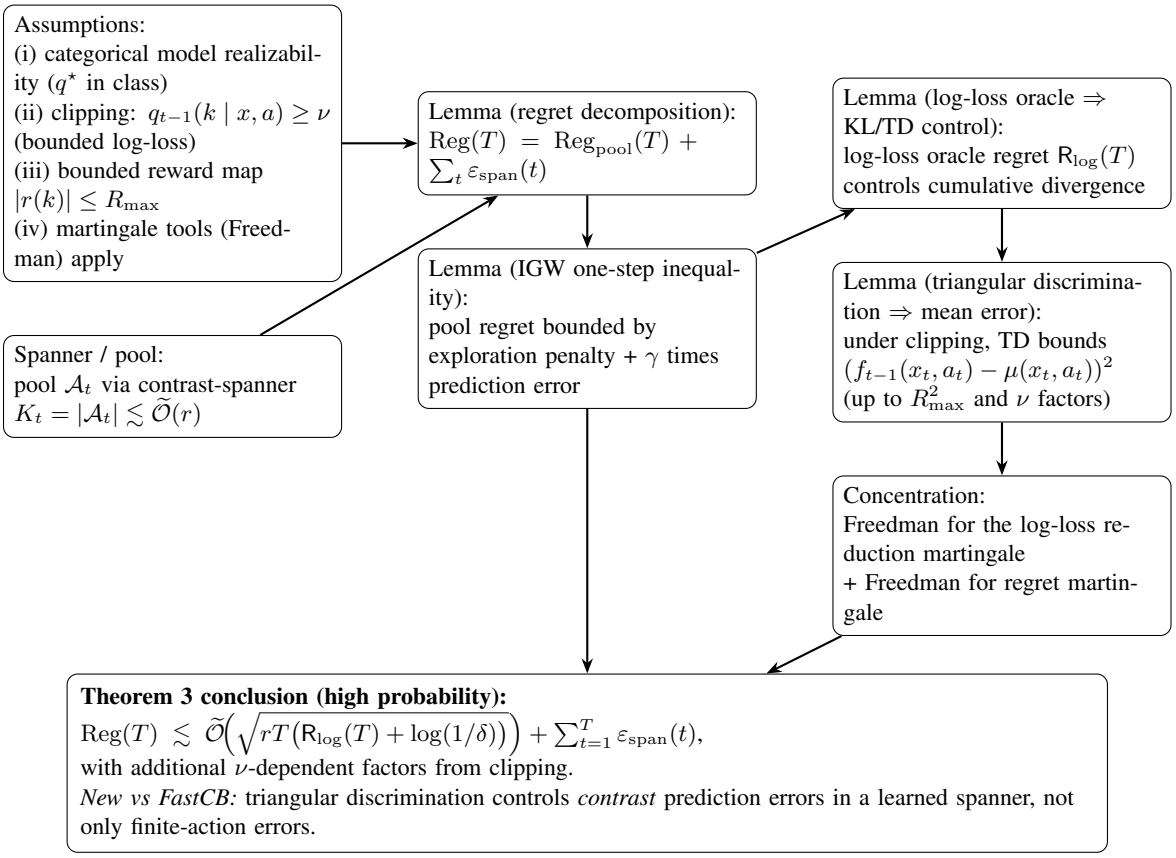


Figure 10: Proof logic flowchart for Theorem 3 (log-loss oracle, stochastic high-probability regret via triangular discrimination).

x, a, k *(equivalently, we post-process the oracle output by clipping to* $[\nu, 1]$ *and renormalizing). In particular, for all* t*,*

$$0 \leq \ell(q_{t-1}; x_t, a_t, Y_t) \leq \log(1/\nu) \qquad \textit{and} \qquad 0 \leq \ell(q^\star; x_t, a_t, Y_t) \leq \log(1/\nu). \tag{78}$$

2. ***(Realizability comparator)*** *There exists a distribution predictor* $q^\star(\cdot \mid x, a)$ *in the log-loss oracle's comparator class such that for all rounds* t*,*

$$\mathbb{P}\big(Y_t = k \,\big|\, \mathcal{F}_{t-1},\, x_t,\, a_t\big) = q^\star(k \mid x_t, a_t) \qquad \textit{for all } k \in [K].$$

3. ***(Fisher curvature / strong convexity proxy)*** *There exists* $\kappa > 0$ *such that for all* x, a *and all distributions* $q(\cdot \mid x, a)$ *in the oracle class,*

$$\mathrm{KL}\big(q^\star(\cdot \mid x, a) \,\|\, q(\cdot \mid x, a)\big) \;\geq\; \frac{\kappa}{2} \big\|q^\star(\cdot \mid x, a) - q(\cdot \mid x, a)\big\|_2^2.$$

Log-loss oracle and clipping. For $q \in \rho([K])$ and outcome $y \in [K]$, define the log loss $\ell(q; y) := -\log q(y)$. In our contextual setting we write $\ell(q; x, a, y) := -\log q(y \mid x, a)$. The log-loss $\ell_{\log}(q; y) := -\log q(y)$ is unbounded if $q(y)$ can approach 0. Assumption 7.2 ensures bounded increments so we can apply Freedman-type concentration to the log-likelihood ratio martingales below. This is the same "curvature stability" issue highlighted in FastCB Foster & Krishnamurthy (2021) (see their discussion around Eq. (16) and Theorem 4).

Remark 8 (When is κ finite?)**.** *Assumption 7.3 is a* curvature *condition ensuring that log-loss excess risk controls an* ℓ_2*-distance between distributions. A universal choice is* $\kappa = 1$*, since Pinsker's*

inequality gives $\mathrm{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2 \geq \frac{1}{2}\|p-q\|_2^2$ *for all* $p, q \in \rho([K])$*. We explicitly retain* κ *because the log-loss literature often states curvature in terms of Fisher information / exp-concavity constants (which become* stable *under clipping), and because our minimax bounds track the dependence on this curvature parameter (e.g.* FastCB *(Foster & Krishnamurthy, 2021, Section 3) and information-geometry treatments such as Amari & Nagaoka (2000)).*

From a distribution predictor to a mean predictor. The log-loss oracle outputs a distribution predictor $q_t(\cdot \mid x, a)$. We convert it into a scalar mean-reward predictor by taking the reward expectation under q_t.

Definition 3 (Mean predictor induced by a distribution predictor)**.** *Given a distribution predictor* $q(\cdot \mid x, a) \in \rho([K])$*, define the induced mean predictor*

$$f_q(x, a) := \sum_{k=1}^{K} r(k)\, q(k \mid x, a). \tag{79}$$

For the oracle output q_{t-1} *we write* $f_{t-1} := f_{q_{t-1}}$.

Lemma 9 (Mean error from distribution error)**.** *For any* (x, a) *and any distribution predictor* $q(\cdot \mid x, a)$*,*

$$\left|f_q(x,a) - \mu(x,a)\right| \leq R_{\max} \left\|q(\cdot \mid x,a) - q^\star(\cdot \mid x,a)\right\|_1 \leq R_{\max}\sqrt{K} \left\|q(\cdot \mid x,a) - q^\star(\cdot \mid x,a)\right\|_2.$$

Proof. By definition (77) and (79),

$$f_q(x,a) - \mu(x,a) = \sum_{k=1}^{K} r(k)\big(q(k \mid x,a) - q^\star(k \mid x,a)\big).$$

Using $|r(k)| \leq R_{\max}$ and the triangle inequality,

$$|f_q(x,a) - \mu(x,a)| \leq \sum_{k=1}^{K} |r(k)| \left|q(k \mid x,a) - q^\star(k \mid x,a)\right| \leq R_{\max} \|q(\cdot \mid x,a) - q^\star(\cdot \mid x,a)\|_1.$$

The ℓ_1–ℓ_2 inequality $\|v\|_1 \leq \sqrt{K}\|v\|_2$ yields the second bound. □

Triangular discrimination and a variance proxy for log loss. We now introduce the divergence that plays the role of "squared loss" in log-loss analyses. This quantity, the triangular discrimination, plays a central role in the regret analysis of (Foster & Krishnamurthy, 2021), and permits a sharpening of the rates that are possible for square losses only. Related identities have gained traction in bandits and reinforcement learning in recent years (Ayoub et al.), and date back to (Topsoe, 2002; Le Cam, 2012)

Definition 4 (Triangular discrimination)**.** *For* $p, q \in \rho([K])$*, define the triangular discrimination (Topsoe, 2002; Le Cam, 2012)*

$$\mathrm{TD}(p, q) := \sum_{k=1}^{K} \frac{(p_k - q_k)^2}{p_k + q_k},$$

with the convention $0/0 := 0$.

Lemma 10 (Elementary comparisons: TD, ℓ_1, and KL)**.** *For all* $p, q \in \rho([K])$*:*

1. *(*ℓ_1 ***controlled by*** TD*)* $\|p-q\|_1^2 \leq 2\,\mathrm{TD}(p, q)$.
2. *(*TD ***controlled by*** KL*)* $\mathrm{KL}(p\|q) \geq \frac{1}{2}\,\mathrm{TD}(p, q)$.

Consequently, $\|p-q\|_1^2 \leq 4\,\mathrm{KL}(p\|q)$.

Proof. (1) By Cauchy–Schwarz,

$$\|p-q\|_1 = \sum_{k=1}^{K} |p_k - q_k| = \sum_{k=1}^{K} \sqrt{p_k+q_k} \cdot \frac{|p_k-q_k|}{\sqrt{p_k+q_k}}$$
$$\leq \Big(\sum_{k=1}^{K}(p_k+q_k)\Big)^{1/2} \Big(\sum_{k=1}^{K} \frac{(p_k-q_k)^2}{p_k+q_k}\Big)^{1/2}.$$

Since $\sum_k (p_k+q_k) = 2$, squaring gives $\|p-q\|_1^2 \leq 2\,\mathrm{TD}(p,q)$.

(2) Write $t_k := p_k/q_k$ (with the convention $0/0 := 1$ on coordinates where $p_k = q_k = 0$). Then $\mathrm{KL}(p\|q) = \sum_k q_k\, t_k \log t_k$. Consider the scalar function

$$\phi(t) := t\log t - (t-1) - \frac{(t-1)^2}{2(t+1)}, \qquad t > 0.$$

A direct calculus check shows $\phi(1) = 0$ and $\phi''(t) = \frac{1}{t} - \frac{4}{(t+1)^3} \geq 0$, hence ϕ is convex and minimized at $t = 1$, so $\phi(t) \geq 0$ for all $t > 0$. Therefore,

$$t\log t - (t-1) \geq \frac{(t-1)^2}{2(t+1)} \qquad \text{for all } t > 0.$$

Multiplying by q_k and summing over k yields

$$\mathrm{KL}(p\|q) = \sum_k q_k t_k \log t_k = \sum_k q_k\big(t_k \log t_k - (t_k-1)\big)$$
$$\geq \frac{1}{2}\sum_k q_k \frac{(t_k-1)^2}{t_k+1} = \frac{1}{2}\sum_k \frac{(p_k-q_k)^2}{p_k+q_k} = \frac{1}{2}\,\mathrm{TD}(p,q),$$

where we used $\sum_k q_k(t_k-1) = \sum_k (p_k - q_k) = 0$ to cancel the linear term. □

A high-probability conversion: log-loss regret ⇒ squared mean prediction error. We now prove the log-loss counterpart of Lemma 4 in the square-loss analysis. The key is to (i) express the cumulative TD divergence as the predictable part of a martingale and (ii) control the martingale via Freedman's inequality with a *variance proxy proportional to* TD, which is exactly where triangular discrimination enters (Topsoe, 2002; Le Cam, 2012).

For a context–action pair (x,a), define the per-round divergences

$$\mathrm{KL}_t(x,a) := \mathrm{KL}\big(q^\star(\cdot \mid x,a) \,\|\, q_{t-1}(\cdot \mid x,a)\big), \quad \mathrm{TD}_t(x,a) := \mathrm{TD}\big(q^\star(\cdot \mid x,a),\, q_{t-1}(\cdot \mid x,a)\big). \tag{80}$$

Also define the clipped log-loss range constant

$$b_\nu := \log(1/\nu)\,, \qquad C_\nu := b_\nu^2\Big(\frac{1+\nu}{1-\nu}\Big)^2. \tag{81}$$

Finally, let $\mathcal{F}_t$ denote the learner's filtration up to time t (as in Assumption 4). Next is the key lemma that reduces cross-entropy/log-loss, via operating upon mean predictions, to a gap in terms of the log regression oracle.

Lemma 11 (Log-loss oracle controls mean prediction error on an adaptive sequence)**.** *Assume Assumption 7. Fix any $\mathcal{F}_{t-1}, x_t, a_t$ as defined prior to Assumption 4 Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \big(f_{t-1}(x_t,a_t) - \mu(x_t,a_t)\big)^2 \leq 8R_{\max}^2\, R_{\log}(T) + \Big(32R_{\max}^2\, C_\nu + \frac{16}{3}R_{\max}^2\, b_\nu\Big)\log\Big(\frac{2}{\delta}\Big), \tag{82}$$

where $b_\nu = \log(1/\nu)$ and $C_\nu = b_\nu^2\big(\frac{1+\nu}{1-\nu}\big)^2$ are defined in (81).

Proof. **Step 1: Martingale decomposition of log-loss excess.** Define the log-loss excess random variable at time t:

$$X_t := \ell(q_{t-1}; x_t, a_t, Y_t) - \ell(q^\star; x_t, a_t, Y_t) = \log\Big(\frac{q^\star(Y_t \mid x_t, a_t)}{q_{t-1}(Y_t \mid x_t, a_t)}\Big).$$

Conditioning on $\mathcal{F}_{t-1}, x_t, a_t$ and using $Y_t \sim q^\star(\cdot \mid x_t, a_t)$, which holds under Assumption 7.2 regarding realizability,

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}, x_t, a_t] = \sum_{k=1}^{K} q^\star(k \mid x_t, a_t) \log\Big(\frac{q^\star(k \mid x_t, a_t)}{q_{t-1}(k \mid x_t, a_t)}\Big) = \mathrm{KL}_t(x_t, a_t). \tag{83}$$

Define the martingale difference sequence

$$\zeta_t := X_t - \mathrm{KL}_t(x_t, a_t), \qquad \text{so that} \qquad \mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}, x_t, a_t] = 0.$$

Summing over t gives the identity

$$\sum_{t=1}^{T} \mathrm{KL}_t(x_t, a_t) = \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \zeta_t. \tag{84}$$

By Assumption 3.2, the log-loss oracle regret bound with comparator $q^\star$ implies

$$\sum_{t=1}^{T} X_t = \sum_{t=1}^{T} \big(\ell(q_{t-1}; x_t, a_t, Y_t) - \ell(q^\star; x_t, a_t, Y_t)\big) \le R_{\log}(T) \qquad \text{almost surely.} \tag{85}$$

Step 2: A deterministic inequality linking mean-squared error to $\sum_t \mathrm{TD}_t$. Recall the definition of TD_t as in (80). By Lemma 9 and Lemma 10(1),

$$\big(f_{t-1}(x_t, a_t) - \mu(x_t, a_t)\big)^2 \le R_{\max}^2 \big\|q_{t-1}(\cdot \mid x_t, a_t) - q^\star(\cdot \mid x_t, a_t)\big\|_1^2 \le 2R_{\max}^2 \,\mathrm{TD}_t(x_t, a_t). \tag{86}$$

Summing over t and letting $S := \sum_{t=1}^{T} \mathrm{TD}_t(x_t, a_t)$ gives

$$\sum_{t=1}^{T} \big(f_{t-1}(x_t, a_t) - \mu(x_t, a_t)\big)^2 \le 2R_{\max}^2 \, S. \tag{87}$$

Step 3: Control S by combining curvature and Freedman. By Lemma 10(2), $\mathrm{TD}_t(x_t, a_t) \le 2\,\mathrm{KL}_t(x_t, a_t)$ [cf. (80)]. Therefore, we may write

$$S \le 2 \sum_{t=1}^{T} \mathrm{KL}_t(x_t, a_t). \tag{88}$$

Combining (84), (85), and (88) yields

$$S \le 2R_{\log}(T) - 2 \sum_{t=1}^{T} \zeta_t. \tag{89}$$

We proceed to construct a version of the difference sequence $M_T := \sum_{t=1}^{T} \zeta_t$ with $\zeta_t := X_t - \mathrm{KL}_t(x_t, a_t)$ so that we may apply a concentration bound. Proceed then by applying the clipping bounds in (78) (which are valid via Assumption 7.1), whereby we have $|X_t| \le b_\nu$ and $0 \le \mathrm{KL}_t(x_t, a_t) \le b_\nu$, so

$$|\zeta_t| = |X_t - \mathrm{KL}_t(x_t, a_t)| \le |X_t| + \mathrm{KL}_t(x_t, a_t) \le 2b_\nu. \tag{90}$$

Second, we bound the conditional variance via TD_t [cf. (80)]. Conditioning on $\mathcal{F}_{t-1}$,

$$\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t] \le \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t] = \sum_{k=1}^{K} q^\star(k \mid x_t, a_t)\, \log^2\Big(\frac{q^\star(k \mid x_t, a_t)}{q_{t-1}(k \mid x_t, a_t)}\Big).$$

Here we use the shorthand notation as $\log^2(j)$ simply means $(\log j)^2$, where $j > 0$ denotes the scalar argument of the logarithm, meaning the right-hand side of the preceding expression is the second moment of the log likelihood ratio with respect to distribution $q^\star$. Fix a coordinate k and abbreviate $p := q^\star(k \mid x_t, a_t)$ and $q := q_{t-1}(k \mid x_t, a_t)$. By clipping, $p, q \in [\nu, 1]$. Define $t := (p-q)/(p+q) \in [-\iota, \iota]$ with $\iota := (1-\nu)/(1+\nu)$ (where ν is as in (81)), so that $p/q = (1+t)/(1-t)$ and $\log(p/q) = g(t)$ where $g(t) := \log\big(\frac{1+t}{1-t}\big)$. The function g is odd and convex on $[0, 1)$, with $g(0) = 0$ and $g(\iota) = \log(1/\nu) = b_\nu$. Therefore, for $|t| \le \iota$,

$$|g(t)| \le \frac{g(\iota)}{\iota}|t| = \frac{b_\nu}{\iota}|t| = b_\nu\Big(\frac{1+\nu}{1-\nu}\Big)\frac{|p-q|}{p+q}. \tag{91}$$

Squaring (91) and multiplying by $p \le p + q$ yields the scalar bound

$$p\, \log^2\Big(\frac{p}{q}\Big) \le C_\nu\, \frac{(p-q)^2}{p+q}, \qquad \text{where} \qquad C_\nu = b_\nu^2\Big(\frac{1+\nu}{1-\nu}\Big)^2. \tag{92}$$

Summing (92) over k gives (with $\mathrm{TD}_t(x_t, a_t)$ as in (80))

$$\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t] \le C_\nu\, \mathrm{TD}_t(x_t, a_t). \tag{93}$$

Let $V_T := \sum_{t=1}^{T} \mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}, x_t, a_t]$ be the predictable quadratic variation. Then (93) implies $V_T \le C_\nu S$, with S as in (89).

We invoke Freedman's inequality for martingale difference sequences with bounded increments (e.g. (Lattimore & Szepesvári, 2020, Theorem 20.1)). Applying it to the martingale $-M_T$ and using (90) and $V_T \le C_\nu S$, we obtain that with probability at least $1 - \delta$,

$$-\sum_{t=1}^{T} \zeta_t \le \sqrt{2\, C_\nu S\, \log\Big(\frac{2}{\delta}\Big)} + \frac{2b_\nu}{3}\log\Big(\frac{2}{\delta}\Big). \tag{94}$$

Plugging (94) into (89) yields, on the same event,

$$S \le 2\mathsf{R}_{\log}(T) + 2\sqrt{2\, C_\nu S\, \log\Big(\frac{2}{\delta}\Big)} + \frac{4b_\nu}{3}\log\Big(\frac{2}{\delta}\Big).$$

Let $L := \log(2/\delta)$ and set $A := 2\sqrt{2C_\nu L}$ and $B := 2\mathsf{R}_{\log}(T) + \frac{4b_\nu}{3}L$. Then the previous inequality is $S \le B + A\sqrt{S}$, which implies $\sqrt{S} \le \sqrt{B} + A$ and hence

$$S \le (\sqrt{B} + A)^2 \le 2B + 2A^2 = 4\mathsf{R}_{\log}(T) + \Big(\frac{8b_\nu}{3} + 16C_\nu\Big)L. \tag{95}$$

Step 4: Conclude the mean-squared error bound. Combining (87) with (95) yields

$$\sum_{t=1}^{T}\big(f_{t-1}(x_t, a_t) - \mu(x_t, a_t)\big)^2 \le 2R_{\max}^2\, S \le 8R_{\max}^2\, \mathsf{R}_{\log}(T) + \Big(\frac{16}{3}R_{\max}^2\, b_\nu + 32R_{\max}^2\, C_\nu\Big)\log\Big(\frac{2}{\delta}\Big).$$

□

We re-derive a few convenient facts regarding IGW distributions in terms of our specific spanner pool action set at time-step t, denoted as $\mathcal{A}_t$, to simplify the subsequent analysis.

Proposition 2 (IGW exploitation and inverse-probability bounds)**.** *Fix a round t and context x_t with pool $\mathcal{A}_t$ of size $K_t := |\mathcal{A}_t|$. Let $\hat{a}_t \in \arg\max_{a\in\mathcal{A}_t} f_t(x_t, a)$ and define plug-in gaps $\hat{\Delta}_t(a) \geq 0$ on $\mathcal{A}_t$ with $\hat{\Delta}_t(\hat{a}_t) = 0$. Let $\vartheta \geq K_t$ and $\gamma_t > 0$, and define $\pi_t(\cdot \mid x_t)$ by (12). Let $\hat{\Delta}_{t,\max} := \max_{a\in\mathcal{A}_t} \hat{\Delta}_t(a)$. Then:*

$$\sum_{a\in\mathcal{A}_t} \pi_t(a \mid x_t)\, \hat{\Delta}_t(a) \leq \frac{\vartheta}{\gamma_t} \log\Big(1 + \frac{\gamma_t}{\vartheta}\hat{\Delta}_{t,\max}\Big), \tag{96}$$

$$\max_{a\in\mathcal{A}_t} \frac{1}{\pi_t(a \mid x_t)} \leq \vartheta + \gamma_t\hat{\Delta}_{t,\max}, \tag{97}$$

$$\sum_{a\in\mathcal{A}_t} \frac{1}{\pi_t(a \mid x_t)} \leq K_t\big(\vartheta + \gamma_t\hat{\Delta}_{t,\max}\big). \tag{98}$$

Proof. Since $\hat{\Delta}_t(\hat{a}_t) = 0$ and $\vartheta \geq K_t$,

$$\sum_{a\in\mathcal{A}_t\setminus\{\hat{a}_t\}} \frac{1}{\vartheta + \gamma_t\hat{\Delta}_t(a)} \leq \frac{K_t - 1}{\vartheta} \leq 1 - \frac{1}{\vartheta},$$

so $\pi_t(\hat{a}_t \mid x_t) \geq 1/\vartheta > 0$ and $\sum_{a\in\mathcal{A}_t} \pi_t(a \mid x_t) = 1$.

For (96), using $\hat{\Delta}_t(\hat{a}_t) = 0$,

$$\sum_{a\in\mathcal{A}_t} \pi_t(a \mid x_t)\hat{\Delta}_t(a) = \sum_{a\neq\hat{a}_t} \frac{\hat{\Delta}_t(a)}{\vartheta + \gamma_t\hat{\Delta}_t(a)}.$$

For any $u \geq 0$, the function $s \mapsto 1/(\vartheta + \gamma_t s)$ is decreasing, hence

$$\frac{u}{\vartheta + \gamma_t u} \leq \int_0^u \frac{1}{\vartheta + \gamma_t s}\, ds = \frac{1}{\gamma_t} \log\Big(1 + \frac{\gamma_t}{\vartheta}u\Big).$$

Applying this with $u = \hat{\Delta}_t(a)$ and then bounding by $\hat{\Delta}_{t,\max}$ gives

$$\sum_{a\neq\hat{a}_t} \frac{\hat{\Delta}_t(a)}{\vartheta + \gamma_t\hat{\Delta}_t(a)} \leq \frac{K_t - 1}{\gamma_t} \log\Big(1 + \frac{\gamma_t}{\vartheta}\hat{\Delta}_{t,\max}\Big) \leq \frac{\vartheta}{\gamma_t} \log\Big(1 + \frac{\gamma_t}{\vartheta}\hat{\Delta}_{t,\max}\Big),$$

which is (96).

For (97), if $a \neq \hat{a}_t$ then $1/\pi_t(a \mid x_t) = \vartheta + \gamma_t\hat{\Delta}_t(a) \leq \vartheta + \gamma_t\hat{\Delta}_{t,\max}$. If $a = \hat{a}_t$, then $\pi_t(\hat{a}_t \mid x_t) \geq 1/\vartheta$, hence $1/\pi_t(\hat{a}_t \mid x_t) \leq \vartheta \leq \vartheta + \gamma_t\hat{\Delta}_{t,\max}$.

Finally, (98) follows by summing:

$$\sum_{a\in\mathcal{A}_t} \frac{1}{\pi_t(a \mid x_t)} = \frac{1}{\pi_t(\hat{a}_t \mid x_t)} + \sum_{a\neq\hat{a}_t} \big(\vartheta + \gamma_t\hat{\Delta}_t(a)\big) \leq \vartheta + (K_t - 1)(\vartheta + \gamma_t\hat{\Delta}_{t,\max}) \leq K_t(\vartheta + \gamma_t\hat{\Delta}_{t,\max}).$$

□

E.4.1 Stochastic regret of FastContrastSpanner (high probability)

We now convert Lemma 11 into a full bandit regret bound. The overall structure mirrors the analysis that gives rise to (Sec. E) Theorem 1, except that the square-loss oracle prediction-error lemma is replaced by its log-loss analogue, through the triangular discrimination, i.e., Def. 4 and Lemma 11). This permits a sharpening of the rate as compared to its vanilla square-loss counterpart, and to the best of our knowledge, is the first time log-loss oracles have been studied in the causal setting. A particular unique aspect of our analysis is the role of contrast pool and the decomposition of terms to within-pool and pool approximation error, which we treat separately in Sec. F

Theorem ((**Restatement of Theorem 3**) High-probability regret of FASTCONTRASTSPANNER (log-loss oracle))**.** *Fix $T \geq 1$ and $\delta \in (0,1)$. Suppose Assumptions 2, 3.2, 4, and 7 hold true. Let $K_t := |\mathcal{A}_t|$ be the pool size on round t and assume the IGW parameter in (12) satisfies*

$$\vartheta \geq \max_{t \leq T} K_t.$$

Run Algorithm 2 with constant thresholds $\varepsilon_t \equiv \varepsilon$ and constant IGW scale $\gamma_t \equiv \gamma$.

Set the learning-rate parameter

$$\gamma := \sqrt{\frac{2\vartheta T}{4R_{\max}^2 \, \mathsf{R}_{\log}(T) + \Big(16 R_{\max}^2 \, C_\nu + \frac{8}{3} R_{\max}^2 \, (b_\nu + 1)\Big) \log\Big(\frac{4}{\delta}\Big)}}. \tag{99}$$

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

$$\begin{aligned}
\mathrm{Reg}(T) &\leq \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + \frac{2\vartheta T}{\gamma} + 4\gamma R_{\max}^2 \, \mathsf{R}_{\log}(T) \\
&\quad + \gamma\Big(16 R_{\max}^2 \, C_\nu + \frac{8}{3} R_{\max}^2 \, (b_\nu + 1)\Big) \log\Big(\frac{4}{\delta}\Big) + 2R_{\max}\sqrt{2T \log\Big(\frac{4}{\delta}\Big)} \qquad (100) \\
&\leq \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) \\
&\quad + 2\sqrt{2\vartheta T\Big(4R_{\max}^2 \, \mathsf{R}_{\log}(T) + \Big(16 R_{\max}^2 \, C_\nu + \frac{8}{3} R_{\max}^2 \, (b_\nu + 1)\Big) \log\Big(\frac{4}{\delta}\Big)\Big)} \\
&\quad + 2R_{\max}\sqrt{2T \log\Big(\frac{4}{\delta}\Big)}.
\end{aligned}$$

Here $b_\nu = \log(1/\nu)$ and $C_\nu = b_\nu^2 \big(\frac{1+\nu}{1-\nu}\big)^2$ are defined in (81).

Proof. Similar to the analysis of Theorem 1, we begin by studying $a_t^\circ \in \arg\max_{a \in \mathcal{A}_t} \mu(x_t, a)$ as the best-in-pool action, with $a_t \sim \pi_t(\cdot \mid x_t)$ as the random action actually played by Algorithm 2 (where π_t is the *vanilla* IGW distribution (12) restricted to $\mathcal{A}_t$). Define the in-pool regret random variable

$$R_t^{\mathrm{pool}} := \mu(x_t, a_t^\circ) - \mu(x_t, a_t) \ \in [0, 2R_{\max}].$$

As in the square-loss section, we decompose the total regret into an approximation (span) term and an in-pool term:

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + \sum_{t=1}^{T} R_t^{\mathrm{pool}} \tag{101}$$

One may simplify $\varepsilon_{\mathrm{span}}(t) \leq \varepsilon$ for constant threshold $\varepsilon_t \equiv \varepsilon$ (cf. Lemma 19 in the square-loss analysis), but we defer this for now.

Step 1: Concentration from realized in-pool regret to its conditional expectation. The sequence $\big(R_t^{\mathrm{pool}} - \mathbb{E}_t[R_t^{\mathrm{pool}}]\big)_{t \leq T}$ is a martingale difference sequence, and $|R_t^{\mathrm{pool}} - \mathbb{E}_t[R_t^{\mathrm{pool}}]| \leq 2R_{\max}$ a.s. By Azuma–Hoeffding, with probability at least $1 - \delta/4$,

$$\sum_{t=1}^{T} R_t^{\mathrm{pool}} \leq \sum_{t=1}^{T} \mathbb{E}_t[R_t^{\mathrm{pool}}] + 2R_{\max}\sqrt{2T \log\Big(\frac{4}{\delta}\Big)}. \tag{102}$$

Step 2: Bound $\mathbb{E}_t[R_t^{\mathrm{pool}}]$ by IGW + squared mean-prediction error. Condition on $\mathcal{F}_{t-1}, x_t$ and apply Lemma 2 on the pool $\mathcal{A}_t$ with $\hat{y}(a) = f_{t-1}(x_t, a)$ and $y^\star(a) = \mu(x_t, a)$. (Importantly, Lemma 2 is purely an *algebraic* property of IGW and does not depend on whether the oracle is square-loss or log-loss.) Writing $e_t := f_{t-1}(x_t, a_t) - \mu(x_t, a_t)$, we obtain

$$\mathbb{E}_t[R_t^{\mathrm{pool}}] \le \frac{2\mu}{\gamma} + \frac{\gamma}{4}\,\mathbb{E}_t[e_t^2]. \tag{103}$$

Summing (103) and substituting into (102) yields: with probability at least $1 - \delta/4$,

$$\sum_{t=1}^{T} R_t^{\mathrm{pool}} \le \frac{2\mu T}{\gamma} + \frac{\gamma}{4}\sum_{t=1}^{T}\mathbb{E}_t[e_t^2] + 2R_{\max}\sqrt{2T\log\Big(\frac{4}{\delta}\Big)}. \tag{104}$$

Step 3: Control $\sum_t \mathbb{E}_t[e_t^2]$ by realized squared error. Since $|f_{t-1}(x_t, a)| \le R_{\max}$ and $|\mu(x_t, a)| \le R_{\max}$, we have $|e_t| \le 2R_{\max}$ a.s. Applying Lemma 5 with failure probability $\delta/4$ gives

$$\sum_{t=1}^{T}\mathbb{E}_t[e_t^2] \le 2\sum_{t=1}^{T} e_t^2 + \frac{32}{3}R_{\max}^2\log\Big(\frac{4}{\delta}\Big). \tag{105}$$

Step 4: Control $\sum_t e_t^2$ using the log-loss oracle via triangular discrimination (Topsoe, 2002; Le Cam, 2012). Applying Lemma 11 to the *played* sequence (x_t, a_t) with failure probability $\delta/2$ yields that, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T}\big(f_{t-1}(x_t, a_t) - \mu(x_t, a_t)\big)^2 \le 8R_{\max}^2\,\mathsf{R}_{\log}(T) + \Big(32R_{\max}^2\,C_\nu + \frac{16}{3}R_{\max}^2\,b_\nu\Big)\log\Big(\frac{4}{\delta}\Big). \tag{106}$$

Step 5: Combine the bounds. On the intersection of the events from Steps 1–4 (which holds with probability at least $1 - \delta$ by a union bound), substitute (106) into (105), and then substitute the result into (104). Finally combine with (101) to obtain the first inequality in (100). The second inequality follows by optimizing the choice of γ in the first line. □

E.4.2 Minimax regret and the role of optimal design (log-loss oracle)

This subsection gives a *minimax / worst-case* high-probability regret guarantee for FastContrastSpanner when the learner uses an *optimal-design reweighting* of IGW. The proof follows the same DEC/OED scaffold as the square-loss minimax analysis (Section E.3) and SpannerIGW Zhu et al. (2022), but the key new ingredient is the conversion from *log-loss excess* to *mean-squared prediction error* via triangular discrimination (Lemma 11). Moreover, we delineate the interaction of triangular discrimination (Topsoe, 2002; Le Cam, 2012) and Fisher-weighted optimal design for the first time.

Fisher-weighted optimal design on contrast directions. Fix a round t and recall the contrast-direction notation from the square-loss minimax section: for the current context x_t and predicted best action $\hat{a}_t \in \arg\max_{a\in\mathcal{A}_t} f_{t-1}(x_t, a)$, define the within-pool direction

$$u_t(a) := z_{x_t}(\hat{a}_t, a), \qquad a \in \mathcal{A}_t \setminus \{\hat{a}_t\}, \tag{107}$$

which matches (59). For log-loss, the natural local geometry is the Fisher metric; we therefore define a *scalar Fisher weight* for the induced mean functional $f_q(x, a) = \sum_k r(k)q(k \mid x, a)$ (Definition 3):

$$w_t(a) := \sum_{k=1}^{K}\frac{r(k)^2}{q_{t-1}(k \mid x_t, a)}. \tag{108}$$

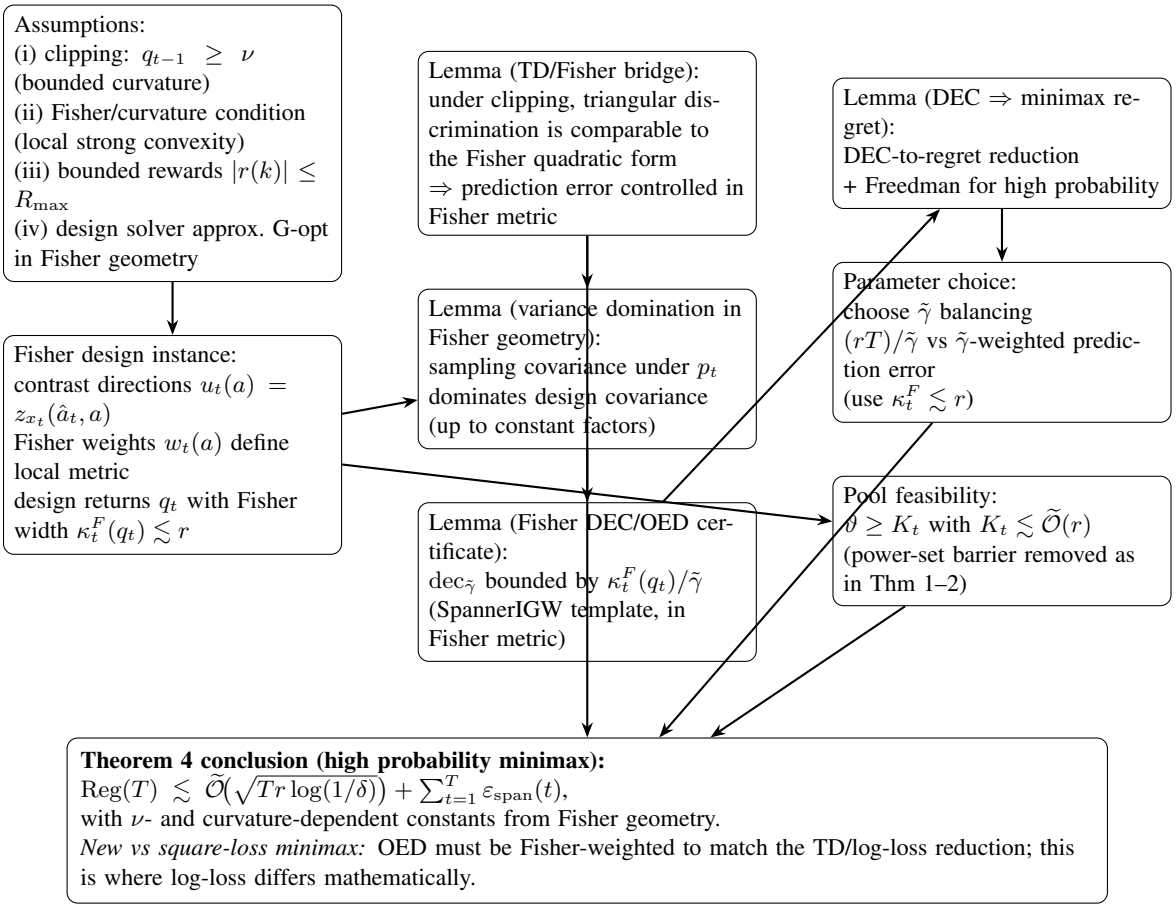


Figure 11: Proof logic flowchart for Theorem 4 (log-loss oracle, minimax high-probability regret via Fisher-weighted OED/DEC).

This quantity is the Fisher information of the one-dimensional functional $q \mapsto f_q(x_t, a)$ in the probability-parameterization of the categorical model (cf. the Fisher metric / information geometry of the simplex Amari & Nagaoka (2000)). Under clipping in Assumption 7.3 (i.e. $q_{t-1}(k \mid x_t, a) \geq \nu$),

$$0 \leq w_t(a) \leq \frac{1}{\nu} \sum_{k=1}^{K} r(k)^2 \leq \frac{K R_{\max}^2}{\nu}. \tag{109}$$

We form Fisher-rescaled directions

$$\tilde{u}_t(a) \coloneqq \sqrt{w_t(a)}\, u_t(a), \qquad a \in \mathcal{A}_t \setminus \{\hat{a}_t\},$$

and solve a (possibly approximate) G-optimal design problem on $\{\tilde{u}_t(a)\}_{a \in \mathcal{A}_t \setminus \{\hat{a}_t\}}$. Concretely, for $q \in \Delta(\mathcal{A}_t \setminus \{\hat{a}_t\})$ define the design matrix and width

$$\begin{aligned} M_t^{\mathrm{F}}(q) &\coloneqq \lambda I + \sum_{a \in \mathcal{A}_t \setminus \{\hat{a}_t\}} q(a)\, \tilde{u}_t(a) \tilde{u}_t(a)^\top, \\ \kappa_t^{\mathrm{F}}(q) &\coloneqq \max_{a \in \mathcal{A}_t \setminus \{\hat{a}_t\}} \tilde{u}_t(a)^\top \big(M_t^{\mathrm{F}}(q)\big)^{-1} \tilde{u}_t(a). \end{aligned} \tag{110}$$

Let q_t be any distribution returned by the design oracle satisfying Assumption 6 for this Fisher-rescaled instance, so that $\kappa_t^{\mathrm{F}}(q_t) \leq C_{\mathrm{opt}} \kappa_t^{\mathrm{F},\star}$. Since Fisher rescaling is by positive scalars and Assumption 2 restricts the directions to an r-dimensional subspace, we still have $\kappa_t^{\mathrm{F},\star} \leq r$ (the

optimal G-opt width is at most the ambient dimension of the span; see, e.g., standard G-opt facts used in SpannerIGW Zhu et al. (2022)). Hence

$$\kappa_t^{\mathrm{F}}(q_t) \le C_{\mathrm{opt}}\, r \qquad \text{for all } t. \tag{111}$$

Design-weighted IGW distribution. Given q_t and the plug-in gaps $\hat{\Delta}_t$ from (11), we use the same design-weighted IGW construction as in the square-loss minimax analysis: define $\lambda_t \in [\frac{1}{2}, 1]$ by (64) and then define p_t by (65). This sampling distribution depends on the single tuning parameter $\tilde{\gamma} > 0$ (the DEC parameter in (66)), and it has no explicit dependence on the pool size K_t. The following lemma is almost identical to Lemma 7, except that we include modifications related to the Fisher geometry.

Lemma 12 (Fisher-design IGW certifies a minimax DEC bound)**.** *Assume* (111) *and define* p_t *by* (65) *(with the same* $\tilde{\gamma}$ *as in* (63) *and* (66)*). Then for every* t*,*

$$\sup_{f\in\mathcal{F}} \sup_{a^\star\in\mathcal{A}_t} \mathbb{E}_{a\sim p_t}\Big[f(x_t,a^\star) - f(x_t,a) - \tilde{\gamma}\big(f_{t-1}(x_t,a) - f(x_t,a)\big)^2\Big] \le \frac{2C_{\mathrm{opt}} r}{\tilde{\gamma}}. \tag{112}$$

Equivalently, $\mathrm{dec}_{\tilde{\gamma}}(\mathcal{F}; f_{t-1}, x_t) \le 2C_{\mathrm{opt}} r/\tilde{\gamma}$.

Proof of Lemma 12. Fix t and abbreviate $x = x_t$, $\hat{a} = \hat{a}_t$, and $\hat{f} = f_{t-1}$. Write $a^\star$ for the maximizer of $f^\star(x,\cdot)$. By definition,

$$\mathrm{dec}_{\tilde{\gamma}}(\mathcal{F}; \hat{f}, x) = \inf_{p\in\Delta(\mathcal{A}_t)} \sup_{a^\star\in\mathcal{A}_t} \sup_{f^\star\in\mathcal{F}} \mathbb{E}_{a\sim p}\Big[f^\star(x,a^\star) - f^\star(x,a) - \tilde{\gamma}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2\Big].$$

We bound the quantity for the specific $p = p_t$ from (65).

Step 1: Decomposition. Add and subtract $\hat{f}$ and group terms:

$$f^\star(x,a^\star) - f^\star(x,a) = \underbrace{\hat{f}(x,\hat{a}) - \hat{f}(x,a)}_{\hat{\Delta}(a)} + \underbrace{\big(f^\star(x,a^\star) - \hat{f}(x,a^\star)\big)}_{T_3} + \underbrace{\big(\hat{f}(x,a) - f^\star(x,a)\big)}_{T_2} + \underbrace{\big(\hat{f}(x,a^\star) - \hat{f}(x,\hat{a})\big)}_{T_4}.$$

Taking expectation under p_t yields

$$\mathbb{E}_{a\sim p_t}\big[f^\star(x,a^\star) - f^\star(x,a)\big] = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \mathbb{E}_{a\sim p_t}\hat{\Delta}(a), \quad T_2 = \mathbb{E}_{a\sim p_t}\big[\hat{f}(x,a) - f^\star(x,a)\big], \quad T_3 = f^\star(x,a^\star) - \hat{f}(x,a^\star), \quad T_4 = \hat{f}(x,a^\star) - \hat{f}(x,\hat{a}) \le 0.$$

(The inequality uses $\hat{a} = \arg\max_a \hat{f}(x,a)$.)

Step 2: Bound T_1. By the definition of p_t,

$$T_1 = \sum_{a\neq\hat{a}} \frac{q_t(a)}{2(\lambda_t + \tilde{\gamma}\hat{\Delta}(a))}\,\hat{\Delta}(a) \le \sum_{a\neq\hat{a}} \frac{q_t(a)}{2\tilde{\gamma}} \le \frac{1}{2\tilde{\gamma}}.$$

Step 3: AM–GM on T_2. By AM–GM,

$$T_2 \le \frac{\tilde{\gamma}}{2}\,\mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2 + \frac{1}{2\tilde{\gamma}}.$$

Step 4: Control T_3 via variance domination and design width. Using Cauchy–Schwarz and AM–GM in the standard self-normalized form, which again is valid under Assumption 1,

$$T_3 = f^\star(x,a^\star) - \hat{f}(x,a^\star) \le \frac{1}{2\tilde{\gamma}} \|u_t(a^\star)\|^2_{\tilde{V}(p_t)^{-1}} + \frac{\tilde{\gamma}}{2} \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2.$$

By Lemma 6 and the assumed design-width bound $\kappa_t(q_t) \le C_{\mathrm{opt}} r$,

$$\|\tilde{u}_t(a^\star)\|^2_{\tilde{V}(p_t)^{-1}} \le 2\,\|\tilde{u}_t(a^\star)\|^2_{\tilde{V}(q_t)^{-1}} \le 2\,\kappa_t^F(q_t) \le 2C_{\mathrm{opt}} r.$$

Hence,

$$T_3 \le \frac{C_{\mathrm{opt}} r}{\tilde{\gamma}} + \frac{\tilde{\gamma}}{2} \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2.$$

Step 5: Combine and cancel. Summing T_1–T_4 (and using $T_4 \le 0$) yields

$$\mathbb{E}_{a\sim p_t}\big[f^\star(x,a^\star) - f^\star(x,a)\big] \le \frac{2C_{\mathrm{opt}} r}{\tilde{\gamma}} + \tilde{\gamma}\, \mathbb{E}_{a\sim p_t}\big(\hat{f}(x,a) - f^\star(x,a)\big)^2,$$

which is exactly (68). □

With these technical results in place, we are ready to present for the first time a minimax regret result for Algorithm 2, which to the best of our knowledge, is the first result of this type for log-loss oracles. In particular, through a novel construction of Fisher design weighting, we are able to formalize the following result, which holds for the class of context distributions and noise processes satisfying Assumption 4.

Theorem ((**Restatement of Theorem 4**) High-probability minimax regret of FastContrastSpanner with Fisher-weighted design)**.** *Fix $T \ge 1$ and $\delta \in (0,1)$. Under Assumptions 2, 3.2, 4, and 7, suppose we run Algorithm 2 with residual thresholds ε_t and, on each round t, compute q_t by Fisher-weighted G-optimal design on (110), and sample $a_t \sim p_t$ where p_t is the design-weighted IGW distribution (65) (with some fixed $\tilde{\gamma} > 0$). Then, with probability at least $1-\delta$, for any class of noise processes satisfying Assumptions 4, and any possibly adaptive/adversarially chosen series of contexts x_t, we have the following regret bound*

$$\begin{aligned}
\mathrm{Reg}(T) &\le \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + \frac{2C_{\mathrm{opt}} r\, T}{\tilde{\gamma}} \\
&\quad + \tilde{\gamma} \sum_{t=1}^{T} \big(f_{t-1}(x_t,a_t) - \mu(x_t,a_t)\big)^2 + 2R_{\max}\sqrt{2T\log\Big(\frac{4}{\delta}\Big)} \qquad (113)\\
&\le \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + \frac{2C_{\mathrm{opt}} r\, T}{\tilde{\gamma}} \qquad (114)\\
&\quad + 8\tilde{\gamma} R_{\max}^2\, \mathsf{R}_{\log}(T) + \tilde{\gamma}\Big(32R_{\max}^2\, C_\nu + \frac{16}{3} R_{\max}^2\, b_\nu\Big)\log\Big(\frac{4}{\delta}\Big) + 2R_{\max}\sqrt{2T\log\Big(\frac{4}{\delta}\Big)}.
\end{aligned}$$

In particular, choosing

$$\tilde{\gamma} := \sqrt{\frac{C_{\mathrm{opt}} r\, T}{4R_{\max}^2\, \mathsf{R}_{\log}(T) + \Big(16R_{\max}^2\, C_\nu + \frac{8}{3} R_{\max}^2\, b_\nu\Big)\log\big(\frac{4}{\delta}\big)}}$$

balances the leading terms in (114) and yields

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + r\sqrt{T\Big(\mathsf{R}_{\log}(T) + R_{\max}^2 \log\tfrac{1}{\delta}\Big)}\right), \qquad (115)$$

where the big-$\mathcal{O}$ hides only absolute numerical constants.

The novel aspects of this proof are the way to intersperse Lemma 12 regarding the Fisher weighting and induced control of the DEC (Def. 66), with the log-to-square reduction in Lemma 11, with the martingale construction for the within-pool regret. See Remark 9. The role of the Fisher weighting gets subsumed into the choice of $\tilde{\gamma}$ to mitigate the effect of κ in (111).

Proof. Recall the exact regret identity (Section F)

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t) + \sum_{t=1}^{T} \Big(\max_{a \in \mathcal{A}_t} \mu(x_t, a) - \mu(x_t, a_t) \Big).$$

It therefore suffices to bound the in-pool regret term.

Step 1: One-step DEC bound on the conditional in-pool regret. Let $a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mu(x_t, a)$ denote the best action in the current pool. Apply Lemma 12 with the comparator $f^\star = \mu$. This is valid because under Assumption 7 and Definition 3, $\mu(x, a) = f_{q^\star}(x, a)$ for the realizable distribution $q^\star$, hence $\mu \in \mathcal{F}$. Conditioning on the history $\mathcal{F}_{t-1}, x_t$ (with the shorthand $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}, x_t]$,

$$\mathbb{E}_t[\mu(x_t, a_t^\star) - \mu(x_t, a_t)] \le \frac{2 C_{\mathrm{opt}} r}{\tilde{\gamma}} + \tilde{\gamma}\, \mathbb{E}_t\big[(f_{t-1}(x_t, a_t) - \mu(x_t, a_t))^2\big]. \tag{116}$$

Summing (116) over $t \le T$ gives

$$\sum_{t=1}^{T} \mathbb{E}_t[\mu(x_t, a_t^\star) - \mu(x_t, a_t)] \le \frac{2 C_{\mathrm{opt}} r\, T}{\tilde{\gamma}} + \tilde{\gamma} \sum_{t=1}^{T} \mathbb{E}_t\big[(f_{t-1}(x_t, a_t) - \mu(x_t, a_t))^2\big]. \tag{117}$$

Step 2: Martingale deviation for the realized in-pool regret. Define the martingale difference sequence

$$D_t := \big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\big) - \mathbb{E}_t[\mu(x_t, a_t^\star) - \mu(x_t, a_t)].$$

Since $\mu(x, a) \in [-R_{\max}, R_{\max}]$ by bounded rewards, we have $|D_t| \le 2R_{\max}$ almost surely. By Hoeffding–Azuma, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} \big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\big) \le \sum_{t=1}^{T} \mathbb{E}_t[\mu(x_t, a_t^\star) - \mu(x_t, a_t)] + 2R_{\max}\sqrt{2T \log\Big(\frac{4}{\delta}\Big)}. \tag{118}$$

Step 3: Oracle-to-prediction conversion for the played actions. Apply Lemma 11 to the adaptive sequence (x_t, a_t) with failure probability $\delta/2$. With probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} \big(f_{t-1}(x_t, a_t) - \mu(x_t, a_t)\big)^2 \le 8R_{\max}^2\, R_{\log}(T) + \Big(32 R_{\max}^2\, C_\nu + \frac{16}{3} R_{\max}^2\, b_\nu\Big) \log\Big(\frac{4}{\delta}\Big). \tag{119}$$

By Assumption 3.2, we may identify $R_{\log}(T)$ with the log-loss oracle regret $\mathsf{R}_{\log}(T)$ appearing in the theorem statement.

Step 4: Combine the pieces. On the intersection of the events from Steps 2 and 3 (union bound gives probability at least $1 - \delta$), combine (117), (118), and (119) to obtain (113) and then (114). Finally, add the span-gap term $\sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t)$. □

Remark 9 (Where Fisher geometry and triangular discrimination enter)**.** *The minimax proof above is intentionally modular. The DEC/OED part (Lemma 12 and Step 1 of the proof) is the same as in SpannerIGW Zhu et al. (2022): optimal design controls the* worst-case *variance of gap estimation through the intrinsic span dimension r, rather than through the ambient pool size. What is new in the log-loss setting is Step 3: Lemma 11 converts log-loss excess into mean-squared prediction error* on an adaptive action sequence *using triangular discrimination. Under clipping, triangular*

discrimination (see (Topsoe, 2002; Le Cam, 2012)) is comparable (up to constants depending on ν) to the Fisher quadratic form on the simplex, which is why Fisher-rescaled designs are a natural analogue of the square-loss OED. In the causal-contrast setting, this is the mechanism by which distribution learning (log-loss) becomes sufficient for minimax regret on interventional mean rewards: the spanner controls which *contrast directions matter (dimension r), optimal design controls* how well *they are explored (width κ), and triangular discrimination links oracle performance back to mean-reward gaps.*

This discussion is formalized in the following lemma, which is provided for completeness.

Lemma 13 (Triangular discrimination induces Fisher geometry under clipping)**.** *Let p, q be distributions on $\mathcal{Y}$ such that $q(y) \geq \nu/|\mathcal{Y}|$ for all y. Then*

$$\sum_{y\in\mathcal{Y}} \frac{(p(y)-q(y))^2}{q(y)} \;\leq\; \frac{2|\mathcal{Y}|}{\nu} \sum_{y\in\mathcal{Y}} \frac{(p(y)-q(y))^2}{p(y)+q(y)}.$$

Moreover, the RHS is (up to constants) the triangular discrimination $\mathrm{TD}(p, q)$, and the LHS is exactly the local Fisher quadratic form at q.

Proof. Since $p(y) + q(y) \geq q(y)$ and $q(y) \geq \nu/|\mathcal{Y}|$, we have

$$\frac{1}{q(y)} \leq \frac{|\mathcal{Y}|}{\nu} \quad \text{and} \quad \frac{1}{q(y)} \leq \frac{2|\mathcal{Y}|}{\nu} \frac{1}{p(y)+q(y)}.$$

Multiply by $(p(y) - q(y))^2$ and sum over y. □

F Low-rank geometry: residual certificates, ε-nets, and when the assumption holds

This section collects the linear-algebraic and geometric facts used in the regret analyses (Theorems 1–4) to justify two key steps: (i) the *residual test* is a certificate that the learned spanner span forms an ε-net for the (possibly enormous) contrast class; and (ii) under additional *span adequacy* conditions, this ε-net is sufficient for near-optimal decision making. We then give sufficient structural causal conditions (linear and nonlinear SEMs) under which the true contrast span is low-dimensional, and we close with an information-theoretic lower bound connected to restricted isometry / incoherence.

F.1 Regret decomposition and the span gap

Throughout, let $\mathcal{A}_t \subseteq \mathcal{A}$ denote the (possibly data-dependent) finite action pool from which the algorithm samples at round t (e.g., the pool induced by the current spanner; see Section F.3.1).

Lemma 14 (Regret decomposes into a span gap term and an in-pool term)**.** *For each round t, let*

$$a_t^\star \in \arg\max_{a\in\mathcal{A}} \mu(x_t, a), \qquad a_t^{\mathrm{pool}} \in \arg\max_{a\in\mathcal{A}_t} \mu(x_t, a).$$

Then the cumulative regret admits the exact decomposition

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \Big(\mu(x_t, a_t^\star) - \mu(x_t, a_t)\Big) \tag{120}$$

$$= \underbrace{\sum_{t=1}^{T} \Big(\mu(x_t, a_t^\star) - \mu(x_t, a_t^{\mathrm{pool}})\Big)}_{\textit{span gap } \sum_t \varepsilon_{\mathrm{span}}(t)} + \underbrace{\sum_{t=1}^{T} \Big(\mu(x_t, a_t^{\mathrm{pool}}) - \mu(x_t, a_t)\Big)}_{\textit{in-pool regret}}. \tag{121}$$

In particular, defining the one-step span gap

$$\varepsilon_{\mathrm{span}}(t) := \max_{a\in\mathcal{A}} \mu(x_t, a) - \max_{a\in\mathcal{A}_t} \mu(x_t, a),$$

we have $\mathrm{Reg}(T) = \sum_{t=1}^{T} \varepsilon_{\mathrm{span}}(t)$ *+ (in-pool regret).*

Proof. Add and subtract $\mu(x_t, a_t^{\mathrm{pool}})$ inside the regret sum:

$$\mu(x_t, a_t^\star) - \mu(x_t, a_t) = \big(\mu(x_t, a_t^\star) - \mu(x_t, a_t^{\mathrm{pool}})\big) + \big(\mu(x_t, a_t^{\mathrm{pool}}) - \mu(x_t, a_t)\big),$$

then sum over $t \le T$. □

The remainder of this subsection provides sufficient conditions ensuring $\sum_{t\le T} \varepsilon_{\mathrm{span}}(t)$ is controlled by the spanner residual thresholds.

F.2 Design matrices, learned spans, and Mahalanobis residuals

We work with the regularized design matrices standard in linear bandit analyses (Abbasi-Yadkori et al., 2011).

Definition 5 (Learned span, design matrix, and Mahalanobis norm)**.** *Fix a regularization parameter* $\lambda > 0$*. Let* $\mathcal{S}_t$ *denote the set (or multiset) of* spanner update indices *up to time* t*. For each update* $s \in \mathcal{S}_t$ *we store the (context-dependent) contrast direction*

$$u_s := z_{x_s}(a_s, a_s') \in \mathbb{R}^d.$$

The learned contrast span *at time* t *is*

$$\widehat{\mathcal{U}}_t := \mathrm{span}\{u_s : s \in \mathcal{S}_t\} \subseteq \mathbb{R}^d,$$

and the associated (regularized) design matrix *is*

$$V_t := \lambda I + \sum_{s\in\mathcal{S}_t} u_s u_s^\top \in \mathbb{R}^{d\times d}. \tag{122}$$

For any $v \in \mathbb{R}^d$*, define the (semi-)norm*

$$\|v\|_{V_t^{-1}} := \sqrt{v^\top V_t^{-1} v}.$$

Finally, for any subspace $U \subseteq \mathbb{R}^d$*, write* Π_U *for the orthogonal projector onto* U *and* $\mathrm{dist}(v, U) := \|v - \Pi_U v\|_2$.

Definition 6 (Residual score)**.** *Given* V_{t-1} *and a candidate contrast vector* $z \in \mathbb{R}^d$ *(typically* $z = z_{x_t}(a, a')$*), its* residual score *at time* t *is*

$$\mathrm{Res}_t(z) := \|z\|_{V_{t-1}^{-1}}.$$

The canonical greedy spanner update rule adds z *to the spanner whenever* $\mathrm{Res}_t(z) > \varepsilon_t$ *for a chosen threshold* $\varepsilon_t > 0$*.*

Intuitively, $\|z\|_{V_{t-1}^{-1}}$ is small exactly when z is well-approximated by the current span $\widehat{\mathcal{U}}_{t-1}$ (made precise below).

F.2.1 From Mahalanobis residuals to Euclidean ε-nets

We first record a standard variational identity (a ridge-regression normal equation) that turns $\|z\|_{V^{-1}}$ into an explicit approximation certificate.

Lemma 15 (A variational form of the Mahalanobis norm). *Let $\lambda > 0$ and let $u_1, \dots, u_m \in \mathbb{R}^d$. Define $V := \lambda I + \sum_{i=1}^m u_i u_i^\top$ and let $S \in \mathbb{R}^{d\times m}$ be the matrix with columns u_i. Then for every $z \in \mathbb{R}^d$,*

$$z^\top V^{-1} z = \min_{b\in\mathbb{R}^m} \left\{ \frac{1}{\lambda} \|z - Sb\|_2^2 + \|b\|_2^2 \right\}. \tag{123}$$

Moreover, the minimizer is $b^\star = S^\top V^{-1} z$.

Proof. Fix z. For any $b \in \mathbb{R}^m$,

$$\begin{aligned}\frac{1}{\lambda}\|z - Sb\|_2^2 + \|b\|_2^2 &= \frac{1}{\lambda}(z - Sb)^\top (z - Sb) + b^\top b \\ &= \frac{1}{\lambda} z^\top z - \frac{2}{\lambda} b^\top S^\top z + b^\top \Big(I + \frac{1}{\lambda} S^\top S\Big) b.\end{aligned}$$

The quadratic form in b is strictly convex, so the unique minimizer satisfies the normal equation

$$\Big(I + \tfrac{1}{\lambda} S^\top S\Big) b = \tfrac{1}{\lambda} S^\top z.$$

Solving gives $b^\star = (S^\top S + \lambda I)^{-1} S^\top z$. Using the Woodbury identity,

$$V^{-1} = \frac{1}{\lambda} I - \frac{1}{\lambda} S (S^\top S + \lambda I)^{-1} S^\top \frac{1}{\lambda},$$

one checks that $S^\top V^{-1} z = (S^\top S + \lambda I)^{-1} S^\top z$, hence $b^\star = S^\top V^{-1} z$. Substituting $b^\star$ into the quadratic objective and simplifying yields $\min_b \{\frac{1}{\lambda}\|z - Sb\|_2^2 + \|b\|_2^2\} = z^\top V^{-1} z$. (See, e.g., Horn & Johnson (2012) for this identity in matrix-analysis form.) □

Lemma 16 (Residual score yields an ε-net certificate). *Let $\widehat{\mathcal{U}} := \operatorname{span}\{u_1, \dots, u_m\}$ and let $V = \lambda I + \sum_{i=1}^m u_i u_i^\top$. Then for every $z \in \mathbb{R}^d$,*

$$\operatorname{dist}(z, \widehat{\mathcal{U}}) \le \sqrt{\lambda}\, \|z\|_{V^{-1}}. \tag{124}$$

Equivalently, if $\|z\|_{V^{-1}} \le \varepsilon$ then $\|z - \Pi_{\widehat{\mathcal{U}}} z\|_2 \le \sqrt{\lambda}\,\varepsilon$.

Proof. By Lemma 15,

$$\|z\|_{V^{-1}}^2 = \min_{b\in\mathbb{R}^m} \left\{ \frac{1}{\lambda} \|z - Sb\|_2^2 + \|b\|_2^2 \right\}.$$

Dropping the nonnegative term $\|b\|_2^2$ only decreases the objective, so

$$\|z\|_{V^{-1}}^2 \ge \min_{b\in\mathbb{R}^m} \frac{1}{\lambda} \|z - Sb\|_2^2 = \frac{1}{\lambda} \operatorname{dist}(z, \widehat{\mathcal{U}})^2.$$

Rearranging yields (124). □

Corollary 1 (Residual certificate implies the ε-net condition). *Fix a context x and let $\mathcal{Z}_x := \{z_x(a, a') : a, a' \in \mathcal{A}\}$ denote the set of all contrast vectors at x. If*

$$\sup_{z\in\mathcal{Z}_x} \|z\|_{V^{-1}} \le \varepsilon \tag{125}$$

for V defined from some update set $\{u_i\}$ as above, then

$$\sup_{z\in\mathcal{Z}_x} \|z - \Pi_{\widehat{\mathcal{U}}} z\|_2 \le \sqrt{\lambda}\,\varepsilon.$$

In particular, the ε-net condition in (8) holds with net radius $\varepsilon_{\text{net}} := \sqrt{\lambda}\,\varepsilon$.

Proof. Apply Lemma 16 pointwise to each $z \in \mathcal{Z}_x$ and take the supremum. □

Algorithmic interpretation. If a residual-based spanner construction has access to an oracle that can (exactly or approximately) maximize $z \mapsto \|z\|_{V^{-1}}$ over the candidate set $\mathcal{Z}_x$, then terminating the greedy update rule at threshold ε certifies (125) and hence produces a $\sqrt{\lambda}\,\varepsilon$-net.

F.3 Greedy spanner growth and low-dimensionality of the learned span

We now upper bound how many times the greedy residual test can add a new direction when all contrast vectors lie in a rank-r subspace (Assumption 2).

Lemma 17 (Matrix determinant lemma)**.** *Let $A \in \mathbb{R}^{d\times d}$ be invertible and let $u \in \mathbb{R}^d$. Then*

$$\det(A + uu^\top) = \det(A)\big(1 + u^\top A^{-1} u\big).$$

Proof. This is standard: write $A + uu^\top = A(I + A^{-1/2}uu^\top A^{-1/2})$ and apply that $\det(I + vv^\top) = 1 + v^\top v$ for $v = A^{-1/2}u$. □

Proposition 3 (Greedy spanner growth)**.** *Assume Assumption 2, and additionally assume the contrast vectors are scaled so that $\|z_x(a, a')\|_2 \le 1$ for all x, a, a' (this is without loss of generality up to rescaling). Consider the greedy residual update rule with a* constant *threshold $\varepsilon_t \equiv \varepsilon \in (0, 1]$. If at most T candidate directions are tested up to time T, then the number of spanner updates satisfies*

$$|\mathcal{S}_T| \le \frac{r}{\log(1+\varepsilon^2)} \log\Big(1 + \frac{T}{\lambda r}\Big). \tag{126}$$

Consequently, the learned span $\widehat{\mathcal{U}}_T$ is a subspace of dimension at most r.

Proof. Let the update times be $t_1 < t_2 < \cdots < t_n$ where $n := |\mathcal{S}_T|$, and write the corresponding update directions as $u_i := z_{x_{t_i}}(a_{t_i}, a'_{t_i})$. Define the sequence of design matrices

$$V_0 := \lambda I, \qquad V_i := V_{i-1} + u_i u_i^\top \;\; (i = 1, \dots, n).$$

Step 1: each update increases $\log\det$ **by at least** $\log(1+\varepsilon^2)$**.** By the update rule, at the moment u_i is added we have $u_i^\top V_{i-1}^{-1} u_i > \varepsilon^2$. Applying Lemma 17 yields

$$\det(V_i) = \det(V_{i-1})\big(1 + u_i^\top V_{i-1}^{-1} u_i\big) > \det(V_{i-1})(1+\varepsilon^2).$$

Taking logs and telescoping gives

$$\log\det(V_n) - \log\det(V_0) > n\log(1+\varepsilon^2). \tag{127}$$

Step 2: low rank implies an upper bound on $\log\det(V_n)$**.** By Assumption 2, all u_i lie in a common subspace of dimension at most r. Therefore $\sum_{i=1}^n u_i u_i^\top$ has rank at most r and has at most r nonzero eigenvalues. Let these nonzero eigenvalues be $\sigma_1, \dots, \sigma_r \ge 0$ (padding with zeros if needed). Then the eigenvalues of $V_n = \lambda I + \sum_i u_i u_i^\top$ consist of $d - r$ copies of λ and r eigenvalues $\lambda + \sigma_j$. Hence

$$\det(V_n) = \lambda^{d-r} \prod_{j=1}^r (\lambda + \sigma_j) = \lambda^d \prod_{j=1}^r \Big(1 + \frac{\sigma_j}{\lambda}\Big).$$

Moreover, $\sum_{j=1}^r \sigma_j = \mathrm{tr}(\sum_i u_i u_i^\top) = \sum_{i=1}^n \|u_i\|_2^2 \le n$. By AM–GM,

$$\prod_{j=1}^r \Big(1 + \frac{\sigma_j}{\lambda}\Big) \le \Big(\frac{1}{r}\sum_{j=1}^r \big(1 + \frac{\sigma_j}{\lambda}\big)\Big)^r = \Big(1 + \frac{1}{\lambda r}\sum_{j=1}^r \sigma_j\Big)^r \le \Big(1 + \frac{n}{\lambda r}\Big)^r.$$

Taking logs and subtracting $\log\det(V_0) = d\log\lambda$ yields

$$\log\det(V_n) - \log\det(V_0) \le r\log\Big(1 + \frac{n}{\lambda r}\Big) \le r\log\Big(1 + \frac{T}{\lambda r}\Big), \tag{128}$$

where in the last step we used $n \le T$.

Step 3: combine and rearrange. Combining (127) and (128) gives $n\log(1+\varepsilon^2) < r\log(1 + T/(\lambda r))$, which is exactly (126). Finally, $\widehat{\mathcal{U}}_T \subseteq \mathrm{span}(U_\star)$ implies $\dim(\widehat{\mathcal{U}}_T) \le r$. □

Remark 10 (Rescaling if $\|z\|_2 \le L_z$)**.** *Assumption 2 allows an arbitrary bound $\|z_x(a,a')\|_2 \le L_z$. To apply Proposition 3 as stated, replace each candidate vector by z/L_z and replace ε by ε/L_z. Equivalently, one may insert L_z^2 into the quantity $T/(\lambda r)$.*

F.3.1 Pools induced by the contrast spanner

At round t, CONTRASTSPANNER constructs a finite pool $\mathcal{A}_t \subseteq \mathcal{A}$ from the current spanner state $\mathcal{S}_t$. Let

$$K_t := |\mathcal{A}_t|.$$

In the most direct "pool = endpoints" instantiation (Algorithm 1), whenever the spanner adds a new contrast direction $z_x(a,a')$ it can add at most two new actions to the pool. Thus, for all t,

$$K_t \le 1 + 2|\mathcal{S}_t|.$$

Combining this with the greedy spanner growth control (Proposition 3 implies that, for any fixed residual threshold $\varepsilon \in (0,1]$,

$$K_t \le K \quad \text{for all } t \le T, \qquad K := 1 + \frac{2r\,\log\!\big(1+\frac{T}{\lambda r}\big)}{\log(1+\varepsilon^2)}. \tag{129}$$

We write the bound so that it depends on the intrinsic rank r and is independent of $|\mathcal{A}|$.

F.4 When does the learned span cover the whole (relevant) space?

The spanner span $\widehat{\mathcal{U}}_t$ is always contained in the true contrast span. Corollary 1 already shows that a uniform residual certificate implies an ε-net guarantee in Euclidean distance. To upgrade such a guarantee into *exact* span recovery, one needs a mild conditioning assumption on a set of spanning contrasts.

Lemma 18 (Exact recovery from a sufficiently accurate net on a well-conditioned basis)**.** *Fix a context x and write*

$$\mathcal{U}_\star(x) := \mathrm{span}\big\{z_x(a,a') : a,a' \in \mathcal{A}\big\} \subseteq \mathbb{R}^d, \qquad r_x := \dim\big(\mathcal{U}_\star(x)\big).$$

Assume there exist contrasts $b_1,\dots,b_{r_x} \in \mathcal{Z}_x$ whose columns form a basis matrix $B := [b_1\ \cdots\ b_{r_x}] \in \mathbb{R}^{d\times r_x}$ with smallest singular value $\sigma_{\min}(B) > 0$. Let $\widehat{\mathcal{U}} \subseteq \mathcal{U}_\star(x)$ be any subspace. If

$$\max_{1\le i\le r_x} \mathrm{dist}(b_i,\widehat{\mathcal{U}}) \le \delta \qquad \text{for some } \delta < \frac{\sigma_{\min}(B)}{\sqrt{r_x}}, \tag{130}$$

then $\widehat{\mathcal{U}} = \mathcal{U}_\star(x)$. In particular, if an ε-net certificate holds uniformly over $\mathcal{Z}_x$ with $\varepsilon < \sigma_{\min}(B)/\sqrt{r_x}$, then the learned span must equal the true span.

Proof. Let $p_i := \Pi_{\widehat{\mathcal{U}}} b_i$ and stack $P := [p_1\ \cdots\ p_{r_x}]$. Then P has all columns in $\widehat{\mathcal{U}}$, and the error matrix $E := B - P$ has columns $e_i = b_i - p_i$ satisfying $\|e_i\|_2 \le \delta$ by assumption. Hence $\|E\|_F^2 = \sum_i \|e_i\|_2^2 \le r_x\delta^2$, so $\|E\|_2 \le \|E\|_F \le \delta\sqrt{r_x}$.

Now take any $\alpha \in \mathbb{R}^{r_x}$. By definition of $\sigma_{\min}(B)$,

$$\|B\alpha\|_2 \ge \sigma_{\min}(B)\,\|\alpha\|_2. \tag{131}$$

On the other hand, if $P\alpha = 0$ then $B\alpha = E\alpha$, so

$$\|B\alpha\|_2 \le \|E\|_2\,\|\alpha\|_2 \le \delta\sqrt{r_x}\,\|\alpha\|_2.$$

Combining with (131) and using $\delta\sqrt{r_x} < \sigma_{\min}(B)$ forces $\alpha = 0$. Thus the columns of P are linearly independent, so $\dim(\widehat{\mathcal{U}}) \geq r_x$. Since $\widehat{\mathcal{U}} \subseteq \mathcal{U}_\star(x)$ and $\dim(\mathcal{U}_\star(x)) = r_x$, we conclude $\widehat{\mathcal{U}} = \mathcal{U}_\star(x)$. □

Remark 11 (Covering the entire ambient space)**.** *If the true contrast span is full ($\mathcal{U}_\star(x) = \mathbb{R}^d$), then $r_x = d$ and the condition (130) shows that approximating a well-conditioned basis of $\mathbb{R}^d$ is enough to certify that the learned span equals the whole space.*

F.4.1 From an ε-net to small span gap: a span-adequacy condition

The ε-net certificate controls the approximation of contrast vectors by the learned span. To translate this into control of the span gap term $\varepsilon_{\text{span}}(t)$ in Lemma 14, we need a compatibility condition between reward differences and contrast approximation.

A minimal (and standard) compatibility is simply boundedness of the contrast parameter. This is a harmless normalization in many models: if $\|w_x\|_2 \leq B$ and $\|z_x(a, a')\|_2 \leq L_z$, then rescaling $z \leftarrow z/(BL_z)$ makes $\|w_x\|_2 \leq 1$ and $\|z\|_2 \leq 1$.

Lemma 19 (Span gap from an ε-net certificate)**.** *Fix a round t and a context x_t. Assume the contrast-realizability model of Assumption 1 holds at x_t: for some $w_{x_t} \in \mathbb{R}^d$,*

$$\mu(x_t, a) - \mu(x_t, a') = \langle w_{x_t}, z_{x_t}(a, a')\rangle \quad \textit{for all } a, a' \in \mathcal{A}. \tag{132}$$

Assume further that $\|w_{x_t}\|_2 \leq B$ for some constant $B > 0$.

Let $a_0 \in \mathcal{A}_t$ be any reference action and suppose the learned span $\widehat{\mathcal{U}}_{t-1}$ satisfies an anchored ε-net condition:

$$\sup_{a\in\mathcal{A}} \big\| z_{x_t}(a, a_0) - \Pi_{\widehat{\mathcal{U}}_{t-1}} z_{x_t}(a, a_0)\big\|_2 \leq \varepsilon_{\text{net}}(t). \tag{133}$$

Define the projected-reward maximizer

$$\tilde{a}_t \in \arg\max_{a\in\mathcal{A}} \big\langle w_{x_t}, \Pi_{\widehat{\mathcal{U}}_{t-1}} z_{x_t}(a, a_0)\big\rangle.$$

If the pool contains such a maximizer (i.e., $\tilde{a}_t \in \mathcal{A}_t$), then the one-step span gap satisfies

$$\varepsilon_{\text{span}}(t) = \max_{a\in\mathcal{A}} \mu(x_t, a) - \max_{a\in\mathcal{A}_t} \mu(x_t, a) \leq 2B\, \varepsilon_{\text{net}}(t). \tag{134}$$

Moreover, if the residual test certifies $\sup_{a\in\mathcal{A}} \|z_{x_t}(a, a_0)\|_{V_{t-1}^{-1}} \leq \varepsilon_t$, then (133) holds with $\varepsilon_{\text{net}}(t) = \sqrt{\lambda}\,\varepsilon_t$ and hence $\varepsilon_{\text{span}}(t) \leq 2B\sqrt{\lambda}\,\varepsilon_t$.

Proof. For any $a \in \mathcal{A}$, write $z(a) := z_{x_t}(a, a_0)$ and $\tilde{z}(a) := \Pi_{\widehat{\mathcal{U}}_{t-1}} z(a)$. Then by realizability (132) with $a' = a_0$,

$$\mu(x_t, a) = \mu(x_t, a_0) + \langle w_{x_t}, z(a)\rangle.$$

Define the "projected" surrogate

$$\tilde{\mu}(x_t, a) := \mu(x_t, a_0) + \langle w_{x_t}, \tilde{z}(a)\rangle.$$

By Cauchy–Schwarz and (133), for all a,

$$|\mu(x_t, a) - \tilde{\mu}(x_t, a)| = |\langle w_{x_t}, z(a) - \tilde{z}(a)\rangle| \leq \|w_{x_t}\|_2\, \|z(a) - \tilde{z}(a)\|_2 \leq B\varepsilon_{\text{net}}(t). \tag{135}$$

Let $a_t^\star \in \arg\max_{a\in\mathcal{A}} \mu(x_t, a)$ be a true optimal action. Then, with $\tilde{\mu}(x_t, a_t^\star)$ as the best reward predictor available from the learned contrast subspace, we can quantify the approximation error from projecting onto the learned subspace as follows.

$$\begin{aligned}\mu(x_t, a_t^\star) - \mu(x_t, \tilde{a}_t) &\leq \big(\mu(x_t, a_t^\star) - \tilde{\mu}(x_t, a_t^\star)\big) + \big(\tilde{\mu}(x_t, a_t^\star) - \tilde{\mu}(x_t, \tilde{a}_t)\big) + \big(\tilde{\mu}(x_t, \tilde{a}_t) - \mu(x_t, \tilde{a}_t)\big) \\ &\leq B\varepsilon_{\text{net}}(t) + 0 + B\varepsilon_{\text{net}}(t) = 2B\varepsilon_{\text{net}}(t),\end{aligned}$$

where the middle term is nonpositive by definition of $\tilde{a}_t$ as a maximizer of $\tilde{\mu}$. If $\tilde{a}_t \in \mathcal{A}_t$ then $\max_{a\in\mathcal{A}_t} \mu(x_t, a) \geq \mu(x_t, \tilde{a}_t)$, so (134) follows.

Finally, if $\sup_a \|z_{x_t}(a, a_0)\|_{V_{t-1}^{-1}} \leq \varepsilon_t$, then Corollary 1 applied to the anchored contrast set $\{z_{x_t}(a, a_0) : a \in \mathcal{A}\}$ gives (133) with $\varepsilon_{\text{net}}(t) = \sqrt{\lambda}\,\varepsilon_t$. □

Remark 12 (How to satisfy the pool condition $\tilde{a}_t \in \mathcal{A}_t$)**.** *Lemma 19 separates two requirements: (i) a geometric ε-net certificate (controlled by the residual test), and (ii) an algorithmic requirement that the pool contain a maximizer for the projected surrogate. Condition (ii) is satisfied, for instance, if the pool is augmented with the output of an optimization oracle that maximizes the current projected predictor over $\mathcal{A}$, or if $\mathcal{A}_t = \mathcal{A}$.*

Next, we formally state this as an assumption, and then provide to conditions under which this condition is guaranteed to hold.

Assumption 8 (Span adequacy / pool condition)**.** *For every round t, the selected pool $\mathcal{A}_t$ satisfies*

$$\operatorname{span}\{z_{x_t}(a, \hat{a}_t) : a \in \mathcal{A}_t\} \supseteq \mathcal{U}_\star \quad \textit{or equivalently} \quad \lambda_{\min}\Big(\sum_{a\in\mathcal{A}_t} u_t(a)u_t(a)^\top\Big) \geq \alpha > 0$$

for some uniform constant α.

Assumption 8 is implied by either the compressive-sensing-type of information-theoretic quantities, as formalized next in Lemma 20 (also, similar conditions on the existence of linear or nonlinear mediators on the structural equation model, or SEM), which we discuss next. Alternatively, one may directly bound the approximation error associated with the pool. While Assumption 8 adopts this later approach for simplicity, we formalize that it holds, next.

To be more precise, Lemma 19 is purely geometric/algebraic, so it is natural to ask when its assumptions (especially the anchored net / residual certificate) are expected to hold. In the linear SEM setting of Proposition 1.(i) (a similar condition is derived on the Jacobian transformation that relates inteventions and outcomes in Proposition 1.(ii)), anchored contrasts take the form $z(u, u_0) = U(u - u_0)$ for a fixed transfer matrix U. If interventions are restricted to be s-sparse and U satisfies an s-restricted isometry property (RIP) or an analogous mutual incoherence bound on the relevant columns, then the map $u \mapsto Uu$ is *bi-Lipschitz* on that sparse class. In particular, recall the restricted isometry property, stated next.

For a matrix $U \in \mathbb{R}^{d\times m}$ and an integer $s \geq 1$, the *s-restricted isometry constant* $\delta_s \in [0, 1)$ is the smallest value such that

$$(1 - \delta_s)\|v\|_2^2 \leq \|Uv\|_2^2 \leq (1 + \delta_s)\|v\|_2^2 \qquad \text{for all } s\text{-sparse } v \in \mathbb{R}^m. \tag{136}$$

See, e.g., Candès (2008); Donoho (2006).

This implies two useful consequences: (a) Euclidean ε-nets in intervention space and in contrast space are equivalent up to constant factors; and (b) once the design matrix is well-conditioned on $\operatorname{col}(U)$ (a restricted-eigenvalue condition that is implied by RIP/incoherence), the Mahalanobis residuals are uniformly small on the sparse contrast class, giving exactly the anchored residual certificate used to invoke Corollary 1. The following lemma makes this quantitative.

Lemma 20 (RIP implies a uniform residual certificate on a sparse contrast class)**.** *Fix a support set $S \subseteq [m]$ of size $|S| = s$ and let $U \in \mathbb{R}^{d\times m}$. Assume U satisfies s-RIP with constant $\delta_s < 1$ in the sense of (136), so in particular the submatrix $U_S \in \mathbb{R}^{d\times s}$ has singular values in $[\sqrt{1-\delta_s}, \sqrt{1+\delta_s}]$.*

Let $V = \lambda I + \sum_{i=1}^n z_i z_i^\top$ be any design matrix built from update directions $z_i \in \operatorname{col}(U_S)$, and suppose there exists $\alpha > 0$ such that the cumulative unregularized design obeys the restricted eigenvalue lower bound

$$\sum_{i=1}^n z_i z_i^\top \succeq \alpha\, U_S U_S^\top. \tag{137}$$

(For example, if among the update directions one includes each basis column Ue_j for $j \in S$ at least n_0 times, then (137) holds with $\alpha = n_0$.) Then for every $v \in \mathbb{R}^m$ supported on S with $\|v\|_2 \le 1$,

$$\|Uv\|_{V^{-1}} \le \sqrt{\frac{1+\delta_s}{\lambda + \alpha(1-\delta_s)}}. \tag{138}$$

In particular, if λ is constant and $\alpha \gtrsim \varepsilon^{-2}$, then $\sup_{\|v\|_2 \le 1,\ \mathrm{supp}(v) \subseteq S} \|Uv\|_{V^{-1}} \lesssim \varepsilon$. Combining with Corollary 1 (applied to anchored contrasts) gives an anchored $O(\sqrt{\lambda}\,\varepsilon)$-net guarantee, and then Lemma 19 yields $\varepsilon_{\mathrm{span}}(t) \lesssim B\sqrt{\lambda}\,\varepsilon$.

Proof. Fix v supported on S with $\|v\|_2 \le 1$ and write $z := Uv \in \mathrm{col}(U_S)$. By s-RIP, $\|z\|_2 = \|U_S v_S\|_2 \le \sqrt{1+\delta_s}\,\|v_S\|_2 \le \sqrt{1+\delta_s}$.

Next, (137) implies $V \succeq \lambda I + \alpha U_S U_S^\top$. Restricting to the subspace $\mathrm{col}(U_S)$, the smallest eigenvalue of $\alpha U_S U_S^\top$ is $\alpha\,\sigma_{\min}(U_S)^2 \ge \alpha(1-\delta_s)$. Therefore, for all $z \in \mathrm{col}(U_S)$,

$$z^\top V^{-1} z \le \frac{\|z\|_2^2}{\lambda + \alpha(1-\delta_s)}.$$

Plugging $\|z\|_2^2 \le 1+\delta_s$ yields (138). □

F.4.2 Conditions on the Causal Model to Ensure Low Rankness

We now give concrete causal models under which the *true* contrast span is low-dimensional. The common object is the (possibly nonlinear) map from interventions to interventional means.

Setup. Let $u \in \mathbb{R}^m$ parameterize interventions e.g., the vector of do-levels on m manipulable nodes, that is, $\mathcal{A} = \{u(a)\} \subset \mathbb{R}^m$. Define the induced (context-free) contrast vectors

$$z(u,u') := m(u) - m(u') \ \in \mathbb{R}^d.$$

(For contextual models, one can let $m_x(u)$ depend on x; the statements below hold pointwise in x.)

Proposition ((**Restatement of Prop. 1**) Low-rank contrast span under linear and nonlinear SEMs)**.**

(i) ***Linear SEM (constant Jacobian).*** *Suppose $X \in \mathbb{R}^d$ obeys the linear structural equation model*

$$X = W^\top X + \Gamma u + \varepsilon, \tag{139}$$

where $W \in \mathbb{R}^{d\times d}$ is (after a topological ordering) strictly upper triangular (i.e., corresponds to a DAG), $\Gamma \in \mathbb{R}^{d\times m}$ injects interventions, and $\mathbb{E}[\varepsilon] = 0$. Then for all u, u',

$$z(u,u') = U\,(u-u') \qquad \textit{where} \qquad U := (I - W^\top)^{-1}\Gamma. \tag{140}$$

In particular, the contrast span equals $\mathrm{col}(U)$ and has dimension $\mathrm{rank}(U)$.

Moreover, fix a set of outcome coordinates *$\mathcal{O} \subseteq [d]$ (e.g., those that enter the reward), and let $P_{\mathcal{O}} \in \{0,1\}^{|\mathcal{O}|\times d}$ denote the coordinate projection. If there exists a node set $\mathcal{M} \subseteq [d]$ of size r such that* every directed path from any intervened node (to which u connects through Γ) to any outcome node in $\mathcal{O}$ passes through $\mathcal{M}$ *(i.e., $\mathcal{M}$ is a directed vertex cut separating interventions from outcomes), then*

$$\mathrm{rank}\big(P_{\mathcal{O}} U\big) \le r. \tag{141}$$

(ii) ***Nonlinear SEM (low-rank Jacobian).*** *Assume $m : \mathcal{U} \to \mathbb{R}^d$ is continuously differentiable on a convex set $\mathcal{U} \subseteq \mathbb{R}^m$. Suppose there exists a fixed matrix $A \in \mathbb{R}^{d\times r}$ such that for every $u \in \mathcal{U}$, the Jacobian of m factors as*

$$Jm(u) = AB(u) \quad \textit{for some } B(u) \in \mathbb{R}^{r\times m}. \tag{142}$$

Equivalently, for all u, the column space $\mathrm{col}(Jm(u)) \subseteq \mathrm{col}(A)$. Then for every $u, u' \in \mathcal{U}$,

$$z(u, u') = m(u) - m(u') \in \mathrm{col}(A), \tag{143}$$

so the contrast span has dimension at most r.

Proof. **(i) Linear SEM.** Rearrange (139) as $(I - W^\top)X = \Gamma u + \varepsilon$. Since W is strictly upper triangular, $I - W^\top$ is invertible and $(I - W^\top)^{-1} = I + W^\top + (W^\top)^2 + \cdots + (W^\top)^{d-1}$ (a finite Neumann series). Taking expectations gives $m(u) = \mathbb{E}[X \mid \mathrm{do}(U{=}u)] = (I - W^\top)^{-1}\Gamma u$. Therefore $m(u) - m(u') = U(u - u')$ with U as in (140), proving the rank claim.

For the mediator cut statement, write the fundamental matrix $F := (I - W^\top)^{-1}$ so that $U = F\Gamma$. Equivalently, we can view Γ as adding directed edges from each intervention coordinate $j \in [m]$ to endogenous nodes $k \in [d]$ with weights $\Gamma_{k,j}$; then $U_{i,j}$ is the total weight of all directed paths from intervention coordinate j to node i in this augmented graph. Formally, expanding $F = \sum_{\ell=0}^{d-1}(W^\top)^\ell$ gives

$$U_{i,j} = e_i^\top F\Gamma e_j = \sum_{\ell=0}^{d-1} e_i^\top (W^\top)^\ell \Gamma e_j,$$

and each term $e_i^\top (W^\top)^\ell \Gamma e_j$ is a sum over all length-ℓ directed paths from j to i of the product of edge weights along the path (Neumann/path expansion; see, e.g., Pearl (2009)).

Now fix $i \in \mathcal{O}$ and $j \in [m]$. By assumption, every such directed path from j to i passes through the mediator vertex cut $\mathcal{M}$. For each contributing path p, let $\kappa(p) \in \mathcal{M}$ denote the first mediator visited by p (along the direction from intervention to outcome). Then p decomposes uniquely as a concatenation $p = p_{\mathrm{in}} \circ p_{\mathrm{out}}$, where p_{in} is a directed path from j to $\kappa(p)$ that does not pass through any mediator in $\mathcal{M}$ except at its endpoint, and p_{out} is a directed path from $\kappa(p)$ to i that does not visit any mediator in its interior. Because path weights multiply under concatenation, $\mathrm{wt}(p) = \mathrm{wt}(p_{\mathrm{out}})\,\mathrm{wt}(p_{\mathrm{in}})$.

Define

$$\beta_{k,j} := \sum_{p_{\mathrm{in}}: j \to k} \mathrm{wt}(p_{\mathrm{in}}), \qquad \alpha_{i,k} := \sum_{p_{\mathrm{out}}: k \to i} \mathrm{wt}(p_{\mathrm{out}}),$$

where the sums range over paths with the above mediator-avoidance property. Then for every $i \in \mathcal{O}$ and $j \in [m]$,

$$(P_{\mathcal{O}} U)_{i,j} = \sum_{k \in \mathcal{M}} \alpha_{i,k}\, \beta_{k,j}.$$

Fix an ordering of $\mathcal{M}$ and form matrices $A_1 \in \mathbb{R}^{|\mathcal{O}| \times r}$ and $B_1 \in \mathbb{R}^{r \times m}$ with entries $(A_1)_{i,k} = \alpha_{i,k}$ and $(B_1)_{k,j} = \beta_{k,j}$. This gives the explicit factorization $P_{\mathcal{O}} U = A_1 B_1$, hence $\mathrm{rank}(P_{\mathcal{O}} U) \le r$.

(ii) Nonlinear SEM. Fix $u, u' \in \mathcal{U}$ and consider the line segment $\gamma : [0, 1] \to \mathcal{U}$ given by $\gamma(t) = u' + t(u - u')$. Since $\mathcal{U}$ is convex, $\gamma(t) \in \mathcal{U}$ for all t. By the fundamental theorem of calculus for vector-valued C^1 functions,

$$m(u) - m(u') = \int_0^1 \frac{d}{dt} m(\gamma(t))\, dt = \int_0^1 Jm(\gamma(t))\,(u - u')\, dt.$$

Using the factorization (142) gives $Jm(\gamma(t))(u - u') = A\big(B(\gamma(t))(u - u')\big) \in \mathrm{col}(A)$ for every t. Since $\mathrm{col}(A)$ is a linear subspace, the integral also lies in $\mathrm{col}(A)$, proving (143). □

Remark 13 (Connection to the Jacobian condition)**.** *Part (ii) makes explicit the desired "Jacobian low-rank" condition: the intervention-to-outcome map $m(u)$ may be nonlinear, but if* all *its infinitesimal changes lie in a fixed r-dimensional subspace, then all finite contrasts lie in the same subspace. The linear SEM case is recovered by $m(u) = Uu$, for which $Jm(u) \equiv U$ is constant.*

F.4.3 Information-theoretic limits: a RIP/incoherence-based minimax regret lower bound

Finally, we record a lower bound showing that, even under favorable conditioning such as a restricted isometry property (RIP) or mutual incoherence of the causal transfer operator, one cannot hope to beat the intrinsic-dimensional $\Omega(s\sqrt{T})$ scaling of minimax regret when the learner is allowed to use s-sparse interventions. The key point is that RIP guarantees that the transfer map embeds an s-dimensional Euclidean ball of sparse interventions into the outcome space with only constant distortion, so the causal bandit contains an ordinary s-dimensional linear bandit as a special case. We do not restrict the action set (which changes the regret benchmark); instead we apply an invertible change of coordinates and restrict the parameter set, which is monotone in minimax.

Theorem 5 (RIP-based minimax regret lower bound)**.** *Fix $s \geq 1$ and suppose a transfer matrix $U \in \mathbb{R}^{d\times m}$ satisfies s-RIP with constant $\delta_s \leq 1/2$. Consider the (context-free) stochastic linear bandit problem with action set*

$$\mathcal{U}_S := \{v \in \mathbb{R}^m : \operatorname{supp}(v) \subseteq S,\ \|v\|_2 \leq 1\}$$

for some fixed coordinate subset $S \subseteq [m]$ of size $|S| = s$. On each round t, the learner chooses an intervention vector $u_t \in \mathcal{U}_S$ and observes

$$R_t = \langle \theta,\ U u_t\rangle + \xi_t,$$

where $\xi_t \sim \mathcal{N}(0,1)$ i.i.d., and the unknown parameter satisfies $\|\theta\|_2 \leq 1$. Then there exists a universal constant $c > 0$ such that for all $T \geq s$,

$$\inf_{\pi} \sup_{\|\theta\|_2\leq 1} \mathbb{E}_{\theta,\pi}\big[\mathrm{Reg}(T)\big] \geq c\, s\sqrt{T}, \tag{144}$$

where the infimum is over all (possibly randomized) bandit algorithms π.

Consequently, if s-RIP holds for some s comparable to the intrinsic contrast dimension, no algorithm can achieve $o(s\sqrt{T})$ minimax regret in general.

Proof. Let $U_S \in \mathbb{R}^{d\times s}$ denote the submatrix formed by the columns of U indexed by S. Any action $u \in \mathcal{U}_S$ can be written as $u = \iota_S(v)$ for some $v \in \mathbb{R}^s$ with $\|v\|_2 \leq 1$, where ι_S embeds v into $\mathbb{R}^m$ by placing its coordinates on S and zeros elsewhere. Thus the reward model restricted to $\mathcal{U}_S$ is exactly

$$R_t = \langle \theta,\ U_S v_t\rangle + \xi_t, \qquad \|v_t\|_2 \leq 1.$$

This is a stochastic linear bandit whose action set in feature space is the ellipsoid $\mathcal{E} := \{U_S v : \|v\|_2 \leq 1\} \subseteq \mathrm{col}(U_S)$.

Let $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values of U_S. By s-RIP with $\delta_s \leq 1/2$,

$$\sigma_{\min}^2 \geq 1 - \delta_s \geq 1/2, \qquad \sigma_{\max}^2 \leq 1 + \delta_s \leq 3/2.$$

Consequently, writing $\mathbb{B}_{\mathrm{col}(U_S)}(\rho)$ for the Euclidean ball of radius ρ in the s-dimensional subspace $\mathrm{col}(U_S)$, we have the inclusions

$$\mathbb{B}_{\mathrm{col}(U_S)}(\sigma_{\min}) \subseteq \mathcal{E} \subseteq \mathbb{B}_{\mathrm{col}(U_S)}(\sigma_{\max}).$$

Let $P \in \mathbb{R}^{d\times s}$ be an orthonormal basis for $\mathrm{col}(U_S)$ and define

$$A := P^\top U_S \in \mathbb{R}^{s\times s}.$$

Since $\sigma_{\min} > 0$, the matrix A is invertible and its singular values coincide with those of U_S (in particular, $\sigma_{\min}(A) = \sigma_{\min}$). Writing $\theta' := P^\top\theta$, we have $\|\theta'\|_2 \leq \|\theta\|_2 \leq 1$, and the rewards can be rewritten as

$$R_t = \langle \theta',\ A v_t\rangle + \xi_t = \langle A^\top\theta',\ v_t\rangle + \xi_t, \qquad \|v_t\|_2 \leq 1.$$

Therefore the problem is equivalent to a stochastic linear bandit in $\mathbb{R}^s$ with *unit-ball* action set $\{v \in \mathbb{R}^s : \|v\|_2 \le 1\}$ and parameter set

$$\Theta := \{A^\top \theta' : \|\theta'\|_2 \le 1\}.$$

Moreover, Θ contains an s-dimensional Euclidean ball of radius $\sigma_{\min}$: indeed, for any $\phi \in \mathbb{R}^s$ with $\|\phi\|_2 \le \sigma_{\min}$, letting $\theta' = A^{-T}\phi$ yields $\|\theta'\|_2 \le \|A^{-T}\|_{\mathrm{op}}\|\phi\|_2 \le (1/\sigma_{\min})\sigma_{\min} = 1$, so $\phi \in \Theta$. Thus,

$$\inf_{\pi} \sup_{\|\theta\|_2 \le 1} \mathbb{E}_{\theta,\pi}\big[\mathrm{Reg}(T)\big] \;\ge\; \inf_{\pi} \sup_{\|\phi\|_2 \le \sigma_{\min}} \mathbb{E}_{\phi,\pi}\big[\mathrm{Reg}(T)\big].$$

The minimax regret for the standard stochastic linear bandit on the unit ball with parameter ball radius $\sigma_{\min}$ is known to be at least on the order of $s\sigma_{\min}\sqrt{T}$ (for $T \ge s$) under sub-Gaussian noise; see, e.g., Dani et al. (2008) or Lattimore & Szepesvári (2020, Ch. 19). Since $\sigma_{\min} \ge 1/\sqrt{2}$ is a universal constant, this yields (144). □

Corollary 2 (A mutual incoherence / coherence variant)**.** *Assume the columns $\{Ue_j\}_{j\in S}$ are normalized and have mutual coherence $\mu := \max_{i\neq j\in S} |\langle Ue_i, Ue_j\rangle|$. Then Gershgorin's theorem implies $\delta_s \le (s-1)\mu$ for the submatrix U_S. In particular, if $(s-1)\mu \le 1/2$, then Theorem 5 applies.*

Proof. Let $G := U_S^\top U_S \in \mathbb{R}^{s\times s}$ be the Gram matrix. By normalization, $G_{ii} = \|Ue_i\|_2^2 = 1$ for all $i \in S$, and by definition of mutual coherence, $|G_{ij}| = |\langle Ue_i, Ue_j\rangle| \le \mu$ for $i \neq j$. Gershgorin's circle theorem therefore implies that every eigenvalue λ of G lies in the interval $[1-(s-1)\mu,\ 1+(s-1)\mu]$. Equivalently, for every $v \in \mathbb{R}^s$,

$$(1-(s-1)\mu)\|v\|_2^2 \le v^\top G v = \|U_S v\|_2^2 \le (1+(s-1)\mu)\|v\|_2^2.$$

This is exactly the s-RIP inequality (136) on the support S with $\delta_s \le (s-1)\mu$. If $(s-1)\mu \le 1/2$, then $\delta_s \le 1/2$ and Theorem 5 applies. □

Remark 14 (Interpretation for causal transfer matrices)**.** *In a linear SEM, the transfer matrix $U = (I - W^\top)^{-1}\Gamma$ maps intervention levels to interventional means, and the reward takes the form $\mu(u) = \langle \theta, Uu\rangle$. RIP/mutual incoherence are standard compressive-sensing conditions ensuring that different s-sparse interventions produce sufficiently distinct outcome directions (i.e., the map $u \mapsto Uu$ is a stable embedding on sparse vectors).*

Theorem 5 makes the reduction to classical linear bandits explicit: under RIP, the reachable set $\{Uu : u \in \mathcal{U}_S\}$ contains an approximately spherical s-dimensional ball, so the causal bandit contains an s-dimensional linear bandit as a special case. Thus, even when the causal transfer is as well-conditioned as one would like from a compressive-sensing viewpoint, the worst-case exploration cost is still $\Omega(s\sqrt{T})$.

Viewed through the lens of low-rank causal mediation, $\mathrm{rank}(P_{\mathcal{O}}U) \le r$ (Proposition 1) limits the number of independent outcome directions *that interventions can steer. The lower bound shows that whenever there exist s sparse intervention directions that survive this transfer without severe distortion (RIP/incoherence), bandit learnability is fundamentally governed by that intrinsic dimension.*